A Generalized Cue-Based Approach to the Automatic Acquisition of Subcategorization Frames

Khaled Elghamry

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Linguistics
Indiana University

May 2004

Acknowledgements

I ask my professors, who know how much I respect them and appreciate their guidance, to kindly forgive me that I commence the acknowledgements by thanking my father who left this world while this dissertation was in the making, and my mother. Their simple parental words of unconditional love, support, and encouragement over the phone have always given my academic steps in the U.S. meaning, substance, and roots.

No words could really express my gratitude to my committee members who have deeply affected the formative years of my academic career at Indiana University. My thesis director, Damir Ćavar, has always been there to answer my repeated questions in every step in the process of writing this dissertation. When I started thinking of a dissertation topic, most of his out-of-class time, which he would rather spend with his family, he spent sincerely helping me find a topic that is worth investing 3 years of my life working on. He has always been patient and ready to help. Writing the algorithms and codes for this dissertation could have been much more difficult without his help. Mike Gasser knows how the whole story has been going from my very first days at IU. I learned a lot from disagreeing with him in the first computational linguistics class I took at IU. His words of encouragement and respect, when he knew I changed my linguistic mind-set 180 degrees, were very valuable, as well as timely. Every conversation I had with him has some impact on one part or another of this dissertation. Steven Franks taught me *good* syntax and tried his best to make a *good* syntactician out of me. Unfortunately, I didn’t seize the opportunity. However, I learned from him the foundations of a good linguistic argument. As a chair, he worked hard to make computational linguistics a minor, and he succeeded. Computational linguistics at IU is now very promising. John Paolillo helped me know what statistics is all about, and how it could be used in linguistic research justifiably and rigorously. His comments on the first drafts of this dissertation added to the coherence and rigor of the arguments.

I would like also to thank Stuart Davis, Yoshihisa Kitagawa, Robert Port, and all the faculty at the Department of Linguistics for everything I learned from them either in class or out of class.

My stay in the U.S. would not have been as fun, enjoyable, and fruitful without the company of fellow graduate students whose eagerness to learn made me like to learn more, and whose friendship made me feel at home away from home.

Without my brothers, Mohammed and Ibrahim, and my sisters, Azza and Iman, coming to the U.S. to study could have been less meaningful. Their faith in me has always given me such a driving force to travel for knowledge and a better education. My wife, Rasha, has always been there for me, though she was also working on her dissertation. Many a time she went out of her way to make writing this dissertation manageable. I could not thank her enough for her love, support, and faith in me.

I would like also to thank my teachers in the Faculty of Al-Alsun, Ain Shams University, who taught me linguistics and made me like it. Without the financial support of my university in Egypt and the Egyptian government for the last five years, this dissertation wouldn’t have been possible.

Ann Baker (administrative assistant), and Marilyn Estep (secretary) were always there to kindly assist me whenever I needed their help.

Abstract

This dissertation has two objectives. The first is to present the formal foundations of a cue-based model of distributional learning and show how it can be used in learning subcategorization frames. The other more general objective is to give further evidence that the input plays a central role in both automatic and human language acquisition. Two implementations of this model are presented. The first comprises a set of learning algorithms that are able to identify arguments, predicates, and subcategorization. This implementation uses as bootstraps proper names and a small subset of pronouns. The other implementation does not assume any initial cues. Learning in this implementation is based only on distributional regularities in the input and the information-theoretic measure of Mutual Information. It presents a procedure for cue extraction, then demonstrates how these cues can be used in categorization and subcategorization. The efficiency of the proposed cue-based model as applied in this implementation is tested on English, German, and Japanese. The performance of the two implementations shows that the model is able to capture language-specific properties based on distributional regularities in the input. The theoretical and practical importance of the cue-based model proposed in this dissertation stems from three main reasons. The first is the need in NLP to acquire maximum lexical and structural knowledge from minimum or zero initial knowledge. The second is the evidence it provides for the possibility of natural language acquisition using a small set of cues in the input by means of distributional analysis. Finally, this model is language-independent, which makes it extendible to other linguistic learning tasks and other languages with little parameterization.

Contents

List of Tables **Page**

List of Figures **Page**

Chapter 1

Background and Motivations

Automatic acquisition of lexical knowledge is a major bottleneck in the production of high-coverage grammars and parsers (Zernik 1991, Manning 1993, Briscoe & Carroll 1993, 1997). Especially important is knowledge about verbs, which are the primary source of relational information in a sentence. The lexical representation of a verb should specify at least one subcategorization frame that generally represents the number, type, and syntactic realization of arguments corresponding to the participants in the event described by the verb.(A more formal definition of frames is given in Chapter 5).

According to Briscoe and Carroll (1993) and Carroll and Rooth (1998), up to 50% of the parse failures on unseen test data were caused either by the inaccuracy or the lack of subcategorization information in the dictionary used by the parser. A parser equipped with this type of information is able to recover the correct predicate-argument relations, which helps the parser to constrain the number of analyses and the space of possible parses for a given structure (Manning 1993). Subcategorization information is also essential in a number of theoretical as well as practical problems in parsing and lexicon construction. These problems include, but are not limited to, PP-attachment, complement-adjunct distinction, and the variation in predicates' argument-taking properties across time and discourse.

For example, an English parser equipped with the subcategorization frames in Table 1 is able to determine the legitimate PP-attachment in the sentences below. Accordingly, the first and the third bracketings are legitimate analyses, whereas the second is not.

(1) I [warned [the man] [of the storm]].

(2) *I [repaired [the motor] [of the car]].

(3) I [repaired [the motor [of the car]]].

Frame	**Structure**	**Example**
Intransitive	0	The woman walked.
Transitive	NP[obj]	John loves *Mary*.
Ditransitive	NP[direct_obj] NP[indirect_obj]	Mary gave *Peter flowers*.
Intransitive with PP	PP	I rent *in Paddington*.
Transitive with PP	NP[obj] PP	She put *the book on the table*.
Sentential complement	Clause	I know *(that) she likes you*.
Transitive with sentential complement	NP[obj] Clause	She told *me he is coming*.

Table 1: Some English subcategorization frames (Manning and Schütze 2003: 105)

When parsing without subcategorization information we are not only confronted with the attachment problem, but also with the problem of distinguishing complements from adjuncts. Adjuncts are usually not mandated by the verb’s syntactic or semantic requirements; they rather provide optional additional, contextual or background information. The following sentence illustrates some typical adjuncts (temporal, locative, manner and reason).

(4) I repaired the car (today /in the backyard /with pleasure /because I had to).

Whereas the number and type of complements is specific to each verb, adjuncts can (in theory) appear with all verbs subject only to semantic compatibility constraints. Complements and adjuncts can look very similar:

(5) She has been waiting *for two friends* (*complement*).

(6) She has been waiting for two hours (*adjunct*).

And not in all cases is the boundary between complements and adjuncts clear-cut. A notoriously difficult case arises with verbs whose function is to mainly focus on any of the above-mentioned contextual information, as illustrated by (7-9).

(7) He lives in Cairo.
(8) The good weather lasted for a week.
(9) She behaved well.

There are various syntactic and semantic criteria for the complement-adjunct distinction, and in some cases their predictions conflict.[1] For example, Meyers et al. (1994) propose some criteria for complement- and adjuncthood, as given in Tables 2 and 3, respectively.[2] It is not hard to realize the problems with most of these criteria. However, as far as this dissertation is concerned, there are two main types of problems.

The first problem is that these criteria are based on vague and problematic concepts. Especially controversial is the concept of thematic role. There is so little agreement as to its nature and definition, and a notable absence of consensus about what thematic roles are (e.g., Dowty 1991: 547, Jackendoff 1987: 371). Without a formal definition, there is no significant purpose for thematic roles to serve (Dowty 1991: 548).[3]

The other is the amount of information these criteria assume. For these criteria to work, the learner should have knowledge of phrases, theta roles, and selection restrictions, among other things. This means that the learner should be able to parse the input structures in order to distinguish complements and adjuncts. It is argued later that these and other criteria that assume similar knowledge to work suffer from a bootstrapping paradox.

[1] This point is discussed in more detail in the following chapter.
[2] XP in these tables refers to maximal projections or phrases in the generative literature.
[3] For detailed discussion of thematic roles and their status in linguistic theory see Dowty (1991), Jackendoff (1987), and references mentioned there.

1	**Obligatoriness**	XP is obligatory for VP to be grammatical or for a particular sense of V to be possible. Ex. *Mary felled the tree.* Mary felled.*
2	**Passive**	XP can only be the subject of the passive if XP is a complement. Ex. *Mary ate the cake. The cake was eaten by Mary.* Only complement PPs can be stranded by pseudo passive. Ex. *Many people lived in that mansion.* *This mansion was lived in by many people.*
3	**Theta Roles**	XP has an argument theta role: theme, source, goal, etc… Ex. *John gave Mary* (recipient, goal) *the book* (theme) .
4	**Implied Meaning**	XP is optional, it is implied if omitted. Ex. *John ate [something]* *John ate.*
5	**Selection Restrictions**	If V imposes selection restrictions on XP, XP is complement.

Table 2: Criteria for complementhood, Meyers et al. (1994)

1	**Frequency**	XP occurs with most verbs with roughly the same frequency and meaning.
2	**Typical Adjuncts**	Purpose clauses, PPs/AdvPs/Subordinate clauses headed by ‘before’, ‘after’, ‘while’, ‘because’, etc…
3	**Selection Restrictions**	An adjunct does not impose selection restrictions on the verb/VP.
4	**WH words**	AdvPs/PPs which can be questioned with ‘why’ or ‘how’.
5	**Fronting**	Adjunct PPs front more naturally than complement PPs.
6	**Island Constraints**	Adjuncts cannot usually violate ‘island constraints’[4].

Table 3: Criteria for adjuncthood, Meyers et al. (1994)

[4] An island in generative grammar refers to a structure out of which constituents cannot be moved by any movement rule.

It has also been shown that predicates change behavior between sub-languages, domains and across time (Korhonen 1997; Roland and Jurafsky 1998). In a series of corpus-based experiments, Roland and Jurafsky (1998) showed that subcategorization frequency variation is caused by factors including discourse cohesion effects of natural corpora, the effects of different genres on verb sense, and the effect of verb sense on subcategorization. For example, they found that in clear cases of polysemy, such as *accuse* and *bill* senses of *charge*, each sense has a different set of subcategorization probabilities. In a series of psycholinguistic experiments, Hare et al. (2003) reported similar results on the effects of sense and discourse on subcategorization variation.

Given this variation in the predicate's frame behavior, it is therefore obvious that frames are indispensable for efficient parsing. A grammar used for parsing must have access to an accurate and comprehensive dictionary encoding (at the very minimum) the number and category of a predicate's arguments.

Several large machine-readable subcategorization dictionaries are already available for English. Some typical examples of these are the ANLT dictionary (Boguraev et al. 1987) and COMLEX Syntax dictionary (Grishman et al. 1994). However, these dictionaries are built either manually or largely automatically from machine-readable versions of conventional dictionaries. Manual development of large subcategorization dictionaries has proved inefficient in compiling fully accurate or comprehensive lexicons.

According to Briscoe & Carroll (1997), the close connection between sense and subcategorization and between subject domain and sense makes it unlikely that a fully accurate static subcategorization dictionary of a language is attainable in any case.

Moreover, given that predicates change behavior between sub-languages, domains and across time, dictionaries produced by hand always lag real language use.

What is needed then is a general method for the automatic extraction of subcategorization information from corpora of unrestricted texts. A subcategorization dictionary obtained automatically from corpora can be updated quickly and easily as different usages develop.

Several methods have been suggested for learning subcategorization frames automatically from text corpora for English (e.g., Brent 1993), Manning (1993), Ushioda et al. (1996), Briscoe & Carroll (1997), Korhonen (1997), and Buchholz (1998).[5] Though the majority of these methods were originally concerned with practical problems in computational linguistics, they have raised fundamental theoretical as well as practical issues regarding (i) the role of the input in learning, (ii) the nature of categorization and subcategorization frames, and (iii) learning procedures.

While much previous theoretical research has been limited to highly idealized artificial input or to *a priori* considerations regarding the feasibility of acquisition mechanisms, these methods are mainly corpus-driven and consequently focus on the distributional regularities in the input as the main source of linguistic information. These methods could thus provide a source of hypotheses for experimental test (Redington & Chater 1997).

[5] For frame identification in other languages, see Zeman and Sarkar (2000) for Czech; Maragoudakis et al. (2000) for Modern Greek; Basili et al. (1997) for Italian; Eckle et al. (1996) for German; and Kawahara et al. (2000) for Japanese.

Seen in this wider perspective, the input-driven mechanisms employed in these methods could be more empirically and theoretically rigorous if they are improved in three aspects.

The first is that these methods did not entertain any formal definitions of frames, and consequently assumed arbitrary and subjective characterizations of these frames. This lack of formalization would definitely result in inconsistent and relatively non-standardized frame and consequently lexical knowledge. Accordingly, frames are formally defined in Chapter 5.

The second is that these strategies assumed a PoS-tagged or partially parsed input, with the exception of Brent's strategy which leveraged minimal initial knowledge in the form of function words and proper names. That is, the learning algorithm is given PoS and structural knowledge in order to extract subcategorization information which will then be used in parsing, and sometimes in categorization. Assuming for now that bootstrapping is attaining new knowledge on the basis of already existing knowledge, these previous methods result in a bootstrapping paradox. This point is discussed in more detail in the following chapter.

Though the most promising of these methods, Brent's method, on the other hand, could be improved in two different ways. Brent does not introduce a formal definition of cues, nor does he show how cues could be learned. Moreover, the frame cues suggested by Brent are English-specific, and will have to be learned by the algorithm to begin with.

In order for these and future input-driven methods to be of more empirical, theoretical, as well as practical value, they should be based on well-defined learning mechanisms that have two main properties, that is, language-independence and minimal or no *a priori* knowledge. This is what this dissertation is presenting.

This dissertation has two interrelated objectives. The first is to present the formal foundations of a model of cue-based distributional learning. This model introduces formal definitions of cues and frames. Assuming minimal initial knowledge, this model demonstrates how these definitions can be used in procedures for learning these cues from corpora and how these cues can be used as bootstraps for learning frames, among other things.

The other more general objective is to give further evidence that the input plays a central role in both automatic and human language acquisition. Given what can be learned using the proposed cue-based model based only on distributional regularities in a given corpus, it is shown that the input contains a set of features that would facilitate human/automatic language acquisition. Two implementations of this model are presented to show how it can be used in lexical acquisition, in general, and frame identification, in particular.

The first implementation comprises a set of learning algorithms that are able to identify arguments, predicates, and subcategorization frames, among other things, via bootstrapping mechanisms. This implementation assumes a minimal set of bootstraps that are limited to proper names and a small subset of pronouns. These elements were chosen based on their logical priority to predicates and their referential nature which would facilitate their identification in the input, and enhance their bootstrapping power. Bootstrapping in this implementation is based on two main assumptions. The first assumption is that the similarity of two or more elements is a function of the similarity of their contexts (Harris, 1951). The second assumption is that two or more elements are distributionally similar if they share more than 50% of their contexts. The results reported for this implementation are based on an English corpus. Testing the efficiency of this implementation of the model with other languages still requires further research.

The other implementation does not assume any initial cues or linguistic knowledge. Rather, using distributional regularities and the information-theoretic measure of Mutual Information, it presents a strategy for learning cues first and then using them in categorization and subcategorization. The Mutual Information statistic is used to measure how much information a given context carries about a certain element, and is expected to yield a more accurate measure of distributional similarity than the relative-frequency measure used in the first implementation.

The efficiency of the proposed cue-based model as applied in this implementation was tested on three languages: English, German, and Japanese. The model was able to capture language-specific properties using only distributional regularities in the input. To establish the language-independent nature of the model as manifested in this implementation still requires further research to test it on other languages.

Both implementations assume no predefined subset of frames, since the learning algorithms are left to identify the set of possible frames, and what is an appropriate frame for a verb based on distributional regularities in the input.

The theoretical and practical importance of the cue-based model proposed in this dissertation stems from three main reasons. The first is the need in NLP to acquire maximum lexical and structural knowledge given minimal or no *a priori* knowledge. The second is the evidence it provides for the possibility of natural language acquisition using a small set of cues in the input by means of distributional analysis. Finally, this model is language-independent, which makes it extendible to other linguistic learning tasks and other languages with little parameterization.

The dissertation is organized into seven chapters. Chapter 1 provides the rationale for this research. Chapter 2 reviews the previous methods for the automatic acquisition of subcategorization frames. Chapter 3 outlines the different approaches to bootstrapping and then discusses some problems in the previous methods in terms of bootstrapping. Chapter 4 surveys the psycholinguistic evidence for the language learner's sensitivity and ability to use different cues in the input. Chapter 5 introduces the two foundations of a cue-based learning model, that is, a cue identification procedure and procedures for establishing distributional similarity. Chapters 6 and 7 present two implementations of this model. The first is a cue-based learner that bootstraps from a set of semantic cues, which are limited to proper names and a subset of pronouns. This implementation shows how the model can be used in identifying noun phrases, verbs, and frames in a given corpus. The other implementation is a more sophisticated version of a cue-based learner that bootstraps from a set of distributional cues that it learns from a given corpus. It shows how cues can be used to set, for example, the head parameter in three different languages (i.e. English, Japanese, and German). It also shows how predicates and arguments can be differentiated in a corpus based on cues. And finally it demonstrates how this knowledge augmented with knowledge about head direction can be used in identifying a set of possible frames in a given corpus. The last chapter concludes with an overall comparison of the performance of the two implementations, and their implications for practical and theoretical issues in computational and human language acquisition.

Chapter 2

Theoretical Issues in Subcategorization Frames Learning

In this chapter, I first review the probabilistic, as opposed to categorical, approach to verb subcategorization. I then present the previous methods for automatic subcategorization acquisition. Finally, I discuss the implications of these strategies both for automatic and human language acquisition.

2.1 Models of Subcategorization Frames

It was briefly mentioned in the previous chapter that, in order to assemble accurate subcategorization information, a distinction should be made between complements and adjuncts. This distinction between complements and adjuncts combining with a head is essential in almost all current formal theories of grammar (e.g. the Minimalist Program (Chomsky 1995), Lexical-Functional Grammar (Bresnan 2001), Head-Driven Phrase Structure Grammar (Pollard and Sag 1994), Categorial Grammar (Morrill 1994), and Tree-Adjoining Grammar (Joshi and Schabes 1997)). According to these frameworks, complements are taken to be syntactically specified and required by the head, whereas adjuncts (of time, place, purpose, etc.) can freely modify a head, subject only to semantic compatibility constraints.

According to this categorical distinction, constituents have to be either selected (as complements) or not. If they are not, they are freely licensed as adjuncts, which in theory should be able to appear with any head and to be iterated any number of times, subject only to semantic compatibility.

However, categorical models of selection have always been problematic (Manning 2003). Many subcategorization distinctions presented in the linguistics literature as categorical are actually counter-exemplified in studies of large corpora of written language use.

According to Pollard and Sag (1994: 105-108, P & S, hereafter), a verb such as *consider* appears with a noun phrase object followed by various kinds of predicative complements (nouns, adjective, clauses, etc.), but not with *as* complements:

(10) a. We consider Kim to be an acceptable candidate.

b. We consider Kim an acceptable candidate.

c. We consider Kim quite acceptable.

d. We consider Kim among the most acceptable candidates.

e. *We consider Kim as an acceptable candidate.

f. *We consider Kim as quite acceptable.

g. *We consider Kim as among the most acceptable candidates.

h. ?*We consider Kim as being among the most acceptable candidates.

However, this lack of *as* complements is counter-exemplified by examples from the *New York Times* corpus in the Linguistic Data Consortium (cited in Manning 2003: 299):

(11) a. The boys consider her as family and she participates in everything we do.

b. Greenspan said, “I don’t consider it as something that gives me great concern.”

c. “We consider that as part of the job,” Keep said.

d. Although the Raiders missed the playoffs for the second time in the past three seasons, he said he considers them as having championship potential.

e. Culturally, the Croats consider themselves as belonging to the “civilized” West…

Moreover, according to P & S, *regard* is the opposite of *consider* in disallowing VP complements, but allowing *as* complements:

(12) a. *We regard Kim to be an acceptable candidate.

b. We regard Kim as an acceptable candidate.

But again there are examples in the *New York Times* where *regard* appears with an infinitival VP complement:

(13) a. As 70 and 80 percent of the cost of blood tests, like prescriptions, is paid for by the state, neither physicians nor patients regard expense to be a consideration.

b. Conservatives argue that the Bible regards homosexuality to be a sin.

P & S describe *turn out* as allowing an adjectival phrase complement but not a present participle VP complement:

(14) a. Kim turned out political.

b. *Kim turned out doing all the work.

But again counter-examples were attested in the *New York Times*:

(15) a. But it turned out having a greater impact than any of us dreamed.

This conflict between the linguist's judgments and the corpus evidence implies that the corpus contains structures that should not, according to the categorical model, be generated by the grammar in the first place.[6] It could be argued that the above counter-examples may be due to possible errors or regional or social variation. However, they should be accounted for by any theory of lexical knowledge, in particular, and language, in general. One possible solution in a categorical model is to expand the model to encompass the new examples. However, by doing that, we are eventually

> "failing to capture the fact that the subcategorization frames that Pollard and Sag do not recognize are extremely rare, whereas the ones they give encompass the common subcategorization frames of the verbs in question", (Manning 2003: 301).

Accordingly, a parser based on P & S's subcategorization information is expected to fail in outputting the correct parses for all of the counter-examples given above.

What is needed then is a more relaxed model that takes into consideration the possible frames permitted by a given verb as well as their probabilities. Furnishing the parser with such information will definitely increase its efficiency and robustness.

According to such a probabilistic model, (Manning 1993, 2003), it is not necessary to categorically divide verbal dependents into subcategorized arguments and freely occurring adjuncts. Rather, subcategorization information is represented as "a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability" (Manning, 2003: 302). The following example demonstrates the dynamics of this model.

[6] Of course, we should assume that these structures are grammatical to start with.

In the *Oxford Advanced Learners Dictionary* (Hornby, 1989), the verb *retire* is subcategorized as a simple intransitive and transitive verb, and as an intransitive verb taking a PP[*from*] or PP[*to*] argument. The following examples from the *Wall Street Journal* corpus in the Linguistic Data Consortium (cited in Manning 2003: 303) show the different contexts where *retire* occurs:

(16) a. Mr. Riley plans to retire to the $1.5 million ranch he is building in Cody, Wyo.

b. Mr. Frey, 64 years old, remains chairman but plans to retire from that post in May.

c. To all those wishing to retire to Mexico, let me offer three suggestions.

d. Donald W. Tanselle, 62 years old, will retire as vice chairman of this banking concern, effective Jan. 31.

e. A worker contributing 10% of his earnings to an investment fund for 40 years will be able to retire on a pension equal to two thirds of his salary.

While prepositional phrases headed by *to* or *from* are common with *retire* (16a-b)—and are arguably arguments by traditional criteria—this does not exhaust the list of putative arguments of *retire*. While *in* most often occurs with *retire* to specify a time point (a canonical adjunct PP), it sometimes expresses a destination (16c), and it seems that these examples demand the same treatment as the examples with PP[*to*]. The same applies to the other examples in (16d-e).

In a model that assumes a categorical distinction between complements and adjuncts, it is not clear how these cases can be accounted for. In a probabilistic model, this subcategorization information is instead represented as a probability distribution over argument frames, with different verbal dependents expected to occur with a verb with a certain probability. Table 4 shows the probabilities of part of the different subcategorization frames for the verb *retire* in the *Wall Street Journal* corpus (Manning, 2003: 303).[7]

Frame	**Probability**
*P(*NP[SUBJ]____|V = *retire*)	0.25
*P(*NP[SUBJ]____NP[OBJ]|V = *retire*)	0.5
*P(*NP[SUBJ]____PP[*from*]|V = *retire*)	0.04
*P(*NP[SUBJ]____PP[*from*]PP[*after*]|V = *retire*)	0.003
…	…

Table 4: Partial frame probabilities for *retire* (Manning 2003)

These probabilities constitute part of the lexical properties of the verb, and can be exploited in parsing or any other task determined by the modeler. For example, Fodor (1978) and Connine et al. (1984) provide evidence that listeners use the probability with which a given verb appears in its various possible frames to guide sentence analysis. Moreover, MacDonald (1994) and Jurafsky (2003) have shown that the probability of subcategorization frames plays an on-line role in the disambiguation of various syntactic ambiguities. This point is discussed in more detail in Chapter 4.

This difference between the two models of subcategorization frames indicates a more fundamental difference between two approaches to language in general. The first approach is nativist and consequently emphasizes the role of innate linguistic knowledge,

[7] Manning considers the probabilities of the complete frame, i.e., including the subject argument.

with the influence of the learner's environment playing a relatively minor role (e.g., Chomsky, 1965, 1981, 1986; Lightfoot, 1991). The other approach, on the other hand, is empirical and accentuates the role of the input in learning. It explores the utility of important classes of language-internal, or distributional information, derived from the relationships between linguistic units such as phonemes, morphemes, words, and phrases (e.g., Harris, 1951). This approach entertains distributional or probabilistic mechanisms, including connectionist networks and conventional statistics, in the discovery of structure in the input.

Regardless of the strongly-held fundamental tenets of each approach, indisputably some aspects of language must be learned. The first and most central of these aspects is *vocabulary*. No matter how great the contribution of innate knowledge to language acquisition, a nativist theory of language is required to show how this knowledge interacts with the input in vocabulary construction. Moreover, a principles-and-parameters interpretation of the nativist approach is required to show how the putatively innate parametric knowledge is tuned (e.g., by parameter setting) to the specific properties of the language to be learned.

A distributional approach, on the other hand, is required to provide mechanisms for learning language from the input, based *only* on language-internal regularities. This dissertation is an attempt in this direction.

Two extreme views concerning the utility of distributional methods have been repeatedly asserted. The first is that distributional methods can learn all of the language. The other is that distributional methods can provide no useful information about any aspect of language. A more moderate view is that distributional methods are valuable in a number of domains, such as word segmentation, morphology, categorization, and lexical semantics. However, other aspects of language (e.g., syntax and compositional semantics) which exhibit highly complex and structured regularities are intractable to any learning method, including distributional methods, and hence require the existence of symbolic linguistic knowledge (Chomsky, 1965). Nevertheless, the view defended by this dissertation is that

> "…the success of distributional methods in the limited aspects of language so far attacked does show that empirical research may produce better results than may be expected from considerations of linguistic theory." (Redington and Chater, 1997: 2)

Consequently, pushing distributional methods as far as possible is an important enterprise, which is likely to illuminate both the value of distributional information and the role of the input in learning.

For the purposes of this dissertation, this issue is discussed in terms of the issues involved in learning subcategorization frames from text corpora. In the following sections, I review some of the attempts at learning these frames through distributional methods.

2.2 Previous Attempts

Several methods have been suggested for learning subcategorization frames automatically from text corpora for English (e.g., Brent 1993; Manning 1993; Ushioda et al. 1996; Briscoe & Carroll 1997; Korhonen 1997; and Buchholz 1998). No more recent methods have been reported for automatic frame identification that present substantial developments of the methods discussed here. Other research that has been done on frame identification in languages other than English is mainly based on one or more of these methods.[8] Consequently, the methods discussed here are assumed to cover the spectrum of techniques available in this area, so far.

These methods implement different learning techniques, yet they share two important features. The first is that they are mainly corpus-driven and consequently focus on the distributional regularities in the input as the main source of linguistic information. The other feature is that they adopt a probabilistic model of subcategorization frames.

2.2.1 Brent

Brent (1991, 1993, and 1994) implements a learning strategy that uses approximate cues in the form of function morphemes – prepositions, determiners, inflection, pronouns, auxiliary verbs, and complementizers – proper names, and punctuation, to determine syntactic structure that is necessary for frame acquisition from pure words.[9] Using function morphemes as starting points for learning lexical syntax is motivated by the fact that these elements share properties that make them salient in the overall segmental and suprasegmental character of the language (Jakobson & Waugh 1987; Gerken 1996; Morgan et al. 1996). Compared to content/lexical words, these words are typically the shortest, most common, most syntactically informative words in a language, and tend to occur at the beginnings and ends of phrases, thus might serve to cue

[8] For frame identification in other languages, see Zeman and Sarkar (2000) for Czech; Maragoudakis et al. (2000) for Modern Greek; Basili et al. (1997) for Italian; Eckle et al. (1996) for German; and Kawahara et al. (2000) for Japanese.

[9] Like parameter-setting models (Chomsky 1981, 1986; Lightfoot 1991), Brent's approach assumes a fixed, finite menu of subcategorization frames from which a lexical entry is selected for each verb. However, Brent is not concerned with whether this menu is innate or acquired – his only concern is that knowledge of the menu is independent of the mechanisms the learning algorithm uses to select from it. It is argued in this dissertation is that there is no predefined set of frames and that they should be learned from the input.

phrase boundaries (Greenberg 1963; Kimball 1973; Clark & Clark 1977; Carter & Gerken 1997; Shady & Gerken 1999). Such cues make it possible to discover relevant syntactic structure in an utterance without already knowing all the words in it.

Brent's approach is based on the following general principles:

(i) Do not try to parse sentences completely. Instead, rely on local morpho-syntactic cues such as the facts about English: (1) the word following a determiner is unlikely to be functioning as a verb; (2) the sequence *that the* typically indicates the beginning of a clause.

(ii) Do not try to draw categorical conclusions about a word on the basis of one or a fixed number of examples. Instead, attempt to determine the distribution of exceptions to the expected correspondences between cues and syntactic frames. Use a statistical model to determine whether the occurrence of a verb with cues for a frame is too regular to be explained by randomly distributed exceptions.

Brent uses as input the untagged Brown Corpus. The syntactic frames targeted by the algorithm are shown in Table 5 (Brent, 1993). Tables 6 and 7 show the cues used for identifying these frames. Table 6 defines lexical categories used in Table 7. Using the lexical categories in Table 6, Brent builds a set of cues for identifying argument phrases. The phrase types for which data are reported here are noun phrases, infinitive verb phrases, and tensed clauses. These phrase types yield three syntactic frames with a single argument and three with two arguments, as shown in Table 5.

Table 5
The six syntactic frames studied by Brent

SF Description	Good Example	Bad Example
NP only	greet them	*arrive them
Tensed clause	hope he’ll attend	*want he’ll attend
Infinitive	hope to attend	*greet to attend
NP & clause	tell him he’s a fool	*yell him he’s a fool
NP & infinitive	want him to attend	*hope him to attend
NP & NP	tell him the story	*shout him the story

Table 6
Lexical categories used in the definitions of the cues

SUBJ :	I | he | she | we | they
OBJ :	me | him | us | them
SUBJ_OBJ :	you | it | yours | hers | ours | theirs
DET :	a | an | the | her | his | its | my | our | their | your | this | that | whose
+TNS :	has | have | had | am | is | are | was | were | do | does | did | can | could | may | might | will | would
CC :	when | before | after | as | while | if
PUNC :	. | ? | ! | , | ; | :

Table 7
Cues for syntactic frames: The category V in Table 3 starts out empty and is filled as verbs are detected on the first pass. “cap” stands for any capitalized word and “cap+” for any sequence of capitalized words

Frame	Symbol	Cues
NP only	NP	(OBJ | SUBJ_OBJ | cap) (PUNC | CC)
Tensed Clause	cl	(that (DET | SUBJ | SUBJ_OBJ | cap+)) | SUBJ | (SUBJ_OBJ +TNS)
Infinitive VP	inf	to V
NP & clause	NP cl	(OBJ | SUBJ_OBJ | cap+) cl
NP & infinitive	NP inf	(OBJ | SUBJ_OBJ | cap+) inf
NP & NP (dative)	NP NP	(OBJ | SUBJ_OBJ | cap+) NP

The cues used by the algorithm must address two problems, i.e., finding verbs in the input and identifying phrases that represent arguments to the verb. The algorithm identifies verbs in two stages, each carried out on a separate pass through the corpus. First, strings that sometimes occur as verbs are identified. Second, occurrences of those strings in context are judged as likely or unlikely to be verbal occurrences. The second step is necessary because of lexical ambiguity.

The first stage uses the fact that all English verbs can occur either with or without the suffix *–ing*. Words are taken as potential verbs if and only if they display this alternation in the corpus. There are few words that meet this criterion but do not occur as verbs, including *income/incoming*, *ear/earring*, *her/herring,* and *middle/middling*. However, the second stage of verb detection, combined with the statistical criteria, prevent these pairs from introducing errors. The algorithm assumes that a potential verb is functioning as a verb unless that context suggests otherwise. In particular, an occurrence of a potential verb is taken as a non-verbal occurrence only if it follows a determiner or a preposition other than *to*. For example, *was talking* would be taken as a verb, but *a talk* would not.

When a putative occurrence of a verb is found, the next step is to identify the syntactic types of nearby phrases and determine whether or not they are likely to be arguments of the verb. Brent’s strategy for determining whether a phrase P is an argument of a verb V has two components:

1. If P is a noun phrase (NP), take it as an argument only if there is evidence that it is not the subject of another clause.

2. Regardless of P’s category, take it as an argument only if it occurs to the right of V and there are not potential attachment points for P between V and P.

For example, suppose that the sequence *that the* were identified as the left boundary of a clause in the sentence *I went to* ***tell him that the*** *idea won’t fly*. Because pronouns like *him* almost never take relative clauses, and because pronouns are known at the outset, the algorithm concludes that the clause beginning with *that the* is probably an argument of the verb *tell*. It is always possible that it could be an argument of the previous verb *want*, but the algorithm treats this as unlikely. On the other hand, if the sentence were *I want to* ***tell the boss that the*** *idea won’t fly,* then the algorithm cannot determine whether the clause beginning with *that the* is an argument to *tell* or is instead to *boss*, as in *I want to fire the boss* **that the** *workers don’t trust.*

An experimental evaluation shows that Brent’s method does well as far as precision is concerned. For most subcategorization frames, close to 100% of the verbs assigned to a particular frame are correctly assigned (Brent, 1993: 255). However, this method does less well at recall. For the six frames covered by Brent (1993), recall ranges from 47% to 100% but these numbers would probably be appreciably lower if a random sample of verb types had been selected instead of a random sample of verb tokens, a sampling method that results in a small proportion of low frequency verbs (Manning and Schütze, 2003: 275). Since low frequency verbs are least likely to be comprehensively covered in existing dictionaries, they are arguably more important to get right than high-frequency verbs.

Brent attributes the errors in frame detection to two main reasons. The first is that the cues are fairly rare, so verbs that occur fewer than 15 times tend not to occur with these cues at all. The other reason is that these cues occur fairly often in structures other than those they are designed to detect. For example, the words *record, recover,* and *refer* all occurs in the corpus with cues for an infinitive, although none of them in fact takes an infinitive argument.

Moreover, there is a ceiling on the number and nature of frames that can be identified using Brent's cue-based approach. The reason for this inextensibility is that this approach has depended upon finding cues that are a very accurate predictor for a certain subcategorization. However, for many frames there are no highly accurate cues. For example, some verbs subcategorize for the preposition *in*, e.g., (17), and the majority of occurrences of *in* after a verb are NP modifiers or locative adjuncts, e.g., (18). There is no high accuracy cue for verbs that subcategorize for *in* (Manning 1993: 3).

(17) a. Two women are *assisting* the police *in* their investigation.
b. We *chipped in* to buy her a new TV.
c. His letter was *couched in* conciliatory terms.

(18) a. He gauged support for a change in the party leadership.
b. He built a ranch in a new suburb.
c. We were traveling along in a noisy helicopter.

2.2.2 Manning

Manning (1993) suggested that the solution to this problem is to collect as much co-occurrence statistics as possible from the text corpus, and then use statistical filtering (e.g., significance test, a mutual information measure, or any other form of statistic) to weed out false cues. He proposed a method for producing a dictionary of syntactic frames from unlabeled text corpora. Kupiec's stochastic part-of-speech tagger was used to tag approximately 4 million words of the *New York Times* newswire. Frame learning was then performed by a program that processed the output of the tagger. This program had two parts: a finite state parser ran through the text, parsing auxiliary sequences noting whether a verb is active or passive, and then it parsed complements following the verb until something recognized as a terminator of subcategorized arguments is reached.[10] Whatever has been found is entered in the histogram. A second process of statistical filtering then took the raw histograms and decided the best guess for what frames each observed verb actually had. The parser does not learn from participles since an NP after them may be subject rather than the object (e.g., *the yawning man*).

The program acquired a dictionary of 4900 frames for 3104 verbs (an average of 1.6 per verb). In general, all the verbs for which frames were determined are in Webster's (Gove 1977), the only noticed exceptions being certain instances of prefixing, such as *overcook* and *repurchase*, as well as verbs such as *fax*, *sensationalize*, and *solemnize*. All in all, the system achieved a token recall of 82%.

2.2.3 Ushioda et al.

Ushioda et al. (1996) also make use of a PoS tagged corpus and finite-state NP parser to recognize and calculate the relative frequency of the same six syntactic frames Brent used. A tagged corpus is first partially parsed to identify NPs and then a regular grammar is used to estimate the appropriate syntactic frame for each verb token in the corpus. Their procedure to automatically find subcategorization frame frequencies is given in Figure 1 (Ushioda et al., 1996: 243).

In an experiment involving the identification of these frames, the system showed an accuracy rate of 83%. The most frequent source of errors in frame identification by

[10] Manning (1993) used a period and subordinating conjunctions as frame terminators.

this system was errors in NP boundary detection. The second most frequent source was misidentification of infinitival purpose clauses, as in *He used a crowbar to open the door*. The phrase *to open the door* is a purpose adjunct modifying either the verb phrase *used a crowbar* or the main clause *he used a crowbar*. But the system incorrectly judged such adjuncts to be complements of their main verbs. The last source of error was caused by verbs that are frequently used in relative clauses without relative pronouns, such as the verb *need*, as in *the last thing they need*. Ushioda et al.'s system was not able to capture this kind of relative clause; consequently, each occurrence of these relative clauses caused an error in measurement.

```
Make a list of verbs out of the tagged corpus.
For each verb on the list (the 'target verb'),
        Tokenize each sentence containing the target verb in the following way: All
        the noun phrases except pronouns are tokenized as "n" by a noun phrase
        parser and all the rest of the words are also tokenized as follows:
                                b: sentence initial marker    e: sentence final marker
                                k: target verb                t: to
                                i: pronoun                    m: modal
                                n: noun phrase                w: relative pronoun
                                v: finite verb                a: adverb
                                u: participial verb           x: punctuation
                                d: base form verb             c: complementizer that
                                p: preposition                s: the rest
        Apply a set of frame extraction rules to the tokenized sentences. These rules
        are written as regular expressions as follows:
                                Frame          Rule
                                NP + NP        k(i|n)n
                                NP + CL        k(i|n(pn)*)c
                                               k(i|n)(i|n)a*(m|v)
                                NP + INF       k(i|n(pn)*)ta*d
                                CL             kc
                                               k(i|n)a*(m|v)
                                NP             k(i|n)/[^mvd]
                                               #pw(i|n(pn)*)a*m?a*k/[^t]
                                INF            kta*d
```

Figure 1: Ushioda et al.'s Frame Identification Procedure

2.2.4 Briscoe & Carroll

Briscoe & Carroll (1997), B & C henceforth, proposed a system for distinguishing 160 verbal frame classes and their relative frequency in English. B & C's system consists of six components, which are applied in sequence to sentences containing a specific predicate in order to retrieve a set of frame classes for that predicate (B & C, 1997: 2):

1. A tagger, a first-order HMM part-of-speech and punctuation tag disambiguator, is used to assign and rank tags for each word and punctuation token in a sequence of sentences;
2. A lemmatizer is used to replace word-tag pairs with lemma-tag pairs;
3. A probabilistic LR tagger, trained on a treebank, returns ranked analyses;
4. A patternset extractor which extracts frame patterns, including the syntactic categories and head lemmas of constituents from sentence subanalyses which begin/end at the end of specified predicates;
5. A pattern classifier which assigns patterns in patternsets to frame classes or rejects patterns as unclassifiable on the basis of the feature values of syntactic categories and the head lemmas in each pattern;
6. A patternset evaluator which evaluates sets of patternsets gathered for a (single) predicate constructing putative frame entries and filtering the latter on the basis of their reliability and likelihood.

To test the performance of their system, B & C took the Susanne Corpus (Taylor & Knowles 1988) and LOB corpora (Garside et al. 1987). The system achieved a token recall of 80.9%, which is comparable to previous approaches. B & C attributed most of the errors to the filtering phase, which they describe as the 'weak link' in the system.[11]

[11] An extension to B & C's system was suggested by Korhonen (1997). Korhonen's method is not discussed in detail here since it is mainly a method for improving the filter component in B & C's, and consequently does not present a new perspective on frame extraction. The basic improvement method was to guide the statistical filter with a more knowledge-based component seeded with general linguistic information. Korhonen made use of *frame alternations*: alternate ways in which verbs can express their

2.2.5 Buchholz

Buchholz (1998) presented an unsupervised learning method for subcategorization acquisition that takes care of what he sees as a shortcoming in the previous methods; that is, they are knowledge-based and thus require either existing tools (e.g., a wide-coverage parser in the case of B & C) or an important amount of time and linguistic expertise to write the necessary patterns, regular expressions, finite-state NP parsers etc (e.g., Brent, Manning, and Ushioda).

In contrast, Buchholz's method only requires PoS-tagged text as input. This method is based on the assumption that subcategorized constituents differ from non-subcategorized ones in terms of frequency (Meyers et al. 1994). This means that the subcategorization property of a verb should somehow show up when enough sentences containing this verb are collected. The idea of unsupervised learning then is to model the global behavior of each verb, and group verbs that behave syntactically similar. These groupings should then ideally correspond to groups of verbs with similar subcategorization properties. The information about the group membership of a verb could therefore be used by a parser when making local decisions, e.g., about the complement- or adjuncthood of a constituent.

The global subcategorization behavior of a verb is extracted using hierarchical clustering (Schütze, 1994; Zavrel and Veenstra, 1995): each word *w* is represented by a high-dimensional vector, and each component *i* of the vector shows the times that another word x_i appeared inside a fixed window around *w* in the text. Buchholz defined the

arguments. The method Korhonen used to determine alternations was based on finding correlations between two patterns of complementation.

window of a verb *v* as stretching only to the right and being delimited by the next verb or the boundary of the sentence.

Following Brent (1993), Buchholz did not count all the words that occur in a window around a verb, he only counted the capitalized words, numbers and closed-class words, and only those that occur at least four times in the whole corpus. After a rough lemmatization of verbs in the corpus, each of the 3531 verb lemmas in the corpus had a 151-dimensional vector. These vectors were then hierarchically clustered, producing a cluster tree that is visualized in Figure 2 (Buchholz, 1998: 5).

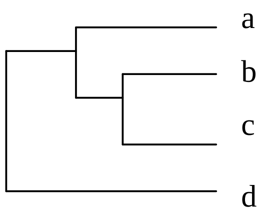


Figure 2: Visualization of a hierarchical cluster of four vectors

In theory, verbs with similar subcategorization behavior should be close in the tree and verbs with different subcategorization behavior should be at a long distance from each other. In practice, however, it is hard to verify this claim by just looking at the tree.

Accordingly, this binary tree was converted into a propositional i.e., feature-value format in the following manner. Each node in the cluster tree was assigned a unique number. Consequently, each vector is characterized by the sequence of numbers through the path from the top node to the vector leaf. These sequences were then used in a memory-based learner[12] to evaluate the influence of the learned subcategorization clusters on the accuracy of learning the distinction between complements and adjuncts. The performance of the learner was tested on data extracted from the *Wall Street Journal*

[12] See Daelemans et al. (1998) for a discussion and implementation of memory-based learning in NLP.

Corpus in the Penn Tree Bank (Bies et al. 1995). The corpus is annotated with PoS tags and parse trees. Buchholz reported a 93.1% accuracy of the memory-based learner on the complement-adjunct distinction.[13]

3. Conclusion

This chapter discussed some theoretical and practical issues in subcategorization frames learning. It started with a discussion of the differences between the categorical and probabilistic approaches to frames. Then it reviewed the previous methods for frame identification, and discussed some of the theoretical and practical issues involved. The mechanisms used in these methods have raised fundamental theoretical and practical issues in language acquisition. They could thus a source of hypotheses for experimental test. Seen in this wider perspective, these methods could be more empirically and theoretically rigorous if they are based on more formal and objective foundations. These issues are addressed in the next chapter in terms of the notion of *bootstrapping*.

[13] No details were given in Buchholz (1998) on how the performance of the learner was evaluated.

Chapter 3

Bootstrapping

This chapter discusses in terms of bootstrapping some issues in the methods for frame learning described above and the assumptions they make. It introduces the concept of bootstrapping, discusses the bootstrapping issues in these methods, and then concludes with the different approaches to bootstrapping.

3.1 Bootstrapping

In general terms, bootstrapping is used here to mean the process of attaining new knowledge on the basis of already existing knowledge. In the context of language acquisition, bootstrapping implies that the learner, on the basis of already existing knowledge and information processing capacities, can make use of specific types of information in the linguistic and non-linguistic input in order to determine the language particular regularities which constitute the grammar and the lexicon of the target language (Weissenborn and Höhle 2001). The central assumption behind the bootstrapping approach is that there is a systematic relationship between properties of the input at one level of representation (i.e. *source domain*), which the learner has access to, and another level of representation (i.e., *target domain)*. In other words, the learner makes use of the regularities that characterize the interface, i.e., the interaction between different linguistic and non-linguistic domains of representation.

Depending on the type of information which the learner makes use of, four main bootstrapping approaches can be distinguished: semantic, syntactic, prosodic, and distributional.[14] These approaches will be discussed in more detail below. A problem with the inter-domain approaches is that there is only a partial, i.e., a non-perfect correlation between the source and the target domains (e.g., Selkirk 1984; Jackendoff 1997). In order to overcome the difficulties resulting from this type of discrepancy, it has been proposed that the learner makes use of different types of information and a correlated set of input cues to bootstrap into a target language (see for example, Hirsh-Pasek & Golinkoff 1996a; Mattys, Jusczyk, Luce, & Morgan 1999; Morgan, Shi, & Allopenna 1996; Christiansen et al. 1998). According to this integrated approach to bootstrapping, the problem of acquisition is easier if multiple cues are taken into account. This suggests that the learner may aim to exploit as many sources of information as possible in order to narrow down the hypothesis space. However, one possible problem with this approach is that recognizing and using correlated sets of cues would require more sophisticated abilities than does use of individual cues (Morgan et al. 1996). In addition to these *inter-domain* and integrated bootstrapping approaches, Cartwright & Brent (1997) introduced the notion of *autonomous* bootstrapping, which applies within a single domain.

Other questions related to the process of bootstrapping are whether and how the bootstrapping strategies and their interrelation may change during development. Such a change is to be expected given the constantly increasing knowledge of the child in the linguistic and non-linguistic domain. For example, the growing lexicon of the child, especially in the domain of the closed class, functional vocabulary which in languages like English, French or German constitute about 50% of the lexical tokens of any given

[14] Syntactic and distributional bootstrapping approaches are traditionally grouped under syntactic bootstrapping. To avoid confusion, *syntactic bootstrapping* will be used to refer to the classical approach, and *distributional bootstrapping* to refer to bootstrapping based on the revived distributional analysis.

text, should considerably facilitate and enhance the lexical (e.g., word segmentation and categorization) and the syntactic (e.g., determination of syntactic boundaries) bootstrapping capacities of the child because of the distributional properties of these items (Weissenborn and Höhle 2001; Clark and Clark 1977; Kimball 1973).

In addition to the dependency on the perceptual and representational capacities of the child in the different linguistic and non-linguistic domains, success of bootstrapping strategies will depend on the availability of information processing capacities like memory and attention which are necessary to integrate the information extracted from the input into the learning mechanisms. Thus, linguistic knowledge acquisition on the basis of distributional learning puts particular demands on memory because of the necessity to keep track of the relevant co-occurrence relations. The existence of such frequency effects in prelinguistic learners points to the importance of memory processes (e.g., Jusczyk et al., 1994).

In order to understand the acquisition process, it is crucial to ask to which extent (and how) the learner uses the information accessed in the learner's rule learning mechanisms (via bootstrapping mechanisms). The fact that the learner is sensitive to a certain property of the input which may be relevant from a theoretical perspective for the acquisition of a particular aspect of linguistic knowledge does not yet mean that the learner actually uses this information to acquire this knowledge (Weissenborn and Höhle 2001).

With this brief introduction about bootstrapping in mind, I discuss below the bootstrapping issues in the previous methods for subcategorization frame learning.

3.2 Bootstrapping Issues in Previous Methods

In the case of Brent's method, cues can be easily identified in the input, for the reasons given in the previous chapter, and frames can be learned based on the information provided by these cues. Verb identification is based on a simple morphological cue in English, i.e., verbs usually occur with or without *_ing*. In the other methods, it is implicitly assumed that frame acquisition is not feasible without PoS and partial parse knowledge. A logical justification for this assumption may be built on compositionality if we interpret a frame as a structural whole whose parts are the PoS tags and NP arguments, and therefore the acquisition of the parts should occur before the acquisition of the whole. Though Brent's initial knowledge seems more plausible than its counterpart in other algorithms, all these mechanisms still suffer from a major learnability paradox (Brent 1994: 435):

> "…young children must infer some of the *lexical knowledge* of their languages – the syntactic facts about individual words – from larger syntactic structures. But it is difficult to see how children could identify syntactic structures in an utterance without already knowing the syntactic functions of some of the words in the utterance. This poses an apparent paradox: *to learn lexical syntax, children must recover the syntactic structure of the input; to recover syntactic structure, they must know lexical syntax – the bootstraps need bootstraps* (Emphasis added)."

Brent assumes that children first learn the syntactic properties of function words that are extremely common and highly informative about syntactic structure. However, these syntactic functions cannot be learned without previous syntactic knowledge that needs function words to be learned in the first place. Verb identification is also similarly circular. For the *_ing* rule to apply, the learner should have the knowledge that this suffix is licensed only with verbs. The PoS and NP knowledge assumed necessary by other methods for frame acquisition may implicitly need this frame information to be extracted. In other words, these methods used contextual information in order to identify verbs as well as other categories and phrases. Given that a verb's frame constitutes a major portion of the context where a verb occurs, we are then faced with a bootstrapping paradox, i.e., for X to be identified, it needs information about Y, yet for Y to be identified, it needs information about X.

Another bootstrapping-related issue in these methods is what can be termed the *Sequentiality Problem*. That is, frame acquisition does not take place until function words, in Brent's, and PoS and NPs, in other systems, have been acquired. This has two implications. The first is that subcategorization occurs only when categorization has been fully completed, and the other is that subcategorization information does not play any role in identifying syntactic categories and NPs in the input. Language acquisition research has shown that this is not the case.

According to Nelson's (1973: 37) study of 18 children with an age range from 16 to 28 months, it was shown that children at this age range were able to produce multiword utterances that exhibit relatively little structural variation. Examples of these utterances are given in Table 8 below, where first names and two-figure numbers indicate the name and age – in months – of the child who produced the utterance concerned (Bloom 1970, 1991; Braine 1976; Radford 1990, among many others). Despite the different explanations of their nature, what is clear from these utterances is that they constitute rudimentary realizations of the subcategorization frames of the verbs involved (For possible syntactic and semantic explanations of these utterances, see O’Grady 1997:34-54).

Utterance	**Child**	**Age**
Miller try	Susan	24
See cow	Eve	25
Doggy bit	Adam	28
Kathryn no like celery	Kathryn	22
Wayne taken bubble	Daniel	21
Hayley draw boat	Hayley	20
Baby ride truck	Allison	22
Want mommy come	Jew	21
Daddy walk	Jonathan	24

Table 8: Examples of early multiword utterances (Nelson 1973: 37)

For example, verbs such as *try, bite, come,* and *walk* can substantiate an intransitive frame, whereas verbs like *see, like, take, draw, want,* and *ride* can realize a transitive frame. At this stage, the child has a vocabulary of about 400 words, 46.8% of which are nouns compared to 8.3% verbs (Bates, Bretherton, and Snyder (1988:153). This means that the child shows sensitivity to subcategorization frames with the earliest emergence of “just enough” information of the categories constituting rudimentary frames.

To summarize, in this section I have discussed some bootstrapping issues in the previous methods for frame learning. The objective of this discussion was to stress the importance of handling these issues in order to strengthen the theoretical and practical value of the learning mechanisms used in these methods. It is argued throughout this dissertation that a learning mechanism that takes into consideration the issue of bootstrapping is more likely to provide a more efficient learning mechanism both theoretically and practically. Accordingly, the following section provides a detailed discussion of the different approaches to bootstrapping and the issues they raise.

3.3 Approaches to Bootstrapping

Theoretical accounts of language acquisition have emphasized the role of innate knowledge, with the influence of the input playing a relatively minor role (e.g., Chomsky 1965). Even if we assume an essential role of some innate knowledge, the quantitative relation between this knowledge and the input aside for now, we still have to explain how this knowledge interacts with the input to bootstrap into the target language. For example, if the space of innate knowledge includes statements such as

> “there exist nouns, verbs, etc...”, such linguistic entities are not marked in the linguistic input to the learner in any way, and the learner must have some way of detecting these elements in the input and mapping them onto the appropriate categories.” (Pinker 1984:38)

This ultimately means that no matter how great the qualitative and/or quantitative contribution of innate knowledge to language acquisition, some input-driven aspects of language (e.g., vocabulary) must be learned, and that putatively innate knowledge must be tuned (e.g., by parameter-setting) to the specific properties of the language to be learned. In other words, what is needed is a set of learning mechanisms that transform the initial state of the learner into a target state/language depending on the information coming from the input. A number of mechanisms have been proposed to handle this transformation, under the umbrella concept of *bootstrapping*, which are reviewed and assessed below.

3.3.1 Semantic Bootstrapping

Under this approach, syntactic entities are *canonical structural realizations* of semantic entities in a word-onto-world fashion (Grimshaw 1981; Macnamara 1982; and Pinker 1984, 1994). The main argument in this approach is that although grammatical entities like noun or verb do not have semantic definitions, nouns and verbs typically refer to distinct, identifiable semantic classes in the input. That is, people and physical objects are referred to with nouns; activities and changes of state with verbs; properties and colors with adjectives, and so on (Croft 1991). Notions like physical objects, agent, and action are therefore available to the learner in the input. Accordingly, the learner starts out with some basic constraints on word learning: there are objects, properties and events which function as the inductive bases; objects map to nouns, properties map to adjectives and events map to verbs. Once the learner has this basic scaffolding of semantically induced information about word syntactic categories and lexical items belonging to these classes, the learner is in a position to acquire the syntactic rules based on these categories.

In addition, propositions with action predicates involving the semantic relations agent-of-action and patient-of-action may be expressed using the grammatical relations SUBJECT and OBJECT.[15] Presumably, such notions as physical objects, physical action, agent-of-action, and so on, unlike nounhood, verbhood, and subjecthood, are available to the learner perceptually and are elements of the semantic representation input to the language acquisition mechanisms (Pinker 1984: 39). If the learner tentatively assumes these syntax-semantics correspondences to hold, it is possible to make the correct inferences. The categorization of words can be inferred from their semantic properties,

[15] For a hypothetically complete list of possible mappings, see Pinker 1984: 41.

and their grammatical relations can be inferred from the semantic relation in the event witnessed.

However, the word-world isomorphism implied by this approach does not hold in many cases. As Quine (1960) has noted there is an infinite set of meanings compatible with any situation, so the learner has an infinite number of perceptually indistinguishable hypotheses about meaning to choose among. For example, all the situations in which a *rabbit* is present are also situations in which an animal is present, an object is present, a set of un-detached rabbit parts are present, and so on.

Moreover, some actions are often not linguistically labeled, e.g., the action of someone opening the door is most probably associated with the linguistic production “I’m home”, and not “I’m opening the door”. Likewise, some aspects of verb meanings challenge a view of verb learning based on observation of events alone. Verb meanings do not simply label categories of events but represent the speaker’s choice of perspectives on events (e.g., Bowerman 1985; Clark 1990; Fillmore 1977; Fisher 1994, 1996; Gleitman 1990; Talmy 1985). For example, *feed* and *eat* denote not different world events but different perspectives on the same events. *Give* and *get* also describe the same scenario.

Furthermore, just as speakers of a single language can choose to describe an event from one perspective or another, different languages make different choices among perspective options (e.g., Bowerman 1985, 1990; Choi & Bowerman 1991; Grimshaw 1994; Talmy 1985). That is, laying out semantic bootstrapping in terms of a word-world isomorphism faces a clear induction problem.

One possible solution to this induction problem is to assume that semantic induction is inherently constrained (Pinker 1994: 378): “not all *logically* possible hypotheses are *psychologically* possible.” Instead, the hypotheses that a learner’s word learning mechanisms make available are constrained in two ways.

The first constraint comes from the representational machinery available to build the semantic structures that constitute mental representations of a word’s meaning: a Universal Lexical Semantics, analogous to Chomsky’s Universal Grammar (e.g., Moravcsik 1981; Markman 1989, 1990; Jackendoff 1990). For example, this representational system would allow ‘object with shape X’ and ‘object with function X’ as possible word meanings, but not ‘all the un-detached parts of an object with shape X’, ‘object with shape X or Buick’, and ‘object and the surface it contacts’ (Pinker 1994: 379).

The second constraint comes from the way in which the learner’s entire lexicon may be built up; on how word’s meaning may be related to another word’s meaning (see Miller and Charles 1991, Miller & Fellbaum 1992). For example, the lexicons of the world’s languages freely allow meronyms (words whose meanings stand in a part-whole relationship like *body-arm*) and hyponyms (words that stand in a subset-superset relationship, like *animal-mammal*) but do not easily admit true synonyms (Bolinger 1977, Clark 1987, Miller & Fellbaum 1992). A learner would therefore not posit a particular meaning for a new word if it was identical to some existing word’s meaning.

From an NLP perspective, semantic bootstrapping has another limitation. Though the extra-linguistic information exploited in this approach plays a central role in the acquisition of language, its utility is difficult to evaluate computationally, because the learner’s representation of the environment is unknown—even if the resources to compile corpora relating language and environment were available, it would still be unclear how the environment should be encoded (Redington & Chater, 1997). It is shown in Chapter 6 that, the utility of this extra-linguistic knowledge can be partially evaluated if we limit the set of semantic cues to names of people and things and a subset of pronouns that are easy to identify in the input.

3.3.2 Syntactic Bootstrapping

The unreliability of the word-onto-world mapping motivated the syntactic bootstrapping approach to verb learning, which advocates the possibility of deducing the word meanings from the semantically relevant syntactic structures associated with a verb in input utterances (Gleitman 1990; Landau and Gleitman 1985). According to this approach, even very partial syntactic information is sufficient to give the learner some sentence-structural cues to the interpretation of verbs (Fisher 1996: 43). This approach also assumes that nouns are acquired prior to predicates (from context), which are then used to learn verb meanings.

Any form of syntactic bootstrapping depends on two fundamental assumptions about verb semantics and syntax, supported by a long tradition of work on the organization of the lexicon (e.g., Bloom 1970; Chomsky 1981; Dowty 1991; Fillmore 1977; Fisher 1994; Fisher, Gleitman & Gleitman 1991; Fisher et al. 1994; Grimshaw 1990; Gruber 1965; Jackendoff 1983, 1987, 1990; Landau & Gleitman 1985; Rappaport & Levin 1988). First, a basic part of the meaning of a verb is a semantic predicate-argument structure specifying how many and what types of participants play out the event described by the verb. *Kick*, for example, has two logical arguments – the kicker and the kicked (Fisher 1996: 43). Second, each verb's argument structure is related to the sentence structures in which it can occur. If a verb describes the motion of an object, for example, it must be able to specify that object as a noun phrase in sentences.

These basic assumptions about lexical organization ensure that the sentence in which a novel verb occurs will be related in principled ways to its meaning, as needed for structure-aided verb learning to work.

In this approach, there are two views on how to draw semantic conclusions from sentence structures. Both views depend on the predicate/argument or structural nature of verb semantics, and on principled relations between meaning and syntax, but differ in their assumptions about the degree of syntactic knowledge needed to recover meaning from structure.

The first procedure depends on *reverse linking* whereby the learner could infer aspects of verb meaning from sentences using rules linking thematic roles and syntax. Thematic roles are categories for participants in semantic structures. These categories represent the similarity among the agents or patients of various causal actions, the themes of diverse motions and changes of state, and so on. Rules linking thematic roles with syntactic categories like subject and object have been proposed to capture strong regularities in the assignment of thematic roles to positions in sentences (e.g., Dowty 1991; Fillmore 1977; Grimshaw 1981, 1990; Jackendoff 1987, 1990; Pinker 1984, 1989, 1994). Causal agents, for example, are very likely to be sentence subjects across languages (e.g., Bates & MacWhinney 1982; Dowty 1991; Grimshaw 1981; Keenan 1976). A partial list of the linking rules given by Pinker (1989: 74), following Rappaport & Levin (1988) and Jackendoff (1983, 1987), is shown in (19).

(19) a. agent → subject

b. patient → direct object

c. theme → subject, or if subject is already linked, direct object

However, this procedure is not without problems. In order to use the reverse linking procedure, the learner must parse the sentence, identifying one noun phrase as the subject and another as the object. The subject of the sentence is identified by various means in different languages, including word order, subject-verb agreement, and case marking particles (e.g., Keenan 1976; Croft 1990). Even supposing that learners begin with the linking rules in (19), they will not be able to apply them in reverse, inferring semantic roles from syntactic positions, until they have learned the surface cues that identify subjects and objects in the target languages. This takes us back to the original bootstrapping paradox.

To circumvent this problem, Fisher (1994, 1996) proposed what she called *The Analogical Mapping Procedure*. The main idea behind this procedure is that learners need only structural cues, and not full structures, of some kind to interpret verbs (cf. Brent 1991, 1993, 1994). Accordingly, the learner could obtain semantic information from a partial or presyntactic representation of a sentence consisting simply of the set of recognizable nouns in the sentence. If so, then sentence structures could bootstrap verb interpretation even before the learner can identify the grammatical parts of a sentence (Fisher 1996: 45). In order for this procedure to work, Fisher (1996) capitalizes on the assumption that semantic structures of verbs are fundamentally of the same kind as the nonlinguistic conceptual structures by which events are represented (e.g., Grimshaw 1990; Jackendoff 1983, 1987, 1990; Pinker 1989; Rappaport & Levin 1988). Both verb semantic structures and conceptual representations of events demand a division between predicates and arguments, and thus between relations and the objects they relate. According to this procedure, sentence interpretation takes the form of mapping one structure onto another: A sentence can be represented as a structure relating a set of noun phrases, while the conceptual representation of an event can be viewed as a structure relating a set of event participants. To the extent that these two distinct representations—syntactic and conceptual—have similar structures, a sentence could provide a partial analogy for its interpretation in conceptual terms (e.g., Gentner 1983). Assuming that conceptual and semantic structures are of like kind, the result of this analogical mapping will be, roughly, a semantic structure (Fisher 1996).

For the purposes of this dissertation, this procedure has some interesting consequences for learning. First, this route from structure to meaning can be used without identifying which noun is the subject and which is the object. To begin mapping two-NP verbs onto two-participant conceptual relations by structural analogy, the learner need only have begun to recognize some nouns. If the learner can draw this kind of inference, then simple structural properties of sentences could influence interpretation before much language-specific syntactic knowledge is acquired.

Second, since the analogical mapping procedure makes no mention of the verb as a formal category, it could guide the interpretation of any argument-taking predicate. This includes verbs, prepositions, and predicative adjectives. Like verbs, prepositions and predicative adjectives take NP arguments and encode semantic relations among those arguments. This means that all predicates should therefore be initially interpreted in the same way. Landau & Stecker (1990) present intriguing evidence that young children interpret a novel word as a semantic predicate if it appears with NP arguments. This finding is consistent with the notion that a sentence, partially represented as a structure containing NP arguments, can serve as a quite general analog of its semantic predicate/argument structure. The generality of the analogical mapping procedure is an advantage for the theory of the acquisition of predicate terms: Not all languages have distinct categories of prepositions and predicate adjectives, but may instead use verbs to convey spatial or attribute meanings (e.g., Croft 1990).

Finally, the generality of the analogical mapping procedure yields a third potential benefit for learners: Once some relational terms are acquired in this structure-sensitive way, they could serve, in turn, as second-order cues for the acquisition of new verbs. This provides support to the view adopted by this dissertation that children show sensitivity to frames with the earliest emergence of "just enough" information of the categories constituting rudimentary frames. Since this conclusion has not been tested using a corpus-based method, it is adopted and tested in the cue-based model in Chapter 6 of this dissertation.

3.3.3 Prosodic Bootstrapping

In the prosodic approach to bootstrapping it is generally maintained that there are phonological and prosodic cues in the input that may point the learner to specific linguistic structures, e.g., clauses and phrases or specific classes of words, such as *open* vs. *closed* words, *lexical* vs. *functional* items, or specific grammatical form classes (e.g., Morgan et al., 1996).

For example, a potential prosodic cue for word segmentation is the canonical patterns of strong and weak syllables exhibited in many languages (Cutler & Cutler 1987;

Cutler & Norris 1988; Jusczyk et al. 1993). One possibility is that they can identify function morphemes as a phonological class based on the fact that function morphemes in a particular language usually share several phonological properties (Jakobson & Waugh 1987). For example, English function morphemes typically contain fricative and nasal consonants, are produced with the reduced vowel schwa (which only occurs in unstressed syllable), and are an integral part of the alternating stress pattern of the language (Gerken et al 1990; Gerken 1994a; Morgan et al. 1996; Mattys and Jusczyk 2001). Such properties might permit syntactically naïve learners to assign words to two major categories, which closely correspond to content words and function words, before discovering the distributional regularities of particular morphemes (Gerken 1994a; Morgan et al. 1996).

Moreover, Kelly (1996) and (Durieux & Gillis 2001) proposed several phonological features of a word itself that could be used to predict its syntactic categories. Kelly (1996) has shown that not a single noun-verb homograph exists in which the verb has first syllable stress but the noun has second syllable stress. If the noun and verb versions of a word contrast in stress at all, the noun always has the first stress syllable and the verb has second syllable stress. An examination of thousands of English disyllabic nouns and verbs, which was not restricted to noun-verb homographs has shown that 90% of words with first syllable stress are nouns, whereas 85% of words with second syllable stress are verbs (Kelly & Bock 1988; Kelly 1992). Durieux & Gillis (2001) have shown that the integration of stress, length, vowel and constant quality leads to a good prediction of the syntactic category for English as well as Dutch words.

It was also shown that timing, lengthening, and pausing in spoken English are systematically related to the geometry of the phrase structure tree, i.e., they tend to occur at clause and phrase boundaries (e.g., Cooper 1975; Cooper & Paccia-cooper 1980; Klatt 1975; Gleitman et al 1988; Jusczyk et al 1992). Moreover, on a higher level, a particular falling intonation contour usually denotes a declarative or imperative sentence, while a particular rising intonation contour usually indicates a yes/no interrogative. Accordingly, if the learner can invert the function mapping syntactic structure onto the prosodic structure and intonation contour, it is possible to recover the syntactic analysis of input sentences without depending on any correspondence between syntax and semantics, and consequently to coin correct rules for the language.

Though prosodic information provides some cues to some syntactic categories and structural configurations, there are still some discrepancies between the prosodic form and the syntactic form that should be accounted for in order for this type of bootstrapping to be efficient. (See Lebeaux (1997) and Nespor & Vogel (1986), for a complete list of these differences).

The first discrepancy is that *"Prehead Specifiers of NP are grouped with the head in the Prosodic Structure, but separated from it in the Syntax"* (Nespor & Vogel 1986). The following examples show the difference between the structural bracketings, in (a) and the prosodic bracketings in (b) (Lebeaux 1997, 2001).

(20) a. $[_{NP}$the $[_{N'}$ picture of Mary]] Syntax

b. [the picture]Φ [of Mary]Φ {Φ = phonological phrase} Phonology

(21) a. [the [tall [cousin of Jeff]]] Syntax

b. [the tall cousin]Φ [of Jeff]Φ Phonology

The second discrepancy is that between the syntactic and prosodic grouping of the auxiliary verbs. In the syntax, the bracketing of these elements is right-branching, while in the phonology, the auxiliary verb forms a sort of complex verb with the main verb. The following bracketings show this discrepancy (Lebeaux, 2001: 91).

(22) a. [has [been [avidly [reading [about NP]]]]] Syntax

b. [has been avidly reading] [about NP] Phonology

The third discrepancy involves the prosodic and syntactic properties of the relative clause. In cases where this structure is recursive, i.e., there are more than one of them, they tend to break into individual intonational units (Chomsky & Halle 1968; Nespor & Vogel 1986). The following example shows the discrepancy (Lebeaux 2001: 93):

(23) a. this is the cat [that ate the rat [that ate the cheese]] Syntax

b. [this is the cat]$_I$ [that ate the rat]$_I$ [that ate the cheese]$_I$ Phonology

Here, the right-branching structure in the syntax breaks into three coordinated units in the phonology.

A fourth type of discrepancy, actually a cluster of three such discrepancies, seems quite systematic in English. This is the cliticization of a closed class head, H, onto a preceding specifier, even though it is grouped syntactically with the following complement. Lebeaux (2001) distinguished three distinct cases in which this occurs: the level of the Determiner Phrase (DP), at the level of Inflectional Phrase (IP), and the level of Complementizer Phrase (CP), as shown in the following examples.

(24) Syntax

a. level of IP: [John] [is going]

b. level of CP: [what] [is that]

c. level of DP: [John][‘s book]

(25) Phonology

a. level of IP: John is going

b. level of CP: what is that

c. level of DP: John’s book

That is, for all categories, the closed class head is cliticized backward, though it forms a syntactic category with the forward element.

3.3.4 Distributional Bootstrapping

This approach is inspired by, and builds on work in structural linguistics, where distributional methods were used as a methodology for deriving linguistic theory, rather than as models of acquisition. Accordingly, it is maintained that grammatical categories/constituents can be discovered on the basis of distributional relations among words; the occurrence of these words relative to each other within a context window (e.g., Bloomfield 1933; Harris 1951; Maratsos & Chalkley 1980; Finch & Charter 1992a, 1992b; Schütze 1996; Mintz 1996; Clark 2000; Klein & Manning (2001) ; Mintz, Newport & Bever 2002).[16] The main idea in this approach is that of *distributional test* and *substitutability* (Bloomfield 1933; Harris1951):

[16] The distributional analysis method can be used in theory with any type of information, e.g., semantic, or phonological. However, the term *distributional bootstrapping* will be used to refer to bootstrapping that uses only pure co-occurrence distributional information with no reference to semantic, syntactic, or phonological information.

(i) if all occurrences of word A can be replaced by word B, without loss of syntactic well-formedness, then they share the same syntactic category;

(ii) a constituent is a sequence of words with variants which can be substituted for that sequence.

The concept of *distributional analysis* is described in detail in the following part, since it is one of the main mechanisms used in this dissertation. Below I review some implementations that use one form or another of this concept in identifying word syntactic categories and linguistic constituents. A common feature in these implementation of distributional analysis is that they entertain a more relaxed version of distributional analysis as proposed in (i) and (ii). Instead of requiring contextual equivalence to establish the categorical similarity of words, these implementations require only a threshold of contextual similarity.

For example, Maratsos & Chalkley (1980) proposed that grammatical categories could be established on the basis of the contextual similarity of words. For instance, in the sentence *The dog is barking at the moon*, the fact that both *dog* and *moon* are preceded by *the*, and are preceded by the same words throughout many sentences, would lead them to be classified together. Other words that fall into the same pattern would be classified in the same category. The resulting category would be nouns.

Assuming distributional analysis, Mintz (1996) and Mintz et al. (2002) also showed that by monitoring the immediate contexts of words, the similarity of those contexts could be used to cluster lexical items and that the clustering coincided with grammatical classes. More specifically, in an analysis of the lexical co-occurrence patterns, Mintz et al. (2002) showed that a window of one word to either side of the target word is sufficient to identify clusters that more or less correspond to nouns and verbs.

Similarly, Finch et al. (1992a, 1992b) and Schütze (1996) expanded the window size to include the two words before and after each the target word as context, and were able to identify more clusters that more or less look like nouns and verbs, as well as pronouns and prepositions.

Van Zaanen (2000) presented an unsupervised Alignment-Based algorithm to bootstrap grammatical constituents based on Harris's idea of substitutability, which states that if two constituents are of the same type then they can be substituted by each other. This algorithm searches for constituents by using a reversal version of Harris's implication: *if parts of sentences can be substituted by each other then they are constituents of the same type*. The process of finding constituents applies in two phases. The *alignment learning phase* finds possible constituents by aligning pairs of sentences to each other. Groups that are different in both sentences are considered possible constituents. The following two sentences show how alignment works: [**Book Delta 128** *from Dallas to Boston*. **Give me all flights** *from Dallas to Boston*.] The italicized words indicate similar parts in the sentences. The dissimilar parts in bold are now considered constituents. When the second sentence is aligned to the third, it receives another constituent, which overlaps with the older constituent. Since the underlying grammar is assumed to be context-free, overlapping constituents are unwanted. The *selection learning phase* eliminates overlapping constituents after the alignment learning phase has finished. The best constituents are selected based on a statistical evaluation function. The probability of each constituent is computed. Using these probabilities, the probabilities of all combinations of all constituents are computed. The set of non-overlapping constituents with the highest probability is chosen to be correct. One problem with this system is that it stores all possible constituents and then after all possible constituents are found, the best constituents are selected. This makes the system slower with larger corpus.

Klein & Manning (2001) presented another unsupervised system for distributional grammar induction using part-of-speech tags as the contextual features.[17] The system uses a distributional notion of context in the following manner. Let α be a PoS tag sequence. Every occurrence of α will be in some context *x* α *y*, where *x* and *y* are the adjacent tags or sentence boundaries. The distribution over contexts in which α occurs is called its *signature*. The similarity of signatures indicates similar syntactic behavior. Accordingly, a metric of similarity is used and an agglomerative clustering process applies over tag sequences. Sequences are compared pair-wise, and the pair with the maximum similarity is merged. Merging two sequences involves the creation of a single new non-terminal category which rewrites as either sequence. Once there are non-terminal categories, the definitions of sequences and contexts become slightly more complex. The input sentences are parsed with the previous grammar state using a shallow parser which ties all parentless nodes together under a top root node. Sequences are then the ordered sets of adjacent sisters in this parse, and the context of a sequence can either be the preceding and following tags, the preceding and following tags, or a higher node in the tree. Merging a sequence and a single non-terminal results in a rule which rewrites the non-terminal as the sequence (i.e., that sequence is added to that non-terminal's class), and merging two non-terminals involves collapsing the two symbols in the grammar (i.e.,

[17] For a similar unsupervised algorithm, see Clark (2001).

those classes are merged). The grammar rules produced by the system are a strict subset of general context-free-grammar rules. As far as this dissertation is concerned, this system suffers from a bootstrapping problem, for the reasons discussed in Section 3.2 above, since it relies on PoS tagged data as input.

In addition to the problems in the specific implementations of distributional learning above, there are some general theoretical problems in distributional bootstrapping that, according to nativist approaches, render a completely input-driven learning procedure implausible for learning the grammar of human languages. For example, a learning procedure that is sensitive to patterns of lexical distribution to induce word order would run into problems due to the variability in sentence construction type in infant directed speech.

Moreover, given the role of equivalence and substitutability in distributional learning, Pinker (1987) argues that this approach would suffer from a categorization problem. For example, given sentences (a-b) below, a distributional learner would postulate that *fish* and *rabbits* belong to the same class.

(26) a. John ate fish.

b. John ate rabbits.

c. John can fish.

d. *John can rabbits.

Then, Pinker argues, when the learner comes across (c), the learner would assume that (d) is also permissible, which is incorrect. Pinker argues that this type of erroneous generalization would be common.

Another argument against a distributional analysis approach to grammar category learning is that often the kinds of distributional regularities that might be important are not local but occur over a variable distance, as in the following sentence "*The big fluffy brown and not so thin dog is barking at the moon*" (Chomsky 1965; Pinker 1987). Here the co-occurrence of *the* and *dog* spans six words. Thus, the problem is how the learner is to know which co-occurrences are important, and which should be ignored. Distributional analyses which consider all the possible relations among words in a corpus of sentence would be computationally unmanageable at best and impossible at worst.

More general arguments about the inadequacy of entirely input-based inductive mechanisms for learning natural languages are based on the formal demonstration that, without information about the types of sentences which are ungrammatical, an entirely input-driven learner will not succeed (Gold 1967; Berwick 1985; Wexler & Culicover 1980; Lightfoot 1991). Due to findings that infant and child language learners do not receive this kind of negative evidence (Brown & Hanlon 1970; Morgan & Travis 1989), many researchers conclude that bottom-up learning algorithms are not what humans utilize. In other words, the task of learning the grammar of a language is impossible, unless negative feedback is provided. Since negative feedback appeared to be unavailable or unused, this meant that language could not be learned without some additional innate constraints.

At its core, most of the search for innate constraints on language learning is grounded on the supposed impossibility of recovery from overgeneralization in an input-based learning mechanism.

In theory, these problems are clearly serious disadvantages in any learning approach. Yet there has never been a demonstration that these problems are actual problems in real speech corpora, and in particular, in speech addressed to infants and young children. The problematic examples might be rare enough that statistically they have no overall effect on learning. Mintz (1996) and (Mintz et al. 2002) have shown that this is the case, and that these potential problems in fact do not make learning by such methods impossible. This suggests that, whatever the strengths of these arguments, they are undermined by the successes of the implementations described above. Nonetheless, the successes of these methods in the limited aspects of language show that empirical research may produce better results than may be expected from considerations of linguistic theory (Redington & Chater 1997).

It is maintained here that with some modifications in the distributional methods, it is possible to construct a more well-defined architecture of an acquisition device that bootstraps a target language from the interaction between input cues and computationally necessary innate knowledge. In such architecture, a possible division of labor between distributional methods and traditional formal learning theory is possible (Osherson, Stob, & Weinstein 1985). In one possible scenario, distributional methods might be used in learning to encode the aspects of the language which are specific to particular languages, so that innate language universal knowledge can be brought to bear. This possibility is considered in detail in the following parts.

3.4 Conclusion

This part discussed some issues pertaining to language acquisition in general and frame identification in particular. Different approaches to the role of language input in learning were approached in bootstrapping terms. The discussion revealed some methodological issues that are central to the work presented in this dissertation.

Firstly, none of the algorithms discussed above has presented a formal and uniform description of bootstrapping and its mechanisms: (i) what is a possible bootstrap? (ii) what is the quantitative relationship between the initial bootstrapping knowledge the algorithm has to use, and the output bootstrapped knowledge it is expected to yield? Intuitively, the information needed to identify and use the bootstraps plus the information provided by them should be quantitatively smaller, using some quantification measure, than the information gained. Otherwise, the whole idea of bootstrapping will be paradoxical, from a language-acquisition perspective, and impractical from an NLP point of view. None of these algorithms has explicitly addressed such constraints.

Secondly, most of these algorithms, particularly those that are distributionally motivated, assume that all the data are present at the same time, i.e., at the start of learning. This of course is not the case in a natural language acquisition situation where the learner receives pieces of the input across an extended period of time. This means that these algorithms implicitly assume that initial learning makes use of input information that cannot be used unless some learning has already taken place.

Thirdly, most of the initial knowledge used by these algorithms is language-specific which reduces the possibility of porting these algorithms to other languages since this will need knowing the language first to decide what knowledge is needed for bootstrapping.

Finally, distributional algorithms for frame identification proceed on the assumption that learning different aspects of a grammar occurs independently of each other, using different mechanisms. For example, there is a mechanism for verb detection that is different and independent from the mechanism used in frame identification, as it is the case with Brent’s algorithm. This implicitly means that the knowledge used in one mechanism is not reusable in another. In addition to this *Autonomy Problem*, these algorithms suffer from another problem that was mentioned above, i.e., *Sequentiality Problem*; that is, frame acquisition does not take place until function words, in Brent’s, and PoS and NPs, in other algorithms, have been acquired. This has two implications: the first is that subcategorization occurs only when categorization has been fully completed, and the other is that subcategorization information does not play any role at all in identifying syntactic categories and NPs in the input. More generally, these two problems imply that the knowledge used in one mechanism cannot be reused by another. Language acquisition research has shown that this is not the case.

Chapter 4

Cues in the Input

Different cues have been used in the different approaches to bootstrapping discussed in the previous chapters. In semantic bootstrapping objects were used as cues of nouns, and actions as cues of verbs. In syntactic bootstrapping, partial structures were used as cues of predicates and their meanings. In prosodic bootstrapping, stress patterns were used as a cue of closed- and open-class words, as well as nouns and verbs, and timing, lengthening, and pausing as cues of clause and phrase boundaries. In distributional bootstrapping, distributional similarity was used as a cue of word classes and constituency. Table 9 summarizes the different cues and the cued events used in different bootstrapping approaches.

Bootstrapping Approach	**Cues**	**Cued Event**
Semantic	Things & People Actions & Relations	Nouns Verbs
Syntactic	Partial syntactic information	Verb meaning
Prosodic	Intonation Contours Stress Patterns Pausing	Sentence Types Open Class vs. Closed Class Phrase/Clause Boundary
Distributional	Distributional Similarity Substitutability	Word Classes Constituents

Table 9: Summary of cues used in different bootstrapping approaches

This chapter summarizes evidence from psycholinguistic experiments that demonstrate children's (as well as adults) attendance to some cues in the input. Until a formal definition of cues is given, it is assumed for now that (i) there are some elements that are intrinsically cues by virtue of some properties that make them distinctly marked in the input, and that (ii) learners are sensitive to these elements. The learner's task then is to discover the events that might be associated with these elements. Psycholinguistic research has shown that, according to this assumption, stress, pausing, lengthening, intonation, entities, properties, events, distributional regularity, and frequency, *inter alia*, are possible cues.

4.1 Cues in the Signal

In a set of experiments on English-learning 7.5-month-olds, Jusczyk et al. (1993, 2001) showed that English-learning infants listened longer to words exhibiting the canonical strong-weak pattern of English words than to words exhibiting a weak-strong pattern.

In a similar set of experiments, Echols (2001) showed that English-learning 9-month-olds attended significantly longer to stimuli containing changes in final syllables, and marginally longer to stimuli containing changes in stressed syllables. It was also shown that the effects of stress and position are additive, that is, infants attended least to changes in unstressed nonfinal syllables, about equally to changes in stressed and in final syllables, and most to changes in syllables that were both stressed and final. (See also Jusczyk & Thompson, 1978; Kuhl, 1983; Morse, 1972). These results tend to support the view that stressed or final syllables are attended to and represented more precisely by 9-month old infants than syllables that are unstressed and nonfinal. Sansavini et al. (1997) reported similar results with newborn Italian infants.

Evidence in support of the claims that stress patterns are fairly diagnostic of grammatical classes was found in stress patterns of disyllabic words (Kelly 1992, 1996). An examination of 3,000 disyllabic nouns and 1,000 disyllabic verbs, drawn from Francis and Kučera (1982), revealed that 85% of words with final stress are verbs and 90% of words with initial stress are nouns. Subsequent experiments in which subjects either had to construct sentences with a disyllabic word which could have either stress pattern, or read target sentences containing a disyllabic non-word in either nominal or verbal position, showed an outspoken preference for linking iambic words with the verb category and trochaic words with noun category. This alternating stress pattern can also be used by a syntactically naïve learner to assign words to two major categories, which closely correspond to content words and function words, before discovering the distributional regularities of particular morphemes (Gerken et al. 1990; Gerken 1994a; Morgan et al. 1996; Jusczyk 2001).

It was also shown (e.g., Cooper & Paccia-Cooper, 1980; Klatt, 1975; Wightman et al. 1992) that learners are sensitive to pre-boundary lengthening, or lengthening of the rhyme, the part of a syllable that does not include the initial consonant/s – for example, [*ae*] in ‘cat’ or [u] in ‘Lou’, at the end of a grammatical unit.

There is also evidence of adult speakers’ sensitivity to pause duration (Cooper & Paccia-Cooper, 1980; Scott, 1982). For example, speakers tend to produce longer pauses at word boundaries when they coincide with clause boundaries.

In an experiment on sixteen infants aged between 6 and 12 weeks with a monolingual French background, Christophe et al. (2003) showed that infants can perceive prominence within phonological phrases.

Beckman & Pierrehumbert (1986) have found that learners are sensitive to the fact that a special kind of intonation pattern or tone pattern may occur at a phrase or clause boundary.

It was also demonstrated that infants are capable of discriminating acoustic properties such as pitch change by 1-2 months old (Morse, 1972). By 4.5 months, infants begin to show sensitivities to certain prosodic markers in fluent speech, preferring passages with artificial pauses inserted at clause boundaries rather than other places in the sentence (Juscyzk, Hohne, & Mandel, 1995; see also Hirsh-Pasek et al., 1987; Kemler Nelson et al., 1995; Morgan et al., 1993).

Further insights about infants' sensitivity to phrase-level prosodic cues (i.e., cues associated with phrase boundaries) were provided in a follow-up study which examined different sentence types (Gerken et al., 1994). Gerken et al. compared sentences such as (27) with sentences such as (28).

(27) (Joe) (kissed the dog).

(28) (He kissed) (the dog).

In sentences of the type exemplified in (27), speakers are likely to produce a prosodic boundary before the verb "kissed", which coincides with the subject/VP syntactic boundary. However, in sentences of the type exemplified in (28), which contains a weak pronoun, speakers either do not produce a salient prosodic boundary, or place the prosodic boundary after the verb "kissed" (e.g., Gee & Grosjean, 1983). Nine-month-old infants showed a preference for passages where the pause was located before the verb when sentences such as (27) were used as stimuli, but showed no preference for placement of the pause either before or after the verb when sentences with pronoun subjects such as (28) were used. In other words, infants demonstrated sensitivity to the syntactic boundary only in the first case, when it more reliably coincided with a prosodic boundary in natural speech.

In addition to the evidence these experiments provide for children's sensitivity to these cues, it has also been shown that children may prefer one cue over the other. For example, Mattys et al. (1999) pitted sequences with good prosodic cues and poor phonotactic cues to word boundaries against ones with good phonotactic cues but poor prosodic cues. English-learning 9 month-olds favored the sequences with the good prosodic cues, suggesting that, at this age, they give greater weight to prosodic cues than to phonotactic cues.

4.2 Cues in the World

Children's sensitivity to entities in the world has been highlighted by experiments on early lexical development that showed the primacy of nouns in the early stages of speech. The primacy of nouns, especially object labels, in the early lexicon has been reported for language communities as different as English, German, Japanese, Kaluli, Mandarin, Turkish, Italian, and Hebrew (Gentner, 1982; Dromi, 1987; Goldfield, 1993; Caselli et al., 1995).

Based on a study of early speech in English and several other languages, Gentner (1982) and E. Clark (1983) showed that nouns have primacy in that the words belonging to this category are acquired first and are predominant in children's early vocabulary. Table 10 gives a history of the words produced by Tad, whose linguistic development was studied by his mother and Gentner. Nouns are clearly the predominant early category here. Not only are all but one of the first dozen words to emerge nouns; this category remains numerically dominant throughout the first months of linguistic development. The second most common word class in Tad's speech – what Gentner calls the 'predicate' category – consists of words that name a property. This category later divides into verbs and adjectives, corresponding roughly to the distinction between action-type properties like 'running' and 'reading' and state-like properties such as 'tall' and 'good'.

Age (month)	Nominal	Predicate	Expressive	Intermediate
11	dog			
12	duck			
13	Daddy	yuk		
	Mama			
	teh (teddy bear)			
	car			
14	dipe (diaper)			
	owl			
	toot toot (horn)			
15	keys			
	cheese			
16	eye			
18	cow	hot		bath
	cup			
	truck			
19	kitty	happy	oops	pee pee
	juice	down	boo	TV
	bottle	up	hi	
	spoon		bye	
	bowl		uh oh	
	towel			
	apple			
	teeth			
	cheek			
	knee			
	elbow			
	map			
	ball			
	block			
	bus			
	Jeep			

Table 10: Tad’s Early Words (from Gentner 1982:306)

Similar findings have been reported in many other studies, including the naming study conducted by Goldin-Meadow et al. (1976) on three children aged between 8 and 26 months. Table 11 shows the results of the production task (naming objects and actions) and comprehension (pointing to objects and acting out actions in response to the experimenter). Consistent with Gentner’s claim, nouns far outnumber verbs in the production data from all three subjects and are the first words used in the children’s own speech. The difference between nouns and verbs is less dramatic in the comprehension task but still favors the noun category by a factor of about 2.

	Number of Different		Number of Different	
Age (mon. wk.)	Nouns	Verbs	Nouns	Verbs
Lexie				
22.0	7	0	35	22
24.2	17	0	54	26
25.0	28	3	58	27
25.1	40	7	61	27
Melissa				
19.1	5	0	22	14
22.1	9	0	40	16
Jenny				
14.0	10	0	27	9
16.0	19	0	33	14
17.0	29	4	38	18
17.1	34	6	45	18

Table 11: Results of Goldin-Meadow et al. (1976)

Further support for the early predominance of nouns can be found in more recent studies as well. For example, based on their longitudinal study of 30 children, Bates et al. (1988: 153) report that at age 20 months, nouns were dominant (46.8% of total vocabulary) compared to verbs (8.3%) and adjectives (7.5%).[18] Drawing on diary data collected from 1803 subjects aged 8 months to 2;6, Bates et al. (1994: 95) report that ‘common’ nouns make up almost 40% of the first 50 words in children’s early vocabulary; the next largest word-class (so-called ‘predicates’) accounts for less than 10% of early vocabulary items. (Bates and her colleagues did not include proper names or places in their calculations; had they done so, the proportion of nouns in early speech would have been even higher.)

[18] These numbers add up to only 62.6%. The authors did not mention anything about the remaining 37.4%.

Nelson et al. (1993) provide information about the subclasses of nouns found in the early vocabulary of children aged between 13 and 20 month; see Table 12. They were especially interested in the contrast between basic level objects (BLOCS), which denoted a category of ‘discrete whole individual objects’ –e.g., *puppy, cheerios, toy, animal*. All other count nouns, including those denoting locations (*beach*, *kitchen*), single actions (*kiss*, *help*), events (*lunch, party*), person roles (*doctor, brother*), natural phenomena (*sky*, *snow*), temporal entities (*morning, day*), parts of objects (*button*), quantities (*drop*), and material (*wood*), were grouped together and dubbed XBLOCS. Excluding words that can belong to more than one category, the mean proportion of nouns in the vocabulary of the children stood at 65%, including a sizeable component (one third of all count nouns) that did not refer to basic level objects.

Word Type	Mean Proportion (%)		
Nouns			65
count nouns		54	
BLOCS	36		
XBLOCS	18		
proper nouns		4	
mass nouns		7	
Dual category[a]			6
Verbs			10
Other			19

a. Dual category items are words that can belong to more than one category (e.g., *drink*, which can function as either a noun or a verb).

Table 12: Mean Proportion of Word Types in Productive Vocabularies at 20 months (based on Nelson et al. (1993: 70))

Nouns not only predominate in the period of first words but also in the period of the vocabulary spurt, which is commonly characterized by an accelerated rate of noun learning (Goldfield & Reznick, 1990; Bates et al., 1994).

Logically, early lexical acquisition in general, and the primacy of nouns in particular, are attributable to child factors, environmental factors, or a combination of the two. Those who focus on child factors include constraint theorists who propose that the task of word learning is simplified by the application of internal linguistic constraints. Early constraints make noun mapping likely. For example, according to the whole object constraint, children initially assume that all words refer to objects and that they refer to the whole object, rather than its parts, attributes, motion, temporary state or other associated properties (Markman, 1987). Others argue for the child's application of principles rather than absolute constraints. Principle theorists suggest that lexical principles are learned to a great extent, hence both child and environmental factors play a role. They view lexical principles essentially as strategies that effectively restrict the search space for the task of word-to-referent mapping (Golinkoff et al., 1994). Similar to the whole object constraint, the principle of object scope (Golinkoff et al., 1994) posits that words label whole objects. Upon hearing a novel word and witnessing an unnamed object in a novel event, children using this principle would likely assume that the novel word refers to the object and not the event.

Gentner’s explanation of the early primacy of nouns is based on semantic considerations, particularly the idea that the referents of nouns tend to have perceptual correlates that are comparatively easy to identify and are therefore more ‘accessible’ to children than those of verbs. In contrast, verbs and other predicates are claimed to have a less transparent relationship with the perceptual world. To illustrate this, Gentner takes the example of a bottle floating down a stream into a cave. Although all languages pick out the bottle as a salient component of the situation and use a single word or phrase to refer to it, there are differences in how the movement is encoded. Whereas English encodes it with the help of a verb and a preposition (e.g., *floated into the cave*), Spanish uses two verbs and a preposition (e.g., *entró en la cueva, flotando*).[19] Gentner takes this as evidence that the types of meanings encoded by verbs are not so obviously 'packaged' as those of nouns, which makes them correspondingly more difficult to acquire.

The plausibility of Gentner's explanation of the early preference of nouns over verbs is further supported by the fact that this preference cannot be attributed to other factors that are likely to be relevant (Gentner 1982).

For example, the primacy of nouns cannot be attributed to inflectional factors (e.g., the fact that in English more verbs than nouns have irregular inflections and therefore do not present the child with a single, fixed root). This is because there is also an early preference for nouns in Mandarin Chinese, in which neither verbs nor nouns are inflected, as well as in Turkish, which exhibits heavy but regular inflection on both verbs and nouns.

It is likewise unlikely that word order is the crucial factor. While nouns can appear sentence-finally in English (this being a highly salient position in the sentence for the child), they normally do not in Japanese, which is uniformly verb-final.

Gentner also claimed that the early emergence of nouns could not be attributed to frequency effects. Relying on data from adult-to-adult speech, she noted that nouns are less frequent than either verbs or prepositions and that among the 100 most frequent words in English, 20 are verbs and only 6 are nouns (1982:316-17). This claim has been supported by other researchers (e.g., Goldfield 1993; Au et al., 1994). For example, Goldfield examined the frequency of nouns and verbs in 17-minute speech sampling

[19] For detailed discussion of this phenomenon, see Talmy 1985.

involving 12 one-year-old children and their mothers in situations involving playing with toys and play with each other (tickling, peek boo, and so forth). She reported that there is an overall frequency advantage of verbs (9.67 tokens per minute during play with toys vs. 7.01 tokens per minute for nouns).

Generally, the evidence presented above for the primacy of nouns in many different languages can plausibly be interpreted as a strong evidence of children's sensitivity to objects (people and things) in the world, which in turn supports using these objects as cues to some linguistic knowledge as maintained by semantic bootstrapping.

4.3 Distributional Cues

By eight months, infants are sensitive to statistical properties of the input (Saffran et al., 1996a) and by 9 months, they are presumably integrating these two cues. The study by Saffran et al. (1996a) has drawn attention to information that infants can extract from speech on the basis of distributional cues alone. Saffran et al. reasoned that because words can be defined as units of sound which consistently co-occur, noting the likelihood of one syllable following another could provide a reliable strategy for segmenting words. For example, the two-word string *prettybaby* consists of four syllables: *pre*, *ty, bay*, and *by*. The first two syllables (*pre* and *ty*) consistently appear together because they form a word. Likewise, the latter two syllables (*bay* and *by*) also tend to occur together. However, the second and third syllables (*ty* and *bay*) occur together relatively rarely. Across a corpus of English, the syllable *ty* follows the syllable *pre* more frequently than the syllable *bay* follows the syllable *ty*, because many different words can follow the word *pretty* (i.e. *pretty flower*), but only a few syllables can follow *pre*. This greater predictability of word internal syllables than syllables spanning word boundaries may be helpful in discovering word boundaries.

Still, noting co-occurrences between syllables will not provide a sufficient cue to accurately segment the speech stream. For instance, if infants were to segment the input simply by noting co-occurrences between syllables, they would be misled to treat commonly occurring syllable pairs, such as *the.dog*, as words. Therefore, Saffran et al. (1996a) proposed that besides tracking the likelihood of one particular syllable following any other particular syllable, infants also track the baseline frequency of the first syllable in the syllable pair. This parsing strategy can be formalized by a statistical relationship: Transitional Probability (= Conditional Probability), where T.P. = (frequency of Y given X)/ (frequency of X). Thus, frequently occurring words such as *the.dog* will not be mistaken as a word because *the* also occurs before many other words.

Aslin et al. (1998) showed that infants respond to transitional probabilities as opposed to simple co-occurrences between syllables. The idea of tracking the probability of one phone following another to detect word boundaries is not new (Harris, 1951; Hayes & Clark, 1970). However, Saffran et al. (1996a) first showed that statistics are a psychologically plausible means for infants to begin to segment words. They familiarized 8-month olds with a 2-minute stream of an artificial language containing 4 tri-syllabic nonsense words: *pabiku*, *tibudo*, *golatu*, and *daropi*. No acoustic cues to word boundaries were present in the speech stream. Only the distributional properties of the sequences of syllables provided cues to the location of word boundaries. After familiarization, the infants were tested for their listening preferences to words versus part-words (tri-syllabic sequences composed of the last syllable of a word and the first two syllables of another word, based on the nonsense words mentioned above, *tudaro* is a part-word). The infants listened significantly longer to the part-words, indicating they can segment the speech stream based on statistics alone.

Saffran et al. (1996) exposed children to a two-minute stream of synthesized speech containing no cues to word boundaries other than the transitional probabilities between syllables. The continuous stream of speech was constructed by concatenating synthesized consonant-vowel (CV) syllables. Saffran et al. (1996) found that infants listened reliably longer towards part-words, indicating that they extracted words defined only by the statistical nature of the speech stream.

Johnson & Jusczyk (2001) provide further evidence. As in Saffran et al. (1996a), J & J found that infants listened significantly longer to the novel part-words, demonstrating their ability to use statistical cues to discover word boundaries in continuous speech. Infants performed nearly identically in this experiment as they did in both the analogous synthesized speech (Saffran et al., 1996a) and tone segmentation (Saffran et al., 1999) tasks.

Learning of the statistical regularities of the language is also suggested by the observation that 9-month-old children listen longer to words that include frequent phonetic sequences than to legal but rare phonetic sequences (Jusczyk, Luce, & Charles-Luce, 1994).

By 9 months, young children prefer to listen to syllables that obey phonotactic rules than to illegal syllables (Friederici & Wessels, 1993; Jusczyk et al., 1993) and they also exhibit a preference for high over low probability phonotactic sequences (Jusczyk et al., 1994). When asked to judge how much nonwords are wordlike (phonological goodness judgment), adults rate nonwords with high transitional probabilities between phonemes as more wordlike than nonwords with low transitional probabilities (e. g., Frisch et al., 2000; Vitevich et al., 1997).

Similar studies using either tone or visual sequences as stimuli revealed that infants' ability to track transitional probabilities is not limited to linguistic stimuli (Aslin et al., 2001; Saffran et al., 1999).

However, there is evidence that 8-month olds show more sensitivity to speech cues than to distributional cues. In an experiment on sixteen 8-month-olds from monolingual English-speaking homes (5 males, 11 females; mean age 35 weeks 2 days; range 33:5 days to 36:5), Johnson and Jusczyk (2001) pitted two competing cues to word segmentation against each other: stress and statistics. In segmenting the familiarization sequence, the infants relied more heavily on the stress cue to indicate word onsets than on the statistical cue relating to the transitional probabilities of successive syllables. Consequently, it appears that although statistical cues are sufficient to segment a simple artificial language, 8-month olds weigh speech cues such as stress more heavily.

4.4 The Frequency Effect

The frequency effect is one of the earliest and most robust effects in psycholinguistics. Frequency plays a role in both the auditory and visual modalities, and in both comprehension and production (Jurafsky 2003).

The earliest work studying frequency effects in comprehension seems to have been by Howes and Solomon (1951). They displayed words, a word at a time for longer and longer durations, to adult subjects who were asked to recognize them. They showed that the log frequency of a word (as computed from corpora of over 4 million words) correlated highly with the mean time subjects took to recognize the word; more frequent words were recognized with shorter presentations. Later, the naming paradigm, in which subjects were read a word out loud, was used to show that high-frequency words are named more rapidly than low-frequency words (Forster and Chambers 1973). In the lexical decision paradigm, in which subjects decide if a string of letters presented visually to them is a word or not, it has also been shown that lexical decisions about high-frequency words are made faster than decisions about low-frequency words (Rubenstein et al., 1970; Whaley 1978; Balota and Chumbley 1984).

Similarly robust results have been found for auditory word recognition. Howes (1957) first found results with speech that were similar to his earlier results with vision: when presented with high- and low-frequency words immersed in noise, subjects were better at identifying high- than low-frequency ones. In an extension to this experiment, Savin (1963) found that when subjects made recognition errors, they responded with words that were higher in frequency than the words that were presented. Grosjean (1980) used the gating paradigm, in which subjects hear more and more of the waveform of a spoken word, to show that high-frequency words are recognized earlier (i.e., given less of the speech waveform) than low-frequency words. Tyler (1984) showed the same results for Dutch.

The effects of lexical frequency on production have also been reported in a number of studies. In two separate studies of the effect of lexical frequency on phonological reduction (number of deleted or reduced phonemes), Fidelholz (1975) and Hooper (1976) showed that frequent words such as *forget* are more likely to have lexically reduced vowels (e.g., [fər]) than less frequent words such as *forgo* (e.g., [fɔr]) see Table13.

Reduced word [fər]		Full vowel [fɔr]	
Word	Count per million	Word	Count per million
Forget	148	forefend	<1
Forgive	40	forgo	<1

Table 13: Lexically reduced vowels in high-frequency words. (After Fidelholz 1975)

While these studies are suggestive of an effect of frequency on a word's phonological makeup, they do not confirm that the effect of frequency on lexical production is on-line and productive. It could be that frequent words have reduced vowels and fewer phonemes because of some diachronic fact statistically reflected in the lexicon that is only related to on-line production in a complex and indirect way (Jurafsky 2003: 45).

To show that frequency plays an active and on-line role in language production, a number of studies have examined whether frequency dynamically affects phonological variation in production. Bybee (2000) examined word-final /t/ and /d/ in a corpus of spoken Chicago English. After excluding the extremely high frequency words *just, went,* and *and*, she classified the remaining 2,000 word tokens into two bins, high-frequency (defined as more than 35 per million in the Brown corpus) and low-frequency (fewer than 35 per million.) She showed that final /t/ and /d/ deletion rates were greater in high-frequency words (54.5%) than in low-frequency words (34.3%). Hay (2000) has shown that for complex words, the ratio of the frequency of the derived word and the frequency of its base is an important predicator of processing time.

Gregory et al. (2000), Jurafsky et al. (2001), and Bell et al. (2001) provided further evidence that these frequency effects on reduction are on-line, by controlling for a wide variety of contextual factors, and also by investigating the effect of frequency on a word's duration, in addition to its phonological reduction. They examined the duration of words and the percentage of final-consonant deletion in a 38,000-word phonetically transcribed sub-corpus from the Switchboard corpus of American English telephone conversations (Godfrey et al. 1992; Greenberg et al. 1996). They first confirmed Bybee's results by analyzing 2,042 word tokens whose full pronunciation ended in /t/ or /d/. After controlling for contextual factors, they found that these final obstruents are more likely to be deleted in more frequent words. High-frequency words were 2.0 times more likely to have deleted final /t/ or /d/ than low-frequency words.

Gregory et al. (2000) and Jurafsky et al. (2001) also investigated the effects of frequency on word duration, using 1,1412 monosyllabic word tokens ending in /t/ or /d/. They found a strong effect of word frequency on duration. Overall, high-frequency words were 18% shorter than low-frequency words.

Pan and Hirschberg (2000) have also shown that conditional bigram probability correlates highly with location of pitch accent; specifically, pitch accent is more likely to occur on low-probability words. Gregory (2001) has extended this result by showing that conditional probability given previous and following words is a significant predictor of pitch accent even after controlling for other contextual factors such as position in the intonation phrase, part of speech, and number of syllables.

Oldfield and Wingfield (1965), for example, showed an on-line effect of word frequency on latency (the time to start producing a word) of picture-naming times.

Presenting subjects with pictures, they found that pictures with high-frequency names were named faster than pictures with low-frequency names. Wingfield (1968) showed that this effect must be caused by word frequency rather than the frequency of pictured objects, by showing that the effect was not replicated when subjects were asked to recognize but not verbalize picture names. These results were also reported for Dutch (Jescheniak and Levelt 1994).

A number of experiments have shown that frequency plays a role in disambiguation. For example, in an experiment by Simpson and Burgess (1985), subjects were first presented with an ambiguous prime word (homograph) that had a more frequent sense and a less frequent sense. Subjects then performed lexical decision on targets that were associated with either the more frequent or the less frequent meaning of the homograph prime. Simpson and Burgess found that the more frequent meaning of the homograph caused faster response latencies to related associates, suggesting that the more frequent meaning is retrieved more quickly. Evidence for the use of word sense frequency in comprehension has also been reported crosslinguistically – for example, in Chinese (Li and Yip 1996).

MacDonald (1993) studied the effect of word-pair (joint) frequencies on comprehension. Investigating the process of a noun followed by a word that is ambiguous between a noun and verb, such as the pair *miracle cures,* she hypothesized that if the noun-noun pair was frequent (like *miracle cures*), its interpretation would be biased toward the noun reading of the second word. She predicted no such bias for infrequent noun-noun pairs (like *shrine cures*). She confirmed this hypothesis by looking at reading time just after the ambiguous word in sentences that were otherwise biased toward a verb reading. For example, subjects spent more time reading the word *people* in (30) than in (29), since the frequent noun-noun phrase in (30) biases the reader toward the noun reading of *cures*, whereas the word *people* is compatible only with the verb reading.

(29) The doctor refused to believe that the *shrine cures* people of many fatal diseases…

(30) The doctor refused to believe that the *miracle cures* people of many fatal diseases…

It has also been shown that the frequency of subcategorization frames plays an on-line role in the disambiguation of various syntactic ambiguities. For example, the verbs *remember* and *suspect* are both subcategorized for either a direct object NP or a sentential complement S, as in (31)-(34) (from Jurafsky 2003:53):

(31) The doctor remembered [$_{\text{NP}}$ the idea].

(32) The doctor remembered [$_{\text{S}}$ that the idea had already been proposed].

(33) The doctor suspected [$_{\text{NP}}$ the idea].

(34) The doctor suspected [$_{\text{S}}$ that the idea would not turn out to work].

While both verbs allow both subcategorization frames, they do so with different frequencies. *Remembered* is more frequently used with an NP complement, while *suspected* is more frequently used with a sentential complement. These frequencies can be computed either from a parsed or transitivity-coded corpus (Merlo 1994; Roland and Jurafsky 1998) or by asking subjects to write sentences using the verbs (Connine et al. 1984; Garnsey et al. 1997).

For example, Trueswell, Tanenhaus, and Kello (1993) tested this effect in an experiment based on cross-modal naming (i.e., the stimulus is auditory while the target is orthographic). Subjects heard a sentence prefix ending in either an S-bias verb (*The old man suspected…*) or an NP-bias verb (*The old man remembered…*). They then had to read out loud ("name") the word *him*. Previous research had shown that naming latencies are longer when the word being read is an ungrammatical or unexpected continuation. In Trueswell et al.'s study, naming latency to *him* was longer after S-bias verbs (*The old man suspected …him*) than after NP-bias verbs (*The old man remembered …him*). This suggests that subjects preferred the more frequent frame of the verb and were surprised when the preference was overturned, causing longer naming latencies.

MacDonald (1994) showed that the effect of subcategorization frame frequency also plays a role in resolving main clause/relative clause ambiguities in garden-path sentences, as first pointed out by Bever's (1970) famous example (*The horse raced past the barn fell*). MacDonald suggested that the subcategorization frequencies proposed by earlier researchers could play a role in explaining processing difficulties in main verb/reduced relative ambiguities. Her test materials used transitive-bias verbs like *push* and intransitive-bias verbs like *move*, in sentences like the following:

(35) The rancher could see that the nervous cattle *pushed* into the crowded pen were afraid of the cowboys.

(36) The rancher could see that the nervous cattle *moved* into the crowded pen were afraid of the cowboys.

MacDonald found that corrected reading times in the disambiguation region *were afraid* were longer for intransitive-bias verbs like *move* than transitive-bias verbs like *push*.

In this section I have summarized some of the psycholinguistic research on learners' sensitivity to a set of prosodic, acoustic, semantic, and distributional cues that provided some strong evidence that some elements are cues by virtue of their intrinsic (perceptual) properties. Below I show how some of these cues can be distributionally learned from a given corpus, and how they can be used in cue-based distributional learning.

Chapter 5

Foundations of Cue-Based Learning

It has been assumed so far that (i) there are some elements that are intrinsically cues by virtue of some properties that make them distinctly marked in the input, and that (ii) learners are sensitive to these elements. Psycholinguistic evidence has shown that this is the case. However, it is maintained here that for some of these cues to be efficiently used in a distributional learner, they should be defined and learned according to some criteria. In this chapter, I first present two procedures for cue extraction. The first is semantic and the other is purely distributional. The main idea in both procedures is to extract, according to some criterion, the smallest subset of elements in the input that provide information about the distributional properties of the maximum number of elements in the input. The criterion used here for selecting this subset is such that every element in the input occurs at least once with at least one element in this subset. Finding this subset is the core of the model of cue-based learning presented in this dissertation. Using the cues extracted by the cue identification procedures, I then present two procedures for how to establish distributional similarity, which provides the basis of the categorization and subcategorization algorithms.

5.1 Semantic Cue Extraction

The logic behind a semantic procedure for cue extraction is rooted in the core of semantic bootstrapping, i.e., nouns and verbs typically refer to distinct, identifiable semantic classes in the input. Accordingly, people and physical objects are referred to with nouns; activities and changes of state with verbs; properties and colors with adjectives, and so on (Croft 1991; Grimshaw 1981; Macnamara 1982; and Pinker 1984, 1994). It was mentioned in Section 4.2 above that children's sensitivity to entities in the world has been highlighted by experiments on early lexical development that showed the primacy of nouns in the early stages of speech.

The way this evidence is used in the semantic criterion for cue extraction is to find the smallest subset of nominal expressions, words that refer to people and things, in the input that is likely to provide information about the distributional properties of other elements in the input. In theory, these cues can be identified according to the criterion in (37a) and the procedure in (37b).

(37) a. Semantic Criterion for Cue Extraction

Let

$W = \{w_1 \ldots w_n\}$be the set of words in a corpus *R*,
$O = \{o_1, \ldots, o_m\}$ be the subset of *W* that refer to objects in the world,

Then

Cues, *K*, are the smallest subset of *O* such that every word in $\{w_1, \ldots, w_n\}$occurs at least once with at least one member in this subset.

(37) b. Procedure for Semantic Cue Extraction

```
Function SUBSET(K,R)
K := Ø;
for i := 0 to m do
get the number of word types N in R;
get the frequency of $o_i$ , $f(o_i)$ in R ;
build a decreasing frequency profile F:= $\{f(o_i) > f(o_{i+1}) > \ldots f(o_m)\}$ ;
get the number of words $|w_{i-}|$ that immediately precede $o_i$ ;
get the number of words $|w_{i+}|$ that immediately follow $o_i$ ;
        $Sum_i := |w_{i-}| + |w_{i+}|$ ;
        $Sum\_total := \sum_{i=1} Sum_i$ ;
        if Sum_total := $\sum_{i=1} Sum_i$ := N
        return K := $\{o_i\}$;
        else
                repeat
                        i := i +1;
                        until (Sum_total := $\sum_{i=0}^{i+k} Sum_i$ := N);
                return K := $\{o_i, .., o_{i+k}\}$
```

To salvage this, we can seed the algorithm with knowledge about the smallest subset of the smallest subset of *O*, as defined in (37). Accordingly, the choice adopted here is to annotate in the input names of people and things, and the smallest subset of pronouns. How names are identified and annotated in the corpus is treated in detail in Chapter (6). The subset of pronouns, on the other hand, is limited to pronouns that could constitute single-word noun phrases (e.g., *he, she, it, I,* etc.), compared to other pronouns which can constitute single NPs or be part of a larger NP (e.g., *her* and *his*), and pronouns that are always part of a larger NP (e.g., *my, your, etc*.). It is assumed that the subset of pronouns that constitute single-word NPs are easier than other pronouns to identify in the input. No claim is made here regarding the accuracy or the psycholinguistic feasibility of this assumption.

Once this subset has been established, it is then used by the learning mechanism to acquire knowledge about other elements in the corpus, which is consequently employed to garner further knowledge, and so on and so forth. The dynamics of this learning mechanism is detailed in Chapter (6).

5.2 Distributional Cue Extraction

Given the difficulty and incompleteness of the semantic procedure for cue extraction, I propose in this section a purely distributional procedure for learning cues from the input, using only language-internal information.

A first approximation to this procedure can make use of the frequency of certain elements or features in the input. Accordingly, a cue can be any member of the set of the highly frequent elements in the input. Consequently, function words, stress, and silence as indicated by utterance boundaries can be possible cues. Utterance boundaries are cues by definition since they indicate the beginning and end of some constituents. Function words are highly frequent in the input, which, among other features, makes them stand out in the input. For that reason, some of the learning methods discussed in previous chapters have used these words as cues (e.g., Brent 1991, 1993, 1994; Mintz 1996, and Buchholz, 1998).

However, detecting cues in accordance with this criterion has two disadvantages. The first, as previously discussed, is that these function words must be learned before they can be used as cues. Using them in this way results in the bootstrapping paradox, discussed above. The other is that applications that entertain this definition of cues usually resort to an arbitrary cut-off point to establish a list of highly frequent words. Mintz (1996), for example, made the cut-off point after the 200th most frequent word in the corpus used.[20] This means that establishing the set of possible cues in these arbitrary and subjective terms is not expected to guarantee an objective and standard mechanism for cue detection.

The alternative mechanism introduced here is based on the assumption that cues should be first learned by the learning algorithm. Cue learning will definitely benefit from the frequency effect, yet indirectly.

Generally, highly frequent elements in the input should provide pieces of information about the distributional properties of a non-trivial subset of elements in the input. It is possible that every element in the input will occur at least once with at least one highly frequent element. Put together, these pieces then represent a summary of the distributional properties of the elements in the input language. What is needed then is a

[20] Mintz used input corpora for 4 different children that were selected from the CHILDES database. The average number of utterances was 12,444.

subset of these frequent elements that carries information about every element in the input. This subset will be referred to as Category Cues to distinguish them from Frame Cues introduced later. In theory, the set of Category Cues can be defined as follows:

(38) Definition of Category Cues (*K*)
The set of Category Cues, *K*, is the smallest subset of the elements $\{k_1,...,k_m\}$in a corpus *R* such that every element in *R* occurs at least once with at least one member in *K*.

Depending on the level of learning pursued, an element in this subset can be a morphological or phonetic feature, a word, a tag, or a constituent.

Assuming cues at the word level, this subset *K* can generally be identified using two different methods. The first method is to find the exact subset according to the following criterion:

(39) Distributional Cue Extraction (1)
Let $W = \{w_1 \ldots w_n\}$ be the set of word in a corpus *R*,
for every word w_i, extract the set of words, S_i, that occur with w_i,
Then
the set of Category Cues for a corpus *R*, K_R, is such that
$K_R \equiv \{S_1 \cap S_2 \cap S_3 \cap \ldots \cap S_n\}$

However, finding such exact minimal subsets in general is known to be intractable (e.g., Hopcroft et al., 2001; Cormen et al., 1995: 916-986; and Mortello and Toth, 1990). This means that identifying cues according to (39) undermines the objective of using cues in learning. Consequently, what is introduced here is a method that approximates the set of possible cues *K* in a given corpus.

The approximate method proposed here makes a direct use of the highly frequent words in a corpus on the assumption that these words would provide information about the distributional properties of other words in the corpus. Unlike (37), the method presented here takes the frequency of a word in a corpus as the main determinant of that word being a cue in that corpus. The core of this method is to find the smallest subset of words in the corpus that co-occur with a number of words that converges to an order of the number of word types in that corpus. This method proceeds as follows.

We start with building a decreasing frequency profile for all the words $\{w_1,\ldots,w_n\}$ in a corpus, R, where w_1 is the first most frequent word in R, w_2 the second most frequent, and so on. The set of cues is $K = \{w_1,\ldots,w_m\}$, such that if we add up the number of words, X_1, that co-occur with w_1 and the number of words, X_2, that co-occur with w_2, until the m-most frequent word, w_m, the number of words $[X_1+X_2+\ldots+X_m]$ converges to an order, α, of n, where n is the number of word types in the corpus. In pseudo-code, this procedure is as follows:

(40) Procedure for Distributional Cue Extraction

```
Function SUBSET(K,R)
K := Ø;
for i := 0 to n do; α := (1,2,…);
get the number of word types $n$ in $R$;
get the frequency of $w_i$ , $f(w_i)$ in $R$ ;
build a decreasing frequency profile $F := \{f(w_i) > f(w_{i+1}) > \ldots f(w_n)\}$ ;
get the number of words $|w_{i-}|$ that immediately precede $w_i$ ;
get the number of words $|w_{i+}|$ that immediately follow $w_i$ ;
    $X_i := |w_{i-}| + |w_{i+}|$ ;
    $X_i\_total := \sum_{i=1} X_i$ ;
    if $X_i\_total := \sum_{i=1} X_i := \alpha n$
    return $K := \{w_i\}$;
    else
        repeat
            i := i +1;
        until ($X_i\_total := \sum_{i=0}^{i+k} X_i >= \alpha n$,
            $X_i\_total := \sum_{i=0}^{(i+k)-1} X_i < \alpha n$);
    return $K := \{w_i,..,w_{i+k}\}$
```

For example, if *the* is the most frequent word in a hypothetical corpus that contains 1000 word types, and the number of words that co-occur with *the* is 1000, then *the* is considered the first-order cue set for this corpus. If *of* and *to*, for example, are the second and third most frequent words in this corpus, respectively, and the number of words that co-occur with *of* and *to* is 600 and 400, respectively, then the second-order cue set for this corpus is {*the, for, to*}, and so on and so forth. Of course, this is an ideal scenario, since in most cases this equivalence is not possible. Consequently, the (15) and (16) lines in the pseudo-code are meant to adjust the number of words, X_i*_total,* that co-occur with the members of the cue set. According to (15) and (16), X_i*_total* can either be greater than or equal to the number of word types in the corpus such that if we remove the number of words that co-occur with the last member in the set from X_i*_total*, X_i*_total* will be smaller than the number of word types in the corpus.

To illustrate with a real corpus, this procedure was run on a random corpus that contained ≈ 110,000 tokens and ≈ 12350 words. Table 14 shows the 20 most frequent words in the corpus and the number of words that co-occurred with each word. It is clear from Table 14 that the sum of the words that co-occur with the first four most frequent words (i.e., 12353) is almost the same as the first order of the number of word types in this random corpus (i.e., 12350). This means that the first-order approximation of the cue set for this corpus is {*the, of, and, to*}. By the same token, the sum of the words that co-occur with the first twenty most frequent words (i.e., 24954) is close to the second order of the number of word types in the corpus (i.e., 24706). This means that the second-order approximation of the cue set for the same corpus is {*the, of, and, to, in, a, that, is, was, it, for, as, on, this, by, with, not, be, but, he*}. It is expected that higher orders of approximation should give more fine-grained information about the distributional properties of the words in the corpus.

Rank	Word	Frequency	Co-occ.
1	the	8705	4010
2	of	4220	3000
3	and	3055	3110
4	to	3049	2233 ←
		Sum[1-4**]**	**12353**
5	in	2629	2030
6	a	2190	1610
7	that	1394	1105
8	is	1160	883
9	was	1090	913
10	it	969	476
11	for	967	943
12	as	847	705
13	on	818	741
14	this	675	456
15	by	664	714
16	with	613	646
17	not	586	377
18	be	548	405
19	but	522	283
20	he	522	314 ←
		Sum[1-20**]**	**24954**

Table 14: The 20 most frequent words in a random corpus in a descending frequency order

Similar approximations can be established using phonemes, stress, etc. Once a cue approximation has been established, it can be used accordingly in different tasks such as segmentation, categorization, constituency detection, and frame identification.

The way this set of cues is used in this dissertation is that it provides the coordinates for the distribution of words in the corpus. This means that the distributional behavior of a word is captured in terms of its co-occurrence with the members of the cue set. Words can then be categorized according to the strength of their association with these cues, as shown in the following section.

5.3 Cued Distributional Similarity

The central concepts in distributional learning are those of distribution, equivalence, and substitutability. These and other related concepts are defined as follows (Harris, 1951):

(41) Distribution

The distribution of an element is the total of all environments in which it occurs, i.e., the sum of all the (different) positions (or occurrences) of an element relative to the occurrence of other elements.

(42) Equivalence

Two utterances or features will be said to be linguistically, descriptively, or distributionally equivalent if they are identical as to their linguistic elements and the distributional relations among these elements.

(43) Substitutability

If all occurrences of word A can be replaced by word B, without loss of syntactic well-formedness, then they share the same syntactic category

(44) Substitution Class

A substitution class is the class of elements which are free variants of each other.

(45) Environment

The environment of an element *E* consists of the elements, within an utterance, before, after, and simultaneous with *E*.

(46) Utterance

An utterance is any stretch of talk, by one person, before and after which there is silence on the part of the person.

An ideal distributional learner should reach word classification and constituency decisions such that all the words in a particular class will have identical distribution. Accordingly, considerable descriptive economy would be achieved if we could replace statements about individual elements with identical distributions by a single statement, applying to a whole set of distributionally similar elements (Harris, 1951: 243). However, such a learner is not distributionally realizable because elements which have precisely the same total environments are not frequently available in a given corpus. This means, Harris (1951: 244) argues, that "if we seek to form classes of words such that all the words in a particular class will have identical distributions, we will frequently achieve little successes".

Given this, Harris proposes what is called here an *Approximate Distributional Learner (ADL)*. The main idea in a distributional approximation procedure is the grouping together of elements which are identical in respect to some stated large fraction of all their environments. This means grouping takes place under distributional similarity and not equivalence. To perform this approximation, we take each element and state all of its environments within the corpus, where the environment is taken to be the whole utterance in which it occurs. We then select one element, and match its range of environments with that of the other element. We do not expect to find many cases of identical ranges, but decide instead upon certain conditions; if an element satisfies these conditions, it will be assigned to the class of our originally selected element.

The conditions may vary with the language system and with our purposes. They may be as crude as requiring that 80% of the environments of the one element should be ones in which the other also occurs (Harris, 1951). In many languages, the case might be that some classes of words are more easily set up first, and others being set up with their aid. Accordingly, we begin with these most frequently occurring elements whose number seems to be small.

The results of each classification can be used in all subsequent classifications. If two elements e_1 and e_2 had been previously assigned to one class E, we would henceforth replace them by that class mark each time they occur. In effect, the occurrence of each class is defined in respect to the occurrence of every other class, rather than defining each element in respect to the occurrence of every other element.

To summarize, a Harrisian approximate distributional learner has the following properties:

(47) ***Properties of an Approximate Distributional Learner (ADL)***

a. Grammatical categories/constituents can be discovered on the basis of distributional relations among linguistic elements.

b. The distributional similarity among elements can be approximated in respect to some stated large fraction of their distributions.

c. The occurrence of each class is defined with respect to the occurrence of every other class.

d. The order of discovery may be language-specific.

Below, I show how this basic learner can be modified to constitute the core of a cue-based distributional learner. Properties (a) and (d) above are assumed here without further discussion. The other properties are further accentuated in terms of a distributional definition of cues.

Given the two procedures for cue extraction introduced above, the distributional similarity among elements can now be approximated in terms of their cued distributions. The cued distribution of a linguistic element is captured in terms of its co-occurrence with one or more cues in the set of Category Cues. That is, two elements are distributionally similar if they have similar cued distributions. These distributions are used in the two proposed models in the following manner.

5.3.1 Relative Frequency

In the first model of cue-based learning which is based on the semantic procedure for cue learning, the distributional similarity of two elements is approximated in terms of their relative frequency in the same cued bigrams. The intuition behind this measure is that: If an element *E* occurs in context *t* more than half of the times *E* occurs in the corpus, *t* most probably represents the distributional signature of *E*. So, if two elements occur more than 50% of the times in the same cued bigram, these two elements are most likely to be distributionally similar. In pseudo-code, this intuition works as follows:

(48) Cue-Based Distributional Similarity

```
FUNCTION D-SIM($E_i$, $E_j$)
for i := 1 to n
        get frequency f($E_i$), f($E_j$), …f($E_n$) ;
for C∈ K, K := {Cues} ;
        C immediately precedes $E_i$ ;
        C immediately precedes $E_j$ ;
        get frequency f(C, $E_i$) ;
        get frequency f(C, $E_j$) ;
        if $\frac{f(C,E_i)}{f(E_i)} > 0.5$, $\frac{f(C,E_j)}{f(E_j)} > 0.5$;
        then D-SIM($E_i$, $E_j$) ;
else
fail.
```

5.3.2 Mutual Information

In the second model, which is based on the distributional procedure for learning cues, distributional similarity is captured in term of the information-theoretic statistic of Mutual Information (MI). MI has been defined in two different yet related ways. The first definition interprets MI as the amount of information provided by the occurrence of one event, *y*, about the occurrence of another event, *x* (Fano 1961: 27). Interpreted this way, the information provided by *y* about *x* consists of changing the probability of *x* from the *a priori* value *P(x)* before the occurrence of *y*, to the *a posterior* value *P(x|y)*, given that *y* has occurred. The measure of change of probability is the logarithm of the ratio of the *a posteriori* probability and the *a priori* probability. Thus, the mutual information between *x* and *y*, which is sometimes referred to as *point-wise mutual information,* is defined as follows (Fano 1961: 27-28; Reza 1961: 140; and Jelinek 1968: 150):

(49)
$$I(x;y) = \log_2 \frac{P(x\,|\,y)}{P(x)}$$
$$= \log_2 \frac{P(x\,|\,y)P(y)}{P(x)P(y)}$$
$$= \log_2 \frac{P(x,y)}{P(x)P(y)}$$

where *P(x,y)* is the probability of observing *x* and *y* together, *P(x)* and *P(y)* are the probabilities of observing *x* and *y* anywhere in the corpus. If *x* and *y* tend to occur in conjunction, their mutual information will be high. If they are not related and co-occur only by chance, their mutual information will be zero. Finally, if the two events tend to 'avoid' each other, their mutual information will be negative.

Another interpretation of mutual information can be given by obtaining the average of the mutual information per event pairs (Fano 1961: 46; Reza 1961: 105; and Cover and Thomas, 1991: 5, 18). Thus, for two random variables, *X* and *Y*, the average value of mutual information provided by Y about X is defined as

$$(50) \quad I(X;Y) = \sum_{X,Y} p(x,y) I(x;y)$$

$$= \sum_{X,Y} p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$$

The average mutual information in (50) can be re-written in terms of entropy. Entropy is the average uncertainty of a single random variable, *X*, and is defined by (51). Mutual information in these terms then is the reduction in uncertainty of one random variable, *X*, due to knowing about another, *Y*, (52).

$$(51) \quad \mathrm{H(X)} = -\sum p(x) \log_2 p(x)$$

$$(52) \quad I(X;Y) = H(X) - H(X|Y)$$

Unlike point-wise mutual information, average mutual information is always a non-negative number.[21]

These two forms of mutual information, both in (49) and (50-52), have been applied successfully to some problems in natural language processing. Research in speech recognition (Jelinek 1985), noun classification (Hindle 1988), constituent boundary identification (Magerman and Marcus 1990), phrase-structure grammars (Clark 2001), and morphological analysis (Cavar et al. 2004), among other areas, have shown that the mutual information statistic provides a wealth of information for solving these and other problems in natural language processing. Magerman and Marcus (1990) and

[21] For further discussion of the mathematical properties of Mutual Information, see the references mentioned in this section.

Clark (2001), for example, have shown that mutual information between words that belong to the same constituent is higher than that between words that do not make up a constituent. Using this criterion, true constituents can be distinguished from distituents, and consequently true phrase boundaries can be easily identified.

However, two interrelated issues have been raised regarding the difference between these two forms of mutual information and how they are used in the computational-linguistic research (e.g., Manning and Schütze 2003: 178-183). The first is that, Manning and Schütze argue, mutual information should be defined as holding between random variables, as defined by average mutual information, not values of random variables, as defined by point-wise mutual information. The other issue is that point-wise mutual information is biased in favor of low-frequency events (Manning and Schütze 2003: 182; Fontenelle et al. 1994: 72). Consequently, other things being equal, bigrams composed of low-frequency words, for example, will receive a higher mutual information score than bigrams composed of high-frequency words. That is the opposite of what we would want a good measure to do since higher frequency means more evidence and we would prefer a higher rank for bigrams for whose interestingness we have more evidence (Manning and Schütze 2003: 182).

As far as this dissertation is concerned, these two issues do not present any problems to how mutual information is used in the proposed cue-based model, or the results that have been achieved based on it. This is due to two reasons. The first is that mutual information is used to measure the strength of association between a set of cues and other elements in a corpus. Given that these cues are a subset of high-frequency elements in the corpus, each of the bigrams of interest in cue-based learning has at least one high-frequency element as a member. Consequently, the issue of the bias to low-frequency events is obviated. The other reason is that the distributional similarity between elements is not established on the basis of the mutual information between one element and another. Rather it is based on the aggregate mutual information between an element and all the cues with which it co-occurs. In other words, the decision of distributional similarity is not based on rare or single events. As a by-product of using mutual information this way, the bias in favor of low frequency event should not affect the result in any significant manner.

Having said that, only the point-wise mutual information as given in (49) was implemented in this dissertation. It is expected that the same results can be attained using either form of mutual information in (49-52), or any other association measure.[22] However, the plausibility of this claim is not tested in this study and is left for future investigation.

To summarize, the way point-wise mutual information is used here is to measure the distributional similarity of linguistic elements in terms of their mutual information with cues identified by the distributional procedure in (38). Using mutual information this way would result in a more gradient and probabilistic model of distributional learning than the relative frequency method in (48). The details of how this measure is implemented in cue-based learning are given in Chapter (7).

[22] See Oakes 1998 for a review of these measures.

5.4 Cue-Based Distributional Frame Identification

The frame identification component in the two models is subject to the following restrictions:

(i) There is no predefined set of possible frames.

(ii) The learning algorithms are left to learn what is an appropriate frame for a verb based on distributional regularities in the input.

(iii) Only cued contexts are visible to the frame identification procedure.

In more specific terms, no *a priori* knowledge is given to the learning mechanism regarding the number and structure of possible frames. Rather, these two pieces of frame information should be learned from the contexts where distributional similarity has been established in accordance with the two procedures in the previous section. In other words, the only information that is accessible or visible to the frame identification procedure is that information yielded by any of the two procedures in subsections (2.3.1) and (2.3.2). For this enterprise to be efficient and consistent, it should build on a cue-based distributional criterion of what constitutes a subcategorization frame.

The criterion introduced here for frame identification is a trivial specification of the distributional cue-based criterion for cue extraction introduced in (38) above. Accordingly, the set of possible frames in a given input is expected to be a subset of the contexts where predicates, in general, and verbs, in particular, occur. The set of possible frame cues can then be approximated as follows:

(53) Frame Cues
Let $P = \{p_1,\ldots, p_n\}$ be the set of possible predicates in a corpus R,
Let $C = \{c_1,\ldots,c_m\}$ be the set of contexts where the members of P occur, then the set of possible frames in R is the smallest subset, C_f, of C such that every predicate in P occurs at least once in at least one context in C_f.

This set of cues can then be extracted using either of the two methods in (39) and (40), as will be shown in the following chapters.

5.5 Conclusion

This chapter introduced the formal foundations of a distributional cue-based model of learning. Cues and frames have been formally defined. In addition, criteria and procedures have been presented for identifying two sets of cues in a given corpus. The first set was based on the semantic properties of some elements in the input that make them easy to identify. The other set of cues was purely based on the distributional properties of high-frequency words in a corpus. It has been proposed that each of these sets of cues can be used to realize a cue-based learner. The details of these two learners are given in the following chapters.

Chapter 6

Implementation [1]:

A Semantically Bootstrapped Cue-Based Learner (CBL-1)

6.0 An Outline

This chapter presents a Semantically Bootstrapped Cue-based Learner (CBL-1, henceforth). This learner is based on two main learning procedures. The first is the Semantic Procedure for Cue Extraction, and the other is the Cue-Based Distributional Similarity Procedure using relative frequency. This learner is semantically bootstrapped, yet it learns distributionally. This means that semantic cues are only used to update the initial state (S_0) of the learner to one intermediate learning state (S_1). Once the learner is in this intermediate state, semantic cues are not triggered any more, and it proceeds from one state to the other based on the distributional properties of the input, until it reaches the final state (S_n). This final state represents learning some *grammar* of the input language. This learner is diagrammed as follows:

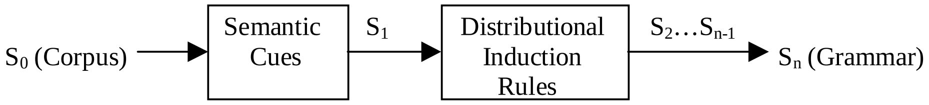


Figure 3: A Semantically Bootstrapped Cue-Based Learner (CBL-1)

This system is fleshed out with procedures for NP detection and verb identification which constitute the basic information required for the automatic acquisition of verbal subcategorization frames.

CBL-1 is data driven and bootstraps from semantic cues that can be easily identified in the input without much *a priori* linguistic knowledge. It assumes an input marked for word and sentence boundaries, and is minimally annotated for nominal expressions, which represent the seed knowledge for the learning procedures. These expressions are limited to a subset of unambiguous pronouns (*i.e., I, he, she, it, we, they, me, him, us,* and *them*) and names of things and people in the corpus. Names were orthographically annotated in the original corpus by capitalizing their first letter. Upper case was not used anywhere else in the corpus.

Restricting the initial knowledge to these nominal expressions is based on the well-known phenomenon that nominal expressions in general are acquired prior to adjectives and verbs (predicates), because while predicates logically presuppose arguments, the reverse is not true (Lenneberg 1967; Gentner 1978, 1982; Markman 1989; Fisher 1992).[23] Based on their longitudinal study of 30 children, Bates et al. (1988) report that at age 20 month, nouns were dominant (46.8% of total vocabulary) compared to verbs (8.3%) and adjectives (7.5%).[24] The early primacy of nouns has also been reported for other languages, including German, Mandarin Chinese, Turkish, and Japanese.[25]

Limiting pronouns to the subset above is based on the fact that these pronouns constitute single-word noun phrases, compared to other pronouns which can constitute single NPs, or be part of a larger NP (e.g., *her* and *his*), which makes them easier to identify in the input. However, the noun-primacy justification of pronominal single-word NPs as bootstraps is used here with a proviso. Given that a noun in this context refers to a specific instance of an object, information that is usually manifested in a prenominal article, demonstrative, or quantifier, it is plausible to think of this noun as a simplified or impoverished NP, given the apparent absence of the function words in the early lexicon,

[23] For a different explanation see Gasser and Smith 1998.

[24] As it was mentioned in footnote 21, these numbers add up to only 62.6%. The authors did not mention anything about the remaining 37.4%.

[25] However, Choi and Gopnik (1993) show that this is not the case in Korean.

which are usually used to build full-fledged NPs. This amounts to claiming that an instance of an object is the equivalent of an NP, and that a NOUN is an abstraction over these NPs/instances. Accordingly, the term NP is used throughout this discussion to refer to an instance of an object, be it a thing or a person. The significance of this point to how the algorithm proceeds is that (i) if these nouns are interpreted as a simple category, then the algorithm would start with detecting more nouns in the input, but (ii) if they are interpreted as simplified noun phrases, the next step would be to detect more NPs, and not nouns. I will assume the second possibility without further discussion. This means that the algorithm proceeds accordingly and detects more NPs first then identifies nouns in a top-down fashion, as will be detailed below.

Given this, these initial NPs provide a priorly learned vocabulary of nominal items to bootstrap verb learning. Accordingly, the algorithm uses the distributional properties of these NPs to detect the maximum number of other NPs in the input before moving to verb detection. The logic behind this is that the early identification of more NPs should result in more NP-cued contexts which would facilitate verb detection.[26] The distributional regularities of these NPs are then used to identify the first examples of verbs in the input. The structural regularities of the contexts of these verbs are then used to identify potential frames. These frames are then used to construct induction rules to infer new NPs and verbs. This NP-Verb-Frame cycle (Figure 4) applies iteratively until no further knowledge can be gained.

[26] In theory, these initial NPs could be enriched with binary semantic properties of nominal expressions such as animate-inanimate, human-nonhuman, and count-mass which would significantly facilitate verb detection, among other things. However, the idea here is to limit the initial knowledge to structural information and to the bare minimum needed for a successful bootstrapping.

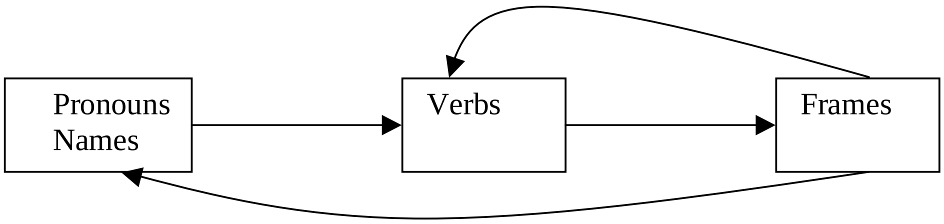


Figure 4. The Learning Cycle in CBL-1

Unlike other strategies, CBL-1 gives frames a more dynamical role in grammar induction. That is, once a frame is identified in the input, it is a part of the induction rules for the identification of more NPs and verbs, and consequently more frames.

6.1 Corpus Description and Pre-processing

The algorithm used the CHILDES database of the child Peter in Bloom (1970). Only the adults' transcribed speech was used. The corpus contained 25148 lines, 156646 tokens, and 3086 words. The corpus was minimally preprocessed in the following manner. Utterances were classified into three main types: Declarative (*UD*), Interrogative (*UQ*), and Exclamatory (*UE*), using the punctuation marks provided in the transcribed speech as approximations. That is, utterances that ended in a period '.', were marked as *UD*, utterances ending in an exclamation mark '!", were marked as *UE*, and utterances ending in a question mark '?', were marked as *UQ*. Figure 5 shows some representative utterances from the corpus after preprocessing.

oh my goodness UE
do you want me to do it UQ
there you go UD

Figure 5: Sample utterances from the CHILDES corpus marked for sentence type

Below I present a detailed description of the three main components of the algorithm above i.e., NP, V, and frame identification, and how distributional induction rules are used to move from one learning phase to the other. The outcome of every phase is used to build a partial grammar (including a lexicon) for the input language. Every time new information is inferred, the grammar, and consequently the corresponding lexicon, is updated to incorporate this information and use it for further learning.

6.2 NP Identification

The NP identification component of the learner starts with initial NPs, which include only a subset of pronouns and names, for the reasons given above. The first step in learning was to replace these expressions in the corpus with a single symbol, i.e., NP, as in (54), where CAP stands for words starting with a capital letter. The application of (54) resulted in replacing 26027 tokens in the corpus as single-word NPs.

(54) {I | he | she | it | we | they | you | me | him | us | them | CAP} → NP

The immediate contexts of these NPs were used to identify the environments where NPs are highly expected to occur, in order to infer new NPs. Accordingly, these NP-cued contexts were brought to bear in the construction of the NP Identification Procedures in (55). The first procedure is based on the Cue-Based Procedure for Distributional Similarity introduced in (48) above. Accordingly, if a word X is followed by an NP more than half of X’s occurrences in the input, it is most probable that the most frequent untagged word following X would indicate the left boundary of another NP. The second procedure detects the right boundary of a single-word NP output by the first procedure. This procedure is implemented as *a demon* which is automatically triggered when particular condition(s) occur (Nilsson, 1982). Whenever a new piece of knowledge is input, this demon (depending on the nature of the particular piece of knowledge) would activate additional pieces of knowledge by applying their sets of inference rules to it. These new pieces in turn might result in more demons being activated.

(55) NP Identification Procedures

```
a.    FUNCTION INITIALIZE-NP(Y, ‘NP[Y’)
      for i := 1 to n,
      get frequency $f(X_i)$ ;
      for every $X_i$, NP immediately follows $X_i$
      get frequency $af_i := f(X_i, NP)$;
            if $af_i < 5$[27] ;
            fail
            else
      get frequency $rf_i := \frac{af_i}{f(X_i)}$;
      if $0.5 < rf_i < 1$;
            for every $X_i$, $Y_i$ immediately follows $X_i$
            get frequency $f(X_i, Y_i)$
            build a decreasing frequency profile $F_i := \{f(X_i, Y_i) > f(X_i, Y_j),\ldots\}$ ;
                  if there is $Y_i$ ;
                        $Y_i$ most frequent in $F_i$ ;
                        $Yi$ most frequent in $F_j$ ;
                  then  INITIALIZE-NP(Y, ‘NP[Y’);
                  else
                  fail
      else
      fail

b.    FUNCTION ‘NP[Y D → NP[Y] D’
      if NP[Y, D,
      D := NP,
            or
      D := sentence-boundary,
      then
      NP[Y D → NP[Y] D .
```

The two procedures were run on the corpus only twice. This number was based on a pilot experiment that showed that no new NP-initializers were identified after the

[27] This threshold on bigram frequency was meant to reduce the effect of transcription errors in the corpus, so that the decision would not be based on rare events that could have resulted from these errors.

second run. Every word X was used only once in inducing NP initializers. That is, if a word X was used in the first induction run, it would be excluded from the second run.

In the first run of the procedure, 824 XNP bigrams were extracted. The bigrams that occurred less than 5 times[28] were excluded, which reduced the number to 242 bigrams. These bigrams were further reduced by the relative frequency requirement to only 56 bigrams, which were used in NP induction. Table 15 shows some of these bigrams and how they were used in induction. In Table 15, the first column shows some of the extracted bigrams, the second and third columns give the absolute frequencies of Xs and the XNP bigrams, the fourth column represents the relative frequency of the XNP bigrams, the fifth column shows the most frequent bigram of X followed by any untagged word Y, and the last column gives the potential NP initializers, using the information in the fifth column.

X NP	**F(X)**	**F(XNP)**	$\frac{f(XNP)}{f(X)}$	**Most Frequent XY Bigram**	**Potential NP Initializer**
what're NP	205	202	0.985	what're the	the
if NP	307	284	0.925	if this	this
let NP	250	227	0.908	let her	her
show NP	97	83	0.856	show her/mommy	her, mommy
gave NP	26	22	0.846	gave her	her
where'd NP	42	36	0.857	where'd that	that
did NP	728	560	0.769	did the	the
give NP	171	129	0.754	give the	the
when NP	196	144	0.734	when that/did	that, did
throw NP	102	61	0.598	throw that	that

Table 15: Some of the bigrams extracted by the NP procedure in (55).

[28] This threshold was necessary because of some transcription errors in the corpus.

Given this information, NP induction applies in the following fashion. For every X, there is an n-number of untagged words Ys that follow X. This means that, for every X there is an n-number of XY bigrams equal to the number of Ys. Every bigram (X_iY_i) has a score S that should be greater than 1 for this bigram to be used in NP induction. The scores of all the XY bigrams for every X were sorted descendingly, and the bigrams with the highest scores were collected. If there is a Y_i such that Y_i is a member in more than one most-frequent bigram, and that the sum of the scores of these bigrams are ≥ 5, then Y_i is interpreted as an NP-initializer.

Applying this part of the procedure resulted in identifying the words *the*, *that*, and *her* as NP-initializers. The scores of these words as the second members of most-frequent bigrams were 12, 11, and 6, respectively. The words *this*, *mommy*, and *did* were also among the members of the set of potential NP initializers, yet they were excluded because they did not have the threshold frequency required by the procedure in (55). Accordingly, whenever any of the words *the*, *that*, or *her* was found in the input an NP was initialized, as in (56). This resulted in marking the left boundary of 9765 NPs. The right boundary of these NPs were identified using the rules in (57), (58), and (59), based on the procedure in (55 b).

(56) a. …the… → …NP[the…
b. …that…→ …NP[that…
c. …her…→ …NP[her…

(57) NP[Y Delimiter → NP[Y] Delimiter; Delimiter is another tag (so far NP) or a sentence/clause boundary.

(58) NP[Y W Delimiter → NP[Y W] Delimiter; if Y does not occur alone between delimiters in the input, i.e., *the*.

(59) NP[X W Delimiter → NP[X W] Delimiter; if W is a second member in the NPs in (58) and X occurs both alone between delimiters as a single-word NP, or as an NP-initializer i.e., *that* and *her*.

These newly identified NPs and the initial NPs (pronouns and names) were then used in the second run of the procedure in (55). All the bigrams that were used in NP induction in the first run were excluded in the second run. This resulted in identifying the words *this, a,* and *your* as NP-initializers. Applying a rule similar to the rule in (56) resulted in identifying the left boundary of 5041 NPs. The right boundary of these NPs was identified using the rules in (57), (58), and (59).

Thus far, 40833 NP tokens have been identified; 26027 of them are pronouns and names, and 14808 were induced using the procedure in (55). These NPs included single as well as double-word NP structures, as given in (60). The emergence of the latter form of NPs can be interpreted as an indicator of the nature of nominal Phrase Structure Rules (PSRs) in English.

(60)a. NP_1 → Word
Word → {I | he | she | it | we | they | you | me | him | us | them | CAP | her | this | that}
b. NP_2 → Initializer Word
Initializer → {the | your | a | her | that |this}
Word → {car | bag | house | box | train | horse | paper | pen | pencil | cow | floor | chair}

It is important to mention here that this strategy of identifying these function words as NP markers is similar to other strategies that were proposed to use function words as reliable cues to constituent structure (e.g., Bever 1970, Fodor & Garrett 1967, Kimball 1973, Watt 1970b, and Clark & Clark 1977). For example, Kimball (1973) has proposed the following strategy: Whenever you find a function word, begin a new constituent larger than one word. More specific strategies were proposed by Clark & Clark (1977), including the following strategy: Whenever you find a determiner (*a, an, the*) or quantifier (*some, all, many, two six, etc.*), begin a new noun phrase (NP). In spite of this ultimate similarity, there is a significant difference between these strategies and the strategy proposed here. While these two strategies are seeded with initial knowledge of function words, the procedure in (55) and the related rules in (56) through (59) inferred this knowledge distributionally. This means that these distributional strategies are able to derive this knowledge and capture the frequency effect of function words without assuming any *a priori* knowledge about these words.

Given the double-word NPs in (60b), two categories were easily differentiated in the input in a top-down fashion, i.e., Determiners and Nouns. (The labels can be anything. These labels are used here for convenience). Determiners are just what have been termed NP-initializers, and Nouns are just the second elements in a double-word NP (i.e., NP[Initializer Word]. This resulted in identifying 640 words with a total of 5208 tokens as Nouns. (See Appendix A for a full list of the nouns identified). These potential nouns and determiners were added to the lexicon which initially included pronouns and names. New tokens of these words were identified using the simple procedure that any instance of these 662 words was tagged as Noun if it occurred immediately after one of the Determiners *the, a,* and *your* in the corpus. These determiners were used because, as it was mentioned above, there was no evidence in the corpus that they could occur alone.

Applying this procedure increased the number of the identified noun tokens to 8445, thus providing an accurate strategy to consume as many input tokens as possible as Nouns. In more specific terms, the NP induction procedures above consumed 14642 tokens of the input either as nouns or determiners. These tokens plus the initial pronouns and names represented ≈ 9% of the 156646-token input. The assumption is that this 9%

would serve as a sufficient bootstrap of predicate identification and frame detection, as will be shown below.

6.3 Predicate Identification

It was observed that after the identification of more NPs, the input did not contain sentences or utterances of the form [NP NP NP], yet there were few examples of utterances of the form [NP NP]. The fact that this and other sequences are not attested in the input could be attributed to two different yet equally possible reasons. The first is that a sequence is ungrammatical; therefore it does not occur in the input. The other is that this sequence is grammatical but does not exist in the particular corpus used in the learning process. In this dissertation, the null hypothesis is that if a certain distributional sequence is not attested in the input, this sequence should be ungrammatical until there is enough distributional evidence in the input supporting the other possibility. In case this evidence is available, the learning procedures should be able to perform accordingly. Consequently, it was concluded that the absence of utterances of the form [NP NP NP] should constitute a distributionally-driven constraint on possible utterance forms in English.

6.3.1 Predicate Identification Procedure

Assuming this constraint, if an utterance is composed of two NPs and a single variable, this variable must be a predicate of some sort in order to avoid the [NP NP NP] sequence that was not attested in the input. This means that in the following configurations, X must be assigned a non-nominal value.

[NP X NP]
[X NP NP]
[NP NP X]

What is interesting about using these configurations in predicate induction is that these inference configurations are neither sensitive to word order nor to case, which gives this procedure of predicate identification a language-independent flavor. Moreover, the distributional basis of this induction has some supporting evidence from language acquisition studies. Landau & Stecker (1990) present intriguing evidence that young children interpret a novel word as a semantic predicate if it appears with NP arguments. This finding is consistent with the notion that a sentence, partially represented as a structure containing NP arguments, can serve as a quite general analog of its semantic predicate/argument structure (Fisher 1996). The generality of this procedure is an advantage for the theory of the acquisition of predicate terms. Not all languages have distinct categories of prepositions and predicate adjectives, but may instead use verbs to convey spatial or attribute meanings (e.g., Croft 1990).

However, this process of variable evaluation is not unconstrained. That is, the utterances/sentences that were used as inference seeds were limited to the declarative type (UD), on the basis of the well-attested psycholinguistic finding that declarative sentences are acquired prior to other sentence types. Building on this, the first phase of the Predicate Identification Procedure is given in (61). Applying (61) to the corpus resulted in identifying the 57 words in (62) totaling 15016 tokens as potential predicates.

(61) **Predicate Identification Procedure**

a. **If** [NP X NP]UD,
[X NP NP]UD, or
[NP NP X]UD,
Then tag X as Predicate.

b. Tag as Predicate all the instances of X in the contexts X NP and NP X.

(62) **Predicates**

[and beg bring brought changed closed closing did do drew dropped eat finished fixed found get give got guess has have heard helped hit hold hurts is knows know like lost made make mean misunderstood move m needs need open pull saw says scared see show take tape tell thank think throw understand want was wiping]

6.3.2 Verb Identification Procedure

It is clear from the set of predicates in (62) that, except for '*and*', all its members can function as verbs. Moreover, some of these verb forms also function as predicative adjectives in the corpus (e.g., *finished, fixed, lost,* and *scared*). Given that these predicates are the *inference base* that would be utilized later in inducing new knowledge about NP arguments and predicates, it is essential to sub-classify them into more fine-grained categories in order to avoid any future errors in tagging/parsing.

The procedure that was used in differentiating verbs from other predicates is based on an assumption similar to that used in the NP Identification Procedure in (55). This procedure works as formulated in (63). Applying this procedure to the set of predicates in (62) resulted in subcategorizing the 48-member subset of the predicates in (64) as verbs.

(63) **Verb Identification Procedure**

If word X is identified as a predicate (P), X is subcategorized as a verb iff

$$\frac{frequency(X = P)}{frequency(X)} > .5$$

(64) **Verbs**

[beg bring brought changed closing did do drew dropped eat found get give got guess has have heard helped hit hold hurts is knows know like made mean misunderstood move m needs need open pull saw says see show take tell thank think throw want was wiping]

Comparing the subset in (64) to the set of predicates in (62), it can be observed that this verb subset excludes the predicates (*and*, *closed, finished, fixed, lost, make, scared, tape,* and *understand*) which include predicates that function as verbs, among other things. However, this does not mean that these predicates are excluded altogether from the set of potential verbs, it rather means that these predicates are excluded from the set of verbs that serve as an inference base, and that the verb status of these predicates would be established at a later phase in induction. On one hand, excluding these predicates from the induction base decreased the cardinality of the induction set, yet on the other hand, it helped, at an early stage in grammar induction, in avoiding future overgeneralizations. This information about verbs was added to the lexicon that has been constructed so far. (See Appendix B for a list of the verbs identified.)

6.4 Frame Identification

Having identified some verbs, the next step was to utilize the distributional regularity of these verbs to capture structural patterns in the input. These patterns would facilitate identifying verbal frame behavior. The set of possible frames in the input was defined by (53) in Section 5.4 in the previous part, repeated below for ease of reference:

> **(53) Cue-Based Frame Definition**
> Let $P = \{p_1,\ldots, p_n\}$ be the set of possible predicates in a corpus R
> Let $C = \{c_1,\ldots,c_m\}$ be the set of contexts where the members of P occur, then the set of possible frames in R is the smallest subset, C_f, of C such that every predicate in P occurs at least once in at least one context in C_f.

6.4.1 Initial Frame Identification

As a result of NP and verb identification more input was tagged and became visible to the learning procedure, and consequently fully or almost fully labeled utterances/sentences started to emerge. Some of these sentences in the corpus are illustrated in (65) below. The emergence of these sentences signaled the emergence of structural patterns in the input, which is a clear precursor of sentential phrase structure rules (PSRs). Accordingly, these patterns were used to extract the primary PSRs in (66). These rules represent possible sentence structures in English. This means that the grammar constructed so far includes a lexicon of Nouns, Predicates, and Verbs, in addition to PSRs describing possible NP and sentence structures.

(65) a. [NP[it] V[hurts]] UD
b. [V[open] NP[it]] UD
c. [NP[I] V[found] NP[it]] UD
d. [NP[I] V[guess] NP[it] V[is]] UD
e. [NP[you] V[have] to V[open] NP[Det[the] N[trunk]]] UD
f. [NP[I] V[want] NP[you] to V[get] NP[it]] UD
g. [V[bring] NP[it] to NP[me]] UD
h. [NP[you] V[give] NP[me] NP[that]] UD

(66) a. S → NP V
b. S → V NP
c. S → NP V NP
d. S → NP V S
e. S → NP V to V NP
f. S → NP V NP to V NP
g. S → V NP to NP
h. S → NP V NP NP

	F_0	F_1	F_2	F_3	F_4	F_5	F_6
V	_	_NP	_NP S	_ to|ta V	_NP to V	_NP to NP	_NP NP
beg	0	1	0	0	0	0	0
bring	0	1	0	0	0	1	1
brought	0	1	0	0	0	1	1
changed	0	1	0	0	0	0	0
closing	0	1	0	0	0	0	0
did	1	1	0	0	0	0	0
do	1	1	0	0	0	0	0
drew	0	1	0	0	0	0	0
dropped	0	1	0	0	0	0	0
eat	0	1	0	0	0	0	0
found	0	1	0	0	0	0	0
get	0	1	0	0	0	0	0
give	0	1	0	0	0	1	1
got	0	1	0	0	0	0	0
guess	1	1	1	0	0	0	0
has	0	1	0	0	0	0	0
have	0	1	0	1	0	0	0
heard	0	1	0	0	0	0	0
helped	0	1	0	0	0	0	0
hit	0	1	0	0	0	1	0
hold	0	1	0	0	0	0	0
hurts	0	1	0	0	0	0	0
is	0	1	0	0	0	0	0
knows	0	1	0	0	0	0	0
know	1	1	1	0	0	0	0
like	0	1	0	1	1	0	0
made	0	1	0	0	0	0	0
mean	0	1	1	1	0	0	0
misunderstood	0	1	0	0	0	0	0
needs	0	1	0	0	0	0	0
need	0	1	0	1	1	0	0
saw	0	1	0	0	0	0	0
says	1	1	0	0	0	0	0
see	1	1	0	0	0	1	0
show	0	1	0	0	0	1	1
take	0	1	0	0	0	1	0
tell	0	1	0	0	1	0	1
think	0	1	1	0	0	0	0
throw	0	1	0	0	0	1	1
want	1	1	0	1	1	0	0

Table 16: Initial Frames identified by CBL-1

Table 16 shows some verbs and their potential frame(s) that were identified using the procedure in (53). Accordingly, '1' under a frame was used to indicate the occurrence of the corresponding verb in this frame, and '0' to indicate its non-occurrence. An underscore '_' in a frame marks the position of the corresponding verb, and if nothing follows this underscore, this means that the corresponding verb is intransitive. An S in a frame, e.g. F_2**,** stands for any sentence structure in (66) and can be replaced by the corresponding structure on the left-hand side of the rule. The 'X|Y' in a frame means X or Y, as in F_3, where a verb can be followed either by *to* or *ta*[29].

This potential frame behavior was added to the lexical entries of the verbs identified so far. The examples in (67) show the primary lexical entries of some of these verbs, enriched with this frame information.

(67) a. bring: Pred, V, F_1(_ NP), F_5(_ NP to NP), F_6(_NP NP)
b. want: Pred, V, F_1(_NP), F_3(_ to V), F_4(_NP to V)
c. see: Pred, V, F_1(_NP)

The lexicon could be streamlined by using a redundancy rule to the effect that 'every verb is a predicate by default', hence the Pred part need not be mentioned in the lexical entry of every verb. The frame information in the lexicon could be further simplified, as given in (68), by adding the frames (*F*) and their structural realizations (*S)* to the lexicon as independent entries, and reducing this information in the lexical entries of the verbs to a pointer to the relevant frame entry.

[29] *ta* resulted from two different sources in the transcription of the corpus. The first was the result of transcribing sequences such as *want a pen* as *want ta pen*. The other was the result of transcribing sequences such as *want to play* as *want ta play*. Given that there were not many instances of the first case in the corpus, *ta* was used as another form of *to*.

(68) $F_\alpha : S_\alpha$

$F_\beta : S_\beta$

$F_\delta : S_\delta$

etc…

(69) $Vi : F_\alpha , F_\beta$, *etc…*

$V_j : F_\alpha , F_\delta$, *etc…*

$V_k : F_\beta$

This representation would result in a more compact and economical lexicon structure, since the frame information would be encoded only once in the frame entry.

6.4.2 Frame-based Induction: The First Phase

Having enriched the lexical entries of the verbs identified so far with their frame information, these frames were used in inducing new arguments as well as new predicates/verbs. The main idea behind using the verbal frame information in induction is that a structure S is the result of satisfying the frame requirements of the verbs involved. Accordingly, if a given structure S contains a variable R and the verbs V_1 and V_2, R is assigned a value/tag that satisfies the requirements of V_1 and V_2. For example, in the sentence S = [X V_1 NP V_2], where the lexical entries of V_1 and V_2 contain F_1 information, the variable X should be assigned an NP tag. Unlike the predicate/verb identification procedures, this induction procedure has no restriction on the sentence types input to the induction rules. This means that this input contains declarative as well as interrogative sentences. And like other induction procedures, the rules below prioritize NP induction over predicate/verb induction.

The first induction rule was based on F_1 in Table 14. This rule works, as mentioned in the example above, in the fashion laid by (70). If there exists a sentence/utterance in the input such that it contains one variable X in the position given in the left-hand side of the rule, and verbs that have transitivity information in their lexical entries, X should be re-written as NP. The input to this induction rule consumed sentences such as those exemplified by (71). Numerical subscripts were used to indicate the frame(s) of the verbs involved in the induction rule.

(70) [X V_1 NP V_1] → [NP V_1 NP V_1]
(71) a. What V[do] NP[you] V[see]] UQ
b. Who V[do] NP[you] V[see]] UQ
c. What V[did] NP[you] V[do]] UQ
d. What V[do] NP[you] V[want]]UQ

The application of (70) to the input distinguished the words *what, who,* and *how* as potential NPs. Other instances of these words were also tagged as potential NPs using the rules in (72), where the rule in (a) identifies the left boundary of these potential NPs, and (b) their right boundary.

(72) a. Delimiter (what|how|who) → Delimiter NP[(what|how|who) .
b. (what|how|who) Delimiter → (what|how|who)]NP Delimiter.

The second NP induction rule was based on F_4 and F_5, namely, *_NP to V* and *_NP to NP*, respectively. These two frames were used in induction in the following manner. Given the structure stretch S = …V X to.., where V is a verb that subcategorizes for either F_4 or F_5, X is re-written as an NP. Rule (73) formalizes this induction process. The dots in the rule are used to indicate elements that do not affect induction.

(73) …$V_{4|5}$ X to …→ … $V_{4|5}$ NP to …

This procedure identified the words *mama, mommy, keys, everyone, people, something, tape, someone, buns,* and *em* as potential NPs. To identify more instances of these words as potential NPs, delimiting rules similar to those in (72) were applied.

New verbs were identified using frames *F_2*, *F_3* and *F_4*, according to the rules in (74), (75), and (76), respectively. If X in the left-hand side of the rule was previously tagged as Predicate, this tag was transformed into the more specific tag, V, as it was the case with *make, tape,* and *understand*. The predicative status of these words is still preserved by the redundancy rule mentioned above. Applying these rules to the corpus yielded the 104 potential verbs in (77), raising the number of potential verbs identified so far from 48 to 152 verbs. Then (61b) of the Predicate Identification Procedure was applied in order to identify more tokens of these verbs. Accordingly, every instance of the verbs in (77), either untagged or previously tagged as Predicate, was also tagged as V if it immediately preceded or followed an NP. These new verbs and their tokens were then used as delimiters in the NP boundary identification demons.

(74) …V_3 (to | ta) X… → …V_3 (to | ta) V…

(75) …V_4 NP to X… → …V_4 NP to V…

(76) … V_2 NP X (] | NP | V) → … V_4 NP V (] | NP |V)

(77) **Newly Identified Verbs**

[am answer are ask be blow borrow break broke build change choose close come cry cut die does draw drink feel find fit fits fix go goes help hole hurt **just** lean learned leave left lick lift likes lock look make **now** opens pack park pat **pen pencil** pick **piece** pitch play pour pull push put re read ride roll said say screw set share sing sit sleep speak spill spilled spread spoil squeeze stand start stay stick stop swim talk tape taste tear threw tinkle trade try turn turned understand unscrew use wait wake walk wants wash waste watch wear will wipe woke work works write]

It can be noticed that with the exception of *just*, *now*, *pen*, *pencil*, and *piece*, all the potential verbs in (77) are actual verbs. By checking the contexts of the last three nouns in these erroneously tagged words, it was found that they occurred in contexts such as *want ta pen|pencil|piece*, which are transcriptions of *want a pen|pencil|piece*, hence the rule in (74) applied. Though this error resulted from transcription errors in the corpus, the mis-tagging of these words as verbs represents a serious weakness in this implementation, i.e., the absence of probabilistic category assignment in order to capture the probabilities of the membership of a given word in more than one category. This drawback is circumvented by the other implementation of cue-based learning as will be shown in the next chapter.

6.4.3 New Frames

As a result of identifying new verbs, new distributional regularities emerged, and consequently the Frame Identification Procedure in (53) was triggered. Remember that the input to this procedure is limited to declarative sentences/utterances. The new configurations where the new and old verbs occurred were both input to the procedure. The output of this procedure created two possibilities. The first was that the new configurations instantiated one or more of the existing frames, in this case frame pointers were added to the lexical entries of the verbs occurring in these configurations, and the relevant induction rules applied. For example, the verbs *try* and *wants* realized the old frame F_3, i.e., *_to|ta V, wants* occurred in F_4, i.e., *_NP to|ta V* as well, *help* instantiated F_2, i.e., *_NP V*, and *make* appeared in F_2 as well as F_6, i.e., *_NP NP*. The other possibility was that some or all of the new configurations did not realize any of the existing frames, in this case new frame entries were added to the lexicon, and pointers to these frames were inserted in the lexical entries of the relevant verbs. Table 17 shows the new potential frames that did not match any of the old frames, and some examples.

The frame results in Table 17 raise two important and interrelated issues. The first is that words such as *off, on, out,* and *up*, which can function as either prepositions or particles, were identified as part of the frames F_7 through F_{10} of the pertaining verbs. Some examples of these items in the corpus are given in (79).

(79) a. [V[turn] NP[it] off]] UD

b. [V[put] NP[it] on]] UD

c. [V[take] NP[it] out] UD

d. [V[pick] NP[it] up]] UD

The other issue is that the frames F_{11} through F_{19} as well contain elements that might be interpreted either as arguments or adjuncts to the corresponding verbs. This can be seen in the examples in (80), where the italicized parts indicate the relevant structures.

(80) a. [NP[I] V[want] ta *V[play] with NP[the N[blocks]N]NP*] UD

b. [NP[you] V[have] to *V[write] with NP[it]*] UD

c. [NP[we] V[have] to *V[take] NP with NP[us]*] UD

d. [NP[I] *V[think] if NP[you] V[look]* in NP[that bag NP[you] might V[find] NP[it] UD

e. [*V[ask] NP[Patsy] if NP[that] V[s]* NP[the N[donkey]N]] UD

f. [NP[I] *V[found] NP[it] for NP[you]*] UD

Again these two issues represent another weakness in this model, which is also circumvented in the other model by resorting to complement- and adjuncthood probabilities, as will be shown later.

	F_{7}	F_{8}	F_{9}	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	F_{15}	F_{16}	F_{17}	F_{18}	F_{19}
V	_np on	_np off	_np out	_np up	_to np	_with np	_ on np	_np with np	_np on np	_np in np	_if S	_np if S	_np for np
ask	0	0	0	0	0	0	0	0	0	0	0	1	0
close	0	0	0	1	0	0	0	0	0	0	0	0	1
come	0	0	0	0	0	0	1	0	0	0	0	0	0
cut	0	0	1	0	0	0	0	0	0	0	0	0	1
do	0	0	0	0	0	1	0	1	1	0	0	0	0
draw	1	0	0	0	0	1	0	0	1	0	0	0	0
drink	0	0	0	0	0	0	0	0	0	1	0	0	0
found	0	0	0	0	0	0	0	0	0	0	0	0	1
get	0	1	1	0	1	0	1	0	0	0	0	0	1
go	0	0	0	0	1	1	1	0	0	0	0	0	0
have	1	0	0	0	0	0	0	0	1	0	0	0	0
hold	1	0	0	0	0	0	0	0	0	0	0	0	1
leave	0	1	1	0	0	0	0	0	1	1	0	0	0
lift	0	0	0	1	0	0	0	0	0	0	0	0	0
made	0	0	0	0	0	0	0	0	0	0	0	0	1
pick	0	0	0	1	0	0	0	0	0	0	0	0	0
play	0	0	0	0	0	1	0	0	0	0	0	0	0
pull	0	1	1	0	0	0	0	0	0	0	0	0	0
put	1	0	1	0	0	0	1	0	1	1	0	0	0
ride	0	0	0	0	0	1	1	0	0	0	0	0	0
see	0	0	0	0	0	0	0	0	0	1	1	0	0
stand	0	0	0	1	0	0	0	0	0	0	0	0	0
take	0	1	1	0	0	0	0	1	0	1	0	0	0
tear	0	1	0	0	0	0	0	0	0	0	0	0	1
tell	0	0	0	0	0	0	0	0	0	0	0	1	0
thank	0	0	0	0	0	0	0	0	0	0	0	0	1
think	0	0	0	0	0	0	0	0	0	0	1	0	0
try	0	0	1	0	0	0	0	0	0	0	0	0	0
turn	1	1	0	0	0	0	0	0	0	0	0	0	0
want	1	0	0	0	0	0	0	0	1	0	0	0	0
wash	0	0	1	0	0	0	0	0	0	0	0	0	0
wipe	0	1	0	0	0	0	0	0	0	1	0	0	0
write	0	0	0	0	0	1	1	0	0	0	0	0	0

Table 17: Examples of the new frames identified by CBL-1

6.4.4 Frame-based Induction: The Second Phase

As for the verbs realizing any of the old frames, induction was carried out on the basis of the induction rules that were based on these frames. In the case of the verbs instantiating new frames, these frames were used to construct the induction rules in (81), in a fashion similar to that adopted with the old frames. The application of the rules in (81) distinguished the words in (82) as possible single-word NPs or Determiners.

(81) a. …$V_{7|8|9|10}$ X (on | off | out | up)…→… $V_{7|8|9|10}$ NP (on | off | out | up)…
b. …$V_{11|12|13}$ (to | with | on) X …→ …$V_{11|12|13}$ (to | with | on) NP[X …
c. …$V_{14|15|16|19}$ X|NP (with |on |in |for) NP|Y…→…V_{12} NP[X] (with |on |in |for) NP[Y…
d. …V_{17} if (X | NP) (Y | V)…→ …V_{17} if NP[X] V[Y]…
e. …V_{18} (X | NP) if (Y | NP) (Z | V)…→ …V_{18} NP[X] if NP[Y] V[Z]…

(82) **New Potential NPs and Determiners**
[an all bed bologna both everything gas home milk my one paper sauerkraut sausage school scissors shoes sleep some their there these those yourself]

6.4.5 Delimiters

So far delimiters have been used in the identification procedures and induction rules without a formal definition of what qualifies as a delimiter. NPs, Vs, Predicates, and sentence boundary were used as delimiters. This arbitrary choice was based on the assumption that these elements mark the beginning and/or the end of NP chunks; i.e., non-recursive NPs that do not overlap with other phrases. Vs and Predicates were used as NP delimiters based on the original assumption that NPs refer mainly to objects (people and things), hence these non-nominal elements indicate the beginning or the end of a nominal expression (i.e., NP). NPs, by default, imply the end of a constituent and the beginning of another.

Having identified the rudimentary frames in Tables 16 and 17, a delimiter can now be formally defined as follows.

(83) An item is a delimiter D iff it immediately precedes or
follows an NP in a given frame. Hence
D = {V | Pred | NP | S | to | off | on | out | up | to | in | with | if | for}

Limiting delimiters to items in the frames identified is consistent with the strategy that has been followed throughout that the input to any component of the learning procedures should be limited to those elements that are visible to the learning algorithm. An element is visible if it is part of the output of an identification procedure or an induction rule. Including the word *to, off, on, out, up, to, in, with, if,* and *for* in the set of delimiters according to (83) is based, in addition to their visibility, on the simple argument that if a word immediately precedes or follows a non-recursive NP, then this word is not part of this NP, and should indicate its left or right boundary. Having expanded the set of delimiters using (83), it was possible to detect the boundaries of more NPs, and consequently all other related procedures and induction rules were triggered.

6.4.6 Saturation

The processes of identifying new frames, new verbs instantiating old frames, and frame-based induction were applied iteratively until no new frames could be identified. This means that the frame identification process reached saturation and a final state was attained. Table 18 shows the new frames that were identified and some of the verbs that realized them. The words in (84) present some of the potential verbs that were identified after the iterative application of the induction rules.

F	*Frame Structure*	*Verbs*
F_{20}	_ V	can don’t doesn’t didn’t can’t could make do must might
F_{21}	_ NP down	pull bring put take
F_{22}	_ NP away	put take took throw threw
F_{23}	_ NP back	got put bring give take push
F_{24}	_ NP over	turn turned pull
F_{25}	_ NP around	turn pull
F_{26}	_ NP from NP	learn take
F_{27}	_ NP under NP	put see
F_{28}	_ NP inside NP	want put
F_{29}	_ into NP	roll look
F_{30}	_ in NP	get sit goes look
F_{31}	_ through NP	look
F_{32}	_ about NP	found think
F_{33}	_ on NP	go ride write got turned sit was get come slide
F_{34}	_ with NP	play do draw goes go write
F_{35}	_ for NP	go need
F_{36}	_ at NP	look sing
F_{37}	_ out NP	wash look
F_{38}	_ up NP	pack open go tape

Table 18: Examples of the new frames identified by CBL-1 after saturation

(84) [**already** been bend **better** breathe buy call can't can care carry catch change choke climb color come complete confine could cry cut didn't doesn't does don't drop dry **ever** fall feel figure fill find finish fold forgotten goes gone gotten had hammer has hear imagine imitate juggle jump keep knock let lick listen lock lose lost mail matter mess might mind missed miss must **only** pedal pinch point pretend reach read realize remember reminded ring rip roll run said say scare seem seen should smash sound spill spoil spread squeak squeeze stuff taken talked taste tickle took touch wait wake walk wants waste will win won works work worry wrap]

The induction rules were constructed following the same strategy used in the induction rules above. Given the frames in Table 18, the ultimate number of potential frames identified was 38. It is obvious that some of these frames could be further compressed into more general frames; however, this step is not pursued in this dissertation.[30]

6.5 Evaluation

The overall performance of the implementation was measured in terms of *precision* and *recall* (Manning and Schütze, 2003:267-269). Precision is defined as $\frac{tp}{tp + fp}$, where *tp* (*true positive*s) are the cases the system got right, and *fp* (*false positives*) are the cases the system got wrong. Recall, on the other hand, is defined as $\frac{tp}{tp + fn}$, where *fn* (*false negatives*) are the correct cases that were not captured by the system.

In order to compute the precision and recall for the procedures in identifying this knowledge, a POS-tagged version of the Peter corpus in the CHILDES database was used. This version contained ≈ 53759 verb tokens and ≈ 1050 types, ≈ 27426 noun tokens and ≈ 1083 types, and ≈ 15156 determiner tokens and ≈ 33 types. The tagged version did not contain frame or NP information, consequently a full evaluation of the algorithms performance on these tasks was not possible.

In this implementation, the learning procedures managed to collect information about potential verbs, nouns, determiners, noun phrases, and frames. The learner was able to identify 310 verb types totaling 27805 tokens, 1062 noun types totaling 8920 tokens, 7 determiner types totaling 10844 tokens, two types of noun phrase structures totaling

[30] For example, all frames that start with a preposition can be compressed in a more general frame of the form P NP.

20084 tokens (in addition to those that were given to the learner as cues), and 38 frames totaling 12512 tokens. Table 19 summarizes some aspects of this knowledge. Table 20 shows the type and token precision and recall ratios for the implementation's performance in these tasks.

Potential Tag	**Types**	**Correct** (*tp*)	**Incorrect** *(fp)*	**Tokens**	**Correct** (*tp*)	**Incorrect** *(fp)*
Verbs	310	291	19	27805	27249	556
Nouns	1062	1041	21	8920	8563	357
Determiners	7	7	0	10844	10844	0
Total	1383	1332	40	47569	46656	913
NP	2	2	0	20084	20084	0
Frames	38	-	-	12512	-	-

Table 19: Summary of knowledge acquired by CBL-1

Potential Tag	**Type Precision**	**Type Recall**	**Token precision**	**Token Recall**
Verbs	≈ 94%	≈ 28%	≈ 98%	≈ 51%
Nouns	≈ 98%	≈ 96%	≈ 96%	≈ 31%
Determiners	100%	≈ 21%	100%	≈ 72%
NP	_	_	≈ 100%	_
Average	≈ 97 %	≈ 48%	≈ 98.5%	≈ 51%

Table 20: Precision and Recall for CBL-1

The precision and recall ratios in Table 20 reflect two basic properties of this semantically-bootstrapped learner. The first is that the learner does much better in precision than in recall. The other is that this learner's performance in noun identification is remarkably good both for types and tokens. This property could be reasonably attributed to the fact that this learner was initially biased in favor of nominal expressions because of the initial cues that were given to it, which were mainly names of things and people. An interesting aspect of this result is that it is consistent with the priority of nouns to verbs in the process of acquisition that has been established by psycholinguistic research as discussed in previous chapters.

As for the learner's performance in learning frames, it was difficult to evaluate directly given the lack of frame information in the tagged corpus. However, given the central role of this information in inducing knowledge about nouns and verbs in the two tables above, it can be assumed that the performance of the learner in identifying these categories can serve as an indication of its performance in frame learning.

6.6 General Discussion and Conclusions

The main assumption behind CBL-1 was that the input provides a set of semantic cues that can be used in bootstrapping a distributional learner. The overall performance of this basic bootstrapping system raises some interesting questions about the role of the input and biased induction in language acquisition, given the very limited initial knowledge it bootstrapped from.

Firstly, the learner started with minimal *a priori* knowledge, mainly a subset of nominal expressions that refer to objects in the world. The bootstrapped knowledge was thus induced almost completely in terms of the distributional regularity in the input, and, yet, was highly accurate. This cue-based learner thus presents strong evidence that the input provides precious language-internal regularities that can, if used methodically, validate an empirical approach to language acquisition. This conclusion carries even more theoretical weight given the very small size of the corpus used by the learner.

Secondly, the learner showed a clear bias towards learning nouns. It is plausible to argue that this is the result of providing the learner with initial knowledge in the form of names of things and people. Yet, it is a result that deserves future investigation in order to understand its causes as well as its consequences. One possible direction is to replace this set of cues with another that includes for example only verbs, and see if the learner will show a similar bias towards verbs. This possibility is partially tested in the second implementation in the following chapter where the learner is not given any initial cues, and consequently is not biased in favor of either verbs or nouns.

Thirdly, from an automatic/human language acquisition perspective, the learner's performance shows the possibility of 'learning a lot given little'. Other systems for grammar induction, in general, and frame identification, in particular, assumed larger and more language-specific initial knowledge, with similar or lower performance.

Finally, the set of initial cues used by the learner is clearly small, and can be easily compiled for any given language from a small corpus of this language. Moreover, the learning procedures contained in this learner proceed according to the distributional regularities in the corpus. These two main features give this cue-based learner an explicit language-independence flavor that makes it highly general and easier to test on other languages.

However, this learner has some weaknesses that should be circumvented in order to obtain a more plausible learner. The first is that, being semantically bootstrapped, it was not based on a fully automatic procedure for cue extraction. The second shortcoming is that it was not able to assign more than one category to each word. The third is that it arbitrarily searched for possible frames in the right-side contexts of verbs, a decision that should be taken based on the distributional information in the corpus. The last shortcoming is that it assigned equal memberships to verbs that selected the same frames, thus ignoring the fact that some verbs clearly prefer some frames over others. Consequently, the following chapter presents a more sophisticated cue-based implementation of the learning procedures introduced in Chapter 5.

Chapter 7

Implementation [2]:

A Distributionally Bootstrapped Cue-Based Learner (CBL-2)

7.0 An Outline

This chapter presents a Distributionally Bootstrapped Cue-Based Learner (CBL-2) based on the Distributional Procedure for Cue Extraction, MI-Based Distributional Similarity, and Cue-Based Frame Identification, introduced in Sections 5.2 and 5.4 in Chapter 5, respectively. The general logic behind this model is (i) that key structural properties of a language can be bootstrapped from the distribution of mutual information in this language, and (ii) that cues provide a simple strategy to capture how this information is distributed.

To circumvent the weaknesses of CBL-1, the learner presented here comprises three main algorithms. The first algorithm presents a simple cue-based method for predicting the head direction given a small size corpus. The logic behind starting with this algorithm is that important structural properties of language should follow naturally from information about head direction. The second algorithm presents another cue-based method for the identification of predicates and arguments using information provided by the first algorithm. Equipped with information about head direction, this algorithm is expected to have some initial ‘intuition’ where to find predicates and arguments. Seeded with the information provided by the first two algorithms, the last algorithm exploits this knowledge to determine the most probabilistic syntactic frames that best describe the lexical syntactic properties of the predicates identified. Together, these three algorithms present a generalized, cue-based, and language-independent system for grammar induction, in general, and frame identification, in particular. Figure 6 visualizes this learner. In the following sections, I present the dynamics of this learner in more detail.

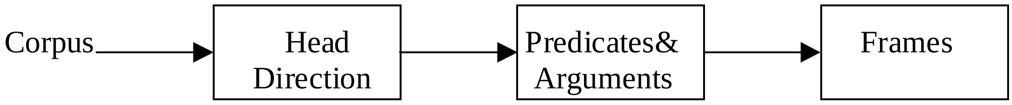


Figure 6: Components of CBL-2

7.1 Probabilistic Parameter Setting: Head First

Setting the head parameter in a target language L provides L-learner with a key feature of this language. Once this parameter is set, the basic word order and the complementation direction(s) in L can be easily identified. Learning that L is *mainly* head-initial will bias the learner towards the right context for information about complements. This learned bias should reduce the search space for an algorithm looking for the arguments of a given predicate, in particular, and any type of dependency, in general. Hence the importance of this step for the identification of subcategorization frames, the main issue of this dissertation.

This means that setting this parameter in an early phase in the language acquisition process using minimal input would significantly help the learner to bootstrap into the rest of the grammar of the input language.

Below, I will introduce a cue-based algorithm for computing the probabilities of any given head parameter setting in any language from a small-size, untagged corpus. Unlike the traditional generative approach to parameter setting, this method assumes that the value V of a given parameter T in language L is probabilistic rather than categorical. Moreover, this algorithm has no preconception of what a ‘head’ is, or what constitutes a set of possible heads in a language. What this algorithm does is estimate the probabilities of head direction in an input language without having any lexical-syntactic knowledge of specific lexical items.

The efficiency and potential language-independent nature of the algorithm was tested on three languages; English, Japanese, and German. However, the main focus here is on English.

7.1.1 Algorithm

The intuition behind this algorithm is that the head direction in any language should affect the distribution of mutual information between the right contexts and left contexts in this language. The mutual information version used in this model is that introduced in (49) in Section 5.3.2, repeated here for ease of reference.[31]

$$(49) \qquad I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Comparing the mutual information in both contexts is then expected to provide a probabilistic measure of the head direction. If a language is *predominately* head-initial**,** the expectation is that the sum of the mutual information between the words in the language and their right contexts should be significantly larger than the sum of their mutual information with their left contexts. A language that instantiates both head values almost equally (i.e., head-initial and head-final) is expected to show similar right and left associations.

[31] Remember that *P(x,y)* is the probability of observing *x* and *y* together, *P(x)* and *P(y)* are the probabilities of observing *x* and *y* anywhere in the corpus. If *x* and *y* tend to occur in conjunction, their mutual information will be high. If they are not related and co-occur only by chance, their mutual information will be zero. Finally, if the two variables tend to 'avoid' each other, their mutual information will be negative.

One way to measure the distribution of mutual information in a language is to sum the mutual information between the words in this language and the words preceding them, and compare it to the sum of mutual information with the words following them.

However, it is shown that language input contains clear and simple language-independent cues that can be used by the learner to uncover the direction of the mutual information flow in the input language. Therefore, we do not need to use the mutual information properties of every word in the input in order to compute the probabilities of the head parameter.

The cues used are utterance boundaries. The psycholinguistic research reviewed in the previous part has provided strong evidence that language learners are sensitive to and use utterance boundaries in language acquisition.

The idea behind using utterance boundaries as a cue to head direction in language is that utterance boundaries should provide key information about the distribution of mutual information in a given language. Languages with different head-parameter values should show a significant difference vis-à-vis the distribution of mutual information utterance-initially and utterance-finally. The probability of a language being head-initial or head-final is computed as a function of the two contexts, respectively.

This algorithm was tested on corpora from English, Japanese, and German. The results of this algorithm showed that it converged to close approximations of the head-parameter in these languages. It is expected that this algorithm should work with any language.

7.1.1.1 Exhaustive Mutual Information

One way to measure the distribution of mutual information in a language is to sum the mutual information between the words in this language and the words preceding them, and compare it to the sum of mutual information with the words following them. The probabilities of any given head parameter setting in this language are then a function of the two sums, in the following manner.

To compute the mutual information between a word *x* and its right context, we first compute the mutual information between *x* and every word *y* immediately following it (including the utterance-right boundary), then we sum these mutual information values. So if *x* is followed by *M* words in the corpus, *x*'s right-context mutual information $I_{x\text{-}r}$ is computed by the formula in (85.a). *x*'s left-context mutual information $I_{x\text{-}l}$ is computed using the words immediately preceding it, in the same way (85.b), where *G* is the number of words preceding *x*.

(85) a $$I_{x-r} = \sum_{i=1}^{M} \log_2 \frac{P(x, y_i)}{P(x)P(y_i)}$$

b. $$I_{x-l} = \sum_{i=1}^{G} \log_2 \frac{P(y_i, x)}{P(x)P(y_i)}$$

To compute the overall left-context mutual information in a given language *L*, ($I_{L\text{-}l}$), we sum the left-context mutual information of every word in *L*. The overall right-context mutual information ($I_{L\text{-}r}$) is computed in the same fashion. If *L* has *N* words, these two values are computed as follows:

(86) $$I_{L-l} = \sum_{i=1}^{N} I(x_{i-l})$$

(87) $$I_{L-r} = \sum_{i=1}^{N} I(x_{i-r})$$

The values in (86) and (87) should provide an estimate of the possible values of the head parameter in any given language. $I_{L\text{-}l}$ provides an approximation of the left dependency in *L*, and consequently can be used to estimate the probability of *L* being head-final, whereas $I_{L\text{-}r}$ estimates its probability of being head-initial.

Assuming that the head parameter is binary-valued, initial or final, the probabilities of the head parameter for a language *L* are computed as given in (88), using the axioms of Probability (Harris and Stocker 1998: 778).

(88)

Let $P(L(H_i))$ be the probability that *L* is head-initial, and
Let $P(L(H_f))$ be the probability that *L* is head-final,
such that

$$0 \leq P(L(H_i)) \leq 1$$
$$0 \leq P(L(H_f)) \leq 1$$
$$P(L(H_i)) + P(L(H_f)) = 1$$

Then

a. $P(L(H_i)) = \dfrac{I_{L-r}}{I_{L-r} + I_{L-l}}$

b. $P(L(H_f)) = 1 - P(L(H_i))$

That is, the probability that *L* is head-initial is computed by normalizing the overall mutual information in the right context by the sum of the overall information in both contexts. The probability that *L* is head-final follows naturally from the axioms above.

The null hypothesis is that the mutual information in any language *L* is distributed symmetrically between the left and right contexts. This means that $P(L(Hi) = P(L(Hf)) = 0.50$. If $P(L(H_i))$ is significantly greater than $P(L(H_f))$, then *L* is biased towards a head-initial value. If $P(L(Hf))$ is significantly greater than $P(L(Hi))$, then *L* is predominantly head-final. Consequently, the null hypothesis is rejected in both cases. If *L* instantiates both head values (i.e., head-initial, and head-final) equally, these two probabilities should be very close, and the null hypothesis holds.

Given the way these two probabilities are computed, it is not likely to assert that a given language is categorically head-final or head-initial. This is due to the fact that in any language there exist words that tend to occur utterance-initially or utterance-finally. This means that the overall mutual information in either context is not expected to be zero. This implies that headedness is gradient in any given language, and that the values of the head parameter are only probabilistic.

7.1.1.2 Cued Mutual Information

After a series of pilot experiments, it was found that head probabilities can be similarly estimated if we limit the left context and the right context in the formulas above to the utterance left-boundary and right-boundary, respectively.

In this case, if *K* is the number of words that occur utterance-initially, and *F* is the number of words that occur utterance finally, (86) and (87) can be re-formulated as (89) and (90), respectively.

(89) $$I_{L-l} = \sum_{i=1}^{K} I(x_i;\$)$$

(90) $$I_{L-r} = \sum_{i=1}^{F} I(x_i;\#)$$

where *$* and *#* indicate the utterance left- and right-boundary, respectively. Using these new estimates, head probabilities are computed using the same formulas in (88).

These new estimates provide a simpler algorithm to compute head probabilities, since they do not require computing the mutual information of every word in the corpus using its co-occurrences with other words. This naturally involves fewer computations than the other strategy, where every context of every word in the corpus is taken into consideration. Consequently, these new estimates were used in the experiments below.

7.1.2 Experiments

This algorithm was tested on untagged corpora extracted from the CHILDES database for English, Japanese, and German. The focus in this chapter, in particular, and the dissertation in general, is mainly on English. Japanese and German were only used to show the efficiency and possible language-independent nature of the algorithm. Therefore, no detailed discussion is devoted in this dissertation to the deeper implications of the head-direction probabilities computed by the proposed algorithm for these two languages.

7.1.2.1 Corpus Description and Preprocessing

The corpus for each of these languages contains the transcription of adults' child-directed speech. Changes to the corpus were minimal. Single word utterances were removed from the corpora. Every line in the corpus was considered an utterance. The utterance-final punctuation (i.e. '.', '!', and '?') was changed into a single utterance-final symbol (#), and an utterance-initial symbol ($) was added.

7.1.2.2 English

The algorithm was first applied to English corpora of adults' child-directed speech in the CHILDES database. The corpus was extracted from the child Peter's files in Bloom (1970). The corpus contained 25148 lines, 206942 tokens (including utterance-boundary symbols) and 156646 tokens (excluding these symbols), and 3086 words.

The values of the overall left-context (I_{L-l}) and the right-context mutual information (I_{L-r}) were computed for this corpus, using the formulas in (89) and (90), respectively. The values were, 100 and 3076, respectively. Using the formulas in (88), the head-direction probabilities were as follows:

(91) a. $P(L(H_i)) = \frac{3076}{3076+100} \approx 0.97$

b. $P(L(H_f)) \approx 1 - 0.97 \approx 0.03$

In other words, these probabilities indicate that in English the majority of words share more mutual information with their right context than with their left context.

7.1.2.3 Japanese

To put the head probabilities computed for English in the previous section in a more meaningful perspective, the same cue-based algorithm was run on Japanese corpora from the CHILDES database. The corpus was the transcription of the adult's child-directed speech in the child Ryookun's corpus in Miyata-Ryo. The corpus contained 26926 lines, 41409 tokens, and 2492 words. The corpus was preprocessed as described in subsection 7.1.2.1 above.

The values of the overall left-context ($I_{L\text{-}l}$) and the right-context mutual information ($I_{L\text{-}r}$) were computed for the Japanese corpus, using the formulas in (89) and (90), respectively. The values were, 2200 and 575, respectively. Using the formulas in (88), the head-direction probabilities were as follows:

(92) a $P(L(H_i)) = \frac{575}{575+2200} \approx 0.21$

b. $P(L(H_f)) \approx 1 - 0.21 \approx 0.79$

In other words, these probabilities indicate that in Japanese the majority of words share more mutual information with their left context than with their right context.

7.12.4 German

To further test the efficiency and the language-independent nature of the probabilistic cue-based algorithm proposed for head parameter, it was run on German corpora. The corpus was extracted from the child Kristin's files in the CHILDES database. It contained 70707 lines, 200535 tokens, and 4949 words. The corpus was preprocessed as described in subsection 7.1.2.1 above.

The values of the overall left-context ($I_{L\text{-}l}$) and the right-context mutual information ($I_{L\text{-}r}$) were computed, using the formulas in (89) and (90), respectively. The values were, 1856 and 5462, respectively. Using the formulas in (88), the head-direction probabilities were as follows:

(93) a. $P(L(H_i)) = \dfrac{5462}{5462 + 1856} \approx 0.75$

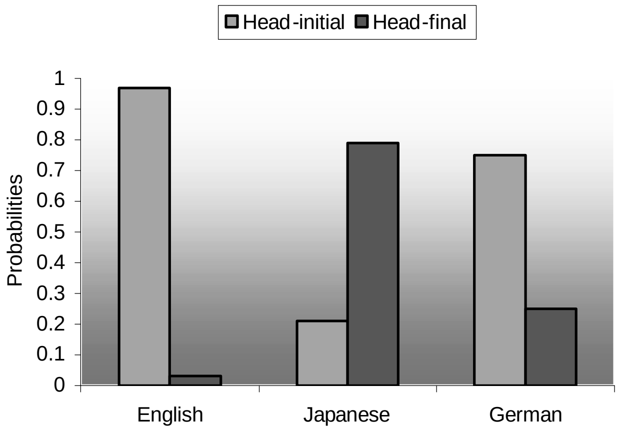
b. $P(L(H_f)) \approx 1 - 0.75 \approx 0.25$

In other words, these probabilities indicate that in German, like in English though with different probabilities, the majority of words share more mutual information with their right context than with their left context.

7.1.3 General Discussion and Conclusion

Figure 7 summarizes the head-direction probabilities for the three languages. These probabilities could be interpreted in different ways.

The first implication is that English is ≈ *0.97* head-initial, and ≈ *0.03* head-final. Japanese is ≈ *0.21* head-initial and ≈ *0.79* head-final. German is ≈ *0.75* head-initial and ≈ *0.25* head-final. The English head probabilities are clearly consistent with the commonly held analysis of English as head-initial. Similarly, German is predominately head-initial, yet with a lower probability than English. Japanese is predominately head-final, and is almost a mirror image of German.



	$P(L(H_i))$	$P(L(H_f))$
English	0.97	0.03
Japanese	0.21	0.79
German	0.75	0.25

Figure 7: Summary of head-direction probabilities in English, Japanese, and German

The second possible interpretation of these probabilities is to think of them as measures of head probabilities as well as word-order flexibility. On this interpretation, these probabilities predict that word order in English should be more fixed than in Japanese and German. Accordingly, higher probabilities in Figure 7 could signify the basic or default word order, and the lower ones signify the *derived* word order, assuming that this distinction exists in the first place.

Another possible interpretation is that these probabilities average the dependency direction of the word types in the input. On this interpretation, these probabilities summarize the distributional properties of the words in a language. That is, these probabilities indicate that some word types have right-side dependency, and others left-side dependency, in accordance with these probabilities. They could also mean that some words have these two kinds of dependency in accordance with the head probabilities for this language.

All these are legitimate and interrelated interpretations. Their significance for this dissertation is that with a simple cue-based, language-independent algorithm for probabilistic head-parameter setting, we are able to differentiate languages, vis-à-vis head direction, using only utterance boundaries as cues, and a small-size, untagged corpus.

The typological and/or psycholinguistic implications of these probabilities for Japanese and German, and other languages, other than English, are not discussed any further in the dissertation. Further investigation is still required in order to establish these probabilities, on the one hand, and what they reflect about the structural properties of these languages, on the other. Moreover, the scope of parametric variation in human languages would require more cues that are able to capture the idiosyncrasies of individual languages. Their implications for the automatic acquisition of the structural properties of English are considered in the following chapters. It is shown in the following chapter that these probabilities could be used to bootstrap another algorithm for the identification of predicates and arguments in English.

7.2 MI-Based Categorization

This section presents an algorithm for the initial categorization of the input into two main classes, i.e., arguments and predicates. The main idea behind this algorithm is to categorize words based on their distributional similarity as measured in terms of their co-occurrence with a set of cues. Cues are extracted using the Distributional Cue Extraction Procedure in Section 5.2 in Chapter 5. Similar to the first algorithm, the association between these cues and other words is measured in terms of mutual information. It is shown below that this algorithm is able to identify a subset of the arguments and predicates in the input with a relatively high level of accuracy.

7.2.1 Algorithm

7.2.1.1 Category Cues

The central part of this algorithm is to extract the set of relevant cues in the input according to the criterion in (38) repeated below.

(38) **Definition of Category Cues (*K*)**
The set of Category Cues, *K,* is the smallest subset of the elements $\{k_1,...,k_m\}$in a corpus *R* such that every element in *R* occurs at least once with at least one member in *K*.

This criterion is the basis of a procedure that makes use of the highly frequent words in order to approximate *K* in the following manner, as introduced earlier. We start with building a decreasing frequency profile for all the words $\{w_1,\ldots,w_n\}$ in a corpus, *R*, where w_1 is the first most frequent word in *R*, w_2 the second most frequent, and so on. The set of cues is $K = \{w_1,\ldots,w_m\}$, such that if we add up the number of words, X_1, that co-occur with w_2 and the number of words, X_2, that co-occur with w_2, until the *m*-most frequent word, w_m, the number of words $[X_1+X_2+\ldots+X_m]$ converges to an order, *α*, of *n*, where *n* is the number of word types in the corpus. Only the first approximation, i.e., α = 1, is implemented in the present algorithm. In pseudo-code, this procedure is as follows:

(40) Procedure for Distributional Cue Extraction

```
Function SUBSET(K,R)
K := Ø;
for i := 0 to n do; $\alpha$ := (1,2,…);
get the number of word types $n$ in $R$;
get the frequency of $w_i$ , $f(w_i)$ in $R$ ;
build a decreasing frequency profile $F := \{f(w_i) > f(w_{i+1}) > \ldots f(w_n)\}$ ;
get the number of words $|w_{i-}|$ that immediately precede $w_i$ ;
get the number of words $|w_{i+}|$ that immediately follow $w_i$ ;
    $X_i := |w_{i-}| + |w_{i+}|$ ;
    $X_{i\_}total := \sum_{i=1} X_i$ ;
    if $X_{i\_}total := \sum_{i=1} X_i := \alpha n$
    return $K := \{w_i\}$;
    else
        repeat
            i := i +1;
        until   ($X_{i\_}total := \sum_{i=0}^{i+k} X_i >= \alpha n$,
                 $X_{i\_}total := \sum_{i=0}^{(i+k)-1} X_i < \alpha n$);
    return $K := \{w_i,..,w_{i+k}\}$
```

7.2.1.2 MI-based Categorization

Once *K* is identified, it can be utilized in word clustering, and perhaps other tasks, using different measures of co-occurrence. The algorithm proposed here uses *K* for categorization in the following manner.

Let the Category Cues for corpus *R* be $K_R \equiv \{k_1,\ldots,k_m\}$, where *m* is the number of words in K_R. The distributional properties of words in *R* are captured in terms of their cued bigrams (henceforth, *K*-bigrams) with the members of K_R.[32] For every word W_i in *R*, we first extract its *K*-bigrams. This means that there are two possible *K*-bigram types: left

[32] Henceforth, the terms ‘bigram’ and ‘*K*-bigram’ will be used interchangeably, unless otherwise specified.

and right. The left bigram for W_i is that where W_i is immediately preceded by a word in K_R. The right bigram is that where W_i is immediately followed by a word in K_R. It is possible that some words could have both types or either. That is, the maximum number of K-bigrams for a word W_i is $2m$, m on each side, and the minimum is 1, either left or right. This is visualized in Figure 8.

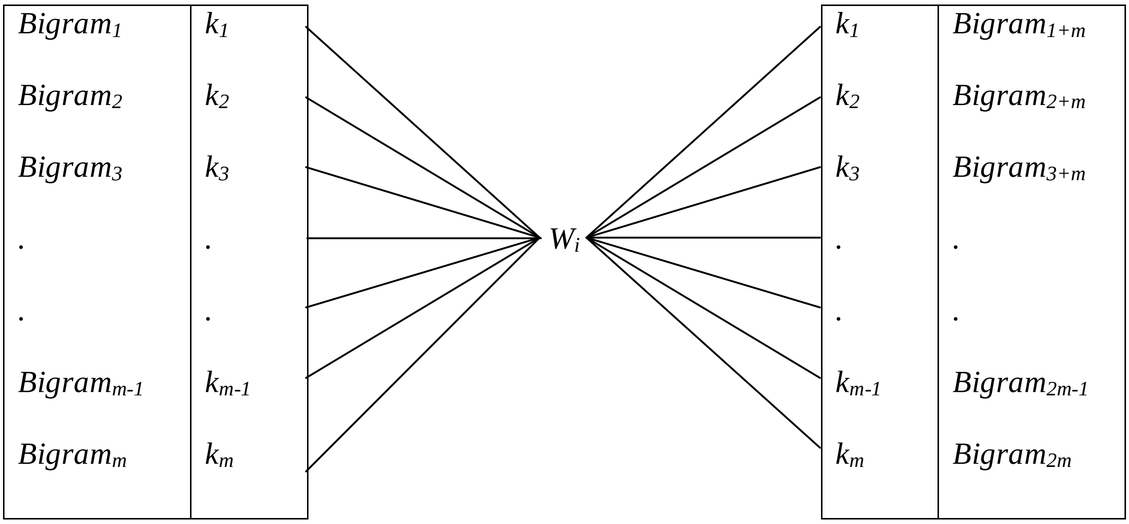


Figure 8: Representation of Distributional Contexts

The hypothesis is that each K-bigram indicates a lexical category of some type. That is *Bigram*$_1$ corresponds to *Category*$_1$, and *Bigram*$_2$ corresponds to *Category*$_2$, and so on. This means that there are 2m possible categories, m on each side. The membership of a word in a certain category is established in terms of MI in the following fashion.

We first compute the MI for every K-bigram. For example, the MI of the left and right K-bigrams for a word W_i is thus computed as given in (94a) and (94b), respectively.

(94) a. $$I(k_i;W_i) = \log_2 \frac{P(k_i,W_i)}{P(k_i)P(W_i)}$$

b. $$I(W_i;k_i) = \log_2 \frac{P(W_i;k_i)}{P(k_i)P(W_i)}$$

Then, for each word W_i, we pick the bigram with the highest MI on each side, if available. That is, each word could either have two bigrams with highest MI, one on each side, or just one on either side. For example, if a word W_i occurs in $Bigram_x$, and this bigram has the highest MI among the left bigrams, it is concluded that W_i belongs to $Category_x$, and so on. If W_i also occurs in $Bigram_y$, and this bigram has the highest MI among the right bigrams, then we conclude that W_i belongs to $\mathrm{Category}_y$. This means that W_i belongs to two categories, *x* and *y*. In case a word has two bigrams on the same side with the same MI, and they happen to be the bigrams with the highest MI on this side, both bigrams are used in categorization.

The decision that a given word belongs to one or more categories is probabilistic. Category probabilities for a word are computed in terms of the MI of its *K*-bigrams in the following manner. Given a word W_i with $Bigram_x$ having the highest MI among the left bigrams of W_i, and $Bigram_y$ with the highest MI among the right bigrams of W_i, the probability that W_i belongs to $Category_x$ is computed by dividing the MI of $Bigram_x$ by the sum of the MI of $Bigram_x$ and $Bigram_y$. The probability that W_i belongs to $Category_y$ follows naturally from the axioms of probability theory. This is formalized as follows:[33]

(95) a. $$P(W_i, Cat_x) \approx \frac{MI(Bigram_x)}{MI(Bigram_x) + MI(Bigram_y)}$$

b. $P(Wi, Cat_y) \approx 1 - P(Wi, Cat_x)$

Using these category probabilities, we can capture the lexical ambiguity of words that belong to more than one category, as it is shown below. Given the probability

[33] If there are two bigrams on one side with equivalent highest MI, the probabilities of possible categories for the target word are computed in a similar fashion by dividing by the sum of the MI in all the related bigrams.

distribution of these categories, we can easily identify the main, as well as, peripheral category to which a given word belongs.

7.2.1.3 Category Compression

The categorization method described above is expected to yield all the logically possible categories given the number of cues. In principle, it is possible that two or more of these initial bigrams/categories are equivalent, which means that they can be collapsed into one more abstract and general class. The equivalence of two initial bigrams/categories is a function of the equivalence in the set of Category Cues. Two or more bigrams/categories are equivalent if their cues are equivalent. The equivalence of two or more cues can be simply established, like other words, using the same strategy described above.

However, it is likely that the bigrams where a cue occurs have negative MI. In this case this strategy cannot be used, and equivalence is established simply using the same MI-based method used in the head-parameter algorithm. That is, for every cue, k_i, we determine its dependency direction using the MI of its bigrams on the two sides. We first add up the MI of the bigrams where it is the left member (right dependency), and the MI of the bigrams where it is the right member (left dependency). We then compute the probability of its left dependency $P(k_i(L))$ by dividing the sum of the MI of its left bigrams by the sum of the MI of both types of bigrams. The probability of its right dependency $P(k_i(R))$ follows from the axioms of probability theory. A cue is mainly left dependent if $P(k_i(L))$ is significantly larger than $P(k_i(R))$. This is formalized as follows:

(96) Let $K = \{k_1,\ldots,k_m\}$ be the set of cues,
Let $\{k_{1,L},\ldots, k_{m,L}\}$ be the sum of the MI of the left bigrams of $\{k_1,\ldots,k_m\}$, respectively,
Let $\{ k_{1,R},\ldots, k_{m,R}\}$ be the sum of the MI of the right bigrams of $\{k_1,\ldots,k_m\}$, respectively, then

$$P(k_i(L)) = \frac{k_{i,L}}{k_{i,L} + k_{i,R}}$$

$$P(k_i(R)) = 1 - P(k_i(L))$$

Having established the equivalence of two or more cues, we then collapse their corresponding bigrams/categories into one class. That is, if we have two bigrams/categories, $\text{Bigram}_x/\text{Category}_x$ and $\text{Bigram}_y/\text{Category}_y$ which have the cues k_x and k_y as their left members, respectively, $\text{Bigram}_x/\text{Category}_x$ and $\text{Bigram}_y/\text{Category}_y$ are equivalent if and only if k_x and k_y are equivalent, given (96). The same applies if the cues are the right members.

Once the new class is determined, the new category probabilities of the words in the new class are the sum of their old probabilities before compression. For example, if a word W_i belongs to categories $Category_x$ and $Category_y$ with probabilities P_x and P_y , respectively, before compression, the probability of the membership of W_i in the new class is the sum of P_x and P_y. This way the category probability of any word that belongs to only one category remains intact, i.e., 1.

It is important to clarify the position of this component of category compression in the overall categorization algorithm. Firstly, *this component is optional.* Its inclusion in the algorithm depends on the level of granularity we intend for the categories. I will compare below the output of the algorithm with and without this component in order to show its effect on the level of category granularity. Secondly, its order of application within the algorithm depends on how we want the categorization to proceed, i.e., bottom-up or top-down. In the first case, compression is applied after the initial categorization has already been applied. In the other case, compression is applied before the initial categorization. Given the way the algorithm works, the order of application is neutral, and both orders should converge to the same classes.

7.2.2 Experiment

7.2.2.1 Corpus Description

The categorization algorithm described in the previous section was tested on English using the same Peter corpus used in the head-parameter algorithm. Changes to the corpus were minimal. Every line in the corpus was considered an utterance. The utterance-final punctuation (i.e. '.', '!', and '?') was changed into a single utterance-final symbol (i.e. #), and an utterance-initial symbol (i.e., $) was added. The corpus contained 25148 lines, 156646 tokens, and 3086 words.

7.2.2.2 Results

The algorithm was applied in three phases. The first phase built a decreasing frequency profile for the words in the corpus. Using the criterion in (40) above, the second phase established the set of Category Cues, *K*. In the third phase, *K* was used to categorize words, using (95), and assign probabilities to their membership in possible categories, using (96).

7.2.2.2.1 Category Cues

Table 21 shows the first ten most frequent words in the decreasing frequency profile built in the first pass for the words in the corpus.

In the second phase, the algorithm converged to the first order of *N* after the fourth most frequent word in the corpus. Accordingly, the first-order *K* contained the words {*the*, *you*, *a*, *it*} as its members. The number of words that co-occurred with these cues was initially 3382, which were distributed among these cues as shown in Table 22. It is clear from Table 22 that, though ‘*the*’ is less frequent than ‘*you*’ in the corpus, as given in Table 21, ‘*the*’ occurred with more words than ‘*you*’ did.

Word	***Frequency***
you	7783
the	6256
it	4502
a	3076
I	2769
that	2181
to	2334
is	2130
in	2123
that's	1907

Table 21: The 10 most frequent words in the corpus

K	*Co-occ. Set*
the	1093
you	810
a	806
it	673
Total	3382

Table 22: First-Order Set of Cues

Given these 4 cues, the maximum number of *K*-bigrams for a word W_i is 2×4 = 8, 4 on each side, and the minimum is 1, either left or right, as discussed above. Each of these cues results in two types of bigrams, i.e., left and right. Remember that each bigram corresponds to a possible category. This means that the maximum number of categories we can infer from these bigrams is 8.

7.2.2.2.2 *K*-Bigrams

In the third phase, the bigrams of words in the corpus with the cues in Table 22 were extracted and their MI’s were computed, as described above. As a result of filtering out bigrams with negative MI, the number of words was reduced from 3382 to 1600, which constitute ≈ *1600/3086* ≈ *0.52* of all the words in the corpus. Note that, negative MI is excluded on the basis that it indicates that the members of the bigram tend not to co-occur. Table 23 shows some of these bigrams and their respective MI's. In Table 23, '_' is used to indicate the position of the target word in the bigram, a blank is used to indicate zero or less MI. $Bigram_1$ is that where 'the' is the left member, $Bigram_2$ is that where 'you' is the left member, and so on.

Bigrams→	$Bigram_1$	$Bigram_2$	$Bigram_3$	$Bigram_4$	$Bigram_5$	$Bigram_6$	$Bigram_7$	$Bigram_8$
Cues→	**the_**	**you_**	**a_**	**it_**	**_the**	**_you**	**_a**	**_it**
be		0.04			0.60		2.72	
box	4.31		1.9					
cover	4.26							2.33
found		3.41			1.55	0.08	2.16	3.62
go		1.66		0.13				
hold		1.13			2.59	0.32		3.5
knock		2.19			1.07			3.14
like		2.71		0.09	0.1		3.49	0.57
monster	1.9		1.9					
off				3.4	1.51			
out				2.16	1.3			
park	4.95				1.37			
quiet			2.7	2.75				
ride	0.16	1.01	3.88		3.27		0.92	
see		2.0			1.52	0.37	0.81	2.26
should		1.4			0.06			
small	2.24		3.75					
spray	3.07	2.61				2.61		
take		1.52			2.49		3.43	3.1
teddy	3.66		3.09					
telephone	4.23		3.45					

Table 23: Examples of words and the MI of their K-Bigrams

7.2.2.2.3 Initial Categories

These 1600 words were then categorized using the strategy described above. Accordingly, for each word we pick the *K*-bigram with the highest MI among the left bigrams, if available, and the *K*-bigram with the highest MI among the right bigrams, if available. For example, for the word '*be*' in Table 23, $Bigram_2$ has the highest MI among its left bigrams, and $Bigram_7$ among its right bigrams. Consequently, it is concluded that '*be*' belongs to the two categories that correspond to these two bigrams, i.e., $Category_3$ and $Category_7$. As expected, the number of categories was 8. Table 24 shows the number of words in each category and their ratios relative to the 1600 words, and some examples of the words in each category.

The distribution of words among these categories deserves some observations. The first clear and important observation is that words that belong to the left-associative categories (i.e., Cat_1, Cat_2, Cat_3, and Cat_4) represent 1404/1600 ≈ 0.88 of all the categorized words. Words that belong to the right-associative categories (i.e., Cat_5, Cat_6, Cat_7, and Cat_8) represent 524/1600 ≈ 0.33 of all the words categorized. Words that are both left- and right-associative comprise 328/1600 ≈ 0.21 of all the words categorized. This result is consistent with the results obtained from the head-parameter algorithm, which established English as mainly head-initial/head-left. Consequently, it is highly predictable that more words are left- than right-dependent.

Secondly, Cat_1 and Cat_2 include more than half, 0.54, of the words categorized, 1600.

Thirdly, words in categories Cat_1 and Cat_3 are mainly nouns (e.g., *airplane, alligator, bath, cat, etc…*) and some attributive adjectives (e.g., *sweet*, *green*). This is natural because these are the categories that correspond to the two bigrams whose left members are *the* and *a*, respectively.

Cat	Count	Ratio	Examples
Cat$_1$	443	0.28	abcs air airplane alligator alphabet ambulance animals annual apartment apple bathroom beach bear bed bedroom beds bell best biggest bike bouncing church closet country cover cows cube diaper dogs door eye family first game grass green heat horse kind last mirror mommy park people pillow screw screws sill sky sleeping spoons spray starting store stuff suitcase suitcases tag taillights teddy telephone telescope things tigers toilet toys, etc…
Cat$_2$	351	0.22	always already almost anyway are ate bang bet blow break bringing calling can coughing could couldn't count did didn't eat eating fitting forgot have haven't having hit knock knocked knocking know just live lock lost missed move moved mustn't need open opened opening ought pack park parking rip saw see seem singing smell speak speaking squeeze spread talk tear understand use using wash were wouldn't write writing wrote, etc…
Cat$_3$	423	0.26	babysitter bacon barrel barrette bat bath cat celery change chest chip choice crayon cream dog doggie donkey drink dresser drill dump earth feeling finger giraffe kiss kite lamb lemon letter lot mafia mail man map mess monkey monster noise night number piece pile pill picnic picture plant plastic plug pocket point ride road round stick strange sweet thumb tire tissue toy traffic wall etc…
Cat$_4$	187	0.12	actually alone along anymore apart around away back because before behind belong belongs broken by called came closed comes coming cost disappear does doesn't down empty fall fit for from gets goes has hasn't hurt in inside into is isn't itches keeps look looks made near out outside over ran roll says scare seems to turns under was wasn't went without won won't work would …
Cat$_5$	125	0.08	about answer around at bat bend between blow bringing bouncing by carrying change chasing cleaning clicking count down drinking eating fasten fitting flushed forgot green hears hit hits hitting holding in inside into latch likes mash moving of off onto opening out outside over park point ride rides roll round scare screws should smell to under use working etc…
Cat$_6$	99	0.06	afraid after ago are ask asked before behind bet can could did distract disturb disturbing fit hope if protect showed spoil spray taught tells then thought tickle told understand were will would yesterday etc…
Cat$_7$	99	0.06	almost always as be been being bought brought called caught cost finding for get gets goes got gotten had has have having hear heard just like live lock looks lost made mail need needs says singing smiles store using wait etc…
Cat$_8$	201	0.13	actually ate bang believe blowing bother break call closing cook cover covered cut decorate does doing doubt drank eat empty feel feeling fill filled filling flush fold folded forget found gave guess heat hid hide hiding mess mixing move open parking pour pretend push pushed pushing reach rip screw screwing shaved sing spill spit squeeze stir tear tearing unwrap unzipped wash etc…

Table 24: The counts, ratios, and examples of words in the 8 categories

Fourthly, words in the other categories are mainly verbs, adverbs (e.g., *actually, ago, always, already, almost,* etc..), and prepositions/particles (e.g., *of, off, over, out, in, into, by* etc…).

Fifthly, some words belong to more than one category. For example, the word *cover* belongs to categories Cat_1 and Cat_8, *spray* and *mail* are members of both Cat_1 and Cat_7, *screws* belongs to Cat_1 and Cat_5. This captures the fact that these words can be both nouns and verbs.

Finally, the categories in Table 24 also capture the fact that the word *open* can be both an adjective and a verb, hence it belongs to Cat_1 and Cat_2.

An expected result in the output is that the cues themselves could not be categorized by the algorithm. This is natural since the members of *K* do not tend to co-occur, and even if some of them do, their co-occurrence is rare and would result in negative MI. This means that category compression should be implemented as described in (96) above.

7.2.2.4 Category Probabilities

In order to establish the degree of membership of a given word in any of these categories, category probabilities were then assigned to every word as proposed in (95) above. Table 25 shows part of the categorized words and the probabilities of their membership in the corresponding categories. (The probabilities in Table 25 were rounded up to the second digit.)

Word	**Category Probabilities**							
	Cat$_1$	**Cat$_2$**	**Cat$_3$**	**Cat$_4$**	**Cat$_5$**	**Cat$_6$**	**Cat$_7$**	**Cat$_8$**
airplane	1							
alligator	1							
almost		0.45					0.55	
always		0.38					0.62	
apple	1							
around				0.56	0.44			
ask		0.74				0.26		
ate		0.58						0.42
bend		0.45			0.55			
bouncing	0.5				0.5			
bring		0.34						0.64
by			0.32	0.68				
called			0.22				0.78	
change			0.45		0.55			
closed				0.48				0.52
comes			0.56	0.44				
could		0.63				0.37		
cover	0.65							0.35
cry		0.89				0.11		
decided		1						
did		0.25				0.75		
do		0.32				0.68		
dog			1					
drink			0.31					0.69
fly			0.57				0.43	
green	0.9				0.1			
grow				1				
hungry			0.64			0.36		
in			0.34	0.66				
mail			0.57				0.43	
mess			0.80					0.20
off				0.69	0.31			
out				0.62	0.38			
over				0.75	0.25			
park	0.78				0.22			
screw	0.59							0.41
screws	0.5				0.5			
spray	0.46					0.54		
store	0.85						0.15	
use		0.24			0.76			
want		0.82					0.18	

Table 25: Category Probabilities of some words in the corpus

The category probabilities in Table 25 are more indicative of the performance of the algorithm. The most obvious examples are words that are always mono-categorical (i.e., words that belong to only one category). For example, the membership probability of words that are always nouns/nominal (e.g., *airport, alligator, apple,* and *dog*) in Cat1 and Cat3, which contain mainly nouns, is always 1. The same applies to words that are always verbs/predicative (e.g., *decided* and *grow*) with a *full membership* in Cat4 and Cat5, respectively, which include mainly predicative or non-nominal words. On the other hand, these probabilities reveal the poly-categorical nature of some words (i.e., words that belong to more than one category). An interesting example is the word *bouncing*, which belongs equally to the nominal category Cat_1 and the predicative category Cat_5. To understand the significance of this result to the performance of the proposed algorithm, I consider the distribution of this word in the corpus in more detail.

The word *bouncing* occurs only twice in the corpus (97a.b.). (Note: As mentioned above, $ and # indicate utterance boundaries.)

(97) a. $ are you *bouncing* the ball #
b. $ follow the *bouncing* ball #

In the first utterance, *bouncing* is a regular verbal *–ing* form. In the second utterance, *bouncing* functions as an attributive adjective. The intricacies and implications of this and similar cases are further discussed in the Discussion and Conclusion Section at the end of the chapter.

Another interesting example in this context is the word *green*. As the probabilities in Table 25 show, this word belongs to the nominal Cat_1 with a 0.9 probability, and to the predicative Cat_5 with a 0.1 probability. This word occurs 24 times in the corpus. It occurs predicatively in only one context, and attributively in all other contexts. (98) shows some of these contexts.

(98) a. $ is it green #
b. $ the green barrel #
c. $ the green one #
d. $ the other green one is inside the blue one #
e. $ what happened to the green chair #
f. $ a green one #
g. $ a green one #
h. $ that's a green lizard #

The obvious observation about the behavior of this word is the proportionality of its category probabilities and its distribution in the corpus.

Similar remarks can be made about other poly-categorical words: (1) words that function as verbs as well as nouns (e.g., *cover, drink, fly, mail, mess, park, screw, screws,* and *store*), and (2) words that can function as prepositions and particles (e.g., *around, off, out,* and *over*).

7.2.2.2.5 Category Compression

It is clear from Tables 24 and 25 that some categories can be collapsed into more comprehensive classes. For example, Cat_1 and Cat_3 contain mainly nouns and attributive elements. Thus it is possible to collapse these two categories into a more general and abstract class of *nominal* words (i.e., nouns and words that occur prenominally). Other categories can be similarly compressed. Remember that category equivalence is a function of cue equivalence. However, it was mentioned above that the cue words (i.e., {*the, you, a, it*}) were not among the words that were categorized by the algorithm. Accordingly, these words were categorized according to the equivalence procedure in

(96) above. According to this procedure, the equivalence of two cues is the function of the aggregate MI of their left and right bigrams.

After applying this procedure to the set of cues, their MI and dependency probabilities were as given in Table 26.

K	**MI(L)**	**MI(R)**	**P(L)**	**P(R)**
the	577	2301	0.2	0.8
you	702	720	0.49	0.51
a	344	2248	0.13	0.87
it	814	581	0.58	0.42

Table 26: The right and left MI and association probabilities for cues

As expected, *the* and *a* are almost always right-associative, *you* and *it* are almost equally left- and right- associative. Given this, the 8 initial categories described above can be compressed as follows. Given that *the* and *a* are both right-associative, it is concluded these two cues are equivalent. Accordingly, the left-associated categories dependent on *the* and *a,* i.e. Cat_1 and Cat_3 could be collapsed into the more general class, $Cat_{1,3}$, that comprises these two categories. Similarly, the right-associated categories dependent on these two words (i.e., Cat_5 and Cat_7) are conflated into one class, $Cat_{5,7}$.

On the other hand, *you* and *it*, are equally left- and right-associative. Consequently, all the categories associated with these two words on both sides (e.g., Cat_2, Cat_4, Cat_6, and Cat_8) are equivalent, and can be collapsed into a new class, $Cat_{2,4,6,8}$. This means that as a result of category compression, the initial 8 categories have been reduced to just 3 classes, i.e., $Cat_{1,3}$, $Cat_{2,4,6,8}$, and $Cat_{5,7}$. The new category probabilities for every word that previously belonged to more than one category were then computed using (96). Tables 27 and 28 show the new categories and their updated probabilities, respectively.

Cat.	Count	Ratio	Examples
Cat$_{1,3}$	866	0.54	abcs air airplane alligator alphabet ambulance animals annual apartment apple bathroom beach bear bed bedroom beds bell best biggest bike bouncing church closet country cover cows cube diaper dogs door eye family first game grass green heat horse kind last mirror mommy park people pillow screw screws sill sky sleeping spoons spray starting store stuff suitcase suitcases tag taillights teddy telephone telescope things tigers toilet toys babysitter bacon barrel barrette bat bath cat celery change chest chip choice crayon cream dog doggie donkey drink dresser drill dump earth feeling finger giraffe kiss kite lamb lemon letter lot mafia mail man map mess monkey monster noise night number piece pile pill picnic picture plant plastic plug pocket point ride road round stick strange sweet thumb tire tissue toy traffic wall etc…
Cat$_{2,4,6,8}$	838	0.53	always already almost anyway are ate bang bet blow break bringing calling can coughing could couldn't count did didn't eat eating fitting forgot have haven't having hit knock knocked knocking know just live lock lost missed move moved mustn't need open opened opening ought pack park parking rip saw see seem singing smell speak speaking squeeze spread talk tear understand use using wash were wouldn't write writing wrote, actually alone along anymore apart around away back because before behind belong belongs broken by called came closed comes coming cost disappear does doesn't down empty fall fit for from gets goes has hasn't hurt in inside into is isn't itches keeps look looks made near out outside over ran roll says scare seems to turns under was wasn't went without won won't work would afraid after ago are ask asked before behind bet can could did distract disturb disturbing fit hope if protect showed spoil spray taught tells then thought tickle told understand were will would yesterday actually ate bang believe blowing bother break call closing cook cover covered cut decorate does doing doubt drank eat empty feel feeling fill filled filling flush fold folded forget found gave guess heat hid hide hiding mess mixing move open parking pour pretend push pushed pushing reach rip screw screwing shaved sing spill spit squeeze stir tear tearing unwrap unzipped wash etc…
Cat$_{5,7}$	224	0.14	about answer around at bat bend between blow bringing bouncing by carrying change chasing cleaning clicking count down drinking eating fasten fitting flushed forgot green hears hit hits hitting holding in inside into latch likes mash moving of off onto opening out outside over park point ride rides roll round scare screws should smell to under use working almost always as be been being bought brought called caught cost finding for get gets goes got gotten had has have having hear heard just like live lock looks lost made mail need needs says singing smiles store using wait etc…

Table 27: Compressed Categories

Word	Category Probabilities	
	Nominal	Predicative
airplane	1	
alligator	1	
almost		1
always		1
apple	1	
around		1
ask		1
ate		1
bend		1
bouncing	0.5	0.5
bring		1
by	0.32	0.68
called	0.22	0.87
change	0.45	0.55
closed		1
comes	0.56	0.44
could		1
cover	0.65	0.35
cry		1
decided		1
did		1
do		1
dog	1	
drink	0.31	0.69
fly	0.57	0.43
green	0.9	0.1
grow		1
hungry	0.64	0.36
in	0.34	0.66
mail	0.57	0.43
mess	0.80	0.20
off		1
out		1
over		1
park	0.78	0.22
screw	0.59	0.41
screws	0.5	0.5
spray	0.46	0.54
store	0.85	0.15
use		1
want		1

Table 28: Category Probabilities

It is obvious in Table 27 that the two largest classes $\mathrm{Cat}_{1,3}$ and $\mathrm{Cat}_{2,4,6,8}$ correspond to nominal and attributive elements, and predicative elements, respectively. $\mathrm{Cat}_{5,7}$, on the other hand, contains mainly words that could function predicatively. For ease of reference $\mathrm{Cat}_{2,4,6,8}$, and $\mathrm{Cat}_{5,7}$ will be referred to as Predicative, and $\mathrm{Cat}_{1,3}$ as Nominal. Accordingly, Table 28 shows the updated category probabilities of the words in Table 25.

Having established category probabilities, this information output by the algorithm was then used in tagging and chunking parts of the corpus in the following manner.

(99) a. Tag word W as Predicative everywhere in the corpus if P(W = Predicative) = 1

b. Tag word W as Nominal everywhere in the corpus if P(W = Nominal) = 1

c. Transform the tag Predicative to Nominal in the context (the|a)_

d. Merge any sequence of nominal expressions into one nominal chunk

e. Mark *the* and *a* as the left boundary of a nominal chunk (given the right-associative nature of these two cues).

7.2.3 Evaluation

The overall performance of the implementation was calculated in terms of *precision* and *recall* (Manning and Schütze, 2003:267-269). Precision is defined as $\frac{tp}{tp+fp}$, where *tp* (*true positives*) are the cases the system got right, and *fp* (*false positives*) are the cases the system got wrong. Recall is defined as $\frac{tp}{tp+fn}$, where *fn* (*false negatives*) are the correct cases that were not captured by the system.

In order to compute the precision for the learner's performance in categorization, a POS-tagged version of the Peter corpus in the CHILDES database was used. The categorization algorithm presented in this section managed to collect information about two main classes: Predicative and Nominal, in addition to Nominal Chunks. Accordingly, the tags in the corpus were collapsed into these two classes in the following manner. Verbs, prepositions, predicative adjectives, and particles were merged together as Predicative. This class contained 66733 tokens and 1163 types. Nouns, pronouns, attributive adjectives, and determiners were merged together as Nominal. This class contained 64311 tokens and 1374 types.

The learner was able to identify 722 predicative types totaling 64049 tokens, and 1018 nominal types totaling 39045 tokens. Table 29 summarizes this information, in addition to Nominal Chunks, that was garnered by the algorithm. Table 30 shows the precision and recall ratios for the algorithm's performance in identifying Nominal and Predicative elements.

Potential Tag	**Types**	**Correct** (*tp*)	**Incorrect** *(fp)*	**Tokens**	**Correct** (*tp*)	**Incorrect** *(fp)*
Predicative	722	693	29	64049	60206	3843
Nominal	1018	1018	0	39045	39045	0
Total	1740	1711	29	103094	99251	3843
Nominal Chunk	-	-	-	21898	-	-

Table 29: Summary of knowledge acquired by CBL-2

Potential Tag	**Type Precision**	**Type Recall**	**Token Precision**	**Token Recall**
Predicative	≈ 96%	≈ 60%	≈ 94%	≈ 90%
Nominal	100%	≈ 74%	100%	≈ 61%
Average	≈ 98%	≈ 67%	≈ 97%	≈ 75.5%
Nominal Chunk	-	-	100%	-

Table 30: Precision and Recall for CBL-2

These results reflect some properties of this distributionally-bootstrapped learner. The first is that, similar to the semantically-bootstrapped learner, this learner shows some bias towards learning nominal expressions than predicative expressions, as reflected by the type recall ratios However, this bias is not as strong as it was in the first learner. The other property is that, though this learner was not provided with any initial cues, the type and token recall ratios for this learner are much higher than those for the first learner, which was given names of things and people as initial cues. Note that in the first learner nouns were not grouped with other nominal elements, and verbs were not grouped with other predicative elements, as it is the case with the second learner.

7.2.4 Discussion and Conclusion

In this section, I have presented an algorithm for the initial categorization of the input into two main classes, i.e., arguments and predicates. The main idea behind this algorithm was to categorize words using their distributional similarity as measured in terms of their co-occurrence with a set of cues. These cues were identified in terms of The Distributional Cue Extraction Procedure. Only the 1st-order approximation of cues was utilized in binary categorization. The distributional similarity of words was measured in terms of their distribution with these cues. The strength of association between these cues and other words was measured in terms of mutual information, which was then used to compute the category probabilities of these words. It was shown that this algorithm was able to identify a subset of the arguments and predicates in the input with a relatively high level of accuracy. It was also demonstrated that this algorithm was able to capture in probabilistic terms the fact that some words belong to more than one category, here arguments and predicates, with different degrees of membership. This algorithm has also demonstrated that categories can be identified in the input using a small set of cues that can be distributionally learned from the input. The performance of this first-approximation algorithm stresses the central role of the input and distributional mechanisms in learning. In the following section, I show how the knowledge yielded by this algorithm can be used in frame identification.

7.3 MI-Based Frame Identification

This section details the frame identification algorithm in CBL-2. This algorithm operates according to the following assumptions.

There is no predefined set of possible frames. This algorithm is completely dependent on the information yielded by the head-direction and categorization algorithms. The head algorithm has established English as a head-initial language. Accordingly, the subcategorization algorithm uses this information to limit its search space to the right-hand context in order to find dependents, which reduces the search space significantly. The categorization algorithm provides the subcategorization algorithm with information about possible predicates, arguments, and chunks in the corpus. Because of this dependence, only words that were tagged by the categorization algorithm are visible to the subcategorization algorithm. The frames identified by this algorithm are very general and are not formalized in the conventional terms associated with subcategorization frames, i.e., verbs, verb phrase, prepositional phrase, etc. Rather they are described in the binary distinction between predicative and nominal expressions as established by the categorization algorithm. It assumes a probabilistic approach to the distinction between complements i.e., arguments required by a predicate, and adjuncts, i.e., elements that occur freely with a predicate. Accordingly, this distinction is captured in terms of a probability function over the kinds of dependents expected with a predicate.

This algorithm works with the surface structure with no presuppositions about any *deeper* levels.

7.3.1 Algorithm

The algorithm used in frame identification is a trivial specification of the distributional cue-based procedure for cue extraction introduced in (38) above. Accordingly, the set of possible frames in a given input is expected to be a subset of the contexts where predicates, in general, and verbs, in particular, occur. Given this, the set of possible frames are generally approximated as given in (53), repeated below.

(53) Frame Cues (1)
Let $P = \{p_1,\ldots, p_n\}$ be the set of possible predicates in a corpus R
Let $C = \{c_1,\ldots,c_m\}$ be the set of contexts where the members of P occur, then the set of possible frames in R is the smallest subset, C_f, of C such that every predicate in P occurs at least once in at least one context in C_f.

Taking the above assumptions into consideration, the algorithm identifies frames piecemeal in two steps. The first distinguishes three basic frame types which are the building blocks of other frames. The second step identifies other frames in terms of these building blocks. More complex frames should result from different combinations of the basic frames. This step utilizes the procedure in (53) in order to capture the smallest number of possible frame combinations that could capture the frame properties of the largest number of predicates.

7.3.1.1 Basic Frames

Given the two word classes identified in the corpus, i.e., nominal and predicative, three basic frames can be distinguished: nominal frames (i.e., frames that begin with a nominal expression, *NE*), predicative frames (i.e., frames that begin with a predicative expression, *PE*), and zero frame (i.e., when a predicate occurs utterance-finally).[34] These basic frames provide the building blocks from which other frames are constructed.

That is, if a predicate PE_i is immediately followed by another predicate PE_j, the first predicate is said to select a predicative frame. If PE_i is immediately followed by a nominal expression, PE_i is said to select a nominal frame. If PE_i occurs utterance-finally, it is said to select a zero frame. This means that for every predicate PE_i, there are three possible frame types it can select from. For example, a predicate such as *want* can select nominal, predicative, and zero frames: PE[*want*] NE[mommy], PE[*want*] PE[to], and PE[*want*] #, respectively.

Predicates are expected to be different vis-à-vis the number of basic frames they select. For example, some predicates tend to select the three types (e.g., *want*), some prefer only nominal frames (e.g., *answer*), some favor both predicative and zero frames (e.g., *go*), others can occur with both nominal and predicative frames (e.g. *to*), and so on and so forth.

It is also expected that predicates should show different degrees of frame selection. For example, *want* selects a predicative frame more frequently than a nominal or a zero frame, whereas *build* shows the opposite preference and selects a nominal frame more frequently than a predicative or a zero frame.

To give concrete examples of the frame information that can be garnered using this simple strategy, Table 31 shows the frame preferences of 100 of the possible 1062

[34] These terms are used throughout this chapter in order to avoid any terminological confusion that could result from using the traditional terms. All terms will be used as defined. NE and PE refer to the binary classes that were identified by the categorization algorithm in the previous chapter. Remember that nominal expressions include all the elements that belong to Cat1,3 (prenominal expressions), while predicative expressions contain all the elements that belong to the other class (verbs, prepositions, particles, etc…).

predicates identified by the categorization algorithm. The ratios in Table 31 represent frame preference in terms of relative frequency. Relative frequency is computed by dividing the number of times a predicate selects a basic frame, by the absolute frequency of the corresponding predicate. These ratios were computed using only tagged words, i.e., if a predicate was followed by an untagged word, this co-occurrence was not added to any of these frames. That is why some of these ratios do not add up to 1.

Though still very rudimentary, the ratios in Table 31 show some interesting regularities as well as irregularities. The most significant observation in this context is that these ratios clearly capture the transitivity distinction among predicates. Transitivity here is defined as the tendency of the predicate to select nominal frames as defined above. Thus, these ratios demonstrate that predicates such as *drink, answer, change, gave, fix, draw, asked, believe, bring, cut, build,* and *found* are predominantly transitive. On the other hand, predicates such *fall, gonna, come, came, belong, go, goes, live, look, run, seem, stay, try,* and *went*, are predominantly intransitive.

However, these ratios are not able to capture the fine-grained idiosyncratic properties of some predicates. To illustrate this, I consider in more detail the cases of the predicates *buy, change,* and *in*.

In the case of *buy,* the ratios give the impression that this predicate does not occur transitively, though the corpus used provides information that can be used to establish the transitivity of this predicate. The source of this shortcoming was that the context that should have provided this information was not tagged by the categorization algorithm.

Predicate	**__PE**	**__NE**	**__Zero**	**Predicate**	**__PE**	**__NE**	**__Zero**
answer		0.88		know	0.25	0.04	0.23
ask	0.08	0.22		leave	0.02	0.63	0.01
asked		0.67		left	0.14	0.47	0.16
ate		0.33	0.17	like	0.2	0.28	0.03
believe		0.67		live	0.67	0.03	0.17
belong	0.69	0	0.31	look	0.76	0.03	0.05
break		0.5	0.23	make	0.03	0.7	0.03
bring	0.03	0.65	0.01	need	0.17	0.31	0.07
build	0.02	0.6	0.09	open	0.05	0.52	0.19
buy	0.16	0	0.26	play	0.66	0.19	0.05
came	0.85	0	0.1	pull	0.01	0.65	0.04
carry	0.52	0.04		push	0.01	0.74	0.04
catch		0.39	0.36	put	0.04	0.6	
change		0.86		read	0.04	0.58	0.04
come	0.85	0.01	0.08	remember	0.26	0.16	0.17
cry	0.32	0	0.63	ride	0.28	0.5	
cut	0.03	0.63	0.16	run	0.78	0.09	0.09
draw	0.05	0.73	0.11	said	0.2	0.16	0.31
drink		1		saw	0.08	0.5	0.17
drive	0.15	0.52	0.21	say	0.16	0.13	0.33
dropped	0.1	0.5		screw		1	
dump		0.2		see	0.14	0.3	0.21
eat	0.14	0.3	0.13	seem	0.82	0.09	
fall	0.85	0.02	0.11	show	0.04	0.36	0.02
feel	0.32	0.47	0.05	sing	0.08	0.25	0.48
fell	0.92	0.03	0.04	sit	0.83	0.01	0.13
find	0.06	0.56	0.04	sleep	0.41		0.55
finish	0.13	0.24	0.18	squeeze	0.05	0.6	0.05
fix	0.02	0.78	0.02	stand	0.85	0.07	0.01
for	0.05	0.44	0.11	start	0.5	0.27	0.13
found	0.02	0.57	0.07	stay	0.81		0.1
gave		0.79		take	0.06	0.58	
get	0.3	0.39	0.03	tell	0.02	0.2	0.02
gets	0.71	0.21		think	0.33	0.21	0.06
give	0.04	0.51		thought	0.06	0.54	0.07
go	0.67	0.03	0.18	throw	0.02	0.63	0.07
goes	0.63	0.14	0.13	to	0.69	0.13	0.02
going	0.66	0.04	0.23	took	0.04	0.68	0.02
gonna	0.9	0.02		touch		0.59	
got	0.21	0.39	0.05	try	0.43	0.3	0.14
guess	0.22	0.17	0.18	turn	0.18	0.71	0.05
had	0.24	0.37	0.08	turned	0.32	0.61	
has	0.21	0.57	0.03	unscrew		0.91	
have	0.34	0.32	0.05	use	0.04	0.49	0.03
hear	0.04	0.36	0.07	used	0.67		0.04
help	0.05	0.26	0.31	wait	0.36	0.31	0.08
hit	0.11	0.4	0.07	want	0.16	0.1	0.02
hold	0.13	0.56		wanted	0.63	0.13	0.08
in	0.21	0.50	0.06	watch	0.15	0.42	0.08
keep	0.35	0.4	0.02	went	0.86	0.03	0.01

Table 31: Relative Frequencies of Basic Frames

As for the predicate *change*, the ratios do not capture the fact that this predicate can also be intransitive. However, by checking the corpus used, there was no single context where this predicate occurred intransitively. Though this conclusion touches upon the issue of data sparseness in corpus-based approaches, the significance of this conclusion in this context is that it does not illustrate a weakness either in the categorization algorithm or the algorithm proposed so far for identifying basic frames. Given the humble size of the corpus used, using a larger corpus is expected to improve the results.

As for the predicate *in*, the irregularities were the direct result of errors in categorization. These ratios reflect that this predicate can select a predicative frame in 0.21 of its occurrences. By checking the corpus, it was found that this conclusion is erroneous and resulted from falsely categorizing the words *here* and *there* as possible predicates.

Despite these shortcomings, it is important to remember here that these ratios are based on the first approximation of the category cues in a humble-size corpus, as was discussed in the previous chapter. Though the results so far are not insignificant, they can be easily improved using a larger corpus and a higher approximation of cues. These ratios were meant to illustrate the frame regularities that could be trivially captured by directly using the information yielded by the categorization algorithm. Below, I will show how these basic frames can be exploited in grasping more fine-grained frames, using MI as a measure of association.

7.3.1.2 Complex Frames

The main idea behind identifying complex frames is that a complex frame should result from the concatenation of two or more basic frames. That is, given the basic frames *NE* and *PE*, a complex frame can be any member in the set {(*NE PE*), (*NE NE*), (*PE PE*), (*PE NE NE*), (*NE PE NE*), etc…}. Figure 9 visualizes this concatenation process, where the utterance boundary sign, #, indicates the end-point of a frame.

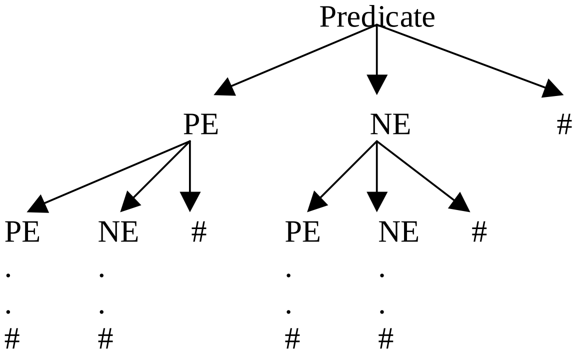


Figure 9: Concatenation of basic frames

For example, a predicate such as *want* would select combinations such as (*PE PE*), (*PE PE PE*), and (*NE PE PE*), as exemplified by the structures *(PE[to] PE[go])*, (*PE[to] PE[scare] NE[mommy]),* and *(NE[mommy] PE[to] PE[stand]),* respectively.

In principle, this strategy would definitely result in a large set of possible frames, and would also fail to capture the relationship between predicates that have similar frame preferences. I will introduce below a formal criterion to limit the set of possible frames to those permitted in a given language, as supported by evidence from the regularities in the corpus.

7.3.1.3 Frame Cues

Possible frames in a given language can be identified using a subset of the possible combinations of basic frames as discussed above. What is needed then is a subset of these combinations that approximates the frame behavior of the predicates in any given corpus. This subset will be referred to as "Frame Cues", K^f, henceforth, and is defined as follows:

> **(100) Frame Cues (K^f) (2)**
> Frame Cues, K^f, are the members of the smallest subset of the possible combinations of basic frames ($f_1,\ldots,f_m$) such that every predicate in the corpus occurs at least once with at least one member in K^f.

Similar to category cue extraction, this subset K^f can be identified using two different methods. The only difference here is that the search space for possible frames is limited to the right context as biased by the head direction in English that was established previously by the head-direction algorithm.

The first method is to extract for every predicate, PE_i, in a corpus, R, the set of basic-frame combinations, f_i, that PE_i selects. K^f for corpus R, K_R^f, is then the intersection set of all the sets $\{f_1,\ldots f_M\}$, where M is the number of predicates in the corpus that were identified by the categorization algorithm. This method is formalized as follows:

(101) Let R be a corpus with M predicates $\{PE_1,\ldots PE_M\}$,
Let $\{f_1,\ldots f_M\}$ be the set of basic-frame combinations, $\{PE_1,\ldots PE_M\}$ select, respectively,
then $K_R^f \equiv \{f_1 \cap f_2 \cap f_3 \cap\ldots\cap f_M\}$

As mentioned before finding such minimal subsets is known to be intractable. Accordingly, what is introduced here is an approximation of this subset.

The method proposed here makes a direct use of the highly frequent frame combinations in order to approximate K^f. We start with building a decreasing frequency profile for all the possible frame combinations $\{f_1,\ldots f_M\}$, in a corpus, R. We then add up the number of predicates, X_1, that select the first most frequent combination, f_1, and the number of predicates, X_2, that select the first most frequent combination, *f2*, and so on. K^f for corpus *R*, K_R^f, is then the set of frame combinations $\{f_1,\ldots f_x\}$, such that:

(102) a. $\{f_1,\ldots, f_x\} \subset \{f_M\}$

b. $K_R^f \equiv \sum_{i=1}^{x} X_i \cong \alpha M$

where α can be any positive number, and M is the number of predicates identified in the corpus. In words, K_R^f is established as the smallest subset of basic-frame combinations in the corpus selected by a number of predicates that converges to an order, α, of *M* (i.e. *1M*, *2M*, etc…). Similar to Category Cues, closer approximations of K_R^f can be obtained by increasing the value of α. Only the first approximation, i.e., α = 1, is implemented in the present algorithm. It is shown below that the algorithm performs well with this approximation. The efficiency of higher approximations is left for future investigation.

Once frame cues are identified, the pieces making up these frames are compressed into one entity with no internal structure. That is, if a frame contains the pieces X and Y, these two pieces are treated as one piece, X_Y. All the calculations below are performed accordingly.

7.3.1.4 MI-based Frame Identification

Once K^f is identified, we can compute the probability that a given predicate selects any of the frames in K^f in the following manner.

Let the frame cues for corpus *R* be $K_R^f \equiv \{f_1,\ldots,f_x\}$, where *x* is the number of frame combinations in K_R^f. For every predicate PE_i in *R*, we first extract the frame combinations it selects. This means that the maximum number of frames for a predicate PE_i is x, and the minimum is 1. This is visualized in Figure 10.

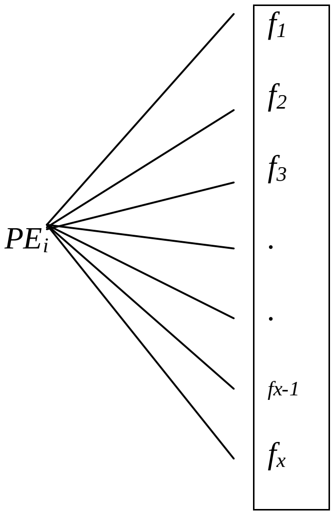


Figure 10: Representation of Frame Distribution Contexts

The probability that a certain predicate selects a given frame is established in terms of MI in the following fashion. We first compute the MI between the predicate and every frame it occurs in, using the same MI formula, repeated in (103).

(103) $$I(x;y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

This formula is used in this context in the following manner. Given a predicate *PE* and a frame *f*, the probabilities in (103) are computed as follows:

(104) a. $$P(PE, f) = \frac{frequency(PE, f)}{T}$$

b. $$P(PE) = \frac{frequency(PE)}{T}$$

c. $$P(f) = \frac{frequency(f)}{T}$$

where *T* is the number of tagged words in the corpus.

The decision that a given predicate selects one or more frames is probabilistic. The frame probabilities of a particular predicate are computed in terms of the positive mutual information between the predicate and the frames it selects, in the following manner. Given a predicate *PE* which co-occurs with the frames [*f1, f2,…,f$_x$*], the probability that *PE* selects any of these frames, f_i, is computed by dividing the mutual information between *PE* and f_i by the sum of the mutual information between *PE* and all the frames it selects (105).

(105) $$P_s(PE, f_i) \approx \frac{I(PE; f_i)}{\sum_{i=1}^{x} I(PE; f_i)}$$

If the mutual information between the predicate and any frame is negative or less than zero, this is an indication that this predicate does not select this frame, and this mutual information is excluded from the formula in (105).

7.3.2 Experiment

7.3.2.1 Corpus Description

The frame identification algorithm described in the previous section was tested on English using the same Peter corpus used in the two previous experiments. Unlike other experiments, this algorithm operates on a binary-tagged corpus output by the categorization algorithm. That is, this corpus contains words that are tagged as predicate expressions (*PE*), and nominal expressions (*NE*).

7.3.2.2 Results

The algorithm was applied in three phases. The first phase built a decreasing frequency profile for all the possible frame combinations in the binary-tagged corpus. The second phase established the frame-cues subset, K^f. In the third phase, K^f was used to establish the frames a predicate selects, and to then assign probabilities to these selections, according to (105).

7.3.2.2.1 Frame Cues

The first phase of the algorithm yielded ≈ *200* possible frame combinations. Table 32 shows the ten most frequent frame combinations and their frequencies in the profile. The frequencies in Table 32 indicate the number of predicate tokens that occur in the corresponding frame. (To understand what these combinations, the reader could interpret *PE* as a verb, a preposition, a particle, or a predicative adjective, and *NE* simply as a noun phrase, NP.) In the second phase, the algorithm converged to the first order of *M*, i.e., the number of tagged predicates in the corpus, after the 8th frame combination, i.e., *NE NE*.

	Frame	**Frequency**	
1	PE	941	
2	PE NE	811	
3	NE PE	691	
4	PE PE	682	
5	NE PE NE	471	
6	Zero	402	
7	NE	366	
8	NE NE	144	←
9	PE PE NE PE	110	
10	NE PE NE PE	99	

Table 32: The 10 most frequent frame combinations in the corpus

Table 33 shows the frame-cues subset, K^f for the corpus, and some examples of the predicates in each frame. The distribution of verbs among these frames shows some frame regularities that deserve some comments.

	Frame	**Examples**
1	Zero	afraid apart ate awake bend bet blow breathe come dance disappear drinking fell finished flying go grow guess hurt jump left live look play reading resting etc…
2	PE	are back be came can could do don't drive fall fell flying get go goes gonna keep look should sit stand stay to turn wake was watch went won't etc…
3	PE PE	go going gonna got have like must need ready supposed to trying want wanted wants won't etc…
4	NE	about answer bless bring build carry catch close do draw drink eat feel find fix found get hear help hit leave like make need open play pull read riding saw etc…
5	PE NE	afraid are bang be belong bite blow broke came can come could fell finish get go gonna jump listening look looking move play put ran rode see sit speak stand etc…
6	NE PE NE	ask bring build drive find fix get give leave mail make put throw write wrote etc…
7	NE PE	believe blow bring broke can carry cut did do does drive fix fold get guess hold leave lift like make pick pull push put etc…
8	NE NE	ask bring brought call called gave give made make read sing taught write etc…
	Total	**4508**

Table 33: Examples of Predicates in different frames

The first frame encompasses predicative adjectives (e.g., *afraid* and *awake*), verbs that do not require complements at all (e.g., *breathe, come, disappear, fell, go,* etc…), and verbs that can occur with or without surface complements (e.g., *ate, drinking, reading,* etc…)[35]. On the other hand, the second frame, PE, where a predicate selects another predicate as its complement, includes mainly auxiliary verbs as well as phrasal verbs. The third frame, also contains auxiliary verbs, in addition to verbs that usually

[35] These are verbs that allow what is traditionally referred to as NP-Drop.

select infinitival complements (e.g., *like*, *need, trying, want*, etc…). The fourth frame is predominantly selected by predicates (prepositions and verbs) that are primarily transitive. The fifth frame is mainly preferred by predicates that tend to select prepositional phrases as complements, in addition to some phrasal verbs. The sixth frame is obviously selected by verbs that require the traditional NP PP complement. The predicates in the seventh frame do not show an overall frame regularity, yet some sub-regularity can be distinguished. This frame contains verbs that could select a tensed clause as a complement (e.g., *believe* and *guess*), and transitive phrasal verbs (e.g., *bring, carry, get, lift, make, pick, pull, push,* and *put*). The eighth frame noticeably contains ditransitive verbs.

7.3.2.2.2 MI-Based Frame Preference

Though these results capture significant frame properties of a considerable portion of the predicates identified, they do not provide accurate conclusions regarding the complement-adjunct distinction by simply using the mere co-occurrence of a predicate with a given frame as a measure of selection. Therefore, the mutual information between the predicate and the frames it co-occurs with was used to obtain more solid conclusions about frame preferences. The mutual information was computed according to the formulas in (103) and (104) above. Table 34 illustrates the frame preference of some predicates as measured in MI terms. In Table 34, '_' indicates the position of the corresponding predicate. For example, the predicate *answer* selects a nominal expression, *NE,* and their mutual information is 6.32, and so on. The negative mutual information is given in this table to demonstrate the point that though some predicates co-occur with certain frames in the corpus, this co-occurrence could be the result of a chance factor and consequently does not provide enough evidence for selection.

Predicate	**_ PE**	**_ PE NE**	**_ NE PE**	**_ PE PE**	**_ NE PE NE**	**_ Zero**	**_ NE**	**_ NE NE**
answer	0	0	0	0	0	0	6.32	0
ask	0	0	2.57	0	0	0	1.12	4.57
belong	3.27	4.57	0	0.19	0	1.59	0	0
bring	0	2.63	3.5	0	2.37	-4	4.13	3.87
build	0	2.8	0.8	0	3.54	-0.12	4.71	0
carry	0	3.31	3.63	0	0	0	4.3	0
catch	0	0	0	0	0	1.82	5.03	0
come	3.14	2.88	0	3.6	0	-0.29	-0.54	0
draw	0.32	0	0.04	0	0	0.15	5.98	0
drink	0	0	0	0	0	0	6.25	0
dump	0	0	0	4.39	0	0	5.41	0
eat	0.68	2.63	0.13	0	2.57	0.32	4.16	0
fall	3.89	0	0	3.65	0	0.1	-1.18	0
feel	2.25	0	3.57	3.65	3.53	-0.96	5.07	0
find	0.71	3.1	3.1	0	3.55	-1.47	5.11	3.43
finish	2.08	1.54	0	0	0	0.85	4.07	0
fix	0	0	1.82	0	3.37	-2.72	5.61	0
get	2.39	3.1	3.2	3.11	2.69	-1.84	4.04	0.98
give	1.1	0	2.61	0	3	0	0.22	5.48
go	2.71	4.07	0	3.33	1.59	0.81	-0.17	0
guess	2.23	2.89	3.18	2.4	3.07	0.83	0.47	0
hear	0	2.98	3.28	0	0	-0.52	4.07	0
hold	1.72	0	3.61	0	2.78	0	4.79	0
keep	2.36	2.72	3.22	2.75	3.49	-2.13	2.9	0
know	1.1	1.62	0	0	1.89	1.16	-0.55	0
leave	0.52	0	4.26	0	2.72	-3.8	1.97	0
like	0.63	1.39	2.18	2.93	2.89	-1.92	3.92	0.97
look	1.74	3.99	0	2.58	1.95	-0.97	-0.63	0
need	1.43	1.95	2.95	1.42	2.45	-0.49	3.89	0
put	0.16	1.91	3.86	0.93	3.56	-5.51	0.8	0
seem	0	0	0	3.58	0	0	0	0
think	2.1	3.17	2.68	2.77	2.2	-0.69	-3.24	0
try	1.9	0	3.92	2.68	0	0.48	3.61	0
turn	1.98	3.11	4.53	0	2.22	-0.96	3.12	2.65
want	0.95	1.07	1.94	1.82	2.6	-2.26	1.66	-3.34

Table 34: Some verbs and their frame MI

The results in Table 34 are self-evident. However, to realize the efficiency of mutual information as a measure of frame preference, I discuss in some detail the frame preferences of some representative verbs from Table 34.

For example, verbs such as *answer* and *drink* can be safely established as almost equally mono-transitive. They have almost the same mutual information with *NE*, and show no preference for other frames, as clear from the zero mutual information. This preference was captured by the subcategorization algorithm as a direct result of the fact that these two verbs occurred 8 and 7 times, respectively, in the corpus, and were followed by a nominal expression in all their occurrences.

Verbs such as *ask* and *give* show a preference for the *NE NE* frame as reflected in the high mutual information between these two verbs and this frame, i.e., 4.57 and 5.48, respectively. Though such a preference is expected, it illustrates an important attribute of mutual information as a measure of association, i.e., assigning high values to rare events. This can be clarified by considering the distribution of these two verbs in the corpus. The first verb, *ask*, occurred 49 times in the corpus. In only 15 contexts, this verb was followed by a tagged word: 11 nominal, and 4 predicative. Out of these 11 nominal contexts, it occurred only once with the *NE NE* frame, which yielded the high mutual information in Table 34. On the other hand, *give* occurred 183 times in the corpus. In only 100 contexts, this verb was followed by a tagged word: 93 nominal, and 7 predicative. Out of the 93 nominal contexts, *give* occurred only 7 times with the *NE NE* frame, which resulted in the high mutual information in Table 34.

This attribute of the mutual information statistic could be tolerated if we take into consideration the fact that these results are based on a humble-size corpus, and on first-order approximations of both categories and frames. Using a larger corpus and closer approximations should establish more solid results.

Below, the last phase in the subcategorization algorithm shows how the mutual information in Table 34 can be used in deriving frame probabilities that better represent the frame preference of predicates.

7.3.2.2.3 Frame Probabilities

Frame probabilities were computed according to the formula introduced earlier in (105), repeated below for ease of reference.

$$(105) \quad P_s(PE, f_i) \approx \frac{I(PE; f_i)}{\sum_{i=1}^{x} I(PE; f_i)}$$

According to (105), the probability that a predicate selects a certain frame is computed from positive mutual information only, by dividing the mutual information between the predicate and the frame by the sum of the mutual information between this predicate and all the frames it selects. Excluding negative and zero mutual information from probability calculation resulted in frame probabilities that would reduce the frame noise for some predicates. This produced a clearer picture of the frame preferences of predicates as illustrated by the frame probability distribution of some predicates/verbs in Table 35. These probability distributions reflect some properties of the frame behavior of these verbs. The first is that these probabilities reflect the fact that some verbs show a clear bias towards some frames. This is clear from the probability distributions for verbs such as *answer, ask, belong, carry, catch, come, draw, dump, fall, finish, fix, give, hear, leave* and *seem*, where the probabilities are concentrated in a relatively small number of frames, with one or two frames carrying probabilities obviously higher than other frames.

Predicate	_ PE	_ PE NE	_ NE PE	_ PE PE	_ NE PE NE	_ Zero	_ NE	_ NE NE
answer							1	
ask			0.31				0.14	0.55
belong	0.34	0.48		0.02		0.17		
bring		0.16	0.21		0.14		0.25	0.23
build		0.24	0.07		0.3		0.4	
carry		0.29	0.32				0.38	
catch						0.27	0.73	
come	0.33	0.3		0.37				
draw	0.05		0.01			0.02	0.92	
drink							1	
dump				0.45			0.55	
eat	0.06	0.25	0.01		0.24	0.03	0.4	
fall	0.51			0.48		0.01		
feel	0.12		0.2	0.2	0.2		0.28	
find	0.04	0.16	0.16		0.19		0.27	0.18
finish	0.24	0.18				0.1	0.48	
fix			0.17		0.31		0.52	
get	0.12	0.16	0.16	0.16	0.14		0.21	0.05
give	0.09		0.21		0.24		0.02	0.44
go	0.22	0.33		0.27	0.13	0.06		
guess	0.15	0.19	0.21	0.16	0.2	0.06	0.03	
hear		0.29	0.32				0.39	
hold	0.13		0.28		0.22		0.37	
keep	0.14	0.16	0.18	0.16	0.2		0.17	
know	0.19	0.28			0.33	0.2		
leave	0.05		0.45		0.29		0.21	
like	0.04	0.09	0.15	0.2	0.19		0.26	0.07
look	0.17	0.39		0.25	0.19			
need	0.1	0.14	0.21	0.1	0.17		0.28	
put	0.01	0.17	0.34	0.08	0.32		0.07	
seem				1				
think	0.16	0.25	0.21	0.21	0.17			
try	0.15		0.31	0.21		0.04	0.29	
turn	0.11	0.18	0.26		0.13		0.18	0.15
want	0.09	0.11	0.19	0.18	0.26		0.17	

Table 35: Some verbs and their frame probabilities

The other observation is that some verbs show ‘frame dispersion’, where probabilities are closely dispersed among a relatively large number of frames. The verbs *get* and *guess* are good examples of these verbs.

The performance of the learner in these representative cases is suggestive of its overall performance in frame learning in general. However, the overall performance of this learner was difficult to evaluate given the fact that the tagged version of the corpus used did not contain frame information. Future investigation is still needed to measure the performance of this learner in frame learning.

7.4 Discussion and Conclusions

This chapter presented the specifics of a distributionally-bootstrapped cue-based learner (CBL-2). CBL-2 comprised three main algorithms. The first presented a simple cue-based method for predicting the head direction given a small size corpus. The logic behind starting with this algorithm was that important structural properties of language should follow naturally from information about head direction.

The second algorithm presented another cue-based method for the identification of predicates and arguments. This method was mainly based on a procedure for learning a set of cues from the corpus. This set was then used to capture the distributional similarity of other words in the corpus. Similarity was based on the strength of the association between a given word and the members of the cue set. Using mutual information to establish degrees of association, this algorithm was able to differentiate the words in the corpus into two main classes, i.e., predicates and arguments. Equipped with the information provided by the first two algorithms, the last algorithm exploited this knowledge to determine the most probabilistic syntactic frames that best describe the lexical syntactic properties of the predicates identified. Together, these three algorithms presented a generalized, cue-based, and language-independent system for grammar induction, in general, and frame identification, in particular.

Unlike CBL-1, the performance of CBL-2 in frame learning could not be evaluated as a function of its performance in head-parameter setting and binary categorization. However, CBL-2 performance in the previous tasks may give an idea of how well this learner is able to handle frames. This said, though both learners achieved almost the same level of performance in the categorization task, CBL-2 presented a more sophisticated implementation of the cue-based model laid out in Chapter 5, that was able to circumvent the weaknesses of CBL-1.

Firstly, CBL-2 learned the cues from the corpus, unlike CBL-1, which was given the semantic cues as its initial free knowledge. Secondly, CBL-2 contained a head-direction setting algorithm that justified searching for complements in the predicate's right-side context. CBL-1, on the other hand, arbitrarily looked for complements in this context. Finally, CBL-2 was more efficient in capturing the gradient nature of some linguistic knowledge. For example, it was able to assign words to their possible categories in terms of the probability distributions of words over these categories. Likewise, it was able to capture the frame preference of predicates in similar probabilistic terms.

These properties of the distributionally-bootstrapped cue-based learner (CBL-2) give it some descriptive edge as a preferable direction for future work on cue-based learning in general, since it offers a mechanism for learning much, given little.

Chapter 8

General Discussion and Conclusions

This dissertation had two objectives. The first immediate objective was to present a cue-based distributional approach to frame identification. The other more general objective was to show the central role of the input in human and automatic language acquisition.

The basic idea behind this approach is that there is a subset of words, i.e. cues, in the input that can be used in bootstrapping significant parts of the grammar of the input language. This subset was formally defined as the smallest subset of words in the input that carry significant information about the distributional properties of other words in the input. It was shown that an approximation of these cues could be distributionally extracted from the corpus.

Building on this basic idea, this dissertation introduced the foundations of a cue-based distributional learning model. These foundations comprised three central procedures for learning cues from the input, establishing distributional similarity in terms of these cues, and identifying frames using the information yielded by the previous procedures.

Two proof-of-concept implementations were presented to test the plausibility of this model. The first implementation was a semantically-bootstrapped cue-based learner (CBL-1) and was based on a set of semantic cues in the input. The other implementation was a distributionally-bootstrapped cue-based learner (CBL-2) and was built on top of a set cues that were automatically learned from a corpus.

Because of the fact that no machine learner has access to objects in the world, CBL-1 was seeded with names of things and people and a small subset of pronouns in the corpus. Using these bootstraps, CBL-1 was able to categorize a considerable subset of words in the input into verbs, nouns, and determiners. CBL-1 then used these categories in learning phrase structure rules and 38 possible subcategorization frames. This learner was able to identify this information with an average type and token precision of 97% and 98.5%, respectively.

In spite of this promising level of performance, this learner suffered from three main weaknesses. The first was its inability to learn semantic cues from the input for the reason mentioned above. The second and more serious problem was its inability to capture lexical ambiguity, where a word could be assigned to more than one part of speech. The third drawback was that this learner was unjustifiably biased toward searching for a verb's complements in its right-side contexts.

The other learner, CBL-2, circumvented these problems by learning cues and head direction based on distributional regularity in the input. Using this information, this learner was able to (i) differentiate words in the input into two main classes, i.e., predicative and nominal, and (ii) to identify 8 subcategorization frames. This learner achieved an average type and token precision in these tasks of 98% and 97%, respectively. Though this average is almost the same as that of CBL-1, CBL-2 has an edge of capturing this information in more probabilistic terms than CBL-1. This way CBL-2 was able to account for lexical ambiguity in addition to degrees of frame preference of different predicates.

The overall performance of these two cue-based distributional learners raises some practical as well as theoretical issues in language acquisition.

Previous approaches to automatic lexical knowledge acquisition assumed part-of-speech tags, partial parses, or a predefined set of cues as their initial knowledge. The two cue-based learners presented here showed that categorization and subcategorization knowledge can be bootstrapped from an untagged corpus, using minimal or even no *a priori* linguistic knowledge. It was shown that bootstrapping this knowledge was possible given a set of cues that could be identified automatically in any given corpus.

Moreover, previous methods for automatic subcategorization frame acquisition also assumed a predefined set of possible frames. The work presented in this dissertation demonstrated that this is not necessary since an approximation of the set of possible frames in a language could be learned distributionally from corpora.

The practical significance of this cue-based approach to knowledge acquisition stems from (i) the evidence it gives to the efficiency of cue-based learning in bootstrapping lexical knowledge from minimal initial knowledge, and from (ii) the formal procedures it introduced for identifying and learning cues and frames from untagged corpora.

The theoretical significance of the work presented in this dissertation resides in the support it gives to the central role of the input in learning and, consequently, to the empirical approaches to language acquisition. It was shown that the input is seeded with a rich set of distributional cues that are accessible to the learner, and that these cues could be easily extracted from a small-size corpus, using simple distributional learning mechanisms. It was also demonstrated that these cues provide the learner with scaffoldings that can be exploited in bootstrapping into the target language. For example, it was shown that the head parameter could be set distributionally on the basis of distributional regularities in a small-size corpus. The efficiency and language-independent nature of the head-parameter setting of the model was partially demonstrated using three different languages: English, Japanese, and German.

The evidence provided by this cue-based model is not meant to deny or assert the existence or centrality of innate knowledge in acquisition, it rather stresses the central role of the input and distributional methods in learning. The ability of the model to accurately capture this parametric variation illustrates its potentiality as a useful and objective tool for the study of (automatic) lexical acquisition, in particular, and grammar construction, in general, in less examined languages.

The model introduced in this dissertation is a pointer to a possible direction of cue-based distributional learning that is worth considering. Future research is still required to investigate the feasibility of this model in learning other aspects of linguistic knowledge such as word segmentation and morphology. Future investigation is also needed to test the efficiency of this model with different languages, particularly those with rich morphological systems. The model is falsifiable, which facilitates the process of testing its feasibility in these tasks as well as revising it in order to increase its efficiency.

Appendices

Appendix A: Words Identified as Potential Nouns by CBL-1

abcs accident address air airplane airport alphabet ambulance animal animals anteater appetite apple apron ark arm arms arrow ay baby back bacon bag bags ball balloon balloons balls banana bandage bandaid barn barrel barrels barrette baseball basket basketball bat bath bathroom bathtub battery beach beagle bear beard beaver bed bedroom beds beginning bell belt bench bendables bicycle biggest bike bird bit bite blackboard blanket block blocks blue board boat boats body bologna bolt bolts book bookcase books boots both bottle bottom bounce bowl box boxes boy boys bracelet breadstick bridge briefcase brother bud buildings bull bulldozer bunny bus button buttons cake calf camera can candy car card carpet carriage cars case castle cat cats ceiling cereal chain chair chairs change checkbook cheek chicken chickens children chimney chin chip church circle circus class clean closet clown coat coats cold collector colt comb cookie corner couch counter country cover cow cowboy cows crack cradle crash crayon crayons crown crumbs crumpled cube cup cushion daddy dark day days deal desk detour diaper diapers difference dinner disaster doctor dog doggie doll dolphin donkey door doorbell downstairs drawer dress dresser drill drum drums dry dryer duck duckie dumping ear ears effort egg eight elbow elbows elephant elevator end engine envelope explosion eye eyebrow eyes face faces fan farmer farmhouse father favorite feeling feet fence ferry few field finger fingernail fingers fire firehouse fireman fireplace fish fisherman flag flood floor flower flowers fly food foot football foremans fork frescade friend friends Frisbee frog front fuel furniture game garage gas giraffe girl girls glass glasses gloves godfather going gonna goodness gorilla gosh grandma grass grasshopper green ground guest guitar gum guy hair hall hallway hamburger hammer hand handle hands hard harmonica hat head headlights hear hearing heat heater help hiccups hippopotamus hole holes home hood hook horn horse horses horsie horsies hose hospital hour house houses hug humidifier hurry hurt husband iceberg idea in ink inside itch jack jacket jersey juice key keys kid kind king kiss kitchen kite kittens Kleenex knife label labels ladder lady lamb lambs lamp laundry least leaves leg legs let letter letters lid lifesaver light lights line lion little look lot luggage lunch machine magazine mail mailbox mailman mama man map mark mask matter meal men mess message microphone microscope middle milk mind minute mirror mistake mittens moment mommy money monkey monster morning most mother motor motorcycle mountain mouse mouth mover music mustache nails name nap napkin neck necklace newspaper nice night noise nose number numbers ocean office on one ones opening operator orange organizer other others outside ow owl own pad page pages pants paper papers parade pardon park party past pedals pegs pen pencil pencils penis penny pens people peoples person pheasant phone piano picnic picture pictures piece pieces pill pillow pinkie pipe place plate playground playhouse playing pliers plug pocket pocketbook pocketbooks point pole policeman pool position pot potty pouch present presents pretzel pretzels problem propeller protest puppet puppets puppy push put puzzle quadracycle question race radiator raft railroad rain raincoat record recorder rectangle red reel reels refrigerator reindeer rest restaurant rhino rhinoceros ride right ring rings road rodeo rollerskate roof room rooms rosie rosy row rug ruler saddle sailboat same sand sandwich saw saxophone scale scarf school scissors scoop scooter scream screen screw screwdriver screws search seat second secret see seesaw seven shades sheep shell ship shirt shoe shoes shore shorts shoulder shower side sides sidewalk sink siren sister sitting size skates sky sled sleigh slide slipper slippers smoke snack snap sneaker sneakers sneeze snorkel sock socks sofa soldier somersault song soup space spanking special spoon spoons spot spray square squeaker squirrel stamp stand star station step steps stick stomach stool stoplight store story stove strange street string stroller stuff subway suit suitcase suitcases summer sun suntan surprise sweetie swim swing sword table tables tabs tag tail taillights tape taperecorder tapes taxi teeth telephone telescope television tent test thing thingamajig things throat thumb ticket ticketman ticktock tiger time tire tires tissue toast toe toes toll tongue tool tools toothbrush top towel tower towers toy toys track tracks tractor trailer train trains trash tray tree trees triangle trick tricycle trip truck trumpet trunk try tube tunnel turkey turn turtle tv uh um umbrella uncle under underpants vacation valentine vehicle violin wagon walk walkietalkie wall wallet walls want wash watch water watermelon waves way weasel weekend weenie what wheel wheels where while whistle who window windows windshield wire wolf woman word work world worm wrench wrist wristwatch write writing yawn yours yyy zebra zigzag zoo

Appendix 2: Words Identified as Potential Verbs by CBL-1

already always am answer are ask be been before beg belongs bend better blow borrow both break breathe bring bringing broke brought build buy call calls came can care carry catch caught change changed choke choose cleans climb close closing color come complete confine could counting cry cut did die diversified do does draw drew drink drive drop dropped dry eat either even ever fall feed feel figure fill find finish fit fits fix fold forgot forgotten found gave get give go goes going gone got gotten guess had hammer happened hardly has have having hear heard help helped hit hold hole hook hurt hurts imagine imitate if is isn’t juggle jump just keep knock knocked know knows lean learn learned leave left let lick lift like likes lining listen lives lock look looking lose lost m made mail make making matter mean meant mess might mind miss missed misunderstood move must need needs not now on one only open opens pack park pat pedal pen pencil pick piece pillow pinch pitch play played point pour poured practice pretend pull push pushed pushing put putting ran reach read realize really remember reminded ride riding right ring rinse rip roll run runs said saw say says scare screw see seem seen set share shaved should show showing sing singing sit sitting sleep slide smash smell smells smiling somebody sound speak spend spill spilled spit spoil spread squeak squeeze stacked stand start stay stepped stick still stop stuff swim swims take taken takes taking talk talked talking tape taste tear tell thank then think though threw throw tickle tinkle to told took tore touch trade traded trading tried try turn turned uh understand unscrew use using wait wake walk want wants was wash waste watch wear went where will win wind wipe wiping woke won work works worry wrap write writing wrote

References

Aslin, R. N., Slemmer, J. A., Kirkham, N. Z., & Johnson, S. P. 2001. *Statistical learning of visual shape sequences*. Paper presented at the meeting of the Society for Research in Child Development, Minneapolis, MN

Au, T. K., Dapretto, M., and Song, Y. K. 1994. Input versus constraints: Early word acquisition in Korean and English. *Journal of Memory and Language*, 30, 567-582.

Balota, D. A., and Chumbley, J. I. 1984. Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance* 10, 340-357.

Basili R., Pazienza M.T. and Vindigni M. 1997. Corpus-driven Unsupervised Learning of Verb Subcategorization frames. *Proceedings of the Conference of the Italian Association for Artificial Intelligenc*e, AI*IA 97, Rome.

Bates, E. and MacWhinney, B. 1982. Functionalist approaches to grammar. In E. Wanner and L. Gleitman (eds.), *Language Acquisition: The State of the Art*. Cambridge: Cambridge University Press.

Bates, E., Betherton, I, and Snyder, L. (eds.)1988. *From First Words to Grammar: individual differences and dissociable mechanisms.* Cambridge: Cambridge University Press.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., Reilly, J., and Hartung, J. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language* 21: 1, 85-124.

Beckman, M. and Pierrehumbert, J. 1986. Intonational Structure in Japanese and English. *Phonology Yearbook*, 3, 255-310

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. 2001. Form variation of English function words in conversation. Manuscript, University of Colorado, Lucent Bell Laboratories, Brown University, and University of Pennsylvania.

Berwick, R. 1985. *The acquisition of syntactic knowledge*. Cambridge, Mass: MIT Press.

Bever, T. G. 1970. The cognitive basic of linguistic structures. In J. R. Hayes (ed.), *Cognition and the development of language*. New York: Wiley, pp. 279-352.

Bies, A., Ferguson, M., Katz, K., and MacIntyre, R. 1995. *Bracketing guidelines for Treebank II style, Penn Treebank Project*, University of Pennsylvania, Philadelphia.

Bloom, L. 1970. *Language development: Form and function in emerging grammars*. Cambridge, Mass: MIT Press.

Bloom, L. 1991. *Language development from two to three*. New York: Cambridge University Press.

Bloomfield, L. 1933. *Language*. Chicago: University of Chicago Press.

Boguraev, B., Briscoe, E., Carroll, J., Carter, D., Grover, C.1987. The derivation of grammatically indexed lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.*

Bolinger, D. 1977. *Meaning and Form*. London: Longman.

Bowerman, M. 1985. What shapes children's grammars? In D. Slobin (ed.), *The crosslinguistic study of language acquisition*, vol. 2: *Theoretical issues*. Hillsdale, NJ: Erlbaum, pp. 1257-1314.

Bowerman, M. 1990. Mapping thematic roles onto syntactic functions: Are children helped by innate linking rules? *Linguistics*, 28: 1253-1289.

Braine, M. 1976. *Children's first word combinations*. Monographs of the Society for Research in Child Development 41.

Brent, M. R. 1991. Automatic acquisition of subcategorization frames from untagged texts. In *Proceedings of the 29th Annual Meeting of the ACL*. 209-214.

Brent, M. R. 1993. From grammar to lexicon: Unsupervised learning of lexical syntax. *Computational Linguistics*. 19:243-262.

Brent, M. R. 1994. Surface cues and robust inference as a basis for the early acquisition of subcategorization frames. *Lingua* 92:433-470.

Bresnan, J. 2001. *Lexical-functional syntax*. Oxford: Blackwell.

Briscoe, T., and Carroll, J. 1993. Generalized probabilistic LR parsing of natural language (corpora) with unification-based methods. *Computational Linguistics* 19:25-59.

Briscoe, E.J. and Carroll, J. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC. 356--363.

Brown, R., and Hanlon, C. 1970. Derivational complexity and order of acquisition in child speech. In J. R. Hayes (eds.), *Cognition and the development of language*. New York: Wiley, 11-53.

Buchholz, S. 1998. Distinguishing complements from adjuncts using memory-based learning. In *Proceedings of the ESSLLI-98 Workshop on Automated Acquisition of Syntax and Parsing*.

Bybee, J. L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In M. Barlow and S. Kemmer (eds.), *Usage-based models of language*. Stanford, CA: CSLI Publications, pp. 65-85.

Carroll, G., and Rooth, M. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, Granada, Spain.

Carter, A. and Gerken, L. A. 1997. Children's use of grammatical morphemes in on-line sentence comprehension. In E. Clark (ed.), *Proceedings of the 28th Annual Child Language Research Forum*. Palo Alto, CA: Stanford University Press.

Cartwright, T., and Brent, M. R. 1997. Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition* 63:121-170.

Caselli, M.C., Bates, E., Casadio, P., Fenson, J., Fenson, L., *Sanderl, L., & Weir, J. 1995. A cross-linguistic study of early lexical development. *Cognitive Development* 10: 159-199.

Cavar, D., Herring, J., Ikuta, T., Rodrigues, P., and Schrementi, G. 2004. *Alignment Based Unsupervised Grammar Induction*. Ms. Indiana University.

Choi, Soonja, and Bowerman, M. 1991. Learning to express motion events in English and Korean: The influence of language-specific lexicalization patterns. *Cognition* 41, 83-121.

Choi, Soonja, and Gopnik, A. 1993. Nouns are not always learned before verbs in Korean: An early verb explosion. Paper presented at the 25th Child Language Forum. Stanford University.

Chomsky, N. and Halle, M. 1968. *The sound pattern of English*. New York: Harper and Row.

Chomsky, N. 1965. *Aspects of the theory of syntax*. Cambridge, Mass: MIT Press.
Chomsky, N. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, N. 1986. *Barriers*. Cambridge, Mass: MIT Press.

Chomsky, N. 1995. *The Minimalist Program*. Cambridge, Mass: MIT Press.

Christiansen, M.H., Allen, J. and Seidenberg, M.S. 1998. Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, 13, 221-268.

Christophe, A., Guasti, M. T., Nespor, M., and van Ooyen, B. 2003. Prosodic structure and syntactic acquisition: the case of the head-complement parameter. *Developmental Science*, 6, 213-222.

Clark, H. and Clark, E. 1977. *Psychology and language: An introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Clark, E. 1983. Meanings and concepts. In P. Mussen (ed.), *Handbook of child psychology*. 3:787-840. 4th ed. New York: Wiley.

Clark, E. 1987. The Principle of Contrast: A constraint on language acquisition. In B. MacWhinney (ed.), Hillsdale, NJ: Erlbaum. pp. 1-34.

Clark, E. 1990. Speaker perspective in language acquisition. *Linguistics* 28, 1, 201-220.

Clark, A. 2000. Inducing syntactic categories by context distribution clustering. In *Proceedings of CoNLL2000*, Lisbon, Portugal. 91-94.

Clark, A. 2001. Unsupervised induction of stochastic context-free grammars using distributional clustering, *Proceedings of the Fifth Conference on Natural Language Learning (CoNLL-2001),* Toulouse, France.

Connine C., Ferreira, F., Jones, C., and Frazier, L. 1984. Verb frame preference: Descriptive norms. *Journal of Psycholinguistic Research* 13, 307-319.

Cooper, W.E., and Paccia-Cooper, J. 1980. *Syntax and speech*. Cambridge, Mass.: Harvard University Press.

Cooper, W. E. 1975. Selective adaptation to speech. In R. Restle, R. M. Schiffrin, N. J. Castellan, H. Lindman, and D. B. Pisoni (eds.), *Cognitive theory* vol. 1. Hillsdale, NJ: Erlbaum. pp. 591-600.

Cormen, T. H., Leiserson, E. C., and Rivest, L. R. 1995. *Introduction to algorithms*. Cambridge, Mass: MIT Press.

Cover, T. M., and Thomas, A. 1991. *Elements of Information Theory*. New York: Wiley.
Croft, W. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Croft, W. 1991. *Syntactic categories and grammatical relations: The cognitive organization of information*. Chicago: Chicago University Press.

Cutler, A., and Carter, D.M. 1987. The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, 2, 133-142.

Cutler, A., & Norris, D. 1988. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* 14, 113-121.

Daelemans, W., Zavrel, J., Van der Sloot, K., and Van den Bosch, A. 1998. TiMBL: Tilburg Memory Based Learner, version 1.0, reference manual. Technical Report ILK-9803, ILK, Tiburg University.

Dowty, D. 1991. Thematic proto-roles and argument structure. *Language*, 67, 547--619.

Dromi, E. 1987. *Early lexical development*. Cambridge: Cambridge University Press.

Durieux G., Gillis S. 2001. Predicting grammatical classes from phonological cues: an empirical test. In J. Weissenborn, and B. Hole (eds.), *Approaches to bootstrapping: phonological, syntactic and neurophysiological aspects of early language acquisition*. Amsterdam: Benjamins, pp. 189-232.

Echols, H. 2001. Contributions of prosody to infants' segmentation and representation of speech. In J. Weissenborn, and B. Hole (eds.), *Approaches to bootstrapping: phonological, syntactic and neurophysiological aspects of early language acquisition*. Amsterdam: Benjamins, pp. 25-46.

Eckle, J. and Heid, U. 1996. Extracting raw material for a German subcategorization lexicon from newspaper text. *Proceedings of the 4th International Conference on Computational Lexicography, COMPLEX'9*6, Budapest, Hungary.

Fano, R. 1961. *Transmission of information: A statistical theory of communications*. New York: MIT Press.

Fidelholz, J. 1975. Word frequency and vowel reduction in English. *Papers from the 13th Regional Meeting, Chicago Linguistic Society,* pp. 200-213.

Fillmore, C. J. 1977. The case for case reopened. In P. Cole and J.M. Sadock (eds.), *Syntax and Semantics 8: Grammatical Relations*. New York: Academic Press. pp. 59-81.

Finch, S., and Charter, N. 1992a. Automatic methods for finding linguistic categories. In I. Alexander and J. Taylor (eds.), *Artificial Neural Networks,* volume 2. Elsevier Science Publishers.

Finch, S., and Chater, N. 1992b. Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, pp 820-825. Hillsdale, New Jersey.

Fisher, C. 1994. Structure and meaning in the verb lexicon: Input for a syntax-aided verb learning procedure. *Language and Cognitive Processes*, 9, 473-518.

Fisher, C. 1996. Structural limits on verb mapping: The role of analogy in children's interpretation of sentences. *Cognitive Psychology*, 31, 41-81.

Fisher, C., Hall, D. G., Rakowitz, S., and Gleitman, L. R.1994. When it is better to receive than to give: Structural and conceptual cues to verb meaning. *Lingua* 92, 333-375.

Fisher, C., Gleitman, H., and Gleitman, L. 1991. Relationships between verb meanings and their syntactic structures. *Cognitive Psychology* 23, 331-392.

Fodor, J. A. 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry*, 9, 427-474.

Fodor, J. A., and Garrett, M. F. 1967. Some syntactic determinants of sentential complexity. *Perception and Psychology* 2, 289-296.

Fontenelle, T., Bruls, W., Thomas, L., Vanallemeersch, T., and Jansen, J. 1994. *Survey of collocation extraction tools*. Technical report, University of Liege, Liege, Belgium.

Forster, K. and Chambers, S. 1973. Lexical access and naming time. *Journal of Verbal Learning and Verbal Behavior* 12, 627-635.

Francis, W. N. and Kučera, H. 1982. *Frequency analysis of language usage*. Boston: Houghton Mifflin.

Friederici, A.D. and Wessels, J.M.I. 1993. Phonotactic knowledge and its use in infant speech perception. *Perception and Psychophysics* 54, 287-295.

Frisch, S. A., Large, N. R. and Pisoni, D. 2000. Perception of wordlikeness: effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42, 481-496.

Garnsey, S. M., Pearlmutter, N. J., Myers, E., and Lotocky, M. A. 1997. The contributions of verb bias and plausibility to the comprehension of temporarily ambiguous sentences. *Journal of Memory and Language* 37, 58-93.

Garside, R., Leech, G., and Sampson, G. (eds.) 1987. *The computational analysis of English: A corpus-based approach*. London: Longman.

Gasser, M., and Smith, L. B. 1998. Learning nouns and adjectives: A connectionist account. *Language and Cognitive Processes*, 13, 269-306.

Gentner, D. 1978, On relational meaning: The acquisition of verb meaning. *Child Development* 49, 1034-1039.

Gentner, D. 1983. Structure mapping: A theoretical framework for analogy. *Cognitive Science*, 7, 2: 155-170.

Gentner, D.1982. Why nouns are learned before verbs: Linguistic relativity vs. natural partitioning. In S. Kuczaj (ed.) *Language Development*, vol.2: *Language, cognition and culture*, 301-34. Hillsdale, NJ: Erlbaum.

Gerken, L. 1994a. A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language*, 21, 565-584.

Gerken, L. 1994b. Young children's representation of prosodic phonology: Evidence from English-speakers' weak syllable productions. *Journal of Memory and Language*, 33, 19-38.

Gerken, L. 1996. Prosodic structure in young children's language production. *Language*, 72, 683-712.

Gerken, L. A., Landau, B., and Remeze, R. E. 1990. Function morphemes in young children's speech perception and production. *Developmental Psychology* 27, 204-216.

Gleitman, L. R. 1990. The structural sources of verb meanings. *Language Acquisition* 1: 3-55.

Gleitman, L. R., Wanner, E. 1988. Current issues in language learning. In M. Bornstein and M. Lamb (eds.), *Developmental psychology: An advanced textbook*. Hillsdale, NJ: Erlbaum. 297-356.

Godfrey, J., Holliman, E., and McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of. ICASSP-92*, pp. 517-520.

Gold, E. M. 1967. Language identification in the limit. *Information and Control* 10, 447-474.

Goldfield, B., and Reznick, J. S.1990. Early lexical acquisition: Rate, content, and the vocabulary spurt. *Journal of Child Language* 17, 171-183.

Goldfield, B. 1993. Noun bias in maternal speech to one-year-olds. *Journal of Child Language* 20, 85-100.

Goldin-Meadow, S., Seligman, M. E. P., and Gelman, R. 1976. Language in the two-year-old. *Cognition* 4, 189-202.

Golinkoff, R., Mervis, C., and Hirsh-Pasek, K. 1994. Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language* 21: 125-155.

Gove, Philip B. (ed.). 1977. *Webster's seventh new collegiate dictionary*. Springfield, MA: G. and C. Merriam.

Greenberg, J. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (ed.), *Universals of Language*. Cambridge, Mass: MIT Press. 73-113.

Greenberg, S., Ellis, D., and Hollenback, J. 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. *Proceedings of the International Conference on Spoken Language Processing* (*ICSLP-96*), Philadelphia, PA, pp. S24-27.

Gregory, M. L. 2001. Linguistic informativeness and speech production: An investigation of contextual and discourse-pragmatic effects on phonological variation. Ph.D. dissertation, University of Colorado, Boulder.

Gregory, M., Raymond, W., Fosler-Lussier, E., and Jurafsky, D. 2000. The effects of collocational strength and contextual predictability in lexical production. *Papers from the 35th Regional Meeting, Chicago Linguistic Society,* pp. 151-166.

Grimshaw, J. 1981. Form, function, and the language acquisition device. In C. Baker and J. McCarthy (eds.), *The logical problem of language acquisition*. Cambridge, Mass: MIT Press. 165-182.

Grimshaw, J. 1990. *Argument Structure*. Cambridge, Mass: MIT Press.

Grimshaw, J. 1994. Lexical reconciliation. In L. Gleitman & B. Landau (eds.), *The acquisition of the lexicon*. Cambridge, MA: MIT Press.

Grishman, R., Macleod, C., and Meyers, A. 1994. COMLEX syntax: Building a computational lexicon. *Proceedings of COLING 1994*, Kyoto, 268-272.

Grosjean, F. 1980. Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*, 28, 267-283.

Gruber, J. S. 1965. *Studies in lexical relations*. Ph.D. dissertation, MIT.

Hare, M., McRae, K., and Elman, J. L. 2003. Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language*, 48, 2, 281-303.

Harris, Z. 1951. *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Harris, J. W. and Stocker, H. 1998. *Handbook of Mathematics and Computational Science*. New York: Springer-Verlag.

Hay, J. B. 2000. Causes and consequences of word structure. Ph.D. dissertation, Northwestern University. (Available at http://www.ling.canterbury.ac.nz/jen).

Hayes, J. R., and Clark, H. H. 1970. Experiments in the segmentation of an artificial speech analog. In J. R. Hayes (ed.), *Cognition and the Development of Language*. New York: Wiley

Hindle, D. 1988. Acquiring a noun classification from predicate-argument structures. Technical Memo. 11222-881017-15, AT&T Bell Laboratories.

Hirsh-Pasek, K., and Golinkoff, R.M. 1996a. *The origins of grammar*. Cambridge, Mass: MIT Press.

Hirsh-Pasek, K., Kemler Nelson, D. G., Jusczyk, P. W., Cassidy, K. W., Druss, B., and Kennedy, L. 1987. Clauses are perceptual units for young infants. *Cognition* 26, 269-186.

Hooper, J. B. 1976. Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie (ed.), *Current progress in historical linguistics*. Amsterdam: North Holland, pp. 96-105.

Hopcroft, J. E., Motwani, R., and Ullman, J. D. 2001. *Introduction to automata theory, languages, and computation*. New York: Addison Wesley.

Howes, D. 1957. On the Correlation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America* 29, 296-305.

Howes, D. and Solomon, R. 1951. Visual duration threshold as a function of Word probability. *Journal of Experimental Psychology* 41, 401-410.

Jackendoff, R. 1983. *Semantics and Cognition*. Cambridge, MA: MIT Press.

Jackendoff. R. 1987. On Beyond Zebra: The Relation of Linguistic and Visual Information. *Cognition*, 26(2):89-114, 1987.

Jackendoff, R. 1990. *Semantic structures*. Cambridge, Mass: The MIT Press.

Jackendoff, R. 1997. *The Architecture of the language faculty*. Cambridge, Mass: MIT Press.

Jakobson, R., and Linda R. Waugh. 1987. *The sound shape of language*. Berlin: Mouten de Gruyter.

Jelinek, F. 1968. *Probabilistic information theory: Discrete and memoryless models*. New York: McGraw-Hill Book Company.

Jelinek, F. 1985. *Self-organizing modeling for speech recognition*. IBM Report.

Jescheniak, J. D., and Levelt, W. J. M. 1994. Word frequency effects in speech production: Retrieval of syntactic information and phonological form. *Journal of Experimental Psychology: Learning, Memory and Cognition* 20, 824-843.

Johnson, E.K. & Jusczyk, P.W. 2001. Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Languag*e, 44, 548-567.

Joshi, A. K., and Schabes, Y. 1997. Tree-adjoining grammars. In G. Rozenberg and A. Salomaa (eds.) *Handbook of formal languages*. Berlin: Springer-Verlag, pp. 69-123.

Jurafsky, D. 2003. Probabilistic modeling in psycholinguistics: Linguistic comprehension and Production. In R. Bod, R. Hay, and S. Jannedy (eds.), *Probabilistic Linguistics*. Cambridge, Mass: MIT Press. 39-98.

Jurafsky, D., Bell, A., Gregory, M., and Raymond, W. D. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. In J. L. Bybee and P. Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, pp. 229-254.

Jusczyk, P. W., Hohne and Mandel 1995. Picking up regularities in the sound structure of the native language. In W. Strange (ed.), *Speech Perception and Linguistic Experience: Issues in cross-language research*. Baltimore: York Press. 91-119.

Jusczyk, P.W., and Thompson, E. 1978. Perception of phonetic contrast in multi-syllabic utterances by two-month old infants. *Perception and Psychophysics*, 23, 105-109

Jusczyk, D., Hirsh-Pasek, K., Kemler Neslon, D. G., Kennedy, L. J., Woodward, A., and Piwoz, J. 1992. Perception of acoustic correlates of major phrasal units by young infants. *Cognitive Psychology* 24, 252-293.

Jusczyk, P. W., Cutler, A., and Redanz, N. J. 1993. Preference for the predominant stress patterns of English words. *Child Development* 64, 675-687.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language* 33, 630-645.

Keenan, E. L. 1976. Towards a universal definition of "subject". In C. N. Li (ed.), *Subject and topic*. New York: Academic Press. 303-333.

Kelly, M.H., and Bock, J.K. 1988. Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 389-403.

Kelly, M. H. 1992. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review* 99, 349-364.

Kelly, M. H. 1996. The role of phonology in grammatical category assignment. In K. Demuth, and J. L. Morgan (eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates: 249-262.

Kemler Nelson, D. G., and 11 Swarthmore College students. 1995. Principle-based inferences in young children's categorization: Revisiting the impact of function on the naming of artifacts. *Cognitive Development,* 10, 347-380.

Kimball, J. P. 1973. Seven principles of surface structure parsing in natural language. *Cognition* 2, 15-47.

Klatt, D. H. 1975. Vowel lengthening is syntactically determined in a connected discourse, *Journal of Phonetics*, 3, pp. 129-40, 1975

Klein, D. and Manning C. 2001. Distributional Phrase Structure Induction, *Proceedings of the Fifth Conference on Natural Language Learning (CoNLL-2001)*, Toulouse, France.

Korhonen, Anna 1997. Acquiring Subcategorization from Textual Corpora. Mphil Dissertation. Department of Engineering, University of Cambridge.

Kuhl, P.K. 1983. Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior and Development*, 6, 263–285.

Landau, B. and Stecker, D. 1990. Objects and places: Geometric and syntactic representation in early lexical learning. *Cognitive Development*, 5, 287-312.

Landau, B., and Gleitman, L. R. 1985. *Language and experience: Evidence from the*

Lebeaux, D. 1997. *Determining the Kernel II: Prosodic Form, Syntactic Form, and Phonological Bootstrapping*. Tech Report, NEC Research Institute.

Lebeaux, D. 2001. Prosodic Form, Syntactic Form, Phonological Bootstrapping, and Telegraphic Speech. In Weissenborn, J. and Hole, B. (eds.). *Approaches to Bootstrapping: Phonological, Lexical, Syntactic and Neurophysiological Aspects of Early Language Acquisition,* Volume 2. John Benjamins Publishing Company: Amsterdam.

Li, P., and Yip, M. C. 1996. Lexical ambiguity and context effects in spoken word recognition: Evidence from Chinese. *Proceedings of the 18th Annual Conference of the Cognitive Science Society (COGSCI-96)*, pp. 228-232.

Lightfoot, D. 1991. *How to set parameters: Evidence from language change*. Cambridge, Mass: MIT Press.

MacDonald, M. C. 1993. The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language* 32, 692-715.

MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* 9, 157-201.

Macnamara, J. 1982. *Names of things: A study of human learning*. Cambridge, Mass: MIT Press.

MacWhinney, B. 1991. *The CHILDES Project: Tools for analyzing talk.* LEA, Hillsdale, NJ.

Magerman, D. M, and Marcus, M. P. 1990. Parsing a natural language using mutual information statistics. In *National Conference on Artificial Intelligence*, pp. 984-989.

Manning, C. D. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, pp. 235-242

Manning, C. D. 2003. Probabilistic syntax. In R. Bod, J. Hay, and S. Jannedy (eds.), *Probabilistic Linguistics*. Cambridge, Mass: MIT Press. 289-342.

Manning, C. D., and Schütze, H. 2003. *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Maragoudakis, M., Kermanidis, K., and Kokkinakis, G. 2000. Learning Subcategorization Frames from Corpora: A Case Study for Modern Greek. In *Proceedings of COMLEX 2000, Workshop on Computational Lexicography and Multimedia Dictionaries*, pp. 19-22, Kato Achaia, Greece, 22-23 September 2000

Maratsos, M.P. & Chalkley, M.A. 1980. The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K.E. Nelson (ed.), *Children's Language* Volume 2. New York: Gardner Press. 127-214.

Markman, E. M. 1987. How children constrain the possible meanings of words. In U. Neisser (ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*. Cambridge: Cambridge University Press. 255-287.

Markman, E. M. 1989. *Categorization and naming in children*. Cambridge, Mass: MIT Press.

Markman, E. M. 1990. Constraints children place on word meanings. *Cognitive Science* 14, 57-77.

Martello, S., and Toth, P. 1990. *Knapsack problems: Algorithms and computer implementations*. New York: Wiley.

Mattys, S. L, and Jusczyk, P. W. 2001. Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, 27, 91-121.

Mattys, S., Jusczyk, P. W., Luce, P. A., and Morgan, J. L. 1999. Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology* 38, 465-494.

Merlo, P. 1994. A corpus-based analysis of verb continuation frequencies for syntactic processing. *Journal of Psycholinguistic Research* 23, 435-457.

Meyers, A., Macleod, C., and Grishman, R. 1994. *Standardization of the complement adjunct distinction*. Proteus Project Memorandum 64, Computer Science Department, New York University.

Miller, G., and Charles, W. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1).

Miller, G. & Fellbaum, C. 1992. WordNet and the organization of lexical memory. In M. Swartz, and M. Yazdani (eds.), *Intelligent Tutoring Systems for Foreign Language Learning*. New York: Springer-Verlag. pp 89-102.

Mintz, T. H. 1996. The roles of linguistic input and innate mechanisms in children's acquisition of grammatical categories. Ph.D. dissertation, University of Rochester.

Mintz, T.H., Newport, E.L. and Bever, T. G. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26, 393-424.

Moravcsik, J. 1981. How do words get their meanings? *The Journal of Philosophy*, 78, 5-24.

Morgan, J. L., and Travis, L. 1989. Limits on negative information in language input. *Journal of Child Language* 16: 531-552.

Morgan, J.L., Shi, R. & Allopenna, P. 1996. Perceptual bases of grammatical categories. In J.L. Morgan & K. Demuth (eds.), *Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisition*. Mahwah, NJ: Lawrence Erlbaum Associates

Morgan, J.L., Swingley, D., and Mitirai, K., (March, 1993). Infants listen longer to speech with extraneous noises inserted at clause boundaries. Poster presented at the Biennial Meeting of the Society for Research in Child Development, New Orleans.

Morrill, G. 1994. *Type-logical grammar*. Dordrecht: Kluwer.

Morse, P. A. 1972. The discrimination of speech and nonspeech stimuli in early infancy. *Journal of Experimental Child Psychology* 14, 477-492.

Nelson, K. 1973. *Structure and strategy in learning to talk*. Monographs pf the Society for Research in Child Language Development 149.

Nelson, K., Hampson, J., and Shaw, L. 1993. Nouns in early lexicon: Evidence, explanations, and implications. *Journal of Child Language* 20: 61-84.

Nespor, M. and Vogel, I. 1986. *Prosodic Phonology*. Dordrecht: Foris.

Nilsson, N. J. 1982. *Principles of Artificial Intelligence*. Berlin: Springer-Verlag.

O’Grady, W. 1997. *Syntactic development*. Chicago: Chicago University Press.

Oakes, M. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Oldfield, R. C., and Wingfield, A. 1965. Response latencies in naming objects. *Quarterly Journal of Experimental Psychology* 17, 273-281.

Osherson, D. N., Stob, M., and Weinstein, S. 1985. *Systems that Learn*. Cambridge, MA: MIT Press.

Pan, S., and Hirschberg, J. 2000. Modeling local context for pitch accent prediction. ACL-2000, Hong Kong. Available at http://www1.cs.columbia.edu/~pan

Pinker, S. 1984. *Language learnability and language development*. Cambridge: Harvard University Press.

Pinker, S. 1989. The bootstrapping problem in language acquisition. In B. MacWhinney (ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum. 399-441.

Pinker, S. 1994. How could a child use verb syntax to learn verb semantics? *Lingua* 92: 377-410.

Pollard, C., and Sag, I. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: Chicago University Press.

Quine, W. V. O. 1960. *Word and object*. Cambridge, Mass: MIT Press.

Radford, A. 1990. *Syntactic theory and the acquisition of English syntax*. Oxford: Blackwell.

Rappaport, M. and B. Levin.1988. What to Do with Theta-Roles. In W. Wilkins (ed.), *Syntax and Semantics 21: Thematic Relations*. Academic Press, New York, NY, 7-36.

Redington, M., and Chater, N. 1997. Probabilistic and distributional approaches to language acquisition. *Trends in Cognitive Sciences*, 1 (7), 273-281.

Roland, D., and Jurafsky, D. 1998. How verb subcategorization frequencies are affected by corpus choice. *Proceedings of COLING-98*, Montreal, pp. 1122-1128.

Rubenstein, H., Garfield, L., and Millikan, J. A. 1970. Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior* 9, 487-494.

Saffran, J. R., Aslin, R. N., and Newport, E. L. 1996. Statistical cues in language acquisition: Word segmentation in infants. *Proceedings of the 18th Annual Conference of the Cognitive Science Society (COGSCI-96)*, San Diego, California, pp. 376-380.

Saffran, J. R., Newport, E. L., and Aslin, R. N.1996a. Statistical learning by 8-month-old infants. *Science* 274, 1926-1928.

Saffran, J.R., Johnson, E.K., Aslin R.N. & Newport, E.L. 1999. Statistical learning of tone sequences by human infants and adults, *Cognition*, 70, 27-52.

Sansavini, A., Bertoncini, J., and Giovanelli, G. 1997. Newborns discriminate the rhythm of multisyllabic stressed words. *Developmental Psychology*, 33(1), 3-11.

Savin, H. B. 1963. Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America* 35, 200-206.

Schütze, H. 1994. Distributional part-of-speech tagging. In *Proceedings of the 7th Conference of the European Chapter of The Association for Computational Linguistics,* Dublin, Ireland.

Schütze, C. T. 1996. *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.

Scott, D. 1982. Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, 71**,** 996–1007.

Selkirk, E. O. 1984. *Phonology and Syntax: The relation between sound and structure*. Cambridge, Mass: MIT Press.

Shady, M., and Gerken, L.A. 1999. Grammatical and caregiver cues in early sentence comprehension. *Journal of Child Language*, 26, 163-175.

Simpson, G. B., and Burgess, C. 1985. Activation and selection processes in the recognition of ambiguous words. *Journal of Experimental Psychology: Human Perception and Performance* 11, 28-39.

Talmy, L. 1985. Lexicalization patterns: Semantic structure in lexical items. In T. Shopen (ed.), *Language typology and syntactic description*, vol. 3: *Grammatical categories and the lexicon*. New York: Cambridge University Press. 57-149.

Taylor, L. J., and Knowles, G. 1988. *Manual of Information to Accompany the SEC Corpus: The machine readable corpus of spoken English*. University of Lancaster.

Trueswell, J. C., Tanenhaus, M. K. & Kello, C. 1993. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19(3), 528-553.

Tyler, L. K. 1984. The structure of the initial cohort: Evidence from gating. *Perception and Psychophysics* 36, 417-427.

Ushioda, A., Evans, D. A., Gibson, T., and Waibel, A. 1996. Estimation of verb subcategorization frame frequencies based on syntactic and multi-dimensional statistical analysis. In H. Bunt and M. Tomita (eds.), *Recent Advances in Parsing Technology*. Dordrecht: Kluwer. 241-254.

van Zaanen, M. 2000. Learning Structure using Alignment Based Learning. In *Proceedings of the 3rd Annual CLUK Research Colloquium*, Brighton, U.K., April 2000.

Vitevitch, M. S., Luce, P. A., Charles-Luce, J., and Kemmerer, D. 1997. Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech*, 40, 47-62.

Watt, W. C. 1970b. On two hypotheses concerning psycholinguistics. In J. R. Hayes (ed.) *Cognition and the development of language*. New York: Wiley. 137-220.

Weissenborn, J., and Höhle, B. (eds.). 2001. *Approaches to bootstrapping: Phonological, lexical, syntactic, and Neurophysiological aspects of early language acquisition*. Amsterdam: Benjamins.

Wexler, K., and Culicover, P. 1980. *Formal principles of language acquisition*. Cambridge, Mass,: MIT Press.

Whaley, C. P. 1978. Word-nonword classification time. *Journal of Verbal Learning and Verbal Behavior* 17, 143-154.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P. J. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of Acoustical Society of America* 91, pp. 1707-1717.

Wingfield, A. 1968. Effects of frequency on identification and naming of objects. *American Journal of Psychology* 81, 226-234.

Zavrel, J. and Veenstra, J.1995. The language environment and syntactic word class acquisition. In F. Wijnen, and C. Koster (eds.), *Proceedings of Groningen Assembly on Language Acquisition (GALA95),* Groningen.

Zeman, D. and Sarkar A. 2000. Automatic Extraction of Subcategorization Frames for Czech. In *Proceedings of the 18th International Conference on Computational Linguistic*s.

Zernik, U. 1991. Introduction. In U. Zernik (ed.), *Lexical acquisition: Exploring on-line resources to build a lexicon*. Hillsdale, NJ: Erlbaum. 1-26.