Part-of-speech Bootstrapping using Lexically-Specific Frames

Richard Eduard Leibbrandt
B.Sc. (Hons), B.Soc.Sci. (Hons), M.Sc.

School of Computer Science, Engineering and Mathematics
Faculty of Science and Engineering
Flinders University of South Australia
July 2009

Thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

The work in this thesis presents and evaluates a number of strategies by which English-learning children might discover the major open-class parts-of-speech in English (nouns, verbs and adjectives) on the basis of purely distributional information. Previous work has shown that parts-of-speech can be readily induced from the distributional patterns in which words occur. The research reported in this thesis extends and improves on this previous work in two major ways, related to the *constructional* status of the utterance contexts used for distributional analysis, and to the way in which previous studies have dealt with categorial *ambiguity*.

Previous studies that have induced parts-of-speech from word distributions have done so on the basis of fixed “windows” of words that occur before and after the word in focus. These contexts are often not constructions of the language in question, and hence have dubious status as elements of linguistic knowledge. A great deal of recent evidence (e.g. Lieven, Pine & Baldwin, 1997; Tomasello, 1992) has suggested that children’s early language may be organized around a number of lexically-specific constructional frames with slots, such as “a X”, “you X it”, “draw X on X”. The work presented here investigates the possibility that constructions such as these may be a more appropriate domain for the distributional induction of parts-of-speech. This would open up the possibility of a treatment of part-of-speech induction that is more closely integrated with the acquisition of syntax.

Three strategies to discover lexically-specific frames in the speech input to children are presented. Two of these strategies are based on the interplay between more and less frequent words in English utterances: the more frequent words, which are typically function words or light verbs, are taken to provide the schematic “backbone” of an utterance. In the first strategy, all frames are schematic structures for full utterances. The second strategy extends this approach to include multi-word sequences that frequently occur embedded inside other frequent multi-word sequences. The third strategy is based around pairs of words in which the occurrence of one word is highly predictable from that of the other, but not vice versa; from these basic slot-filler relationships, larger frames are assembled.

These techniques were implemented computationally and applied to a corpus of child-directed speech. Each technique yielded a large set of lexically-specific frames, many of which could plausibly be regarded as constructions. In a comparison with a manual analysis of the same corpus by Cameron-Faulkner, Lieven and Tomasello (2003), it is shown that most of the constructional frames identified in the manual analysis were also produced by the automatic techniques.

After the identification of potential constructional frames, parts-of-speech were formed from the patterns of co-occurrence of words in particular constructions, by means of hierarchical clustering. The resulting clusters produced are shown to be quite similar to the major English parts-of-speech of nouns, verbs and adjectives. Each individual word token was assigned a part-of-speech on the basis of its constructional context. This

categorization was evaluated empirically against the part-of-speech assigned to the word in question in the original corpus. The resulting categorization is shown to be, to a great extent, in agreement with the manual categorization.

These strategies deal with the categorial ambiguity of words, by allowing the frame context to determine part-of-speech. However, many of the frames produced were themselves ambiguous cues to part-of-speech. For this reason, strategies are presented to deal with both word and context ambiguity. Three such strategies are proposed. One considers membership of a part-of-speech to be a matter of *degree* for both word and contextual frame. A second strategy attempts to discretely assign multiple parts-of-speech to words and constructions in a way that imposes internal *consistency* in the corpus. The third strategy attempts to assign only the minimally-required multiple categories to words and constructions so as to provide a *parsimonious* description of the data.

Each of these techniques was implemented and applied to each of the three frame discovery techniques, thereby providing category information about *both* the frame and the word. The subsequent assignment of parts-of-speech was done by combining word and frame information, and is shown to be far more accurate than the categorization based on frames alone. This approach can be regarded as addressing certain objections against distributional part-of-speech bootstrapping that have been raised by Pinker (1979, 1984, 1987).

Lastly, a framework for extending this research is outlined that allows semantic information to be incorporated into the process of category induction.

I certify that this thesis does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

Richard Eduard Leibbrandt
21 July 2009

Acknowledgements

A number of people have helped to make the eventual completion of this thesis possible.

I would like to express my thanks first of all to my thesis supervisor David Powers, for his friendship and encouragement throughout these last few years. David gave me free rein to develop my own ideas, engaged with me on the scientific issues while respecting my viewpoint when we disagreed, served as an inexhaustible source of technical (and sometimes esoteric) information, and helped me to develop into a scientific researcher.

To my friends and colleagues in the Artificial Intelligence Laboratory and the School of Computer Science, Engineering and Mathematics at Flinders University, Graham Bignell, Aaron Ceglar, Denise de Vries, Sean Fitzgibbon, Shawn Haggett, Trent Lewis, Martin Luerssen, Takeshi Matsumoto, Anna Shillabeer, Kenneth Treharne and Dongqiang Yang, I give my thanks for helping to create a welcoming and supportive community in which to work.

Thanks to Darius Pfitzner for your friendship through almost five years of despair, gnawing self-doubt, etc., etc. while we struggled with our respective PhDs, having interminable cuppas and spending many afternoons and evenings enjoying meals at your home (with the additional forbearance of Susan, Mellion and Pheobe). Thank you for showing me the value of sticking at it and pushing through, and that you can achieve more if you attempt more. We dunnit mate.

Lastly, thank you to my parents for supporting me financially during my undergraduate studies. This is the culmination of that process (I promise. It's the last one you can get).

1 Introduction

The problem of part-of-speech induction

The parts-of-speech of a language (word classes such as nouns, verbs and adjectives) are of crucial importance in describing its grammar. A vast amount of research has aimed to delineate the processes by which language-learning children acquire the parts-of-speech of their native language. The work reported in this thesis will attempt to investigate aspects of this question by means of a computational approach.

While most theories of linguistics assign a pivotal role to parts-of-speech, some controversy exists about how these classes should be defined. In particular, there is a tension between a definition based on the meanings of words, and one based on the patterns of usage of words in various linguistic contexts. This theoretical dichotomy is mirrored in the language acquisition literature, with some researchers holding that parts-of-speech are learned on the basis of *semantic* similarities between words, while others maintain that these classes are formed from words that are used in similar context *distributions*.

The work presented in this thesis follows in the distributional tradition, and aims to show the feasibility of making use of particular distributional sources of information in order to determine the part-of-speech of a word in context. The main contributions of this thesis fall into two categories.

(i) A number of explicit, psychologically feasible techniques will be presented to automatically discover pertinent linguistic contexts in which words occur in child-directed speech. In contrast to much previous research, these contexts are also arguably legitimate objects of linguistic knowledge in their own right. The approach I will take here is to focus on *lexically-specific frames*, combinations of specific words with variable slots; such frames are believed to play a crucial role in children's early language development.

(ii) In addition, these contexts are utilized in order to form categories that correspond to parts-of-speech; several techniques for doing so are presented and evaluated. It will be necessary to deal adequately with the pervasive part-of-speech *ambiguity* of both words and the contexts in which they occur. A useful strategy is to combine categorial information from both the context and the word in focus.

1.1 Outline of this thesis

In the next three chapters I will review experimental and theoretical work from linguistics, developmental psychology and cognitive science that are relevant to the concerns of the current work. These reviews will necessarily be highly selective, and my focus will be on the sources of information (especially distributional information) that children might use to discover the parts-of-speech of their native language. Chapter 2 is concerned with the role of parts-of-speech in linguistic theory, Chapter 3 with language development research into the acquisition of parts-of-speech, and Chapter 4 with previous computational work that has attempted to account for the learning of these categories.

Subsequent chapters present the empirical contribution of the thesis. Chapter 5 outlines the frame-based approach taken here, and Chapter 6 presents the first technique by which lexically-specific frames corresponding to schematic structures for full utterances (referred to as *full-utterance frames*) may be identified in the language input. The frames produced by this process are then subjected to a *clustering* analysis in order to induce the parts-of-speech, and these categories are evaluated.

In Chapter 7, the technique for inducing parts-of-speech from frames is modified so as to account for the fact that some contexts are ambiguous, in accommodating words from more than one part-of-speech. Three *co-clustering* techniques to deal with ambiguity are explored and evaluated.

Chapter 8 expands on the full-utterance frame structures of Chapter 6 by also identifying *nested frames*, which are smaller structures embedded in full utterances. Chapter 9 takes a different frame discovery approach, in which a *prediction-based frame* is composed of a set of elements that mutually predict each other. The frames produced by these two processes are also subjected to the standard clustering analysis and the co-clustering techniques, and the resulting categorizations are evaluated.

In Chapter 10, I attempt to position the current work in terms of previous computational and language developmental work. Lastly, in Chapter 11 I describe a number of ways in which the current framework may be extended; of particular importance is allowing the framework to take *semantic* information into account.

2 Parts-of-speech and their role in grammar

2.1 Introduction

In this chapter I will review work in the field of linguistics in order to describe what parts-of-speech are, which parts-of-speech exist in the world's languages and in English in particular, and to consider whether there are parts-of-speech that are *universal*, i.e. common to all languages. The focus of this chapter will be on considering the sources of information from which parts-of-speech can be identified.

I will start off by drawing on research in linguistics to describe how parts-of-speech are conventionally viewed in the field, and to attempt to enumerate a number of the main categories exhibited in the languages of the world. In doing this, I will draw in particular on a relatively theory-neutral typological survey by Schachter & Shopen (2007).

Parts-of-speech are typically defined partly according to the contexts in which particular words occur. This opens the possibility that lexical categorization may be related to the notion of the linguistic *construction*, an abstract characterization of the various multi-morpheme structures found in any language. A relatively recent approach in linguistics, known as Cognitive Grammar or Construction Grammar (Bybee, 1985; Croft, 2001; Goldberg, 1995; Kay & Fillmore, 1999; Langacker, 1987; referred to hereafter as Construction Grammar), places the construction at the centre of linguistic theory. I will briefly sketch the main tenets of Construction Grammar, by way of contrast with the principles of the dominant linguistic approach of the last 50 years, transformational generative grammar (TGG; Chomsky, 1957, 1965, 1981; Ouhalla, 1999), and describe how parts-of-speech may be viewed in Construction Grammar.

Lastly, a great amount of controversy exists around the question of whether particular parts-of-speech are universal across languages. I will review work by two Construction Grammar theorists, Langacker (1987) and Croft (2001), that will be particularly useful in shining light on this issue.

2.2 Parts-of-speech in traditional linguistic theory

Most modern theories of linguistics agree that there are basic free-standing elements in every language that can be combined in various ways so as to produce acceptable and meaningful utterances; these basic elements are the *words* of a language. Furthermore, different words exhibit underlying similarities that allow them to be grouped together into categories, traditionally referred to as *word classes* or *parts-of-speech.*

The parts-of-speech of a language are made up of words that share certain grammatical properties. Schachter and Shopen (2007) characterize words from the same part-of-speech as being similar (i) in their distribution (mostly defined in terms of other words or parts-of-speech that may grammatically occur in certain positions relative to the words in question), (ii) in their functional roles (e.g. English nouns can function as subjects while verbs cannot) and (iii) in their "categorizations", or morphological markings (e.g. English nouns are categorized/marked for number but not tense, while verbs are marked for both number and tense).

Although these criteria for defining parts-of-speech are language-internal, Schachter and Shopen (2007) acknowledge that the *name* given to a particular category is often determined by universal semantic criteria, so that, for example, if a category contains a majority of words for persons, places or things, then that category is likely to be named the noun category of that language, while if it contains mostly words for actions and events, the category is named the category of verbs.

A distinction should be made between *word types* and *word tokens*. A word type corresponds to a word (or a distinct word sense) in a dictionary, such as "drink", "blue", "mean", etc. A word token corresponds to an individual act of using a particular word type in an utterance. If parts-of-speech were defined according to word types, all tokens of a word type such as "drink" would implicitly be assigned to the same category. Most linguistic approaches, however, recognize that there is a great amount of categorial *ambiguity* attached to many word types, so that "drink" in "What would you like to drink?" can be regarded as a verb, whereas "drink" in "Here's your drink" is a noun.

Hence, part-of-speech categorization operates on particular word tokens as they are used in context.

A major high-level distinction can be made between the open and closed word classes of a language. The closed-class categories are so named because they do not normally accept, or only occasionally accept, new members. These classes are typically used mainly to serve grammatical functions in a language (and hence are also termed *function word classes*), rather than to carry specific semantic content. The open classes, on the other hand, are usually semantically contentful (and are hence also known as *content word classes*), and can accept new members fairly easily.

2.2.1 Open word classes

Schachter and Shopen (2007) state that there are only four open classes in the languages of the world: nouns, verbs, adjectives and adverbs. These classes can be characterized in terms of their prototypical meanings. Nouns typically designate the names for people, places and things; verbs designate actions and processes; adjectives describe properties of noun referents, and adverbs typically modify verbs, but can also be used to modify adjectives, other adverbs and some other linguistic constituents.

Whereas it has previously been claimed that some languages, notably Nootka and Tagalog, do not distinguish between a noun and a verb class, Schachter and Shopen (2007) review evidence that this is probably incorrect, and that nouns and verbs do not have an identical profile of allowable marking in either of these languages. The authors conclude that the noun-verb distinction is one of the few universal properties of languages.

When a language possesses a class of adjectives, the key concept identifying a word as an adjective seems to be modification of a noun. However, Schachter & Shopen claim that there are many languages that either have no class of adjectives, or else have a closed class of only a few adjectives with no possibility of adding items to the class. Among languages with no adjectives, the semantic concepts expressed in English via adjectives are instead expressed either by means of nouns or verbs.

As one example, many languages use a single set of grammatical devices to express concepts that are expressed in English by means of either stative verbs or adjectives. For instance, in Mandarin we have (this example from Schachter and Shopen, 2007):

piaoliang de nüaizi
beautiful REL girl ("a beautiful girl")

liaojie de nüaizi
understand REL girl ("a girl who understands, an understanding girl")

The example also shows that stative verbs may have something of the character of atemporality, so that they may be similar in meaning to adjectives and adverbs. Note that, in these two cases, the construction seems to call for words that perform the function of modifying a noun; in this sense, the words *piaoliang* and *liaojie* could be seen as belonging to a single class of modifiers.

In similar fashion, the function of modification is something that potentially links nouns and adjectives together. Again, this is reflected in some languages; as demonstrated in the following examples from Quechua (Schachter & Shopen, 2007):

chay hatun runa
that big man

chay alkalda runa
that mayor man ("that man who is mayor")

And in fact a similar phenomenon occurs in English, in the "compound noun" construction ("diesel truck", "wind turbine"); the second noun represents the semantic essence of the compound, and the first noun modifies the second.

Schachter and Shopen (2007) tentatively characterize the class of adverbs as consisting of all words that modify linguistic elements other than nouns; however, this typological classification seems to run into difficulty given evidence that in some languages, adjectives are used to modify both nouns and verbs (cited in Schachter & Shopen, 2007, p. 22).

In languages lacking a distinct adverb class, modification of non-noun elements can be performed by nouns, adjectives or verbs. Adverbs in English seem to form a rather incoherent category (Van der Auwera, 1994). Many subtypes of English adverbs might more accurately be termed closed classes, e.g. time adverbs (e.g. “yesterday”, “today”, “tonight”, “tomorrow”) or directional adverbs (e.g. “home”, “away”). Of the remaining truly open adverbs, many are derived from adjectives or present participial forms of verbs in combination with the suffix “-ly”. In addition to the traditional view that adverbs modify verbs and adjectives, Van der Auwera (1994) points out that it could well be argued in many cases that modification of a verb phrase or an entire clause takes place.

It might perhaps be cogent to propose a part-of-speech corresponding to *modifiers*, a class that would cover words that modify any linguistic element, and that would hence include both adverbs and adjectives. It might, for a particular language, be possible to subdivide the modifier class in terms of the kinds of elements which are modified (if we presume the existence of nouns and verbs, there are three logical possibilities: a modifier can modify a noun, a verb or another modifier).

The main open word classes (nouns, verbs, adjectives, adverbs) of a particular language are sometimes further subdivided on the basis of the criteria distribution, function and morphological marking as before; however, these subclasses are not generally termed parts-of-speech. Some of these subdivisions can be substantiated on semantic grounds, e.g. the distinction between count nouns (for differentiable objects) and mass nouns (for undifferentiated substances).

Other open subclasses are only partly semantically based and are otherwise purely linguistic, e.g. the distinction between nouns of the masculine, feminine and neutral genders in German, a distinction which is only partly based on the sex of humans and animals.

It is worth mentioning that several theorists (especially in the tradition of Categorial Grammar, e.g. Pullum, 1994) have taken the opposite view from the one espoused here, that there are only a few parts-of-speech in a particular language. In the view of these researchers, basic categories can be composed out of the combinatorial possibilities produced by certain feature values that a word or phrase may exhibit (e.g. third person plural feminine noun), and these categories can be described in terms of their relations to others (e.g. present tense verb phrase agreeing with third person plural feminine noun), so that the number of categories in a language may be arbitrarily large (Pullum, 1994).

2.2.2 Closed word classes

Surveying the cross-linguistic data in order to provide a comprehensive description of language typology, Schachter and Shopen (2007) conclude that the closed word classes are far more varied and disparate across languages, and that there may be no closed classes that are present in every language (with the possible exception of the class of interjections). On the other hand, Schachter and Shopen (2007) do not find evidence that there are languages with no closed classes. Some prominent closed classes identified across languages are the following:

Pro-forms (including pronouns, pro-verbs, pro-adjectives, pro-adverbs and pro-sentences): These are words that stand in the place of more elaborated open-class items which are understood in the utterance context. Pronouns are the most commonly-found forms across languages, and include the subtypes personal, reflexive, reciprocal, demonstrative and relative pronouns. The function of pro-forms may be to promote understandability by reducing the amount of information in an utterance, especially when some of that information is redundant and can therefore be replaced with an empty "placeholder" pro-form. In the so-called pro-drop languages, it is possible simply to omit these redundant elements altogether.

Noun adjuncts: This broad class consists of closed-class words that are associated with nouns, and is further broken down by Schachter & Shopen (2007) into:

<u>*Role markers*</u>: These are markers for case (indicating syntactic/semantic roles, e.g. agent, subject) markers for discourse function (e.g. topicality) and other adpositions (prepositions and postpositions). The only role markers in English are the prepositions (case is marked on the pronouns, rather than by means of separate words). These markers denote the roles of different elements in the utterance as a whole, so that it may not be strictly accurate to describe them as noun adjuncts only, and indeed Schachter & Shopen (2007) note that role markers may be associated with verbs in some languages.

<u>*Quantifiers:*</u> These include numerals as well as words meaning "some" "few", "much", "all", "each", etc. Some of these quantifiers take different forms when modifying nouns with different semantics (e.g. one quantifier for human and another for nonhuman nouns).

<u>*Classifiers:*</u> In some languages, when a noun is modified by a numeral, quantifier or demonstrative, it is obligatory to add a closed-class word indicating class membership of the noun. These classes are clearly of interest for lexical categorization, as they seem to provide clear distributional information according to which nouns can be subdivided into classes. As is also the case with gender systems, the classification is often arbitrary and only to a limited extent semantically based.

<u>*Articles:*</u> Schachter & Shopen (2007) include in this category not only the definite and indefinite articles "the" and "a", but also the demonstratives "this" and "that".

<u>Verb adjuncts</u>: These are closed-class words associated with verbs, and can be categorized in turn as:

<u>*Auxiliaries*</u>: These words express the tense, aspect, mood, voice or polarity of a verb, and English examples include "will", "should", "might", "could", etc.

<u>*Particles*</u>: These are words that co-occur with certain verbs and essentially have the effect of turning them into new verbs with different meanings. Examples include "wake *up*", "calm *down*", "take *off*". These words can sometimes be separated from their verbs (e.g. "wake me up at five").

Conjunctions: These are words that connect words, phrases or clauses, and can be used to connect elements of equal status (*coordinating* conjunctions such as “and” and “or”), or to connect a pair of elements where one is clearly subordinate to the other (*subordinating* conjunctions such as “that”, “who”, “before”, etc).

Other classes: Additional closed classes include clitics, copulas, predicators, emphasis markers, existential markers, interjections, mood markers, negators and politeness markers.

2.3 The role of constructions in assigning parts-of-speech

Given the important role of distribution and function in delineating the parts-of-speech, it is worthwhile to look at a linguistic approach that places these two aspects in centre stage, namely the various theories that can be described under the label of Cognitive Grammar or Construction Grammar. The main tenets of these theories can best be illustrated by contrast with those of Transformational Generative Grammar.

2.3.1 Transformational Generative Grammar

The field of Transformational Generative Grammar (TGG; Chomsky, 1957, 1965, 1981) has been the dominant paradigm in Western linguistics for the last approximately 50 years. TGG aims to describe the grammatical knowledge by means of which a person can determine whether a particular sentence is grammatical or not. Because of the presumed complexity of language, combined with the speed and accuracy with which children acquire it, it is assumed that much knowledge of language is governed by universal principles which are innately present in the human mind (the argument from the “poverty of the stimulus”).

TGG concerns itself with accounting for what it regards as the core of a language – the relatively systematic and abstract set of rules governing certain kinds of sentences which are regarded as lying at the heart of language. Accounting for non-core parts of a language, such as words, idioms and semi-fixed idiosyncratic expressions, is regarded as falling outside the purview of a grammatical theory. Typically, the emphasis is on describing the constituents of a sentence in the most abstract possible way, so as to provide the simplest and most elegant explanation for the largest amount of phenomena.

In most formulations of TGG (e.g. Chomsky, 1957, 1965, 1981), a distinction is made between so-called deep structure, said to encapsulate the basic semantic relations in the sentence, and the surface structure, which is essentially the produced phonological forms of the sentence. TGG proposes that a set of transformational rules transform deep structure into surface structure, although no claim is made that this process actually takes place during human language processing.

With regard to parts-of-speech, Chomsky (1970) introduced the features +/-N and +/-V to mark words as possessing the “noun” and “verb” features (adjectives are said to be +N and +V). Baker (2003) attempts to flesh out these features, and describes a noun as an item that bears a referential index, and a verb as an element that takes a specifier (a kind of subject).

Within the generative approach to language acquisition, the problem of mapping words to their parts-of-speech needs to be solved before innate grammatical knowledge can be exploited, as the rules of syntax are thought to take the parts-of-speech as their domain, rather than individual words. Under a TGG view, parts-of-speech are therefore regarded as innate (see also Pinker, 1984).

2.3.2 Cognitive Grammar and Construction Grammar

In the linguistic approaches that are known under the rubric of “Usage-based linguistics”, “Cognitive Grammar” or “Construction Grammar”, the essence of language is its symbolic nature, i.e. the way that forms correspond to meanings (e.g. Bybee, 1985, 1995; Croft, 2001, 2005; Goldberg, 1995, 2003; Kay & Fillmore, 1999; Lakoff, 1987; Langacker, 1987, Tomasello, 2003, 2006). This applies not only to words, but also to grammatical constructions. A construction, according to Croft (2005, p.1), is “an entrenched routine that is generally used in the speech community, and involves a pairing of form and meaning”. Goldberg & Casenhiser (2006) define constructions as “patterns that systematically combine any morphological or phrasal elements”. Tomasello (2006) defines a construction as “prototypically a unit of language that comprises multiple linguistic elements used together for a relatively coherent communicative function, with sub-functions being performed by the elements as well”.

Not only full-utterance structures, but also phrases and even words and morphemes are regarded as constructions. Examples in English include the passive construction, the noun phrase and verb phrase, and the “way” construction (as in “He fought his way through the angry crowd”). Constructions are also thought to exist at varying levels of schematicity, where the description of a construction becomes less schematic as more and more elements are specified as individual morphemes. So, for example, the “way” construction is also an instance of the maximally abstract transitive construction [Subject Verb Object Location]. Likewise, the declarative passive construction could be described as the partially schematic [Subject *BE* Verb-*en by* Obl.], where the elements in italics (*BE*, -*en*, *by*) are specific, and the other elements are abstract.

In Construction Grammar, the goal of language learning is to acquire the constructions of a language. An inventory of these items is often termed a *constructicon*, an allusion to the more common term *lexicon* used to describe the list of words in a language. This inventory is usually taken to be a structured network, with inheritance links between constructions in the network that indicate greater or lesser schematicity.

Under a generative approach, constructions are not taken to be central, and indeed are not even elements of the theory, as they are considered to be epiphenomena that result from the interaction of universal rules which are disconnected from the meanings of the individual elements (Chomsky, 1981). In Construction Grammar, however, constructions are the primary elements of linguistic knowledge.

The Construction Grammar approach treats all elements of grammatical knowledge in a uniform manner, in contrast to TGG which has separate theoretical components for syntactic rules, morphology, syntactic categories, idioms and the lexicon (Croft, 2005). In addition, Construction Grammar is a monostratal theory and does not postulate any distinction between deep and surface structures.

According to Construction Grammarians, linguistic knowledge is first and foremost *knowledge* (Goldberg, 1995, p. 5), and so exhibits properties found in other kinds of knowledge, e.g. prototype effects and organization in associative networks. This viewpoint also strongly implies that linguistic knowledge is acquired rather than innate.

A crucial idea in Construction Grammar is that the use of an element embedded in a construction has the ability to direct the interpretation, or *construal*, of that element; in other words, meaning is defined by context. For example, the word "mean" is a verb in the sentence "What do you *mean*?", but an adjective in "That's pretty *mean*", and a noun in "calculating the *mean* of the values". This effect is typically more obviously demonstrable with open-class words than with closed-class words: the latter are less susceptible to categorial ambiguity, perhaps because their associated meaning is more "syntactic" rather than "substantial" in nature. One might expect, therefore, that constructions would play the dominant role in determining the part-of-speech of a word in Construction Grammar.

2.4 A closer look at open word classes

In the case of closed word classes, there are several divergent classes which perform various grammatical functions, but no classes that are common across all languages. However, we have seen that the classes of nouns and verbs appear to be present universally. The claim that some languages do not have a class of adjectives is perhaps surprising to native speakers of languages that do have such a class. One reason why this claim seems so strange is that the function of ascribing properties to nouns seems to be a fundamental one in language. This touches on a deeper issue in the definition of parts-of-speech: intuitively, the *semantic function* performed by members of some classes such as adjectives seems to go to the essence of that class (whereas other classes, such as the different gender classes of nouns, are semantically more arbitrary, but see Lakoff, 1987). It is instructive to consider the way in which two different CG approaches, namely that of

Langacker (1987) and Croft (2001, 2005), have attempted to accommodate meaning and function in their treatment of parts-of-speech.

2.4.1 Langacker's Cognitive Grammar

Langacker explicitly argues (1987; pp. 183-274; p.420) that the main parts-of-speech in English, such as nouns, verbs and adjectives, can be defined entirely in semantic terms. The definitions of these categories are made not in terms of what they "objectively" are, but in terms of the *cognitive events* that constitute their *conceptualization*. In other words, Langacker does not regard the distribution profiles of the main parts-of-speech as germane to *defining* the parts-of-speech. Note that Langacker does not claim here that distribution is necessarily irrelevant in *acquiring* parts-of-speech; only that, when we have complete knowledge about a word in usage context, *including* its intended meaning, then it is the meaning of the word that determines the correct categorization. In Langacker's view, the grammatical properties of linguistic elements are *symptomatic* of the categories to which they belong (ibid., p. 255), rather than definitional of those categories.

Langacker (1987) distinguishes between *nominal* and *relational predications*[1]. A nominal predication designates a *region* in some conceptual domain, which in turn is defined as a set of interconnected entities: these entities become interconnected "when the cognitive events constituting their conception are coordinated as components of a higher-level event" (p.198). Langacker also describes nominal predications as "things", and nominal predications are represented in English by nouns.

Relational predications, on the other hand, "profile [i.e. make conceptually prominent] the interconnections among conceived entities" (p. 219); these entities can be things, but can also be composed of relations themselves.

[1] Note that Langacker uses the term "predication" to mean "the semantic pole of an linguistic expression" (1987, p. 97); this is a completely different sense from the more traditional one of referring to an event or process, as it is used by e.g. Croft (2001) (see below).

Langacker distinguishes between two kinds of relational predications, on the basis of how they are mentally conceived. Temporal relations, or *processes*, are designated in English by verbs. The evolution of a process through conceived time is sequentially scanned, i.e. there is a series of relations between a trajector and a landmark at various sequential points in time, each of these relations being scanned in sequence in order to constitute the conceptualization of a verb. In the case of *atemporal relations*, the series of relations is not serially scanned, but represented mentally at once as a Gestalt. These concepts are conveyed by adjectives, adverbs and prepositions.

Langacker (1987) also discusses a possible semantic foundation for finer subclass divisions in English, such as the distinction between count and mass nouns. Count nouns are construed as referring to a bounded area in a domain, whereas the referents of mass nouns are unbounded. Langacker suggests that a similar distinction holds between active and stative verbs, i.e. verbs that respectively denote actions that indicate change over time and verbs that do not; in Langacker's view, active verbs are temporally bounded, while stative verbs are unbounded.

Some support for a semantically-based view of parts-of-speech comes from work in neurophysiology. Pulvermüller (1996, 1999) summarizes a large body of research to suggest that the major word classes may be distinguished on neurophysiological grounds that correlate with the meanings of the words. The phonological forms of all words can be shown to activate a number of cell assemblies located in the perisylvian cortices (and strongly lateralized to the left hemisphere). In the case of function words, these are the only brain areas in which assemblies are activated. Highly concrete, imageable content words, on the other hand, activated additional neurons in both hemispheres, with words referring to visual stimuli activating assemblies in the visual areas of the brain, and words referring to actions activating areas of motor cortex.

2.4.2 Croft's Radical Construction Grammar

In Croft's Radical Construction Grammar theory (2001, 2005), only the constructions of a language are primitives – there are no presupposed syntactic/parts-of-speech, but instead categories are derived from the constructions whose slots they fill, so that the constructions are taken to be definitional for the parts-of-speech. Nevertheless, Croft does offer a definition of a language-universal set of parts-of-speech, namely nouns, verbs and adjectives. These are not taken to be the categories of any particular language; instead they are *functional prototypes* which are *typologically* defined.

A central dimension of Croft's theory of parts-of-speech is the notion of the pragmatic functions that utterances serve, labeled by Croft as "*propositional act functions*". He makes particular reference to three such functions: reference, predication and modification. The psychological reality of the pragmatic functions is justified by Croft in terms of the distinct cognitive operations that are carried out during the processing of each particular function. Reference establishes a cognitive file for the referent, while predication does not, but instead "ascribes something to the referent" (p.66) – typically "transitory states of affairs, often in a narrative sequence". Modification adds additional features to the referent's cognitive file.

The other central dimension in Croft's theory is the semantic distinction between words for objects, properties and actions. These three classes are defined in terms of four underlying semantic properties: relationality, stativity, transitoriness and gradability.

Croft defines a two-dimensional *conceptual* space from these two dimensions (propositional act function vs. semantic class), and uses this space as the basis for his definition of the parts-of-speech. Each of the three semantic classes is taken to be the *prototypical* filler of the role in each of the three kinds of propositional act constructions: objects are the prototypes of referring, properties are the prototypes of modifying, and actions are the prototypes of predicating. To the occurrence of each of these semantic classes in their corresponding functions, Croft assigns the terms noun, adjective and verb respectively.

The justification for this purported prototypicality is that these semantic classes are structurally unmarked when fulfilling these roles, whereas the coercion of any of the other classes into a “less-comfortable” function is often structurally coded, particularly by the use of grammatical morphemes, whether affixes, particles, or other function words.

Croft stresses that the categories of any particular language will not be the parts of speech thus defined, but will instead be language-specific. Croft’s prototypical parts-of-speech provide the core of a category, but the boundaries of the category are determined by the distributional facts of the individual language.

2.4.3 The universal status of the main open classes

Both Croft and Langacker seem to provide theoretical support for a threefold view of the open-class parts-of-speech, with Croft emphasizing three prototypical conjunctions of function with meaning, and Langacker pointing to three kinds of mental conceptualization. These two theories are reconcilable: reference can be made only to a concept that can be “pointed” to and hence is reified, or thought of as a “thing”; predication may well require conceptual operations that sequentially scan a series of discrepant situations; modifying an entity is very similar to ascribing a property to it, which is likely to be best represented as an atemporal relationship.

Furthermore, the prototypical meanings identified by Croft (objects, actions and properties) may be the concrete examples in which the three kinds of construals proposed by Langacker are grounded. For instance, when we think of an “idea” as something thing-like (I can get an idea, I can share it with you, you can steal one of mine, my head can become crowded with ideas), the sequence of conceptual operations required to represent “idea” mentally as a delimited region in some domain are, for Langacker, the same as those used to represent cookies and elephants. It may be that these more concrete representations are acquired first. More abstract nouns could then make use of these preestablished cognitive routines; the cue to indicate that they should be conceived of as nominal predications comes from their use in a nominal construction.

The three functions proposed by Croft and the three basic forms of predication proposed by Langacker therefore roughly map onto each other, but are somewhat at odds with suggestions that, while all languages distinguish between nouns and verbs, some languages lack a class of adjective. As suggested earlier, it may be possible to consider an alternative view, in which there are three classes: nouns, verbs and modifiers. The latter class consists of words that modify nouns, words that modify verbs and words that modify other modifiers, and so subsumes not only adjectives and adverbs, but also the "subordinate" noun in noun compounds such as English "*diesel* truck", and also some usages of verbs as modifiers, e.g. "It's a *pretend* aeroplane". Recall that the relevant function as proposed by Croft is that of modification. This viewpoint would also be consistent with Langacker's (1987) view of these modifiers as words that foreground atemporal relations. And in fact, much of the linguistic data used by Schachter & Shopen (2007) to argue against the existence of adjectives in some languages revolves around the use of supposed non-adjectives in a modifier role.

2.4.4 Constructions as clues to word meaning

In order for a hearer to establish a veridical mental representation of what a speaker intends to convey, he or she needs to extract certain information from the utterance which may be assumed to be expressible in all languages.

- The hearer needs to know which elements denote the entities that are implicated in the current utterance (nouns).
- The hearer also needs to know which elements indicate the actions/processes in which these entities are involved, whether really or hypothetically (verbs).
- If there is any additional information supplied in the utterance which will modify the entities or processes or any other elements which themselves modify entities or processes, then the elements indicating this information need to be identified as well (modifiers, verb adjuncts, noun adjuncts).

In addition, there are a number of other kinds of semantic information which may be optionally distinguished in a particular language, including:

- whether entities are to be conceived of as spatially bounded, or processes as temporally bounded. (mass vs. count nouns, active vs. stative verbs)
- relationships between entities which are prototypically spatial, but may be metaphorically extended to non-spatial domains. (prepositions)

- other items of information that may be relevant to communication in specific language communities. (other closed-class words)

The information above is conveyed by indicating that particular elements in an utterance belong to particular categories. There are two possible strategies for conveying this information. The one way is to use words which belong *unambiguously* to only one part-of-speech. The other way is to carry this information in the *constructions* of the language, so that a particular usage of a word is a noun usage whenever it is used in some construction in a position reserved for words belonging to the noun category.

On this view, then, constructions define parts-of-speech; however, their purpose is not merely to allow a hearer to allocate a word to some abstract, linguistically-defined category that plays a role in some linguistic theory. Instead, the purpose is that the hearer may know what construal to place on the particular word. The semantic function of the parts-of-speech is therefore crucial to their definition.

When a construction is viewed as essentially an acceptable sequence of parts-of-speech, then identifying the part-of-speech of a particular word from the construction in which it occurs is a difficult task. Given the categorial ambiguity of many words, a language learner would need to entertain all possible sequences of all of the categories to which each word could belong, which would threaten to produce a combinatorial explosion of possibilities. It would seem difficult for the learning of constructions and parts-of-speech to get started under such circumstances.

Importantly for the work in this thesis, though, the two strategies for conveying category membership are not mutually exclusive. The identification of a particular construction, and hence the assignment of categories to other words, would be greatly facilitated by the presence of a number of words which do not need to be classified on the basis of their constructional context, because they only belong to one category anyway. In English, the words that best fit this description are the function words. Words like “the”, “from”, “but”, “you” and “if” do not easily change their class membership depending on the construction in which they occur. Instead, these words tend to impose category interpretations on other (mostly open-class) words; note how the surrounding function words change the part-of-speech of “water” in, e.g. “the *water*” vs. “to *water* it”

The learner’s task would also be greatly simplified if certain positions in a number of constructions were reserved for *specific* words, rather than for all words from a particular category. Hence, during identification of a construction, the learner would only need to recognize the specific word, rather than listing the categories to which the word could belong, thereby reducing the size of the combinatorial problem.

A major working assumption in this thesis is that English has a great number of these lexically-specific frames and constructions, in which the fixed elements are mostly function words which are unambiguous as to their part-of-speech.

3 Children's knowledge and discovery of parts-of-speech

3.1 Introduction

This Chapter presents a selective review of the psycholinguistic evidence regarding children's discovery of the parts-of-speech of their native language. I will briefly consider the dichotomy between theories that postulate that parts-of-speech are acquired on the basis of their meaning, on the one hand, and on the basis of their use in context, on the other hand.

Subsequently, I will focus on evidence that a great deal of children's early linguistic knowledge is based around very specific constructional frames. Many of the first utterances children are able to produce follow a small number of patterns with specific words combined with variable slots. Many of these specific frames are quite useful as cues to the part-of-speech of the words that occur in their slots. During the first three years of life, children become increasingly able to make use of these frames to guess the part-of-speech and hence the meaning of novel words.

These patterns are often constructed around function words, and evidence is reviewed to show that, despite often omitting function words from their first sentences, children have early knowledge of the phonological forms of function words, and of the positions in which they occur in utterances.

The mechanisms by which these frames are learned are still unknown, and I will review a number of proposals in the literature, as well as empirical evidence that casts light on this issue.

3.2 Bootstrapping theories

There are two main schools of thought about how children get started in discovering the parts-of-speech of their first language, depending on whether the initial source of information about word classes is taken to be the meaning of words (their *semantics*), or their pattern of usage in relation to other words that occur with them in utterances (their *distribution*). As reviewed in Chapter 2, these two potential sources of part-of-speech information for the child mirror the two kinds of information used by linguists to make the same distinction.

3.2.1 Semantic bootstrapping

Theories of semantic bootstrapping appeal to the prototypicality of concrete objects, properties and events as representatives of the main content word classes. The suggestion is that children form their first word categories by grouping together words that refer to the same dimensions of concrete meaning, as embodied in everyday percepts with which they are familiar, such as objects and actions. So for instance, the child might form separate categories of words referring to physical objects, words referring to concrete actions, words referring to properties of entities and words referring to spatial relations (Pinker, 1984); these become the proto-categories out of which the adult categories of nouns, verbs, adjectives and prepositions, respectively, will form.

In a nativist view of semantic bootstrapping such as that of Grimshaw (1981) or Pinker (1984), these semantic categories are innately given, as are the linguistic part-of-speech categories themselves and the mapping between the semantic and linguistic categories. The task for the language learner is then that of *linking* the specific words of his/her language with the innately-given syntactic categories.

In an empiricist version of semantic bootstrapping (e.g. Macnamara, 1982), no innate knowledge is postulated. Rather, the child is presumed to divide the world into conceptual categories such as actions, objects and properties, possibly on the basis of their experience with the world, and then to attempt to map words onto these categories.

The semantic bootstrapping approach is often criticized on the grounds that many of the first words produced or understood by children are not the concrete examples that would be regarded as prototypical of a particular category: for instance, some early nouns are not the words for physical objects, and some verbs are not words for concrete causal actions. But this argument is logically flawed: we should take our evidence not from the early words that are nouns, verbs, etc. for linguists, but from the words that are nouns, verbs, etc., *for the children themselves*. If a child is able to produce utterances with the non-concrete verb “think”, this does not mean that she automatically treats the word “think” as a verb; it would be quite possible for her to be developing a category of verbs, based on concrete actions that she encounters in everyday experience, while at the same time using “think” in a number of utterances without connecting it to her developing part-of-speech. A bootstrapping theory is only intended to show how categories might get started for a child who has no categories yet; it is not required that all eligible words be assimilated into the category.

3.2.2 Distributional bootstrapping

In an alternative view of the origin of the parts-of-speech, a part-of-speech is defined by the various *contexts* in which its member words appear. Typically, not only one but a pattern of several contexts is considered, so that we can speak of the contextual *distribution* of the words. On this view, what makes a noun into a noun is not its function of reference, but the fact that it is used in the same contexts in which nouns occur; as we have seen, this is a standard way of defining parts-of-speech. The hypothesis of distributional bootstrapping posits that even young language-learning children are able to keep track of the distributions of words, and to merge words into categories if they tend to occur in the same contexts. As Maratsos & Chalkley (1980) put it: “… [I]t is not that children learn how verbs act, as though they begin with the notion of verbs. Rather, they come to learn that a certain set of terms may appear in correlated uses.” (p. 133).

In a purely distributional approach, the only information tracked is the various contexts in which a word appears. For this reason, it becomes important in such a theory to specify exactly how a “context” is to be defined, i.e. in terms of immediately adjacent words, immediately adjacent phrasal constituents, co-occurrence with other words in any position in the same utterance, etc. One possibility that is in keeping with a Construction Grammar approach as considered in the previous chapter, is that the relevant contexts for words are the abstract or semi-abstract *constructions* in which a word is used; this proposal is the basic premise of the work in this thesis.

While Maratsos & Chalkley (1980) are often cited as proponents of distributional bootstrapping, their position is *not* that word classes are learned merely by grouping together words that share common formal distributional patterns; instead, the meanings of words are regarded as of equal importance. So, for instance, verbs are words that occur in combination with the suffix *–ed* in certain environments, but in addition, when they occur in these environments, they are words that denote relations which have occurred in the past. On this view, a word class such as verb is defined as a set of terms which can appear in a certain set of correlated *semantic-distributional* patterns.

Similarly, Tomasello (2006) argues that it is the communicative *functions* (e.g. reference, predication) performed by the different parts-of-speech, as well as the distributional characteristics of words in these categories, that are crucial to the discovery of the categories, rather than the meanings that words of a category have in common.

Few psycholinguists would endorse a purely distributional view of part-of-speech bootstrapping that does not take word meaning into account (although some linguists have done so, e.g. Fries, 1952). However, in concrete implementations of the distributional idea, (a number of these studies will be reviewed in Chapter 4), it has been customary to ignore the semantic characteristics of these categories and to make use only of word co-occurrence patterns in textual data. The reason for this is likely to be purely pragmatic: we have huge amounts of textual data available detailing the *words* that were spoken in the presence of children (e.g. the corpora that comprise the CHILDES database; MacWhinney, 2000); by contrast, there are no equally large databases detailing the *meaning* of what was being said (and no prospect even of a notation to represent semantics that would be acceptable to most researchers).

One way in which a process of semantic-distributional bootstrapping might work is a *word-centric* one: as words are used in speech around the language-learning infant, a great variety of aspects of the speech signal and of the real-world concomitants of certain words become available to be connected with the word via *associative learning*. In this way, a word might become associated with aspects of reality that co-occur with its use, as well as with the profile of contexts in which the word occurs. Over time, words may become grouped together to the extent that they share features in both their semantic and contextual distribution.

In the converse, *context-centric* process, the context itself becomes an object of linguistic knowledge, and the words that are used in its slots are stored in memory, along with their semantics. Then when the context is next encountered, these features are accessible, so that eventually contexts can be clustered on the basis of the similarity in their word profiles, and also on the basis of highlighting the same semantic dimensions. And in fact, both of these processes may well operate simultaneously.

The distributional proposal has been heavily criticized by Pinker (1979, 1984, 1987) among others. Pinker asserts that the task of part-of-speech induction from distributional evidence is intractable, *inter alia* because of the large amount of ambiguity prevalent in everyday language. Given the sentences 'John eats rabbits, 'John eats fish' and 'John can fish', Pinker suggests that a child following a distributional strategy might erroneously accept 'John can rabbits' as a valid sentence, due to the ambiguity of the word 'fish' which acts as a noun in one sentence and a verb in another. By the same token, contexts can also be ambiguous; a distributional analysis that starts from 'John eats meat', 'the meat is good' and 'Jane eats slowly' would supposedly accept 'the slowly is good' as a valid sentence, because the frame 'Jane eats X' does not uniquely pinpoint the category of the word occupying the X slot.

Ambiguity is indeed pervasive in language, even in the speech that children are likely to hear. Nelson (1995) shows that many words that occur regularly in the input to children are ambiguous with regards to the part-of-speech to which they belong. For instance, Nelson analyses the usages of "call", "drink", "help", "hug", "kiss" and "walk" in a corpus of child-directed speech, and finds that each of these words is used as a verb on some occasions and a noun on others. Conwell & Morgan (2008, submitted) confirm this with an extensive analysis of six corpora of child-directed speech. In addition, Conwell &

Morgan found that children themselves frequently use certain ambiguous words as both nouns and verbs, and that their pattern of use correlates with that of their caregivers.

Pinker's (1979, 1987) critique, however, is based on a "straw man" version of distributional bootstrapping; few serious distributional proposals are as brittle as Pinker suggests. It is entirely possible that children are able to make use of evidence across the range of contexts in which a word is used in order to determine its category membership, rather than drawing conclusions from single utterances. The point Pinker makes is that words and their contexts can both be ambiguous. But a learning process which *explicitly* considers words and contexts to be potentially ambiguous, and attempts to determine a word's part-of-speech on the basis of a wider range of information than just the identity of the word, may be able to overcome the objections that Pinker raises.

Note, for instance, that in the 'John can rabbits' example, there is likely to be a great deal of distributional information from other utterances to suggest that 'John can X' is a frame that favours verbs only, whereas 'rabbits' is nearly always used as a noun. Combining these two sources of information might be enough in itself to resist the generalization to 'John can rabbits', as the context and word together would be in conflict. Furthermore, hearing 'fish' appear in the same context ('John eats X') as the reliable noun 'rabbits', and subsequently in the reliable verb context 'John can X', could prompt the child to explicitly flag the word 'fish' as ambiguous, and therefore an unreliable basis for categorial generalization. Hence, the child could avoid extrapolating from 'John can fish' to 'John can rabbits'. These considerations suggest that combining category information from both the word and the context in which it occurs may provide for a more accurate categorization strategy than taking only one of these two sources of information into account.

3.2.3 Other bootstrapping theories

It should be noted that other bootstrapping proposals exist in the literature. Kelly (1992) reviews evidence of several correspondences between the *phonological* properties of some words and the word categories to which they belong, and proposes that these properties may play a role in the acquisition of parts-of-speech. For instance, in disyllabic

English words, stress tends to fall on the first syllable for nouns and on the second syllable for verbs; also, the stressed syllables of nouns contain more back vowels and those of verbs more front vowels.

Morgan & Newport (1981) have proposed that the *prosodic* structure of an utterance provides valuable evidence about the syntactic constituents it contains, which could provide indirect evidence about the category to which words in the utterance belong.

3.3 Children's early use of lexically-specific frames

A very significant discovery in the literature on language development, and one which has gained even more currency with the rise of "usage-based approaches" to language acquisition (e.g. Tomasello, 2000; 2003), is that much of children's early linguistic knowledge is organized around very specific words, and is therefore highly *lexically-specific* in nature. This contrasts with a nativist view in which grammatical categories are usually supposed to exist as part of the child's biological endowment of linguistic knowledge.

These approaches are of course also highly compatible with the Cognitive Grammar/Construction Grammar approaches discussed in the previous chapter (e.g. Bybee, 1985; Croft, 2001; Goldberg, 1995; Langacker, 1987). In these linguistic theories, lexically-specific frames may be seen as special cases of the general set of constructions in a language.

While usage-based theorists tend to agree on the highly lexically-specific nature of early language learning, there seem to be two distinct interpretations of the concept of lexical specificity. One interpretation, extensively articulated in the work of Tomasello (e.g. Akhtar & Tomasello, 1997; Olguin & Tomasello, 1993; Tomasello, 1992, 2003), places individual words, and in particular verbs, at the centre of language development, and views the constructions in which particular verbs occur merely as knowledge about how particular words should be used. The other interpretation is represented in the work of Lieven, Pine and colleagues (e.g. Lieven, Pine & Dresner-Barnes, 1992; Pine & Lieven, 1993; Lieven, Pine & Baldwin, 1997), and characterizes children's early language as being organized around certain frequently-occurring, lexically-specific frames or constructions, in which a variety of words can be embedded.

3.3.1 Braine's pivot grammars

One of the most influential studies on early child grammars and parts-of-speech was that of Braine (1976), who showed that a lot of early systematicity in child productions could be expressed in terms of semi-abstract constructions (*limited-scope formulae*) consisting of one fixed element (the *pivot*) and another element which could vary. The descriptions of these child grammars were termed *pivot grammars*. Examples include the celebrated "more X" frame, e.g. "more juice", "more read", "more hot", and the "X off" frame: "shoe off", "hat off", etc. These constructions were said to be restricted to accepting only a limited number of words into the variable slot on the basis of their semantic properties (hence "limited-scope").

Braine (1976) suggested that children who have acquired a set of fixed word patterns that are similar in form might, from these fixed patterns, abstract a productive word pattern with slots, by noticing the similarities between the patterns. The work on pivot grammars constitutes some of the earliest evidence that children could abstract recurring, positionally-defined frames from the input, could deduce what the semantic implications of a frame were, and use the frames productively.

3.3.2 Tomasello's Verb Island hypothesis

Tomasello (1992) has proposed the Verb Island Hypothesis, according to which verbs are the main organizing principle for children's early productions of constructions. Tomasello showed that the earliest uses of each individual verb in the productions of his own two-year-old daughter were restricted to a small number of constructions particular to that verb, with little generalization or systematicity of usage across verbs. So for instance, his daughter would use "cut" only in the construction "cut _", while the word "draw" was used in a variety of constructions (for example, "draw _", "draw _ on _" , "draw _ for _", "_ draw on _"). Furthermore, the range of constructions in which a particular verb was used was extended in an incremental, piecemeal fashion, e.g. by adding a single argument or a tense marking; this again suggests that each verb developed on its own timescale, independently of other verbs. This is the essence of lexical specificity in Tomasello’s work: the ways in which a verb can be used depend on which particular verb it is.

3.3.3 Lieven, Pine and colleagues: lexically-specific frames

A large body of research has documented that the first constructions used by children are typically very concrete in nature, rather than reflecting any grammatical knowledge of abstract rules. This idea has been promoted especially strongly in the work of Lieven and colleagues, who have argued that much of the structure of children’s early multi-word combinations can be explained in terms of frames composed of specific words in combination with open slots that may be filled by variable material.

Lieven et al. (1997) show that between 50% and 70% of the utterances produced by a group of eleven English-learning children could be accounted for by their first 25 constructed productive patterns. (To be regarded as constructed in this scheme, a pattern needs to be attested three times with independently occurring filler elements, after which it is always treated as a constructed pattern.) Lieven et al. (1997) describe grammatical phenomena, and parts-of-speech in particular, as phenomena that emerge from the co-occurrence of these frames with particular semantic or pragmatic contexts.

Lieven, Pine & Rowland (1998) have suggested that, in the case of verb learning, verbs need not be the only elements around which constructions are formed (as proposed by Tomasello, 1992); instead, some constructions may be based on other lexical material such as pronouns and morphological affixes. Lieven et al. (1997) note that many of the patterns exhibited by the children in their study contained verbs followed by the pronoun “it” (e.g. “I carry it”, “my mend it”), and suggest that this may show evidence for the emergence of a *class* of verbs.

Pine & Lieven (1993; see also Lieven, Pine & Dresner Barnes, 1992) provide evidence that there are individual differences between children in the way that they develop their early constructions: some children assemble multi-word utterances by combining two or more familiar single words together, while others start with longer, “frozen”, unanalysed phrases and proceed to analyse these into fixed parts with variable slots, with these last becoming productive positional patterns in which various elements can be used to fill the slot. These patterns form fairly disparate sets across different children. Some of the patterns discovered in Pine & Lieven (1993)'s longitudinal study of five children from 0;11 to 1;8 were "X on", "Mummy X", "Oh X ", "X cat", "No X", "It's a X", "Oh don't X", "Wanna X", "The X", "X shoe", "More X", "X gone", "I X", "X bird", "That X", "There's the X" and "X car".

As corroborating evidence for the "frozen phrase" route into productive pattern acquisition, Pine & Lieven (1993) show that the number of productive patterns is highly correlated with the number of frozen phrases which a child knows. Pine & Lieven suggest that the acquisition of these frozen phrases may therefore play a crucial role in children's language development, and suggest a mechanism by which this might happen: "Such examples presumably reflect a process whereby a phrase is initially segmented as a unit, but is subsequently reanalysed as a result of regularities perceived by the child in other similar words or phrases, resulting in a flexible lexically defined formula…" (p. 567).

In the work of this group of researchers, the frames are described as lexically-specific because there is no assumption that two apparently similar constructions such as "Can you X it?" and "Will you X it?" are linked in terms of the words that can appear in their X slots, just because "can" and "will" are both auxiliary verbs. Two similar frames can, in principle, follow very different developmental pathways.

As Cameron-Faulkner et al. (2003) cogently argue, if a mother used a great number of questions of the structure "Are you …", but used no other auxiliary verbs, it would be unlikely that her child would use a great number of other auxiliary verbs; instead the child's productions are likely to reflect the constructions and words that she has heard used. This argument suggests that children's early language may be structured around highly word-specific constructions. Even if adults can see the correspondence between the frames "Are you ..", "Were you …", "Was he …", not to mention "Have you …", "Can you …", etc., the child may not make an association between these frames, or generalize readily from one of these frames to another.

Usage-based approaches therefore reject the traditional assumption of generative grammar that early knowledge of language is abstract from the onset. This assumption has been used to support the belief that "you can't get there from here" (from the state of linguistic knowledgelessness that usage-based theorists presume to exist in young children, to adult linguistic competence, via processes of learning), and underlies the generativist conclusion that most linguistic knowledge must be innate. Generative grammar assumes, for instance, that once children have determined that theirs is an SVO language, they should be able to use any verb in an SVO construction. However, a number of studies by Tomasello and colleagues have attempted to show that on the contrary, children are quite conservative in their usage of verbs, and will not use them in an SVO construction if they have not heard them so used (e.g. Akhtar & Tomasello, 1997; Olguin & Tomasello, 1993).

While some generativists acknowledge the prevalence of lexically-specific frames in children's speech, they are generally regarded as a "dead-end" in language development, providing a way for the child to communicate before learning "proper" adult syntax (e.g. Radford, 1990). Generative grammar would regard the constructions "A _" and "The _" as essentially the same pattern (Determiner Noun), and one which arises out of the interplay between abstract innate linguistic rules, rather than from merely memorizing the two patterns as coherent units. However, it has been shown that the two sets of nouns that occur embedded in children's uses of these two constructions are quite distinct, and also that children's early productions have a more positionally-fixed word order than would be predicted if productions were generated by abstract syntactic rules (Lieven, Pine & Baldwin, 1997).

3.4 Children's understanding of constructions and frames as cues to word classes

As Braine (1976) and several others (including Lieven et al., 1997) have noted, merely observing that the words that children use in particular pivot constructions belong to the same underlying part-of-speech does not show that these categories exist as organizing principles in the child's linguistic knowledge; they may merely reflect similarities in the things that the child wishes to talk about.

What is important is not so much the utterances that children produce as the utterances that they are able to comprehend. This is because production may be a much harder task for the child: in order to use a particular construction correctly, the child needs to know what it means, in the sense that she has to check that all the semantic facts are true that would license the use of that construction. For instance, the use of the word "the" in the noun phrase "the <Noun>" requires that the referent of the noun phrase is already known to the speaker and hearer, possibly because it has been semantically implied to exist in the context of the things that have already been discussed, or because it has been explicitly introduced earlier in the conversation. The use of "a <Noun>", on the other hand, implies that the referent has not yet been introduced in this way. Use of the appropriate article in one of these two noun phrase forms requires the speaker to consider which of the two situations holds.

By contrast, if we are interested in the child's ability to make use of a particular construction used by someone else, as a guide to interpreting the meaning of certain words that occur in it, then it is necessary merely that the child should be able to *recognize* the construction and to be aware of the semantic implications for a word which occurs in one of its slots (in the case of both "the X", and "a X", the implication is that the "X" word is the label for something which may be an object or person, but is more generally just some entity which is *conceived of* as a thing).

Strong evidence of the role played by constructions in defining categories therefore comes from experimental studies investigating whether children are able to correctly *understand* certain utterances that other people produce. In order to gauge whether children's understanding derives from the constructions themselves, rather than the words that occur in the slots of the constructions, these experiments fill the slots with *novel* words, and attempt to discover how children interpret these words.

An experiment by Brown (1957) was one of the first studies to demonstrate that language-learning children are able to make use of nothing more than the linguistic context in which a novel word occurs in order to guess at its meaning. Three-year-olds and 4-year-olds were exposed to a target picture of, for example, a pair of hands performing an unusual kneading motion on an unfamiliar substance in an oddly-shaped container. Three additional test pictures each contained only one of the components of the original picture (the motion, the substance or the container). Children were introduced to a novel word (say, "sib") in one of three linguistic frames: mass noun ("here you can see *some sib*"), common noun ("here you can see *a sib*"), or verb ("here you can see *sibbing*"), and when asked to pick out another instance of "some sib", "a sib" or "sibbing" from the test picture set, reliably chose the unusual substance, the container, or the kneading motion respectively.

Ever since Brown (1957), it has been taken for granted that children are able to guess the meaning of a novel word from context, and attention has come to focus instead on the course of development of these abilities for each of the main (content) parts-of-speech, and on subtle features of the semantic interpretation of these words (see e.g. Bloom & Markson, 1998, for a broad review). This experimental paradigm therefore implicitly accepts one of the tenets of Construction Grammar, that syntactic structures have a meaning of their own, independent of the words that occur in them (Goldberg, 1995; Kako & Wagner, 2001, Langacker, 1987).

A number of key publications in this paradigm are reviewed in the remainder of this section. Of particular interest are children's abilities to infer that frame slots that accept nouns as their fillers are clues to nominal, "thing-like" meanings, that verb frame slots single out processes, and that adjective frame slots point to properties of objects.

3.4.1 Nouns

Smith and colleagues (e.g. Jones & Smith, 1998; Jones, Smith & Landau, 1991; Landau, Jones & Smith, 1992; Landau, Smith & Jones, 1988; Samuelson & Smith, 1999; Smith, 2001; Yoshida & Smith, 2005) have accumulated a great deal of evidence on the role of linguistic context (and other situational factors) in acquiring nouns.

Smith (2001) argues that, if a cue in stimulus material is reliably associated with paying attention to a particular property, then that cue will eventually become associated with the property and selective attention will obligatorily be directed to that property when the cue is present. Smith suggests that language-learning children are repeatedly exposed to specific linguistic contexts while paying attention to specific properties in the world, and hypothesizes that these linguistic contexts may come to be associated with those properties and eventually come to serve as cues that direct selective attention. In other words, the attentional highlighting of certain aspects of the world contingent on linguistic cues is learned in a non-explicit, automatic manner, as a result of basic domain-general associative learning processes.

In the case of common noun names for objects, Smith and colleagues have argued that the *shape* of a referent object is particularly germane to the meaning of the name, and that the linguistic context may progressively come to draw selective attention to object shape.

Landau, Smith and Jones (1988) introduced children of 2 and 3 years of age to a novel object and named it with a novel word, in the frame “This is a *dax*”, and then asked, for each of a number of test objects, the question, “Is this a dax?” Children were willing to extend the novel word to objects of the same shape as the original object, rather than differently-shaped ones of the same substance and size. This shape bias extended even to objects that were made of other substances than the familiarization object, or that were many times larger. When asked in the same experiment to pick out items that were “like” the familiarization object, children immediately “reverted” to using overall similarity, picking items made of the same material and of the same size, often ignoring shape. This indicates that it was the *act* of naming (*inter alia* by the use of a specific linguistic frame) that led to the shape bias, rather than any intrinsic characteristic of the experimental object itself.

That this bias to shape *develops* has been demonstrated using both cross-sectional and longitudinal methods (Jones, Landau and Smith; reported in Smith, 2001). Young children are initially insensitive to the implications of the "This is a _" frame, but come to make use of these implications over time. Children between the ages of 18 and 24 months with small noun vocabularies are less likely to make a shape interpretation of a novel word when hearing, "Look, this is a dax. Give me another dax," during exposure to the exemplar than children with large noun vocabularies. Furthermore, as individual children's vocabularies develop, they become more likely over time to make a shape interpretation (Smith, 2001).

English syntactically marks subclasses of nouns, distinguishing between common nouns, proper nouns and mass nouns. Investigating the proper noun-count noun distinction, Gelman and Taylor (1988) presented 2-year-olds with a set of four toys (two stuffed animals and two block-like, clearly non-animal toys), and provided the children with a name for one of the toys, in either a common noun frame ("This is a *zav*") or a proper noun frame ("This is *Zav*"). Children were prepared to extend a new common noun name to the other object from the same category as the originally-named toy. However, when the new name was introduced in a proper noun frame, and was originally used to label an animal, it was less readily extended to the other animal, and when it was used to label a block-like toy, it was not applied to the other block toy, but was applied to one or other of the animal toys almost as often as to the originally-named block toy.

The mass noun - count noun distinction in English is marked syntactically by e.g. the use of frames such as "This is a mell" versus "This is some mell". This distinction correlates quite well in English with a corresponding semantic distinction, namely between shape-based categories for count syntax and material-based categories for mass syntax; in addition, most count nouns are names for solid things while most mass nouns are names for non-solid things (Samuelson & Smith, 1999). However, while young children reliably associate novel count noun names for solid objects with shape-based object categories, they are initially less likely to associate novel mass noun names for non-solid objects with material categories. Samuelson & Smith (1999) suggest that knowledge about the names for non-solids emerges only by about the age of three (see also Yoshida & Smith, 2005).

The linguistic frame is not the only relevant factor that children use in interpreting a novel noun. Sometimes when a linguistic context is ambiguous, the child is able to make use of properties of the exemplar object in order to determine which meaning is appropriate (Soja, Carey and Spelke, 1991). For instance, a novel word in the frame "This is my *mell*" is interpreted differently depending on the degree of rigidity of the exemplar object. If the object is made of a non-rigid substance (e.g. foam), then 2- to 2 and a half-year-old children are more likely to extend the word *mell* to other objects made from the same substance (suggesting a mass noun interpretation) than to rigid objects of a different substance with the same shape; the opposite tendency was exhibited when the exemplar object was itself rigid (suggesting a count noun interpretation).

Even for unambiguous count noun frames such as "This is a *dax*", children's readiness to extend novel words to test objects depends in quite subtle ways on the characteristics of the exemplar object. For instance, when the exemplars were equipped with eyes, the texture of the objects was no longer treated as an irrelevant feature, but was taken to be a necessary condition for belonging to the category identified by the novel word (Jones, Smith and Landau, 1991). One way in which this can be made understandable is to think of the eyes as cues to animacy; in this case, the texture of an animal's external body is taxonomically important, i.e. some animals are furry, scaly, etc.

Smith (2001) proposes that a shape bias in the presence of a context such as "That's a _" is developed first, and that other more contextual refinements are developed later, such as attending also to shape and texture in the presence of cues such as having eyes (or wearing shoes: Jones & Smith, 1998). On the other hand, perhaps a more general explanation that accounts for the results of both Landau et al. (1988) and Jones et al. (1991) could be that children understand that the use of a common noun highlights considerations around the essence of an object category; i.e. those facts about an exemplar that are relevant to defining the *type* of object that it is are brought under the spotlight.

In the case of inanimate objects, plausibly shape is the most important dimension to attend to, relating as it does to function. In the case of animate objects, i.e. living creatures, however, one might need to delve more deeply and consider other factors related to the creature’s bodily appearance, including colour or textured patterns on the body surface. It is plausible that the presence of eyes or shoes in the exemplar objects used by Jones et al. (1991) and Jones & Smith (1998) served as a cue to animacy, thereby prompting a more detailed search for characteristic properties. This world knowledge may well have been developed in the context of the particular linguistic frames that Smith (2001) refers to, however. For instance, one might imagine that learning the difference between, say, tigers and lions could be mediated by utterances such as “This is a tiger” and “This is a lion”, combined with attention on the part of the child to the striped markings on the body of the tiger or the mane of the lion. Children would then be triggered by the “This is a X” frame to search for characteristics that define the essence of the referent, but would need to learn which particular features are relevant for different kinds of referents.

Waxman and colleagues (reported in Waxman, 2002) exposed 14 month-olds to a set of similarly-coloured exemplar toy animals named by a novel word in a noun frame (“These are *blickets*”). At test, children were allocated to one of two conditions in which they had to choose between a target exemplar from the category and a distractor which was not in the category, when asked, “Can you give me the *blicket*?” Children were more likely to select the correct target exemplar (e.g. a purple horse) when the distractor differed in object category (a purple chair) than when it differed only in colour (a blue horse), suggesting that 14-month-old children have some knowledge of the conceptual dimensions that are *not* relevant for the common noun frame[2].

[2] This experiment also featured an analogous test with what Waxman et al. regard as an adjective frame (“These are *blickish*. This one is *blickish* and this one is *blickish*”). However, the suffix “-ish” is not entirely reliable as a cue that a word is an adjective, nor do a large proportion of adjectives end with”-ish”; in this case, as Labelle (2005) points out, the frame could just as well be regarded as “This one is _”, and

3.4.2 Verbs

As Gleitman (1990) points out, it is far more difficult for the child to guess what a novel verb means from its ostensive context alone, than to do the same for a novel noun. This is because a given scene can be interpreted in multiple ways, so that the meaning of a verb requires that a certain *perspective* be taken on the events, essentially in terms of the entities involved, and "who-does-what-to-whom". Even more problematic are verbs for which there is no ostensive referent.

Gleitman proposes that the meaning of a verb must instead be inferred on the basis of its syntactic arguments (typically noun phrases). So for instance, since the verb "put" is typically used with three arguments (the putter, the thing that is put somewhere, and the place to where the putting is done), using "put" with three arguments could help to indicate to a language-learning child that "put" refers to an act of putting rather than looking (which typically requires at most two arguments).

Lederer, Gleitman & Gleitman (1995) showed that information about the syntactic structures in which verbs occurred in child-directed speech (the set of syntactic arguments, plus some information about morphological inflection) were sufficient to allow a clustering analysis to form overlapping clusters that corresponded closely to a set of clusters formed from adult native speakers' intuitions of verb semantic similarity.

Often, English subjects and objects are taken from a very small set of words, which might facilitate the process of determining the meaning of the verb. Laakso & Smith (2004) investigated the range of subjects and objects of different verbs in child-directed speech taken from several CHILDES corpora, and showed that a large proportion of these subjects and objects are taken from the small set of English pronouns. Most verbs were found to take either a complement clause (in which case they were usually psychological verbs such as "think", "know", etc.) or else took the pronoun "it" (these included verbs of motion or transfer). Verbs taking "I" as a subject tended to form fixed-phrase-like

the slot may just as easily be filled by a proper noun. Hence, no conclusion can be drawn from the "adjective" condition.

utterance preambles that indicated the epistemic status of the utterance as a whole ("I guess", "I bet", "I think"). Verbs taking "you" as subject indicated the deontic status of the subsequent clause, and included "like", "want" and "need". Laakso & Smith (2004) suggest that the co-occurrences of particular verbs with pronoun objects and subjects could help children to identify subclasses of verbs.

Sethuraman and Goodman (2004) examined a corpus of child-directed speech for the most frequently-used subjects and objects in the transitive construction in English, and found that about 90% of all subjects and over 40% of all objects were pronouns. In addition, 39% of all SVO utterances were Pronoun Verb Pronoun frames, and 50% were Pronoun Verb Noun or Noun Verb Pronoun frames, with the 3 most frequently-occurring frames being "you _ it", "I _ it" and "we _ it". Childers & Tomasello (2001) have shown that 2-and-a-half-year-olds are able to use a nonsense verb productively in a transitive sentence when they have been trained on sentences of the form "He's [verb]-ing it", but not when they have heard only example sentences where both agent and patient are nouns. Childers & Tomasello suggest that English-learning children may build their early constructions around specific configurations of pronouns.

Naigles & Kako (1993) showed that 2-year-old children's interpretations of novel verbs in "neutral" syntactic frames may be shifted by introducing the verbs in transitive or intransitive frames instead. Comparing nonsense verb interpretations in an ambiguous scene (e.g. a rabbit pushes a duck's head forward while both are making the same circular waving motion with their arms) when coupled with neutral syntax ("Look! *Gorping*!"), transitive syntax ("The rabbit is gorping the duck") or intransitive syntax ("The rabbit and the duck are gorping"), Naigles & Kako (1993) found evidence that the use of the transitive frame inclined children towards an interpretation where one character was affecting the other.

As mentioned in Section 3.3.3, a number of experiments by Tomasello and colleagues (e.g. Akhtar & Tomasello, 1997; Olguin & Tomasello, 1993) have shown that children are disinclined to use verbs in a transitive frame when they have only heard them used in intransitive frames. Furthermore, when exposed to a novel verb in a transitive frame, 2-year-olds appear not to be able to make use of the information in the utterance in order to determine who did what to whom (Olguin & Tomasello, 1993). The experiments by Tomasello and colleagues on children's early understanding of verbs have been widely taken to show that 2-year-old children do not have a linguistic category of verbs. However, I believe this result downplays the amount of knowledge that children do have about verb frames.

What is required, in order to demonstrate basic knowledge of the semantic implications of a verb frame is not that the child should be able to use every novel word in the full range of possible contexts in which any English verb can occur, but merely that the child should be able to infer that the word embedded in the construction refers to an action or process, rather than to a physical object or a property. To the extent that children do understand this, they can be said to have minimal knowledge of the verb frame.

A study by Mintz (reported in Mintz 2006a) suggests that children as young as 12 months may have some knowledge of the different frames in which English nouns and verbs occur. Each child was familiarized with four nonsense words, two occurring in plausibly familiar English verb frames (e.g. "You can _", "She wants to _ it") and two in familiar noun frames (e.g. "That's your _", "I see the _ in the room"). At test, children were exposed to grammatical and ungrammatical sentences, i.e. sentences in which words kept their original category, and sentences in which the words switched category. The children listened longer to ungrammatical sentences than grammatical sentences. This effect was due to the verb-frames only; in other words, if children had heard a word in a noun frame during familiarization, they listened longer when the word occurred at test in a verb frame than when it occurred in a noun frame (but there was no analogous effect for words initially introduced in verb frames). This result is remarkable, as it suggests that children may have the ability to make some kind of categorial distinction between English words based solely on their context of usage, at a far earlier age than has previously been believed.

Höhle, Weissenborn, Kiefer, Schulz and Schmitz (2002, 2004) have replicated these results with 15-month-old German-learning infants. If these infants had heard a novel word introduced in a noun frame and then heard it used in several verb frames in a test passage, they listened longer to this passage than to one in which the word continued to be used in noun frames. As in the experiment by Mintz (2006a), no analogous result was found when nonsense words introduced in verb frames were subsequently used in noun frames. Nouns were modeled as following a determiner, whereas verbs were modeled in utterances following a personal pronoun. Höhle et al. suggest that the reason for the difference in results between nouns and verbs may be due to the fact that the determiner is a better predictor of a following noun in German than the personal pronoun is a predictor that the following word is a verb.

3.4.3 Adjectives

To investigate children's knowledge about the linkage between adjectives and object properties, Taylor & Gelman (1988) introduced their 2-year-old subjects to a novel word describing a stuffed toy in either a noun-syntax frame ("This is a mef") or an adjective-syntax frame ("This is a mef one"). Taylor & Gelman found that, if the toy was, for instance, a green dog, and children were allowed to play with a green bird and a yellow dog, they tended to extend the novel word to the other dog when a noun frame had been used and to the other green object when an adjective frame had been used. This result indicates an awareness of adjective frames.

Taylor & Gelman (1988) also found a *familiarity effect* for noun frames, in that children were less likely to extend the new noun to objects other than the originally labeled object when they already knew a label for the object kind (possibly indicating that they interpreted the word as a proper name, or as referring to a subordinate object kind). Taylor & Gelman (1988) failed to find an analogous familiarity effect for adjective frames with their 2-year-olds; that is, interpreting a new adjective as referring to a property was not facilitated by using an object with a known label.

Hall, Waxman & Hurwitz (1993) replicated the study of Taylor & Gelman (1988), and also found no familiarity effect for 2-year-olds. However, 4-year-olds did exhibit a clear familiarity effect: using a familiar object as the focus of a new adjective increased the number of object property interpretations.

Hall et al. (1993) review evidence that, in general, children will make an object kind interpretation when confronted with a new word labeling an object for which the object kind name is unknown, but will tend to make an attribution other than object kind when the object kind name is known. This tendency occurs, to some extent, regardless of the frame in which the word is used. Hall et al. propose that children learn words for concepts such as object parts, properties, etc. in the context of hearing the words used for objects for which they already have a name.

Hall et al. (1993) interpret their results as showing that children have a default assumption that words will refer to object kinds; simultaneously counteracting the effects of this assumption are the lexical contrast principle, which makes an object-kind interpretation less likely with familiar objects, and a growing sensitivity to syntax, which makes a property interpretation more likely with an adjectival frame. These last two factors seem to emerge between the ages of 2 and 4, and had an effect in Hall et al.'s (1993) study only when both were present at the same time (a familiar object described by an adjective). Still, as Taylor & Gelman (1988) did find a preference for property interpretations for adjectives with 2-year-olds, it may be that results depend crucially on the properties of the specific materials used.

For the attributive form of an adjective (e.g. "That's a *feppy* one"), the nature of the noun or pronominal form being modified has an influence on the interpretation of the adjective. Mintz (2005) investigated the circumstances under which children would interpret an attributive adjective as referring to a *property* of the object referred to by a (pro)nominal form, as opposed to referring to the *kind* of the object. There was an interesting interaction between the frame used in the test question related to the novel adjective ("which is the *stoof* one?" vs. "which is the *stoof* thing?") and the familiarity or unfamiliarity of the target object for the child. Thirty-six-month-old infants were able to make an object-property interpretation of a novel adjective in the test question "which is the *stoof* one?" when the object kind was familiar to them, but showed no preference for either an object-property or object-kind interpretation when the object category was unfamiliar. For the test question "which is the *stoof* thing?", this pattern was reversed: children made an object-property interpretation when the object kind was unfamiliar, but showed no preference for object-property over object-kind when the object category was familiar. Mintz interprets these results as suggesting that these 36-month-old children might be exhibiting a sensitivity to the pragmatic constraints peculiar to the two constructions: "one" implies that the speaker has a certain category in mind, and that the hearer has access to this category, so that interlocutor and child share a common frame of knowledge about the intended category; by contrast, "thing" implies that the speaker has no particular object category in mind. It is therefore to be expected that an object-property interpretation could be facilitated when the object category was known to the child, and the word "one" was used. Similarly, the use of "thing" would be natural in the case of an unknown object category. Mintz argues that children may simply have been uncertain about the category intended by the speaker when using "one" with an unfamiliar category, or confused by the pragmatic oddness of using "thing" with a familiar category.

Using the "canonical" adjective frame "This is a *dax* one", Smith, Jones and Landau (1992) were able to draw 36-month-old children's attention away from the shape of an exemplar object and focus it on its glittering colour, *only when* the exemplar and test objects were displayed under special lighting which emphasized the glittering nature of the exemplar and one of the test objects. This suggests that adjectival interpretations are more difficult for the child than nominal interpretations if the relevant semantic dimension is not explicitly highlighted.

Waxman & Markow (1998) were able to coax property interpretations from a group of 21-month-old infants, but only when the test objects were of the same kind as the original object to which the novel adjective referred. If children heard the word being modeled while seeing a yellow spoon, for example, and were tested with a green key and a yellow key, children responded at chance level, i.e. showing no preference for one object over the other. Possibly, using a single object kind throughout helps to focus children's attention on the perceptual similarities and differences between objects.

It seems that English-learning children do understand that adjectives and nouns in these experiments both single out physical attributes of the target objects; however, it may be that it takes a number of years to learn which particular attributes are relevant to a noun interpretation and which to an adjectival interpretation. When familiar objects with known names are used, the space of possible meanings for a novel word is reduced; and when the same objects are used at test as during familiarization, the similarities in the referent objects facilitate a comparison between the two objects and hence the determination of the meaning if the novel adjective.

3.5 The interaction between constructions and embedded words

A number of studies have shown that, when a word occurs embedded in a constructional frame, both the word and the frame may influence the interpretation of the embedded word, and that children also have some knowledge of which words "go with" which frames.

An experiment by Goldberg & Casenhiser (2006; see also Casenhiser & Goldberg, 2005) provides evidence that in some case, specific words used in a construction may "lend their meaning" to the construction as a whole. Adult native English speakers were exposed to a set of exemplars of a novel English construction (Subject Object Verb-*ed*, with all novel verbs in this construction ending with the "morphological affix" *–o*), which was paired in each case with a video depicting the meaning of the utterance. All scenes depicted an entity coming into view suddenly, so that the abstract meaning of the construction was (apparently) something like "Subject causes Object to appear". Subjects were assigned to one of two conditions. In the Balanced condition, a variety of novel verbs were used in the exemplar sentences, with each verb occurring approximately equally frequently. In the High Token Frequency condition, one particular verb occurred far more frequently than the others. At test, subjects were more likely to have "understood" the abstract constructional meaning and to extend it to new exemplars in the High Token Frequency condition than in the Balanced condition. Crucially, the novel words used at test had not been used during familiarization. This suggests that subjects were better able to attach the “appearance” meaning to the one frequently-used verb than to the construction itself; however, once this association had taken place, the construction seemed to take on some of the “appearance” meaning from the word that most frequently occurred in it.

It needs to be taken into account, however, that these results pertain to adults who already have an essentially complete knowledge of English. Amongst other things, these adults may have had access to *meta-knowledge* to the effect that the set of constructions is a closed class, while the set of verbs is open, and hence may have attended to only the verb as a possible focus for the novel meaning. To my knowledge, this experiment has not yet been replicated with language-learning children.

The findings by Goldberg & Casenhiser (2006) relate to *semantic* generalization. In a similar fashion, Ninio (1999) has suggested, based on observational data of children’s productions in both English and Hebrew, that a number of “path-breaking verbs” facilitate the learning of *syntactic* generalizations.

Goldberg, Casenhiser & Sethuraman (2005) provide evidence that the construction is sometimes a better prediction of utterance meaning than the verb. Based on examination of a corpus of child-directed speech, Goldberg et al. examined the link between the “dative” Verb Object Object construction and the semantic meaning of transfer, and between the “locative” Verb Object $\text{Object}_{\text{location}}$ construction and the meaning of movement from one place to another. The conditional probability of the specific meaning given the verb, and of the meaning given the construction, were both high, but the conditional probability of using the “locative” construction given that the utterance meaning involved movement, and of using the dative construction given a meaning of transfer, were higher than the maximum conditional probability of any particular verb given either of these meanings. This is obviously to be expected given that there are many verbs, and only a handful of constructions to be considered. But this result does show the higher predictive value of an element taken from a closed class (the set of English constructions) rather than an open one (the set of English verbs).

Matthews, Lieven, Theakston & Tomasello (2004) have shown that three-year-old children who are asked to repeat a novel construction which appears to have the same meaning as a construction with which they are already familiar, are able to automatically correct the construction to its conventional form. For example, when shown a video of "Bear" bumping into "Elephant" while being told "*Bear Elephant ramming*! Did you see what happened? *Bear Elephant rammed*!", three-year-old children who were asked what happened, corrected the word order to the conventional SVO form by saying "Bear rammed Elephant". An interesting entrenchment effect was shown with two-year-olds, however; these children corrected word order to the conventional SVO order for high-frequency verbs such as "push", but repeated the novel word order that the experimenter had modeled in the case of rare verbs like "ram". This would suggest that the SVO construction is associated with the more common verbs like "push", presumably because these verbs have occurred very often together with SVO utterances in the child's experience. On the other hand, a word such as "ram" has had few opportunities to be used in the SVO or any other construction, and so the link between "ram" and the SVO order is not as well-established. A slightly different possibility is that children may already have formed early categories, corresponding to parts-of-speech, at this stage, and both frequent verbs and the SVO construction may have been associated, not with each other, but each with the underlying category.

Rarer words such as "ram" may not yet have been associated with any word category, and hence there would have been few guidelines on how these words should be used in context. Similar results were obtained by Theakston (2004); when children and adults were asked to rate the grammaticality of a set of sentences (which were all ungrammatical), low-frequency verbs were found more acceptable than high-frequency words when occurring in the same contexts.

The fact that the three-year-olds managed to use the correct SVO construction with the infrequent words suggests that the construction has some psychological reality for children of this age that is independent of particular (in this case unfamiliar) verbs. This is in line with Goldberg's (1995) proposal that constructions exist as psychological units in their own right for adult native speakers, and with the work reviewed in earlier sections in which even novel words are interpreted correctly according to their context.

3.6 The role of function words in lexically-specific frames

3.6.1 Early knowledge of function words

On the face of it, the functors of English (closed-class words, clitics and morphological affixes) should provide an extremely useful source of distributional information for the purpose of part-of-speech bootstrapping. Gerken, Landau & Remez (1990) point out that function words could be crucial in the two tasks of *word segmentation* and *word labeling* (category assignment). Function words are potentially useful in segmentation because recognizing the relatively small number of function words makes it easier to separate out the far more heterogeneous open-class words that are interspersed between them. Function words could also aid labeling, because they occur in very stereotypical positional relations to open-class words, for instance, "the" is often followed by a noun (or sometimes by an adjective which is followed by a noun), and "-ing" is usually preceded by a verb root.

However, it is a very robust finding that English-learning children tend to omit function words from their early utterances (e.g. Bloom, 1970; Bowerman, 1973), resulting in a distinctive "telegrammatic" style of speech. Even when children are asked merely to repeat the words of a speaker, they still omit the functors (e.g. Eilers, 1975).

This phenomenon is often taken as evidence that children simply do not process function words. It has been proposed that this is due to the reduced salience of function words, either semantically or phonologically. Function words are less "contentful" than content words, and so it may be the case that children initially attend selectively only to open-class words with concrete real-world referents (e.g. Brown, 1973). Function words are also phonologically less salient, being typically unstressed, shorter in length and with reduced vowels (usually schwa), and so may be less likely to be detected and processed (e.g. Gleitman & Wanner, 1982).

Of course, this logic is flawed: the evidence that children do not *produce* function words is no more than that; we cannot conclude without further experimentation that they also do not *process* these elements when they hear them, and in fact there is an overwhelming body of evidence to demonstrate that they do process them.

It may still be the case that the phonologically and semantically diminished substance of function words is indeed the explanation for why these elements are initially not produced in speaking, even though they are processed during hearing.

For instance, Gerken (e.g. 1991, 1994) has put forward a metrical explanation for children's omissions of function words, pronominal sentential subjects, and unstressed syllables at the beginnings of multisyllabic words, according to which English-speaking children make use of a strong stress – weak stress frame for speech production, and attempt to make their productions fit against this frame. Gerken (1987, reported in Gerken, 1991) has shown that children who are asked to repeat the sentence "Pete pushes the dog" will be more likely to omit the article "the" than the inflectional morpheme "-es", and argues that this is because the stress pattern of the sentence yields a strong-weak foot followed by a weak-strong foot; this latter foot does not fit the frame and so the initial weak syllable is omitted. Gerken (1991) suggests that the strong-weak preference may be due entirely to motor constraints on the alternation between strong and weak syllables.

Of course, function words are also semantically less contentful than open-class words. As Labelle (2005) points out, children's omission of function words in their productions may be due to the fact that they haven't worked out what the meanings or functions of these words are yet. In adult English, "the" contrasts with "a" in indicating that the referent of the subsequent noun is known to the speaker and the hearer, either because it has previously been referred to, or because it is understood from the pragmatic situation. An English-learning child may be able to make use of distributional facts, e.g. that “the” is often followed by a noun, without yet knowing the semantic distinctions that should be considered before deciding to use “the”.

Morgan, Shi & Allopenna (1996) have shown that content and function words in child-directed speech can be distinguished from each other based on significant differences on a number of phonological and acoustic properties, including syllable complexity, vowel diphthongization, vowel duration, syllable amplitude and vowel quality. Shi, Werker & Morgan (1999) have shown that newborn infants have some sensitivity to the differences between English function and content words (which is presumably based on the afore-mentioned phonological and acoustic properties); neonates habituated to hearing a list of function words dishabituated more strongly to a list of content words than to a new list of function words.

A number of studies by Shi and colleagues have also attempted to track the trajectory of function word recognition in continuous speech in various languages. So for instance, 8-month-old French-learning infants exhibited different listening times to passages containing an embedded function word to which they had been familiarized previously compared to passages not containing the function word. Six-month-olds were able to do this only for function words that were very dissimilar phonologically (Shi, Marquis & Gauthier, 2006; Shi, 2007). Similar results have been found for 7- to 9-month-old (but not 6-month-old) German-learning infants (Höhle & Weissenborn, 2003). Shi, Cutler, Werker & Cruickshank (2006) showed that 11-month-old English-learning infants could use familiar function words (such as “the”) to segment out a nonsense noun in continuous speech. Eight-month-olds were also able to make use of “the” to identify and segment out the noun, but responded similarly when the functor was the nonsense functor “kuh”, suggesting that these infants’ phonological representation of function words was still underspecified.

Shi, Werker and Cutler (2006) showed that when 8-, 11- and 13-month-olds were exposed to English "functor + content word" sequences and "nonsense functor + content word" sequences, the 13-month-olds showed an ability to distinguish between the two kinds of sequences, while 8-month-olds could not, and 11-month-olds were intermediate in their abilities, as would be expected if the recognition of functors developed over this period.

Gerken, Landau & Remez (1990) found that 2-year-old children were better able to imitate content words if they occurred adjacent to English functors rather than to nonsense functors (e.g. "push" in "Pete pushes the dog" is more accurately repeated than "push" in "Pete pusho na dog"), supporting the idea that functors may aid in word segmentation and therefore in recognition of content words. Furthermore, children were more likely to omit English functors than nonsense syllable "functors" occurring in the same position in the sentence, even when these were superficially phonologically similar to English functors. This suggests that children have a fairly detailed segmental representation of these elements, contrary to what one might expect from their frequent use of filler syllables in the place of function words (e.g. Gleitman & Wanner, 1982).

3.6.2 Early knowledge of the relationship between function words and other words

Although children's speech may sound telegrammatic, it has long been established that speaking back to children in a telegrammatic manner impairs their comprehension of what is being said. Children respond correctly more often to utterances that contain function words ("give me the ball") than to utterances that omitted them ("give ball"; Petretic & Tweney, 1977; Shipley, Smith & Gleitman, 1969).

Gerken & McIntosh (1993) took this result further, to attempt to show that children might be able to use their knowledge of English functors to label constituents in a sentence. This could happen only if children are aware not only of the function words themselves, but also of the contexts in which they are licensed to occur. So, for instance, if children were familiar with the function words "the" and "was", but not with the patterns in which they are allowed to occur in sentences, then they might find the sentence "find was bird for me" just as acceptable (and comprehensible) as "find the bird for me". Gerken & McIntosh (1993) exposed 2-year-olds to one of four variants of a grammatical English sentence. Each sentence requested the children to identify an item in a picture book that the experimenter and child were reading together. The four variants differed in the element which appeared in the position occupied by "the". One variant contained a nonsense syllable ("Find gub bird for me"), another contained a legitimate function word which was not acceptable in the particular context ("Find was bird for me"), another contained no word at all ("Find bird for me"), and the fourth control variant was the original utterance. Children more readily responded by pointing to the requested object in the grammatical control utterance than to both the utterance with the nonsense function word and the utterance with the misplaced function word. However, this study failed to replicate the advantage shown in the earlier studies for the grammatical condition over the omitted function word condition.

Kedar, Casasola and Lust (2004), replicating the work of Gerken & McIntosh (1993) with the more sensitive experimental paradigm of preferential looking, were also able to show an advantage for grammatical utterances over utterances with omitted function words. In terms of the latency of responses (the time taken to look toward the target), children looked significantly more quickly to the target in the grammatical than in the omitted-word condition.

An experiment by Santelmann & Jusczyk (1998) examined children's knowledge of non-adjacent dependencies in English. Fifteen- and 18-month-old children heard passages containing several instances of either a grammatical discontinuous dependency, made up of the word "is", followed by an optional adverb, a main verb and the inflectional morpheme "-ing" (e.g. "everybody *is cheerfully baking* bread"), or an non-dependency with "can" instead of "is" ("everybody *can cheerfully baking* bread"). The 18-month-olds could distinguish between grammatical and ungrammatical sentences, whereas the 15-month-olds could not. However, this ability was also conditional on the length of the material intervening between the two dependent elements: when the intervening material exceeded three syllables in length, the 18-month-olds no longer distinguished between grammatical and ungrammatical sentences. Santelmann & Jusczyk interpret the result as showing that 18-month-olds can track discontinuous dependencies between function morphemes, but that they can do so only over a window with a limited size.

German offers a far larger number of possible structures that may legally appear between an auxiliary and a verb inflection than English does. Höhle, Schmitz, Santelmann & Weissenborn (2006) attempted to replicate the results of Santelmann & Jusczyk (1998) with German materials containing various forms of the German present perfect with intervening adverbs (e.g. “Der Hamster hat leise gequiekt”, “the hamster squeaked softly”, vs. “*Der Hamster kann leise gequiekt”, “the hamster can squeaked softly”). However, unlike the English-speaking children studied by Santelmann & Jusczyk (1998), their 18-month-old subjects did not show a differential listening preference between grammatical and ungrammatical sentences. However, when the intervening adverb was replaced by a noun phrase consisting of an article followed by a noun (e.g. “Der Hamster hat/kann das Korn genascht”, “the hamster nibbled/can nibbled the grain”), 18-month-olds were able to distinguish between the grammatical and ungrammatical sentences. Höhle et al. (2006) interpret this result as indicating that the presence of the very familiar article indicated to the child that the intervening material (the noun phrase) was a unit of German, thereby allowing the material to be classified, and so facilitating the recognition of the surrounding dependency between “hat” and the “ge-X-t” verb structure. Arguably, this was not possible with the adverb, as German adverbs are not marked with obvious cues to their category membership. Höhle et al. (2006) suggest that because English adverbs often end with the common morpheme “-ly”, this may facilitate the induction of the adverb category in English, which is later extended to adverbs with similar distributional properties, regardless of whether they have the “-ly” ending, and that this may have brought about the difference in results between German- and English-learning children when adverbs were used. An alternative explanation not considered by Höhle et al. (2006) is that the German-learning children may have acquired an explicit schematic frame for the entire utterance, of the shape “Der X hat der Y ge-Z-t “, and recognized this frame in the stimulus materials. The additional function word material may have made this frame more readily recognizable than the “adverb” utterance frame “Der X hat Y ge-Z-t”; this is in line with evidence (discussed in Section 3.6.4) that having a larger number of convergent cues to category membership facilitates categorization.

An important set of studies by Shady (1996) delineates the extent to which very young children are aware of the difference between function words and content words in English, and of the allowable patterns in which function words occur in utterances. Infants of the age of ten and a half months were offered a choice of listening to a grammatical English passage, or to the same passage with the function words replaced with nonsense syllables. An example pair of sentences from the unmodified and modified passages would be (p. 27):

There was once a little kitten who was born in a dark cozy closet.

There [haI] once [I] little kitten who [haI] born in [I] dark cozy closet.

Infants showed a significant preference to listen to the grammatical material, even though the nonsense syllables were designed to be phonologically similar to natural English morphemes. This suggests that even infants of this very young age had fairly detailed segmental knowledge of the function words of English, and could notice the difference between a passage that contained these familiar words, and one which did not. This suggests that children of this age may be sensitive to the "texture" of English sentences, as constituted out of the function words in utterances, and may be able to notice when this texture is violated.

In another experiment, the passage was modified so as to retain the original function words, while changing several of the content words to nonsense words. For example:

There was once a little kitten who was born in a dark cozy closet.

There was once a little [mafIt] [gə] was [tɛk] in a dark cozy closet.

By contrast to the nonsense function-word condition, infants showed no preference for either the unmodified or modified passage in this condition. This may be viewed in the light of the “texture” interpretation that I offered above: the modified utterances in this case preserved the basic, “background” structure of English utterances. Even though the substituted words were unfamiliar nonsense words, they occupied acceptable “grammatical” positions where real content words might have occurred, and so this material was not unusual for young children in the way that the nonsense function-word material was.

In further experiments, Shady attempted to delve deeper into the specificity of children’s knowledge about function words in English. These experiments were designed to gauge whether children were aware that specific function words could only appear in specific positions in an utterance. It would, in theory, have been possible for children to be finely aware of the segmental structure of English function words, but nevertheless to treat all *valid* function words as essentially substitutable for each other in their different positions in an utterance. Examples of modified and unmodified sentences in this experiment were

This man has bought two cakes.

Has man this bought two cakes.

Ten and a half and 12 and a half month-olds showed no preference for either passage; however, 16-month-olds preferred the grammatical passage. These results taken together suggest that while children may at 10 and a half months be able to recognize function words in continuous English, and be aware of where function words should occur in relation to the other words in utterances, they may still have underspecified knowledge of exactly where *particular* function words are allowed to appear, and that this knowledge develops somewhere between the ages of 12 and a half to 16 months.

3.6.3 How do lexically-specific frames develop?

The question of how children come to learn the lexically-specific frames of their language is currently under-researched. However, a small number of researchers have carried out theoretical and empirical research in this area; this work is reviewed here.

In contrast to the widely-held view that children initially learn individual words and then attempt to combine them into utterances, Tomasello (2006) argues that on the contrary, children attempt from the beginning to emulate complete adult communicative utterances, as evidence for instance in the use of conventional intonational patterns associated with requests, questions and commands. Tomasello (2003, 2006) suggests that *utterance-level constructions* play a prominent role in language development: these are verbal expressions that can be used as complete utterances, and that are associated in a routinized way with certain communicative functions. Peters (1977) proposes that children may follow one of two quite different strategies in language learning. In the so-called *analytic* strategy, the child starts with basic elements such as single words, and combines these so as to form increasingly larger units. By contrast, the *Gestalt* strategy starts from a full utterance and proceeds to discover its parts.

There are a number of factors that are believed to influence the degree to which a particular construction is productive in accepting particular elements as slot fillers. Unsurprisingly, two of these factors relate directly to how productively the construction is actually *used* in the input to the child. A construction's *type frequency* (e.g. Bybee, 1985) refers to the sheer number of different element types that have been attested to occur as slot fillers in that construction. A related concept is that of the *openness* of the construction, i.e. the degree of variability between elements that occur in the construction (Bybee, 1995). Intuitively, exposure to a large number of filler elements may help the child not only to focus on the abstract properties that these elements have in common and that possibly license their use in the construction, but also to ignore properties of individual slot fillers which are not common to all slot fillers and hence irrelevant to occurring in the particular construction. Lastly, the process of *preemption* or *blocking* also plays a role: if a child already has mental access to a construction-filler pair that express a particular meaning, she will be disinclined to use the filler in a different construction with apparently the same meaning (Goldberg, 1995).

One of the few proposals in the literature attempting to spell out the explicit strategy by which children might learn the constructional frames of their language is that by Peters

(1983). It is worth considering her suggestions in some detail, as they are similar in several ways to a number of computational approaches to be reviewed in the next chapter (e.g. Adriaans, 1992; Van Zaanen, 2001), and also to the work that will be presented in later chapters of the current work.

Peters is concerned with determining what the natural units of language are for language-learning children. In line with her earlier (1977) proposal regarding analytic and Gestalt language learning strategies, she emphasizes that some children may extract single words form connected speech while others extract phrases.

Peters suggests that the basic starting point for children will be to start with entire utterances, and to hypothesize that they are linguistic units. Utterances are then stored in memory, together with the salient features of their situational context. Note that this does not commit children to an early Gestalt strategy; on the contrary, many utterances spoken to children are composed of single words, and words spoken in isolation may be more likely to be learned than others (Brent & Siskind, 2001). Hence, extracting utterances from the speech stream may be a way to learn both full-utterance structures and single words, as well as intermediate units such as clauses and phrases: essentially any elements that are sufficiently coherent and autonomous for adults that they should utter them independently may in this way become putative units for the child.

Peters proposes a number of phonologically-based heuristics for identifying utterances: an utterance is bounded by silence, utterances are stretches of speech that are suprasegmentally delimited (e.g. by word-initial stress, utterance-final pitch, etc.), an utterance is a speech tune or melody, and an utterance is a rhythmic pattern of speech.

Next, these early units may be segmented into smaller ones by employing a number of heuristic strategies aimed at producing putative subunits. These are: segment off the beginnings and ends of utterances, segment off stressed syllables, segment the utterance at rhythmically or intonationally salient places and segment out subunits that are repeated within a unit.

In addition, Peters offers two segmentation strategies that begin to bring what the child currently hears into a relationship with her previously extracted units. Using the SG:MATCH1 heuristic, the child will postulate that if the beginning or end of a unit is another unit, the remainder of the larger unit is also a candidate for being a unit. More generally, the SG:MATCH2 heuristic states that if two units overlap in having shared phonological material, the material that they have in common may be a unit, as well as any residues that the two units do not have in common. This strategy is also the essence of Van Zaanen's (2001) Alignment-Based Learning system, to be discussed in detail in Chapter 4.

Putative units are evaluated by the child in terms of their utility, according to yet another set of heuristics, namely: prefer units that have been produced by more than one segmentation heuristic; prefer frequently-proposed units; prefer units with clear associated meanings. A particularly important heuristic states that, if a child produces a putative unit and is not understood (including the case of "sounding funny" to the child herself), then this may prompt a reanalysis of the compositional structure of the unit.

Peters provides an account of how children might discover the morphosyntactic frames of a language. This is said to happen when children have identified a set of units that can be subdivided (by any of the segmentation heuristics), such that one subunit occurs in all the units in the set, while the other units may vary. This leads to the postulation of a frame with a fixed element, combined with a slot that may be filled with variable material. Crucially, in Peters' theory a frame can only be discovered in this way once the child has also determined which *semantic* features are common to all of the instances of the frame. Hence, Peters' theory insists that frame formation must always be mediated by semantics, and will consequently make quite different predictions about the trajectory of frame discovery than "purely distributional" theories.

The more often the segmentation heuristics give rise to a particular frame, the stronger will be the evidence that the frame is a valid unit in the language. Here too, the frame may be evaluated by being produced, as was the case for smaller non-abstract units. Lastly, Peters proposes that the child will also look for frames that generalize the ones she already has, whether by moving from a frame with one slot to one with multiple slots, turning slot fillers themselves into frames that are hierarchically embedded within the larger frame, and moving towards wholly abstract frames.

Lieven, Behrens, Speares and Tomasello (2003) attempted to trace the developmental trajectory of one two-year-old child's lexically-specific frames by showing how her later productions could be derived from earlier ones by one of five simple change operations: substituting one element for another, adding on an element, dropping an element, inserting an element in a non-final position, and swapping the positions of two elements. The authors found that 74% of the child's utterances in the final session recorded could be formed from earlier utterances by a single "editing" operation. The results suggest that children might be able to extend the range of their productions by a number of simple operations on elements which are already part of their linguistic repertoire. The idea of linking an utterance to a precursor utterance that is a minimal number of editing operations away is highly reminiscent of one version of Van Zaanen's (2001) ABL system (reviewed in Chapter 4), in which utterances are "aligned" (matched as in Peters' (1983) SG-MATCH2 heuristic) in such a way as to minimize the *edit distance* between the two utterances (the number of insertions, deletions and substitutions required to turn one utterance into the other).

3.6.4 Artificial and other language learning experiments

Methodologically, it would be useful to be able to know exactly which language input a child has been exposed to when he or she demonstrates a particular linguistic behaviour. Given the difficulty of achieving this level of knowledge in the case of a child's first language, researchers have instead conducted experiments using *artificial languages*, in which adults and children are exposed to strings from an artificially-created language, typically generated by a grammar for that language. Occasionally, real, unfamiliar foreign languages take the place of the artificial languages. In this paradigm, a child's exposure to the language can be controlled precisely.

A well-established result in the artificial language literature is that language learners are unable to induce parts-of-speech which are distinguishable from one another on only one cue per training item; at least some of the training items need to have convergent (at least two) cues to category membership. In an early experiment by Smith (1966), subjects were exposed to sentences following either a MN or a PQ pattern, where M, N, P and Q represent categories of words, and at test were required to identify strings that they had heard during familiarization. Subjects as readily accepted test strings of the form MQ and PN as they did MN and PQ sentences, suggesting that they had formed categories corresponding to words that occurred in first and in last positions, but had not tracked word co-occurrences.

Braine (1987) found that subjects could distinguish between provided real-world referents for his artificial words, with half of the words in one category having male referents, while half of the words in the other category had female referents. Other studies have shown successful categorization in adults when marking words in the N and Q categories with affixes as well as final positional order (Frigo & McDonald, 1998).

In a similar vein, Gerken, Wilson & Lewis (2005) familiarized English-speaking children aged 1;5 with examples of the Russian noun paradigm of masculine vs. feminine nouns. Subsequently, infants looked longer in the direction of a sound speaker producing sequences of ungrammatical items than one producing grammatical ones, but only when some of the words presented during familiarization provided *two* cues to either masculine or feminine gender, rather than one.

Braine (1987) distinguishes between two phases of category learning. In the first phase, learners associate *individual elements* from the M and P categories with cues that distinguish between the N and Q category. At this stage, there is not yet the ability to treat the M or P categories *as* categories, i.e. by generalizing from, say, an N element's occurrence with one M element to expect that the N element may co-occur with any M element. This knowledge is said by Braine (1987) to develop during the second phase, when several M- and P-elements are categorized based on their co-occurrence with the N or Q cues. This step can be seen as the formation of *clusters* of contexts, one in which N elements could occur and another in which Q elements could occur.

Gómez & Lakusta (2004) successfully showed that 12-month-olds are able to exhibit the first of these two forms of learning. The infants in their study heard sentences of either “a X” or “b Y” structure, in which the “a” and “b” are presumably intended to convey that these elements play a similar role in the artificial language to function words (being few in number, and being the elements that serve as cues to the category), while the “X” and “Y” elements are like open-class words. There were two “a’ and two “b” elements, and whether a word belonged to category “a” or “b” predicted whether the following word was monosyllabic or disyllabic. Infants were able to distinguish between valid from invalid utterances after training. This ability was statistical rather than absolute in nature, as it persisted even when a sixth of the training sentences were ungrammatical, but not when the proportion of invalid sentences was increased to a third.

In work by Mintz (2002), adults were exposed to sentences from an artificial language which for the most part exhibited an “x Y z” pattern, where the “x” and “z” elements served as framing elements for the more variable Y elements. The artificial language contained two classes of words that could be distinguished on the basis of the different sets of frames in which they occurred. During test, subjects were required to choose between sentences that matched a medial (Y) word from one word class with one of the sentence frames of its “own” family of frames, and sentences featuring these same frames with words from the other category. Subjects had heard neither of the two sentences, but significantly often preferred the frame/medial word match to the mismatch. (Interestingly, they also showed a significant preference for the mismatch sentences over random sentences.) Mintz (2002) suggests that the frame/medial word combination provides a figure/ground distinction which makes it clear which are the words over which a category is to be induced. According to Mintz, this may have been crucial in allowing the distributional analysis to succeed in this case, where it has failed in many others. Note also that in this experiment, the induction of categories was done not only with classes of words but also classes of frames that were reliable cues to the category.

Gómez (2002) explored this idea further. Her artificial language consisted of strings of the “a X b” variety, but with the set size of the X category varied to contain 2, 6, 12 or 24 elements. It was hypothesized that it would be easier for learners to notice the non-adjacent dependency between “a” and “b” as set size increased, because these elements would become salient as more or less table anchoring points against the variability of the X category. In fact, Gómez suggests that the decrease in predictability of the X element given “a” might trigger a search for a non-adjacent dependency with higher reliability. By contrast, a more orthodox view of learning would predict that high-order elements are built by chunking together adjacent elements to form a lower-order chunk, which is itself based on a similar adjacent dependency between chunks, all the way down to atomic elements. On such a view, non-adjacent dependencies cannot be learned, and the best learning outcome in this experiment would occur with a set size of 2, where some strong adjacent dependencies between “a” and the X element still exist. In fact, the results supported Gómez’s hypothesis that longer-range dependencies can be learned when adjacent dependencies are absent. Learners were much better able to judge as “correct” strings from the artificial language as the X category size increased, with a sharp discontinuous jump in accuracy between set sizes 12 and 24.

It should be noted, however, that it may be more accurate to say that this study demonstrated the learning of a number of frame-style constructions, rather than the induction of a word class to fill the X slot. To show this, it would have been necessary to have two X classes, with different sets of frames for each class, as was done in Mintz’s (2002) study. Another point to note is that there were convergent cues in this experiment to indicate which elements constituted the frame and which the “focal”, “content-like” word: the “a” and “b” elements were monosyllabic, while the “X” word was disyllabic, something which would have been a reliable cue to closed-class vs. open-class membership in English.

Gómez & Maye (2005) investigated the developmental trajectory of the ability to learn non-adjacent dependencies using the same paradigm as in Gómez (2002), and found that infants are able to track these dependencies at 15 months, but not yet at 12 months.

3.7 Synthesis: a possible role for lexically-specific frames in part-of-speech bootstrapping

In this chapter, I have reviewed evidence from studies on children's language development that suggest that familiar lexically-specific frames are items of linguistic knowledge for the child, not only appearing in their productions (Lieven et al., 1993, 1997; Tomasello, 1992) but also shaping their comprehension of utterances. Evidence from a variety of sources has converged to suggest that children become able, during the course of development, to make use of lexically-specific frames, made up of specific words plus slots, to determine the meaning of the words that occur in the slots (Brown, 1957; Naigles & Kako, 1993; Smith, 2001, Taylor & Gelman, 1988). This is in line with the notion in Construction Grammar that constructions serve to place specific construals on embedded elements (Bybee, 1985, 1995; Goldberg, 1995; Langacker, 1987).

From the literature, there are guidelines to indicate some of the properties that frames should have in order to facilitate the learning of categories:

- The frames should occur frequently (Bybee, 1985)
- There should be considerable variation in the fillers that can appear in the slots (Bybee, 1995)
- There should be multiple converging cues to indicate category membership for slot fillers; these plausibly serve to distinguish the frame from the filler (Braine, 1987; Smith, 1969)

In addition, the specific words in these lexically-specific frames are very often function words (Laakso & Smith, 2004; Lieven et al., 1997), and children are able to recognize these elements from an early age (Gerken, 1987; Gerken et al., 1990; Shi et al., 2006; Shi, 2007), are sensitive to the patterns in which function words co-occur in utterances (Gerken & McIntosh, 1993; Shady, 1996), and can use function words as cues to the parts-of-speech of proximate words (Höhle et al, 2002, 2004; Mintz, 2006a).

This idea has implications for distributional approaches to part-of-speech bootstrapping. Such approaches hold that parts-of-speech are induced from the contexts in which words occur, but it is an open question which particular contexts are to be used. It would be useful to propose explicit mechanisms by means of which children may discover the lexically-specific frames that occur in their native language, and to show that these frames are efficacious in allowing the parts-of-speech to be bootstrapped.

In the next chapter I will review several explicit, computational approaches that have attempted to discover the parts-of-speech and identify common structures in English from large corpora of natural English text.

4 Computational models of automatic part-of-speech induction

Several researchers have proposed computational models for the automatic discovery of both parts-of-speech and syntactic structures of a natural language, using a sample of spoken or written text from that language. This section provides a selective review of some of the most influential approaches and models.

4.1 Lexical categorization from contextual distribution

A very important set of studies in the 1990s managed to demonstrate that it was possible to group words together into categories that are very similar to traditional linguistic categories, purely on the basis of the distributional contexts in which those words occur in a sample of the target language.

4.1.1 Finch

Finch (1993; Finch, Chater & Redington, 1995) investigated whether parts-of-speech could be induced merely by grouping together words that appeared in similar contexts in a corpus. In Finch's work, the context of a target or *focal word* (this term will also be used in this thesis) was based on four "streams" of information, obtained from the words before and after a focal word, and the words two before and two after a focal word.

Finch's algorithm collected co-occurrence statistics from a corpus of natural language (logs of computer newsgroups). For each of 2000 focal words, it determined the frequency with which each of the 147 most common words in the corpus occurred in each of the positions 2 words before, 1 word before, 1 word after and 2 words after the focal word. This produced a set of 4 usage vectors for each word, corresponding to each of the 4 contexts. These vectors were concatenated to produce a single vector of distributional statistics.

Next, words were subjected to a hierarchical clustering analysis. Hierarchical clustering (described in more detail in Chapter 5) produces groups of items by grouping together items that are similar, in that they possess certain shared characteristics. In this case, words were grouped/clustered together if they were used in similar contexts, i.e. had similar usage vectors.

Finch (1993) presents the hierarchical trees produced by this clustering process. The groupings found by this purely unsupervised process were highly intuitive, grouping together, for instance, modal verbs (“would”, “should”, “can”, “won’t”, “doesn’t”, “hadn’t”), prepositions (“on”, “at”, “in”, “among”, “beyond”, “above”), both accusative (“it”, “me”, “them”, “her”, “us”) and nominative (“I”, “they”, “we”, “he”, “she”) forms of pronouns, as well as categories corresponding closely to determiners, nouns and adjectives, and many others.

It is remarkable that such a detailed and successful categorization of words can be obtained from such very local and limited contextual information. It is important to note, however, that Finch’s model allocates each word type to one and only one cluster. This means that every token of a particular word type is categorized into the same category, (even though many word types are ambiguous), so that word tokens from the same type belong to different categories depending on their context. Finch considers this issue of ambiguity, but concludes that it is negligible, citing a finding by Church (1992) that 90% of word tokens belong to the majority category associated with their word type.

Having allocated words to categories, Finch next considered longer linguistic constituents made up of several words. Rather than collecting distributional information for a huge number of individual word sequences, Finch replaced all words in the corpus with labels corresponding to the categories of the words as determined by the lexical clustering, and collected all category label sequences of length 1, 2 and 3 in the corpus. For the 3000 most common such sequences (likely to correspond to linguistic units such as phrases), Finch again collected contextual information, this time based on the category labels of the 2 words before and the 2 words after the focal item, and again a clustering analysis was performed. This produced a categorization of what Finch terms X-level short sequences. These sequences could very often be interpreted linguistically. One class appeared to correspond to noun phrases (“her status”, “the following section”, “her favourite colour”), and another to word sequences elaborating on and modifying the copula BE ("was simply", "just shouldn't be", "am probably not"). A very interesting finding was that some of the sequences were not directly mappable to traditional linguistic categories such as one of the phrasal constituents of English, but could perhaps be more compatible with approaches such as Categorial Grammar. For instance, one class contained the sequences "use the", "break into these", "add an". When a noun phrase is appended to each of these, a verb phrase is formed, so that one could perhaps describe this category in a Categorial Grammar framework as something like a "VP \ NP" category.

Next, the most common sequences of X-level sequences were discovered, by considering the 3000 most common X-level sequence sequences of length up to 3, and again taking as context the 2 adjacent X-level sequences on either side. This time, clustering managed to move up to the level of simple sentences, as well as noun phrase, verb phrases and infinitival complements ("to accept this attitude", "to be at an end", etc.).

Redington, Chater, Huang, Chang, Finch and Chen (1995) replicated the results of Finch (1993) with word categorization for Mandarin, and again found that the main linguistic categories were automatically discovered by this method.

4.1.2 Redington, Chater and Finch

The work by Finch (1993) was later replicated and extended by Redington, Chater and Finch (1998), who made use of child-directed speech taken from the CHILDES database, and varied aspects of the algorithm including the kinds of input used, and the definition of context. Among their results were the findings that:

- near contexts (immediately adjacent words) are more informative than distant ones (second-next/previous words and further);
- combining preceding and following contexts is better than using just one or the other;
- it is beneficial to use only the most frequent words as context words (the 150 most frequent were found to be optimal) - this is likely due to their being function words rather than content words;
- removing function words from the corpus adversely affected performance;

- nouns are the easiest class to identify using this method, followed by verbs, while function word classes tended to be poorly identified;
- performance is improved by taking utterance boundaries into account and not collecting contextual information that straddles such boundaries;
- the technique worked equally well with adult-directed natural speech.

It is important to note that the quantitative evaluation performed by Redington, Chater and Finch did not penalize their model for its disregard for word type ambiguity, as in each case they evaluated their empirical classification against a gold standard which specifically also ignored ambiguity: a word type was assigned to its major category according to the Collins Cobuild lexical database. Hence, an ambiguous word such as "empty", for example, is assigned exclusively to the category of verb, because that is its main category in the opinions of the linguists who compiled the Cobuild database.

4.1.3 Mintz, Newport and Bever

Mintz, Newport and Bever (2002) replicated the basic approach of Finch and colleagues with a number of child-directed corpora from CHILDES. Mintz et al. made use of the 200 most common words as both focal words and context words, using the cosine distance as a distance measure, and using their own *purity* measure to evaluate success.

In one experiment, Mintz et al. varied the size of the context window on either side of the word, between 1, 2 and 8, and found that a size-8 window improved the algorithm's ability to identify the cluster of verbs (although nouns were identified readily at any window size). A further manipulation made use of the positions of function words in the input: instead of using a fixed-size window, a context was used which stretched from the last function word before the focal word (inclusive) to the first function word after the focal word (exclusive). This had the effect that function words were used as phrasal delimiters. These contexts proved to be at least as good as and sometimes better than using all words, for identifying both nouns and verbs. Lastly, the authors found that verb categorization was improved by collapsing all function words into a single token, i.e. by not distinguishing between different function word types (at least for a 1-word context). This result is somewhat at odds with the work of Redington, Chater and Finch (1998) who found the opposite result: when they replaced all function words with a single token, categorization accuracy declined in their experiment. It is unclear what the reason for this discrepancy might be.

A major problem with all of the experimental studies reported above is that they make the simplifying assumption that all tokens of a particular word type belong to the same category. But in fact, word types are ambiguous, and tokens may belong to one category or another depending on the context in which they are used. The distributional clustering approaches reviewed in the following three sections attempt to rectify this shortcoming.

4.1.4 Clark

Clark (2000, 2001) takes a similar approach to that of Finch (1993). However, in Clark's model, the context of a focal word is defined as the *conjunction* of the two words on either side of a focal word (e.g. in "the mouse ran up the clock", the context of "mouse" is "the … ran"), and the context vector used for clustering expresses the distribution of a focal word into all flanking word pair contexts in which it occurs. This is in contrast to the work of Finch and colleagues (Finch, 1993; Redington, Chater and Finch, 1998) and Mintz et al. (2002), who used context vectors where left and right context words were treated as independent contexts of a word, so that a word which occurred 100 times with word *a* on its left, and 100 *different* times with word *b* on its right, would be indistinguishable from a word which occurred 100 times with the word pair *a* ... *b* flanking it. A word is taken to define a probability distribution over all contexts, which is estimated from a corpus; words with similar distributions are clustered together.

Clark's method can account to some extent for word type ambiguity: once prototypical context distributions have been obtained for each of the clusters, the context distribution for a particular word is modelled as a linear combination of these prototypes. Note, however, that Clark's method will give an abstract breakdown into the various categories for a particular word type, but cannot be used directly to categorize any particular instance (token) of a word based on its context, as humans can do. This is because contexts are not treated as linguistic objects that can themselves be subject to categorization. As a result, Clark's approach is also not able to categorize a novel or rare word based on its context: instead, Clark gives each rare word a default category profile based on the overall frequency of occurrence of the various categories in the corpus.

Clark reports good results for his algorithm, although his quantitative evaluation is rather different from that of the other researchers mentioned in this section, so that a direct comparison is difficult.

4.1.5 Frequent Frames

In Mintz's (2003, 2006a, 2006b) Frequent Frames model, as in all of the other studies reviewed so far, the local context surrounding a word is taken to be significant for lexical categorization. Frequent frames are defined as a disjunct frame made up of the word immediately preceding the focal word, a slot for the focal word, and then the word immediately following the focal word, so that all frequent frames have the form *a* _ *b*, with *a* and *b* standing for fixed words, and the underscore being a slot that can accept variable material. Hence, Frequent Frames are the same contexts that were used by Clark (2001) in the work reviewed in the previous section (but are put to different use, as discussed below).

Once all frames of this form have been collected from a corpus, only the most *frequent* ones are retained for the purpose of categorization. This reflects the intuition that, if a pair of words co-occur frequently on either side of another word in utterances, this is likely to be due to some meaningful linguistic relationship between them.

Mintz aims to improve on these models of Finch (1993), Redington et al (1998) and others by making use only of contexts that (i) have been shown to be attended to by children and (ii) are reliable indicators of part-of-speech.

Firstly, Mintz (2003) defends the psychological plausibility of these structures in part on the basis of the results of Gómez and Maye (2005), where children were able to learn about disjunctive word co-occurrences in an artificial language where all sentences had *a* _ *b* structures; hence it is known that children are able to attend to these structures.

Secondly, Mintz (2003) argues that making use of the *non-adjunct pair* of words flanking a focal word provides a far more reliable indicator of the part-of-speech of the focal word than considering only the context on the left or on the right of the word in isolation from each other, as was done in Finch (1993), Redington et al. (1998), Mintz et al (2002) and Clark (2001).

The most significant aspect of Mintz's work is that it seems to be one of the few implemented distributional clustering models to acknowledge that word types are ambiguous, and that the part-of-speech of two tokens of the same word type might differ according to context. Instead of assigning all instances of a word type to the same category, as done in all the other studies reviewed so far, Mintz leaves it to the *specific* contextual frame to determine the part-of-speech of a word token. The Frequent Frames approach therefore replaces clusters of word types with clusters of frame types.

Another important aspect is that the frames are frequent; Mintz does not make use of all frames of the form *a* _ *b*, but only of the ones that recur regularly, and hence are arguably of linguistic significance; it is the input that guides the selection of frames.
During categorization, all words that occur in the same frequent frame are treated as belonging to the same category. In all of Mintz's reported simulations, the number of frames actually used for categorization is low: for example, only the 45 most frequent frames are used in Mintz (2003). These frames provide a very accurate categorization of the words that occur in them, and Mintz (2003) reports accuracy scores in excess of 0.97. Nevertheless, completeness scores are low, as a result of the large number of frames. The process is therefore extended by amalgamating two frames into a larger group whenever they share more than 20% of their filler words in common. This process is carried out transitively, so that if A and B satisfy the criterion, and B and C satisfy the criterion, then A, B and C are merged. After this process, all focal words occurring in the same group of frames are allocated to the same category. Evaluation of this categorization still yields high accuracy, and completeness is greatly increased (both in excess of 0.9 in Mintz (2003)).

4.1.6 Schütze

Schütze (1993) attempted to extend the earlier work of Finch (1993) by considering the categorization of word tokens in context, rather than word types. Instead of forming clusters based on distribution vectors that represent the words occurring on either side of a focal word, Schütze made use of a concatenated vector made up of four different context vectors that represented the left context vector of the focal word, the right context vector of the focal word, and also the right context vector of the word to the left of the focal word, and the left context vector of the word to the right of the focal word. Clusters were then formed based on similarity between these concatenated vectors for each focal word. In this way, two focal words in context are considered to be similar not only if they tend to occur in the same contexts, but also if the context of one word occurs with the same kinds of words as the context of the other.

When applied to the Brown corpus (Kucera & Francis, 1967), this technique was extremely successful in finding clusters corresponding to the traditional parts-of-speech in English. This high degree of success would seem to be due to the fact that, instead of clustering only word types together, Schütze clustered word tokens in context, combining information about both the distribution of the word and the distribution of its context simultaneously.

4.1.7 Cartwright & Brent

Cartwright and Brent (1997) propose an incremental strategy, implemented in a computer simulation, by which children could group words into categories. This strategy is based on the principle of finding an optimal model of a language, and involves the child coming up with an explicit mapping from each of the words in an utterance to their respective categories (i.e. the sentence “the cat slept” might be mapped to “Determiner Noun Verb”). This mapping is effected on the basis of an underlying model of the language, which lists the words and categories of the language, the possible categories to which words may belong, and the possible full-utterance frames that are permissible in the language (these frames are described in terms of a sequence of parts-of-speech only, rather than making reference to any specific words).

The model is updated after every new utterance that the child hears, and the learning process aims to minimize the amount of entropy (i.e. complexity) in the model. While Cartwright and Brent's category learning strategy is solidly founded on the goal of entropy minimization, they also show how pursuing this goal can have the effect that the learning system adheres to a number of psychologically intuitive heuristics, including: minimizing the number of frames, minimizing the number of categories, creating frames with the highest possible frequency, minimizing the number of words that can belong to more than one category, minimizing the number of words in a category, and favouring the use of large categories in the frames. Clearly, all of these heuristics contribute to producing a parsimonious language model. Cartwright and Brent's model was tested on real child-directed speech from the CHILDES database, and was shown to produce highly accurate categories of words (although completeness was low due to the large number of categories produced).

4.2 Syntax learning

In this section I review a number of computational models aimed at discovering the syntax of a language rather than its parts-of-speech *per se*, but which nevertheless have some relevance for the current experiments with lexically-specific frames, because they need to tackle the problem of part-of-speech induction along the way.

Most of these approaches are concerned with the discovery of *paradigmatic* and *syntagmatic* patterns. Syntagmatic patterns are "horizontal" and involve sequences of linguistic symbols (e.g. phonemes or words) that occur frequently enough in sequence *in single sentences* to warrant postulating that they form a larger unit. Paradigmatic patterns are "vertical" and correspond to sets of elements that occur frequently enough in the same context *across different sentences* to warrant postulating that they belong to the same category.

4.2.1 EMILE

EMILE is an algorithm developed by Adriaans (1992, 1999; Vervoort, 2000) for the purpose of discovering the grammatical structure of a natural language from a corpus, and expressing that structure in terms of a categorial grammar. A categorial grammar describes the various grammatical constituents and words of a language as belonging to certain categories; these categories are defined in terms of the potential of constituents to combine with constituents from other categories to form compound constituents. Expressions that belong to the same category can be substituted for each other in any context. EMILE has been extended since its initial formulation, and the current version is 4.1 .

In EMILE, an equal role is played by a word and the context in which it occurs, in line with the categorial grammar approach. For a specific word in context to be assigned to a specific category, both the word ("expression" in the EMILE terminology) and its context would need to be associated to that category. Category assignment is therefore highly context-sensitive: another expression occurring in that same context, or that same expression occurring in another context, might be assigned to an entirely different category.

Categories are therefore defined in terms of both a set of expressions, and a set of contexts. The cross-product of these two sets defines a set of context-expression pairs that each belong to the category.

The algorithm attempts to discover the structure in a language from a corpus sample of sentences from that language. For every sentence, the algorithm generates every possible division of the sentence into three parts *a b c*, where *b* is non-empty and one of either *a* or *c* is allowed (but not required) to be empty. The string *b* is the *expression* which enjoys the current focus, and the *context* of *b* is created from the concatenation of *a*, a placeholder slot (.), and *c*. So for example, from the sentence "John loves Mary", the following context-expression pairs are formed:
["(.) loves Mary", "John"], ["John (.) Mary", "loves"], ["John loves (.)", "Mary"], ["(.) Mary", "John loves"], ["John (.)", "loves Mary"], ["(.)", "John loves Mary"].

Next, a co-occurrence matrix is formed, listing all the combinations of contexts and expressions that have occurred together in the corpus. An example is shown in Table 1.

	play	cry	school
Do you want to (.) ?	×	×	
Are you going to (.) ?	×	×	×
He’s in (.)			×

Table 1. An example of the working of the EMILE clustering algorithm. The shaded areas are clusters created from the example sentences. The potential cluster outlined with a dashed box is discarded, as all its examples are covered by the other two clusters.

From the matrix, the algorithm attempts to find larger groups of expressions that occur in the same contexts, and large sets of contexts that accept the same expressions in their placeholder slots. These groups are *clusters* of expressions and contexts. By adding a new expression (respectively context) to the cluster, it is implied that that expression (context) can appear in all of the contexts already added to the cluster (can accept all the expressions already added to the cluster as slot-fillers). The algorithm attempts to increase the size of the cluster as much as it can.

The algorithm creates clusters by starting from a single context-expression pair that is not covered by the current set of clusters, and randomly adds either contexts or expressions to the cluster. The rectangle delimited by the set of contexts and expressions implicitly defines context-expression pairs that are presumed to be valid, and belong to the category corresponding to the current cluster. The process of expanding the cluster by adding expressions or contexts is bound by the constraint that a certain proportion of the context-expression pairs defined by the cluster should actually have been attested in the corpus. This proportion can be specified by the researcher using a system parameter.

All clusters that are completely covered by one or more other clusters are discarded at the end of the cluster discovery phase. In practice, even if there are a small number of cells in a cluster that are not covered by other clusters, the cluster will still be discarded. A parameter of the EMILE system, *type_usefulness_required*, states how many cells a cluster needs to cover on its own in order to avoid being discarded.

In Table 1, the context “Are you going to (.) ?” has appeared in the corpus with the expressions “play”, “cry” and “school”. It is an ambiguous context that can take both verbs and nouns. At the same time, the unambiguous context “Do you want to (.) ?” appears with “play” and “cry” only, and likewise the unambiguous context “He’s in (.)” appears with “school” only. The algorithm would produce the clusters

[(“Do you want to (.) ?”, “Are you going to (.) ?”), (“play”, “cry”)]
[(“Are you going to (.) ?”, “He’s in (.)”), (“school”)]
[(“Are you going to (.) ?”), (“play”, “cry”, “school”)].

However, the last cluster (demarcated by a dashed line in Table 1) is completely covered by the other two, and so the only two clusters produced would be composed out of the cells in the shaded areas in Table 1, covering

[(“Do you want to (.)?”, “Are you going to (.) ?”), (“play”, “cry”)] and
[(“Are you going to (.)?”, “He’s in (.)”), (“school”)],

corresponding nicely to the categories of nouns and verbs.

After forming clusters, the algorithm attempts to make use of these clusters to induce the rules of a context-free grammar. EMILE performs quite well (Van Zaanen and Adriaans, 2001) in discovering syntactic constituents in two small structured corpora (the ATIS (Marcus, Santorini and Marcinkiewicz, 1993) and OVIS (Bonnema, Bod and Scha, 1997) corpora), but is less successful in finding structure in larger corpora.

4.2.2 ABL

The Alignment-Based Learning (ABL) system by Van Zaanen (2001; see also Geertzen & Van Zaanen, 2004) was developed for the purpose of discovering the syntactic structure of a language purely from an unannotated corpus of sentences from that language. The basic idea of ABL, dating back at least to the work of the American Structural Linguists (e.g. Harris, 1954), is that, if two sentences have some amount of phonological material in common, and some material that differs, then the differing material is likely to be a linguistic *constituent* of the language in question. For example, in the pair of sentences

Cookie Monster sees the red apple
Big Bird sees a pear

the word “sees” is common to both sentences, and so it is possible to hypothesize that “Cookie Monster”, “Big Bird”, “the red apple” and “a pear” are all constituents of English. Furthermore, it can be postulated that “Cookie Monster” and “Big Bird” belong to the same grammatical category, since they are *substitutable* for each other. The same can be said for “the red apple” and “a pear”. The two sentences can be said to be *aligned* with each other on the shared word “sees”.

ABL works by firstly (in the so-called *alignment learning* phase) considering every pair of sentences in the corpus, and attempting to align them on their shared structure. The unequal parts that differ between the two sentences are then taken to be hypothesized constituents. The constituents are annotated in place with a symbol, corresponding to a non-terminal symbol in a phrase-structure grammar. So for instance the sentences

Oscar sees Bert and
Oscar sees Big Bird

would yield the annotated sentences

[*Oscar sees* [*Bert*]$_1$]$_0$ and
[*Oscar sees* [*Big Bird*]$_1$]$_0$

with the 1 serving as the non-terminal symbol which represents the context “Oscar sees” in which the constituents were encountered, and the 0 indicating the full-sentence context. Each hypothesis in ABL therefore entails postulating that a putative constituent (inside the brackets) occurred in a particular context (surrounding the brackets).

In the particular algorithm followed in Van Zaanen (2001), the annotation symbol represents the nonterminal in the grammar to which the two constituents are hypothesized to belong. In other words, any two constituents that occur in the same context are regarded as being substitutable for each other in the grammar. Van Zaanen points out (2001, p.31) that this is not an intrinsic feature of the ABL framework, and that other ways of clustering constituents together could be considered.

In practice, there are often a number of different possible alignments for a pair of sentences. One variant of ABL accepts all possible alignments between sentences as valid hypotheses, under the assumption that more evidence for the hypothesis will accrue later if it is a valid one, whereas an invalid hypothesis will be swamped by other, better hypotheses (this competition between hypotheses takes place during the subsequent selection learning phase of ABL). Another variant uses only the alignment that requires the fewest number of transformations to change one sentence into the other (as measured by an edit distance metric).

The algorithm proceeds to discover more and more structure between alignments and to annotate the corpus in this way. It is also possible to discover structures nested to several hierarchical levels within each other.

During the alignment learning phase, it is possible for the system to postulate partially overlapping hypotheses on the same sentence. In a phrase-structure grammar description, these hypotheses would be contradictory, because there would be no way to construct a tree diagram for the sentence that accommodates all the hypotheses at once. Therefore, a subsequent *selection learning* phase filters through the sets of hypotheses for each sentence and tries to remove overlapping hypotheses. Three methods to do this are explored in Van Zaanen (2001). The least empirically successful method favours an hypothesis that was derived early on in the execution of the algorithm over any newer conflicting hypothesis. The two other methods prefer those hypotheses that are the most probable, in terms of their support in the analysed corpus. Recall that an hypothesis has two components: a constituent and a context. One method favours the hypothesis that proposes the constituent candidate that occurs the more frequently in the corpus; the other favours the constituent candidate that occurs the more frequently *in that particular context* in the corpus.

Finally, ABL derives a grammar for the language, which can take the form of either a stochastic context-free grammar or a stochastic tree-substitution grammar. ABL also gives good results in deriving bracketed structures when tested on the ATIS and OVIS treebanks (Van Zaanen and Adriaans, 2001).

4.2.3 SNPR/ICMAUS

There have been many computer models which approach the problem of syntactic structure discovery in language as one of *information compression*. In these models, utterances of a language are viewed as exhibiting a mixture of highly repetitive structure with occasional idiosyncratic elements. The purpose of data compression is to produce the most parsimonious redescription of the data (typically an unstructured corpus in the case of language structure discovery). The corpus is rewritten according to a set of rewrite rules (rules to replace a longer pattern in the data with a shorter one). A parsimonious model is one which is *minimally complex*; this notion is generally expressed in terms of *Minimum Description Length (MDL)*. Under MDL, complexity is measured in terms of both the length (size) of the data after applying the rewrite rules, and the length of the rules themselves (written in some appropriate description language). Li and Vitányi (1997) provide a thorough discussion of MDL and the related concept Minimum Message Length, under the general banner of Kolmogorov complexity. MDL principles are believed to be in operation in a variety of aspects of human cognition and learning (see Chater & Vitányi, 2003, for a review).

One of the most influential syntax learning models has been the SNPR model of Wolff (1982). The model successively reads in samples of a corpus of a natural language, employing a number of heuristics to iteratively compress a grammar for the language.

During parsing of the corpus, the system attempts to redescribe each sentence in terms of the grammatical constructs that it currently possesses. The set of rewrites that comes closest to successfully accounting for the sentence (analogous to deriving the sentence from an initial sentence nonterminal S in a context-sensitive grammar) is further exploited in order to add new rules to the grammar. New rules entail postulating either syntagmatic relationships (i.e. concatenating two or more terminals or nonterminals) or paradigmatic relationships (postulating that an element belongs to a certain class if it occurs in a context in which another item from that class has previously occurred). The system attempts to parse the sentence in a top-down fashion, starting from rules that have the full-sentence symbol on their left-hand side, and favouring rules that have been applied recently.

Grammars are made more compact by replacing recurrent syntagmatic patterns with a single symbol (thereby shortening rewrite rules), or by replacing several different elements that occur in the same context with the same symbol (thereby dispensing with a number of rewrite rules). The heuristics used in this pattern extraction phase entail, on each iteration, selecting the most frequent syntagmatic sequence of elements (conditional to a particular context) for rewriting as a syntagmatic element, and selecting the most frequent pair of elements appearing in the same context for rewriting as a paradigmatic element. This bias to frequency is intended to balance the bias towards recency encapsulated in the parsing phase.

Rewrite rules may be rebuilt, if they postulate the occurrence in a particular context of an element or syntagmatic pattern which is never attested in that context. In this case, a new paradigmatic class is created which does not contain the absent element, and all syntagmatic patterns involving the old class are rewritten to refer to the new class.

The most recent incarnation of this system is the Information Compression by Multiple Alignment, Unification and Search (ICMAUS) model (Wolff, 2001, 2002a, 2002b, 2002c). Wolff views the ICMAUS framework as potentially providing a way to unify all disciplines concerned with data representation, including psychology, computer science (e.g. fuzzy pattern recognition, information retrieval and probabilistic reasoning), and also mathematics and logic. Whereas the heuristics used in SNPR were designed to have the expected effect of compressing the grammar in the long run, ICMAUS calculates the amount of compression provided by a particular rewrite rule directly in order to select the most powerfully compressing rules.

4.2.4 GRIDS

Langley and Stromsten (2000) present their GRIDS model as "a rational reconstruction of Wolff's SNPR". GRIDS starts with a set of rewrite rules for a particular corpus, consisting of rules of the form S → X … Y for every sentence, and rules of the form X → *w* for every word type (terminal) *w*. The algorithm makes use of a metric for grammatical simplicity based on both the size of the model (i.e. the lengths of all rules) and the length of the data in the corpus, expressed as the summed length of derivations for each corpus sentence using the rules of the grammar. GRIDS then attempts to incrementally modify its grammar so as to maximize the simplicity of the grammar. The constraint of taking into account the size of the model prevents over-specific, under-generalizing grammars that merely generate every sentence from its own nonterminal, as such a model would not make use of any redundancy in the data. The constraint of taking into account the length of the derivations guards against over-generalizing grammars, as such a model would require many long, detailed rules in order to distinguish valid from potential invalid sentences.

GRIDS repeatedly performs a series of paradigmatic merges (adding a rule that merges two right-hand sides together into a new nonterminal), choosing all merges that increase simplicity, until no such merges are available. At this point, the algorithm switches to syntagmatic merges, adding rewrite rules with length-2 strings of symbols on their right-hand-side and new nonterminals on their left. Again, only merges that increase simplicity are chosen, and this process repeats until no more such merges exist. The algorithm switches back and forth between paradigmatic and syntagmatic merging until no further merge of either kind is possible. This algorithm is able to discover the underlying grammars for artificially created test data sets, but has not been evaluated on real data. Similar approaches have been taken by Grünwald (1996), and by Stolcke (1994) who attempted to select a probabilistic context-free grammar that maximizes posterior Bayesian probability.

4.2.5 ADIOS

The ADIOS system by Solan (2006) has a great number of similarities to the work of Wolff (1982, 2001). It also attempts to redescribe sentences from a corpus by substituting consecutive elements with syntagmatic symbols, and to merge paradigmatically substitutable elements into equivalence classes. In ADIOS, the criterion for merging items into a syntagmatic pattern is somewhat more complicated than in SNPR. Starting from any element, the system calculates the leftward conditional probability, i.e. the conditional probability of the item on the left of the central element. Concatenating these two elements, the system next calculates the conditional probability of the item directly to the left of the concatenated string formed from the two elements, given that the concatenated string has occurred. This step is iterated to produce a sequence of leftward-facing probabilities that are each conditional on the growing concatenated sequence to their right. This sequence is "pushed out" from the original element in this way, until the difference between two successive conditional probabilities in the probability sequence drops below a pre-specified constant threshold. The same process is then performed in a rightward direction. The sequence of elements between the two boundaries where the conditional probability drops below the threshold on either side is now taken to be a potential syntagmatic pattern. Next, the binomial probability of this sequence is calculated in a leftward and rightward direction, and the sequence with the highest significance on these binomial tests is selected as the "leading pattern" for that sentence, and is rewritten using a new symbol. The algorithm therefore searches for units that are disjunct from their context, and only treats them as patterns when they occur in that context.

At the same time, ADIOS takes a broad view of what constitutes a syntagmatic pattern: at each slot in the developing pattern, the system considers not just the element which occurs there in the sentence, but also all other elements which can be substituted in that same context to produce other sentences in the corpus. A geometric mean of the probability of occurrence for all of these elements given their context is calculated instead of the simple conditional probability. If the pattern involving this set of substitutable elements is chosen to be the leading pattern, then these elements are merged into a paradigmatic class, and the pattern is rewritten so as to make reference to the symbol for that class. The system will reuse a previous paradigmatic class if it has more than 65% overlap with the new class; otherwise, a new class is created.

SNPR and ADIOS clearly take a broadly similar approach to syntactic structure discovery, with their main points of difference being the probability-based criterion for pattern extraction in ADIOS, and the rule-rebuilding step in SNPR. While the individual principles of these two algorithms are relatively simple, the way that these principles interact when operating on an actual corpus of natural language may well lead to highly complex and emergent differences between them. Therefore it is somewhat difficult to compare the two models analytically. In addition, neither model has been tested on a large child-directed corpus such as those in the CHILDES repository, specifically for the purpose of determining whether their paradigmatic classes correspond to the parts-of-speech one would expect in English, or whether their syntagmatic structures correspond to constructions (although Solan (2006) has applied ADIOS to corpora from CHILDES (MacWhinney, 2000), and has evaluated its ability to complete sentences in a standardized test of second-language English competence).

4.2.6 MOSAIC

The MOSAIC model of Freudenthal, Pine and Gobet (2002, 2005; Gobet, Freudenthal & Pine, 2004; Gobet & Pine, 1997) is an extension of the well-known EPAM model of perception and learning (Feigenbaum & Simon, 1984), designed to account specifically for phenomena of language learning. MOSAIC once again encapsulates the ideas of syntagmatic and paradigmatic patterns. Items (initially words, but later also syntagmatic chunks of words and other chunks) are entered as nodes in a discrimination network (similar to a decision tree in machine learning). Nodes which occur together with a frequency greater than a threshold value are chunked together syntagmatically. Nodes with more than 20% overlap in both their immediately preceding and immediately following context are linked in the model (equivalent to joining them paradigmatically).

A unique feature of MOSAIC is that utterances from a corpus are entered into the network from their back to their front, i.e. the final words of an utterance are added to the network before the first words are added. This reflects empirical evidence to suggest that children are also better able to learn material that occurs at the end of an utterance. Hence, a node can be created for a word in an utterance only when all material following it has already been added to the network. Even when this has occurred, a node will only be added with a certain probability, dependent on the distance from the word to the end of the sentence, and on the amount of the corpus that has been processed already (this probability increases as more and more nodes are added).

The MOSAIC research group have shown that the model is able to produce results that are compatible with developmental data on language acquisition regarding phenomena such as the production of optional infinitives and the omission of sentential subjects.

The bias to utterance-final material in MOSAIC and the bias to recency in SNPR are the most important substantial differences between these two models. (MOSAIC operates in strict iterative fashion, while the working of SNPR may be batch-like, but not much is likely to hinge on this distinction.) Again, it is difficult to assess how the two models would fare when compared directly on the same corpus.

4.2.7 Powers

Some of the earliest studies in unsupervised language learning were undertaken by Powers (1983, 1989, 1991, 1997a). Powers's work revolves around the central notion of agglomerating smaller units occurring in sequence into larger units, based on the contexts in which they occur.

In the system of Powers (1983), each token (initially a word) is considered to form a putative phrase with the tokens before and after it, as well as to the classes of these tokens. The result of running this algorithm on a small artificial corpus was that punctuation symbols emerged first as a class, then articles, then punctuation-article sequences, and finally a class consisting of the open-class words. The algorithm determined on-the-fly whether classes should be merged into a new class or an existing one (similar to the merge phase in GRIDS).

The Morpholearn system of Powers (1991) attempts to induce grammatical categories from a raw corpus of natural language. A core idea in Morpholearn is the notion that when two elements appear contiguously in a sentence, it is often the case that there is a relationship between them, and furthermore that one of them is typically substitutable by a relatively larger class of elements than the other one. Often, the element with relatively lower opportunities for substitution is a closed-class element, and the other an open-class element. The paradigmatic example of this phenomenon is the noun phrase structure *Determiner Noun*, where there are relatively many nouns which can be substituted for, say, “cat” in “the cat” (“dog”, “idea”, “misanthropy”, and many millions of other words), but relatively few words that can substitute for “the” in the determiner slot (“that”, “this”, “a”, and a handful of other words).

Morpholearn passes through the corpus, collecting frequency counts for all sequences of consecutive words up to a certain length, i.e. the *n-grams* of the corpus. Next, it attempts to amalgamate the left-hand-sides of these n-grams into a set and the right-hand-side into a coset. This is done by combining left-hand sides that occur with the same right-hand sides. It is not required that every element in the left-hand set should co-occur with every item in the right-hand set. Instead, it is required that every element in one set should co-occur with SEVEN ± TWO elements from the other set, where SEVEN and TWO are parameters of the system, inspired by the famed “7±2” limit on cognitive capacity first identified by Miller (1956), and which do in fact by default take on their eponymous values, but can be manipulated by the experimenter. SEVEN typically takes on a low value, so as to enforce the constraint that one coset should be relatively smaller than the other. The use of TWO acknowledges that it is unlikely that all valid sentences should be present in any particular sample of a language, and hence some degree of variability is tolerated.

When applied to the text of an English-language novel, Morpholearn is able, amongst other things, to discover the distinction between the relatively closed class of vowels and the relatively open class of consonants, and to derive frequently occurring patterns of English syllabic structure by rewriting members of the co-occurring vowel and consonant classes with nonterminal symbols, then repeating the process of class extraction. In this way, Morpholearn is eventually able to parse the corpus to up to 8 levels of structure. Among the structural units discovered by Morpholearn are units corresponding to function words and functional prefixes emerge as classes that become attached at a higher level to adjacent open-class words.

The Differential Grammar approach of Powers (1997a) attempts to describe all the statistically significant contexts in which a particular word can occur in a corpus. The algorithm does this by considering the immediate context surrounding a word, and expanding it in either direction for as long as the statistical significance of the co-occurrence between focal unit and its context is above a certain value, and each environment is significantly different from the next smallest environment. This process is reminiscent of the way in which ADIOS forms patterns that are units of the language description.

4.3 Other models

4.3.1 Yuret

Yuret (1998) puts forward the notion of *lexical attraction* as a basis for postulating that two words have a specific linguistic relation to one another. In Yuret's model, syntactic relations between words (e.g., subject-verb relations) are taken to be the primitives of linguistic description; hence his model is based on *dependency*. Two words are said to attract each other in Yuret's terms if they tend to co-occur, and this co-occurrence is likely to be due to the existence of an underlying syntactic relation between them. The set of syntactic relations in which a word is embedded in a particular instance becomes the basis for defining its usage context.

Yuret makes use of *mutual information* to define the strength of attraction between two words. His algorithm processes a corpus iteratively, gathering more and more information regarding co-occurrence probabilities, and hence refining its estimate of the mutual information between words. Yuret's model is therefore a fully lexical one; only relationships between specific words are considered, and words are never replaced with tags representing their parts-of-speech.

For each sentence in the corpus, the algorithm attempts to describe the structure of the sentences in terms of the knowledge it already has. If two words have a certain amount of mutual information, a link between them is postulated. The algorithm makes use of two heuristics: cycles of links are not allowed in a sentence (e.g. from A to B, from B to C, and then from C to A); and links are not allowed to cross. In cases where these conflicts arise, the weakest links are discarded until the problem is resolved.

Initially, only links between adjacent words are considered; however, once such a link is established, the two words in question are treated as if they had collapsed into one, and links from the word before and the word after the pair are considered, to *either* of the words in the pair (hence potentially allowing links between non-adjacent words).

In this way, Yuret's algorithm postulates a set of dependency relations that might exist in a sentence. The algorithm performs fairly well in finding these relations, when tested on a corpus with explicitly coded dependency relations (using only relations between content words).

An important point to note about Yuret's model is that all dependency links are *undirected*: when thought of as a link from a head to a dependent element, then it is not specified in this model which element is the head and which the dependent element. This results mainly from Yuret's theoretical approach, under which a dependency structure is treated as a Markov network that expresses the joint probability of a sentence; joint probability is independent of the direction of any links.

4.3.2 PARSER

The PARSER model by Perruchet and Vinter (1998) attempts to describe how a speech stream may be segmented into individual words, making reference only to processes which are known to operate in human memory, such as associative learning, interference and forgetting. Starting from a set of primitive linguistic elements, the algorithm combines frequently-recurring sequences of these primitives by a process of chunking. This requires that two or more elements simultaneously receive the focus of attention, which leads to the putative unit (the chunk) being stored in memory.

Each unit has a certain activation strength, which is increased every time the unit recurs in the input. Units are subject to memory interference, in that a currently perceived unit causes the activation strength of its potential competitors to decline, where competitors are defined as any units that have any phonological material in common with the unit in focus. Only units that have attained a certain level of activation strength are able to cause memory interference. In addition, forgetting is modeled by subjecting all units to a constant rate of decay in their activation strength, at the end of every iteration of the model. Units with activation strengths above the interference threshold are said to belong to the lexicon of the model – they are the items that form the model's current knowledge of the language.

PARSER is able to account, among other things, for the experimental data from the experiment by Saffran et al. (1996) on the segmentation of a speech stream into (nonsense) words.

4.4 Comparison of computational models

The models of automatic language learning reviewed in the previous sections differ along various dimensions. In particular, the following questions are helpful in distinguishing between models:

1. Does a model identify syntagmatic patterns, i.e. commonly occurring sequences of more primitive elements that may be regarded as linguistic units?

2a. Does a model identify paradigmatic classes, i.e. classes of elements that are treated as similar due to certain similarities (notably, occurring in similar contexts in utterances)?

For models that do identify paradigmatic classes, there are a number of subsequent questions.

2b. Are these classes compatible with traditional categories in linguistic theory? To be more precise, we are particularly interested for the current purpose in whether the model is able to account for the "main" parts-of-speech of Noun, Verb and Adjective (NVA).

2c. Does the model acknowledge that a word type may be ambiguous, in that individual tokens of that word type may belong to different classes depending on the context in which they are used?

2d. Does the model acknowledge that the same may be true for the context itself, i.e. that not all elements (words) used in the same context are necessarily of the same class?

2e. Does the model explicitly treat the contexts in which paradigmatic classes occur as linguistic units in their own right? This is particularly important if we are interested in providing an account of the constructions of a language.

2f. Are these contexts lexically-specific, i.e. do they make use of specific individual words, or are they defined in terms of allowable sequences of paradigmatic categories only? This consideration may be important in providing a way for the child to bootstrap into parts-of-speech without treating all words as potential representatives of certain categories, leading to a combinatorial explosion of possibilities to consider.

Table 2 lists the models considered in this chapter in grid form, showing the answers to these questions for each model.

4.4.1 Syntagmatic patterns

Out of the models aimed purely at lexical categorization, reviewed in Section 4.1, only three models identified syntagmatic patterns explicitly. The work by Finch (1993) forms lower-level categories (starting with word classes), then successively considers syntagmatic sequences of these lower-level categories and merges these sequences into higher-level paradigmatic categories. In the Frequent Frames model of Mintz (2003, 2006a, 2006b), a frequent frame may arguably be considered a syntagmatic pattern. And in the model by Cartwright and Brent (1997), all utterances are described in terms of frames for full-utterance structures, defined as valid sequences of word categories.

All the syntax learning models reviewed in Section 4.2, as well as the model of Yuret (1998) and the PARSER model of Perruchet & Vinter (1998), aim to discover syntagmatic patterns in a language. In the case of the syntax models, syntagmatic patterns exist both at the level of specific word sequences that together form units (such as, for instance, some of the most common collocations), and at the level of sequences of paradigmatic categories, or mixtures of paradigmatic categories and words. Being able to identify regular patterns is an important property for a computational model to exhibit, because one of the goals of language learning is to become familiar with the constructions of that language.

4.4.2 Paradigmatic categories

At the same time, the ability to group elements into paradigmatic categories is what allows a language-learning child to generalize from utterances in her experience to utterances that she has not encountered before. In the models reviewed here, these categories are invariably formed by combining elements which occur in the same surrounding context, or similar sets of contexts, across different utterances.

A distinction can be made between paradigmatic categories that operate at various levels of detail in a language system. *Parts-of-speech* are categories of individual words, i.e. they are the traditional parts-of-speech such as nouns, verbs and adjectives. *Syntactic categories* consist of units that play a specific role in syntactic structures, for example noun phrases, verb phrases or prepositional phrases. At the most general level, the term *grammatical category* is used to describe any category of items that perform the same role in a language, and can include lexical and syntactic categories as special cases. All the models reviewed in Section 4.1 are aimed at forming strictly parts-of-speech, while the syntax models of Section 4.2 are potentially able to account for grammatical categories in general, although in practice they have been applied only to finding lexical and syntactic categories.

4.4.3 The three main parts-of-speech

A "correct" division of the words of a language into their individual parts-of-speech is a matter of great controversy, and will perhaps never be resolved. Nevertheless, there does seem to be some consensus that English at least makes a distinction between the three "major" classes of nouns, verbs and adjectives. These categories may have their root in the way in which they are mentally construed, as in Langacker's (1987) analysis, or as the prototypical intersection between linguistic function and meaning, as in Croft's (2005) Radical Construction Grammar. It would be desirable for these three categories to be accounted for by a theory of lexical categorization.

Nearly all of the models specifically aimed at lexical categorization are able to account for these three classes (with the possible exception of Cartwright and Brent). The syntactic models, on the other hand, are typically concerned with postulating any categories that will allow them to describe the syntactic structure of a language successfully, and so usually do not have any mechanisms for reducing the number of categories produced. None of the syntactic models discussed here have to my knowledge been directly evaluated purely on their ability to discover the traditional parts-of-speech from a large natural corpus, so that it is difficult to determine how many categories would actually be produced. Nevertheless, it seems likely that they would typically produce a proliferation of categories, each contingent on a particular context or small set of contexts, rather than a robust division of content words into the three main categories.

Certainly, the main categories can be subdivided into very fine subclasses (past, present and past participial, root and continuous forms of verbs; mass, common or proper nouns; predicative or attributive adjectival forms, etc.). However, the level of distinction between words referring to entities, processes or attributes seems to be a very basic and important one to make, in addition to all the finer distinctions that can be made *simultaneously*. I would argue that a model of language learning which does not account for these basic, high-level word categories is not a complete one.

4.4.4 Word type ambiguity

Not all tokens of a particular word type belong to the same part-of-speech, as shown by Pinker's (1979, 1987) example of "John can *fish*" versus "John eats *fish*". Successful language models need to take this into account. Certainly, all the syntactic models reviewed here are capable of treating word tokens flexibly depending on context.

However, this is often an area of difficulty for pure lexical categorization models, and many of the reviewed models treat all instances of a word type as belonging to the same category (Finch, 1993; Redington et al., 1998; Mintz et al., 2002; Clark, 2001). Other, more sophisticated models do allow tokens of the same word type to belong to different categories depending on the context in which they are used (Mintz, 2003, 2006a, 2006b; Schütze, 1993; Cartwright and Brent, 1997).

4.4.5 Contextual ambiguity

Conversely, the idea that the context in which a word is used determines its part-of-speech can also be taken too far, as shown in the example by Pinker (1984) of "*John eats* fish" versus "*John eats* slowly". Occasionally, the context in which a word is used is ambiguous or uninformative, and a model should rely on other information (including the identity of the word itself) before categorizing the word.

The lexical categorization models that lumped all tokens of a word type into the same category manage to avoid this pitfall, precisely because they do not make use of the particular context of a word token in order to guess at its category. On the other hand, the Frequent Frames model of Mintz (2003, 2006a, 2006b), in aiming to address the ambiguity of words by categorizing on the basis of context, is paradoxically prone to the criticism that it does not consider the ambiguity of contexts (while the reported accuracy figures in Mintz's work are high, they are not 100%). Only the models of Schütze (1993) and of Cartwright and Brent (1997) flexibly combine both context and word in order to arrive at a categorization.

Most of the syntactic models also treat context as potentially ambiguous. In ABL, however, all words and word sequences that occur in the same context are placed in the same category; the same happens in ADIOS for all elements which are "disjunct" from their context according to the pattern distillation metric used. In these two models, context is taken to be definitional for a particular category, and hence not ambiguous.

4.4.6 Contexts as explicit linguistic units

Related to the question of syntagmatic pattern extraction is the question of whether the contexts used to allocate words to particular paradigmatic categories are themselves explicitly identified in the model as linguistic units. The aim here is more specific than merely identifying constructions (which could simply be collocations and specific fixed phrases, without any possibility of generalization); we can now go further and use the occurrence of a word inside a construction to guess at its part-of-speech.

All syntactic models potentially allow contexts of paradigmatic categories to be listed explicitly, whether through an analysis of rewrite rules, trees in a tree grammar, or other means. EMILE in particular creates categories out of both words and their explicit contexts. Most of the lexical categorization models, however, create categories of word types based on similarities in their *profiles* of contextual usage, rather than attempting to categorize individual instances of words in context. Only the Frequent Frames model of Mintz (2003, 2006a, 2006b) and the model by Cartwright and Brent (1997) treat the context of a word as an explicit unit which is stored in linguistic memory and has some degree of autonomous existence.

4.4.7 Lexically-specific contexts

Even more specific than the question of whether contexts are listed explicitly is the question of whether some contexts can be defined in a lexically-specific way. Lexically-specific frames potentially provide a way to “get there from here”, as Tomasello (2003) puts it, because they avoid the combinatorial complexity of considering every word in an utterance as belonging to some part-of-speech.

Again, all of the syntactic models are potentially able to yield lexically-specific frames for a paradigmatic category, although none are constrained to produce *only* lexically-specific frames. Cartwright and Brent’s contexts are based purely on sequences of categories, rather than specific words. Only in the Frequent Frames model are all contexts explicitly stored in memory as items of linguistic knowledge *and* constrained to be entirely lexically-specific. However, as Mintz (2006b) notes, many Frequent Frames are not constructions as such (although they might be used to discover constructions).

4.4.8 The current work

The three models that will be described in this thesis attempt to improve on the work reviewed in this section by providing an affirmative answer to each of the seven questions above. My aim is to show how a language-learning child may discover lexically-specific contextual frames in the speech input that may themselves be constructions of the language, and which may be used for the purpose of assigning words to parts-of-speech. Both word and contexts will be treated as potentially ambiguous. The resulting lexical categorization will be evaluated in terms of its ability to account for the main three categories of nouns, verbs and adjectives.

One broad observation that can be made about the models reviewed here is that there is a large incongruity between the models aimed at discovering parts-of-speech and the models that have syntax as their goal. Out of the lexical categorization approaches, none are particularly concerned with categorizing words on the basis of occurrence in some of the main *constructions* of English. Mintz's Frequent Frames comes closest to this goal and identifies some English constructions, but many other frames in that model are not constructions.

Importantly, except for Frequent Frames and the work by Schütze (1993), none of the models is able to account for the ability of children to correctly interpret a novel word in a familiar frame, as exhibited in e.g. Brown (1957), because they are all "word-centric" models that describe the typical context in which a particular word can occur, as determined by averaging over many exposures to that word; with novel words, there is nothing to average over.

On the other hand, the syntactic models can all be regarded as providing a very detailed description of the patterns present in English, and so arguably one could treat their output as containing information about the constructions in English. However, because these models are not specifically constrained to account for parts-of-speech, they typically postulate a very large number of equivalence classes. The linguistic considerations of Chapter 2 suggest that there is some reason to expect there to be three or four major open classes in most languages, and in any case there is ample linguistic evidence for the existence of nouns, verbs, adjectives and adverbs *in English*. But the syntactic models examined here typically lack a mechanism to constrain them to produce such a small number of word classes.

It will also be an aim of this work to attempt to bridge the gap between these two subfields by proposing a model which can arguably identify some of the main constructions of English, and use these to discover a small number of parts-of-speech that correspond well to the main classes that we would expect on linguistic grounds.

	Syntagmatic patterns	Paradigmatic classes	NVA	Word type ambiguity	Context ambiguity	Explicit contexts	Lexically-specific contexts
Finch	Y	Y	Y	N	Y	N	-
Redington, Chater & Finch	N	Y	Y	N	Y	N	-
Mintz, Newport & Bever	N	Y	Y	N	Y	N	-
Clark	Y	Y	Y	Y	Y	N	-
Frequent Frames	Y	Y	Y	Y	N	Y	Y
Schütze	N	Y	Y	Y	Y	N	-
Cartwright & Brent	Y	Y	N	Y	Y	Y	N
EMILE	Y	Y	N	Y	Y	Y	Y
ABL	Y	Y	N	Y	N	Y	Y
SNPR/ICMAUS	Y	Y	N	Y	Y	Y	Y
GRIDS	Y	Y	N	Y	Y	Y	Y
ADIOS	Y	Y	N	Y	N	Y	Y
MOSAIC	Y	Y	N	Y	Y	Y	Y
Powers (Morpholearn)	Y	Y	N	Y	Y	Y	Y
Yuret	Y	N	-	-	-	-	-
PARSER	Y	N	-	-	-	-	-
Full-utterance frames	Y	Y	Y	Y	Y	Y	Y
Nested and full-utterance frames	Y	Y	Y	Y	Y	Y	Y
Prediction-based frames	Y	Y	Y	Y	Y	Y	Y

Table 2. A tabular comparison of some characteristics of (unsupervised) computational models of language learning. The last three rows refer to models in the current work.

5 The approach based on lexically-specific frames

5.1 Introduction

The next couple of chapters present the empirical research carried out for this thesis. Central to this empirical work is the question of how a language-learning child might discover a number of linguistic contexts, preferably linguistic constructions in their own right, that provide clues to the part-of-speech of a word that appears in that context (using the terminology introduced in Chapter 4, words in context will be called focal words).

In particular, the approach taken here is to find ways of identifying *lexically-specific frames*, i.e. schematic construction structures that occur reliably in the language, and that consist of a number of specific words combined with one or more slots in which variable material may be inserted.

In this work, therefore, I am considering the induction of parts-of-speech based on only the formal linguistic contexts (semi-abstract constructions) in which words occur, while completely neglecting the role played by semantics. This is not due to any conviction that semantics is irrelevant to part-of-speech induction, and that the task can be performed without recourse to semantic information. Rather, the main aims of this thesis are:

(i) to explore procedures by which lexically-specific frames may automatically be discovered

(ii) to show that parts-of-speech may be induced purely on the basis of the co-occurrence of words with these frames, and

(iii) to address the question of the ambiguity of both single words and their linguistic contexts.

A key notion in this thesis is that it is the context in which a word is used that largely determines its part-of-speech; these contexts are explicitly identified as frames. A major preoccupation of this work will therefore be to answer the question of *how* the context of a word should be defined for the purpose of lexical categorization. In Section 5.2, I provide the rationale for the frame approach taken in this thesis, while Section 5.3

describes the approach in outline, and Section 5.4 is concerned with some aspects of the methodology that will be followed.

5.2 Rationale for the frame approach

5.2.1 The formation of construction frames

The basic intuition about language in general, and English in particular, that is adhered to in this work is that many utterances in natural spoken language are instances of constructions that are not completely abstract, but instead consist of a number of *specific* words, in combination with a number of other elements (slot fillers) that are characterized abstractly in terms of the grammatical category to which they belong. So for example, there are constructions covering a number of common questions, requests or assertions: “What X did you use?”, “Don’t X it”, “Here’s the X”, where X indicates a position in the phrase where variable material can be inserted. Other constructions represent phrases or other linguistic constituents: “the X”, “under a X”, “to X it up”, “not very X”.

The specific words in the constructions constitute the more reliable and predictable portion of the construction. Of course, these words may, at a higher level of abstraction, be members of parts-of-speech, potentially allowing several constructions to be subsumed into a single wholly abstract construction. For instance, the very common phrase structure “the X” has a slot that is typically filled by a single noun. Other phrases that accept single nouns include “a X”, “another X”, “this X”, “that X”, etc. Many linguists would have described these phrase structures more abstractly as the generic phrase structure [Determiner][Noun]; indeed it is presumed in strict formulations of Generative Grammar that this is the only valid way to represent these phrases, connecting as it does with a set of innately-given parts-of-speech that includes the categories Determiner and Noun.

By contrast, in usage-based approaches, it is not inconsistent for these phrases to exist at several levels of abstractness, e.g. both as semi-abstract constructions such as “the X” and as instances of the abstract construction [Determiner][Noun].

The guiding assumption I will make is that instances of some *particular* semi-abstract constructions such as “the X” occur so frequently in the input to the child that the constructions are salient and noticeable in their own right, precisely because of the fact that part of the construction is a highly specific word (e.g. “the”) which can be recognized directly, and does not first need to be “translated” into the overarching abstract category Determiner. Having recognized “the”, the child may then also discover, after being exposed to several instances of the construction, that “the” does not appear alone, but is invariably followed by additional material, allowing the semi-abstract pattern “the X” to be discovered (and likewise for “a X”, “that X”, etc.). The specific words therefore serve as “hooks” by means of which the constructions can be discovered.

I also assume that a semi-specific construction discovered in this way is explicitly stored in the child’s memory in such a way that the memory trace can become reactivated whenever an instance of that construction is encountered. It is not suggested that the child necessarily knows what the construction *means*, only that the construction is *recognized* when it reoccurs.

5.2.2 Amalgamation of frames into categories

Recognition of these semi-abstract constructions from their specific words is what allows grammatical (and particularly lexical) categories to be discovered. Whenever the construction reoccurs in the input, it is plausible that a memory trace of the particular word or words used to fill the X slot is also stored (these slot-filler words may already be stored as memory items in their own right, or they may *become* familiar as a result of their repeated occurrence as slot fillers in semi-abstract constructions.). In this sense, the words that occur in the slot of a construction can be regarded as *co-occurrence features* of that construction. At this point, general memory processes related to category and prototype formation come into play (Kruschke, 1992; Rosch, 1983). Categories are formed from items that are similar in some way, e.g. by having certain characteristics in common. If two constructions take many of the same words into their respective slots (i.e. they have co-occurrence features in common), then they can be regarded as being relatively similar to each other, compared to two constructions that have no filler words in common. In this way, it is possible that categories of constructions may be formed, on the basis that members of the category tend to accept roughly the same words as fillers in their variable slots. (Likewise, one would also predict that corresponding categories of words might be formed, according to the sets of constructions in which they serve as slot fillers.)

An alternative model of this process would hold that *all* words should be regarded as potential representatives of parts-of-speech, so that all words are treated abstractly from the outset. Under this view, for example, the determiners “the”, “a”, “this” and “that” could be grouped into one category, and the nouns “doggie”, “fish”, “bridge”, “tower”, etc. into another category simultaneously, on the basis of the occurrence of each (or most) of the 16 possible determiner-noun combinations in input utterances. This is the approach taken by Powers (1991) in his Morpholearn model. Cartwright & Brent (1997) follow a similar approach that considers only utterance structures made up of allowable sequences of parts-of-speech.

In the current work, by contrast, we “break into” lexical categorization not by considering all words to be potential members of abstract categories, but instead by being given a schematic *frame* description of an utterance (or partial utterance) in which some of the words are fixed and other words (the slot-fillers) are allowed to vary. The fixed words do not get assigned to any part-of-speech; only the slot-filler words are categorized. The fixed words can be regarded as the “background” to the “figure” represented by the variable words, reinforced by the fact that the background words are likely to be the semantically diminished *function* words of English, while the variable words are likely to be the semantically rich, informative, *content* words. The frame can be viewed as a kind of substrate in which the more informative, less predictable filler words are embedded.

The model of language processing that underlies this approach is therefore not one in which the task is to learn which combinations of parts-of-speech are legitimate (as in e.g. Cartwright & Brent, 1997). Instead, it is more compatible with usage-based theories, with the frames being regarded as basic constructions (or proto-forms of constructions). As in all other areas of human cognition, anything that is processed frequently becomes automatized (see e.g. Logan, 1988), and hence may be stored as a unit in memory in its own right. Very frequently occurring utterance frames become entrenched as memory units (elements in the "constructicon"), regardless of whether they could also have been described (by linguists) as instances of some more abstract construction.

In this way, processing of an utterance that has been described by means of a frame entails first of all perceiving the specific words of a unit that is likely to be a linguistic construction, whether it is the schematic description of a full utterance, or of a smaller linguistic constituent (e.g. a phrase) that has been embedded in a longer utterance. In terms of the discussion in Section 4.4.6, contexts in the current work are therefore *explicitly* represented in the child's linguistic knowledge.

An important objection to the entirely abstract approach of e.g. Cartwright & Brent (1997) is that there is evidence (reviewed in Chapter 3) that children do not treat all words similarly, but instead seem to be sensitive to the figure-ground separation between function and content words in English utterances. Some of the most telling evidence comes from the work by Shady (1996), who found that 10.5-month-old infants noticed when all the function words in English utterances that they heard were replaced with nonsense words, but were unperturbed when the function words were left intact and all the content words were replaced with nonsense instead. This seems to suggest that the infants were already able to recognize utterances conforming to the "background texture" of English, as subtended by function words; hence utterances that preserved this texture were acceptable to the infants and were preferred during listening, while utterances that violated it rather grossly by containing no familiar English function words were not preferred, presumably due to their unfamiliarity.

Additional evidence comes from the work by Gerken & McIntosh (1993), who demonstrated that language-learning children's understanding of English sentences was impaired when function words were omitted, or replaced with other material, and by Gerken, Landau & Remez (1990), who showed that English-learning children's tendency to omit function words is not due to a lack of processing of these elements, but that function words are more likely to be omitted than other phonologically-similar nonsense words, indicating that they are recognized by the child and then selectively omitted.

The computational models that will be described in this thesis do not start from an externally-provided list of the function words of English, but rather identify the structure-building, specific words of constructional frames by other means. Nevertheless, it will be shown that the specific words of the frames that are identified in these models are, for the most part, function words.

Additional points should be made about the possible process of formation of these frames. While the focus of the current work is on presenting and evaluating computational techniques which will be shown to produce a large number of frames which are useful for part-of-speech induction, it is of course necessary for these techniques to connect with the abilities of a language-learning child. Whether these techniques are actually an accurate description of the processes that occur in learning a language is something that remains to be determined through empirical experimentation. Nevertheless, the emphasis throughout the current work is on psychological plausibility, and on the use of techniques that are compatible with what is known about language learning (e.g. the studies reviewed in Chapter 3). As each technique is presented and described in each of the following chapters, remarks will be made to attempt to situate the particular computational technique in a psychological context, appealing to certain basic processes such as chunking and associative learning.

Essentially, the frame-based approach advocated here is congruent with proposals in the learning literature that elements that are the focus of simultaneous attention become associated with each other in memory (see e.g. Logan, 1988; Logan & Etherton, 1994; Pacton & Perruchet, 2008; Treisman & Gelade, 1980). So, for instance, Perruchet & Vinter (1998) have suggested that some units in language, such as words, are formed by *chunking* smaller elements together to form larger units. In the case of frames, the lexically-specific items (fixed words) of a frame may become associated with each other by simultaneously receiving selective attention. This configuration of elements then becomes an element in its own right.

Many of the frames identified in child speech by researchers such as Lieven et al. (1997) and used experimentally by Santelmann & Jusczyk (1998) contain sequences of non-adjacent elements, with slots between them (e.g. "you _ it"). Chunking models traditionally operate only on adjacent elements, so that discontinuous dependencies may be thought to be problematic (see e.g. Perruchet & Pacton, 2006). Recently, however, Pacton & Perruchet (2008) have shown that adults exposed to material containing both adjacent and non-adjacent dependencies can learn the specific relations that they have been required to attend to in the course of performing an experimental task. The authors suggest that the privileged status given to the learning of relationships between contiguous elements may merely stem from the fact that contiguous elements are often the simultaneous targets of attention.

One way to reconcile the frames that are of interest here with an associative approach is to consider that even a "discontinuous" frame such as "you _ it" is made up of three adjacent elements: the first and third elements are "you" and "it", and the second element is phonologically entirely *unspecified* (or partially unspecified, if we incorporate possible morphological behaviour). Therefore, the "place" represented by the slot is an integral part of the frame: "you it" is not an instance of "you _ it". Even if we cannot specify the middle element explicitly, it still has phonological substance, and in the case of a variable frame slot, this should be specified.

5.3 Basic approach

The experiments reported in Chapters 6 to 9 follow a similar basic plan, outlined in this section. The focus is on automatic methods of finding lexically-specific frames, i.e. commonly-occurring contexts in which focal words may be embedded. Three distinct such methods will be presented.

Chapter 6 presents a method for finding full-utterance frames. In Chapter 8, this technique is extended to include hierarchically nested part-utterance frames. Both of these methods rely heavily on the establishment of a basic dichotomy between frequent and infrequent words, the former (in English) tending to be function words and the latter content words. Chapter 9 is concerned with an attempt to discover frames without making use of this dichotomy, focusing instead on the strength and direction of association between pairs of words; the resulting frames are termed *prediction-based frames*.

All three of these chapters provide psychologically plausible methods that an English-learning child might use to discover frame structures present in the language she hears around her. Many of these frames are reliable cues to the part-of-speech of the focal word. Given that there are only a few parts-of-speech, many frames are associated with the same part-of-speech, and so it would be useful to amalgamate these frames together, as discussed in Section 5.2.2. An obvious way to do this is to form *clusters* of frames by grouping together these frames that accept broadly the same words into their frame slots. The computational methods of *cluster analysis* therefore play a prominent role in this research. For each of the three different frame discovery methods, the frames that are found are subjected to clustering. The clusters that form in this way can be seen as large lexical "paradigms" corresponding closely to the traditional parts-of-speech such as nouns, verbs and adjectives. Words which can be found in the context of several frames in a cluster may plausibly occur in all other frames in that cluster.

Assigning a category to a focal word on the basis of the context that it appears in is an appropriate way to overcome the problems with some of the earliest word clustering research (e.g. Finch, 1993), where a word was assigned to one and only one category, even though many words can actually function as members of more than one category. However, when examining some of the frames produced by the computational procedures outlined here, it will become apparent that several of these frames are not good indicators of part-of-speech at all. Effectively, not only the focal word, but also the context itself may be uninformative with regards to part-of-speech, which is the point made by Pinker (1984) in his "John eats X" example. (That construction is effectively two constructions: "John eats [Noun]" and "John eats [Adverb]".)

There are two important points to consider in dealing with frame ambiguity. Firstly, it would be useful to be able to distinguish between frames that are and are not informative of part-of-speech, or at least to enumerate the parts-of-speech that can occur in each category. Secondly, by focusing only on the context a word occurs in, we are ignoring a key insight from earlier clustering studies such as those of Finch (1993), that many words are not ambiguous *in their common usage*. Hence, we are sacrificing a large amount of valuable information about part-of-speech by not taking note of *which word* is actually in focus. It would seem to be useful to try to combine information from both the frame and the word in order to arrive at a more reliable guess about the part-of-speech. This is part of a broader conceptualization of language learning, in which the child makes use of any and all sources of regular information available to her in order to learn language (and indeed, any other cognitive skills). These ambiguity-related issues are explored in Chapter 7. I present three ways in which frame and word information may be combined in order to achieve lexical categorization, in the context of the full-utterance frames introduced in Chapter 6, and evaluate the results of implementing these three methods. These methods are then used in Chapters 8 and 9 to improve on the categorization results obtained with the nested and prediction-based frames.

The entire treatment of the frames produced by a particular method (discovering the frames, collecting word-frame co-occurrence data, frame clustering, and then also word and frame co-clustering to accommodate ambiguity) is therefore demonstrated first for full-utterance frames, in Chapter 6 and Chapter 7. Once this machinery has been developed and presented, it is applied again in turn for the nested frames (Chapter 8) and the prediction-based frames (Chapter 9).

The aim is emphatically not to look for simple local contexts that serve as cues to part-of-speech, as was done in e.g. Redington et al. (1998), or Mintz’s Frequent Frames approach (Mintz, 2003, 2006a, 2006b). In that work, it seems that the child is overtly engaged in the task of assigning a tag to a word, and exploits any available local contextual cues in order to perform this assignment. By contrast, in the current work the overt task is to identify autonomous units in a language, and the child will be mainly occupied in learning to perceive these units directly. Somewhat incidentally, then, some of these units will have the compositional structure of a lexically-specific frame, i.e. with some words specified exactly and others left as abstract slots. And by an additional process of category formation, slots which are similar will become grouped together in the child's mind, whether that similarity is distributional, i.e. based on the same groups of words occupying the slots, or semantic, i.e. based on similarities in the meanings of those slot-filling words. On this view, parts-of-speech are discovered naturally during the course of learning about syntactic constructions. The alternative view would be that parts-of-speech are primary, and constructions are discovered by concatenating parts-of-speech into larger sequences. This begs the question of how to deal with words which are ambiguous, such that the ambiguity can only be resolved by considering the larger context (i.e. the construction) in which the word occurs.

5.3.1 Syntagmatic contextual patterns

When describing the process of discovery of lexically-specific frames, it seems that one of the most important questions to be addressed is how these frames are separated out from the speech stream, i.e. what the cues to their boundaries are. One way in which this could be achieved is to make use of the most salient unit boundaries in connected speech: in most cases in natural speech, there are pauses between utterances in a conversational turn by one partner. Hence utterance boundaries are perceivable by a child, and so one way to delimit potential units of a language is to consider each single utterance at a time as a unit.

Another possibility occurs when some lexically-specific frames already exist: the slot-fillers that go into those slots can be regarded as units in their own right, and in cases where these slot-fillers can themselves be given a partially lexically-specific structure, this would justify the slot-fillers also being treated as lexically-specific frames. So for both of these possibilities, units are recognized against a surrounding *context*: in the former case it is the context of occurring inside utterance boundaries as indicated by silence, and in the latter case it is the context of occurring embedded inside a previously-learned lexically-specific frame.

A third, different possibility is not to assume that there are any predefined boundaries such as those between utterances. Instead, words merely become associated with each other if they co-occur often, giving rise to learned configurations of words; the positions in an utterance where a particular learned configuration starts and ends are then quite simply the segment boundaries. This proposal therefore hinges on how associative links are formed.

Each of these three segmentation strategies is explored in depth in its own chapter in this thesis, respectively in Chapters 6, 8 and 9.

5.3.2 Paradigmatic parts-of-speech

The second main theoretical issue has to do with *how* parts-of-speech are induced from knowledge of lexically-specific frames. One way (investigated by researchers such as Finch, 1993; Mintz, 2003, 2006a, 2006b) is to group words together into a category if they tend to occur in the same sets of frames. However, this approach presumes that words are unicategorical; in fact, as we have seen (e.g. Nelson, 1995), it is common even in the input to children for the same word type to belong to different parts-of-speech according to context. A better approach (taken by e.g. Mintz, 2003), is to group the frames themselves into categories, which are interpreted as frame categories for the particular parts-of-speech: any word that occurs inside, say, the "noun frame" category is presumed to be a noun. In other words, the context imposes an interpretation on the word. This approach is taken in Chapter 6.

However, even under this approach, it will become apparent that some frames are simply not informative about the part-of-speech of the words that occur inside their slots. More generally, it seems that combining information about both the frame and the word which occurs in it would provide the most complete and therefore the most accurate information for lexical categorization. Three techniques to explore this possibility are presented in Chapter 7, and are applied to each of the three frame discovery procedures.

5.4 Methodological issues and assumptions

5.4.1 Corpus and preprocessing

All experiments reported here were carried out on the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001) of child and child-directed speech, taken from the database of the CHILDES project (MacWhinney, 2000). The Manchester corpus was collected in the course of a longitudinal study of 12 children aged between 2 and 3 years, all first-borns from predominantly middle-class families in the Manchester and Nottingham areas. Children were recorded in 34 pairs of half-hour sessions (i.e. 34 hours in total for each child, although some sessions are missing). During the first half-hour children played with their own toys, and during the second half-hour they played with toys provided by the experimenters.

There is a wealth of information to be found in CHILDES corpora, in the form of annotation in various "tiers", added according to the CHAT markup specification (MacWhinney, 2000), potentially including speaker identity, phonetic information, and part-of-speech information. For the current purpose, only the orthographic words used are required, and only child-directed speech is required. In fact, for simplicity, these experiments made use of only the mother's speech, which made up the vast majority of child-directed speech.

In order to turn the sentences spoken by mothers into simple, uniformly-formatted sentences that could be batch-processed by a computer program (referred to hereafter as the "cleaned-up corpus"), it is necessary to perform a certain amount of preprocessing. CHAT makes use of a variety of non-alphabetic characters for corpus annotation. In addition, some sentences are not necessarily complete, and it needs to be decided which ones represent usable data. Non-alphabetic characters need to be removed, and unsuitable utterances discarded, in order to produce a "clean" corpus on which experiments can be carried out.

Specific details of how the corpus was preprocessed can be found in Appendix 1. Two points bear repeating here. Firstly, all punctuation is removed, except for commas and question marks (utterances are presumed to end with a full stop by default, and hence this is omitted). Secondly, the name of each child is replaced with the token *childname*. This is done because a child's own name may well have a significant role in the spoken utterances that the child hears. It is likely to occur very frequently, and to occupy a significant position in sentence structures (in many cases, the child's name is used almost as a synonym for the second-person pronoun *you*). In order to allow the computational procedures to pick up this regularity when the data from all twelve mothers is pooled, it is necessary to use a standard token for the child's name. Summary counts displayed in Table 3 give an indication of the size of the corpus after preprocessing.

Number of files	800
Number of lines (utterances)	334806
Number of words	1321591

Table 3. Summary counts of the cleaned-up version of the Manchester corpus as used in this thesis.

5.4.2 Segmentation into orthographic words

As can be inferred from the previous section, the data on which all the algorithms in this thesis work is assumed to be already segmented into orthographic (i.e. dictionary) words. It is a debatable point whether this is psychologically plausible; nevertheless, the models outlined here assume that the child is in some way able to segment the speech signal into discrete units, and that these units are likely to correspond to orthographic words. The task of a frame-finding algorithm is then purely to discover common patterns subtended by these elements.

There is some evidence of recognition of some common nouns as early as 6 months of age (Tincoff & Jusczyk, 1999; see also Jusczyk & Aslin, 1995), while infants can recognize their own name at as early as 4 and a half months of age (Mandel, Jusczyk & Pisoni, 1995). The work on function words cited in Section 3.6.1 shows that these elements are also familiar to the child from an early age and can be segmented out of the speech stream (Höhle & Weissenborn, 2003; Shi, 2007; Shi, Cutler, Werker & Cruickshank, 2006; Shi, Marquis & Gauthier, 2006; Shi, Werker and Cutler, 2006). Certainly, by the age of the children involved in the Manchester corpus, children are able to identify and understand several words. It is not completely necessary for the success of the computational models described in this thesis that the segmentation should be *complete*; however, the models do rely heavily on the assumption that at least the most *frequent* words should be *familiar* (i.e. available in *recognition memory*). This is certainly plausible; frequency in the input is strongly related to familiarity for the child, as shown in the references cited above. And to the extent that some words used by the algorithms that are discussed in this work are not familiar to the child, the performance of the algorithms may be expected to degrade gracefully rather than catastrophically.

A more pertinent concern is whether dictionary words comprise the correct level of granularity for the discovery of frames. Certainly, much of English orthography is arbitrary and conventional; there is no obvious reason why, for instance, *into* should be spelt as one word and *out of* as two, and many English word sequences are translatable into single words in other languages and vice versa. Furthermore, morphological inflection is a very valuable clue to part-of-speech, so that one might imagine that a segmentation into morphemes rather than words might be more informative for part-of-speech induction.

The position I take here is that it is highly likely that children at the age of about 18 months are able to segment continuous speech into at least *some* constituent elements, and that these are likely to correspond reasonably closely to words. Hence, starting from a corpus of segmented words is a reasonable approximation to the early input provided to children.

5.5 Evaluation of clustering results

5.5.1 Quantitative evaluation against a gold standard

In order to evaluate the outcome of the categorizations provided by the experiments in this thesis, it is necessary to compare them against a “gold standard” describing the “correct” categorization of each focal word. The Manchester corpus comes provided with part-of-speech markup information, so that this categorization can be used as the gold standard categorization.

In order to make this task somewhat simpler, it was decided to restrict the analysis and tagging to focal words that belong to the categories Noun, Verb and Adjective (NVA) only. Decisions about other word classes (in particular function word classes) are likely to be distinctly more problematic and less clear-cut than for these three classes. Words deemed to belong to other categories were, of course, included in the data during clustering (the child cannot be presumed to discard them when they occur in the input), but they are simply ignored for the purpose of evaluation.

Nouns are tagged in the Manchester with "n", adjectives with "adj", and verbs with "v". The following additional categories were also mapped to one of "n", "adj" or "v":

- "n-prop": proper nouns were mapped to "n"
- "n-pt": "plural-seeming" nouns such as "pants" were mapped to "n"
- "n-let": letters of the alphabet were also mapped onto "n"
- "n-v": nouns apparently derived from verbs were mapped to "n"
- "n-gerund": gerunds were mapped to "n"
- "adj-v": adjectives apparently derived from verbs became "adj"
- "adj-n": adjectives apparently derived from nouns became "adj".
- "aux": auxiliaries were treated as verbs
- "part": participial forms of verbs (past and present) were treated as verbs
- "v~neg" and "aux~neg": compounds of verbs plus negation ("couldn't") were treated as verbs

All other combination words, indicated by several categories concatenated with tildes (e.g. "Mommy's" in "Mommy's had enough" would be "n~v"), were not interpretable as belonging to just one part-of-speech, and were left out of the analysis.

5.5.2 Quantitative evaluation measures

Evaluating the results from an experiment in unsupervised clustering against a "gold standard" categorization requires some care.

One issue to be addressed is the question of which particular measure to use to express the degree of fit between a gold standard and the empirically derived clustering result. Traditionally, researchers wishing to give quantitative expression to the match between reality and the predictions from a model have made use of one of a variety of association measures, calculated on a two-by-two *correspondence table*, as represented in Table 4, which is normally derived for one gold-standard category at a time. The columns represent the number of items that were (left column) or were not (right column) in the particular gold-standard category of interest, while the rows represent the number of items that a model did (top row) or did not (bottom row) assign to the category. When we label the cells *a*, *b*, *c* and *d* as in the table, we can see that *a* and *d* represent correct categorization on the part of the model, and *b* and *c* represent errors.

	Belonging to category (Gold Standard)	Not belonging to category (Gold Standard)	Total
Belonging to category (Model)	*a*	*b*	*a* + *b*
Not belonging to category (Model)	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

Table 4. A correspondence table, showing allocation of items to a particular category according to a gold-standard (columns) against allocation of items to that category by a model (rows).

5.5.2.1 Accuracy, Completeness and F

There are a great number of different measures that can be calculated from a correspondence table (for reviews, see Hayek, 1994; Pfitzner, Leibbrandt & Powers, 2009). A popular choice is to report the measures known in psychology as *accuracy* and *completeness* (respectively named *precision* and *recall* in computer-science-oriented fields such as data mining, information retrieval and computational linguistics). Accuracy is defined as

$$accuracy = \frac{a}{a+b} \tag{1}$$

and expresses the proportion of items that were assigned to the category in question by the model (*a* + *b*) and that were in fact in that category according to the gold standard (*a*). Completeness is defined as

$$completeness = \frac{a}{a+c} \tag{2}$$

and expresses the proportion of items that belonged to the category in question according to the gold standard (*a* + *c*) and that were in fact assigned to that category by the model (*a*). Intuitively, then, accuracy is the proportion of our shots at target that were on target, and completeness is the proportion of targets that we managed to hit.

There is typically a trade-off between accuracy and completeness; either measure can be artificially inflated at the expense of the other. For this reason, it is customary in the computational linguistics literature to attempt to merge accuracy and completeness into a single measure, the most widely-used being *F*, the harmonic mean of accuracy and completeness. F is given by

$$F = \frac{2a}{2a+b+c}.$$

A second issue relates to the difficulty in establishing a correspondence between accepted linguistic categories and the clusters produced by unsupervised clustering. Say that we have a gold standard categorization of a set of focal words in context, which assigns each of the words to a particular part-of-speech (however this gold standard may have been arrived at). The clustering algorithms used in the experiments that are reported here assign instances of focal words in context to one of a fixed number of clusters. However, when we try to determine whether the clustering algorithm “got it right”, we have no way of knowing which cluster is “meant to” correspond to verbs, which one to adjectives, etc., because an unsupervised algorithm such as clustering has no access to these labels. All we have is a partition of the set of words into cluster 1, cluster 2, etc.

As a way of addressing this problem, it is fast becoming standard practice, in the case of categorization experiments in psycholinguistics (e.g. Cartwright & Brent, 1997; Mintz, Newport & Bever, 2002), to derive accuracy and completeness scores from a 2-by-2 table obtained by *pair counting*. Pair counting evaluates the implicit correctness of a clustering by considering *pairs* of focal words that are assigned to the *same cluster/category*, and pairs of focal words assigned to different categories. These pairs are identified in both the gold standard model and the clustering model.

The number of pairs of words that are assigned to the same category in the gold-standard, and *also* in the clustering model, is entered into *a* in the table. The number of pairs of words that are assigned to the same cluster in the model, but to different categories in the gold-standard goes into cell *b*, and the number of pairs of words that belong to the same category according to the gold-standard but are allocated to different clusters in the model, goes into cell *c*. Lastly, *d* contains the number of potential word-pairs that do not belong to the same category in the gold standard, and are also not allocated to the same cluster by the model.

To the extent that a cluster corresponds to, say, the category of nouns, it should contain (i) the nouns, (ii) all the nouns, and (iii) nothing but the nouns. The model will therefore (i) receive "plus points" in cell *a* for all the pairs made up of the nouns that the cluster does contain. To the extent that some nouns are not in the cluster, the model will (ii) "miss out" on obtaining credit it would have received had it paired up the missing nouns with the nouns that are present, and will score "minus points" in *c* (causing completeness to decrease). To the extent that the cluster contains words from another category, say verbs, the model will also (iii) score "minus points" in *b* for every word pair made up from a noun paired with a verb (thereby causing accuracy to decline).

It is not necessary to report these numbers per category or per cluster; we can add the numbers of pairs in each cell together for the entire set of focal words to obtain one accuracy and one completeness score.

There are a number of problems with the accuracy and completeness measures; notably, they can only be used on 2×2 contingency tables, rather than tables of higher dimensionality; both measures can be artificially inflated (completeness by putting too many word pairs together and therefore producing very few clusters, and accuracy by proliferating the number of clusters so as to avoid the penalty of incorrectly putting items together); also, there is a tradeoff between the two measures, and they should always be interpreted together. An additional shortcoming of these two measures, as well as F, is that values obtained on different experiments and by different researchers cannot be directly compared against each other, but should first be interpreted against a random baseline which differs for every experiment. Partially in order to compensate for the shortcomings of these very popular measures, Powers (2003) developed the Bookmaker measure, described in the following section.

5.5.2.2 Bookmaker

Powers describes the *Bookmaker* measure as quantifying the extent to which a prediction is an informed one, such that a person placing a bet on the predicted outcome against fair odds would make a profit.

When considering predictions of outcomes, a central concept is that of the *conditional probability* of event B given event A, namely the probability, if A has already occurred, that B will follow. The conditional probability therefore expresses to what extent an organism can predict that an event will occur, given information about what has occurred before.

These notions can be explained in terms of the contingency table displayed in Table 4. If the rows of Table 4 are taken to be the occurrence or non-occurrence of the earlier event A, and the columns to be the occurrence or non-occurrence of the later event B, then the formula for conditional probability is

$$P(B \mid A) = \frac{a}{a+b}.$$

Likewise, the conditional probability of outcome A given outcome B is expressed as

$$P(A \mid B) = \frac{a}{a+c}.$$

Conditional probability is a very commonly used measure in evaluation; note that *P(B|A)* is identical to accuracy as defined earlier, and *P(A|B)* is equal to completeness. However, a more reliable indicator of the extent to which A predicts B is given by the function *delta-P* (ΔP), which discounts the probability that B will follow A by the probability that B will occur even when A does not occur. This measure gives the true value of A as a predictor of B; if B is more likely to occur after A than after the absence of A, then A is a good predictor of B. In terms of the cells of Table 4, ΔP is given by

$$\Delta P = \frac{a}{a+b} - \frac{c}{c+d}.$$

Many learning theorists (e.g. Shanks, 1995) have accorded centre stage in learning theory to ΔP, regarding it as the normative measure of contingency in learning. Powers (2008) refers to ΔP as *markedness*. He also defines an analogous function of *informedness*, given by

$$I = \frac{a}{a+c} - \frac{b}{b+d}.$$

In a betting context, Informedness expresses the proportion of time that an outcome has occurred, and we have placed a bet on its occurring, discounted by the proportion of time that the outcome has not occurred, but we still had a bet riding on its occurrence. If we are perfectly “informed” about which details are relevant to the outcome, we would be able to place a bet only in cases where the outcome does end up occurring, in which case we would never lose money, and Informedness would take on a value of 1. On the other hand, if we had no relevant information about the contingencies of the situation, we would bet incorrectly as often as correctly, and end up losing as much money as we won (the two terms in the function would cancel out), yielding an Informedness value of zero. In other words, unlike accuracy, completeness and F which have a varying random baseline, Bookmaker always has a random baseline of zero, and so results from different experiments, including experiments carried out by different experimenters using different approaches, can be compared against each other directly.

Informedness is the foundation of the Bookmaker measure. When we have a 2-by-2 contingency table, Bookmaker is exactly informedness as defined. However, Bookmaker is also defined for square contingency tables of arbitrary dimensions (another advantage over completeness and accuracy). A table of dimensions M×M, with the rows and columns in correspondence, can be turned into a 2×2 table for each row R by collapsing together all columns except R and all rows except R. Informedness can then be calculated on each of these tables. Bookmaker is then calculated as a weighted sum of each of these informedness values for each 2×2 table, where the weight is the ratio of the number of cases in which the predictor R occurs over the number of all cases.

Finally, in the current work I generalize Bookmaker to work with unsupervised clustering results and an arbitrary $M \times N$ table of empirical categories on the rows and gold standard categories on the columns, as follows. Let d be the contingency table describing the co-occurrence of empirical and gold standard categories, such that $d_{ij} = k$ means that k elements from gold standard category j were allocated to empirical category i. Let the marginal totals of each column in d, i.e. the numbers of elements allocated to each of the gold standard categories, be denoted by $g_1, \ldots, g_M$, and let the marginal totals of each row, i.e. the numbers allocated to each empirical category, be denoted by $e_1, \ldots, e_N$.

Furthermore, let z: $\{1, 2, \ldots, M\} \rightarrow \{1, 2, \ldots, N\}$ be the mapping from gold standard categories to empirical categories that maps each gold standard category onto the empirical category which contains more elements from that gold standard category than any other empirical category, i.e. z(a) = a' ↔ there does not exist any b in $\{1, 2, \ldots, N\}$ such that $d_{ba} > d_{a'a}$[3]. Then we can construct, for each gold standard category a, a 2×2 contingency table that distinguishes between elements belonging or not belonging to gold standard category a on the columns, and belonging or not belonging to empirical category a' on the rows (see Table 5, where T = the total sum of all cells in d).

[3] If there is more than one such a category, an average may be taken over all of these categories.

	a	*~a*	*sum*
a′	**$d_{a'a}$**	**$e_{a'} - d_{a'a}$**	$e_{a'}$
~ a′	$g_a - d_{a'a}$	$t - g_a - e_{a'} + d_{a'a}$	$t - e_{a'}$
sum	**g_a**	**$t - g_a$**	t

Table 5. A two-by-two contingency table for an arbitrary gold standard category *a*.

Now, Bookmaker can be calculated by obtaining an Informedness value for each 2×2 table for each gold standard category (from the cells in bold in Table 5), then weighting each of these Informedness scores by the weight of that particular empirical category. In other words, Bookmaker is given by

$$BM = \sum_{a=1}^{M} \frac{e_{a'}}{t} \left(\frac{d_{a'a}}{g_a} - \frac{e_{a'} - d_{a'a}}{t - g_a} \right), \text{ where } a' = z(a).$$

When applied to a table of empirical versus gold standard categories, Bookmaker expresses the degree to which membership of a word-frame instance in each individual empirical category predicts membership of the corresponding gold standard category. If a child conceives of a word in context as belonging to one of the empirical categories that she has developed, to what extent will her treating the word in this way receive a "payoff" as a result of being congruent with the true category of the word?

As discussed above, Bookmaker requires a mapping between empirical and gold standard categories. In the experiments presented in this thesis, the empirical categories are so similar to the gold standard categories that this mapping can in fact be effected. In these cases, it would be possible to express the correctness of categorization using the Bookmaker measure directly on the contingency table, in addition to accuracy, completeness and F using a pair counting approach as suggested previously (note that accuracy and completeness cannot be directly applied to the contingency table, as these measures are defined for 2×2 tables only).

The approach that I will take in all experiments is to report accuracy, completeness and F using a pair counting approach, and Bookmaker on the contingency table directly. The use of accuracy, completeness and F is mainly for continuity with the existing literature. Bookmaker as discussed above provides a more statistically sound way to measure the correctness of a categorization.

An additional note should be made about the calculation of the baseline values for completeness, accuracy and F. Many researchers (e.g. Redington et al., 1998) calculate these by randomly allocating individual words to one of the available categories, then calculating accuracy and completeness on these random allocations. However, this is unnecessary, as the baseline values of accuracy and completeness (and hence of F) can be determined *analytically* from the marginal totals in the contingency table. If, say, 30% of all possible word pairs in the data set are pairs that are placed into the same category by the gold standard, then, if there is no meaningful pattern behind the way that the empirical categorization works, so that it merely allocates words randomly, it can be expected that 30% of the pairs it puts together will be correct ones, so that the baseline accuracy is 30%. (Of course, 30% of the pairs it does *not* put together will also be correct.) In other words, the baseline for accuracy is given by (a+c) / ($a + b + c + d$). Likewise, if the empirical categorization is such that 60% of all possible word pairs that can be put together in a cluster are in fact put together, then 60% of all the correct pairs according to the gold standard will be covered by the empirical categorization (as will 60% of all the incorrect pairs). Hence, the baseline for completeness is given by (a+b) / ($a + b + c + d$).

5.5.2.3 Statistical significance of quantitative results

In addition to reporting the obtained values for the measures accuracy, completeness, F and Bookmaker, it will also be useful to state the level of significance of these values against a null hypothesis that the categories assigned to focal words by the techniques presented in this thesis have no relationship to the actual categories that these words belong to (as given by the gold standard). In addition, as a number of different clustering algorithms will be examined, it will also be useful to be able to compare the obtained measure values for different algorithms and express the significance of the difference between them.

While accuracy, completeness and F are standard measures of correctness in many fields in computational linguistics and psycholinguistics, very few studies ever report significance on these measures. For instance, many studies in computational linguistics present the performance of a new algorithm on a *standard* problem set, and in such cases it is considered sufficient merely to show that the new algorithm produces higher measure values than the previously best-performing algorithm on that problem set. Of course, such an improvement in measure values does not show that the improvement is statistically significant.

In general, gauging significance for a measure such as F is difficult, as no specific parametric assumptions can be made about the distribution of these values; in particular, normality may not hold, so that a standard t test of significance making use of the standard deviation of the sample may not be appropriate.

The approach that will be taken in this thesis is to make use of *randomization methods* (Edgington, 1995; Manly, 1997) to assess significance. In essence, these methods assess the significance of some statistic obtained from a data set by producing a distribution of that statistic, generated by randomly reordering the items in the data set. The original value obtained for the statistic is compared against the distribution in order to determine whether it is a typical value under that distribution, in accordance with the null hypothesis, or an atypical and significant value. Significance is indicated by the proportion of values in the distribution that are at least as extreme as the obtained statistic.

Two randomization methods will be used in the empirical chapters of this thesis, one to report significance of the deviation of an obtained value from a random sample generated according to the null hypothesis, and one to compare the significance of a difference between measure values for two different algorithms. As F can be regarded as a measure that summarizes accuracy and completeness, only F and Bookmaker values will be subjected to the significance tests.

The main step in the randomization test is to generate a value for F (alternatively, Bookmaker) that is selected randomly from an appropriate population of F values. This population does not consist of all possible F values; instead, it is appropriate to consider only F values that would be obtained for a random categorization of the focal words that maintains the *same marginal total values* as the categorization that produced the F value that is to be tested, i.e. the new categorization respects the inherent *bias* of the algorithm to favour some categories over others.

In other words, if a particular algorithm allocates focal words to three categories in the proportions 1: 2: 3, then for the randomization test we generate a random categorization (a reordering) of the focal words that also allocates words to these categories in the proportions 1: 2: 3 (individual words will of course be allocated to different categories than the ones they received under the test categorization). This can be regarded as a randomly drawn categorization from the distribution of categorizations that display the same 1: 2: 3 bias as the original categorization. The value of F is then calculated on this categorization, and added to a set of such randomly-generated F values. As this set becomes sufficiently large, it begins to approximate the complete distribution (in the limit, when *all* possible categorizations have been generated, it is of course *identical* with the complete distribution).

Significance can then be calculated directly from this sample, by simply determining the proportion of F values in the sample that are greater than or equal to the empirically-obtained F value for the algorithm in question; i.e. if 15 values in a randomly-generated sample of 1000 F values are as extreme as the empirical F value, yielding a percentage of 15/1000 = 0.015, then the obtained F value is significantly different from the randomly generated sample at an estimated significance level of $p = 0.015$.

Determining the significance of a difference in F values for two algorithms can be accomplished with a straightforward extension of this method. Two sample distributions are created for each of the algorithms as before. One element is then drawn at random from the sample for one algorithm, and one element from the sample for the other algorithm, and their difference is calculated. This process is repeated in order to produce a sample of difference values, and the significance of the original observed difference in values can be determined as before, as the proportion of difference values in the sample that are at least as extreme as the observed difference.

When reporting the results of all empirical experiments in this thesis, I will provide significance levels for deviations from random baselines and for differences between algorithms for F and Bookmaker, according to the methods just described. In all cases, the generated samples of F and Bookmaker values contain 1000 items each.

In all cases, the proportion of values as extreme as the obtained value is reported. However, because we are dealing with samples rather than the complete distribution, it is appropriate to be more conservative in drawing a conclusion about the actual level of significance demonstrated. I will report the conventional significance levels p = 0.05 and p = 0.01. Manly (1997, pp. 82-83) tabulates several 99% confidence limits for significance levels estimated with randomization tests. For a sample of 1000 items, if the "real" significance of the obtained data against the full distribution is 0.05, then the estimated significance level from the sample will fall between 0.032 and 0.068, 99% of the time. For a "real" significance level of 0.01, the estimated significance will fall between 0.002 and 0.018.

In order to be conservative, I will make use of these lower limits and treat an estimated significance level equal to or less than 0.032 (32 or fewer items out of 1000) as significant at the 0.05 level, and a significance level equal to or less than 0.002 (2 or fewer items out of 1000) as significant at the 0.01 level.

6 Full-utterance frames

6.1 Introduction

In this section, I present a technique to automatically identify lexically-specific frames that can stand on their own as full utterances. While this technique is an extremely simple procedure to discover frames, it will turn out to be remarkably successful in categorizing focal words. A crucial element in this procedure is the establishment of a fundamental dichotomy between the most frequently-occurring words and all other, less-frequent words.

The structure of this chapter is as follows: first, the procedure is described, and a possible interpretation in psychological terms is discussed. Subsequently, I present results from an implemented simulation on the Manchester corpus, and evaluate these results against the "gold standard" reference part-of-speech assignment provided with the corpus.

6.2 Automatic discovery procedure

6.2.1 Outline

The essential notion guiding the computational procedure presented in this chapter is that the words that are specific in a lexically-specific construction, and that provide the structure to the construction, are typically taken from a very small set of word types. These words are very often function words, such as "the", "of", "it", etc. Conversely, the content words of English are relatively less frequent. It is these words that we would like to allocate to familiar parts-of-speech such as noun, verb, adverb, adjective, etc.

One characteristic that these words have is that they are very *frequently-occurring* words. The frame-discovery procedure presented in this chapter makes use of a very simple heuristic in order to exploit these frequent words and discover lexically-specific frames. It compiles a list of the most frequently-occurring words in the corpus, and then rewrites every utterance in such a way that words on the list of frequent words are retained as they are, while all other, less-frequent words are replaced with the symbol X, standing for a slot placeholder. The remaining rewritten utterance is then taken to be the frame structure for that utterance. Take for instance the question

Can you hold it?

Some of the most common words in English (and, as will be shown, in the Manchester corpus) are “can”, “you” and “it”, while “hold” is relatively rare. As a result, this question would be rewritten in frame form as

Can you X it?

Likewise, the utterance

That’s not a yellow one.

contains the common words “that’s”, “not”, “a” and “one”, while “yellow” is relatively rare, so that the procedure presented here yields the frame

That’s not a X one.

Looking at the schematic frames for these two utterances, it seems intuitive that the X slot in “Can you X it?” is likely to be occupied by a verb, and the slot in “That’s not a X one” by an adjective. For these two frames at least, the notion that a frame might provide sufficient information to categorize the word occupying its slot seems quite promising.

Next, the algorithm selects the most frequently-occurring frames that accommodate the *widest range* of different words, and attempts to form groups of these frames, by grouping together frames that accept roughly the same sets of words into their slots. In this way, it might be possible to identify a whole set of frames that accommodate, say, nouns into their slots, and to group them together into a noun frame cluster (and to do the same for verbs and adjectives).

Some details about the internal structure of the frames should be considered. Given that we are interested in finding *lexically-specific* frames, the frame would need to contain a fair amount of lexically-specific material. For this reason, frames consisting of only X's and punctuation, e.g. *X, X X ?* or *X X X X X* are not allowed. In practice, all frames used in this procedure are required to contain at least one lexically-specific word.

Furthermore, another constraint imposed on frames is that sequences of slots are not allowed. Hence, an utterance such as "the glass broke", which would be represented as "the X X", would not be considered in these experiments, because of the sequence of two X slots. All X filler words should therefore be "isolated", in the sense that they are flanked on either side by either a frequent word or an utterance boundary.

6.2.2 Details

The procedure followed in the experiments of this chapter is as follows:

- Identify the *N* most frequently-occurring words in the child-directed portion of the corpus, where *N* is a parameter of this model.
- Rewrite all utterances in the corpus, retaining only the words on the list of most frequent words, and replacing all other words with an X. Each utterance is therefore expressed in its skeletal form, as determined by the sequential order of frequent words plus placeholders for other words. These structures are now treated as potential lexically-specific frames, with the most frequent words making up the lexically-specific portion of the frame, and the X's indicating the positions of the variable frame slots.

  Some frames are likely to recur in the input, and it is possible (in fact, highly likely) that they will recur with different words filling the X slots. To the extent that this is true, these lexically-specific frames provide an opportunity for the child to discover that there might be a class of words that are licensed to occur in the particular slot, and to try and find out what the commonalities between these words are.

  The most obvious area of shared similarity between words would be their meaning: the words that fill the X slot in, e.g., "don't X it" are likely to be words for actions, these actions may be ones that the child is able to perform herself, and, they may be actions that are often met with disapprobation by the mother. However, meaning is completely ignored in the work presented in this thesis; as noted previously, this is probably a serious shortcoming. Chapter 11 outlines ways in which this issue may be tackled in future work.

- Collect frequency counts of all co-occurrences of particular frames with particular focal words. This provides us with a *co-occurrence data matrix*, which will be the data representation on which the clustering process of this chapter and the ambiguity resolution processes of the next chapter will work. The data rows in the data matrix correspond to frames and the columns to focal words. Each cell at the intersection of a row and a column contains the frequency with which the particular corresponding focal word occurred in the particular corresponding frame slot in the Manchester corpus.
- At this point, it is necessary to filter the data matrix in order to make use of only the most *reliable* frames and focal words. There are two main ways to determine reliability:

  (i) Reliable frames and words *occur reasonably often*, because rare words and especially rare frames provide unreliable evidence (e.g. the frames may have been misanalysed, words may have been miscoded during corpus transcription, words that occur only once in a corpus may be misleading if that one occurrence entails an anomalous sense, etc.)

  (ii) Reliable frames and words are productive units of the language and so *combine with a wide variety* of words and frames respectively. Words and frames that are not flexible in this way are more usefully thought of as subcomponents of fixed phrases, and may not really be productive units in their own right (Bybee, 1985; Goldberg, 1995).

  These goals can be achieved by restricting the data matrix so that it includes only:

  - frames that occur at least F_T times in the data matrix
  - words that occur at least F_W times in the data matrix
  - frames that accept at least V_T different focal words in their slots
  - words that occur as the focal words in the slots of at least V_W different frames

(where F stands for “frequency” and V for “variety”; these notions are also commonly described by the terms *token frequency* and *type frequency* respectively).

For simplicity, one could make use of only the two V parameters, which is effectively the same as taking $F_T \leq V_T$ and $F_W \leq V_W$. To simplify the model even further, one could consider only $V_T = V_W$, so that effectively only one parameter V would be required in the model. In fact, V is not manipulated in any of the experiments reported here, and its value is fixed at an arbitrary value of 5, i.e. the data set is restricted to all frames that occur with at least 5 different word types in their slots and all words that occur in at least 5 different frames (this will be called the ‘5-5 criterion’).

- Lastly, data clustering analysis methods are used to form clusters of frames, clustering together frames that take roughly the same kinds of words into their slots. In this way, sets of contexts are created that are similar to each other in the words that they accept, so that one might expect that a word which occurs in one context of the set may easily be acceptable in any other context in the set. As previously stated, it is one of the main hypotheses of this work that these large frame clusters will correspond to traditional parts-of-speech such as noun, verb and adjective.

6.3 Psychological considerations

The set of complete utterances is an appropriate domain for the discovery of frequently-used constructional frames. Firstly, full utterances are usually delimited on either side by silence on the part of the speaker, a clear indicator of their status as complete, autonomous units (i.e. constructions) in the language. Secondly, utterances (at least, simple and short ones) often tend to cohere suprasegmentally in having a single intonation contour. Thirdly, an utterance often serves as the vehicle for a single pragmatic intent on the part of the speaker, e.g. conveying a single message, question or request, which would reinforce treating an utterance as a single coherent unit. As mentioned earlier, Tomasello (2006) has argued that *utterance-level constructions* play a prominent role in language development: these are verbal expressions that can be used as complete utterances, and that are associated in a routinized way with certain communicative functions. (While the current work does not make use of information about meaning or communicative intent, and so cannot be said to be identifying utterance-level constructions directly, it does aim to discover some of the most prominent full-utterance structures in the corpus by taking isolated utterances as its starting point.)

Starting from the full utterance means that this approach is compatible with the Gestalt language learning strategy of Peters (1977). However, any particular utterance is of course made up of entirely concrete material; in order to account for semi-abstract, lexically-specific *frames* that contain some fixed words and some variability in slots, it is necessary to describe how such structures may be abstracted out from the set of utterances.

One technique by which this could happen is for the child to store in memory essentially all specific utterances that she hears (or at least those that occur frequently), and then, when encountering an utterance that has some material in common with another utterance in memory, to recall the previously stored frame, align it against the current frame and postulate a shared frame consisting of the shared material plus a slot for the non-shared material. As more evidence accrues that this analysis is correct, the frame structure will be reinforced, and gain in psychological status.

This strategy is not new; it has been proposed and implemented experimentally by Van Zaanen in his ABL system (Van Zaanen, 2001), and will not be considered in this chapter. The strategy is compatible with exemplar theories of categorization (Goldinger, 1996, 1998; Hintzman, 1996; Kruschke, 1992); however, it does require that all or most utterances should be able to be retrieved verbatim from memory, which may make a large demand on the language-learner's memory retrieval abilities.

The technique proposed in the previous section suggests a different way of developing full-utterance frames, one which builds these structures from individual, highly familiar words. In the course of being exposed to language input, it can be expected that the child will initially recognize no words, and at later stages will be able to recognize an increasing number of words. This would happen as the child becomes aware of certain phoneme sequences that occur fairly regularly in the input. These would be common words and word collocations (recall that in the current work it is assumed that the speech signal has already been successfully segmented into words). It seems likely that the first words she will be able to recognize from their phonological strings alone will be the most *frequent* words.

If the child is also able to notice co-occurrence patterns between words in an utterance, she will, once again, most likely start with the co-occurrence patterns between the most frequent words. Suppose that at some stage the child can recognize the very familiar words “you”, “can’t” and “that”, but not yet the less frequent word “chew”. When faced with the utterance “you can’t chew that”, what the child can recognize out of the utterance could be represented as “[you] [can’t] […] [that]”. Given more extensive experience of this pattern, possibly with different slot fillers (“eat”, “drink”, “have”, etc.), the child may eventually discover the co-occurrence pattern between the frequent words, so that the larger pattern “you can’t … that” may become a familiar one. As these words and collocations frequently recur in several utterances, therefore, they may become associated with each other into a larger configuration of words, with the positions of the variable slots between them being part of the mental representation. Here, the actual identity of the intervening material plays no role in the abstraction of the frame (but the work described in Chapter 9 will take a different approach).

Such a process may account for the results obtained by Gómez and Maye (2005) in an artificial language learning paradigm: it could be argued that children had abstracted out a kind of frame for the stimulus sentences, consisting of two words linked by a disjunct dependency, by attending to co-occurrences between the most commonly-occurring elements.

At the same time, the continuous phonological contour typical of many utterances, as well as the demarcation of the utterance by silence on either side, may serve to tie the elements of the utterance together into a coherent whole.

Note that I refer here to mere recognition only, and to a process by which the "texture" of English utterances becomes familiar to the language-learning child; it is *not* required that the child should know what any of these structure-building words mean. However, the linguistic validity of the full-utterance frame would become even stronger if it was accompanied by predictable semantic concomitants, as in the work by Smith and colleagues (Jones & Smith, 1998; Jones et al., 1991; Landau et al., 1988; Samuelson & Smith, 1999; Smith, 2001; Yoshida & Smith, 2005; reviewed in Section 3.4.1) suggesting that certain linguistic frames serve to draw attention to object shape.

6.4 Preliminary considerations

6.4.1 Variant treatments of the frequent/infrequent word dichotomy

With regards to the role of the most common words, there are at least three variant ways of handling the frequent/infrequent word dichotomy, with slightly different psychological interpretations.

(1). Strict, with no replacement: Frequent words form a strict dichotomy with non-frequent words, in that all tokens of frequent words are used only as potential frame-building words, and are never considered as potential slot-fillers. If frequent words are not taken up into frames, they are not returned to the pool and can serve neither as frame-building elements nor as fillers.

(2) Strict, with replacement: Frequent words form a strict dichotomy with non-frequent words, but if a particular frequent word type is not used to form part of any frame, it is returned to the pool of words which can function as fillers, and so is no longer part of the frame-building element set. Under a psychological interpretation, this means that the child is able to *keep track* of which words are likely to form part of a frame, and hence are involved in establishing the "texture" of English, and which words are not, despite their high frequency of occurrence. This requires some initial exposure to English for the sake of becoming familiar with a number of different words, followed by a phase of becoming familiar with frames, and an explicit or implicit marking of the constituents of

these frames as frame-building words. Words that are marked as frame-building words cannot also serve as fillers of frames; hence the dichotomy is described as strict.

(3) Broad: The same two processes that were assumed under “strict with replacement” are at work; however, once frames have been formed, any word (including any frame-building word) can play the role of a filler. Such an approach would investigate the extent to which some frame-building words may play dual roles or have two different semantic functions, one “contentful” function where the word functions as a typical content word and the other more grammatical, with the word playing the role of a typical function word.

In the current work, I will investigate only the option “strict with replacement”. This means that words in the list of the top *N* most frequent words are “returned to the pool” if they do not form part of any frames after filtering out frames and words (by setting the parameter *V* equal to 5 as discussed above). This means that the returned words are treated as “content words”, i.e. they can function as focal words whose part-of-speech is to be determined.

6.4.2 Clustering analysis

After the data matrix is compiled, it is subjected to a clustering analysis. The purpose of the clustering step is to attempt to create clusters of frames that together comprise a contextual paradigm from which the classes of say, verbs, nouns, or adjectives may be induced. All focal words that occur in a frame that has been allocated to a particular cluster are said to belong to the same part-of-speech.

In this experiment, therefore, I am investigating the possibility that the part-of-speech of a word may be determined entirely from its appearance in one of a number of highly familiar contexts that have the status of linguistic units (plausibly constructions) for the child. I am hypothesizing that the main categories (noun, verb and adjective) of English may be bootstrapped out of only a *small set* of these lexically-specific contexts. This experiment is therefore the direct converse of the early experiments by Finch (1993) and others (e.g. Mintz et al, 2002; Redington et al., 1998), in which words were clustered together if they appeared in similar contexts, although the contexts themselves were never the target of conscious focus. In this instance, it is the contexts that are clustered together, rather than the words. Subsequent chapters will take a more sophisticated approach.

Many different clustering algorithms have been proposed in the data analysis literature. Some of the most popular approaches include hierarchical clustering, prototype-based methods such as K-means clustering, and density-based algorithms (see e.g. Tan, Steinbach & Kumar, 2006, for an overview). The essence of clustering is that members of a set of items are placed together into groups or *clusters* on the basis of shared characteristics. Items with similar characteristics, typically represented as numerical values in a data vector, are placed into the same cluster, while items with dissimilar characteristics end up in different clusters. The clusters are often, but not always, mutually exclusive, and membership of a cluster can be all-or-nothing, or a matter of degree.

Insofar as clustering can be regarded as the formation of an abstract category of items based on shared characteristics of the items that go into the category, clustering is entirely compatible with standard psychological models of category formation (Goldinger, 1996, 1998; Hintzman, 1986; Kruschke, 1992; Rosch, 1983). All experiments in this work will start from a basic set of categories produced by *hierarchical clustering.* In the current chapter, this categorization will be used “as is”; later chapters will present more sophisticated elaborations to hierarchical clustering.

Hierarchical clustering operates on a data matrix where the rows of the matrix are the data vectors that represent the characteristics of the items that are about to be clustered (one row per item). All items are initially allocated to “singleton” clusters of their own. Clusters are subsequently merged into larger clusters by combining the two clusters with the most similar characteristics, according to a predefined *distance function* (the two clusters with the smallest distance between them are combined). The distance function is defined over two data vectors representing individual elements.

Another significant parameter in hierarchical clustering is the *linkage function,* which specifies how to apply the distance function in order to calculate the distance between two clusters with *more than one* element. Hierarchical clustering proceeds by repeatedly merging together the pair of clusters with the smallest linkage function value.

Powers (1997b) provides a very thorough and systematic evaluation of the effects of combining different clustering distance functions and linkage functions, in the context of producing a hierarchical clustering on the alphabetic letters of English text. In the current context of part-of-speech induction, Finch (1993) obtained the most satisfying results by using *Spearman's rank correlation* as the distance function, and *average linkage* as the linkage function. Preliminary testing showed this combination of functions to produce the best results in the current work as well, so that they are the two functions used in all work reported here.

Prior to clustering, the data rows are preprocessed in order to offset the effects of frame frequency. Recall that each row corresponds to a frame slot, and each cell of the row contains the number of times that each particular focal word was encountered in that slot in the Manchester corpus. A row can therefore be regarded as the "word profile" vector for that particular frame. The distance function discussed above is applied to word profile vectors of different frames. When two frames have similar profiles, their distance value will be low, and they are more likely to be clustered together than two frames with disparate word profiles.

If the data matrix is provided to hierarchical clustering in its raw frequency form, then the absolute token frequency of a frame will have a large influence on the magnitude of the distance of its word profile vector to that of another frame, artificially increasing the distance for some distance functions, and decreasing it for others. Given that we regard these frames as linguistic units with psychological reality for the child, it would be preferable to treat all frames equally, regardless of their particular token frequency in the corpus. Hence, the data matrix is *L1-normalized* to correct for frame frequency: the value of each cell in the data vector for a particular frame is divided by the *sum* of the values of all cells in that row (i.e. by the overall token frequency of the frame itself). The result is that each cell in a particular row contains the *conditional probability* that each particular focal word will occur in the *given* frame slot corresponding to the row. The hierarchical clustering process will now cluster frames together based on their profiles alone, with no effect from the token frequencies of the frames.

A further parameter in this model is the number of clusters *K* that are produced by the clustering algorithm. Hierarchical clustering produces a tree of relationships between the subclusters that it forms, and two clusters that are merged into a larger cluster are represented as two sibling nodes in the tree, with the merged cluster as their parent. In order to produce *K* clusters, the tree is "cut" at a certain level, so that a number of unconnected top-level nodes are produced; these correspond to the individual clusters. When using hierarchical clustering software, one supplies the program with the required number of clusters, and the tree is automatically cut at the desired level.

Specifying the value of *K* is somewhat undesirable from a modeling point of view, because one would prefer the number of clusters to emerge automatically from the model, rather than being artificially imposed from outside by the researcher. A number of techniques have been proposed for determining the optimal cutoff level in a hierarchical cluster tree. However, in the current situation we have a rough idea on theoretical grounds about the number of clusters we would like to produce. The three most prominent categories of content words in English are nouns, verbs and adjectives, and a preliminary inspection of the words that appear in slot positions in the current experiment shows that these categories make up the vast majority of all focal words. This suggests that the number of clusters should be low, and roughly of the order of 3 to 7. Ideally, the clustering process should produce one cluster each corresponding to the nouns, adjectives and verbs, but this will not necessarily happen in each instance. Often, hierarchical clustering produces small and "idiosyncratic" clusters consisting of a few items that happen to be closely related (in this case, they would be a set of frames that happen to take the same words into their slots, while not having a large amount of similarity to many members of the larger clusters). In these cases, one might expect some of the major classes to be merged together for low values of *K*, because the idiosyncratic classes are regarded as more dissimilar to the other classes than the two merged major classes are to each other. The major classes would then "separate out" only for larger values of *K*. In the current work, I will in each case choose the lowest value of *K* that produces 3 clusters corresponding to nouns, verbs and adjectives. In nearly all cases reported in this chapter, the value of *K* is in fact equal to 3. In all other cases, the point where the three main categories emerge is also the first point at which three *sizeable* clusters have formed (i.e. clusters that cover more than 1% of the instance tokens in the data set). This seems to indicate quite strongly that the three main categories are an intrinsic feature of the English language, at least in the input to language-learning children.

6.4.3 The number of candidate frame-building words

The process of frame formation is based on finding skeletal structure made up of the *N* most commonly-occurring words. It therefore becomes important to ask what value *N* should take on. If the success of this process was to depend heavily on the specific value of *N*, then the process could hardly be said to be a robust one, and this would make it doubtful that children learning a language could follow a similar strategy. It is desirable that the process should produce good and roughly comparable results for any reasonably large value of *N*.

The choice of *N* does not, perhaps, directly influence the quality of the results obtained in these experiments, but it does influence a number of other variables that can arguably have an effect on the quality of the results. For instance, *N* determines:

- the number N' of words that are actually used to form frames (after filtering)
- the number of different frames in the data set (after filtering)
- the number of different focal words in the data set (after filtering)
- the proportion of all utterances in the corpus that are accounted for by combining a frame and a focal word from the data set (i.e. word-frame *instance tokens*)
- the number of different frame-focal word combinations in the data set (i.e. word-frame *instance types*)

Because of the potentially important role played by *N*, the experiments in this chapter will be repeated over a range of specific values of *N*. The values that will be used are *N* = 80, 150, 180, 240, 290, 410, 450, 520, 610, 690. A detailed examination of the effect of *N* on various parameters, and a justification for the use of these specific values, will be deferred until Section 6.5.3.

6.5 Implementation

6.5.1 Qualitative results

6.5.1.1 Most frequent words in the Manchester corpus

Table 6 shows the top 290 most common words in the Manchester corpus, arranged from most to least frequent. The value of *N* = 290 was chosen as a reasonably "average" and representative value in the range to be considered. This will also be the fixed setting of *N* chosen for experiments in later chapters. Words in bold take part in lexically-specific frames as the specific, structure-building elements. Most of the structure-building words were closed-class/function words, although a number of words towards the less-frequent part of the range were open-class/content words. Out of these contentive structure-building words, some were light verbs such as "make", "give" and "take", or verbs that typically appear in fixed expressions which formed part of a larger utterance, such as "matter", "happened". Surprisingly, a large number of adjectives (including the five most prominent colour words) were on the list of frame-building elements, as were a number of very concrete object names such as "dog", "car" and "train". These words may well provide a way to bootstrap into discovering the rather common [Adjective Noun] structure in English, by means of the frames that these words support, such as "X car" (taking adjectives) and "blue X" (taking nouns). In this way, even highly contentive words can serve as a focus around which productive frames can form.

you, **the**, **it**, **a**, **to**, **oh**, **that**, **what**, **is**, **I**, **on**, **and**, **do**, **there**, **are**, **in**, **we**, **that's**, **no**, **one**, **your**, **it's**, **have**, **don't**, **childname**, **can**, **right**, **he**, **going**, **not**, **this**, **go**, **got**, **put**, **well**, **then**, **look**, **yeah**, **want**, **now**, **think**, **what's**, **of**, **with**, **like**, **for**, **they**, **all**, **did**, **you're**, **yes**, **here**, **get**, **isn't**, **me**, **see**, **come**, **some**, **them**, **she**, **shall**, **up**, **out**, **okay**, **be**, **just**, **mmhm**, **know**, **was**, **at**, **there's**, **her**, **mummy**, **he's**, **very**, **good**, **you've**, **where**, **bit**, **little**, if, **because**, **didn't**, **down**, gonna, **off**, **does**, **doing**, **big**, so, **back**, **him**, **I'm**, **can't**, **his**, hmm, **make**, about, **where's**, **they're**, **why**, doesn't, **more**, **say**, **my**, **nice**, play, **again**, **these**, dear, **but**, **over**, **car**, **thankyou**, **who's**, **aren't**, else, **two**, **what're**, **has**, **let's**, **or**, **baby**, **another**, **who**, **other**, **those**, **haven't**, when, **daddy**, **how**, **take**, **I’ll**, **gone**, **she's**, **need**, will, **please**, **were**, **find**, **train**, better, **any**, way, pardon, **away**, had, **too**, sit, **an**, we'll, **round**, **eat**, whoops, something, **which**, tell, **aswell**, **mummy's**, done, **would**, **box**, **give**, goes, **red**, alright, really, might, **I've**, remember, house, **from**, **girl**, willn't, **boy**, could, **color**, **we've**, let, **wasn't**, fit, you'll, **time**, won't, **darling**, **blue**, **green**, **man**, things, **having**, book, hasn't, **we're**, went, sleep, aah, been, hair, getting, coming, dolly, top, as, **three**, horse, **one's**, animals, **yellow**, **only**, **many**, head, **um**, **first**, looking, today, **said**, tea, careful, help, **bridge**, **called**, sure, **draw**, **thought**, **turn**, playing, through, ones, **happened**, though, **orange**, **looks**, **hello**, much, toys, **here's**, pull, **next**, bricks, **silly**, stuck, fall, try, **lots**, hey, anything, enough, minute, sorry, **naughty**, being, drink, bed, Caroline, **cake**, water, keep, **build**, putting, **door**, **matter**, **under**, Anna, should, nose, **still**, **poor**, stop, cow, wants, tiger, **making**, **hurt**, stand, **funny**, **yet**, **dog**, work, read, **move**, show, **eggs**, into, broken, elephant, leave, says, shopping.

Table 6. The 290 most common words in the Manchester corpus, with frame-building words indicated in bold.

6.5.1.2 Example frames

Having selected the top N words for some value of N, the next steps in the algorithm are:

- to find all full-utterance frames made up of these words,

- to collect co-occurrence data describing which full-utterance frames occur together with which focal words, and
- to filter the resulting data matrix to contain only frames with at least 5 different words occurring in their slots, and words that occur in the slots of at least 5 frames (the 5-5 requirement)

There are two reasons for the 5-5 requirement, related to the token frequency and type frequency respectively of the frames. Firstly, only presumed frames that occur fairly often (high token frequency) are likely to be reliable features of the input and hence likely to be “real” elements of the language. The same remark can also be made from a pragmatic point of view: larger frequencies in the cells of the data matrix are more likely to provide a reliable expression of the distance between two frames, and hence a better clustering of frames. Consequently it is useful to require there to be at least 5 non-zero entries in each row and column.

Secondly, if frames take only a limited set of focal words into their slots, then it may be more appropriate to describe them as fixed expressions rather than flexible and productive frames. The same holds in cases where certain words only occur in one or two frame contexts. The 5-5 requirement ensures that only frames and words which combine productively with each other (high type frequency) will be included in the final data set.

The details of the algorithm are as follows: We start with a set of candidate frames and a set of candidate words that have occurred, in the corpus, in the slots of these frames. First, all frames with fewer than 5 different words occurring in their slots are discarded from the frame set; next, all words that, as a result of the previous step, occur in fewer than 5 different frames are discarded from the word set. These two steps are repeated until convergence of both the frame and word sets, i.e. until no additional frames and no additional words are discarded.

The resulting set of frames appears to contain a number of intuitively comprehensible and familiar sentence frames, which could quite plausibly be used for part-of-speech induction. Table 7 shows the 100 most frequently-occurring frames produced by this process for *N* = 290, and Table 8 shows a selection of frames together with the words that occur in the slots of each one.

oh X; that's X; are you X ?; a X; it's X; it's a X; X then; that's a X; well X;
can you X ?; is it X ?; is that X ?; a X ?; X it; the X; where's the X ?; X me;
Z the X; not X; poor X; that's the X; X the Z; Z your X; X your Z; there's the X;
what X ?; it's not X; X a Z; do you X ?; do you like X ?; can you say X ?;
i'm X; they're X; be X; the X ?; what're you X for ?; and X; he's X; is he X ?;
Z a X; in the X; you X; X down; two X; X up; there's a X; it's not a X;
on the X; some X; where's your X ?; X it ?; just X; X and Z; no X; and the X;
it X; that's X, isn't it ?; is she X ?; very X; that's a good X; Z and X; is it a X ?;
and a X; what's X ?; there's X; don't X; your X; another X; that X; which X ?;
you're X; X you; are they X ?; don't X it; what about X ?; X what ?; Z you X ?;
what're you X ?; big X; say X; that's a X , isn't it ?; X again; what a X;
where's X ?; that's not a X; what do you X ?; X , childname; more X; she's X;
it's X , isn't it ?; it's a X , isn't it ?; that's not X; X that ?; your X ?;
there's your X; you like X , don't you ?; is that a X ?; it's X , is it ?; X you ?;
it is X

Table 7. The 100 most frequently-occurring full-utterance frame slots in the Manchester corpus, for N = 290.

X it off : [finish, knock, pull, shake, slide, wipe]

X , was it ? : [Alice, horrible, juice, motorbike, yesterday]

a X , isn't it ? : [cat, cup, dinosaur, doggy, frog, hotdog, lemon, lid, lift, mouse, panda, piano, radiator, shirt, top, tractor]

a X cake : [birthday, chocolate, jam, pretend, yummy]

a bit X , isn't it ? : [different, difficult, easier, fiddly, stiff]

a blue X : [ball, balloon, brick, knife, monster, slide, square]

a very X one : [fast, huge, old, small, tiny]

and her X : [arms, cheek, clothes, hand, hat, head, toys, trousers]

are we X ? : [comfortable, done, ready, sorted, stuck, tired]

are you X your Z ? : [counting, eating, getting, stamping, washing, writing]

are you going to X ? : [bed, blow, count, dance, drive, hammer, help, hide, jump, listen, nursery, paint, pull, sing, sleep, start, swim, talk, tip, town, watch, work, write]

are you making a X ? : [farm, fence, house, mess, sandwich, tower, wall, windmill, zoo]

be X : [careful, gentle, quick, quiet, sick]

because he's X : [broken, hard, poorly, sad, Thomas, waving, working]

can I have a X ? : [bite, cuddle, digger, kiss, lick, piece, pig, play, spoon, strawberry, truck]

can you X that ? : [catch, feel, hear, hold, manage, pull, read, remember, sing, squash]

do you want to do some X ? : [coloring, cooking, drawing, rolling, writing]

don't X me : [ask, bang, bite, forget, help, hit, kick, Mummie, push, tell, tickle]

from X : [Duplo, FatherChristmas, Grandma, MacDonalds, Mark, OldBear]

has it X ? : [broke, broken, crashed, popped, stuck]

he is X : [clumsy, coming, cross, crying, grey, happy, lovely, odd, outside, pink, stuck, upset]

how many X have we got ? : [animals, books, candles, penguins, tins]

I X it : [brought, caught, done, drop, dropped, had, love, made, mean, missed]

I don't know where X is : [Caitlin, duck, Heidi, Henry, horsie, Tigger]

let's have a look at your X : [feet, finger, nose, teeth, toes]

Mummy X it : [ate, broke, catch, drive, dropped, fit, fix, fixed, had, hide, hold, keep, mend, open, pull, push, roll, sort, stop, wipe]

nice and X : [clean, cold, comfortable, dry, flat, gently, hard, hot, quiet, straight, sweet, tidy, warm]

some more X ? : [beans, bricks, chips, dogs, fence, fun, gates, peas, spaghetti, tomato, toys, water]

that X be Z : [might, must, should, will, willn't]

that Z be X : [better, easier, fine, interesting, new, sticky]

that's a baby X : [calf, chicken, cow, duck, goat, horse, lion, pig, sheep, tiger]

train X : [coming, goes, horse, thing, track]

turn the X over : [basket, cakes, lid, numbers, tape, top]

we X : [could, had, might, saw, went, will, willn't, won't]

what X is that ? : [animal, letter, noise, piece, shape, song]

what have you X ? : [bought, done, dropped, forgotten, found, lost]

what're we going to X ? : [buy, cook, drink, fix, play]

why are you X ? : [busy, crying, hiding, laughing, sad, shouting, stamping, tired, whispering]

you X a Z : [bang, choose, had, made, play, read, sing]

you Z a X : [book, drink, drum, goal, lorry, rainbow, tower]

Table 8. A selection of frames from the Manchester corpus, produced from the top 290 most frequent words.

Note that, because of the way that the frames are constructed, it is perfectly possible that one particular frame contains more than one slot (e.g. “Did you X a X ?”). In collecting data for the data matrix, each of these slots is tracked *independently*, so that there are two different and independent frames: “Did you X a Z ?” and “Did you Z a X ?”, where the X in each case represents the active slot for which we are collecting filler data, and the Z represents the inactive slot. The reader should remain aware that the term “frame” refers to a particular configuration of words with one active slot, not to the underlying schematic utterance structure (e.g. “Did you X a X?”) from which these configurations arose.

The assumption of independence, however, is potentially unrealistic. If a frame has, say, two slots, each of which could be occupied by members of two different parts-of-speech, then it is likely that the two slots are not independent, but that a choice between the two alternatives in the first slot would determine the choice in the second slot. Nevertheless, these potential interactions are ignored in this simple model, in accordance with the idea that “you can’t abstract two things at the same time”.

6.5.2 Clustering results

Hierarchical clustering was carried out on the data matrix obtained as described above, using Spearman’s rank correlation as a distance measure between rows (frames), and using the average linkage algorithm of Sokal and Sneath (1963) in order to form clusters.

The clusters obtained when the algorithm is asked to produce 3 clusters are shown in Table 9. Intuitively, it seems as if the clustering process has produced 3 very convincing sets of contexts for, respectively, verbs, adjectives and nouns. The frames in cluster 1 seem to be for the most part contexts in which one would expect to see verbs, and exhibit a range of argument structures (e.g. “X it”, “X it off”, “X it to me”, “X on my Z”, “X the Z down”), in a variety of questions, imperative utterances and declarative utterances. The frames in Cluster 2 seem to be likely contexts for adjective fillers, especially in conjunction with frequently-used nouns (“X baby”, X box”, “X car”), the copula (“are you X again?”, “I’m X”), and modifiers (“a bit X, isn’t it?”, “it’s very X”, “not too X”, “still X”). By far the largest of the three clusters is Cluster 3, which appears to be a

Cluster 1 (233 frames)	Cluster 2 (324 frames)	Cluster 3 (908 frames)
X again	X , is it ?	X at the Z
X down	X baby	Z at the X
X her up	X box	Z in your X
X him	X car	Z with a X
X it	X girl	a X
X it off	Z it X	a baby X
X it to me	X some Z	a green X
X me	a X one	all these X
X on my Z	a bit X , isn't it ?	and another X
X over	all X	are you having a X ?
X that one	and they're X	back to the X
X the Z down	are you X again ?	called a X
X the car	be X	can i have a X ?
X with your Z	because he's X	can you find me the X ?
X your Z	going X ?	did you Z your X ?
are you going to X ?	has it X ?	do you want a X ?
can you X a Z ?	he's X , is he ?	don't put your X in there
can you X it ?	i know it's X	get the X out
did she X ?	I’m X	give him a X
do you want me to X it ?	is it a X one ?	it's for X
don't X	is it too X ?	it's not very X , is it ?
give it a good X	it's X , isn't it ?	more X
I’ll X that	it's very X	on his X
let's X it	make it X	put her X on
mummy X it ?	not too X	shall we draw a X ?
shall we X again ?	still X	some X
she X	that's a X one	that was my X
to X	this one's X	that's a X
what did you X ?	what have you X ?	the X
why don't you X ?	what're you X for ?	there's lots of X
you X it then	who's X ?	what X do you want ?
you can X it	you are X , aren't you ?	what's happened to X ?
you're going to X	you're very X	your X

Table 9. Representative frames from each cluster for full-utterance frame clusters, N = 290, with 3 clusters, together with their associated slot-fillers.

cluster of frames that accept nouns into their slots. Apart from the “archetypal” noun contexts “a X” and “the X”, there are a great number of basic utterance structures illustrating the ways in which nouns and noun phrases can be used in various argument structures, e.g. “that’s a X”, “do you want a X?”, “on his X”, “give him a X”, etc.

Not all frames assigned to the clusters fit with the general trend of the cluster. For instance, the “verb” Cluster 1 contains “give it a good X”, which accepts as slot fillers the nominalized verb forms “pull”, “rub”, “squeeze”, “wash” and “wipe” which should have been classified as nouns in this context. In Cluster 2, in particular, there are many non-adjectival frames, especially as a result of the prevalence of many past and present participial verb forms such as “stuck”, “broken”, “exciting”, etc., which are arguably adjectival in contexts such as “is that X?”, “a X one”, but which are also used in verbal form in frames such as “has it X?” and “what have you X?”; these frames end up being grouped with the other adjectival frames in Cluster 2 because of the large overlap in their filler sets. There are also many ambiguous frames in Cluster 2, mainly due to the presence of the copula, which can be used in conjunction with not only adjectives, but also proper names, mass nouns and pluralized common nouns. Lastly, some assignments of frames to clusters are simply anomalous: there is no clear reason why, for instance, “It’s not very X, is it ?” is grouped with the “noun” frames in Cluster 3.

It is perhaps difficult to gain a clear impression of the nature of these clusters, and their putative relationship to traditional parts-of-speech, by inspecting a list of frames that form the clusters. It may be more helpful also to display these qualitative results in a format derived from the *words* that go into the slots of the frames that make up these clusters. This is done in Table 10. The word lists were obtained for each frame cluster by, for each word in the data set, counting the number of frames from that cluster in which the word occurred as a filler. The list was then sorted from highest frequency to lowest, so that, for instance, the word which occurred as a filler for the highest number of *different* Cluster 1 frames (“open”) appeared at the top of the Cluster 1 list, and so too for the other two clusters (see Box 1 for a description of the algorithm). Table 10 shows the top fifty words that occur in the greatest number of frames from each cluster. As expected, the words most closely affiliated with a wide range of Cluster 1 frames are (mostly morphologically unmarked forms of) verbs, and the word list for Cluster 3 contains nouns exclusively. In the case of Cluster 2, some of the problems identified earlier with frames are confirmed in the word list. The list contains a great number of verbal present and past participial forms, some of which are arguably adjectival (“stuck”, “broken”, “tired”), but also some which are probably not (“playing”, “swimming”, “hiding”) and this is presumably due to the inclusion of a great number of frames in which the X slot is associated with the copula, as discussed above. There is also one outright proper name (“Thomas”) on the list. Nevertheless, it can clearly be seen from the majority of the words in Cluster 2 that it is a cluster that favours adjectives.

Cluster 1	Cluster 2	Cluster 3
open, pull, sing, hold, try, leave, watch, fix, push, read, catch, count, keep, remember, will, play, press, sit, use, drive, help, jump, blow, stop, tell, wash, willn't, hear, hide, roll, throw, pick, bite, close, drink, stand, break, might, wipe, bring, cut, fall, fit, write, doesn't, dropped, finish, kick, kiss, show, shut, tickle, tip, undo	stuck, broken, alright, done, tired, better, hot, cold, eating, dirty, crying, hiding, lost, lovely, clever, coming, hungry, hard, poorly, wet, finished, sleeping, pink, asleep, happy, yours, sad, ready, heavy, outside, purple, Thomas, clean, playing, pretty, small, white, upstairs, cross, difficult, drawing, empty, fine, found, horrible, swimming, wrong, dry, love, lucky, mine, noisy, running, sorry, through	horse, cow, fish, house, tiger, monkey, pig, book, dolly, hat, cat, chicken, elephant, boat, sheep, panda, penguin, bricks, head, drink, foot, ball, bag, nose, teddy, tower, water, animals, duck, hair, tea, egg, giraffe, juice, cheese, chips, hand, rabbit, trousers, milk, table, top, tractor, dress, eyes, lady, lion, feet, hippo, bus, chair, Thomas

Table 10. The words that occur in the largest set of frames from each full-utterance frame cluster, N=290, with 3 clusters.

Data structures:

- *data*: the original *I*x*J* data matrix, where *I* is the number of frames and *J* is the number of words in the data set. The cell *data[i][j]* contains the number of times that frame *i* and word *j* have occurred together in the corpus.
- *totals*: a two-dimensional *K*x*J* matrix, where *K* is the number of clusters. Each cell *totals[k][j]* will contain a pair <*x, f*> where *x* is the index of the word in *data*, and *f* is the frequency value, giving the total number of frames from cluster *k* that have occurred with word *x* in the corpus.
- *clustering*: the original 1x*I* hard clustering vector; *clustering[i]* contains the index of the cluster to which frame *i* has been assigned by the hard clustering algorithm.

Algorithm:

```
1. Let k be the index of each frame cluster. Then, for each frame cluster, do the following:
   a. Let j be the index of each word. Then, for each word, do the following:
      1. Set the value of the index-frequency pair in totals[ k ][ j ] to <j, 0>.
      2. Let i be the index of each frame. Then for each frame, do the
         following:
               If clustering [ i ] = k (i.e. frame i belongs to cluster k), AND if
               data[ i ][ j ] > 0 (i.e. frame i and word j have occurred together
               in the data set), then increase the frequency value of
               totals[ k ][ j ] by 1.
   b. Sort the row totals[ k ] from highest frequency value to lowest. This sorts
      the words of the data set from most to least strongly associated with the
      cluster k.
```

Box 1. Algorithm to determine how strongly a word is associated with a frame cluster (as based on frame type frequency).

Note also that, because of the way that the lists were constructed, it is perfectly possible for some (ambiguous) words to appear on more than one list; for instance, “Thomas” appears on the lists for Clusters 2 and 3, and “drink” appears (correctly) on the lists of Clusters 1 and 3.

Taking the clustering analysis further reveals even finer categorial distinctions. Table 11 shows the most closely-associated words when 20 clusters are created. (Only the eleven clusters composed of more than 15 frames are shown).

This table shows that the adjective/proper noun cluster has successfully split into separate clusters for proper noun, participial verb, and all other adjective frames (Clusters 3, 2 and

1 respectively). The participial-seeming forms such as “broken” and “stuck” in Cluster 1 are actually present in the list due to their adjectival usages, and in fact the truly verbal usages of “broken” lead to its also appearing in Cluster 2.

Cluster 1	stuck, alright, broken, cold, tired, hot, dirty, better, crying, lovely, coming, hungry, clever, hard, hiding, wet, poorly, sleeping, pink, yours, asleep, happy, sad, ready, purple, finished, heavy, Thomas, white, clean, done, outside, small, difficult, pretty, upstairs
Cluster 2	done, eating, lost, found, broken, building, had, holding, dropped, drawing, driving, getting, made, riding, wearing, been, bought, missed, being, caught, drinking, forgotten, spilt, washing, writing
Cluster 3	dolly, Thomas, panda, Gordon, yours, Caroline, Henry, James, Andy, hippo, Anna, driver, monkey, grandpa, granny, penguin, Percy, Pingu, baba, Edward, rabbit, teddy, Toby, black, mummie, pink
Cluster 5	animals, bricks, cars, babys, horses, things, pigs, trains, wheels, bits, duplo, tins, cows, doll, monkeys, penguins, pieces, puzzles, toys, vehicles, bees, biscuits, candles, dolly, elephants, gates, letters, men, mine, money, ones, pennys, people, rings, sheep, starfish, teddys
Cluster 6	fish, chips, juice, cheese, chicken, water, grapes, beans, milk, bananas, bricks, strawberrys, tea, chocolate, peas, sweetcorn, bread, meat, pears, soup, tomato, apples, biscuits, carrots, fruit, icecream, lettuce, money, penguins, sheep, toast, food, oranges, sausages, cabbage, cows, paper, shoes
Cluster 8	open, pull, sing, hold, try, leave, watch, fix, read, count, push, catch, keep, play, press, remember, sit, use, help, blow, drive, stop, jump, tell, wash, hear,

	hide, roll, throw, pick, bite, close, drink, stand, break, wipe, bring, cut, fall, fit, write, dropped, finish, kick, show, shut, tickle, tip, undo, fetch, kiss, lift, mend, missed, reach
Cluster 11	horse, cow, book, tiger, pig, house, cat, monkey, boat, hat, drink, ball, penguin, bag, giraffe, sheep, fish, table, tower, duck, lion, elephant, lady, picture, tractor, chicken, tunnel, basket, bus, rabbit, top, whale, truck, banana, bull, dolly, goat, hole, panda, snake, biscuit, dress, bird, chair, gate, tree, way, window, digger, lemon, circle, hippo, lid, spoon, teddy, wheel, brick, farm, fence, story, balloon, doll, piece, sausage, thing, foot, lorry, pottie, tomato
Cluster 12	elephant, egg, apple, icecream, onion, teddy, aubergine, aeroplane, b, c, chicken, d, f, h, igloo, j, m, monkey, n, umbrella, w
Cluster 14	head, nose, hair, feet, eyes, foot, hand, trousers, tummy, tea, mouth, face, finger, legs, knee, shoes, ears, hat, name, toes, fingers, leg, arm, arms, dinner, dress, toe, clothes, tail, thumb, bottom, knees, house, juice, money, toys, bed, favorite, neck, teddy, teeth, top, bib, chair, ear, hands, pants
Cluster 16	will, willn't, might, won't, could, must, doesn't, hasn't, couldn't, shouldn't, weren't, says, wouldn't, crashed, had, should, fits, goes, leave, saw, went
Cluster 18	through, love, so, gently, nightnight, pet, sorry, aah, later, listen, sweetheart, carefully, hey, watch, better, byebye, careful, dear, nicely, phone, properly, steady, wrong, actually, ah, alright, bye, er, fine, inside, nearly, open, quick, thanks

Table 11. The words that occur in the largest set of frames from each full-utterance frame cluster, N = 290, with 20 clusters.

Other small clusters are associated with plural count nouns (Cluster 5), mass nouns (Cluster 6), body parts and clothing (Cluster 14), and modal verbs (Cluster 16). Two slightly anomalous clusters seem to be based on frames where the slot mostly follows the word “an” (Cluster 12) or precedes “for”, and one where the filler is to some degree detached from or “tacked onto the end of” the frame (Cluster 18), and so may be an interjection such as “thanks” or “alright”, a vocative such as “sweetheart” or “pet”, or an adverb such as “gently”.

Most of the 9 smaller omitted clusters are not obviously coherent, although one seems to favour numerals such as “five” and “six”, and two others are associated with the names of places, such as “shops”, “school” and “playgroup”.

6.5.3 Effect of number of frequent words used

This section examines the effects of manipulating *N*, the number of most frequent words that are considered as potential frame-building elements, on various properties of the resulting data set.

Recall that frequent words that are candidates for forming frames but that do not end up forming part of any frames are treated as potential slot-fillers instead. It might therefore be possible that, as we increase the value of *N*, there might be a point at which the number of words *N’* actually *used* in frames levels off so that further increases in *N* do not increase *N’*, and hence contribute no further additions to the set of frames that are discovered. This is to some extent the case: Figure 1 illustrates the effect on *N’* of increasing *N*. For low values of *N*, every new word added to the list is taken up into a frame structure. At about *N* = 90, the first words appear in the list that do not participate in any frames.

At *N* = 390, *N’* reaches its maximum value of 195. From this point on, *N’* actually declines in value. This is because, during the discovery process, if too many lower-frequency words are treated as potential structure-building elements, and hence are not considered as X filler elements as they should be, then many useful frames are “blocked” from being recognized in cases where their slots are filled by these lower-frequency

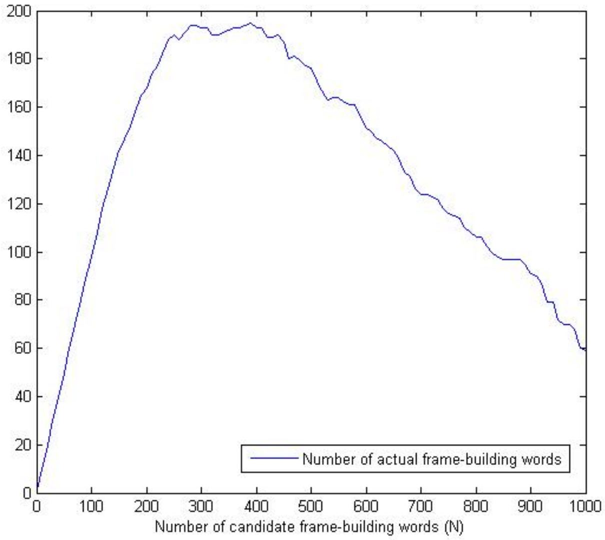

Figure 1. The effect of the number of candidate words on the number of words used to form frames.

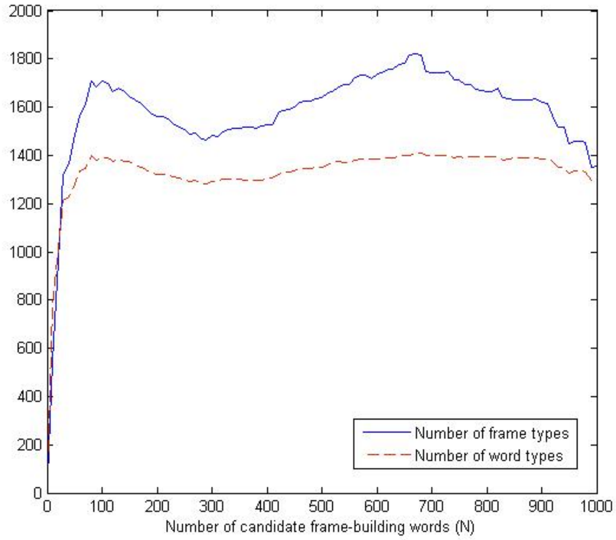

Figure 2. The effect of the number of candidate words on the number of frame types and focal word types in the final data set.

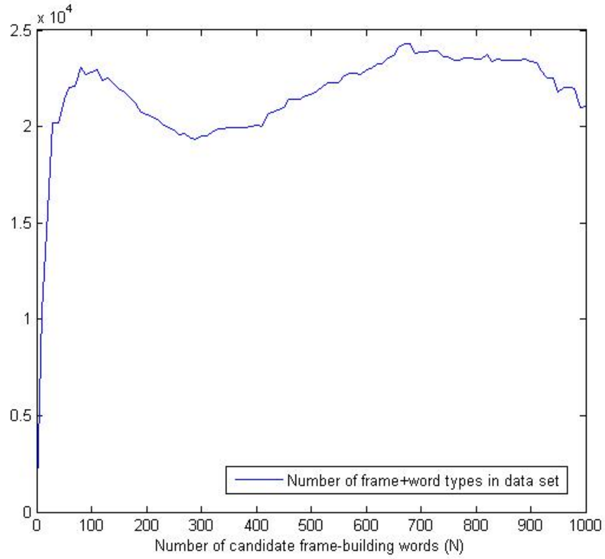

Figure 3. The effect of the number of candidate words on the number of frame+word instance types in the final data set.

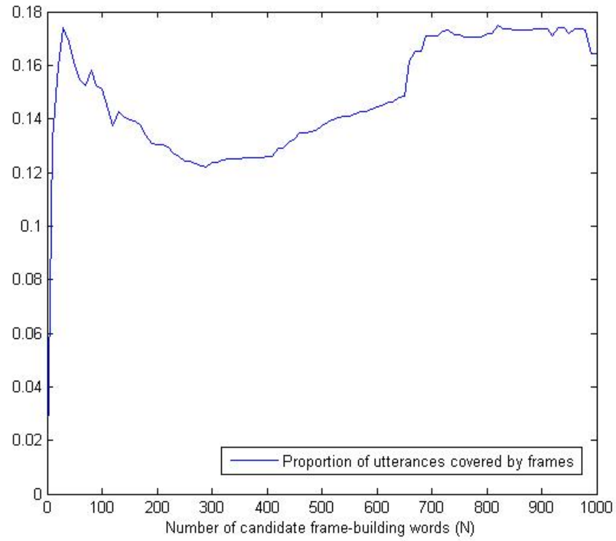

Figure 4. The effect of the number of candidate words on the proportion of all utterances covered by the data set.

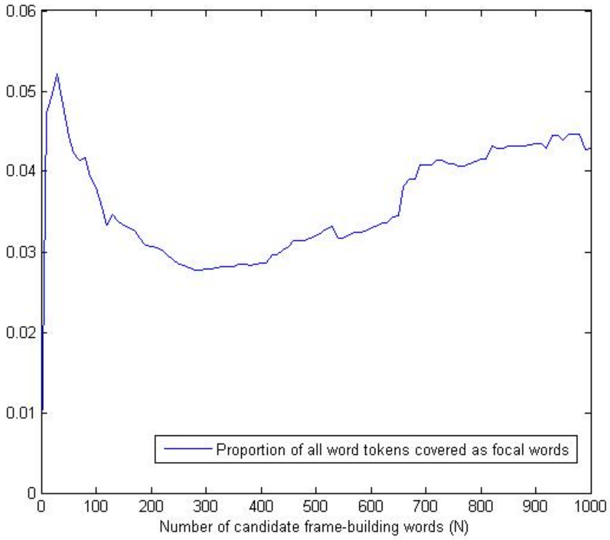

Figure 5. The effect of the number of candidate words on the proportion of all word tokens in the corpus covered by the data set (as focal words).

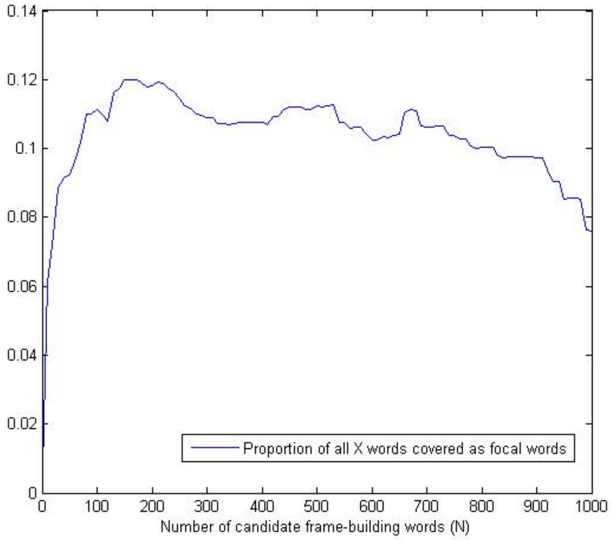

Figure 6. The effect of the number of candidate words on the proportion of all X words in the corpus covered by the data set (as focal words).

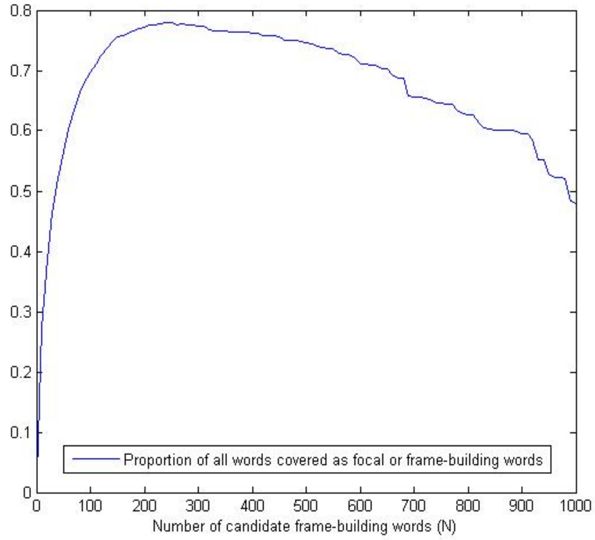

Figure 7. The effect of the number of candidate words on the proportion of all word tokens in the corpus categorized by the complete model (as either focal words or frame-building words).

words. Hence, the only frames that "survive" are the ones that occur so frequently that they still manage to be filled by a sufficient number of even-lower-frequency words. **Figure 2** shows the effect of *N* on the number of frame types and focal word types that are used in the experiment after filtering the dataset. Both the number of words and the number of frames attain two distinct peaks, at *N* = 80 and at *N* = 690 for both words and frames, and a single trough between the peaks at *N* = 280 for words and *N* = 290 for frames. Figure 3 shows the effect on the number of distinct frame-word combinations covered by the dataset; this value also peaks at *N* = 80 and troughs at *N* = 290.

In Figure 4, we see how *N* influences the proportion of all utterances that are covered by a frame-word combination in the data set, and Figure 5 shows the proportion of all word tokens that are treated as focal words in the data set. These two graphs therefore show, respectively, the proportion of utterances and of word tokens in the corpus that are actively categorized by the model.

A different way to gauge coverage is to take into account that the data set can only account for word tokens that have been rewritten as X's. Hence, a reasonable measure of coverage might be the proportion of all X's that are covered by the data set. This graph is shown in Figure 6, with values ranging between 8-12%, peaking at *N*=170.

The proportions shown in Figures 4 to 6 are rather low (mostly ranging between 12-18% and 2.5%-5%, respectively). However, they may under-represent the extent to which the model covers the corpus. The most frequent, structure-building words are not subjected to tagging, but are treated as themselves in the frames. Hence, they are arguably tagged as belonging to a category with one element, namely themselves. A fairer description of the number of words that are tagged by the model might be obtained by adding the number of word tokens appearing as focal words (as shown in Figure 5) to the number of word tokens belonging to the set of frame-building word types. This graph is depicted in Figure 7, and provides a much higher estimate of the proportion of the corpus accounted for by the model, peaking at 77% when *N* = 290.

In the experiments reported in this chapter, the value of *N* is set to one of a number of "key" values, in order to gauge the effect of manipulating *N* on the final outcome. From the graphs, it is clear that 290 is a key value, with the number of frames, words and frame-word instances being at a low point, while the number of frame-building words reaches its peak. Paradoxically, the value *N* = 290 represents the point with the largest number of frame-building word types, but where the smallest number of frames have actually been built. The converse situation holds at the two values of *N* = 80 and *N* = 690, where the numbers of frames, focal word types and frame-word instances reach their local peaks, and so these two values are used to delimit the range of considered values of *N*. Attaining the maximum amount of coverage (at *N* = 80 and *N* = 690) is not necessarily desirable in and of itself: the additional frames and words made available at these two points may be of lower quality compared to those in the smaller sets used for *N* = 290. Only a quantitative evaluation of the resulting categorization will determine whether any values of *N* produce markedly better results than others .

Intermediate values of *N* will also be considered, and are chosen on the basis of the number of frame-word instances they yield; with *N* = 80, 290 and 690 producing approximately 23, 19 and 24 thousand instances, respectively, intermediate values of *N* are chosen that interpolate between these values in increments of a thousand instance tokens at a time. The additional values of *N* are 150 (22K instances), 180 (21K), 240 (20K), 410 (20K), 450 (21K), 520 (22K) and 610 (23K).

6.5.4 Quantitative evaluation

The 3-cluster clustering obtained in Section 6.5.2 can also be evaluated quantitatively, by assigning a category to each of the focal words occurring in the utterances that were used in the data matrix. In the current experiment, the categorization process is very simple: there is one category for every cluster, and if a full-utterance frame is allocated to a particular cluster, every instance of a focal word that occurs in that frame in the corpus is assigned to the category corresponding to that cluster. Taking examples from Table 8, if the frame "what X is that?" is allocated to, say, cluster 1, then all the focal words that occur in that frame are allocated to category 1; that is to say, "animal" in the utterance “what animal is that?”, “letter” in “what letter is that?”, etc., will all be allocated to the same category.

Once a category has been assigned to each focal word instance, this experimental categorization can be compared to a gold standard categorization. As mentioned, for the purpose of evaluation we make use of only the three main categories in the gold standard: nouns, verbs and adjectives. This decision is justified by the fact that other categories represent only a small portion of the categorized words. Table 12 shows how the word tokens in the portion of the corpus considered for N = 290 are distributed into categories. The main three categories make up 86.4% of word tokens between them. The only other categories with a sizeable number of tokens are adverbs (3.2%) and communication words such as “oh” and “yeah” (2.5% for “co” and 5.3% for “co-voc”), with all other categories together making up around 3% of the total. The three main categories therefore make up the lion’s share of all tokens.

Category	# tokens	% of total
adjective	**3608**	**9.9**
adv	1162	3.2
chi	16	0.0
co	904	2.5
conj-subor	23	0.1
co-voc	1928	5.3
det-num	161	0.4
fil	76	0.2
int	39	0.1
n	**19026**	**52.0**
n~v	62	0.2
on	22	0.1
post	101	0.3
prep	124	0.3
pro	7	0.0
pro-indef	135	0.4
pro-poss	192	0.5
qn	31	0.1
v	**8984**	**24.5**

Table 12. The gold standard category distribution of the words in the full-utterance frame data set, for N = 290. Category labels are from CHILDES. Bold text indicates the three main categories considered for analysis.

We make use of the pair counting method, as discussed in Section 5.5.1, in order to evaluate the performance of the hierarchical clustering process in categorizing focal word instances. Recall that the pair-counting method is based on comparing the pairs of word-frame instances that are assigned to the same category by the clustering process, against the instance pairs assigned to the same category by the gold standard.

The pair counting approach is appropriate for a method such as clustering, where we do not strictly know which clusters correspond to which actual categories in the gold standard. However, it is possible in the current case to effect such a mapping between categories, allowing the Bookmaker measure (Powers, 2003) to be used. Recall that Bookmaker requires a contingency matrix in which the columns represent the gold standard categories, and the rows represent the empirically-derived categories, with cells of the matrix representing the number of word-frame instances belonging to one particular gold standard category that have also been allocated to one particular empirical category. In the case of clustering, we therefore need to map gold standard categories to empirical ones.

As discussed in Section 5.5.2.2, this mapping can be achieved by taking each of the gold standard categories in turn, and identifying the empirical category which contains the largest number of instances from that gold standard category. The gold standard category is then mapped onto that empirical category. (Under this scheme, it would be possible for more than one gold standard category to map onto the same empirical category.) This mapping process allows us to identify the empirical category into which instances from each gold standard category "should go", allowing the use of a measure such as Bookmaker.

This mapping was carried out for each of the key values of *N*, and the mapping was used in conjunction with the categorization contingency table to calculate Bookmaker values for each of the final categorizations.

The results for accuracy, completeness F and Bookmaker are shown in Table 13, for each of the values of N, with random baseline scores for the first three measures in italics (the random baseline for Bookmaker is always zero).

	Accuracy	Completeness	F score	Bookmaker
N = 80	0.797 (*0.480*)	0.647 (*0.389*)	0.714 (*0.430*)	0.694
N = 150	0.818 (*0.528*)	0.639 (*0.412*)	0.718 (*0.463*)	0.695
N = 180	0.837 (*0.535*)	0.690 (*0.440*)	0.756 (*0.483*)	0.709
N = 240	0.827 (*0.555*)	0.713 (*0.478*)	0.766 (*0.514*)	0.693
N = 290	0.844 (*0.559*)	0.774 (*0.513*)	0.808 (*0.535*)	0.708
N = 410	0.863 (*0.571*)	0.764 (*0.506*)	0.810 (*0.536*)	0.716
N = 450	0.871 (*0.576*)	0.771 (*0.510*)	**0.818** (*0.541*)	**0.734**
N = 520	0.870 (*0.575*)	0.762 (*0.503*)	0.812 (*0.536*)	0.727
N = 610	0.850 (*0.564*)	0.782 (*0.519*)	0.814 (*0.540*)	0.703
N = 690	0.840 (*0.548*)	0.746 (*0.487*)	0.791 (*0.516*)	0.708

Table 13. Quantitative evaluation scores obtained from "hard" frame clustering of full-utterance frames, for various values of N. Number of clusters produced is 5 for *N*=450, 4 for *N*=690, and 3 for all other values of *N*. Random baseline values are shown in italics.

Clearly, categorization was highly successful, and robust across all values of *N*, as indicated by the high values obtained. Accuracy, completeness and F were well above their random baselines, with F reaching a maximum of 0.818 at *N* = 450 against a baseline of 0.541. Bookmaker was also well above its random baseline of zero, and reached its highest value of 0.734 at N = 450 (although the confidence interval for Bookmaker at N = 450 overlapped with that at N = 520). Given that the clustering process was halted at three clusters in most cases, these results confirm the impression from the qualitative results that this entirely automatic process was indeed successful in separating the intuitive “big three” content categories of nouns, verbs and adjectives.

It is also possible to report significance, using randomization tests as described in Section 5.5.2.3. The procedure described in that section was applied to the outcomes of each clustering for each of the target values of N, to produce samples of 1000 F and 1000 Bookmaker scores. For all values of N, the actual F and Bookmaker scores obtained exceeded *all* randomly-generated values in the sample set, indicating that the empirically-obtained values were greater than would be expected if focal words had been randomly assigned to categories, at a significance level of p = 0.01 (using the lower 99% confidence limit of 0.002 for the estimated significance, as discussed in Section 5.5.2.3).

It has therefore been shown that the very simple algorithm for distributional bootstrapping of parts-of-speech that was considered in this chapter was able to discover the three main categories of noun, verb and adjective, without the use of any semantic information, or any external guidance (other than the specification of the number of large clusters to be formed), and was able to provide a highly accurate categorization of words in context based solely on the full-utterance frames in which they occurred.

6.6 Discussion

The results obtained in this chapter clearly show that it is feasible for a child to induce the parts-of-speech of English purely on the basis of the distributional co-occurrence of words and a set of the most basic utterance frames in natural child-directed speech. These frames have been discovered in a fairly straightforward way, by postulating a basic dichotomy between frequent and less-frequent words, and collecting the most prevalent and flexible full-utterance frames that can be built up out of the frequent words. It is remarkable that a simple clustering process on these frames can produce clusters of frames that correspond very closely to nouns, verbs and adjectives. As shown in Table 13, allowing these frame clusters to dictate the category to which their filler words are assigned produces a highly successful categorization of these words as measured against the reference lexical categorization of the Manchester corpus.

Nevertheless, it may be that in considering only the information from the frame context in which a focal word occurs, we are needlessly neglecting useful information from the word itself. The point of the work of Finch (1993) and others (e.g. Redington et al., 1998; Mintz et al., 2002) is that we can obtain rough parts-of-speech also by clustering together words according to the contexts in which they appear. While it is true that that work neglected the ambiguity of words by assigning a word to only one category, the current model does exactly the same for frames.

Greater correctness might be achieved by using a different solution, one which *combines* information from both the frame and the word in order to arrive at a more reliable categorization. This requires something more than merely clustering frames independently, and then words independently, because we would then have the problem of determining which word cluster (if any) corresponds to which frame cluster - in the general case, there need not be any relationship between the two sets of clusters. Attempting to combine information about both the frame and the word in an adequate way will be the focus of Chapter 7.

7 Resolving ambiguity with co-clustering

7.1 *Introduction*

The experiments reported in Chapter 6 achieved some success in assigning focal words to the categories of noun, verb and adjective. Nevertheless, some errors of categorization were made. The main reason for this may well have been that many of the full-utterance frames found are ambiguous, just as many words are. The focus of this chapter is on finding procedures for *combining* information from frames and words in order to arrive at an improved categorization of focal words in frame contexts.

7.1.1 Constraints on the representation of frame and word ambiguity

Prior to devising these methods, however, there are questions to be answered about the basic structure of an appropriate categorization model that can deal with linguistic ambiguity:

1. In what form should information about the categorical ambiguity of words and frames be expressed?
2. What implications does the categorical ambiguity of words and frames have for the categorization of individual *instances* of words in frame context?

Consideration of these questions provides us with two constraints that can be applied to all models of lexical categorization presented in this thesis.

1. The ambiguity information is expressed *independently* for words and frames. Both individual frames and individual words are regarded as potentially ambiguous with regards to their part-of-speech, and any one frame or any one word can be associated with or belong to multiple categories. This still leaves open the question of whether membership in a category is all-or-nothing, or a matter of degree, and the models considered in this chapter will explore this issue.
2. However, it is assumed that the potential ambiguity of a frame and a word (when considered in isolation) "collapses" when the two are actually used in combination in an utterance instance. Consequently, every instance of a focal word used in frame context (every cell in the co-occurrence matrix) will have only one part-of-speech assigned to the focal word. Consider the example in Table 14. The frame "That's X, isn't it?" can accept both adjectives and nouns

into its slot, and the word “mean” may function as either a verb or an adjective. However, the result of combining these two constraints is that when the frame and word come together in the utterance “That’s mean, isn’t it?”, the only possible categorization for “mean” is that it is an adjective. (Not all cases will be as clear-cut as this; for such cases, additional mechanisms will need to be employed.)

that’s X, isn’t it?	1	Noun	0	mean
	0	Verb	1	
	1	**Adjective**	**1**	

Table 14. How ambiguity may be resolved when an ambiguous frame and an ambiguous word come together in an utterance.

In other words, in the models discussed in the current work, all ambiguity will be relegated to the level of the frames separately, and the words separately, and the nature and extent of ambiguity will be explicitly characterized for each of these elements. But at the level of the co-occurrence matrix, there will be no ambiguity in a particular cell that represents the occurrence of a focal word in a frame slot. Whenever a particular frame and a particular word that occur together in a corpus utterance exhibit some degree of potential ambiguity individually, therefore, it will be necessary to resolve that ambiguity.

From a psychological point of view, the full-utterance frame model of Chapter 6 is now being extended to include both frame and word information. In that chapter, a part-of-speech was said to form out of a set of constructions (frames) which were able to accommodate very similar fillers into their variable slots. The words were regarded as features of the particular frames.

In the current extension to the full-utterance frame model, not only the frames but also the words that are used in the frames become associated with the part-of-speech. Words are presumed to be linguistic units or constructions on their own, and so are the frames. Each of these constructions may have very specific linguistic information associated with it; in addition, each node or entry in the “constructicon” may be presumed to contain information about associations with more abstract constructions, in this case each of the parts-of-speech. Membership of or association with a category would not be an exclusive relationship; because of the phenomenon of ambiguity that is the focus of this chapter, words and frames are likely to be associated with various categories at once, with varying degrees of association strength.

The current view of parts-of-speech is rather “bloodless” and formal, in that categories are assumed to be based purely on distributional information about the constructional forms that are associated with them. In order to make this model compatible with Construction Grammar approaches, it will eventually be necessary also to consider meaning. It is very likely that a large component of the substance of a part-of-speech is also tied up in the semantic implications of a category, i.e. in the “notional” criteria for category membership. In Langacker’s (1987) theory, it is the cognitive attitude taken by a speaker towards a linguistic unit (a word) that determines its category; so for instance what makes a particular word a verb is the speaker’s construal of it as a process, rather than an entity or atemporal relationship. It seems very likely that these semantic notions also become associated with the part-of-speech as it develops. Under this view, a part-of-speech is the nexus of a large amount of both distributional and semantic information that defines the category.

A very important issue to be addressed is how a simultaneous clustering of words and frames should be effected. It would be possible to cluster words on the basis of the frames in which they occur, and then to cluster frames according to their filler words, but we would then have difficulty in matching up a word cluster with its corresponding frame cluster. What is required is a *co-clustering technique* that clusters words and frames to the *same* categories.

7.1.2 Co-clustering and biclustering

The terms *co-clustering* and *biclustering* are often used in the field of genetics, where they refer to a group of data mining techniques for finding patterns in genomic data (see Madeira & Oliveira, 2004, for a review). Often, genes are expressed only under certain circumstances, captured in experimental conditions. The purpose of biclustering in this context is to find the combinations of genes and experimental conditions that together lead to a high level of activation of the gene. For this purpose, the typical data representation is a co-occurrence data matrix of genes against experimental conditions. Biclustering techniques are then applied in order to find large *biclusters* of matrix cells where certain genes intersect with certain experimental conditions so as to produce high activation values in the data matrix cells. The term biclustering refers to the fact that the purpose is to find clusters that are defined simultaneously in terms of genes and experimental conditions.

On the face of it, then, biclustering techniques would appear to be useful for the current problem, words and frames being clearly analogous to the genes and conditions of genetic data biclustering. However, few established biclustering methods adhere to the two constraints identified in the previous section. In some methods, either all genes or all conditions or both are assumed to be unambiguously associated with their clusters. In others, both genes and conditions may belong to more than one cluster, but they are allowed to form overlapping biclusters, such that individual cells in the matrix are assigned to multiple categories.

The work in this chapter investigates ways of obtaining co-clusters of frames and words together. The established methods of bi-/co-clustering are not suitable, and so other methods were devised. I present three solutions to the problem of combining frame and word information.

The first solution is concerned with turning a hard/all-or-nothing one-dimensional clustering of frames into a fuzzy/graded two-dimensional *co-clustering* of frames and words together. It may be possible to regard every word and frame as a potential member of *every* category, and to express the *degree* of membership numerically. The category membership of individual word-frame instances is then determined by combining these numbers for both the word and the frame. This is the approach taken in the "fuzzy" co-clustering algorithm of Section 7.2.

The other two solutions presented in this chapter attempt to deal with ambiguity in a discrete rather than a continuous way, by explicitly *enumerating* all the parts-of-speech in which each frame and each word may take part, and then *combining* these two separate sources of information during categorization.

If a particular frame or word is believed to potentially belong to more than one category, and it happens to be the case that one of the categories is redundant, in the sense that all of the word-frame instances that it covers can also be covered by a different category, then the redundant category can safely be discarded from the set of categories to which the frame or word belongs. This idea forms the basis of the "parsimony-based" co-clustering algorithm, presented in Section 7.3.3.

The co-occurrence of a word and a frame is a valuable piece of information, because it indicates that the word and the frame should have at least one category in common. If this is not yet the case, it indicates that there is a conflict between the ways that the categorical possibilities of the frame and of the word have been described. This conflict needs to be resolved, by allowing either the word or the frame to partake in one of the categories of the other item. This is the essence of the so-called "conflict-based" co-clustering algorithm, presented in Section 7.3.4.

7.2 Fuzzy co-clustering

In this section, I present the first of the three co-clustering algorithms, which assigns to each word and frame a graded degree of membership of each of the parts-of-speech. The degree of membership is interpreted as the *probability* that a particular item is associated with a particular part-of-speech.

7.2.1 Procedure

We attempt, for each word and for each frame, to determine a *probability distribution vector* for that item over each of the part-of-speech categories. Each cell in the vector expresses the probability that the item belongs to category 1, category 2, etc. During categorization we determine the joint probability of the word co-occurring with the frame. Because of the assumption of independence between words and frames, this can be done by simply *multiplying together*, for each category c_k, the probability that the word belongs to c_k and the probability that the frame belongs to c_k. The focal word is then assigned the category with the highest product probability.

7.2.1.1 Probability distribution vectors

The process for calculating the probability that a word or frame is associated with a particular cluster starts with the hard clustering obtained as in the previous chapter, where only frames were clustered together. Let the resulting frame clusters be referred to as c_x, where x ranges over [1, 2, …, K], and K is the number of clusters. From this hard clustering, we obtain the conditional probability P(c_k | w_j) that word w_j is associated with cluster c_k (rather than any other cluster) by considering only frames that have been allocated to c_k, adding together the frequencies with which the word in question occurred in each of those frames, and dividing by the frequency with which the word occurred overall. This conditional probability expresses the probability, for a given word w_j, that the word occurred in a frame from cluster c_k rather than any other cluster. More formally, according to Bayes' rule we have

$P(c_k \mid w_j) = \dfrac{P(c_k, w_j)}{P(w_j)}$, which can be written as

$P(c_k \mid w_j) = \dfrac{\sum_{i \in c_k} D_{ij}}{\sum_i D_{ij}}$, where D is the original data matrix.

P(c_k | w_j) can now be interpreted as a probability vector describing the probability that w_j is associated with each of the K clusters. Subsequently, we can attempt to express the probability of an association between each frame and each cluster, induced from
P(c_k | w_j). As before, Bayes' rule gives

$P(c_k \mid f_i) = \dfrac{P(c_k, f_i)}{P(f_i)}$.

Now, in order to expand this equation, we wish to determine how frequently a particular frame f_i has been used in the data matrix in combination with a focal word belonging to category c_k. If we divide that frequency by the frequency with which f_i occurs in the data matrix overall, that will give us an expression for P(c_k | f_i).

Say that a particular word w_j occurs x times in the context of frame f_i (in other words, $D_{ij} = x$). If w_j was associated unequivocally with only c_k, then it would contribute an amount of x to the total frequency of occurrence of c_k in the context of f_i. However, because w_j is split between the various clusters according to P(c_k | w_j), it follows that we also have to split the frequency x among the clusters in the proportions given by P(c_k | w_j). Doing this gives

$$P(c_k \mid f_i) = \frac{\sum_j P(c_k \mid w_j) D_{ij}}{\sum_j D_{ij}}.$$

Note that dividing the frequency in this way is not the same as actually *allocating* some of the x cases to category 1, some to category 2, etc., in contravention of our earlier stated constraint 2. We are instead dividing our *amount of certainty* between the categories. The categorization of w_j in the context of f_i will be entirely unequivocal, as will be shown later.

These equations took the initial hard frame clustering as their starting point. This starting point is clearly an asymmetric one, in that word clusters were not involved at all, and reflects the intuition that to some extent, it is the frames that are really responsible for placing a particular part-of-speech construal on a word. When distributional information is used to bootstrap parts-of-speech, as was done in the previous chapter, it should be the contexts of the words that should be amalgamated into groups or clusters that define the category, rather than the words themselves. This is because context is a far less ambiguous cue to part-of-speech than a typical English content word, which can be molded quite readily into whatever "shape" the speaker wishes: a noun in one context, and adjective in another. For this reason, the work in Chapters 6 to 8 is distinctly asymmetrical in its approach, and always starts with a hard clustering of frames.

The procedure described above can be viewed as a kind of softening of the initial hard clustering to produce a fuzzy co-clustering of words and frames - actually two separate fuzzy clusterings of words and of frames, which, crucially, make reference to the same categories.

7.2.1.2 Categorization

The above procedure yields two *allocation matrices*, which can be labeled A^W and A^F, such that $A^W{}_{kj}$ gives the probability of a given word w_j belonging to category c_k, and $A^F{}_{ki}$ gives the probability of a given frame f_i belonging to category c_k (i.e. $A^W{}_{kj} = P(c_k \mid w_j)$ and $A^F{}_{ki} = P(c_k \mid f_i)$).

This section describes how the process of categorization is carried out using A^W and A^F in combination. According to the constraints stated earlier, this categorization should assign a single category to every word in frame context.

The combination process is very simple: for frame f_i there is a corresponding column i in the A^F probability matrix, and for word w_j, there is a corresponding column j in A^W. These columns are simply multiplied together. This entails, for each category c_k, multiplying together the probability that the frame belongs to c_k and the probability that the word belongs to c_k. The category with the highest product of probabilities is the "winner", and the word-frame instance is assigned to that category[4].

Table 15 shows an example taken from the execution of this algorithm. The word "mean" is ambiguous, as it can be used as either a verb or an adjective, depending on context. The fuzzy co-clustering algorithm assigns it an approximately equal probability of being either. Hence, the frame context has to cast the deciding vote. When "mean" occurs in "What do you mean?", the frame "What do you X?" is heavily biased towards accepting verbs, and so the product of the frame and word probabilities is the highest for the verb category. In the utterance "That's mean", however, the frame heavily favours adjectives, and so the product of probabilities ends up in favour of an adjective categorization. In this way, the algorithm is able to categorize "mean" appropriately depending on its frame context.

[4] Note that previous presentations of results from this work (Leibbrandt & Powers, 2007, 2008) made use of the *sum* of word and frame probability, rather than their product. During the evaluation of the work in this chapter, reported in Section 7.4, it was found that there was no significant difference between using the sum and using the product, in terms of the resulting categorization. As the probability product is interpretable in a sensible way as a joint probability, it is the only one of the two functions that will be reported here.

	What do you X?		mean		What do you mean?		That's X		mean		That's mean
N	0.03		0.04		$1.2e^{-3}$	N	0.26		0.04		$1.04e^{-2}$
V	**0.70**	×	**0.51**	=	**$3.57e^{-1}$**	V	0.01	×	0.51	=	$5.1e^{-3}$
A	0.27		0.45		$1.22e^{-1}$	A	**0.73**		**0.45**		**$3.29e^{-1}$**

Table 15. An example of the resolution of part-of-speech for the ambiguous word "mean", using the fuzzy co-clustering approach.

7.2.2 Psychological considerations

Fuzzy co-clustering is broadly compatible with a psychological outlook that regards the membership of a particular category by a particular item to be a matter of degree rather than an all-or-nothing, yes/no affair, and this outlook is indeed compatible with standard theories of categorization (Kruschke, 1992; Rosch, 1983). Under fuzzy co-clustering, a particular word or frame is associated with a particular cluster to a certain degree. The strength with which each word is associated with a cluster exerts an influence on the strength of association of each frame with that same cluster, and vice versa.

The association strength may also quite naturally be related to activation strength. It could be postulated that the process of combining word and frame information for the purpose of categorization relies on the differing degrees to which each of the clusters is activated by the frame and the word, as follows: When a word and a frame are combined into an utterance, the hearer needs to allocate a part-of-speech to the word using both the identity of the word and of its frame context. The word *activates* (makes more readily available for further cognitive processing) each of the available parts-of-speech in accordance with the association strength from the word to the category, so that the most strongly associated category is activated the most strongly. The frame does the same. Hence, the category that is most strongly activated (by the joint activation from word and frame) is regarded as the "correct" category for the word in frame context.

The actual steps taken in this algorithm to arrive at the fuzzy word and frame memberships may, however, not necessarily be psychologically veridical. There is an initial hard clustering of all frames, followed by a "softening" of the clustering for all words, followed by another "softening" for frames, and these three steps are carried out in "batch mode" fashion, with any entire phase having to complete before the next one can commence. It seems unlikely that this is literally the process by which a child arrives at fuzzy memberships of parts-of-speech for words and frames. However, this does not preclude the existence of some psychological process which is *functionally equivalent* to the fuzzy co-clustering process described here.

Although this algorithm is simulated in batch mode and therefore does not represent an iterative model, it is nevertheless quite possible that this kind of co-clustering could be established on the basis of distributional information alone, under a description that appeals only to processes of associative learning. One possible way in which this could happen is as follows. When the child already has some knowledge of a few (content) words of her language, as well as of some of its constructions, she can begin to associate words with constructions when they co-occur. In this way, each construction will develop associative links to its slot-filling words, and each word to its co-occurring constructions. This is compatible with the first phase of category learning as outlined by Braine (1987). Once a number of words and constructions have been associated in this way, we might expect that mental activation of a word also weakly activates its associated constructions, and vice versa.

Subsequently, we need to account for the development of the ability to generalize across constructions, so that words that have not been encountered in a construction before can be regarded as acceptable by the child, in accordance with Braine's (1987) second phase of category learning. This can be done by forming clusters of constructions, clusters of words, or ideally co-clusters of words and constructions together. In order to deal with ambiguity, it is required that any word and any construction can potentially be associated with more than one cluster.

Accounting for this phase of the developmental process is slightly tricky. Possibly, when a word and construction co-occur, the word also weakly activates all the other constructions in which it has occurred, and the construction likewise activates all its other slot-filling words. All of these elements are now available to be associated with each other, allowing *second-order associations* to form between constructions that accept the same word and words that occur in the same construction. Over time, the eventual effect of these second-order associations will be that when a construction-word pair is encountered, all similar constructions and words are simultaneously weakly activated. This then allows for *third-order associations* to be formed between each of the construction-word pairs. It is these third-order associations that constitute a co-cluster.

The existence of ambiguous words and constructions poses a threat to the model sketched above, as they would seem to create the danger that eventually, all words and constructions will end up being associated with each other. However, one might argue that in practice, if an ambiguous word in a verb context activates its associated verb and noun contexts, the increase in association between the noun contexts and the current verb construction will be relatively far smaller than the total increase in association that will occur between the construction and other verb constructions over the course of development, so that these latter associations will predominate. Of course, it should be stressed that these remarks are merely speculative; implementing such an iterative model fell outside the scope of this thesis, but this would be the only way to evaluate whether the ideas outlined in the last few paragraphs are feasible.

Since we are essentially considering associative links between nodes in some representational framework, where this framework could take on any form from a subsymbolic connectionist model to a high-level network of linked mental constructs, it is possible that *the category itself* may constitute a node in this framework, and hence be available for association with words and constructions. A more satisfying account of co-clustering may then be obtained by positing that words and constructions are not associated directly with each other, but rather indirectly by way of their mutual direct associations with the category. If these associations are allowed to be bidirectional, so that a category can activate constructions and words, and vice versa, then there is no need for second- or third-order associations: words and constructions are simply associated with a category when both they and the category are simultaneously activated. When a word and construction co-occur, they each activate the various categories according to their associative strengths, and the most strongly activated category is the "winner" to which the focal word is allocated. This proposal is closer to the fuzzy co-clustering algorithm I have outlined above.

It still remains to be explained how a node corresponding to the abstract category is formed. In a neural network model, and in fact in the "real" neural networks of human associative cortex, this may well happen by accident. It is possible that, at least in the early years, specific neuronal assemblies may have haphazard synaptic connections to many other neurons and assemblies, which will be pruned during the course of development and learning. Then, if a word and construction co-occur, the sets of neurons that they activate may, purely by chance, have some overlap (cf. the case of "Wickelfeatures"; Wickelgren, 1979).

Another possibility, considered again in Section 11.2, is that the substance of a category may consist of the mental operations involved in representing the *semantic* aspects of the category.

7.3 Discrete co-clustering algorithms

In this section, I present two solutions to the problem of combining word and frame information that make use of discrete rather than graded category membership: for each category, whether a word or a frame belongs to the category is an all-or-nothing affair, and for each item, the categories to which it can belong are listed exhaustively in a binary *allocation vector* for that item. The allocation vector of an item is a 1-dimensional vector, each cell of which corresponds to a category. The cell contains a 1 if the item in question can potentially belong to that category, and a 0 otherwise. For ease of exposition in what follows, I use the term *item* to refer to either a word or a frame, and introduce the concept of *co-items* of an item, which is, for a word, the set of frames in which it has occurred as a filler, and, for a frame, the set of words that have occurred as fillers in its slot.

Both algorithms proceed in a similar fashion. In each step, one item (either a frame or a word) is selected, at random in the case of the parsimony-based algorithm (Section 7.3.3), or, in the case of the conflict-driven algorithm (Section 7.3.4), according to a heuristic which will be described later. The allocation vector for that item is updated to reflect the categories to which the item can belong, using information from the allocation vectors of its co-items in the corpus (i.e. the words that have filled the frame, if the target item is a frame, or the frames in which the word has occurred if the target item is a word). In subsequent steps, the updated allocation vector for the target item can now be used as information to update the allocation vectors of its co-items, when those co-items become the target items in turn. In this way, the allocation vectors for each of the words and frames are adjusted so as to converge onto the “correct” allocation. When no more changes can be made to the allocation vectors, the algorithm halts.

Both these algorithms start with an initial set of allocation vectors for the words and frames. This initial set consists of a number of reliable allocations, based on words and frames that unambiguously belong to only one category with high probability. These are the so-called “seed” frames and words. From these seeds, the correct categorical allocation “crystallizes out” as the algorithm proceeds. Sections 7.3.1 and 7.3.2 describe the derivation of the seed words and frames, and the data structures used in this experiment. Sections 7.3.3 and 7.3.4 provide details of, respectively, the parsimony-based and conflict-driven algorithms.

7.3.1 Seed words and frames

The first step in both discrete co-clustering processes is to derive a set of *seed words* and *seed frames*. These are words and frames that are highly prototypical of a particular category, and have a high probability of belonging to that one category only. The process for doing this is outlined below, and summarized in Box 2.

The process starts off similarly to the fuzzy co-clustering algorithm, in this case taking a purely “type-frequency-centric” approach. Starting with a hard clustering of frames, all words that occur in any of the frames belonging to a particular category c_k are sorted in descending order of the number of different frames in which they occur. This means that we are interested in finding the words with the highest frame diversity, on the assumption that they are the most prototypical words associated with that category (they most comprehensively capture what the category is “about”). Recall that this was the process by which the word lists in Table 10 were generated.

The resulting sorted histogram of distinct frame counts forms a roughly Zipfian distribution (Zipf, 1949) for most categories. We would like to take as seed words the top few words with the highest distinct frame counts; however, we need to bear in mind that the clusters are potentially highly disparate in size, so that it would not be appropriate merely to select an absolute number of words from each category.

The solution followed is therefore to obtain the sorted histogram of distinct frame counts, and to add words from left to right until the cumulative *proportion* of distinct frame counts (the proportion of the entire graph accounted for so far) is greater than or equal to a fixed proportion η (η is a parameter of this model, although I do not manipulate it in these experiments, and keep its value fixed at 0.25).

The process described above can be considered to produce similar results to a psychological process of association between clusters and words, where the strength of association between the cluster and the word is strengthened each time the word is used in a frame that is strongly associated with that cluster already. Each distinct frame is considered to contribute an equal amount of activation strength to the word, regardless of its own frequency of occurrence in the input, so that this association process is sensitive to the type frequency of frames co-occurring with the word in question, rather than to the token frequency. A wider range of co-occurring frames counts as more robust evidence that the word does indeed belong with the cluster (and most likely possesses many of the semantic attributes that are associated with the cluster).

<u>Data structures</u>:

- *data*: the original *I*x*J* data matrix, where *I* is the number of frames and *J* is the number of words in the data set. The cell *data[i][j]* contains the number of times that frame *i* and word *j* have occurred together in the corpus.
- *totals*: a two-dimensional *K*x*J* matrix, where *K* is the number of clusters. Each cell *totals[k][j]* will contain a pair <*x, f*> where *x* is the index of the word in *data*, and *f* is the frequency value, giving the total number of frames from cluster *k* that have occurred with word *x* in the corpus.
- *seedWords*: a 1x*K* vector of sets of the most reliable seed words for each of the clusters.
- *seedFrames*: a 1xK vector of sets of the most reliable seed frames for each of the clusters.
- *sum:* a whole-number value used to count the total number of word-frame instances associated with a particular frame cluster.
- *cumulativeProportion*: a floating-point value indicating the proportion of the total set of instances associated with a cluster that have been accounted for by adding instances involving the most strongly associated word each time.
- *wordIndex*: an index that keeps track of the number of words that have been added.
- *useOwn* and *useOther*, two boolean (true or false) values.
- η, a floating-point parameter (set to 0.25 in this simulation)

<u>Algorithm</u>:

```
1. First execute the algorithm of Box 1 to obtain ordered lists of word index- word frequency
   pairs for each cluster, and place the result in totals. (Recall that these lists are sorted
   from highest frequency value to lowest).
2. Let k be the index for each of the clusters. Then for each cluster do the following:
   |  a. Set sum equal to the sum of all the frequency values of all the cells in row
   |     totals[ k ].
   |  b. Divide the frequency value of every cell in totals[ k ] by sum. The
   |     frequency value of each cell totals[ k ][ j ] now contains the proportion of all
   |     word-frame instances associated with cluster k that involve word j.
   |  c. Set cumulativeProportion to 0.
   |  d. Set wordindex to 0.
   |  e. Do the following until cumulativeProportion ≥ η:
   |     |  1. Increase wordIndex by 1.
   |     |  2. Obtain the index-frequency pair <x, f> = totals[ k ][ wordIndex ].
   |     |  3. Increase cumulativeProportion by f.
   |     |  4. Add word x to the set seedWords[ k ].
   |
3. From each of the sets in seedWords, remove all words that occur in both seedWords[ a ]
   and seedWords[ b ] for some a ≠ b.
4. Let k be the index for each of the clusters. Then for each cluster do the following:
   |  a. Let i be the index for each of the frames. Then, for each frame, do the following:
   |  |  1. Let j be the index of each word. Then, for each word, do the following:
   |  |     |  a. If data[ i ] [ j ] > 0, do the following:
   |  |     |     |  1. If seedWords[ k ] contains word j, set useOwn to true.
   |  |     |     |  2. If seedWords[ m ] contains word j, where m ≠ k, set useOther to
   |  |     |     |     true.
   |  |     |
   |  |
   |  b. If useOwn is true and useOther is false, add frame i to seedFrames[ k ].
   |
```

Box 2. The algorithm for identifying the seed words and frames of each of the clusters.

The foregoing process produces a list of strongly associated words for every cluster. In order to find words that are also *distinctive* for that cluster, all words are discarded that occur in the strongly-associated-word list of more than one cluster.

This can be interpreted psychologically as a competition process between clusters for a particular word. Possibly, attending to a word automatically activates all categories with which it is associated, so that words that evoke several categories are not regarded as "pure" examples of a category (or of a certain set of semantic attributes that are closely associated with that category). Such a process might be simulated in a connectionist model by means of lateral inhibition.

At this point, a list of unambiguous and prototypical words has been produced for each of the initial parts-of-speech. From the sets of seed words, sets of seed frames can be induced: the seed frames for a particular category are those frames that accept seed words from the category in question and do not accept seed words from any other category.

7.3.2 Allocation matrices

As mentioned above, the crucial data structures for the discrete co-clustering algorithms are the *allocation vectors* for each of the words and frames, concatenated to form an *allocation matrix* A^F for frames and an allocation matrix A^W for words. These are binary-valued versions of the continuous-valued allocation matrices used in the fuzzy co-clustering algorithm. A cell $A^F{}_{ki}$ in the frame allocation matrix has the value 1 if frame f_i can potentially belong to category k, and is 0 otherwise (and the analogous situation holds for A^W). Each column in an allocation matrix represents the allocation vector for a particular frame or word, listing the categories to which the frame or word belongs.

A^F and A^W are initialized to contain only the seed frame and seed word information, i.e. for every seed word w_j that belongs to category c_k, $A^W{}_{kj}$ is set to 1, and all other cells contain 0 (and the analogous initialization is performed for frames).

7.3.3 Parsimony-driven co-clustering

The first discrete co-clustering algorithm examines, for each item, the possible category memberships of all of the item's co-items, and attempts to find the most *parsimonious*

brush	drink	eat	mean	say	sing	wear		What do you X?
1	**1**	0	0	0	0	0	**N**	**0**
1	**1**	**1**	**1**	**1**	**1**	**1**	**V**	**1**
0	0	0	**1**	0	0	0	**A**	**0**

Table 16. An example taken from the execution of the parsimony-based co-clustering algorithm. The frame *What do you X?* occurs with several ambiguous words; nevertheless, the Verb category covers all of these, and hence the algorithm treats the frame as unambiguous.

are you going to X?	can I have a X?	that's a nice X	have you got a X?	shall I X it?	did you X it?	are you going to X them ?		kiss
1	**1**	**1**	**1**	0	0	0	**N**	**1**
1	0	0	0	**1**	**1**	**1**	**V**	**1**
0	0	0	0	0	0	0	**A**	**0**

Table 17. An example taken from the execution of the parsimony-based co-clustering algorithm. The word *kiss* occurs in a set of frames that can only be described as Noun frames, and a *different* set of frames that can only be described as Verb frames, and hence the algorithm concludes that *kiss* is ambiguous, and can function as either a noun or a verb.

explanation for the pattern of data in order to decide to which category/ies the item belongs.

7.3.3.1 Procedure

The parsimony-driven process is a stochastic algorithm. An item is selected at random, and an attempt is made to update the allocation matrix for that element. The algorithm inspects the allocation vectors for all co-items of the item, in order to find the most parsimonious or economic explanation for their occurrence with the target item.

Consider the example in Table 16, taken from the actual execution of this algorithm. The table shows the allocation vectors for all words that occur as fillers in the frame “What do you X?”, as they are at this particular point in the execution of the algorithm. Some of these words (“say”, “sing”, “wear”) are always used as verbs only and so have received allocations for only that category in the allocation matrix, whereas others such as “brush” and “drink” are described as ambiguous, i.e. they can function as either a noun or a verb. Likewise, “mean” can function as a verb or an adjective. On the basis of this evidence, however, it can be seen that the category Verb covers all of the instances of the frame “What do you X?”, as every word that occurs in “What do you X?” has received the potential allocation Verb. Hence, the most parsimonious explanation for the data in Table 16 is that “What do you X?” is an unambiguous frame context for verbs, and so the frame receives the allocation 1 for the category Verb, and 0 for the categories Noun and Adjective. This updated allocation vector for “What do you X?” can now be used as a piece of evidence for determining the category allocations of each of its filler words, the next time they are randomly selected to be the targets for updating.

By contrast, Table 17 shows an example where the algorithm determines that a word is ambiguous. The word “kiss” occurs in a number of different contexts, some of which are clearly noun contexts (according to the evidence available to the algorithm at that time), while others are unambiguous Verb contexts. However, there is no one category that covers all of the frames in which “kiss” occurs; hence the only explanation for the data in the table is that “kiss” is ambiguous and can be either a verb or a noun, and so the allocation vector for “kiss” is updated to reflect this. (Note that the ambiguous frame

"Are you going to X?" provides no information to decide between Nouns and Verbs, and hence plays no role in the allocation of categories to "kiss").

In more detail: once an item has been selected, we collect the allocation vectors for all of its co-items, and then, for each category c_k, count the number of co-items that have a 1 in their allocation vector for category c_k. The category that has the highest count is the winning category from that round, and the target item receives an allocation to that category. The co-items that were covered by that category (i.e. that had a 1 in their allocation vectors for that category) are removed from the pool, and the process is repeated on the remaining co-items, until none remain. At this point, all co-items have been covered by a category in the target item's new allocation vector.

	CoItem$_1$	CoItem$_2$	CoItem$_3$	CoItem$_4$	CoItem$_5$	CoItem$_6$	Item
C$_1$	**1**	0	0	**1**	0	0	**1**
C$_2$	0	0	0	0	0	0	0
C$_3$	**1**	**1**	**1**	0	0	0	**1**

Table 18. Example to illustrate parsimony-based allocation updating, part 1.

	CoItem$_1$	CoItem$_2$	CoItem$_3$	CoItem$_4$	CoItem$_5$	CoItem$_6$	Item
C$_1$	**1**	0	0	**1**	0	0	0
C$_2$	0	0	0	0	0	0	0
C$_3$	**1**	**1**	**1**	**1**	0	0	**1**

Table 19. Example to illustrate parsimony-based allocation updating, part 2.

	CoItem$_1$	CoItem$_2$	CoItem$_3$	CoItem$_4$	CoItem$_5$	CoItem$_6$	Item
C$_1$	**1**	0	0	**1**	0	0	0
C$_2$	0	0	0	0	0	**1**	**1**
C$_3$	**1**	**1**	**1**	**1**	0	0	**1**

Table 20. Example to illustrate parsimony-based allocation updating, part 3.

It may happen that more than one category accounts for the maximum number of category assignments. In this case, we simply allocate all of these categories to the item, and remove all co-items that have any of these categories allocated in their allocation vectors.

This process can delete previous allocations to a particular category, as well as adding new ones. Suppose that the situation for Item is as depicted in Table 18. Category C_1 obtains a count of 2, and C_3 a count of 3, so that C_3 is the winner. When the 3 cases covered by C_3 are removed, there is still one case ($CoItem_4$) that is not covered by category C_3, and so C_3 cannot be the only category allocated to Item. The counting process is repeated with the remaining co-items; only $CoItem_4$ remains, and it can only belong to C_1. Hence, C_1 is added to the allocation vector of the item. No items remain, and the algorithm step is complete. Now suppose that at some later stage $CoItem_4$ became the target item, and its allocation vector was updated to include an allocation to the C_3 category. Now, when Item again becomes the target item (Table 19), the category C_3 covers all of its co-items including $CoItem_4$, and so Item's C_1 allocation is deleted. If, later still, $CoItem_6$ has been allocated to the category C_2, then when Item once again becomes the target item (Table 20), the category C_2 will need to be added to Item's allocation vector in order to cover $CoItem_6$.

In this way, the category profile of an item is progressively elaborated, based on all the possible categories that its co-items may take on. As the algorithm progresses, more and more evidence is accumulated, so that category allocations can "wink in and out of existence" based on the state of information at each time. In each case, the algorithm looks for the simplest explanation for an item's category profile that will cover all co-items that have occurred with it. An item may be selected as the target more than once during the execution of the algorithm, and its allocations may change if the information from its co-items has changed.

In fact, the actual allocation value is not a binary value of 0 or 1, but a floating-point value between 0 and 1. The size of the number is equal to the proportion of allocated co-items that are covered by a particular category. The algorithm makes use of a threshold θ. If an item's allocation with a particular category has a value below θ, the allocation will be

Data structures:

- *D* : the frame-word co-occurrence data matrix, with dimensions $M \times N$
- F_{target} : the target frame, with dimensions $1 \times K$, where *K* is the number of categories
- *W* : the set of words that have occurred in F_{target} according to *D*
- *allocationSum* : a $1 \times K$ integer vector
- *max* : an integer
- *winners* : a set of categories
- *explained* : the set of words that are covered (explained) as belonging to the categories in *winners*
- *frameAllocation* : an $M \times K$ floating-point matrix
- *wordAllocation* : an $N \times K$ floating-point matrix
- θ : a floating-point threshold value
- η : a floating-point value used to test for convergence

Algorithm:

```
1. Initialize frameAllocation and wordAllocation so that wordAllocation[ k ][ j ] = 1 if word $w_j$
   is a seed word, 0 otherwise, and frameAllocation[ k ][ i ] = 1 if frame $f_i$ is a seed frame, 0
   otherwise.
2. Repeat the following until convergence (i.e. until the proportion of cells changed in both
   frameAllocation or wordAllocation in the last 100 iterations falls below $\eta$):
     a. Select a frame or word at random as the target item. (The rest of the
        description assumes that the target item is a frame. If it is a word, replace
        "frames" with "words" and vice versa in what follows.) Set $F_{target}$ = the
        target frame.
     b. Set W = the set of words $w_j$ that have occurred with the target frame $F_{target}$
        in the input (so that D[ target ][ j ] > 0), and that have wordAllocation[ k ][ j ]
        > $\theta$ for at least one category $c_k$. (i.e. ignore words that have not been
        allocated to any categories yet).
     c. Repeat the following until there are no words left in W:
          1. For each category k, count the total number of words in W that
             could potentially belong to that category, i.e. the number of words
             $w_j$ in W such that wordAllocation[ k ][ j ] > $\theta$. Set allocationSum[k] to
             this total. Set max = the maximum value in allocationSum after all
             categories have been counted.
          2. Set winners = the set of all category/ies that have attained the
             maximum score max in allocationSum.
          3. For each category k in winners:
               a. Remove from W all words $w_j$ such that wordAllocation[ k ][ j ]
                  > $\theta$, and add these words to explained. (These words have
                  now been covered by at least one category in winners).
               b. Set frameAllocation[ k ][target] = (size of explained) / K.
     d. Normalize the column frameAllocation[target], by dividing each cell in
        frameAllocation[target] by the sum of all cells in that column.
```

Box 3. Parsimony-driven discrete co-clustering algorithm (update phase).

treated as zero. Say for instance that the word *mean* is allocated to Verbs with a strength of 0.75, to Adjectives with a strength of 0.22, and to Nouns with 0.03. For θ = 0.05, *mean* does not contribute its supposed Noun character to any frame chosen as a target. Only if the value of the Noun allocation were to rise over 0.05 would there have been enough evidence accumulated to accept that *mean* might be a noun. Full details of the parsimony-based algorithm are shown in Box 3.

7.3.3.2 Categorization

During the update phase, allocations take on graded values, as described above, and represent something like the degree of evidence that a word or frame belongs to a category. During categorization, however, the allocation vector for a word or frame is effectively binarized, with any value over θ being treated as adequate evidence that the item belongs to the category in question. Categorization of an instance of a frame-focal word pair is done by intersecting the allocation vectors of the frame and word (*after* setting all values below θ to zero), and choosing the remaining category. For instance, in Table 21, the utterance "not happy" is broken up into its frame "not X" and filler "happy", and the two allocation vectors are combined (after thresholding against θ). Although "not X" is an ambiguous frame, the only category that the word and frame have in common is Adjective, and so that is the category to which the focal word "happy" in "not happy" is finally assigned.

not X		happy
1	N	0
0	V	0
1	A	1

Table 21. The categorization of the utterance "not happy" in the parsimony-based co-clustering approach.

In the event of a tie, we fall back on the categorization method of fuzzy co-clustering, and multiply the allocation values for each category, then choose the category with the highest product. In the current case there would be less justification for treating the allocation values as probabilities, as was the case with fuzzy co-clustering; they are more validly interpreted as merely an expression of the strength of association between a frame or word and a category.

7.3.3.3 Psychological considerations

The parsimony-based algorithm instantiates a human tendency towards conservatism in the assumptions that are made: allocations are not made to a particular item, unless it is absolutely necessary to do so. The approach can thus also be seen as an evidence-based form of reasoning. A particular frame, say, which has accepted a number of filler words that are unambiguously nouns, will be taken to be a noun frame. It will not also be regarded as a verb frame unless there is insurmountable evidence that this is the case; this evidence would come from instances of the frame occurring with fillers that are unambiguously verbs.

The algorithm also has the power to reverse earlier decisions: if, in retrospect, a frame which had been thought to be ambiguous can now be seen to account for all of its co-occurring words with only one category allocation, it will receive only that allocation.

It is also possible, if an ambiguous word occurs in the frame and the frame has no allocations in common with the word, for the word to lend all of its allocations to the frame. These cases may be expected to be relatively rare.

7.3.4 Conflict-driven co-clustering

This section presents the third of the co-clustering algorithms investigated in this chapter. This *conflict-driven co-clustering* algorithm attempts to “resolve conflicts” between incompatible words and frames, in order to account for the word-frame data from the corpus.

7.3.4.1 Procedure

In conflict-driven co-clustering, we deal with allocation vectors of the same format as in the parsimony-driven process: an allocation vector is a 1-dimensional vector with *K* cells, one for each category that a word or a frame may belong to. In the update phase, these vectors are treated as if their values were discrete binary variables that can be on or off only, rather than the floating-point numeric values that they were in the parsimony-driven algorithm. During categorization, however, they are again converted into floating-point vectors; this is done purely to allow resolution of ties, in the same way as during categorization for the parsimony-driven process.

If we are given a certain discrete allocation of frames to categories, and a similar allocation for words, then it may happen that there are word-frame instances in the corpus such that the word and the frame allocations are inconsistent, i.e. that there is a *conflict* in their allocation vectors. We can define a conflict to exist for a word-frame pair when there is no single category such that both the word and the frame can be assigned to it. In other words, we cannot say to which category the word in context belongs, because there are no candidates that are acceptable to both the word and the frame.

	brush	Shall I X it?	brush	There's your X	brush	Don't X it
N	0	0	**0**	**1**	1	0
V	**0**	**1**	**1**	**0**	1	1
A	0	0	0	0	0	0

Table 22. Examples of potential conflicts in category assignment between words and their frames. Conflicts are in bold.

Conflicts therefore arise when the binary intersection (the AND) between the frame allocation vector and the word allocation vector is zero. This includes the case where either the word or the frame has no categories assigned to it yet (but not when *both* are as yet unassigned).

The conflict-driven co-clustering algorithm attempts to find a conflict-free allocation of categories to words and frames. It does so by repeatedly removing the largest existing conflict until no conflicts remain. Conflicts between items and their co-items are removed by simply allocating those additional categories to items that they would need in order to no longer be in conflict with the co-items. Conflicts are not resolved in random order; instead, the conflict resolution option that has the most evidence in its favour is chosen at every step.

Examples of conflicts are shown in Table 22. In the left-hand figure, “brush” has not yet been allocated to any category, and so when it is encountered in the context of the frame “Shall I X it?”, which has been allocated to the category Verb, a conflict arises. The conflict can be resolved by allocating the category Verb to “brush”. Suppose that this has in fact been done. Then, when the child later encounters the utterance “There’s your brush”, another conflict occurs, this time because “There’s your X” has so far been allocated to Noun only, while “brush” has been allocated to Verb only, a situation depicted in the middle figure of Table 22. The two ways to resolve this conflict are to assign an additional allocation of Noun to “brush”, or an additional allocation of Verb to “There’s your X”. The way in which the conflict will be resolved will be determined by the number of other words and frames placing pressure on “brush” and “There’s your X” to receive allocations as a result of other conflicts. Suppose that in this case, the resolution chosen is to allocate Noun to “brush”, so that it is now represented as a potentially ambiguous word (having both Noun and Verb as potential categories). Now, when the utterance “Don’t brush it” is encountered (as depicted in the right-hand figure), no conflict is experienced, because “Don’t X it” has the potential allocation of Verb, which is compatible with the potential Verb allocation of “brush”.

The algorithm works in batch mode, considering the entire set of relevant data at once. For every item (whether word or frame), the set of co-items that are currently in conflict with the item is collected. Using the current allocation vectors for each of the conflicting co-items, the algorithm allows each co-item to cast a vote for every category to which it is currently allocated (i.e. co-items cast votes to have particular categories added to the item’s allocations). Per definition, these are categories that the target item does not have in its allocation vector, so that adding that allocation to the item’s allocation vector would resolve the conflict between the item and that particular co-item; however, the point of voting is to find the single change that would result in the *largest number* of conflict resolutions at once. The number of votes for each category is determined in this way for every target item (every word and every frame). The category allocation that has received the largest number of votes is designated the "winner", and the category in

Data structures:

- D : the frame-word co-occurrence data matrix, with dimensions $M \times N$
- *frameAllocation* : an $M \times K$ binary matrix
- *wordAllocation* : an $N \times K$ binary matrix
- *conflict*: an $M \times N$ binary matrix, whose cells will be set to 1 when the frame corresponding to the row and the word corresponding to the column are in conflict
- *frameVotes*: an $M \times K$ integer matrix used to tally the votes that each frame receives to add a category to its entry in *frameAllocation* (votes are cast by words that are in conflict with the frame)
- *wordVotes*: an $N \times K$ integer matrix used to tally the votes that each word receives to add a category to its entry in *wordAllocation* (votes are cast by frames that are in conflict with the word)
- η : a floating-point value used to test for convergence

Algorithm:

```
1. Initialize frameAllocation and wordAllocation so that wordAllocation [ k ] [ j ] = 1 if
   word $w_j$ is a seed word, 0 otherwise, and frameAllocation[ k ] [ i ] = 1 if frame $f_i$ is a
   seed frame, 0 otherwise.
2. Repeat the following until convergence (i.e. until the proportion of cells changed in both
   frameAllocation or wordAllocation in the last 100 iterations falls below $\eta$):
      a. For each word $w_j$ and frame $f_i$ which occurred together as an instance in
         the data matrix D, set conflict[ i ][ j ] to 1 if and only if $w_j$ and $f_i$ are in
         conflict. Conflict is detected as:
         Step through all the categories;
         If there is any category $c_k$ such that frameAllocation[ k ][i ] = 1
             and wordAllocation[ k ][ j ] = 1, then conflict[ i ][ j ] is 0.
         If no such category has been found, then conflict[ i ][ j ]is 1.

      b. Now tally the votes for category additions in frameVotes:
           For each frame $f_i$:
                For each category $c_k$ that $f_i$ has not been allocated to (i.e.
                    frameAllocation[ k ][ i ] = 0):
                  Move through all the words $w_j$ such that conflict[ i ][ j ] = 1.
                  Add up the number of words that vote for category k to be
                      added to frameAllocation (by having
                      wordAllocation[ k ][ j ] = 1).
                  Set frameVotes[ k ][ i ] equal to this sum.

      c. For each word $w_j$, do the same to fill up the wordVotes matrix.
      d. Now find the largest sum in either frameVotes or wordVotes. This is the
         largest current conflict.
      e. Resolve the largest current conflict, by assigning the allocation in question
         (i.e. adding the category in question to the allocation matrix for either the
         frame or the word).
```

Box 4. The conflict-resolution discrete co-clustering algorithm.

question is added to the allocation vector of the item in question. The co-clustering algorithm is detailed in Box 4.

The algorithm always chooses as its next step the option that offers the highest immediate gain. It can therefore be regarded as a *hill-climbing* algorithm, and may be prone to a well-known vulnerability of such algorithms, namely converging on a *locally* optimal solution rather than the *globally* optimal solution. One solution to this problem is to start the solution search from a variety of different starting positions. Note that in the current experiment, this is not done; the algorithm starts with the same seed frames and words as used for the parsimony-driven process. Informal prior experimentation showed that the quality of the seed allocation is vital to the success of the algorithm. Unlike many other algorithms such as gradient descent learning, which can usually start from almost any position in solution space and converge onto a reasonably satisfactory solution, the current algorithm adheres to the computer science maxim of "garbage in, garbage out", and delivers poor categorization results from poor starting positions. Also note that, given a particular set of seed words and frames, this algorithm is entirely deterministic, unlike the parsimony-driven process.

One of the benefits of the voting system is that it is self-correcting. Suppose that at some stage during the execution of the algorithm, for whatever reason, the allocation vector for the frame "the X" incorrectly states that the frame cannot be a noun frame, but can potentially take verbs as fillers. There are a great many words that can go into "the X", and each of them receives an incorrect vote from "the X" to add the verb category allocation to them. However, most specific instances of words occurring in "the X" are nouns, and so we can be reasonably confident in expecting that most of these words will also occur in other noun frames. As a result of having occurred in these frames, the words may be expected to have received the noun allocation at some time (also recall that the process of discovering seed words identifies a number of words that are nearly always nouns, and assigns them this allocation, and no other, during initialization). Therefore, most of these words may be expected in their turn to cast votes to assign the noun category to "the X". Because there are many of these words, and only one "rogue" frame "the X" that has been misallocated, the incorrect verb votes from "the X" are dispersed over all the words, but the correct noun votes from the words are concentrated on “the X”. Hence, adding the noun category to “the X” will receive many more votes than adding the verb category to any of the words, and so is more likely to be the step eventually taken.

Like the parsimony-based algorithm in Section 7.3.3, the conflict-driven algorithm could be said to search for a parsimonious explanation for the observed data, by adding only the minimum number of allocations required to attain “harmony”. It allocates the categories that provide the greatest pressure for allocation first, thereby avoiding making any allocations lower in the order that may turn out to be unnecessary in the long run; consequently, it does not “proliferate hypotheses unnecessarily”. Another way of putting this is to observe that the algorithm strictly respects the *quantity* of evidence in shaping its “beliefs” about the categorial possibilities of the frames and words.

7.3.4.2 Categorization

The categorization of individual word-frame instances is mostly carried out in the same way as for the parsimony-based algorithm (see Section 7.3.3.2): if there is only one unique category that is “acceptable” to both word and frame, then the word-frame instance is allocated to that category.

In the case when there is more than one acceptable category, however, the algorithm falls back on information about the *amount of support* for each category. This is calculated from the co-items of an item. The algorithm examines the (discrete) allocation vectors of each of the co-items, and determines what proportion of all the “on” cells (allocated categories) for all co-items combined belongs to category 1, 2, 3, etc. This co-item profile vector is then used as a (continuous-valued) allocation vector for the item itself. This is done for both the word and the frame. The two allocation vectors are then combined as in the fuzzy co-clustering approach, by multiplying together the word and frame allocation values for each individual category, then picking the category that yields the largest product.

7.3.4.3 Psychological considerations

The conflict-based algorithm captures the idea that encountering a frame and a filler word together, where the two items are not though to have any categories in common, gives rise to a kind of "conceptual unease" or perhaps "curiosity", which needs to be resolved by making the two sets of allocations compatible with each other (i.e. by allocating a category from either the frame or the word to the other item).

The order in which conflicts are tackled is crucial in this algorithm. The allocation that would solve the largest number of conflicts at once is tackled first. This reflects the idea that inconsistencies are not reacted to as they occur; instead, evidence for a conflict resolution accumulates, and when enough evidence has been gathered, the resolution is taken. This kind of process could be readily simulated in an iterative version of this algorithm: instead of choosing the conflict with the largest number of votes first, we would resolve any conflict with more than a certain threshold number of votes.

As was the case with the parsimony-based algorithm, the conflict-based algorithm instantiates the idea that human cognition is conservative, and will only add category assignments to items if there is sufficient evidence for doing so.

7.4 Evaluation

Each of the three co-clustering algorithms outlined in the previous sections was implemented and applied to each of the data sets used in the hard frame clustering experiments of Chapter 6, and the resulting categorizations of word-frame instances were evaluated against the gold standard categorization, using the same evaluation measures as in that chapter. This section presents the results of this evaluation process. Results are presented in a compact tabular form from Table 23 to Table 32, allowing different algorithms to be compared side-by-side. The following paragraphs describe the column headers for each algorithm that was evaluated.

All the algorithms started from the hard clustering of frames into all-or-nothing categories. The same range of key values of the parameter N used in Chapter 6 was also explored here. The results from Chapter 6 for categorization using the hard clustering are repeated for comparison, under the column heading **Hard F**.

Starting from the hard clustering, the probability matrices P(W|C) and P(F|C) were obtained, as described in Section 7.2.1.1, for the purpose of performing fuzzy co-clustering. These matrices described the probability that each word and each frame belonged to each category. The fuzzy co-clustering process was then carried out using these two matrices, by allocating to each word-frame instance the category that attained the highest product of word and frame probability. This algorithm is represented in the results tables by the heading **Fuzzy F×W**.

In addition, it may be of interest to consider using only the fuzzy word probability matrix P(W|C), or only the fuzzy frame matrix P(F|C) to categorize with. This is because, if the fuzzy co-clustering performs better in the evaluation than the hard clustering, it will still be an open question whether this is due to the "fuzziness" of the category allocations (acknowledging word and frame ambiguity), or due to the use of co-clustering (combining word and frame information). For this reason, two additional categorization algorithms were used: one that categorized a word-frame instance as belonging to the most probable category for the frame in question, according to the fuzzy probability matrix P(F|C), and another which categorized the instance in an analogous way according to the most probable category for the word, using P(W|C). These two algorithms are represented in the results tables under the column headings **Fuzzy F** and **Fuzzy W**, respectively.

The hard frame clustering was also used to obtain a set of seed frames and a set of seed words, as described in Section 7.3.1. Starting from these seed sets, the two discrete co-clustering algorithms were executed. The conflict-resolution algorithm was executed until convergence for each data set, and the results appear under the heading **Confl**. The parsimony-driven algorithm, being nondeterministic, was executed in a series of 50 simulations for each data set, and the average evaluation measures from each set of 50 is displayed in the column **Pars**.

Table 23 to Table 32 display the results for each algorithm under the various starting values for *N*. Bolded entries indicate the best performing F and Bookmaker scores (the scores that exceed their random baseline values by the greatest margin).

The obtained F and Bookmaker scores were compared against their baseline values using the randomization method described in Section 5.5.2.3, for $N = 290$ only. All F and Bookmaker scores were higher than any scores that were produced in the 1000-item sample sets, so that all scores for all algorithms attained estimated significance at a level $p < 0.001$. In order to be conservative, we take these results to indicate actual significance only at p = 0.01, as before.

A number of broad trends can be described. First and foremost is the observation that *all* the co-clustering techniques improved categorization over the hard frame clustering, for all values of N. This shows that these more sophisticated clustering techniques are efficacious in providing a more accurate categorization.

Roughly speaking, the extreme ends of the range of N values produced the poorest results, and the middle values the best, with measures peaking at N = 450. However, the range of variation is not extreme, and the algorithms are robustly successful at all values of N considered in this experiment.

The conflict-driven algorithm Confl performs better than the fuzzy algorithms and the parsimony-based algorithm for values of N between 150 and 240, but Fuzzy F × W outstrips Confl and Pars for values of N from 290 up.

In all cases, the fuzzy combination of frame and word information Fuzzy F × W performed better than fuzzy clustering using only word or frame information (Fuzzy W and Fuzzy F respectively).

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.797 (*0.480*)	0.811 (*0.480*)	0.848 (*0.480*)	0.863 (*0.480*)	0.864 (*0.480*)	0.856 (*0.480*)
Completeness	0.647 (*0.389*)	0.664 (*0.392*)	0.743 (*0.420*)	0.749 (*0.416*)	0.758 (*0.421*)	0.740 (*0.415*)
F score	0.714 (*0.430*)	0.730 (*0.432*)	0.792 (*0.448*)	0.802 (*0.446*)	**0.808** (*0.448*)	0.794 (*0.445*)
Bookmaker	0.694	0.714	0.778	0.791	**0.792**	0.785

Table 23. Unsupervised evaluation scores for full-utterance frames, *N*=80, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.818 (*0.528*)	0.839 (*0.528*)	0.870 (*0.528*)	0.881 (*0.528*)	0.885 (*0.528*)	0.875 (*0.528*)
Completeness	0.639 (*0.412*)	0.695 (*0.437*)	0.748 (*0.454*)	0.758 (*0.454*)	0.889 (*0.530*)	0.785 (*0.474*)
F score	0.718 (*0.463*)	0.760 (*0.478*)	0.804 (*0.488*)	0.815 (*0.488*)	**0.887** (*0.529*)	0.828 (*0.499*)
Bookmaker	0.695	0.724	0.776	0.792	**0.808**	0.788

Table 24. Unsupervised evaluation scores for full-utterance frames, N=150, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.837 (*0.535*)	0.841 (*0. 535*)	0.873 (*0. 535*)	0.885 (*0. 535*)	0.893 (*0. 535*)	0.882 (*0. 535*)
Completeness	0.690 (*0.440*)	0.720 (*0.457*)	0.763 (*0.467*)	0.784 (*0.473*)	0.888 (*0.532*)	0.818 (*0.496*)
F score	0.756 (*0.483*)	0.775 (*0.493*)	0.815 (*0.499*)	0.831 (*0.502*)	**0.891** (*0.533*)	0.849 (*0.514*)
Bookmaker	0.709	0.715	0.774	0.792	**0.816**	0.793

Table 25. Unsupervised evaluation scores for full-utterance frames, *N*=180, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.827 (*0.555*)	0.835 (*0. 555*)	0.871 (*0. 555*)	0.878 (*0. 555*)	0.896 (*0. 555*)	0.880 (*0.555*)
Completeness	0.713 (*0.478*)	0.731 (*0.485*)	0.790 (*0.503*)	0.802 (*0.507*)	0.893 (*0.553*)	0.833 (*0.525*)
F score	0.766 (*0.514*)	0.780 (*0.518*)	0.829 (*0.528*)	0.838 (*0.530*)	**0.895** (*0.554*)	0.856 (*0.540*)
Bookmaker	0.693	0.709	0.775	0.786	**0.817**	0.788

Table 26. Unsupervised evaluation scores for full-utterance frames, *N*=240, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.844 (*0.559*)	0.846 (*0.559*)	0.894 (*0.559*)	0.900 (*0.559*)	0.888 (*0.559*)	0.895 (*0.559*)
Completeness	0.774 (*0.513*)	0.799 (*0.528*)	0.865 (*0.541*)	0.886 (*0.551*)	0.911 (*0.574*)	0.876 (*0.548*)
F score	0.808 (*0.535*)	0.822 (*0.543*)	0.879 (*0.550*)	**0.893** (*0.555*)	0.899 (*0.566*)	0.885 (*0.553*)
Bookmaker	0.708	0.715	0.803	**0.814**	0.800	0.804

Table 27. Unsupervised evaluation scores for full-utterance frames, *N*=290, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.863 (*0.571*)	0.862 (*0.571*)	0.900 (*0.571*)	0.910 (*0.571*)	0.896 (*0.571*)	0.896 (*0.571*)
Completeness	0.764 (*0.506*)	0.808 (*0.535*)	0.871 (*0.553*)	0.894 (*0.561*)	0.912 (*0.582*)	0.875 (*0.558*)
F score	0.810 (*0.536*)	0.834 (*0.552*)	0.885 (*0.562*)	**0.902** (*0.566*)	0.904 (*0.576*)	0.886 (*0.564*)
Bookmaker	0.716	0.725	0.803	**0.822**	0.801	0.797

Table 28. Unsupervised evaluation scores for full-utterance frames, *N*=410, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.871 (*0.576*)	0.873 (*0.576*)	0.904 (*0.576*)	0.917 (*0.576*)	0. 900 (*0.576*)	0. 902 (*0.576*)
Completeness	0.771 (*0.510*)	0.798 (*0.526*)	0.865 (*0.551*)	0.893 (*0.561*)	0. 915 (*0. 586*)	0. 878 (*0. 561*)
F score	0.818 (*0.541*)	0.834 (*0.550*)	0.884 (*0.563*)	**0.905** (*0.568*)	0. 908 (*0. 581*)	0. 890 (*0. 568*)
Bookmaker	0.734	0.743	0.805	**0.833**	0.809	0.807

Table 29. Unsupervised evaluation scores for full-utterance frames, *N*=450, Full-utterance Frames, 5 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.870 (*0.575*)	0.871 (*0. 575*)	0.906 (*0. 575*)	0.916 (*0. 575*)	0.902 (*0. 575*)	0.905 (*0. 575*)
Completeness	0.762 (*0.503*)	0.802 (*0.529*)	0.866 (*0.549*)	0.882 (*0.553*)	0.914 (*0.582*)	0.872 (*0.554*)
F score	0.812 (*0.536*)	0.835 (*0.551*)	0.885 (*0.562*)	**0.899** (*0.564*)	0.908 (*0.578*)	0.888 (*0.564*)
Bookmaker	0.727	0.738	0.810	**0.826**	0.811	0.809

Table 30. Unsupervised evaluation scores for full-utterance frames, *N*=520, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.850 (*0.564*)	0.853 (*0. 564*)	0.896 (*0. 564*)	0.900 (*0. 564*)	0.895 (*0. 564*)	0.896 (*0. 564*)
Completeness	0.782 (*0.519*)	0.800 (*0.528*)	0.880 (*0.554*)	0.893 (*0.560*)	0.907 (*0.571*)	0.876 (*0.552*)
F score	0.814 (*0.540*)	0.826 (*0.546*)	0.888 (*0.559*)	**0.896** (*0.562*)	0.901 (*0.568*)	0.886 (*0.558*)
Bookmaker	0.703	0.714	0.797	**0.807**	0.804	0.798

Table 31. Unsupervised evaluation scores for full-utterance frames, *N*=610, Full-utterance Frames, 3 clusters.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.840 (*0.548*)	0.841 (*0. 548*)	0.893 (*0. 548*)	0.896 (*0. 548*)	0.889 (*0. 548*)	0.893 (*0. 548*)
Completeness	0.746 (*0.487*)	0.773 (*0.504*)	0.855 (*0.525*)	0.867 (*0.530*)	0.890 (*0.549*)	0.870 (*0.534*)
F score	0.791 (*0.516*)	0.806 (*0.525*)	0.874 (*0.536*)	**0.881** (*0.539*)	0.889 (*0.549*)	0.881 (*0.541*)
Bookmaker	0.708	0.719	0.801	**0.811**	0.799	0.804

Table 32. Unsupervised evaluation scores for full-utterance frames, N=670, Full-utterance Frames, 4 clusters.

Table 33 and Table 34 show the significance of the differences in F and Bookmaker scores (respectively) between different pairs of algorithms, for *N* = 290 only. The tables do not show all possible comparisons, but only a few that are of interest. The three main co-clustering algorithms (Fuzzy F × W, Pars. and Confl) are compared against each other, and against the simple hard clustering algorithm of Chapter 6. Fuzzy W and Fuzzy F are compared against Fuzzy F × W only. The arrows displayed in the significant cells point in the direction of the better-performing algorithm (an up arrow indicates the algorithm represented by the column, and a left arrow the algorithm represented by the row). For instance, the top lefthand cell in Table 33 indicates that Fuzzy F × W had a significantly higher F score than Hard F, with significance at < 0.001, or fewer than 1 score in 1000 (taken conservatively as indicating significance of p = 0.01 only).

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← 0.001**	0.042
Confl.			0.063
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 33. Significance levels of differences in F scores for various clustering algorithms, for full-utterance frames, N=290.
*** significant at p=0.05, ** significant at p = 0.01.**

	Fuzzy FxW	Confl.	Pars.
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← 0.025*	0.098
Confl.			0.205
Fuzzy F	↑ < 0.001**		
Fuzzy W	0.064		

Table 34. Significance levels of differences in Bookmaker scores for various clustering algorithms, for full-utterance frames, N=290.
*** significant at p=0.05, ** significant at p = 0.01.**

The results confirm that all algorithms perform better than Hard F, and that Fuzzy F × W performs better than Confl. There is no significant difference between Pars and Fuzzy F × W, or between Pars and Confl. Fuzzy F × W outperforms Fuzzy F, but is only superior to Fuzzy W in terms of its F score, not its Bookmaker score.

The small magnitude of the performance advantage of Fuzzy F × W over Fuzzy W was surprising. It is possible that the higher rate of success for Fuzzy F × W over the other algorithms might mainly be due to the "fuzzification" of the hard clustering, rather than to the fact that two sources of information are used for the purposes of categorizing instances. Nevertheless, combining word and frame information is an inextricable part of the conflict-resolution algorithm Confl, which performed better than any of the fuzzy algorithms for lower values of N.

7.5 Other issues

7.5.1 The robustness of the three main categories in the co-clustering approaches

One of the most important issues in the clustering approach taken throughout this thesis is determining the number of clusters that should be formed. I have simply evaluated each of the approaches on the lowest cluster total that produces three sizeable clusters. However, this is obviously a constraint that is imposed from the outside. One possibility is that this is indeed what happens, and that the constraint comes from semantics; the semantic categories corresponding to things, actions and properties of things may be so salient that they provide an anchor for the distributional category induction process.

Another possibility is that the categories can in fact self-organize in the clustering process on the basis of purely distributional information. The co-clustering algorithms may be taken to be essentially conservative in nature, as they require that both the word and frame information should point to the same category before they assign a word-frame instance to that category. It may be possible, therefore, that the co-clustering algorithms *automatically* identify only the most important categories and ignore the minor categories. In this case, any of the algorithms might start with the clustering results for, say, 20 cluster categories, but allocate words and frames to only a small number (ideally 3) of these categories.

In order to examine this question, the three co-clustering algorithms were run on the results from the hard clustering, as before, while the number of clusters *K* produced by the initial hard clustering was systematically varied from 3 to 20. The algorithms were then evaluated according to the number of distinct categories to which word-frame instances were assigned during categorization. The results are displayed in Table 35.

There are three values reported for each value of *K* (each row). These are (i) the number of categories that have any frame-word instances allocated to them ("Any"); (ii) the number of categories receiving more than 1% of all instance allocations ("1%"); and (iii) the number of categories receiving more than 5% of all instance allocations ("5%"). In the case of the parsimony-based algorithm, the numbers shown are means over 50 iterations.

A very striking result is that, even from a starting position that theoretically allows frame-word instances to be allocated to any of *K* = 20 clusters, in practice only a small number of clusters out of the full range are used. In particular, with the conflict-based algorithm, for all values of *K*, only 3 categories are extensively used (greater than 1% of instances) during categorization, and they correspond to the familiar categories noun, verb and adjective. Even though we may start from a high number of clusters, therefore, the conflict-based algorithm “self-organizes” so as to reduce the number of clusters used to three (and note that it stops there, and does not proceed to merge these into two categories, or one). This behaviour may be understood in terms of the voting mechanism: because the main categories have a numerical advantage from the start, they are able to cast a larger number of votes, so that the conflict-based algorithm is biased to allocate most instances to the majority categories; the initial bias to the main categories is therefore amplified so that only those categories are used to a significant extent.

This result strongly suggests that the three main categories in English are robustly present in the data and can be uncovered by a conservative co-clustering algorithm such as the conflict-based algorithm, and that it may not be necessary to make use of other mathematical techniques to determine the optimal “cut-off level” of the hierarchical clustering tree (but see Mintz, 2000, for one possible approach to doing so).

However, it should be noted that the parsimony-based and fuzzy product algorithms did not exhibit this behaviour as clearly as the conflict-based approach. While these algorithms greatly reduced the number of categories used compared to the original *K* value, the number of categories used increases with higher values of *K*.

It is also possible, for the discrete algorithms, to ask the same question for the frames and words separately, as these algorithms produce separate allocation matrices for words and frames. We can examine the allocation matrix for, say, words, and count the number of categories such that at least one word received a 1 in its allocation vector for the category in question, and likewise determine which categories received 1% or 5% of the allocations. The results for the parsimony-based algorithm are displayed in Table 36, and the results for the conflict-based algorithm are displayed in Table 37. In addition to the questions asked about the instance categorization, the tables also show the number of categories such that there exists at least one word (or frame) for which that category is the *only* allocated category (“Unq”).

As was the case with the categorization of instances, the conflict-based algorithm was far more conservative than the parsimony-based algorithm in constraining the number of categories used for both frames and words. The number of categories to which more than 5% of words and frames were assigned was nearly always 3 (except for a short range of *K* values in the case of frames). The number of unique categories increased slowly with *K*. Rather surprisingly, the number of unique categories for the parsimony-based algorithm, for both frames and words, was very high: the algorithm was not able to allocate all items to the main categories.

These results therefore suggest that there may be independent support for the reality of the main categories noun, verb and adjective in the input to the child, confirming linguistic intuitions as reviewed in Chapter 2. Out of the three co-clustering algorithms, the conflict-based approach is best able to reduce the set of categories to these three categories.

	Fuzzy FW Product			Parsimony-based co-clustering			Conflict-based co-clustering		
K	Any	1%	5%	Any	1%	5%	Any	1%	5%
3	3	3	**3**	3	3	**3**	3	3	**3**
4	4	3	**3**	3.92	3	**3**	4	3	**3**
5	4	3	**3**	4.98	3	**3**	5	3	**3**
6	5	3	**3**	5.98	3	**3**	4	3	**3**
7	7	4	**4**	7	4	**4**	5	3	**3**
8	8	4	**4**	7.14	4	**4**	5	3	**3**
9	9	4	**4**	8.12	4	**4**	5	3	**3**
10	10	5	**4**	9.04	5	**4.84**	6	3	**3**
11	11	6	**5**	10.02	6	**5.32**	8	3	**3**
12	12	6	**5**	11	6	**5.28**	9	3	**3**
13	13	6	**5**	11.86	6	**5.22**	9	3	**3**
14	14	6	**5**	12.86	6	**5.54**	10	3	**3**
15	15	6	**5**	13.8	6	**5.36**	10	3	**3**
16	16	6	**5**	14.82	6	**5.16**	9	3	**3**
17	17	6	**5**	15.74	6	**5.34**	9	3	**3**
18	18	7	**5**	16.4	7	**5.04**	9	3	**3**
19	19	7	**5**	16.18	6.98	**4.22**	9	3	**3**
20	20	8	**5**	18.38	7.98	**4.28**	8	3	**3**

Table 35. The number of categories used during categorization following each of the three co-clustering algorithms, for full-utterance frames, N = 290. "Any" = number of categories that account for at least 1 allocated frame-word instance; "1%" = number of categories accounting for at least 1% of instances; "5%" = number of categories accounting for at least 5% of instances. Values for parsimony-based co-clustering are means over 50 iterations.

	WORDS								FRAMES							
	SEEDS				FINAL				SEEDS				FINAL			
K	Any	1%	5%	Unq	Any	1%	5%	Unq	Any	1%	5%	Unq	Any	1%	5%	Unq
3	3	3	3	3	3	3	**3**	**3**	3	3	3	3	3	3	**3**	**3**
4	4	4	3	4	3.94	3.04	**3**	**3.04**	4	3	3	4	3.92	3	**3**	**3.92**
5	5	5	4	5	4.98	4	**3**	**4**	5	4	3	5	4.98	3.44	**3**	**4.98**
6	6	6	5	6	6	5	**3.56**	**5.02**	6	4	3	6	6	4.34	**3.74**	**5.98**
7	7	7	5	7	7	6	**4.04**	**6.04**	7	6	4	7	7	5.66	**4.02**	**7**
8	8	7	5	8	7.3	6	**4.04**	**6.02**	8	6	4	8	7.28	5.78	**4.02**	**7.12**
9	9	8	5	9	8.24	7	**4.04**	**7.04**	9	7	4	9	8.16	6.58	**4.02**	**8.1**
10	10	9	6	10	9.22	7.86	**5**	**8**	10	8	5	10	9.14	7.3	**5**	**9**
11	11	10	7	11	10.16	9.04	**6**	**9.04**	11	9	6	11	10.2	8.26	**6**	**9.98**
12	12	11	7	12	11.2	9.48	**6**	**9.04**	12	9	6	12	11.16	8.52	**6**	**11**
13	13	12	7	13	12.06	9.18	**5.94**	**9.04**	13	9	6	13	12.02	8.38	**6**	**11.84**
14	14	13	7	14	13.08	10.22	**6**	**9.18**	14	10	6	14	13.12	8.6	**6**	**12.82**
15	15	14	7	15	13.88	10.24	**6**	**10.26**	15	10	6	15	13.94	8.54	**6**	**13.8**
16	16	15	7	16	15.02	10.14	**5.98**	**10.18**	16	10	5	16	15.08	8.38	**6**	**14.82**
17	17	16	7	17	16.04	10.12	**6**	**10.7**	17	10	5	17	16.06	8.6	**6**	**15.74**
18	18	17	8	18	16.78	11	**6.54**	**11.66**	18	12	5	18	16.76	9.6	**6.82**	**16.32**
19	19	18	8	19	17.08	10.78	**6.64**	**11.74**	19	12	5	19	17	9.4	**6.72**	**16.06**
20	20	19	8	20	18.96	11.66	**6.5**	**13.7**	20	13	6	20	18.94	10.34	**7.5**	**18.22**

Table 36. Effect of the initial number of clusters on mean number of large categories produced in the parsimony-based co-clustering algorithm, full-utterance frames, N=290. Key as for Table 35, plus "Unq" = number of categories that are the only category allocated to at least one item.

	WORDS								FRAMES							
	SEEDS				FINAL				SEEDS				FINAL			
K	Any	1%	5%	Unq	Any	1%	5%	Unq	Any	1%	5%	Unq	Any	1%	5%	Unq
3	3	3	3	3	3	3	**3**	**3**	3	3	3	3	3	3	**3**	**3**
4	4	3	3	4	4	3	**3**	**3**	4	3	3	4	4	3	**3**	**3**
5	5	5	4	5	5	3	**3**	**3**	5	4	3	5	5	3	**3**	**3**
6	6	6	5	6	6	4	**3**	**3**	6	4	3	6	6	4	**3**	**3**
7	7	7	5	7	7	4	**3**	**4**	7	6	4	7	7	5	**4**	**3**
8	8	7	5	8	8	4	**3**	**4**	8	6	4	8	8	5	**4**	**3**
9	9	8	5	9	9	4	**3**	**5**	9	7	4	9	9	6	**4**	**4**
10	10	9	6	10	10	5	**3**	**5**	10	8	5	10	10	6	**4**	**4**
11	11	10	7	11	11	6	**3**	**4**	11	9	6	11	11	7	**3**	**5**
12	12	11	7	12	12	6	**3**	**4**	12	9	6	12	12	7	**3**	**5**
13	13	12	7	13	13	6	**3**	**4**	13	9	6	13	13	7	**3**	**5**
14	14	13	7	14	14	6	**3**	**4**	14	10	6	14	14	7	**3**	**5**
15	15	14	7	15	15	6	**3**	**4**	15	10	6	15	15	7	**3**	**5**
16	16	15	7	16	16	6	**3**	**4**	16	10	5	16	16	7	**3**	**5**
17	17	16	7	17	17	6	**3**	**4**	17	10	5	17	17	7	**3**	**5**
18	18	17	8	18	18	6	**3**	**4**	18	12	5	18	18	8	**3**	**5**
19	19	18	8	19	19	5	**3**	**4**	19	12	5	19	19	8	**3**	**5**
20	20	19	8	20	20	6	**3**	**5**	20	13	6	20	20	9	**3**	**6**

Table 37. Effect of the initial number of clusters on mean number of large categories produced in the conflict-driven co-clustering algorithm, full-utterance frames, N=290. Key as for Table 36.

7.5.2 Evaluation of independent frame and word ambiguity

The parsimony-based and conflict-driven co-clustering algorithms produce models of the ambiguity inherent in the frames and the words independently, by means of the separate allocation matrices for frames and for words. Because these matrices list the possible categories that can be assigned to each frame and each word, they express the degree and kind of ambiguity inherent in each item. It is therefore possible to ask *which* individual items were identified as ambiguous. In order to look more closely at this issue, we can focus on the words that have been explicitly identified as ambiguous by the discrete co-clustering algorithms. Because of the practical difficulty in this case of describing the average behaviour of the parsimony-based algorithm, I focus only on the conflict-based algorithm.

For comparison, we can list the words that have been allocated to more than one category in different usage contexts, according to the gold standard. This listing is shown in Table 38, and contains word lists for each of the four combinations of the main three categories. These should, in theory, be the ambiguous words in the dataset according to the gold standard. Unfortunately, deliberately focusing on the most ambiguous words brings us up against the limitations of the gold standard part-of-speech tagging. Not all words described as ambiguous in the table are actually used in more than one category in the dataset, and the gold standard contains several tagging mistakes, in my opinion.

For instance, in the following dialogue (Manchester corpus, file *john18a.cha*, lines 178-189), "parrot" is misclassified as a verb, although it is clearly a noun given the context:

*MOT: right. (adverb)
*CHI: right. (adverb)
*CHI: right. (adverb)
*MOT: you parrot. (pronoun verb)

And in fact, about half of the words on the lists in Table 38 are used in only one category in the portion of the corpus under consideration, but have been tagged incorrectly by the gold standard in some contexts. Words that were, in my opinion, incorrectly tagged as ambiguous are shown in italics.

NOUN/VERB/ ADJECTIVE	*black*, *crash*, *cross*, fit, *side*, *upset*
NOUN/VERB	answer, arm, arms, *bag*, balance, *ball*, bang, *bear*, *being*, bend, bite, blow, *bottle*, brush, building, bump, *buy*, call, *chalk*, change, *chicken*, *chin*, clap, *clock*, *colors*, *cook*, *corner*, *count*, cover, *crawl*, cut, dance, dress, drink, drive, *duck*, *excuse*, fall, *farm*, *field*, *fingers*, *fire*, *fish*, *fishes*, *fits*, fly, *foot*, *game*, *ham*, hand, *hands*, help, *hide*, *hit*, hold, *hole*, *home*, *house*, *juice*, jump, kick, kiss, knock, *leave*, lift, lock, love, *mat*, *measure*, *mess*, *milk*, mind, miss, *number*, *numbers*, nurse, *page*, paint, *pants*, *parcel*, *pardon*, park, *parrot*, pass, pat, *pay*, peel, phone, *picnic*, *picture*, *piece*, *pig*, *pile*, play, *pocket*, point, pop, post, press, prod, pull, push, *puzzle*, race, rest, ride, ring, rock, roll, rub, run, *sauce*, *shake*, *shampoo*, *shop*, shopping, show, *sign*, sink, sleep, slide, smell, smile, snap, sneeze, sort, spot, squash, squeeze, *stand*, start, stay, *steps*, stick, *stocking*, stop, *stroke*, swim, talk, tape, telephone, *thumb*, *tin*, toast, *toys*, *track*, try, video, *wake*, walk, wash, washing, watch, *water*, weewee, wind, wipe, work, *works*, writing
ADJECTIVE/VERB	awake, broke, *brown*, clean, clear, close, dry, empty, *gentle*, left, *long*, lost, mean, open, pretend, *rough*, *shy*, *slow*, warm, wee, wet
ADJECTIVE/NOUN	beautiful, *billy*, bottom, cold, cream, dark, *dead*, drunk, *fast*, *fat*, *flat*, *full*, fun, grey, head, high, *key*, kind, last, light, lucky, magic, *minute*, *present*, purple, quiet, *safe*, *sick*, square, *stable*, sticky, *stiff*, stripy, *super*, sweet, tiny, top, white

Table 38. Words in the currently-used data matrix that are ambiguous according to the gold standard part-of-speech tagging of the Manchester corpus. Words incorrectly tagged in the gold standard as ambiguous are in italics.

The words identified as ambiguous by the conflict-based algorithm are shown in Table 39, for each combination of the main categories. Words that appear in both this list and the gold standard ambiguous list for the same category combination are highlighted in bold. Here, too, there are several instances where words have in my opinion been used as members of more than one category, but this is not reflected in the gold standard tagging. These words are marked in italics. Some of these may be debatable, but there seems to be little justification for accepting some uses of “bang” as nouns, but not “crash”, or accepting “kiss” as a noun but not “cuddle”.

NOUN/VERB/ ADJECTIVE	**fit**, pretend, through, *washing*
NOUN/VERB	**bang**, **bite**, **blow**, **brush**, by, **change**, **cover**, *crash*, *cuddle*, **dress**, **drink**, goes, **help**, **hold**, **kiss**, **lift**, **lock**, **love**, mean, **mind**, **nurse**, open, **paint**, **park**, **pass**, **pat**, **play**, **point**, **pop**, **post**, **pull**, remember, **ride**, **rock**, **roll**, *saw*, sit, **sleep**, **slide**, *smack*, so, **sort**, **spot**, **squeeze**, **stand**, **start**, stay, **stick**, *tickle*, **try**, **walk**, **wash**, **watch**, **wind**, **wipe**, **work**
ADJECTIVE/VERB	behind, biting, **broke**, bumped, **clean**, **close**, crashed, *done*, driving, dropped, **dry**, eating, fallen, *finished*, *fixed*, found, had, **lost**, made, maybe, missed, moved, near, popped, read, really, *shut*, spilt, *squashed*, standing, stop, *tidy*, **warm**, won, works
ADJECTIVE/NOUN	actually, *Anna’s*, bedtime, better, *biggest*, *black*, *both*, *bright*, broken, *brown*, building, carefully, **cold**, coloring, crying, *daddy's*, dalmatians, **dark**, drawing, enough, **fun**, gently, grandma, gumdrop, *home*, inside, **kind**, *left*, *long*, lovely, **lucky**, mine, morning, much; orangejuice, painting, *pink*, playing, **purple**, pushing, quick, sitting, sleeping, steady, **sticky**, *straight*, **stripy**, stuck, sweetcorn, Thomas, **tiny**, walking, **white**, writing, *wrong*, *yours*, yummy

Table 39. All words deemed as ambiguous by the conflict-based co-clustering algorithm for N=290. Correctly identified ambiguous words (i.e. that are also ambiguous according to the gold standard) are in bold. Ambiguous words that were not marked as ambiguous in the gold standard are in italics.

Having stated these reservations, it nevertheless seems that the conflict-based algorithm performed quite well in identifying ambiguous verb/nouns; only a few of the words identified as being both verbs and nouns were not so used in the corpus. Note that in the case of “wind”, there are actually two different pronunciations for the verb and noun forms, so that these would not have been confused with each other by a child. In the case of adjective/verb pairs, the algorithm was much less successful; it identified many verb participial forms which are arguably adjective-like, in that they often occur in constructions with the copula (“are you *standing*?”). Among the successful cases, note again the presence of the phonologically disambiguable word “close”. The adjective/noun group was also less successful than the noun/verb group. Some words on the list may be surprising, but they are indeed used in the corpus in unusual ways, e.g. “is it a *sticky*?”, “what does *Stripy* say?”, “in the *bright*”. Note in this group also five present participles which are clearly nouns (most of them names for objects) in certain contexts: “building”, “coloring”, “drawing”, “painting” and “writing”. These should perhaps have been classified as verb/noun ambiguous words, but because of the affinity of participial forms for the class of adjective, they have been termed ambiguous noun/adjective words instead.

While the gold standard tagging is likely to be useful for the purpose of roughly gauging the classification success of the algorithms discussed in this thesis, it runs into clear problems when dealing with the fine-grained detail of ambiguous words. Given these categorization problems, the analysis above is not repeated for frames.

7.5.3 Frames with multiple-X sequences

The frames considered so far have the characteristic that all X slots are isolated from each other: an X is always preceded and followed by either a frequent word or an utterance boundary. Cases where we have two or more X’s in succession, e.g. *the X X*, are not considered.

These cases are in fact somewhat problematic for part-of-speech induction. To take the example of “the X X”, the fillers for the slots could be an adjective followed by a noun: “the dirty glass”, “the ugly duckling”, “the hungry lion”. However, the slots could also be filled by a noun followed by a verb, as in “the glass broke”, “the duckling cried”, “the lion sleeps tonight”. Deciding between the two cases can be done if the part-of-speech of at least one word is known. It is really the presence of both “the” and “Noun” in the structure “the Adjective Noun” that licenses the presence of “Adjective”. Likewise, “the” and “Adjective” license “Noun”.

Categorizing the adjective in this structure would have been possible under a lexically-specific approach, if, say, “the X glass” was a prominent frame; however, it is not, because of the relative rarity of the word “glass”, and this is in fact true for most specific nouns. The child would need to know that “glass” is a noun first before being able to learn the “the Adjective Noun” pattern. But when the child is still learning the conditions for recognizing the various parts-of-speech, she will not yet have the required knowledge on which to base the decision.

One way of looking at lexically-specific frames is that the specific material *licenses* the use of the category associated with the variable slot, e.g. *the* in *the X* licenses the presence of a noun in the X slot. Note that this relationship is between a specific word (*the*) and the *category* of nouns, not any specific noun. In other words, only when the category has been identified can we describe the relationship in the utterance correctly. (This relationship between elements is known in linguistics as *dependency*.) The "no-multiple-slots" constraint may be seen as a way to take dependency into account.

However, it is worth examining how crucial the constraint of isolated X slots is for a successful categorization outcome. If dropping this constraint results in much poorer performance on the evaluation measures, then the constraint is a vital one. On the other hand, if the evaluation still yields fairly high scores on the measures, then that allows for the possibility that children are able to make use of frames with arbitrary lengths of X sequences, and hence that the isolated-X situation is contiguous with the multiple-X sequence situation rather than being qualitatively different.

For this experiment, I consider only N=290, as a reasonable representative of the full range of N. Words were divided into frequent and less-frequent words, and all utterances in the corpus rewritten as before. This time, however, word-frame co-occurrence data was collected for words occurring in any frame, not just ones with X's in isolation, with the proviso that the frame had to contain at least one non-X word.

For example, the new set contained the frames "Do you like X Z?" and "Do you like Z X?", where the X indicates the active slot, and the Z the inactive one. These frames matched against utterances such as "Do you like baked beans?", "Do you like Grandma's cakes?" and "Do you like eating spaghetti?". Note that these two frames are taken to be independent from each other, so that "eating" is taken to be a filler of "Do you want X

Z?” and “spaghetti” is a filler of” Do you want Z X?”, and there is nothing in the data set to indicate that “eating” and “spaghetti” occurred together.

The rest of the experiment was conducted as before, with a hard clustering of the frames, followed by execution of the set of co-clustering algorithms. The results are shown in Table 40.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.787 (*0.526*)	0.803 (*0.526*)	0.866 (*0.526*)	0.872 (*0.526*)	0.852 (*0.526*)	0.858 (*0.526*)
Completeness	0.655 (*0.438*)	0.719 (*0.471*)	0.780 (*0.474*)	0.811 (*0.489*)	0.818 (*0.577*)	0.804 (*0.493*)
F	0.715 (*0.478*)	0.759 (*0.497*)	0.821 (*0.499*)	**0.841** (*0.507*)	0.835 (*0.516*)	0.830 (*0.509*)
Bookmaker	0.623	0.657	0.760	**0.775**	0.690	0.752

Table 40. Evaluation of frames with multiple-X sequences, N=290, 4 clusters.

Randomization tests of significance showed that all differences of interest are significant, with Fuzzy F × W performing the best on both F (Table 41) and Bookmaker (Table 42) scores, followed by Pars, then Confl and then Hard F.

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	← < 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 41. Significance levels of differences in F scores for full-utterance frames with multiple-X sequences, N=290.
*** significant at p=0.05, ** significant at p = 0.01.**

	Fuzzy FxW	Confl.	Pars.
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	← 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ 0.001**		

Table 42. Significance levels of differences in Bookmaker scores for full-utterance frames with multiple-X sequences, N=290.
*** significant at p=0.05, ** significant at p = 0.01.**

As is apparent from comparison with Table 27, relaxing the constraint that all X's should be isolated adversely affected the correctness of the resulting categorization to some degree. All evaluation scores are lower than they were with "isolated-X" frames. Nevertheless, these scores are still high in absolute terms, so that it would be fair to say that even multiple-X sequence frames are an adequate basis for discovering the parts-of-speech of English. The constraint of isolated X's will be maintained, however, on the grounds that these frames may be based on dependency relationships which obtain between specific word categories and dependent frequent words; these relationships cannot be calculated between members of parts-of-speech before the categories have been discovered.

For multiple-X frames, the most successful clustering algorithm was Fuzzy F × W, followed by Pars. (and Fuzzy W), and then Confl.

7.5.3.1 An alternative treatment of material in successive X slots

The no-multiple-slots constraint was justified above because it takes into account that sometimes only the *category*, rather than the *identity*, of a contextual word is informative of the category of a word adjacent to it. Apart from this, there is also a problem of *segmentation*. With the most common words, we can assume that the words can be segmented out from the speech signal fairly easily due to their familiarity. This facilitates segmentation of the rarer words if we apply the no-multiple-slots constraint; they are simply the sections in between the frequent words. In the case of a *sequence* of rarer words, however, we cannot assume that the child already knows how to split the sequence into its individual words.

An alternative solution to the problem of frame slots in succession is to investigate the constituency of the utterances that the child hears. Arguably, the sequence *Adjective Noun* in utterances such as “the dirty glass”, “the ugly duckling”, “the hungry lion”, can be regarded as a coherent unit, something that is more difficult to claim for *Noun Verb* sequences, e.g. “lion sleeps” in “the lion sleeps”. Suppose now that we alter the definition of the X slots, so that any sequence of non-frequent words is rewritten as a *single* X. In other words, X means “slot filler” rather than “single-word slot filler”. This would mean that all the sentences discussed above would be described as “the X”, with fillers “dirty glass”, “glass broke”, “ugly duckling”, “duckling cried”, etc. Now bear in mind that frequency of occurrence plays a major role in entrenching units, including the putative constituents “hungry lion” and “lion sleeps”. To the extent that “hungry lion” is more of a constituent than “lion sleeps”, we can expect it to occur more frequently and hence become more entrenched as a unit in the child’s lexicon. In this way, “hungry lion” will survive as a unit, while “lion sleeps” will be discarded, and hence not be considered as a filler for the frame “the X”. Furthermore, “hungry lion” is treated as a filler on a par with just the word “lion”, obtained from “the lion”.

Now, at a later stage, the child can use this knowledge to redescribe utterances such as “the hungry lion sleeps” as, say, “the W X”, where W represents the previously learned category consisting of nouns and adjective-noun sequences. This then facilitates learning what kinds of words can occupy the new X slot (e.g. “sleeps”). Thus the problem is dealt with in two stages.

Of course, it is an empirical question whether children do follow such a developmental path during language learning, and one which has to my knowledge not been investigated yet. From a purely practical standpoint, it is desirable in the current work to evaluate the outcome of categorization using the part-of-speech tagging provided with the Manchester corpus, which assigns a tag to each individual word and would not assign a single category to the unit "hungry lion", hence evaluating the results of such an experiment would be difficult on pragmatic grounds. The approach just discussed is not examined in this thesis, but may well become the focus of future explorations.

7.6 Discussion

In this chapter, I presented a number of techniques to obtain a categorization of word-frame instances that were more sophisticated than the strategy followed in Chapter 6 of merely imposing the category of the frame on the instance, with frames presumed to be associated with one and only one category. These techniques (i) made use of information about the identity of the word as well as the frame, combining these two sources of information in order to arrive at a better categorization, and (ii) moved away from an all-or-nothing allocation, where words and frames are assigned to only one category, to an allocation that directly addresses the inherent ambiguity of linguistic elements. This was done by either expressing category membership in probabilistic terms, as in fuzzy co-clustering, or by enumerating all categorical possibilities explicitly, as was done in the parsimony-driven and conflict-resolution co-clustering algorithms. These techniques proved to be highly successful in categorizing word-frame instances, attaining the exceptionally high evaluation scores of an F value of 0.905 (over baseline of 0.568) and a Bookmaker value of 0.833 in the best case.

However, it may be necessary, in accurately describing the set of frames to which the child is exposed during language learning, to move beyond using only frames that capture the structure of a full utterance. Many of the full-utterance frames identified in Chapter 6 are highly redundant with each other: Frames such as "Find the X", "Are you Z the X ?", "Are you going to Z the X ?", "Can I have the X ?", etc., are all assigned to the "noun" cluster; yet we might suspect that it is just the local noun phrase structure "the X" that is doing the work in these cases of identifying the word in the X slot as a noun.

One could surmise that the prevalence of “the X” in the above contexts and many others is due to the fact that it is a linguistic *constituent*, i.e. it is a coherent unit which can be embedded in a variety of contexts. It would be of great use, in learning about parts-of-speech, to be able to identify these nested constituents. If the phrase “the X” was identified as a nested constituent in all of the above larger frames, then the frames themselves could be discarded in favour of “the X” only, and their word occurrence data, which had been divided among a set of independent frames, could now be credited to the single “the X” frame. Taking this approach would make the description of utterances more compact and more rational, and could provide the clustering process with more accurate information. In addition, a larger part of the corpus might be covered by the frame model if frames were able to cover partial utterances as well as full utterances. This approach will be the focus of Chapter 8.

8 Hierarchically nested frames

8.1 Introduction

In this chapter, I will extend the full-utterance frame approach of the previous chapters, by presenting an algorithm that discovers smaller frame-like structures nested inside larger utterances. The algorithm is based on the notion of discovering linguistic constituents. Constituents are elements that are able to appear as whole units in utterances, in positions where they can often be interchanged with other constituents. This criterion is enshrined in the substitution test of structural linguistics (e.g. Harris, 1954).

I will set out a procedure for automatically discovering such embedded frames, making use of a modified substitution criterion which notes that constituents can often be *replaced by single words* without doing violence to the syntactic structure of an utterance (see e.g. Clark & Clark, 1977). So, for instance, the noun phrases in brackets in the first two sentences below can be replaced by a single word, e.g. “Maurice”:

It’s [*her voice*] I can’t stand.
It’s [*the standing in line waiting*] I can’t stand.
It’s [*Maurice*] I can’t stand.

And so one might imagine that one could line up sentences with the same basic sentence structure, but which differ in having variable material occurring in the same position in the structure across sentences. The varying material, if repeated often enough in several different contexts, is likely to represent the structure of a linguistic constituent, and it is these nested structures that will be treated as so-called *nested frames*. This chapter details a computational implementation of this idea.

The idea of lining up sentences against each other in order to discover shared versus different structure is not new; in fact, it has been explored in depth by van Zaanen (2001) in his work on Alignment-Based Learning (ABL). The work in the current chapter follows a general approach similar to that of ABL. Doing so is of interest in that it complements and extends the full-utterance frame approach of Chapter 6 to finding lexically-specific frames. However, there are also a number of significant differences between the current approach and ABL, which will be highlighted in Chapter 10.

There is at least a reasonable justification for the fairly conservative approach of the previous chapters, namely searching for linguistic constituents by considering only full-utterance frames: we are interested in discovering items that can be clearly segmented out from their surrounding context, and full-utterance frames, arguably either having no surrounding context, or otherwise being separated from the surrounding discourse context by silence, are clearly segmented-out units in this sense.

Furthermore, the full-utterance frame approach also gives us access to structures that would possibly not have been apparent in a pure embedded-fragment approach. Notably, some question frames (*what do you X ?*, *why have you X it ?*) have a structure which shapes the utterance as a whole, and these frames are unlikely to be embedded in larger structures.

Nevertheless, finding nested frames allows us to describe the regularities in child-directed language more accurately: instead of making use of several full-utterance frames that each contain the substring “the X” around the focal word, we can make use of the single frame “the X” and merge the co-occurrence data from all of these frames together to obtain a richer set of information about word-frame co-occurrence. The nested-frames approach also increases coverage, allowing us to categorize words that would not have been within reach of the full-utterance frame approach, merely because the entire utterance structure would not have matched a frame on its own.

8.2 Frame discovery procedure

The procedure for finding nested frames starts with the Manchester corpus rewritten in the same way as for full-utterance frames, with the most frequent 150 words retained as they are and all other words replaced with X. Next, the set of *all* full-utterance frames is collected as a substrate on which the substitution process will operate. No filtering step is applied to remove frames that do not co-occur with a minimum number of different word types, as was the case in the previous chapters. All frames that occur more than a fixed number of times overall in the corpus are included for consideration; in the current simulation, a frame needed to occur only twice in the corpus. This relatively lenient parameter setting reflects the intuition that utterances that occur only once may be spurious, or may have been incorrectly coded. Frames that occur at least twice are probably reliable features of the input.

Next, all pairs of utterance frames are examined, in order to determine whether the two utterances align in such a way that an X slot in one member M_1 of the pair can be *expanded/elaborated* in order to produce the other member M_2 of the pair. If so, then the X slot in M_1 is treated as a slot that can potentially accept material of more than one word, and we can say that M_1 is *schematic for* M_2. So for instance, the X slot in "do you want X?" can be expanded into "another X"; hence, "do you want X?" is schematic for "do you want another X?". If these two utterances were encountered in the corpus, their alignment would suggest that the phrase "another X" is a potential linguistic constituent, and hence a candidate to be considered as a *nested frame*. The frame "do you want X?" is called the *nesting frame* of the nested frame, and is written as "do you want Y?", where the Y indicates a slot that can accept material of more than one word.

We can write the above alignment using a bracketing, where a pair of brackets delimits a putative syntactic constituent:

do you want [another X]?

As the algorithm progresses through the corpus, aligning all pairs of sentences, it collects data about all pairs of nested and nesting frames, storing the data in a frame-frame co-occurrence matrix, analogous to the frame-word matrix of the previous chapters.

If schematicity is represented as an ancestor-descendant relationship in a graph, then it is possible to build up an acyclic *schematicity graph*, detailing the ancestor-descendant links between nodes that represent frames. The computational procedure builds such a network as it progresses through the set of full-utterance frames. The relationship of schematicity is transitive: if T_1 is schematic for T_2, and T_2 is schematic for T_3, then it necessarily follows that T_1 is schematic for T_3. For this reason, nodes are only linked to their most immediate ancestors, as representing explicit links to non-immediate ancestors would be redundant. (The algorithm works by adding a schematicity link whenever schematicity is discovered, then removing redundant links in a subsequent cleanup phase.)

Enforcing transitivity has the benefit that we can actually analyse an utterance in finer detail than would have been the case had we represented only individual schematicity relationships. A schematicity chain provides a kind of structural bracketing for the last utterance in the chain; the bracketing can be constructed by placing each putative constituent in a pair of brackets, potentially producing several levels of nested structure in an utterance. Consider the bracketing

do you [want [me to [get your X]]] ?

This bracketing was produced in the actual simulation of the nested frame discovery algorithm, because the corpus contained the utterance structures "do you X?", "do you want X?", "do you want me to X?", and "do you want me to get your X?".

From a bracketed structure, we can collect frame-frame co-occurrence data at all levels. In this example, "do you Y ?" can be filled by "want Y", "want Y" can be filled by "me to Y", and "me to Y" can be filled by "get your X".

It is quite common for an utterance to produce more than one bracketing (i.e. there is more than one path upwards in the graph from a particular node). In this case, we simply collect co-occurrence data from all possible paths. It is assumed that reliable information will be more prevalent than spurious information when data from the entire corpus has been collected.

The alignment process does in fact produce a large number of incorrect or spurious bracketings, with many nested frame candidates that are not constituents at all. Once a data matrix of nesting frames and nested frames has been collected, it is therefore necessary to discard these misanalyses. This is done by enforcing the “5-5 constraint” from the full-utterance frame approach. Only nested frames that have been nested inside at least 5 different nesting frames are considered, and only nesting frames that have contained 5 or more nested frames are considered legitimate contexts. All other nested and nesting frames are filtered out.

At this point, we have a data matrix that indicates which local frames reliably appear nested inside which surrounding contextual frames. These nested frames are the ones that are likely to be relevant to lexical categorization of the words that, in turn, occur inside them. The next step is to examine the co-occurrence of words and these local nested frames.

There are at least three ways in which utterances can be parsed using the found local frame contexts:
Context-free: if a local frame occurs anywhere in the utterance, count the co-occurrence of that frame and the word that occurs in it. Such an approach is inherently able to generalize beyond the particular instances in which the frame was initially discovered.
Immediately context-sensitive: In a stricter approach, we only recognize nested frames when they actually occur in one of the nesting frames where they were discovered, i.e. in a nesting frame from the frame-frame co-occurrence matrix. So the recognition of the frame is only legitimate if there is an appropriate context surrounding it to indicate that it is in fact the frame that we think it is.
Fully context-sensitive: In the strictest version of this parsing regime, the constraint of context-sensitivity is not just that a frame should appear inside another nesting frame, but also that the nesting frame should in turn be nested inside another frame, and so on recursively until the full-utterance level is reached. The entire utterance must therefore be analysable into a set of frames nested inside other frames (with allowable combinations as sanctioned by the frame-frame matrix). This option is not explored here.

The corpus is parsed once again using one of these three methods, this time for the purpose of creating a frame-word matrix, where the frames are the set of accepted nested frames. This time round, if “get your X” is one of the nested frames, then when the utterance “do you want to get your rolling-pin?” is encountered, the frame context of “rolling-pin” is “get your X”, not “do you want to get your X?” as it was under the full-utterance frame approach.

Once the corpus has been parsed and the frame-word co-occurrence matrix has been derived, it is again filtered as was the case with the frame-word co-occurrence matrix in Chapter 6, so that only slots and fillers that occur more frequently than a fixed threshold are used. The clustering process proceeds as before.

8.3 Psychological considerations

Essentially, this process is intended to be one in which frames that have already become familiar through the full-utterance frame process now serve as “pathbreaking” frames for discovering more complex utterance frames that have the same lexically-specific structure, but contain multi-word expressions in their slots instead of single words. These multi-word expressions can themselves be regarded as nested frames in many cases.

Under the full-utterance frame approach, if a child has encountered a sentence such as *is that Thomas?*, then it is presumed that the frame of the utterance is schematically represented as *is that X?*. If the child encounters other utterances such as *is that Pingu?*, *is that Percy?*, etc., then those utterances will be similarly schematically represented, and the schematic representation will become reinforced in memory with each repetition (token frequency) and also with each different word filler (type frequency). When the schematic structure has been successfully entrenched in this way, it is reasonable to assume that it will be activated whenever it is encountered in the input, with a one-word filler.

From here, it does not take a large leap of generalization to consider the possibility that the frame may be applied even when the filler material is longer than one (infrequent) word. In this way, an utterance such as *is that your teddy?*, schematically represented as *is that your X?*, could be analysed as *is that X?* with the multi-word schematic filler *your X*. Note that the original *is that X?* frame, by being imposed onto a different utterance structure from the one in which it was discovered, allows for a *segmentation* (or a bracketing) of the utterance into hierarchically nested constituents. The new fragment *your X* is now potentially available to be entrenched in memory as a unit in its own right, provided once again that it exhibits high token frequency and type frequency, where the latter in this case entails not only that many different word types should occur in the X slot of *your X*, but also that *your X* itself should occur nested inside several different larger contexts other than *is that X?*. If the only context in which *your X* ever regularly occurred was the utterance *is that your X?*, then it would have been more accurate to regard the larger utterance as the "true" linguistic unit, and to describe the context frame of *is that your teddy?* as *is that your X?*. However, *your X* occurs embedded inside a great many larger structures, e.g. *with* [*your X*], *that's* [*your X*], *want* [*your X*]?, with each of these larger nesting frames already known to accept single-word fillers, and so there is a great deal of evidence for its status as an independent unit.

The purpose of carrying out this process, in the context of part-of-speech induction, is that it is the *local* context in which a focal word occurs that constrains the word's part-of-speech. In *is that* [*your X*]*?*, it is the immediate surrounding frame "your X" which constrains, say, *teddy*, to be a noun, not the larger frame *is that your X?*, and so this approach leads to a more parsimonious way of describing the constructions in English, and hence potentially a more accurate source of information about parts-of-speech. It would have been possible for the different full-utterance frames *is that your X?*, *with your X*, *that's your X* and *want your X?* to be allocated to different clusters under the approach of Chapter 6. However, when all of these utterances are redescribed as essentially "being" the smaller frame *your X*, then no such confusion can occur. Even though the sets of words that occurred in the various larger frames may not have overlapped perfectly, the words are now "drawn closer together" because they all occur in *your X*, and hence are more likely to be treated as belonging to the same category (at least when they occur in the context *your X*). The same holds to a lesser extent at the next degree of separation;

words which would have clustered together with each of the focal words of *is that your X?*, *with your X*, etc., are now also "drawn closer" to each other. Therefore, acknowledging nested frames provides a more compact and perspicuous representation and potentially a better source of information for part-of-speech induction.

An important difference between the current algorithm and Van Zaanen's (2001) ABL is that, in ABL, all possible alignments between two utterances are treated as potential information about linguistic constituency, so that for instance the two utterances *you want a cookie?* and *you want another cookie?* could potentially align because of the shared *you want Y* context. In the nested frames approach, this can happen only if the frame *you want X* (i.e. with a single-word filler) has also been attested in the corpus, and so the current approach does not allow general alignments as ABL does.

The reason for this is that in *you want X*, the filler slot (the *X*) is implicitly segmented away from the rest of the frame, by reason of being occupied by an infrequent word. Therefore the presence of the slot is regarded as a surface clue that variable material can be entered into the slot (the slot has implicitly been identified as a "growth point"). By contrast, in *you want a X?* there is no surface clue to indicate that the word string at the beginning of the frame should be segmented after *want*.

From an ABL point of view, it could be argued that this segmentation knowledge arises as soon as the child hears the structure *you want another X?*, because the simultaneous activation of both utterances in memory will indicate the appropriate point of segmentation, by a process of alignment and contrast. This is possible; but note that this requires not only the simultaneous activation in memory of the two utterances (the current utterance as well as the past one), but also the postulation of a new frame context *you want Y?* and two filler frames *a X* and *another X.* Such a process may well take place for pairs of utterances spoken with only a short time interval between them, as the utterances may then both be easily accessible in memory. In the general case, though, this process would require perfect recall of potentially any utterance previously encountered (and indeed several different utterances may align simultaneously), as well as the postulation of alignment structures for every pair of matches, which may place an undue burden on the child's memory retrieval and processing abilities.

In the current approach, the child is presumed to already possess the *you want X?* frame, and is merely *applying* or *recognizing* it during processing of *you want a X?*, in order to discover the filler constituent *a X* (and will apply it again when she encounters *you want another X?*). The nested frames approach requires less "going-out-on-a-limb" on the part of the language learner than ABL does (ABL can however discover a potentially larger amount of structure by having a wider range of alignments to draw on). Letting single-word slots guide the alignment process allows for a more conservative but more controlled approach.

Although this process has been outlined as an incremental one, with knowledge developing as more and more of the corpus has been processed, the actual implementation is a "batch-mode" algorithm, that finds these "substitution alignments" between the full set of full-utterance structures. An incremental version would be more psychologically accurate, but is left as a possibility for exploration in future work.

8.4 Implementation

The procedure described in the previous sections was implemented on the Manchester corpus. Table 44 and Table 45 show some summary statistics for locally context-sensitive and context-free parsing, respectively. (The corresponding numbers for full-utterance frames are shown in Table 43, for comparison.) It is apparent that there are more frames and filler words in the final data set in the case of context-free parsing than in the case of locally context-sensitive parsing. With the less stringent context-free parsing method, which recognizes frames wherever they occur, there are many more opportunities to recognize both frames and filler words, and hence it is easier for items to meet the 5-5 criterion for inclusion into the data set. The effect of this, also shown in the tables, is that there are many more word tokens that are covered as focal words by the final data set under context-free parsing, and that there are many more utterances that contain at least one instance of a nested or full-utterance frame.

Number of frame types	1465
Number of slot-filler types	1284
Number of focal words covered	36601 (2.8%)
Number of utterances covered	40885 (12.2%)

Table 43. Summary numbers regarding coverage of the Manchester corpus by the full-utterance frame approach, for N=290.

Number of frame types	643
Number of slot-filler types	2454
Number of focal words covered	98321 (7.4%)
Number of utterances containing at least one nested or full-utterance frame	86677 (25.9%)

Table 44. Summary numbers regarding coverage of the Manchester corpus by the nested-frame approach, with locally context-sensitive parsing, for N=290.

Number of frame types	923
Number of slot-filler types	3356
Number of focal words covered	131426 (9.9%)
Number of utterances containing at least one nested or full-utterance frame	108422 (32.4%)

Table 45. Summary numbers regarding coverage of the Manchester corpus by the nested-frame approach, with context-free parsing, for N=290.

Comparing the full-utterance frame values in Table 43 with the results from the nested frame approach in Table 44 and Table 45, it is clear that the nested-frame approach greatly reduced the number of frames used in the model. Several full-utterance frames were replaced by a far smaller number of nested frames. At the same time, the number of focal word types increased; this change can be attributed to the fact that the criterion for recognizing a nested frame is less stringent than that for a full-utterance frame, so that frames can be recognized in a variety of contexts, without needing to be the frame structure for a full utterance. Hence, there are more opportunities to recognize nested frames than full-utterance frames; this is confirmed by the far greater numbers of word tokens and utterances covered by the nested frame model. The new nested model is more compact (in terms of number of frames) and simultaneously covers a larger portion of the corpus. Whether the categorization produced by the nested model is better than that of the full-utterance frame model still remains to be determined, and will be examined in the rest of this chapter.

It should be pointed out that coverage is not as important as correctness of categorization. This is because the current work is not aimed at categorizing every word in the corpus, but at bootstrapping a set of parts-of-speech from frame information alone. These initial categories can then be developed further, for instance by incorporating semantic information.

Table 46 shows a selection of the bracketed structures extracted from the starting set of utterances. Where multiple possible bracketings were found for one utterance, every option is shown on a new line. Many of these bracketings are apparently intuitively correct, for instance “give it to [your X]”, “let’s [find X]”, “I’m [not [a X]]”. In other cases, though, a number of spurious alignments occur, that in turn may lead to the postulation of equally spurious nesting and nested frames. For instance, “I’ve X that one” aligns correctly with “I’ve X” to produce “I’ve [X that one]”, but also aligns incorrectly with “X one” to produce “[I’ve X that] one”. And it seems that only one of the three bracketings for “shall we do X again?” (i.e. “shall we [[do X] again]?”) is correct.

However, the algorithm is intended to be self-correcting to some extent. As long as these incorrect nesting or nested frames do not occur in a *systematic* way, i.e. with the same frames occurring in several utterances, they will not be added to the final data set. And even if some of them do get added, the correct data should outnumber the incorrect data sufficiently for the clustering algorithm to produce valid clusters. Section 8.6 will aim to examine whether this has in fact happened, by looking at the quantitative results of using nested frames.

Original utterance	**Bracketed utterance**
lots of X	lots of X *(remains the same)*
give it to your X	give it to [your X]
can you see a X ?	can you [see [a X]] ?
I've X that one	I've [X that one]
	[[I've X] that] one
you're X it	[you're X] it
	you're [X it]
you're very X , aren't you ?	[you're [very X]] , aren't you ?
let's find X	let's [find X]
you like X ?	you [like X] ?
I'm not a X	I'm [not [a X]]
that X a good X , X it ?	[that X a good X] , X it ?
I don't want that in my X	I [don't [want [that in my X]]]
oh he's a bit X , isn't he ?	[oh he's [a bit X]] , isn't he ?
you've X the X , have you ?	you've [X the X] , have you ?
don't X them on X	don't [[X them] on X]
	[don't [X them]] on X
shall we do X again ?	[shall we [do X]] again ?
	shall we [[do X] again] ?
	shall we [do [X again]] ?

Table 46. Some examples of bracketings found using the nested-frame algorithm.

that's **a X**; it's **a X**; going **to X**; do you **want X** ?; are you **going X** ?; it's **not X**;
that's **the X**; **that's X** , isn't it ?; where's **the X** ?; want **to X**; i **don't X**; not **a X**;
there's **the X**; **it's X** , isn't it ?; **X in** the X; there's **a X**; that's **not X**; **X in** there;
X on the X; are you **going to X**; X in **the X**; you **want X** ?; a **bit X**; X on **the X**;
do you **like X** ?; you've **got X**; in **the X**; what **do you X** ?; got **a X**; what **a X**;
you **X it**; **what X** that ?; what **X that** ?; **it's X** , is it ?; a **good X**; i'm **not X**;
on **the X**; **you X** it; **don't X** it; a **big X**; can you **say X** ?; i **think X**; **X that** one;
don't **think X**; you're **not X**; it **is X**; and **the X**; is it **a X** ?; don't **X it**; and **what X** ?;
this **is X**; a **little X**; can you **see X** ?; look **at X**; where's **your X** ?; want **a X**;
what **are you X** ?; what're you **X for** ?; have you **got X** ?; a **Z of X**; that **was X**;
that's X , is it ?; the **other X**; and **a X**; a **X one**; like **a X**; you **can X**; you **don't X**;
is **a X**; **i X** you; what **did you X** ?; can i **have X** ?; **X it** up; **X in** the X ?;
is that **a X** ?; **what X** one ?; what **X one** ?; i **can't X**; **X that** one ?; what's **the X** ?;
and **there's X**; it's **the X**; **that X** there; that **X there**; is **the X**; what **color X** ?;
are you X the X ?; are you **X the X** ?; want **some X**; what **Z of X** ?; a **nice X**;
that's **your X**; there's **your X**; do you **think X** ?; there's **some X**;
what's **X doing** ?; that **one X**; what's **that X** ?; **are you X** a X ?; about **the X**;

Table 47. The top 100 most frequent combinations of nested frames inside nesting frames occurring in the Manchester corpus.

Next, a nesting frame- nested frame co-occurrence matrix was obtained from the bracketing data. The most commonly-occurring combinations of nested and nesting frames are shown in Table 47. In addition, some selected examples of nesting and nested frames are shown in Table 48 and Table 49, organized around the nesting and the nested frame respectively. In each case, the nested frames are intended to appear in the Y slots of the nesting frames. Although some nesting frames favour nested frames from one particular phrasal type, e.g. “Can I Y?” which is the nesting context for a variety of frames which seem to be verb phrases (“X a X”, “X it”, “X up”), this is clearly not true for all nesting frames: for instance, just as the full-utterance frame “is it X?” contained verbs, adjectives and nouns, the nesting frame “is it Y?” accommodates noun phrases (“a X”, “the X”, “your X”), adjectival phrases (“too X”, “very X”) and items which could be either verb phrases or prepositional phrases (“X the X”, “X your X”). A number of

Nesting frame	Nested frames

all Y	X and X; X down; X now; X out;
	X up; a bit X; in X; nice and X;
	on the X; right X; that X; the X;
	these X; this X; those X; your X

can I Y ?	X a X; X her; X here; X in there;
	X it; X that one; X the X; X this X;
	X up; X you; X your X; have X

make Y	a X; another X; it X; some X; the X
why Y ?	X I; are they X; are you X; is he X; is it X; is she X

is it Y ?	X again; X now; X on; X on X;
	X or X; X the X; X there; X to X;
	X yet; X your X; a X; all X; an X;
	big X; called X; for X; going X; her X;
	his X; in a X; in the X; in your X;
	like X; my X; nice X; not X; on the X;
	still X; that X; the X; too X; very X;
	your X

Table 48. Some nesting frames and the nested frames that occur in them.

Nested frame	Nesting frames
too X	Y, is it?; Y now; are Y; are you Y?; bit Y; get Y; he's Y; I'm Y; is he Y?; is it Y?; isn't Y; it's Y; not Y; one's Y; that's Y; they're Y; what's Y; you're Y

your X	Y are Z; Y car; Y doing; Z in Y;
	about Y; all Y; do Y; find Y; have Y;
	I'm Y; in Y?; is she Y?; not Y;
	that's Y; there's Y; who's Y?

X it	Y a bit; Y again; Y for Z; Y in; Y over; Y to Z; are you Y; can Y; did you Y?; going to Y; haven't Y; I'm Y; is she Y?; let's Y; shall we Y?; she Y; to Y; who Y?

Table 49. Some nested frames and the nesting frames in which they occur.

missegmentations persist: for instance, “all right X” was missegmented as “all [right X]”, and “make it X” as “make [it X]”. Likewise, the context “your X”, a prototypical noun phase that can be expected to take nouns in its slot, was correctly segmented out from surrounding contexts such as “[your X] are Z”, “Z in [your X]” and “all [your X]”, but was also identified in the context “[your X] car”, where it is not a constituent and the single-word fillers may be expected to be adjectives. Nevertheless, most nested frames seem to correspond to plausible constituents.

As described above, the corpus was parsed again, and a word-frame co-occurrence matrix was obtained under the two parsing regimes. The context-free parsing routine used the nested frames as contexts for the words wherever they appeared. The context-sensitive parsing routine took nested frames to be the contexts of words only when those nested frames appeared inside the context of one of their “own” nesting frames, i.e. a nested frame was recognized only when it appeared in the context of one of the nesting frames with which it co-occurred when the nesting frame-nested frame matrix was created.

The full-utterance frames obtained in the previous chapter were retained in this experiment as a “fallback option”: if no nested frame matched any particular part of an utterance, a match was sought against the full-utterance frames. This reflects the view that both full-utterance frames and nested frames may be two different kinds of frames that form part of the child’s knowledge of English, and so the full-utterance frames are also included in the final data set, provided that they are used frequently enough to meet the 5-5 criterion. The word-frame matrices were then subjected to the various co-clustering algorithms introduced in Chapter 7.

8.5 Qualitative results

For locally context-sensitive parsing, application of the “3-sizeable-clusters” rule from the previous experiment on full-utterance frames did not produce a set of clusters corresponding to nouns, verbs and adjectives. Instead, the first three large clusters corresponded to a cluster of nouns, a cluster of modal verbs, and a merged cluster of verbs and adjectives. It was only at 8 clusters that adjectives differentiated out from verbs, and these are the results reported here. Selected frames and most-closely-associated words for the 4 sizeable clusters out of the 8 are shown in Table 50.

For context-free parsing, the clustering for 8 clusters was the first point at which 3 sizeable clusters were formed. Again, one cluster was a mix of verbs and adjectives, while another was a cluster of (singular count) nouns. The third cluster in this case was one of plural nouns and mass nouns. The examples shown in Table 51 are from 9 clusters, where verbs differentiated out from adjectives.

Cluster	Frames	Words
Cluster 1	X , is it ?; X baby; X girl; Z it's X; a bit X; all X; are they X; be X; has it X ?; I'm X; is X; it isn't X; nice and X; not X; one's X; still X; that Z be X; that's X; they're X , aren't they ?; too X; very X; you're X	stuck, broken, hot, tired, cold, dirty, asleep, better, coming, hiding, poorly, alright, sleeping, sad, hungry, wet, happy, crying, outside, ready, done, eating, looking, upstairs, behind, driving, getting, lovely, sitting, clever, cross, hard, pink, wrong, gonna, horrible, inside, running, yours, clean, dark, home
Cluster 2	X another Z; X away; X back; X her Z; X him; X it all; X it; X me; X out; X round; X some Z; X that; X the Z; X with Z; can X; come and X; did X; didn't X; don't want to X; going to X; got to X; have you X; having a X; i X; I'll X; I'm not X; I've X; in X; let's X; mummy X; shall i X; shall we X; to X; want to X; we've X; what're you X	sit, play, pull, read, hold, keep, bring, getting, push, stand, cut, leave, watch, buy, had, pick, stop, fit, open, tell, try, done, use, sing, blow, made, eating, let, press, throw, drive, help, lost, roll, bite, found, stay, fix, jump, stick, drink, fall, knock, run, tip, wear, show, wipe, break, brought, shut, talk, wash, write, clean, lift, putting, sleep, taking, walk, work

Cluster 3	Z , X it ?; X , look; X be; X come; X darling; X go; X gone; X have X; X just; X man; X put; X they ?; X we ?; X what; Z with the X; X you ?; X you Z ?; are you Z to X ?; is it Z the X ?; mummy X ?; or X; think X; what does a X do ?; you Z the X	will, might, willn't, won't, could, so, doesn't, wouldn't, you'll, couldn't, gonna, he'll, should, we'll, must, probably, you'd, Anna, better, that'll, aah, childname's, daddy's, er, mummy'll, sheep, always, I’d, it'll, shouldn't, teddy, they'll, never, really, she'll, this'll, actually, ah, car's, Caroline’s, daddy'll, gotta, hasn't, he'd, horse, panda's, pig, they've, Thomas, wanna
Cluster 4	X , aren't they ?; Z a X; Z another X; Z at the X; Z be in the X; X do you want; X doing; X for X; Z for X; X in it; Z it on the X; Z it up X; Z some X; a Z of X; a X; all the X; an X; and what about X ?; are you going to Z the X ?; are you making a X ?; big X; build a X; can you find me the X ?; come on then , X; do you think Z like X ?; doing X; draw X; for X; funny X; get your X; give me the X; good X; her X; how many X ?; in a X; is that X; little X; make X; more X; mummy's X; need X; nice X; one X; other X; put the X on; red X; some X; that X; the X; this X; those X; what X; what do X eat ?; with X; your X;	dolly, fish, animals, horse, house, book, milk, hat, tea, teddy, water, bricks, cat, things, cars, eyes, egg, hair, Thomas, panda, shoes, trousers, bag, duck, sheep, toys, tractor, way, tower, cow, dress, drink, money, monkey, elephant, feet, picture, head, ball, bed, face, hand, nose, tiger, babys, bits, boat, chair, chicken, chips, lady, legs, something, table, cheese, dinner, ears, food, juice, pig, bottom, piece, clothes, cows, icecream, thing, top, tree, window, letters, paper, rabbit, socks, truck, banana, people, story, foot, mouth, name

Table 50. Some representative frames and words from the hierarchical clustering of the combined nested and full-utterance frames, N=290, locally context-sensitive parsing, 8 clusters.

Cluster	Frames	Words
Cluster 1	X a Z; X another Z; X her; X him; X it; X me; X one; X some Z; X the; X this; X what; are you going to X; can X; can i X; come and X; did X; did it X; did you X; do you want me to X it ?; don't X; from X; give it a good X; go and X; going to X; have to X; he can't X; i can't X; I’ll X; let's X; mummy X; she X; they don't X; to X; want to X; we X; what can you X ?; what did you X ?; what do you X ?; what do you want to X ?; what shall we X ?; what're we going to X ?; why don't you X ?; you have to X; you're going to X	play, buy, bring, pull, read, sing, cut, keep, tell, open, sit, use, watch, pick, stand, throw, fit, hold, leave, push, wear, break, stick, try, fix, help, drive, blow, stop, drink, show, wipe, hear, knock, let, ask, catch, jump, talk, wash, write, bite, count, fall, paint, remember, hide, lift, stay, tip, carry, finish, work, had, hit, sleep, cook, really, roll, start, call
Cluster 2	X , was it ?; X again; X baby; X boy; X can you see; X going; X like; X naughty; X right; X we ?; a very X one; and it's X; and what's X ?; are you a bit X ?; because he's X; can you see X ?; doing X; find X; have you X; he was X; he's very X; i can see X; i think X; i thought it was X; is it X; is this X ?; it's a bit X; like X; not that X; oh is she X ?; on X; she's not X; that one's X; that was X; there you go , X; they're all X; very X; what X; what does X say ?; what's he X ?; where's he X ?; yeah X; you were X	so, really, better, something, Thomas, coming, dolly, yours, broken, childname's, stuck, gonna, daddy's, cold, getting, Anna’s, when, quite, done, teddy, anything, actually, hot, as, dirty, Gordon, much, probably, Caroline, if, dolly's, everything, looking, outside, playing, things, wet, Anna, pink, mine, panda, crying
Cluster 3	X , aren't they ?; Z big X; X eat Z; Z some X; all the X; all those X; any more X ?; can i have some X please; do you like X ?; does Z like X ?; eat X; have you got some X ?; how many X ?; i haven't got any X; look at all those X; lots and lots of X; more X; no X; some X; these X; three X; what about these X ?; what do X say ?; you say X	animals, bricks, grapes, babys, cows, fish, things, cars, chicken, horses, bananas, people, wheels, chips, money, shoes, water,biscuits,letters,men, peas, sheep, trains, bits, bread, cheese, eyes,food, milk, monkeys, pennys, books, colors, pieces, toast, toys, birds, flowers

Cluster 4	Z a X; a X; a big X; a green X; another X; are they going to the X ?; are they having a X ?; are you a X ?; baby X; blue X; called a X; can i have the X please ?; can you see a X ?; did you like the X ?; do you need a X ?; does your X hurt ?; don't put your X in there; for your X; give me the X; have a X; i can see the X; I’ve got a X is he in the X ?; is there another X ?; it was a X; it's got a X; let's have a look at your X; move the X; not in the X; on my X; out of the X; poor little X; put them in the X; shall we get the X out ?; that one's a X; that's a good X; the other X; there's one X; there's only one X; this is a X; turn the X; we've got the X; what a X; what did the X do ?; what're you doing with your X ?; what's a X ?; what's happened to his X ?; where shall we put the X ?; who's on the X ?; with a X; you have a X	horse, house, hat, cow, tractor, book, cat, tiger, pig, bag, monkey, ball, chair, picture, truck, dress, fish, drink, boat, piece, lion, tower, cup, hand, nose, bus, duck, egg, garage, digger, foot, penguin, top, dolly, sheep, table, way, elephant, spoon, head, tree, face, fireengine, hair, rabbit, bath, teddy, leg, bottle, lorry, thing, bottom, giraffe, lady, basket, new, panda, tunnel, brick, chicken, trousers, lid, banana, bed, bird, doll, driver, farm, hippo, hole, mouth, shopping, white, baba

Table 51. Some representative frames and words from the hierarchical clustering of the combined nested and full-utterance frames, N=290, context-free parsing, 9 clusters

8.6 Quantitative results

Table 52 shows the results of categorization of nested frames under locally context-sensitive parsing. As was the case with full-utterance frames, all F scores and categorization Bookmaker scores are high, indicating that the algorithms were successful in categorizing word-frame instances from full-utterance plus nested frames. However, all scores were clearly lower than the full-utterance frame scores. The “trade-off” is that the coverage of the nested frames approach is higher, i.e. that more focal words were categorized this way.

	Hard	Fuzzy F	Fuzzy W	Fuzzy FxW	Confl	Pars
Accuracy	0.705 (*0.468*)	0.712 (*0.468*)	0.797 (*0.468*)	0.818 (*0.468*)	0.744 (*0.468*)	0.830 (*0.468*)
Completeness	0.563 (*0.374*)	0.606 (*0.399*)	0.654 (*0.385*)	0.743 (*0.426*)	0.825 (*0.519*)	0.744 (*0.420*)
F score	0.626 (*0.416*)	0.655 (*0.431*)	0.719 (*0.422*)	0.779 (*0.446*)	0.782 (*0.492*)	**0.785** (*0.442*)
Bookmaker	0.542	0.564	0.652	**0.719**	0.709	0.717

Table 52. Evaluation scores for nested and full-utterance frames, N=290, Locally context-sensitive parsing, 8 clusters.

By contrast, the results for full-utterance and nested frames under context-free parsing, shown in Table 53, are much lower than those for locally context-sensitive parsing, with F scores seldom exceeding their baselines by more than 0.15. Worst of all, the Confl algorithm failed catastrophically, its poor performance being due to its allocating all items to a single cluster (the tell-tale sign of this being the very high baseline score for

	Hard	Fuzzy F	Fuzzy W	Fuzzy FxW	Confl	Pars
Accuracy	0.696 (*0.527*)	0.667 (*0. 527*)	0.708 (*0. 527*)	0.716 (*0. 527*)	0.530 (*0. 527*)	0.605 (*0. 527*)
Completeness	0.442 (*0.334*)	0.479 (*0.378*)	0.440 (*0.327*)	0.514 (*0.379*)	0.992 (*0.985*)	0.571 (*0.499*)
F score	0.541 (*0.409*)	0.557 (*0.440*)	0.542 (*0.404*)	**0.599** (*0.441*)	0.691 (*0.687*)	0.587 (*0.512*)
Bookmaker	0.400	0.393	0.452	0.492	0.018	**0.525**

Table 53. Evaluation scores for nested and full-utterance frames, N=290, Context-free parsing, 9 clusters.

Completeness). The Bookmaker score is quite near zero, indicating that this algorithm performed at the same level as would be achieved by random guessing. The reason for this behaviour is probably that the context-free information was so “muddy” (i.e. contained so many uninformative words and especially frames) that there seemed to be reasonable evidence during conflict resolution to link essentially every item with every other item.

Because of the large data sets involved, *all* F and Bookmaker scores in Table 52 and Table 53 were significantly higher (as determined by randomization tests) than their random baselines. Even the disappointingly weak results for Confl were nevertheless shown to be much better than would have been achieved by chance, at $p = 0.01$.

The results of the comparisons between F scores and Bookmaker scores for the various algorithms are shown below. Table 54 shows the significance of differences in F scores, and Table 55 the significance of differences in Bookmaker scores, for locally context-sensitive parsing. Table 56 and Table 57 show the results for F and Bookmaker respectively, for context-free parsing.

For locally context-sensitive parsing, the three co-clustering algorithms performed significantly better than hard clustering, and Fuzzy F × W performed better than Fuzzy F or Fuzzy W. Fuzzy F × W also performed better than Confl at the 0.01 significance level for F, but only at the 0.05 level for Bookmaker. Pars performed significantly better than Confl in terms of F scores (but not Bookmaker scores), and the difference between Fuzzy F × W and Pars was not found to be significant.

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	0.088
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 54. Significance levels of differences in F scores for various clustering algorithms, for nested frames, N=290. locally context-sensitive parsing.
*** significant at p=0.05, ** significant at p = 0.01.**

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	↑ < 0.001**	↑ < 0.001**
Fuzzy FxW		← 0.029*	0.173
Confl.			0.106
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 55. Significance levels of differences in Bookmaker scores for various clustering algorithms, for nested frames, N=290. locally context-sensitive parsing.
*** significant at p=0.05, ** significant at p = 0.01.**

For context-free parsing, the results were more clear-cut. The ordering suggested by the absolute F and Bookmaker scores was confirmed by the randomization tests of significance, with Fuzzy F × W performing best, followed by Pars, then Hard F and lastly Confl. Fuzzy F × W performed significantly better than both Fuzzy F and Fuzzy W.

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	← < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	← < 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 56. Significance levels of differences in F scores for various clustering algorithms, for nested frames, N=290, context-free parsing.
*** significant at p=0.05, ** significant at p = 0.01.**

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	← < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	↑ < 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 57. Significance levels of differences in Bookmaker scores for various clustering algorithms, for nested frames, N=290, context-free parsing.
*** significant at p=0.05, ** significant at p = 0.01.**

These results seem to show the value of context-sensitive interpretation of the nested frames. It is better to take into account the greater context in which a word is being used, and even to avoid making a categorization decision when no additional knowledge is available from the surrounding context (i.e. preferring to make no decision rather than making an incorrect context-free decision).

As is apparent, the results for the combined set of nested and full-utterance frames are not as good as those obtained for full-utterance frames alone. This is surprising, given that many full-utterance frames point to the same "active ingredient", e.g. the noun phrase "the X" in "can you see the X", "that's the X", etc., so that pooling the information about which words go into these disparate frames should have strengthened clustering and categorization.

Part of the problem lies in the fact that some units discovered by the substitution test are not in fact viable constituents, and others are constituents only when they occur in specific contexts. For instance, one of the nested frames discovered was "X out", which is a valid constituent when it takes the form of a separable verb ("take out", "pull out", "hold out", etc.), but not when it occurs as part of a larger structure in which a noun argument precedes the particle of such a separable verb, e.g. "animals out" in "take the animals out", "hand out" in "pull your hand out", etc. These structures pass the substitution test, because "take the X" and "pull your X" are acceptable structures in their own right that can take single-word nouns as objects for the verbs in question; hence "X out" can be substituted for X in these cases, so that "take the X" and "pull your X" are treated as nesting frames for "X out". In fact, though, we are really dealing with a different (separable) verb.

In several cases, the nested frames are valid constituents, but are ambiguous as to the category of the words that occupy their slots. For instance, one of the discovered nested frames "X the door" is a prepositional phrase when it takes slot fillers such as "behind", "outside" and "towards", but a verb phrase (occasionally in the imperative form) when it contains the fillers "close", "lock", "mind" and "shut". This issue of frame ambiguity is of course already familiar from the previous discussion on full-utterance frames.

Nevertheless, the correctness of the categorization using nested frames (especially in locally context-sensitive mode) is still high, and a fair proportion of the corpus was covered. The frames considered here were a sufficient basis for bootstrapping the three main parts-of-speech of English.

The combined set of discovered nested and full-utterance frames includes such "prototypical" phrase structures as "don't X it", "too X", "very X", "the X", "your X", "this X", "another X", etc. It should be emphasized, however, that the identification of these structures was entirely due to the heuristic that phrases are often substitutable by single words, which is true for English but may not necessarily hold for other languages. One might want to postulate that all languages have this property; it would seem to facilitate learnability by making the "joints" in an utterance easily discernible.

The way in which this heuristic operates in practice is not always straightforward. Phrases can occasionally be substituted by single-word items from a different phrasal category: for instance, a noun phrase can be replaced by an adjective in sentence pairs such as "Are you [a little choo-choo train] ?" and "Are you [hungry] ?". In such a case, it might be argued that the algorithm was "right for the wrong reason". I would offer the counter-argument that, in fact, the shared use of the copula BE in these constructions indicates a deeper underlying semantic similarity between adjectival phrases and noun phrases. In Langacker's (1987) terms, for instance, adjectives (and by extension presumably adjectival phrases) signify atemporal relationships, while nouns (noun phrases) are represented as entities. What these two kinds of representations have in common, though, is their *static* nature; they are not conceived of as changing over time, and this seems to be the semantic characteristic that licenses the use of the copula BE.

This argument could also make more palatable the frequent conflation of adjectives and proper nouns that the algorithms in this thesis exhibit; recall that the presence of a large number of frames containing the copula often forces these two categories of words together into a cluster. On this view, a shared frame indicates similarity at some level, even if that level is an overarching one that recognizes stativity versus dynamicity, rather than the currently preferred level of nouns, verbs and adjectives. At the same time, the hypothesis that the appearance of two items embedded in the same construction always indicates some level of semantic similarity seems to be compatible with Croft's (2001) Radical Construction Grammar proposal (reviewed in Section 2.4.2), that the constructions are primary in determining grammatical category, and with Langacker's view (1987; reviewed in Section 2.4.1) that categories are grounded in the different mental operations involved in representing category members. On this view, the distinction between categories is ultimately underwritten by notional (semantic) criteria, but these criteria might be to a great extent reflected in the distributional data. The corollary is that finer distributional distinctions allow for finer semantic distinctions, and this is to some extent reflected in the current set of experiments, with full-utterance frames at 20 clusters producing groupings corresponding to mass nouns, plural count nouns, body parts and clothing, places and modal verbs, and with the adjective/proper noun cluster splitting into respective clusters for proper nouns, possessive forms of proper nouns, past and present verb participles, and other adjectives.

It would be a satisfying conclusion to this argument to be able to state that dividing the frames into two clusters lumps the nouns and adjectives together. Unfortunately, when two clusters are produced for full-utterance frames, for instance, the outcome is in fact that nouns stay separate, while verbs and adjectives are lumped together. This is almost certainly due to the "bridge" between these two categories provided by the large number of participial verbs that are commonly used in adjectival constructions with the copula, e.g. "Is it *broken*?", as well as verbal constructions such as the perfective "You've *broken* it".

The nested and full-utterance frame approach can be seen as an attempt to explicitly catalogue some of the most basic linguistic constructions in the input to the English-learning child. The notion of nesting structures inside others also allows the nesting structures, not used directly in the categorization experiments, to come to the fore as potential constructions that can accept material of arbitrary complexity into their Y slots. It is an interesting question whether this catalogue of constructions corresponds to the constructions that children may plausibly be familiar with at a comparable age. In fact, a manual listing of such common constructions was compiled by Cameron-Faulkner, Lieven and Tomasello (2003), based on the constructions attested in the Manchester corpus. Hence, it is possible to directly compare the frames produced by the current approach against the constructions identified by Cameron-Faulkner, Lieven and Tomasello (2003). The results of this comparison are presented in the next section.

8.7 A comparison with Cameron-Faulkner et al. (2003)

Cameron-Faulkner, Lieven and Tomasello (2003; for the remainder of this section, CFLT) manually analyzed a section of the Manchester corpus, in order to identify some of the most frequently-occurring frames in the input to children. In doing so, they showed that 51% of all utterances to children began with one of 52 specific sequences of 1, 2 or 3 words, and that 45% of all utterances began with one of 17 specific words.

CFLT give a very detailed listing of some of the most common item-based frames in the child-directed portion of the Manchester corpus. Because I have argued that the frames identified by automatic techniques in this thesis are often likely to be constructions of English, it will be instructive to investigate whether these frames produce an analysis of the corpus that is roughly similar to one derived manually by human experts. The work by CFLT is ideal for such an investigation, as it is also based on lexically-specific frames, and is derived from the same corpus. In this section I outline some differences between the CFLT approach and the current frame model, and then compare the structures discovered by the frame model against those manually identified by CFLT.

8.7.1 Differences from the current work

How frames are constructed: The work by CFLT takes a view that is quite different from that of the current nested-frame model regarding the criteria according to which a particular frame should be treated as a frequent item-based frame. Frames derive their identity from their initial 1, 2 or 3 specific words, and the remainder of the utterance is taken to be a slot that can be filled by any material. There are no restrictions on the set of words that may initiate a frame. Examples include “How many _?”, “There’s _”, “What shall _?”, with the underscore indicating a slot for filler material of arbitrary length. It does not seem to be critical for the authors that frames should be identified by their starting words only, and indeed they note that their approach misses frames that have item-specific material only at a later point in the utterance; the use of only the first three words may have been intended purely to simplify the analysis.

The lexically-specific frame model, on the other hand, builds its structures out of a specific and limited set of frame-building words, and these words can occur at any position in the utterance structure. One consequence of these two differences is that the frame model is likely to produce a great number of structures not considered by CFLT (notably, any frame that starts with a variable slot). Furthermore, many relatively general CFLT frames may be present in the frame model only in more lexically-specific forms that can be seen as special cases of the CFLT frames.

Semantics: CFLT characterize the speech of mothers to their children in terms of constructions. These constructions are often defined using categories for the fillers, e.g. "It's [NP/Adj]" is assumed to be one construction, and "It's [VP]" another. Presumably, the authors intend these to be constructions that would be available to an *adult* speaker of English.

In order to learn constructions as CFLT have defined them, it seems that the child needs to be sensitive to two kinds of information. Firstly, they need to keep track of word sequences which occur very frequently in the input, and will form the fixed part of the frame; as noted, these lexically-specific sequences always occur at the beginning of an utterance in the current formulation; however, a more general approach might be to take note of recurrent sequences that appear at any position in an utterance, and to construct a frame around these. The second source of information to which children should be alert is semantic information about the meanings of the fillers, so that they can distinguish between constructions that have exactly the same specific words, but which differ in the category of the filler (as in the "It's _" example above, where children would need to be able to identify a filler word or phrase as either a VP or an NP/Adj in order to distinguish between the two constructions).

It is not clear whether CFLT intended their set of frames to be viewed as an implicit model of the kinds of constructions that children should be learning. However, since a key aspect of CFLT's work is the correlation between adult and child productions, this assumption seems reasonable. In that case, both formal and semantic aspects would be germane to language learning in their model (although the analysis seems to emphasize the formal aspects).

By contrast, the lexically-specific frame model ignores semantics in its current formulation, attempting to discover the categories such as Noun, Verb and Adjective from their distributional usage across frames only *after* these frames have been identified, rather than by presupposing that these categories have already been sorted out on semantic grounds and can be used to distinguish between two otherwise identical frames. In practice, this means that the set of frames of the current work will include items such as "It's Y", collapsing across the CFLT distinction between "It's [N/Adj]" and "It's [VP]".

Occurrence frequency: The two models also differ in their requirements regarding the frequency of a frame and/or its fillers. CFLT made use of all frames such that at least one mother used a particular frame at least four times in a two-hour sample taken from each child; this was termed their "4+ criterion". The authors do not require that the set of filler material needs to exhibit any particular level of variability. In the current work the entire Manchester corpus was used (containing up to 34 hours of data for each child). The criterion for a nesting (resp. nested) structure to be included in the data set is merely that it should have occurred as a nesting (resp. nested) frame for at least *V* nested (resp. nesting) frames. In the experiments reported in this thesis up to now, *V* was set to a value of 5; for the current comparison to CFLT, *V* is set to 2. It may be argued that the CFLT criterion is somewhat stricter for frames (though less strict for fillers), and so one might expect that the set of frames in the current approach would be a superset of the CFLT set.

8.7.2 Comparison of frames

The frames listed by CFLT are all full-utterance frames. When comparing these frames against the current frame approach, there are three kinds of frames that can be considered to represent the structure of full utterances:

- A frame which has been identified as the *nesting* frame for several other nested frames
- Any (potentially recursive) embedding of nested frames inside their corresponding nesting frames; for the current analysis I will not consider recursive embedding, only single-level embedding, in line with the locally context-sensitive parsing approach used above
- A full-utterance frame (i.e. single-word fillers only), as obtained via the methods of Chapter 6.

The following examples illustrate some of the issues involved in mapping between CFLT and the current full-utterance and nesting/multiply-nested frames:

(1) The "best match" possible for a frame against a CFLT frame would be a nesting frame, as nesting frames are (theoretically) able to accept arbitrary material. As an example, one of CFLT's frames is "draw [NP]", and one of the nesting frames produced by the current model is "draw Y", which has been attested to take as its filler a variety of nested frames. Note firstly that in the current model, not all instances of CFLT's "draw _" would be regarded as instances of the current "draw Y", because recognition of the nesting frame is conditional on recognition of these specific nested frame fillers. Secondly, as noted above, not all fillers can be expected to be noun phrases, because semantic information is not being used. Thirdly, it should also be borne in mind that nesting frames can themselves occur nested inside large contexts, so that they are not *necessarily* descriptions of full-utterance structure, while the CFLT frames are.

(2) Another CFLT frame is "that one's _". The corresponding nesting frame in the current approach would have been "that one's Y"; however, this frame does not occur as a nesting frame. Instead the nesting frame "that Y" contains in its filler set the nested frame "one's Y", so that the recursive combination "that [one's Y]" covers the CFLT

frame (even though a syntactic segmentation at the juncture suggested by this frame would seem to be incorrect).

(3) A different example is “how did _”. The corresponding nesting frame in the current approach would have been “how did Y?”, which, once again, does not occur as a nesting frame. The nesting frame “how Y” does not take the nested frame filler “did Y” either; however, it does take the frame “did you Y”, producing the recursive combination “how [did you Y]”. While this is a more specific frame than the CFLT frame, it is “as close as we’ll get”. The fact that the frame structures may be more specific than that identified by CFLT does not necessarily indicate a difference in the actual utterances covered; for instance, it may be that a majority of the utterances covered by their “How did _ ?” are actually cases of “How did you _?”. (Nevertheless, note that this would necessarily be at best a majority of cases, as CFLT would otherwise have chosen the frame “How did you _” instead).

(4) The CFLT frame “What does _” is not covered by the nested frame-approach at all; however, there are several full-utterance frames that are more specific than that frame and are therefore specializations of the frame (with single-word fillers). These include “what does X say?” and “what does a X do?”.

Cutting across the division into four kinds of frame matches, therefore, we have the distinction that some frames will cover a CFLT frame exactly (as in examples 1 and 2), while for other frames, only specializations of those frames are produced by the lexically-specific frame approach (examples 3 and 4). In the following comparison, I will describe each CFLT frame according to whether it is covered exactly by the frame approach, covered only by specification, or not covered at all.

CFLT divide their frames into the following categories: fragments, Wh-questions, yes/no questions, imperative constructions, copula constructions, transitive constructions, intransitive constructions and complex constructions. In this section, I will show the detailed comparison for Wh-questions only, followed by a summary comparison for all categories of CFLT frames. Tables showing the detailed comparison for all categories can be found in Appendix 2.

8.7.2.1 Detailed comparison of Wh-questions

There are 31 Wh-question frames in CFLT, of which 11 were identified by the researchers as “core” frames (occurring 4 or more times in the speech of 6 or more of the 12 mothers). The leftmost column in Table 58 displays the CFLT frame. The second column contains the nesting frame(s) that match the CFLT frame, if any. The third column contains multiply-nested frames that cover the CFLT frame, if any, and the fourth column contains full-utterance frames that cover the CFLT frame. Frames that match the CFLT frame exactly are displayed in bold. In total, only 7 of the Wh-question frames are covered exactly, 19 are covered indirectly and 5 (“what were _”, “whose [N]”, “why not _”, “what kind of _” and “what number _”) are not covered at all.

Wh-questions			
Cameron-Faulkner et al. (2003) frame	**Nested frame**	**Multiply-nested frame**	**Full-utterance frame**
what’s _	**what’s Y?**		what's X ? what's X doing ? what's X now ? what's X there ? what's a X ? what's <child’s name> X ? what's happened to X ? what's happened to his X ? what's happened to your X ? what's he X ? what's in that X ? what's in the X ? what's in this X ? what's in your X ? what's it X ? what's on the X ? what's on your X ? what's she X ? what's that X ? what's that X called ? what's that X doing ? what's the X ? what's the X called ? what's the X doing ? what's the matter with your X ? what's this X ? what's this X doing ?

what're _	what're we Y? what're you Y?		what're we going to X? what're you X? what're you X about? what're you X for? what're you X now? what're you doing with your X? what're you going to X?
what do _		what [do you Y]?	what do X do? what do X eat? what do X say? what do you X? what do you want for your X? what do you want to X?
what did _	what did you Y?	what [did you Y]? what [did we Y]?	what did X do? what did the X do? what did you X?
what colour _			what colour are the X? what colour is X? what colour is the X?
what (ha)s _	same as for "**what's _**" above		
what about _		**what [about Y]?**	what about X ? what about a X ? what about her X ? what about his X ? what about some X ? what about that X ? what about the X ? what about the other X ? what about these X ? what about this X ? what about your X ?

what shall _			what shall we X?
what can _			what can you X? what can you see on the X?
what does _			what does X say? what does a X do? what does a X say? what does the X say?
what happened _		what [happened to Y]?	what happened to the X? what happened to your X?
what were _	**Not covered**		
what’ve _		what [have you Y]?	what have you X?
what kind of _	**Not covered (see text)**		
what number _	**Not covered (see text)**		
where’s _	**where’s Y?** where’s the Y?		where is X ? where is the X ? where is your X ? where's X ? where's X going ? where's X gone ? where's a X ? where's he X ? where's her X ? where's his X ? where's my X ? where's that X ? where's that X gone ? where's the X , <child’s name> ? where's the X ? where's the X going ? where's the X going to go ?

			where's the X gone ? where's the X then ? where's the little X ? where's the other X ? where's your X ? where's your X gone ?
where'd _		where [did we Y]?	where did the X go?
where're _		where [are you Y]?	where are you X? where are the X? where are your X?
where shall _			where shall we put the X?
who's _	**who's Y?**		
whose [N]	**Not covered**		
who're _		who [are you Y]?	
who did _		who [did you Y]?	
why don't _			why don't you X?
why do _		why [do you want your X]?	
why's _		why [is he Y]? why [is it Y]?	why is he X?
why not _	**Not covered**		
how many _			how many X? how many X are there? how many X have we got? how many X have you got?
how did _		how [did you Y]?	
which one _		**which [one Y]?**	which one's X?

Not covered	which Y? who Y? why Y? how Y?		where was the X? which X? which X is it? who X a X? who X it? who X the X? why are you X?

Table 58. A comparison of the Wh-question frames identified by Cameron-Faulkner et al. (2003) against the frames produced by the current approach.

The exercise of comparing the current frame approach against that of CFLT uncovered a slight shortcoming in the algorithm used for full-utterance and nested frames, which could be rectified in future extensions to this algorithm. This was highlighted by the inability of these algorithms to discover the (intuitively important) frames “What kind of _” and “What number _”. The data set for full-utterance frames contains a large number of two-slot frames such as “what X of X is it?”, “what X did you X?”, where the first slots are typically filled by words such as “kind”, “sort”, “type”, etc, and so one would expect, say, “What kind of X is it?” to have been a frame. There were two factors working against this outcome, however. First and foremost, “kind” is not a frame-building word, and so the frame could not have been formed for consideration. However, one might still have expected “what X of X is it?” to have been produced as a frame, with words such as “kind”, “sort” etc. as fillers for the first slot. In fact, there was a frame based on these utterances which ended up in the final data set; however, this frame was based on the *second*, not the first slot (“what Z of X is it?”). The frame based on the first slot was dropped from the data set because it could accept only a narrowly restricted set of fillers (specifically, “kind” and “sort”) in that slot.

This case shows that the algorithm should be modified in the case of frames with multiple slots. A frame with multiple slots can still be “saved” as a useful frame if one of its slots does not accept a variety of fillers; the variability would have to be located in its other slots instead. The solution in the case of “what X of X is it” would be, whenever a frame fails the 5-5 criterion, to generate new frames for each of the particular words that go into the first slot, i.e. generate the two new frames “what kind of X is it” and “what sort of X is it”, and then proceed as before: if these frames now take the required number of different fillers into their (only) slot, then they are viable as frames for the final data set; if not, they are dropped from consideration. This approach would probably have allowed “what kind of X is it” to have entered the final data set, and hence the CFLT frame “What kind of _” would have been covered by the frame approach.

At the end of Table 58, there is a row of frames covered by the current approach but not by the CFLT approach. In most cases, CFLT identified other frames that started with the first words of these frames, but these particular frames simply did not reach the 4+ criterion for inclusion.

8.7.2.2 Summary comparison of all frames

Table 59 indicates the breakdown of all CFLT frames into the different frame types of the last few chapters. A majority (102 out of 152) of the CFLT frames were found directly in the set of frames produced under the full-utterance-plus-nested-frame approach. Only 23 of the 152 frames used by CFLT were not covered by any frame in the current approach.

Frames covered directly	Frames covered indirectly	Frames not covered
102	**27**	**23**

Table 59. Summary of coverage of Cameron-Faulkner et al. (2003)’s frames.

For the most part, the non-covered frames were ones where one or more of the lexically-specific words were not in the list of the most frequent, frame-building words used by the frame approach, and so could not have been formed (e.g. “whose _”) In a few other cases, the frame could not have been recognized as a frame, because it consisted entirely of frequent words, and so would have been treated as a fixed expression. Table 60 breaks down the reasons for failure in the 23 cases not covered by the frame approach. Two frames (“[Pron] isn’t” and “there [Pron] go”) were not found because they consisted entirely of frame-building words in the current context (all the nominative and possessive pronouns were members of the set of frequent words), and were therefore regarded as fixed expressions. Twelve others were not identified because their lexically-specific components contained words that were not on the list of frame-building words. The reasons why the remaining 9 cases were not identified are mostly unknown.

Fixed expressions	Non-frame-building words	Other
2	**12**	**9**

Table 60. Reasons why some frames identified by Cameron-Faulkner et al. (2003) were not generated by the frame approach.

When considering only the core frames (Table 61), all but 2 of the 52 core frames are covered by the frame approach. These two were the already-mentioned “there [Pron] go”, which was treated as a set of fixed expressions, and the complex frame “if _”, which was not covered because “if” is not a frame-building word. This may in turn have been due to the fact that utterances starting with “if” tend to be rather heterogeneous and complex in structure, while the current approach requires a reasonably-sized set of single-word fillers for “if X”, something which was clearly not present in the corpus. In addition, 7 core frames (all Wh-questions) were represented indirectly rather than directly by the frame approach. All other core frames were represented directly.

Core frames covered directly	Core frames covered indirectly	Core frames not covered
43	**7**	**2**

Table 61. Summary of coverage of Cameron-Faulkner et al. (2003)’s core frames.

An important point should be made about the implicit use of the nested frame approach as a generative model. In the above example, “have you Y” takes a number of nested frames as potential Y slot fillers, including some fillers which are appropriate for “have you Y?” as a full utterance, but not for “have you Y” in a nested context such as “What [have you Y]?”. Examples include “X it”, “got X” (which are more likely to take direct objects when they are embedded in the full utterance “Have you Y?” than indirect objects as in “What have you Y?”), and “any X”. Therefore the model over-generates, producing sentences that are not English, such as “What have you any X?”. This is not a fault of the model, but of the common assumption in linguistics that generative models are the *sine* *qua non* that any researcher should aim for. On this view, the language-learning child is apparently engaged in randomly generating “well-formed sentences” that are disconnected from their meaning. But in fact, the child is likely to produce only utterances by which she means to achieve a communicative purpose, and usually this means assigning meaning to the components of the utterance as well as the structure of the utterance as a whole. If the child can assign no coherent meaning to, say, “What have you any wool?” that she could not assign to some other simpler utterance that she has actually encountered before (such as “Have you any wool?”), then the incorrect structure is likely to be “blocked” by the previously-heard correct structure. This kind of blocking mechanism is of course not part of the current model, nor is semantics of any kind.

Instead, I would suspect that children will produce errors that occur as a result of nesting frames inside others, but that these errors will only occur when there is a coherent meaning that could plausibly be ascribed to the new production, and the child has not yet associated that meaning with a different (correct) form which would block the new production.

As for comprehension, obviously children will interpret only utterances that they hear, and the malformed utterances generated by unconstrainedly combining frames are not likely ever to be uttered; hence there is no problem with a model that could potentially find these non-utterances acceptable if they should occur.

8.8 Discussion

This chapter has presented the results of a technique to extend the set of full-utterance frames from Chapter 6. The technique generalizes the single-word slots of pre-existing full-utterance frames so as to allow multi-word fillers. Setting up a hierarchy of frames that are schematic for each other in this way allows the identification of nested and nesting frames; the most flexibly-used nested frames that occur in the most flexibly-used nesting frames are taken to be reliable frames for the purpose of lexical categorization. When used in a locally context-sensitive manner, these frames are successful in categorizing focal-word instances into each of the three main parts-of-speech.

Quantitative evaluation measures are not as high as for the full-utterance frame approach; on the other hand, a larger proportion of the corpus is covered.

The last three chapters have explored the use of a heuristic that dichotomizes words into a frequent and a less-frequent word group. It may also be fruitful to consider other ways in which frames may arise. The next chapter considers an approach in which frames are based on predictable relationships of co-occurrence between pairs of words.

9 Prediction-based frames

9.1 Introduction

In the previous chapters, a number of algorithms were presented that took as their starting point a basic dichotomy between frequent and infrequent words (loosely corresponding in practice to function and content words respectively). Frames are presumed to form as utterance skeletons constructed out of frequent words alone. On the other hand, the words occurring in the X slots of frames play no role in the creation of these frames, other than serving in the role of slot-fillers.

Under this view of frame formation, a noun phrase structure such as "the X" is discovered by postulating that "the" is a special, structural word in English, and then noticing that "the X" is used as the structure of several full utterances and nested linguistic constituents. It is presumed that the child is able to identify the frame-building words, based largely on the basis of their very frequent occurrence in the corpus.

An entirely different model of frame formation would be one where the child does not divide words into two groups and treat them qualitatively differently, but where frames form as a result of the child noticing regular co-occurrences between pairs of words, regardless of which particular words they are. With repeated exposure to these word pairs, these word configurations become elements of memory in their own right.

Whereas in the previous frame discovery procedures, all frames (made up out of frequent words) that occurred frequently enough, and with a wide enough range of fillers, were considered as viable frame candidates, the procedure outlined in this chapter will be based on the statistical predictability of one word from another. If word *x* can be statistically predicted to occur if word *y* has occurred, based on the frequency with which *x* and *y* occur together in utterances, then it is likely that this co-occurrence of the two words is not due to chance, but reflects underlying linguistic structure in the utterances in question; the two words are likely to be part of a larger unit.

These larger units may either make up part of a frame, or part of an extended filler. So for instance, the words “television set” often occur together (expressions such as these are also termed *collocations*), and may constitute a phrase than can be the filler of the frame “the X”. On the other hand, the sequence “that’s a lot of” is also a collocation, but one that one could imagine to provide the structure for the phrase “that’s a lot of X”.

The other aspect in which the current frame discovery procedure differs from the full-utterance frame and nested frame procedures is that those techniques started with full utterances, discovered a number of frames for these and then divided them into smaller nested constituents, in line with Peters’s (1979) “Gestalt” approach to language learning. The current, *prediction-based* frame discovery technique will take a more “analytic” approach: units will be built up from smaller to larger units. Hence, the technique will not make use of the segmentation boundaries provided by the edges of the utterances, but will have to determine its own frame boundaries; these will be determined from the statistical predictability of elements for each other, as will be discussed below.

9.2 Misconceptions about predictability

There are a number of unwarranted assumptions associated with the term “predictability” which need to be avoided. The first issue relates to the role of statistical predictability in human cognition. Perruchet and colleagues (e.g. Perruchet & Vinter, 1998; Perruchet, 2008) have pointed out that behavioral sensitivity to statistical properties of the input does not necessarily mean that humans overtly *calculate* statistical properties during processing. Instead, the behavioral sensitivity may be an emergent effect of underlying processing. For instance, Perruchet and Vinter (1998) have shown in their PARSER model how the abilities of infants to segment statistical words from an extended speech stream (Saffran et al., 1996) may be simulated by a process that chunks together units (initially syllables) if they co-occur (producing putative words), and exhibits interference between chunks that have material in common (so that these chunks are effectively different candidate words that compete for the same phonological material).

In the current work, I am using the statistical predictability of one item from another as a *heuristic* guide to the strength of association between the items, regardless of whether this association is grounded in underlying calculation of statistics, or in the chunking together of these items.

The second unjustified assumption is that the only valid numerical formulation for the predictability of one event from another is *conditional probability*. In fact, Shanks (1995) has argued and provided a great amount of empirical evidence to show that a more valid expression of predictability can be obtained by using ΔP, a measure which *discounts* the probability that B will follow A, by the probability that B will occur even when A does not occur. In terms of the cells of Table 4 in Section 5.5.2, ΔP is given by

$$\Delta P(A,B) = \frac{a}{a+b} - \frac{c}{c+d}.$$

This measure gives a truer value of A as a predictor of B than conditional probability does; if B is more likely to occur after A than when A is absent, then A is a good predictor of B. ΔP will decrease if there are several other predictors of B than just A alone. If, say, 30% of all occurrences of A are followed by B, and 30% of all non-occurrences of A are also followed by B, then it is clear that there is no real relationship between the events A and B – B just occurs 30% of the time anyway, regardless of whether A occurred, and ΔP will be zero.

In a discussion of ΔP and related statistics, Perruchet and Peereman (2004) point out that an analogous measure may be postulated to describe a kind of backward ΔP, where the roles of events A and B are reversed (so that we "back-predict" A on the basis of B). They name this measure ΔP′, defined as

$$\Delta P'(A,B) = \frac{a}{a+c} - \frac{b}{b+d}.$$

This measure expresses the probability that B is preceded by A, discounted by the probability that the non-occurrence of B is also preceded by A. If A is more likely to occur before B than before the absence of B, then B is a good predictor of A. If, say, 30% of all occurrences of B are preceded by A, and 30% of all non-occurrences of B are also preceded by A, then there is no real relationship between the events A and B – A just occurs 30% of the time anyway, and the occurrence of B does not predict A, so that ΔP will be zero.

Note that ΔP and ΔP' correspond respectively to Powers's (2008) Markedness and Informedness, discussed in Section 5.5.2.2. Mathematically, it should be clear that ΔP' (A, B) is just equal to ΔP(B, A) when the rows and columns are interchanged, i.e. ΔP' is just the "right-to-left" counterpart of the "left-to-right" ΔP. Perruchet and Peereman (2004) have shown that both ΔP and ΔP' (equivalently, both "left-to-right" and "right-to-left" ΔP) are implicated in human language processing.

It is possible to calculate ΔP in an explicit iterative simulation, by updating the strength of an associative link between two words in accordance with a *delta learning rule* similar to that formulated by Widrow and Hoff (1960). For instance, Shanks (1995) formulates the delta rule as stating that the change in association ΔV from element A to element B is given by

$$\Delta V = \alpha\beta(\lambda - \Sigma V),$$

where α and β are learning rate parameters that increase with the salience of A and B respectively, λ is 1.0 when B is present and 0.0 if it is absent, and ΣV is the association strength of all cues present on the trial (i.e. in the current simple case, neglecting background cues other than A, it is equal to the association strength V from A to B). The value of V is updated only on trials where A is present (i.e. when α is non-zero), so that only cases allocated to cells *a* and *b* in the contingency table alter the value of V. It can be shown that the asymptotic value of V under the delta learning rule is equivalent to ΔP (Chapman & Robbins, 1990).

The process of strengthening an associative link when it is confirmed and weakening it when it is disconfirmed is also the essence of the PARSER model (Perruchet & Vinter, 1998), so that the expected values of associative links in PARSER might arguably be expected to be proportional to ΔP (or perhaps to correlation, which is the geometric mean of ΔP and ΔP'), and evidence from Perruchet & Peereman (2004) suggests that this may be the case. As will be discussed later, an important difference between PARSER and the current work is that it will be suggested here that associative links can be asymmetric, whereas links are implicitly symmetric in PARSER.

By updating the associative strength from A to B in this way, it might be possible to devise an iterative computer simulation of the formation of frames from associative links. Note that it is controversial whether chunking can in fact be reduced to associative mechanisms (Shanks, 1995; see also Perruchet & Pacton, 2006).

The third unjustified assumption is that prediction necessarily entails making a prediction about "what will happen *next*", i.e. that we are using the past in order to anticipate/prepare for an event in the future. In fact, a more encompassing meaning of prediction entails simply using some available information as a premise from which to draw the conclusion that some other statement about the world is also true, regardless of which of the two events is taken to occur first. This is in line with the idea that when items are associated with each other, the statistical predictability expresses the degree or strength of association.

Association can, of course, be used for prediction in certain circumstances. So for instance, during speech processing, even while words are being processed, some information about the phonological details of previous words is still available for a limited time. If a listener finds part of an utterance unclear, he or she can use contextual information about the rest of the utterance in order to infer what was said during the unclear section, even using material which was spoken after the unclear section. Given most of the remainder of the utterance, hearers can use that context to predict what the unclear material would have been. This happens under normal hearing conditions too; for instance, Bard, Shillcock & Altmann (1988) let subjects hear samples of spontaneous conversation, in stretches which were incremented by a word at a time. Twenty percent of all words could only be recognized when further context (i.e. subsequent words) was provided. In this chapter, the concept of prediction has this more general, atemporal meaning.

9.3 Predictive relations as indicators of frame-filler relations

The intuition captured by the frame discovery algorithms presented in the previous chapters is that, to a large extent, the function words of English form the backbone of the language. In everyday phrases such as *is sleeping*, *the bell*, *very wrong*, it seems that the structures of these phrases can be captured by the abstractions *is X –ing*, *the X*, *very X*. These function words and morphemes are clear indicators of the part-of-speech of the word that they abut. In a sense, these functors owe their presence to the part-of-speech of the adjacent word: in linguistics, the relationship between the so-called head of a phrase and associated functors is described as one of *dependency*, in that, for instance, *the* in *the bell* seems to be licensed to occur only because of the occurrence of a noun soon after it in the phrase, so that *the* is dependent on the noun *bell*.

The concept of linguistic dependency is closely related to that of predictability. If we know that we are dealing with a word used as a noun, there are a number of dependent elements that we might expect with fairly high probability to see occurring close to it, such as *the*, *your*, *another*, etc. before the noun, or the plural or possessive markers *–s* and *'s* after it. In other words, there is a certain amount of *predictability* about the dependent elements given the non-dependent element (or at least, its part-of-speech).

However, as will be shown later, the concept of predictability considered here is not exactly the same as dependency in linguistics. Part of the focus in this chapter will be to combine basic predictability relationships into a sequential pattern that constitutes a frame. One step in that process will be to attend to words that predict *the same* other word, and linking them into a larger frame. In linguistics, this would be tantamount to suggesting that the mutually-predicted word is dependent on two other words simultaneously, which would not be regarded as coherent (for instance, this would make it impossible to depict the situation by means of a syntactic tree).

The extent to which one word predicts another can easily be determined by applying the ΔP and ΔP' functions discussed in the previous section to contingency tables derived from the frequency of co-occurrence of the two words.

Take as an example the archetypal construction "the X", where X stands for a noun; and to take a concrete example of "the X", consider the phrase "the dog". We might expect that if the word "dog" occurs, there is a reasonably high probability that the word preceding it is "the". This is because there are only a handful of determiners and quantifiers (such as "the", "a", "another", "your", some", any", etc.) that are highly likely to precede "dog". Hence, if "dog" occurs, it is a reasonably successful predictor for the (prior) occurrence of "the". (Of course, the prediction is far from certain; however, it is a great deal stronger than most predictive relationships between words.) By contrast, when given that the word "the" has occurred, there are a great many words (nouns as well as adjectives) that could follow "the", and so it is not easy to predict from the occurrence of "the" that "dog" will follow.

This asymmetry in predictability seems to capture the essence of a lexically-specific frame: the lexically-specific element is predicted by the slot-filler element, but starting from the lexically-specific element it is not possible to predict the slot-filler with any degree of reliability.

Following the convention that the predictor appears on the rows of the contingency table and the predicted item on the columns, we can use the contingency table depicted in Table 62 to determine whether, in a two-word sequence with the word "the" in the left-hand position and "dog" in the right-hand position, we can reliably predict the occurrence of "the" from the occurrence of "dog".

	the	~ the
dog	a	b
~ dog	c	d

Table 62. The contingency table to determine whether "dog" predicts "the". The tilde indicates logical negation, i.e. "~ dog" indicates any word in second position other than "dog".

In line with the above discussion, the conditional probability $\frac{a}{a+b}$ of the occurrence of "the" given the occurrence of "dog" is likely to be high, while the conditional probability $\frac{a}{a+c}$ of the occurrence of "dog" given the occurrence of "the" is likely to be low. At the same time, the term $\frac{c}{c+d}$, representing roughly the probability that words in general (i.e. other than "dog") are preceded by "the", is not likely to take on as high a value as $\frac{a}{a+b}$, because there are many words (notably verbs) that are never preceded by "the". Hence, the value of ΔP from "dog" to "the", given by $\frac{a}{a+b}-\frac{c}{c+d}$, will be relatively high, while the value of ΔP from "the" to "dog", given by $\frac{a}{a+c}-\frac{b}{b+d}$ (which is the same as ΔP' from "dog" to "the"), is relatively low (the remaining term $\frac{b}{b+d}$, roughly giving the probability for words in general to precede "dog", is likely to be negligible).

Hence, we can hypothesize that an operational technique for identifying frame-filler relationships is to look for asymmetries in the ΔP value, i.e. for two elements A and B such that there is a high ΔP value from A to B and a low ΔP value from B to A. In this case, B is the filler item for the partial frame constituted by A. The two items can then be written in frame fashion, where B is the lexically-specific word and A is the slot-filler. So the current example would yield the frame "the X". Because several of these asymmetric word pairs will eventually be combined into a single frame, a single pair of words exhibiting asymmetry in their ΔP value will be called a *frame primitive*.

It is worth noting that the ΔP relationships that will be calculated here can hold for elements that are not adjacent to each other. This is in line with Gómez's (2002; Gómez & Maye, 2005) work on non-adjacent dependencies, and on work by Pacton & Perruchet (2008) that shows that adults are aware of the relationships between elements of a sequence if they have attended to the elements, regardless of whether the elements occur adjacently or not.

In this chapter, when it is stated that an associative link exists from one word to another, what is meant is that the strength of the association, as calculated using the formula for ΔP, is *relatively high*. In practice, an arbitrary threshold value is applied, and only associations stronger than the threshold value are recognized.

9.4 Assembling complex frames from individual associative links

The discussion above considered the case of a single associative link from one word to another word. In this simple case, where only the two words in question are presumed to make up the entire utterance, the frame is a frame primitive, made up of those two elements only (with one serving as a slot). In this section, we consider how more than one frame primitive in a particular utterance may be combined in order to form a larger frame.

In these complex cases, there might be several associative links originating from an element, or terminating at a certain element, and some pairs of elements might have strong mutual associations between each other.

There are a number of issues to consider:

1. Equivalence relationships: In some cases, two elements each have strong associative links to each other. For instance, in the frame “out of X”, the words “out” and “of” might have high ΔP values for each other. Intuitively, the two words seem to form a kind of collocational unit when they occur together, with “out of” having more the character of a “long word” than of a combination of two independent words (compare “into”). In such cases, “out” and “of” can be said to exhibit an *equivalence relationship*, rather than one of prediction.

   Equivalence-linked words may be disjunct; for instance in phrases such as “a kind of fruit”, “a lot of cereal”, “a sort of aeroplane”, there is likely to be an equivalence relation between “a” and “of”, and the frame in this case would be “ a X of X”.

Not only pairs of frame-building words, but also pairs of slot-fillers may be linked by an equivalence relation; in phrases such as “take it out”, “hand it over”, “kiss it better”, there is likely to be a mutual association between “take” and “out”, between “hand” and “over”, etc., and the frame would be “X it X”.

In all of these cases, the words that enjoy equivalence relations to each other should be treated as units, so that if one word is included in a frame, the other should be too.

2. Transitivity: Complicated situations may arise with extended chains of words that exhibit predictive or equivalence relations to each other. For instance, it may happen that word A is equivalent to word B, and word B is equivalent to word C, without there being an equivalence relationship from A to C. The question is whether we treat equivalence as a transitive relation (in which case we would accept that A is equivalent to C even if this is not borne out by the ΔP values between them). Other situations which need to be considered include ones where A predicts B and B predicts C but A does not predict C, or where A and B are equivalent and A predicts C, but B does not predict C. Particularly problematic are *cycles*: for instance, A predicts B, B predicts C, and C predicts A.
3. Implicit links: In some cases, two words which are not directly linked via equivalence or predictive relations can nevertheless be implicitly linked via a third word to which they both are linked. One example would be a situation where A and B are not linked, but both A and B predict a third word C, which serves as the implicit link between them. Another situation is where word A predicts both words B and C, which are not linked to each other.

9.4.1 Basic frame patterns

In order to create a specific implementation of the prediction-based frame idea, however, certain decisions need to be made about how several linked element pairs are combined into a larger frame. In this section, I will outline the specific decisions made in the current work with regards to the issues outlined above. Alternative implementations may make different choices from the ones made here.

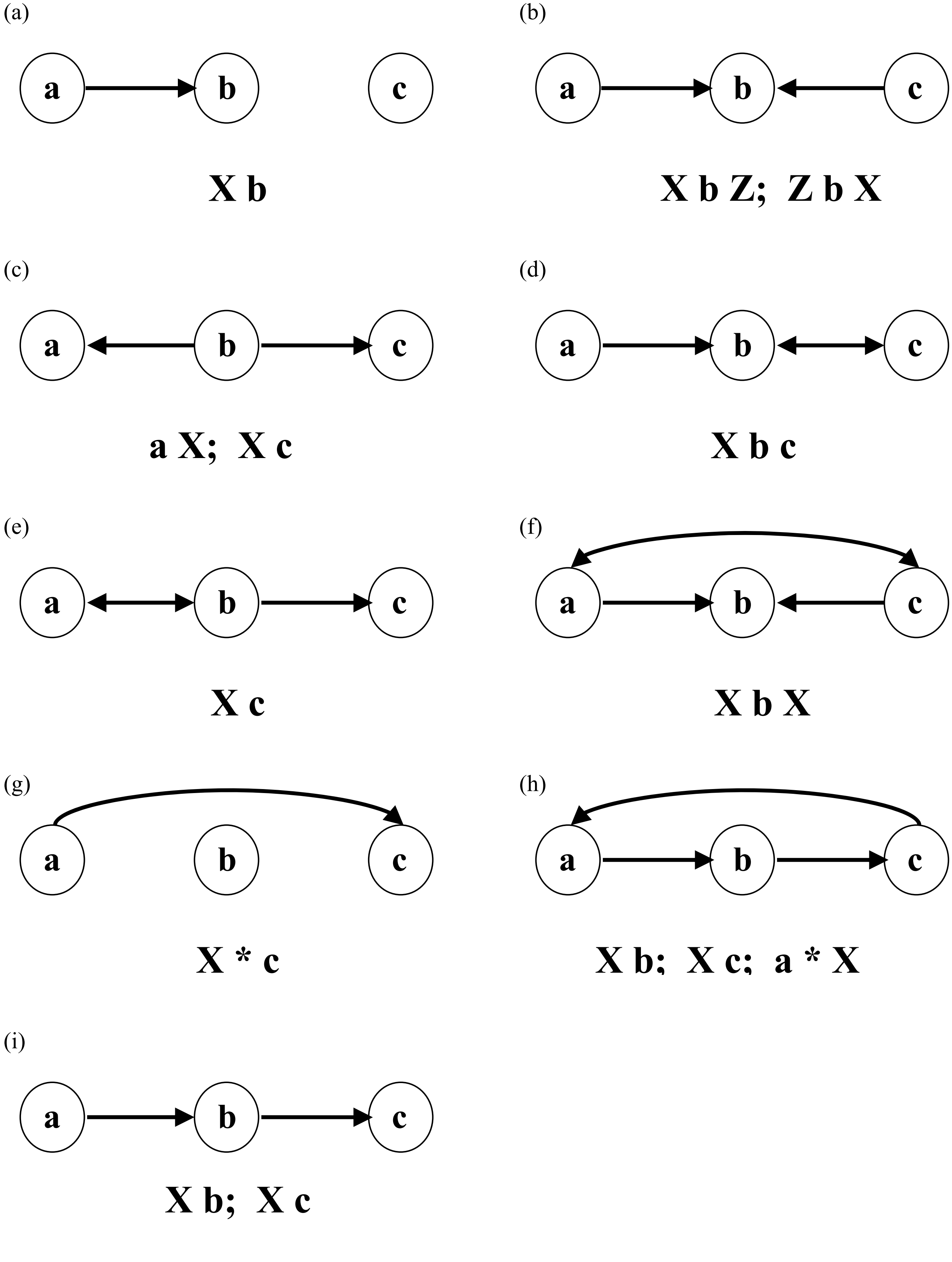

Figure 8. Examples of prediction-based frames. The arrows shown here point towards the predicted element.

In order to demonstrate the effect of these decisions, I identify a number of "primitive" frame patterns that are the building blocks of the frames that will be identified in this experiment. For each pattern, we have an utterance or utterance fragment consisting of the three words "a b c" in sequence. The relationships that exist between these words determine the topology of the frames that we identify.

Each of the patterns is illustrated diagrammatically in Figure 8a-i. The relationships between words are denoted schematically in the figures by means of arrows. Single-headed arrows indicate a high *unidirectional* ΔP value in the direction of the arrow, i.e. there is a predictive relationship from the source word to the target word. Double-headed arrows indicate high ΔP values in both directions, i.e. there is an equivalence relationship and the two words are treated as part of the same unit.

In Figure 8(a), *a* predicts *b*. Hence, there exists a relationship in which *b* is the lexically-specific element in a frame *X b*, with *X* representing the frame slot as in previous chapters. Note that *c* is not involved in any relations with either *a* or *b*, and so does not form part of the frame.

In Figure 8(b), *a* and *c* both predict *b*, but there is no relationship between *a* and *c*. Once we think of *b* as the lexically-specific element in a frame, however, it can be regarded as the "link" between *a* and *c*. Even though *a* and *c* are not related, they are both involved in a filler-frame relationship with *b*, and so we can take the frame to cover the entire string *a b c*: in fact, we have here a single frame with two frame slots, which can be represented as *X b Z* and *Z b X*, with *X* being the focal slot and *Z* an "out-of-focus" slot.

In Figure 8(c), by contrast, the element *b* predicts both *a* and *c*; the decision made in this case is not to merge these two frame primitives, but to take them as independent frames. This is because *a* and *c* are likely to be independent of each other, so that *b* can be expected to occur on occasion with *a* only, and on other occasions with *c* only. If *a* and *c* were not independent, then there would have been an equivalence relationship between them.

Figure 8(d) and Figure 8(e) demonstrate situations where there are equivalence relations in the fragment. Elements which are equivalent to each other can be treated as collocations, i.e. they can be regarded as functioning as a unit, regardless of the fact that they are written as more than one dictionary word. Collocational units can be either fillers or lexically-specific elements in a frame, depending on the direction of the arrow. In Figure 8(d), *a* predicts *b*, which is equivalent to *c*. Note that, in this case, it is not required that *a* also predicts *c* in order to construct a frame out of all three elements, namely *X b c*. Equivalence "lends its transitivity" to the predictive relationship from *a* to *b*. In Figure 8(e), we have a filler consisting of more than one word. This is something not yet encountered in the work presented in the previous chapters. The frame in this case is just *X c*, and the filler is the "phrase" *a b*.

Figure 8(f) demonstrates a similar situation to that of Figure 8(e); however, in this case the two equivalent words making up the filler are not even contiguous. This could occur for instance in an utterance like *pick it up*, where the parts of the "separated verb" *pick up* occur as a filler in a frame with two slots on either side of *it*. The two slots are not independent, though (as in Figure 8(b)); instead there is an interaction because the two words are equivalent (predict each other). Hence they constitute a kind of collocational filler. In this case, the frame is written as *X b X*, and the filler is *a c*.

Figure 8(g) represents a situation similar to that of Figure 8(a), in that there is one predictive relationship between two elements, and another element that is "isolated" from the other two. However, the isolated element occurs between the other two elements. The disjunction of *a* and *c* is regarded as a major feature of the frame, so that the space occupied by *b* needs to be represented. However, because *b* does not engage in any relations with the other elements, it is not a filler of the frame. In this case, the only frame filler is *a*, and the frame is written as *X* * *c*, with the * representing the "unused" position in the fragment.

Figure 8(h) represents the situation where there is a cycle in the diagram. In such a case, there is no easy way to include all the relevant relationships into one consistent frame, and so we list each sub-frame separately. Figure 8(h) represents a situation where *a* predicts *b*, *b* predicts *c*, and *c* predicts *a*. These three relationships give rise to the frames *X b*, *X c*, and *a* * *X* respectively, following the examples given in previous Figures.

Lastly, Figure 8(i) illustrates the point that predictive relationships are not handled transitively; when *a* predicts *b* and *b* predicts *c*, it does *not* necessarily follow that *a* predicts *c* (unless this is explicitly the case in the ΔP matrix). From the diagram, we can deduce that there is evidence only for the frames *X b*, with *a* as the filler, and *X c*, with *b* as the filler. The two predictions give rise to two separate frames, and do not merge into one larger frame with two levels of nesting (although this possibility could be explored in future versions of this approach).

9.5 Psychological considerations

As with the other frame discovery procedures, it is assumed that the specific words in a prediction-based frame are associated with each other by virtue of their co-occurrence in several utterances in the input to the child; in this way, the configuration of words may itself become a unit of linguistic knowledge.

These frames may start out as fairly verbatim sequences of words that co-occur, involving both the frame and its slot-filler; for instance, in the example phrase of "the dog", as used before, the presence of "dog" can be taken to be associated with the word "the" occurring before it. It may thus seem problematic to account for how a slot-filling word is eventually separated out from its frame, so as to allow the frame to become abstract. But as stated before, the associative links between words are not necessarily equally strong in either direction. Whereas the associative link from a predicting word to its predicted frame (as modelled by ΔP) is strong, the links from the frame words to the predicting word are weak. This may occur due to interference incurred by the many other fillers that have appeared in the slot, as in the PARSER model (Perruchet & Vinter, 1998). Unlike the situation in PARSER, however, it may be that interference operates on *asymmetric* associative relationships. If a frame has been encountered with filler A and later with filler B, it may be that only the link from the frame to filler A suffers a decrease in strength due to interference, while the link from A to the frame remains unchanged (because A has not been encountered, its outgoing links are not modified). In this way, the frame may become more abstractly represented, whereas a particular filler is still associated with each of its potential frames. It may be this asymmetry which eventually allows a frame to be segmented off from its fillers.

It is therefore predicted that a particular word may be associated with each of the frames in which it could occur, in a kind of "halo" of combinatorial possibility[5]. Individual frames, on the other hand, are not strongly associated with particular fillers.

As suggested earlier, this approach also accounts for the existence of collocations, which are sets of words that are in an equivalence relationship to each other by virtue of occurring often together. These collocations either form part of the lexically-specific part of a frame, or are multi-word fillers of certain frame slots.

This approach therefore allows for both symmetrical predictive relationships (collocations), and asymmetrical relationships where item A is associated with item B, but not vice versa (the so-called frame primitives). Particular configurations of both of these kinds of relationships constitute the frames produced under the prediction-based approach.

During processing, each frame plus all of its fillers are then recognized directly in the input, and the beginning and end of this set in the utterance are taken to constitute the segmentation boundaries of the frame, in contrast to the approaches in the earlier chapters, which started from the segmentation boundaries provide by the edges of the utterance.

[5] An extension to the current work would be to incorporate not only predictive relations between words, but also between word roots and affixed morphemes.

It should be emphasized that, as with similar remarks on the other frame discovery techniques, these remarks are speculative; it remains to be determined empirically whether children do exhibit this form of learning behaviour.

9.6 Frame discovery procedure

9.6.1 Obtaining frame probabilities

Conditional probabilities and ΔP statistics can be obtained by collecting counts of *n-grams*, sequences of *n* consecutive items occurring in a corpus. For *n* equal to 2, n-grams are termed *bigrams*, and for *n* equal to 3, they are called *trigrams*.

We are interested primarily in highly *local* predictive relationships, on the assumption that they make up the vast majority of predictive relationships in a language. A stronger version of this assumption is to assume that languages develop in such a way as to make them be easily learnable by young children, and to propose that local relationships should be the norm in most languages, given that they narrow down the space of possibilities that need to be considered. Local relationships also seem to embody the intuitive notion that concepts that go together are expressed through words that occur together. This too is as would be expected if the learning of relationships is mediated by learned associations between the words: if related words occur together, that makes it more likely that the learner will represent them simultaneously in short-term memory at some point, thereby facilitating the forging of an associative link between them.

For this reason, I will only consider relationships between two consecutive words, or between two words separated by a third word. I will also assume that in the second case the separating word is a necessary prerequisite for the recognition of the relationship (regardless of the identity of the separating word), so that a dependency from A to B when they appear consecutively is different from a relationship from A to B with a word intervening.

Whereas in the previous chapters, the corpus was first rewritten by replacing the relatively infrequent words with X's, the experiments in this chapter make use of the original ("cleaned-up") corpus directly, so that all words are treated equally. In collecting the relevant starting information for this experiment, the *bigram frequency matrix* was obtained, consisting of all bigrams (all two-word sequences) in the corpus, and also the *disjunct-by-1 frequency matrix*, consisting of the "disjunct-by-1 bigrams", all word pairs attested in the corpus where the second word occurs two positions after the first word (essentially, this is the set of trigrams, collapsed along the dimension of the middle word).

All bigrams had to occur inside a single utterance, i.e. bigrams were not allowed to straddle utterances. Bigrams containing commas were also disallowed, as were bigrams containing question marks (recall that full stops were removed from the corpus). The same constraints were applied to the disjunct bigrams.

After bigram statistics were collected, the next step was to calculate ΔP values from both frequency matrices, in both the "leftward" and "rightward" directions (i.e. ΔP and ΔP'), in accordance with the formulas given in Section 9.2. At this point, we have a set of relationships between words which will be the basis of the set of frame primitives. For every utterance in the input, we have the strength of the predictive (ΔP) relationship between every pair of adjacent words (both left-to-right and right-to-left) from the bigram frequency matrix, and the strength of the predictive relationship between every pair of words that occur with one word intervening (both left-to-right and right-to-left) from the disjunct-by-1 bigram frequency matrix.

Next, only the ΔP relationships that are *significant* are retained. Significance is determined by comparison against a fixed threshold: only ΔP values that are above that threshold are regarded as significant. While the choice of this threshold is somewhat arbitrary, we can see here a major benefit of using ΔP over the more traditional conditional probability: with conditional probability, we would have no way of choosing a single threshold for all bigrams. This is because there is no way to set a fixed "baseline" that would hold for all conditional probability values; each particular conditional probability has a different baseline. By contrast, the baseline for ΔP is zero across all pairs of items; any ΔP value greater than zero is indicative of an implicational relationship. In practice, the threshold for significance was set not at zero, but at 0.005, a value that proved to yield good results in preliminary testing.

At this point, we have a set of significant ΔP relationships; these constitute the set of frame primitives.

9.6.2 Parsing the corpus for frames

Next, the frame primitives are used in order to parse the corpus again, this time assembling frames from the frame primitives present in each utterance. These frames are then used for collecting frame-word co-occurrence data.

Having found all the significant predictive ΔP relationships from one word to another, the first parsing step is to identify all *equivalence groups* in the utterance, defined as a set of elements (words) from the utterance such that, for any element in the group, there is at least one other element in the group with which it has an equivalence relationship (i.e. two symmetric significant ΔP relationships, one from element *a* to *b* and one from element *b* to *a*); equivalence is therefore treated as a transitive relation. If an element enters into no equivalence relationships with any elements, it is placed in an equivalence group on its own.

Take as an example the utterance “I heard you were speaking”. In the implementation discussed in the next two sections, there is a high value for ΔP at an offset of 1 (i.e. ignoring the intervening word) from “I” to “you”, and also from “you” to “I”. Consequently, the words “I” and “you” in this utterance are placed in an equivalence class. All other words in the utterance are placed in equivalence classes of their own.

Next, we look for *predictive relationships between equivalence groups*, which are considered to exist if *any* element in one equivalence group is predictive of *any* element in another equivalence group. Under this definition (as we have seen in Figure (h)), it is perfectly possible to have *cycles* in the graphs, where equivalence group *E* predicts equivalence group *F*, and vice versa.

In the example, the word “heard” strongly predicts “I” (there is a high ΔP from “heard” to “I”), and the word “speaking” strongly predicts “you”. Because “I” and “you” are placed in an equivalence class, these predictive relationships are treated as one prediction from the equivalence class containing only “heard” to the equivalence class consisting of “I” and “you”, and another from the equivalence class containing “speaking”, also to the equivalence class of “I” and “you”. There is no predictive relationship between “heard” and “speaking” in either direction. Hence, the resulting relationship diagram is as in Figure 8(b).

Subsequently, for every equivalence group *E*, collect the set of all equivalence groups that predict *E*. At this point, we have a set of predicting groups and a single predicted equivalence group. These are all the words that will be involved in the current frame.

If we take *E* to be the equivalence class consisting of “I” and “you” in the example, then the set of predicting groups consists of the two equivalence classes that consist respectively of only the element “heard” and only the element “speaking”.

Now “flatten out” the groups (both the “predicting” groups and the “predicted” group *E*) into their constituent words, and select the word w_α that occurs earliest in the utterance from among all the words in all the groups, and the word w_ω that occurs last. The frame then stretches from w_α to w_ω. The words belonging to *E* are taken to be the lexically-specific words in the frame, and each predicting equivalence group as a whole corresponds to a single filler in the frame. In cases where there is more than one word in a predicting equivalence group, the filler is a multiword filler constructed by taking each of the words in the group in left-to-right sequence, separated by spaces. In cases where the members of an equivalence group are not contiguous, they are still listed from left to right in order to produce a multi-word filler.

This process performed on the example sentence produces a frame that stretches from “I” to “speaking”. The words “I” and “you” are retained as lexically-specific words, and “heard” and “speaking” become slot-fillers.

Any words that occur between w_α and w_ω but which are not members of either E or any of its predicting groups are still included in the frame, but they do not form fillers of any frame slot and their positions are indicated by *.

The intervening word "were" is not linked predicatively to any of the other four words, but because it is located between the first linked word "I" and the final linked word "speaking", it is included as a non-functional slot and indicated with a *.

The resulting two frames are therefore "I X you * Z" and "I Z you * X", where the X indicates the active slot, filled by "heard" and "speaking" respectively.

In this way, the algorithm parses the entire corpus for frame structures based on ΔP, and collects filler-frame data in the same way as in the previous chapters. The resulting data matrix is then subjected to clustering analysis and subsequent co-clustering phases, in the same way as in the previous chapters.

9.7 Implementation

The prediction-based algorithm as outlined above was carried out on the Manchester corpus. The resulting data matrix was subjected to co-clustering in the same way as in previous chapters. Table 63 shows the numbers of frames and slot fillers (single-word and multi-word) in the matrix, after applying the 5-5 criterion, as well as the number of words treated as slot-fillers in the model, and the number of utterances that contained at least one frame.

Number of frame types	2923
Number of slot-filler types	4186
Number of focal words covered	239603 (18.1%)
Number of utterances containing at least one frame	142706 (42.6%)

Table 63. Summary numbers regarding coverage of the Manchester corpus by the prediction-based frame approach.

the X; X it; a X; X you; it's X; that's X; Z the X; X the Z; you X; and X;
it's a X; your X; that X; are you X; oh X; X that; X now; X there; that's a X;
the X Z; Z a X; i X; X on; X then; it X; and * X; X a Z; the Z X; X in; he's X;
you're X; is it X; Z the Z X; you X Z; they're X; X up; Z your X; Z X it; don't X;
X Z the Z; is that X; that's the X; Z that X; where's the X; X * you; X aswell;
X here; not X; Z in the X; X again; X it Z; X your Z; well X; you Z X; Z it X;
X that Z; Z the X Z; X Z it; X to; in the X; X out; it's not X; X me; X * it;
X you Z; there's the X; mummy X; it's * X; X in the Z; X the Z Z; Z you X;
just X; oh * X; this X; is he X; on the X; two X; that's X Z; X one; X what;
he X; no X; what X; Z X on; do you like X; i'm X; and X Z; that's Z X;
there's X; a X Z; X the; to X; some X; Z X the Z; i X Z; there's a X; a Z X;
Z to X; like X; it X Z;

Table 64. The top 100 most frequently-occurring prediction-based frames in the Manchester corpus.

The top 100 most frequently-occurring prediction-based frames that were discovered in the Manchester corpus are shown in Table 64. In addition, a number of selected frames along with their fillers are shown in Table 65; a couple of points are illustrated by these examples. Firstly, a frame such as “a bit of X”, which is also included in the full-utterance frame and nested frame approach, takes a fairly reliable set of noun or gerund fillers (including multi-word fillers such as “fuzzy felt”). By contrast, the very similar frame “a bit of * X” takes mostly nouns, but also a number of adjectives. For members of some frame “couples” such as “I’ve Z my X” and “I’ve X my Z”, the slot fillers fall into very clear categories, namely verbs and noun phrases respectively; but for many other frame couples such as “are they X Z” and “are they Z X”, the fillers are not of a coherent class. An example of a frame couple that involves “separable” verbs is “did you X Z X” and “did you Z X Z”. While “did you X Z X” accepted a fairly consistent set of fillers such as “build bridge” and “put on”, the fillers for the “middle” slot (“did you Z X Z”) were a mixed set including nouns, pronouns and the determiner “a”. Other pairs such as “shall I X it Z” and “shall I Z it X” demonstrate how the two parts of a separable verb may sometimes be identified independently, rather than as the two parts of a single collocational filler. And lastly, some frames such as “X Z today” or “more * X” are simply not informative about the word class of the filler at all, as evidenced by the disparate set of fillers for that frame.

FRAME	FILLERS
Z * Z with the X	cows, juice, prodding stick, roller, sugar, whale
Z * you X Z * Z	bought, could, never, picking, should, won't
X Z today	almost, bedroom, being, jelly, quite, really horrible, Terence, the water
X at the bottom of the Z	balls, cows, probably, sleep, worm
Z X Z for	another, shoe, two, very, wet
a bit of X	chicken, clown, fluff, fuzzy felt, newspaper, running, settee, spaghetti
a bit of * X	bumpy, crazy, dead, disaster, hat, hole, maybe, misery, puzzle, tantrum
are they X Z	both, dry, eating, falling, getting, hungry, kissing, molly's
are they Z X	biscuits, bones, ducks, each other, everywhere, sukie's, their
did you X Z X	build bridge, eat banana, put in, put on, read story
did you Z X Z	a, butter, her, him, salt, stamps, that, them, your
I've X my Z	brought, changed, drunk, eaten, finished, lost
I've Z my X	cup of tea, dinner, earring, memory, name, spade, teddybear, yoghurt
more * X	insist, less, ow, pat, rectangle, scratching, shorts, twenty pounds, windows
shall I X it Z	cut, finish, fix, hatch, open, read, roll, send, spin, start, straighten, throw, tie, tip, undo, wipe
shall I Z it X	away, back, for, now, off, out, over, round, then, through, under, while

Table 65. Some example prediction-based frames and their slot-fillers .

9.7.1 Qualitative results

Because all words are now allowed to function as slot-fillers, there are a great many word types in the final data set for prediction-based frames that were not in the data set for full-utterance frames or nested frames. So for instance, determiners, pronouns and modal verbs are very frequent as slot-fillers. This means that it might not be possible to obtain just three clusters using hierarchical clustering that correspond to nouns, verbs and adjectives. And in fact, the first three sizeable clusters that form (when 5 clusters are created) are not interpretable in this way. Two of these clusters do correspond to verbs and nouns, but the third cluster is an amalgamation of adjectives, possessive pronouns, proper names and one or two determiners. It is not until ten clusters are formed that the class of adjectives separates out from the closed-class words and proper names in that cluster. Some examples of frames and the most prevalent words associated with particular clusters are shown in Table 66. These details are for 10 clusters (with just the 5 sizeable clusters shown).

Cluster	Frames	Words
Cluster 1	Z * a bit X; Z * are X; Z * got X; Z X in a minute; Z a X one; X all the time; X bag; X ball; X chair; X hair; X little boy; X people; X than; a X boy; a X one; are they X; because it was Z X; do you think she's X; gonna be X; he hasn't X; he is X; if it's X; it's not very X; it's too X; nice and X; no X today; quite X; really X; should X; that one's not X; what were you X; where's he X; which one's X; you are X	poorly, broken, dirty, cold, crying, black, jumping, naughty, new, wet, lovely, hot, hungry, special, busy, drawing, horrible, asleep, hiding, funny, noisy, taking, sleeping, dark, different, driving, happy, tired, fluffy, sick, cross, eating, saying, stuck, thinking, wearing, building, holding, quick, real, tiny, white, cheeky, coloring, drinking, sad, still, thirsty, watching, bad, cutting, high, reading, sitting, sticky, frightened, full, playing, running, singing, squashed

Cluster 2	Z * it's X; Z * like X; Z * with X; Z X back; Z X in the water; X * clean; X Z for; X at the weekend; X doesn't want; X foot; X going to sleep; X on the car; X said * Z; and that's X; and who's X; are you gonna put X; because they're X; called X; can i have X; can you remember X; come on X; do it X; do you have X; do you think X Z; does X like Z; give it to X; has X got a Z; have you seen X; here's X; i don't think X; i know it's X; is that what X Z; it is X; look what X; pick X up; play with X; push X; remember X; saw X; talk to X; that's not X; the one with the X; there's a little X; what did X do; what does X say; you can't Z X	mummy's, childname's, daddy, something, yours, daddy's, mummy, Thomas, yellow, they're, childname, those, green, he's, blue, my, Caroline, red, still, Pingu, Henry, James, orange, Anna’s, dolly, an, who's, Gordon, her, pink, purple, two, another, anything, everything, right, mine, Caroline’s, his, like, milk, Nana, panda, Percy, Anna, four, Andy, baba, grandpa, better, grandma, probably, dolly's, him, their, enough, teddy, yourself
Cluster 3	Z * Z X the Z; Z * you X Z * Z; Z X for; Z he X; Z if i X; Z you X them; X * bag; X * ball; X * last night; X a Z one; X a cow; X at the bottom of the Z; X away; X for him; X her; X him out; X if you're Z; X it; X it to me; X lots of Z; X me; X me * Z; X my Z; X off then; X on that; X on your Z; X on your own; X one of those Z; X out of the way; X sleep; X some; X that; X that one; X the Z; X the green one; X the monkey; X them off; X these things; X up then; X with this one; X your bottom; X yourself; all the X are; and then you X; and what did you X; are you X me; are you going to X it * Z; because you Z X; better X; broken X; can X; careful you don't X; did X; do you want to X; don't X him; give him a X; going to Z; have they X; i didn't X that; must X; run out of X; shall we X this; try and X; what did they X; you can X Z	watch, keep, push, made, open, move, hold, bring, read, use, break, draw, stick, still, cut, pull, sing, always, bought, call, won't, gave, leave, said, wouldn't, brought, stop, jump, catch, couldn't, try, eat, found, walk, wear, ask, missed, will, carry, dropped, roll, throw, blow, hide, hurt, could, willn't, never, listen, press, really, wash, and, asked, giving, bite, hit, just, kick, lost, stay, built, done, has, lift, forget, moved, should, took, getting, given, left, sat, say, told, buy, make, probably, actually, build, paint, pretend, turn, up

Cluster 4	Z * in the Z X; Z Z get X; Z X for him; Z for the Z X; Z like that X; X , did you; X at the bottom; X if you like; X Thomas; and that's Z X; are we going X; aren't they X; can you see the Z X; do you think X; don't Z it X; get it X; go * X; i Z it X; i don't know X; i don't think there are any more X; I’ve got your X; is it going X; it is Z X; mummy do it X; perhaps it's X; she's going X; that's it X; the car X; wasn't it X; what do you like X; you can do it X; you've got a Z X	yeah, though, aswell, now, first, before, but, down, so, today, darling, up, maybe, outside, yes, through, love, yet, again, anyway, or, over, please, because, yesterday, actually, already, later, perhaps, sometimes, somewhere, from, swimming, like that, er, properly, says, shopping, then, things, fast, fits, home, too, anywhere, off, round, soon, will, bit, by, together, upstairs, car, goes, into, made, okay, pet, really, tomorrow
Cluster 5	Z * Z a X; Z * Z with a X; Z Z a nice X; Z X for you; Z any X; Z up to the X; X all over the floor; X back on; X can; X doing Z; X don't Z; X for a walk; X goes in the Z; a Z in the X; a baby X; a nice X; all those X; are those X; are you Z a X; big X; bit of a X; blue X; can i have Z X; can you find another X; can you see a X; coming out of the X; do they eat X; do you need a X; do you want some X; draw * X; eat * X; get a X; got X on; he's going to the X; here's your X; i don't like X; into the X; it is a X; keep * X; like a X; look at all those X; look like a X; make some X; move * X; naughty X; no more X; on my X; poor X; put your X down; shall we get the X out; that X; that is a X; the X; there's another X; this X; this is a Z X; through the X; what X do you want; where's your X; you did Z X; you don't like X; your X	cat, tractor, dog, tiger, house, pig, sheep, truck, dolly, lion, fish, duck, ball, cake, digger, lady, fireengine, picture, cow, giraffe, rabbit, bus, hat, penguin, dress, egg, wheels, bag, hole, panda, driver, boat, chicken, lorry, thing, snake, teddy, bottle, eggs, story, book, horse, people, spoon, key, leg, tree, whale, zoo, hippo, trailer, cars, horsie, orange, bed, garage, noise, fire, spider, trains, bird, bull, farm, tomato, banana, letters, money, tower, balloon, doll, icecream, milk, tunnel, babys, bike, biscuit, strawberry, wheel, daddy, goat, lemon, man, chair, crayons, van

Table 66. Some representative frames and words from the hierarchical clustering of the prediction-based frames, showing representatives for 5 out of 10 clusters

9.7.2 Quantitative results

The frame-word matrix was subjected to the set of hard clustering and co-clustering algorithms in the same way as for the data matrices of previous chapters. The quantitative evaluation results are shown in Table 67.

	Hard F	**Fuzzy F**	**Fuzzy W**	**Fuzzy FxW**	**Confl.**	**Pars.**
Accuracy	0.679 (*0.471*)	0.690 (*0. 471*)	0.820 (*0. 471*)	0.813 (*0. 471*)	0.508 (*0. 471*)	0.808 (*0. 471*)
Completeness	0.403 (*0.279*)	0.528 (*0.360*)	0.565 (*0.325*)	0.636 (*0.369*)	0.888 (*0.824*)	0.620 (*0.361*)
F score	0.506 (*0.351*)	0.598 (*0.409*)	0.669 (*0.385*)	**0.714** (*0.414*)	0.647 (*0.600*)	0.702 (*0.409*)
Bookmaker	0.349	0.441	0.557	**0.596**	0.184	0.553

Table 67. Evaluation of prediction-based frames, 10 clusters.

It can be seen that the quantitative performance of the prediction-based frame approach yielded a reasonably good lexical categorization, with Bookmaker scores of up to 0.596 (for the Fuzzy FxW co-clustering algorithm). Nevertheless, it is also clear from a comparison of this table with the results from previous chapters that this particular implementation of the prediction-based approach does not yield as accurate a categorization as do the full-utterance frame and (locally context-sensitive) nested frame approaches that start off from a dichotomy between slot-filler words and frame-building words. It should also be remembered, however, that the coverage in this case is far greater than that of the full-utterance frames and nested frames (18% of words and 42% of utterances versus 9% and 32% in the context-free nested frame case). Also note that the categorization here is much more successful than was the case for nested frames using context-free parsing.

Another interesting result is that, as was the case with context-free parsing of nested frames, the conflict-based algorithm failed to categorize instances adequately for prediction-based frames (the values of all measures are close to their random baselines). Further examination of the performance of this algorithm (not shown here) revealed that this was due to its assigning almost every word and frame to the noun category (Cluster 5). Out of the three co-clustering algorithms, the conflict-based algorithm is by far the most sensitive to an initial data set of poor quality. When the data set contains many frames that do not reliably indicate the part-of-speech of a focal word embedded in them, then it is likely that a majority of words occurring in each frame will belong to the majority category (which is the noun category in this case) and hence that a majority of conflicts can be solved by allocating all or most of the remaining items in the data set to that majority category.

Randomization tests of significance (Table 68 and Table 69) revealed that the parsimony-based and Fuzzy F × W co-clustering algorithms significantly out-performed the hard clustering approach for prediction-based frames, as was found for all other frame discovery approaches (hard clustering was significantly better than conflict-based clustering). These results also show that the Fuzzy F × W co-clustering algorithm achieves a significantly better categorization than the parsimony-based algorithm.

	Fuzzy FxW	**Confl.**	**Pars.**
Hard F	↑ < 0.001**	← < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	← < 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 68. Significance levels of differences in F scores for various clustering algorithms, for prediction-based frames, N=290.
*** significant at p=0.05, ** significant at p = 0.01.**

	Fuzzy FxW	Confl.	Pars.
Hard F	↑ < 0.001**	← < 0.001**	↑ < 0.001**
Fuzzy FxW		← < 0.001**	← < 0.001**
Confl.			↑ < 0.001**
Fuzzy F	↑ < 0.001**		
Fuzzy W	↑ < 0.001**		

Table 69. Significance levels of differences in Bookmaker scores for various clustering algorithms, for prediction-based frames, N=290. * significant at p=0.05, ** significant at p = 0.01.

9.8 A comparison with Cameron-Faulkner et al. (2003)

As was done for the combined nested-and-full-utterance-frame approach in Section 8.7, it is possible to compare the prediction-based frames against those manually identified in the Manchester corpus by Cameron-Faulkner et al. (CFLT, 2003).

Part of the point of the work by CFLT was to demonstrate that many of the utterances that children hear start with a small number of word sequences, which may hence be thought to be predictable features of the input. The prediction-based approach could be regarded as compatible with this idea if it could be shown that the utterance preambles identified by CFLT are also present in the lexically-specific portion of the prediction-based frames identified for the Manchester corpus. This would mean, amongst other things, that CFLT's preambles are equivalence groups according to the prediction-based approach. (Note that, of course, prediction-based frames are not constrained to represent the *beginnings* of utterances, as CFLT's frames were. Nevertheless, when confronted with a sentence starting with a CFLT frame preamble, the prediction-based approach would still be likely to treat the preamble as a coherent equivalence group, because of the high mutual ΔP values between the words.)

Table 70 indicates the number of CFLT frames that were also covered by prediction-based frames. In brief, a majority (139 out of 152) of the CFLT frames were found in the set of prediction-based frames. Only 13 frames were not covered. As was the case for the nested-plus-full-utterance-frame approach, most of the errors came from the set of Wh-questions. In addition, every one of the 52 core frames was covered by a prediction-based frame. Full details about the correspondence between prediction-based and CFLT frames can be found in Appendix 2.

All frames covered	Core frames covered
139 / 152	**52 / 52**

Table 70. Summary of coverage of Cameron-Faulkner et al. (2003)'s frames by the set of prediction-based frames.

9.9 Problems with prediction-based frames

While the above categorization was fairly successful, and the frames produced by the prediction-based discovery procedure were for the most part convincing ones (including the basic frames that one would expect to see, such as "a X", "the X", "this X", "that X", "it's X", "your X", "a X one", "X it", "X me", "going to X it", "can you X it", etc.), a major problem with this approach is the question of generalizing the discovered frames to the corpus as a whole. Recall that, because these frames are presumed to be constellations of words that are simultaneously activated in working memory, the presence of the predicting slot-filler word is crucial to activating the entire frame. If an unknown word were to be used as a filler in the same frame structure, there are no predictive relationships in place from that word to the lexically-specific elements that would facilitate their activation. It was presumed earlier that the structures themselves would become familiar to the language-learning child as units in their own right. However, consider the following situation: there are, in the current data set, two frames "the X" and "the X X", where the slot of the first frame corresponds to a slot for nouns, and the two slots of the second frame are typically filled by an adjective and a noun respectively. When an utterance is encountered that contains the word "the" followed by two or more novel words, which frame should be applied? In the case where only the first following word predicts "the", the answer would be simple, and the frame would be "the X".

Similarly, when both slot-filling words predict “the”, the frame would be “the X X”. However, with novel words, it is difficult to decide whether the correct frame to apply is “the X” or “the X X”, and therefore it is undecidable whether the word following “the” should be categorized with the slot fillers of “the X” (likely to be nouns), or the slot-fillers of the first slot of “the X X” (likely to be adjectives).

A possible way out of this dilemma would be to extend the current work by taking an approach similar to that taken with the full-utterance frames and nested frames, namely to recognize frames only in the context of an appropriate surrounding environment that indicates the boundaries of the frame (full-utterance frames were recognized only in the context of surrounding silence, and nested frames only when embedded in the context of an already-recognized frame). This would require expressing the predictability structure of an utterance as a whole, which would mean amongst other things listing allowed patterns of consecutive, unlinked frames, taking cognizance of words that are never linked into any frames, and most importantly, considering frames hierarchically nested inside others, as was alluded to in Section 9.4.1 when discussing the kind of relationship depicted in Figure 8(i).

9.10 Discussion

The results of this chapter show that it is possible to discover lexically-specific frames in the input to children based purely on reliably predictive relationships between words, conceived of as learned associations, the strength of which is given by the ΔP function, and which indicate a dependency relationship between two words. The results also show that parts-of-speech can be induced from these discovered frames, although the resulting categorization is less successful than that from the techniques from earlier chapters where only a circumscribed set of words were allowed to function as frame-building words.

This chapter concludes the reporting of empirical results from implemented experiments. The next two chapters take a broad view of the work presented here, firstly situating it against other computational models of lexical categorization, and then outlining ways in which the current framework could be extended.

10 Summary and comparison to other computational models

In this chapter, I will first attempt to summarize the main contributions of this thesis. Next I will compare the current work against the computational models reviewed in Chapter 4. This comparison will focus on two main aspects of the current models, namely (i) the way in which contexts for the induction of paradigmatic categories are discovered, and (ii) the way in which context and word information are used in order to arrive at a categorization of a focal word in context.

10.1 Summary of empirical results

One of Pinker's (1984, 1987) major objections to distributional bootstrapping of parts-of-speech is that the number of possibly significant distributional properties to which the child should attend is essentially unbounded. Pinker (1987) suggests properties such as occurring in seventh serial position in a sentence, co-occurring in a sentence with the word "mouse", etc.

The response to this should be (see also Redington et al, 1998) that the child need not be presumed to be attentive to every conceivable distributional property, but only to a small number that are psychologically salient. The traditional approach in implemented distributional treatments of part-of-speech bootstrapping has been to demonstrate that attending to one such property, namely occurrence of a word within a very small window of words surrounding a focal word, can provide enough information for a highly successful part-of-speech categorization (e.g. Clark, 2000, 2001; Finch, 1993; Finch et al., 1995; Mintz, 2003, 2006a, 2006b; Redington et al., 1998, Schütze, 1995).

In the current work, I have taken a different starting point, namely that there are a number of semi-fixed sentence and sentence fragment patterns in the input to the child, that are easily detectible by mechanical means, and which are arguably constituents or constructions of the language, either because they represent full utterances, or because they correspond to smaller units such as phrases, which can be embedded inside larger frames. These constituents are composed of specific words, which facilitates learning the constituents, together with slots in which a variety of items can occur. When these items are single words, their occurrence in such a frame is often a reliable indicator of the part-of-speech of the word.

In order to justify this account, it is necessary to demonstrate a plausible mechanism by which these frames can be discovered. One approach that was taken here was to regard some of the words in the input to the child as special in some way, and base the frames around them. If we presume that the ability to recognize a word (e.g. by neuronal assemblies in the brain; Pulvermüller, 1996) is reinforced every time the word is recognized, then it follows that the most frequently-occurring words in the input will be the ones that are most readily and effortlessly recognized. These words are therefore also the most likely substrate from which the first co-occurrence relationships between words are learned.

In the speech stream, the most frequent tokens are typically function words and a number of semantically light open-class words. These words are marked by diminished semantic content, and, in the case of English, often by reduced phonological content. These factors may allow these words to retreat into the background so as to allow more attentional and processing resources for the more contentful, relatively rarer words (typically open-class words).

When several of these words co-occur in an utterance, their co-occurrence pattern is presumed to be learnt by virtue of the fact that these words are simultaneously attended to and hence associated with each other and stored in memory as a unit (in line with proposals by e.g. Logan & Etherton, 1994; Pacton & Perruchet, 2008; Treisman & Gelade, 1980). I have proposed that it is not only the frequent words, but also the positions of the slots, that are simultaneously represented and associated together as a unit.

The first mechanism considered, the full-utterance frame technique, took the approach that constructional frames can be identified as complete utterances that follow a familiar pattern, with this pattern defined by the configuration of frequently-occurring English words that are found in the utterance. In this view, these constructions were identifiable as units because they were delimited in the speech stream by fairly long pauses before and after them (as well as by single coherent prosodic contours, and other linguistic and non-linguistic cues).

The second technique explored the idea that some informative constructions may occur nested inside others, and that these constructions may quite easily be segmented out of the speech stream, because the surrounding context is already familiar from having being encountered before with simpler material filling the slot. Two variants of this nested frame idea were investigated: one where the surrounding context was required to be present in order for the nested construction to be reliably identified (locally context-sensitive parsing), and one in which the surrounding construction was used only during the learning phase, but where the nested construction was thought to be a reliable enough unit that it could be reliably recognized in whichever context it occurred (context-free parsing).

The full-utterance frames are thus the pathbreakers for discovering certain smaller, phrase-like structures which occur in them in the same positions in which single words have previously been encountered.

In contrast to the full-utterance and nested frame approach, another possibility is that there are no words that are special and marked from the start as frame-building words. Instead, frames are built out of associations between any words that happen to co-occur reliably in the input. This is the approach taken in creating the prediction-based frames. These frames are based around a fundamental asymmetry in the predictive relationships between a pair of words, such that one word predicts another, but not vice versa. The word configurations predictable from a particular word were described earlier as constituting a "halo" of possible contexts in which the predicting word could occur. These haloes are taken to be frames which become stored in memory, again by way of associations between mutually-predicting words as well as between the predicted words and the unspecified slot-fillers.

Prediction-based frames also differ from full-utterance and nested frames in departing from the idea that units can only be learnt or recognized through being noted to occur in certain privileged surrounding contexts. In the prediction-based approach, words occurring in sequence cohere into larger units due to the learned probabilistic associations between them; the words themselves are therefore responsible for the constructional nature of the unit, not whether or not these word sequences act as units against a particular contextual background.

These three frame discovery processes were applied to the Manchester corpus. The constructional status of the full-utterance frames seems strongly supported: these structures were all schematic representations of valid English utterances. In the case of the nested frames, however, it was noted that while several of the phrase structures that one would have hoped to see emerge from this process were in fact produced, many other nested frames were not valid constituents. Much the same was true of the prediction-based frames. The frames from each of the three processes were also compared against those produced by Cameron-Faulkner et al. (2003) in a manual analysis of the Manchester corpus. Most of the manually-identified frames were also discovered by the automatic processes.

The frames so produced were shown to provide a solid basis for part-of-speech induction when the frames were grouped together into clusters on the basis of the sets of words which occurred in them, thereby dealing with the inherent ambiguity of words. However, it was noted that categorization could be improved by taking into account that frames could also be ambiguous. The category of a particular word in a frame can be determined by combining category information from the word with category information from the frame.

Three co-clustering techniques were proposed to deal with this issue. Fuzzy co-clustering explored the idea that words and frames belong to several different categories with specific probabilities, and that assigning a part-of-speech to a word in context is a case of choosing the part-of-speech that has the greatest joint (product) probability between the word and the frame.

The discrete co-clustering algorithms, on the other hand, produce a list of all the categories to which a particular word or frame can belong. The conflict-driven algorithm is based on the idea that the child is trying to make coherent meaning out of her experience, and that she knows that a word and the context in which it occurs should be consistent with each other. Hence, an allocated word occurring in a frame slot that has already been allocated to a category is likely to belong to that same category. Evidence of category membership is assumed to accumulate, and the more evidence exists for a particular frame or word to be allocated to a particular part-of-speech, the more likely it is that that allocation will take place in the child's mind.

The essence of the parsimony-based algorithm is the notion of cognitive conservativeness: a word (or frame) should be taken to be ambiguous only if this conclusion is forced on us by the data concerning the frames (words) with which it co-occurs, i.e. we seek the smallest set of potential categories to which the word could belong that could account for its pattern of co-occurrences. A language-learning child may resist assigning a word or construction slot to a particular part-of-speech if the evidence does not warrant it, and may even reexamine previous conclusions to decide whether they are still correct in the face of current knowledge.

The various construction frame discovery methods and clustering algorithms were evaluated in terms of their ability to assign words correctly to the parts-of-speech of nouns, verbs and adjectives. The first observation that should be made is that constructional frames are indeed a good source of information about the part-of-speech of a word embedded in them. In addition, the categories noun, verb and adjective appear to be quite robust features of English, that self-organize out of the word-frame co-occurrence input.

Another observation is that, in the current simulation at least, contextual information is most reliable when the construction context has itself been identified as occurring in a specific context: the context-free parsing of nested frames and the prediction-based frames performed worse at part-of-speech allocation than the full-utterance frame approach and context-sensitive parsing of nested frames.

Thirdly, combining word and frame information is always more accurate than using frame information only. The fuzzy co-clustering approach was most successful in most cases, followed by the parsimony-based algorithm (although these two performed comparably for nested frames). The conflict-based algorithm performed well with construction frames that had been identified in surrounding contexts of their own, but when the initial frame information was less reliable (obtained in a context-free manner), the conflict-based algorithm was not robust enough to allocate parts-of-speech correctly.

This work is therefore somewhat similar to the Frequent Frames approach of Mintz (2003, 2006a, 2006b), and can be seen as an attempt to extend the frames considered in that approach to more general, construction-like frame contexts, thereby potentially creating a bridge to work by Lieven and colleagues (Lieven et al., 1992, 1997, 1998; Pine & Lieven, 1993) on the importance of basic constructions for early language development.

10.2 Differences in the method of context discovery

10.2.1 The nested frame approach compared to ABL

The nested frames algorithm described in Chapter 8 can be viewed as broadly similar to van Zaanen's (2001) ABL framework. However, that approach considers every possible pair of sentences in a corpus in an attempt to discover possible "alignments", i.e. ways to match up shared structure between the two sentences. The approach outlined here is able to extract a reasonable amount of knowledge about the main constituents of English without considering nearly the same number of alignments. This is mainly due to its adherence to a basic dichotomy between frequent and less-frequent words, which happens to coincide to a large extent with the function-word/content-word distinction. Since function words are central to abstract structure in English, one would expect an approach that treated the function words as scaffolding for the embedding of content words to be successful in locating common English frames. The nested frames approach differs from ABL in the following ways:

- The corpus is not used “raw”, but instead is rewritten in terms of frequent words and variable slots, thereby compressing away a very large proportion of the information content in the corpus, so that there are far fewer potential alignments to consider (different utterances are “collapsed” together as a result of this move).
- Not all utterance structures are considered for alignment, but only the ones that have recurred in the input at least a certain number of times.
- On top of that, the final nested and nesting structures are required to meet the criterion of flexible combination, in that nested structures have to appear in a certain number of different nesting structure contexts, and vice versa, further restricting the search space to only promising possibilities.
- Lastly, not all alignments between sub-utterance fragments are considered; only fragments that occur in contexts in which a single word can occur are considered as potential nested frames. Whether this feature of the nested frames algorithm is advantageous, or whether many valuable constituents are actually missed in this way is something which would need to be investigated. However, this does of course further reduce the search space for the nested frames algorithm.

In ABL, it is presumed that every utterance that the system (the child) has encountered has been memorized and can be recalled to some extent, and ABL attempts to align every possible pair of utterances. By contrast, by reducing each sentence to its “bare-bones” structure, the current approach actually compresses the data set, by collapsing several utterances together into one if they are the same utterance in terms of their frequent words. This serves to reduce the space of possibilities for finding full-utterance frames and for aligning sentences in order to find nested constituents, and makes this process a more tractable one.

However, this step is not merely a heuristic to be applied prior to performing a process that is very similar to ABL. The process is fundamentally different, because I have tried to adhere to the principle that the algorithm should take only very simple steps, each one justified in terms of a mental architecture that is limited in such resources as memory, attention and processing speed. Recall of previously encountered utterances, for the sake of comparison against a current utterance, is done only for stored utterances that have recurred frequently enough to be regarded as potential units of the language.

As noted before, the combined set of discovered nested and full-utterance frames includes such "prototypical" phrase structures as "don't X it", "too X", "very X", "the X", "your X", "this X", "another X", etc. Van Zaanen has noted (2001, p.76) that ABL does not find structures directly corresponding to these phrase structures. Instead, it discovers structures such as, for instance, "noun phrase without a determiner" (which is of course a coherent category in Categorial Grammar). The fact that the nested frame approach is able to identify some of the major phrase structures in English may therefore be viewed as an advantage over ABL. It should be emphasized, however, that the identification of these structures was entirely due to the heuristic that phrases are often substitutable by single words, which is true for English but may not necessarily hold for other languages.

It would be possible to argue that the smaller search space of the nested frames approach is an advantage over ABL in terms of psychological plausibility, as ABL is heavily memory-based and, as a literal theory of language development, would require that a child aligns every pair of sentences it has ever heard – something which would perhaps be an unreasonable burden on a child's supposed limited cognitive resources.

On the other hand, one could counter this line of argument by pointing out that ABL is a broad framework for essentially discovering structure via variants of the substitution test, and that the reference implementation (Van Zaanen, 2001) is merely one instantiation of this framework.

Presumably, the psychological model behind ABL is similar to that underlying modern exemplar theories of conceptualization (e.g. Goldinger, 1996. 1998; Hintzman, 1986; Kruschke, 1992): all experiences are thought to be stored as memory traces, and can be reactivated in some way to influence current mental processes. ABL takes this approach to a logical extreme: when a new utterance is encountered, it is matched against every utterance ever encountered. This is clearly not a realistic proposal if it is believed that this comparison should be carried out *consciously*, with each pair of utterances in turn becoming the focus of conscious attention. Instead, a feasible solution may be a model in which all sentences that a child encounters are stored in memory as exemplars, and any new utterance *subconsciously* activates memory traces of all utterances that share some phonological overlap with it. In, for instance, a connectionist-inspired model, the process governing alignment could be simply a matter of incrementally strengthening the memory circuits that are responsible for maintaining the memory trace of the portions of the two sentences that are shared, and creating weak new circuits for the maintenance, as separate new memories, of the portions that are not shared. These memory traces would then become stronger with repeated use, or decay with disuse.

By contrast, the current approach seems, at first glance, to be more compatible with a prototype model (e.g. Rosch, 1983), in that an explicit schematic frame is formed fairly early on, out of the pattern of frequent words that an utterance contains, and newly-experienced utterances are matched against these abstract frames only. However, the process of frame formation is unlike that of prototype formation, in that the abstract frame does not arise out of a comparison between two experiences that entrenches their similarities and effaces their differences; instead, certain words (the most frequent words) are regarded as familiar signposts in the input stream, and patterns of co-occurrence between them are observed, and entrenched in memory every time they occur. The frame structure of an utterance is therefore perceived “directly” at the moment it is encountered (requiring only the identification of the frequent words and the registering of their positional configuration, including the positions of intervening non-frequent words), and “resonates” with a stored frame in memory only if the two are identical. Hence, there is no heavy load placed on memory processes: comparison of a current utterance against the memory of a previous utterance does not enter into this process at all.

It may be possible, then, to describe the work on full-utterance and nested frames as merely ABL with a number of heuristics. I would disagree with this characterization, however. In this work I am instead pursuing a model where there is at first a process of unconscious "sifting" and a familiarization with some of the more common structures in the input (syllables, words, even whole phrases or utterances), after which these structures are assumed to be potentially available to conscious awareness. At this stage processes such as alignment can take place on these already familiar items; these are taken to be *conscious* processes (as opposed to the account based on putative implicit processes that I have just offered to support ABL), because, as further discussed in Chapter 11, the child is continuously trying to make *meaningful sense* of the language that it is exposed to. Hence, the child is biased towards forming linguistic categories that also have semantic correlates. This conscious process is then made tractable only by limitations to the search space of the kind considered in this thesis.

10.2.2 The full-utterance frame approach compared to EMILE

The EMILE model was designed to discover categories that could be compatible with a Categorial Grammar approach. It takes full-utterance frames as its starting-point, just as the work in Chapter 6 did. However, frames are generated in EMILE by taking each sentence from a corpus and turning it into a set of one-slot frames, each one generated by replacing each word of the sentence in turn with a slot for variable material. For this reason, the number of contexts considered by EMILE is very large compared to the current full-utterance frame approach. EMILE's contexts are roughly a superset of the full-utterance frames, except that a full-utterance frame would be able to contain more than one slot, something not possible in EMILE at the moment.

EMILE therefore makes even larger demands on memory (both in a computer simulation and on the part of a putative human learner) than ABL. Every possible context for every possible word is considered, and there is no use of a heuristic to constrain the context discovery process to only the most frequent or most flexibly-used contexts and words, as done in the current work. As was the case with ABL, EMILE is possibly compatible with exemplar-based theories of memory. However, it may be more psychologically plausible for only the most useful contexts and words to be retained, rather than every combination from every utterance.

10.2.3 The current approach compared to Frequent Frames

The current frame approach is very close in spirit to the work by Mintz (2003, 2006a, 2006b) on Frequent Frames. One similarity between the two approaches is that contexts are explicitly listed, and treated as part of a child's knowledge of language. Also, both approaches aim at finding lexically-specific frames, i.e. the contexts are based on specific words rather than abstract word categories. Another similarity is the fact that the frame imposes a category on the word that is embedded in its slot, in contrast to work that allocates only a single category to a particular word type (e.g. Finch, 1993; Redington et al., 1998; Mintz et al., 2002).

However, the approach taken here to *defining* the context of a word is very different. Mintz (2003, 2006a, 2006b) defines the context in what might be called a "topological" manner, i.e. in terms of the relative positions of words, using a particular "shape" of context that is the same for all focal words. Frequent frames are defined as a disjunct frame of the form "a X b". Mintz therefore takes the approach that it is to a large extent the *local* context in which a word appears that determines its part-of-speech, and is more concerned with local cues to part-of-speech in the language input than with constructions as such. Mintz justifies the Frequent Frames technique as being supported by, for instance, the artificial language learning results of Gómez and Maye (2005), where *a X b* structures were used exclusively.

In the current frame approach, on the other hand, contexts are defined in a more "functional" manner – they are all utterances or partial utterances that are plausibly used as autonomous units, presumably for the purpose of serving a particular communicative function that pertains to the unit as a whole, and indeed they are intended to be constructions of the language. For this reason, frames of all topological shapes are acceptable. Furthermore, the final set of frames is selected on a criterion of flexible usage, i.e. that the frames should accept a variety of different word types into their slots, rather than on their frequency of occurrence only.

It might seem worthwhile to try to compare the empirical categorization success of the two models (e.g. in terms of accuracy, completeness, and F). Mintz (2003) reports accuracy and completeness scores of 0.90 and 0.91 for Frequent Frames, suggesting that the two models have comparable degrees of success in categorization (at least for the full-utterance frame approach of Chapters 6 and 7). It is, however, problematic to compare the two models directly. This is because there are several other factors (mostly parameters of the models) that are presumably tangential to the question of which frame approach is "better" for categorization, but which may contribute to a higher or lower success rate for each model. For instance, there is a parameter in the Frequent Frames model that determines how frequently a frame should appear in the corpus, and hence how many frames are used. In Mintz's reported results, typically around 50 frames are used. Comparing this with the 1465 frames of the full-utterance frame approach or 2923 of the prediction-based approach makes it seem likely that the difference in context set sizes might influence the respective outcomes of the two models. Similarly, the current approach uses the "5-5 rule" to pick out only the most flexible frames and words; however, the value of this parameter could have been set to a higher value than 5, and the number of frames would have dropped sharply.

Frequent Frames are undoubtedly highly informative contexts for predicting the part-of-speech of the slot-filler word, as shown by their empirical results (Mintz 2003, 2006a, 2006b). I would like to make two arguments in favour of the current approach, however:

(i) The first point is that the current approach is more closely integrated with syntactic learning, in that the lexically-specific frames discovered by the techniques presented in this thesis are intended to be *constructions*. The procedure has been designed to find frames that are likely to be whole constructional units. The child may be presumed to be learning these items anyway in the course of acquiring syntax, as shown by the work of Lieven and colleagues (e.g. Lieven et al., 1992, 1997, 2003; Pine & Lieven, 1993), and learning constructions is arguably central to what language learning is all about. Mintz is not concerned with the discovery of frames that are constructions, but rather with finding local cues to category membership, and acknowledges (Mintz, 2003) that many

Frequent Frames are not constructions. Mintz (2006b) has recently begun to show how Frequent Frames may be *used* to discover constructions in language, in cases where a Frequent Frame is embedded inside a larger construction. As an example, Mintz suggests that attending to a Frequent Frame such as "what's _ doing" may facilitate learning about larger constructions such as "what's X doing Y". Nevertheless, the current approach still seems more parsimonious in that it attempts to show how constructions may be learned directly, as evidenced by the comparison to Cameron-Faulkner et al's (2003) manual analysis of the Manchester corpus.

(ii) The second point relates to the exclusive use of "a X b" – style frames. Mintz argues that the Frequent Frames approach shows that lexical classes can be induced even from very limited local contextual information, and earlier, I referred to the definition of context used in Frequent Frames as a "topological" definition. However, the "a X b" frame is not the most primitive topologically-defined context that a word could have; the most primitive contexts would rather be "a X" or "X b". These simpler frames are often problematic for the purpose of lexical categorization. Take for instance the frame "the X". Many of the slot-fillers of this frame are nouns; indeed, when the frame occurs as a full-utterance frame, then essentially all its fillers are nouns. However, when we accept as instances of "the X" *all* cases where the frame occurs in an utterance (as would happen in a Frequent Frames-style treatment), it often happens that "the X" forms the beginning of a phrase of the form "the <Adjective> <Noun>", in which case the filler is an adjective rather than a noun. Hence, the frame "the X", when identified in a context-free manner wherever it occurs in the corpus, is not a reliable cue to the part-of-speech of the filler word. This is in fact true of most of the primitive "a X" and "X b" frames. It is precisely the *combination* of two disjunct contextual words on either side of the focal word that makes the Frequent Frames technique so successful. In the "the X" example, when we know that the word after the slot is "and", the category is pinned down fairly accurately as that of noun. But this begs the question of why the child would automatically consider only these useful contexts. We as adult speakers of English know that it happens to be the case in English that a two-sided context is very useful for determining part-of-speech, but how does the child know that it should pay attention only to “a X b” contexts, and ignore the simpler and more obvious “a X” and “X b” contexts, as well as any contexts with a more complex shape?

It should be noted that Mintz has indicated that a strict adherence to “a X b” structures is not necessarily an indispensable component of the Frequent Frames approach. Viewed in this way, the current work can be regarded as an extension of Mintz’s Frequent Frames work, and one which allows a greater variety of topological context shapes to be considered, under the constraint that they should plausibly form linguistic units (i.e. constructions).

10.2.4 The frame approach compared to Yuret (1998)

The probability-based frames of Chapter 9 are clearly closely related to the work of Yuret (1998). In Yuret’s work, all links between words are undirected, and the probability-based frames were in fact developed with the purpose of extending Yuret’s approach so as to give an explicit statement of the directionality of links between words. For this reason, an *asymmetric* measure of association, ΔP, is chosen. Just as with Yuret’s work, links are postulated between specific words, based on their probability of co-occurrence in the corpus. But in this case, a high ΔP from element A to element B indicates that A is predictive of B.

Another difference between the current probability-based frames and the work of Yuret is that once it has been exposed to a sizeable corpus, Yuret’s model tends to link all words in an utterance into a single large dependency structure, which is desirable as his model is specifically designed for the purpose of finding all such dependencies. In the current work, I am concerned only with identifying a number of very prominent constituent frames which constitute the local context of a focal word and can be used as a clue to its part-of-speech. For this reason, frames are extended only as far as a chain of significant (i.e. suprathreshold) ΔP values stretches. The underlying model is that of activation in consciousness, which spreads from element A to element B if a strong associative link exists from A to B. This allows the current model to pick up slot-filler relationships between words, where the predicted elements are explicitly identified with the specific words in a frame, and the predicting elements with the slot-fillers.

Both Yuret's model and the current probability-based approach suffer from the rather serious problem that they are based on the co-occurrence probability between asymmetric elements (dependent and depended-on element) which is specific to the *pair* of elements in question. In the current case, it is difficult to extend the frame relation to cover other slot-fillers for which there is little evidence of a high co-occurrence probability.

10.3 Differences in the method of focal word categorization

10.3.1 Clustering in EMILE

EMILE takes the approach that both frames and the focal words that occur in them are associated with grammatical categories, an approach which was also pursued in this thesis. However, EMILE induces clusters of frames and focal words in a different way from the techniques that were outlined here. The closest clustering algorithm to the one used in EMILE is the conflict-based co-clustering algorithm, and so it will form the basis for comparison against EMILE in what follows.

Given any particular cluster of word-frame instances, EMILE will attempt to extend the cluster by *randomly* adding contexts or words to the cluster, within the constraint that, if for instance a context is added, a *certain proportion* of the words already in the cluster should appear in that context in the corpus. The proportion specified is a manipulable parameter of the EMILE model. Once a context (word) is added to the cluster, it is assumed that all words (contexts) already in the cluster are compatible with it, and can potentially co-occur with it, even if these co-occurrences have never actually taken place in the corpus.

By contrast, the conflict-based co-clustering algorithm adds on each iteration the context (i.e. frame) or word that will cover the *largest absolute number* of *attested* utterances. If (without loss of generality) a context is added to a cluster, the other words already in that cluster but which have not appeared in that context are not taken into consideration at all.

Both EMILE's constraint of not generalizing a context too far, and the conflict-based algorithm's constraint of only adding the context with the largest amount of support, are heuristics towards conservativeness. EMILE is concerned with making the cluster of *word-context instances* coherent; the conflict-based co-clustering algorithm clusters words and frames *independently*, but clusters them into a shared set of categories and uses as a guide only the data from instances that have actually occurred.

Another crucial point to note is that individual word-frame combinations in EMILE can potentially be assigned to more than one category. (Presumably, actual utterances would be parsed using the grammar as derived from the clusters; the latest version of EMILE (4.1) is derived without considering context ambiguity, however.) The work outlined in this chapter takes a completely different view: it is assumed that a focal word that occurs in a frame slot belongs to one and only one category, and that the category can be determined completely from the identities of the frame and the word. Hence, even though frames and words on their own may be ambiguous with regards to the parts-of-speech with which they are associated, the ambiguity is resolved when the frame and word are combined into an utterance. At the level of an individual cell in the co-occurrence matrix, it is possible to make a unique category assignment for each cell.

This is a modeling assumption, made in order to come up with a workable computational solution to the ambiguity problem. In reality, there are a few cases, even in the Manchester corpus, where the assumption does not hold. For example, the word *work* in *Are you going to work?* is truly ambiguous when the utterance is taken in isolation, without either discourse or situational context. The word *work* could be a verb when the utterance is taken to mean "Are you about to/do you intend to do work in the immediate future?", but it could be a noun if the meaning of the utterance is "Are you travelling to your place of work?". Nevertheless, I would argue that such cases are extremely few in number, and that the assumption of *unicategoricity* for matrix cells holds for the vast majority of frame-word combinations.

Adriaans (1999, p. 30) acknowledges that the categories obtained with EMILE are very seldom traditional parts-of-speech. It may well be EMILE's tolerance of ambiguity *at the matrix cell level* that prevents it from finding these categories.

The EMILE system offers a large number of parameters for the researcher to manipulate, which provides a great deal of flexibility, but makes it difficult to explore the parameter space exhaustively. There is a parameter *type_usefulness_required* in EMILE which allows the user some control over the degree of cluster overlap. Any cluster needs to cover at least *type_usefulness_required* cells in the data matrix not covered by any other cluster in order to be retained in the final clustering. If the parameter value is set to 0, all clusters are retained, even if all their cells are accounted for by other clusters. Setting the parameter value to higher numbers allows the user to discard clusters that overlap to a great extent, potentially making it possible to come up with a small number of large non-overlapping clusters. However, the problem is that *type_usefulness_required* is an *absolute* parameter. Its value is the actual number of cells that need to be contributed by a cluster, and would hence need to be determined anew for each new dataset. More to the point, in a situation where clusters have highly divergent sizes (as is the case for the traditional parts-of-speech in at least the Manchester corpus), it is not possible to provide a "one-size-fits-all" value for the parameter. Had *type_usefulness_required* been a *relative* parameter, i.e. if it was the *percentage* of a cluster's cells that were required to be non-overlapping, then setting that parameter to a value near 100% (i.e. specifying maximally non-overlapping clusters) may well have provided results similar to the clustering results obtained here.

10.3.2 Clustering in Frequent Frames

An important difference in clustering between Frequent Frames and the current approach is that the co-clustering algorithms take into account the possibility that the frames themselves may be ambiguous cue to parts-of-speech, whereas Frequent Frames are taken to be unambiguous. Some evidence to suggest that this might not be the case in all languages comes from Erkelens (2008), who found that the top 45 Frequent Frames from a corpus of Dutch child-directed speech contained a variety of different parts-of-speech as slot fillers. One way to address this issue might be to make use of wider contexts, i.e.

the lexically-specific frames considered here. Another approach could be to apply the co-clustering techniques presented here to Frequent Frames, in order to address both word and frame ambiguity.

In contrast to the hierarchical clustering and co-clustering processes used in the work presented in this thesis, Mintz (2003, 2006a, 2006b) takes a clustering approach where two frames are clustered together if they have more than a certain proportion (20% in Mintz, 2003) of their words in common with each other. This difference in clustering methods may also have an effect on the results of both models. In fact, the way that these issues of frame selection and frame clustering are handled is an important component of each model; for instance, if the current full-utterance frames were to be clustered according to Mintz's approach, it is very likely that they would all end up being bundled into a single, large category (especially since there are so many more full-utterance frames than Frequent Frames, and Mintz's 20%-overlap rule is applied transitively), and hence the model would fail to categorize correctly.

10.3.3 Criticisms of category formation via clustering

10.3.3.1 Freudenthal et al.

Freudenthal et al. (2005) have criticized the Frequent Frames work of Mintz (2003) on grounds that apply to the current work as well, and so it is worthwhile to address their objections here. Freudenthal et al. (2005) note that, while the Frequent Frames model identifies the class of verbs, for instance, it groups together both the root forms of verbs and their inflected forms, which would lead to syntactic errors if, say, an inflected form were to be substituted for a root form in some contexts. Freudenthal et al. (2005) use their MOSAIC model to generate sentences, randomly selecting a particular word and a particular context in which that word occurs, and then substituting it with another word that is paradigmatically linked to the first word. They argue that the low rates of grammatical error produced by this process (as judged by native English speakers) indicate that MOSAIC is a superior model of language acquisition to Frequent Frames, and indeed to any model that "use[s]… co-occurrence statistics to derive syntactic categories" (p. 17).

While it is true that a system which is not able to distinguish between these different kinds of verbs would have serious shortcomings, there are a number of problems with Freudenthal et al.'s argument.

First is their insistence that only generative models are valid ones. I would argue on the contrary that it is incoherent to expect a model derived from heard speech input, and hence usable for listening and comprehension (or at least recognition) to be identical to one used for speaking. The task of speaking is more difficult than that of listening, because it requires a larger number of decisions to be made by the speaker. Most notably, the speaker is concerned with expressing *meaning*, and initiates the process of assembling an utterance from the starting-point of the intended meaning, selecting utterance constructions and words accordingly. This is a different process from merely *randomly* combining words and contexts as is done by Freudenthal et al. (2005) in their evaluation of MOSAIC. I would submit that, whether or not this process produces legitimate English sentences is irrelevant, as this is not how a speaker would produce an utterance in the first place.

Secondly, the fact that the current work, and also the published results of Mintz (e.g. 2003), show large clusters corresponding to verbs, nouns, and adjectives, regardless of any finer subdivisions that may exist in these classes, is due to factors other than the particular contexts used for classification. In the current work, the system was explicitly constrained to produce three sizeable clusters. Likewise, Mintz's clustering approach apparently led to a small number of clusters being created. However, this does not mean that finer distinctions are not obtainable from the frames used. In fact, as shown in Table 11, for larger number of clusters, more fine-grained groupings emerge, including participial versus root forms of verbs, plural count nouns, mass nouns, body parts and clothing, and modal verbs[6].

[6] Likewise, my own informal simulation on the Manchester corpus revealed that applying the same treatment to Frequent Frames as was applied to the frames in the current work (hierarchical clustering and the 5-5 criterion) produced a set of clusters that, at the level of 20 clusters, showed very fine-grained and sharp distinctions between groups of words, including, notably, separate clusters for root forms of verbs,

Finer part-of-speech distinctions are, therefore, available in the current set of models. It seems quite plausible that any particular word, expression, construction, etc. will necessarily be represented at more than just one level of specificity, e.g. the word "milk" in "are you drinking your milk?" is simultaneously a noun, a mass noun, a name for a food substance, etc. To the extent that these details are discernible in the clustering results, the current models are at least in theory capable of accounting for these distinctions.

However, my purpose in this thesis was to account for the three major parts-of-speech of nouns, verbs and adjectives; hence the constraint of producing three large clusters. There is nothing akin to the major three categories in the results from EMILE, for instance, as Adriaans (1999) acknowledges. The work by Cartwright and Brent, as noted, also produces far more than three clusters, and it is likely that the same is true of most of the syntactic models reviewed in Section 4.2, such as ADIOS, MOSAIC, SNPR, etc. While those models produce categories that are exquisitely sensitive to the context in which they occur, there are few mechanisms available to merge many small categories into a few large, robust ones. In my view, a model of language development in English which does not account for the development of the three major categories does not adequately describe language development.

10.3.3.2 Cartwright and Brent

Cartwright and Brent (1997) criticize the use of hierarchical clustering in models of the distributional discovery of parts-of-speech. Their first criticism is that hierarchical clustering potentially produces a huge number of categories, leaving it unclear how many clusters should be used, i.e. at what level the hierarchy should be "cut". One of the major strengths of Cartwright and Brent's (1997) model is that it aims to self-organize the number of categories (as opposed to being constrained to produce three sizeable categories, as was done throughout the current work). This is also to some extent problematic for its quantitative evaluation, because it results in very low completeness values (less than 0.18) when the model is evaluated against a natural language corpus,

for transitive versus intransitive participial forms, for modal verbs, and for the root forms of verbs that take a clause as direct object ("know", "remember", "think", etc).

apparently because it made use of too many categories. I agree that a better case could be made for the discovery of parts-of-speech on distributional grounds if the number of categories in the current model were allowed to self-organize. The results in Section 7.5.1 suggest that the three main categories of nouns, verbs and adjectives are sufficiently prominent "features of the landscape" in child-directed English that the co-clustering algorithms automatically converge on them; notably, the conflict-based algorithm makes use of only 3 categories closely corresponding to these three classes, even when starting out from a division into 20 categories. On the other hand, it is possible that the number of categories may not be provided by distributional information alone, but may be imposed by the semantics of the situation.

Cartwright and Brent also criticize hierarchical clustering on the grounds that they regard it as a highly unfeasible model when it comes to *incrementally* updating the hierarchy of categories. This is a valid criticism, and many familiar algorithms for hierarchical clustering that operate in "batch mode" (i.e. coming up with a clustering for the entire corpus at once), such as the average linkage clustering algorithm of Sokal and Sneath (1963), are probably not adequate models of the updating process for psychological content. However, there are other algorithms that can plausibly be viewed as incremental hierarchical clustering methods. For instance, self-organizing maps (SOMs; Kohonen, 1995) map a multi-dimensional space onto a lower-dimensional one in such a way that similar items in the multi-dimensional space are mapped onto proximate positions in the lower-dimensional space. Consequently, items that are roughly similar occupy the same wide area, and items that are more precisely similar share smaller territories. In this way, one can "zoom in" from large and more abstract groupings to smaller and more specifically related groupings, just as one could move down a hierarchical tree from large, broad clusters to small, compact ones.

In any event, the decision was made in the current context to make use of hierarchical clustering, purely because *some method* of grouping on the basis of distributional similarity was required, and hierarchical clustering has the benefits of being fast to execute and being deterministic, hence expediting the analysis. Details about the dynamics of the categorization process were not considered relevant in the current work, although of course a complete model should adequately describe these phenomena as well.

.

11 Future extensions of the part-of-speech discovery framework

The work presented in this thesis has shown that a number of lexically-specific frames may provide a reliable substrate for the discovery of parts-of-speech, and has also shown that a distributional bootstrapping process might benefit from recognizing the pervasive ambiguity of linguistic items, both focal items and their contexts. There are a number of ways in which this bootstrapping process might be extended; discussion of these will be the focus of the rest of this chapter.

11.1 Extending distributional bootstrapping

11.1.1 Extended coverage of the corpus: generalizing beyond the current data set

The current categorization process provides a categorization for reasonably familiar and commonly-occurring words that occur in similarly familiar frames. But overall coverage of all utterances occurring in the corpus is still fairly low. In general, the categorization process should be able to handle unfamiliar words in familiar contexts, and familiar words in unknown contexts. It is fairly straightforward to extend the current approach to do so.

In the three co-clustering algorithms, frame and word information is combined in order to arrive at a final part-of-speech. In cases where either the word information or the frame information is missing, the categorization algorithm could merely select the most-strongly-associated category for the item that is present. In this way the algorithms can also account for utterances where one of the frames in the data set is used with a word that is not in the data set, and utterances where a word in the data set occurs outside of any of the frames in the data set. These constitute context-free identifications of frames and words, and the best guess possible is the majority category of the item in question.

11.1.2 Heuristics about the structure of frames

There are two heuristics that, when applied to the frame-discovery procedure, yield more selective coverage but greatly improved evaluation results. One of these is the prohibition of multiple X consecutive slots, used in the full-utterance frames and nested frames (but not in the prediction-based frames). Relaxing this constraint impairs correctness (see Section 7.5.2). This constraint was motivated by arguing that it takes into account that a specific adjacent word is a more reliable context than a variable slot for the purpose of categorization, and possibly also that it is easier to segment a speech stream into words (and hence recognize a frame) when the rare words are flanked by frequent ones. This constraint is not in force in the case of prediction-based frames: here, slots can occur in sequence if there is enough evidence of predictability in the corpus for specific instances where the parts of the entire frame can be predicted. It is quite possible that the poorer performance of the prediction-based frame categorization may be partly due to the absence of this constraint.

The current full-utterance frame approach could of course be extended by first inducing categories as was done here, then relaxing the no-multiple-slots constraint and allowing known words to stand both as Xes and as their *majority* category under the original categorization. This allows us, for instance, to *guess* that "glass" in "the empty glass" and "the glass broke" is a noun in either case. This allows us to find the frames "the X Noun" and "the Noun X". Because of the ambiguity of word types, this will be a heuristic rather than a hard-and-fast rule. Nevertheless, if the frame is a valid one, we would expect a larger amount of evidence for it than for spurious frames.

The second heuristic that would have improved categorization was harder to motivate and hence was not exploited in this thesis. Informal experimentation (not reported here) has indicated that a very significant increase in correctness is obtained by considering only frames that *start* with a specific word (rather than a slot). This heuristic was also seen at work in the frames considered by Cameron-Faulkner et al. (2003): in their manual analysis of the Manchester corpus, they considered only frames that were defined by their initial two or three words. This constraint seems to exploit the fact that dependent words tend to occur before their heads in English (e.g. "*to* ask", "*the* caterpillar", etc.). However, prior knowledge of this tendency is not something that one would want to build into the model.

11.1.3 Morphological awareness

An important extension to the current approach would be one that is sensitive to *morphological inflection* in deriving frames. For instance, a very salient clue in English that a particular word is a verb is its frequent occurrence before the suffix “-ing”. Being able to exploit this kind of cue would greatly facilitate identifying verbs from context. Some morphemes are ambiguous too: consider “-s” which can be used to pluralize a common noun, to form the possessive of a noun (when we ignore conventional English spelling, which is not part of the input to the language-learning child anyway), or to form the third person singular form of a verb. This means that we would have to consider the morpheme in the context of a larger frame: in “That one X-s X”, the filler before the “-s” is likely to be a verb, in contrast to the frame “Are they your X-s?” where it is likely to be a noun.

11.1.4 Finding functional categories

In the current approach, because most of the function words in English ended up as frame-building words, very few of these words were placed into categories. This was true to some extent for the prediction-based frames, but especially for the full-utterance frame and nested frames approaches. However, it might be the case that this highly lexically-specific approach is too strict. The determiners and possessive pronouns in the frames “That’s a X”, “That’s the X”, “That’s another X”, That’s your X”, etc. clearly serve a very similar purpose, and occur in highly similar contexts. It might be possible also to discover function word categories, based purely on evidence from the existing set of frames. This could be done by grouping together frame-building words that can be substituted for one another from one frame to the next, as the words before the X slot in the examples above can. Again, it would be desirable to cluster function words together not on the strength of one or two pairs of aligned frames, but to consider the weight of the available evidence for putting any two function words together into a category.

11.1.5 Extending the prediction-based approach

It would also be worthwhile to explore other ways of creating and combining frame primitives in the prediction-based approach. An especially fruitful technique might be to adapt the approach to render an account of hierarchically nested frames. One way in which this might be achieved could be to follow the approach followed by Yuret (1997) with undirected dependency graphs based on mutual information, where constituents were linked in order according to the relative *strengths* of associative relationships, and linked elements were chunked up to form units that could then form associative links with other words. This is appropriate in the current case as well: as constituents can appear in many different contexts, the associative relationships within a constituent may be expected to be stronger than the associations from the constituent to the surrounding context.

11.1.6 Other languages

All experiments in this thesis focused on English exclusively. It is *not* crucial for the acceptability of the models of lexical categorization presented here that these techniques should work for all languages. Most likely, the language-learning child exploits any information available in the environment opportunistically, noticing regularities wherever they can be found and learning associations between any features that co-occur, whether linguistic, semantic or phonological. The function words provide a very reliable source of information regarding parts-of-speech in English, and a wealth of evidence, as reviewed in Chapter 3, shows that children are sensitive to the combinatorial implications of English function words. To the extent that another language exhibits a similarly salient "backbone" formed by a set of specific words, the techniques discussed here should work for that language also. Languages that are relatively impoverished in terms of function words may not be analysed as easily. It still remains to be seen whether the current approach would be successful on corpora from other languages.

The most important and fruitful extension to the current system is likely to involve incorporating semantic information with the existing distributional information; this is the subject of the rest of this chapter.

11.2 Semantic aspects: MicroJaea

The experiments in this thesis have considered distributional approaches to part-of-speech induction only, and have shown that distributional information is rich enough to form the major classes of Noun, Verb and Adjective, and to deal with ambiguity by combining word and frame information.

Nevertheless, the algorithms discussed in the previous chapters still did not provide a perfectly accurate categorization of words in context, even with the limited set of contexts that were actually considered.

An important issue has to do with the substantive content of the categories that are induced. The experiment by Brown (1957) shows that children are able to draw semantic inferences about a novel word based on the context in which it occurs, something that is clearly not possible in an approach that neglects meaning entirely.

Furthermore, all the experiments in this thesis have at their core the notion of forming clusters. It is presumed that frames are treated as being similar to other frames based on the complete pattern of words used in those frames in the child's prior experience.

This notion suggests that there is a mental entity corresponding to the cluster, possibly physically instantiated as a neural circuit, which is activated by frames belonging to that cluster, or by frames with a usage history that makes them similar to the typical pattern of usage for that cluster (whether stored in memory as exemplars of frame histories, or a single prototypical word usage pattern).

Any frame which has been explicitly associated with a category cluster by whatever means will activate the mental representation corresponding to the category itself (and possibly weakly activate representations for the other frames associated with the category as well).

However, this need not be the only way in which a category is activated, and, given the suggestion by Langacker (1987) that parts-of-speech can be defined entirely in semantic terms, it seems unlikely that only distributional information is relevant. For Langacker, the way in which we are *intended to construe* a word is what makes that word a noun, verb, adjective, etc. If this information is available even for a few words, it would seem to be very important information to store in association with the cluster representation. Under this view, *any* usage of a word where the word is clearly meant to refer to an entity, whether physical, abstract or figurative, activates the representation of the noun category, and allows the new word usage to be associated with that category. Thus the semantic aspects become part of the substance of the category, in that the semantic knowledge can evoke the category, and can also be invoked by activation of the category, allowing semantic inferences to be drawn.

Because of these considerations, the major focus of this research project in the future will be to provide computer models with semantic information in conjunction with distributional information. The focus in this thesis fell on techniques for finding and evaluating frames and for creating clusters of word-frame instances. As a complement to this, I have developed a virtual-world simulation software system intended for use in language learning experiments, which will form the basis of the research to be carried out on lexically-specific frames in the future (this system was developed as part of my thesis work, and the original intention was to use it to collect data to investigate the development of parts-of-speech on semantic grounds). I will now describe the virtual-world software and outline the experimental approach to be taken.

11.2.1 The MicroJaea system

The virtual-world software system MicroJaea was developed for exploring grounded language learning in general. It is implemented in Java3D (version 1.3.1), and can therefore run on any underlying computer platform, and with either DirectX or OpenGL graphics support. MicroJaea was based on the specifications for the Magrathea virtual-world system (Hume, 1984).

11.2.2 Scripting language

The purpose of MicroJaea is to allow a researcher to create short animated scenes by writing scripts in a custom scripting language (MicroJaeaScript) and to play the scenes in real-time on a computer display. MicroJaeaScript allows a script writer to (i) define the shapes of all objects that will populate the virtual world, (ii) define the actions that these objects can carry out, if any, and (iii) create a script which places selected objects on the "stage", and defines points in time at which various objects carry out their actions. During rendering, these objects are placed in a three-dimensional animated scene, and perform their actions at the designated times.

Object Type Definitions: The major component in a MicroJaeaScript script is the definition of object types. These are abstract definitions of types of objects, both inanimate (chairs, tables, footballs) and animate (animals, people), both in terms of their shapes and their behaviour.

> Object Shapes: MicroJaea has access to only four primitive geometric shapes: cylinders, ellipsoids, cones and blocks. However, objects can be composed out of any number of these shapes, allowing an arbitrary number of objects (including people and animals) to be constructed. Primitive shapes are attached to each other at joints, and the kind of motion allowed at the joints can be defined. MicroJaea supports three kinds of joint movement: free rotational movement, rotation around an axis, and translation along a fixed line. Primitive shapes can be defined to have specific physical sizes and colours.
>
> Object Actions: There are two kinds of primitive actions, translation and rotation. However, primitive actions can also be combined into more complex actions. Action sequences consist of a number of actions (primitive or complex) carried out one after the other. Action routines consist of a number of actions carried out simultaneously in parallel. Action routines and sequences can be nested inside other action sequences and routines. In addition, there is a "signal" action which allows objects to emit a signal to which other objects may react when it occurs. This is one way in which coordination between the actions of different objects may be effected.

Scene Scripts: After defining object types, the remainder of a script in the MicroJaea scripting language is devoted to declaring a number of instances of the object type, together with their initial positions and orientations in the scene (analogous to a cast of characters and stage instructions), and then to define the events that occur during the movie scene (analogous to an actor's script). Events can also be viewed as stimulus-response pairs: one type of event describes the action taken by a character when it receives a signal from a particular object or another character; another merely schedules an action to take place at a certain point in time after the start of the animation.

11.2.3 A framework for language learning experiments

The purpose of MicroJaea is to facilitate language learning experiments. This can be achieved by combining visual information obtained from an animated scene with utterances in the target language that describe the events in the scene (Li Santi, Leibbrandt & Powers, 2007a, 2007b; Powers, Leibbrandt, Li Santi & Luerssen, 2007). In other words, the intention is to use MicroJaea for grounded learning of a subset of a language, in the manner of the well-known L0 language learning problem (Feldman, Lakoff, Stolcke & Weber, 1990). Prior work has shown the feasibility of language learning experiments using this platform (Li Santi, 2007).

11.2.3.1 Visual scene information output

During playback of an animated MicroJaea movie, the system produces a sequence of output records. Each record is produced after a certain periodic time interval, and provides details of the objects that are present in the scene at the time, their parts, the locations in space of each of the object parts, and any other visual information such as the sizes of the parts and their colours. The record is therefore a static snapshot of the objects in the scene at that particular point in time, and hence arguably represents (a very stylized version of) the visual information available to the child during the event portrayed in the scene.

Table 71 shows an example of a visual record for a simple compound object consisting of a red sphere perched on top of a blue cylinder. The record gives details of the names of the object (Fred), its type (VeryBasicMan) and of its parts (Head and Body), and specifies where the parts are in three-dimensional space, as well as their colours and dimensions (*shapeParameters*), their original rotational attitudes (*rollPitchYaw*), and the points at which the Head is joined to the Body.

```
{RECORD TRIGGER TIME_PERIOD AT TIME 578

      {OBJECT REF# 7 NAME Fred TYPE VeryBasicMan

            {PART REF# 10 NAME Fred.Head
                  parent Fred.Body join ( 0, 0.2, 0 )
                  position ( 0.25, -0.2, 0 )
                  shape sphere
                  shapeParameters ( 0.05 )
                  colour ( 1, 0, 0 )
                  rollPitchYaw ( 3.14159, -0, 0 )
            }

            {PART REF# 9 NAME Fred.Body
                  position ( 0.2, 0, 0 )
                  shape cylinder
                  shapeParameters ( 0.0080, 0.4 )
                  colour ( 0, 0, 1 )
                  rollPitchYaw ( 3.14159, -0, 0 )
            }
      }
END RECORD}
```

Table 71. An example visual output record from MicroJaea.

11.2.3.2 Language input/output

The MicroJaea scripting language also allows script writers to add language information to the movie. One can imagine that the scenario is one that a language-learning baby would experience, where there is something happening visually, and a speaker of the target language (e.g. a caregiver) is speaking to the child at the same time. The content of at least some utterances can be expected to be in some way relevant to the scene being experienced, whether overtly commenting on events or commenting on aspects of the social interaction. This linguistic information is also inserted into the output record of the unfolding movie. The intention is that a language-learning computer program can read in the visual and time-synchronized language information, and attempt to work out the meanings of words and sentence structures based on what is visible in the scene[7].

There are a variety of techniques in the literature for learning grounded meanings of words from simultaneous language and sensory information. For instance, a very

[7] Of course children have access to a wide array of sensory modalities when experiencing reality, and nothing in theory precludes the addition of other sensory modalities to the system.

influential model has been the cross-situational learning paradigm of Siskind (1996). In Siskind's model, utterances are combined with statements in a first-order-logic-like description language, which state the meaning of the utterance using semantic primitives such as BALL, WALK, etc. Learning consists in attempting to map words onto their referents. This is done by compiling for each word a list of hypotheses about its possible meanings, and then filtering this list by a process of elimination as more and more utterance-meaning pairs are processed, in accordance with a number of heuristic language learning rules. For instance, for any particular word in an utterance, any currently hypothesized meanings which are not present in the corresponding utterance meaning descriptor may be discarded. Also, if a particular semantic primitive is not present in the hypothesis set of any of the words of the utterance, it needs to be added to each one. This cross-situational learning process is continued until each word has exactly one associated meaning.

It is possible to apply this framework to the current problem of lexical categorization. Here, we would be interested not just in the meanings of specific verbs, nouns, and adjectives, but rather in the abstract semantic concomitants associated with the class of verbs, the class of nouns and the class of adjectives. In other words, what is at issue is what has been termed the "notional" meanings associated with a part-of-speech. Particularly appealing is Langacker's (1987) view of parts-of-speech as being grounded in the mental interpretations that are imposed on a word depending on whether it has been used as a noun, verb, adjective, etc.

Also relevant in this regard is work on language development by Smith and colleagues (e.g. Landau et al., 1988, 1992; Jones et al., 1991; Smith, 2001). Smith argues that the act of *naming* an object becomes a familiar situation for young children, presumably in terms of not only the situational and social correlates of object naming, but also the stereotypical frame in which the object name is embedded (e.g. "That's a _", "This is a _"), and so comes to serve as a cue which directs selective attention towards the aspects of reality (in this case, the *shape* of the named object) that are relevant to the meaning of the novel word.

It is an interesting but open question whether children's developing knowledge of verb and adjective frames as reviewed in Section 3.4 can also be interpreted in terms of learned cues for selective attention. So, for instance, a verb frame may draw attention to the kind or path of motion that a living entity performs, or to a comparison between an earlier and a later state of an object or entity (which is carried out in memory). An adjectival frame may focus attention on properties of an object *other* than its shape. Solving this general problem of learning where to direct attention is a special instance of what is known as the "frame problem" in artificial intelligence: how does a computer program or an organism know how to define the scope and boundaries of the problem that it has to solve?

Note that, before the motion or properties of the referent object can be attended to, it must first be determined *that* a referent object needs to be singled out, and then *which* object it should be. The referent is most often determined by contextual knowledge, e.g. it is the referent on which the child and caregiver are already focused, or it is the one to which the caregiver has now directed his or her gaze. Correctly attending to the environment in order to learn the meaning of a verb or adjective therefore also presupposes the application of implicit knowledge about how to focus attention to locate the referent, even prior to preparing to attend to its motion or properties.

At the same time, following Langacker (1987), the child should also be learning how to set up the correct representation in her mind that will allow her to understand the meaning of the utterance. It is interesting to speculate that a linguistic context or frame may also serve as a learned cue to set up the basic required semantic representations in the mental domain. In the case of a verb, there should be a representation corresponding to what Langacker calls an entity, and another representation corresponding to a process, with the two representations linked.

Only a small number of computational models have explored the grounded learning of the semantic correlates of parts-of-speech. A study by Roy (2002) is arguably one of the benchmark works in this area. Roy collected data from participants who were asked to verbally describe a geometric shape on a computer screen (e.g. "The narrow purple rectangle below and to the right of the blue square"). A description of the referent was created in terms of 8 semantic dimensions (such as height/width ratio, area, etc.). Subsequently, Roy defined two distance metrics, one for expressing the semantic distance between the referent objects of two words, and the other for expressing the linguistic (distribution ally-based) distance between two words, and clustered words together based on a distance measure that was a weighted sum of the semantic and linguistic distance. Some of the classes formed in this way were readily interpretable linguistically: for instance a class consisting of the elements *pink*, *yellow*, *salmon*, *orange*, *grey*, *red*, *green*, *purple*, *colored*, *blue* and *brown* was clearly a class of colour words. Also, *leftmost* and *rightmost* formed a cluster together, as did *lowest* and *highest*. At the same time, it was possible to obtain semantic correlates for each of the classes in terms of the semantic variables that were relevant to the meanings of the class members. For instance, for the colour words, the important semantic features were the red, green and blue components of the colour of the object. By implication, for any new word which behaved in the same way distributionally as words in the colour word class, the referent would be likely to have salient red, green and blue colour values.

The work by Roy (2002) certainly captures the flavour of what is proposed here. However, a number of important issues were neglected. Notably, the frame problem was already solved for the learning system in advance: in each case, the referent was already known, and the relevant word in the utterance which made reference to it was identified manually. It would be preferable for the system to solve the frame problem itself.

Secondly, essentially the only kinds of words used in this experiment were nouns and adjectives, thereby allowing word meanings to be learned from a static representation of the situation, rather than a dynamically unfolding one. But we would like to be able to learn about all parts-of-speech without restriction, including verbs, which would require handling dynamic visual information.

The language model used by Roy was extremely crude, and more sophisticated models (including the ones considered in this thesis) should improve the system's ability to learn about parts-of-speech.

In the experiment proposed here, the meanings of individual words will be learned first, by some process such as cross-situational learning. At the same time, the meanings of the open-class categories such as nouns, verbs and adjectives are slowly abstracted out from these individual words.

The process of understanding the world, including the language spoken in it, entails creating a mental representation of that world. Initially, the representation is constructed from what can be perceived through the senses; later, such world representations may be assembled from the imagination, or in response to hearing or reading a sentence that describes a particular situation.

One of the abilities that a child acquires during the course of development is to construct such a mental representation from linguistic input. This can be regarded as a constructional *procedure*, triggered by cues in language, and may therefore be subject to the kinds of learning processes often implicated in behavioral learning, for example reinforcement learning.

In addition, the child needs to learn how to pay selective perceptual attention to the parts of the world that are relevant to the meanings of utterances. This may be presumed to be driven by two sources of motivation: firstly, curiosity about how language works and how it is related to the world, and secondly, a drive to obtain more information about the world itself, with language understanding being one powerful way of doing so. The tuning of selective attention may also be presumed to be a learning process; a plausible candidate might be reinforcement learning (Sutton & Barto, 1998).

Desirable (reinforceable) behaviour corresponds to performing all the steps required to attend appropriately to the environment in order to determine the meaning of a novel word, and successfully establish a mental representation of its meaning. Unsuccessful behaviour, which should be inhibited, would be incorrectly attending to aspects of the sensory input, leading to no or an incorrect hypothesis about the word meaning, or otherwise setting up a mental model which turns out to be incongruent with later semantic information (often prompting a search for a better representation).

For this reason, the system should be equipped with a number of primitive operations for respectively directing selective attention, and building mental representations. These primitives can be combined into more complex procedures by performing several primitive operations in sequence, and the processes by which these sequences are created may be hypothesized to be subject to the same learning processes governing the learning of other human behaviours such as motor sequences. The attentional primitives, when operating on a visual output record as in Table 71 could be as simple as the operation to select one of the fields for attention. In a more developed model, one could more accurately model human gaze, and primitives might include moving towards a local minimum on a saliency map. Representational primitives include creating an entity record, creating a process record, or annotating these with property information.

Suppose that the proposed computer model already has in its lexicon the lexically-specific construction “that’s a X”. Now consider, for example, a virtual world depicting a collection of toy animals on a tabletop, representing the visual scene when a child is sitting with its caregiver and playing with the toys. The system would then be exposed to a number of utterances such as “That’s a hippo”, “That’s a camel”, etc., with the toy hippo, camel, etc, being the focus of the caregiver’s attention at each stage. (There is ample evidence suggesting that children are sensitive to the focus of a speaker’s attention when they speak, and will form hypotheses about the meanings of novel words that are congruent with where the speaker is looking, rather than, say, where they themselves are looking (e.g. Baldwin, 1993). How the focus of the speaker’s gaze would be indicated in the proposed system will remain to be decided; initially, it may be expedient merely to indicate explicitly that attention is focused on one particular object.)

The “correct” way to process an utterance of the form “That’s a X” is to find the target object of the speaker’s gaze, scan it visually in order to determine its shape only (ignoring other potential visual properties such as colour, size or texture), then create in working memory a representation record of a single entity, with an associated property record that will contain information about its shape. Successfully bringing about this state of affairs in response to encountering the linguistic cue “That’s a X” would indicate understanding of (part of) the abstract meaning of the object name slot in the “That’s’ a X” construction. Initially, however, one might expect that this process is errorful and inefficient, with many incorrect aspects of reality being attended to and inadequate representations being set up.

The feedback about whether the current set of operations was correct or not is likely to come from the *expectation* that the system has about reality. Recall that, in Siskind’s cross-situational learning system, a set of meaning hypotheses is maintained for each word. A natural way to implement this in the current scheme may be to associate with the word and utterance context the steps required to set up a mental representation of all possible aspects of reality (present in the scene) to which the word may refer. Then, when this representation is reenacted at a later stage and one or more of the aspects of the representation are not present in reality, the operations contributing to those representations are inhibited, while the operations contributing to the representations that are in line with reality are reinforced (broadly in accordance with the principles of reinforcement learning). This corresponds to discarding incorrect hypotheses from the hypothesis set when they are disconfirmed by reality.

While this system is learning the meanings of individual words such as “hippo”, “camel”, etc., it is also associating these meanings with the context (in this case the frame “That’s a X”) in which they occur. The end result will be that the aspects of the eventual mental representation that these utterances have in common will be strongly associated with the frame in which they all occur, whereas the aspects of meaning peculiar to the individual words will cancel each other out. Hence, the frame may be expected eventually to have associated with it the operations identified above, which focus on the shape of an object, and place an entity representation with shape information in working memory. The specific operation associated with “hippo” will then embellish this representation with the specific shape details required for the concept of a hippo.

It is important to note here, in accordance with Langacker (1987), that it is not the fact that the entities in question are *physical things in the world* which is relevant here, but the fact that they are *conceptually* represented as bounded entities which may metaphorically take on some of the characteristics of objects. Hence, even abstract nouns could be accommodated in this framework.

In the above proposal, the frames are supposed to be known to the system already, and could have been derived according to any of the three procedures outlined in previous chapters. In an alternative (converse) approach, it might be possible for scenes to be perceived “wordlessly” to start with, and for a number of semantic categories such as Entity, Process and Atemporal Relation to have self-organized out of the original mental representations of these scenes. Then at a later stage, distributional information comes into play, and frames and words are associated with the semantic clusters. Yet a third possibility is for semantic and distributional information to be available simultaneously, and for clusters to form out of expression-meaning pairs that have both semantic and distributional information in common. This last proposal would be the one that is most strongly compatible with Construction Grammar.

To the extent that the basic sentence frames of a language (see Goldberg, 1995) are vital to an understanding of the language, it is important for the child to come to grips with the semantic implications of each of these possible frames. However, the task of understanding, say, an utterance using a transitive frame, would be made simpler if the child (i) had some rough notion of the meaning of the verb, and of the words for any entities that are mentioned in the utterance, and (ii) had a number of examples at their disposal of how the arguments of a verb are arranged in *lexically-specific* frames that instantiate the subcategorization frame.

Currently, it is envisaged that a researcher would make up scenes and associated dialogue. However, it might also be possible to elicit linguistic information from native English-speaking informants. This could be done by devising a number of animated movies, then inviting caregiver-child dyads to view the movies together, and recording the comments of the caregiver to the child as the movie is played.

Appendix 1: Preprocessing performed on the Manchester corpus

In order to turn the sentences spoken by mothers into simple, uniformly-formatted sentences that could be batch-processed by a computer program (referred to hereafter as the "cleaned-up corpus"), it is necessary to perform a certain amount of preprocessing. CHAT makes use of a variety of non-alphabetic characters for corpus annotation. In addition, some sentences are not necessarily complete, and it needs to be decided which ones represent usable data. Non-alphabetic characters need to be removed, and unsuitable utterances discarded, in order to produce a "clean" corpus on which experiments can be carried out. The following phenomena are at issue:

Interruptions: When utterances are interrupted (whether by the speakers themselves or others), they are marked by the symbols [+/. +/? +//. +//?] at the end of the line. These utterances are therefore necessarily incomplete, and hence do not provide promising material from which to extract valid full-utterance constructions. Interrupted utterances may arguably differ from complete ones in their intonation contour, given that the typical intonation contours associated with both statements and questions are most distinctly marked only at their endings (by falling and rising intonation respectively). Nevertheless, the Manchester corpus does not provide information about intonation, and we cannot assume that children (especially at the early stages of language acquisition) can competently distinguish between finished and unfinished utterances. Hence, interrupted utterances should be included in the cleaned-up corpus.

Trailing off: Utterances where the speaker trails off without completing the utterance (and without having been interrupted) are terminated with the symbols [+.. +.? +!?]. The same comments as for interruptions apply, and trailing-off utterances should be included in the cleaned-up corpus as well.

Retracings: In some instances, speakers terminate a particular "path" in their utterances, fall back to an earlier point in the utterance and resume the utterance with new words.

The material which did not form part of the final "approved" utterance is followed in CHAT by [/] or [//] (and is surrounded by angle brackets < > when it is longer than one word). Arguably, the intended utterance may be recoverable by the child, by "working out" which part was corrected or altered and assembling the intended utterance from the rest of the spoken material. But again, we cannot assume these abilities on the part of the language-learning child, and so need to include the entire utterance, including the part which was "retraced over".

Omitted words: On occasion, researchers include words which should have been part of an utterance but were in fact omitted, and mark these omitted words with a 0 (zero) before the word. Given the "warts and all" approach taken here, however, it is more appropriate to omit these words from the cleaned-up corpus.

Phonological words: Some utterances contain non-words that are phonologically coded by the researchers, and are preceded by an & (ampersand). These words should be included in the final utterance, even though they are not accepted English words, as they formed part of the input.

Unintelligible words and utterances: When a coder is unable to make out a word or sometimes an entire utterance, the word or utterance is represented by one of the symbols [xx, xxx, yy, yyy, www]. In these cases, we truly cannot make use of the utterance, because we don't actually know what it was (and it is reasonable to consider that the child might not have known either). An utterance that contains any of these markers should be discarded from the data set.

Quoted speech: Quoted (or reported) speech is marked in CHAT by placing the quoted material in double quotes, sometimes combined with other non-alphabetic characters. These utterances are not really complete utterances in their own right, but rather either two concatenated utterances, or one utterance sandwiched into another. Once again, however, it cannot be assumed that the child parses the utterance in the "correct" adult way, and hence these utterances are included in the cleaned-up corpus as they occurred (after removing the quotes).

Pauses: the symbol # indicates a pause, and is omitted from the data set.

Compound words and collocational phrases: Compound nouns (such as “diesel truck”) and collocational phrases (such as “once upon a time”) can be indicated as such in CHAT by joining their constituent words by plusses and underscores respectively. In preprocessing, these symbols are simply omitted, and the constituent words are concatenated into one word, effectively acknowledging the coders’ intuition that they form a single “long word”.

Special form markers: Some words are followed by descriptors to indicate that they belong to specific types of word (e.g. letters of the alphabet, onomatopoeic sounds, etc). These markers should, of course, be omitted when creating the cleaned-up corpus.

Contracted words: Related to the omitted words are cases where parts of words are omitted. This is of course conventional in spoken English for some words, and the position of the omitted material is indicated by an apostrophe (e.g. we’re, that’s, etc.). In CHAT, the omitted material is occasionally included in parentheses. These parenthetical elements should be removed and replaced by apostrophes.

Punctuation: Utterances are terminated in CHAT by either a full stop, an exclamation mark or a question mark. I have taken the view that utterances ending with question marks typically exhibit a distinct intonational contour identifying them sonically as questions, but that utterances ending with full stops and with exclamation marks are not so reliably distinguishable from each other, and indicate perhaps nothing more than increased volume or heightened affect in the case of exclamation marks. While these characteristics may indeed be detectable by the child, they seem to be less important linguistically, compared to the function played by rising intonation in identifying

questions. For this reason, the only “marked” punctuation marker in the cleaned-up corpus is the question mark, while all exclamation marks are turned into full stops.
Commas are retained in the cleaned-up corpus, with the qualification that we do not need to distinguish between tag questions and other syntactic junctures (indicated in CHAT by respectively two commas and a single comma) – these are all coded as a single comma.

Fortunately, the CLAN software, developed for working with CHAT data and available free-of-charge from the CHILDES website, provides a built-in command, FLO, that performs most of the preprocessing described above. The purpose of FLO is to convey the “flow” of speech as it occurred, in a slightly more readable format than the full CHAT-annotated text on the main tier line. FLO “strips out markers of retracing, overlaps, errors, and all forms of main line coding” (CHILDES CLAN manual, p. 71).
FLO removes nearly all non-alphabetic characters, but leaves interruptions, trailing-off utterances, retracings and quotes as they are. Omitted words and special form markers are omitted, phonological words are included, and contracted words are contracted (with the addition of apostrophes).

On the other hand, FLO retains all instances of unintelligible words, certain non-alphabetic markers, and double commas. For this reason, all files are preprocessed in two steps, firstly by issuing a FLO command on the original data file, and then by running a specific Java preprocessing program on the data file that was output by FLO. The preprocessing program removes utterances containing markers for unintelligible words, removes certain CHAT characters not removed by FLO, changes double commas to single ones, omits utterance-final exclamation points and full stops, and changes all upper-case letters to lower-case. All utterances appear in a single line on their own in the cleaned-up corpus.

In addition, the name of the child in each file is automatically changed to the token *childname*. This is done because a child’s own name may well have a significant role in the spoken utterances that the child hears. It is likely to occur very frequently, and to occupy a significant position in sentence structures (in many cases, the child’s name is used almost as a synonym for the second-person pronoun *you*). In order to allow the computational procedures to pick up this regularity when the data from all twelve mothers is pooled, it is necessary to use a standard token for the child's name. The resulting cleaned-up corpus contains only the lower-case alphabetical letters, the question mark, apostrophe, comma and space. Each line contains exactly one utterance.

Appendix 2: Full comparison of current full-utterance and nested frames against the frames identified by Cameron-Faulkner et al. (2003)

This Appendix shows the ways in which the frames manually identified Cameron-Faulkner et al. (2003; CFLT) match against those produced by the procedures presented in this thesis. Core CFLT frames are in bold and preceded by asterisks.

Nesting (as determined in Chapter 8) and full-utterance frames (as determined in Chapter 6) are regarded as a single coherent approach; multiply-nested frames are shown only when no nesting frame match exists. Prediction-based frames are as determined in Chapter 9.

In a few cases, it was difficult to determine prediction-based matches, especially for CFLT frames which required the fillers to be of a particular grammatical category. The frames produced here do not, of course, make reference to a particular category when they are produced, as it is exactly the category of the filler which remains to be determined (and the frame itself may be ambiguous for that reason, as discussed). In such cases, a prediction-based frame was taken to match a CFLT frame if there was a reasonable way to extend the prediction-based frame (by appending a sequence of words) so as to produce the required arguments. For instance, to the CFLT frame “he’s [VP NP]”, we can match the prediction-based frame “he’s been X”, on the grounds that this frame could occur e.g. in an utterance such as “he’s been fighting a fire”.

For full-utterance frames, not only exact matches, but also matches that are somewhat more specific than the CFLT frame are shown.

Fragments

Cameron-Faulkner et al. (2003) frame	Nesting frame	Multiply-nested frame	Full-utterance frame	Prediction-based frame
***a [N]**	N/A[8]		a X	a X
***the [N]**			the X	the X
[Adj] one			X one	X one
that [N]			that X	that X
not [N]			not X	not X
this [N]			this X	this X
some [N]			some X	some X
poor [N]			poor X	poor X
another [N]			another X	another X
more [N]			more X	more X
big [N]			big X	big X
***[Numeral] [N]**			one X two X three X	one X two X three X four X

[8] Some frames identified by Cameron-Faulkner et al. are marked as "Not Applicable" to nested frames, because they make reference to single words of a particular part-of-speech. The nested frames are not restricted in this way, but can accept material of any length.

***[Poss][N]**			his X her X your X my X	his X her X your X my X our X their X
[Colour][N]			red X yellow X blue X green X orange X	red X yellow X blue X green X orange X pink X white X
not [VP]	not Y		not X not X it	not X
put [NP]	put Y		put the X	put X put the X
don’t [VP]	don’t Y		don’t X don’t X her X don’t X him don’t X it X don’t X it don’t X me don’t X that don’t X the X don’t X them don’t X too X don’t X your X don’t put your X in there	don’t X don’t X it X don’t X him don’t X it don’t X it X don’t X that don’t X them don’t be X don’t get X

got [NP]	got Y		got X got no X	got X got a X got no X got some X got your X
make [NP]	make Y		make a X make some X	make X make a X make some X
can’t [VP]	can’t Y			can’t X
have to [VP]		have [to Y]		have to X
draw [NP]	draw Y		draw X	draw X
***in [NP]**	in Y		in X in a X in his X in that X in the X in your X	in X in a X in his X in that X in the X in this X in your X
on [NP]	on Y		on X on a X on her X on his X on my X on that X on the X on your X	on X on her X on his X on my X on that X on the X on this X on your X

with [NP]	with Y		with X with a X with his X with the X with your X	with X with a X with her X with his X with the X with this X with your X

for [NP]	for Y		for X
			for a X
			for the X
			for your X
				for X
				for a X
				for the X
				for your X

over [NP]	over Y			over the Z
at [NP]	at Y		at X at the X	at X at the X
like [NP]	like Y		like X like a X	like X like a X
very [Adj]	N/A		very X	very X
[Pronoun] isn’t	not covered			X isn’t

Table 72. Comparison between Cameron-Faulkner et al. (2003)'s fragments and the current frame approach.

Wh-questions

Cameron-Faulkner et al. (2003) frame	Nesting frame	Multiply-nested frame	Full-utterance frame	Prediction-based frame
***what's _**	what's Y?		what's X ? what's X doing ? what's X now ? what's X there ? what's a X ? what's <child's name> X? what's happened to X ? what's happened to his X? what's happened to your X? what's he X ? what's in that X ? what's in the X ? what's in this X ? what's in your X ? what's it X ? what's on the X ? what's on your X ? what's she X ? what's that X ? what's that X called ? what's that X doing ? what's the X ? what's the X called ?	what is X what is it X what is that X what's * X what's X X what's X what's X doing what's X got what's childname X what's happened to X what's happened to the X what's he X what's his X what's in the X what's mummy X what's she X what's that * X what's that X X what's that X what's the X what's the X X what's this X

			what's the X doing ? what's the matter with your X ? what's this X ? what's this X doing ?	
***what’re _**	what’re we Y? what’re you Y?		what’re we going to X? what’re you X? what’re you X about? what’re you X for? what’re you X now? what’re you doing with your X? what’re you going to X?	what're * X what're you X X what're you X what're you doing X what're you doing to * X what're you going to X what are * X what are X what are we X what are you X what are you X X
***what do _**		what [do you Y]?	what do X do? what do X eat? what do X say? what do you X? what do you X? what do you want for your X? what do you want to X?	what do X X what do X do what do X say what do you X X what do you X what do you like X what do you mean X what do you think * X what do you think X what do you want X

***what did _**	what did you Y?	what [did you Y]? what [did we Y]?	what did X do? what did the X do? what did you X?	what did * X what did X do what did he X what did she X what did they X what did you X X what did you X
***what colour _**			what colour are the X? what colour is X? what colour is the X?	what color X what color are * X what color are X what color is * X what color is X what color's * X what color's X what color's that X what color's this X
***what (ha)s _**	same as for "what's _" above			
***what about _**		what [about Y]?	what about X ? what about a X ? what about her X ? what about his X ? what about some X ? what about that X ? what about the X ? what about the other X ? what about these X ? what about this X ? what about your X ?	what about * X what about X what about that X what about the X what about the other X what about this X what about your X
***what shall _**			what shall we X?	what shall i X what shall we X

what can _			what can you X? what can you see on the X?	what can you X
what does _			what does X say? what does a X do? what does a X say? what does the X say?	what does * X what does X what does X do what does X say what does that X
what happened _		what [happened to Y]?	what happened to the X? what happened to your X?	what happened to the X
what were _	Not covered			what were * Z what were you Z
what’ve _		what [have you Y]?	what have you X?	what've you X what've you got X what have you X
what kind of _	Not covered			Not covered
what number _	Not covered			Not covered
***where’s _**	where’s Y? where’s the Y?		where is X ? where is the X ? where is your X ? where shall we put the X ? where was the X ? where's X ? where's X going ? where's X gone ? where's a X ? where's he X ? where's her X ? where's his X ?	where's * X where's X where's X going where's X gone where's childname's X where's he X where's her X where's his X where's my X where's the * X where's the X where's the X X where's the X then

			where's my X ? where's that X ? where's that X gone ? where's the X , <child’s name> ? where's the X ? where's the X going ? where's the X going to go? where's the X gone ? where's the X then ? where's the little X ? where's the other X ? where's your X ? where's your X gone ?	where's the little X where's the other X where's your X where's your X gone
where’s [has]	same as for “where’s _” above			
where’d _		where [did we Y]?	where did the X go?	where did the X where did you go X
where’re _		where [are you Y]?	where are you X? where are the X? where are your X?	where're you going X
where shall _			where shall we put the X?	where shall we put the X
who’s _	who’s Y?		who’s X? who's X it ? who's X the X ? who's in the X ? who's on the X ?	who's * X who's X who's gonna X who's that X

whose [N]	Not covered			whose X whose X is it whose X is that
who’re _		who [are you Y]?		Not covered
who did _		who [did you Y]?		who did you X
why don’t _			why don’t you X?	why don't you X why don't you X X
why do _		why [do you want your X]?		why do you X X why do you X
why’s _		why [is he Y]? why [is it Y]?	why is he X? why is it X?	why is he X why is it X
why not _	Not covered			Not covered
how many _			how many X? how many X are there? how many X have we got? how many X have you got?	how many X how many X are there how many X has he got how many X have you got
how did _		how [did you Y]?		Not covered
which one _		which [one Y]?	which one’s X?	which one X which one's * X which one's X

Not covered	which Y? who Y? why Y? how Y?		where was the X? which X? which X is it? who X a X? who X it? who X the X? why are you X?	where was the X which * X which X which is * X which is X who * X who X who X it who X you who did you X who else X who was X why X why are you X

Table 73. Comparison between Cameron-Faulkner et al. (2003)'s wh-questions and the current frame approach.

Yes/No-questions

Cameron-Faulkner et al. (2003) frame	**Nesting frame**	**Multiply-nested frame**	**Full-utterance frame**	**Prediction-based frame**
***are you _**	are you Y?		are you X , <child’s name> ? are you X ? are you X a X ? are you X again ? are you X her X ? are you X it ? are you X me ? are you X now ? are you X the X ? are you X them ? are you X to X ? are you X your X ? are you a X ? are you a X boy ? are you a bit X ? are you going X ? are you going to X ? are you going to X a X ? are you going to X her ? are you going to X it ? are you going to X it up? are you going to X that ? are you going to X the X?	are you * X are you X a X are you X are you X X are you X him are you X it are you X me are you X them are you doing X are you going X are you going to X are you going to X X are you going to X her are you going to X him are you going to X it are you going to X it * X are you going to X that are you going to X them are you going to be X are you going to come and X are you going to do some X are you going to have a X are you going to make a

			are you going to X them ? are you going to X your X? are you going to have a X? are you going to make a X? are you going to put the X on ? are you having a X ? are you making a X ? are you still X ? are you the X ?	X are you gonna X are you gonna put X are you just X are you making a X are you not X are you putting * X
are they _	are they Y?		are they X ? are they X now ? are they all X ? are they going to the X ? are they having a X ?	are they * X are they X X are they X are they all X are they going X are they going to the X
are we _	are we Y?		are we X?	are we X are we going X are we going to X
aren’t you _			aren’t you X?	aren't you X

***can you _**	can you Y?		can you X ? can you X a X ? can you X him ? can you X it ? can you X it X ? can you X that ? can you X the X ? can you X them ? can you find a X ? can you find another X ? can you find me the X ? can you find some X ? can you find the X ? can you say X ? can you see X ? can you see a X ? can you see some X ? can you see the X ?	can you X can you X X can you X it can you X that can you find X can you find a X can you find another X can you find me * X can you find the X can you pass me * X can you remember X can you remember what * X can you say X can you see * X can you see X X can you see X can you see a X X can you see a X can you see any X can you see some X can you see the X can you see the X X
can I have _			can i have X ? can i have X then ? can i have a X ? can i have a X please ? can i have some X ? can i have some X please ? can i have the X please ?	can i have * X can i have X can i have X X can i have a X can i have some X can i have some X please

can’t you _	can’t you Y?		can’t you X?	Not covered

***do you _**	do you Y?		do you X ?
			do you X the X ?
			do you know where the X is ?
			do you like X ?
			do you like mummy's X?
			do you like that X ?
			do you like the X ?
			do you need a X ?
			do you need some X ?
			do you think X like X ?
			do you think it's a X ?
			do you want a X ?
			do you want another X ?
			do you want me to X it ?
			do you want some X ?
			do you want the X ?
			do you want the X out ?
			do you want to X ?
			do you want to X it ?
			do you want to X on the X?
			do you want to X the X?
			do you want to X with the X ?
			do you want to do some X?
			do you want your X ?
			do you want your X out?
				do you X X
				do you X
				do you have X
				do you know X
				do you know what X
				do you know where * X
				do you like * X
				do you like X
				do you like X X
				do you like that X
				do you mean X
				do you need a X
				do you remember * X
				do you remember X
				do you think * X
				do you think X
				do you think X X
				do you think it's X
				do you think it's a X
				do you think she's X
				do you wanna X
				do you want * X
				do you want X
				do you want a X
				do you want another X
				do you want it X
				do you want me to X
				do you want me to X it
				do you want some X
				do you want that X

				do you want this X do you want to X do you want to X X do you want to X it do you want to do X do you want to do some X do you want to make a X do you want to play with the X do you want your X
don’t you _	don’t you Y?		don’t you like X?	don't you X don't you X X don't you like X don't you want * X
did you _	did you Y?		did you X ? did you X it ? did you X the X ? did you X your X ? did you like the X ? did you say X ?	did you X did you X X did you X it did you X that did you get X did you go X did you have X did you like * X did you like X did you say X
did we _	did we Y?			did we go to the X did we see any X
does it _			does it X?	does it X does it look like a X

***have you _**	have you Y?		have you X ? have you X a X ? have you X it ? have you X your X ? have you got X ? have you got a X ? have you got any X ? have you got some X ? have you got the X ? have you got your X ?	have you X have you X X have you X it have you X it X have you X them have you been X have you got * X have you got X have you got X X have you got a X have you got a X X have you got any X have you got some X have you got that X have you got your X have you seen X
has it _			has it X? has it got a X?	has it X has it got X has it got a X
***is it _**	is it Y?		is it X ? is it X now ? is it X or X ? is it X the X ? is it X yet ? is it a X ? is it a X one ? is it a big X ? is it a bit X ? is it an X ? is it in the X ? is it still X ?	is it * X is it X is it X X is it X X * X is it a X is it a X X is it a X one is it a big X is it going X is it going to X is it not X is it your X

			is it the X ? is it too X ? is it your X ?	
***is that _**	is that Y?		is that X ? is that X is that a X ? is that a nice X ? is that another X ? is that her X ? is that his X ? is that my X ? is that one X ? is that the X ? is that your X ?	is that * X is that X is that X X is that X there is that a X is that a X X is that one X is that what * X is that what X is that what X X is that your X
***is he _**	is he Y?		is he X ? is he X a X ? is he X his X ? is he X now ? is he X the X ? is he a X ? is he having a X ? is he in the X ?	is he * X is he X is he X X is he a X is he going X is he going to X
is this _	is this Y?		is this X ? is this a X ? is this the X ? is this your X ?	is this * X is this X is this X X is this a X is this your X
is she _	is she Y?		is she X? is she X now?	is she X is she going X

***shall I _**	shall I Y?		shall I X ? shall I X it ? shall I X the X ? shall I X you ? shall I X your X ?	shall i * X shall i X shall i X X shall i X it shall i X you shall i do X shall i draw * X shall i get the X X
***shall we _**	shall we Y?		shall we X again ? shall we X her X ? shall we X it ? shall we X some X ? shall we X the X ? shall we X this one ? shall we X with your X? shall we build a X ? shall we draw a X ? shall we find the X ? shall we get the X out ? shall we have a X ? shall we make a X ? shall we make some X ? shall we put the X on ? shall we take the X off ?	shall we X shall we X it shall we X the X shall we X this shall we do X shall we do a X shall we do some X shall we do the X shall we do the X X shall we find the X shall we get X shall we get the X shall we get the X out shall we have X shall we have a X shall we make a X shall we make some X shall we make the X shall we put X shall we put the X shall we put the X on shall we put this X shall we see if * X

shall mummy _	Not covered			Not covered
should we _	Not covered			Not covered
Not covered			aren’t they X ? do they eat X ? does X like X ? does <child’s name> like X ? does he like X ? does she like X ? does your X hurt? has she X her X? is there a X ? is there a X in there ? is there another X ? was it X? was it a X ? was that X ? were you X? would you like a X ? would you like some X ?	are there X are there any X are these X are those * X are those X aren't they * X aren't they X aren't we X can i X can i X X can i do X did X come did daddy X did he X did i X did she X did they X didn't X do they X do they eat X has he X has he got X has he got a X has she X has she got a X have they X have we got a X is there X is there a X

				is there a X X is there a X in there is there another X isn't it X was he X was it * X was it X was it X X was that X was there * X wasn't it X were they X would X would you X would you like X would you like a X would you like some X

Table 74. Comparison between Cameron-Faulkner et al. (2003)'s yes-no questions and the current frame approach.

Imperative constructions

Cameron-Faulkner et al. (2003) frame	Nesting frame	Multiply-nested frame	Full-utterance frame	Prediction-based frame
***come _**	come Y		come here X come on, X come on then, X	come * X come X come and X come here X come on X
***look _**	look Y		look X look at all these X look at all those X look at that X look at the X look at this X look at your X	look * X look X look at * X look at X look at all the X look at all these X look at all those X look at that X look at the X look at this X look at those X look what X
***let’s _**	let’s Y		let’s X it let’s X the X let’s X let’s have a look at your X	let's X let's X this let's get the X let's make * X let's put the X

***put _**	put Y		put her X on put it in the X then put it in the X put it on the X put some X on put the X away put the X in the X put the X in put the X next to the X put the X on the X put the X on put the X put them in the X put your X down put your X in put your X on put your X there	put * X put X put X on put it X put it in the X put it on X put it on the X put some X on put the X put the X X put the X away put the X back put the X in put the X in the X put the X in there put the X on put them in the X put this X put your X put your X away put your X down put your X in put your X on
***don’t _**	don’t Y		don't X her X don't X him don't X it , <child’s name> don't X it X don't X it don't X me don't X that	don't X don't X X don't X him don't X it don't X it X don't X that don't X them don't be X

			don't X the X don't X them don't X too X don't X your X don't X don't put your X in there don't you like X ?	don't get X
*go _	go Y		go and get your X	go X go and X go in the X
get _	get Y		get X get a X get the X out get your X	get * X get X get a X get her X get it X get the X get the X out get your X
let me _	Not covered			let me X
see _	see Y			see * X see X see X X see if * X see if you can X see the X
take _	take Y		take the X out take your X off	take * X take X take the X take the X X take the X out
turn _			turn the X over	turn * X

make _	make Y		make a X make it X make some X	make * X make X make a X make a X X make it X make some X make sure * X make this X
watch _	Not covered			watch * X watch X watch out X
leave him _	Not covered			Not covered
press _	Not covered			Not covered
have a look _		have a [look X] have a [look at this X]		Not covered
Not covered			do your X draw X draw a X find the X give him a X give it a good X give me the X have a X move the X move your X	(several frames start with unmarked forms of verbs, but not necessarily occurring in utterance-initial position)

Table 75. Comparison between Cameron-Faulkner et al. (2003)'s imperative constructions and the current frame approach.

Copula constructions

Cameron-Faulkner et al. (2003) frame	Nesting frame	Multiply-nested frame	Full-utterance frame	Prediction-based frame
***that's [NP/Adj]**	that's Y		that is X that is a X that's X, <child's name> that's X as well that's X for X that's X that's a X, <child's name> that's a X of X that's a X one that's a X that's a baby X that's a big X that's a bit X that's a funny X that's a good X that's a little X that's a mummy X that's a nice X that's a very good X that's an X that's another X that's for X that's for the X that's her X that's his X that's mummy's X	that's * X that's X that's X X that's a * X that's a X that's a X X that's a X one that's a big X that's a bit X that's a bit of a X that's a funny X that's a good X that's a little X that's a nice X that's a very good X that's all X that's all the X that's an X that's not * X that's not X that's not X X that's not a X that's the * X that's the X X that's the X that's the X X that's the X one

			that's my X that's not X that's not a X that's not a baby X that's not an X that's not the X that's not your X that's one X that's some X that's the X one that's the X that's the right X that's very X that's your X	that's the little X that's the right X that's very X that is * X that is X that is a X
***it’s [NP/Adj]**	it’s Y		it's X , <child’s name> it's X a X it's X again it's X in the X it's X now it's X on the X it's X the X it's X to X it's X it's a X , <child’s name> it's a X of X it's a X one it's a X yeah it's a X it's a baby X it's a big X it's a bit X	it's * X it's X X it's X X X it's X a X it's X it's a * X it's a X and X X it's a X it's a X X it's a X one it's a big X it's a bit X it's a bit of X it's a funny X it's a good X it's a little X it's a nice X

			it's a little X it's a nice X it's a red X it's all X it's an X it's another X it's for X it's her X it's his X it's in the X it's in your X it's like X it's like a X it's my X it's not X it's not a X it's not the X it's not your X it's on the X it's the X it's the other X it's too X it's very X it's your X	it's a red X it's all X it's an X it's just * X it's just X it's just a X it's like X it's like a X it's like a X X it's no good X it's not * X it's not X it's not X X it's not a X it's not a X X it's not very X it's only * X it's only X it's too X it's very X it is * X it is X it is X X it is a X it is a bit X
***there's [NP/Adj]**	there's Y		there's X there's a X here there's a X of X there's a X on the X there's a X there there's a X	there's * X there's X there's X X there's a * X there's a X there's a X X

			there's a little X there's an X there's another X there's her X there's his X there's lots of X there's more X there's my X there's no X there's one X there's only one X there's some X there's the X , look there's the X there's the baby X there's two X there's your X	there's a X in the X there's a little X there's an X there's another X there's his X there's lots of X there's more X there's no X there's not X there's one X there's only one X there's some X there's some X here there's some more X there's the * X there's the X there's the X X there's two X there's your X there is X there is a X
***he’s [NP/Adj]**	he’s Y		he's X he's a X he's in the X he's not X he's very X	he is X he’s X he's a big X he's a bit X he's just X he's not * X he's not X he's very X

***here’s [NP/Adj]**	here’s Y		here's X here's a X here's a little X here's another X here's some X here's the X , look here's the X here's your X	here's * X here's X here's another X here's some X here's your X
***this’s [NP/Adj]**	this is Y		this is X this is a X this is my X this is the X one this is the X this is your X	this is * X this is X this is X X this is a X this is a X X
that one’s [NP/Adj]		that [one’s X]	that one’s X that one’s X, is it? that one’s X, isn’t it? that one’s a X that one’s the X	that one is X that one's * X that one's X that one's a X that one's got X that one's got a X that one's not X
this one’s [NP/Adj]		this [one’s X]	this one’s X	this one X X this one's X
***they’re [NP/Adj]**	they’re Y		they are X	they are X

			they're X the X they're X they're a bit X they're all X they're in the X they're not X they're the X	they're * X they're X they're X X they're a bit X they're all * X they're all X they're called X they're going X they're just X they're not * X they're not X they're very X
***you're [NP/Adj]**	you're Y		you are X you're X you're a X you're all X you're going to X you're not X you're too X you're very X	you are X you're * X you're X you're a X you're a X X you're a bit X you're just X you're not X you're not X X you're not a X you're very X
it was [NP/Adj]	it was Y		it was X it was a X	it was X it was X X it was a X it was a X X it was a bit X it was the X
that was [NP/Adj]	that was Y		that was X	that was * X

			that was a big X that was a bit X that was my X	that was X that was X X that was a X that was a X X that was a big X that was a bit X that was a good X that was a nice X
your [N] is your [N] are	N/A		Not covered	your X is your X are
Not covered				he was X he was a bit X it isn’t X it isn’t a X it wasn’t X it wasn’t a X they were X

Table 76. Comparison between Cameron-Faulkner et al. (2003)'s copula constructions and the current frame approach.

Declarative (transitive, intransitive and complex) constructions				
Cameron-Faulkner et al. (2003) frame	Nesting frame	Multiply-nested frame	Full-utterance frame	Prediction-based frame
*you [VP NP]	you Y		see Table 78[9]	see Table 78
*you’ve [VP NP]	you’ve Y		you've X a X you've X it , have you ? you've X it ? you've X it now you've X it you've X my X you've X one you've X what ? you've X your X you've got X you've got a X ? you've got a X in your X you've got a X you've got lots of X you've got the X you've got your X on you've got your X	you've X X you've X you've X it you've X it now you've been X you've got * X you've got X you've got X X you've got X on you've got a X you've got a X X you've got it X you've got lots of X you've got no X you've got one X you've got some X you've got to X you've got two X you've got your X you've got your X on you've just X you've just X it you've not X

[9] Some CFLT frames have a great number of matches in the sets of full-utterance frames and prediction-based frames. For compactness, they are displayed in a more concise format in Table 78.

you’re [VP NP]	you’re Y		you're X get X you're X it you're X me you're X the X ? you're X your X	you're * X you're X you're X X you're doing X you're getting * X you're getting X you're going X you're going to * X you're going to X X you're going to X you're going to X it you're gonna X you're gonna X it you're gonna X me you're gonna get X you're having a X you're just X you're making X you're making a X you're not X you're not X X you're taking * X

***I [VP NP]**	I Y		I X it I X the X I can see X I can see a X I can see the X I can't X it I don't like X I don't want X I haven't got a X I haven't got any X I know it's X I know it's your X I know you like X I like X I said X I want a X I want my X	see Table 78
I’ll [VP NP]	I’ll Y		I’ll X it I’ll X that I’ll X the X	I will X I’ll X I’ll X it I’ll X you I’ll be X I’ll have * X I’ll have X I’ll have this X I’ll have to X I’ll just X

I’m [VP NP]	I’m Y		I’m X it I’m X you	I'm * X I'm X I'm X X I'm going to X I'm gonna X I'm just X I'm making * X I'm not * X I'm not X
I’ve [VP NP]	I’ve Y		I’ve X a X I’ve X it I’ve got X I’ve got a X I’ve got the X	I've X I've got * X I've got X I've got a X I've got to X I've got two X I've got your X I've just X
***we [VP NP]**	we Y		we X some X, didn’t we? we haven’t got any X	we * X we X X X we X the X we X we X X we X X * X we X it we are X we can X we can't X we could X we didn't X we don't X we don't want X

				we have X we haven't X we haven't got X; we haven't got a X we haven't got any X we need the X we saw X we went to X we went to the X
*it [VP NP]	it Y		Not covered	it * X it X X it X X X it X a X it X it X X * X it can't be X it does X it doesn't X it goes X it looks X it looks a bit like X it looks like X it looks like a X it looks very X it might X it might be X it must be X it will X it willn't X

it’s [VP NP]	it’s Y		it’s X a X it’s X the X it’s got X it’s got a X it’s got X on it’s got X in it	it's * X it's X X X it's X a X it's X it's X X it's going X it's going to X it's gonna X it's got X X it's got X it's got X on it's got a X it's just * X it's just X it's not * X it's not X it's not X X it's only * X it's only X
that [VP NP]	that Y		Not covered	that X that X X that X X X that X a X

he’s [VP NP]	he’s Y		he’s got X he’s got a X	he's * X he's X a X he's X he's X X he's been X he's going to X he's going to X X he's going to the X he's gonna X he's got * X he's got X he's got a X he's got a X X he's got a big X he's got big X he's got his X he's got no X he's just X he's not * X he's not X
she [VP NP]	she Y		she X a X she X her X	she * X she X she X X she can X she doesn't like X she likes X

they (ve) [VP NP]	they X		Not covered	they * X they X they X X they can X they don't X they eat X they have X they just X they look X they look like X
mummy [VP NP]	mummy Y		mummy X it mummy X the X	mummy * X mummy X mummy X X mummy X it mummy and daddy X mummy didn't X mummy do it X
mummy has [VP NP]	mummy’s Y		mummy’s got a X	mummy's * X mummy's X mummy's got * X mummy's got X mummy's got a X mummy's not X
Anne [VP NP]	<child’s name> Y		<child’s name> X <child’s name> X it	<child’s name> * X <child’s name> X <child’s name> X it <child’s name> likes X
***it [VP]**	it Y (as above)		it X	it * X

			it looks like a X	it X X X it X a X it X it X X it X X * X it does X it doesn't X it goes X it is * X it is X X it is X it is X X it isn't X it looks X it looks a bit like X it looks like X it looks like a X it looks very X it might X it might be X it must be X it was X X it was X it was X X it wasn't X it will X it willn't X
that one [VP]			that one X	that one * X

			that one is X that one's X , is it ? that one's X , isn't it ? that one's X that one's a X that one's the X	that one X that one is X that one's * X that one's X that one's not X
it’s [VP]	it’s Y (as above)		it's X , childname it's X , is it ? it's X , isn't it ? it's X ? it's X again it's X in the X it's X now it's X on the X it's X to X it's X it's going to X it's not X , is it ? it's not X	it's * X it's X X X it's X it's X X it's going X it's going to X it's gone X it's gonna X it's just * X it's just X it's not * X it's not X it's not X X it's only * X it's only X
***you [VP]**	you Y (as above)		you X with the X you X you can X you can't X you did X , didn't you ? you didn't X you do X you don't X you have to X you want to X ?	you * X you X X you X X X you X have X you X you X X * X you X to you can X you can X X you can't X

			you want to X with your X ?	you can't X X you did X you did X X you didn't X you do * X you do X X you do X you don't * X you don't X you don't X X you don't get X you go X you go and X you have to X you have to X X you haven't X you just X you should have X; you want to X
you’re [VP]	you’re Y (as above)		you are X , aren't you ? you are X you're X , are you ? you're X , aren't you ? you're X ? you're X on my X you're X on the X you're X you're going to X you're not X , are you ? you're not X ? you're not X	you're * X you're X you're X X you're a X you're a X X you're a bit X you're being X you're going X you're getting X you're going to X you're gonna X you're going to the X

				you're just X you're not X you're not going X you're sitting on X
you’ve [VP]	you’ve Y (as above)		Not covered	you've X X you've X you've been X you've got to X you've just X you've not X
***I [VP]**	I Y (as above)		I X I can’t X	I * X I X X I X X X I X I X X * X I can X I can't X X I can't X I didn't X I don't X X I don't X; I have X I haven't X I want to X
he [VP]	he Y		he X	he * X

			he can’t X	he X he X X he X in he can X he can't X he did X he didn't X he doesn't X he has X he hasn't X he went X he willn't X
he’s [VP]	he’s Y (as above)		he’s X a X he’s X his X he’s X it he’s X the X he’s X you he’s X your X he’s got X he’s got a X	he's * X he's X he's X X he's been X he's going X he's going to X he's going to X X he's gone X he's gonna X he's just X he's not * X he's not X
she’s [VP]	she’s Y (as above)		she’s X she’s X her X she’s going to X	she's * X she's X she's going X she's going to X she's not X
the phone [VP]	Not covered			Not covered
***there [pro] go**	Not covered			there you go X

***I think _**		I [think Y] I [think Y like] I [think it Y] I [think she Y] I [think you Y]	I think he's X I think it's X I think it's a X I think that's X I think that's a X I think they're X I think you've X it	I think * X I think X I think X X I think he's X I think it X I think it might be X I think it was X I think it's * X I think it's X I think it's X X I think it's a X I think she's X I think that X X I think that's X I think that's X X I think that's a X I think that's the X I think they're X I think this X I think we X I think you X I think you X X I think you're X
***I don’t think _**		I [don’t [think Y]]	I don’t think it’s X	I don't think * X I don't think X

				I don't think X X I don't think it X I don't think it's X I don't think it's a X I don't think that X I don't think there are any more X I don't think there is a X I don't think there's X I don't think they X
I thought _		I [thought you X]	I thought it was X I thought it was a X	I thought X I thought it was X I thought it was a X I thought you were X
think _	Not covered			think X think X X think it's X think it's a X
I don’t know _			I don’t know X I don’t know where X is I don’t know where the X is	I don't know X I don't know what * X I don't know where * X I don't know where X is
you know _	Not covered			you know * X you know X you know it's X

***if _**	Not covered			if * X
				if X
				if X X

				if I X
				if I X X
				if I X you
				if it's X
				if you X
				if you X X
				if you X it
				if you're X

***because _**	because Y		because he’s X because it’s X	because * X because X because X X because he's X because i X because i X X because i'm X because it was X X because it's * X because it's X because she X because she's X because that's X because they X because they're X because you X because you X X because you X it because you were X because you're X because you've X
when _	Not covered			when * X when X

		when X X when it's X when she X when they X when we go to the X when you X when you were X when you're X

Table 77. Comparison between Cameron-Faulkner et al. (2003)'s declarative constructions and the current frame approach.

Cameron-Faulkner et al. (2003) frame	**Lexically-specific frame matches**
You [VP NP]	**Full-utterance frame matches:** you X a X; you X her; you X him; you X his X; you X it ?; you X it X; you X it then; you X it; you X me; you X mummy; you X that X; you X that one; you X that; you X the X; you X them; you X your X; you X; you can X it; you can have X; you can't X it ?; you can't X it; you don't like X , do you ?; you don't like X ?; you don't like X; you don't want X; you find me the X; you find the X; you have a X; you hurt your X ?; you like X , do you ?; you like X , don't you ?; you like X ?; you like X; you like the X , don't you ?; you need a X ?; you need a X; you put the X in; you say X; you want X ?; you want X; you want a X , do you ?; you want a X ?; you want a X; you want me to X it ?; you want some X , do you ?; you want some X ?; you want some X; you want the X ?; you want the X; you want to X ?; you want to X it ?; you want to X with your X ?; you want to have a X ?; you want to make a X ?; you want your X ?
You [VP NP]	**Prediction-based frame matches:** you * X; you * a X; you X X; you X X X; you X a X X; you X have X; you X it * X; you X it X; you X me * X; you X that X X; you X; you X X * X; you X him; you X it; you X it off; you X it out; you X it then; you X me; you X some; you X that; you X that X; you X that one; you X them; you X them out; you X to; you X what; you X your X; you are X; you can X; you can X X; you can X it; you can do it X; you can have * X; you can have X; you can have it X; you can see X; you can't X; you can't X X; you can't X it; you can't eat X; you can't have X; you could have X; you did X; you did X X; you didn't X; you do * X; you do X X; you do X; you do it X; you don't * X; you don't X; you don't X X; you don't X it; you don't get X; you don't like * X; you don't like X X; you don't like X; you don't need * X; you don't need X; you don't want X; you don't want a X; you find me * X; you find the X; you get X; you get the X out; you go X; you go and X; you got * X; you got X; you got a X; you got it X; you had X; you had a X; you had a X X; you have * X; you have X X; you have X; you have a X; you have to X; you have to X X; you have to X it; you have to X it X; you haven't X; you haven't got X; you haven't got a X; you haven't got any X; you just X; you like * X; you like X; you like X X; you like that X; you mean X; you need * X; you need X; you need a X; you put the X; you put the X in; you say X; you see X X; you see X; you should have X; you want * X; you want X; you want a X; you want a X X; you want it X; you want me to X; you want me to X it; you want some X; you want that X; you want to X; you want to X it; you want your X

I [VP NP]	**Prediction-based frame matches:** I * X; I X X; I X X X; I X; I X X * X; I X it; I X it X; I X you; I X you * X; I can X; I can have * X; I can have X; I can see * X; I can see X; I can see the X; I can't X X; I can't X; I can't X it; I can't do it X; I didn't X; I didn't X that; I don't X X; I don't X; I don't like X; I don't want * X; I don't want X; I don't want any X; I don't want to X; I have X; I haven't X; I haven't got a X; I haven't got any X; I know X X; I know X; I know it's * X; I know it's X; I know it's a X; I know what X; I know you like X; I said X; I want X; I want a X; I want my X; I want to X

Table 78. Selected declarative constructions from Cameron-Faulkner et al. (2003) and their lexically-specific frame matches.

References

Adriaans, P. (1992). *Language Learning from a Categorial Perspective.* Unpublished PhD thesis, University of Amsterdam.

Adriaans, P. (1999). *Learning Shallow Context-Free Languages under Simple Distributions* (Technical Report No. PP-1999-13): Institute for Logic, Language and Computation, University of Amsterdam.

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*(6), 952-965.

Baker, M. C. (2003). *Lexical categories: verbs, nouns, and adjectives*: Cambridge University Press.

Baldwin, D. (1993). Infants' ability to consult the speaker for clues to word reference. *Journal of Child Language, 20*(2), 395-418.

Bard, E. G., Shillcock, R. C., & Altmann, G. T. M. (1988). The recognition of words after their acoustic offsets in spontaneous speech: effects of subsequent context. *Perception and Psychophysics, 44*, 395-408.

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

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Sciences, 2*(2), 67-73.

Bonnema, R., Bod, R., & Scha, R. (1997). *A DOP model for semantic interpretation*. Paper presented at the 35th Annual Meeting of the Association for Computational Linguistics.

Bowerman, M. (1973). *Early syntactic development: a cross-linguistic study with special reference to Finnish*. Cambridge: Cambridge University Press.

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

Braine, M. (1987). What is learned in acquiring word classes - a step toward an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65-87). Hillsdale: Erlbaum.

Brent, M. R., & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition, 81*, B33-B44.

Brown, R. (1957). Linguistic determinism and the part of speech. *Journal of Abnormal and Social Psychology, 55*(1), 1-5.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Bybee, J. L. (1985). *Morphology: a study of the relation between meaning and form*: John Benjamins.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*(5), 425-455.

Cameron-Faulkner, T., Lieven, E., & Tomasello, M. (2003). A construction-based analysis of child directed speech. *Cognitive Science, 27*, 843-873.

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

Casenhiser, D. M., & Goldberg, A. E. (2005). Fast mapping between a phrasal form and meaning. *Developmental Science, 8*(6), 500-508.

Chapman, G. B., & Robbins, S. J. (1990). Cue interaction in human contingency judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17*, 837-854.

Chater, N., & Vitányi , P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences, 7*(1), 19-22.

Childers, J. B., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the English transitive construction. *Developmental Psychology, 37*(6), 739-748.

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

Chomsky, A. N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, A. N. (1970). Remarks on nominalization. In R. Jacobs & P. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184-221). Waltham, MA: Ginn.

Chomsky, A. N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

Church, K. W. (1992). *Parts of speech tagging.* Paper presented at the 5th Annual CUNY Conference on Human Sentence Processing.

Clark, A. (2000). *Inducing syntactic categories by context distribution clustering*. Paper presented at the CoNLL-2000.

Clark, A. (2001). *Unsupervised language acquisition: theory and practice.* Unpublished PhD thesis, University of Sussex.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language : an introduction to psycholinguistics*. New York: Harcourt Brace Jovanovich.

Conwell, E., & Morgan, J. L. (submitted). When parents verb nouns: Resolving the ambicategoricality problem. *Language Learning and Development*.

Croft, W. (2001). *Radical construction grammar : syntactic theory in typological perspective*. Oxford: Oxford University Press.

Croft, W. (2005). Logical and typological arguments for Radical Construction Grammar. In M. Fried & J.-O. Östman (Eds.), *Construction Grammar(s): cognitive and cross-language dimensions*. Amsterdam: John Benjamins.

Edgington, E. (1995). *Randomization tests.* New York:Marcel Dekker.

Eilers, R. (1975). Suprasegmental and grammatical control over telegraphic speech in young children. *Journal of Psycholinguistic Research, 4*, 227-239.

Erkelens, M. A. (2008). *Restrictions of frequent frames as cues to categories: the case of Dutch.* Poster presented at the 32nd Boston University Conference on Language Development, Boston, MA.

Feigenbaum, E., & Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8*, 305-336.

Feldman, J., Lakoff, G., Stolcke, A., & Weber, S. (1990). *Miniature language acquisition: a touchstone for cognitive science*. Paper presented at the 12th Annual Conference of the Cognitive Science Society.

Finch, S. (1993). *Finding structure in language.* Unpublished PhD thesis, University of Edinburgh.

Finch, S., Chater, N., & Redington, M. (1995). Acquiring syntactic information from distributional statistics. In J. P. Levy, D. Bairaktaris, J. A. Bullinaria & P. Cairns (Eds.), *Connectionist models of memory and language* (pp. 229-242). London: UCL Press.

Freudenthal, D., Pine, J. M., & Gobet, F. (2002). *Modelling the development of Dutch optional infinitives in MOSAIC*. Paper presented at the 24th Meeting of the Cognitive Science Society.

Freudenthal, D., Pine, J. M., & Gobet, F. (2005). Resolving ambiguity in the extraction of syntactic categories through chunking. *Cognitive Systems Research, 6*, 17-25.

Fries, C. C. (1952). *The structure of English: an introduction to the construction of English sentences*. London: Longmans.

Frigo, L., & McDonald, J. L. (1998). Properties of phonological markers that affect the acquisition of gender-like subclasses. *Journal of Memory and Language, 39*, 218-245.

Geertzen, J., & Van Zaanen, M. M. (2004). Grammatical inference using suffix trees. In *Lecture Notes in Computer Science* (Vol. 3264, pp. 163-174). Berlin: Springer-Verlag.

Gelman, S. A., & Taylor, M. (1988). How two-year-old children interpret proper and common names for unfamiliar objects. *Child Development, 55*, 1535-1540.

Gerken, L. (1987). *Function morphemes in young children's speech perception and production.* Unpublished PhD thesis, Columbia University.

Gerken, L. (1991). The metrical basis for children's subjectless sentences. *Journal of Memory and Language, 30*, 431-451.

Gerken, L. (1994). A metrical template account of children's weak syllable omissions from multisyllabic words. *Journal of Child Language, 21*, 565-584.

Gerken, L., Landau, B., & Remez, R. E. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology, 26*(2), 204-216.

Gerken, L., & McIntosh, B. J. (1993). Interplay of function morphemes and prosody in early language. *Developmental Psychology, 29*(3), 448-457.

Gerken, L., Wilson, R., & Lewis, W. (2005). Infants can use distributional cues to form syntactic categories. *Journal of Child Language, 32*, 249-268.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition, 1*(1), 3-55.

Gleitman, L., & Wanner, E. (1982). The state of the state of the art. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: the state of the art* (pp. 3-48). Cambridge, MA: MIT Press.

Gobet, F., Freudenthal, D., & Pine, J. M. (2004). *Modelling syntactic development in a cross-linguistic context.* Paper presented at the 20th International Conference on Computational Linguistics (COLING 2004), Workshop on "Psycho-Computational Models of Human Language Acquisition", Geneva, Switzerland.

Gobet, F., & Pine, J. M. (1997). *Modelling the acquisition of syntactic categories.* Paper presented at the 19th Annual Meeting of the Cognitive Science Society.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar approach to argument structure*. Chicago: University of Chicago Press.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *Trends in Cognitive Sciences, 7*(5), 219-224.

Goldberg, A. E., & Casenhiser, D. M. (2006). English Constructions. In B. Aarts & A. McMahon (Eds.), *Blackwell handbook of English linguistics*. London: Blackwell.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics, 14*, 289-316.

Goldberg, A. E., Casenhiser, D. M., & Sethuraman, N. (2005). The role of prediction in construction learning. *Journal of Child Language, 32*(2), 407-426.

Goldinger, S. D. (1996). Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition, 22*(5), 1166-1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

Gómez, R. (2002). Variability and detection of invariant structure. *Psychological Science, 13*(5), 431-436.

Gómez, R., & Lakusta, L. (2004). A first step in form-based category abstraction by 12-month-old infants. *Developmental Science, 7*(5), 567-580.

Gómez, R., & Maye, J. (2005). The developmental trajectory of nonadjacent dependency learning. *Infancy, 7*(2), 183-206.

Grimshaw, J. (1981). Form, function and the language acquisition device. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition* (pp. 165-182). Cambridge, MA: MIT Press.

Grünwald, P. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff & G. Scheler (Eds.), *Connectionist, statistical and symbolic approaches to learning for natural language processing*. Berlin: Springer-Verlag.

Hall, D. G., Waxman, S. R., & Hurwitz, W. M. (1993). How two- and four-year-old children interpret adjectives and count nouns. *Child Development, 64*, 1651-1664.

Harris, Z. S. (1954). Distributional structure. *Word, 10*(23), 146-162.

Hayek, L. C. (1994). Analysis of amphibian biodiversity data. In W. R. Heyer, M. A. Donnelly, R. W. McDiarmid, L. C. Hayek & M. S. Foster (Eds.), *Measuring and monitoring biological diversity: standard methods for amphibians*: Smithsonian Institute Press.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93*(4), 411-428.

Höhle, B., Schmitz, M., Santelmann, L. M., & Weissenborn, J. (2006). The recognition of discontinuous verbal dependencies by German 19-month-olds: evidence for lexical and structural influences on children's early processing capacities. *Language Learning and Development, 2*(4), 277-300.

Höhle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science, 6*(2), 122-127.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2002). *The assignment of word forms to syntactic categories in early language acquisition.* Paper presented at the 13th Biennial International Conference on Infant Studies.

Höhle, B., Weissenborn, J., Kiefer, D., Schulz, A., & Schmitz, M. (2004). Functional elements in infants' speech processing: the role of determiners in the syntactic categorization of lexical elements. *Infancy, 5*(3), 341-353.

Hume, D. (1984). *Creating Interactive Worlds with Multiple Actors.* Unpublished BSc Honours, Electrical Engineering and Computer Science, University of New South Wales.

Jones, S. S., & Smith, L. (1998). How children name objects with shoes. *Cognitive Development, 13*, 323-334.

Jones, S. S., Smith, L., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62*(499-516).

Jusczyk, P. W., & Aslin, R. (1995). Infants' detection of sound patterns of words in fluent speech. *Cognitive Psychology, 29*, 1-23.

Kako, E., & Wagner, L. (2001). The semantics of syntactic structures. *Trends in Cognitive Sciences, 5*(3), 102-108.

Kay, P., & Fillmore, C. J. (1999). Grammatical constructions and linguistic generalizations: the *What's X doing Y?* construction. *Language, 75*, 1-33.

Kedar, Y., Casasola, M., & Lust, B. (2004). *24-month-olds' sensitivity to the syntactic role of function words in English sentences: noun phrase determiners*. Paper presented at the 32nd Stanford Child Language Research Forum.

Kelly, M. H. (1992). Using sound to solve syntactic problems: the role of phonology in grammatical category assignments. *Psychological Review, 99*, 349-364.

Kohonen, T. (1995). *Self-organizing maps*. Berlin: Springer.

Kruschke, J. K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22-44.

Kucera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence: Brown University Press.

Laakso, A., & Smith, L. (2004). *On a possible role for pronouns in the acquisition of verbs.* Paper presented at the 20th International Conference on Computational Linguistics (COLING 2004), Workshop on ”Psycho-Computational Models of Human Language Acquisition”.

Labelle, M. (2005). The acquisition of grammatical categories: A state of the art. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 433-457). Amsterdam: Elsevier.

Lakoff, G. (1987). *Women, fire and dangerous things: what categories reveal about the mind*. Chicago: Chicago University Press.

Landau, B., Jones, S. S., & Smith, L. (1992). Syntactic context and the shape bias in children's and adults' lexical learning. *Journal of Memory and Language, 31*, 807-825.

Landau, B., Smith, L., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299-321.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar, Vol. 1: Theoretical prerequisites*. Stanford, CA: Stanford University Press.

Langley, P., & Stromsten, S. (2000). *Learning context-free grammars with a simplicity bias.* Paper presented at the 11th European Conference on Machine Learning.

Lederer, A., Gleitman, H., & Gleitman, L. (1995). Verbs of a feather flock together: semantic information in the structure of maternal speech. In M. Tomasello & W. E. Merriman (Eds.), *Beyond names for things: young children's acquisition of verbs*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Leibbrandt, R. E., & Powers, D. M. W. (2007). *Learning lexical categories using lexically-based templates.* Paper presented at the Australian Society for Cognitive Science Conference, Adelaide, Australia.

Leibbrandt, R. E., & Powers, D. M. W. (2008). *Grammatical category induction using lexically-based templates*. Poster presented at the 32nd Boston University Conference on Language Development (BUCLD 32).

Li, M., & Vitányi , P. (1997). *An introduction to Kolmogorov complexity and its applications*. New York: Springer-Verlag.

Li Santi, M. (2007). *Acquiring spatial semantics using artificial 3D worlds.* Unpublished Masters' thesis, Flinders University.

Li Santi, M., Leibbrandt, R. E., & Powers, D. M. W. (2007a). *Desiderata and Trade-offs in HxI for Immersive Language Learning,.* Paper presented at the Joint HCSNet-HxI Workshop on Human Issues in Interaction and Interactive Interfaces, Sydney, Australia.

Li Santi, M., Leibbrandt, R. E., & Powers, D. M. W. (2007b). *Developing 3D Worlds for Language Learning.* Paper presented at the Australian Society for Cognitive Science Conference, Adelaide, Australia.

Lieven, E., Pine, J. M., & Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language, 24*, 187-219.

Lieven, E., Pine, J. M., & Dresner Barnes, H. (1992). Individual differences in early vocabulary development: Redefining the referential-expressive distinction. *Journal of Child Language, 19*(287-310).

Lieven, E. V. M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language, 30*, 333-370.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95*(4), 492-527.

Logan, G. D., & Etherton, J. L. (1994). What is learned during automatization? The role of attention in constructing an instance. *Journal of Experimental Psychology: Learning, Memory and Cognition, 20*(5), 1022-1050.

Macnamara, J. (1982). *Names for things: a study of child language*. Cambridge, MA: MIT Press.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk.* (3 ed. Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics, 1*(1), 24-45.

Mandel, D. R., Jusczyk, P. W., & Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science, 6*, 314-317.

Manly, B. F. J. (1997). *Randomization, bootstrap and Monte Carlo methods in biology*. London: Chapman & Hall.

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* (Vol. 2). New York: Gardner Press.

Marcus, M., Santorini, B., & Marcinkiewicz, M. (1993). Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics, 19*(2), 313-330.

Matthews, D., Lieven, E., Theakston, A. L., & Tomasello, M. (2004). *The role of frequency and distributional regularity in the acquisition of word order.* Paper presented at the 32nd Stanford Child Language Research Forum.

Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Mintz, T. H. (2000). *Unique entropy as a model of linguistic classification.* Paper presented at the Twenty-Second Annual Cognitive Science Society Conference.

Mintz, T. H. (2002). Category induction from distributional cues in an artificial language. *Memory and Cognition, 30*(5), 678-686.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition, 90*(1), 91-117.

Mintz, T. H. (2005). Linguistic and conceptual influences on adjective acquisition in 24- and 36-month-olds. *Developmental Psychology, 41*(1), 17-29.

Mintz, T. H. (2006a). Finding the verbs: Distributional cues to categories available to young learners. In K. Hirsh-Pasek & R. M. Golinkoff (Eds.), *Action meets word: How children learn verbs*. Oxford: Oxford University Press.

Mintz, T. H. (2006b). Frequent frames: Simple co-occurrence constructions and their links to linguistic structure. In E. V. Clark & B. F. Kelly (Eds.), *Constructions in acquisition*. Stanford: CSLI Publications.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Morgan, J. L., & Newport, E. L. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior, 20*, 67-85.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (pp. 263-283). Hillsdale, NJ: Lawrence Erlbaum Associates.

Naigles, L. G., & Kako, E. (1993). First contact in verb acquisition: defining a role for syntax. *Child Development, 64*, 1665-1687.

Nelson, K. (1995). The dual category problem in the acquisition of action words. In M. Tomasello & W. E. Merriman (Eds.), *Beyond Names for Things: Young Children's Acquisition of Verbs* (pp. 223-249). Hillsdale, NJ: Lawrence Erlbaum Associates.

Ninio, A. (1999). Pathbreaking verbs in syntactic development and the question of prototypical transitivity. *Journal of Child Language, 26*, 619-653.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development, 8*, 245-272.

Ouhalla, J. (1999). *Introducing transformational grammar: from principles and parameters to minimalism*. London: Arnold.

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology: Learning, Memory and Cognition, 34*(1), 80-96.

Perruchet, P., & Pacton, S. (2006). Implicit learning and statistical learning: One phenomenon, two approaches. *Trends in Cognitive Sciences, 10*(5), 233-238.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics, 17*, 97-119.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of Memory and Language, 39*, 246-263.

Peters, A. M. (1977). Language learning strategies: does the whole equal the sum of the parts? *Language, 53*(3), 560-573.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge: Cambridge University Press.

Petretic, P. A., & Tweney, R. D. (1977). Does comprehension precede production? The development of children's responses to telegraphic sentences of varying grammatical adequacy. *Journal of Child Language, 4*, 201-209.

Pfitzner, D. M., Leibbrandt, R. E., & Powers, D. M. W. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*, 361-394.

Pine, J. M., & Lieven, E. V. M. (1993). Reanalysing rote-learned phrases: Individual differences in the transition to multi-word speech. *Journal of Child Language, 20*, 551-571.

Pinker, S. (1979). Formal models of language learning. *Cognition, 7*, 217-283.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum.

Powers, D. M. W. (1983). Neurolinguistics and psycholinguistics as a basis for computer acquisition of natural language. *SIGART, 84*, 29-34.

Powers, D. M. W. (1991). *How far can self-organization go? Results in unsupervised language learning.* Paper presented at the AAAI Spring Symposium on Machine Learning of Natural Language and Ontology, Stanford.

Powers, D. M. W. (1997). *Learning and application of Differential Grammars.* Paper presented at the Conference on Natural Language Learning (CONLL97) Workshop on Computational Natural Language Learning, Madrid, Spain.

Powers, D. M. W. (1997). Unsupervised learning of linguistic structure: an empirical evaluation. *International Journal of Corpus Linguistics 2*(1), 91-131.

Powers, D. M. W. (2003). *Recall and precision versus the Bookmaker*. Paper presented at the 4th International Conference on Cognitive Science (ICCS).

Powers, D. M. W. (2008). *Evaluation evaluation.* Paper presented at the 18th European Conference on Artificial Intelligence (ECAI 08), Patras, Greece.

Powers, D. M. W., Leibbrandt, R. E., Li Santi, M., & Luerssen, M. H. (2007). *A multimodal environment for immersive language learning - space, time, viewpoint and physics.* Paper presented at the Joint HCSNet-HxI Workshop on Human Issues in Interaction and Interactive Interfaces, Sydney, Australia.

Powers, D. M. W., & Turk, C. (1989). *Machine Learning of Natural Language*. Berlin: Springer-Verlag.

Pullum, G. (1994). Categories, linguistic. In E. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (pp. 478-482). Oxford: Pergamon Press.

Pulvermüller , F. (1996). Hebb's concept of cell assemblies and the psychophysiology of word processing. *Psychophysiology 33*, 317-333.

Pulvermüller , F. (1999). Words in the brain's language. *Behavioral and Brain Sciences, 22*, 253-336.

Radford, A. (1990). *Syntactic theory and the acquisition of English syntax: the nature of early child grammars of English*. Oxford: Blackwell.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Redington, M., Chater, N., Huang, C.-R., Chang, L.-P., Finch, S., & Chen, K.-J. (1995). *The universality of simple distributional methods: identifying syntactic categories in Mandarin Chinese.* Paper presented at the 4th International Conference on Cognitive Science and Natural Language Processing.

Rosch, E. (1983). Prototype classification and logical classification: The two systems. In E. Scholnick (Ed.), *New Trends in Cognitive Representation: Challenges to Piaget's Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Roy, D. (2002). Learning words and syntax for a visual description task. *Computer Speech and Language, 16*(3), 353-385.

Saffran, J., Newport, E. L., & Aslin, R. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language, 35*, 606-621.

Samuelson, L. K., & Smith, L. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond? *Cognition, 73*, 1-33.

Santelmann, L. M., & Jusczyk, P. W. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition, 69*, 105-134.

Schachter, P., & Shopen, T. (2007). Parts-of-speech systems. In T. Shopen (Ed.), *Language typology and syntactic description, Volume 1: clause structure*. Cambridge: Cambridge University Press.

Schütze, H. (1995). *Distributional part-of-speech tagging.* Paper presented at the 7th Conference of the European Chapter of the Association for Computational Linguistics.

Sethuraman, N., & Goodman, J. C. (2004). *Children's mastery of the transitive construction.* Paper presented at the 32nd Stanford Child Language Research Forum.

Shady, M. E. (1996). *Infants' sensitivity to function morphemes.* PhD thesis, State University of New York at Buffalo.

Shanks, D. R. (1995). Is human learning rational? *Quarterly Journal of Experimental Psychology A, 48*, 257-279.

Shi, R. (2007). *Infants' recognition of function words in continuous speech.* Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany.

Shi, R., Cutler, A., Werker, J. F., & Cruickshank, M. (2006). Frequency and form as determinants of functor sensitivity in English-acquiring infants. *Journal of the Acoustic Society of America, 119*(6), EL61-67.

Shi, R., Marquis, A., & Gauthier, B. (2006). *Segmentation and representation of function words in preverbal French-learning infants.* Paper presented at the 31st Boston University Conference on Language Development, Boston, MA.

Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy, 10*(187-198).

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*, B11-B21.

Shipley, E., Smith, C., & Gleitman, L. (1969). A study in the acquisition of language. *Language, 45*, 322-342.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition, 61*, 39-91.

Smith, K. H. (1966). Grammatical intrusions in the recall of structured letter pairs: mediated transfer or position learning? *Journal of Experimental Psychology, 72*, 580-588.

Smith, L. (2001). How domain-general processes may create domain-specific biases. In M. Bowerman & S. C. Levinson (Eds.), *Language acquisition and conceptual development* (pp. 101-131). Cambridge: Cambridge University Press.

Smith, L., Jones, S. S., & Landau, B. (1992). Count nouns, adjectives and perceptual properties in children's novel word interpretations. *Developmental Psychology, 28*(273-286).

Soja, N. N., Carey, S., & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meanings: object terms and substance terms. *Cognition, 38*, 179-211.

Sokal, R. R., & Sneath, P. H. A. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman.

Solan, Z. (2006). *Unsupervised learning of natural languages.* Unpublished PhD thesis, Tel Aviv University, Tel Aviv.

Stolcke, A. (1994). *Bayesian learning of probabilistic language models.* Unpublished PhD thesis, University of California, Berkeley.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning : an introduction* Cambridge, MA: MIT Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. Boston: Pearson.

Taylor, M., & Gelman, S. A. (1988). Adjectives and nouns: Children's strategies for learning new words. *Child Development, 59*, 411-419.

Theakston, A. L. (2004). The role of entrenchment in children's and adults' performance on grammaticality judgment tasks. *Cognitive Development, 19*, 15-34.

Theakston, A. L., Lieven, E., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language, 28*, 127-152.

Tincoff, R., & Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological Science, 10*, 172-175.

Tomasello, M. (1992). *First verbs: A case study in early grammatical development*. Cambridge, UK: Cambridge University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition, 74*, 209-253.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. (2006). Acquiring linguistic constructions. In D. Kuhn & R. Siegler (Eds.), *Handbook of Child Psychology*. New York: Wiley.

Treisman, A. M., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology, 12*, 97-136.

Van der Auwera, J. (1994). Adverbs and adverbials. In E. E. Asher & J. M. Y. Simpson (Eds.), *The encyclopedia of language and linguistics* (Vol. 1, pp. 39-43). Oxford: Pergamon Press.

Van Zaanen, M. M. (2001). *Bootstrapping structure into language: Alignment-based learning.* PhD thesis, University of Leeds.

Van Zaanen, M. M., & Adriaans, P. (2001). *Alignment-based learning versus EMILE: a comparison.* Paper presented at the Belgian-Dutch Conference on Artificial Intelligence (BNAIC).

Vervoort, M. R. (2000). *Games, walks and grammars: problems I've worked on.* Unpublished PhD thesis, University of Amsterdam.

Waxman, S. R. (2002). Early word-learning and conceptual development: Everything had a name, and each name gave birth to a new thought. In U. Goswami (Ed.), *Blackwell handbook of childhood cognitive development*: Blackwell.

Waxman, S. R., & Markow, D. B. (1998). Object properties and object-kind: twenty-one-month-olds' extension of novel adjectives. *Child Development, 69*, 1313-1329.

Wickelgren, W. A. (1979). Chunking and consolidation: a theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review, 86*, 44-60.

Widrow, B., & Hoff, M. E. (1960). *Adaptive switching circuits.* Paper presented at the Institute of Radio Engineers, Western Electric Show and Convention.

Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication, 2*(1), 57-89.

Wolff, J. G. (2001). Information compression and multiple alignment as unifying concepts in AI and computing. *Expert Update (Bulletin of British Computer Society Specialist Group on Knowledge-based Systems and Applied Artificial Intelligence), 4*(3), 22-36.

Wolff, J. G. (2002). *Mathematics and logic as information compression by multiple alignment, unification and search*: Technical report, cognitionresearch.org.uk.

Wolff, J. G. (2002). *Neural mechanisms for information compression by multiple alignment, unification and search*: Technical report, cognitionresearch.org.uk.

Wolff, J. G. (2002). *Unsupervised learning in a framework of information compression by multiple alignment, unification and search*: Technical report, cognitionresearch.org.uk.

Yoshida, H., & Smith, L. (2005). Linguistic cues enhance the learning of perceptual cues. *Psychological Science, 16*(2), 90-95.

Yuret, D. (1998). *Discovery of linguistic relations using lexical attraction.* Unpublished PhD thesis, Massachusetts Institute of Technology.

Zipf, G. K. (1949). *Human behavior and the Principle of Least-Effort*: Addison-Wesley.