The Marker Hypothesis
A Constructivist Theory of Language Acquisition

James Michael Pake

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
1997

Abstract

This thesis presents a theory of the early stages of first language acquisition. Language is characterised as constituting an instructional environment - diachronic change in language serves to maintain and enhance sources of structural marking which act as salient cues that guide the development of linguistic representations in the child's brain. Language learning is characterised as a constructivist process in which the underlying grammatical representation and modular structure arise out of developmental processes. In particular, I investigate the role of closed-class elements in language which obtain salience through their high occurrence frequency and which serve to both label and segment useful grammatical units. I adopt an inter-disciplinary approach which encompasses analyses of child language and agrammatic speech, psycholinguistic data, the development of a developmental linguistic theory based on the Dependency Grammar formalism, and a number of computational investigations of spoken language corpora. I conclude that language development is highly interactionist and that in trying to understand the processes involved in learning we must begin with the child and not with the end-point of adult linguistic competence.

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

September 1997

Acknowledgments

I feel privileged to have studied at the University of Edinburgh, where I originally took an MSc course in the Department of Artificial Intelligence before moving to the Centre for Cognitive Science. Both of these places provided stimulating and supportive environments and I feel that many people there have helped me along the road to this thesis. In particular, I would like to express my gratitude to my primary supervisor, Richard Shillcock, who has been friendly, helpful and supportive through good times and bad and has made a significant contribution to my learning over the past few years. I would also to thank my second supervisor, Matt Crocker, who has given me some very helpful advice and useful pointers along the way. And of course a big 'cheers' to my friends in Edinburgh who have helped oil the wheels of cognition.

I would not have been able to study at Edinburgh without the financial assistance which I received from the Department of Education in Northern Ireland and for which I am truly grateful.

Thanks to my parents for all their love and support and for first getting me interested in language all those years ago. And finally, I would like to send my deepest love to Sara, my wife, and Jacob, our baby boy, for the happiness and fulfillment which you have given me.

Contents

List of Tables

1. Introduction

Is it possible that certain constraints on language development are inherent in the language rather than the learner? In this thesis I propose a model of language acquisition which gives a greater role to language itself. This view (The Marker Hypothesis) holds that natural languages evolve to support the human learner in a complex interaction in which the development of a child's linguistic representations is guided by the informational structure of language in such a way as to ground later development in terms of earlier acquired constraints.

A key claim is that a pivotal role is played by closed-class words in the acquisition of grammar. It has been suggested that these words are bound in an intimate relationship with the grammatical structure of the sentence (Abney, 1987; Garrett, 1982; LaPointe, 1985). Others have argued that they play an essential role in the learning of the language by acting as explicit markers of structure in the speech stream (for example Braine, 1963; Green, 1979; Valian & Coulson, 1988; Gerken, Landau & Remez, 1990). It is the latter concept which forms the basis for the Marker Hypothesis.

Some of the main claims made in this thesis are:

1. Language structure is cued by the input *i.e.* language is not, at a surface level, merely a series of abstract symbols, but rather forms an instructional environment by packaging language into structural units through the use of various salient cues which are detectable by the learner in a naive state.

2. One such cue is provided by function morphemes - the salience of which lies, initially, in their high frequency and whose use as structural cues by the learner is motivated by concerns of efficiency and constraints on memory. These morphemes, I claim, act both to delimit the bounds of, and indicate the grammatical structure of, the lower frequency morphemes with which they alternate in the speech stream.

3. Languages are subject to certain evolutionary processes which lead to the marking of structural units: languages gain 'fitness' through their learnability and their facilitation of communication. Consequently, utterances in the language can be perceived as carrying a dual signal: one which is communicative and one which is instructional. Diachronic changes in language serve to maintain and extend the role of structural marking.

4. This view of language as instructional subverts arguments concerning the (un)learnability of language and consequent nativist assumptions - I argue that by treating language as an instructional environment we can reduce the complexity and the linguistic specificity of primitives posited as innate.

5. Language acquisition is best viewed within a constructivist framework - representational structures are developed incrementally under environmental guidance and earlier structure provides a constraint on later development. Consequently, I argue that to understand the acquisition of adult language we must consider the gradual development of language in the child.

1.1 Methodology

This thesis adopts a range of methodologies. I investigate various aspects of the marker hypothesis through computational corpus analysis, child language data, psycholinguistic research and a theoretical linguistic analysis.

While a considerable amount of research has been undertaken in the induction of linguistic information from corpora much of it has come from the language engineering approach. Often, this work, which is aimed at developing better and more robust parsers, does not

consider the particular problems faced by a child. We might ask a number of questions of any such model:

1. How unsupervised is it? Many learning systems use input which has been hand-annotated by linguists with information such as phrase-bracketing and part-of-speech tags. The current approach is concerned with how much can be learned 'from scratch' with only a fairly minimal initial state. The desire to keep the learning mechanism simple stems from a belief that to allow out investigations of child language to be guided by theories of adult lingustic knowledge is a perilous route to take. This view is summarised by Braine (1993):

> *Nativism has prospered because discovering what is cognitively and linguistically primitive is one of the fundamental tasks of the study of cognition and development, and great progress has been made on it in recent years. However, nativism is ultimately insufficient because it systematically ignores the other major task of the study of development, which is to account for developmental change, including the ontogeny of primitives... To the extent that we can explain the origin of some putative primitives in terms of learning based on others, we reduce the total number of innate primitives and thereby also the burden on a developmental theory of accounting for the origin of cognitive primitives.*
>
> Braine (1993, p.29)

2. Incrementality. Many of the existing learning systems are designed to achieve a particular end without worrying too much about how they get there. As such only the end result is considered. The theory I present places incremental development in a central position. Why does child language develop along a particular route? I argue that in order to understand how humans learn language it is essential to consider such developmental progressions.

1.2 Content

In Chapter 2, I review some of the arguments concerning the Logical Problem of Language Acquisition which underlies nativist accounts of language acquisition. I distinguish between what I call the Enigma Hypothesis and the Marker Hypothesis that offer two very different conceptions of language structure and learnability. The Enigma hypothesis is the view that

language structure from the point of view of the uninitiated is obfuscated with little indication of the abstract structure of the underlying grammar. According to this view, language learning is akin to breaking a code by selecting the correct grammar from a multitude of possibilities and the task is only made possible by giving the learner a considerable amount of language specific innate knowledge. The Marker Hypothesis suggests that, instead, language contains a number of sources of information in the form of distributional and phonological cues which serve to guide the learner in their task. Morgan and Demuth (1996) is a recent collection which gives a good overview of this area particularly with respect to phonological cues.

I then outline an argument framed in terms of linguistic change in which I argue that languages *evolve* to be learnable. This idea is discussed by Christiansen (1994) who presents a view of language changing so as to accommodate the learner rather than vice-versa. I then consider the specific case of how language change seems to preserve high frequency markers of structure and propose the view that we should see language (or utterances within a language) as providing a *dual-signal* - one signal is the familiar communicative function of language while the second acts as a guide to learning and acts primarily through the placement of high frequency morphemes in the speech stream in such a way as to structure the input in a manner that facilitates learning (although I do not deny the importance of phonological cues this thesis focuses primarily on distributional markers of structure).

In chapter 3, I review previous research which has been concerned with the role of marker elements. I discuss Harris's (1951) account of structural linguistics which emphasized the importance of frequently occurring functional elements in language as a source of syntactic information. I also look at how the Marker Hypothesis has been explored through artificial grammar learning. I then consider the overlap between high frequency and functional status in language and consider Garrett's (1975) model of speech production which presents a view of language production in which a distinction is made between syntactic frames composed of function morphemes and lexical fillers.

Chapter four considers how young children in the early stages of language development show evidence of using high frequency markers. Firstly, I consider the apparent problem to the current theory in the fact that, in the early stages of language development, children make little use of precisely those high-frequency functional elements on which this theory rests. Turning this apparent negative evidence around I review psycholinguistic evidence which shows that children who do not use function morphemes in their own speech are nevertheless capable of perceiving them, distinguishing between them and show awareness of their correct usage in language. I then argue that children's omission of function morphemes is the result of an active processing of language rather than a passive failure to encode them. I provide an informational account of why children might omit function morphemes from their own productive language while at the same time using them in structuring the linguistic input which they receive.

Using an operational definition of this theory, I then show how a very simple distributional account can make very similar predictions to a theoretical linguistic account of omissions in both child language and agrammatic language. Unlike a linguistic account, my approach puts the emphasis on learning and has little need for innate linguistic knowledge and, I argue, the simple distributional processes which underlie the model may also account for a functional-lexical distinction in terms of gradual modularisation along the lines put forward by Karmiloff-Smith (1992). Under this account, the existence of distinct functional and lexical components in the language processor is not the result of innate maturational processes but is rather a constructivist response to the structure of the language signal.

In chapter 5, I consider the interplay between functional morphemes and grammatical structure and argue that the traditional phrase structure account of syntax (as embodied in GB-theory, X-bar theory and the more recent Minimalist approach), while possibly offering the benchmark for understanding the structure of adult language, is not the most appropriate formalism for understanding how children represent language in the early stages of learning. Instead, I argue that the dependency grammar formalism is better suited to this task.

The argument I develop constitutes a constructivist approach to grammatical development. Within dependency grammar the concept of constituency is not a primary concept but a defined one and a number of possible, plausible, definitions of constituency can be defined 'on top of' a dependency grammar. I suggest that the child's grammatical representation is characterised by a succession of different and more powerful definitions of constituency. Importantly, this account preserves continuity - no aspect of the grammar has to be discarded and new representations are built on top of the existing ones.

In the first stage is the flexible definition of continuous dependency constituency given by Barry and Pickering (1990) which admits as a constituent any contiguous sequence of words which forms a connected sub-graph of the dependency structure of a sentence. Most importantly I argue that it is such units which are marked by both functional and phonological cues in language and thus serve as the initial focus of a learner's analysis of their language. By receiving such packaged input, the learner is able to constrain their analysis to those domains in which dependencies occur.

In the second stage, the definition of constituent is changed to the 'traditional' dependency constituent - a constituent consists of a word and all of its direct or indirect dependents. This produces non-overlapping constituents which are similar to those in X-bar theory but with shallower parse trees. An important difference is that modifier attachment is simpler so that, for example, no distinction is made between adjuncts, specifiers and complements in their relations within a noun-phrase. Also, the subject-predicate division which is central in phrase structure accounts is not present - subject and object noun phrases are attached equally to the verb of a sentence.

Finally, Covington has shown how a DG account can be made identical to an X-bar account by allowing the stacking of nodes and the differential attachment of modifiers.

This account allows the learner to begin with a representation which is more closely related to the structure of the input and to gradually introduce greater complexity which requires taking account of higher order features of the input.

Having outlined the distributional and grammatical underpinning of the theory, chapter 6 considers some empirical studies using computational corpus analysis to test three main questions related to the current theory.

The first study examines the question of whether high frequency words segment utterances into dependency constituents in spoken language. Secondly I consider how the child might extract useful syntactic marker frames from the input and use these to define lexical categories.

Finally, I consider how ,through the use of syntactic frames, the child might begin to 'detect' phrasal units in their input - marker frames serve to bound words into grammatical units; at this level no distinction is made between individual words and multi-word phrases and so this offers a simple form of structure dependent analysis as opposed to the strictly word-based accounts given in other distributional learning models. By being able to establish the equivalence of phrases and words at an early stage of learning we are able to gain entry into a simple phrase structure system and to use a structural linguistic approach to enable the identification of syntactic heads

Finally, Chapter 7 presents the conclusion to the thesis.

2. Language as an Instructional Environment

'Language is a form of human reason and has its reasons which are unknown to man.'

Claude Lévi-Strauss, The Savage Mind (1962).

This chapter explores some general background issues concerning the learnability of language which underlie the syntactic marker hypothesis. I examine traditional claims that natural languages are non-learnable from the available input and an alternative constructivist approach which considers the learner as being non-stationary and developing under the guidance of environmental stimuli interacting with non-domain-specific learning mechanisms. I then consider the nature of the 'environment' - in this case language, and argue that linguistic structure, rather than being essentially arbitrary from the point of view of the naive learner, is actually structured in ways which specifically guide the learning process. I argue that we can see the development of such structure in historical analyses of language and that the process is best viewed within the framework of evolutionary selective pressures acting on language itself.

2.1 The Logical Problem of Language Acquisition

The logical problem of language acquisition addresses the following problem: A fluent speaker of a language (the teacher) knows a grammar for their language, a discrete combinatorial system which, through the use of self-referential structures, can produce an

infinite set of sentences. A learner must reach a final state in which they can produce the same set of sentences as the teacher but they must acquire this ability through exposure to some finite subset of the teacher's infinite set of sentences. At the end of the learning process, both the teacher and learner should be able to agree on whether a particular sentence is grammatical or not even if neither of them has encountered the sentence previously.

Chomsky (1957, 1963) offered a framework for the classification of the complexity of languages. The `Chomsky hierarchy' proposes levels of linguistic complexity ranging from finite-state languages to the recursively-enumerable languages including the classes of context-free and context-sensitive languages. He argues that natural languages fall within the set of recursively enumerable languages. Gold (1967) found that only regular languages (as produced by a finite state automaton) could be learned with exposure to positive data only. In Gold's proof, he makes certain assumptions about the nature of the learning situation, considering two ways in which the language may be presented to the learner. In *informant presentation*, the learner is given access to grammatical and ungrammatical utterances, both of which are appropriately `labelled' - by using the knowledge concerning ungrammatical utterances, the learner is able to correct over-general rules which may have been hypothesised and thus restrict the grammar appropriately. In the other condition, *text presentation* the learner receives only positive, or grammatical, data and, as a result, is not able to use negative examples in order to restrict the grammar with the consequence that over-general rules may be developed *i.e.* the grammar will not only allow all the grammatical sentences but will also `overgenerate', producing extra, ungrammatical sentences. In Gold's method, at each time step, the learner is presented with a sentence from the language, and constructs a guess for the grammar, which may be based on all the sentences presented up to that point. If the learner reaches a stage where its guesses are all equivalent to the grammar being learned within a finite time, then the language is said to have been *identified in the limit*. Gold's proof showed that this method would only succeed if the set of possible languages contained only those defined as finite in the Chomsky hierarchy - if non-finite

languages also had to be considered the learner may adopt one of these as the correct grammar even it was over-general. Since natural languages are considered to be non-finite (Chomsky, 1957) and it is generally assumed that children receive little or no direct negative evidence (i.e. examples of ungrammatical sentences which are marked as such) we are forced to ask how the child manages to avoid learning a grammar that over-generates.

Gold suggested that there were three possibilities for overcoming the acquisition problem:

1. There is indirect negative evidence
2. There is ordered presentation
3. There is innate knowledge

It has been argued that, although the child appears to receive little direct negative evidence, they may nevertheless receive a considerable amount of indirect negative evidence if we assume that they adopt a statistical process which can interpret non-occurrence of a particular feature as negative evidence (Horning 1969). Even approaches which assume innate linguistic knowledge have had to acknowledge that statistical evidence is important.

In *ordered presentation* sentences are presented to the learner in an order which serves to guide development. Elman (1991) and Carroll and Charniak (1992) both support this idea - they show that, in computer modelling of artificial language acquisition, a language which is unlearnable by a particular system when sentences are presented in random order can be made learnable if sentences are presented in such a way that simpler sentences appear first, where simplicity may be defined, for example, in terms of sentence length (as in Carroll and Charniak) or, additionally, by some measure of linguistic complexity (as in Elman). However, there is little evidence that children can rely on receiving such tailored input. The motherese hypothesis is one such claim, but there are doubts concerning its usefulness or universality in learning and , as Elman notes, there is little evidence to suggest that modifications made in motherese affect the grammatical structure of language. An alternative to the ordered presentation of sentences may be possible in the form of constraints on the

learner or explicit markers in the language which serve to focus the learners attention on particular features of the language at different stages of the learning process. This possibility will be considered later.

The third solution suggested by Gold is that the learner contains innate knowledge which serves to constrain their hypotheses and thus prevent overgeneralisation. This has been the solution adopted in the theory of Universal Grammar.

2.1.1 Universal Grammar

Chomsky sees the solution to the learnability problem as lying in the genetic endowment of an innate language faculty - a mental organ which is pre-determined in much the same way as physical organs. This mental organ contains a highly structured rule system which constrains the language learner to a limited set of possible languages of which all natural human languages are members. Chomsky argues that

> *'these rule systems cannot be derived from the data of experience by abduction, abstraction, analogy or generalization in any reasonable sense of these terms, any more than the basic structure of the mammalian visual system is inductively derived from experience.'*
>
> (Chomsky 1987, p. 420)

This view is based on the Poverty of the Stimulus argument which claims that the language input underdetermines the grammar which must be acquired and therefore some alternative source of information must be available to the child. It is proposed that this is provided by an innate knowledge of grammatical principles which are common to all human languages: this Universal Grammar constrains the set of possible grammars which must be considered by the child. The child must learn the vocabulary items but the problem of grammar acquisition simply involves the setting of parameters. Baker (1979) argues that the innate linguistic

mechanism must be constrained so that it only considers rules which can be learnt on the basis of positive evidence.

Amongst the linguistic principles which are considered to be innate in Chomsky's theory are those of stucture dependency, the projection principle and headedness - it is argued that these are domain-specific features which could not be derived from the input and must therefore be part of the innate knowledge with which the learner begins.

Another question related to the learnability of language is that of how the child initially enters the linguistic system for a particular language- this is the Bootstrapping problem.

2.1.2 The Bootstrapping Problem

The Bootstrapping problem (Pinker 1987) can be characterized thus: the utterances of a language which act as input to a learner consist of strings of words but the final result of learning is coined in terms of abstract linguistic properties, such as word categories, which are not themselves indicated in the speech stream. There is a mutual interdependence between linguistic rules and categories because linguistic rules act on abstract categories, and these abstract categories only make sense if we consider them in the light of linguistic rules. Thus, we are faced with a classic 'chicken and egg' problem - how can children acquire either categories or rules without having first acquired the other? The Bootstrapping problem concerns the process by which the learner 'enters' the linguistic system by breaking this deadlock.

Various solutions have been proposed to this problem, and these vary both in terms of the type of information which they consider and the amount of pre-specified knowledge which they assume. The three main sources of information which I review here are semantic, prosodic and distributional. With regards to innate knowledge - some adopt a minimalist, developmental approach, while others assume the existence of substantive linguistic principles, such as X-bar theory. The various permutations that can be produced by

combinations of these assumptions can lead to a tangled web of possibilities, so I shall look at semantic, prosodic and distributional bootstrapping in isolation.

2.1.2.1 Semantic Bootstrapping

The main basis for the semantic bootstrapping hypothesis is that the child can exploit universal relationships between semantic entities in the environment and syntactic entities in their Universal Grammar in order to provide them a basic core of grammatical knowledge. These universal relationships include, for example, the fact that objects in the world are referred to by nouns, actions by verbs, properties by adjectives and so forth. Pinker (1987) provides a more comprehensive list of such relationships.

Maratsos and Chalkey (1981) have questioned the role which semantics can play in the acquisition of syntax, arguing that certain syntactic features display only an arbitrary relationship to the semantic reference field. In Pinker's model, innate syntactic knowledge, such as the principles of X-bar theory, supposedly avoids such problems by providing the details which cannot be extracted from the combined semantic and linguistic input from the environment.

It is worth noting that there are certain methodological problems with attempts to model semantic bootstrapping. In computer simulations, the programmer provides a set of example sentences and a set of semantic representations of those sentences. The problem is that the semantic representations may be over-generous in the information which they provide by, for example, providing a structuring of the data which is isomorphic to the syntactic structure to be acquired or by giving no more, and no less, information than is necessary to interpret the sentence, thereby avoiding the confusing mass of possible mappings that may be possible in the real world. These arguments do not, of course, in any way question the validity of the semantic bootstrapping hypothesis as a model of child language acquisition but are cause for doubting the relevance of such computer models in providing evidence to support the hypothesis.

2.1.2.2 Prosodic Bootstrapping

Another potential source of information to the learner is the sound structure of a language. Such features as word stress and pauses may coincide with linguistic features such as categories and clause boundaries. For example, in bisyllabic English words, nouns tend to have stress on the first syllable while verbs have it on the second syllable. So in the sentence *'I want to record a record'* the first occurrence of *record* has the weak-strong stress pattern characteristic of verbs, while the second has the strong-weak pattern which is more characteristic of nouns. Also, closed-class items more often appear in unstressed form while open-class items generally have at least one stressed syllable. Other prosodic features of language such as pausing and pitch may be correlated with phrase boundaries (Cooper and Paccia-Cooper, 1980 - although see Chapter 6). Pinker (1987) argues that theories of prosodic bootstrapping are not sufficiently developed to explain how the child can extract phrase structure from prosodic structure. I will return to the this topic at various points throughout the thesis.

2.1.2.3 Distributional Bootstrapping

Distributional bootstrapping (or correlational bootstrapping) deals in terms of the relationships which hold between surface elements in the speech stream. Such factors as the co-occurrence of two words or of a word and an inflection, or the serial position of a word in an utterance may be recorded. This approach is based on the assumption that words of the same syntactic category will appear in similar types of context, for example, a word which precedes the word *of* is likely to be a noun but unlikely to be a verb, while a particular set of inflections tend to be associated with verbs.

Various accounts of distributional bootstrapping are proposed by Braine (1963), Kiss (1976), Maratsos and Chalkey (1981) and Finch & Chater (1992). The increasing availability of corpora and powerful computers has enabled the claims of this hypothesis to be put to the test.

For example, Finch & Chater (1992) performed a distributional analysis of a large corpus of USENET news articles. A set of the 2000 most common words in the corpus were treated as focal items, to be categorised, and the 150 most common words were used as context items. The distributional statistics were collected in a contingency table which recorded the numbers of times a context word co-occurred with a focal item in a particular position. The system considered four relative positions - the two positions immediately to the right and the two positions immediately to the left of the focal item. So for example, if the focal item (f) is the word *win* in the sentence 'Who will win the election?' the system would record that *Who* appears in position [f-2], *will* appears in position [f-1], *the* appears in position [f+1] and *election* appears in the position [f+2]. Each of the focal items is represented by a 600 element vector which records the number of times that each of the 150 context items appeared in each of the four relative positions. Once the corpus has been analysed, these context vectors are normalised in order to rule out the effects of frequency and the resulting unit vectors are used as the input to the categorization process.

The categorisation process exploits the assumption that particular word categories tend to appear in particular contexts - thus similarity of category will be correlated with similarity of context vectors. There are many ways to assess the similarity of vectors; Euclidean distance for example, or, as Finch uses, the Spearman's rank correlation metric. A hierarchical clustering algorithm (see, for example, Sokal and Sneath, 1963) is used to group words into nested groups of categories. The resulting category groupings seem to reflect linguistic categories such as noun, verb, determiner *etc*. Huckle (1996) has shown that such an analysis can also extract semantic categories to some degree.

An interesting point about this approach is that words are categorised purely on the basis of surface-level characteristics without reference to other knowledge sources - no linguistic knowledge is 'programmed-in' nor is there any mapping to external semantic reference. Maratsos and Chalkey (1981) see this as a necessary feature of a syntactic acquisition model,

because of the existence, in natural language, of syntactic features which have no correspondence in the environment (for example, gender agreement in German).

They also note that such distributional approaches break what may otherwise be a circularity, words which appear in particular contexts belong to particular categories while membership of a particular category provides license for a word to appear in a particular context, by allowing words whose contexts display a signigicant overlap to be generalised into the same category. So, even if we have not seen a word appear in a particular position we may nevertheless infer that it can occur in that position if it bears sufficient similarity to other words which appear in that position - e.g. if words *X*, *Y* and *Z* occur in contexts C1, C2 and C3 but not in C4 and C5 and words *X* and *Y* appear in context C6, then we may infer that *Z* can also appear in C6. But it is important to realise that any such example is a simplification - in practice, such models consider a wide body of evidence across many words occurring in different contexts and categorisation takes into account the weight of evidence provided by different sources - decisions are not based on single occurrences in a particular context. I say this because one criticism of the distributional approach seems to make this over-simplification - Pinker (1979;1982;1987) argues that a distributional learner could infer, when given the sentences *John ate fish*, *John can fish* and *John ate rabbits*, that *John can rabbits* is a valid sentence. But this argument depends on a learning mechanism which focuses only on these contexts and which disregards contrary evidence, for example, the word's *can* and *ate* will, given a reasonably representative set of occurrences, be differentiated - they will not always appear with ambiguous neighbours such as *fish* (Finch's results support such a conclusion). Thus, the sentences above would be distinguished on the basis of their category labels. It seems strange that Pinker attacks such a straw man when his own theory of bootstrapping (Pinker 1987) recognises the importance of considering an overall weight of evidence rather than making decisions based on small, potentially misleading set of examples - any learning mechanism can be led to fail by being given hand-tailored, misleading examples.

However, it is true that the distributional approach has to cope with a reasonable amount of noise and that, ideally, we would like to reduce this noise. An example of such a noise problem can be considered in terms of a typical distributional approach.

2.1.2.4 Structure Dependence

One of the main criticisms of the distributional approach is that the dependencies which it considers are not the linguistically relevant ones - language does not deal in serial word positions but in abstract phrase structures. For example, given the inputs *The old man* and *The man*, a distributional approach such as Finch's will record the same relationship between *The* and *old* in the first utterance as it does for *The* and *man* in the second utterance. But *old* and *man* are not of the same category so the fact that they both occur in the same context will lead to noisy input being passed to the categorisation process.

Now, Finch's results show that adjectives and nouns can be distinguished between, given enough evidence, as the statistical basis of the system can overcome such noise, but this does show how serial positions are not a great indicator of linguistic dependencies - the difference can be highlighted by a prediction problem: given the word *the* what is likely to follow? The distributional approach could provide the relevant probabilities of particular words, or word classes, occurring next (such prediction problems are common in the use of distributional statistics in speech recognition) but could not specify a definite entity. Armed with linguistic knowledge, though, we would say that what follows will be a Noun Phrase - an abstract linguistic category that is not dependent on serial word position - this prediction will be correct in either of the above examples or any other grammatical sentence and thus it may be considered a more relevant dependency than that used in the distributional approach.

2.1.3 Miller and Chomsky's Criticism of Statistical Approaches

Miller and Chomsky (1963) proposed a number of criticisms of contemporary distributional accounts of language acquisition. Such arguments paved the way for modern theoretical linguistics as they showed alternative behaviourist accounts were inadequate as models of language learning or use. One such model was based on simple probabilistic word chains and

characterised the Stimulus-Response approach of Behaviourist psychology. In this view, behaviour could be explained as a series of learned responses to environmental stimuli - the classic example being the salivation of dogs in response to food, or an associated stimulus such as a dinner bell. Applied to language, the process of Stimulus-Response was held to underlie the production of sentences - a word would act as both a response to a previous stimulus and as a stimulus to the next response, or word. This account can be modelled by a probabilistic word chain, in which the probability of a word occurring depends on the word or words which have preceded it.

The training of such a model involves processing a body of utterances in a language and recording the probability of different words occurring given a particular prior context. We might decide, for example, to limit the prior context to sequences of two symbols in which case we would store the likelihood of a particular word appearing given a particular prior context of two symbols. So after analysing the sentence *The big dog chased the big cat* we would record that given the sequence *The big* there was an equal proabability of the next word being either *dog* or *cat.* Given a sufficient amount of input, such a model will produce a statistical approximation of the language. Such models constitute K-limited stochastic sources, in which K equals the number of previous contexts that are considered.

Miller and Chomsky highlight a number of problems with such an account. The first is the exponential increase in memory requirements as we increase the number of states considered: If W is the number of words being considered and n is the number of previous words considered, then where are W^n possible states to consider. Low order approximations require less memory but produce utterances which are ungrammatical whereas higher order approximations appear to produce more grammatical utterances, but with a concomitant loss in variation and increase in memory load. For example, a system which considers no prior context but simply recorded the chance of a particular word occurring, will produce a wide range of 'sentences' including novel, grammatical ones but it will also produce a high proportion of ungrammatical strings - it will produce grammatical sentences in much the

same spirit of an infinite number of monkeys banging away on typewriters. On the other hand a system which takes into account a long prior context (such as the previous twenty words) will produce mostly grammatical sentences - but only ones which it has seen: it will effectively record the input and not have the ability to generate novel sequences.

Miller and Chomsky note that 'there are many grammatical utterances that are never uttered and so could not be represented in any estimation of transitional probabilities' (p. 425). So although, higher order approximations capture a greater degree of grammaticality, the sentences which they they produce are more constrained by the training corpus. Chomsky gave the sentence 'Colourless green ideas sleep furiously' as an example of a sentence which is grammatical but which, nevertheless, has transitional probabilities which are very low. Only a very low order approximation would be able to produce this sentence, but only at the expense of producing mostly ungrammatical sequences.

Another problem for probabilistic models is the existence of long distance dependencies in language. Miller and Chomsky illustrate this with the example sentence *The people who called and wanted to rent your house when you go away next year are from California* where there is a dependency between the second word *people* and the seventeenth word *are*. They point out that even if we reduce the problem by dealing with the probabilities between word categories rather than individual words, and we assume only 4 categories exist then the system will have to record 4^{15} (= 10^9) different states in order to capture such a dependency. Such a model would require a vast amount of input to estimate reliable statistics for such an example, or as Miller and Chomsky say: 'we cannot seriously propose that a child learns the values of 10^9 parameters in a childhood lasting only 10^8 seconds'(p. 430).

It is important to realise the distinction between the statistical models which are criticised by Miller and Chomsky and models such as those proposed by Finch. One difference is that the latter makes use of a form of *contextual generalisation*. Braine (1963) noted that statistical word chains were not sufficiently powerful to handle language and that 'there must exist generalisation mechanisms in language learning whereby a word learned in one context

generalises to another context, even though no associations may have previously formed between the word and its new context.'

Finch's categorisation model enables such generalisation by allowing the learner to deal with sequences of word categories rather than word tokens. In such an account, a sentence such as *Colourless green ideas sleep furiously* which has a low probability of occurrence, can be tagged with category labels to produce a more probable sequence of word categories (e.g. adj., adj., noun, verb, adv.) By adopting a layered approach to learning, in which later developments are built upon category discoveries made at an earlier stage, we can avoid many of the pitfalls of the simple word-chaining devices criticised by Miller and Chomsky.

However, distributional accounts have by no means solved the language acquisition problem - they have merely shown that simple non-language specific processes can begin to uncover linguistic structure. Such work can be seen as complementing the theory of Universal Grammar by showing the way in which the non-innate component of language can be acquired. Alternatively, if it can be shown that a particular feature, normally considered to be innate, can be derived from the environment, then it casts doubt on the need for such a feature to be innately specified.

The current work addresses the latter issue in providing a theory which shifts a considerable burden of the language learning problem out of the domain of innate knowledge and onto the interaction between the learner and the environment. The remainder of this chapter will be concerned with the underlying assumptions of the theory - suggesting an alternative to traditional conceptions of language and learning.

2.1.4 The Enigma Hypothesis

After Saussure, it is generally assumed that words are completely arbitrary signs with no relation to that which is signified other than an intermediate, representation in the brain i.e. the word *cat* contains no quality which identifies it with the concept cat, only through a

lexical representation which serves to relate the phonological properties of the word to the corresponding context, does the relationship acquire any meaning.

In the logico-mathematical view of grammar structure expounded by Gold and Chomsky there is a similar arbitrary relationship between the surface realization of language and the underlying grammatical structure; both the words and the ordering of words assume an arbitrary form which gives no indication of the underlying grammatical structure, there are few clues to aid the learner in discovering the grammar for the language, consequently the nativist approach makes the assumption that the child must already have much of the structure and be given the innate predisposition to focus their learning activities on the relevant aspects of the input as defined by the Universal Grammar.

This view seems to be based on an assumption that I shall call the *Enigma hypothesis*. This is the idea that language assumes its complex form in order to facilitate powerful communicative abilities between users of the language. This knowledge, while allowing the expression of original and powerful concepts between initiates of the language is not easy to come by - the linguistic knowledge constitutes an arbitrary and abstract system of arcane knowledge whose acquisition must proceed under the guidance of a well-informed innate acquisition device which contains the phylogenetic Universal Grammar. Ontological development involves the setting of parameters on these innate principles and the, not inconsiderable, task of lexical acquisition. Thus, the argument goes, the learner which did not have the requisite innate knowledge would be faced with an impossible task: unable to converge on the appropriate grammar in the finite time available on the basis of the input data.

The Enigma Hypothesis is that language is as unintelligible to the uninformed learner as a message that has been deliberately obfuscated by its encryption into a secret code and the solution is to provide the 'code-breaker' with sufficient clues to help break the code, just as the British code-breakers at Bletchley Park in WWII were only able to break the German

Enigma code because they had knowledge of the machine used to produce it, thus enabling them to restrict their search to a manageable set of possibilities.

This portrayal of language as a cryptogram is made explicit in the following quote from Pinker who, in discussing how a combination of world knowledge and innate grammatical knowledge may aid the learning process, states:

> *'If children can infer parent's meanings they do not have to be pure cryptographers trying to crack a code from the statistical structure of the transmissions. They can be a bit more like the archaeologists with the Rosetta Stone.' (Pinker 1994, p 278)*

But is this view correct? We know that the sheer complexity of language may militate against its learnability but is the 'statistical structure of the transmissions' comparable to a secret code? The alternative view which I want to develop is that, far from being akin to a secret code, language is structured so as to facilitate learning - in Pinker's terms, the Rosetta Stone is included in the transmission.

2.2 Language as Teacher

2.2.1 Syntagmatic Iconicity

In order to begin to frame this idea let us consider another of the potential problems with the distributional approach that is raised by Pinker (1987) - that there are far too many possible properties for a learner to pay attention to all of them (he suggests potential candidates such as 'occurring in seventh serial position , occurring in the same sentence as the word *mouse*, occurring to the left of the word sequence *the cat in.*' (Pinker 1987)) How then does the child decide which properties to attend to? This raises a couple of points - firstly, some contexts are more useful than others - a context which does not appear very often is not very useful for the purposes of comparison so we should use frequently occurring contexts (infrequent contexts may include *e.g.*, appearing in fifty-first serial position , occurring to the right of *Colourless green.*) Even amongst those contexts which appear

frequently there is considerable variation in their usefulness as predictors of syntactic category, so a word appearing to the left of *of* is likely to be a noun while a word appearing to the left of *and* may be one of many categories. Ideally, a distributional approach should have the ability to assess the usefulness of various contexts as predictors of syntax - this is an issue which I will address in Chapter 6. The second interesting aspect to the question of which properties the learner should attend to lies in the answer typically provided by distributional theories - in the use of local contexts in preference to more distant ones. Finch's model makes use of the immediate neighbours of a word to provide its context and this is the norm in such approaches. The reason for considering local relationships is based on the insight that the majority of syntactic dependencies are realized locally in the surface structure of utterances and while long-distance dependencies do occur they are in the minority.

Hudson (1993) supports this view with the following figures, based on a dependency grammar analysis of a set of written sentences (approximately 900 words), which show that syntactic dependencies are more common between words which are close together (I provide a similar analysis on a larger sample of spoken language in Chapter 6).

- 71% (±8%)of all dependents will follow their heads.
- 67% (±3%) of dependents will be next to their heads
- 79% (±4%) of dependents will be no more than one word away from their heads.

Matthews (1991) calls this phenomenon 'syntagmatic iconicity' - iconicity in this sense corresponds to Haiman's (1980) definition of *iconic diagrams* namely, 'a systematic arrangement of signs, none of which necessarily resembles its referent, but whose relationship to each other mirrors the relationships of their referents' - in this case the signs are the words and the referents are their syntactic categories - the distance between words mirrors, imperfectly, the syntactic relationships between their categories.

The relationship may be imperfect, and perhaps trivial, but it nevertheless constitutes a non-arbitrary link between abstract linguistic structure and surface realization - because of its imperfection it does not fit neatly into a logico-mathematical model of language but it can be exploited by a learner who takes account of occurrence statistics.

That such a relationship exists is not surprising given the importance of locality in linguistic rules - this does explicitly imply a correspondence in the surface string, for example the local syntactic relationship between a noun and its determiner may be separated by an arbitrarily long string of adjectives on the surface level, but it does not require us to posit particularly sophisticated statistics to ensure that local syntactic relationships will *generally* be realised by a corresponding proximity on the surface level.

The fact that surface proximity correlates with syntactic relatedness does not explain why the learner should give preference to local relationships but it is not difficult to envisage that they would - language is by no means the only domain in which events that occur together in time are likely to be related in some other way *e.g.* the cause and effect of an action such as sticking a pin in a balloon and hearing a bang. It may not be a matter of choice - Elman (1991) argues that initial cognitive restrictions, in memory span for example, will force the learner to focus on important local dependencies before progressing to more global relationships; he argues that such a restriction on the cognitive abilities of the learner may in fact be a necessary part of language learning - a result which may be counter-intuitive and which I will examine later.

2.2.2 Introducing the Marker Hypothesis

Syntagmatic iconicity is one way in which underlying linguistic structure may leave its mark on the surface level of language but local relationships on the surface are not entirely reliable - as already noted, strict serial dependencies between words give us a rather noisy source of information about linguistic structure and do not take account of the structure dependency of language. One way in which the learners efforts may be made easier would be through the

provision of some form of bracketing information that would identify structural units above the level of the word.

I have already mentioned the prosodic bootstrapping hypothesis which makes use of another form of iconicity - that linguistic structure is in some way mirrored by the sound structure of language - that sound structure may be used to identify phrase boundaries. Such a source of information would give the learner an important extra constraint, enabling them to focus on the more useful syntactic dependencies in language while ignoring the more misleading serial dependencies.

Another source of such information is proposed in the marker hypothesis (Braine, 1963; Valian and Coulson ,1988; Green, 1979; Morgan, Meier and Newport, 1987; Stankler, 1991) which will form the basis for the work reported in this thesis. This hypothesis holds that all natural languages contain a small number of frequently occurring, grammatical items - function words such as determiners, prepositions etc. and inflections which serve to mark various grammatical features. These items may provide useful information in two ways: firstly they can provide a source of categorial information because they offer a few-to-many mapping between surface form and syntactic category e.g. the determiner *the* will occur in a wide range of, otherwise phonologically distinct, noun phrases and thus it provides a reliable marker of that structure. Secondly, such elements often appear at the beginning or end of phrases and thus may be used to provide information about phrase boundaries (e.g. Kimball (1973) suggests a parsing strategy that makes this assumption). It is suggested (e.g. Braine 1963) that a learner would initially focus on such elements because of their high frequency of occurrence relative to other elements, thus they would form the initial foci for associations. This view may seem at odds with the observation that in the early stages of language development children's productive speech is largely bereft of precisely these functional items. However, as I shall discuss in Chapter 4, there is evidence of an asymmetry between the production and perception of functional elements and I shall argue that children's failure to produce function morphemes in their own speech is not evidence of a failure to encode

them but is a direct result of processing biases which result from their distinctive distributional properties.

Over the next chapters, I will consider the marker hypothesis in detail - considering statistical, linguisic and developmental issues which, I will argue, support a view that the marker hypothesis has implications that stretch beyond the question of learnability and into the domain of cognitive architecture. However, let us now consider a precedent for such a theory in another area of cognition.

2.2.3 Instructional Environments

Returning to Chomsky's quote above, we see the view that the data to the child underdetermines the underlying rule systems which must be acquired; general learning mechanisms are not enough to extract the structure present in the utterances heard by the child, but the reference to the mammalian visual system is an interesting one because of a theory which suggests that innate endowment may not be sufficient to determine the development of the visual system but that its development but must be guided by the structure present in the environment. Evidence for this comes from studies of selective sensory deprivation on new-born kittens which were raised in environments that contained either horizontal lines only or vertical lines only (Blakemore and Cooper, 1970). Kittens raised in an environment consisting of only vertical stripes were found to be unresponsive to horizontally orientated stimuli and vice-versa. This sensory neglect was found to be reflected in the receptiveness of neurons in the visual cortex. Various other experiments have shown similar effects caused by selective deprivation of aspects of the visual environment which lead to a corresponding 'gap' in the visual cortex. One explanation for this effect is the 'instructive acquisition hypothesis' which suggests that the development of neurons takes place under the guidance of the environment and consequently if some aspect of the 'syllabus' is absent, as in the selective rearing experiments, then the corresponding cortical responses, lacking sufficient innate specification, will be unable to develop. Under this view, the development of the visual system is still innately specified to some degree, but it is very

much an underspecification and puts the burden of the developmental process on the 'tuition' provided by cues in the environment; the innate component can therefore be much simpler and more general than if a complete specification of the processes was required. The strategy is an effective one, and one that has proved reliable in the millions of years of evolution in which most mammals have been free from the interventions of neurophysiologists. There are criticisms of the instructive acquisition hypothesis - for example, it is claimed that although the cognitive development is malleable it is not so malleable that it could develop beyond the normal limits of the visual system if exposed to a radically different environment. The environment will have remained mostly constant through the evolutionary span and so once a visual system had evolved that was good enough to develop given the usual environmental input, there was no need for it to evolve further - the innate component has developed to take advantage of the constant environment. Now, when we turn to language my claim is that language itself constitutes the instructional environment but there is a radical difference from the visual context because now we have an environment which does not remain constant but is itself capable of evolution. Furthermore, whereas the presence of cats is unlikely to change the nature of the visual environment, the linguistic environment is bound in an intimate relationship with the language learner.

2.2.4 Constructivism

The role that the environment plays in guiding development is raised by Quartz and Sejnowski (in submission) who argue that the developmental process must be viewed as a `process of dynamic interaction between the informational structure of the environment and neuronal growth mechanisms, allowing the representational properties of cortex to be constructed by the nature of the problem domain confronting it '. They argue that the view of learning as expressed, for example, by Chomsky (1980), that development may be viewed as an instantaneous process without affecting its key qualities, is wrong and that it is instead necessary to consider the `non-stationarity' of the learner *i.e.* changes in the structure of the learner over time which serve to constrain later development . They cite the work of Elman

(1991), which demonstrated how initial restrictions in the learner (perhaps equivalent to cognitive restrictions in the child) could actually improve the learning ability of the system - a finding which contrasts sharply with the assumption that a learner can only be hindered by such restrictions. They see Elman's work as a particular example of a more general learning mechanism in which the representational space of the learner is shaped by the environment and they argue that the ability to shape the representation in response to the informational structure of the environment confers a very powerful and general learning ability which surpasses the kind of traditional hypothesis-testing model that was employed by, for example, Gold (1967).

Quartz and Sejnowski also consider neurophysiological evidence, concerning the effect of the environment on development and the apparent lack of pre-specification of this development, as running contrary to nativist approaches and instead suggesting that, although based on innate mechanisms, the structure of the brain develops through the interaction between general principles of cognitive growth and the input provided by particular domains in the environment. Concerning the nature of the environmental `instruction' they state:

> *How these cues are integrated into this view of learning as structural development, and how the system's actions on its environment and their consequences shape a theory of the development of cognitive representations remain as outstanding problems in the study of cognition.* (p. 32)

This thesis will show an example of how structured cues in the environment can shape the development of cognitive representations. I will consider a view of language learning in the context of a learner whose internal representations develop under the guidance of structural cues in the language. This argument will have implications in a number of areas including theories of natural language universals and of modularity in cognitive architecture.

2.2.5 Learning with Limitations

One question that arises when we consider the way in which an uninformed learner may approach the problem of language acquisition is that of why they should be more likely to prefer one particular approach over another. I have already considered how syntagmatic iconicity may exploit a bias in the learner to consider local relationships over long-distance ones - a bias which may result from memory limitations in the learner.

Another way in which cognitive limitations may result in the adoption of particular learning strategies is illustrated by observations of learning in neural networks. Seidenberg (1994) considers a case in which a multi-layered neural network is trained to 'auto-associate' a set of phonetic feature matrices *i.e.* given a particular feature matrix, the network is simply required to reproduce the same matrix on its output units. Now, if given a sufficiently large hidden layer, the network will simply act as a memorizing device - remembering which output to give in the case of a particular input. If, on the other hand, we limit the number of hidden units to prevent such rote-learning from occurring, the network is forced to compress the information through the hidden layer - the only way it can do this while still maintaining a reasonable success rate is to extract higher-level features from the input matrices and using these to make generalisations about the input. Such a limitation may also underlie a language learning strategy which focused on high-frequency grammatical markers - if the learner is unable to consider all possible dependencies in the input it would be natural for them to concentrate on those elements which occurred most frequently. So language input may act as an instructional environment by being structured in such a way as to exploit natural biases in the learner which stem from the need for cognitive economy and efficiency.

2.2.6 Modularity

Fodor's Modularity of Mind presents a nativist view in which cognitive functions are handled by innately programmed, function specific modules - the environment merely acts as a catalyst in the development of these modules.

However, modularity and innateness need not go hand in hand, as argued by Karmiloff-Smith in *Beyond Modularity (1992)*. She sees the modular architecture of the brain as being the end product of development rather than its starting point. Modularization is a gradual process which is a result of a dynamic interaction between environmental input and development of the brain.

A similar contrast in views will be presented later in this thesis. Here I will consider Radford (1990) in which syntactic acquisition is seen as taking place against a background of innate, modular linguistic systems. Radford argues that the grammatical form of early child grammar can be explained by the availability of lexical subsystems and the absence of functional subsystems. I will present a contrasting view, in which the lexical/functional distinction arises as a developmental response to the structure of language rather than being innately specified - I will argue that natural languages exhibit universal features which serve to promote such a distinction. Thus, my argument is that language development is structured but that this structure is more of a reaction to the structure present in language than the result of an innate specification. I will finish this chapter by addressing the question of how languages could come to exhibit such structure if it is not the result of innate knowledge in the brain.

2.3 Evolution of Language

The marker hypothesis suggests that there is a sense of order in natural language - an order which serves to guide the learning processes of the child. This raises the question of where such order comes from - one answer may be that the order is a result of innate structure provided by the human genome but the availability of such information reduces the need for the availability of such instructional cues in the first place. For example, Elman's (1991) findings suggest that limitations in the learner may serve to aid learning by forcing them to focus on the syntactic dependencies in language in the correct order *i.e.* by local relationships first- but this, in turn requires that languages be so structured that they exploit this bias in the learner by ensuring that utterances in the language tend towards syntagmatic iconicity. One

way in which this may be ensured is for there to be innate linguistic constraints which serve this purpose but if such innate knowledge is available then it reduces the need for language to mark surface structure in the first place. An alternative is that languages themselves are subject to evolutionary processes which select for language to serve as an instructional environment.

The idea that language itself can evolve dates back to Darwin's original work on evolution. A more recent proponent of this view is Richard Dawkins:

> *'Cultural transmission is analogous to genetic transmission in that, although basically conservative, it can give rise to a form of evolution. Geoffrey Chaucer could not hold a conversation with a modern Englishman, even though they are linked to each other by an unbroken chain of some twenty generations of Englishmen, each of whom could speak to his immediate neighbours in the chain as a son speaks to his father. Language seems to 'evolve' by non-genetic means, and at a rate which is orders of magnitude faster than genetic evolution.'*
>
> Dawkins (1976), p. 203.

Obviously, the relationship between the evolution of living organisms and that of language is not an isomorphic one but there are useful parallels particularly in the evolutionary maxim that random change with non-random selection produces complex, ordered structure. Christiansen (1994) provides a useful analogy, describing language as a *non-obligate symbiant* i.e. there is a symbiotic relationship in which the language user gains from the communicative ability conferred by knowledge of the language while the language itself gains from having a 'host' to inhabit and from which it can be reproduced.

Some may worry about the apparent intentionality that this description seems to ascribe to language but the idea that a language 'benefits' from being reproduced is no more fanciful than the idea that a DNA molecule benefits from replicating, the apparent intentionality is simply a natural feature of any evolutionary process, for example the apparent 'desire' of all living creatures to reproduce no matter how simple they are or how much they lack

cognitive structures comes from the fact that only organisms which fit this pattern do reproduce and therefore continue to exist.

Continuing this argument, we can say that only languages which are learnable would exist so the apparent ease with which children learn language is not surprising. This hypothesis makes the further claim that, because, as Dawkins claims, cultural evolution is much faster than biological evolution, we might expect that it would be more likely for languages to have evolved to exploit the available cognitive abilities of human beings rather than vice-versa.

Consider the analogy between a simple organism and a simple communication system: a mutation in an organism which causes its forbears to survive longer or reproduce more effectively will proliferate through subsequent generations, however a mutation which made an organism more able to survive but, in the process, prevented it from reproducing would not be carried on to subsequent generations, likewise a mutation which improved reproductive ability but had a high chance of causing early death would also not proliferate. Similarly a random change in language which increased communicative ability amongst adults but which was unlearnable for the child would not last, a change which made the language more learnable but cost communicative ability would confer little use to the learner and would not offer any advantage, but a change which independently improved communication or learnability without any loss to the other function would proliferate as would a change which served to boost both abilities.

So, as genetic evolution leads to a world full of complex reproducing survivors almost by default, we would expect similar processes of cultural evolution to produce complex, communicatively-useful and learnable languages. According to this view, language would not be so much an Enigma as a teacher, we would expect it to provide not just a communicative function but also an instructional function that would serve to guide the learning mechanisms of the child in the same way as proposed by the instructional environment hypothesis of vision but to a much greater extent because the linguistic environment can, unlike the visual environment, adapt itself to the structure of the brain.

Accordingly, we might dispense with the need for the child to be equipped with complex, innate, language-specific rule systems and instead focus on identifying the general learning mechanisms and the 'instructional' cues in natural language which may conspire to give rise to grammatical development.

The idea that language is both communicative and instructive implies that the 'transmissions' or utterances contain a dual signal. This suggests that as well as studying language structure in the usual method, by looking for the syntactic and semantic structure of sentences, we should look for accompanying cues which could serve to guide the learner and we should expect, for reasons of efficiency, to see structures which served both purposes. In the following chapters, I will examine the *marker hypothesis* which states that certain elements in language serve to provide cues to syntactic features of the language such as syntactic categories and phrase structure. Other researchers (e.g. Cutler 1993, Cutler and Mehler 1993, Jusczyk *et al* 1992) have argued that the phonological features of language, such as word stress, pitch and rhythm, also provide cues that aid the identification of word units, the categorisation process and the identification of phrase structure. Such work also falls into the general area of language as an instructional environment, and I will explore it further. Before continuing with the study of language structure, however, I will look at some historical evidence for the current hypothesis.

2.3.1 Language Change

Whereas the prevalent trend in modern linguistics has been the synchronic study of language i.e. the study of language as a 'snapshot' in time irrespective of its historical context, there has been a significant, though less fashionable, body of research devoted to diachronic, or historical, linguistics. The historical study of language is relevant to the current research in two ways: firstly, with regard to the evolutionary view of language, we may look for changes in language that may serve to maintain its learnability and, secondly, the historical development of language is guided by the intervention of generations of child language learners and so the history of language and the learning of language are closely related. One

way in which diachronic processes may serve to maintain the learnability of language is to ensure that structural features of the language are reliably marked by explicit cues in the speech stream; an example of such a cue would be an inflectional marking such as the *-ing* ending in English which is reliably associated with verbs. The use of such markings provides the language learner with a few-to-many mapping from inflection to word i.e. there are many verbs but only a small number of different inflectional endings which serve to mark them - according to the view that language is an evolving instructional environment, we would expect the occurrence and reliability of such cues to be maintained or improved over time. In fact, historical linguists have proposed several 'self-correcting' mechanisms whereby language maintains or updates structure overtime. A number of such processes come under the general description of *analogy*. Several examples of Analogical processes are described by McMahon (1994).

2.3.2 Analogical Extension

The process of analogical extension involves the generalisation of some morphological feature from positions in which it is already established to new positions. An example of this can be seen in the historical development of the *-s* inflection on plural and genitive nouns in English. Old English contained a much richer system of inflectional marking than Modern English; marking gender, number and case in a similar way to Latin, but many of these inflectional markings gradually broke down as various sound changes in the language led to the loss of their phonological salience. Some nouns, however, used the more robust *-s* ending in certain inflections and this became established as the possessive and plural marker in this group of nouns, while the other, less robust, markings were lost leaving only the lexical stem as the phonologically realised form for singular forms. Other nouns lost their inflections completely as a result of these sound changes and at this point the process of analogical extension came into play, generalising the *-s* ending from the small group of nouns where it

had traditionally occurred to those nouns which now lacked any plural or possessive inflection. Such analogical extensions are represented by a proportion of the form

stan: stanes = sunne:X (X=sunnes)

Which describes the generalisation of the -s ending from its established occurrence with the word *Stan* ('stone') to the unmarked *Sunne* ('sun'). *Sunne* previously took the various plural forms *sunnan, sunnena, sunnum* which indicated case, gender and number features and which were lost as a result of phonological changes occurring in the language. As well as the analogical extension of the -s ending to 'fill the gap' made by the loss of these endings, the change in marking which was accompanied by the loss of phonological marking of case led to the more fixed word order of Modern English. If we view such inflectional markers as structural cues which are used by the learner in the acquisition of structure, then we may view the process of analogical extension as one which serves to maintain the instructional nature of the linguistic signal. The details of how such markers are used by the learner will be considered throughout this thesis, but for now, let it be said that the processes which maintain, or create, such structures are not to be viewed as part of the grammar of the language, a view expressed by Lightfoot (1979):

> *analogy is a principle governing the construction of grammars, influencing the form of grammars, but in no sense directly represented in those grammars.*
>
> (p. 371)

Rather, we may view the process as an extra-linguistic one, which is manifested in the grammar indirectly through the interaction between the linguistic input and the mechanisms of the learner. As regards these mechanisms, a key theme of this thesis is the role of marker frequency - in order for an element to serve as a marker it is not sufficient for it to be merely

a grammatical or functional item but it must display properties which cause it to be adopted as a marker element by the learner - one such property is for it to exhibit a higher frequency of occurrence in the input language than other elements. Again, the field of historical linguistics provides support for the importance of frequency in linguistic change.

2.3.3 The Importance of Marker Frequency

McMahon (op cit) considers several areas in which frequency is implicitly or explicitly tied to the processes of diachronic change in language. Concerning analogical processes, she notes that 'In general, the connection of resistance to analogy with frequency seems to hold' (p. 73). That is to say, that irregular forms which occur with high frequency are less prone to be generalised by the process of analogy - for example an irregular verb past-tense is more likely to be generalised by analogy if it occurs with a low frequency.

A striking effect of frequency can be seen in the process of grammaticalisation that has been proposed as occurring in the 'creolization' of languages. Givon (1979) has noted that auxiliary verbs in Creoles very often develop out of a small set of lexical verbs in the originating Pidgin language. For example, the lexical items meaning `want' and `go' often become future tense auxiliaries and `finish' and `have' often become markers of perfective or past categories. This suggests that the underlying semantics of these verbs leads to their frequent use in particular contexts associated with the passage of time and this frequency, in turn, leads to their grammaticalisation as syntactic markers of tense.

Another example of the relationship between frequency and function is observed by Singleton and Newport (1993) in their study of deaf children learning ASL from parents whose late learning of the language had resulted in an imperfect grasp of the grammar. Such children display a more rigid adherence to the grammar of the language than their parents, for example, in cases where parents may only use the correct grammatical 'morpheme' in 65% of contexts in which it was obligatory, their child actually used it in 90% of obligatory situations. Morgan and Newport suggest that this results from a process of 'frequency boosting' during learning *i.e.*the difference between adult and child grammar is a matter of

degree - the learner simply makes more frequent use of structures that already display a high relative frequency of occurrence in the input language - they suggest that a similar effect may underlie the phenomenon of creolization.

Such effects of frequency and function are not restricted to exceptional circumstances such as the process of creolization, for example McMahon describes the process of grammaticalization that has taken place within the French negative construction. The negative form '*ne... pas*' was originally signalled using the word *'ne'* alone - the *'pas'* extension was originally a lexical item that was used to add emphasis (meaning *'not a step'* from the Latin *passus*) in the same way , as in English, we might say `I don't like it (one bit)'. Prior to the existence of written language, the history of this process is somewhat speculative: Price(1984) argues that the use of *pas* would have originally played a more semantic role - being associated with verbs of motion and being expressed in a nominal roles, as in the sentence '*Je ne vais un pas*'. By the time of the earliest written languages *pas* could be attached to any verb and appeared as an adverb, but its use in negative constructions was not obligatory and it shared the role with a number of similar negative particles *e.g. point* (from Latin *punctum* meaning 'place' or 'spot'), *goute* (from Latin *gutta* 'drop') and *mie* (Latin *mica* meaning 'crumb'). While some of these forms are still used in rare colloquial French, *pas* has become an obligatory marker of the negative and, McMahon notes, it is actually beginning to be used on its own as the only negative marker as *ne* is increasingly dropped in the spoken language.

This example displays two ways in which languages may change to support the marking of syntactic structure. As, gradually, the '*pas*' extension became grammaticalised it lost its lexical meaning and flexibility and become a fixed, functional marker of the negative. Thus we see how grammatical markers may be created out of tokens which originally serve a semantic function. The second point is that a single marker replaced a number of original markers which has the effect of increasing the relative frequency which the surviving marker exhibits and thus increasing its salience and its syntactic iconicity. Frequency may also have

played a role in the original process of grammaticalisation - from its original shift from an oft-used emphatic item to an obligatory negative marker. Sankoff and Brown (1976) describe the process by which the relative clause marking particle *-ia* in the creole language Tok Pisin seems to be a functionalized reduction of the tag question 'Hear?' which served the discourse function of topicalising a referent in the Pidgin language. Examples such as these seem to indicate that frequency is an essential component of the creation of syntactic markers out of lexical morphemes - items which reach a certain critical frequency of usage in a particular linguistic context are promoted to functional elements.

Other theorists in historical linguistics have suggested that the process of structural marking is an ongoing one. For example, Kurlyowice (1949) and Manczak (1958) have suggested , on the basis of dictionary-based etymological studies of languages, that marking becomes more pronounced over time, or as Hock (1986) puts it: 'More overt marking is preferred'. Manczak claims that syntactic marking in languages display certain 'tendencies' when viewed diachronically, for example:

1. Longer inflectional forms are 'preferred' over shorter ones.
2. There is a tendency for alternative stems to be reduced to single stems.
3. Zero endings are replaced by full endings and mono-syllabic endings by polysyllabic endings.

These examples show how languages change in such a way as to preserve, or enhance, the use of grammatical marker elements which is what we would expect to find if linguistic structure was under selective pressure to be learnable and if syntactic markers serve to enhance learnability.

In summary, this review has provided evidence that languages exhibit the necessary diachronic processes required to give rise to instructional cues in language.

2.4 Conclusion

I began by considering some arguments about the learnability of language and argued that the supposed need for substantive innate linguistic knowledge had been motivated by the poverty of the stimulus argument which holds that grammatical structure is underdetermined by surface level utterances. I suggested that this view had arisen from a view which I call the Enigma hypothesis *i.e.* the assumption that grammatical structure is akin to a secret code - hiding its true underlying structure on the surface level - and that this structure is designed purely for communicative purposes. I proposed an alternative assumption - that surface level utterances are structured so as to serve not only communication but also to guide the development of the learner. I proposed that languages evolve to exhibit such structure and that we can see potential evidence of such processes in the historical study of language change. Learning is seen within a constructivist framework - the development of cognitive structures is a result of an interaction between the mechanisms of the learner and the structure of language rather than the result of an innate program, and progressive stages of development are constrained by previously established structures. One implication of this view is that it is vital to consider the incremental development of language - a theory of language acquisition should explain the progressive stages in the development of child language and I will consider this in later chapters. The theory also has implications for the modular model of cognition - I will consider this issue in relation to the development, as opposed to the innate existence of, distinct lexical and functional systems in language. Furthermore, I will argue that a constructivist approach is also required at the level of linguistic theory - proposing that the initial syntactic representation and use of constituency in the learning process is best described in terms of a theory based on dependency syntax but that adult grammar is better described by a phrase structure based theory, however this conception does not demand a sudden shift in representation or a loss of information which would entail an unnappealing break in continuity but may rather take place gradually - building new representations on top of the existing substrate under the guidance of linguistic input.

3. The marker hypothesis

In the previous chapter I presented the view that language acts as an instructional environment by being structured in such a way as to guide the development of the learner. This chapter will explore a specific instantiation of this view which claims that certain morphemes in natural language act as structural markers and serve to provide the learner with information regarding the scope and nature of phrasal units.

3.1 Language as Discovery

Harris (1951) proposed a set of discovery procedures that could be used by the field linguist in uncovering a structural description of a language, a task that has obvious similarities to that faced by the language learner. Harris adopted a layered approach in which structural description at one level was based on those features discovered at an earlier level - I will consider his method for uncovering syntactic categories or morpheme classes, an approach that is mirrored in the substitution tests used in modern linguistics, as Harris puts it:

> *We equate any two sequences of classes if one of them is substitutable for the other in all utterances in which they occur. (p 263)*

Thus, words which appear in the same environments are treated as belonging to the same class. We can tabulate morpheme strings together with the environments in which they occur to form a more efficient representation than a listing of the separate sentences, for example:

left context	target	right context
Did you	[eat]	the stuff
He'll	[get]	it later
	[find]	them for me please

This illustrates how the three verbs *eat, get* and *find* may appear interchangeably in the three sentential contexts '*Did you ___ the stuff?*', '*He'll ___ it later*' and '*___ them for me please*'. However, it is not possible to perform such a test perfectly in practice because of a lack of common sentential contexts and so some approximation of the procedure must be used. Harris therefore proposes the use of 'environments shorter than the full utterance', or *short environments*, which are reliably associated with particular word classes - these might include particular suffixes or other grammatical markers. But, Harris notes, environments may give rise to a number of possible classifications, for example the environment '*The ____*' which may be associated with nouns, will also admit words such as *very, large* etc., while the environment '*the large ____*' admits *and beautiful* as well as *man*.

Harris suggests that certain short environments may occur with all the morphemes of a class, such as '*They will ____*' but that 'it may be impossible to devise a procedure for determining which short environments over what limits should be set up as the differentiating ones for various substitution classes', unless the linguist already has knowledge of the language or access to an informant.

Obviously, if a learner could somehow establish a set of useful linguistic environments they would aid in the task of syntax acquisition. The use of such environments is available to adult speakers as is shown by the ability to make some sense of nonsense verse such as Lewis Carroll's *Jabberwocky* - when we hear the words *the slithy toves* for the first time we can use the recognisable morphemes *the,*

-y and *-s* to establish the grammatical properties of the unknown morphemes. Such an ability would be extremely useful in the early stages of learning when a large number of lexical items and syntactic forms must be acquired. Pinker (1987) proposes that such distributional information is used, but only after the child has already established a sufficiency of syntactic knowledge through the combination of innate grammatical constraints and the use of

semantic reference. In his approach the child will have already acquired a core set of vocabulary items for each of the major lexical classes through their cooccurence with concrete semantic referents. Only once such a state has been reached can the child then note that particular classes appear in certain grammatical contexts and then use this knowledge to identify the class of elements for which concrete semantic reference is not available (*e.g.* nouns with abstract referents such as *idea* or *situation*). In order to explain the use of grammatical contexts, or short environments, within the distributional approach it is necessary to explain how the learner could isolate the necessary information purely on the basis of exposure to utterances of the language. (It is also necessary, if this theory is to extend beyond the merely possible, to then ask if children can and do use such information - this will be the focus of the next chapter).

3.1.1 Marker Frames

Braine(1963) suggests that the language learner develops grammatical frames, which are similar to the short environments of Harris. '*In English a grammatical frame consists of an arrangement of closed-class morphemes and dashes, such as THE ___-S ___LY, which completely determines the part of speech going in each vacant position*'. Whilst Harris believed that knowledge of a language is required in order to identify useful short environments, Braine proposes an approach that would enable a learner to extract grammatical frames from the input without access to syntactic knowledge of the language.

He suggests that these frames may be developed on the basis of associative links between morphemes; he argues that given the high frequency of closed-class morphemes such items will form the strongest links between one another while less strong connections will be developed between closed-class items and open-class items. The strong associations between the high-frequency closed-class morphemes will give rise to grammatical frames consisting of closed class morphemes, such as '*IS* ___*-ING*' and '*HAS* ___*-ED*' or '*OF THE* ___ ##' (where ## is used to indicate the beginning or end of an utterance), while the slots in the frames will contain lower frequency items whose grammatical category may be determined

by the type of frame in which they occur. Given a sentence containing unknown words, the learner may use these frames to identify the syntactic category of these items - they can use contextual generalisation to identify other contexts in which they may occur. Indeed, as so many open-class words can play more than one syntactic role (for example *dog, table, rose, book* can all be either nouns or verbs) the context provided by functional morphemes will be required to categorise even known words.

In this view, the development of frames is gradual - as the learner listens to the utterances of a language their attention will initially focus on those items which occur most frequently and on the associations between these items and this will lead to them acquiring sets of strongly associated grammatical morphemes. Lower frequency lexical items will then be classified syntactically in terms of their co-occurrence within these frames. Braine suggests two mechanisms here - firstly there are the associations between the lexical items and the frames in which they occur, and secondly he argues that lexical elements which occur together within a frame will develop stronger associative bonds with each other than they will with other lexical items outwith the frame. This raises the following two roles which are served by grammatical frames and which are explicitly stated by Gerken, Landau and Remez (1990):

1. **Labelling:** Closed class items can serve to label phrases with which they occur. Distributional approaches to the bootstrapping of syntactic categories (see Chapter 2) assume that words can be categorised on the basis of their co-occurrence with frequently occurring items. For example, noun phrases may contain a large variety of possible word tokens - a variety which would prove confusing for a learner, but there is a much smaller set of possible determiners, *the, a, an, etc.* which can act as more reliable cues. Because the same determiner may occur across a wide range of noun phrases it can serve as a common link between them. Grammatical markers thus offer a more invariant relationship between syntactic structure and surface structure than is offered by open-class, lexical items.

2. **Segmentation:** As noted by Harris, it is necessary to establish the limits over which a marker's effect holds - language structure is not based on the serial positions of words as used by most distributional approaches but between abstract linguistic entities. A second claim of the marker hypothesis is that grammatical markers can serve to mark the boundaries between useful grammatical units because they tend to appear at the beginning or end of a phrase (Clark & Clark,1977; Kimball, 1973). For example, in the sentence *The big dog is happy*, the words *big dog* which occur between two closed-class words will be treated as a linguistic unit - thus the range of the linguistic environment which is labelled by *The* is delimited by the next closed class item in the sentence. Kimball (1973) proposes that closed-class items can be used in parsing to signal the creation of a new phrase node and thus we would expect it to be easier to parse sentences which use markers to signal new phrases than those which do not. So the sentence *I know that John is guilty* would be easier to parse than *I know John is guilty*. Several early parsers relied entirely on closed-class information in order to assign structure to a sentence (e.g. Thorne 1968)

Both of these claims may be considered statistical arguments - they both depend on a weight of evidence rather than one-off linguistic examples. One way of testing such claims is by applying them to a computational analysis of a body of natural language - this will be considered in chapter 6 of the thesis.

The rest of this chapter will be concerned with previous work which has explicitly used the marker hypothesis in the field of artificial language learning experiments and to addressing the relationship between the statistical properties of markers - their high occurrence frequency - and their grammatical properties - their status as closed-class or functional elements.

3.1.2 The Role of Word Frequencies

Is it possible that word frequency could play such an important role in language acquisition? The distributional bootstrapping approach used by Finch and Chater (1991) and described in Chapter 2 also makes use of word frequency - words were categorised on the basis of their

cooccurrence with a set of context words which comprised the 150 most common words in the corpus. There were two reasons for using the most common words - firstly there is the question of economizing memory - if all words were allowed to act as context items then there would a huge number of possible dependencies to consider - if we use a reduced set of words to form the context then it makes sense to choose those which occur most frequently as they will give rise to more cooccurrence data. A second reason is that the 150 most common words in English are predominantly closed-class items and these play a pivotal role in syntax.

3.1.3 Zipf's Law

Zipf (1935, 1949) noted that there was a logarithmic relationship governing the occurrence frequencies of words - if we rank the words in a corpus and then divide each word's rank by its frequency of occurrence we will obtain a value which remains fairly constant for each word. This means that a small number of word types will account for a large number of word tokens in a corpus. Zipf also noted that the more frequent words tended to be shorter. He attributed these phenomena to 'The Principle of Least Effort', the details of which we need not go into, except to say that it suggested some underlying drive for efficiency which manifested itself in many areas of human endeavour - he noted a similar relationship holding across a range of diverse areas of human activity such as the population of cities and the distribution of income amongst a population. Miller and Chomsky (1963), however, examined the claims for a meaningful relationship holding between word frequency and word length and came to a rather different conclusion. They demonstrated how the phenomena noted by Zipf could arise as the result of a purely random process. To illustrate this, consider that we have a device which outputs a string of symbols which are each chosen at random from a list of the 26 letters of the alphabet plus a space character. If we define a 'word' as any sequence of letters occurring between two spaces we can calculate the probability of a particular word occurring and thus an indication of its frequency in the symbol generator's output. WIth this alphabet we will obtain 26 words with an occurrence

probability of $1/27^2$, 26^2 words with the probability $1/27^3$ etc. Essentially the same relationship holds between word rank and frequency and word length and frequency as in natural languages. The authors thus conclude that there is no reason to view such relationships as indicating anything meaningful about human languages and they suggest that:

> *any grammatical rule regulating word lengths must be regarded with considerable suspicion - in an English grammar at least.* (p. 430)

They suggest that word frequency is a side effect of other underlying processes, but they note that there are some peculiarities in that there are differences in the distributions which follow particular classes of words and they conclude that:

> *There is nothing in our present parsimonious theory of the rank frequency relation that could help us to explain these apparent deviations from randomness.* (p. 434)

One feature of English, which Zipf noted, is the disparity of type:token ratios between word classes, this is the well documented fact that closed-class, or function words, generally occur much more frequently than the open-class, or content words. If we look at a ranked frequency list for English words we will find that the top 150 words consist almost entirely of closed-class words and that most of the closed-class words occur in these positions. This would suggest a non-random connection between word frequency and one aspect of linguistic function, namely closed-class status. Various statistics have been published on this phenomenon, for example Cutler (1993) notes that 1% of word types that a speaker knows account for 50% of the word tokens which they hear and Shillcock *et al* (submitted) report that in the London-Lund Corpus of English conversation (Svartik & Quirk, 1980), which contains some 494,000 tokens, 52% of the words were closed-class (although such figures depend on what we define as being closed-class). Such figures suggest that the development

of grammatical frames through associations between frequently occurring closed-class items, suggested by Braine (1963), may be feasible.

3.1.4 The Role of Closed-Class Words

Given the figures above, we can see that potentially a learner could use a simple word frequency count in order to give a rough differentiation between open- and closed-class items. In fact, we may imagine that a learner who operates under restrictions of processing power and memory would be forced to pay special attention to the most frequent words - by remembering only the most frequent 1% of words they would 'recognize' 50% of the words which they heard. Thus, the learner would initially build up some store of occurrence frequencies for words. This need not be particularly detailed or accurate as we are only interested in a gross distinction between very high frequency (VHF) morphemes and other (non-VHF) morphemes. There are two ways in which such a distinction may be made - either on the basis of a frequency count or on some measure of the time between a word's tokens re-occurrence - Zipf refers to this as the wavelength of a word. In either case, the end result is that VHF morphemes will be treated differently from other morphemes. Once such a distinction has been made it will result in a bipartite split in the analysis of utterances - Figures 3.1a shows a graph in which the words of the sentence *The old man is walking his dog down the street* run along the abscissa and the occurrence frequencies of the words (taken from a corpus of spoken natural language) are plotted on the ordinate. Figure 3.1b shows the same sentence using rank frequencies and figure 3.1c shows the raw frequencies on a logarithmic scale. We can distinguish between the morphemes on the basis of a frequency cut-off point - those which lie above the cut-off are treated as grammatical marker elements, while those below the cut-off are treated as lexical elements to be analysed in terms of their occurrence within grammatical frames.

If we set the cut-off point at the 150th rank frequency then our grammatical frame would consist of '*the* ____ *is* _____ *his* ____ *down the* ____' and this would isolate the lexical elements '*old man*', '*walking*', '*dog*' and '*street*'.

This example shows how a distinction based purely on word frequencies may be reasonably effective in distinguishing between function words and lexical words. This example is limited to whole words rather than morphemes but we might expect that at the morpheme level word endings such as *-ing* would stand out as high frequency elements.

We could also use a mechanism which took more account of the context of a particular sentence by utisilising *changes in frequency* between adjacent items *i.e.* the decision as to whether to use a word as a marker in a particular context would depend on the degree to which it stood out by virtue of being more frequent than the surrounding words. In either case, we would expect that such an approach would tend to isolate functional or closed-class items as belonging to the grammatical frame and lexical or open-class items items would then be analysed within the context of these frames.

A useful analogy is with that of a figure-ground distinction in a visual scene - the grammatical frames are akin to the ground or background of an image and the lexical groupings are akin to the figures or objects of a scene - contiguous sequences of lexical elements which appear together (such as *old man* in the above example) are seen as being more closely related than those which are separated by grammatical elements. The learner is thus analysing lexical objects in terms of their occurrence against the grammatical background. The next chapter will present an implementation of such a *frequency filter* and the results of its application in a study of child language.

Fig. 3.1a

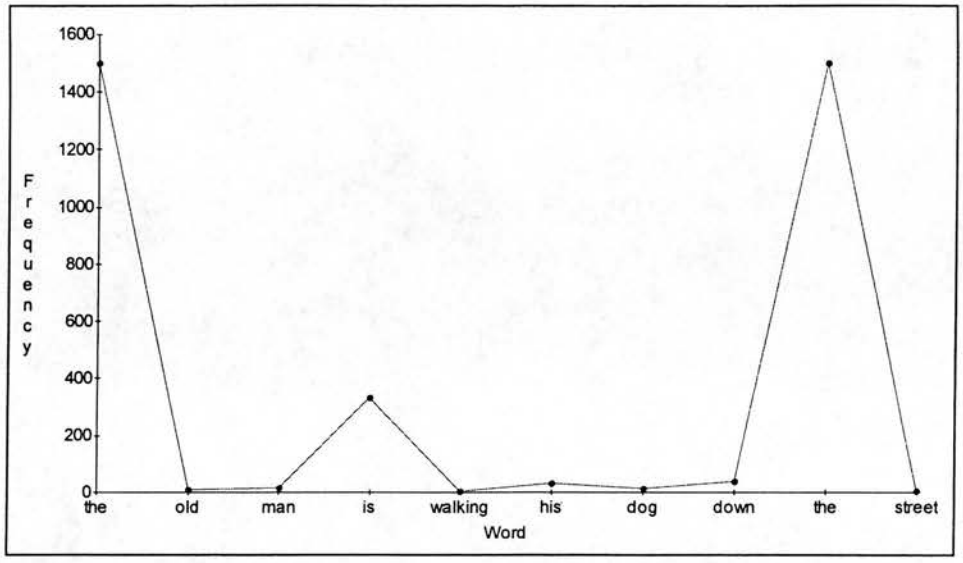


Fig 3.1b

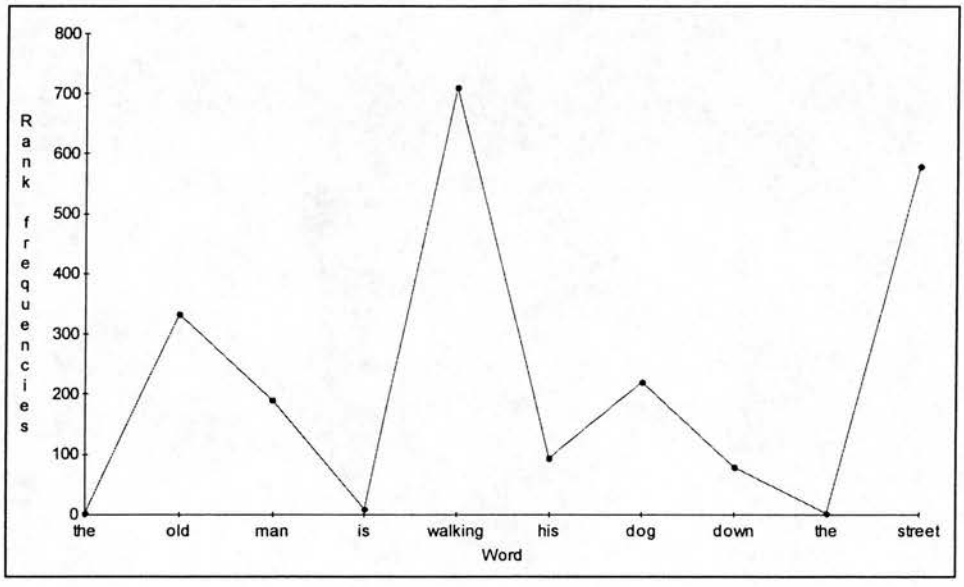


Fig. 3.1c

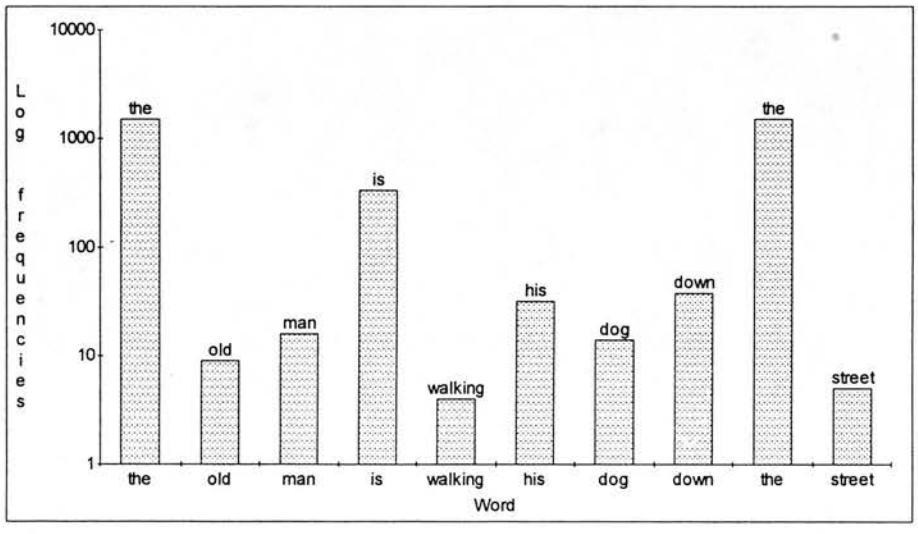


Figure 3-1 Frequency structure of the sentence 'The old man is walking his dog down the street'

3.2 Evidence from Artificial Language Learning

The most explicit use of the marker hypothesis in previous research has been in the field of artificial language learning experiments. In such work, a simple artificial grammar is used to randomly generate a set of symbol strings (consisting of, for example, letters or nonsense words). A subset of these sentences is then used as the training set for a grammar-learning experiment. There are a number of paradigms used in this field (for example, experiments vary in whether or not the task is presented explicitly as a grammar learning task or implicitly as a memory task) but they all share the common goal of testing a subject's ability to learn the underlying grammar that produced the sentences rather than the sentences themselves - this is tested by requiring the subjects to distinguish between grammatical and ungrammatical sentences which were not present in the original training set.

Moeser (1977) argues that the use of such *miniature artificial languages* (MALs) in language-learning experiments enables us to control different aspects of the learning environment which would be impossible in natural-language-learning studies. Certainly, the interest in the experiments which I review lies in their variation of different aspects of the language or the learning environment which cannot be studied independently in naturalistic settings. However it is necessary to approach this research with some caution for we cannot make direct comparisons between the MALs that are used and natural languages. It is also important to note that most of the studies covered here involved the use of adult subjects who have already acquired a language and have generally passed the supposed critical period for language learning. Indeed, this work seems to be founded on the assumption that language learning employs only general purpose learning mechanisms and that these mechanisms are available to adult subjects. But this assumption is highly controversial and even if true we have to take into account that older subjects bring to bear a body of linguistic knowledge which is not available to the first time language learner. MALs are invariably much simpler than natural languages; rarely containing more than fifty vocabulary items which are assembled into sentences by a small number of very basic grammatical rules. The interest in

these experiments from the current viewpoint is that they provide the most explicitly stated claims of the marker hypothesis so far.

In one early MAL experiment, Smith(1966) used a simple grammar of the form:

$$S \rightarrow M\ N$$

$$S \rightarrow P\ Q$$

Where M, N, P and Q represent word classes which are each realized in the surface strings as one of a number of word tokens. Smith found that subjects had little problem learning that M and P words appeared in sentence-initial positions and N and Q words appeared in sentence-final position but they failed to capture the dependency between the classes, *i.e.* they were as likely to consider the 'ungrammatical' sequence MQ as being grammatical as MN or PQ sequences although they would reject NM sequences.

Thus the subjects appeared to learn the absolute positions of words in a string but not the dependency between word classes. Smith's work can be thought of as providing a base case - what is learnable purely on the basis of a distributional analysis without alternative sources of information. Real language learning demands that learners *do* master the syntactic dependencies which exist between word classes, and the suggested solutions to this problem within the work in MAL learning experiments have certain parallels to the debate over the kind of information used in bootstrapping, as reviewed in chapter 2, in that phonological, semantic and distributional approaches have been proposed.

Moeser and Bregman (1972) argued for the necessity of a semantic reference field as an aid to learning. In their study, the different sentences produced by a MAL each had a corresponding semantic referent which consisted of an image of an arrangement of shapes. There was an isomorphic relationship between the words of a sentence and the particular arrangement of objects in the corresponding image. Without the semantic reference field subjects were unable to learn the syntactic dependencies of the language, as in Smith's experiment. However, when provided with the semantic reference for each sentence they did

learn the underlying grammar. This result was presented as evidence that semantic information played a vital role in the acquisition of syntax.

Green (1979) provided an alternative source of information for the learner in the form of syntax markers. In his experiment, subjects in different conditions were required to learn different 'dialects' of a MAL which differed with respect to their use, or non-use, of syntactic markers. In the unmarked dialects, sentences consisted of nonsense words which were seen as being equivalent to open-class items *i.e.* each class contained a number of possible tokens.

The marked dialects used the same form as the unmarked dialect but added marker 'words' to phrases - these marker words may be considered as equivalent to closed-class items because they have fewer possible forms (In fact only one form for each phrase type). For example, in the unmarked dialect, there would be sentences such as '*Deech hift*' and '*Tasp ghope*' while the equivalent marked sentences would be '*Ong deech ush hift*' and '*Ong tasp ush ghope*'.

Green found that the dialects which did not use markers were virtually unlearnable but that languages which used markers to indicate both words and multi-word phrasal units were learnable. Comparing these results to the findings of Moeser and Bregman, he suggests instead that even a semantically empty language may be learnable provided that it contains markers to its structure.

Thus, the problem of learning dependencies between abstract syntactic classes that was identified by Smith may be solved by providing either a semantic reference field or grammatical markers. One problem with the semantic solution is that the idealised relationship between the semantic reference and the syntactic relationships which existed in Moeser and Bregman's study is unlikely to be mirrored in a natural language-learning setting. Different languages vary greatly in the syntax which they use to encode similar semantic forms and some syntactic dependencies seem to have no obvious correlate outwith the syntax of the language. For example, syntactic gender systems seem to have little

relationship to semantic properties of the world (Maratsos and Chalkey, 1980) and in such systems, children appear to initially develop the syntactic and semantic components separately (Karmiloff-Smith,1979).

Morgan *et al* (1987) see the role of grammatical markers as being just one instance of what they call 'structural packaging' - the idea that language, through a variety of systems, is packaged into grammatical structural units which are focused on by the language learner. So function words, for example, display a range of distinctive features including high frequency, weak stress and a tendency to be monosyllabic and their occurrence may be reinforced by other cues such as prosody. The current research focuses mainly on the role played by high frequency elements as markers but this is for practical reasons rather than a theoretical assumption that frequency is the only important factor, indeed I will later discuss the position of word-frequency as just one of a number of related synax markers in natural spoken language. Morgan *et al* suggest that learning may require the correlation of various cues which are not, by themselves, sufficient to aid learning.

With regard to the role of word frequency, Valian and Coulson (1988) added to the debate by comparing the learnability of MALs in which the relative frequencies of markers are varied. In their experiment, they compared MALs which differed in that one used markers that had a relative frequency that was six times that of the 'open-class' items whilst in the other the ratio was 3:2. Only the language which used the higher frequency markers was learnable. They, like Braine, contend that children employ a domain-general process (distributional analysis) to acquire the domain-specific knowledge of language typified by syntactic categories. They make explicit the requirements for an element to act as a syntax marker (or *anchor* as they call it):

1. It must be reliably associated with just one type of syntactic structure.
2. It must have a high frequency relative to surrounding items.

They point out that these requirements are independent - a determiner, for example, may be reliably associated with Noun Phrases but it will only be useful as a marker if it also occurs with a very high frequency.

3.2.1 Do Markers work together or in isolation?

Valian and Coulson stipulate that markers be 1) reliably associated with a single structure and 2) of relatively high frequency. Concerning the first point, we may ask if a marker should be considered as a single element or a collection of elements which conspire to cue a single structure. Consider the -s ending in English which has certainly got the frequency requirement but is not reliably associated with one structure. For example, consider the homophony of bracketed phrases in the following sentences:

The (dogs) are friendly.

The (dog's) bowl needs cleaning.

He (dogs) my every move.

There are several points to note here:

1) The *-s* marker serves to restrict the possibility of the word category to Verb or Noun, as opposed to say Adjective.
2) The determiner 'The' serves as a more reliable marker for the Noun category but may combine with the -s ending to provide additional support for the categorisation choice.
3) Although the plural 'dogs' and the possessive 'dog's' differ syntactically - the category of the lexical stem 'dog' is the same in both cases.

Markers in natural language are less clear-cut than those used in the artificial languages which I have described. We might expect language to show a tendency to avoid such ambiguous markers either through their removal over time or through the provision of

disambiguating cues. In the examples above, the additional functional elements surrounding the lexical item *dog* should be sufficient to distinguish between its use as a noun or a verb.

Research in artificial language learning provides considerable support for the marker hypothesis but is this research reliable? There are good reasons to question the applicability of these experiments to the study of natural language learning. Firstly, none of the experiments used pre-linguistic subjects and so we can question their relevance to natural language learning - one assumption that has cropped up is that language learning is just an example of general learning but this position is controversial, and if children actually make use of an innate, specialised language acquisition device which follows a strict temporal series of stages (see for example, Radford 1990) then we can argue that these experiments on adults and older children tell us nothing about natural language acquisition. Even if we assume that language learning makes use of domain-independent processes we must acknowledge that such processes will be partly driven by previously acquired knowledge and, thus, the language-acquainted subject can bring to bear top-down processes which may not be available to the pre-linguistic subject. If the subjects themselves are a problem to the relevance of this work, then the problem is greatly compounded by the use of artificial languages themselves.

As, I have explained, MALs are, by necessity of the learning task, very simple compared to natural languages both in vocabulary size and grammatical complexity. Although they offer the advantage of giving us a heavily controlled learning task, it is questionable as to whether we can extrapolate the results based on these languages to real language learning. For instance, consider the use of phrase markers in MAL learning experiments - we have seen experiments which suggest that they are necessary in the absence of a reliably correlated reference field but how can we be sure that natural language employs markers in the same way? Valian and Coulson make the stipulation that anchor points be associated reliably with just one type of structure and that they have a high relative frequency, but this raises the problem of identifying such elements in natural language, for example, in English, the word

and meets most of the requirements for a marker element (i.e. monosyllabic, unstressed, hi-frequency, closed-class) and yet its occurrence is definitely not restricted to just one type of structure (consider *Jim is tall and skinny, I went and bought a magazine, John and Mary went to Paris and London*) but in a typical MAL the marker elements have been used in a very specific and simplified way and do not appear to give the learner the problem of identifying which markers are reliable and those which are not and so they may over-simplify the task.

On the other hand, many artificial grammars in both human learning tasks and computational learning studies have made no use of marker elements (e.g. Elman 1990) and so, although these grammars are, in many ways, simpler than natural grammars they may be adding a complication that is not present in real language.

So artificial grammars as well as being much simpler than natural languages may either simplify or complicate the learning task in ways that appear to be impossible to quantify. Although the research reviewed in this section has raised some interesting points that are in accord with the current thesis, it only really helps in the generation of hypotheses as regards natural language. The computational approach which I have adopted gives two main benefits over these experiments:

1) it enables us to test our hypothesis on a large body of natural language and
2) it allows us to view the internal representation that is developed by the model in the process of learning which enables us to make comparisons with the development of language in the human learner.

The aim of the research in this thesis is to test the marker hypothesis by investigating ways in which markers may be identified and used in the acquisition of grammatical structure.

3.3 Closed-Class Elements in Language Processing

The discussion so far has raised two ways in which grammatical markers may be characterised. Statistically, they are items which appear with a very high frequency and this

is one feature which may motivate a learner to treat them distinctly. Linguistically, they generally correspond to closed-class or functional elements. Statistical features of language are viewed with suspicion in traditional theoretical linguistics - Miller and Chomsky's account of word frequency effects in language sees them as an unimportant side-effect of the linguistic rule system and warns against the notion that grammars may encode such features directly. On the other hand, the Marker Hypothesis views the frequency of grammatical markers as being essential to their role in language learning. I have already noted that, side-effect or not, the words which occur most frequently in a natural language corpus are generally closed-class elements. Consequently, I will now turn to a consideration of the role of closed-classes in language.

3.3.1 The Marker Hypothesis and Adult Grammars

Given the proposal that grammatical frames are used by the language learner and that the learner distinguishes between functional and lexical items, on the basis of frequency, in order to make an initial identification of these frames we might expect to see some evidence of such distinctions in the adult speaker because:

1) The adult grammar has been constructed upon the child grammar and therefore we might expect to see artifacts from the early distinction based on functional frames and the lexical units delimited by these frames.

2) The adult would have to produce such frames in their speech in order for the child to make use of them.

Note that neither of these reasons necessitate that the adult grammar displays such features explicitly *i.e.*there need not be a specific and distinctive level of processing devoted to the production of such frames - a simple CFG can give rise to the appropriate structures if it is suitably constructed. Similarly, although the child may employ frames in learning the language, this does not imply that the frames are themselves an explicit part of what is learnt. It is important to note that although this model proposes that the child initially makes a

frequency based distinction between lexical and functional items, this does not imply that the adult grammar employs such distinctions, rather frequency is presented as the 'entry point' into the functional/lexical distinction but as linguistic knowledge develops the frequency based distinction will be augmented or superseded by other sources of variation between the classes, or higher level features in the input. For example, as distributional information is accumulated, the class of marker items may be redefined in terms of its members' occurrence patterns. Frequency just serves to 'bootstrap' the marker elements, but the adult grammar will differentiate between members of this class on the basis of more abstract features which are better described in conventional linguistic terms. So the distinctions made in terms of frequency in the learner will be manifested in terms of linguistic entities such as closed- and open-classes in the adult.

We might then ask the question 'What processes ensure that languages exhibit the required marker structure?' The marker hypothesis suggests a distinction in processing of language by the learner which exploits structural cues in language. Is it necessary for the production of such structural cues to be explicitly stated in the rule system of the adult? With regards to this question I will turn first to Garrett's theory of adult speech production which offers an account which seems to provide a natural complement to the use of markers in learning.

3.3.2 Garret's Model of Speech Production

Garrett's model was developed out of an analysis of adult speech errors. He claims that systematic patterns in the nature of speech errors can be explained by the adoption of a particular model of normal production.

Amonst the types of speech error which Garrett identifies are:

1) Word exchange errors, such as

'Is there a cigarette building in this machine?'

'This spring has a seat in it'

Where the underlined words have been exchanged from their intended positions.

2) *Stranding* exchanges, such as:

'I thought the park was truck-ed'

'Fancy getting your model re-nose-d!'

In these exchanges, the underlined lexical items have been exchanged but bound function morphemes remain, stranded, in their correct position rather than following the exchanged item with which they should be associated.

3) Sound exchanges, such as:

'Sot holdering iron'

'Show snovelling'

Where the underlined sounds have been exchanged.

Garrett notes that there are marked differences between the different kinds of errors, for example, in sound exchanges we find that generally:

1) exchanged items are metrically and phonologically similar.
2) exchanged sounds tend to be from the same position in their respective word (*e.g.* word initial, word final)
3) Words involved in a sound exchange tend to be members of the same major phrase.
4) Words involved in a sound exchange tend to be of different syntactic category.

On the other hand, word exchanges differ in important ways:

1) they more often involve words of the same grammatical category.
2) they tend to occur between, rather than within, phrases.

Furthermore, in all of the error types, the exchanged elements are almost invariably open class elements and this is emphasised by stranding exchanges such as "It wait-s to pay" and "You have to square it face-ly" where lexical items are exchanged but the bound closed-class morphemes remain in their intended postions. Garrett argues that these distinctive error profiles suggest a multi-level speech production system with different levels being responsible for the different kinds of errors observed.

Garrett (1975) proposes the following outline of the various levels in the production process:

1 Message source

2 Functional level

3 Positional level

4 Sound level

The functional level involves the creation of a syntactic frame based on the structure of the message to be produced. Garrett argues that the closed-class items are immanent in the frame i.e. the actual phonological forms of the words are fully realised in the frame with place holders for the lexical items to be inserted. So, for example, in producing the sentence "The dog is chasing the ginger cat", the functional frame would be *The* ____ *is* ___*-ing the* _____. The full sentence would be generated by inserting the appropriate lexical items into the place holders in this frame to give the final sentence - stranding errors occur whenever these lexical items are inserted into the wrong placeholders e.g. "The chase is catt-ing the dog" and the bound morphemes, which are already in place in the frame, are left unnaffected.

There are obviously parallels between Garrett's model and the current model, the latter being practically a reverse of the other - in the learning model, the high frequency frame which is first identified corresponds to Garrett's functional frame and is used to isolate the embedded lexical items, although initially only localised frames, rather than full sentence frames are used. The two models complement one another rather well, for they make the same

distinctions in structure. While Garrett's model proposes that a sentence is constructed out of different levels, the marker hypothesis holds that the learner processes perceived sentences on different levels - given the sentence *The cat is chasing the dog*, the learner will treat the high frequency morphemes as marker elements to yield the structure *The __ is __ -ing the __*. We may be tempted to say that the model of speech production proposed by Garrett is complementary to the marker hypothesis in learning because both models propose a similar distinction between grammatical frames and lexical slot-fillers - this might support the view that the adult explicitly produces the grammatical frames which the learner employs. However research by Bock (1989) has questioned the privileged role of closed-class items in Garrett's model, specifically she claims that closed class words are not immanent in the structural skeleton of sentences. Her syntactic priming experiments suggests that it is not particular closed-class elements which are primed, but rather the associated abstract grammatical structures. She suggests that these results may be consistent with a weaker form of the closed-class hypothesis as embodied in the grammatical model of LaPointe (1985) in which there exists a one to many mapping between grammatical frames and associated closed-class elements.

In some senses LaPointe simply provides details which Garrett's model left unspecified. His model deals with the issue of verb form production in agrammatic speech and is intended to explain the disparity between the production of lexical and functional items in this condition. In his model, the grammatical frames consist of tree fragments which are associated with the maximal projections of head categories.

LaPointe suggests that appropriate lexical items are inserted into the frame, while functional items are retrieved from a distinct store and are themselves represented as tree fragments. Rather than being associated with particular closed class items, these tree-fragments contain slots for classes of items such as auxiliaries.

Relating these different accounts to the present model, I would argue that the learner's initial representation of an utterance is more similar to that of Garrett's model *i.e.* frames are based

on specific functional items rather than more abstract categories. As I have already suggested, the initial gross distinction made on the basis of word frequency must be refined through the use of higher level, more linguistically relevant, properties of the input so that functional items will gradually be differentiated as new syntactic knowledge is acquired. So the more abstract representation proposed by LaPointe is not incompatible with the current model, rather we might say that the notion of syntactic frames which is used in Garrett's model embodies the structural distinctions of the early learner while Lapointe's account reflects the more informed and abstract adult model built upon the earlier account. Initially then, the syntactic representations would be directly associated with particular closed-class items (i.e. these items would be immanent in the frames used by the learner) but as these items were abstracted or categorised, the frames with which they are associated would also become abstracted, so that frames which shared many features would be abstracted into a single representation replacing particular closed class items with category markers pointing to the appropriate items.

3.3.3 Implicit Production of Frequency Distinctions

It is important to realise that in stressing the importance of word frequency distinctions in learning we do not have to posit a statistically based model of language in the adult speaker. To illustrate this point I will consider two simple text book grammars. A context free grammar (CFG) of the form:

S -> NP VP
NP -> Det, Noun
VP -> V, NP
Det -> the, a
Noun -> dog, cat, man, woman, boy, girl
V -> loves, likes, chases, respects, surprises, distracts

can be used to parse and generate simple sentences such as *The boy distracts the dog* - we could generate sentences by either producing all alternatives in sequence or by choosing an

alternative for each rule at random during the grammatical expansion. This grammar contains no statistical bias on the rules - each alternative is as likely as any other and there is no inherent statistical model, by contrast a Stochastic CFG (SCFG) attaches probabilities to the likelihood of choosing each alternative so that for example we may add the following probabilities to the final rule of the grammar above:

V-> loves (0.5), likes (0.1), chases (0.2), respects (0.1), surprises (0.07), distracts (0.03).

So, in this case, the verb of the setence will be *loves* 50% of the time and *distracts* only 3% of the time. However, it is not necessary to use a stochastic grammar to produce the kind of frequency effects that are exploited in the current theory, the non-stochastic CFG above is sufficient for this purpose. If we make the basic assumption that each alternative has an approximately equal chance of occurrence then the type/token disparity between the different word classes will ensure that some words occur more frequently. In the example above the NP always consists of a Determiner followed by a noun - the fact that there are only two possible determiners and six possible nouns means that, given a random choice of each alternative, we can expect a particular determiner to occur three times more often than a particular noun. If we increase the number of possible nouns but keep the number of determiners constant then the disparity between the frequency of the words of these two classes will increase. There are several ways in which such a frequency effect can be maintained in natural language:

1. The fact that words are closed class ensures that the number of different words in that class cannot be increased, thus preventing their 'dilution'. For example, imagine that we were to introduce a new definite article into English to be used with plural Noun Phrases instead of *the* which would be restricted to singular cases only - there would a considerable drop in the occurrence of *the* that would correspond to the proportion of plural definite noun phrases. However, such a hypothetical situation would not arise because of the resistance of the closed-classes to new members. One reason for this

resistance may be to preserve the high frequencies of these words and, somewhat cicularly, the resistance of these classes to new members may be because of their members' frequencies. I have already suggested that closed-classed items may initially be treated differently because of their high frequency, one effect of this may be to increase the closed-class system's resistance to new members.

2. The lack of concrete semantic reference of these words enables their use across all domains of discourse , whereas open-class, content words are restricted in use to particular subjects or by the meaning of the utterances. We might use the words *Polar Bear* in a conversation about the Antarctic or in the Zoo but they would be unlikely to occur in a conversation about football. By contrast, closed class function words such as *the, to* or *in* can occur in any context (except perhaps telegrams and tabloid newspaper headlines). Such a relationship between the occurrence frequency of a word and its promiscuity across different semantic domains has been a consideration in the study of techniques for classifying documents. Luhn (1958) provided a measure of this relationship, in order to provide some indication of the usefulness of a word for characterising the content of a particular document. He defines a distribution which admits words of a median frequency while excluding those which occur with a particularly high or low frequency. Figure 3.2 shows such a distribution over a Zipfian rank-frequency curve. This distribution acknowledges the fact that infrequent words are of little use because of their low occurrence, while high-frequency words are of little use because of their semantic promiscuity. This feature of closed-class words also makes them ideal for marking the purely syntactic structures of language.

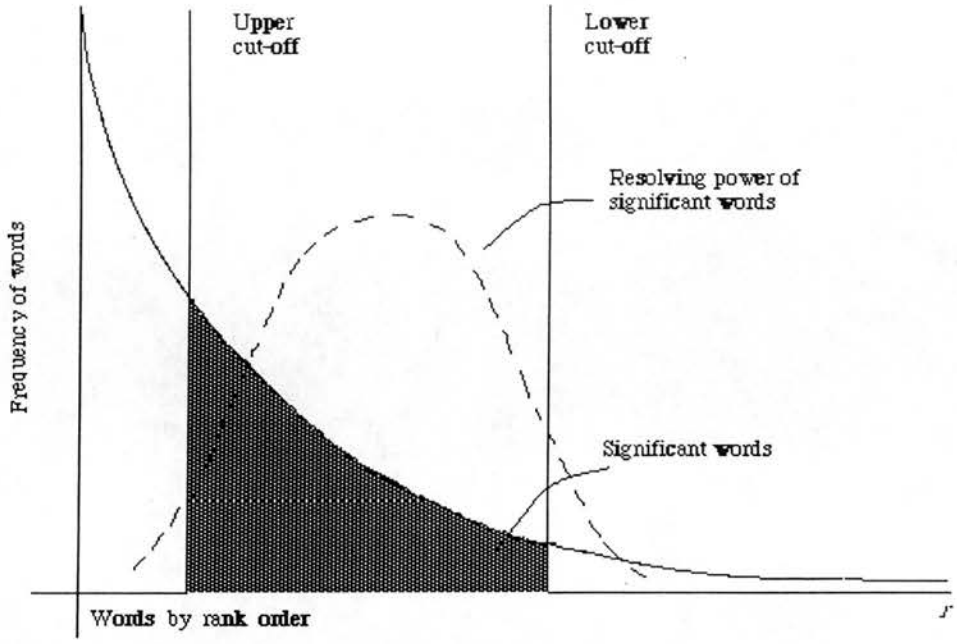


Figure 3.2 Resolving power of words based on their rank frequency (Adapted from Schultz (1968), p. 120)

Thus the very nature of grammatical markers (functional rather than lexical and closed- rather than open-class may be features which preserve their role in language acquisition.

Note that I am not saying that precise frequencies are of interest - the exact number of times that a word is encountered will vary greatly between people. However I am interested in the leftmost end of the Zipf curve whose inhabitants tend to remain fairly constant across speech samples and are kept this way by the features mentioned above. The relative frequencies of content words heard by children will vary greatly according to their situation (A child who has a pet dog will hear `dog' more than a child who has a pet cat) but the higher frequencies of the function items will not vary nearly so much across these different contexts. Although the relatively high frequency of certain items may arise out of a purely symbolic grammar (as discussed above) that does not explicitly encode probabilities, I am making the claim that this feature is not merely a side-effect but an important aspect of the learnability of languages. I have already suggested a certain circularity here - the special processing of

closed-class elements may be initially motivated by the statistical properties of these words, while this special processing may in turn perpetuate their high-occurrence frequency.

3.3.4 Closed-Class Words in Modern Theories of Grammar

In recent years, there has been a new importance placed on the role of functional elements in language. While previously major phrasal categories were seen as the projection of lexical heads, new interpretations propose a model in which functional categories act as heads. For example, traditionally the phrase *The old man* would have been treated as a Noun Phrase, but Abney's (1987) version of GB-theory treats it as a Determiner Phrase in which a noun phrase (*old man*) is projected into a DP by the functional element *The* (Hudson (1984) also argues for determiners as heads in his conception of dependency grammar). This reflects a general trend towards placing a greater responsibility for the syntax of language onto functional categories. For example, Chomsky (1988) has suggested that the major source of parametric variation between languages lies within the functional items while lexical elements tend to show a much greater degree of similarity across different languages. The idea that functional elements carry a greater burden of the syntax of a language is compatible with their use as marker elements in language acquisition while their differential processing, their resistance to new members and their applicability across diverse semantic contexts all serve to preserve the statistical properties which enable their use in this role.

3.3.5 Differential Processing of Open- and Closed-Class Words

A number of researchers have proposed that closed- and open- class elements are processed differently in the brain.

Much of the research has been geared towards providing an explanation of agrammatism. For example, LaPointe (1985) suggests that lexical and functional elements are stored and processed differently while Shillcock and Tait (1994) have shown that the syntactic loss in

agrammatic speech can be predicted by a loss of functional projection in a GB-theoretic account of syntactic production such as that proposed by Abney.

Shillcock and Bard (1993) have provided evidence that open- and closed-class items differ with respect to their lexical access, closed class items show evidence of top-down syntactic influence while there is no such effect for open-class items. They suggest that this difference in the degree of modularity between the two word systems suggests that modularity is an adaption to the different processing requirements of the two classes and that this suggests that modules develop in response to the nature of language. Such a view is consistent with the constructivist approach adopted here in which linguistic structure serves to guide cognitive development. According to the theory presented here, morpheme frequency may be one factor which acts as a catalyst in the division of processing between the open- and closed- classes - this initial distinction will be considerably simpler than the adult representation which will be based on more subtle and higher-order features of the different word classes. I stress once again, that the proposed role of word frequency in the initial stages of learning does not necessitate an explicit frequency-based distinction in the adult's linguistic representation - there need be no rules mediating frequency of production. Rather, the various properties of the closed-classes which I have reviewed - their resistance to new members, their roles as heads of phrases and their promiscuity with regard to different semantic contexts, can ensure that they will be produced with a high frequency and serve as markers of syntactic structure.

3.4 Universality

As noted by Green (1979), the marker hypothesis requires that markers be a universal feature of languages. Morgan *et al* (1987) argue that this is the case, although they note that language differ in their use of markers. For example, some languages use concordance morphology (where syntactically related words share the same word ending) while others do

not. The main research in this thesis is focused on English but I will briefly address the issue of universality of marker structure.

Zipf's law has been found to apply across many languages and at different levels of linguistic description - given the fact that such a statistical pattern may emerge from a stochastic process this may be viewed of little import. However it does have important ramifications for the marker hypothesis. Firstly, it ensures that the most frequent elements in a language will stand out as such *i.e.* it will not require a very accurate frequency count to detect the set of most frequent elements as the distinction will be a gross one. Secondly, the fact that a small number of the most frequent word types will account for a high proportion of the word tokens which are heard has important implications for the differential treatment of these items from the perspective of computational efficiency. Both of these points have important implications for the distinct types of processing suggested by the marker hypothesis. Aside from the statistical requirements for markers, we also need to show that these frequent items meet the grammatical requirements for their use as structural markers.

It does seem to be the case that the most frequent words in different languages are closed-class elements. For example the following table shows the 20 most frequent words in English, French and German newspaper writing (Crystal 1987):

Rank	1	2	3	4	5	6	7	8	9	10
English	the	of	to	in	and	a	for	was	is	that
French	de	le	la	et	les	des	est	un	une	du
German	der	die	und	in	des	den	zu	das	von	für

Table 3-1 Ten most frequent words in written English, German and French

However, it is important to account for the notion of frequency with respect to morphemes rather than words. When dealing with written language, it is easy to focus on words as the main linguistic unit. Grosjean and Gee (1983) have warned against adopting such a view of language based on what they call the written dictionary word. The marker hypothesis , like Garrett's model of speech production, gives equal status to closed-class words and bound morphemes as in, for example, the grammatical frame THE ____ IS ____-ING THE ____. Garrett noted that both kinds of morpheme are resistant to movement in stranding exchanges such as 'I thought the park was truck-ed'. The computational model presented in this thesis does not, for practical reasons, make use of bound morphemes but this approach is not ideal and is only possible because English is not highly agglutinative as is, for example, Swahili. A word based approach may be most fruitfully applied to an isolating language such as Vietnamese, but it is important to realise that, in order to be applied universally, the marker hypothesis must deal in terms of morpheme frequency. For this reason, it will be necessary ultimately to place the marker hypothesis within the context of a model of speech segmentation which could capture the intuitive notion that bound grammatical markers will be amongst the most frequent morphemes.

The linguistic role played by marker elements varies across languages - I have already mentioned the case of concordance morphology, but languages also vary in the types of functional categories which they employ and the relationship between functional elements and the phrases with which they appear. In Japanese, postpositions act as marker elements, for example in the sentence:

Watashi wa Igirisujin desu

I [subject] English am

'I am English'

the postposition 'wa' serves to mark the subject of the sentence. These marker elements share certain features with markers in other languages, notably:

1) They have high-frequency.
2) They mark particular syntactic categories.
3) They are omitted in early child language (An issue which will be addressed in the next chapter).

However, whereas in English noun phrases are marked by determiners and number inflections, Japanese lacks both of these functional systems (as well as a case system); using markers for roles such as Subject and Object instead. The general trend across different languages to use markers that signal different functional properties militate against an account which uses such linguistic properties in the initial acquisition process, instead, I suggest, marker elements are best characterised by a number of (fairly transparent) properties that emerge from a statistical analysis of utterances, the most obvious being the frequency of the element (See Cutler (1993) for other salient properties). Therefore, the structural marking provided by these elements is distinct from, and available earlier than, their functional role in the linguistic system of the language. Thus the child may initially use different functional categories (determiners, subject markers, case markers etc.) in the same way on the basis of their statistical salience and only later distinguish between them in terms of their linguistic function.

The Japanese example also highlights that markers may vary in their position relative to that which they mark - whereas an English Noun phrase is generally marked by a determiner at the start of the phrase, Japanese postposition markers appear after that which they mark. The important point from the perspective of the marker hypothesis is that languages are consistent on this point - generally being head-first or head-last. This consistency reduces the number of possible alternatives that a learner may have to consider - they merely have to note a general trend in the relative position of markers to phrases. It is arguably unneccesary

for the learner to have an innate language-specific constraint in order to note that a high-frequency marker will appear consistently before or after that lexical phrase with which it is associated.

3.4.1 Free Word Order

Languages vary with respect to the freedom of their word order. This raises some additional insights into the marker hypothesis. In languages with a more free word order (for example Latin) there appears to be a greater reliance on bound morphemes - this would ensure that lexical units will still be reliably associated with a particular marker element even when their position in a sentence may be variable - in a sense, the lexical unit takes its marker with it wherever it goes. This is further emphasized by the fact that even unbound closed-class items seem to be much less 'free' than other words in free-word order languages. An example of this is provided by Russian prepositions which are constrained to appear before their associated phrase, even though the language as a whole allows great variation in word order (Covington 1990). Indeed, it may be the case that it is only because of rigid marking requirements such as this that such languages can be 'free'. The historical changes in English word marking, which I discussed in the last chapter, show that a loss of case-marking was associated with a more rigid word order - when explicit case markers were used word order was more variable but when they were lost from the language the information which they had provided was replaced by word order information. This suggests that the learner can make use of the position of an element in a sentence or relative to a particular marker. The relationship between markers and word-order is also illustrated by the concept of basic word order in a language. Generally we find that the greater the departure from the basic word order the greater the degree of marking. In fact, Matthews (1981) defines basic word order as that which is least marked in a language. For example, consider the basic sentence 'John loves Mary', which contains the single marker element *-S*, with the passive equivalent 'Mary is loved by John' which contains the markers *IS*, *-ED* and *BY*. This suggests that more

complex linguistic forms require greater marking and indeed this extra marking may, in turn, serve as an indication to the learner of the syntactic complexity of a particular utterance.

In conclusion, I would suggest that the marker hypothesis offers us a valuable new insight into linguistic universals - from its fairly simple premises we can see that languages which are otherwise rather different appear to adhere to a few basic requirements that are necessary, according to the marker hypothesis, for learning to take place. The hypothesis makes sense of a fact which is otherwise hard to reconcile. To give an example, in English, the determiners *the* , *a* and *an* are amongst the most frequently used words in spoken or written language which may lead us to believe that their linguistic role is an important one. However, if their role in marking the distinction between definite and indefinite reference is so important as to merit their such common use in English, why is it the case that there are other languages (such as Japanese or Latin) which have no comparable functional marking - distinguishing between definite and indefinite only in certain special circumstances. There are numerous examples of how different languages make obligatory use of frequently occurring morphemes to signal linguistic distinctions that are not considered necessary in other languages. Why make so much use of items if they are not strictly necessary for conveying meaning? The marker hypothesis offers a better explanation for the frequent occurrence of these elements than a conventional linguistic description. The reason that such markers are so frequent and obligatory is that they are vital for language acquisition, while their linguistic role may sometimes seem arbitrary, their role as structural markers is essential. So different languages seem to vary in terms of what functional information they signal but they agree in the necessity of using frequently occurring marker elements to signal phrase structure. Thus although an English determiner and a Japanese post-position subject marker differ greatly with respect to their linguistic properties, they are essentially the same in terms of being frequently occurring marker elements which are reliably associated with a particular type of phrase.

3.5 Conclusion

This chapter has considered general features of the marker hypothesis and the relationship between frequency and word function. I have presented the marker hyptothesis which holds that high-frequency elements in natural language serve to focus the attention of the learner on important syntactic information by both segmenting speech into linguistically relevant units and labelling such units with category information.

In the next chapter, I will consider the status of the current hypothesis with regards to children. Do children have the ability to make use of grammatical markers, and is there evidence of them using an approach such as that outlined in this chapter?.

Chapter 5 will consider the grammatical pre-requisites for the theory in more detail - what kind of linguistic representation best explains early acquisition and how does this relate to adult language.Chapter 6 will then present the results of a computational analysis of a natural language corpus based on the use of marker elements.

4. The Frequency Filter

In Chapter 2, I considered the adoption of a constructivist theory of language acquisition, one consequence of such an approach is that the learning process must be considered incrementally. Later, more complex, constructions are learned in the context of previously acquired structure - the learner's representational space is changing gradually and the shape of this space affects the way in which input is treated. So, a one-year-old and a three-year-old will apply different analyses to the same input and this difference lies in the changes which occur in the representational space in the intervening period.

One of the key claims of the Poverty of the Stimulus argument is that the learner must deal with 'rare and complicated input' an example of which is provided by the two sentences 'John is eager/easy to please' - the question being, how could a learner acquire the distinction in syntactic structure of the two forms? A problem with such a question is that it presents the problematic construction in isolation from the rest of language development - it makes the claim that only by having some innate knowledge of language could the learner cope with such forms but within the constructivist framework we can argue that rather than requiring innate knowledge, the answer lies in the representations which the learner will have developed prior to attempting to learn such sentences. Children take time to master such difficult forms and it is futile to attempt to address the question of how such mastery arises without first addressing the question of how simple sentences are mastered.

If children used such rare and complicated forms from an early stage it would support the adoption of a nativist account but such forms are used to support nativism in a different way - by presenting examples which are difficult to explain from an empirical standpoint. However, the constructivist approach would suggest that we cannot tackle questions of later grammatical development before showing how the child's representational space is moulded in the early stages of development. Thus, we should not set out by going down the unreasonably long road of providing a 'batteries included' acquisition model (something which no approach has succeeded in providing). Instead we should try to develop an understanding of the incremental nature of language learning.

The solution adopted here is to provide an account of early child language. If successful, this account would show that the structure of such language may be predicted by a distributional approach and although this would not explain the acquisition of full adult grammar, it would nevetheless provide strong grounds for arguing that humans use the learning approach.

Child grammar differs from adult grammar. Within the constructivist approach we may argue that the structure of the early grammar will reflect the representations and processes which have been acquired by a certain stage and which underpin later development. I will argue that the marker hypothesis can enable us to make certain predictions about the nature of early syntax which can then be compared against examples of actual use.

Such predictive approaches to the structure of child language have been presented before - for example Brown (1963) related the time course of the development of certain linguistic structures to their semantic complexity and Radford (1990) applies a GB-based account to the structure of early child syntax.

The task of analysing child language is a huge one which may be approached from many angles. Luckily, there exist many previous studies and I will focus on the analysis provided by Radford (1990) and offer a reanalysis of his data in terms of the current theory. Any analysis of corpus data will be selective, and I hope that by basing my study on examples

chosen by Radford, whose theory is radically different from mine, I will avoid the potential pitfall of handpicking data which may suit my argument.

Early child language is different from adult language is certain systematic ways. I will argue that it is characterised by a selectivity which arises from an interaction between the structure of language and limitations in the learner.

Children do not analyse adult sentences as whole entities, rather we can think of the child as viewing sentences through a window which is constrained by locality and by marker elements. This windowing process enables them to focus on constrained areas of the input and the objects of this focus provide them with an initial set of syntactic 'building blocks'.

The coverage in this chapter is not aimed at explaining the reason for such structures to be acquired first but to demonstrate that a simple distributional approach based on the marker hypothesis may explain the difference between adult and child forms. Neither do I consider the form in which the child represents their initial linguistic structure - this question will form the basis of the next chapter.

If the marker hypothesis is to extend beyond the confines of the artificial learning situation (whether through MAL experiments or computational models) it is necessary to show that children can and do use marker elements in the early stages of syntactic acquisition. In this chapter I will consider evidence to support such a view.

Few language learning models have attempted to explain the incremental development of syntax in humans. Generally, the aim of such models is to explain how a grammar may be acquired - little attention has been paid to explaining the incremental patterns of acquisition observed in children. The constructivist approach which I have adopted argues that the structure of language guides the development of grammar.

4.1 Child language

4.1.1 Telegraphic Speech

One of the most striking observations about children's early language, particularly from the current perspective, is its telegraphic nature *.i.e.* children generally omit closed-class elements including function words and inflections. Thus they say 'hit ball' rather than 'hit the ball' and 'daddy car' rather than 'daddy's car'. These omissions occur both in their own spontaneous speech and in their apparent attempts to imitate adult speech - an example of the latter is given in the following example:

ADULT: I can see a cow.

CHILD: See cow.

The lack of functional elements in child speech has been seen as evidence that they cannot use them. Several possible reasons have been suggested as underlying the lack of functional elements.

4.1.2 Theories for omission of functional elements

4.1.2.1 Prosodic

One argument is that strongly stressed syllables are more salient to the child and that they therefore tend to focus mainly on open-class words which generally contain strongly stressed syllables and either fail to perceive, or do not analyse, closed-class elements which are generally realised as weakly stressed single syllables and are often phonologically reduced in speech. An extension to this theory argues that omitted morphemes are not merely those which are unstressed but are ones which appear in a certain context or stress pattern.

4.1.2.2 Semantic

Semantic accounts focus on the differences in complexity and concreteness between lexical and functional categories. Brown (1973) suggested that children don't use grammatical morphemes because they have a greater semantic complexity than content words. He showed

that the appearance of the first 14 grammatical morphemes can be explained in terms of increasing complexity.

Brown analysed children's early sentences in a number of languages. He suggested that they can mostly be classified as belonging to one of eight broad groups based on the semantic relationship which they encoded.

action-object: 'Hit ball'

agent-action: 'mummy kiss'

agent-object: 'mummy doll'

action-locative: 'sit chair'

entity-locative: 'cup table'

possessor-possession: 'daddy car'

entity-attributive: 'big car'

demonstrative-entity: 'that car'

Pinker (1982) argues that the essentially lexical nature of early utterances reflects the primacy of semantic reference in early acquisition, thus children initially learn words which have real-world referents and only later do they begin to analyse functional elements.

4.1.2.3 Syntactic

Radford (1990) claims that children's early grammars lack functional categories systems. This account is based on that version of GB theory which argues that functional elements act as heads of phrases which was reviewed briefly in Chapter 3 (see Abney, 1987; Cook & Newson,1996). According to Radford's analysis, children's early language consists purely of lexical categories and lacks functional categories and their associated projections - so while an adult might use the DP 'The big dog', the child would use only the lexical NP 'big dog'. I will be considering Radford's account in some detail later in this chapter.

4.2 Perception vs Production

All of these views equate the lack of functional elements in productive speech with a general inability to process such elements. Such an inability would obviously render the marker hypothesis untenable as a theory of early syntax acquisition. Given the frequency and potential usefulness of function words in language acquisition there is something inelegant about the idea that children cannot use them - as Anne Cutler says, concerning the processing of closed-class elements in adult speech:

> *...it would be highly surprising if they were hard to process; recall that closed-class words make up more than 50% of all word tokens occurring in typical speech samples (Cutler & Carter, 1987). If this high a proportion of all words we hear were to cause processing difficulty, then at the very least one might feel that our processing mechanism was not functioning optimally.*
>
> (Cutler, 1993, p. 117).

The same appeal to processing efficiency can be made regarding their role in acquisiton - there has been in the past a tendency to dismiss the role of these elements in language - each of the theories of non-use mentioned above share the common feature that they interpret the child's non-use of function elements in their own speech as evidence of a failure to use them in the analysis of adult speech. However, an alternative view, is that while children may not produce function words in their own speech, they may nevertheless be capable of perceiving and utilising function words. Evidence for this view comes from a number of experiments which have shown that children who do not use function words in their own speech are, nevertheless, able to both perceive and make use of function words in understanding language.

4.2.1 Psycholinguistic evidence for comprehension

Peretric & Tweney (1977) and Shipley, Smith & Gleitman(1969) found that children who omitted function items from their own speech were more likely to respond appropriately to strings containing functors ("Give me the ball") than without ("Give ball"). Katz, Baker &

Macnamara (1974) and Gelman & Taylor (1984) found that 18-24 month olds would distinguish between proper and common nouns on the basis of whether a determiner was present ('The dax') or absent ('Dax') and Eilers (1975) found that children were less likely to omit functors in sentences with scrambled word order.

Gerken & McIntosh (1993) found that children whose own speech was telegraphic nevertheless performed better in a picture identification task whenever the target word was preceded by a grammatical article than an ungrammatical auxilary. This effect was facilitated by the use of a female voice.They argued that this showed that prosodic and morphological cues work in tandem with an interplay occurring between the prosodic characteristics of the female voice and functional elements. This study suggests that children can distinguish between grammatical and ungrammatical placement of function morphemes in perceived speech even when such elements are not used productively in their own speech.

Gerken, Landau & Remez (1990) studied children's omissions in imitative speech. They studied two groups of children - those with a high mean length of utterance (MLU) and those with a low MLU (typically telegraphic speakers). Children were asked to imitate a range of sentences which contained either real or nonsense functors and real or nonsense content words as illustrated in the following table:

Content Word	**Functor**	**String**
English	English	*Pete pushes the dog*
English	Nonsense	*Pete pusho na dog*
Nonsense	English	*Pete bazes the dep*
Nonsense	Nonsense	*Pete bazo na dep*

Table 4-1 Example stimuli from Gerken et al *(1990)*

Different experiments were run using both human speech and speech generated by the DECtalk speech synthesis system in order to control for phonological factors. The three experiments obtained similar results. The low MLU children omitted English functors more often than the nonsense functors even though the latter were matched for phonology and occurrence positions. Both groups of children omitted significantly fewer of either the

English or nonsense content words than the English functors. They also found that the presence of English functors helped the imitation of content words for both low and high MLU children.

These studies suggest that children who do not use function words in their speech are, nevertheless, capable of using such words in understanding adult speech and provides counter-evidencec to many of the arguments for the omission of functors presented in the previous section. The fact that in the Gerken *et al* (1990) study children produced the nonsense functors more than the real functors, despite the fact they were controlled for stress and position suggests that the phonological argument is dubious - children did produce the nonsense functors despite their weak stress. If anything, the phonological argument would predict the opposite result since the nonsense functors suffered the additional disadvantage of unfamiliarity while children would have had much more exposure to the real functors and thus much more of a chance to overcome the disadvantage incurred by weak stress.

The semantic argument is also called into question as the nonsense functors and content words had no semantic content but were imitated by the children significantly more than the English functor words which have at least some, albeit abstract, semantic content.

Additionally, the work by Gerken & McIntosh shows that children are able to distinguish between, and make use of, real English functors which runs contrary to the idea that children cannot encode these elements.

4.2.2 The Child's Metalinguistic Theory

An additional source of evidence which is relevant to the current account comes from studies of children's metalinguistic awareness of the words in their language. Karmiloff-Smith (1992) describes how children up to the age of about 6 do not seem to have explicit knowledge that closed-class words are words - for example in tasks that require children to count the words in a sentence or to repeat the last word spoken they often seem to overlook closed-class items. This phenomenon occurs for a considerable time after children show clear

evidence of being able to use closed-class elements. In the context of the current theory, this distinction may be explained in terms of the separate processing requirements for lexical and marker elements - children, according to this view, distinguish between 'words' (open-class words) and things which mark 'words' (closed-class words).

4.3 A Distributional, Frequency-Based Account of Omissions

Gerken, Landau & Remez (1990) suggested a phonological explanation for children's omission of functors. They cite Allen and Hawkins' (1980) argument that children show an initial difficulty in alternating between weak and strong syllables in speech production and that, in particular, they have a preferred production pattern of a strong syllable followed by a weak syllable so that, for example, *giraffe* is more likely to be reduced to *raffe* than *monkey* to *mon*. Gerken *et al* had found that their subjects, when given a sentence such *Pete pushes the dog* were more likely to omit the second functor *the*, which occurs in a weak-strong pattern, than the first functor *-es*, which occurred in a strong-weak pattern. In order to explain the finding that children omitted real functors more often than nonsense functors (which occurred in equivalent stress patterns) they argue that the nonsense functors were treated as part of the adjacent content word - this reduced the number of morphemes and therefore the complexity of the utterance. So omission, they claim, is explained by a combination of stress pattern and morphemic complexity.

4.3.1 An Informational Account of Omissions

An alternative account of function word omissions may be provided within the context of the marker hypothesis. I stated earlier that children seemed to omit from their speech precisely those elements which are considered to be marker words. Given the findings described above which suggest that children are capable of perceiving functional elements even when they do not produce them we can retain the marker hypothesis. Furthermore, we can postulate that children specifically omit marker elements from their own utterances while at the same using these elements in order to isolate and label phrases in adult speech. For example, I suggest

that children isolate grammatical frames and use these to analyse lexical elements - the adult sentence 'The man is happy' will thus be deconstructed into the grammatical frame 'The ___ is ___' and the lexical units 'man' and 'happy'. It is only these lexical units which are used in the child's own equivalent sentence 'Man happy'. To continue with this analyis we have to identify the criteria by which a child would identify marker elements in the speech stream.

At around the age of 12 months children begin to utter single words, this stage shows no sign of word combination and therefore little evidence of syntax and so most of the studies of this stage have focused on the semantics of the words used. Alison Bloom's speech at ages up to the age of 22 months contained mostly words from the four lexical categories - noun, verb, preposition and adjective (Bloom, 1973). However, as Radford (1990) suggests, the use of words at this stage may be acategorial as there is little evidence of syntactic combination. One thing that we can claim is that such lexical elements are the most useful semantically.

It has been noted previously (Anglin,1977; Brown, 1958) that children's first nouns tend to be those which are at a level which is most useful semantically - usually an intermediate category, so they will tend to use 'dog' rather than the more general 'poodle' or 'terrier' or the less general 'animal'. Rosch *et al*'s (1976) basic-level categories are those which allow the greatest degree of semantic functionality according to a number of measures. It would therefore seem appealing to assume that the child would use such categories first because, given a limited vocabulary, they would offer the most useful set of lexical elements - this is, after all, implicit in the term 'telegraphic' . There is, however, a problem with such an account - by describing children's early words in terms of their semantic usefulness we assume that the child has learned the correct meaning of the word from the beginning, however there is much evidence to suggest that this is not the case. Words are often over-extended in meaning, so that, for example, 'dog' may be applied not only to dogs but to other animals or other objects which share some feature of dogs or even the situation in which dogs have been observed. (Clark, 1983). Children also underextend a words meaning - for example using 'dog' to refer only to the family pet. (In fact children's errors do not appear to

fit neatly into a simple hierarchical model but often seem to involve certain shared features between objects). The fact that children misapply words to higher level categories, lower level categories or across adult category boundaries raises the question of how they learn to first use words at an optimal level of functionality. If a child has not acquired the correct semantics for a word then they cannot employ those semantics as the criteria by which they select words of basic level categories first.

The argument that the child's early vocabulary choices are based on their being at a functionally-optimal categorial level depends on the child having a correct (i.e. adult equivalent) semantic representation for those words and yet this would appear not be the case. If the child's own semantic representation does not underlie their initial vocabulary choices then what other source of information could?

4.3.2 Luhn's Theory

The analysis of Luhn, introduced in chapter 3 may prove useful here. Luhn provided a simple measure for establishing a set of useful keywords to be used in the process of automatic document classification (Luhn, 1958; Schultz, 1968). The aim of establishing a set of useful keywords is to enable the grouping together of documents with related topics and the discrimination of documents with different topics. Luhn argued that words of median frequency tended to be the most useful keywords for this task. Words of a very high frequency tend to occur across all documents and will therefore be of little use in distinguishing between documents while words of a low frequency will not be sufficiently common to enable the cross comparison across documents. The key point here is that Luhn used a simple statistical measure to isolate a set of words which were particularly useful in distinguishing between document topics - no recourse is made to the semantic content of the words. Luhn's analysis can be applied to the problem of how children select a set of semantically 'useful' words prior to actually establishing the relevant semantics. By replacing the notion of document topic with that of conversational topic or referential focus we can apply the analysis to the current problem. Words which occur infrequently may be

useful to a person who knows the meaning of the word - but the child or the automatic document classifier, relies on cross-comparisons across different uses of a word in order to establish meaning, and such cross-comparisons are only possible for words which are sufficiently frequent to occur in a number of different contexts. On the other hand, words of a very high frequency will be so 'promiscuous' with respect to the contexts in which they appear that they will not allow sufficient semantic discrimination between them. This leaves us, again, with words of median frequency which appear in a number of contexts which is large enough for cross comparison and yet small enough for discrimination between topics. While the original analysis extracted useful keywords, this analyis extracts a set of useful vocabulary items - in both cases the process requires no knowledge of the meaning of the words.

To establish what kind of range we are talking about when using the term median frequency, recall that the top 150 or so words in English are predominantly closed-class - the words which immediately follow these are the higher-frequency lexical items and it is these which, according to the current analysis, will constitute the most useful vocabulary items for the inital stages of language. Furthermore, we would expect those object names which occur in this range to be those which are at the most useful conceptual level ('dog' rather than 'poodle' or 'animal') because they will tend to be used more by adult speakers precisely because of their semantic utility.

4.3.3 Frequency-Based Omissions

Can high-frequency explain children's omission of function words? Let us first consider the one-word stage. The following lists show the first 50 recorded spoken words from two American children, Daniel and Sarah (from Crystal, 1987). Beside each word is listed its rank frequency from the adult to child speech of the CHILDES corpus (zero indicates a rank below 600).

	Rank Frequency	**Word**	**Rank Frequency**
no	20	oh	12

up	39	no	20
big	80	whats	27
nice	83	down	63
baby	91	baby	91
more	99	more	99
daddy	103	daddy	103
my	106	mommy	113
mommy	113	book	116
hello	115	ball	143
ball	143	car	144
hi	148	hi	148
horse	158	doggie	161
bye	164	bye	164
nose	169	nose	169
night	222	door	196
kitty	227	box	201
bunny	234	hat	207
milk	261	dolly	219
juice	296	night	222
block	319	kitty	227
apple	333	bird	235
walk	345	duck	258
hot	352	milk	261
shoe	382	juice	296
orange	418	done	316
light	419	apple	333
bottle	420	hot	352
woof	421	teddy	365
uhoh	429	shoe	382
eye	432	orange	418
cookie	439	bottle	420
rock	483	eye	432
rock	483	cookie	439
whats that	0	bath	464
wow	0	paper	473
banana	0	rock	483
moo	0	cracker	0
quack	0	alex	0
clock	0	cheese	0
sock	0	wow	0
bubble	0	button	0
fire	0	alldone	0
yogurt	0	coat	0
pee	0	bib	0
whack	0	ear	0
frog	0	toast	0
yuk	0	O'toole	0
ernie	0	leaf	0
nut	0	lake	0

Table 4-2 First 50 words of two children ordered by rank frequency

In both lists, only a small number of words are amongst the 150 most frequent (which account for a disproportionate number of the word *tokens* heard by a child) and there are also relatively few from the range below the 600 most frequent (which account for most of the

word *types* heard by the child). Most of the children's words occur in a range between the 100th-600th rank positions. The highest frequency words used are not typical of other high frequency words in terms of their distribution (for example they are more likely to be used in isolation (*no!, oh!, up!, more?)* or in conjunction with an even higher frequency word (*a baby).* Given the fact that the first 100 word types account for approximately 60% of all word tokens heard and that children appear to be able to perceive and make use of functional elements, their tendency not to use high frequency words may lend some support to the view that the use of such elements in productive speech is suppressed due to their very high frequency. Combined with the relative lack of lower frequency elements this data would seem to support the idea that children's first words are predominantly those of median frequency.

4.3.4 Input Filtering and Input Streams

Obviously, there is so much more to the knowledge of a word than its frequency but what I am suggesting is that word frequency - both the individual frequency of a word (or morpheme) and the differences in frequency between adjacent words in the speech stream - can act as the basis for an initial filtering of the speech stream into two separate streams - a high frequency stream which corresponds closely to a syntactic frame and a low frequency stream which consists primarily of lexical elements. The low frequency stream provides the raw material for the child's semantic learning and contains the elements which are 'cross-referenced' with the environment while the high-frequency stream is used to structure the elements in the low frequency stream and thus corresponds to syntactic information; as Newport, Gleitman & Gleitman (1977) put it:

> *The child has means for restricting, as well as organising, the flow of incoming linguistic data; he filters out some kinds of input and selectively listens for others.* (p. 140)

4.4 The Frequency Filter Analysis of Child Language

I will now consider how we may apply the theory outlined in the previous section to an analysis of children's early utterances through the use of a computer program which will simulate the way in which children might distinguish markers from non-markers in perceived speech. The basic components of the theory are:

- The learner treats a sentence on two levels - a 'background' of functional, marker elements which bounds a foreground of lexical 'objects'.
- Markers are chosen on the basis of their frequency and the frequency of neighbouring elements.
- Children first use elements which are marked in the input and omit items which are themselves markers.

The 'frequency filter' approach is not a learning model but is designed to filter out the marker elements from a sentence. It can thus be used to examine the claim that children omit marker items from their speech.

How does the learner identify marker elements? One answer would be to simply treat any word whose overall occurrence frequency exceeds some threshold as a marker (e.g. the top 150 words). Given the fact that some lexical elements occur within such a high frequency bracket and are used by children, such an approach may be too simplistic. We might instead consider the relative frequencies of words in a sentence - the status of a word may depend on the relative frequency of the surrounding words so that a particular word may sometimes act as a marker and at other times act as a marked element. There is a parallel between this idea and the use of intensity changes to detect object boundaries in vision research in which relative light intensities are more useful than absolute ones. It is this approach which I have followed in producing the following program.

The frequency filter program (hereafter referred to as FreqFilt) is extremely simple. The only data used by the program is a ranked list of the 200 most frequent words from the CHILDES

corpus adult to child speech. The actual frequency counts are not included - the program generates a frequency for a word based on its rank according to a Zipfian function. The following is an outline of the program:

- A sentence is read in as a list of words and a marker symbol (##) is added to the beginning and end.
- Each word is then checked to see if it is one of the 200 most frequent words - if it is, it is assigned a frequency value which is a calculated as 200/rank. The begin and end symbols are given a frequency value of 200 and all other elements are given a value of 0. The process by which the system distinguishes markers from non-markers is described by the following pseudo-code:

for each word (n) in sentence: change(n)=freq(n)-freq(n-1)
for each word (n) in sentence: second(n)=change(n)-change(n+1)

words are allowed to pass through the filter (ie. treated as lexical):
if change(n)<0 and change(n+1)>=0
or
if the last word was lexical and change(n)>=0 and change(n+1)>0

Finally, a 'confidence factor' is calculated by:
if freq(n)=0 then cf=100+second(n)
else
cf=second(n)/freq(n)^2

This value gives some indication of how likely a word is to be treated as a marker element.

In any implementation such as this there will be arbitrary decisions made, but the program is intended to capture the following notions:

- Only approximate frequencies for the most frequent 200 words are 'known' - all other words are considered equal except in terms of their occurrence in a particular sentence.
- The only other distributional information used is the change in frequencies between adjacent words in the current sentence - no record is kept of this between sentences nor does any other form of learning take place. The status of an element as a marker depends only on its frequency and the frequency of the adjacent words.

- FreqFilt is not a model of learning, but is simply intended as a 'first pass' filter which distinguishes marker elements from non-markers.

- It is hypothesized that marker elements will be omitted from children's language.

Thus the only language specific knowledge contained in the program is a list of the 200 most frequent words. FreqFilt outputs the non-marker elements of a sentence as a simulation of child imitative speech. It outputs a value with each word which constitutes a (rather arbitrary) confidence factor of that word being a non-marker (lexical element). For example, given the input sentence is 'I want a cup of tea', FreqFilt outputs the following:

want (1) cup (171) tea (304)

In this case, the words *I, a* and *of* have been identified as marker elements and been filtered out of the output, for the remaining words the figure in brackets indicates the degree of 'lexicality' or 'open-classedness' attached to that word - so *cup* and *tea* are consider more likely to be lexical items than *want*.

4.4.1 Relative versus absolute frequency

The decision to employ a measure based on relative frequencies between words rather than simply on a word's absolute frequencies was motivated by a number of factors.

Human perception. The perception of the intensity of a stimulus is affected by the context in which it appears. The Just Noticeable Difference (JND) between two stimuli is a constant fraction of the size of the stimuli, so that for example it easier to judge the heavier of two weights weighing 5 grams and 10 grams than for two weights of 100 grams and 105 grams even thought the absolute difference is the same in each case (Goldstein, 1980). In visual perception, the perceived brightness of an object depends on the brightness of the background against which it appears and edge detection depends on the relative light intensity between adjacent areas rather than their absolute light intensity (Marr, 1982).

Evidence from artificial language learning. Valian and Coulson (1988), discussed in Chapter 3, suggest that it is the relatively high frequency of marker elements compared to

other elements which gives them their salience. The high frequency elements within phrases act as 'anchor points' for those phrases. If a child is using a strategy of assigning structure to phrases based on markers they may use the best marker available based on that element which has the highest frequency. The results from Valian and Coulson's experiment suggest that adults make use of items of relatively high frequency within the task domain even though the frequency with which they have been exposed to those items (nonsense words) will be much less than for words in their natural language. This suggests that people have the ability to quickly establish the relative frequency of items within a given context and make use of these distinctions in distributional learning.

4.4.2 Radford's Maturational Theory

The FreqFilt program can be used to provide a more concrete demonstration of how the marker hypothesis may explain the structure of early child utterances. Radford (1990) provided an account of child language in terms of a nativist, GB theory based account in which he argues that early language use displays "the acquisition of lexical category systems and the concomitant non-acquisition of functional category systems" (p 83). Radford considers each of the functional systems in GB theory in turn and presents evidence that such systems are not realised in children's speech.

The current approach divides things up differently - the overarching theory is that marker elements are omitted and this is subdivided according to particular marker elements.

I will provide a non-nativisit alternative to Radford's theory which I believe will be superior in that :

- it is simpler.
- it explains why children use the forms they do.

- it explains the ontogenetic component, which I argue is primary but which is not addressed by Radford, despite the fact that it is still necessary in a nativist account.
- I will provide an account of specific differences in use and non-use of particular forms which Radford's account cannot explain.

Radford considers the following systems in turn

1. Determiner system
2. COMP system
3. Inflectional system
4. Case system
5. Grammar of missing arguments

I will deal with each of these separately but argue that they can be explained by the same simple process in each case.

4.4.3 Absence of a Determiner system

According to Radford, the child uses simple NPs where the adult would use a DP. The child NP is lexical and lacks the functional projection into DP. So whereas the adult speaker might say [dp [d THE] [np OLD MAN]] the equivalent child form would consist simply of the bare NP - [np OLD MAN]. Radford argues that all elements that serve a determining function (a/the/this/that/some/any/no/much etc.) will be omitted from the child's purely lexical constructions. The lack of functional category systems will also result in the non-use of the genitive - *'s* and *of*.

The marker hypothesis offers an alternative explanation. In Radford's examples there is a small set of omitted functional elements - determiners such as 'the' and 'a', the genitive '-s' and 'of' all of which are high frequency elements which will be treated as markers, while nouns and adjectives are less frequent and will be treated as lexical units. So 'The old man' will be analysed into the frame [The ____ ##] and the lexical unit 'old man' - only the lexical unit is used in the child's own speech. Thus the child's production of NPs will be internally mediated by the implicit marker elements - a frame such as [The ____ ##] will be

associated with a number of lexical elements e.g. old man/man/dog/cat/house. Such an account is similar to Garrett's model of speech production (see previous chapter) in which utterances are generated by slotting lexical elements into a functional frame, the difference being that in this case the functional frame is not phonologically realised but is simply used to control, or structure, the production of the lexical elements. This account does not require us to posit innate functional and lexical systems, rather such systems are seen as 'progressively modularising' in response to the structure of the input. Furthermore, initially the child's internal syntactic representation would not be framed in terms of traditional linguistic categories but would be tied to individual marker elements (a 'the' phrase, an 'a' phrase etc.) and the fusing of these constructs into a single 'DP' form would not be a linguistic 'given' but would arise from further distributional analyis on the marker elements (the details of such an account will be considered in Chapter 6). Note that according to this argument, simple NPs such as 'big car' would be treated as single lexical units while genitives and 'of' forms would consist of two.

The big dog -> The ____ ##

The jar of coffee -> The ___ of ___ ##

mummy's car -> ## ___ 's ___ ##

Frequency Filter Analysis

Radford gives numerous examples of child utterances coupled with adult equivalent utterances. These adult equivalents are either actual sentences which a child has attempted to imitate or are inferred from what is thought to be the child's intended meaning. By feeding these adult versions through the FreqFilt program we can compare the predictions it makes with the child's reduced form.The following are a set of such examples from page 85 of Radford - they show the adult form, the child's utterance and the reduced form of the adult version made by FreqFilt (with confidence factors in brackets after the words):

Adult Equivalent **Child Utterance** **FreqFilt output**

did you drop your tea	**Drop tea**	did (12) drop (311) tea (311)
is there a baby in there	**Baby**	there (0) baby (16) there (0)
you want the top off	**Top off**	want (4) top (201) off (51)
has he got any legs	**Legs**	has (113) legs (301)
this is a kitten	**Kitten**	this (1) is (-1) kitten (366)
teddys in the boot	**Boot**	teddys (318) in (0) boot (400)
is it a duck or a chick	**Duck**	Is (0) it (0) duck (167) or (39) chick (366)
is that a drum	**Drum**	that (0) drum (366)
did you go down the slide	**Slide**	did (12) down (10) slide (400)
thats a bag look	**Bag**	thats (3) bag (171) look (7)
mummy will take the nut off	**Nut out**	mummy (138) will (0) take (16) nut (201) off (51)
you were playing in the water	**In water**	playing (119) in (0) water (400)
thats a goat	**Goat**	thats (3) goat (366)
tell him hes a naughty boy	**Naughty boy**	tell (59) hes (14) naughty (168) boy (57)
thats a cup	**Cup**	thats (3) cup (366)
paint a rabbit	**Rabbit**	paint (366) rabbit (366)
its a brush a hairbrush	**Brush**	its (11) brush (233) hairbrush (366)
thats an apple	**Apple**	apple (301)
thats a hippo	**Hippo**	thats (3) hippo (366)
its the sun	**Sun**	its (12) sun (400)

Examples from Radford (pg 91):

a cup of tea	**Cup tea**	cup (171) tea (304)
a bottle of juice	**Bottle juice**	bottle (171) juice (304)
a picture of gia	**Picture Gia**	picture (171) gia (304)
i want a piece of the chocolate bar	**Want piece bar**	want (1) piece (171) of (4) chocolate (200) bar (300)
i want to have a drink of orange	**Have drink orange**	want (0) have (1) drink (171) orange (304)
i want a cup of tea	**Want cup tea**	want (1) cup (171) tea (304)
a picture of kendall	**Picture Kendall**	picture (171) kendall (304)
the colour of the crate	**Colour crate**	colour (204) of (4) crate (400)
the colour of my new shoes	**Colour new shoes**	colour (204) new (101) shoes (300)

Examples from Radford (pg 91):

i'm getting my glasses so i can read the book	**Read book**	getting (103) glasses (104) so (1) read (209) book (99)
you can put your finger through mummys ring	**Finger**	put (0) finger (111) through (100) mummys (100) ring (300)
i wonder if bear could do it	**Want bear do it**	wonder (115) bear (101) could (120) do (0) it (0)
shall we read more books	**Read books**	shall (82) read (107) books (302)
shall we go find her	**Go find her**	shall (82) we (0) find (6) her (35)
shall i open the little one	**Open little**	shall (86) open (214) little (6) one (4)
shall i open the big one	**Big one**	shall (86) open (214) big (16) one (4)
do you think he will wake up	**Wake up**	do (0) think (23) wake (106) up (7)
mommy wont fit in the refrigerator	**Mommy fit refrigerator**	wont (101) fit (118) in (0) refrigerator (400)

would you like some help	**Help**	would (194) help (303)
that must be the dining room	**Dining room**	must (130) be (24) dining (200) room (300)

The FreqFilt analysis predicts many of the omissions made by the child and although the output is often longer than the child utterance the extra words tend to have lower confidence factors. Additionally, the child utterances often consist of the final portion of the FreqFilt output - perhaps illustrating a recency effect in their imitation of adult forms that acts in addition to the filtering of the input. Most importantly for this analysis is the fact that the program reliably imitates the pattern of omission claimed by Radford - reducing adult DPs to lexical NPs - but without recourse to linguistic knowledge.

4.4.4 Absence of a Complementizer System

Radford argues that early child grammars lack a Complementizer system (C-system) and an Inflection system (I-system). He draws a comparison between children's early clause structure and small clauses in adult speech. In contrast to ordinary clauses, adult small clauses also lack a C-system or I-system. For example consider the adult ordinary clause:

I consider [*that* this candidate *would* be unuitable for the post]

In this clause, *that* acts as head of the C-system and *would* acts as head of the I-system. By contrast, the small clause equivalent lacks both these functional systems:

I consider [this candidate unsuitable for the post]

Children's clauses also appear to lack a C-system and an I-system (as well as the D-system as already mentioned) and this is illustrated by examples such as the following:

Sausage bit hot.

Wayne in bedroom.

Teddy want bed.

However, it is possible to extend the frequency filter analysis to these forms by continuing the argument that child forms can be characterized as equivalent to adult forms from which marker elements have been omitted.

The examples below show a frequency filter analysis, the middle row(in bold) shows a child sentence, the first row (underlined) shows my own grammatical expansion of that sentence and the bottom row is the FreqFilt analysis of this expansion:

1. the sausage is a bit hot
   Sausage bit hot
   sausage (240) is (-1) bit (166) hot (300)
2. wayne is in the bedroom
   Wayne in bedroom
   wayne (340) in (0) bedroom (400)
3. *teddy wants to go to bed*
   Teddy want bed
   teddy (300) wants (125) go (0) bed (151)
4. *wayne has taken the bubble*
   Wayne taken bubble
   wayne (301) taken (201) bubble (400)
5. *mummy is doing the dinner*
   Mummy doing dinner
   mummy (165) doing (17) dinner (400)

The FreqFilt examples differ from the child version in only two word tokens and these are given a very low confidence factor -we could exclude these by setting a threshold on confidence factors although this would also omit the preposition in example 2. Omitting elements with negative confidence factors would reduce the difference to one token. Differences in inflection cannot, as already stated, be handled by FreqFilt.

For these examples, FreqFilt produces results which are very similar to both the child utterances as well as to the predictions that would be made by Radford's GB analysis.

Wh-questions (From Radford, p. 123)

where does daddy go	**Daddy go**	daddy (2) go (1)
where shall i go	**Go**	shall (6) go (1)
where does it go	**Go**	does (5) go (2)
where did the bow wow go	**Bow-wow go?**	where (13) did (3) bow (200) wow (110) go (1)
what have you got	**You got?**	have (4) got (34)
what is mummy doing	**Mummy doing?**	what (0) mummy (28) doing (25)
where is the car going	**Car going?**	where (15) is (0) car (52) going (11)

what is he doing there	**Doing there?**	what (0) doing (2) there (0)
where have my shoes gone	**My shoes gone?**	where (13) shoes (103) gone (112)
what is the mouse doing	**Mouse Doing?**	what (0) is (0) mouse (202) doing (25)

Some of the words which are not present in the child utterances but which FreqFilt allows through such as 'What' and 'Where' are given low CFs. A second point to note is that while they pass through the filter on some occasions, they are blocked on others - thus their status may be considered as wavering between marker and non-marker. The factor which distinguishes between the two cases is the frequency of the following word - when this is high the element is more likely to be treated as a non-marker and included in the sentence.

Another factor which may affect the performance of the program is related to dialect differences between speakers - while some words, such as determiners, are unlikely to vary greatly, in terms of rank frequency, between speakers - others may be affected. So, for example, some speakers may tend to use 'will' rather than 'shall' - *What will/shall we do?* and there will also be variation between the amount of reduction of function morphemes by different speakers - '*We'll / We will go to the shops'* this variation may also be introduced by the person who transcribes the speech. Such variations will have an affect on the analysis as the frequency data used by FreqFilt is based on a different sample of language from that which a particular child will hear. It follows, of course, that the current approach will predict differences in the relative use of those function words which are prone to such variation between children. For example, a child whose caregivers tend to produce the reduced form *we'll* will treat the word *shall* as a lexical element while a child who is regularly exposed to the expanded form will be more likely to treat it as a marker and therefore omit it.

Furthermore, in the current analysis the status of a particular element as a marker depends on its context - the analysis could be augmented with information concerning the proportionate use of an element as a marker so that a word's use a marker on one occasion would increase the likelihood of it being used as a marker again.

For the examples of Wh-questions given above, we can make the additional generalization that the child forms simply consist of the right-most portion of the adult equivalent. While such a generalisation cannot, by any means, capture all of the qualities of child language it does seem to be a common feature of Radford's example. One explanation for this may be that lexical elements tend to occur more often at the end of sentences - certainly English sentences tend to end with lexical elements and many constructions (such as Wh-questions) place a number of functional elements at the beginning. Another possibility is that the child is constrained by the capacity of their phonological memory and that this will tend to focus their analysis on the most recently heard portion of a sentence.

4.4.5 Absence of an Inflection System

Radford argues that the absence of an inflectional system in the child is supported by examples such as the following which demonstrate the omission of modals:

mr miller will try	**Miller try**	mr (300) miller (101) try (154)
i will read the book	**Read book**	read (201) book (99)
i can see a cow	**See cow**	see (1) cow (366)

As we can see both the child and FreqFilt omit the modals from the adult sentence.

4.4.6 Use of Prepositions

Prepositions are often considered to be one of the four main lexical category systems (N, V and A being the others) althought they seem to occupy something of a grey area between the functional and lexical extremes. Radford argues that (because there is no case system in early child grammars) "children are not 'aware' at this stage of the *requirement* (imposed by Case Theory) for a case marking proposition to be used in prepositional contexts.... This means that if they use prepositions in such contexts, they will only do so *sporadically*" (p. 189).

However, the marker hypothesis lets us make a more specific claim. In the examples given by Radford to demonstrate this sporadic use of prepositions the children seem to have used prepositions in contexts where, in their adult equivalent, they would appear in high frequency

contexts such as 'is __ the' and 'is ___ there' while they have omitted pronouns in other, more varied and less frequent contexts. Most of the examples of a child using P occur in contexts where the adult would use the form NP is P the NP. 'Is' and 'the' are both high frequency markers. In the cases where the children have omitted P it tends to be preceded by a lexical verb, *e.g.*:

'go to school' -> 'go school'

where 'go' and 'school' are less frequent elements. We might therefore claim that children tend to use P specifically in those contexts where it would be marked by frequently occurring neighbours in adult speech. In Section 4.5 I will consider this point in more detail.

4.4.7 The Case System

Radford studies the status of the case system in early child language through an analysis of personal pronouns (the only elements which are overtly marked for case in English). English pronouns show contrasting case in such nominative/objective pairs as I/me, he/him, she/her etc.

Many children at this stage use only nominal forms but some use both nominals and pronominals and it is within this latter group that Radford searches for evidence of a case distinction.

He notes that such children appear to correctly use objective pronouns in positions which require them, such as 'Geraint hit me', 'Put them on', 'pick him up' but that they also use the objective form where a nominative form is required, for example 'Me talk', 'Me sit there', 'Me play', 'Him gone', 'Her climbing ladder', ' Her gone school'.

There are examples where children whose speech is predominantly lexical in nature use nominative pronouns but, Radford argues, they show the signs of being 'set phrases' rather than of displaying productive syntax.

The fact that some children at this stage appear to be using objective pronouns productively may suggest that, contrary to the claim that they lack functional systems, they have acquired

some knowledge of case marking. However, Radford argues that while they use objective pronouns children show no productive use of the corresponding nominative or genitive forms and therefore no signs of a systematic case contrast:

> *'It would seem reasonable to suppose that the absence of any systematic contrast between objective forms and nominative/genitive forms implies the absence of* case *as a formal property in early child English'*. (p. 181)

He continues to suggest that these child pronouns are not actually case marked objective forms but that they have the status of caseless pronominal NPs. This view is not incompatible with a distributional approach as objective pronouns tend to occur in similar contexts to people's names *e.g.* 'Give it to Mary/John/him/her/me', and the evidence would suggest that children apply objective pronouns in other (innappropriate) positions where names occur (e.g., from Radford p.181, 'Big Geraint.... big him').

One question that is raised by, and unanswered by, Radford's account is that of why children only use objective pronouns productively and not nominative forms *i.e.* why do they say 'Geraint hit me' and 'me hit Geraint' but not 'I hit Geraint' and 'Geraint hit I'. Without such an account we are left with what appears to be a case contrast - in order to selectively omit nominative forms the child must somehow distinguish them from objective forms and it, might be argued, this would require them to make a case-based distinction. Given the argument that children lack knowledge of case and are treating pronouns as caseless nominals, we would expect them to use nominative and objective forms in much the same way. The fact that they specifically seem to omit nominative forms is a curious distinction to be made by one who has no knowledge of a nominative/objective distinction. Consider an analogy: if a student had to sit a multiple choice exam on linguistics in which they were required to choose between three alternative answers for each of 100 question and they were found to have got all of them wrong, a naive observer might come to the conclusion that they

knew nothing about linguistics, but a mathematician would note that the probability of scoring 0 by chance would be vanishingly small and that the more likely conclusion would be that the student had known the correct answers and had deliberately failed the test. Similarly, the child's use of objective forms and exclusion of nominative forms might suggest that rather than lacking knowledge of case, they must have such knowledge in order to distinguish between the two forms - unless some other explanation can be provided for their selectivity.

The marker hypothesis may provide such an alternative. The analysis which will be presented in section 4.5 suggests that distributional differences between pronouns of different cases may underlie their use.

The hypothesis is that nominative pronouns will be more likely to act as marker elements (and thus not enter into initial productive speech) than objective pronouns (which will be used in productive speech) - nominative pronouns will be filtered out while objective pronouns will remain.

According to this account, the child is treating subjective pronouns predominantly as marker elements and is treating objective pronouns predominantly as lexical elements. The objective forms are thus entered into their own productive speech and are categorised according to the marker frames in which they appear. The following shows FreqFilt's analysis of a number of sentences containing both objective and nominative personal pronouns:

Input sentence	FreqFilt output
do i know him	know (1) him (21)
do they know them	know (0) them (22)
do you know me	do (0) know (25) me (9)
does he know her	does (22) her (35)
does she know me	does (22) know (0) me (9)
geraint hit me	geraint (300) hit (104) me (9)
he gave it to me	gave (156) me (10)
he hit me	hit (111) me (9)
he will ask her	ask (103) her (35)
help me out	help (304) out (19)
i asked them	asked (117) them (22)
i hit geraint	hit (114) geraint (300)
i hit her	hit (116) her (35)
i hit him	hit (117) him (21)
i want to eat them	want (0) eat (8) them (22)

i will ask him	ask (104) him (21)
paula put them on	paula (308) them (2) on (0)
she chased me	chased (107) me (9)
she hit me	hit (107) me (9)
she will ask them	ask (104) them (22)
they chased me	chased (107) me (9)
they chased them	chased (106) them (22)
they chased us	chased (103) us (300)
they gave it to them	gave (153) them (25)
they gave it to us	gave (153) us (325)
they like them	they (18) them (22)
throw them in	throw (302) them (1) in (0)
we asked them	asked (108) them (22)

FreqFilt does seem to display a marked bias towards using objective forms resulting from differences in their frequencies and in the contexts in which they appear. In section 4.5 I will give a more detailed account of this phenomenon.

4.4.8 Missing Arguments

Children omit arguments which are required in adult grammar. I would suggest that, rather than being the result of a grammar in which arguments are allowed to be omitted, child speech consists of partial constructions. That is to say, there is no imperative on the child to produce full sentences in adult terms. I have already suggested that, in many cases, the child imitation, or equivalent, of an adult utterance differs in that it omits marker elements but also in that it is incomplete. Child utterances may consist of phrasal units below the full adult sentence. Where such utterances do not correspond to phrasal units in conventional phrase structure accounts, they are nevertheless 'meaningful' units. In the next chapter I will consider the issue of how such units may be considered in terms of syntactic theory and how they could underlie the development of adult grammar.

The difference between the output of FreqFilt and the child utterances may be one of quantity rather than quality - both show a lack of functional categories but the FreqFilt 'sentences' are longer. Such a simple difference may be put down to memory capacity.

4.5 The Frequency Filter, GB-Theory and Functional Impairment

4.5.1 Explanation

In order to provide a more concrete comparison between the predictions made by FreqFilt and those of a GB account in which functional categories are disabled, I will now offer a direct comparison between the two approaches through a reanalysis of data provide by Shillock and Tait (1994). In this work, the authors suggested that agrammatic speech may be predicted by the simulation of a GB-based account with impaired functional projection. They used data from Menn & Obler (1990) which provides examples of speech by an agrammatic patient, "Mr Eastman", and fully grammatical expansions of these utterances which are intended to reflect the equivalent form that would be expected to be produced by a normal adult speaker. In order to measure the ability of GB theory to predict this data: "Three syntacticians working at research level in GB-theory were given Menn's grammatical expansion and asked to generate a version which "would be produced by the complete impairment of functional projection as in recent GB accounts"". These impaired versions were then compared with the original agrammatic speech of Mr Eastman.

The results produced by the linguists who took part in this experiment allow us to make a direct comparison between the predictions made by GB theory and the Frequency filtering approach. Although the aim of this experiment was to study agrammatism, the underlying premise of functional impairment is the same as that used in Radfords's theory - in both cases the impaired or child form is considered to be equivalent to an adult form in which functional projection is not available.

Below are two examples from the 24 sentences: the first line contains the full grammatical expansion from Menn & Obler, the second line is Mr Easton's sentence, this is followed by the three syntacticians sentences from Shillcock and Tait's paper, and finally the output from FreqFilt.

I was in a wheelchair at Hahnemann Hospital for a week.
Uh. Wheelchair ... uh . Hahnemann Hospital . a week, a week. uh ...
in wheelchair at Hahnemann Hospital (for) a week
be in wheelchair at Hahnemann Hospital for week
wheelchair Hahnemann Hospital week
was (4) in (0) wheelchair (69) Hahnemann (3) Hospital (3) for (4) week (266)

She shaved me, the nurse
shaved me, nurse.
shaved (me), nurse *shave, nurse* *shave nurse*
shaved (7) me (4) nurse (300)

4.5.2 Comparison with FreqFilt

The Frequency filter program (FreqFilt) was run on the same stimulus sentences used in Shillcock and Tait's experiment and results were produced in the same way - by recording the number of occurrences of each of a set of grammatical categories in the set of reduced (filtered) sentences. The results for FreqFilt were calculated without taking any account of the confidence factors associated with the words. Of the original 9 categories used by Shillcock & Tait, one (verb suffix) was omitted from the results due to the inability of FreqFilt to handle bound morphemes. The following table gives the results - the table is the same as that used in the original experiment but with the row for verb suffix removed and the addition of FreqFilts results in the second column:

	Mr Eastman	FreqFilt	Linguists	Expansion
Det	4	0	0.3	9
Aux	0	0	0.7	3
Prep	6	7	12.3	20
Pron	9	10	2.6	27
Conj	6	3	2.6	9
V	13	17	14.6	22
A	8	7	8	8
N	40	43	42.6	43

Table 4-3 Comparison of predictions made by linguists and frequency filter

The following table gives the above values as a proportion of the grammatical expansion:

	Mr Eastman	FreqFilt	Linguists	Expansion
Det	.444	0	0.0333	1
Aux	0	0	0.23333	1

Prep	.3	0.35	0.615	1
Pron	.3333	0.37037	0.096296	1
Conj	.66667	0.33333	0.2888889	1
V	.5909	0.772727	0.6636364	1
A	1	0.875	1	1
N	.930232	1	0.99069	1

Table 4-4 Proportion of each syntactic category omitted by Mr Eastman, FreqFilt and linguists.

Looking at these values we see some interesting results - FreqFilt removed all occurrences of Det & Aux. It is very close to Mr Eastman's data in preposition and pronoun use - In both cases approximately one-third of the required elements are included. Bear in mind the sporadic use of Prepositions and pronouns in child language.

In other respects the freqfilt program produces a more pure lexical/functional distinction than the linguists - it maintains a higher degree of V and N and omits all determiners and auxilary verbs.

The similarity between the four sets of values was measured for each pair using the Euclidean distance metric. The following are the pairwise distances sorted in ascending order:

Pairwise Euclidean distances (sorted)
Distance of Linguists from FreqFilt : .4800884
Distance of FreqFilt from Mr Eastman : .6050323
Distance of Linguists from Mr Eastman : .7283747
Distance of Expansion from Mr Eastman : 1.589472
Distance of Expansion from Linguists : 1.762394
Distance of Expansion from FreqFilt : 1.825008

Looking at the distances we can see that the frequency filter is closer to Mr Eastman than the linguists results *i.e.* FreqFilt is slightly better at predicting the agrammatic speech than the functionally impaired GB analysis.

Equally interesting is that the closest pair are the GB analysis and that of FreqFilt. In most respects there is little variation between the two sets of data - the main source of difference lies with prepositions and pronouns. As I have already mentioned, Radford notes the sporadic use of prepositions and pronouns in child language - FreqFilt also produces these classes 'sporadically' - including them in about a third of the cases.

4.5.3 Prepositions and Pronouns

In order to see if there was in any pattern in the use and non-use of pronouns by FreqFilt, I examined the pronouns individually. There were 27 pronoun tokens in the sample and their pattern of omission or inclusion by FreqFilt is shown in the following table:

Pronoun	Total	Omitted	Included	Proportion Included
I	10	9	1	0.1
My	7	6	1	0.143
It	1	1	0	0.0
She	1	1	0	0.0
Me	4	0	4	1.0
Who	1	0	1	1.0
Its	2	0	2	1.0
Mine	1	0	1	1.0

Table 4-5 Omission and Inclusion of different pronouns by the Frequency Filter

The single occurrence of 'I' resulted from a case in which it appeared alone in the expansion - it was omitted in all normal sentential contexts - thus on a larger sample its occurrence rate would probably approach zero. 'Its' is a problematic case for FreqFilt because of the inability to handle bound morphemes. What is most interesting is the general omission of the personal pronouns in nominative and genitive forms ('I', 'She' and 'My' with only 2 inclusions out of 18 tokens) and the inclusion of the objective pronoun 'Me' (all 4 tokens included). These results fit with the pattern observed by Radford in early child use of personal pronouns.

The predictions made by FreqFilt offers some degree of explanation for the variability of categories such as P. Abney (1987) notes that "like all major grammatical distinctions, there is a substantial gray area between thematic and functional elements; there are thematic elements with some properties of functional elements , and vice versa" (pp. 64-65) and that "P seems to straddle the line between functional and thematic elements." (p. 63). Cutler (1993) also mentions differences in the degree to which different word classes may be considered open or closed:

Thus, for example, nouns form the largest of the open classes, and are also in a sense the "most open" in that new nouns are formed more frequently than new verbs (Kelly 1992). Likewise, within the closed class, forms such as prepositions and pronouns are more numerous than, say, articles and complementisers, and the former are more likely to be semantically highlighted in a sentence by being deployed in a contrastive construction ("in the world but not of it") -although of course such constructions are far more common again with open-class words ("Stirred not shaken").(p. 110)

The various degrees of openess described by Cutler accord well with the results above in which the proportion of items which remain untouched by the frequency filter descend in the order Noun > Verb > Pronoun > Preposition >Determiner. So as well as underlying a gross distinction between functional and lexical elements, the frequency filter approach may also go some way towards explaining the gradations between the two extremes.

4.6 Conclusion

In this chapter I have claimed that early child language is not typified by a lexical-functional distinction but by distributional distinctions which underlie the later development of a lexical-functional distinction. Children do not omit determiners because they are a functional category but because of simple distributional features which typify members of that category - while the status of D as functional, from the perspective of the learner, arises out of the fact that it displays certain distributional features.

Children's omission of words is better predicted by the simple distributional properties outlined here than by a functional GB analysis but there is a large degree of overlap between the two approaches. The frequency based approach has the dual advantage of explaining the ontogenetic source of these omissions while greatly reducing the reliance on a phylogenetic source (Whereas any approach that depends heavily on the latter must still explain the former).

Accounts such as Radford's are to some degree motivated by a desire to maintain continuity between the child and adult grammars. But continuity may be maintained without assuming that the child's system is the same as that of an adult - rather we might argue that the adult grammar is the product of a series of progressions which do not themselves violate continuity but whose beginning and end are very different. This argument depends on a consideration of the syntactic representation employed by the child which will be the subject of the next chapter but for now I will simply say that the current hypothesis holds that the use of marker elements provides a basic set of representational tools with which the child builds later representations. It is the route followed by the child which makes learning possible - only by developing the right kind of initial representation are they able to develop later complexity. The principles outlined in this chapter provide the mechanism by which the structure of language serves to guide the early development in the right direction.

The account proposed so far has some parallels with the following comment made by Matthew's (1981):

> *'if a child of five says 'Mummy wears a hat to keep warm', we may have no compelling reason to assign this to a complex sentence construction... rather than a simple collocational schema... in which at successive points an open or closed set of items can be substituted. To be precise, the question of what construction it has, or whether it is a simple sentence or a complex sentence, involve categories of 'construction', 'sentence' and so on which are not appropriate to the schemata with which learning begins.'*
>
> (Matthews, 1981, p. 187)

The current work holds that childrens initial schemata are built on fixed syntactic frames into which a set of related items may be inserted.

We can think of the early productive speech of a child as the product of contextual generalisation of open-class items mediated by closed-class syntactic frames. These syntactic

frames may or may not correspond to full sentences. The process of contextual generalization which enables items to be generalised between frames is described in the following description from Braine (1963):

> *"there must exist generalization mechanisms in language learning whereby a word learned in one context generalizes to another context, even though no associations may have previously formed between the word and its new context"* (p.324)

and this, by Maratsos and Chalkey (1981):

> *"Productivity arises from the fact that form class members are used in highly overlapping contexts and these contexts (operations) become correlated by association with common terms"* (p. 133)

The marker hypothesis provides us with the mechanism by which useful contexts may be acquired.

The evidence presented in this chapter suggests that the lexical character of early child speech can be explained by simple distributional properties of language.Children's early utterances can be understood as simple collocations of elements under the mediation of high frequency marker elements.It is entirely possible that in the early stages of acquisition, linguistic 'rules' are tied to particular functional elements.There is little evidence at this stage of deeply nested linguistic structure - for example in terms of modifier/complement attachment. Nor is there evidence of any significant reworking of language input based on high level syntax - the utterances used by the child can be explained in terms of contextual generalisation between marker frames. Thus at this stage, invention is limited - the child is a relatively conservative linguist who does not stray from simple variations on perceived utterances.

Children's imitations seem not far removed from their adult counterparts. The functional elements which are omitted are all phonologically realised. Where child syntax is productive

it would seem possible to explain it in terms of contextual generalisation between frames. Many of the child utterances in Radford's examples can be grouped together around the common syntactic frames which are omitted from the child's speech but which may be mediating their structure:

Geraint naughty, Lisa naughty, Daddy away, Me nice, Him gone, That Ashley -> ## ___ is ____##

Wayne in bedroom, It in bag, ## ____ is __ the ___##

Mouse in window, Bubble on dungaree -> ## The ___ is ___ the ___ ##

Cup tea, bottle juice, picture Gia -> ## a __ of ___ ##

Paula good girl, Mommy girl, Daddy boy , This hand -> ## __ is a ___ ##

whilst not all sentences with the same syntactic structure will have the same marker frame many do and this will serve to indicate the similarity between constructions. Such an ability will enable a powerful form of generalisation between different context - even before lexical elements have been categorised it will be possible to establish syntagmatic equivalence between different phrases and even sentences. For example, the phrases 'a drink of orange' and 'a bottle of juice' have the same marker structure ('the __ of __ ##') even though they contain different lexical elements - this is particularly useful as the rarity of lexical elements makes them difficult to categorise and even if they are categorised they can give rise to a great deal of amiguity e.g. 'drink' is more likely to occur as a verb than a noun but such a categorisation would be wrong in the phrase 'a drink of orange'.

Marker elements may also offer a direct measure of syntactic complexity - more markers equates to more complexity (in a more predictive way than would be provided by a word count alone). For example consider the following sentences:

Sentence	**Words**	**Markers**
'**The** rabid dog bit **the** old man'	7	2
'**The** rabid dog **was** bitten **by the** old man'	9	4
'**The** boy hit **the** girl'	5	2

'**the** boy **was** hit **by the** girl'	7	4

Table 4-6 Markers as indicators of syntactic complexity

While the number of words is not a good indication of the complexity of these sentences, the number of marker elements is. Basic sentences tend to be those which are least marked while increases in syntactic complexity are matched by an increase in the number of markers, thus a child will, in a sense, be given warning of complex forms and this may enable them to simplify their analysis in the early stages of learning by helping them avoid what could be confusing syntactic forms.

5. Grammatical Development

In the last chapter, I showed how a simple model based on the use of high frequency functional markers could go a considerable way towards explaining the early utterances produced by children. This approach showed how the child may divide morphemes in perceived speech into marker elements and lexical, or marked, elements and that this dichotomy can explain the use and non-use of different morphemes. However, aside from considering how a process of contextual generalisation may lead to the use of morphemes in novel positions, the model did not consider the process by which a child could build a syntactic representation of the language - I argued that the omission of functional elements in early speech can be explained in terms of a simple, pre-linguistic, statistically-based analysis of the input language and that the same process enabled the child to isolate useful phrasal units. In this chapter I explore the syntactic nature of these units and how they can underlie early syntactic development.

The primary claim of the marker hypothesis is that the language heard by a child contains various salient cues which serve to direct the child's attention to important grammatical features of the language. There are two main claims made concerning the kind of information provided by these markers:

1. Certain syntactic categories may be marked by cues which are both salient and, if not invariant, of low variance. For example, in English certain patterns of prosodic stress may distinguish verbs and nouns. Also certain highly frequent functional morphemes may be reliably associated with particular categories.

2. Certain cues (phonological and functional) may serve to bracket utterances into linguistically useful units thereby constraining the child's attention to co-occurrences which are syntactically useful.

This chapter addresses the second of these issues. In particular, I will ask what form these 'linguistically useful units' take.

Ideally, different types of structural cue should reinforce one another and I will look at cases where different cues agree on the structures which they mark, I will also consider cases where structural markers appear to give unreliable 'advice' to the learner

I will consider two broad classes of grammatical theory which have been proposed to account for the structure of language: Phrase Structure Grammar and Dependency Grammar and I will compare the advantages and disadvantages of each representation from the perspective of the learner and the similarities between them and also look at issues which cut across both theories.

It is quite common for research into language acquisition to adopt a traditional phrase structure grammar as the framework for investigation and where results fail to match the phrase structure account to see this as a failure of the learning mechanism. In this chapter I will propose that the phrase structure approach may be misleading to research into acquisition and that a better linguistic framework is provided by dependency based grammars.

5.1 Marked Structures in Language

I have already outlined the theory that various readily available cues in language may serve to 'package' the input to the learner by delimiting and marking useful grammatical units so that the learner's efforts are restricted to the important grammatical relationships. The two types of cues which I have considered are the intonational structure of the speech stream and

the occurrence of high frequency elements both of which may be considered to be available to a learner in a naive state.

We might expect that both these sources of structural information would demonstrate a large degree of similarity in the structures which they mark, or at least that they would not disagree in any way that might cause the learner to have to choose between two differing interpretations; we would also expect the structures which they marked to be linguistically useful in that they either matched those structures used by the adult speaker of the language or would provide a useful stepping stone towards the development of such structural knowledge in the learner. Regarding this last point, it has usually been accepted that for it to be useful, structural marking should approximate as closely as possible to phrase structure bracketing as anything else is presumed to be misleading.

5.1.1 Marker-Syntax Mismatches

Yet there would appear to be one area in which such structural cues seem to consistently provide the same misleading information, this is the division of the subject and predicate of a sentence. In the traditional approach, the sentence *John loves Mary* will receive the constituent bracketing (John (loves Mary)) - the division between subject Noun-phrase subject and Verb-phrase predicate is the most basic division of the sentence in phrase structure grammar and we would perhaps then expect it to be mirrored by a similarly basic division by structural marking. However, this does not appear to be the case. If we take a very basic English sentence such as the *The dog chased the cat* we can see that whether we assume that the marker elements are function words or high frequency words the same analysis will be given - the word *The* will be considered a marker element under either approach and consequently the sentence will be split into the units (dog chased) and (cat). Rather than providing us with a neat subject/predicate division, the subject is instead joined to the verb and the object is left stranded. If we assume phrase structure grammar as the goal for the learner then such packaging of the input appears to be highly misleading. As Morgan *et al* (1987) note "...in English, phrases may sometimes lack initial bracketing function words

or final vowel lengthening" and that "individual cues may sometimes provide misleading bracketing information" (p. 535).

Morgan *et al* note a similar effect with concordance morphology - such as that which occurs in Latin where words which agree have homonymous markers and which they consider to be another potential source of structural marking, but they note that it may "cut across phrase boundaries as is the case in subject-verb agreement." (p. 536) which may inidicate a grouping of, rather than a split between the subject and verb of a sentence. So, while concordance may sometimes indicate words which go together in a phrase such as when an adjective and noun share the same ending, it may also relate two words which are not in the same phrase. Morgan *et al* consider such examples as being misleading to the learner but argue that a combination of different sources of structural marking may conspire to overwhelm the misleading cues with the correct information.

But a third source of structural marking has also been noted to cut across the conventional phrase boundaries. The intonational structure of a sentence has long been observed to differ from the phrase structure. For example, Selkirk (1984) describes how the sentence '*Mary prefers corduroy*' can have the intonational structure ((Mary prefers) (corduroy)) .

Jusczyk (1993) notes that "There is good reason to believe that reliance on prosodic features of the input will only take the infant so far in discovering the syntactic organization"(p. 56). Prosodic boundaries do not always correspond to syntactic boundaries and the kinds of boundary mismatches which occur are likely to be quite common in the language directed to children who are beginning to acquire language.While in the sentence '*Ellen / threw the ball'* the prosodic boundary is likely to occur between the subject and predicate, in '*She threw / the ball'* the weakly stressed subject pronoun is likely to be joined to the verb with the prosodic boundary occurring between the verb and object. Jusczyk suggests various ways in which the child may avoid this problem but these solutions rely on the learner having access to existing

syntactic information or being able to use evidence from other sources to overcome misleading cues.

These examples show that at least these different sources of structural marking appear to agree on this issue thereby meeting one of the requirements suggested above but within a phrase structure account they appear to be agreeing on the wrong structure - can this problem be resolved?

5.1.2 Retaining the Marker Hypothesis

5.1.2.1 Mapping between Syntax and Prosodic Structure

Generally the syntactic and prosodic structures of language have been seen as similar but not identical and various approaches have been proposed to enable mapping between them. For example Chomsky and Halle (1968) propose a set of readjustment rules for mapping from the syntactic to the prosodic structure and Gee and Grosjean (1983) proposed a set of transformation rules for the same purpose.

However the need for such rules does not fit easily with the marker hypothesis as it would severely complicate the process of using intonational structure as a cue to syntactic structure and would beg the question of how the learner could acquire the necessary transformational rules without the prior knowledge of the syntactic structure of the sentence.

5.1.2.2 Conspiracy theories

Another approach which has been suggested by several authors (Morgan *et al*, 1987; Jusczyk, 1993) is that the learner relies not on a single source of structural marking information but on a number of different sources which, in conjunction with one another, conspire to overcome the shortcomings of any single source. Thus the learner could make use of correlations between cues to reach a final picture of the marked structure. Assessing the practicality of such an approach is very difficult as it, by definition, involves the analysis of a great deal of information. Intuitively, though, the approach may be problematic - the examples above show how a number of different cue sources may point to the 'wrong' structure in quite a basic sentence form. The complexity of the issues involved make it

difficult to reject the conspiracy approach but that same complexity makes it difficult to see how the language learner could balance the incoming information so as to hit on the 'right' answer. While there is no doubt that multiple sources of information could be of great value to the learner, it would seem preferable for these sources to agree, thus offering some redundancy, rather than disagreeing and thus adding a new obstacle in the learner's path. This seems especially so in the early stages of learning when structural marking is arguably most useful but also when the learner is in the most naive state and unable to draw on top-down syntactic knowledge in order to resolve discrepancies between cues.

5.1.2.3 Change the Grammar

Morgan & Demuth (1996) point out that:

> *"An alternative view of such evidence is that the presence or absence of phonological cues to particular distinctions or the existence of particular perceptual capacities in infants may provide clues to the nature of grammar ... or to aspects of grammar whose early acquisition is of prime importance"*
>
> (p. 4)

Rather than trying to find ways to explain away these mismatches between structural markers and phrase structure grammar, we may instead reconsider the grammatical formalism which we adopt.

Such an approach has been proposed by Steedman (1991) who argues that syntax and intonational structure become homomorphic if we adopt Combinatory Categorial Grammar (CCG) as our syntactic model rather than conventional phrase structure. CCG allows a more flexible approach to constituency which more readily explains the different intonational structures which can be allowed in a sentence thus obviating the need for transformational rules to map between the two structures. This is an attractive proposition from the current perspective, for it provides a direct route by which the learner may utilise the structural marking offered by intonation to enter the syntactic system of the language. In this Chapter I

will consider an alternative syntactic model based on the dependency grammar formalism which is described by Barry and Pickering (1990) and which, they suggest, offers a similar explanatory power to Steedman's account. They argue that their account of dependency constituency:

> *... appears to be relevant to an account of intonational structure. Steedman (1990) shows that the range of possible intonations is greater than those that are most obviously allowable by traditional phrase structure, but that there are still many strings which do not form possible intonational units... we believe that the range of possible intonational units can be derived from the notion of dependency constituency.* (pp. 19-20)

I will argue that such a syntactic representation is recommended by its superiority in explaining a number of issues in early syntactic development. However, in order to explore this avenue it is first necessary to digress into a consideration of dependency grammar.

5.2 Consideration of Grammatical Formalisms in Learning

5.2.1 Phrase Structure Grammar and Dependency Grammar

The predominant linguistic tradition of recent years has been that based on the idea of constituency which has formed the basis for the work of Bloomfield (1933) and Chomsky (1957) and the bulk of the ensuing work in the field of linguistics. The constituency model can be characterised by its emphasis on the relationship of parts to a whole, it posits that a sentence is represented by a hierarchical structure which groups words into phrases and phrases into higher-level phrases up to the sentence level. Many of the relationships thus formed act between phrases rather than individual words. I will say little more about this and assume that the reader is familiar with the basis of phrase structure grammar. I will simply note the important point that within PSG the definition of constituency is a primary concept - it is difficult to imagine how one could remove the concept of constituency from PSG.

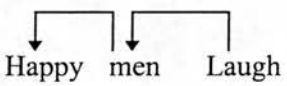
5.2.2 Dependency Grammar

An alternative to the constituency model is provided by the tradition of dependency grammar which is often traced back to the work of Tesniere (1959) (although Covington (1994) describes its practice in the Middle Ages). While the constituency model is based on the part-whole constituent relationship, dependency grammar is based on the relationships between parts (words or morphemes).

Thus, within dependency grammar the notion of constituency is not basic as it is in the constituency model (although constituency can be defined within it and this is central to the ensuing account).

Within Dependency Grammar (DG), pairs of words are joined by a syntactic relationship known as a dependency. For example in the sentence *Happy men laugh* there are two dependency relations - one between *Happy* and *men* and the other between *men* and *laugh*. The question of which words are related by dependencies is mediated by the grammar and has been described by Hudson (1984) as 'constrained co-occurrence' - the constraint being provided by the syntax of the language. In a correct representation, every word of a sentence must be part of a single dependency diagram. The language of graph theory can be useful here - if we think of the words of the sentence as the nodes of a graph and the dependencies as arcs between them then a dependency grammar representation of a sentence should constitute a fully-connected graph. In the example above, the dependencies reflect the syntactic (and semantic) relations that hold between the words - *Happy* is linked to *men* because it modifies it, *men* is linked to *laugh* as its subject. There is no dependency between *Happy* and *laugh* because there is no direct meaningful relationship between them, although they are indirectly connected through their membership of the same syntactic unit (the sentence) and this is reflected in their being indirectly joined within the overall dependency diagram. Of course, the term dependency does not suggest an equal relationship and this reflects the fact that one member of each linked pair is seen as being dominant over the other. The dominant element is often referred to as the *head* but this may cause confusion with that

term's use in other linguistic theories so I shall follow Matthews (1981) in using the term *controller* for the dominant element while using the conventional term *dependent* to refer to the subordinate element of the pair. Applying the controller/dependent relationship to the above example, we obtain a directed graph in which the arrows point from the controller to the dependent of each pair:



This shows that *Happy* is dependent on *men* which is, in turn, dependent on *Laugh*. This representation has certain properties common to all dependency diagrams in classical dependency grammar. Firstly, there is only one independent element, or root *i.e.* an element which has no controller. In this case, as is usual, it is the verb which is the root of the sentence. The dependents of the verb are those words whose categories it selects - in this case it selects *men* as its subject. *Happy* modifies *men* and is thus dependent on it. Another way of examining these relationships is to consider their dependence in terms of *presupposition* - an element **a** is dependent on an element **b** if the presence of **a** presupposes that of **b**. Thus, in the sentence *Happy men laugh*, *Happy* may only appear because of the occurrence of the word *men* - we could not have *Happy laugh* as a sentence but we could have *Men laugh* thus *Happy* is said to presuppose the presence of *men* and is therefore dependent on it. Following a similar line, we can observe that the verb *laugh* may appear on its own, as in the imperative *Laugh!* but that *Men* cannot form a sentence on it's own, thus we can say that *laugh* is the head of *men* and *men* is dependent on *laugh*. Hudson likens such presupposing relationships to that which holds between a house and its dustbin - although the two appear together, the dustbin only appears because of the presence of the house and not vice-versa, the presence of the dustbin therefore presupposes that of the house and is thus dependent on it. Another way of looking at this would be in terms of cause and effect relationships such as that expressed in the proverb 'There's no smoke without fire' - when two events occur together one may be

caused by the other and similarly when two words appear together one may be only present as the result of another.

Further evidence that two words are connected by a dependency comes from the number agreement between them so we cannot say *Happy men laughs* because *men* and *laughs* must agree for number. The use of agreement to justify dependencies is easier in languages which have a richer, phonologically-realised agreement system such as Latin.

There are no foolproof methods for deriving the dependency structure of a sentence and there are disagreements in the field but this situation is not terribly different from that in the constituency based approach. For example, classical DG holds that a noun is the controller of its determiner so that, for example *men* would be the head of *the* in the sentence *The men laugh*. However Hudson (1984) argues that the determiner controls the noun. In a sense, this contribution to dependency grammar is similar to that made by Abney (1987) to the constituency based approach in which he adopts functional heads and reformulates the traditional Noun Phrase as a Determiner phrase.

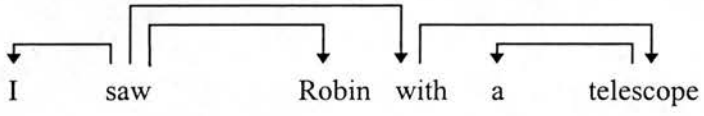
There are other disagreements about the fundamental structure of a dependency diagram. Classical DG makes the stipulations that:

- One and only one element may be independent
- All others depend on some element
- No element depends directly on more that one other
- (The Adjacency Principle) If A depends directly on B and some element C intervenes between them, then C directly depends on A or B or some other intervening element.

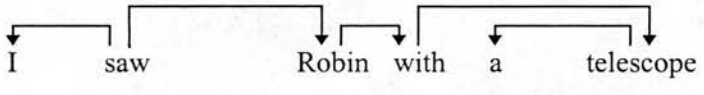
Together these principles mean that for a sentence of N words there will be exactly N-1 dependency pairs. These principles of Dependency Grammar are not set in stone and various alternatives to them have been proposed. In Hudson's (1984) *Word Grammar* - a variant of Dependency Grammar - an element is allowed to be controlled directly by more than one

other. Also, the adjacency principle has problems with languages such as Dutch in which crossed dependencies are allowed. For the current work, the principles of classic dependency grammar will be followed unless otherwise indicated.

By looking at a sentence with an ambiguous syntactic structure *I saw Robin with a telescope*, we can compare the representations provided by a DG account and a constituency-based phrase structure grammar account:



(I (saw Robin (with a telescope)))



(I (saw (Robin (with a telescope))))

From these diagrams we can see that the two distinct phrase structure representations of the sentence have corresponding dependency structures. Also, the two dependency structures differ only in one dependency relationship. In the first example, which corresponds to the meaning "I had a telescope with which I saw Robin", the word *with* is dependent on the verb. In the second example, which corresponds to the meaning "I saw Robin who had a telescope", the word *with* is dependent on *Robin*. Of course, this change in the direct dependency also affects the indirect dependencies in the structure - so that when the controller of *with* changes, the overall configuration of the words *the telescope* within the context of the whole sentence also changes although their direct controller is the same. Such indirect connections brings us to the fact that although constituency is not basic to dependency grammar it may nevertheless be defined within it, and it is the nature of this definition which I will now consider.

5.2.3 Constituency in Dependency Grammar

The conventional definition of constituency in DG (which dates back to Tesniere) is that a constituent consists of a word and all of its subordinates (a subordinate being a word which is either directly dependent on the word, or is dependent on a subordinate of the word, to use a recursive definition) - a further stipulation may be made, at least for English, that constituents must be continuous. Looking again at the previous example, we can see that, in either structure, the word *a* has the sole subordinate *telescope* and thus *(a telescope)* is a constituent, the word *with* has both *a* and *telescope* as its subordinates and thus *(with a telescope*) is also a constituent. Together this means that *with a telescope* has the structure *(with (a telescope))* which is the same analysis that we would obtain under a phrase structure representation, however the DG account of the complete sentence does differ from the phrase structure account. In the first example the full constituents (other than the individual words) are

(*a* ***telescope***), (***with*** *a telescope*) and (*I* ***saw*** *Robin with a telescope*).

In the second example we have:

(*a* ***telescope***), (**with** a telescope), (***Robin*** *with a telescope*) and (*I* ***saw*** *Robin with a telescope*)

In each case the word in bold is the head of the constituent. As we can see, the DG account produces fewer constituents than the corresponding PSG account and consequently it produces trees with a 'flatter' structure.

5.2.4 DG Theory and X-Bar Theory

It should be noted that there are many similarities between DG and PSG, for example the notion of head in X-bar theory is very similar to that in DG, also case frames encode much of the information that PSG does not handle directly but which are inherent in the DG system.

In fact this account produces structures which are very similar to an X-bar account in which (from Covington, 1994):

- There is only one non-terminal bar level
- Apart from the bar level, X and the X-bar immediately dominating it cannot differ in any way, because they are "really" the same node;
- There is no "stacking" of X-bar nodes (an X-bar node cannot dominate another X-bar with the same head).

Covington notes that the third of these restrictions mean that conventional DG cannot handle certain problems that X-bar theory handles well, specifically the proform *one* and the semantics of multiple modifiers (as in 'a long haired student of physics from Oxford'). Covington's solution to the problem is to produce a way for mapping dependency structures into stacked X-bar trees. This analysis distinguishes between specifier, modifier and complement dependents and attaches them under stacked nodes in a controlled way. This analysis narrows the gap between DG and X-bar theory to the point where the difference between them seems minimal (if not non-existent). These additions depend on two modifications to the classical model: 1) the use of different Surface Syntactic Relationships on dependencies - this involves taking the dependencies of a classic dependency grammar and assigning them more specific syntactic labels, and 2) changes to the way in which dependencies are built which take into account these syntactic labels and use them to build 'stacked' structures. Many of the developments in PSG since its original conception, such as X-bar theory and the concept of governent have served to bring that account closer to dependency grammar, while reinterpretations such as Covington's serve to narrow (or even close) the gap from the other direction.

An important feature of Covington's modifications from the current perspective is that they can be built on top of an existing, traditional DG account rather than requiring such an

account to be torn down and rebuilt. This feature is important for our constructivist model of syntactic development and its significance will be outlined shortly.

But while Covington's account narrows the gap, I now want to consider another account which highlights the differences between the two approaches and which I will argue offers an account of the early grammatical representations used by the child.

5.3 Flexible Constituency in DG

As noted above, the concept of constituency is not basic to dependency grammar but can be defined within it and I have so far considered two definitions: the conventional approach which produces flatter structures, and Covington's (1994) reanalysis which brings it in line with an X-bar account.

The concept that I want to look at now is that of Flexible Dependency Constituency (Barry and Pickering, 1990; Pickering and Barry, 1993) which provides yet another definition of constituency within DG. Barry and Pickering first of all note that traditionally constituency has been viewed as *rigid*, meaning that, although a constituent may be fully embedded within another, two constituents cannot overlap. So in a sentence with the elements (A B C) we could not have the constituents (A B) and (B C) because they share some, but not all, common elements. This is the *single mother condition,* as stated by Borsley (1991) :

> *..it is generally assumed that no expression can be a constituent of two different expressions unless one is a constituent of the others. This means that no node can have more than one mother.* (p. 21)

Barry and Pickering argue that there is no principled reason as to why constituents should not overlap and that the restriction is largely a product of the fact that context-free phrase structure grammars can only produce such non-overlapping constituents. They instead propose that we can adopt a theory of flexible constituency in which constituents can overlap and argue that standard tests for constituency support this account. For example in the

sentence *John loves Mary*, the sequences *John loves* and *loves Mary* can both serve as conjuncts as in the sentences *John [loves Mary] and hates Sue* and *[John loves] and Bill hates Mary*. The latter example can only be generated within conventional PSG through the use of transformations, but Barry and Pickering argue that it is strange that such a basic strucure cannot be produced by the grammar without the the use of additional mechanisms.

John loves can also appear as an apparent unit in the fronted sentence *Mary, [John loves].* They also consider sentence fragments such as:

Q:What will John do? A: Go to the shops.

Q: Who will go to the shops? A: John will go.

and elliptic constructions:

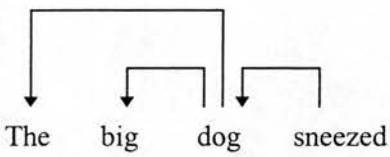
Q: Who will go to the shops? A: John will.

Q: Where will John go? A: To the shops.

arguing that in each case the evidence supports the adoption of flexible constituency because *(John will go)* and (*go to the shops*) are overlapping sequences and yet seem to behave as linguistic units in the examples above. They derive a theory of flexible constituency from dependency grammar which they call dependency constituency but which I shall refer to as *flexible dependency constituency* to distinguish it from the traditional definition of dependency constituency given above which I shall *call classical dependency constituency*.

The definition of flexible dependency constituency can be made simply using concepts from graph theory. If we take a dependency diagram for a sentence and treat it as a graph, as detailed above, then the Flexible Dependency Constituents (FDCs) of that sentence are those words which form connected sub-graphs of the dependency diagram *i.e.* any group of words in the sentence forms an FDC if it is possible to trace a path from one of the words to any of the other words along the dependencies between them (ignoring the direction of the dependencies). We can also impose the further constraint that these FDCs are continuous *i.e*

that their words form a contiguous sequence within the sentence. For example, if we have the sentence:



The FDCs are (the big dog), (big dog), (big dog sneezed) and (dog sneezed) plus the sentence as a whole and each of the individual words. (The dog) is not an FDC because it is not continuous although it would be acceptable if we were to accept discontinuous sequences. (The big) is not a constituent under any analysis because its elements do not form a subgraph and are only linked whenever *dog* is introduced as an intermediary.

Pickering and Barry argue that FDCs offer the right type of syntactic unit to characterize syntactic phenomena such as coordination, ellipsis, gapping and extraction. Pickering (1991) also applies it to explaining incremental interpretation in human sentence processing, arguing that the FDC provides a more parsimonious unit for incremental interpretation than conventional constituents. Perhaps most importantly for the current purposes, they also draw parallels between FDCs and intonational phrase structure.

5.3.1 Flexible Dependency Constituency and Intonational Structure

In order to pursue this last point I will first consider Selkirk's (1984) explanation of intonational structure and then consider the application of flexible dependency constituency to the same task.

Selkirk argues that any apparently syntactic conditions on where 'breaks' in intonational phrasing may occur 'are, we claim, ultimately to be attributed to the requirement that the elements of an intonational phrase must make a certain kind of semantic sense.'

In order to capture this notion, she defines the 'sense unit' as follows:

Two constituents C_i and C_j form a sense unit if (a) or (b) is true of the semantic interpretation of the sentence:

a) C_i modifies C_j (a head)

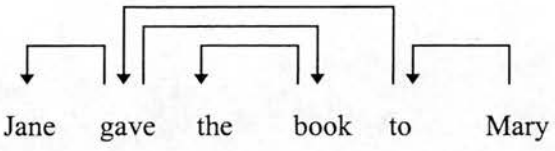
b) C_i is an argument of C_j (a head)

She demonstrates how the intonational structure of sentences can be described with these sense units. For example, she outlines the following possibilities for the intonational structure of the sentence *Jane gave the book to Mary* which she argues also match the sense units for the sentence.

1. (Jane gave the book to Mary)
2. (Jane) (gave the book to Mary)
3. (Jane gave the book)(to Mary)
4. (Jane gave) (the book) (to Mary)
5. (Jane) (gave the book) (to Mary)
6. (Jane) (gave) (the book) (to Mary)
7. *(Jane) (gave) (the book to Mary)
8. *(Jane gave) (the book to Mary)

The two marked (*) examples are not valid intonational structures because they contain the unit (the book to Mary) which is not a valid sense unit. On the other hand, example 4, is a valid intonational structure because although the unit (Jane gave) is not a valid phrase structure constituent, it is a valid sense unit. Selkirk, arguing that her semantic account of intonational structure offers a better explanation than any syntactic account, asserts that "the burden of proof is now on the advocates of the syntax based approach to show that the facts reviewed here reduce to some essentially syntactic generalisation"(p. 329). Just such a syntactic generalisation has been provided by Pickering and Barry. The sense unit proposed

by Selkirk seems to be very similar, if not identical, to the flexible dependency constituent. The dependency diagram for the sentence used in Selkirk's example looks like this:



Remembering that an FDC is a contiguous sequence of words which forms a connected sub-graph of the dependency diagram, we get the FDCs (Jane gave)(Jane gave the book)(gave the book)(gave the book to Mary)(the book) (to Mary) as well as the individual words and the full sentence. These match the intonational constituents in Selkirk's example - for example, they include the grouping (Jane gave) but not (the book to Mary).

This example, at least, would seem to support Barry and Pickering's suggestion "that the range of possible intonational units can be derived from the notion of dependency constituency" (p. 19-20) and further, that the intonational structure of a sentence can be described syntactically. Thus there may indeed be a direct correlate of syntactic structure available directly to the learner in the input provided that we characterise syntactic structure in terms of flexible dependency constituents.

Turning to the case of functional markers; I have already drawn a parallel between the high-frequency marker approach and the functional frames in Garrett's model of speech production. Of the latter, Garman (1990) notes that word exchanges (at the positional level) "frequently occur within the same constituent phrase (which may better be characterised in terms of phonological criteria - a phonological phrase)" (p. 392). Like the intonational phrases in Selkirk's example, we may be best able to describe the phrases enclosed by functional frames using flexible dependency constituency. Certainly, the problem of markers cutting across phrase boundaries as shown by the subject-predicate examples earlier, may be solved if it can be shown that the phrases which are 'bracketed' by such structures do observe phrasal boundaries within this more flexible version of constituency - which is certainly the

case with the grouping together of the subject and verb of a sentence. A more detailed test of this hypothesis is presented in Chapter 6.

Under this analysis we begin to meet the criteria mentioned earlier, namely that we would expect that different kinds of structural cues would lead the learner to converge on the same representation and that such a representation would be useful in the acquisition of syntax.

It is interesting that despite the similarities between Selkirk's 'Sense Unit' account and the Dependency Grammar account, the former is characterized as a semantic argument whilst the latter is framed as syntactic. Pickering (1991) has argued that Flexible Dependency Constituents correspond closely to what, intuitively, we may chararacterise as the minimal units for which some semantic representation may be produced in incremental sentence processing.

5.3.2 The Use of FDCs in Language Learning

How would the learner benefit from receiving structurally packaged FDCs in the input? By definition, the FDC is a grouping which forms a connected subgraph of the full sentence's dependency diagram, therefore given an FDC, we know that each word in it must be connected to at least one of the other words in the group. In the simplest example, if the group consists of two words then there must be a dependency between them. Note that this is not true of any pair of words in a sentence, or even any pair which appear together. Thus, if presented with such information in the input, the language learner could constrain their analysis to those groups of words between which dependencies actually do exist - rather than giving equal attention to all co-occurrences as is the case in the 'window in time' approach of models such as Elman's (1991), the co-occurrences are constrained by the cues in the input - which is reminiscent of Hudson's (1984) description of dependencies as 'constrained co-occurrences'. The question of how this knowledge can help the learner will be addressed in more detail later.

While Barry and Pickering argue that flexible constituency is very useful for describing phenomena in adult language, it is important to stress that the syntactic representation used by the child when initially learning their language is not necessarily the same as the syntax of the full adult language. If we posit that the adult grammar is more powerful or complex than that of the child then it is necessary to consider how we can viably progress from the latter to the former.

5.4 A Constructivist View of Grammatical Development

5.4.1 The Continuity Hypothesis

An account which constrained the learner to too simple a grammar may prove useful in explaining early child language but would fall down when trying to explain the full adult grammar (such problems blighted the pivot grammar approach, see for example Ingram, 1989). Also, we want to avoid an account in which the transition from adult to child involves a radical re-learning of the grammar - there is no point in learning one grammatical representation if it must be thrown away in order to start again. Some reinterpretation of the syntax during the course of learning may be acceptable but generally it is preferable to provide an account in which new forms are built on top of, rather than replacing, existing forms.

Pinker (1987) argues that the syntactic representation used by the child must be upwardly-compatible with that of the adult. His Continuity Hypothesis (Pinker, 1984) proposes that children adhere to UG principles throughout. An alternative view is given by Susan Ervin (1964) who argues that children produce sentences which have regular patterns that are different from those of adult language - she suggests that we might best characterise child language as employing a series of different grammars.

If we assume that the power of a phrase structure grammar such as is characterised by GB/X-bar accounts is necessary to explain adult language then a theory of language acquisition should show how the child can acquire rules couched in these terms. Cook and Newson (1996) distinguish between strong and weak versions of the continuity hypothesis as well as

maturational hypotheses such as Radford's. They describe Gleitman's discontinuity view in which the child's representation of language metamorphoses from a semantic to a grammatical phase in a way that is analagous to the change from a tadpole to a frog. Rather than maturation or discontinuity we may view changes in the child's grammatical representation as a constructivist process.

5.4.2 A Non-Stationary View of Constituency

One problem with the learning of syntax is the apparent need to begin with a representation which is close to that of the final state. I argue that, on the contrary, the power of a grammar can increase as learning takes place as a result of learning. As Quartz and Sejnowski (in publication) explain:

> *"the constructivist learner builds this hypothesis space as it learns, and so is characterized as a process of activity-dependent construction of the representations that are to underlie mature skills"* (p. 29)

The theory must be upward-compatible - there are problems with theories which break such a continuity hypothesis. However, there is more than one way in which continuity may be maintained - it is not necessarily the case that children must use PSG from the start. Such an assumption requires by definition that children's constituents would be those of PSG for constituency is at the heart of the theory - it is a primary feature of the theory.

In fact, just such a progression may be possible within the grammatical framework which has been outlined thus far.

I have already discussed three different definitions of constituency within dependency grammar. To recap, the notion of constituency is, unlike with PSG, not basic to DG but constituency can be defined in terms of dependency structures. The classic dependency constituency models produces structures which are similar to X-bar trees but which are 'flatter' and which lack stacked nodes. Covington has shown how DG can be made

equivalent to X-bar by treating different kinds of dependency relationship with different attachment properties thus allowing full equivalence between the two. Finally, the concept of flexible dependency constituency provides a direct link between the marked structure in the surface level of language and the dependency syntax which underlies it.

The use of FDCs in the input can help the child learn the dependencies that exist between words in the language but this would not mean that the child would be restricted to such a definition of dependency. Given a knowledge of the dependency grammar for a language it would be a simple matter to develop the classical model of constituency within dependency grammar - remember that this model states that a constituent consists of any word plus all of its subordinates. The adoption of such a definition would not require the disposal of any of the knowledge so far acquired but could simply be added on to the existing representation to give a new, rigid concept of constituency which, as has been mentioned earlier, provides a restricted form of X-bar theory. Furthermore, Covington's (1994) reinterpretation of DG shows how we can further redefine this classical dependency constituency to bring it much closer to standard X-bar constituency, and again this re-definition does not throw away the existing representation, but rather builds upon it.

So we can propose an account in which:

1. the initial unit of syntactic representation is the FDC, this is used to help the learner acquire the correct dependencies for the language.
2. Once the learner has acquired the ability to formulate a dependency representation for sentences, it is a simple matter to use this knowledge to derive classical dependency constituents from the input.
3. Further refinements to this knowledge, through the addition of extra information concerning the different types of modifier which can exist and restrictions on the way in which these are attached to their controller can further refine the syntactic representation to the point where it is equivalent to a X-bar account.

This gives us the following progression in the learner's syntactic theory:

Flexible dependency constituents -> Dependency Relationships-> Classical dependency Constituents -> X-bar Constituents

If the course of acquisition follows such a path we might ask why the learner would have need for a dependency representation of language at all rather than simply beginning with an X-bar type grammar.

5.5 Dependency Grammar and Learnability

5.5.1 Why is DG more Suited to Learning?

X-bar theory and GB both require knowledge of dependencies between words - consider that there is of course a dependency between the subject and verb of a sentence, and this relationship is not purely a semantic one but is also exhibited by the number agreement between the two elements. If a child focused attention on the relationships emphasised by the PSG bracketing they would be led to overlook this important syntactic restriction. As Joshi (1991) points out, the PSG can only encode this relationship indirectly:

> *A CFG cannot locally (that is, in one rule) encode the dependency between a verb and its arguments and still keep the VP node in the grammar.* (p. 1244)

He suggests that lexicalised, or word based grammars are better suited for `integrating structural and statistical information in a uniform manner.' A similar view is expressed by Lafferty, Sleator and Temperley (1992) who argue that a lexically based grammar can more easily capture the 'proclivities' of individual words.

I argue that the learning mechanism should start bottom-up in that higher-level structure should be grounded in constraints induced from the information that is immediately available in the speech stream. Even without considering the marker hypothesis, a dependency based representation forms a more appropriate beginning for the task of language acquisition. De

Marcken (1996) provides an analysis of the problems that underlie traditional unsupervised learning approaches to language: such as the use of the inside-outside algorithm for estimating the parameters of stochastic context free grammars. One problem he identifies is that the search space is more complicated than it need be due to the effect of different interacting rules, he suggests that this problem can be overcome if we 'flatten' the search space by using a dependency-style grammar - in this case he suggests the link grammar of Sleator and Temperley (1991). Link grammar is similar to dependency grammar in that it deals in relationships between words although it does not make the contoller-dependent distinction but instead uses bidirectional links, labelled with surface syntactic relations, between words.

Intuitively, we can say that because the DG representation of a sentence generally contains fewer nodes than the PSG equivalent it provides a smaller search space. Also, the nodes of the DG representation are all realised phonologically as words - unlike the phrase structure representation which contains higher level nodes which have no such direct realization. Thus the child's initial search will be restricted to relationships between words rather than more abstract, higher level nodes. This approach makes sense because the correct configuration of the higher level nodes cannot be made in the absence of information from the lower levels and generally unsupervised learning algorithms make "the naive assumption that nonterminal expansions are statistically independent.....(which) causes many problems for statistical induction algorithms" (De Marcken, 1996, p. 14).

Thus, initial learning may take place in this more restricted search space and the knowledge thus acquired can be used to constrain the development of higher level nodes at a later stage.

5.5.2 Syntactic Heads

The concept of head has been applied differently within different traditions so it is necessary to clarify what is meant, ultimately I want to maintain an approach which is as theory-neutral as possible -I do not want to propose a model of learning which depends critically on the idiosyncracies of a single theory but would rather show how such a model can discover

syntactic information which would be useful within a variety of approaches. So, for example, above I showed how the knowledge of a dependency grammar can be useful in the generation of a PSG. Similarly, I want to clarify the notion of grammatical head so that we may avoid 'burning our boats' as far as reliance on a particular theory is concerned. Bloomfield (1933) gives the following definition of head:

> *In subordinate endocentric constructions the resultant phrase belongs to the same form class as one of the constituents, which we shall call the head,: thus ' poor John ' belongs to the same form class as 'John ' which we accordingly call the head; the other member, in our example 'poor', is the attribute."* (p. 232)

The terms used here are somewhat circular - a construction is endocentric if one of its elements is of the same form class, or can be subsituted for, the whole. An endocentric construction is'subordinate' if it has only one such member. Compare Bloomfield's description to that of Pollard and Sag (1987, p. 53):

> *Each phrase contains a certain word which is centrally important in the sense that it determines many of the syntactic properties of the phrase as a whole.*

In early structuralist linguistics, the head of a construction was that element which shared the distribution of the whole construction, whereas in contemporary theories it is whichever element that projects its categorial features onto the phrase a a whole. These two characterizations will no doubt result in a lot of overlap but the structuralist approach is based on a descriptive approach whereas the modern conception is more concerned with the abstract features of thegrammar. In dependency grammar the idea of a head is often characterised in terms of pre-supposition, for example, given the phrase C=A+B then if A presupposes B but B does not presuppose A then B is the head. Another view is that B licences the appearance of A and provides the connection between A and the rest of the sentence. For example, in the sentence "Happy people laugh", the appearance of"Happy" is licenced by the appearance of "people" which is its only connection with the rest of the

sentence - it is only indirectly connected to the verb. To highlight a difference between the two approaches we can take an example from Pollard and Sag - "Sandy likes bagels", in this case "likes" is the head of the VP and the lexical head of theVP and the sentence. However, whereas in DG we can say that"likes" is the head of "Sandy likes", this would not be a valid question in PSG because heads operate within phrases and "Sandy likes" is not a valid PSG phrase. This relates back to the discussion on Subject-Verb agreement and for now I will simply re-iterate that it would be preferable if the language learner is able to directly ask questions such as "What is the head of 'Sandy likes'?" because it enables the direct use of co-occurrences within the speech stream without having to first master the complexity of the hierarchical phrase structure which rules out such questions.

5.5.2.1 Functional Heads

Approaches such as that of Abney(1987) which advocate functional heads have a natural correlate in the current theory - marker elements, which tend to be functional elements, are used to syntactically label the lexical elements with which they co-occur and also mediate the syntactic relations between phrases. Thus phrases obtain their syntactic characteristics from functional elements and this suggests that the concept of functional headedness may arise out of the statistical structure of the language.

5.5.3 Induction of Head-Dependent Relations

The approach that I adopt is somewhat similar to that of the structuralists in that headedness is assumed to be a property of distribution - this is because such an approach does not preseume existing knowledge of grammar. More precisely, the input to the learner will be treated as pre-packaged dependency constituents. For example, the child may hear minimal pairs such as 'big dog','kiss mummy' etc. and can treat them as dependency constituents which implies that the two members of the pair are related by a dependency. Furthermore, if the child never hears the word 'big' on its own, but does hear 'dog' on its own then it can be established that 'big' presupposes, and is therefore dependent on, 'dog'. However, while there is no doubt that children do hear such minimal pairs, such examples form only a small

number of the utterances which they hear, and they also suffer from the lack of a discriminating context i.e 'kiss daddy' and'big dog' have no obvious distinguishing structural features which mark them as different kinds of phrase. There is also the problem that although these phrases do have a head element, these minimal pairs offer insufficient contextual information to enable us to reliably decide which element shares the distributional characteristics of the entire phrase. For example, the verb 'Kiss' seems no more, or even less, likely to occur in isolation than its dependent 'daddy'. 'Isolation' is the key word here - minimal pairs provide the useful fact that a dependency exists between the two members, but the isolation from a sentential context and the structural cues provided by such a context, means that much potential information as the syntactic category of the phrase is unavailable. So while other learning models make much use of such highly simplified input as a key to acquisition, I argue that 1) to rely on such simple sentences would greatly reduce the number of utterances which are used in learning and, 2) such utterances actually provide less useful syntactic information than more complex sentences. But more complex sentences give us a wealth of possible dependencies so how does the learner avoid the computational complexity and the irrelevant 'noise' that would result if all possible co-occurrences were considered? The answer, I suggest, is in the use of structural cues and, specifically in the context of the current work, the information provided by high frequency morphemes. The specific predictions are that high-frequency morphemes:

1. serve to segment, or bracket, the input into phrasal units which are, specifically, dependency constituents and
2. that they serve to mark these phrases for syntactic category.

The syntactic category of a word or phrase is defined by its pattern of occurrence within these high-frequency frames and so can be classified and compared in order to decide on headedness relations and other syntactic features which I will outline.

Given the sentence *This is a big house*, the learner can segment the unit 'big house' and label it as appearing in the context 'a ____ ##'. Assuming that marked units are FDCs, we can thus establish that there is a dependency between *big* and *house*. Furthermore, using the labelling information, we can compare the phrase *big house* with other elements which occur in the context 'a ___ ##' for example *man, old man, dog, house, old house.*

- We can establish equivalence classes of words and multi-word phrases - thus *man, dog, house* and *old house* can be grouped together.
- Through comparison within these equivalence classes, certain simple recursive structures can be established e.g. given a category *C1* which has amongst its members *house* and *big house* we can establish that a C1 category can consist of a C1 category member preceded by another element.
- By comparing the occurrence patterns of the individual words which constitute a phrase, some knowledge of headedness may be acquired *e.g.* given that *house* and *old house* can both appear in the context 'a ___##' and that *old* cannot appear alone in such a context, we can establish that the presence of *old* in this context depends on, or presupposes, the presence of house - thus fulfilling one of the traditional criteria for head-dependent relationships.

The analysis above dealt only in terms of individual contexts in order to keep the description simple, however, in practice, it is desirable that the initial analysis by the learner is based on cross-comparisons between a number of different contexts. Thus the equivalence of two forms will be measured on the basis of their occurrence across a range of different contexts and can take account of the statistics of the input in order to iron out the effects of noisy input. Such an analysis will be considered in depth in the next chapter. However, one problem with such statistical approaches is their reliance on a large number of examples - while this may not be a problem for developing a core vocabulary of reasonably frequent words, it causes problems for rarer constructions which may include not only infrequent

words but also multi-word phrases whose occurrence frequency will be less than the individual words which comprise them. Ideally the learner should be able to make syntactic decisions about words on the basis of limited examples - that adults have such an ability is shown by the fact that we can make grammatical sense of nonsense verse such as Jabberwocky, so that on hearing '..the slithy toves' we can use the word *tove* in novel sentences such as 'what is a tove?'. Such an ability would be particularly useful in the early stages of language acquisition before grammar has been acquired and the use of markers as outlined in this thesis may offer a pre-grammatical approach to this task but this is not without problems. The contexts provided by the high-frequency markers are highly variable in their usefulness at providing unambiguous syntactic marking. So, for example the context 'The ___ is' is very reliably associated with lexical NPs and would thus be very useful to the learner, but other contexts are not so useful - for example, a plethora of different syntactic forms may appear in the context 'and ____ ##' (For example, 'They go and **jump**' , 'He likes Peter and **John**', 'It is big and **yellow**', 'They looked inside and **outside**' where the words in bold appear in the same local context.). Employing wider contexts may help overcome such problems but this leads to a vast increase in the number of relationships that need to be considered and therefore the amount of data that needs to be analysed. Ideally, therefore, we would like our theory to provide some means by which the learner may, prior to acquiring the grammar of a language, be able to compare the relative usefulness of different contexts so that they may form useful decisions about the syntactic form of a construction on the basis of limited examples. This issue moves away from the theoretical linguistic to the more statistical side of the current theory and so its treatment will be deferred until the next chapter.

To summarise: dependency relationships are basic to both DG and PSG, it is possible to proceed smoothly (i.e. not discontinuously rather than easily) from a simple dependency representation to a GB/ X bar grammar. Given dependency relations and Heads we can

impose a phrase structure representation on the basis of later development without the need to discard what has been acquired.

5.5.4 Structure Dependence

One important feature of natural language which is used to support the idea that langauge is represented by a formal grammar and sometimes to support the nativist hypothesis (*e.g.* see Cook & Newson (1996)) is that linguistic operations (transformations for example) operate on structures rather than words or numerical positions. For example, in the transformational approach (the principles outlined here are equally applicable to more recent accounts in PSG) the question *"Is John fishing?"* is formed by a transformational operation on the kernel sentence *"John is fishing"* whereby *is* is moved to the front of the sentence. Similarly *"Can John fish?"* and *"Was John fishing?"* are formed by fronting *can* and *was*. The straw-man which is set up for the argument of structure dependency is the idea that we can provide a general rule for this transformation which says that we move the second word in the kernel sentence to initial position. Such an argument is easily falsified by presenting examples such as "The men can fish" which would be transformed into "Men the can fish", and "The group of men are fishing" which would produce "Group the of men are fishing". The point is that the transformation is not acting on absolute postitions but, rather, on grammatically defined structures - *John* is equivalent to *The group of men* rather than *men*. Arguments of this kind were used by Chomsky to demonstrate the inadequacy of simple behavioural or statistical models for characterising grammatical structure. N-gram models, markov chains and other typical statistical models tend to operate on relationships between absolute positions. So for example a K-means procedure may be used to generate language by producing words based on the probability of occurrence of words following a certain number of prior words. There are actually two problems here, one is that language tends not to operate on absolute positions, the other is that grammars operate on word classes rather than particular words. The latter problem gives rise to the famous example 'Colourless green ideas sleep furiously'

a sentence which is perfectly grammatical but which consists of sequences of words which would have a very low rate of occurrence in language (at least, would have had if Chomsky had not made it such a famous example). In fact, simple statistical models face a two-fold problem when it comes to measuring their parameters - either we restrict ourselves to only local co-occurrences (bigrams of adjacent words) in which case we have no chance of capturing long-distance dependencies in even quite simple sentences, or we extend our scope to measure the probability of long sequences of words in which case we face a massive increase in the number of possibilities that must be considered while at the same time restricting the possibility of what can be produced - we reduce the number of ungrammatical sentences accepted by the model but we also reduce the number of grammatical sentences (such as 'colourless green ideas...') which will be acceptable. Some of these problems are eased if we consider relationships between classes of words rather than words, in which case 'colourless....' becomes A+A+N+V+ADV - a fairly commmon sequence. Furthermore, simple statistical models have shown some degree of success in categorising words from raw text (Finch & Chater, 1991) and thus provide a means for overcoming the bootstrapping problem by providing tagged input for later stages of learning. That such models are successful despite their use of absolute rather than structurally dependent relationships is no doubt largely due to the noise-resistance of statistical methods and the fact that function words provide a frequent and reliable indication of the syntactic class of neighbouring elements. The fact that function words are used in these models is generally not because of a specific theoretical position like the one adopted in the current work - but simply because such words provide the most frequent cues, so again the child need not have an innately specified instruction to focus on function words but may simply have a general learning procedure which makes use of frequent items purely because of their frequency. Considering the lack of structure dependence in statistical models we may consider a predictive Markov model (or neural net) which is attempting to predict the next word in a series Given the word 'The...' the model will most likely predict that it will be followed by a noun, but there will be some chance of the word being an adjective, similarly, given'The big...' we might again

predict a noun as being most likely, but again another adjective is possible. The prediction provided by the system would be more accurate, powerful and useful for evaluation by the learner, if it could predict that following 'the..' we expect an NP category of indeterminate length. So rather than saying N with 80% chance, we would say NP with 99% chance. The example holds for the process of classification too, nouns are often immediately preceded by a determiner but adjectives also have this property - statistical systems can overcome this noise but it would be better if they could reduce the noise. The use of high-frequency items to segment the input into units can, if the theory of structural packaging is correct, provide this increase in power. For example, the occurrence of N-bar categories such as'dog', 'big dog', 'big hairy dog' are usually marked at the beginning by a high-frequency determiner and are often marked at the end by either a plural stem -s, the end of a sentence, or another function word such as 'is'. This tendency has been noticed by several researchers (See for example Kimball, 1973) and has been used in designing parsers. The aim here is to test the claim on a large corpus of natural language and this will be covered in the next chapter, but for now I will consider some further implications of the theory. In a sense, we can say that, under this approach the concept of structural dependence is strongly cued by the input. For example,we would expect the system to detect the syntactic or distributional equivalence of 'dog' and 'big dog' - a fact that may be discovered by more brute-force statistical strategies but only by considering a much greater set of possible relationships with the concomitant computational complexity. Given this information, there are a number of resulting benefits:

- The very concept of structure dependence and that a word and a phrase can be equivalent arise naturally from the marker hypothesis.

- Following on from this, is the concept of head-dependent relations- there is a dependency between 'big' and 'dog' and it is an unequal one - 'dog' and 'big dog' occur in similar frames but 'big' only occurs on its own within different frames,thus the occurrence of 'big' in the context 'the big dog'is only made possible by the occurrence of 'dog' and thus big pre-supposes 'dog' and is, consequently, dependent on it. Thus,'dog' is the head and

'big' the dependent. Notice how the provision of contextual information provided by the marker item'the' gives us more than was available in minimal pair examples('kiss mummy', 'good boy' etc).

- Given the fact that we can can establish the equivalence of phrases such as 'dog', 'big dog' 'cat' and 'friendly cat' and that, by their occurrence in different contexts, we can establish the non-equivalence of 'big' and 'friendly', it becomes possible to learn simple recursive structures.

5.5.5 The Naturalness of Grammatical Relations

Dependency operates between individual words and thus offers a more direct representation of the input encountered by the child, there are certain respects in which the features of DG can be considered to have natural equivalents in other learning and perceptual domains. For example, the notion that two words are connected is equivalent to their meaningful co-occurrence (as Hudson points out, it is a constrained co-occurrence) . For example, a pet cat may learn to associate the rattling of its food bowl with the arrival of its dinner- the reliable association of the two events leads to a meaningful dependency to develop between them. There may be other events which sometimes co-occur with the arrival of food, the honking of a car horn outside, the opening of a door, someone switching on a television - but only the rattling of the food bowl is reliably associated with the appearance of food and leads to the formation of a meaningful co-occurence relation or dependency to form between them. Similarly,we take such a 'natural' approach to the concepts of head and dependence. Hudson (1984) uses the example of houses and dustbins -the fact that a dustbin appears is dependent on the presence of the house and not vice-versa similarly the expression 'no smoke without a fire' describes a dependency relationship in which smoke is caused by, or dependent on, fire. What I wish to emphasise is that certain key aspects of dependency grammar can be treated as perceptually basic *i.e.* they can be thought of in terms of general cognitive capacities. In this case, the reliable co-occurrence of two words will cause an association or dependency to

develop between them and the direction of the dependency can be considered in terms of a cause and effect relationship - of two words in a relationship, one can occur on its own and it is its presence which licenses the occurrence of the other one.

There are two key points - more powerful representational mechanisms than that provided by simple associative/behaviourist accounts are not unique to language. Such representation may be learned, particularly if the language is suitably structured in order to encourage such learning.

5.6 A Case Study

5.6.1 Carroll and Charniak's Model

Carroll and Charniak (1992) experimented with a computational model of language learning in which they attempted to derive a dependency grammar for an artificially generated corpus. Their work is especially relevant here, not just because of its adoption of the dependency grammar formalism but for the way in which the author's assumptions differ from those of the current theory.

The input to their model was an artificially generated, tagged, corpus, *i.e.* the input stream consisted of word category tokens rather than actual words, thus they make the implicit assumption of a strictly layered approach to language acquisition in which the ability to categorise words is assumed to be complete before any grammatical development occurs, an assumption which I question and will discuss shortly. Carroll and Charniak adopt dependency grammar because it offers a more constrained search space for the learning algorithm (the Inside-Outside algorithm) than a context-free phrase structure grammar due to the smaller number of nodes required to represent a sentence.

However despite the extra constraint, the search space for the task still offers a huge number of possibilities. A further constraint is added to the system by ordering the input corpus so that shorter sentences are presented first, under the further assumption that '....children are

exposed to simple language before seeing more complex language', another questionable view (Elman 1990, 1991) discusses this issue and offers a more plausible alternative in the form of a restricted input window). Rules which are rejected on the basis of exposure to the shorter sentences in the corpus are not allowed to re-emerge to explain the longer sentences - the initial input thus acts as a kind of 'kindergarten' search space that provides a basis for a more restrained approach to the later, more complex input. A further constraint limits the number of symbols which can appear on the right hand side of a rule in the grammar, thus reducing the chances of memorization taking place.Despite all of these constraints, the initial experiment proved highly disappointing - on the first trial the grammar settled on an incorrect grammar, supposing that this may have been an unlucky descent into a local minima, Carroll and Charniak repeated the experiment 300 times with a different random starting condition on each trial. They discovered that on each of the 300 trials the system settled on a different grammar, not one of which was correct! For their next experiment it was decided to further constrain the learning process by providing the model with a set of restrictions which would prevent particular dependency relationships from being formed. We can consider such constraints in terms of a binary matrix in which each cell stores the legality of a particular pair of categories standing in a head-dependent relationship - such relationships can be either allowed or forbidden. After experimenting with such restrictions Carroll and Charniak found that the system could learn the correct grammar provided that it was forbidden from using a small set of restricted relationships. The grammar used 7 word categories in all, thus allowing 49 possible dependency relationships of which 11 were actually needed to encode the correct grammar and 6 which were dissallowed in order to sufficiently narrow the search. The authors suggest that an equivalent set of restrictions in the child may be semantically based, for example a child may know from their general semantic knowledge that a particular property of an object depends on the object rather than vice-versa *e.g.* in the case of a red ball, the colour of the object is dependent on the existence of the object rather the object's existence being dependent on its colour.

The current theory differs from Carroll and Charniak's approach in a number of important ways which highlight the implications of adopting the marker hypothesis.

5.6.2 Windowing

Carroll and Charniak's model orders the corpus so that shorter and therefore simpler sentences are presented first. This is an undesirable option given that there is serious doubt that children are guaranteed exposure to specially simplified language. Elman (1991) provides an alternative to simplification of the input in the form of reduced capacity in the learner. In his model, the learner begins with a limited `attention span' which is only able to consider a limited number of words, thus effectively simplifying the input. For example, if the learner's `window' is restricted to only 4 words, then it will only be able to consider sentences of that length or less, while a longer sentence such as `cats chase dogs who chase mice who squeak' would be treated as a series of window-sized `chunks' such as `cats chase dogs who', `chase dogs who chase', `dogs who chase mice' etc .This method provides a more easily-justified alternative to the ordering of the input corpus but as Elman says, it causes the input to be more `noisy' as not all of the input chunks will constitute sentences or even phrasal units. Like Elman's approach the use of high-frequency marker frames serves to provide a window on the incoming language, however the marker hypothesis contends that there is an extra constraint in that the windows so created will tend to focus the learner on a much higher proportion of useful grammatical units and will therefore reduce the amount of noise which the system will be exposed to.

5.6.3 The Layered Approach

A second feature of Carroll and Charniak's experiments which differs from the current theory is their adoption of a strictly layered approach to the induction of linguistic units (In this respect they follow the standard of Zellig Harris discussed at the beginning of Chapter 3). By their use of category tags, rather than words, as input they make the implicit assumption that the learner only begins to consider the syntax of the language after they have

fully achieved the ability to categorise words in the language. By contrast, in the current approach, the linguistic units isolated by the marker items are of variable length and considerations such as the syntactic dependencies that occur between words and the relationship between one word and multiple word phrases occurs in tandem with the categorisation process. Thus the model departs from the strictly linear grammatical development of the layered approach.

5.6.4 Use of in Built-Constraints vs Grounded Constraints

In order to provide sufficient constraint on their learner, Carroll and Charniak provided a small set of restrictions on the allowable dependencies which could occur, such constraints could be viewed as innate linguistic knowledge -Carroll and Charniak suggested that they could instead be based on learned semantic knowledge of the world, although their only example of this was based on the correspondence between adjectives and nouns and their associated properties and objects - it is not clear that such an example would extend to cover other necessary syntactic restrictions. I take the view that the use of markers provides cues to the most important dependency relationships in the language and that this process takes place alongside the categorisation process.

The learner cannot learn everything at once. Markers serve to provide an initial set of useful syntactic building blocks which may then serve to constrain further learning. A dependency grammar offers a more constrained search space. Carroll and Charniak found that it was still necessary to provide constraints on possible dependencies in order for the correct grammar to be acquired but the marker hypothesis offers a means by which such constraints may be provided in the form of packaged input.

Morgan *et al* (1987) provide an interesting example. They note that in structures such as "This is the cat that ate the rat that stole the cheese...", the prosodic cues to structure would suggest phrase boundaries after *cat*, *rat* and *cheese*. Firstly they note that such mismatches are rare and that such sentences deviate from the the norm. Secondly they note that the prosody "suggests that the sentence has a 'flatter' structure than is the case. However the

units demarcated by prosody (and redundantly by the function word *that*) are phrases, even though they are not the proper constituents of the sentence."

This provides yet another example of the 'less is more' view that was highlighted by Elman's approach. Although a sentence may have a high-level strucure which is complex, the markers can serve to focus the learner on the more basic relationships which need to be acquired first. In addition, as I considered in Chapter Four, the sheer presence of a large number of marker in such sentences is enough to indicate that the sentence has a complex structure and could thus cause further analysis to be 'put on hold' in the earlier stages of learning.

In the next Chapter I will present a computational instantiation of the ideas outlined so far as well as dealing with certain issues which can only be explained clearly in the light of empirical evidence.

6. Corpus Analysis

In this chapter I describe three separate computational analyses designed to test aspects of the marker hypothesis. Each of these studies makes use of the English adult to child speech transcriptions from the CHILDES corpus (MacWhinney & Snow, 1985, 1990; MacWhinney 1991, 1993). The first study is an investigation of the theory that high frequency morphemes serve to segment the speech stream into useful grammatical units - defined in terms of the concept of flexible dependency constituency described in Chapter 5. The second study looks at how syntactic frames may be extracted from the speech stream and used to categorise lexical units; it also considers how the child may learn to associate particular contexts with particular categories and to evaluate the usefulness of different contexts. The third and final study considers how the learner may be able to exploit the marker structures in language in order to acquire syntactic knowledge of multi-word constituents including equivalence classes and head-dependent relationships without having to first undergo a separate stage of categorising individual words as is normal in 'layered' approaches to grammar induction.

6.1 Estimation of Cue Reliability for High Frequency Morphemes as Markers of Dependency Syntax

This study set out to test the claim that high frequency words serve to segment language into useful grammatical units. The basic procedure involved 1) producing dependency grammar representations for a set of adult utterances from the Childes corpus, 2) segmenting these utterances into groups of words bounded by high frequency words and 3) testing the

constituency of these groups of words according to the definition of flexible dependency constituency given in Chapter 5.

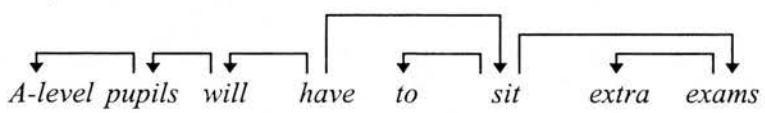
6.1.1 Overview of Method

6.1.1.1 The Parser

The sentences were analysed using a robust, large-scale English Dependency Parser (Järvinen & Tapanainen, 1997a & 1997b). Below is an example of the output of the parser for the newspaper headline *A-level pupils will have to sit extra exams*:

```
<Root> #0
<A-level>
   N SG @A> attr:>2
<pupils>
   N PL @SUBJ #2 subj:>3
<will>
   AUXMOD @+FAUXV #3 v-ch:>4
<have>
   INF @-FMAINV #4 main:>0
<to>
   INFMARK> @INFMARK> pm:>6
<sit>
   INF @-FMAINV #6 obj:>4
<extra>
   A @A> attr:>8
<exams>
   N PL @OBJ #8 obj:>6
<$.>
<$<s>>
```

The #symbol followed by a number marks a syntactic head and the > symbol followed by a number indicates that a word is dependent on the correspondingly numbered head. So, for example, the noun *exams* is dependent on the verb *sit*. This output can be used to produce a dependency graph for the sentence:



The parser is robust and will generate partial representations of a sentence if a full analysis cannot be given. This is important given the number of partial and errorful constructions in spoken language.

6.1.1.2 Definition of Constituency

A program was written which would apply a constituency test to any given sub-sequence of contiguous words from a parsed sentence. This program treats the words of the sequence as nodes in an undirected graph and the dependencies as arcs between them. It returns **true** if the resulting graph is a connected whole and **false** otherwise. Thus it can be established for any sequence whether that sequence is a flexible dependency constituent. Note that the constituency test takes into account the context in which a sequence occurs and not merely the words in the sequence. For example, the sequence *the ugly* is a valid FDC in the title *The good, the bad, and the ugly* but not in the NP *The ugly Duckling*.

6.1.1.3 Definition of Cue

The *N* most frequent words were labeled high -frequency (HF) markers and all other words were labeled low-frequency (LF) (The beginning and end of a sentence were also tagged with HF markers). A word sequence of length *L* was said to be *cued* if it consisted of *L* contiguous LF words bounded on either side by a HF marker.

6.1.2 Evaluation of Sub-sequences

Each of the parsed sentences in the corpus was analysed in a series of stages:

1) All possible sequences of length 2 or more words were extracted from the sentence. For example, in the sentence *The old man likes the woman*. The 15 possible sequences are given in the table below. The table also shows the result of the constituency test.

Sequence	FDC
The old	N
old man	Y
man likes	Y
likes the	N
the woman	Y
The old man	Y
old man likes	Y
man likes the	N
likes the woman	Y
The old man likes	Y
old man likes the	N
man likes the woman	Y
The old man likes the	N
old man likes the woman	Y
The old man likes the woman	Y

Table 6-1 - contiguous word sequences of length 2 or more in the sentence *The old man likes the woman.*with result of a constituency test

2) Each sequence was checked for constituency.

3) Sequences which were cued by high frequency markers were identified and separate statistics were kept for them as for the sequences as a whole.

In another analysis the physical length of each dependency was measured and the total proportion of dependencies at each length were recorded. The results are given in the following table:

Distance	number	proportion
1	2884	0.63
2	1112	0.24
3	372	0.08
4	128	0.03
5	39	0.008
6	17	0.004
7	16	0.003
8	7	0.002
9	3	0.0007
10	5	0.001
11	6	0.001
12	1	0.0002
13	0	0

Table 6-2 Proportion of dependencies of different lengths (Total dependencies = 4590)

These results are similar to those given in Chapter two for a smaller sample of written English and provide further support for the idea of relative distance as a form of syntagmatic

iconicity. This result is important for the view of learning outlined so far which assumes that locality is an important feature in syntax learning. This fits with the view of using a small input window in the early stages of learning.

The marker hypothesis, of course, makes further claims, specifically that linguistic input tends to be 'bracketed' by marker elements. This theory was tested on the same corpus of sentences.

6.1.3 Selecting an Appropriate Measure

There are a number of issues to consider concerning the measures which we adopt to assess the usefulness of markers. We want to know what the use of markers would 'buy' the learner and that demands a number of points to be taken into consideration.

1. What is the *cue reliability*? How often does a marker give the learner correct information?
2. What is the *cue availability*? Are a sufficient percentage of structures marked in order for the use of marking to be more than a mere sideshow in language acquisition?
3. Does the use of markers give us any significant advantage over other simpler measures such as randomly choosing any contiguous sequence of words?

Fernald and McRoberts (1996) have been critical of related research into the role of prosody as a source of syntax marking. One of the main criticisms concerns the way in which cue reliability has been evaluated. They cite the definition of Cue Reliability given by Bates and MacWhinney (1987) - 'the ratio of the cases in which a cue leads to the correct conclusion, over the number of cases in which it is available.' This definition gives us the probability of a structure given a cue - P(Structure|Cue) - which is the correct way to measure cue reliability but Fernald and McRoberts argue that much of the research which has been used to support the Prosodic Bootstrapping Hypothesis has used the quite different measure P(Cue|Structure). For example, Cooper and Paccia-Cooper (1980) found that there was a high probability of a syntactic boundary being accompanied by a pause. Such results, while necessary, are not sufficient to support the use of pauses as cues to syntactic boundaries for

the simple reason that they do not tell us about the proportion of pauses which are at positions other than syntactic boundaries. Such information is vital 'Because the infant must proceed from cue-to-structure rather than from structure-to-cue..' and thus if there are a high proportion of misleading cues then the learner would be suitably misled. Cooper and Paccia-Cooper's study was not examining the bootstrapping question but has been cited in support of it, their data did not consider 'misleading' cues but Fernald and McRoberts summarize a number of studies which together suggest that 'only about 50% of pauses occur at sentence, clause, and phrase boundaries'. This means that about half of the time, the occurrence of a pause acts as a misleading cue.

Another question reviewed by Fernald and McRoberts is that of whether IDS contains more reliable prosodic marking than ADS. Here the data appears to show a much higher cue reliability, but, the authors caution, this may be due to a very lax definition of syntactic constituency in which all utterances are defined as sentences or clauses. Fernald *et al* (1989) found that 40% of utterances were sub-clausal and a high proportion were also sub-phrasal.

The current study cannot be directly compared to these results because of the use of a non-standard view of constituency. However, it does enable us to make measures of cue validity within the current paradigm. How then do we evaluate the results? It is easier if we consider sequences of a particular length separately. In a given corpus there will be a certain number of sequences of length L.

We can analyse the following variables.

How many contiguous sequences of L words are there in the corpus? (S)
How many of these sequences are cued by high frequency markers? (C)
How many of the sequences (S) are FDCs?
How many of the cued sequences (C) are FDCs?

Any given sequence can be grouped into one of four categories on the basis of two binary features - 1) constituency - does the sequence constitute a sub-graph of the dependency representation of the sentence in which it occurs? and 2) whether or not the structure is cued. This gives us the following contingency table:

	Constituent+	Constituent-
Cue+	A	B
Cue-	C	D

In the current analysis, a cued structure is a contiguous sequence of lexical items bounded by markers. The measure of Cue Validity is given by A/(A+B) - the number of times that a cue correctly marks a constituent over the total number of cued sequences. Cue availability is given by A/(A+C) - the number of Consituents which are marked over the total number of constituents which occur.

Another measure of interest is the probability of a sequence chosen at random being a valid constituent. This value is given by (A+C)/(A+B+C+D) - the number of sequences which are constituents over the total number of sequences. This measure acts as a baseline with which to compare the utility of the high frequency markers and we can use the χ^2 statistic to gives us a measure of the statistical significance of our result.

The baseline measure (the probability of a sequence chosen at random being a constituent) can in itself be thought of as a cue validity for the cue of proximity. Because of syntagmatic iconicity, the chance of adjacent words forming a constituent is higher than the chance of words chosen at random from a sentence forming a sub-graph as it is more likely that a word will form a dependency relationship with a neighbouring word than a more distant one. As I have already argued this proximity effect can itself be considered a form of structural marking.

6.1.4 Results

Originally, 1000 utterances of Adult speech from the CHILDES corpus were selected at random. Only utterances with four or more words were used in order to try and eliminate short, stock expressions such as *hello* or *good boy* and to avoid biasing the data with isolated phrases which would be most likely to improve the cue validity - the average length of the resulting utterances was 7.25 words. Due to a technical error in using the parser 46 of the original utterances were lost. The final analysis made use of the remaining 954 utterances of

which 269 were fully parsed (proportion = 0.28). This figure belies the fact that the majority of the unparsed sentences were assigned a partial parse which included most of the words in the sentence in two or more dependency subgraphs. The language in the corpus contained a mixture of structurally simple child-directed utterances and more complex adult-directed utterances. Because only sequences of length four or more were considered, a large proportion of the simplest utterances in the corpus were filtered out.

For the parsed corpus the following values are given: for each sequence length *n*, the total number of sequences of that length, the proportion of sequences which are FDCs, the number of sequences which are cued, and the proportion of these which are FDCs. Three separate analyses were run with the number of markers set at N=57, N=30 and N=10. As we can see from the following tables, the measure of cue reliability given in the table as p(fdc|cue) gave a significant improvement over the baseline for sequences consisting of two to four words.

Seq. Length	Segments	p(fdc)	Cued Segments	p(fdc|cue)	χ^2
2	5298	.544	504	.758	102.6**
3	4353	.459	222	.604	19.7**
4	3411	.394	91	.571	12.4**
5	2470	.372	55	.4	0.2
6	1742	.313	14	.5	2.3

*Table 6-3 Cue validity using 57 marker items (** = $p<0.005$, otherwise n.s.)*

Seq. Length	Segments	p(fdc)	Cued Segments	p(fdc|cue)	χ^2
2	5298	.544	493	.720	67.8**
3	4353	.459	273	.652	43.7**
4	3411	.394	149	.557	17.4**
5	2470	.372	95	.505	7.5*
6	1742	.313	39	.385	.96

*Table 6-4 Cue validity using 30 marker items (** = p<0.005, *=p<0.01, otherwise n.s)*

Seq. Length	Segments	p(fdc)	Cued Segments	p(fdc|cue)	χ^2
2	5298	.544	410	.717	53.5**
3	4353	.459	280	.689	63.9**
4	3411	.394	203	.562	25.5**
5	2470	.372	150	.5	11.3**
6	1742	.313	70	.414	3.5

*Table 6-5 Cue validity using 57 marker items (** = p<0.005, otherwise n.s.)*

Reducing the number of markers did not greatly affect the results, although for longer sequences it increased the cue availability.

Overall these results support the marker hypothesis. Children who made use of high frequency items to segment the speech stream into units for analysis would significantly increase their 'hit-rate' for isolating grammatically useful sequences.

The Cue Reliability values are not as high as might have been hoped - a considerable proportion of marked structures would not constitute FDCs. However, there is no doubt that the use of such markers will yield a significant benefit to the learner. Although the Cue Reliability is higher than that cited by Fernald and McRoberts for pauses as prosodic markers (approximately 0.5), the results cannot be directly compared as the studies which they analysed dealt in terms of traditional phrase structure constituents rather than the dependency constituents used in the current work. An interesting area for further study would be to test prosodic marking within the current syntactic paradigm.

Because this analysis made use of only those utterances which consisted of four or more words, there can no 'whole utterance' effect for the shorter sequences *i.e.* stock expressions such as 'good boy' or 'all gone' which would tend to be FDCs are omitted from the analysis and so cannot be boosting the Cue Reliability. However such short sequences may themselves be a useful source of FDCs for the learner.

In order to provide further comparison, another analysis was carried out in which the definition of the cue was reversed so that cued sequences now consisted of sequences of high frequency words bounded by low frequency words:

Seq. Length	Segments	p(fdc)	Cued Segments	p(fdc|cue)	χ^2
2	5298	.544	256	.199	129**
3	4353	.459	72	.347	3.7
4	3411	.394	21	.190	3.7
5	2470	.372	7	.429	0.1
6	1742	.313	4	0	1.8

*Table 6-6 57 markers - reversing normal method i.e. consider sequences of high-freq items bounded by low-freq items. (** = $p<0.005$, otherwise n.s.)*

This analysis provided a surprising result - adjacent pairs of high frequency items are considerably less likely to form an FDC than word pairs chosen at random. It is worth considering this result in relation to ideas mentioned earlier in the thesis. In many ways, the simplest principle for the learner to adopt is that adjacent words are related - in the absence of any other guiding factor this would be the most commonsense option and it makes sense in light of the fact that dependencies most occur between immediate neighbours. However, this result suggests that such an approach would not work so well for high-frequency items. This fits with the idea that the number of high frequency items in an utterance give some measure of its syntactic complexity and gives added justification to the theory that children's initial efforts to construct utterances should focus on low-frequency items.

A final analysis was made which regarded any sequence of low frequency words, whether bounded by high frequency words or not, to be cued. As the results below show, this cue again performed significantly better than the baseline but only for two word sequences and cue reliabilty scores were lower than the bounded sequences for all sequence lengths.

Seq. Length	Segments	p(fdc)	Cued Segments	p(fdc|cue)	χ^2
2	5298	.544	1584	.648	97.7**
3	4353	.459	687	.480	1.5
4	3411	.394	294	.405	0.2
5	2470	.372	123	.350	0.3
6	1742	.313	43	.326	0

*Table 6-7 57 markers - sequences of low frequency items without consideration of bounding (** = $p<0.005$, otherwise n.s.)*

6.1.5 Conclusion

By making use of longer utterances and a mixture of ADS and CDS this analysis provided quite a tough test for the marker hypothesis. It also did not distinguish between markers or make use of any constraints which might serve to indicate to the learner that some markers are more useful than others.

The pattern of the results indicate that high-frequency markers can indeed serve to focus the learner's attention on domains in which dependencies tend to occur.

This supports previous suggestions (Braine, 1963; Valian and Coulson, 1988; Morgan *et al,* 1987; Gerken *et al,* 1990*)* but most of the supporting evidence for such claims has been based on artificial grammar learning or on hand-picked example sentences and hunches. The current paradigm allows us to undertake a quantitative analysis of the availability and reliability of markers in spoken language.

Equally encouraging is the finding that adjacent pairs of high frequency markers are considerably less likely to constitute an FDC. This gives further motivation for the frequency filter approach outlined in chapter four. Function words tend to be associated with more complex sentence structures and this also seems to apply to the more localised internal structure of sentences. The overall pattern of results suggest that the best strategy for the learner to follow in seeking syntactic dependencies is to focus on contiguous sequences of lower frequency words, preferably bounded by high frequency words.

6.2 Extraction and Evaluation of Marker Frames

The previous study supports one claim of the marker hypothesis namely that markers serve to divide the speech stream into non-arbitrary syntactic units. The second main claim made for markers is that they serve to label lexical units.

Brill (1993, p 47) describes the 'layered' approach to language analysis suggested by Harris (1951) in his Structural Linguistics i.e. first approaching the problem of finding word categories and then learning about how these categories combine. There are a number of reasons for this:

1. Sparse data - there are more examples of word class co-occurrences than of word co-occurrences.

2. The 'chicken and egg' problem - rules refer to classes, classes are only needed if we have rules - unsupervised categorisation enables us to break into this system.

3. Sentences which may have a low occurrence probability at the word level, such as 'colourless green ideas sleep furiously', can have a high probability of occurrence in terms of grammatical structure i.e. Adj-Adj-N-Verb-Adv or even NP-VP depending on the level of our analysis.

The approach outlined here to some degree, conflates the two separate stages by allowing similar treatment for individual words and small phrases. Any lexical sequence which is bounded by two marker elements can be treated as a single unit of analysis *i.e.*

<M1_X_M2>

where M1 and M2 correspond to particular markers and X represents a variable lexical sequence which appears between them.

Braine(1963) argues that because of their frequency, functional items will more readily form associations with one another. Because lexical categories have a much higher type-token ratio, lexical positions will tend to be filled by a much wider range of elements and thus the

frequency of any particular morpheme in a lexical position will tend to be diluted. (consider the number of forms 'The ... is', 'The man is', 'The dog is', 'the bucket is', 'the old lamp is..'. These marker frames are obviously similar to Zellig Harris's short environments except that they must be extracted from language, without the aid of a linguist, by an unsupervised learner.

6.2.1 Discovery of frames

We can operationalise Braine's suggestion as follows:

1. Select the N most frequent words or morphemes as our set of markers X_1 to X_N - in the current study we used the 200 most frequent words. We also include a marker (##) to represent the beginning and end of utterances.

2. Next we scan through a corpus and count the occurrence of all possible sequences X_m-* - X_n where X_i is a marker and * is a wild-card which matches any non-marker element.

3. Finally, we select the 200 most frequent marker sequences as our set of marker frames.

6.2.2 Initial categorisation

The approach which I adopt for obtaining the initial categorisation of words is based on that used by Finch and Chater (1991; 1992). In this research a set of focal words were categorised on the basis of the contexts in which they appeared. Finch and Chater defined context in terms of the co-occurrence with a set of context words in one of 4 serial positions (-2, -1, +1 & +2). In this study the context is defined in terms of the occurrence of a word within a marker frame.

We first of all select a set of focal items - these are the words which we want to categorise. We then scan a corpus, recording all occurrences of one of our focal items within one of our marker frames. We record these occurrences in contingency table in which each row corresponds to a focal item and each column to a marker frame. A row in the table therefore constitutes a vector which records the occurrence of a particular focal item within the range of marker frames.

For example, the sentence '## **The** man **is** big ##' would contain two focal-frame pairings - (*man* in the context ***the*** __ ***is***) and (*big* in the context **is**__##). The following table shows how a small example contingency table might look after processing this example:

	The__##	**The__is**	**a__##**	**is__##**
man	0	**1**	0	0
house	0	0	0	0
nice	0	0	0	0
dog	0	0	0	0
big	0	0	0	**1**

Table 6-8 Example contingency table

The context vector for *man* would thus be [0,1,0,0] and that for *big* would be [0,0,0,1]. This process continues until the whole input corpus has been analysed.

It is then possible to assess the relative similarity of focal elements by the similarity of their vectors - this is based on the assumption that words of the same category will occur in similar contexts.

One method of assessing this similarity is to think of the context vectors as representing the position of each of the focal items in a multi-dimensional *context space* and assuming that the Euclidian distance between words in this space corresponds to their similarity in terms of syntactic class.

At this stage, however, it is necessary to remove frequency effects from the table by normalising the vectors - we are interested in the direction of the vector rather than the absolute position in the space - say that one word appears in a particular context 5 times while another appears in the same context 10 times - although the two words appear in the same context, the more frequent element will be distanced from the other element along the same dimension. By normalising the two vectors we remove this kind of frequency based disparity. The resulting vectors act as record of their words occurrence profile in terms of the different marker frames.

6.2.3 Correlation Matrix

6.2.3.1 The labelling problem.

In Chapter 3, I described the use of marker elements in human artificial grammar learning. Valian & Coulson (1988) came to the conclusion that markers should be both of a high frequency relative to the elements which they mark, and reliably associated with a particular type of phrase. In English, the determiner *the* fulfills both these criteria but the case for other potential marker elements is not so clear cut. For example, *and* is notoriously promiscuous in what it can conjoin - not only a wide range of different constituent types but also a wide range of units which are not traditional phrase structure units :

She went to Rome and Paris
I have a bucket and spade
It helps you work, rest and play
It is orange, green and red
It is small and fast
It went up and down
He went home and ate
Read it and weep

The point is that *and* alone does not constitute a reliable marker and although the wider sentential context may serve to aid the labelling of unknown words this demands that either we already know a sizeable portion of the language or we make use of an unfeasibly large distributional context in our learning procedure. Statistical approaches can circumvent this problem by taking into account the range of contexts in which a word appears. However, adult speakers can reliably make these kinds of decisions on the basis of single examples (and the research reviewed in Chapter 2 suggests that fairly young children can too). But much of this ability assumes the existence of a knowledge of the language, the problem that we face is how to bootstrap this ability. How does the child know which contexts are reliable and which are not?

In order to evaluate the contribution made by each template to the categorisation of a particular focal item I adopt the following procedure:

1. Select the first focal item F.

2. Measure the distance from F to each of the other focal items - write these values to a new column in the contingency table C0.

3. For each column in the contingency table (Cx (x=1 to N)) calculate the Spearman correlation coefficient between C0 and Cx - write the resulting value into (row F, column x) of the correlation table.

4. Repeat steps 1-3 for the next focal item.

This procedure gives us a correlation matrix which records the degree to which a focal item's categorisation profile is correlated with each of the context items. Each row in the table represents the degree to which a focal item's category is predicted by each of the contexts. Each column in the table shows the relation between a particular context and each of the focal items.We can thus find, for example, the best predicting context for a particular focal item (the word `dog' is best predicted by the context `The X Stop')

The correlation table can be thought of as representing the degree to which a focal item's category is determined by a particular context, it is somewhat similar to Quinlan's ID4 system which extracts efficient decisions trees from raw data by determining the degree to which different features act as determinants of category membership.

The next step is to look at the spread of values associated with a particular context. We extract the column in the correlation table which corresponds to the template. This gives us a list of values corresponding to the association between that template and each of the focal items. We then sort this list into ascending order, this gives us a continuum of values with the most highly associated elements at the top. For example, the template `The X STOP' produces a list with nouns at the top.

We can draw a graph for these values which will give us a downward sloping curve - the x-axis corresponds to the sorted focal items and the y-axis to the association with the template.

Taking the second derivative of this graph gives us a visual representation of sudden changes in the associative values. The correspoding graph for the template `The X STOP' show a marked `spike'. Examining the word list reveals that this sudden change in associative values corresponds to the `boundary' of the noun category - all the nouns in the sample appear to the left of the spike.

This method is similar to processes used in models of vision to detect object boundaries - in this case the boundary marks the edge of a category.

The correlation table can be thought of as representing the degree to which a focal item's category is determined by a particular context, it is somewhat similar to Quinlan's ID4 system which extracts efficient decisions trees from raw data by determining the degree to which different features act as determinants of category membership.

6.2.4 Finding zero crossings

Each of the templates is associated with each focal item to some degree - this means that each template gives a linear continuum of associativity. We can thus order focal items according to their level of association with a template and, adopting an idea from computational vision we can look for sudden changes in the level of association which occur in this linear series by taking the second derivative of the series. The following example shows a graph of the second derivative for the template (the X STOP) (An example of which would be X=`big dog' in the sentence `She likes the big dog'):

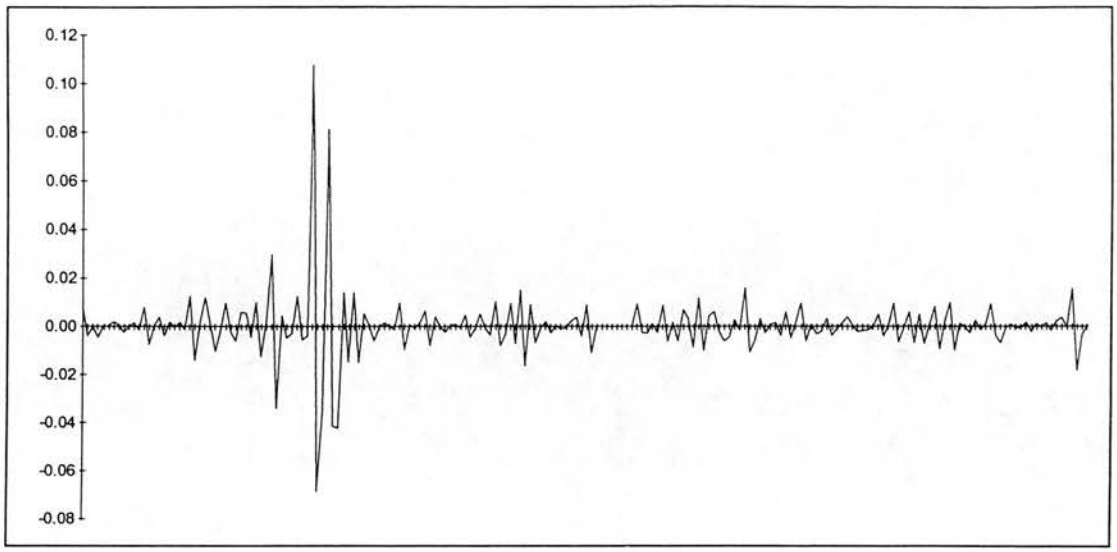


Table 6-9 Second derivative of correlations for the frame 'The ___ ##'

Looking at the sorted list of words associated with these values, we can see that the occurrence of the sudden jump in this graph corresponds quite closely to the boundary of the noun category. (The break in the list shows the position of the largest zero-crossing in the graph:

THE_X_STOP word associates
floor, matter, bag, door, box, table, story, lady, tree, house, truck, car, book, water, ball, dog, paper, chair, baby, cup, other, girl, horse, man, top, mommy, daddy, boy, time, bed, way, picture, morning, room, cat, piece, kind, color, back, name, minute, thing, milk, mouth, lot, tiger,
.
school, ride, hand, big, too, two, um, , nice, huh, which, him, very_good, alright, sure, hi, long, them, uh, thank_you, hmm, who, good, where, more, really, else, uhhuh, hm, mmm, mmhm, yep,

A similar separation is found for the frame <THE X OF>:

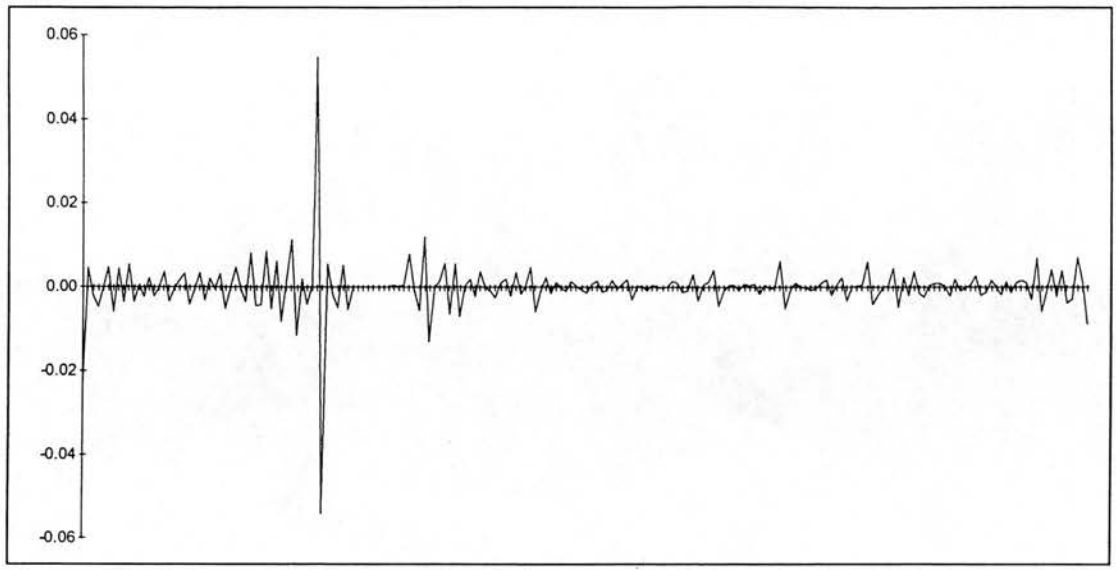


Table 6-10 Second derivative of correlations for the frame 'The ___ of'

THE_OF word associates
kind, piece, lot, way, lady, thing, bag, name, picture, minute, time, matter, tree, story, room, top, house, dog, book, car, ball, box, girl, cup, table, chair, paper, tiger, door, floor, truck, color, man, horse, boy, morning, mouth, baby, bed, ride, water, too, back, mommy, cat, daddy, other,

looks, happened, hand, first, else, school, wait, whatis, again, train, big, milk, them, him, if_you, over,

and we see a similar result with the frame <YOUR X STOP>:

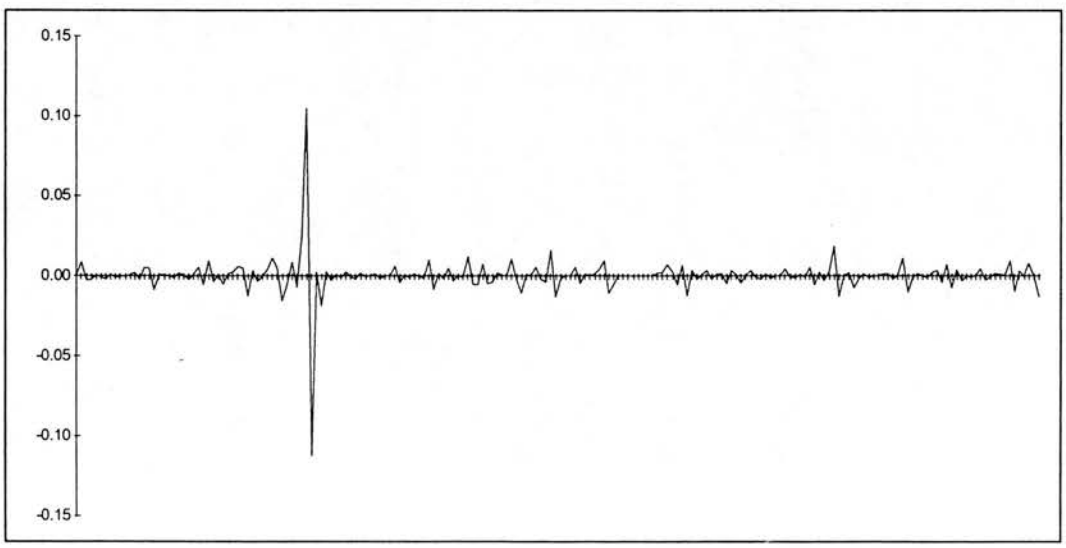


Table 6-11 Second derivative of correlations for the frame 'Your ___ ##'

<YOUR_X_STOP> word associates
room, mouth, mommy, paper, name, horse, chair, book, cup, bag, daddy, bed, car, other, house, ball, dog, story, door, way, truck, water, matter, box, table, tree, lady, girl, baby, top, picture, man, time, floor, boy, school, milk, morning, back, hand, cat, piece, kind, tiger, thing, minute, lot, color, ride,

big, too, huh, very_good, alright, which, sure, thank_you, hi, nice, hmm, hm, long, uhhuh, mmm

Contexts which are associated with verbs produce less reliable results. In the following graph for the frame <TO X IT> there is no obvious division corresponding to a definite category boundary:

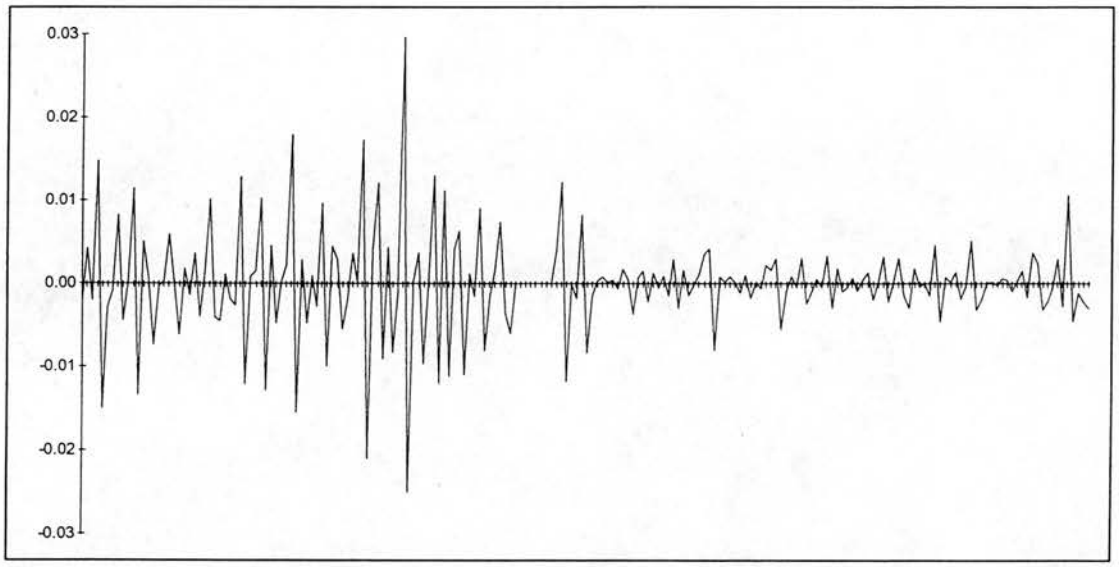


Table 6-12 Second derivative of correlations for the frame 'To __ it'

If we look at the associated words, however, we can see that they consist predominantly of verbs:

<TO_X_IT> word associates
get, use, read, hold, take, leave, give, open, show, tell, make, you_get, wheres, try, look_at, you_tell, how_about, heres, you_got, turn, got, let, whos, theres, need, thatis, does, was, had, eat, thought, play, at, itis, or, isnt, be, has, said, if, theyre, ride, doesnt, im, say, you_doing, write, hes, just, then, first, youre, but, shes, because, maybe, you_going, so, talk, come, those, youre_going, looks, over,

should, him, watch, her, about, guess, school, these, going, could, sarah, them, you_say,

didnt, you,

These result suggest that based on the analysis of a fairly small amount of input data the learner could begin to develop the ability to use marker contexts to label syntactic categories and, to a certain degree, to find 'natural' category boundaries

6.2.5 Rating templates

In order to rate the ability of templates at discriminating between categories we can calculate the Standard Deviation of the correlations associated with each template. Higher SD values will correspond to a greater spread of associations and thus, to some extent, the degree to which the template distinguishes between categories.

Below is a list of the best and worst templates (rated by the standard deviation of their correlations with the focal words). The best associate is that word which has the highest (inverse) correlation with the template.

SD	TEMPLATE	Best Associate
27.127783	the X STOP	floor
24.478167	a X STOP	tiger
23.708300	START X the	wheres
15.494007	your X STOP	room
15.435098	START X a	it is
14.843438	this X STOP	way
14.194753	the X and	box
13.828815	that X STOP	thing
13.462290	the X in	book
13.060888	START X that	whos
13.047486	START X it	leave
12.968562	the X to	car
10.418649	the X with	story
10.158907	START X your	leave
10.023393	the X on	truck

The most discriminating contexts

1.032995	do X like	ya
1.007707	think X STOP	so
0.968307	a X a	lot
0.957667	did X do	you
0.930961	put X STOP	them
0.929207	know X STOP	how
0.870676	do X want	ya
0.870676	do X think	ya
0.870676	do X know	ya
0.799485	what X to	happened
0.755000	START X for	youre_going
0.719718	whats X STOP	doing
0.570668	START X know	you
0.530562	and X and	play

The least discriminating contexts

6.2.6 Assessing templates

Intuitively, we would expect that the better a template was at discriminating categories, the greater the spread of associative values that it would give. We can get a measure of this spread by calculating the Standard Deviations of each column in the correlation table - this gives us a measure of the extent of the variability of the associations connected to each template. By sorting the templates by this measure we get the following list and we would expect the elements at the top to be better predictors of class than those lower down.

The ability to assess the utility of templates in defining category is important to meet the conditions expressed by Valian and Coulson (1978) that a marker be not only frequent, but also reliable. Some markers are not so reliable, for example <and X and> can be associated with many categories and appears at the bottom of the list above, whereas <the X STOP> is reliably associated with noun phrases and appears at the top.

These results suggests how an analysis of the more frequently occurring words in a corpus can guide the development of representations which can enhance the ability of the learner to acquire further information from the input.

This is an important ability since the statistical/distributional approach generally requires a lot of data in order to reach a conclusion and yet many words occur infrequently, this problem is alleviated by the ability to produce a set of reliable templates and core categories with which they are associated.

This approach shows how the learner may detect 'natural' category boundaries from rather noisy data. It suggests how a symbolic representation may arise out of a more statistical representation. One motivation for applying such a process would be the need for an efficient representation of the input data. This ability constitutes a form of meta-knowledge - rather than simply using markers, the learner can assess the relative usefulness of different markers and this state of knowledge can be achieved without explicit instruction.

6.2.7 Conclusion

The process of categorisation involves a series of steps. Initially, the different dimensions of the context vectors are treated equally and are used to formulate an initial categorisation for a relatively small set of median frequency items. Once this has been achieved, it becomes possible to identify the degree to which the different dimensions contribute to the categorisation of particular items.

Several advantages emerge from this approach:

1) The ability to detect discrete category boundaries.

2) The ability to identify reliable contexts.

Consider the position of the child learning a language -

The fact that two words appear in the same immediate context does not necessarily imply that they are of the same category.

Over time the child will form a representation of the contexts in which different words appear (it seems likely that such representations may make use of the same kind of information which is used for speech segmentation). These representations are in some sense 'iconic' in that they reflect the raw cooccurrence data from the input - they contain no indication of which cooccurrences are meaningful and which are merely 'noise'. However, there is sufficient regularity in the input for these representations to form the basis for an initial categorisation of words. It then becomes possible to determine the degree to which different contexts are correlated with the categorisation profile of a word.

This ability to associate contexts with words and to detect discontinuities in these associations enables the learner to form discrete category boundaries and also to assess the relative discriminatory power of different contexts.

Keil(1986) and Keil & Kelly (1987) outline a theory of the way in which children categorise physical objects in the world which appears to have certain similarities with the current approach. Consider the following description from Harnad (1987a):

> *Keil & Kelly suggest that in the holistic stage the child is categorizing on the basis of prototypes: characteristic features with integral dimensions. The only way he can sort things is by their overall similarities. In the later, analytic stage, representations consist of separable, defining features. The child is then sorting things the way we do, and can usually even verbalize the basis for his categorization in the form of a rule or definition.* (p. 40)

This description parallels aspects of the current approach. Initially, the child is trying to classify a set of lexical 'objects' whch occur with sufficient frequency to provide a reasonably broad base of occurrence contexts - these items are treated holistically, in that they are compared on the basis of the contexts in which they occur, in a somewhat indiscriminate manner - the only way in which certain contexts are paid more attention is on the basis of their frequency - low frequency contexts will generally be paid little attention.

However, within the set of frequently occurring contexts, no discrimination is made between contexts which might reliably indicate a category ('The ___') and those which are less defining ('and ___'). Until some learning has taken place, the particular morphemes which act as markers cannot be distinguished between - either in terms of the categories with which they are associated or in terms of their reliabilty.

The child can initally sort lexemes on the basis of their overall similarity on the basis of a holistic representation in which no distinction is made between the different contexts (dimensions) which record a word's occurrence profile but such initial categorisations can then form the basis for a refinement of the representation which identifies the particular features or contexts which are reliably associated with particular categories. This process leads to a more symbol based representation which establishes a set of defining features which may be used to filter the input on the basis of separable contextual cues.

6.3 Estimation of Head-Dependent Relations

One feature of the frequency segmentation approach is that multi-word phrases and individual words can be classified together, this enables us to implement a structuralist approach to the identification of the heads of constituents (remember that we make the assumption that all items delimited by high-frequency items are flexible dependency constituents).

The initial approach simply treats a phrase as a lexical unit - it is categorised in an identical way. However we can takes things a stage further by examining the internal structure of these phrases.

The concept of headedness and the projection principle mean that a phrase observes the syntactic properties of its head. This means that we might expect a phrase to have similar

distributional properties to its head. For example 'old men' is interchangable with 'men' in all of the contexts in which the latter may appear.

All words in the top 200 of the rank frequency list are treated as marker items (again, this includes a marker for the beginning and end of an utterance. Because of the problem of sparse data, it was decided not to use the syntactic frames that were described in section 6.2 but rather to treat the left and right contexts as separate dimensions - this means that our table has rows consisting of 400 items consisting of the 200 possible markers occuring at the left hand side of the focal item and the 200 to the right.

The entire corpus was segmented into triplets consisting of the focal item and the left and right context items. For example `the big dog is happy' would be seperated into the two pairs:

(big dog) (the-is)

(happy) (is-STOP)

There are potentially 40000 unique contexts (200 left marker * 200 right marker). However only 18177 of these actually occurred in the corpus and only 7622 appeared more than 4 times. The top ten contexts accounted for 92502 occurrences out of the total of 412377 items (22.4%).

In the next stage, the number of occurrences of each focal item (word or phrase) were recorded and the 4000 most common retained for analysis. A contingency table was created recording the number of occurrences of each context item in each template. If stored normally, this would produce a matrix containing (4000×400 =) 1600000 items however since a large majority of these elements would be empty, a sparse matrix representation was used and this enabled the table to be stored relatively compactly.

6.3.1 Analysing short phrases

The next stage of the analysis focused on the multi-word focal items. All of the focal items which consisted of two words were analysed in a number of ways:

The 4000 focal elements consisted of one or two word phrases.

Given a two word phrase the program would calculate the following:

1. The phrase's nearest neighbour in the vector space - no distinction was made between words and phrases so that a phrase's neighbour can be either a single word or another phrase.

2. The distance between the phrase as a whole and each of its constituent words. For example, given the phrase 'big dog' we would calculate Distance('big dog', 'big') and Distance('big dog', 'dog').

3. The number of contexts shared by a phrase and its constituents. This gives us some indication of how reliable the distance calculations are.

Step 2 was intended to test the system's ability to identify the head of a phrase: for each phrase (A) which contained two words (B, C) we measured the Euclidian distance from the vector for A to the vectors for B and C - the closest element out of B or C is marked as the head element. This is an operationalisation of the structuralist approach to headedness described in Chapter 5 *i.e.* the head of a phrase is that element whose distribution matches

the distribution of the phrase. The further requirement that the phrase be endocentric can be simulated by ensuring that the minimum distance between the phrase and its elements is below a threshold value, otherwise we can assume that the phrase has a different category from either of its components.

6.3.2 Results

Below is a table showing the analysis for the top 100 phrases. The listing is sorted according to the inverse of the distance between the focal phrase and its nearest neighbour. This measure was used because it was assumed that words which were too far from any other element in the set would be unlikely to yield good results:

Focal Phrase	Head	Neighbour	Distance	Shared contexts
last-part-	last	floor	0.0231	2
white-ones-	ones	black-ones-	0.0236	4
black-ones-	black	white-ones-	0.0236	4
street-light-	light	race	0.0273	1
toy-bag-	bag	shore	0.0278	2
swimming-pool-	pool	merrygoround	0.0283	0
whole-week-	week	smartie	0.0288	3
our-town-	town	wellfleet	0.0304	1
jill-found-	found	once-upon-	0.0305	0
ah-ha-	ah	yep	0.0329	29
tape-recorder-	recorder	microphone	0.0333	40
washing-machine-	machine	kitchen	0.0351	4
next-page-	page	wrong-side-	0.0360	3
daddys-gone-	gone	james-wanted-	0.0360	1
toy-box-	box	bedroom	0.0421	4
whole-lot-	lot	lot	0.0433	1
ready-yet-	yet	finished-yet-	0.0454	3
finished-yet-	yet	ready-yet-	0.0454	4
whole-bunch-	bunch	lot	0.0490	3
letter-box-	box	milkman	0.0501	3
hello-peter-	hello	byebye	0.0505	3
flat-tire-	tire	mistake	0.0509	4
wrong-place-	place	air	0.0543	1
magic-word-	word	toy-bag-	0.0547	3
m-hm-	hm	whee	0.0547	3
hold-still-	still	whoops	0.0591	3
post-office-	post	telephone	0.0613	2
fire-engine-	fire	snowman	0.0615	7
bath-tub-	tub	bathtub	0.0622	4
minute-ago-	minute	minute	0.0652	5
bunny-rabbit-	rabbit	lion	0.0660	5
mouth-full-	mouth	birthday	0.0689	5
garbage-truck-	truck	tunnel	0.0699	3
favorite-color-	favorite	birthday	0.0711	5
todays-visit-	visit	awhile	0.0716	0
birthday-present-	present	cindy-doll-	0.0728	6
straight-line-	line	minute	0.0744	0
christine-wants-	wants	whereis	0.0763	0
god-bless-	bless	whatre	0.0798	0
bed-early-	bed	sleep	0.0812	3
coffee-pot-	pot	monster	0.0818	3

front-door-	door	kitchen	0.0827	3
itis-hard-	hard	james-wanted-	0.0876	1
milk-bottles-	milk	milkbottles	0.0958	5
uhoh-trouble-	uhoh	woop	0.0961	12
bad-word-	word	smartie	0.0964	3
door-open-	door	sidewalk	0.0995	5
kitty-cat-	kitty	screwdriver	0.0999	29
gas-station-	station	barn	0.1015	2
babys-hair-	hair	toy-box-	0.1033	2
bobby-pin-	pin	moment	0.1047	0
white-sheep-	sheep	skunk	0.1070	4
birthday-party-	party	sandwich	0.1072	12
milk-bottle-	bottle	street-light-	0.1115	4
funny-noise-	noise	prize	0.1117	2
lunch-box-	lunch	birthday	0.1178	3
being-silly-	silly	welcome	0.1199	2
choochoo-train-	train	lion	0.1209	2
cookie-monster-	monster	letter-box-	0.1228	5
cheese-sandwich-	sandwich	sandwich	0.1286	8
fire-truck-	truck	screwdriver	0.1288	5
by-yourself-	yourself	by-himself-	0.1300	12
by-himself-	by	by-yourself-	0.1300	6
high-chair-	chair	plate	0.1304	2
keep-still-	still	comeon	0.1324	5
black-sheep-	sheep	skunk	0.1395	6
best-part-	best	middle	0.1396	3
teddy-bear-	bear	camera	0.1406	5
green-coat-	coat	brown-suit-	0.1419	4
brown-suit-	suit	green-coat-	0.1419	2
sleeping-bag-	bag	nap	0.1455	2
hi-shem-	hi	sarah-marie-	0.1540	4
belly-button-	belly	birthday	0.1572	3
christmas-tree-	tree	spoon	0.1614	5
id-love-	love	james-wanted-	0.1615	0
same-size-	same	same-color-	0.1663	6
same-color-	same	same-size-	0.1663	7
wont-work	work	works	0.1717	3
birthday-cake-	cake	spoon	0.1749	7
hey-pete-	hey	ooh	0.1756	18
spare-tire-	tire	monster	0.1792	4
funny-face-	face	runny-nose-	0.1809	3
same-idea-	same	pooh	0.1877	3
whole-story-	story	middle	0.1881	3
rubber-band-	band	moment	0.1893	1
likes-smarties-	likes	hurt-himself-	0.1926	1
hurt-himself-	hurt	likes-smarties-	0.1926	3
paper-towel-	towel	nap	0.1969	9
youll-fall-	fall	frances	0.2013	4
red-ball-	ball	tunnel	0.2016	5
after-dinner-	after	careful-adam-	0.2053	7
game-before-	game	year	0.2124	5
happy-birthday-	birthday	daddys-gone-	0.2127	4
dump-truck-	truck	popsicle	0.2138	2
finished-eating-	eating	dare	0.2151	4
comes-next-	next	happens	0.2161	0
keep-turning-	keep	stop	0.2173	9
new-game-	game	passy	0.2208	3
itis-called-	itis	thereis	0.2217	3
lunch-yet-	lunch	lunch	0.2229	4

Of these 100 phrases the majority are n-bar categories and of these the head noun is correctly identified in the majority of cases. In those cases where the wrong choice is made of head, most are the result of one of two factors:

1) The phrase contains a closed-class item which exhibits unusual distributional properties (e.g. *ones, few, much*)

2) The phrase is a compound noun. In such cases it is impossible, given the method being used, to distinguish reliably between the two choices because they both share the category of the phrase as a whole.

Only five of the phrases are definitely non-constituents according to the definition of flexible dependency constituency, which is better than we might expect from measurements of cue validity given in section 6.1. Six of the phrases are of the form Subject-verb (*Christine wants*, *God bless* etc.) and in all six cases the verb is correctly identified as the head. There are also a number of VP categories in which, again, the verb is correctly identified as head. Finally there are two PPs (by himself, by yourself) in which the head is not reliably identified.

The nearest neigbour results are encouraging: most of the neighbours are of the same or a very similar category. Again, as in the previous study, noun groups are the most reliably categorised but the verb based phrases also produce reasonable results. In most cases a phrase containing a verb is categorised with another phrase containing a verb although in several cases it is not a direct syntactic equivalent.

In order to look at the performance on simple n-bar constituents more closely, a set of the best 21 phrases containing colour adjectives was extracted:

Focal Phrase	Head	Neighbour	Distance	Shared contexts
white-sheep-	sheep	skunk	0.1070	4
black-sheep-	sheep	skunk	0.1395	6
green-coat-	coat	brown-suit-	0.1419	4
brown-suit-	suit	green-coat-	0.1419	2
red-ball-	ball	tunnel	0.2016	5
grey-dog-	dog	grey-cat-	0.2469	0
grey-cat-	cat	grey-dog-	0.2469	0
green-light-	light	seesaw	0.2666	5
red-light-	light	windmill	0.2672	8
red-light-	light	windmill	0.2672	8
red-ones-	red	barrels	0.2810	4
red-spots-	spots	green-wheels-	0.2934	3
green-wheels-	wheels	red-spots-	0.2934	3
orange-juice-	juice	juice	0.3191	20
yellow-ones-	yellow	rice	0.3757	4
blue-shirt-	shirt	light	0.5660	5

blue-ones-	blue	latch	0.7080	8
green-beans-	beans	ten-dollars-	0.9948	9
brown-cat-	cat	repeat	1.2073	3
blue-bag-	bag	heater	1.2379	3
blue-eyes-	eyes	claws	1.2659	3

This analysis is very encouraging - in each case the nearest neighbour is of the same syntactic category as the target item and the head has been correctly identified except in the phrases containing the word *ones*.

7. Conclusion

In this thesis I have presented a constructivist account of language intended to address the incremental nature of child language development within an instructional environment. This account demands that we try to understand the way in which the child uses language not in the terms used in explaining adult language processing but in terms of the ontogenesis of structure through a complex interaction between the constantly changing mind of the child and the complex signal carried in langauge.

In chapter 2, I contrasted the idealised instantaneous view of language acquisition with the constructivist approach. The marker hypothesis suggests that the child's early learning experience is focused on particular elements - these elements form building blocks for later stages of acquisition. Within such a view, we can consider the child's representational space as being markedly different from that of an adult - in a way which is more fundamental than the mere lack of some component.

The initial stages of learning will offer us the purest insight into the use of such an approach. As the learner develops so will their representation and treatment of the input they receive. For example, a learner may initially focus on the relationships between words but as their representation of language becomes more complex, they will begin to consider relationships between word categories. Trying to understand such processes will require us to first understand the categorisation which the child is using. Thus if we were to try and draw parallels between a particular learning model and its final state (full adult grammar) we would need to be sure that our model mimicked human behaviour at each stage of

development - deviations in the early stages of learning will result in greater differences as the learning process progresses.

The approach takes no account of alternative sources of information, there are a number of reasons for this. The first was to see how far simple distributional information could go in explaining the structure of early child language. Secondly, there were practical restrictions on the degree to which different sources of information could be combined in such an approach - either because of a lack of such information or because of the difficulty of integrating it. The omission is not intended to suggest that such information is not used. For example, various phonological cues may serve to differentiate open- and closed-classes. (See Cutler 1993 for a review). Morgan, Meier and Newport (1987) suggest that a variety of different cues may conspire to produce a final result which is superior to the sum of its parts. Similarly, it is possible that in addition to the simple frequency based approach used here, that various other sources of information would underlie the choice of marker elements in a sentence.

However, word frequency offers a motivation for treating such elements differently while their phonological properies do not - high frequency elements act as more reliable and efficient markers in an obvious way whilst no such claim can be made for weakly stressed elements. Therefore phonological information is secondary to distributional information and the motivation for differential treatment of open- and closed-classes comes primarily from their distributional properties.

Secondly, if phonology does contribute to the omission of certain elements in child speech it is not because such elements are more difficult to perceive or process - Gerken *et al's* (1990) results suggest otherwise - rather it may be because particular phonological properties act as secondary cues to a word being a marker element and it is their status as markers which underlies the omission.

The frequency-based approach suggests that several phenomena which are normally discussed in terms of functional properties can be characterized by a much simpler explanation. I claim that the functional-lexical distinction, rather than being the result of an innately specified modular structure, arises during the learning process in response to the statistical structure of language. The fact that different categories display varying degrees of open- and closed- classedness is a natural consequence of this approach.

The fact that children do not use function words has often been taken as evidence that they cannot use them in aiding language acquisition. The marker hypothesis suggests a dichotomy between elements in language which closely corresponds to that between lexical and functional elements. Children distinguish between marker and non-marker elements in the sentences which they hear. Marker elements act as a frame for lexical elements but are not used productively. In Chapter 4, I contrasted the Frequency Filter approach with the GB theoretic account of which Radford (1990) states:

> *"From a theoretical and descriptive point of view, perhaps the most important result is that the Government-and-Binding (GB) model offers us a particularly insightful perspective on the nature of early child grammars"*
>
> Radford, 1990 p. 289

Radford claims that his GB-based account of early child language is supported by the fact that 'Apparently unconnected features are intrinsically connected and reducible to a single postulate'. However the view of learning expressed in this thesis is similarly parsimonious but has the added advantage of explaining learning. Radford's account assumes complex underlying mechanisms and does not consider learning. e.g. How does the child know what is functional? Why do they prefer objective to nominative pronouns?

The sudden spurt of functional acquisition is explained by the fact that children have already acquired functional items. The apparently sudden acquisition is the result of a transition from non-production to production (akin to realisation of two separate languages

in bilingual children).The lexical-functional distinction need not be innate and therefore we require less complex innate structure. My account is equally predictive, more parsimonious with respect to innate structure and more explanatory (of the developmental component). The simplicity of Radford's theory is misleading.

- it is the tip of an iceberg in that it assumes the existence of a complex and largely undefined underlying model (UG). It also fails to acknowledge the ontogenetic development of the child and, in doing so, fails to explain certain phenomena. In explaining the ontogenetic component we dismiss the need for the innate machinery of UG to explain these phenomena.

At the same time, it does not conflict with the description of adult language provided by GB theory - it shows how a functional lexical distinction may arise without the need for innate specification. The normal GB-UG theory is an example of innate modularity, the current account is more akin to the progressive modularisation of Karmiloff-Smith (1992).

In Chapter five I considered how such a constructivist account could be applied to grammatical representation itself. In this account I showed how a child's representational mechanism for syntax could be characterized as a set of refinements of an initial model of dependency syntax. Within this account the basis of Phrase Structure Grammar is not innate but is a response to the structure of language. Constituency is not primary but is built out of simpler linguistic building blocks.

References

Abney, S. P. (1987) *The English Noun phrase in Its Sentential Aspect*, Unpublished PhD Thesis, MIT.

Allen, G.D., & Hawkins, S. (1980). Phonological rhythm: Definition and development. In G. Yeni-Komshian, J.F. Kavanagh, & C.A. Ferguson (Eds.), *Child Phonology: Vol 1. Production.* New York: Academic Press.

Allen, J. (1995) *Natural Language Understanding (2nd Edition).* Redwood City, CA: Benjamin/Cummings.

Anglin, J.M. (1977) *Word, Object and Conceptual Development*. New York: Norton.

Baker, C.L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry,* 10, 533-581.

Barry, G. & Pickering, M. (1990). Dependency and constituency in categorial grammar. In G. Barry and G. Morrill (eds.), *Edinburgh Working Papers in Cognitive Science, Volume 5: Studies in Categorial Grammar,* Centre for Cognitive Science, University of Edinburgh.

Bates, E., & MacWhinney, B. (1987) Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Blakemore, C., & Cooper, G. F. (1970). Development of the brain depends on the visual environment. *Nature*, 228, 477-478.

Bloomfield, L. (1933). *Language.* New York: Holt.

Bock, K., (1989). Closed-class immanence in sentence production. *Cognition*, 31, 163-186.

Borsley, R.D. (1991) *Syntactic Theory: A Unified Approach.* New York : Routledge.

Braine, M. D. S. (1963). On learning the grammatical order of words. *Psychological Review,* 70, 323-348.

Braine, M. D. S. (1994). Is nativism sufficient? *Journal of Child Language,* 21, 9-31.

Brill, E. (1993) *A Corpus Based Approach to Language Learning*. PhD Thesis, Department of Computer and Information Science, University of Pennsylvania.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65,* 14-21.

Brown, R.W. (1973). *A First Language: The Early Stages*. Harvard University Press.

Carroll, G. & Charniak, E. (1992). Two experiments on learning probabilistic dependency grammars from corpora. In *Proceedings of the AAAI Workshop on Probabilistic-Based Natural Language Processing Techniques*, 1-13.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

Chomsky, N. (1963). Formal properties of grammars. In R. B. R. Luce, & E. Galanter (Eds), *Handbook of mathematical psychology*. New York: Wiley Inc.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* Cambridge, Mass.: MIT Press.

Chomsky, N. (1987). Language: Chomsky's theory. In R.L. Gregory (ed.), *The Oxford companion to the mind*. Oxford: Oxford University Press.

Chomsky, N. (1988). Some notes on economy of derivation and representation, Ms., MIT. Cited in Radford (1990).

Christiansen, M.H. (1994). *Infinite Languages, Finite Minds: Connectionism, Learning and Linguistic Structure.* PhD Thesis.: Centre for Cognitive Science, University of Edinburgh.

Clark, E.V. (1983) Meanings and concepts. In J.H. Flavell and E. Markman (eds.), *Handbook of Child Psychology*, Vol. 3: *Cognitive Development*. Chichester: Wiley.

Clark, H. H. & Clark, E. V. (1977). *Psychology and Language: An Introduction to Psycholinguistics*. New York: Harcourt Brace Jovanovich.

Cook, V.J., & Newson, M. (1996) *Chomsky's Universal Grammar.* Oxford: Blackwell.

Cooper, W. E. & Paccia-Cooper, J. (1980). *Syntax and Speech*. Cambridge, MA: Harvard University Press.

Covington, M.A. (1990a). A dependency parser for variable-word-order languages. University of Georgia Research Report AI-1990-01.

Covington, M.A. (1990b) Parsing discontinuous constituents in dependency grammar. *Computational Linguistics*, 16, 234-236.

Covington, M.A. (1994) An empirically motivated reinterpretation of dependency grammar. University of Georgia Research Report AI-1994-04.

Crystal, D. (1987). *The Cambridge Encyclopedia of Language*. Cambridge University Press.

Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, 22 (2), 109-131

Cutler, A. & Carter. D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language,* 2, 133-142.

Cutler, A. & Mehler, J. (1993). The periodicity bias. *Journal of Phonetics*, 21, 103-108.

Dawkins, R. (1976). *The Selfish Gene.* Oxford.

Eilers, R (1975). Suprasegmental and grammatical control over telegraphic speech in young children. *Journal of Psycholinguistic Research,* 4, 227-239.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science,* 14, 179-211.

Elman, J. L. (1991). Incremental learning, or The importance of starting small. In *Proceedings from the Thirteenth Annual Conference of the Cognitive Science Society*, 443-448. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ervin, S. M. (1964). Imitation and structural change in children's language. In E. H. Lenneberg (ed.), *New Directions in the Study of Language*, pp. 163-89, MIT Press, Cambridge, Mass.

Fernald, A. & McRoberts, G. (1996) Prosodic bootstrapping: a critical analysis of the argument and the evidence. In Morgan, J.L. & Demuth, K. (eds), *Signal to Syntax:Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Finch, S. & Chater, N. (1991). A hybrid approach to learning syntactic categories. *Artificial intelligence and Simulated Behaviour Quarterly,* 78, 16-24.

Finch, S. & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society,* (pp. 820-825). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fodor, J.A. (1983). *The Modularity of Mind.* Cambridge, Mass.: MIT Press.

Garrett, M. F. (1975). The analysis of sentence production. In G. H. Bower (Ed.) *The Psychology of Learning and Motivation*, Vol 9, pp 133-178.

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language Production Vol. 1: Speech and Talk.* London: Academic Press.

Gerken, L. A., Landau, B., & Remez, R. E. (1990), Function morphemes in young children's speech perception and production. *Developmental Psychology,* 26 (2), 204-216.

Gerken, L.A. & McIntosh, B.J. (1993). The interplay of function morphemes and prosody in early language. *Developmental Psychology,* 29, 448-457.

Gerken, L.A. (1996) Phonological and distributional information in syntax acquisition. In Morgan, J.L. & Demuth, K. (eds), *Signal to Syntax:Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Givón, T.(1979). *On Understanding Grammar.* New York: Academic Press.

Gold. E. M. (1967). Language identification in the limit. *Information and Control*, 16, 447-474.

Goldstein, E.B. (1980) *Sensation and Perception*. Belmont: Wadsworth.

Green, T. R. G. (1979). The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behaviour*, 18, 481-496.

Haiman, J. (Ed.) (1980), *Iconicity in Syntax.* Amsterdam: Benjamins.

Harnad, S. (1987a). Psychophysical and cognitive aspects of categorical perception: A critical overview. Chapter 1 of S. Harnad (Ed.) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harnad, S. (1987b). Category induction and representation. Chapter 18 of S. Harnad (Ed.) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harris, Z. S. (1951). *Structural Linguistics*. Chigaco University Press.

Hock, H. H. (1986). *Principles of Historical Linguistics*. Mouton de Gruyter: Amsterdam.

Horning, J. J. (1969). *A Study of Grammatical Inference*. Stanford University, Computer Science Department.

Hudson, R.A. (Mailing list). Posted on the Dependency Grammar mailing list DG-LIST on the Internet.

Hudson, R. A. (1984). *Word Grammar*. Basil Blackwell, Oxford.

Hunt, A.J. (1995). Syntactic influence on prosodic phrasing in the framework of the Link Grammar. In *Proceedings of Eurospeech*.

Ingram, D. (1989) *First Language Acquisition: Method, Description and Explanation*. Cambridge: Cambridge University Press.

Järvinen, T., & Tapanainen, P. (1997a). A dependency parser for English. Technical reports No. TR-1, Department of General Linguistics, University of Helsinki, Finland.

Järvinen, T., & Tapanainen, P. (1997b). Dependency parser demo. Descriptions of System Demonstrations and Videos. in the 5th Conference on Applied Natural Language Processing (ANLP'97). pages 9-10. Association for Computational Linguistics, Washington, D.C.

Joshi, A.K. (1991). Natural language processing. *Science*, 253, 5025, 1242-1249.

Jusczyk, P. W., Kemler Nelson, D. G., Hirsh-Pasek, K., Kennedy, L., Woodward, A., & Piwoz, J. (1992) Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology*, 24, 252-293.

Jusczyk, P. W. (1993). Discovering sound patterns in the native language. In *Proceedings from the Fifteenth Annual Conference of the Cognitive Science Society*, 49-60. Hillsdale, NJ: Lawrence Erlbaum Associates.

Karmiloff-Smith, A. (1992). *Beyond Modularity: ADevelopmental Perspective on Cognitive Science*. Cambridge, Mass. MIT Press.

Katz, B., Baker, G., & Macnamara, J. (1974). What's in a name? On the child's acquisition of proper and common nouns. *Child Development*, 45, 269-73.

Keil, F.C. (1986). On the structure dependent nature of stages of cognitive development. In I. Lewin (Ed.) *Stage and Structure: Reopening the Debate*. Norwood, NJ: Ablex.

Keil, F.C. & Kelly, M.H. (1987). Developmental changes in category structure. In S. Harnad (Ed.) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press

Kimball, J. (1973). Seven principles of surface structure parsing in natural language. *Cognition,* 2, 15-47.

Kiss, G.R. (1972). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, 7, 1-41.

Kurlyowice, J. (1949). La nature des procès dits 'analogiques'. *Acta Linguistica,* 5, 15-37. (cited in McMahon, 1994).

Lafferty, J., Sleator, D., & Temperley, D. (1992) Grammatical trigrams: A probabilistic model of Link Grammar. Proceedings of the AAAI Conference on Probabilistic Approaches to Natural Language, October, 1992..

Lapointe, S. G. (1985). A theory of verb form use in the speech of agrammatic aphasics. *Brain and Language*, 24, 100-155.

Lightfoot, D. (1979). *Principles of Diachronic Syntax*. Cambridge, England. Cambridge University Press.

Luhn, H.P. (1958) The automatic creation of literature abstracts. *IBM Journal of Research and Development,* 2, 159-165.

MacWhinney, B. & Snow, C. (1985) The child language data exchange system. *Journal of Child Language,* 12, 271-96.

MacWhinney, B. & Snow, C. (1990) The child language data exchange system: an update. *Journal of Child Language,* 17, 457-472.

MacWhinney, B. (1987). The competition model. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1991) *The CHILDES Project:tools for analyzing talk.* Hillsdale, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1993) *The CHILDES database: second edition.* Dublin, OH: Discovery Systems.

Manczak, W. (1958). Tendences générales des changements analogiques. *Lingua 7: 298-325, 387-420.* (Cited in McMahon, 1994).

Maratsos, M.P., & Chalkey, M. (1981). The internal language of children's syntax: the ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language,* Vol. 2. New York: Gardner Press.

de Marcken, C. (1996). Lexical heads, phrase structure, and the induction of grammar, In 1996 Workshop on Very Large Corpora.

Marr, D. (1982) *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman.

Matthews, P. H. (1981). *Syntax.* Cambridge: Cambridge University Press.

Matthews, P.H. (1991) *Morphology.*(2nd edition). Cambridge: Cambridge University Press.

McMahon, A.M.S. (1994). *Understanding Language Change.* Cambridge: Cambridge University Press.

Menn, L., & Obler, L.K. (1990). *Agrammatic Aphasia*: *A Cross Language Narrative Sourcebook.* Amsterdam Benjamins 1990.

Miller, G. A. & Chomsky, N. (1963). Finitary models of language users. In R. B. R. Luce, & E. Galanter (Eds), *Handbook of Mathematical Psychology.* New York: Wiley Inc. (pp. 419-491).

Moeser, S. D. (1977). Semantics and miniature artificial languages. In J. Macnamara (Ed.), *Language Learning and Thought.* New York: Academic Press.

Moeser, S. D., & Bregman, A. S. (1972). The role of reference in the acquisition of a miniature artificial language. *Journal of Verbal Learning and Verbal Behaviour*, 11, 759-769.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology*, 19, 498-550.

Morgan, J.L. & Demuth, K. (1996) Signal to syntax: an overview. In Morgan, J.L. & Demuth, K. (eds), *Signal to Syntax:Bootstrapping from speech to grammar in early acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Newport, E.L., Gleitman, H., & Gleitman, L.R. (1977) Mother I'd rather do it myself: Some effects and non-effects of maternal speech style. In C.E. Snow and C.A. Ferguson (eds.), *Talking to Children: Language Input and Acquisition.* Cambridge: Cambridge University Press.

Peretric, P.A., & Tweney, R.D. (1977). Does comprehension precede production? The development of children's responses to telegraphic sentences of varying grammatical adequacy. *Journal of Child Language*, 4, 201-209.

Pickering, M. & Barry, G. (1991) Sentence processing without empty categories. *Language and Cognitive Processes*, 6, 229-259.

Pickering, M. & Barry, G. (1993) Dependency Categorial Grammar and coordination. *Linguistics*, 31, 855-902.

Pierce, J.R. (1961) *Symbols, Signals and Noise: The Nature and Process of Communication.* New York: Harper.

Pinker (1982). A theory of the acquisition of lexical interpretive grammars. In J. Bresnan (Ed.), *The Mental Representation of Grammatical Relations*. Cambridge, MA: MIT Press.

Pinker, S. (1979). Formal models of language learning. *Cognition*, 1, 217-283.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pinker, S. (1994). *The Language Instinct.* London, England: Penguin.

Pollard, C., & Sag, I.A. (1987) *Information Based Syntax and Semantics: Vol. 1 Fundamentals.* CSLI lectures notes number 13. Chigaco: Chigaco University Press.

Price, G. (1984). *The French Language: Present and past.* London: Grant and Cutler.

Quartz, S.R. & Sejnowski, T.J.(In submission) The neural basis of cognitive development: A constructivist manifesto.

Quinlan (1979) Discovering rules by induction from large collections of examples. In D. Michie (ed.) *Expert Systems in the Micro-Electronic Age.*

Radford, A. (1990) *Syntactic Theory and the Acquisition of English Syntax: The Nature of Early Child Grammars of English.* Oxford: Basil Blackwell.

Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8, 382-439.

Schultz, C.K. (1968) *H.P. Luhn: Pioneer of Information Science - Selected Works.* London: MacMillan.

Selkirk, E.O. (1984). *Phonology and Syntax: The Relation Between Sound and Structure.* Cambridge, MA: MIT Press.

Shillcock, R. C., & Bard, E. G. (1993). Modularity and the processing of closed-class words. In Altmann, G. T. M. & Shillcock, R. C. (Eds.), *Cognitive models of speech processing*. The second Sperlonga Meeting. Erlbaum.

Shillcock, R.C. & Tait, M.E. (1994). The processing of closed-class morphemes in normal and agrammatic language use: convergence of psycholinguistic and formal approaches. Poster presented at TENNET, Montreal, May, 1994.

Shillcock, R.C., Hicks, J., Cairns, P., Levy, J., & Chater, N. (1995). A statistical analysis of an idealised phonological transcription of the London-Lund corpus. Submitted to *Computer Speech and Language*.

Shillcock, R.C., Kelly, L., Buntin, L. & Patterson, D. (1996). Visual processing of content and function words: similarities and differences in the Clarity Gating Task.

Shipley, E.F., Smith, C.S., & Gleitman, L.R. A study in the acquisition of language: Free responses to commands. *Language,* 45, 322-342.

Singleton, J. L. & Newport, E. L. (1993). *When learners surpass their models: The acquisition of American Sign Language from impoverished input*. Ms. Department of Psychology, University of Illinois at Urbana-Champaign.

Sleator, D.D.K., & Temperley, D. (1991). Parsing English with a link grammar. CMU tech-report CMU-CS-91-196. Carnegie Mellon University.

Sokal, R.R. & Sneath, P.H.A. (1963). *Principles of Numerical Taxonomy.* Freeman.

Stankler, J. (1991). *Phonological Distinctions as Morphological Signals*. Unpublished Master of Philosophy thesis, Department of Engineering, University of Cambridge, Cambridge, England.

Steedman, M. (1991) Structure and intonation. *Language,* 67, 260-296.

Steedman, M. (1996) Phrasal intonation and the acquisition of syntax. In Morgan, J.L. & Demuth, K. (eds), *Signal to Syntax:Bootstrapping from Speech to Grammar in Early Acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Svartik, J., & Quirk, R. (eds.) (1980). *A Corpus of English Conversation.* Lund Studies in English 56. Lund: Lund University Press.

Tesnière, L. (1959) *Elements de Syntaxe Sructurale*. Paris: Klincksieck.

Valian, V. V. & Coulson, S. (1988). Anchor points in language learning: the role of marker frequency. *Journal of Memory and Language*, 27, 71-86.

Zipf, G. (1935). *The Psychobiology of Language.* Houghton Mifflin.

Zipf, G. (1949). *Human Behaviour and the Principle of Least Effort.* New York: Hafner.