THE UNIVERSITY OF CALGARY

Language Inference from a Closed-Class Vocabulary

BY

Anthony Clive Smith

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

CALGARY, ALBERTA

March, 1993

ISBN 0-315-83252-5

Canada

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled, "Language Inference from a Closed-Class Vocabulary," submitted by Anthony Clive Smith in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor Ian H. Witten
Computer Science

Brian Gaines
Computer Science

Gary Libben
Linguistics
University of Alberta

Date APRIL 6th, 1993

Abstract

Language surface structures demonstrate regularities that make it possible to learn a capacity for producing an infinite number of well-formed expressions. This thesis outlines a system that uncovers and characterizes regularities through principled wholesale pattern analysis of copious amounts of machine-readable text. The system uses the notion of *closed-class* lexemes to divide the input into phrases, and from these phrases infers lexical and syntactic information. The set of closed-class lexemes is derived from the text, and then these lexemes are clustered into functional types. Next the open-class words are categorized according to how they tend to appear in phrases and then clustered into a smaller number of open-class types. Finally these types are used to infer, and generalize, grammar rules. Statistical criteria are employed for each of these inference operations. The result is a relatively compact grammar that is guaranteed to cover every sentence in the source text that was used to form it. Closed-class inferencing compares well with current linguistic theories of syntax and offers a wide range of potential applications.

Acknowledgements

My sincerest thanks and appreciation go first to my advisor Ian Witten, whose guidance and patience have made this thesis possible. His insight was often a bright light in very dark days. It was always nice to hear that rollercoasters go up sometimes.

I would also like to say thank you to Gary Libben for introducing me to the most fascinating world of language and the mind. His breadth of knowledge and keen insight in so many areas always provided much food for thought. It never seemed like much of a fair deal though—he kept his door open for me, and I erased all the files on his portable computer. Sorry about that, Gary!

Many thanks are due to John Lewis, William O'Grady, Bruce MacDonald, Eithnie Guilfoyle, Mark James, John Aldwinkle, Kevin Jameson, Michael Huston, Brian Gaines, and a host of others who lent books, advice, assistance, and an ear throughout the research and writing period.

Many thanks are due my faithful companion, Max, who sat at my side through many long days at the computer terminal at home.

Most of all, I would like to thank my wife, Linda. Little contributed more to the success of this project than did her patience, moral support, good food, and affection.

Contents

List of Tables

List of Figures

Chapter 1

An Introduction to Language Processing

Everybody uses language. We begin acquiring it very early in life, master it in a relatively short period of time, and rely on it almost unwittingly for the remainder of our years. We use it for communication, for expression and, in many instances, as an aid to our thought processes. Yet there is a peculiar contradiction about language that only becomes obvious when we try to develop theoretical and computational descriptions of its form and function. This paradox has been both spur and deterrent to linguistic research, and is perhaps best presented as a question—How is it that we seem to know so very little about something that we do so well?

The desire to build an effective computational system that can adequately process language has been a significant motivation in AI research—chiefly because of the central role that language is presumed to play in cognition [9]. Prescriptive efforts towards designing language processors with initially comprehensive accounts of what constitutes a natural language grammar have not demonstrated very effective levels of performance as a consequence of two inherent limitations: 1) an insufficient lexicon, and 2) an inadequate structural description for the language in question. An alternative approach, which will be pursued in this thesis, has been inspired by the observation that every child does eventually achieve a sufficient performance level with a natural language—without ever being fully cognizant of the grammatical rules that define it [43]. This method involves the passive inference of grammatical information from positive instances of well-formed expressions.

Natural languages have evolved culturally, and perhaps even biologically, as suitable systems for abstract communication. Language surface structures demonstrate regularities that make it easy to learn capacities for the production and comprehension of an infinite number of well-formed expressions. This thesis discusses the development of a system that uncovers and characterizes many of these regularities through wholesale principled pattern analysis of copious amounts of machine readable text. The system uses the simple notion of *closed class* lexemes to infer lexical and syntactic information. Closed class inferencing compares well with current linguistic and psycholinguistic theories of syntactic structure, and offers a wide range of practical applications.

1.1 Formalizing language

It is in some respects very difficult to understand why no one has yet been able to develop a complete characterization of a single natural language. What makes this apparent failure so surprising is that the principal task addressed by a grammarian can, to some degree, be likened to that confronted and accomplished by a child acquiring his or her first language: the discovery of a general characterization for a particular grammar. True, it is not fair to say that the grammarian's task is exactly the same for unlike the developing child he is not trying merely to acquire that characterization, but instead is faced with the challenge of finding a formal account of it.

There are many reasons for supposing that such a formalization is possible. First and foremost is the simple observation that languages demonstrate an extremely

high degree of regularity. The existence of such regularity is not simply a matter of perception, but must be a matter of fact. Consider the acquisition process for the child learning their first language. It is presumed that children learn their first language by listening to example utterances from their parents and isolating semantic and structural generalities [43, 11]. The presumption is strongly supported by two indisputable facts: after exposure to a finite number of well-formed sentences of a language, children demonstrate 1) a capacity to understand a potentially infinite number of novel sentences, and 2) a capacity to produce the same. It is difficult to explain the acquisition of such skills if language did not readily submit to generalization.

1.1.1 The fundamental elements of language

To develop a formal description of language it is first necessary to determine what qualities of language are generalizable. That is, what is it that a child learns when he learns language?

> It is somewhat misleading to speak of children acquiring a language. In fact what is acquired is a grammar—a set of rules and elements that allows people to speak and understand a language.
>
> Cho and O'Grady ([49], page 288)

Such an account of language is useful because it allows a language to be defined by its own properties without reference to its speaker/hearers. These properties are collectively known as a grammar, and if language acquisition is the procuring of a grammar then we must assume that such properties are demonstrated conspicuously enough for them to be learnable.

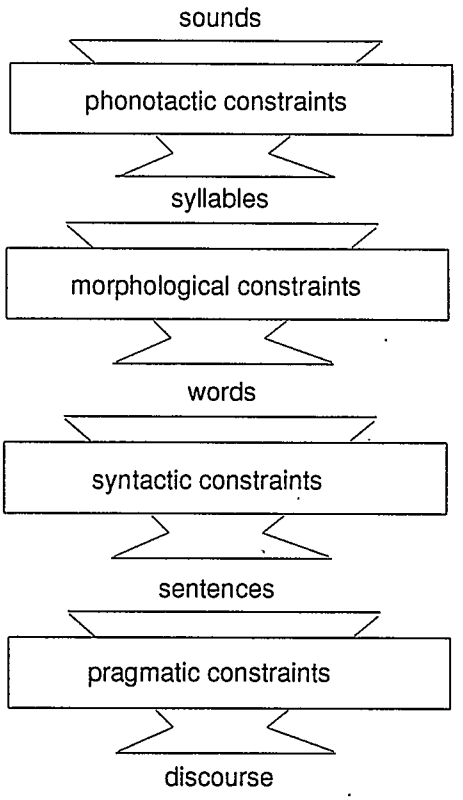

Figure 1.1: Hierarchy of language elements.

From the perspective of linguistics, a grammar is a collection of language elements at various levels of abstraction. As Figure 1.1 depicts, sounds go together to form syllables, syllables compose into words, words into phrases, and phrases into sentences—each level subject to its own grammaticality constraints. Language as a whole is perceived as a hierarchical composition of individual subsystems. From this we can view language acquisition as a multi-dimensional learning process, and infer that many levels of linguistic analysis are necessary for a comprehensive language learning system. We can reduce the complexity of the problem by reducing our definition of a grammar according to the static and dynamic characteristics of a language.

1.1.2 Dynamic language elements

Language is what people do, and in this respect we can say that language is, in a certain context, a property of people. The acoustic peculiarities of a language, including stress and intonation, vary from person to person, and from time to time, at the moment of production, and thus are part of the dynamic properties of language as a social action. Differences in dialect, accent, and even physical differences in an individual's speech organs can be so extreme that one might be able to construct an argument that no two people speak the same language. Even so, we do have some sort of notion that individuals with widely different accents or speech peculiarities are still speakers of the same language because we can circumvent any acoustic impasse by simply asking each speaker to write down whatever they want to communicate.

1.1.3 Static language elements

Particular productive differences need not be accounted for in a static definition of what constitutes a language. What allows us to recognize that individuals of differing accent and vernacular are speakers of the same language is, as illustrated by Figure 1.2, that their vocabularies and sentence structures are intersecting subsets of a common grammar. We may regard these grammatical elements as static not because they are unchanging, for that is clearly not the case, but because they can be expressed and analyzed without reference to a speaker/hearer. For example, insofar as a book contains language it does so without a record of the acoustic peculiarities of the writer. To a certain extent the letters used to compose the words are emblematic of the sounds used to produce them, but it is unnecessary to reproduce these sounds to recognize the language or divine the meanings embodied in the words and sentences of the text. Moreover, the capacity to understand spoken language rests on the ability to identify individual words and propositions.

Obviously our capacity to understand language will ultimately fall on our ability to identify the *meaning* of what is being said or written. We can, therefore, be sure that any attempt to acquire language without an account of its semantics will be fundamentally inadequate. But the understanding of language is, like productive peculiarities, what people bring to a language, and not a demonstrable property of the language itself. The question addressed by this thesis is how much can be learned about a language based strictly on an analysis of just its basic units of representation (i.e. the words) and its basic units of expression (i.e. the sentences)?

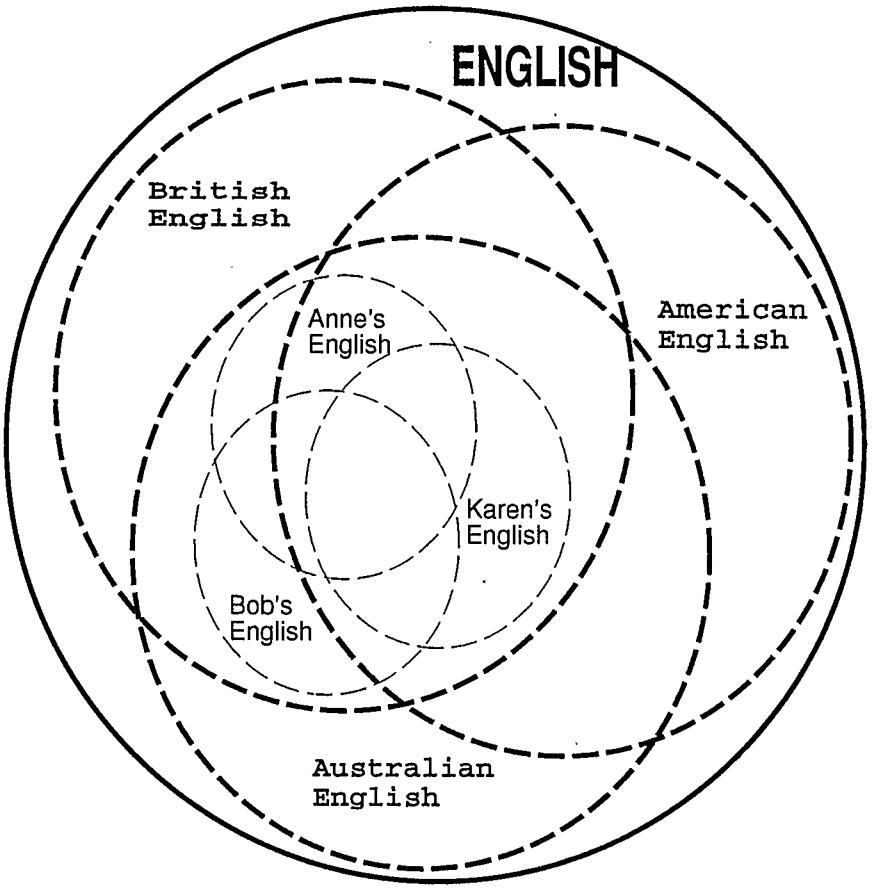

Figure 1.2: Intersecting subsets of vocabulary and syntax.

1.1.4 The basic units of representation

It is not necessary here to defend or debate the role of words as symbols of meaning, rather we need only concede that at a certain level of abstraction words constitute the fundamental building blocks of language. That is, the word is the smallest unit of meaning "that can occur in isolation and/or whose position with respect to neighbouring elements is not entirely fixed" ([49], page 128). For example, even though a word like "lamps" can be reduced to the smaller meaning bearing units (morphemes) of the root form "lamp" and the plural marker "s", the intensionality of the plural marker is lost if that morpheme occurs in isolation (e.g. "*lamp s"), or if its position with respect to the root is changed (e.g. "*slamp"). Thus words can be regarded as the smallest free form elements of a language.

Such a definition need not entail that words themselves be meaningful, but it seems clear that words are, by virtue of their use, suitable tokens to which meaning can be assigned. Isolation of words as the fundamental unit of meaning is a natural response to their psychological function. For the same reasons that led our species to delimit words with spaces when we developed our systems of writing, it is the intuitive familiarity of words that permits us to recognize them as the basic elements of language.

> Twice I have taught intelligent young Indians to write their own language according to the phonetic system which I employ. They were taught merely how to render accurately the sounds as such. Both had some difficulty in learning to break up a word into its constituent sounds, but none whatever in determining the words. This they both did with spontaneous and complete accuracy ... Such experiences with naive speakers and recorders do more to convince one of the definitely plastic unity of the word than any amount of purely theoretical argument.
>
> Edward Sapir

1.1.5 The basic units of expression

Words in isolation cannot convey all the meanings we wish to communicate via language. If the purpose of language is to express the infinite concreteness of our experiences, then clearly the finite vocabulary of even the richest language could not be sufficient for accomplishing that goal. Such an inherent limitation is counterbalanced by the infinite number of configurations into which the free form elements of a language can be arranged. For example, the sentence *The dog bit the postman* uses the same words as *The postman bit the dog*, allowing the same set of words to be used for conveying two separate ideas. In one sense, the meaning of each word is essentially unchanged in both sentences (e.g. *postman* still means "one who delivers mail", etc.), but the semantics of each sentence is (at least) the product of a meaningful arrangement of meaningful tokens.

Just as the word has both a logical and a psychological existence as the fundamental token of meaning, so too does the sentence have a similar existence as the major functional unit of speech. It is defined as the linguistic expression of a proposition, whose finished form is of a particular sentence type with fixed formal characteristics. Certain sentence types can be combined and overlaid with considerable freedom, allowing the opportunity for individual style, but the basic forms are "as rigidly given by tradition as are the words"[52].

We understand the essence of the linguistic proposition to be something like *who* did *what* to *whom*. We expect therefore to have to isolate and identify these basic constituents if we are to interpret a linguistic expression. They may be expressed with a one-to-one relationship between a single word and its referent. But there are

most often several words employed to properly express each component; and in some instances they may be expressed implicitly without any words at all. Further, there may be additional elements present to identify who did what to whom *where*, *at what time*, *with what*, and so on.

If each constituent of a proposition is expressed as a kind of word packet (where some packets may be empty) then it must be evident to the hearer which words belong in which packet, otherwise the content of the proposition may become ambiguous or nonsensical. Successful communication is accomplished by the speaker's adherence to a set of rules for sentence construction that will be familiar to the hearer. Knowledge of these rules is essential for both production and comprehension of natural language expressions and, therefore, is an indispensable component of any language processor—natural or synthetic.

1.2 Syntactic Description

The primary occupation of linguists is uncovering universal principles for natural languages. Such principles are, for the most part, generalizations of grammaticality judgments made by speaker/hearers of particular languages. "Roughly speaking, a sentence is considered grammatical if speakers judge it to be a possible sentence of their language" ([49], page 91). The goal of a syntactic theory is to establish a formal description of the regularities observed in sentential word order. Its success is often evaluated on the basis of 1) how many well-formed expressions of a language it can account for, and 2) how well it conforms to a psychological model of language production and comprehension. Though the latter is understandably dif-

ficult to assess, subtle indications—like the principle of compositionality described in Section 2.1.2—insist that psycholinguistic conformity be part of the evaluation criteria.

1.2.1 Prescriptive accounts

Constructing a theory of syntax is often accomplished in a kind of ad hoc fashion. Principles of syntactic structure are postulated to account for a few example utterances, and subsequent expressions are tested against the model as they are encountered [20, 58]. The theory is then tweaked and twisted to admit sentential structures not accounted for by the original description. Until recently, this process was conducted manually and, as a consequence, it often took a considerable amount of time before the weaknesses of a particular theory were made apparent.

With the advent of the general purpose computer, early language processors relied on manually constructed theories for their syntactic components [45, 61, 62, 21]. Though some degree of success was achieved from such processors, unparsable sentences were encountered so rapidly that the number of sentences rejected by the processor often exceeded the number it would accept. Unwilling to wait for the syntactician's remedy, the task of tweaking and twisting was quickly taken up by computer programmers who were willing to trade away psychological well-foundedness for high volume parsing.

Some language processors adopted techniques that side-stepped syntactic formalisms altogether. For example, Joseph Weizenbaum's Eliza [59] used simple keyword analysis and sentence inversion to simulate a Rogerian psychotherapist. By doing away with the notion of sentence rejection based on ungrammaticality, Eliza

could respond to any language input. It presumed that the intention of the speaker was to produce a well-formed expression, and thus viewed the input sentence as a communicative act rather than a procedural one.

1.2.2 Descriptive accounts

The idea that productive grammaticality arises from conditions of effectiveness rather than rule-based constraints is becoming more widely accepted by language theorists.

> An important change in generative theory has been the conceptual shift from viewing grammar as a set of rules, to viewing grammar as a set of well-formedness conditions.
>
> Steven Abney ([1], page 1)

The nature of these well-formedness conditions is beyond the scope of this thesis. Instead, we focus on a method for discovering descriptive characterizations of syntax under the assumption that well-formedness is demonstrated in sample utterances, and discernible through a principled wholesale pattern analysis of machine readable texts. Such a premise is congruent with theories of first language acquisition which view language learning as an inductive process from positive instances. Moreover, application of induction to complete texts allows the presentation expressions to represent positive instances of "communication" more closely than disembodied sentences do.

1.2.3 Inferring syntactic descriptions

Induction can be defined as the "process of reasoning to a conclusion about all the members of a class from examination of only a few members of the class; broadly,

reasoning from the particular to the general" (The Houghton Mifflin Canadian Dictionary of the English Language). In everyday terms, one can view induction as a kind of learning. For instance, a professional baseball player might claim that his ability to hit home runs was the result of his having learned how to hit tennis balls with a broom handle when he was a kid. That is, his ability to hit a baseball with a bat was the result of his having applied what he as a child had learned about hitting near-spherical objects with near-cylindrical ones—learned inductively through extrapolation from specific experiences with a tennis ball and broom handle. More than this, his ability to hold the broom handle the first time he tried was presumably the result of his having learned how to hold more general kinds of things prior to that—once again, inferred from specific experiences.

This kind of breakdown illustrates that induction is a process of developing progressively higher levels of generalization about observed instances so that they may be used to account for subsequent novel instances. Of course, the inferencing device must have some idea of what properties of the observed instances are to be generalized. In other words, inductive inference requires that something is known a priori about the goal of the reasoning process. To build an inference engine for acquiring linguistic information, therefore, requires an initial idea of what constitutes such information. Moreover, as with any inferencing device, a grammatical induction mechanism must, at the outset, have a procedure for forming generalizations about this information. It must have an induction algorithm. In principle, what one wants to infer is an arbitrary choice based on the level of abstraction desired; but the less initial information needed to infer the same final state of generalization, the more robust the induction algorithm.

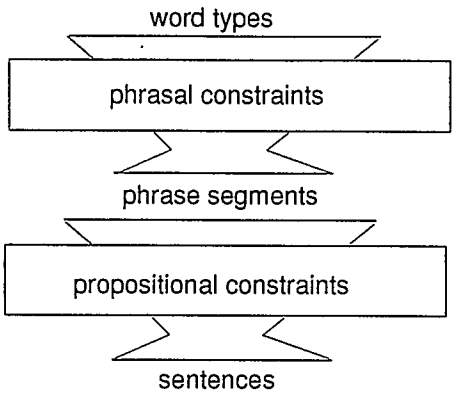

Figure 1.3: Sentence substructures.

1.2.4 The objective of syntactic inference

The first step towards developing a formal description of syntax is to recognize that the words of a language can be assigned to a small group of lexical categories according to certain properties they exhibit—some of which pertain to meaning, some to structure, and some to usage. For the syntactician, the lexical characteristic of greatest interest is the combinatorial property—the manner in which a word combines with other words to form larger linguistic units.

As Figure 1.3 shows, characterization of words according to their combinatorial property propagates into generalizations about larger linguistic units. Thus phrase segments are generalized according to the word types that compose them and how those word types are positioned with respect to each other, and characteristic sentence forms are similarly constructed according to their constituent phrase types. Therefore, the goal of a syntactic inference mechanism is the abstraction of general word, phrase, and sentence types based on their characteristic combinatorial properties.

1.2.5 The base case information

The induction of sentential structures from machine readable text has been attempted many times [8, 32, 50], and with varying degrees of success in terms of the first evaluation criterion: how many well-formed expressions are accounted for by the terminal characterization. But such attempts seem most often to have been motivated by the high level of interest that syntax induction presents as an inferencing problem in general. As a consequence, little (if any) consideration has been given to the second evaluation criterion: how well the characterization conforms to a psychological model of language production and comprehension.

Syntactic theory has attempted to reconcile structural principles to psychological accounts by focusing on the positions in which predicates and arguments occur. That is, insofar as a sentence is essentially a proposition, it seems reasonable to expect its underlying form to capture a fundamental relationship between a predicate and its arguments. This expectation lies at the heart of Generalized Phrase Structure grammars [29], Lexical-Functional grammars [36], and X-Bar theory [40]—all of which concentrate on the meaning-laden elements of linguistic expression.

The net result of this attention to predicates and arguments has been a fairly clear understanding of the structural positions in which their respective lexical categories appear. But the constraining principles governing the balance of categories have remained largely unknown. Recent theories of syntax have begun to correct this imbalance by focusing on lexical categories whose role in sentence structure appears to be primarily functional rather than semantic. For example, Abney's work on DP-Theory [1] allows some functional elements to head phrase structures, thus

offering an entirely new perspective on X-Bar theory. Subsequent developments of DP-Theory [55, 26] describe how functional elements regulate and contribute to the interpretation of their thematic complements. Furthermore, results from psycholinguistic research, like Garrett's positional model for sentence production [28], provide evidence that these so-called function words (and some inflectional morphemes) characterize sentence structures in the production process.

Function word categories differ from thematic categories in many ways, one of which is that functional elements constitute closed lexical classes. That is, while people show little hesitation in creating new nouns, verbs, and adjectives when a situation warrants it, they are reluctant to admit into their everyday vocabulary new determiners, prepositions, auxiliary verbs, or any other category of function word. The semantic content of function words tends to be more abstract, less picturable, and less ambiguous than that of nouns, verbs, adjectives, and so on. Further, function words demonstrate a significantly higher frequency of use in common parlance than that exhibited by their more semantically-laden counterparts. The fixed number of functions words, their high frequency, and the significant role they play in sentence structure, suggests they would be a reasonable choice as base case information for grammar induction.

1.3 Aim and Objectives

The aim of this thesis is to infer a set of function words from a statistical analysis of large machine readable texts, and discover the extent to which this closed-class vocabulary can be used to infer lexical and syntactic information. This can be broken

down into the following objectives.

1. To analyze concepts of syntactic structure offered from linguistic theory.
2. To analyze the concept of closed-class words as structural elements and psycholinguistic exceptions.
3. To identify the requirements of an inference mechanism, and review some algorithms that have been applied to grammar induction.
4. To infer a set of words which can be categorized as closed-class.
5. To group closed-class words into categories according to a statistical evaluation of their syntactic function.
6. To assimilate and categorize open-class words according to their syntactic relationship to closed-class categories.
7. To infer a general syntactic description for phrase structures demonstrated in a text, where each structure is characterized with a production rule headed by a closed-class category.
8. To analyze the degree to which production rules headed by closed-class categories conform to psychological and theoretical models of language structure.
9. To examine possible practical applications of the closed-class inferencing process.
10. To determine the limitations of the closed-class inferencing process.

1.4 Synopsis of the thesis

The aim and objectives of the thesis are met by the design and testing of a system that infers lexical and syntactic characterizations of texts based on the structural roles of closed class lexemes. The underlying motivation for the research is the exceptional role functional elements play in language surface structures, and the apparent special treatment they receive in language acquisition and cognitive processing—peculiarities that have been made evident through psycholinguistic research, and now form an integral part of current theories of syntax. Results from empirical testing of the system indicate that grammatical inferencing from closed class words compares well with other methods.

Chapter 2 meets the first three objectives of the thesis through an overview of syntactic theory and grammar induction. The basic principles of syntactic structure are described, along with the goals of language theory in general. Characteristics of closed-class words are also outlined to support their usefulness as base case information for grammatical inferencing. Finally, the principles and prerequisites of grammar induction are presented with some examples of existing systems.

Chapter 3 outlines the lexical acquisition component of the inferencing system. The statistical methods used to derive and categorize the closed-class elements are presented, along with the final set of function word categories used by subsequent induction procedures. The remainder of this chapter describes how the proximity and juxtaposition of functional categories are employed in the categorization of open-class words.

Chapter 4 describes the syntax induction algorithm adopted by the system.

Aimed at producing a context-free grammar for a given text, phrase boundaries are defined in terms of closed-class categories so that production rules can be derived that generalize sub-phrase, phrase and sentence structural characteristics.

Chapter 5 presents and compares the results obtained from applying the induction process to a variety of machine readable texts. Four evaluation criteria are used to assess the merit of the process results in terms of intuitive and formal expectations, and practical and esoteric applications and extensions.

Chapter 6 summarizes the thesis and reflects on the extent to which each objective is satisfied.

Chapter 2

Principles of Syntax and Syntax Induction

Inductive learning is essentially the process of forming generalizations about observed specific examples. What makes induction such a powerful learning tool is its ability to compress a collection of data into a set or sequence of generalizations to which certain properties or principles apply. This reduces the amount of knowledge that learners must possess for the recognition and comprehension of subsequent information by raising the level of abstraction at which they are attempting to understand.

First language acquisition is an inductive process. Upon hearing example utterances, a child will form semantic and structural generalizations that will allow him or her to understand and produce subsequent novel expressions. Over time, the child gradually develops an essentially "correct" grammar for that language.

Machine learning researchers have attempted to apply principles of inductive inference to develop a computational system for the purpose of acquiring correct language grammars [10]. Such systems require that the designer first 1) identify what the system is attempting to reason about, 2) establish the initial information required by the system for the reasoning process, and 3) develop an appropriate induction algorithm for the domain space. This chapter addresses these three components with respect to the grammatical inferencing system described in this thesis. The discussion includes an overview of syntactic theory, an analysis of the psychological and structural importance of function words in linguistic expression, and a review of some induction methods that have been applied to grammatical inferencing in

general.

2.1 Syntactic theory

The goal of linguistics is to construct explicit descriptions of particular languages from which a general theory of language structure can be developed [42]. These theories are expressed using a formal framework restrictive enough to permit nontrivial, falsifiable claims about what constitutes a grammar [29]. Establishing a suitable framework within which theories of sentential structure can be couched depends on a clear notion of the fundamental properties constituting such structure and the existence of a formal metalanguage capable of representing syntactic elements for a class of language grammars such that it is clear exactly what elements can and cannot be expressed with it [51].

2.1.1 Context-free grammars

In general, words are regarded as the basic elements of sentence structure (although some effort has been directed towards describing language syntax in terms of more basic morphological components [39]). The most transparent format for describing sentence structure with respect to words is achieved by equating correct structure with a specific sentence whose grammaticality is unchallenged. For example, if S is an intended generalization for correct English syntax, then

1a) $S \Rightarrow$the dog slept

states that the sequence of words *the dog slept* is well-formed. The notation is the Backus-Naur Form (BNF), a well-defined metalanguage used to express general

structure descriptions for the class of context-free grammars (CFGs). In isolation, 1a) restricts the set of well-formed expressions to the sequence *the dog slept*; but it can easily be extended by adding more expressions to the grammatical description. For example,

2a) $S \Rightarrow$the dog slept
2b) $S \Rightarrow$the canary slept

loosens the restriction of correct syntax to either of the word sequences explicitly defined by 2a) or 2b). In principle, we could trivially construct a formal BNF description of English syntax by including every sentence judged to be well-formed, if it were not for the fact that the number of such sentences is infinite. To overcome this obstacle, BNF descriptions can be generalized through the introduction of variables. Words are replaced by non-terminal symbols which allow the phrase structures to be characterized with so-called production rules. For example, 2a) and 2b) can be generalized to

3a) $S \Rightarrow$A B C
3b) $A \Rightarrow$the
3c) $B \Rightarrow$dog, canary
3d) $C \Rightarrow$slept

permitting S to capture the fundamental syntactic structure exemplified in the sample expressions without specific reference to either one. Further generalization is possible by replacing sequences of non-terminal symbols with more production rules. Thus, 3a)–3d) can be rewritten as

4a) $S \Rightarrow NP\ VP$
4b) $NP \Rightarrow D\ N$

4c) $D \Rightarrow$the
4d) $N \Rightarrow$dog, canary
4e) $VP \Rightarrow V$
4f) $V \Rightarrow$slept.

The capacity for English to produce an infinite number of grammatical sentences through processes like conjunction can be accounted for by introducing recursive production rules. For example, the addition of

4g) $VP \Rightarrow V\ C\ S$
4h) $C \Rightarrow$and

to the CFG description would admit to the grammaticality of sentences composed of conjunctive coordinate phrases. Similar such rules can be constructed to account for subordinate conjunction and dependent clauses.

It is worth noting that there is a great deal of controversy about whether or not natural languages belong to the class of CFGs and, thus, whether or not their syntactic structures can be depicted using BNF. Objections notwithstanding, BNF is still a widely adopted notation for describing phrase structures, and one would be hard pressed indeed to find any book on syntax that does not make use of it.

2.1.2 The principle of compositionality

Production rules are linear representations of tree structures which allow sentence forms to be depicted in terms of constituent elements. For example, Figure 2.1 shows the tree structure for 2a) as described by the rules 4a)–4f), wherein each non-terminal (i.e. each production rule) is graphically portrayed as a component sub-tree of the overall sentence structure. Trees allow a theory of syntactic structure to be evaluated in terms of the widely accepted principle of compositionality: that the meaning of an

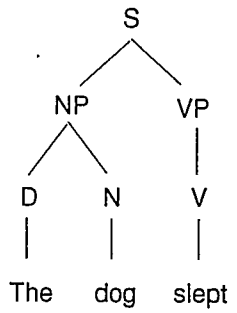

Figure 2.1: Grammar 4a)–4g) tree structure for expression 2a)

expression is a function of the meaning of its component parts [47]. This principle is adopted under what is often called the *rule-to-rule hypothesis* which insists on a direct association between every syntactic rule in a grammar and a corresponding semantic rule which determines the meaning of the specified syntactic component [4].

The principle of compositionality is an important constraint under which a general phrase structure description of a natural language is devised. Consider the grammar described by the following production rules:

5a) $S \Rightarrow D\ MP$
5b) $D \Rightarrow$the
5c) $MP \Rightarrow N\ V$
5d) $N \Rightarrow$dog, canary
5e) $V \Rightarrow$slept

These rules provide no less an accurate structural account of the expressions 2a) and 2b) than do the production rules 4a)–4f). In fact, there are an infinite number of CFGs that will describe these two sample expressions. General scientific canons such as maximizing simplicity and generality can help to reduce the number of descriptions which might be considered when attempting to settle on a final BNF grammar, but

the principle of compositionality requires that the grammar support a psychological explanation of linguistic structures. That is, syntactic components should correspond to semantic ones.

2.1.3 Constituency test

Section 1.1.5 stated that the basic semantic elements of a linguistic proposition were something like *who* did *what* to *whom*, with possible additional elements like *with what*, *where*, and *at what time*. The supposition that such semantic elements are mirrored in syntactic components tends to imply that genuine structural bonds exist between the words used to express given semantic elements. Evidence of these bonds can be discovered from the so-called *constituency test*—sometimes referred to as the substitution test.

Consider the phrase structure grammars

A:				**B:**			
	S	$\Rightarrow$	$SP\ OP$		S	$\Rightarrow$	$NP\ VP$
	SP	$\Rightarrow$	$NP\ V$		NP	$\Rightarrow$	$D\ N$
	OP	$\Rightarrow$	$NP\ P\ NP$		VP	$\Rightarrow$	$V\ NP\ PP$
	NP	$\Rightarrow$	$D\ N$		PP	$\Rightarrow$	$P\ NP$
	D	$\Rightarrow$	the, a		D	$\Rightarrow$	the, a
	N	$\Rightarrow$	girl, bicycle, park		N	$\Rightarrow$	girl, bicycle, park
	V	$\Rightarrow$	rode		V	$\Rightarrow$	rode
	P	$\Rightarrow$	in		P	$\Rightarrow$	in

Each of these grammars provides an account for the syntactic structure of the sentence *the girl rode a bicycle in the park*, and Figure 2.2 shows the two tree structures corresponding to each grammar. Both **A** and **B** are expressed with the same number

of rules and the same number of non-terminal and terminal symbols. Yet we can claim that **B** provides a more valid description because its non-terminal symbols reflect packets of words that have a demonstrable psychological bond. If we ask "What did the girl do?", we may answer that "*She* rode a bicycle in the park." Substitution of the word *she* for *the girl* indicates that the symbol NP in

$$S \Rightarrow NP\ VP$$

represents a single cohesive constituent of the proposition. Similarly, answering "Who rode a bicycle in the park?" with "The girl *did.*" indicates that the symbol VP in the same rule represents another cohesive constituent, and "The girl rode *it there*" demonstrates constituency for NP and PP within the rule

$$VP \Rightarrow V\ NP\ PP\ .$$

No such substitutions are possible for SP or OP in grammar **A**. Because its rules fail to reflect an apparently genuine binding relation within constituent elements of the linguistic expressions it describes, grammar **A** is rejected in accordance with the principle of compositionality.

2.1.4 X-Bar theory

The set of possible phrase structure grammars includes those with rules that do not occur in natural language [42]. Chomsky [18] incorporated X-bar theory as a set of constraints on the class of possible phrase structure grammars, and many of its principles have become fundamental to modern theories of language. A satisfactory exposition of X-bar is well beyond the scope of this thesis, but a brief summary is

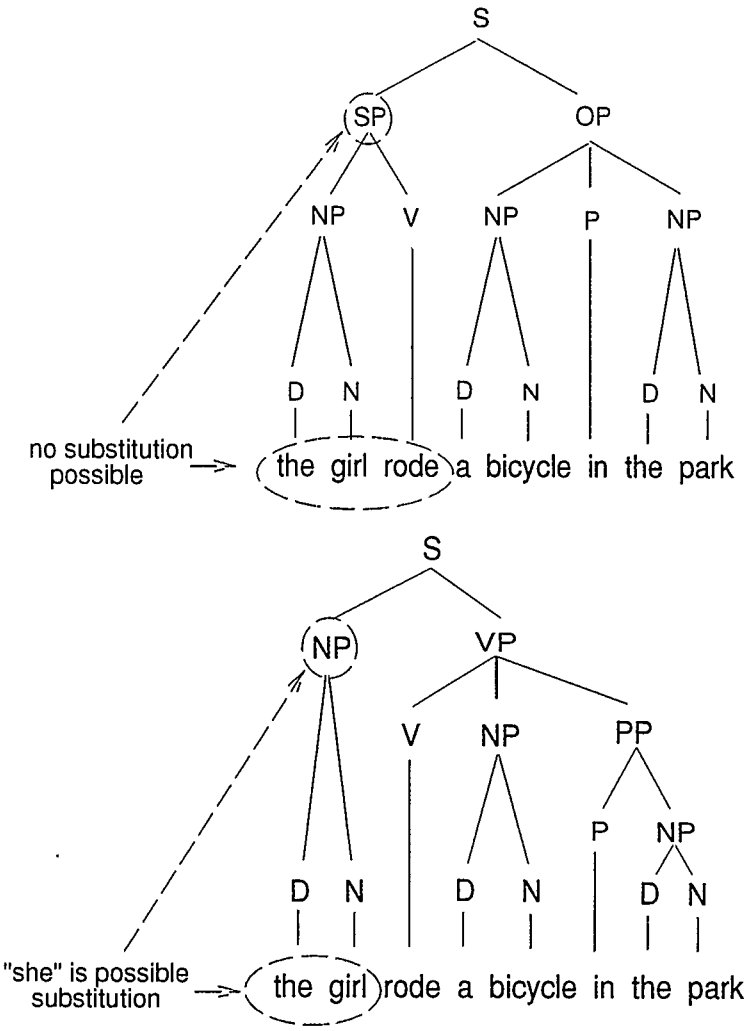

Figure 2.2: Constituents within tree structures

expedient for our purposes. More thorough discussions can be found in Chomsky [19], Stowell [56], and Farmer [23].

X-bar attempts to capture certain cross-categorial generalizations within natural language phrases. Consider the following simple production rules for verb phrases and prepositional phrases.

$$VP \Rightarrow V\ NP$$
$$PP \Rightarrow P\ NP\ .$$

These rules display a similarity that allows both to be generalized with a single description:

$$XP \Rightarrow X\ \text{COMP}\ .$$

Under this characterization, each phrase is comprised of a head element (the X) and a complement structure. That is, a verb phrase is comprised of a verb and a complement noun phrase, and a prepositional phrase is comprised of a preposition and a complement noun phrase. This uniformity has important implications for language acquisition processes in general.

> The complements of English phrases are highly systematic. All complement phrases can be expressed as XP ⇒X NP PP* (SS). If all NL grammars have this structure then what a child must learn is quite trivial [9].

In X-bar, all hierarchical substructures of linguistic expression can be expressed in terms of head and non-head nodes. Such cross-categorial generalizations allow theories about language acquisition to move "away from the view that the speaker of a language 'knows' specific PS [phrase structure] rules, making instead the claim that

the speaker of a language 'knows' the innate principles of [say] X-bar theory"([55], page 34). In other words, first language acquisition is more properly viewed as "making sense" of intrinsic language properties that serve to define or restrict general syntactic structures.

Formal versions of X-bar theory state that "underlying [syntactic] structures arise by means of projection from the lexicon"([55], page 35). In general, this has meant that conditions of well-formedness are not determined by phrase structure rules per se, but by head term constraints over strings of non-head terms [56]. First language acquisition is thereby reduced to the inference of head categories and the underlying principles of their projection over internal structures.

2.1.5 Head categories, projection, and rightward selection

In general, empirical studies of X-Bar determine headship according to characteristics of the semantic elements within a phrase—the meaning-laden words such as verbs and nouns [57, 38, 35, 44]. The implication to language acquisition is that the learner extrapolates syntactic categories and structures from the semantic features of these keyword categories.

> To learn a language, we must also learn its principles of sentence structure, and a linguist who is studying a language will generally be more interested in the structural principles than in the vocabulary per se. It is especially interesting, then, that in recent years linguistic research within quite diverse frameworks has been converging on the idea that sentence structure is to a large extent a reflection of the properties of lexical items.
>
> Margaret Speas ([55], page 1)

According to the so-called Head Feature Convention [30], each category of head is defined by its association with a designated set of feature values: boolean features

which decompose major parts of speech according to verbal or nominal attributes, or provide for distinction between such things as plurality/singularity; and qualitative features for identifying attributes like case, number, or verb paradigm (e.g. infinitive, passive, etc.). X-bar maintains that each phrasal element is marked by the feature values of the head itself—in effect, head and complement must carry the same feature values at any level of syntactic structure.

X-bar further asserts that complements are structurally constrained by, among other things, the *thematic* function of the head element—its semantic contribution to the propositional content of the expression. That is, the structural head of a phrase is also the semantic head—it is the lexical source of the descriptive content within the structure.

In English, the head is usually phrase-initial and the complement structure is subject to selection (generally rightward) by the thematic requirements of the head—a relation projected as a lexical property [38]. Unlike other structures, however, the head of a noun phrase (i.e. the noun) occurs in the final position, as in *the little brown fox.* The fact that determiners appear exclusively in noun phrases suggests that there is selection between the noun and determiner. But if selection in English is "generally rightward", one must assume that the determiner selects the noun.

The desire for uniform treatment of the nominal system within X-bar has contributed to the development of DP-Theory—a formalization of the view taken by Fukui and Speas [26] and argued for extensively by Abney [1, 2] that selection is functional.

2.1.6 DP-Theory

DP-Theory distinguishes between thematic assignment and functional selection through the notion of descriptive content. The noun *goat*, for example, has descriptive content in that it maintains a more or less unambiguous semantic relationship to a type of entity in the world. Similarly, the verb *hit* has descriptive content because it is clear what action it describes. In comparison, determiners like *the* or modals such as *will* lack descriptive content because they do not describe any "picturable" aspect of the world. DP-Theory maintains that since these words have no intrinsic descriptive content they must acquire it from a complement.

In DP-Theory, the determiner is the functional head of the noun phrase. "Its function is to specify the reference of the phrase. The noun provides a predicate, and the determiner picks out a particular member of that predicate's extension" ([1], page 3). In the verbal system, DP-Theory maintains that tense, or inflection, is the functional head of the verb phrase. Tense locates a particular event in time from the class of events predicated by the verb.

Determiners, auxiliary verbs, prepositions, pronouns, and other so-called minor lexical categories all serve as functional heads within DP-Theory. Phrases that do not have a free-standing functional element may be headed by an inflectional morpheme—like the possessive marker *'s* of prenominal genitives, or the tense marking suffixes *-ed* and *-ing* in participial constructions. In situations where no functional element is present at all, DP-Theory supports a null category that heads the phrase as a trace element similar to that adopted by transformational grammars [13].

2.1.7 Garrett: Positional level representation

The importance of functional elements within linguistic substructure is also supported by evidence from psycholinguistic research. For example, some slips-of-the-tongue in everyday speech reveal the possibility that functional elements are formed into syntactic structures prior to the insertion of any major lexical items—the nouns, verbs, adjectives, and other more "meaningful" lexemes. "Word exchanges", such as *he is planting the garden in the flowers*, and "stranding" errors, such as *he is schooling to go*, were among the corpus of thousands of naturally occurring speech errors that led Garrett [27] to develop a psychological model of sentence production in which functional elements establish sentence form.

Garrett's model describes a sentence planning process where the choice and location of free-standing function words and inflectional morphemes are determined apart from processes determining what semantic elements are to appear. The stages of sentence planning in the model are illustrated in Figure 2.3 (reproduced from [28], page 174). The message level representation consists of basic conceptual structures for the semantics of the expression being planned. At the functional level, lexical items are found which correspond to the conceptual structures, but without appropriate phonological information for their production. Thematic information—what is the action's agent, what is its instrument, etc.—is included at this level to establish how many arguments are associated with each verb in the expression, but information about the syntactic form, such as active or passive, is not included.

Basic word order is specified at the positional level. Here, the syntactic structures contain functional elements, but the precise phonological representations of semantic

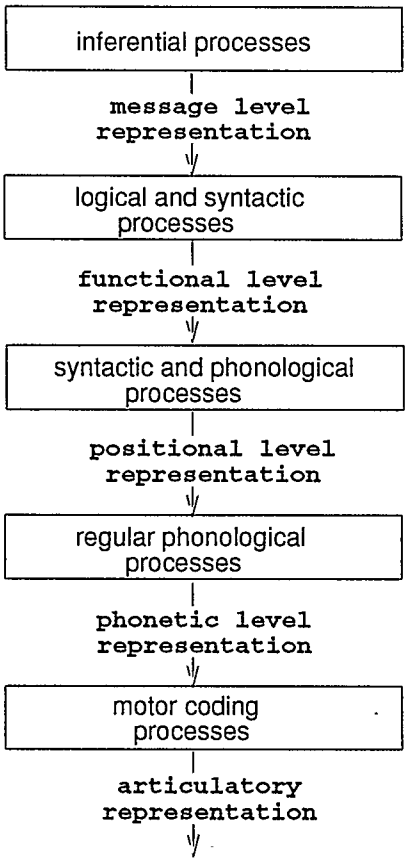

Figure 2.3: Garrett's model of the sentence production process.

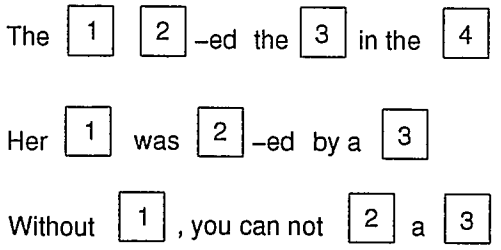

Figure 2.4: Possible functional element phrase structures.

items are not yet determined. It is at this point of the production process that Garrett proposed insertion errors could occur to produce the slips-of-the-tongue he was studying. In contrast with the traditional notion offered by *complement theory*, Garrett's model requires that functional elements be established within a syntactic "skeleton" structure with places left for semantic items to be inserted. Even though lexical semantic information determined at the functional level may establish some of the requirements for the eventual sentence form, the basic syntactic structures are constructed with functional elements.

Garrett's model attempts to describe cognitive processes that might be involved in sentence production, and thus provide an account of particular speech errors. That is, his focus is towards offering a psycholinguistic description of a system that generates language—not a generative description of the language itself. Extrapolation of a basic syntactic structure wherein functional elements define form follows as a consequence of the model. Figure 2.4 illustrates what some of these structures might be like.

2.2 Function words

DP-Theory and Garrett's model of sentence production are, in part, responses to a large body of evidence suggesting functional elements make a significant contribution towards establishing syntactic structure. The aim of this thesis is to discover the extent to which they can be used as the base case information for grammar induction. To avoid issues that arise from attempts at morphological decomposition, this thesis focuses on free-standing functional elements, sometimes called *function words.*

2.2.1 The structural role of function words

We can use standard grammatical categories to make an informal distinction between "content" words and "function" words. Content words consist of nouns, verbs, adjectives, most adverbs and perhaps some prepositions. Function words are exemplified by determiners, auxiliary verbs, pronouns, conjunctions, and a few other items. The class of function words is so named because of the particular role that function words play in syntactic structure [15]. They "function" to introduce certain grammatical forms like relative clauses (1a), verbal complements (1b) and questions (1c).

1a. The mouse *that* the cat chased.
1b. Alice knows *who* can bring some candles.
1c. *Why* did Ted have to leave this morning?

Function words may also serve to introduce such semantic notions as possession (2a) and tense (2b).

2a. The owner *of* this book must be a student.
2b. Melissa *will* arrive on Saturday.

2.2.2 The semantic role of function words

We may also distinguish between content and function words by their semantic contributions to linguistic expression. Content words convey meaning that is more or less concrete and picturable–that is, they are token referents to objects, actions, and attributes in the real world. In contrast, the semantic information conveyed by function words is generally more abstract and less referential than that of content words. In general, their principal role is more syntactic than semantic. Function words serve primarily to clarify relationships between their more meaning-laden counterparts.

2.2.3 Function word peculiarities

Function words demonstrate many distinctive properties. Unlike content words, they tend not to enter freely into word-formation processes. That is, they resist affixation and are seldom compounded with other words to form new ones—perhaps because of their inability to combine semantic notions with content words.

Function words tend not to carry stress in everyday speech. They do, however, work to place stress for more meaningful words and thus help to clarify semantic notions embedded within particular linguistic constructs [33]. When function words *are* stressed it generally reflects a deliberate effort to make these clarifications more conspicuous.

The set of function words constitutes a "closed class". That is, even though new words are added to the vocabulary of English nearly every day (e.g. fax, photocopy, workstation), the number of function words is fixed at about 500. This may have important consequences for psychological processing. Once an individual has learned the function words of a particular language, that person need not learn any more [15].

No one tries to invent a new determiner, or a new preposition—perhaps implying something about the sufficiency of the set of function words.

Function words are the last lexemes to show up in the demonstrative vocabulary of children learning to talk [43]. This may not be surprising given that there are no referents with which the child can establish semantic associations for the function words. But what may be surprising is that function words are also often the first items to be lost from the vocabulary of aphasics—individuals who have an impaired language facility due to brain damage. One particular class of aphasic symptoms, known as agrammatism, is characterized by difficulty in the production and comprehension of functional elements [5]. Patients suffering from agrammatism may experience a sense of familiarity with function words, but their inability to understand the contribution such words make in linguistic expression reduces sentences to something like newspaper headlines [41].

Finally, function words are generally the most frequently used vocabulary items—"... of the 100 most common words in the English language, nearly all are function words" ([15], page 267). It seems at least plausible that the high-frequency nature of function words is a direct consequence of their importance in language structure.

The characteristics of function words are summarized as follows. They are

1. resistant to word-formation processes.
2. unstressed—often functioning as clitics.
3. closed class.
4. late to appear productively during first language acquisition.
5. quick to disappear in agrammatism.

6. high frequency.

The combined effect of all these peculiarities is to suggest that function words may possibly be subject to different cognitive processes than are more meaning-laden words. Even further, they suggest that function words may have strong potential as base case information for grammatical inference.

2.3 Grammatical inference

Grammatical inference is the discovery of an acceptable grammar for a language based on a finite set of sample strings constructed from a finite alphabet. The problem is perhaps best stated by Chomsky:

> The strongest requirement that could be placed on the relation between a theory of linguistic structure and particular grammars is that the theory must provide a practical and mechanical method for actually constructing the grammar, given a corpus of utterances. Let us say that such a theory provides us with a discovery procedure for grammars [17].

Inferring natural language grammars is a principal task of linguists, but the notion of a grammar in linguistic research involves a description of (at least) the phonetics, syntax, semantics, and pragmatics of particular languages. In general, grammatical inference by automata is concerned with recognizing only the syntactic regularities of the sample strings, though other approaches have attempted to infer grammatical rules for mapping legal sentences onto corresponding meaning structures [9]. However, the formidable problems that accompanied syntactic analysis when semantic relationships had to be included led to the generally accepted notion that context free grammars were adequate enough models for generalizations or hypotheses about

natural languages [43].

2.3.1 Identification in the limit

The inference process requires that the inference device be presented with an increasingly larger corpus of example strings from the grammar being inferred. At each presentation, the device must simultaneously make guesses of the underlying rule being exemplified. By recognizing regularities within the sample set, the inference mechanism must be able to form a generalization of the data that will permit the prediction of future data. In a keystone paper on grammatical inference [31], Gold identified three possible results that can be expected from this approach, which can be formulated into a rough criterion for success:

1. The hypothesis will converge to a single description that correctly identifies the grammar—in which case the inference is correct.
2. The hypothesis will oscillate indefinitely—which implies that the inference has failed.
3. The hypothesis will converge to an incorrect description of the grammar—in which case the inference is incorrect.

The important aspect of this realization is the idea of convergence. That is, assuming the induction is tending towards a type 1 hypothesis, the inferencing mechanism will move ever closer to a correct description of the grammar—the point in time when the grammar may be considered as having been *identified in the limit.* Unfortunately, such convergence also implies that at no point in the inductive process can the device assert that the grammar it has inferred is correct, since further evidence may prove it incorrect. In practice, the mechanism can only detect a point in time

when its hypothesis has not been changed for a significant number of consecutive sample strings, at which point it may be deemed to have reached a sufficiently high probability of correctness.

2.3.2 Formulating the inference algorithm

> One may arrive at a grammar by intuition, guess-work, all sorts of partial methodological hints, reliance on past experiences, etc.. It is no doubt useful to give an organized account of many useful procedures of analysis, but it is questionable whether these can be formulated rigourously, exhaustively, and simply enough to qualify as a practical and mechanical discovery algorithm [17].

Since an induction mechanism can never know when its target grammar has been correctly inferred, it is clear that the success of the inference cannot be measured by the resulting hypothesis. Rather, it is determined by the criteria used to construct the inference algorithm. Feldman [24] stated that all attempts to formalize grammatical inference must include precise formulations of the following principal components:

- the hypothesis space: the general set of rewrite rules that are to be considered—i.e. the candidate classes of grammars.
- a measure of adequacy: the hypothesis must account for all sample strings plus all other strings generated by the grammar.
- the presentation criteria: in what order will examples be presented—n.b. each string must eventually be seen.
- a criterion for success: when has the limit been reached?
- minimum complexity requirement: how tightly should the inferred grammar fit the target language.

Each formulation must be established in terms of the overall requirements of the inference mechanism: 1) the device must be able to infer a correct grammar for any language generated by a grammar in the class being considered, 2) the device must infer the correct grammar for all sequences and stop in finite time with the correct answer, 3) the device must yield the best grammar for each language, and 4) the device must make a *best guess* after each presentation.

Clearly, Feldman's notions of "correct answer" and "best guess" are the source of difficulty in assessing the adequacy of an algorithm, particularly in light of the uncertainty involved with attempting to identify a language in the limit. Generally speaking, tractability is the principal concern for developing an algorithm, and the lowest possible complexity for generating the presentation is often used to assess correctness [25].

2.3.3 Enumerative methods

The most effective technique for generating the hypothesis space is the enumeration of all possible grammars in the class of interest. Enumeration has the advantage that it can often be shown to arrive at an optimum answer with a minimum amount of data. Gold has pointed out that no other algorithm can uniformly reach a correct answer in less time for all grammars in the class, and for all presentations. In practice, however, the combinatorics associated with enumerative techniques are computationally prohibitive [31, 3, 6].

The amount of enumeration can be substantially reduced by applying certain restrictions. Equivalent grammars that may be generated can be eliminated by appropriate search techniques, and the search can be constrained to simple grammars (e.g.

regular grammars). If the order of presentation is established with some discretion, then the rejection of grammars which are not compatible with previously observed examples can greatly reduce the search space. That is, the induction method should avoid inferring any hypothesis that can be deductively falsified by previous observations. Thus, entire classes of grammars can be deleted on the basis of a test on just one of them [12].

Horning [37] developed an enumerative inference program that successfully employed theses techniques by organizing a class of grammars into a tree based on the number of non-terminals within each grammar. A grammar with n non-terminals would represent a node in the tree, and each grammar with $n+1$ non-terminals that is covered by the first grammar is linked to it. The search process begins at the top of the tree and moves downward creating grammars at each node, after determining its compatibility with the observed data. If any grammar fails to produce the entire presentation set then it is eliminated—as are all its descendants.

2.3.4 Oracles and teachers

Pruning can also be applied to an enumerative search space if the presentation permits both positive and negative examples to be considered by the inference mechanism.

Pao [50] developed a finite search algorithm which used the grammar covering concept as a pruning technique, but also employed a "teacher" as a guide for the inference process. She constructs the simplest finite state machine that will accept the set of sample strings by gradually synthesizing states of other machines. At each presentation, the inference mechanism constructs a finite state machine that

will accept the current sample string, and its states are then merged with those of previously created machines. All machines that are created are ordered in a finite lattice so that each machine at one level of the lattice is connected to each machine at a higher level. In this way, the grammar associated with a particular machine at one node covers the grammar of machines connected to it at lower levels. Two connected machines are selected and a string is found that is accepted by one machine, but not the other. The "teacher" is asked if the string being considered is in the language. If the answer is yes, then the machine that rejects the string, and all machines below it in the lattice, are eliminated from the search space. If the answer is no, then the other machine is rejected along with all connected higher machines. The inference process terminates only when one machine (or several equivalent machines) are left, and the grammar associated with it is proffered as the solution.

2.3.5 Constructive methods

Alternative to enumerative methods are the constructive methods. This kind of approach was first formulated by Solomonoff [54] and others early in the history of grammar induction. It involves the construction of rules that account for regularities detected in the sample strings, rather than a systematic search through all candidate grammars. Solomonoff claimed that most of the interesting structure of a grammar involves the cycles in its language and the conditions under which they occur. His claim was that the detection of the *basic cycle form* was the end of the inference process.

Solomonoff's method required substituting repeating patterns within the sample strings with super-symbols. In this way, the language was characterized by sets of

combinators and incidental residue (i.e. leftovers). Pattern analysis was performed recursively on the resulting characterizations producing larger and larger combinators (e.g. non-terminals for strings of length 2, then 3, then 4, etc.). When the inference process failed to produce any new non-terminal combinators for an arbitrary number of consecutive presentations, then the resulting set of rewrite rules was deemed a correct grammar.

Berwick and Pilato [9] formalized another constructive approach that was designed for the inference of *k-reversible* languages. A *k-reversible* language is one where whenever two *prefixes* of expressions from a language *whose last k words match* (where k is any non-negative integer) also have a tail in common, then the two prefixes have all tails in common. For example, given a language that has *Mary bakes pies* and *John bakes pies* as acceptable sentences, if it also has *Mary bakes cakes* then it must also have *John bakes cakes.* Thus a new string is inferred for the language in order to maintain (in this case) its *1-reversibility.*

The inference is attained by generating a prefix-tree automaton that accepts only those strings provided in the sample. The tree is an acyclic directed graph with a single root. Every node (except the root) contains a word from the sample strings, and every sentence can be constructed by tracing a path from the root node to a terminal node. To generalize the grammar in such a way that new strings can be added while maintaining the grammar's *k-reversibility*, equivalent states of the prefix tree are collapsed together, working backwards from the terminal nodes. If the resulting graph produces a k-length segment that can be accessed from different nodes, *and that ends in a terminal node*, then any access to the start of that segment infers a new string for the grammar.

Constructive approaches are attractive for many reasons. The most prominent advantage that they offer is an improvement of the tractability for the inference algorithm, due to the fact that they need only create characterizations of the grammar to account for the examples seen up to a given point. They do not try to construct every possible set of rewrite rules according to some enumerating generative procedure. This removes the combinatoric explosion that happens with the systematic (and consequently more thorough) methods of testing prospective grammars. An important drawback to constructive methods, however, is that they cannot be subjected to any evaluation as "best" accounts of a grammar because they are usually only expected to produce one set of production rules. Thus, they often overlook more simple characterizations of the language.

2.3.6 Identifying successful inferences

Choosing a criterion for success is perhaps the most difficult aspect of formulating an algorithm for grammatical inference. Certainly we can confidently claim that an inference machine M identifies a grammar G if it eventually guesses only one grammar—the grammar that generates exactly $L(G)$, the language of G. But, because the machine cannot effectively prove that a grammar generates exactly $L(G)$, it cannot choose one grammar and then cease consideration of new data. A weaker form of learning is required in which for each presentation y at time t, the guesses A_t would be increasingly closer approximations of the grammar G. That is, a metric is required for evaluating convergence.

Feldman [24] defines a notion of *approachability* for evaluating convergence that says a machine M approaches the grammar G if:

a) For any $y \in L(G)$ there is a time τ such that $t \succ \tau$ implies $y \in L(G)$.

b) For any H such that $L(H) - L(G) = \phi$ there is a time τ such that $t \succ \tau$ implies $A_t \neq H$.

In other words, any expression y that can be generated by a grammar will always be generatable by that grammar (i.e. for any time t after time τ), and any hypothesized grammar that generates all and only the expressions generatable by the target grammar *at some point in time* will always generate all and only those expressions.

Both a) and b) require that conditions hold for an infinite number of $t \succ \tau$. In practice, however, these conditions can only be evaluated in terms of the number of successive presentations where $L(A_t) - L(A_{t-1}) = \phi$. This would allow the inference device to decide it has probably guessed A for the following condition:

c) There is an A such that $L(A) = L(G)$ and for an infinite number of t, $A_t = A$.

2.3.7 The spectrum of correct grammars

The decision to assume that $L(A) = L(G)$ is, as Gold has pointed out, a necessarily arbitrary decision. It is also deceiving to hope that conditions a) and b) are sufficient to assign any sort of probability to the correctness of the decision to terminate the inference process. There are possibly an infinite number of grammars that can be inferred that will contain the language of the grammar being sought and so be deemed correct grammars. The set of correct grammars can be seen as a spectrum, with the grammar generating the largest language that contains $L(G)$ on one end, and the grammar that generates all and only the presentation set on the other end. We can be sure that the grammar inferred falls somewhere on this spectrum, as does the

target grammar; and thus the final choice is arbitrary according to the tightness of the fit desired [12]. The degree of arbitrariness can be tempered if we apply some sort of complexity metric to the inferred grammar, or if we select the grammar resulting from a constructive algorithm that converges on a correct grammar at the fastest rate. Any final evaluation will ultimately rest on a probablistic model.

Chapter 3

Lexical Acquisition

3.1 Categorizing words

Human beings are chronic organizers—organizers in the sense that we have an instinctive desire to sort things into groups according to their particular characteristics. Provide a young child with a variety of toy blocks and before long they will be sorted into small groups—sometimes according to size, sometimes by colour or shape, and sometimes according to properties which escape the unenlightened onlooker.

The propensity to sort is not only instinctive for human beings, it is also essential for the process of language acquisition. That we are able to take the sentence "the cow jumped over the moon" and substitute *goat* for *cow*, without affecting the grammaticality of the sentence, tells us that we acknowledge a characteristic similarity between the two words that enables them to fulfill precisely the same syntactic role. Further, if Lewis Carroll or e. e. cummings had written "the tweedlebom jumped over the moon" we would have no difficulty at all accepting the idea that *tweedlebom* enjoys this same intrinsic quality, despite our inability to say exactly what a tweedlebom is (or are).

3.1.1 Identifying lexical characteristics

Recognition of characteristic similarity between words allows us to sort them into categories and generalize about their grammatical roles—an essential process for

language acquisition.

> The first step in syntactic analysis is the identification of the categories to which the words of a language belong. If words could not be assigned to a small group of categories, it would be very hard to learn or use a language. Each of the ten thousand or so lexical items in the average person's everyday spoken vocabulary would have its own set of properties that would have to be memorized—a rather daunting task.
>
> (O'Grady, 1987, page 92)

But how do we determine these categories? What *salient properties* do the words "cow" and "goat" share? We could try some standard epistemological explanation like "both *cow* and *goat* are token references to substantive entities—things, or types of things—and that while a *tweedlebom* is not substantive in the physical sense, it does, like a *unicorn*, have its substance in a kind of Platonic form."

A peculiar difficulty with semantic interpretation arises in languages that allow a word to be used in several syntactic roles. A word like *hit*, for example, which can be used in its more conspicuous verb form in sentences like "Babe Ruth hit a home run", or in its more colloquial noun form in sentences like "Babe Ruth got a hit". What this means is that the syntactic capacity of a word is not determined directly by the word itself. Rather, the syntactic role is assigned to the word according to its context and associated expectancy. Consider the following stanza from Lewis Carroll's *Jabberwocky*:

> Twas brillig and the slithy toves
> did gyre and gimble in the wabe.
> All mimsy were the borogoves,
> and the mome raths outgrabe.

Even though many of the words that Carroll uses are nonsense, we recognize that "toves" are without question *things*. More specifically, they are things that have the

property of *slithy-ness*. We have no mechanism for conjuring up images of a "tove", nor can we look them up in a dictionary, yet we feel instinctively that they have at least the same sort of referent existence as does a Platonic unicorn. We know this not from a semantic interpretation, but from a syntactic one. Leaving aside the morphological information indicating plurality, we know that "toves" is a noun and we know this because it is used as a noun with a calculated expectancy according to our knowledge of the syntax of poetry. This expectancy is not just used for nonsense. We can use similar knowledge to help us interpret an unfamiliar word that we come across whilst reading.

If we are to persist in a lexical acquisition procedure that refuses any semantic assistance then we must rely on more static word traits. More specifically, we must focus on the characteristics of word usage in a structural context.

3.1.2 Finding categories

Any process of word categorization that begins with a fixed number of lexical categories to which each word of the language will be assigned too readily sacrifices the criterion for a robust induction algorithm described in Section 1.2.3—prescribe as little as possible. Ideally we would like to discover how many categories are demonstrated syntactically as legitimate structural roles. We may assume that there are at least two categories or the premise of syntactic analysis is metaphysically unsubstantiated. Therefore, a sound method of inductive categorization might first separate the entire vocabulary of a language into two groups by some coarse method of differentiation, and then continue to filter each of these groups recursively into smaller and smaller categories according to a distinction criterion that becomes increasingly

more discriminate. The soundness of the process hinges on the soundness of the separation criterion, which should reflect a balanced measure of arbitrariness and reasonability. Fortunately, there exists a method of distinction that has this balance; one that can also be applied with increasing particularity. It is a distinction of functionality, and provides an initial division of the vocabulary into two important categories: closed-class and open-class words.

3.2 Closed-class words

Creation and assimilation of new words are ongoing activities for any modern language. Borrowing terms from another language, blending words and parts of words into new ones, and creating them from scratch constitute principal processes of language evolution. But for most languages, there is a group of word types to which new terms are neither added nor created—a so-called closed class. Its reluctance towards accepting new words stems from the apparent sufficiency of its present membership. Words from this class serve to clarify relationships between the meaning bearing words of an expression—that is, their function is more syntactic than semantic. In fact, within linguistic literature they are often referred to as *function words*.

The existence of a collection of words that can be characterized as closed-class is not a point of contention for most linguistic researchers. But there is no consensus on exactly which words should be deemed to belong to that collection. As a consequence, the kind of class description usually found in linguistic literature is most often vague enough to allow the definition to pass uncontested. A description like *determiners, pronouns, auxiliary verbs, conjunctions, and some prepositions* is commonly proffered

so that the reader is able to cultivate a general understanding of what the function word essence is like, without becoming bogged down with issues of precisely which words possess it. However, computational theories that depend on the notion of a function word for constructing systems that will actually process language must (for the sake of tractability) include an explicit list of words that share in that essence, even if the definitional constraints held by the theorist have to be relaxed.

3.2.1 Identifying closed-class words

Clearly one of the first tasks that must be faced before implementing such a processor is to assemble the list of words which satisfy the closed-class criterion. There are two prescriptive methods of doing this. The first is to look systematically at each word of the language and decide whether or not its intrinsic properties merit its inclusion in the closed-class category. Apart from the obvious tedium involved with this approach, its principal shortcoming is the degree of arbitrariness that accompanies the decision of whether to accept or reject any given word. The same lack of precision that compels authors of linguistics texts to avoid precise definitions of the function word would necessarily be present in a formal account of this method.

Conversely, the second method requires that the vague definitions of the literature be hardened into firm boundaries, fixing the criterion to include all words that traditionally belong to certain standard grammatical categories—for instance, all determiners, conjunctions, and prepositions. This approach is very attractive from a computational standpoint since it allows the set of closed-class words to be quickly gleaned from any computer readable dictionary. Its failing is that it presumes the complete subsumption of those standard grammatical forms into the more general

closed-class category—a point of contention amongst linguists.

3.2.2 Discovery criterion

Insofar as inferencing is a discovery process, the criterion used for determining membership in the closed class category should be one that allows function words to be discovered according to their usage. Moreover, the criterion should be established in accordance with their definition. However, as can be seen from the following review of the function word definition, not all function word properties can be evaluated in a static analysis of language.

- low semanticity;
- low stress—frequently phonological clitics;
- late production appearance during first language acquisition;
- marked production, comprehension, and recognition difficulty in agrammatism;
- closed class;
- high frequency.

The psychological aspects of low semanticity, late production appearance, and aphasic usage difficulty are properties which cannot be regarded as salient without a cognitive account of language. The stress and intonation properties are part of the production peculiarities and, consequently, not a demonstrable characteristic of static language elements. Obviously, since the closed class attribute is the property we are attempting to determine, we are left only with the high frequency nature of function words as a discovery criterion.

Far From The Madding Crowd		Moby Dick	
word	occurrences	word	occurrences
the	7746	the	13982
and	4285	of	6427
a	3911	and	6263
of	3782	a	4597
to	3591	to	4517
in	2349	in	4041
I	2123	that	2915
was	1970	his	2481
it	1566	it	2374
that	1534	I	1993
you	1468	but	1796
her	1465	he	1751
he	1391	as	1712
she	1266	with	1681
as	1191	is	1676
had	1157	was	1602
his	1145	for	1586
for	989	all	1510
with	969	this	1375
at	948	at	1297
vocabulary	11589	vocabulary	16832

Table 3.1: Most frequent words in *Far From the Madding Crowd* and *Moby Dick*

3.2.3 High frequency metric

If we are to identify function words according to their frequency then we must first decide how often a word need occur for it to be considered a functional element. That is, what ratio of one word's occurrence with respect to the occurrence of all words constitutes a high frequency? The degree of statistical frequency necessary for a word to merit inclusion in the closed class category is ultimately determined arbitrarily. However, we can draw on the literature to develop a rough guideline for what constitutes a statistical significance.

> There are approximately 500 or so function words in English, and, of the 100 most common words in English, most are function words.
>
> (Caplan, 1987, page 267)

number of words	vocabulary items represented	fraction of vocabulary	total usage	fraction of text
1	{the}	0.01%	7,746	5.5%
2	{and, the}	0.02%	12,031	8.5%
3	{a, and, the}	0.03%	15,942	11.3%
5	{a, and, of, the, to}	0.04%	23,315	16.6%
10	{a, and, I, in, it, ...}	0.09%	32,857	23.4%
15	{a, and, as, I, in, ...}	0.13%	39,638	28.2%
115	{a, about, again, all, am, ...}	0.99%	75,688	53.8%
11589	{aaron, abandon, abasement, ...}	100.00%	140,632	100.0%

Table 3.2: Vocabulary distribution in *Far From the Madding Crowd*

The average person's everyday vocabulary consists of about 10,000 words. If, as Caplan claims, most of the 100 most common English words are function words then the the top 1% of most frequently used words from a typical vocabulary is a reasonable first approximation to the closed class—assuming that the functional importance of the other 400 words diminishes proportionately with their declining frequency.

Table 3.1 provides partial lists of the most common words from the vocabularies employed by Thomas Hardy in *Far From the Madding Crowd* and Herman Melville in *Moby Dick*. A cursory analysis reveals that words used with the highest frequencies fit well with our intuitive notion of the function word.

3.2.4 The everyday vocabulary

The principal practical obstacle to the 1% cutoff is how to ascertain an "average person's everyday vocabulary." One might object to the suggestion that Hardy or Melville exemplify common parlance—despite the fact that their demonstrated vocabularies are of an appropriate size. But we assume that closed-class elements are

a	being	Gabriel	into	night	say	they	well
about	Boldwood	go	is	no	see	think	went
again	but	good	it	not	she	this	were
all	by	had	its	now	should	time	what
am	came	have	know	Oak	so	to	when
an	can	he	Liddy	of	some	too	which
and	come	her	like	on	such	Troy	who
any	could	here	little	one	than	two	will
are	did	him	man	only	that	up	with
as	do	his	me	or	the	upon	woman
at	don't	how	more	other	their	very	would
Bathsheba	face	I	much	out	them	was	yes
be	down	if	my	must	then	way	you
been	for	in	never	over	there	we	your
before	from			said			

Table 3.3: The top 1% most frequent words in *Far From the Madding Crowd*

functionally significant for the language itself, and will therefore be statistically dominant in any individual's vernacular, including Hardy's or Melville's. For example, Table 3.2 shows that the top 1% of Hardy's vocabulary accounts for almost 54% of the text in *Far From The Madding Crowd.* These 115 words are listed in Table 3.3 and only about 16 of them fail any sort of intuitive test as function words.

Table 3.1 shows tremendous commonality between the most frequently used words of Hardy and Melville. Sixteen of the top twenty are the same, the first six differing only in their order. This similarity proceeds beyond the words listed here. But there are also some significant discrepancies. For instance, there are no feminine pronouns in the 80 most frequently used words of *Moby Dick*, with *she* appearing in the relatively distant 217th position, though it is the 14th most common word in Hardy's novel. Moreover, *whale* is the 28th most common word in *Moby Dick* yet it never occurs in *Far From the Madding Crowd*; similarly *Bathsheba*, Hardy's 38th

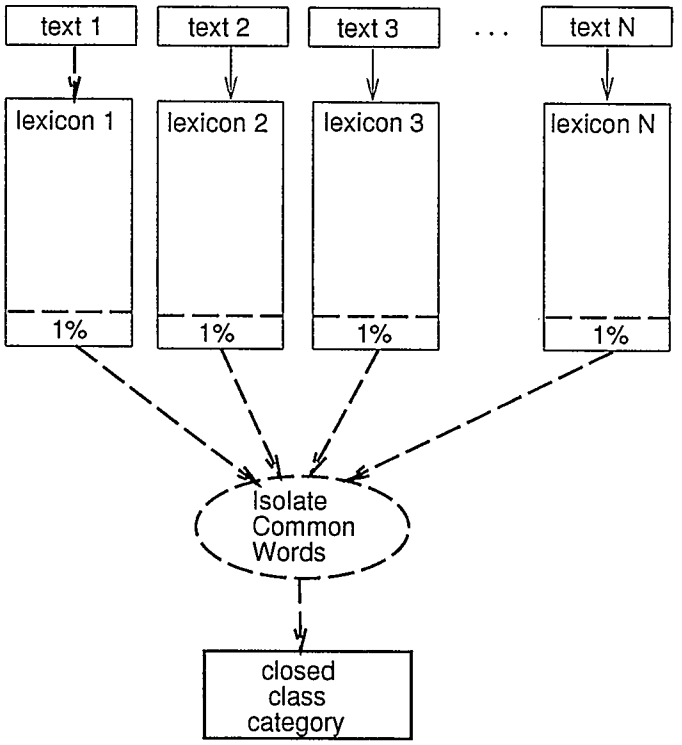

Figure 3.1: Procedure for determining closed class category.

most frequently used word, does not appear in Melville's book.

3.2.5 Removing lexical peculiarities

Of course, neither *Bathsheba* nor *whale* conforms with our intuitive notion of a function word and should be removed from the class, whereas it would be unfortunate if feminine pronouns were overlooked. Therefore neither vocabulary is entirely suited to be the paradigm. But we can capitalize on their similarities by intersecting the vocabularies before taking the top 1%. Lexical items peculiar to any one text are discarded and, as a consequence, function words that may otherwise be overlooked are moved higher up in the frequency ordering. This process, outlined in Figure 3.1,

a	for	it	or	to
about	from	its	out	up
all	had	like	said	very
an	have	me	so	was
and	he	more	some	we
are	her	my	that	were
as	him	no	the	what
at	his	not	them	when
be	I	now	then	which
but	if	of	there	who
by	in	on	they	with
do	into	one	this	would
down	is	only	time	you

Table 3.4: The closed class, inferred from Hardy, Melville and Carroll

will, in principle, isolate a more reliable closed class category as more vocabularies are considered, though additional vocabularies must be adequately large. Table 3.4 lists the function words obtained by applying this method to the vocabularies of Hardy, Melville, and that employed in a collection of works by Lewis Carroll (*Alice in Wonderland*, *Alice Through the Looking Glass*, and *The Hunting of the Snark*). Unfortunately "she" still does not appear in the list, though "he" and "her" do. Nevertheless, the final intersection set contains those words statistically significant in the vernacular of all three authors and, to that extent, presumably reflects a class of words functionally significant for English in general.

3.2.6 Categorizing function words

We have assumed that the relative high frequency of words ostensibly low in semanticity implies that their structural roles are functionally significant. It follows that each closed-class lexeme is either used to perform a specific and unique functional role, or is representative of one of a number of functional categories.

There are many reasons to prefer the second conclusion, even though the first permits stronger inferences. Perhaps the most compelling evidence is the intuitive notion of the functional role performed by what is called the determiner. We recognize a certain functional similarity between the words "a" and "the". In general terms, "the" is a kind of existential quantifier indicating a specific referent, whereas "a" works as a kind of universal quantifier indicating a representative of a general class of referent. Moreover, determiners like "his", "some", "many", and "all" permit reference at greater and lesser degrees of specificity.

It seems that closed-class words fall into functional categories. This is attractive because it greatly reduces the number of syntactic roles in a language. However, in keeping with a static analysis, we seek to achieve such generalization without relying on semantic or psychological properties. Once again, frequency analysis provides a solution.

The frequency-based method for discovering closed-class words can be regarded as a kind of zero-order test which considers the usage of words in isolation. It takes no account of the structural usage demonstrated by a word—its proximity and juxtaposition with respect to neighbors. But if closed-class words represent functional categories, then words from the same category might be expected to demonstrate similar structural usage. This can be determined by comparing the number of times each one is used in a structural context similar to that of another.

Define the "first-order successors" of a function word to be the set of words that immediately follow it in a particular text. (To extend the idea further, the "second-order successors" can be defined as the set of words following second after it, and so on.) The relative size of the intersection of the first-order successors of

two function words is a measure of how often the words are used in similar syntactic structures. Where two closed-class words share an unusually common structural usage, we assume that they are functionally similar.

To determine whether two function words have a unusually large degree of commonality in their first-order successors, assume that closed-class words play no part in establishing functional roles. Then the words following each particular closed-class lexeme in a text would represent a more or less random sampling of the vocabulary.

By counting the number of different words that occur after two particular closed-class words, the expected number of different words that will appear after both can be calculated, under the assumption of random sampling. In fact, the degree of commonality is often very much higher than expected. This is no doubt partly due to the breakdown of our simplifying assumption. However, in some cases the degree of commonality—measured as the probability of this much commonality occurring by chance—is so extremely high that it indicates a substantial similarity between the syntactic roles of the two closed-class words being considered.

What is the probability that the intersection between two randomly-chosen sets is as large as a given value? Consider sets S_1 and S_2 of given sizes n_1 and n_2, whose members are drawn independently and at random from a set of size N. Denote the size of their intersection, $|S_1 \cap S_2|$, by the random variable I. It can be shown that I is distributed according to a hypergeometic distribution, and the probability that it exceeds a certain value n, $Pr[I \geq n]$, can be determined. Unfortunately, the calculation is infeasible for large values of n_1, n_2 and N. Various approximations can be used to circumvent the problem, such as the binomial, Poisson and Normal distributions.

For example, suppose that for a particular corpus with a vocabulary of 10000 words ($N = 10000$), two particular function words are both followed by 2000 different words ($n_1 = 2000$, $n_2 = 2000$). Suppose that these two sets have 700 words in common ($n = 700$). Then the Normal approximation has mean $\mu \approx 400$; in other words one expects only 400 words to be in common if the sets were randomly chosen. Its standard deviation is $\sigma \approx 16$, and so the actual figure of 700 is 19 standard deviations from the mean. It follows that the probability of I being at least as large as it is, $Pr[I \geq 700]$, is very tiny—about 10^{-80}. (In fact tables of the Normal distribution do not generally give values for $z \geq 5$—they end with $Pr[z > 4.99] = 3 \times 10^{-7}$.)

To estimate the probability $Pr[I \geq n]$ in general, several approximations are possible. It was decided to split the problem into three cases depending on the size of n, n_1 and n_2. First, when $n = 0$, use $Pr[I \geq 0] = 1$. Second, when either n_1 or n_2 is large (say n_1 or $n_2 > 100$), use the Normal approximation to the hypergeometric distribution, employing standard mathematical tables to approximate the integral that is involved. Otherwise, when both n_1 and n_2 are small (i.e. ≤ 100), calculate an approximation directly from the hypergeometric distribution and evaluate it using precomputed factorials up to 100 stored in a table.

Table 3.5 lists the probabilities calculated for intersection sizes of the first-order successors for some of the function words in the novel *Far From the Madding Crowd*. The first line shows that "I" and "you" were followed by 231 and 293 different words respectively, of which 110 are in common. Considering the vocabulary size of 11,589 words, it is very unlikely that as many as 110 would be in common had the successors been randomly chosen—the probability is in fact only 10^{-316}! "I" and "you" thus

word	first-order successors	word	first-order successors	intersection size	log probability	apparent association
I	231	you	293	110	−316.0	strong
we	71	you	293	45	−238.0	strong
her	557	you	293	55	−27.7	weak
he	348	they	138	71	−253.0	strong
her	557	my	243	99	−149.0	strong
him	113	me	104	27	−149.0	strong
her	557	his	562	149	−138.0	strong
him	113	he	348	20	−18.9	weak
his	562	he	348	13	−0.1	weak
had	341	have	205	80	−211.0	strong
had	341	was	641	115	−117.0	strong
is	229	was	641	93	−117.0	strong
from	126	was	641	32	−23.1	weak
about	63	at	124	24	−184.0	strong
at	124	from	126	29	−127.0	strong
on	147	from	126	28	−101.0	strong
have	205	at	124	15	−18.9	weak
was	641	at	124	26	−15.2	weak

Table 3.5: Probabilities for intersection sizes (vocabulary: 11,589 words)

seem to perform similar functions. So do "we" and "you", whereas "her" and "you" are much less strongly associated. The remaining blocks of the table give samples of other associations, both strong and weak. Possessive pronouns, for example, show strong associations with each other, as do pronouns in the same case (i.e. nominative, objective, etc.). Relatively weak associations are indicated by comparisons across such class boundaries. Auxiliary verbs also show strong associations with each other, and prepositions do as well, yet these two categories offer little statistical evidence of any relationship between them.

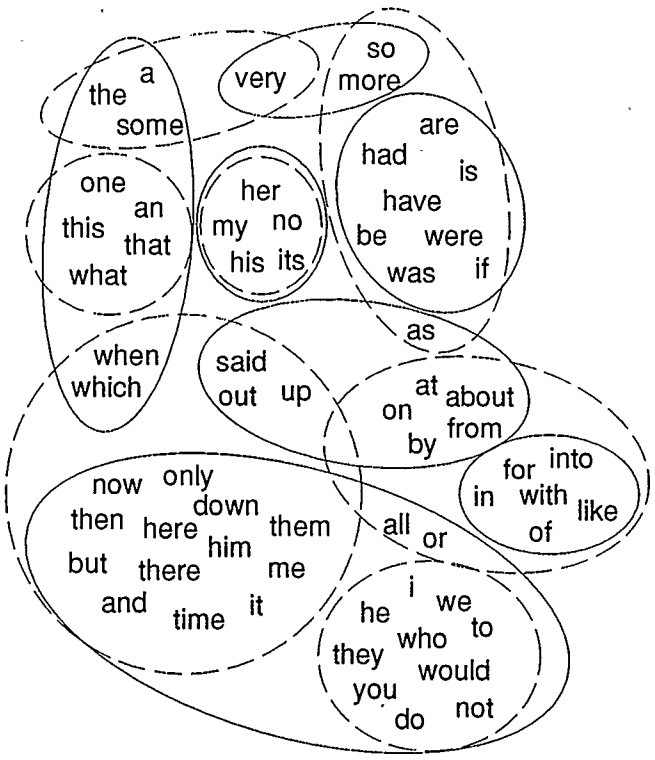

Figure 3.2: Categorization clusters for Hardy (solid lines) and Melville (dashed lines)

3.2.7 Clustering function words

Function words can be divided into syntactic categories by assuming that the strongest associations are between those whose first-order commonality is most unlikely to have arisen by chance. First, calculate the probabilities for the first-order successors' intersection sizes observed between each pair of function words. Then, place each particular word into the same syntactic category as the one to which it most strongly associated, where "strength" is measured by the unlikelihood that the two words would demonstrate such similarity in usage accidentally.

This scheme works well for most of the closed-class lexemes. However, due to a

category	elements			
fw_0	a	an	her	his
	my	no	one	that
	the	this	what	your
fw_1	he	I	she	then
	they	we	who	you
fw_2	are	be	had	has
	have	if	is	was
	were			
fw_3	can	could	did	do
	does	might	must	should
	will	would		
fw_4	here	him	it	me
	them	there	us	which
fw_5	all	and	as	but
	how	not	now	only
	or	than	to	when
fw_6	more	much	so	some
	very			
fw_7	about	after	at	by
	for	from	in	into
	like	of	on	out
	up	with		

Table 3.6: Function word categories

phonetic peculiarity, the words "a" and "an" exhibit a very poor first-order relationship and consequently do not end up in the same functional category. This undesirable situation could be avoided if the second-order successors could be brought into the categorization procedure, but to do this in a general way would require a scheme for weighting each of the n-order probabilities. Alternatively, if both "a" and "an" were compared with "the" before being compared with each other, they would all be categorized together. However, this would require artificial manipulation of the order of comparisons.

A third, less contrived, solution is to reassess the initial groupings to check whether each function word is in its best category and, if not, reassign it. For every function word, the distance is calculated to each category by averaging its first-order association probability with every word in the category. It is then reassigned to the closest category. The procedure is iterated until no reassignments occur. Figure 3.2 shows the final categories obtained by applying this clustering technique to the texts of Hardy and Melville. These categories do reflect functional similarities for closed-class words, particularly in the case of determiners, auxiliary verbs, prepositions, and pronouns.

It is important to note that results obtained through any clustering technique can be highly susceptible to variation depending upon initial groupings. Sets of final clusters obtained using probability distancing, though different in detail, demonstrate a remarkably high level of consistency from a variety of initial states. An analysis of the robustness of this probability distancing algorithm is presented in Chapter 5.

One function-word categorization that was obtained is summarized in Table 3.6. The remaining stages of the inferencing process are based directly on derived functional categories. Therefore, the categories listed in Table 3.6 are used as the basis for the content-word classification and grammar induction procedures described in the following sections.

3.3 Open-class words

Every word that does not qualify for membership in closed class categories is, by default, an open class word. As a matter of consequence, discovering such words is

trivial once the closed class words have been identified, but since approximately 98 percent of the vocabulary demonstrated in a large text will be construed as open class, the chore of determining syntactic categories within this class and assigning each of the 10,000 or so remaining lexical items to such categories is significantly more difficult.

In contrast to the syntactically functional roles that we have supposed are fulfilled by the closed class elements, we assume that the roles for open class words are ones which supply content, or meaning, to text. Drawing on notions of what may be called *classical* categories, we sustain a certain expectation that the categories to which open class words are to be assigned correspond to general types of referents. That is, we expect each content lexeme to convey a particular kind of referential information, and it is the nature of its *kind* that defines the category to which the lexeme belongs. For example, we have an intuitive sense that some meaning-laden words function semantically as referents to specific objects or classes of objects whose existence is real, surreal or imaginary. Other meaningful elements function as referents to qualities attributable to such objects—qualities like colour, texture, shape, or temperature—and still others refer to actions perpetratable by or to existent objects. We have a nomenclature for such classical categories—Noun, Adjective, and Verb respectively—and their character is for the most part clear [23, 40]. But a static analysis of language precludes immediate access to any sort of semantic information that would help derive these categories and assign open class words to them. Even so, consideration of the conspicuous traits of classical categories does offer some insight into how reasonable approximations of these categories can be formed.

3.3.1 Categorization

Thus far, we have used the proximity of particular open class words to closed class words in a very straightforward manner for establishing structural and categorial relationships between those closed class elements. This was possible primarily because the large number of open class elements allowed us to assign significant probability values to the chance occurrence of such relationships. Though in principle we could apply this same technique for establishing relationships between open class elements, the sheer size of the open class category renders comparisons of first order successors for every possible pair an immense computational problem.

In keeping with our premise of the importance of closed class lexemes in syntactic structure, we can reduce the number of comparisons that must be made by assuming an association between each instance of open class element and a function word category. Initial open class categories can be defined by a proximity relation between each content word and a functional category. Probabilities for the intersection sizes of each pair of initial categories are calculated and, in a single pass, each category is merged with the group that demonstrates the strongest relationship. The resulting categories are deemed the final open class categories.

The effectiveness of this simple strategy hinges on the soundness of the proximity relation used to form the initial groups. That is, the initial categories must be formed according to relevant criteria. Though such criteria are determined in a somewhat arbitrary manner, a careful examination of the ways in which content words are used offers some useful insight towards establishing a reasonable proximity relation.

3.3.2 Classical expectation

We would expect inferred lexical types to match classical categories fairly closely, and it is this expectation that leads us to look for certain regularities in the usage of elements from a particular classical category as a guide towards developing categorization criteria. Consider the following examples of noun usage:

1a) *The little brown* ***fox*** was quite lost.

1b) *An old* ***man*** slept on *the* ***sidewalk.***

1c) ***He*** left after eating *Alison's* ***lobster.***

1d) *Many* ***people*** have foolishly fed *wild* ***animal*** from *their* ***cars.***

The position of nouns in these examples demonstrates their consistent occurrence as the last word of the noun phrase structures. Further, most of these example noun phrases begin with one of the closed class elements from the fw_0 category listed in Table 3.6. We need not be too concerned with whether these apparent regularities hold for all English noun phrases, rather we ought only consider whether they are useful observations prima facie (n.b. phrase termination is, in fact, not true for noun usage as a whole, but it is common).

What these examples indicate is that the direct proximity relation between the fw_0 word and noun is very weak. However, the respective positions of the two words within the noun phrase suggest that the structural roles of the words are perhaps constrained by the requirements of the phrase itself. That is, the consistent use of fw_0 words in the initial position and nouns in the final position of noun phrases is a possible characterization of the phrase structure. Formalization of such regularities is the objective of the inferential process and they are not included as base case

information. It is therefore incumbent on us to establish a means by which phrase structure itself can be generalized before positional constraints on its constituents can be established.

3.3.3 Rightward selection

The fact that determiners appear exclusively in noun phrases suggests that there is indeed a relationship between determiner and noun [1]. Moreover, whenever determiners appear they mark the onset of a noun phrase. Insofar as the fw_0 category can be likened to determiners, fw_0 elements also indicate the onset of some kind of phrase. From this we may define the left bound of that phrase type as the fw_0 element itself. Extrapolating, we may assume that any function word element indicates the onset of some phrase type and, consequently, that the phrase is bounded on the right by either another function word (indicating the start of a new phrase) or the end of the linguistic expression (since the end of a sentence is the end of a phrase qua a phrase). Such a definition is very attractive from a computational standpoint in that it allows for straightforward isolation of phrase structures. Conveniently, the definition is also congruent with a regularized X-Bar theory in that every phrase complement can be expressed in terms of rightward selection from the phrase head [2].

3.3.4 Creating non-function-word categories

Each instance of a phrase presumably demonstrates a sequence of open class categories characteristic for that particular phrase type. But, from this same syntactic perspective, each category is characterized by its ability to fulfill a particular struc-

tural role within certain phrase structures. Though somewhat circular, we can use these assumptions to create definitions for both phrase types and content word categories. A phrase type is defined by its three static attributes: 1) the function word category that heads it, 2) the number of words comprising the phrase, and 3) what follows it. From this, we can further define a structural role for all open class words by noting *a*) the phrase type(s) in which it is used and *b*) the position therein that it occupies. If the position of a syntactic role within each phrase type is fixed then a statistical analysis of where a given open class word occurs within particular phrase types will allow us to determine a categorial relationship between that word and others demonstrating a similar usage.

The first stage of categorization requires that each open class word be assigned to an initial category. The category is individuated according to the function word category heading the phrase in which the open class word appears, what follows that phrase, the length of the phrase and the relative position occupied by the open class word within it. For example, the sentence

A tiny bird sat in the tree

has the functional phrase structure

fw_0 tiny bird sat fw_7 fw_0 tree.

This allows its constituent content words to be assigned to temporary categories such that

1a) tiny $\in cw(fw_0, fw_7, 1, 3)$

1b) bird $\in cw(fw_0, fw_7, 2, 3)$

1b) sat $\in cw(fw_0, fw_7, 3, 3)$

1b) tree $\in cw(fw_0, fw_\phi, 1, 1)$

where $cw(fw_0, fw_7, 2, 3)$ is interpreted to mean the set of content words appearing in the second position of a phrase with length three, where the phrase is headed by a word from the fw_0 category and followed by a phrase headed by a word from the fw_7 category. Similarly, the word "tree" has been assigned to $cw(fw_0, fw_\phi, 1, 1)$, the open class category for words appearing in the first position of a phrase with length one, where the phrase is headed by a word from the fw_0 category and followed by the end of a sentence (i.e. the empty phrase fw_ϕ). As each sample string from the presentation sequence is processed, previously unseen content words are added to existing sets or new categories are created for them. No provision is made to prevent words from being assigned to multiple categories, though duplicates are removed from within each category.

3.3.5 Category generalization

When applied to *Far From The Madding Crowd*, this procedure creates about 90,000 initial categories. Each is subsequently compared against all others in the same manner as the first-order successors for function words were compared. That is, the strength of the association between two categories is determined by the probability that the sets have an intersection of the size exhibited. The larger the intersection, the more likely it is that the categories share the same lexical function. After probabilities have been calculated for all category pairs, each category is collapsed into a single set with the category to which it is most strongly related. Amalgamation is carried out in a single pass and a record is kept of each merger so that initial categories whose strongest relation has already been subsumed can be added to that same set. Once again, no provision is made to prevent words from existing

category	elements		
cw_{41}	pulled	sent	drew
	wrong	formed	asked
	visible	returned	short
	used	closed	
cw_{44}	certainly	merely	entirely
	already	apparently	sometimes
	really	nearly	hardly
cw_{57}	doing	beginning	able
	coming	next	began
	feeling	looking	having
	going		
cw_{58}	miles	circumstances	pounds
	clothes	hours	arms
	feet	neighbours	thoughts
	horses	trees	features
	lips	days	others
	sort	hands	minutes
	things	times	people
	sheep	women	years
	words	men	

Table 3.7: Some content word categories from *Far From the Madding Crowd*

as members of more than one set, though duplicates within each set are removed.

3.3.6 The final categories

Table 3.7 shows some of the 61 final content word categories derived using this technique. Category cw_{44} exemplifies a fairly sound collection of adverbs, and cw_{41} and cw_{57} are reasonably consistent sets of past tense and present participle verbs respectively. Category cw_{58} includes many of the plural nouns from *Far From The Madding Crowd.* These groupings represent some of the more coherent open class categories; however, they do not demonstrate complete collections of the classic grammatical forms they exemplify. For example, most of the present participle verbs

used in Hardy's novel are found in groupings not listed here, often mixed in with words from a variety of standard syntactic categories. Of the 61 categories, 58 contain fewer than 170 words, each of which tends toward a particular grammatical class. Unfortunately the three largest sets contain over 3000 words and do not submit to characterization under traditional syntactic forms. In general, the larger the group the more difficult it is to interpret using standard grammatical terminology.

Though 61 final categories constitutes a significant improvement from the original 90,000, it is still quite a few more than the dozen or so standard classical categories of nouns, prepositions, and the like. This apparent discrepancy arises from syntactic properties attributed to such things as inflection, tense and number agreement. A more thorough evaluation of the final content word categories is discussed in Chapter 5. However, it is worth noting here that the effects of inflection and agreement on syntactic structure are readily acknowledged in linguistics literature, and any system (such as this one) that avoids transformational analysis of deep structures would have to be disregarded if no concession is made for it at the surface level.

3.3.7 Category symbol conventions

The type of function word that follows a phrase in which a content word appears is incorporated into the initial category information to allow for lexical distinctions to be made based on such things as number agreement, passivization and verb valency (e.g. transitivity and tense). Once specific lexical groups are collapsed and condensed into their general categories this information need no longer be preserved, and each final category can be labeled more simply. However, it is essential that the syntactic inferencing component described in the following chapter be able to

distinguish between function and content word category symbols. As has already been seen in this chapter, a notation has been adopted where a category symbol of the form fw_i indicates a function word from the closed class category indexed by i. Similarly, a category symbol of the form cw_j indicates a content word from the open class category indexed by j. This notation is employed in the remainder of this thesis.

Chapter 4

Syntax Induction

The definition for a grammar presented in Section 1.1.1 viewed language as a collection of language elements at various levels of abstraction, each level subject to its own grammaticality constraints. Grammatical inference is specifically concerned with uncovering generalizations about constraints at the word level—the syntax of a language. Syntax comprises several levels of abstraction within the hierarchy of linguistic structure. As Figure 4.1 shows, words combine to form sub-phrasal elements, which combine to form phrase segments, which combine to form sentences. Syntax induction involves the compilation of a formal description for linguistic structures at each of these levels. The method by which this so-called Phrase Structure grammar is derived is the subject of this chapter.

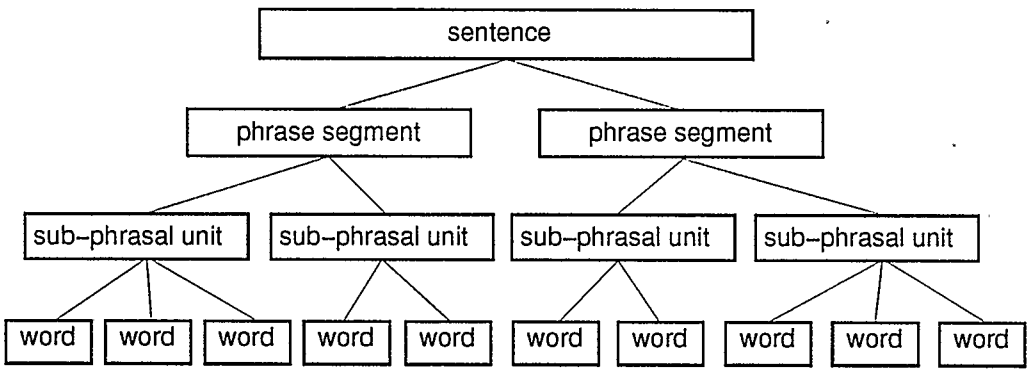

Figure 4.1: Levels of syntax.

4.1 Generalizing linguistic structures

In the broadest sense, a Phrase Structure grammar is a description of word sequence regularities detected in linguistic expressions. In general, efforts to produce such grammars through automatic induction have done little to incorporate theoretical principles of language presumed to constrain syntactic forms. As a consequence, the results obtained describe word patterns present in sample expressions, but neither provide nor reflect extensible language properties. That is, they attempt to explain well-formedness as a product of the attributes of individual words. Given the nearly infinite set of orderings possible for the words of any natural language, this kind of approach entails a description of syntax that is inherently exponential and unusable.

Theoretical linguistics assumes that the syntactic regularities of interest are not those demonstrated by particular words so much as those demonstrated by sequences of word categories. Although it may be of passing interest to know that the phrase "a spotted dog" occurs in certain positions and with a certain frequency in typical English discourse, it is inherently more valuable to study occurrences of the sequence "determiner-adjective-noun" instead. Category sequences represent a higher level of generalization about language and presumably reflect deeper knowledge of the principles that govern it.

It is not surprising that grammar induction systems have shied away from seeking generalizations about category sequences. Chapter 3 argues that syntactic categories are not systemically the properties of individual words, rather they are attributed to words when they function in particular structural roles. That is, a syntactic category is a property of a place within a syntactic structure, and that category is attributable

to any word that assumes that place in the production of a well-formed expression. This subtle distinction implies that syntactic categories should not be included as a priori information within a grammatical inferencing system. They should be derived from the sample text.

The categorization procedures outlined in Chapter 3 provide a means for establishing lexical categories according to structural positions. The procedures depend on two assumptions: 1) that the sample text consists entirely of well-formed expressions, and 2) that the syntactic category of every functional element is unambiguous. The validity of these assumptions is analyzed more critically in Chapter 5, but their strengths are plain. While the first is made as a matter of necessity, the second allows for fixed points to be established within sample utterances from which other syntactic categories can be defined by proximity relations. Once lexical categories have been established, the way is clear to uncover the syntactic structures they combine to form.

4.1.1 Variable substitution

Phrase structure grammars do not distinguish between terminal and non-terminal symbols in terms of their membership of the vocabulary being considered [29]. Since word categories are represented within the vocabulary as non-terminal symbols, any process of induction that can be applied to specific word sequences can be applied to sequences of category symbols in the same manner. The approach adopted in this thesis is a process of variable substitution. The process entails replacing repeating patterns of category symbols with super-symbols while recording the substitutions as production rules. Rule classes are defined according to the combinatorial properties

of the pattern sequences and rewritten as disjunctive expressions. The final set of production rules is a context-free grammar for the sample text.

4.1.2 Pattern constraints

The patterns of interest are not simply those haphazardly found distributed through the sample expressions. Following from the principle of compositionality explained in Section 2.1.2, the surface form of a linguistic expression is (transformations notwithstanding) a combination of phrase segments, which are themselves combinations of words. Section 2.1.3 argues that the unity within each phrase segment stems from a genuine psychological bond constraining its composition and form—constraints different from those that bind phrases into whole expressions. Inasmuch as the induction process should avoid the formation of production rules that compromise this difference, a distinction is made between infra-phrasal patterns (regularities within phrase boundaries) and supra-phrasal patterns (regularities across phrase boundaries). The variable substitution technique described in the following sections focuses on infra-phrasal patterns first, and incorporates the results into generalizations of broader sentence forms.

4.1.3 Overview of the inferencing process

The syntax induction procedure outlined in the remainder of this chapter is a multi-stage process which, when applied to a single text, yields a CFG for that text. Unlike most grammatical inference methods, which form successively broader generalizations of syntax through a sequential analysis of sample expressions, the process described below applies wholesale principled pattern analysis to entire texts.

Figure 4.2 outlines the induction procedure. The first stage is the substitution of each word in the sample expression by its corresponding category symbol derived through the categorization procedure described in Chapter 3. Each symbol designates a particular word category, but also serves to establish the word's membership in one of two broad lexical classes: 1) function words and 2) content words.

Instances of function words are presumed to indicate phrase boundaries. The second and third stages of the process use this information to form generalizations about sequences of category symbols within these boundaries. In stage two, sentences are dissected into phrase segments and recorded as production rules. Contiguous strings of content word symbols are extracted from the phrase segments and sorted by length. Each shorter sequence found to occur within a longer one is replaced with a super-symbol and a production rule is recorded for that substitution during the third stage of the induction process.

In the final stage of the induction process, the infra-phrase structures are categorized according to their combinatorial properties, and rewritten as disjunctive rules. The final set of production rules is output in BNF as a context-free grammar for the text.

4.2 Function word phrases

According to the hierarchy of syntax shown in Figure 4.1, sentences are composed of phrase segments, which are themselves composed of sub-phrasal elements. Inasmuch as smaller linguistic units combine to form larger ones, it follows that the first step towards generalizing sentence structure is to capture the regularities present at the

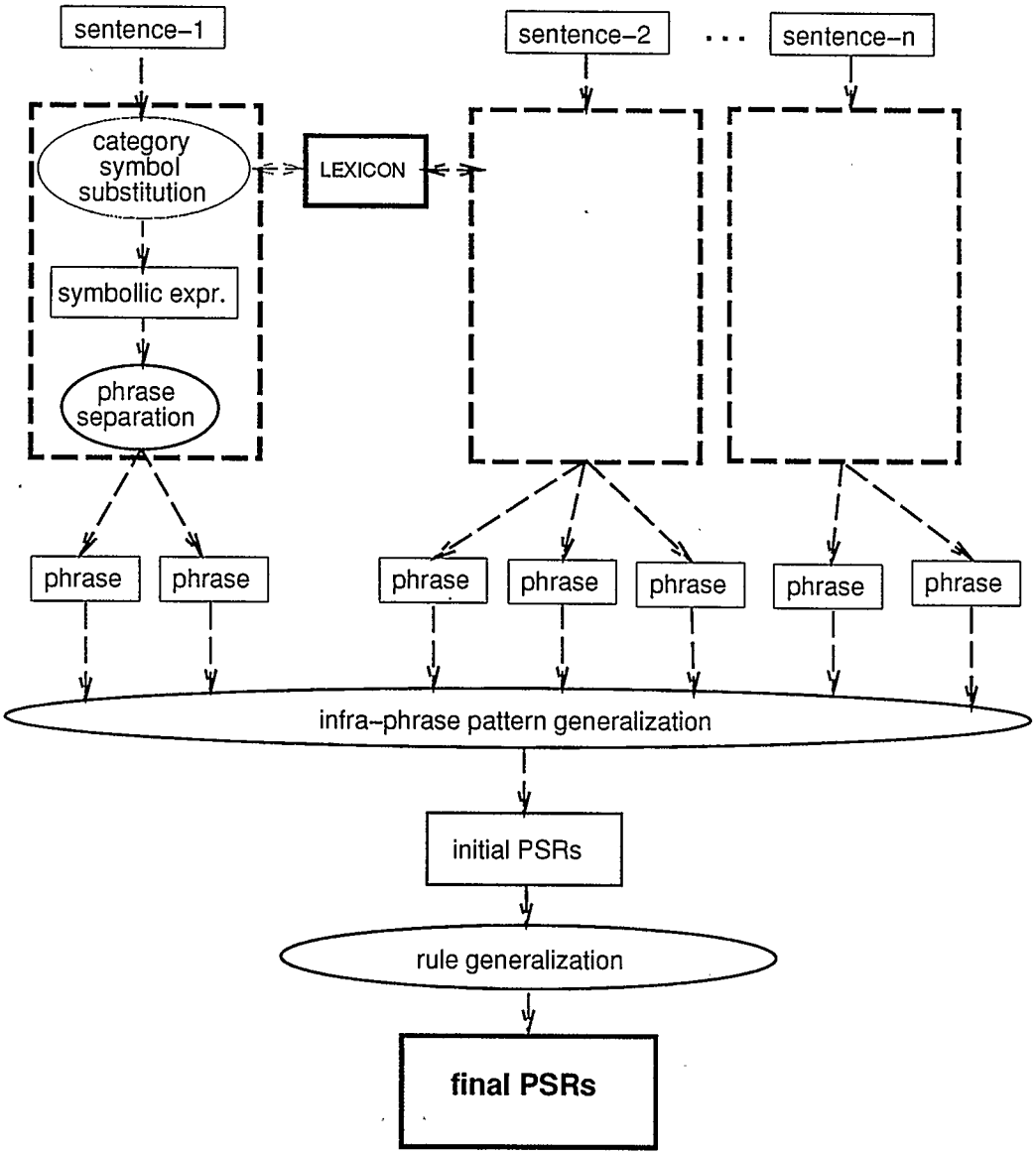

Figure 4.2: Overview of the induction procedure

lowest level of abstraction—the infra-phrase patterns. Thus, the first stage of the induction process is the creation of a generalized phrase structure grammar for fw-phrases alone.

4.2.1 Isolating phrases

Once all words of the sample strings have been replaced with their appropriate category symbols, the strings are dissected into phrases according to the boundary definitions outlined in Section 3.3.3—that is, each phrase segment is bounded on the left by a function word (i.e. headed by a function word) and bounded on the right by a subsequent function word or the end of the sentence. For example, the expression

the tiny bird sat in a hollow tree

could, via the categorization and symbol substitution procedures, yield the string of category symbols

$fw_0\ cw_{24}\ cw_{51}\ cw_{40}\ fw_7\ fw_0\ cw_{24}\ cw_{51}$.

This symbolic expression is then broken into *function word phrases* (fw-phrases), such that each fw-phrase is comprised of a string of category symbols headed by an initial functional element and terminated by a function word symbol that specifies the type of fw-phrase that follows the phrase in question. The expression above decomposes into

$fw_0\ cw_{24}\ cw_{51}\ cw_{40}\ fw_7$
$fw_7\ fw_0$
$fw_0\ cw_{24}\ cw_{51}\ fw_\phi$.

Unlike the other category symbols, the terminating symbol for each segment does not denote a substituted word. It merely indicates the type of fw-phrase that follows the segment in question and serves to preserve fw-phrase links within the grammar. As before, the symbol fw_{ϕ} represents a null fw-phrase and is used to mark the end of a sentence. No distinction is made between full-stop characters of different sentence moods—thus periods, question marks, and exclamation points are treated uniformly.

Each unique fw-phrase is recorded as a production rule, and its phrase type is codified by the super-symbol. Phrase types are differentiated by their head element category and by the category of the head element in the fw-phrase that follows—contextual information that is ultimately used to categorize fw-phrase types according to more discriminating structural properties. The fw-phrases above are rewritten as

$$Fp_{0,7} \Rightarrow fw_0\ cw_{24}\ cw_{51}\ cw_{40}\ Fp_{7,0}$$
$$Fp_{7,0} \Rightarrow fw_7\ Fp_{0,\phi}$$
$$Fp_{0,\phi} \Rightarrow fw_0\ cw_{24}\ cw_{51}\ Fp_{\phi}.$$

Each rule expresses a fw-phrase defined by its functional head, content-word substring, and fw-phrase continuation. For example, the rule $Fp_{0,7}$ defines a fw-phrase headed by a fw_0 element, continued by words from each of the content categories cw_{24}, cw_{51} and cw_{40}, and followed by a fw-phrase headed by a fw_7 element. Similarly, the rule $Fp_{0,\phi}$ defines a fw-phrase headed by a fw_0 element, continued by words from the cw_{24} and cw_{51} content-word categories, and followed by the *null* fw-phrase—the end of the sentence.

4.2.2 Generalizing infra-phrase sequences

In keeping with the hierarchical view of syntax, stronger structural bonds are presumed to exist within phrase boundaries than across them. The next stage of induction, therefore, seeks to form generalizations of symbol sequences within phrase segments. For example, the sequence

$cw_{24}\ cw_{51}$

is present within the first and last of the three fw-phrases above—a repetition that invites further generalization.

Because of the way fw-phrases are defined, only sequences of content word symbols can exhibit such patterns. Strings of contiguous content word symbols are extracted from the initial rules and sorted by decreasing length. Each unique sequence is rewritten as a production rule and compared against longer sequences in case they form a substring of another rule. If so, the symbol for the shorter sequence is substituted into the longer one. Comparison continues for shorter and shorter sequences down to substrings of length two. The result of this stage of processing is a context-free grammar for the fw-phrase segments. Table 4.1 shows the fw-phrase grammar for the example sentence.

4.2.3 The headless phrase

It is conjectured that the high frequency nature of closed-class elements in linguistic expression is a direct consequence of their functional importance [5, 33, 15]. Even so, it is not unusual in large texts to find some sentences that do not contain words from any of the derived functional categories. Some may contain one or two of the 400 or

the tiny bird sat in a hollow tree		
S	$\Rightarrow$	$Fp_{0,7}$
$Fp_{0,7}$	$\Rightarrow$	$Cp_1\ Fp_{7,0}$
$Fp_{7,0}$	$\Rightarrow$	$fw_7\ Fp_{0,\phi}$
$Fp_{0,\phi}$	$\Rightarrow$	$fw_0\ Cp_2\ Fp_\phi$
Cp_1	$\Rightarrow$	$Cp_2\ cw_{40}$
Cp_2	$\Rightarrow$	$cw_{24}\ cw_{51}$
cw_{24}	$\Rightarrow$	tiny || hollow
cw_{40}	$\Rightarrow$	sat
cw_{51}	$\Rightarrow$	bird || tree
fw_0	$\Rightarrow$	the || a
fw_7	$\Rightarrow$	in
Fp_ϕ	$\Rightarrow$	.

Table 4.1: A context-free grammar for the example sentence.

so less frequently used function words that escape detection through the discovery procedures outlined in Chapter 3. Such sentences are, in principle, consistent with a functional view of syntax. However, it is still incumbent on us to provide an account for those sentences that contain no function words at all.

Most anomalous sentences can be reconciled to the premise of a function word syntax if inflectional morphemes, such as the tense marking verb suffixes *-ed* and *-ing*, are also treated as functional elements—a view taken by Abney [1] in DP-theory, and by Garrett [28] in his positional model of sentence production. In fact, within DP-theory the possessive morpheme *'s* is regarded as the functional head of genitive noun phrase structures.

The syntax induction mechanism described in this thesis has no morphological component to identify inflectional morphemes—a shortcoming that indicates a possible extension to the system. But since the induction relies on closed-class elements to mark the onset of phrase structures, the addition of such a component would

not provide a means for dealing with sentences that do not begin with a functional element. Conformance for such discrepancies can, however, be achieved through broad interpretation of the *null category* that forms part of DP-theory. Section 2.1.6 noted that every noun phrase must have a functional head specifying the reference of that phrase. Noun phrases that do not demonstrate an explicit head element are attributed headship from the null category. In such instances, the head is implied as a trace element.

The syntax induction procedures outlined in this chapter express each sentence as a chain of production rules that form a left-associative [34] sequence of category symbols from S (the sentence start) through to Fp_{ϕ} (a full stop). Most often S rewrites immediately to some fw-phrase symbol, but may on occasion rewrite to a content word phrase symbol resulting from sentences that do not begin with a function word. In principle, a function word trace category could be created to prefix these anomalous sentence starts, thus creating a uniform fw-phrase description for all infra-phrase structures. But the null category already included in the syntactic description can also serve to mark this trace element.

Every sentence structure terminates with the Fp_{ϕ} symbol. Viewed another way, every sentence start is also preceded by the Fp_{ϕ} symbol since all but one sentence of a text is followed by another. In this respect, the fw_{ϕ} category can be regarded as a de facto head element for all sentence structures. In practice, however, there is no pressing need to incorporate a specific head element into sentence structures that begin with a content word symbol. They are, therefore, left as headless phrases.

4.3 Phrase classification

Results from the procedures outlined in the previous section represent the lowest level of infra-phrase pattern analysis achieved by the inferencing system—generalization of basic content word sequences. The final stage of inferencing is the characterization of general sentence structures according to the phrase segments that combine to form them—in effect, generalization of the combinatorial properties of fw-phrase types.

4.3.1 The over-generalized grammar

At this point in the induction process, each sentence of the source text rewrites to a single production rule from which a chain of fw-phrase segments can be traced through to a full stop. As a consequence, some sequences of fw-phrase segments are expressed as novel chains within the grammar without being exhibited in the text. This is, to a certain extent, a desirable property of the grammar in that it allows for an account of well-formed expressions that have not been observed. It is, however, also an undesirable property because it allows for an account of malformed expressions as well. As noted in Section 2.3.1, it is generally accepted that construction of a grammar that accepts all and only the expressions of the target grammar is impossible. Even so, we can make the inferred grammar more discriminate by introducing rules that limit the ways in which fw-phrase segments can be combined.

Consider the sentences

1a) The salesman looked up the account.
1b) The policeman walked up the street.

The following grammar may be inferred as a characterization of their structural similarities.

2a) $S \Rightarrow Fp_{0,7}$
2b) $Fp_{0,7} \Rightarrow fw_0\ Cp_{19}\ Fp_{7,0}$
2c) $Fp_{7,0} \Rightarrow fw_7\ Fp_{0,\phi}$
2d) $Fp_{0,\phi} \Rightarrow fw_0\ Cp_{37}\ Fp_{\phi}$
2e) $Cp_{19} \Rightarrow cw_{31}\ cw_{42}$
2f) $Cp_{37} \Rightarrow cw_{31}$
2g) $fw_0 \Rightarrow$ the
2h) $fw_7 \Rightarrow$ up
2i) $cw_{31} \Rightarrow$ salesman || policeman || account || street
2j) $cw_{42} \Rightarrow$ looked || walked.

As a consequence of the generalization, the grammar asserts the grammaticality of some unseen expressions, like

3a) The salesman looked up the street
3b) The policeman looked up the account
3c) The salesman walked up the street.

But, it also asserts well-formedness for a few nonsense expressions, like

3d) The salesman walked up the account
3e) The street looked up the policeman
3f) The account looked up the street.

This does not necessarily indicate a problem with the grammar since it is easily argued that sentences 3d)–3f) *are* syntactically well-formed.

The inadequacy of the grammar is made apparent when the rule set is expanded to account for such constructs as restrictive clauses. For example, to account for the sentence

1c) The salesman up the street looked up the account

the following rule is added to the grammar

2k) $Fp_{0,7} \Rightarrow fw_0\ Cp_{37}\ Fp_{7,0}$.

This new rule allows the grammar to account for an even larger number of unobserved well-formed expressions. Unfortunately, it will also assert the grammaticality of malformed sentences like

3g) The salesman up the account.
3h) The street up the policeman.

Such discrepancies reflect the fact that the grammar does not dictate sufficient constraints on the combinatorial capabilities of fw-phrases.

4.3.2 Refining fw-phrase types

Each fw-phrase rule rewrites to a functional element, some sequence of content word symbols, and a fw-phrase continuation. At this stage of processing, fw-phrase types are individuated according to their head element and continuation symbol—information that is partially encoded in the phrase symbol. For example, the symbol $Fp_{0,7}$ represents a fw-phrase headed by a fw_0 element and followed by a fw-phrase headed by a fw_7 element. There are, of course, many fw-phrase rules that have these characteristics and under this formulation all are examples of a particular class of fw-phrase. For example, the following rules

4a) $Fp_{0,7} \Rightarrow fw_0\ Cp_{28}\ Cp_{29}\ Fp_{7,1}$

4b) $Fp_{0,7} \Rightarrow fw_0\ Cp_{28}\ Cp_{31}\ Fp_{7,3}$

4c) $Fp_{0,7} \Rightarrow fw_0\ Cp_{55}\ Cp_{31}\ Fp_{7,3}$

are fw-phrases of the same type. They have many common features—most notably, they all begin with the same function word category. The first two rules begin with the same pair of symbols, the last two rules end with a common sequence. In fact, the last two differ by only one symbol. To improve the conditions by which fw-phrase types are defined, we must decide which, if any, of these attributes dictate their combinatorial property.

4.3.3 Constraints on combinatorial properties

An intuitive assessment of rules 2b) and 2k) of Section 4.3.1, and their infra-phrasal components, suggests a possible source for combinatorial constraints on fw-phrase types.

2b) $Fp_{0,7} \Rightarrow fw_0\ Cp_{19}\ Fp_{7,0}$
2k) $Fp_{0,7} \Rightarrow fw_0\ Cp_{37}\ Fp_{7,0}$

2e) $Cp_{19} \Rightarrow cw_{31}\ cw_{42}$
2f) $Cp_{37} \Rightarrow cw_{31}$

2i) $cw_{31} \Rightarrow$ salesman || policeman || account || street
2j) $cw_{42} \Rightarrow$ looked || walked.

The only difference between the two $Fp_{0,7}$ rules is their content-phrase symbol. The infra-phrase segments represented by those content-phrase symbols, in turn, differ in that one has an additional content-word category appended to it. That particular category symbol is the distinguishing feature of the two fw-phrases and thus is quite possibly the source of their different combinatorial properties.

This conjecture is more strongly supported when the fw-phrases are described using standard syntactic terminology. Both phrases are headed by a determiner that is followed by a noun complement. In this respect, they can be likened to the determiner phrases of DP-Theory described in Section 2.1.6. One segment, however, includes a past tense verb which, under DP-Theory, would normally be part of a different phrase segment—one whose structural head is the tense marking suffix *-ed*. A reasonable extrapolation from this comparison is that the fw-phrase type headed by the fw_0 element should ideally end with the nominal element—the cw_{31} category.

Using standard grammatical analyses to distinguish between phrase types is neither practical nor desirable for the induction technique adopted in this thesis. It does, however, indicate a simple computational method that achieves, albeit roughly, the same end.

4.3.4 Phrase rules as directed graphs

Each fw-phrase rule is a linearly ordered subset of the non-terminal symbols that comprise the inferred grammar. Each rule can be expressed as a directed graph, or *digraph*. The digraphs for rules 4a)–4c) are pictured in Figure 4.3. The combinatorial relationships between these production rules can be easily identified when their corresponding digraphs are overlaid.

Figure 4.4 shows the digraph that results when those from Figure 4.3 are overlaid. At the node labeled fw_0 the path through the digraph splits as a reflection of the structural differences between the rules 4a)–4c). At the node labeled Cp_{28} the digraph splits again. One path rejoins the digraph of another rule at the node labeled Cp_{31} while the other path moves toward a different terminal node altogether.

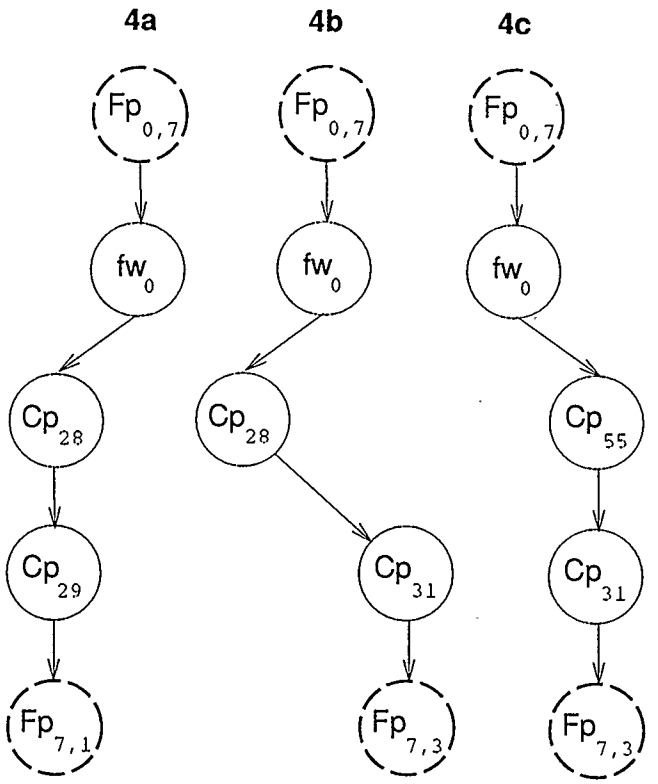

Figure 4.3: The digraphs for fw-phrase rules 4a)–4c).

The two paths between nodes fw_0 and Cp_{31} constitute optional routes through the digraph—routes with identical start and finish points. In effect, they reflect optional substructures within one general rule type. Under this interpretation, the path ending at the node labeled $Fp_{7,1}$ does not reflect an optional substructure and therefore cannot be of the same rule type.

If we view the overlaid digraphs as a single lattice, then each class of fw-phrase constitutes a sublattice. We attach the condition that the greatest lower bound of the sublattice must be a content-phrase symbol. This is necessary because terminal nodes do not correspond to actual grammatical elements but are, in effect, merely pointers to fw-phrase classes.

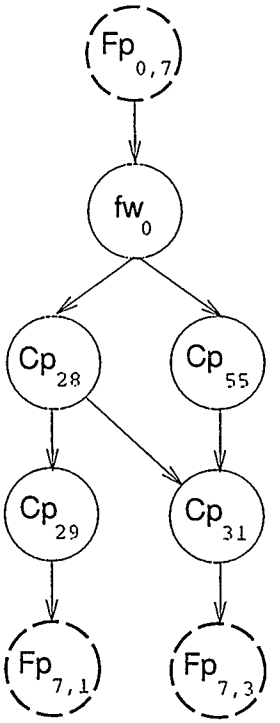

Figure 4.4: Overlaid digraphs for rules 4a)–4c).

In practice, such sublattices are easily detected. All production rules with the same head element, terminal content-word symbol and continuation symbol must be of the same type. By grouping related fw-phrase rules according to these conditions much of their redundant information can be removed. Each class can be expressed with one head element symbol and one continuation symbol. Optional substructures are expressed as a disjunction. The following production rules express the general fw-phrase types exemplified by rules 4a)–4c).

$$Fp_{071} \Rightarrow fw_0\ Cp_{28}\ Cp_{29}\ Fp_{710}$$
$$Fp_{073} \Rightarrow fw_0\ \{\ Cp_{28} \parallel Cp_{55}\ \}\ Cp_{31}\ Fp_{730}$$

The categorization of fw-phrase types according to their combinatorial properties obviates the need to preserve the head and continuation information. For this reason,

each class of fw-phrase is associated with a simple and arbitrary symbol. However, the left-associative properties of the grammar rules are maintained through preservation of the appropriate continuation symbol. That is, sentence forms are still expressed as a chain of phrase structures linked by the last symbol in each rule.

4.4 The Function Word Grammar

The result of the syntax induction process is a phrase structure grammar consisting of two general production rule types. Patterns of content word symbols are expressed as explicit rewrite rules. This type of rule does not formulate any contextual restriction for the content word sequence being expressed. In contrast, sentence structures are expressed as implicit chains of fw-phrase symbols. Each fw-phrase symbol expresses a class of structures defined by the combinatorial properties of its members. Rule classes are expressed as production rules whose right-hand side may include one or more disjunctions. The combined set of production rules is a context-free grammar for the sample text.

Chapter 5

Evaluation and Application

This chapter presents the results obtained from application of function word inferencing to a diverse collection of large, machine readable texts. The induction mechanism is evaluated according to four general criteria: 1) how well do its inferred characterizations correspond with our intuitions about language structure, 2) how well do these characterizations conform to various principles of linguistic theory, 3) to what extent can the induction methods be applied to practical language processing tasks, and 4) how can the procedures and results be incorporated into more esoteric applications.

It would be inappropriate to measure the success of the induction system through an analysis of its end product alone. Apart from the underlying notion of a closed-class vocabulary of function words, the various levels of generalization presented in this thesis result from quite dissimilar processes. Each component constitutes an independent subsystem worthy of assessment in light of its own strengths and weaknesses. These assessments are the subject of this chapter.

5.1 The sample texts

The various stages of induction outlined in the previous chapters trace a path from a given sample text through to the lexical categories and syntactic descriptions derived from it. Before we analyze each of these stages in turn, a few words about the texts

are merited.

Creating machine readable versions of large texts is a substantial undertaking. As a consequence, their selection and availability has for the most part been quite limited. In fact, the results presented in this thesis are drawn primarily from *Far From The Madding Crowd* and *Moby Dick* simply because electronic versions of these texts were readily accessible.

Fortunately, the paucity of electronic texts is beginning to change as others not only afford their time and effort to the task, but also make their texts available to the research community at large (e.g. Project Gutenberg, Illinois Benedictine College, Lisle, Illinois; electronic access to mrcnext.cso.uiuc.edu via telnet 128.174.201.12). However, this ever increasing collection of ASCII texts has been accompanied by a similar increase in the difficulty with which the texts can be used.

Apart from the actual words of the texts, the ASCII versions usually encode presentation control sequences as a means of preserving type styles and format instructions—thus allowing the texts to be subject to a wide variety of analyses. Moreover, each author's particular flair for depicting dialect, colloquial expressions, and so on, are also faithfully maintained in the electronic versions.

Before machine readable corpora can be processed by the induction procedures outlined in this thesis, format instructions and some author idiosyncrasies must be dealt with in an appropriate fashion. To accomplish this, texts are subject to a pre-processing stage that removes format instructions, expands many contractions and abbreviations, extricates dialogue from its "quote" environment, and generally transforms the text into a collection of uniform expressions that can be easily analyzed by subsequent processes.

a	about	all	an	and
are	as	at	be	but
by	do	down	for	from
had	have	he	her	him
his	I	if	in	into
is	it	its	like	me
more	my	no	not	now
of	on	one	only	or
out	said	so	some	that
the	them	then	there	they
this	time	to	up	very
was	we	were	what	when
which	who	with	would	you

Table 5.1: The closed-class derived from Hardy, Melville and Carroll

Unfortunately, no universal standards have been established for encoding format instructions, nor is there any limit to the number of presentation styles available to a given author. As a consequence, the fidelity of the text is sometimes compromised in small ways during preprocessing. Even so, it is believed that any loss of information, or misinterpretation of control sequences, that result from preprocessing have had a negligible effect on the results presented here.

5.2 Derivation of the closed-class

The first stage of inferencing is the derivation of a closed-class vocabulary to be used by subsequent induction procedures. Focusing on the high frequency nature of function words, the technique adopted involves taking the intersection of the top 1% of the vocabularies from several large texts. The soundness of this approach can be assessed by comparing results obtained from various combinations of these vocabularies.

Table 5.1 shows the closed-class words presented in Section 3.2.5—those derived from Thomas Hardy's *Far From The Madding Crowd*, Herman Melville's *Moby Dick*, and a collection of works by Lewis Carroll that includes *Alice in Wonderland*, *Alice Through the Looking Glass*, and *The Hunting of the Snark*. On the whole, these represent an intuitively sound set of function words, consisting almost exclusively of articles, pronouns, auxiliary verbs, conjunctions, and so on—the traditional closed-class elements. However, as discussed in Chapter 3, there are a few discrepancies and oversights. The paucity of feminine pronouns, for instance, and the presence of the verb *said* and the noun *time* indicate possible weaknesses of a frequency based algorithm.

According to Sapir [52], *time* is the most frequently used English noun. And, according to Caplan [15], only the majority of most frequently used English words are function words. These claims indicate that the set of words listed in Table 5.1 is the result of an algorithm that effectively identifies the most frequent English words, but one that may be unsuitable for determining the closed-class. However, it must not be overlooked that the process is highly susceptible to the kinds of texts from which the frequency lists are compiled.

5.2.1 Undesirable absence

Section 3.2.5 argues that the absence of some feminine pronouns in the final set of function words is due to the influence of Melville's vernacular on the inferencing formula. Melville was a whaler and, not surprisingly, many of his stories center on the lives of seamen at a time when women were not a usual part of everyday life—in fact, within *Moby Dick* the word *she* is most often used in reference to either a ship

a	above	after	all	also	an
and	any	are	as	at	be
because	been	being	both	but	by
can	could	did	do	does	done
each	for	from	had	has	have
he	her	here	hers	him	his
how	I	if	in	into	is
it	many	me	might	more	much
must	my	no	not	of	on
one	or	other	over	she	should
so	some	such	than	that	the
their	them	then	there	they	this
to	up	us	was	we	were
what	when	where	which	while	will
with	within	would	you	your	

Table 5.2: The closed-class derived from Hardy, research papers and news articles

or the sea. This example demonstrates how aspects of a particular vocabulary can have considerable impact on the result of the intersection process.

5.2.2 Undesirable presence

Other problems emerge when the sample vocabularies are drawn from texts that reflect a similar genre or period of literature. In the case of Hardy, Melville and Carroll, the three source texts are all examples of English fiction from the late 1800's. Furthermore, all three are narratives. The preponderance of dialogue and descriptions of settings in this style of literature can account for the anomalous verb and noun in the final closed-class.

5.2.3 Improving the closed-class

Consider the set of words shown in Table 5.2, which are derived from applying the 1% solution to *Far From The Madding Crowd*, some 20th century newspaper articles, and a collection of current research papers written by university faculty and graduate students from a variety of fields. This somewhat more diverse collection of literary styles, periods, and genres results in a much more comprehensive function word vocabulary. All nouns, verbs, adjectives and other more semantic elements have been extricated from the closed-class, and many function words that were previously overlooked are now included.

It is important to note, however, that the texts used to produce the closed-class vocabulary shown in Table 5.2 were arrived at through trial and error. Many combinations were tried until a strong set of function words emerged from the intersection process. The improvement stems chiefly from the selection of texts whose frequency lists produce better stand-alone examples of the closed-class in their top 1%.

In some ways this kind of manipulation undermines the "hands-off" principle expected for induction procedures. It does, however, still allow the texts to determine the precise membership of the function word category, and is not, therefore, as contrived a method as, say, manually adding or removing words from the derived set.

The improvements still beg the question: why not be more discriminate in the selection of the source texts? In fact, if the derived set of function words is always going to be fundamentally flawed, why not construct it from a machine-readable dictionary using the standard grammatical categories of determiner, preposition, auxiliary

verb, and so on—a set that would almost certainly be better than that which can be inferred? The answer is that the aim of this thesis is not to solve the grammatical inferencing problem. Rather, it is to determine what sample texts dictate about their own general characteristics based simply on the notion of functionally significant words. For the time being, the 1% solution offers a straightforward algorithm for deriving the closed-class, and results from different sample texts demonstrate a high degree of similarity.

5.3 Determining functional categories

Section 3.2.6 outlines the method by which function word categories are constructed. Initial groupings are established by assigning each function word to the same category as the word that demonstrates the strongest contextual similarity. Initial categories are reassessed to check whether each function word is in its best category. If it is determined that a word is most strongly related to a group other than the one it is in, it is reassigned to that group. This so-called *clustering* process is iterated until no changes are made to the categories.

The soundness of any clustering algorithm is generally dependent on two factors: 1) the method by which initial data groups are established, and 2) the criterion used to measure goodness-of-fit. The topic of clustering is worthy of study in its own right, and somewhat beyond the scope of this discussion. However, an analysis of the technique adopted in this thesis is expedient.

5.3.1 Strength and distance

The extent to which two data elements are related can be determined in any number of different ways. The measure chosen in this thesis is their similarity of first-order successors.

Section 3.2.6 outlines how the *strength* of relationship between two function words is determined according to the similarity of their respective first-order successors—the set of words that occur immediately after a particular function word in the sample text. The intersection is taken of the first-order successors for two function words and, assuming a more or less random sampling for these sets, the strength of the relationship between the function words is set as the inverse of the probability that the intersection is as large as it is.

Each function word is assigned to the same initial category as the word to which it is most strongly related. Using the first-order successors from *Moby Dick* creates the groupings on the left side of Table 5.3. Though many words of the same standard grammatical category demonstrate a strong enough first-order relationship to end up in the same derived group (e.g. possessive pronouns, prepositions), the categories themselves are often split into several groups (e.g. auxiliary verbs, determiners, objective case pronouns, etc.). Section 3.2.6 noted the undesirable situation of "a" and "an" ending up in different categories because of a phonetic peculiarity. Similar aspects of English may cause other divisions in standard grammatical categories.

A reassessment of the intial groupings is carried out to correct some of these apparent discrepancies. A *distance* is calculated between each function word and every initial grouping. Each distance is determined as the average probability of first-

initial categories				final categories			
had	have			a	the	some	very
all	and	but	do	and	but	down	here
for	here	him	it	him	it	me	now
me	not	only	or	only	out	said	them
out	said	them	there	then	there	time	when
time	up	which	with	which	up		
are	as	be	if	are	as	be	had
is	more	so	very	have	if	is	more
was	were	when		so	was	were	
he	I	then	they	do	he	I	not
who				they	to	we	who
				would	you		
to	we	would	you	an	one	this	that
				what			
her	his	its	my	her	his	its	my
				no			
in	of						
an	no	one	some	about	at	all	by
that	this	what		for	from	in	into
				like	of	on	or
				with			
a	the						
about	at	by	down				
from	into	like	on				

Table 5.3: Clustering derived categories from Melville.

order successor intersections for a function word and all other words in a grouping. If a function word is determined to be "closer" to another grouping, it is reassigned to it—a process that is iterated over new groupings until no reassignments occur. Applied to the initial groupings from *Moby Dick* yields the final categories listed on the right side of Table 5.3.

Partial unification is achieved for some of the divided categories—like auxiliary verbs and prepositions. This is perhaps more easily seen in the corresponding cluster

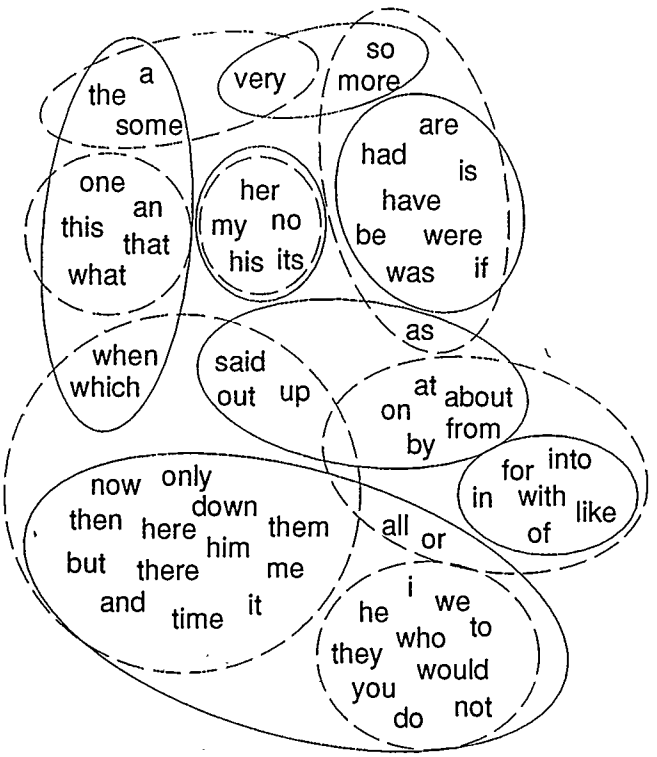

Figure 5.1: Clusters for Hardy (solid lines) and Melville (dashed lines)

diagram shown in Figure 5.1. Sadly, in the case of *Moby Dick*, "a" and "an" are still in separate categories which may indicate a weakness in the general approach. However, as can be seen with the other clusters in the figure, this problem does not emerge when first-order successors are drawn from *Far From The Madding Crowd.*

5.3.2 Incorporating n-order successors

To achieve satisfactory results from any text, the algorithm might be improved by including broader contextual information into the strength and distance formulae. For example, the phonetic peculiarity that dissociates "a" and "an" does not influ-

ence their second-order successors to the same extent. It follows that the strength of relationship between two function words could be calculated using a formula that incorporates the probabilities for intersection sizes of second-order successors, third-order successors, and so on. Experimentation with n-order successors was, in fact, attempted during the design stage of the system, but some difficulty was encountered in trying to establish an appropriate weighting scheme for each probability. Moreover, as n increases, the intersection size of the n-order successors very quickly approaches the expected size associated with random sampling. For want of an appropriate schema for including broader contextual information, the first-order successor comparison was adopted in its simplest form.

5.3.3 Robustness

Despite any shortcomings the clustering technique demonstrates, the uniformity of the final clusters for *Far From The Madding Crowd* and *Moby Dick* suggests some merit in the approach. The robustness of the technique, however, is a separate issue from its validity. If the algorithm genuinely captures the strength of relationship between two function words, we might expect to obtain similar clusters from a variety of initial groupings.

To test this, function words are assigned randomly to initial categories and the clustering process is repeated. The initial groupings and final categories from this experiment are listed in Table 5.4, and the corresponding terminal clusters are illustrated in Figure 5.2. Although different in detail from the clusters in Figure 5.1, these also reflect functional similarities between closed-class words. Further testing is needed to fully determine the robustness of the clustering algorithm; however, several

initial categories				final categories			
and	only			on	out	up	
are	here	for		more	so	very	
we	you	then	so	are	be	had	have
if				if	is	was	were
had	have	him	me	as	but	do	here
on	would	one	about	him	me	now	or
very				said	time	when	
who	said	down	its	an	one	some	that
her				this	what		
be	it	he	now	a	her	his	its
in	of	out	do	my	no	the	
or	all						
were	which	at	from	not	only	to	would
there	his	my	them	and	he	I	it
that	this	not	into	them	then	there	they
as	to	time	what	we	which	who	you
more	some						
is	was	I	they	about	at	by	for
but	a	the	when	from	in	into	like
up	like			of	with		
with	by	no		all	down		

Table 5.4: Clustering from random categories.

other random groupings were tried without significant variation in the results.

5.4 Categorizing open-class words

The sheer size of the open class makes it difficult to incorporate statistical properties for each pair of words as the basis for their categorization. However, barring any sort of semantic judgment about such words, some statistical analyses of contextual information must be used. By adhering to the principle that a syntactic role can be associated with a particular place in a particular syntactic structure, approximate

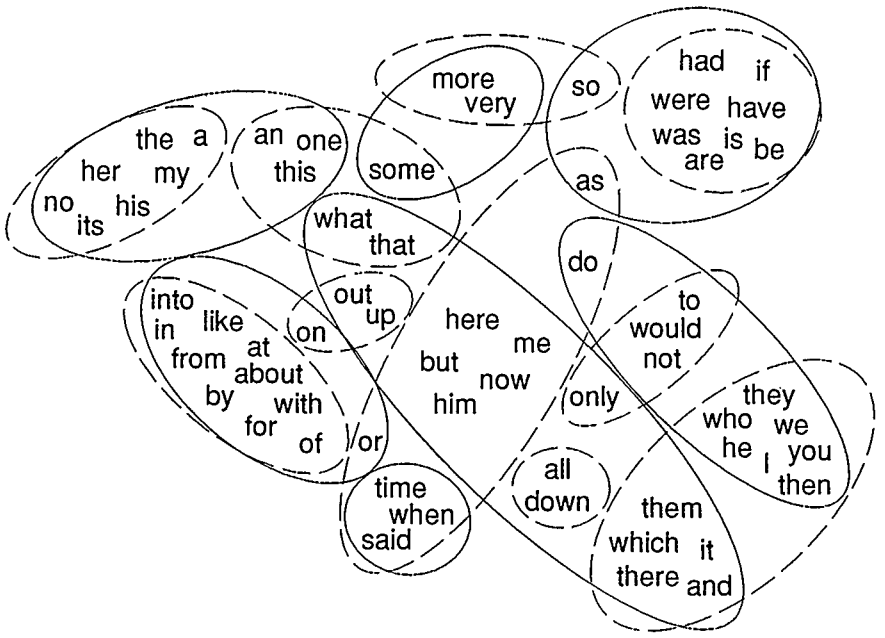

Figure 5.2: Clusters derived for Hardy and Melville from random initial groups

groupings are created by defining each category in terms of the discernable attributes of such places. The attributes used are: the functional category heading the phrase in which an open-class word is used, the functional category heading the phrase that follows, the length of the phrase in terms of the number of open-class words in it, and the relative position of the word within the phrase.

For example, based on the functional categories listed in Table 3.6, Table 5.5 lists a portion of the words in *Far From The Madding Crowd* that appear in the first position of a phrase with length one, where that phrase is headed by a fw_0 category and followed by a phrase headed by a fw_7 category. Though not without some discrepancies, the table shows a fairly consistent set of nouns. In fact, many initial groupings indicate a notable similarity with standard grammatical categories,

$cw(fw_0, fw_7, 1, 1)$ – 1486 members					
abruptness	absence	absurdity	accident	accidents	account
accuracy	act	action	adaptation	adieu	adjustment
admission	advance	advice	affection	affections	age
agent	agreement	aid	aim	air	aisle
ally	alternatives	altitudes	anger	angle	angles
anguish	angularity	ankles	antagonist	anticipations	anxiety
anzieties	apex	appeal	appearance	appeared	appointment
apprehensions	approach	arc	arch	archway	argument
arise	arm	arms	arrangement	arrival	article
articles	ascendant	ascent	ashes	...	

Table 5.5: Partial list of an initial content-word category from Hardy

and it is tempting to use them *as is* for the syntax induction process. However, the contextual information employed tends to yield a very large number of open-class categories—slightly more than 90,000 in the case of Hardy's novel. To reduce this number, the decision was made to unify categories demonstrating marked similarity.

Once again, similarity was determined according to the probability that two categories would have the intersection size they do as the result of chance. For example, the content word category $cw(fw_3, fw_5, 2, 2)$ also consists largely of nouns, as do a few other initial groupings. Consequently, the sizes of the intersections taken for pairs of these categories are improbably large. Categories with the most improbable similarity are collapsed together into a single content-word category. The number of categories that result from this technique is well under a hundred. Unfortunately, many of their similarities to standard categories are lost due to dilution.

5.4.1 Dilution

The fact that three of the 61 final categories derived from *Far From The Madding Crowd* consist of over 3000 words each while the remaining 58 categories contain fewer than 170 words reflects a dilution effect in the amalgamation process. Initial categories that do not closely resemble any particular standard grammatical category can demonstrate a close relationship with a variety of different groupings. The distinguishing characteristics of many sets are lost when they are drawn into these melting pot categories. A possible solution to this problem might be to improve the quality of such nondescript groupings.

5.4.2 Improving initial groupings

Though some of the initial categories derived from the sample text demonstrate a certain similarity with classical categories, it should be made clear that many are riddled with disagreeable exceptions. For example, Table 5.6 lists the open-class category $cw(fw_0, fw_7, 4, 4)$—a veritable grab-bag of nouns, verbs, adverbs, and other syntactic categories. It appears that, in general, the more distant the proximity of a structural position to its functional head, the more poorly the resulting group conforms to a single standard category.

Viewed another way, shorter fw-phrases yield stronger initial categories. Consider the words "just", "again", and "am" included in Table 5.6. Our intuitions about functional elements suggest that these words might be more appropriately placed into a closed-class category. In some instances, a larger set of function words would have the effect of shortening fw-phrases, thus possibly improving the initial content word categories by strengthening proximity relations.

$cw(fw_0, fw_7, 4, 4)$ – 123 members					
adding	again	am	arose	around	bathsheba
beat	bosom	bound	breakfasting	bustling	came
caused	ceased	close	coloured	connected	consisting
crawled	crocketed	curling	day	distress	downs
downstairs	dressing	emphatically	entreaty	everdene	expected
face	far	fed	feller	fitted	fixed
forth	getting	gods	goings	hanging	heaps
heedless	idiotically	idly	images	indulged	instance
inversion	jack	just	keenly	...	

Table 5.6: Partial list of an initial content-word category from Hardy

5.4.3 Tense, inflection and number

Manual inspection of the final categories reveals a number of sets comprised of words that are inflectional forms of those in other sets—forms which would not be separated under standard classical criteria. For example, past tense and active verbs are separated from their infinitive forms; and forms for irregular verbs are often scattered hither and thither. Many plural nouns are sequestered away from their singular counterparts into a category of their own, due presumably to problems stemming from number agreement. A process of affix stripping could be undertaken prior to the generalization procedure to remedy this discrepancy. However, it is not at all clear whether the resulting assimilation is actually desirable for the purposes of inferring syntactic descriptions. Without any transformational account of language surface structures, attributes that stem from things like tense and number *will* produce an effect on the sentence form—an effect that should be preserved in the open-class categories.

Further experimentation is required to identify precisely which aspects of the

category definition and assimilation process are responsible for the anomalous results, and what steps should be taken to correct them. Until then, some comfort may be taken from the fact that there is no single computational or theoretical account of language that does not succumb to the odd exception.

5.5 Inferring the function word grammar

The decision to seek generalizations of infra-phrase patterns first and broader sentence forms afterward is based on the widely held perception of language as a hierarchical structure. The decision to delimit phrase segments at each functional element is based on a broad interpretation of the structural heads incorporated into DP-Theory. A defense for these decisions has already been profferred in Chapter 2. Their success is the subject of this section.

5.5.1 Oversights in infra-phrase generalization

Assignment of unique content word strings to super-symbols is a trivial process—as is determining which shorter strings are expressed a substrings of longer ones. However, another kind of substring pattern is entirely overlooked by the infra-phrase generalization process—the substring patterns not expressed as stand-alone sequences.

Consider the following pair of production rules.

$$Cp_{25} \Rightarrow Cp_{12}\ Cp_{63}\ Cp_{20}$$
$$Cp_{44} \Rightarrow Cp_{12}\ Cp_{63}\ Cp_{33}$$

The combination of symbols $Cp_{12}\ Cp_{63}$ occurs in both rules. Each combination must translate to the same sequence of content word category symbols—a non-unique

sequence of greater length—and thus could be rewritten as an independent content-phrase rule. This does not happen because the sequence never appears in isolation between two functional category symbols.

To generalize these patterns, every possible substring of every content word sequence must be extracted and checked to see whether it is expressed as a substring of another sequence. That is, given a string of content word symbols with length k, for some $k > 2$, each of its $k - 1$ length substrings is compared against longer content word sequences to determne whether it is a repeating pattern. If it is, a new rule is written for it and its corresponding symbol is substituted into the appropriate longer sequences. The process is iterated for $k - 2$ length substrings, $k - 3$ length substrings, and so on, for all $k - n$ length substrings, where $k - n > 2$, of every content word sequence.

This process was included in an earlier version of the infra-phrase generalization component, but was abandoned for two reasons. First, the algorithm entails a tremendous number of comparisons—a computational complexity that borders on intractability. Second, the patterns generalized by the process have no relation to the functional elements and, therefore, say nothing about the principle being investigated.

5.5.2 Principled inference

In Section 4.1, two assumptions are made upon which the grammar induction process is based: 1) that the sample text consists entirely of well-formed expressions, and 2) that the syntactic category of every functional element is unambiguous. The first of these is taken to be self-evident, while the second is not. Though determiners always

function as such, prepositions may also function as verb particles—a subtly different grammatical role.

The need to assume unambiguous functions for closed-class elements is necessary in order to fix grammatical points within sentences structures. Certainly all inferred categories could be defined using proximity relations with respect to sentence length alone. This, however, would reduce the syntax induction process to unprincipled pattern generalization. It appears, then, that an evaluation of the syntactic component described in this thesis must include an evaluation of the principles it is trying to uphold.

There are essentially two tenets that form the basis of the grammar induction—one linguistic, and the other computational. They are the *principle of compositionality* described in Section 2.1.2, and the *notion of approachability* described in Section 2.3.6. Though neither is fully realized by the final grammar, some level of conformity is achieved.

5.5.3 The principle of compositionality

The principle of compositionality requires that phrase boundaries be established in a meaningful way. That is, established phrase structures must reflect an appropriate level of psychological unity—unity that can be partially verified using the constituency test outlined in Section 2.1.3.

The following production rules are taken from the grammar derived from *Alice in Wonderland.*

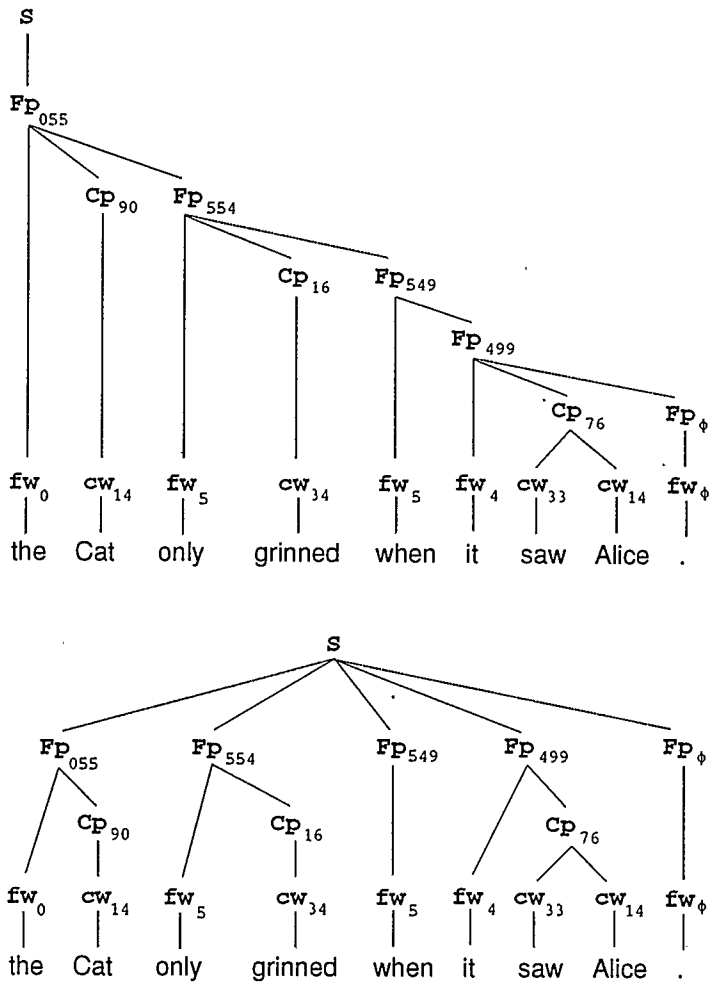

Figure 5.3: Two phrase structure trees for a sentence from *Alice in Wonderland*

$$S \Rightarrow Fp_{055}$$
$$Fp_{055} \Rightarrow fw_0\ Cp_{90}\ Fp_{554}$$
$$Fp_{554} \Rightarrow fw_5\ Cp_{76}\ Fp_{549}$$
$$Fp_{549} \Rightarrow fw_5\ Fp_{499}$$
$$Fp_{499} \Rightarrow fw_4\ Cp_{16}\ Fp_{\phi}$$

These rules account for the fw-phrase structure of the sentence "The Cat only grinned when it saw Alice", and their corresponding tree structure is shown in the upper half of Figure 5.3. This left-associative representation can also be interpreted as a

sequence of fw-phrases, giving the equivalent *flat* tree structure shown in the lower half of the figure.

Substitutions are possible for some constituents of the lower tree. For example, the word "it" can substitute for the node Fp_{055} to produce "*it* only grinned when it saw Alice"; and, if we ask "What only grinned when it saw Alice?", the answer "The Cat *did* when it saw Alice" indicates a genuine bond within the substructure Fp_{554}.

The principle of compositionality also predicts a substitution that will produce "The Cat only grinned *then*". Unfortunately, no single node of the lower tree in Figure 5.3 can be replaced with "then". Such a substitution is, however, possible for the node Fp_{549} in the upper tree structure. In other words, though neither tree supports the principle of compositionality completely, every substitution predicted by the principle can be satisfied by some aspect of one tree or the other. While we could produce some combination of the two representations that would account for all possible substitutions, such a measure is, at present, indefensible.

5.5.4 Approachability conditions

Feldman's [24] notion of approachability entails the following conditions:

> a) For any $y \in L(G)$ there is a time τ such that $t \succ \tau$ implies $y \in L(G)$.
> b) For any H such that $L(H) - L(G) = \phi$ there is a time τ such that $t \succ \tau$ implies $A_t \neq H$.

The induction process described in this thesis satisfies the first of these by limiting $L(G)$ to the set of expressions comprising the sample text. More plainly, $L(G)$ is fixed to a given set of expressions such that there is no expression $y \in L(G)$ that is not seen by the inferencing mechanism.

The second condition is partially satisfied by similarly fixing the maximum value of t, where $t \succ \tau$. That is, because the presentation set is finite, the hypothesized grammar will reach a point when it need no longer change.

We can be sure that H can generate at least $L(G)$. However, there is no *np*-complete method by which we can test the condition that $L(H) - L(G) = \phi$ since H may be an over-generalization of G. In practice, we can quite easily show that H is an over-generalization of G by simply using H to generate an expression that does not appear in the source text—a result that is, in fact, difficult to avoid. We could relax the limitation that $L(G)$ is equal to precisely the expressions of the source text, and thereafter postulate that any expression generatable by H is also generatable by G. However, this would prohibit the fixing of t and thus prevent satisfaction of the first condition.

5.6 Applications

Possible applications for any language processing system are many and varied. Grammars produced from syntax induction are inherently generative to the extent that they can be used to reproduce *at least* the set of expressions from which the rules were inferred. This has practical implications for day-to-day computing with improved data compression techniques, and more esoteric applications in computer generation of prose and poetry. This kind of grammatical analysis may provide a new tool for attacking authorship puzzles for anonymous texts, and the use of function word grammars for semantic-free language processors may have prospects in artificial intelligence.

processing stage	number of rules	symbols/rule	grammar size
original text	7281	19.31	140,632
phrase grammar	8801	4.46	39,285
final grammar	6328	5.25	33,212

Table 5.7: Stages of grammar reduction for *Far From the Madding Crowd*

5.6.1 Text compression

The substitution and decomposition procedures uncover a tremendous amount of similarity within the expressions of a text. These similarities reflect general syntactic structures characterized as a context-free grammar. If we express the original text of *Far From the Madding Crowd* as a grammar such that each sentence is equated with a production rule, then the entire text requires 7281 rules to describe its 7282 sentences ("I must go." is the only duplicate sentence), with each rule averaging 19.31 symbols (i.e. words) in length. The same text can be expressed by 8801 fw-phrase structures with an average length of 4.46 symbols, and although the disjunctive representation of the final grammar increases the average length of each rule, nearly 2500 rules are eliminated. The number of rules and symbols per rule in the various grammars is summarized in Table 5.7. The total size of the grammar in symbols is the product of these two quantities. It seems likely that the generalizations captured by these grammars can be used to compress the text through standard encoding techniques [7], and this possibility is presently being investigated.

5.6.2 Text generation

There has been much interest over the years in the "creative computer," using programs to create prose, poetry, and other forms of literature [53, 46]. One of the key problems in this area is the immense amount of labor required to develop a system to

An soothingly were perceived miss laid of the hour. It hope what which have brought of accident. And gloves to such stream and in the herself and the board inexpressibly stirred of two and inflamed any liddy. He reach window of such juno. I has good the people plainly cajolery for mossy the little whistling to crack about frankly and tarried of a with his christmas ingenuity you must keep to the multiplying no her dark try know the omen with the running rest to oldest girls on some enough to one tartly off all but it health in he leafless on he revealed shivering in age evil and meeting to of a matter not to not. As stream at coggan and a winter in the boys. From at his high two fog water.

Table 5.8: Text generated randomly from the grammar for *Far From the Madding Crowd*

create text in a particular genre. The ability to infer a grammar from a given text and then use it for generation opens up new possibilities for the automatic writing of text within a particular genre. Table 5.8 shows a sample of text generated randomly from the grammar inferred from *Far From the Madding Crowd*. The quality of this extract is somewhat disappointing, and tends to imply that the system in its present state has not been successful in capturing the essence of Hardy's grammar. In all fairness, however, it is also characteristic of the text generated from compression schemes in general [60]. In any event, studying the shortcomings of randomly-generated text is an excellent device for focussing attention on the quality of the grammar that is inferred.

5.6.3 Authorship analysis

Statistical techniques have often been employed to identify authors of anonymous texts, or to challenge authorship claims [14, 22]. O'Donnell [48] outlines statistical analysis of sentence length, vocabulary size, distribution of sentence complexity, and other "stylistic variables" to evaluate the proposal that Thackeray and Dickens were

one and the same author, and similarly for Shakespeare and Marlowe. Grammatical inference allows such analyses to examine the more microscopic details of sentence structure.

Recently, law enforcement agencies have begun to use statement analysis as a field tool for interrogation [16]. The technique statically examines the use of determiners, connectives, tense, and possessive pronouns to evaluate the sincerity of witness statements and to provide indications for further questioning. The method is based upon conjectures of an indissoluble relationship between language and thought. Because statement analysis focuses primarily on functional elements, closed-class inferencing might be an effective basis for automating the process.

5.6.4 Functional language processing

A computational account of language that focusses on functional elements is not a novel idea. Dewar *et al.* [21] describe a system that isolates syntactic components using grammatical information about a limited number of words: prepositions, articles, auxiliaries, conjunctions and pronouns.

As noted in Section 2.1.6, linguistics literature regards the class of functional elements to include both function words and inflectional morphemes. The system described in [21] also includes a number of inflectional morphemes (e.g. *-s, -ed, -ing*) that were considered syntactically important. The program uses this dictionary of functional elements to identify syntactic relations, such as the subject, object, and indirect object of input expressions. It identifies semantic heads, and can even parse some inverted sentence forms. Ultimately, the system identifies each expression as declarative, imperative, interrogative, or indirect.

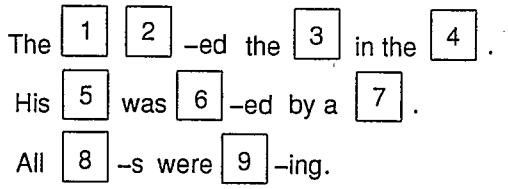

Figure 5.4: Functional element phrase structures

A possible extension to a functional parser would be to use the inferencing mechanism we have described as part of a semantic-free question-answer system. The induction would create syntactic templates along the lines of those implied by Garrett's positional model of sentence production discussed in Section 2.1.7. A few possible templates are shown in Figure 5.4, where the numbered boxes map onto appropriate semantic elements isolated from the source text. The question-answering component would accept a query of the form "Where was the 3 2-ed?", where 3 and 2 are content words present in the original text. The functional structure derived from the generalized text would permit a response to the query in the form "In the 4." without a need for assistance from any sort of semantic structures. Implementation of such a system would, of course, require a sound method for affix stripping.

5.7 A final word

It is difficult to assess the precise strengths and/or benefits of a functional approach to language inference. This chapter has outlined many ways in which the results of the process conform to our intuitive and theoretical expectations of language structure, and the list of possible applications is certainly longer than the one presented here.

In defending the precepts that inspired this research, the thesis may often leave the reader with the impression that it is somehow trying to redefine our notions about

natural language structure. This is certainly not its intention. However, the thesis *is* an attempt to provide computational perspectives on some aspects of theoretical and psycholinguistic research that *are* changing our perspectives on the underlying principles of language. The extent to which this thesis is successful in achieving its goal is ultimately left for the reader to decide.

Chapter 6

Summary, Future Developments and Conclusions

This chapter provides a summary of the thesis in terms of the aim and objectives outlined in the beginning. Possible extensions and future developments to the function word inferencing system are included, along with some concluding remarks.

6.1 Summary of Chapters: meeting the objectives

The aim of this thesis is to determine the extent to which a closed-class vocabulary can be used to infer lexical and syntactic information from machine readable texts. Ten objectives were identified in Section 1.3, and it is worthwhile now to provide a summary of the chapters as a review of how these objectives have been met.

6.1.1 Analysis of syntactic structure

Chapter 2 provides an overview of the three general requirements for any inferencing system, and relates these requirements to the problem of automatic grammar induction. The implementation of such a system is accomplished by 1) identifying what the system is attempting to reason about, 2) establishing the initial information it requires, and 3) choosing an appropriate algorithm by which the induction can take place.

The goal of syntax induction is to discover fundamental syntactic structures and develop a formalism that will capture their basic characteristics. Before this can be

accomplished, it is essential that we first establish a suitable framework and notation by which these structures can be expressed. The most transparent format for describing sentence structure is to equate grammaticality with expressions that are assumed to be well-formed—an assumption based on grammaticality judgements of the speakers of a language. This format allows us to establish a top level description for syntax simply by defining well-formedness as whatever structure is present in the expressions of a text taken to be grammatical. That is, we equate a grammar to the expressions of a text, and then seek to make generalizations about regularities present within those expressions. The effect is to place natural language into the class of context free grammars (CFGs)—a placement that allows us to use a well-defined metalanguage for the notation of our formalism: Backus-Naur Form (BNF).

For any given set of expressions there are an infinite number of CFGs that will describe their structure. Rather than allow the inference engine to settle on an arbitrary grammar, we look to linguistic theory to help us build in some constraints that will guide the mechanism towards a more psychologically plausible account of language. The most important of these constraints is the principle of compositionality.

We are led to believe, through such evidence as that provided by the constituency test, that there are genuine psychological bonds between the words of an expression. Moreover, these bonds exist at various levels of abstraction over the structure of the expression—each level expressing a different degree of cohesion. The first level of sentence decomposition that follows from this observation is the dissection of a sentence into constituent phrases. The individual structures of these phrases tend to demonstrate a strong cross-category regularity—where each phrase is comprised of a

head element and a complement sub-expression. In part, awareness of this regularity led to the widely accepted X-Bar theory of linguistic structure.

X-Bar theory maintains that individual phrase structures are restricted according to features associated with their respective head elements. In general, empirical studies of X-Bar determine headship according to characteristics of highly semantic elements within a phrase—nouns, verbs, adjectives, and so on—while regarding lesser meaning word categories like determiners, prepositions, and conjunctions as subordinate. But evidence from psycholinguistic research indicates that these so-called minor lexical items offer a significant contribution towards establishing syntactic structure. Such significance, in fact, that they have come to be known as function words.

6.1.2 The closed class: function words

Function words are characterized according to many peculiar properties. They tend not to enter freely into word-formation processes, nor are they subject to much intonational stress. They show up late in the demonstrative vocabulary of children learning to speak their first language; and they are often the first vocabulary items lost in agrammatism. The most conspicuous of their traits, however, is that they demonstrate very high frequency of use in common parlance, and they constitute a closed class.

Each of their peculiarities contributes to the suggestion that function words may possibly be subject to different cognitive processes than are highly semantic items. Some slips-of-the-tongue by non-aphasics, for instance, display evidence that they are positioned into syntactic structures before any major lexical items are selected

from the mental lexicon. Word exchanges like *he is planting the garden in the flowers*, and "stranding errors" like *he is schooling to go* were amongst the corpus of speech errors Garrett used to develop a model of sentence production wherein syntactic superstructures are expressed exclusively by function words (and inflectional morphemes) with places left for the subsequent positioning of content words.

The role of function words in sentence structure has become the cornerstone of modern linguistic theories regarding syntax. For example, DP-Theory is an attempt to reconcile an apparent discrepancy within X-Bar, that the noun phrase is the only major base structure that is not head-initial. To compensate for this, DP-Theory maintains that the determiner is the structural head within noun phrases. DP-Theory similarly proposes headship for pronouns, modals, and some inflectional morphemes.

The importance of function words in syntactic structure, as revealed in both psycholinguistic research and contemporary linguistic theory, is used to justify their use as the base case information in a grammatical inference mechanism.

6.1.3 Grammatical inference

The final stage of preparation for implementing a system for automatic grammar induction is the selection of an appropriate algorithm. Inference algorithms can be grouped into several broad classes: 1) enumerative methods, which are computationally intractable, 2) oracular methods, similar to enumerative techniques except for the use of a teacher to guide the process away from incorrect hypotheses, and 3) constructive methods, which seek to create characterizations for regularities within sample expressions rather than search through candidate grammars.

Syntactic regularity is expressed in terms of lexical categories, rather than individual words. Before the actual inference of syntactic structure can be undertaken, therefore, a method for placing words into appropriate categories must be developed. Grammatical inference begins with the inference of word classes.

6.1.4 Inferring the closed class

Chapter 3 describes the first step towards inferring a grammar: the creation of lexical categories to which the words of a text can be associated. The entire induction procedure is based on the importance of closed class elements, and the discovery of this class is the first task to be addressed.

Of all the characteristics attributed to closed-class elements, the most conspicuous is their high frequency of use—a property which further suggests their importance in language structure. Frequency lists are compiled for the vocabularies of a number of texts, and their intersection is taken to remove lexical items that may be peculiar to any one text.

6.1.5 Categorizing closed-class words

Once the closed-class has been established, its members are further generalized according to usage similarities. To determine whether two function words demonstrate similar usage, the first-order successors for each are collected. Under the assumption of random sampling, probabilities are calculated for the intersection sizes of each pair of first-order successors. The degree to which two function words are related is measured as the inverse of the probability that their first-order successors have an intersection size as large as they do. Each closed-class element is tentatively assigned

to the same functional category as its closest relative. Once initial categories have been established, they are reassessed to determine whether each function word is in its best category. Where it is not, it is reassigned. The final groupings show a marked similarity to the standard grammatical categories associated with functional elements—i.e. determiners, prepositions, pronouns, and so on.

6.1.6 Inferring open-class categories

Words not admitted to the closed-class are, by default, open-class lexemes. Given the large number of open-class words, initial categories must be established using a much simpler method than comparison of their first-order succesors. Phrase structure boundaries are set at each functional element, and tentative open-class categories are defined according to the position of content words with respect to phrase boundaries. For example, one open-class category might consist of all content words that appear in the second position of a phrase with length three, where the phrase is bounded on the left by a closed-class element of a particular type, and is bounded on the right by another closed-class category.

Initial open-class categories tend to reflect the standard grammatical categories of semantic elements such as nouns, verbs, and adjectives; though some separation does result that reflects differences in number and tense. An extremely large number of initial categories is produced by this process, largely because of the way these categories are define. In the hope of obtaining a better generalization of content word categories, initial groupings are subjected to an amalgamation process.

6.1.7 Amalgamating open-class categories

Once all content words have been assigned to initial categories, probabilities are calculated for the intersection sizes of every pair of categories. Each category is collapsed into a single grouping with the category to which it is most strongly related, where strength is measured as the inverse of the probability that their intersection is as large as it is.

Some unified categories manage to maintain a certain level of similarity to standard grammatical categories. However, weak initial categories tend to show a strong relationship with a variety of other groupings. As the weaker groupings are combined those that are more sound, the distinguishing charactersitics of intial categories become increasingly diluted.

6.1.8 Infra-phrase generalization

Once all words have been assigned to lexical categories, each sentence in the sample text is replaced with its corresponding sequence of category symbols. The sequences are dissected into phrases at each functional element. The left most function word symbol is regarded as the head of the phrase. The right most symbol is maintained to preserve linking information for subsequent phrase generalization. Unlike the head element, the terminal symbol is a copy of the head of the next phrase and does not represent a substituted word.

Phrases are defined according to their head element, their length, and the head of the phrase that follows it. The contiguous sequence of content word symbols within each phrase is replaced with a super-symbol and the substitution is recorded as a production rule. Content word sequences are sorted by length and checked

against longer sequences to see if they are expressed as a substring. If they are, their corresponding super-symbol is substituted into the longer sequence. The result of this process is a context-free grammar for the phrase structures.

6.1.9 Generalization of phrases

To limit the ways in which phrase rules can be combined, they are generalized into classes according to their combinatorial properties. Phrases of the same type are identified by overlaying their corresponding digraphs. Every sublattice of the larger digraph depicts a class of phrases if its greatest lower bound is a content phrase symbol. The sublattices are rewritten as disjunctive production rules, and the combined set of rules for content word sequences and phrase types are output as a context-free grammar for the complete text.

6.1.10 The success criteria

The success of the lexical component is determined according to how well the resulting categories reflect standard grammatical categories. An intuitive assessment of the functional categories indicates fairly sound divisions between prepositions, determiners, possessive pronouns and auxiliary verbs. Similarly, content word categories do, to some degree, preserve distinctions between nouns, verbs, adverbs, and so on. However, in both cases, large nondescript initial groupings produce significant discrepancies by maintaining a strong relationship with a wide of variety of other categories. When initial groupings are reassessed, the weaknesses of nondescript categories propagate into the stronger ones.

The success of the syntactic component is measured according to how well it con-

forms to the tenets on which it is based: the principle of compositionality and the conditions of approachability. Some substructures of the final grammar do satisfy the basic requirements of the constituency test described in Section 2.1.3.. That is, tree structures corresponding to the grammar will support some of the substitutions predicted by the principle of compositionality. To support the remaining predictions, however, the grammar must be translated into a "flatter" phrase sequence description, as outlined in Section 5.5.3.

Similarly, Section 5.5.4 describes how only one condition of approachability can be satisfied at any given time. By fixing $L(G)$ to the set of sample expressions, the hypothesized grammar will reach a point where it no longer changes. However, if the hypothesized grammar is an over-generalization of G, then the condition that $L(H) - L(G) = \phi$ requires that we allow $L(G)$ to include all expressions generatable by the hypothesized grammar. Thus, $L(G)$ cannot be fixed to the presentation sequence.

The extent to which the inferred grammar captures principles of DP-Theory would require a thorough account of such things as licensing and projection, both in terms of DP-Theory and in terms of the inferred grammar. Such an analysis is well beyond the scope of this discussion. We can, however, detect some degree of similarity between the substructures of a function word grammar and the precepts of DP-Theory, if only at an intuitive level. Favorable comparisons may also be made between the function word grammar and Garrett's positional level model of sentence production, provided that we are not too circumspect.

6.1.11 Applications to language processing tasks

Closed-class inferencing has a variety of possible applications in practical language processing tasks. For example, it could introduce a whole new set of "stylistic variables" in the field of authorship analysis that was presented in Section 5.6.3, or provide a general framework for a semantic free question-answering system. Some effort is already being extended towards incorporating the function word induction process into the areas of text compression and generation. The final evaluation may have to wait until the jury is in.

6.2 Future developments

The results obtained from the various components of the induction system described in this thesis suggest several avenues of improvement. Further, the possible application of function word grammars to practical language processing tasks such as those outlined in Chapter 5 suggest further research based on closed-class inferencing. Future improvements and extensions to the system are the subject of this section.

6.2.1 Morphological analysis

Both DP-Theory and Garrett's positional model of sentence production suggest that inflectional morphemes must also be considered functional items. Many of the system's shortcomings also seem to stem from its failure to incorporate these morphemes into the closed-class. The addition of a morphological component that could identify inflectional affixes would allow such morphemes to be analyzed in the same manner as the free-standing functional elements.

Adding inflectional morphemes to the list of candidate functional elements could increase the size of the closed-class and shorten many of the content word sequences. Section 5.4.2 argues that this may strengthen the similarities between derived content word groupings and standard syntactic categories. The difficulty associated with affix stripping resulted in its omission from the current system. But its potential contribution to the soundness of the overall approach merits its serious consideration as a future development.

6.2.2 Parser

Application of function word grammars to text compression, statement analysis, question-answering systems, and so on, will require that an appropriate parsing mechanism be available. In fact, parsers are usually standard stock in grammatical inferencing mechanisms, and the addition seems a natural extension to the system described in this thesis. Furthermore, as an inductive method, function word inferencing can be evaluated using many of the standard metrics outlined in Section 2.3. But judging the quality of the grammar inferred is not at all straightforward. A function word parser would seem a reasonable first step in achieving such analyses.

6.3 Conclusion

We began by noting a paradox about language that has been both spur and deterrent to linguistic research—How is it that we seem to know so very little about something that we do so well? This seems a particularly poignant question when we consider the act of inferring a grammar. It is apparently so trivial that even a child can do

it. What is so frustrating to linguistic researchers is the apparent corollary that in fact *only* a child can do it. This should not be terribly surprising however, for those in the field of artifical intelligence have long been aware of the ease with which good chess playing programs can be written when compared with the difficulty of getting a robot to play hopscotch.

There is something about the simplicity of function words that makes us feel we ought to be able to figure out their precise purpose. It seems clear from the evidence of psycholinguistic research that functions words do have a special role in the cognitive aspects of language production and comprehension. It also seems clear from their statistical presence that their importance in abstract communication is substantial—even fundamental. Nevertheless, their precise contribution to language seems likely to remain cloaked in mystery for some time, and writers on the subject will have to continue to wave their hands in speculation.

Though there were times during the course of this research when it seemed that tremendously significant insights were beginning to emerge, there were many others when the words "barking up the wrong tree" seemed quite pithy. Such is the nature of an exploratory study. Indeed, the project should be looked upon as basic research if any of its results are to be considered useful. The aim of the thesis was to determine the extent to which the notion of function words could be used to infer lexical and syntactic information. The extent to which we have been successful is not entirely clear because we are not sure what we are looking at—at least we are not sure what metrics to apply when evaluating the results.

This thesis represents a novel approach to lexical and syntactic inferencing. Its techniques offer a wide range of practical applications and possible extensions. It

achieves some results that might be surprising, and others that might seem trivial. To whatever extent it is considered successful, it is hoped that at least one thing is achieved—that a little more is known about this thing we do so well.

Bibliography

[1] Steven Abney. Functional elements and licensing. presented to GLOW, Gerona, Spain, April 1986.

[2] Steven Abney. *The Noun Phrase in its Sentential Aspect.* PhD thesis, MIT, 1987. unpublished.

[3] D. Angluin. Inductive inference of formal languages from positive data. *Information Control*, 45:117–135, 1980.

[4] Emmon Bach. An extension of classical transformational grammar. In *Problems in Linguistic Metatheory, Proceedings of the 1976 Conference at Michigan State University*, pages 183–224, 1976.

[5] B. Badecker and A. Caramazza. On consideration of method and theory governing the uses of clinical categories in neurolinguistics and cognitive psychology: the case against agrammatism. *Cognition*, 20:97–125, 1985.

[6] G. E. Barton, R. C. Berwick, and E. S. Ristad. *Computational Complexity and Natural Language.* The MIT Press, Cambridge, Massachussetts, 1987.

[7] T.C. Bell, J.G. Cleary, and I.H. Witten. *Text Compression.* Prentice Hall, Englewood Cliffs, New Jersey, 1990.

[8] R. C. Berwick. *The acquisition of syntactic knowledge.* MIT Press, Cambridge, Mass, 1986.

[9] R. C. Berwick and S. Pilato. Learning syntax by automata induction. *Machine Learning*, 2(1):9–38, 1987.

[10] R. C. Berwick and A. S. Weinberg. *The Grammatical Basis of Linguistic Performance: Language Use and Acquisition.* The MIT Press, Cambridge, Massachussetts, 1984.

[11] D. Bickerton. Pidginization, creolization: language acquisition and language universals. In A. Valdman, editor, *Pidgin and Creole Linguistics*, pages 49–69. Indiana University Press, Bloomington, 1977.

[12] A. Bierman and J. A. Feldman. A survey of grammatical inference. In S. Watanabe, editor, *Frontiers of Pattern Recognition.* Academic Press, New York, 1972.

[13] Joan Bresnan. A realistic transformational grammar. In Morris Halle, Joan Bresnan, and George Miller, editors, *Linguistic theory and psychological reality*, chapter 1, pages 1–59. MIT Press, Cambridge, Mass., 1978.

[14] Claude S. Brinegar. Mark Twain and the Quintus Curtius Snodgrass letters: A statistical test of authorship. *American Statistical Association Journal*, March 1963.

[15] D. Caplan. *Neurolinguistics and Linguistic Aphasiology.* Cambridge University Press, Cambridge, 1987.

[16] Sgt. Robert Chamberlain. private communication, RCMP Serious Crimes Division, Prince George, Canada, April 1990.

[17] N. Chomsky. On certain formal properties of grammars. *Information Control*, 2:137–167, 1959.

[18] Noam Chomsky. *Lectures on Government and Binding.* Foris Publications, Dordrecht, 1981.

[19] Noam Chomsky. *Barriers.* MIT Press, Cambridge, Mass., 1986.

[20] G. Curme. *A grammar of the English language*, volume 2 of *Syntax.* Barnes and Noble, Boston, 1935.

[21] Hamish Dewar, Paul Bratley, and James Peter Thorne. A program for the syntactic analysis of English. *Communications of the ACM*, 12(8):476–479, August 1969.

[22] Alvar A. Ellegard. *A Statistical Method for Determining Authorship.* Goteborg, Holland, 1962.

[23] Ann Farmer. *Modularity in Syntax.* MIT Press, Cambridge, Mass., 1984.

[24] J. A. Feldman. Some decidability results on grammatical inference and complexity. AI Memo 93.1, Computer Science Dept., Stanford University, Stanford, California, 1970.

[25] J. A. Feldman, J. Gips, J. J. Horning, and S. Reder. Grammatical complexity and inference. Technical Report CS 125, Computer Sience Dept., Stanford University, Stanford, California, 1969.

[26] Naoki Fukui and Peggy Speas. Specifiers and projection. *MIT Working Papers in Linguistics*, 8:128–172, 1986.

[27] M. F. Garrett. Syntactic processes in sentence production. In R. Wales and E. Walker, editors, *New Approaches to Language Mechanisms.* North-Holland, Amsterdam, 1976.

[28] M.F. Garrett. The organization of processing structure for language production. In D. Caplan, A.R. Lecours, and A. Smith, editors, *Biological Perspectives on Language.* MIT Press, Cambridge, Mass., 1984.

[29] Gerald Gazdar, Ewan Klein, Geoffrey Pullam, and Ivan Sag. *Generalized Phrase Structure Grammar.* Basil Blackwell, Oxford, UK, 1985.

[30] Gerald Gazdar, Geoffrey K. Pullum, and Ivan A. Sag. Auxiliaries and related phenomena in a restrictive theory of grammar. *Language*, 58:591–638, 1982.

[31] E. M. Gold. Language identification in the limit. *Information Control*, 10:447–474, 1967.

[32] R. C. Gonzalez. *Syntactic Pattern Recognition, An Introduction.* Addison-Wesley, Reading, Massachussetts, 1978.

[33] B. Gordon and A. Caramazza. Lexical decision for open- and closed-class items: failure to replicate differential frequency sensitivity. *Brain and Language*, 15:143–60, 1982.

[34] R. Hausser. Left-associative grammar: An informal outline. *Computers and Translation*, 3(1):23–67, 1988.

[35] Teun Hoekstra. *Transitivity: Grammatical Relations in Government-Binding Theory.* Foris Publications, Dordrecht, 1984.

[36] George M. Horn. Lexical-functional grammar. In Werner Winter, editor, *Studies and Monographs 21: Trends in Linguistics.* Mouton Publishers, Berlin, 1983.

[37] J. J. Horning. *A study of grammatical inference.* PhD thesis, Computer Science Dept., Stanford University, Stanford, California, 1969.

[38] Norbert Hornstein. S and X-Bar convention. *Linguistic Analysis*, 3, 1977.

[39] Richard Hudson. The power of morphological rules. *Lingua*, 42:73–89, 1977.

[40] Ray S. Jackendoff. *X-bar Syntax: A Study of Phrase Structure.* MIT Press, Cambridge, Massachusetts, 1977.

[41] M. L. Kean. The linguistic interpretation of aphasic syndromes: agrammatism in Broca's aphasia, an example. *Cognition*, 5:9–46, 1977.

[42] John Lyons. *Introduction to Theoretical Linguistics.* Cambridge University Press, Cambridge, 1968.

[43] B. MacWhinney. Basic processes in syntactic acquisition. In S. A. Kuczaj, editor, *Language Development: Vol. 1, Syntax and Semantics.* Lawrence Erlbaum, Hillsdale, New Jersey, 1982.

[44] Alec Marantz. *On the Nature of Grammatical Relations.* MIT Press, Cambridge, Mass., 1984.

[45] M. Marcus. *A Theory of Syntactic Recognition for Natural Languages.* MIT Press, Cambridge, Mass., 1979.

[46] K. McKeown. Discourse strategies for generating natural language-text. *Artificial Intelligence*, 27:1–42, 1985.

[47] Richard Montague. Formal philosophy. In R. H. Thomason, editor, *Selected Papers of Richard Montague*. Yale University Press, New Haven, CT, 1974.

[48] Bernard O'Donnell. *An Analysis of Prose Style to Determine Authorship*. Mouton & Company, The Hetherlands, 1970.

[49] William O'Grady and Michael Dobrovolsky, editors. *Contemporary Linguistic Analysis*. Copp Clark Pittman Ltd., Toronto, 1987.

[50] T. W. Pao and J. W. Carr. A solution of the syntactical induction-inference problem for regular languages. *Computer Languages*, 3:53–64, 1978.

[51] Paul Postal. Limitations of phrase structure grammars. In Jerry A. Fodor and Jerrold J. Katz, editors, *The Structure of Language: Readings in the Philosophy of Language*, pages 137–51. Prentice-Hall, Englewood Cliffs, 1964.

[52] Edward Sapir. *Language*. Harcourt, Brace & World, New York, 1949.

[53] Tony C. Smith and Ian H. Witten. A planning mechanism for text generation. *Literary & Linguistic Computing*, 6(2):119–126, 1991.

[54] R. Solomonoff. A new method for discovering the grammars of phrase structure languages. *Information Processing*, pages 258–290, June 1959.

[55] Margaret Speas. *Phrase structure in natural language*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[56] Timothy Stowell. Subjects across categories. *The Linguistic Review*, 2:285–312, 1983.

[57] Frits Stuurman. *Phrase Structure in Generative Grammar.* Foris Publications, Dordrecht, 1985.

[58] F. T. Visser. *An historical syntax of the English language. Part One: Syntactical units with one verb.* Brill, Leiden, 1963.

[59] Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1):36–45, January 1965.

[60] I.H. Witten and T.C. Bell. Source models for natural language text. *International Journal Man-Machine Studies*, 32(5):545–579, May 1990.

[61] W. A. Woods. Procedural semantics for a question answering system. In *1968 AFIPS Conference Proceedings*, volume 33, pages 457–471, 1968.

[62] W. A. Woods. Transition network grammars for natural language analysis. *CACM*, 3(10):591–606, 1970.