Working Paper Series
ISSN 1170-487X

Language inference from function words

by Tony C. Smith & Ian H. Witten

Working Paper 93/3

January, 1993

Department of Computer Science
The University of Waikato
Private Bag 3105
Hamilton, New Zealand

Language Inference from Function Words

Tony C. Smith

Department of Computer Science, University of Calgary, Calgary T2N 1N4, Canada
Email: tony@cpsc.UCalgary.CA; phone: +1 (403) 220–6015; fax: +1 (403) 284–4707

Ian H. Witten

Department of Computer Science, University of Waikato, Hamilton, New Zealand
Email: ihw@waikato.ac.NZ; phone: +64 (7) 838–4246; fax: +64 (7) 838–4155

Abstract

Language surface structures demonstrate regularities that make it possible to learn a capacity for producing an infinite number of well-formed expressions. This paper 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.

January, 1993

1 Introduction

Syntactic analysis of natural language has generally focused on the structural roles fulfilled by the thematic elements of linguistic expression: agent, principal action, recipient, instrument, and so on [4, 30]. This has produced theories in which the noun and verb are the primary constituents of every utterance, and syntactic structure emerges as a projection from these major lexical categories [15, 19, 27, 24]. Language processors developed under this prescriptive tradition face two serious practical limitations: an inadequate lexicon and an incomplete structural description for the language in question [6, 8, 20]. In contrast, the present paper investigates an alternative methodology that passively infers grammatical information from positive instances of well-formed expressions.

Grammar induction has often been employed to find descriptive formulations of language structure [3, 6, 9, 10, 20, 35, 33]. Such efforts commonly adopt straightforward sequence inference techniques without regard to lexical categories or syntactic modularity. This paper discusses the development of a system that uses the simple notion of *closed-class* lexemes to infer lexical and syntactic information, including lexical categories and grammar rules, from statistical analyses of copious quantities of machine-readable text. Closed-class inferencing compares well with current linguistic theories of syntax and offers a wide range of potential applications.

1.1 Language properties

From the perspective of linguistics, a grammar is a collection of language elements at various levels of abstraction. Sounds combine 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 a hierarchical composition of individual subsystems. Accordingly, language acquisition is a multi-dimensional learning process, and many levels of linguistic analysis are necessary for a comprehensive language learning system. The complexity of the problem can be reduced by distinguishing between the "static" and "dynamic" characteristics of a language.

Dynamic language elements

Language is what people do, and in this respect can be regarded as a property of people. The phonetic peculiarities of language, including suprasegmental properties like stress and intonation, vary from person to person and from time to time at the moment of production, and thus form part of the dynamic properties of language as social action. Differences in dialect, accent, and even physical idiosyncracies in speech organs can be so marked that one could argue that no two people speak the same language. Even so, we do regard individuals with widely differing speech habits as having a common language.

Static language elements

We recognize that people with different speech habits speak the same language because the vocabularies and sentence structures they exhibit are intersecting subsets of a common grammar. These grammatical elements are regarded 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 text.

Obviously our capacity to understand language rests ultimately on our ability to identify the *meaning* of what is being said or written—thus, any attempt to acquire language without taking account of its semantics will be fundamentally inadequate. But understanding, like productive peculiarities, is what people bring to a language, and not a demonstrable property of the language *per se*. This paper asks 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)?

1.2 Requirements of grammar induction

The design of any inferencing system requires explicit, *a priori* identification of 1) the goal of the reasoning process, 2) the prior information to be used as a basis for inference, and 3) an algorithm for the induction procedure. In the case of grammatical inference, each of these is further constrained by the inherent limitations of static language analysis. We discuss the first two here; the third is the subject of the following sections.

The goal of language induction

After exposure to a finite number of well-formed expressions, children demonstrate a capacity to both understand and produce a potentially infinite collection of novel sentences. They presumably accomplish this by isolating semantic and structural generalities of the sample utterances [32]. Since semantic information cannot be gleaned from a purely static analysis, structural generalities are the target of grammar induction.

Syntax is the system of rules and categories that allows words to combine into sentences. The rules do not pertain directly to words, but rather to the lexical categories to which words are assigned. Grammar induction must, therefore, address the formation of these categories before generalizations about sentence structure can be made.

Lexical categories are defined by delicately intertwined notions of meaning and form. Nouns, for instance, are such because they are token referents to substantive, abstract, or imaginary things. But they are also nouns because they are used structurally in a uniform and fairly well understood manner. This uniformity allows them to be identified as a particular syntactic class through a crude statistical analysis of structural contexts. This process can, in principle, be applied to all lexical categories. Once the categories have been established, the way is clear to identify syntactic rules for how they combine.

The prior information

To infer lexical categories one might first divide the vocabulary of a language into two groups by some coarse method of differentiation, and thereafter continue to filter each of these groups recursively into smaller categories according to a criterion that becomes increasingly more refined. A suitable criterion is the distinction of *functionality*.

A great many languages, in particular the Indo-European languages, allow for division of their vocabulary elements into two major categories: "content" words and "function" words. Content words consist of nouns, adjectives, verbs, and so on—words whose meaning is more or less concrete

and picturable. In contrast, function words are exemplified by prepositions, articles, auxiliary verbs, pronouns, and such—words whose principal role is more syntactic than semantic. Function words serve primarily to clarify relationships between the more meaning-laden elements of linguistic expression, or to introduce certain syntactic structures like verbal complements, relative clauses and questions.

Function words demonstrate many distinctive properties. Though not entirely without meaning, their semantic contribution is generally more "abstract" and less referential than that of content words. They tend not to carry stress in everyday speech. They are often the last vocabulary elements to appear in the productive language of children learning their first language. Moreover, a particular type of aphasia known as "agrammatism" is characterized by marked difficulty in the production, comprehension, and recognition of function words.

Compared to other vocabulary items, function words demonstrate high frequency usage. They tend not to enter freely into the word formation process. That is, they resist affixation and are seldom compounded with other words to form new ones. Similarly, though new content words are added to the vocabulary of a language almost daily, the number of elements in the function word class remains fixed.

The fact that the set of function words is a "closed class" of vocabulary elements that demonstrate extremely frequent usage suggests to linguists an importance in psychological processing [12, 25, 22, 23]—an importance that further underlines the usefulness of function words as a basis for grammar induction.

2 Function words

Most linguists accept that there is a set of function words that can be characterized as "closed class." But there is no consensus on exactly which words this comprises. Since language inference is a discovery process, membership in the closed class should be based on a criterion that identifies function words by analysing their usage patterns. Of the previously mentioned characteristics of function words, relative high frequency is the only one that can be determined from a static language analysis.

2.1 Identifying function words

We define function words operationally as those that occur more frequently than a certain predetermined threshold. The value of the threshold is ultimately determined arbitrarily. However, we can draw on the literature to develop a rough guideline. Caplan [12] claims that "there are approximately 500 or so function words in English, and, of the 100 most common words in English, most are function words." The average person's everyday vocabulary consists of about 10,000 words. Thus 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 along with their declining frequency.

Table 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.

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
total	11589	total	16832

Table 1: Most frequent words in *Far From the Madding Crowd* and *Moby Dick*

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 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 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 and only about 16 of them fail any sort of intuitive test as function words.

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

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 two vocabularies before taking the top 1%. This removes lexical items peculiar to any one text and, as a consequence, moves function words that

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 2: Vocabulary distribution in *Far From the Madding Crowd*

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: The most frequent 1% of words in *Far From the Madding Crowd*

may otherwise have been overlooked higher up in the frequency ordering. Table 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.

2.2 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

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 4: The closed class, inferred from Hardy, Melville and Carroll

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.

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

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 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 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.

2.3 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 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 1 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.

Slightly different function-word classes are obtained depending on exactly how the procedure is carried out. For example, it is interesting to apply the iterative reassignment procedure starting from randomly-chosen initial categories. This generates the final categories shown in Figure 2. Although rather different in detail from Figure 1, these also reflect functional similarities between closed-class words. The language inference procedures should be robust under such variation, and we believe that they are—though this has not yet been fully tested. A further function-word categorization that was obtained is summarized in Table 6, and this is in fact the one that is used as a basis for the content-word classification described next.

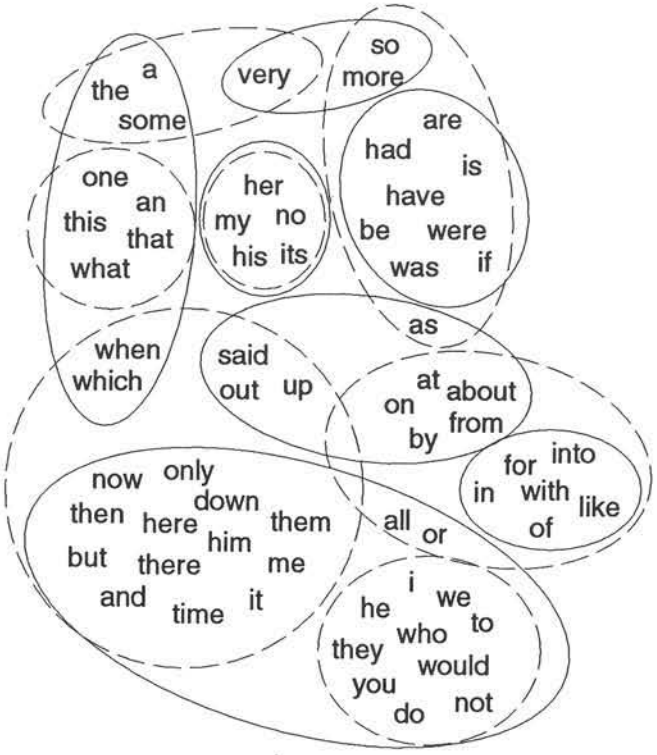

Figure 1: Categorization clusters for Hardy (solid lines) and Melville (dashed lines)

3 Categorizing open-class words

Every lexeme that does not qualify as closed class is, by default, an "open-class" word. Around 99% of the vocabulary falls into this class, and it is necessary to determine a syntactic category for each of these words.

3.1 Classical categories

In contrast to the syntactically functional roles that we have supposed are fulfilled by the closed-class words, the role of open-class words is to supply content, or meaning, to text. According to classical linguistics, the categories of open-class words correspond to general types of referent. Each content lexeme conveys 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, some meaning-laden words seem intuitively to function as referents to specific objects or object classes whose existence is real, surreal or imaginary. Others refer to qualities attributable to such objects—qualities like colour, texture, shape, and temperature. Still others refer to actions that can be perpetrated by or to objects. We have a nomenclature for such classical categories—nouns, adjectives, and verbs respectively—and their character is for the most part clear [1]. But our self-imposed restriction to a static, rather than any kind of dynamic, analysis of language

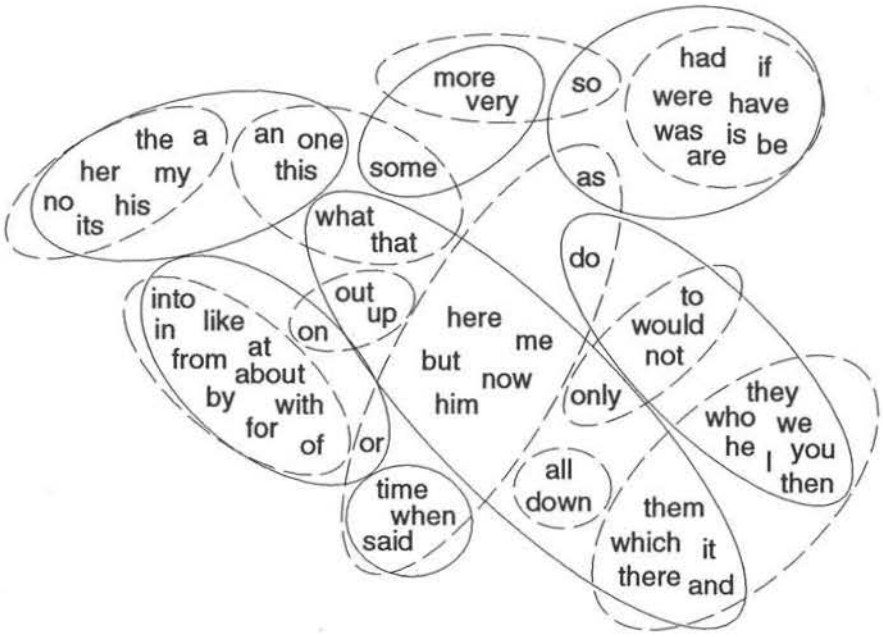

Figure 2: Clusters derived from initial random groupings

precludes access to any sort of semantic information that would help to assign open-class words to these categories.

Regularity within phrases

Ideally, inferred lexical types should correspond closely to classical categories. Consequently, the regularities used in classical syntactic analysis prove a practical guide to the development of a suitable categorization procedure. Consider the following examples of noun usage:

> *The little brown* **fox** was quite lost.
> *An old* **man** slept on *the* **sidewalk**.
> **He** left after eating *Alison's* **lobster**.
> *Many* **people** have fed *the* **bears** from **car windows**.

The noun positions (in bold) demonstrate consistent occurrence as the last word of noun phrase structures (in italics). Note further that most noun phrases begin with one of the closed-class elements from the fw_0 category of Table 6.

These examples exhibit only a weak proximity relation between an fw_0 word and the corresponding noun, because other words often intervene. However, the word positions within each noun phrase suggest that the structural roles of the words are constrained by the requirements of the phrase itself—phrases are characterized by consistent use of fw_0 words in the initial position and nouns in the final one. In order to characterize the positional roles of their constituents, a means must therefore be established to delimit phrases.

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 6: Function word categories

The function word phrase

Determiners appear exclusively in noun phrases, and this suggests a relationship between determiner and noun [1]. Moreover, whenever determiners appear they mark the onset of a noun phrase. Consequently, since the fw_0 category can be likened to determiners, fw_0 elements can be taken to indicate the onset of some kind of phrase—a function word phrase or "fw-phrase". The phrase's left boundary is the fw_0 element itself. Generalizing, *every* function word can be taken to indicate the onset of some fw-phrase type. Consequently, phrases are bounded on the right either by another function word, indicating the start of a new fw-phrase, or by the end of the linguistic expression.

Three attributes define the type of a fw-phrase: the function word category that heads it, the number of words comprising it, and the function word that follows it.

3.2 Creating content word categories

Every content word can be characterized by the ability it demonstrates to occupy certain structural positions in particular fw-phrase types. A structural role can be identified for each open-class word by noting the type of phrase in which it occurs and its position within that phrase. A categorial relationship can be inferred between a given open-class word and others demonstrating similar usage by analyzing the types of phrase it appears in, and which positions it occupies.

Initial categories

The first stage of categorization requires that each open-class word be assigned to a temporary category. This is identified by 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 fw_ϕ

(where fw_ϕ marks the end of a sentence). This allows the open-class words to be assigned to temporary categories as follows:

$cw(fw_0, fw_7, 1, 3) = \{\text{tiny}\}$
$cw(fw_0, fw_7, 2, 3) = \{\text{bird}\}$
$cw(fw_0, fw_7, 3, 3) = \{\text{sat}\}$
$cw(fw_0, fw_\phi, 1, 1) = \{\text{tree}\}.$

For example, *bird* is assigned to the set of words appearing in second position of a phrase of length 3 headed by a word from the fw_0 category and followed by a phrase headed by a word from fw_7. Similarly, *tree* is assigned to an open-class category for words appearing in the first position of a phrase of length 1 headed by fw_0 and followed by the end of a sentence (i.e. the empty phrase). As each sentence is processed, previously unseen content words are added to existing sets, or new categories are created for them. A word can be assigned to several categories, though duplicates are removed within each category.

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. Probabilities are calculated for all pairs before any are combined, and amalgamation is performed in a single pass. Once again, no provision is made to prevent a word from occupying several categories.

Table 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.

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

Inflection and agreement

The generalization procedure reduces the number of open-class word categories from 90,000 to about 60. This is still more than the dozen or so standard linguistic categories of nouns, adjectives, and the like. However, manual inspection reveals that a number of categories are comprised of inflectional forms of words in other sets, forms that would not be distinguished in a standard linguistic analysis. For instance, past tense and active verbs are separated from their infinitive counterparts, and irregular verb forms are often scattered. Many plural nouns have a category different from their singular counterparts, which is presumably due to effects of number agreement.

A process of affix stripping could be undertaken prior to the generalization procedure to remedy these discrepancies. However, it is not at all clear that the resulting assimilation is desirable for the purpose of inferring syntactic descriptions. The effects of inflection and agreement on syntactic structure are readily acknowledged by linguists, and any system that avoids transformational analysis of deep structures has to accommodate transformations at the surface level.

4 Inferring syntactic rules

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. Words combine to form sub-phrasal elements, which combine to form phrases, which combine to form sentences. Syntax induction involves compiling a formal description for linguistic structure at each of these levels.

4.1 Variable substitution

Grammar induction generally seeks to characterize not the syntactic regularities demonstrated by particular words but rather 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.

Like others (e.g. [35, 9]), we use a phrase structure grammar to express the sequential regularities that linguistic expressions exhibit. Such grammars do not distinguish between terminal and non-terminal symbols in terms of their membership of the vocabulary under consideration [24]. Any process of induction that can be applied to word sequences can just as well be applied to sequences of category symbols.

The approach we adopt is a process of variable substitution, whereby repeating patterns of category symbols are replaced with super-symbols. Such substitutions are then recorded as production rules. This process is applied iteratively to the resulting patterns of super-symbols. The final set of production rules is a context-free grammar for the sample text.

4.2 Pattern constraints

Not every pattern that the sample text exhibits is syntactically interesting. The surface form of a linguistic expression is (transformations notwithstanding) a combination of phrase segments, which are in turn combinations of words. The unity of a phrase segment stems from a genuine psychological bond constraining its composition and form—constraints different from those that bind phrases into whole expressions [30, 14, 16, 27]. The induction process should avoid forming production rules that compromise this unity by distinguishing infra-phrasal patterns, that is, regularities within phrase boundaries, from supra-phrasal ones, that is, regularities across phrase boundaries.

To support this distinction, the inference mechanism is focussed on patterns within phrase boundaries first. Once infra-phrase patterns have been assessed, attention is shifted to seek regularities in patterns expressed by the phrase segments. The resulting set of production rules reflects a modular view of syntax that more closely corresponds to general phrase structure grammars than unprincipled pattern detection techniques.

4.3 The inferencing process

We adopt the syntax induction procedure outlined in Figure 3. This is a multi-stage process which, when applied to a text, yields a context-free grammar for it. Unlike most grammatical inference methods, which form successively broader generalizations of syntax through a sequential analysis of sample expressions, variable substitution applies wholesale pattern analysis to an entire text.

The first pass grammar

The first stage of the process is to substitute the corresponding category symbol for each word in the sample expression. For example, the expression

the tiny bird sat in a hollow tree

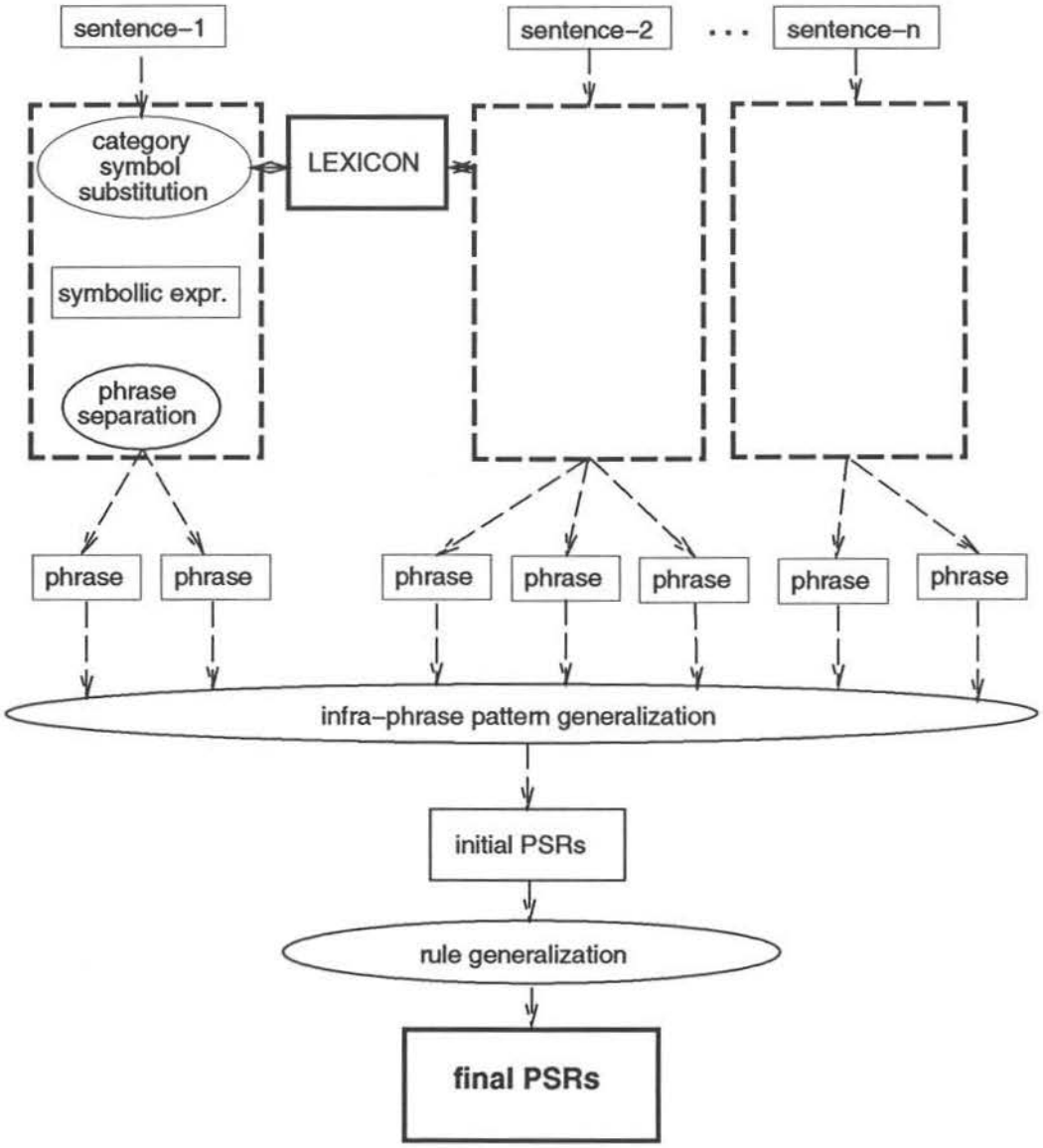

Figure 3: Overview of the induction procedure

The [1] [2] –ed the [3] in the [4] .
His [5] was [6] –ed by a [7] .
All [8] –s were [9] –ing.

Figure 4: Functional element phrase structures

the tiny bird sat in a hollow tree					
S	$\Rightarrow$	$Fp_{0,7}$			
$Fp_{0,7}$	$\Rightarrow$	fw_0	Cp_1	cw_{40}	$Fp_{7,0}$
$Fp_{0,\phi}$	$\Rightarrow$	fw_0	Cp_1	Fp_ϕ	
$Fp_{7,0}$	$\Rightarrow$	fw_7	$Fp_{0,\phi}$		
Cp_1	$\Rightarrow$	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 8: A sample context-free grammar

yields the string of category symbols

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

Symbols not only designate a particular word category, but also distinguish function from content words. In stage two, sentences are dissected into fw-phrase segments. 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 at the end of sentences.

The second pass grammar

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, is to form generalizations of symbol sequences within phrase segments. For example, the sequence

$$cw_{24}\ cw_{51}$$

is present within the first and the last of the three 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. Duplicates are rewritten as new production rules.

Content word sequences are compared against longer ones in case they form a substring of another rule. If so, a new rule is created for the substring and its symbol is substituted into the

processing stage	number of rules	symbols/rule	grammar size
original text	7281	19.31	140,632
first pass	8801	4.46	39,285
second pass	10455	2.96	30,947

Table 9: Stages of grammar reduction for *Far From the Madding Crowd*

longer sequence. Comparison continues for shorter and shorter sequences down to substrings of length two. Finally, the fw-phrase rules are presented in Bachus-Naur Form as a context-free grammar describing the text. Table 8 shows the grammar for the example sentence.

5 Evaluation and conclusions

There are two principal measures by which the induction procedure outlined above can be assessed: its utility for practical language processing tasks, and whether its suppositions and results reflect current linguistic theory.

5.1 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. We briefly outline each possible application in turn.

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 grammar generalization stage creates about 1650 new rules, the rules' average length decreases significantly, to just under 3 symbols. The number of rules and symbols per rule in the various grammars is summarized in Table 9. 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.

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

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 [34, 29]. One of the key problems in this area is the immense amount of labor required to develop a system to 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 10 shows a sample of text generated randomly from the grammar inferred from *Far From the Madding Crowd*. We find the quality of this extract disappointing, although, to be fair, this is characteristic of the text generated from compression schemes in general [37]. It indicates that the system in its present state has not been successful in capturing the essence of Hardy's grammar. We plan to investigate this deficiency and, if possible, remedy it. Studying the shortcomings of randomly-generated text is an excellent device for focussing attention on the quality of the grammar that is inferred.

Authorship analysis

Statistical techniques have often been employed to identify authors of anonymous texts, or to challenge authorship claims [11, 18]. O'Donnell [31] 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 in 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 [13]. 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.

Functional language processing

A computational account of language that focusses on functional elements is not a novel idea. Dewar *et al.* [17] describe a system that isolates syntactic components using grammatical information about a limited number of words: prepositions, articles, auxiliaries, conjunctions and pronouns.

In linguistic terminology, functional elements include both function words and inflectional

morphemes—affixes that change the subclass of a word without affecting its grammatical category, like the plural marker *-s* or the past tense suffix *-ed*. The system described in [17] 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.

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 shown in Figure 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. The notion of functional templates is not incongruent with psycholinguistic theories of sentence processing. We describe this more thoroughly in the following section.

5.2 Linguistic theory

Grammatical descriptions that result from syntactic inference may offer little direct assistance to theories of syntax, because many linguists have abandoned any notion of natural languages submitting to expression as context-free grammars. In fact, generative theories of syntax have undergone a major conceptual shift away from rule-based explanations of grammar, towards viewing grammar in terms of well-formedness conditions [36, 1]. Even so, it is still difficult to find any syntax literature that does not make use of phrase structure representations to analyze certain principles of grammar.

The idea of a function word grammar is supported by two important aspects of language theory. It represents a practical extension to DP-Theory, a response to the desire to reconcile the noun phrase within X-bar. Also, it reflects the apparent psychological importance of function words in sentence production and comprehension.

DP-Theory

Chomsky incorporated X-bar into his theory of grammar to capture cross-categorial generalizations that are true of natural language phrase structures. A satifactory exposition of X-bar is well beyond the scope of this paper, but we can for our purposes express a crude summary by the following formulation:

$$XP \Rightarrow \{COMP \mid SPEC\}\ X\ \{COMP \mid SPEC\}.$$

In X-bar, all hierarchical substructures of linguistic expression are labelled as head or non-head nodes, where non-head nodes are constrained by thematic projections from their respective head nodes. For English, the following formulation generalizes XP for verb phrases and prepositional phrases:

$$XP \Rightarrow X\ NP\ PP^*\ (SS).$$

For verb phrases, such as *invited the man with the bald head*, and prepositional phrases, such as *down the street*, the head is phrase-initial and the complement structure is subject to selection

(generally rightward) by, among other things, a thematic relation to the head—a relation projected as a lexical property [27]. That is, the structural head (the X) is also the semantic head—it is the lexical source of the descriptive content within the structure.

Unlike other structures in English, the head of a noun phrase 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 [21] and argued for extensively by Abney [1, 2] that selection is functional.

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.

Like DP-Theory, the function word approach to grammar induction uses functional elements to indicate the onset of a new phrase type, and generalizes phrase structure as a rightward continuation from that functional head.

Psycholinguistics

The peculiar properties exhibited by function words indicates that they receive a rather different treatment in cognitive language processes than do content words. Their late appearance in the productive vocabulary during first language acquisition seems to imply that the use of function words involves inferring somewhat more abstract grammatical knowledge than that required by other lexical items.

Psycholinguistic research of aphasics, individuals that suffer from language processing deficits due to brain damage, indicates that the function word vocabulary may exist as a separate mental lexicon from the rest of an individual's vocabulary. Goldstein [26] noted that one class of aphasics, those suffering from a condition known as agrammatism, demonstrate selective impairment in using one class of vocabulary elements—the "little words," or "grammatical words." Kean [28] further noted that the omission of function words in agrammatism is often accompanied by inflectional omissions—a characterization confirmed by Badecker and Caramazza [5].

Some slips-of-the-tongue by non-aphasics reveal the possibility that functional elements are composed into syntactic structures prior to the insertion of any major lexical items. 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 [22] to develop a psychological model of sentence production wherein functional elements establish sentence form.

Garrett's model describes a sentence planning process in which the choice and location of function words and inflectional morphemes are determined apart from processes that determine what content words are to appear. The model indicates that the syntactic level of sentence production consists of functional representations similar to those shown in Figure 4. Though the semantic intention of the production may influence which representation is to be selected, the semantic elements are inserted after the basic syntactic structure has been established.

The extent of psycholinguistic evidence that indicates a special status for functional elements in inflectional languages warrants a sincere effort to incorporate this notion into a computational account. The research presented in this paper represents such an effort.

6 Acknowledgments

This research has been supported by the Natural Sciences and Engineering Research Council of Canada. We gratefully acknowledge the help of Ingrid Rinsma in locating approximations to the hypergeometric distribution.

References

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

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

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

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

[14] Noam Chomsky. *Aspects of the Theory of Syntax.* MIT Press, Cambridge, Mass., 1965.

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

[16] V. J. Cook. *Chomsky's Universal Grammar.* Basil Blackwell Ltd., Oxford, England, 1988.

[17] 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.

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

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

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

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

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

[23] 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.

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

[25] N. Geschwind. The paradoxical position of Kurt Goldstein in the history of aphasia. *Cortex*, 1:214–224, 1964.

[26] K. Goldstein. *Language and Language Disturbances*. Grune & Stratton, New York, 1948.

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

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

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

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

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

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

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

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

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

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

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