LANGUAGE AND COGNITIVE PROCESSES, 1998, *13* (2/3), 129–191

Connectionist and Statistical Approaches to Language Acquisition: A Distributional Perspective

Martin Redington and Nick Chater

Department of Experimental Psychology, University of Oxford, UK

We propose that one important role for connectionist research in language acquisition is analysing what linguistic information is present in the child's input. Recent connectionist and statistical work analysing the properties of real language corpora suggest that a priori objections against the utility of *distributional* information for the child are misguided. We illustrate our argument with examples of connectionist and statistical corpus-based research on phonology, segmentation, morphology, word classes, phrase structure, and lexical semantics. We discuss how this research relates to other empirical and theoretical approaches to the study of language acquisition.

INTRODUCTION

The acquisition of natural language can be viewed as the result of a complex interaction between two sources of information.

1. The innate knowledge, both language-specific and general, possessed by the infant.
2. The infant's environment, both linguistic, and extra-linguistic.

Traditionally, linguists have emphasised the role of innate knowledge in language, with the influence of the child's environment playing a relatively

Requests for reprints should be addressed to M. Redington, Dept. of Psychology, University College London, 26 Bedford Way, London WC1E 6BT, UK.

Emails: m.redington@ucl.ac.uk and nick.chater@warwick.ac.uk

This research was supported by UK Economic and Social Research Council (ESRC) Research Studentship R00429234268, Research Grant SPG9029590 from the Joint Councils Initiative in Cognitive Science/Human Computer Interaction, and ESRC Research Grant R000236214. Some of this work was performed while the authors were members of the Psychology Department and the Centre for Cognitive Science, University of Edinburgh, and visitors at the Center for Research in Language, University of California, San Diego. Thanks to Steve Finch for helping develop the ideas discussed in this article, and to Kim Plunkett, Ulrike Hahn, and three anonymous reviewers for their useful comments and suggestions on an earlier version of the article.

minor role. In contrast, psychologists studying language development have to explain how the interaction of innate knowledge and the child's environment account for the developmental progression of language ability.

No matter how great the contribution of innate knowledge to language acquisition, some aspects of language (e.g. vocabulary) must be learnt. Even strongly nativist accounts of language require the setting of parameters (specifying, for example, the particular phonology of the language, or particular grammatical constructions permitted), which can only be determined through exposure to the language. Therefore, within the developmental literature, there is a strong emphasis on learning. However, discussion of learning and learnability within the developmental literature is generally naive with respect to the capabilities of formal learning methods (for example, Pinker's, 1984 critique of distributional approaches, which we discuss later). Similarly, empirical demonstrations of the utility of particular sources of information, for particular aspects of language, are generally absent from the developmental literature. We propose that disciplines such as *machine learning* and *statistics* (which, as we shall see, are closely related to connectionist approaches), which are specifically concerned with learning, can usefully inform developmental research.

We argue that the connectionist and statistical approaches to learning are of great relevance to the study of language acquisition, for three reasons:

1. They provide principled conceptions of learning and learnability.
2. They provide potential learning mechanisms for particular aspects of language. Applying such mechanisms to corpora of real language allows empirical measurement of the utility of potential sources of information, for particular aspects of language. Additionally, the results of such analyses can provide empirical predictions about the time course and profile of acquisition, and suggest new avenues for experimental work.
3. These approaches allow inferences concerning the nature and extent of innate knowledge, either in terms of innate learning mechanisms (as embodied by such models), or innate knowledge per se (in terms of what knowledge may be required to supplement the information that such learning mechanisms can provide).

Regarding potential sources of information, we shall generally be concerned with language-internal, or *distributional* information, derived from the relationships between linguistic units such as phonemes, morphemes, or even words. Such information can be readily extracted by a range of learning mechanisms, including connectionist networks and statistical models, which we shall collectively term *distributional learning mechanisms*. Our primary claim in this article is that distributional

information may be highly useful in the acquisition of many aspects of language.

Distributional information contrasts with extra-linguistic sources of information that infants might exploit, including features of the physical environment, the meaning of an utterance, or its communicative value. Such information undoubtedly plays a major role in the acquisition of language. Although the kinds of learning mechanism we shall discuss can be applied to exploiting this kind of information (e.g. Mareschal, Plunkett, & Harris, 1995; Plunkett, Sinha, Møller, & Strandsby, 1992) we do not discuss this here.

Since advocacy of distributional methods is sometimes associated with a number of logically distinct views, it is worth making explicit what we are *not* arguing for. First, we do not propose that distributional information is relevant to every aspect of language acquisition. Second, where distributional information is relevant, we do not argue that it is the only, or even the major, source of information available to the child. Third, the view that distributional information is important in language acquisition is compatible with the innateness of both domain-specific language learning mechanisms and knowledge of many universal properties of language (see e.g. Kirsh, 1991; Plunkett, 1996). Finally, we are not committed to the view that children use any of the particular distributional methods that we will discuss. However, we do advocate the general utility of distributional information and learning mechanisms.

The structure of this article is as follows. Distributional Methods, Statistics and Connectionism briefly discusses these ideas and the relationship between them. In Why Distributional Information have been Ignored, we discuss reasons why language acquisition theorists have neglected distributional approaches, and rebut a priori objections to the utility of distributional information. In Distributional Analysis must be Assessed Empirically we illustrate the role of, and guiding principles for distributional approaches in language acquisition research. Finally, in Empirical Research, we illustrate the utility of distributional information, and distributional learning mechanisms, for a range of linguistic phenomena.

DISTRIBUTIONAL METHODS, STATISTICS, AND CONNECTIONISM

Consider a simple distributional property of a corpus: Co-occurrence statistics. Given the corpus *to be or not to be*, the co-occurrence statistics for adjacent words in this corpus are that *to be* occurs twice, while *be or*, *or not*, and *not to* all occur once. Simple co-occurrence statistics of this kind can be a useful cue concerning the syntactic category of words, as we see below. Such statistics can be easily represented in a *contingency table* (Fig. 1).

Co-occurrence statistics can also be easily captured by a connectionist network. In the network shown in Fig. 2,units in the first layer are activated

$word_n$ \ $word_{n+1}$	to	be	or	not
to		2		
be			1	
or				1
not	1			

FIG. 1. A contingency table for corpus *to be or not to be.*

to represent the "current word", and units in the second layer are activated to represent the "next word". The connections between two units are strengthened whenever both units are active (i.e. a form of Hebbian learning). The weights of the network will reflect the co-occurrence statistics of the corpus in exactly the same way that the contingency table does.

Clearly there are many other possible distributional properties. A more complex property is the presence/absence of different combinations of phonetic features in the spoken form of a word. Rumelhart and McClelland (1986) show how a single-layer connectionist network can map from present to past tense for both regular and irregular English verbs, using this kind of distributional representation. The problem of optimally training a single layer neural network is directly analogous to a conventional statistical technique: Multiple linear regression. So, Rumelhart and McClelland's (1986) model can be interpreted as picking up simple distributional statistics.

Moreover, at a more general level, many connectionist learning algorithms can be viewed as implementing general statistical principles, such

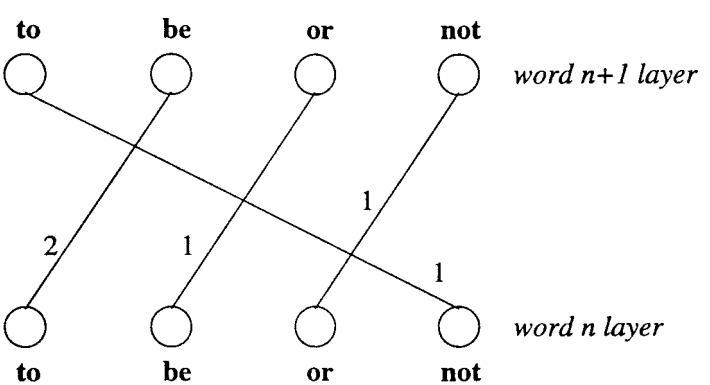

FIG. 2. A Hebbian network, whose weights reflect the same statistics as the contingency table shown in Fig. 1. For clarity, only non-zero weights are shown.

as maximising the probability of the weights chosen according to Bayesian principles (e.g. Mackay, 1992), or minimising description length (e.g. Zemel, 1993).

This does not mean, however, that connectionism offers no new distributional methods in addition to conventional statistics. For example, multilayer networks trained by back-propagation do not correspond directly to any standard statistical method. Such networks have been used to model many important aspects of language acquisition and processing (e.g. Plunkett & Marchman, 1991; Seidenberg & McClelland, 1989). Furthermore, multilayer networks with recurrent connections trained by back-propagation (for example, Elman's 1990 simple recurrent networks), which have no direct relation to existing statistical techniques, have been widely applied to language related tasks (e.g. Abu-Bakar & Chater, 1993; Christiansen & Chater, 1994).

Conversely, while many statistical methods can be directly implemented as connectionist networks, this may not be possible in all cases. For example, nonparametric statistical methods (such as rank correlation, which is used in a model described later) do not readily translate into connectionist networks. None the less, it may be possible to approximate some standard statistical methods using connectionist networks, as we shall see.

Overall, given these relationships it seems appropriate to consider the utility of distributional learning methods in general, both connectionist and statistical. The arguments levelled against the relevance of distributional information apply equally to connectionist and conventional statistical methods, and our case studies draw on examples of both kinds.

WHY DISTRIBUTIONAL INFORMATION HAS BEEN IGNORED

Language acquisition researchers appear to have downplayed distributional information for three reasons. First, historically, distributional analysis has been associated with now discredited ideas, including structuralist linguistics, behaviourist psychology, and positivist epistemology. Second, researchers have been influenced by a priori objections to the notion that any interesting aspects of language can be learnt at all. Finally, the proposal that distributional methods can provide useful information about linguistic structure has been heavily criticised. If valid, these objections would damn both statistical and connectionist approaches. We now rebut each in turn.

Historical Associations: A Failed Programme

Within linguistics, distributional analysis is most closely associated with the structuralists. From Bloomfield to Harris, distributional methods were central to linguistic methodology, providing a set of (distributional)

discovery procedures through which the field linguist could discover the nature of an unknown language. Starting with the smallest linguistic units, the linguist attempted iteratively to discover high-level units and their relationships. The data used as the input to the discovery procedures was minimal: Corpora of observed utterances, with the sole addition of native speaker judgements of sameness and difference between pairs of utterance tokens.

Structural linguists did not assume that such discovery procedures had any relationship to psychological processes of language acquisition. Indeed, structuralists assumed that psychological aspects of language (and its acquisition) stood outside the concern of linguistics, and could be dealt with within the framework of behaviourist psychology. Structural linguistics was also explicitly allied to a positivist conception of epistemology, in which each science is assigned a particular, circumscribed domain of data to be explained, and uses a set of inductive procedures for finding regularities in that data (see Fodor, 1981, for discussion).

With the Chomskyan revolution, the structuralist, behaviourist, and positivist views associated with distributional analysis were undermined. The nature of linguistics as a subject, regarding its scope, data and theory, were changed almost beyond recognition. Generative grammar replaced the structural description of language, providing a much more rigorous and far-reaching account of syntax and many other aspects of language structure. Linguistics was viewed not as independent from, but as part of, psychology, concerned with characterising the knowledge of language involved in production and comprehension (Chomsky, 1980).

Furthermore, behaviourist (and by extension all associationist) psychology of language was devastatingly criticised (Chomsky, 1959). These developments went alongside a rejection of the positivist assumptions of the structuralists. As part of psychology, linguistics was no longer merely a framework for developing abstract descriptions of utterances in each given language. There were no longer any methodological prescriptions concerning how theories should be constructed: In principle, data from linguistic informants, data on acquisition, data from experimental studies, and neuropsychological constraints were all considered relevant. Outside the study of language, behaviourism and positivism were also discredited and replaced wholesale (see Neisser, 1967; Quine, 1953).

As part of this revolutionary change, the distributional analysis of corpora was overshadowed by the use of native speaker grammaticality intuitions as the primary source of linguistic data. In the psychology of language, the failure of behaviourist methods to give a satisfactory theory of language as a whole brought into disrepute associationist accounts of any aspect of language. In the wake of this upheaval, distributional methods are frequently dismissed (in conversation) as hopelessly outdated, conclusively

falsified, doomed to failure, and deeply incompatible with everything that has been learnt about language and mind in the last 40 years.[1]

In fact, such dismissals are largely based on guilt by association. Distributional analysis, as a way of learning aspects of language, is wholly compatible with generative grammar, cognitivism, and modern epistemology. Even if children are innately equipped with a universal grammar, there are still many aspects of the particular of their native language that must be picked up by experience. Distributional learning mechanisms provide a potentially useful means which might contribute to this process. As potential psychological models, although some distributional methods may be couched in associationist terms (such as the connectionist and statistical methods described later), this requires no general commitment to associationism as a theory of the entire cognitive system. Finally, the positivist aspects of distributional methods do not apply in the context of viewing distributional analysis as one of many sources of information that the child may use in acquiring language.

A Priori Objections to Language Learning in General

The Poverty of the Stimulus

The poverty of the stimulus argument proposes that the vast majority of children acquire language so rapidly and so well, and the input they receive is of such variable quality, that most of their knowledge of language must be innate (Chomsky, 1965). This argument, if correct, directly counts against an extreme tabula rasa empiricist position, in which all language structure is learnt by general cognitive mechanisms.

But the claim that *some* interesting aspects of language can be learnt does not imply the claim that *all* aspects of language can be learnt from scratch. Indeed, as discussed previously, empiricist and nativist positions alike are compatible with distributional learning mechanisms. Also, as noted earlier, some aspects of language clearly must be learnt, and hence some kind of learning mechanism, whether distributional or otherwise, must succeed despite the poverty of the stimulus.

The poverty of the stimulus argument, when applied to some specific aspect of language, asserts that this aspect of the language cannot be learned consistently from the varying and possibly very poor inputs received by children. Disconfirming this argument in a particular case requires the construction of some learning mechanism that robustly learns the relevant aspect of language across the variety of inputs children receive. To validate

[1]This dismissal of distributional methods was never quite complete. They have retained some importance as discovery procedures (for linguists) in phonology and morphology.

the poverty of the stimulus argument, is is necessary to attempt such disconfirmation. Although we might find it difficult to see how some particular aspect of language can be learnt, the poverty of the stimulus does not stand as already established, and able to count against the plausibility of distributional and other learning methods. Rather, only by pursuing such learning methods as vigorously as possible can the poverty of the stimulus argument gain any credibility in regard to particular linguistic phenomena. This gives an additional reason why those who emphasise innate knowledge over learning mechanisms should be interested in, rather than dismiss, distributional learning mechanisms.

The Absence of Negative Evidence

Language learning mechanisms, including distributional analysis, are sometimes argued to be infeasible, because the child does not receive "negative evidence" (e.g. Baker, 1979).[2] The child generally receives only grammatical speech input, and those ungrammatical sentences it does hear are not marked as such. Furthermore, in production, children are not reliably informed whether their utterances are grammatical or not (although there is some debate over whether children receive "noisy" feedback, Marcus, 1993, we will assume this is not the case for the present discussion). How the child manages to determine the appropriate linguistic generalisations, and manages to avoid or retreat from inappropriate (over-)generalisations, despite this lack of negative evidence, is a central problem in the study of language acquisition (e.g. Bowerman, 1987, 1988).

This point is related to the poverty of the stimulus argument, in that the lack of negative evidence is a particular way in which the input that the child receives counts as poor, and it is unpersuasive for the same reasons. At a general level, discounting the possibility of language learning because of the lack of negative evidence leads to a reductio ad absurdum. Almost all interesting learning from experience occurs without negative evidence, from finite sets of observations. For example, scientific theories are entirely grounded in observations of what *does* happen. None the less, scientific progress seems possible. In learning about the physical structure of the world children, too, see only positive evidence. Yet they appear to learn a great deal about the world from this evidence alone.[3] Since almost all interesting problems of learning from experience involve no negative evidence, and can manifestly be solved successfully, there seems no reason,

[2]Ironically, the "no negative evidence" problem was first identified by Braine (1971), as an obstacle to Chomsky's (1965) nativist account of language acquisition.

[3]Of course, children may receive some "negative evidence" in verbal instruction by parents, perhaps that objects do not fall upwards, but presumably, as in the language learning case, this is relatively unimportant.

at a general level, to assume that language learning from experience faces any special difficulties.

Gold (1967) is sometimes misinterpreted as having provided formal proof that language learning without negative evidence is impossible. In fact analogues of Gold's results apply to all interesting learning problems (those with an infinite number of possible hypotheses). If Gold's result really implied that learning language from positive evidence alone is impossible, it would also rule out the possibility of scientific endeavour and human learning in almost every interesting domain. In fact, Gold's result does not have this epistemologically catastrophic conclusion. Gold takes an extremely simple model of learning, and examines how data is presented to the model, and shows that learning the correct solution is not always possible "in the limit" given this idealisation. The appropriate reaction to Gold's result is to search for a richer idealisation of learning, and the data upon which learning occurs.

The kind of problem that Gold appears to raise is familiar from the philosophy of science, in which it is a commonplace that any finite set of data is consistent with an infinite number of hypotheses.[4] Choosing among these hypotheses requires criteria additional to consistency with the observed data (and this cannot itself be taken as absolute, since observations may be inaccurate or untrustworthy). The problem of characterising these criteria is central to epistemology and philosophy of science—criteria such as simplicity, generality of explanation, consistency with past theorising, and so on have been suggested, but prove difficult to define formally.

In the study of language acquisition, there are no easy or straightforward solutions to the problems posed by the absence of negative evidence. Even strongly nativist accounts face great difficulties in accounting for children's ability to arrive at appropriate linguistic generalisations (see Bowerman, 1987, 1988, for discussion).

We suggest that in studying distributional approaches, language acquisition researchers should seek to take advantage of research in mathematical statistics, machine learning, pattern recognition, and related disciplines. Within these fields, the question of how one of an infinite set of hypotheses is chosen given a finite amount of data is the central research question, and a large range of formal approaches have been proposed, both in general, and for specific classes of problem. The connectionist and statistical learning methods discussed in this article are examples of learning methods that find considerable structure in language from positive evidence

[4]Gold (1967) is concerned with language learning in the limit, rather than learning given a finite set of data. Of course, all actual learning, including that involved in language acquisition, must be based on a finite set of data, and hence the "underdetermination of theory by data" applies.

alone. Without careful investigation of the specifics of the input that children receive (such as that outlined later), and the range of learning mechanisms, distributional and otherwise, that they might use, claims concerning what children can and cannot learn are simply unfounded.

A Priori Objections to Distributional Language Learning Mechanisms

So far, we have considered general objections to the possibility that interesting aspects of language structure can be learned by any mechanism. We now turn to objections specifically targeted at distributional learning mechanisms. We consider a number of claims, due to Pinker (1984), that distributional analysis cannot yield information useful for language acquisition.

Some Properties of Language Must be Deduced

Pinker (1984) describes an argument, which he attributes to Grimshaw, that allegedly shows that distributional analysis cannot successfully lead to the acquisition of word classes, such as *noun* and *verb*. Since this argument (Pinker, 1984, pp. 48–49) is quite complex, we summarise it here.

1. Some properties of language cannot be learnt without negative evidence (e.g. adults never perform extraction from complex noun phrases—the complex noun phrase constraint, CNPC).[5]
2. Therefore, correlations of such properties with other properties (e.g. the co-occurrence statistics of nouns) cannot be observed.
3. Therefore, these correlations must be deduced (i.e. the child must first know that an element is a noun, in order to predict that it also obeys the CNPC).
4. Thus, such properties (as the CNPC) cannot be used as part of a discovery procedure (e.g. for identifying which words are nouns).

The first point concerns the possibility of learning in the absence of negative evidence. Although we have already discussed this, Pinker's exact words (1984, p. 49) neatly illustrate the general misconception in this area:

> The only possible clue in the input that nouns have such a property [the CNPC] is that adults never use sentences involving extraction from complex noun phrases . . . The child cannot use this absence as evidence since . . . the very next sentence in the input could have extraction from a complex noun phrase, and

[5]For instance, given the sentence *John made the claim that he saw Bob*, the noun *Bob* cannot be extracted from the noun phrase *the claim that he saw Bob*: The question **Who did John make the claim that he saw?* is ungrammatical.

> their absence until then could have arisen from sampling error or a paucity of opportunities for the adult to utter such sentences.

This argument is clearly fallacious. Consider the well-attested generalisation that the sun rises in the East. By Pinker's argument, we could conclude that this hypothesis could never be learnt from positive instances alone, and the absence of counterexamples. For, to paraphrase the original argument, the very next morning could see the sun rising in the West, and indeed, paucity of experience and sampling error raise the possibility that it may already have done so, but we slept in.

Given the *assumption* that the CNPC cannot be learnt by the child on the basis of positive evidence alone, the remainder of Pinker's argument is correct: The CNPC cannot be used to aid in the discovery of nounhood. In fact, this assumption, and Pinker's conclusions, are justified on quite different grounds: If the child does not know which words are nouns, and presumably has an equally poor knowledge of other syntactic categories, they will not be able to observe, and utilise, properties (such as the CNPC) that are defined in terms of word classes.

However, this conclusion, that learners cannot utilise relationships that are not apparent in the surface structure of the language, is irrelevant to the utility of distributional learning. Generally, distributional learning mechanisms exploit relationships that *are* apparent in the surface structure. The important question is, "Are these, easily observable, relationships, sufficient to adequately categorise words as nouns?" This question must be addressed empirically, rather than by a priori argument.

Distributional Methods are Unconstrained and Uninformative

Pinker (1984) also provides another three reasons why distributional information is uninformative for language acquisition. First, Pinker argues that relationships that are apparent to the learner in the surface structure of language cannot usefully be exploited by distributional methods. He claims that the vast number of possible relationships that might be included in a distributional analysis is likely to overwhelm any distributional learning mechanism in a combinatorial explosion.

Second, he claims (in answer to the question posed), easily observable properties of the input are in general linguistically uninformative (Pinker, 1984, pp. 49–50):

> Most linguistically relevant properties are abstract, pertaining to phrase structure configurations, syntactic categories, grammatical relations, … but these abstract properties are just the ones that the child cannot detect in the

> input prior to learning … the properties that the child can detect in the input—such as the serial positions and adjacency and co-occurrence relations among words—are in general linguistically irrelevant.

Third, Pinker argues that "even looking for correlations among linguistically relevant properties is unnecessarily wasteful, for not only do languages use only certain properties and not others, they sanction only certain types of correlations among those properties".

Pinker's second point relies on equivocation over what is meant by "linguistic relevance". Uncontroversially, generative grammar does not capture the structure of language in terms of serial position, adjacency, and co-occurrence relations. However, this is not to say that such relations are not linguistically relevant, in that they carry useful information about the structure of language. Indeed, contrary to Pinker's assertion, all three of the examples he gives can provide information about a word's syntactic category, for English at least. The utility of distributional learning mechanisms, as a technique for investigating language acquisition, is that they allow empirical tests of such assertions. As should be clear from the above, a priori intuitions on such matters, even with the benefit of experience, cannot be trusted.

Pinker's first and third points, the danger of a combinatorial explosion, and the wastefulness of a mechanism which cannot be applied to certain languages, are also misguided. The first point appears to assume that distributional learning mechanisms will search blindly for relationships between a vast range of properties. Although this may be a fair criticism of early, unimplemented distributional proposals (e.g. Maratsos & Chalkley, 1980), the kinds of learning mechanisms that contemporary researchers have considered and implemented tend to focus on highly specific properties of the input. Although combinatorial explosion is a possible issue for some distributional learning mechanisms (especially those dealing with raw speech), the existence of learning mechanisms that can demonstrably deal with realistically sized (millions of words) corpora suggests that it is not an obstacle to distributional learning mechanisms in general.

Pinker's third point starts from reasonable premises: As languages vary in many respects, it seems likely that different learning mechanisms will be recruited, and that their contributions might differ from one language to the next. But this cannot be condemned as "unnecessarily wasteful". Since the child is able to learn any language, but in actual fact generally faces only one, its learning apparatus is "wasteful" by necessity. Even a strict universal grammar account is "wasteful", in that some parameter settings will go unused.

Spurious Correlations

A final objection to distributional analysis is that "spurious correlations" will arise in local samples of the input. For example, the child could hear the sentences *John eats meat*, *John eats slowly*, and *the meat is good* and then conclude that *the slowly is good* is a possible English sentence (Pinker, 1984).

Although this may be a fair criticism of early and underspecified distributional proposals, an important aim in the study of distributional learning mechanisms is to avoid such spurious generalisations. The fact that a brittle and extraordinarily naive approach to distributional analysis, which draws conclusions from single examples, falls prey to such errors cannot be taken as damning for the class of distributional approaches. Without consideration and empirical assessment of more sophisticated approaches such objections are premature.

DISTRIBUTIONAL ANALYSIS MUST BE ASSESSED EMPIRICALLY

It should now be clear that distributional methods have not been ruled out on a priori grounds. Many of the objections to distributional methods simply assert that such methods cannot succeed, rather than demonstrating that this is the case, either in general, or for specific aspects of language. We believe that the extent to which distributional methods can, and do, contribute to language acquisition is an empirical question.

The relevant evidence here is of two main kinds. First, the extent to which distributional learning mechanisms, applied to realistic input, can acquire particular aspects of linguistic knowledge. Second, the extent to which infants are sensitive to the properties of language exploited by distributional models, and the fit between the developmental profile of infants' linguistic knowledge, and that predicted by the models.

In practice, many other kinds of evidence are also likely to be relevant, including but not limited to studies (and simulations of) language impairment, adult learning of artificial grammars, distributional learning mechanisms applied to artificial input, neurophysiological evidence, etc., but in this article we will concern ourselves with those already stated.

The application of distributional learning techniques to language has generally been restricted to engineering and computer science, and only recently have distributional methods, particularly connectionist networks, had even a small influence on developmental research. This situation is regrettable, given the potential relevance of such investigations to the study of language acquisition. We now outline the sort of role that we see for the investigation of distributional methods within the psychology of language development.

General Methodology—An Analogy with Vision

We suggest that the study of distributional methods in language acquisition has a parallel in the study of computational approaches to vision (Marr, 1982; Richards, 1988).

Early pioneering work in computational vision concentrated on simple artificial problems (e.g. Waltz, 1975). However, it is now widely accepted that work on simplified problems did not shed much light on human visual processing. Only by studying the natural problem of vision is it possible to identify the relevant regularities in the environment. Gibson (1979) argued that a central component of the psychology of vision should be the study of the relationships between the structure of the optic array and the structure of the environment that make perception possible. Marr (1982) adapted and extended this approach to computational vision, suggesting that computational research focus on analysing the relation between the natural world and natural images. In studying natural images, low-level problems such as stereopsis (Marr & Poggio, 1979), edge detection (Marr & Hildreth, 1980), and shape from shading (Horn, 1975) must be addressed before more complex problems, such as object recognition, can be tackled. In studying edge detection, for instance, computational vision is concerned with finding out what edges really look like, that is, what are the visual properties that generally distinguish real edges from other kinds of boundaries, such as shadows or marking. To assess which cues are informative requires analysis of typical natural images, especially their statistical properties.[6]

Many basic phenomena in vision are influenced by both low-level and high-level information. Segmenting the visual image into coherent parts is importantly constrained by low-level factors, such as texture and colour segregation, depth cues, and so on, but is also influenced by high-level information, such as the identity of the object being viewed. For example, highly degraded perceptual stimuli (such as the well-known dalmatian dog on a spotted background; Gleitman, 1991, p. 224) are more rapidly recognised, and thus segmented into meaningful parts, when subjects are told the content of the stimulus. Whereas researchers generally recognise that both low- and high-level factors are relevant in image segmentation, computational vision has concentrated on low-level factors, because there is no clear conception of how high-level information is represented.

---

[6]In practice, research does not always require exhaustive statistical analysis of every relationship between the image and the natural world, because many regularities are known to hold from optics, and other aspects of natural science. Thus, knowledge of optics allows the construction of a detailed mathematical theory of how shape can be derived from shading, without having to discover this relationship anew using statistical analysis. In the computational study of language acquisition, it is presumably also valuable (and in practice unavoidable) to guide research using constructs from existing theory, in this case linguistics.

This approach to computational vision has successfully tackled limited but important problems. It has proved possible to build computational models that are directly constrained by existing psychophysical and neurophysiological data, and suggest new directions for experimental research. These successes complement, rather than oppose, the study of higher level processes in vision. The value of theoretical and empirical research on low-level and high-level aspects of vision is recognised by all researchers. Researchers studying low-level aspects of vision do not claim to be able to deal with all aspects of vision; they simply choose particular problems where existing techniques can tractably be applied. Researchers studying high-level aspects of vision do not ignore low-level research as irrelevant, but are concerned with how the results of low-level research may inform the interaction between low- and high-level processes.

The application of distributional learning mechanisms to the problem of language acquisition shares many similar concerns, problems, and aims with the study of computational vision. Because distributional information refers specifically to properties of the input, studies of artificial data are likely to be of limited use. A central question for such research is, "What information is available in the natural input to the child?"

Just as computational vision research works with natural images, so studies of distributional learning mechanisms should work with real corpora of natural language. This is not to say that any work that does not utilise raw speech signals is of no interest. In practice, some idealisations will always be necessary. Typically, researchers might work with corpora transcribed at the phonemic level, or at the level of words, or with lexicons rather than corpora. These idealisations assume that the child has solved the problem of mapping raw speech to phonemes, or the problem of segmenting the input into distinct words. Generally we should aim to work on "real-world" problems, utilising an approximation to the child's input[7] that is as faithful as possible, given our techniques and technology, and with any simplifying assumptions clearly stated. Naturally, the learning mechanisms that we consider must be able to scale up to deal with similar quantities of natural language input to that received by real language learners.

Just as perceptual phenomena in vision can be influenced by both low-level and high-level factors, so are many linguistic phenomena. For instance, top-down semantic influences can encourage listeners to segment

[7]Of course, there may be considerable differences between the input received by the child and the input that is attended to and encoded. However, because we can only directly observe the input received by the child, an approximation of the linguistic input received by the child is assumed to be an approximation to the input to the relevant learning mechanisms. This is not a hard and fast rule. For example, Elman (1993) and Newport (1990) have suggested that limitations of memory span may sometimes make detecting linguistic regularities more straightforward, and computational models of learning might reflect such proposals.

an utterance as "wreck a nice beach" or "recognise speech". Although connectionist and statistical learning methods can in principle be applied to language external factors, such as semantics and pragmatics, there is no clear conception of how to appropriately represent these language-external properties of the environment. Therefore, as in computational vision, researchers must necessarily concentrate on relatively low-level, language internal features of the environment. The availability of large corpora of natural language, both of text, and of transcribed speech, make studies of language internal cues relatively feasible. For studies at or below the level of individual words (e.g. connectionist studies of morphology), corpora are not necessarily required; lexicons of words, possibly with frequency information may suffice.

A number of other methodological concerns are also relevant. In assessing the success of a particular distributional method in acquiring some particular aspect of language, it is important to have some quantitative measure of success. Previously, many distributional learning mechanisms have been considered successful on the grounds that qualitatively, their output appears to be "linguistically relevant". However, for particular methods to be meaningfully assessed across different kinds of input (text, transcribed speech, across languages, etc.) a quantitative measure is required. Where this measure is simply the extent that the method does acquire a particular kind of information, this assesses the *feasibility* of the learning mechanism. Where a measure takes into account the correspondence with the child's linguistic knowledge throughout development, this potentially provides much stronger evidence for the utilisation of distributional information, and/or particular learning mechanisms, by real language learners.

Quantitative measures will also be required in order to study the effect of combining possible sources of information. Generally studies of distributional methods in language have considered particular cues in isolation. The general strategy of starting with the simplest possible assumptions, in terms both of learning apparatus, and innate knowledge, is one that we strongly support. However, a single language-internal cue is unlikely to provide a complete solution to any particular problem. Indeed it is possible that cues which are wholly uninformative when considered alone may be highly informative when combined with other cues or innate knowledge. The study of combinations of cues is a natural progression for work of this kind.

We believe that by concentrating on low-level, tractable problems, distributional learning mechanisms will be able to provide empirical evidence regarding the potential, and actual, contribution of distributional information to language acquisition processes. The validity of such evidence will be crucially determined by the extent to which the learning mechanism's

input reflects the properties of the child's linguistic environment. We believe that this kind of approach is complementary to, and should both inform and be informed by, empirical work with infants, and studies of the influence of higher-level factors in language acquisition, such as semantics and pragmatics.

We conclude this section with a caveat. We believe that connectionist and statistical learning methods are powerful tools for the study of language acquisition. However they are by no means universal panaceas. To state that a particular aspect of language is acquired from distributional information has, by itself, no more explanatory power than to say that a particular aspect of language is known innately. What is required are specific learning mechanisms, or kinds of distributional cue. Identifying appropriate mechanisms and cues, and demonstrating their validity, is by no means a trivial task. It is likely to require careful consideration of the aspect of language to be acquired, the appropriate idealisation of the input, the statistical properties of the input that are likely to be informative, and, especially for connectionist models, the appropriate kinds of input and output representation for the learning mechanism.

EMPIRICAL RESEARCH

In this section we present numerous examples of the application of distributional learning methods to language learning. These illustrate the role that connectionist and statistical learning methods can play in investigating a wide range of language acquisition phenomena, from the acquisition of phonology, and the segmentation of the speech signal, through to elementary phrase structure and lexical semantics.

Because these models receive realistic input (generally corpora of natural text or transcribed speech), and demonstrably do acquire linguistic knowledge, they serve to refute the criticisms of distributional methods discussed earlier. Where the work we discuss has borne psychological issues in mind, we have tried to make this clear, and to point out the connections to the relevant psychological and developmental issues accordingly. However, for some examples, especially those emerging from the study of distributional methods and language as technology, no clear psychological or developmental predictions can be derived. We include such examples because they serve as feasibility proofs that a particular source of information can be informative about particular aspects of language. Generally, we believe that there is much more scope for connections between distributional learning mechanisms and the evidence from child studies. We hope that this article will encourage developmentalists to make those connections.

Finally, while the examples we discuss are wide-ranging, they by no means constitute a comprehensive review of the psychological work in this area, let alone potentially relevant technologically oriented work, from speech recognition or language engineering. Such an ambitious project is beyond the scope of this article. Rather they represent a sample of the work in this area, and a biased one at that, as we have tended to choose examples with which we were familiar, or involved in ourselves. Nevertheless, we think that they serve to illustrate our argument.

When discussing the application of distributional methods to each area of linguistic knowledge below, we first describe the problems faced by the language learner in acquisition and summarise the relevant developmental evidence, then discuss potential sources of information, both distributional and otherwise, before presenting the case studies themselves.

Phonology and Prosody

The Problem

A fundamental problem from the very earliest stages of language acquisition concerns the speech sounds allowable in the language, and the rules governing how they are combined, that is, the child must learn the phonological structure of their native language.

Development of Phonology

Learning the sound patterns of one's native language begins very early, perhaps even before birth. Mehler, Jusczyk, Lambertz, Halstead, Bertoncini, and Amiel-Tison (1988) showed that newborns can distinguish their native language French from Russian across a wide variety of utterances. This discrimination persisted when speech was low-pass filtered, suggesting that the information used is primarily prosodic (e.g. rhythm, stress, and intonation).

The ability to distinguish between prosodically similar languages (such as Dutch and English), on the basis of their differing phonological properties develops between six and nine months; nine-month-old infants will turn their heads preferentially to their native language, but this effect disappears if the input is low-pass filtered, eliminating phonological cues (Jusczyk, Friederici, Wessels, Svenkerud, & Jusczyk, 1993). At around the same age, sensitivity to the sequential structure of phonemes develops. Jusczyk, Luce, and Luce (1994) found that nine-month-old infants listen for longer to lists of words that contain frequent combinations of speech sounds, rather than words containing infrequent sound patterns.

This development of discrimination between speech sounds in the native language appears to occur at the expense of discriminations that are relevant

only in other languages; that is, children become insensitive to contrasts that do not occur in their own language (Best, McRoberts, & Sithole, 1988; Werker & Tees, 1984).

Possible Sources of Information

Learning the phonological structure of the language appears to be a prerequisite of acquiring higher order aspects of language, such as words, morphology, and syntactic categories. Therefore, the early acquisition of phonology is presumably based primarily on language-internal properties of the speech input, in combination with innate constraints. However, the problem of translating speech to phonemes has proved very difficult. Research on a computer speech recognition (as opposed on acquisition) has enjoyed only relatively modest success, after decades of intensive and computationally sophisticated work (suggesting that considerable innate knowledge—possibly in the form of highly sophisticated language specific, distributional learning mechanisms—may be brought to bear in the child's solution to the speech recognition problem).

Case Studies

In the light of the problems encountered in computer speech recognition, and the difficulties of dealing with raw speech signals, developmentally oriented work in this area has until recently been relatively scarce. The few brave exceptions include a number of connectionist networks that aim to extract relatively low-level features of the speech stream, which may help identify phonetic features and phonemes (e.g. Cottrell, Nguyen, & Tsung, 1993; Nakisa & Plunkett, this issue).

Other research applying distributional learning methods to the acquisition of phonology has used idealised speech input. For example Abu-Bakar and Chater (1993) used representations of artificially constructed formant transitions in investigating how a connectionist network can discriminate phonemes using duration based cues (e.g. voice onset time) when these durations are sensitive to overall speech rate in a complex, nonlinear way. However, to assess the validity of these findings with real speech would require large corpora of speech, transcribed at a very low level of detail, and a significant scale-up of the connectionist network (which may not be feasible for the particular connectionist architecture used).

Given the difficulties involved in converting speech to phonemes, and of working with raw speech, a more fruitful line of research has taken a solution to the "speech-to-phoneme" problem as a given, and used a phonemic or phonetic representation of the input as a starting point. This research has

focused on finding linguistically relevant patterns in strings of phonemes or bunches of phonetic features.

For example, Ellison (1992) conducted a distributional analysis on a restricted lexicon of words (represented as strings of phonemes) for a wide range of languages. Using a Bayesian statistical analysis, he showed that a number of important aspects of phonology could be derived automatically, including the distinction between vowels and consonants, learning sonority hierarchies, and vowel harmony. This work is not directly psychologically motivated, but provides a feasibility proof that distributional statistics are potentially highly informative concerning phonological structure (see Vroomen, van den Bosch, & de Gelder, this issue, for related work using a simple recurrent network).

Shillcock and colleagues (e.g. Shillcock, Hicks, Cairns, Levy, & Chater, in press) have developed a phonetically transcribed corpus of conversational English, which can be used for studying the informativeness of a variety of distributional cues. The corpus is a transcription of the London-Lund corpus (Svartvik & Quirk, 1980) of conversational English, which was originally transcribed at the level of individual words. These words were automatically converted into strings of phonemes, and these phonemes were converted into bundles of phonological features. To mirror real speech more closely, some phonetic features were "reduced" in line with standard rules for phonological reduction. This corpus is therefore a useful, although rough, approximation to what would be obtained from direct phonetic transcription of the original speech by a linguist.

This corpus has been analysed both by collecting co-occurrence statistics and a recurrent connectionist network (Cairns, Shillcock, Chater, & Levy, 1995; Shillcock, Lindsey, Levy, & Chater, 1992). Both approaches show that sequences of phonetic features (and sequences of phonemes) can be predicted successfully from recent phonological context, which suggests that local distributional statistics are informative about phonological regularities.

This work has potentially important consequences for the interpretation of experimental data that appear to give strong support for the interactive view of spoken word recognition (Elman & McClelland, 1988). While theories of adult spoken language recognition are not directly connected to developmental work, they obviously place a strong constraint on developmental theories (and vice versa), and may be highly relevant to related developmental problems, such as learning segmentation.

The central issue here is the extent to which phoneme detection is influenced, as the interactive view supposes, by the feedback of information from the lexical to the phonemic level. Elman and McClelland (1988) take as their starting point the apparent word superiority effect for phoneme restoration: Degraded phonemes are perceptually restored more strongly in

words than in phonologically regular nonwords. At first glance, it might appear that these lexical effects directly demonstrate top-down feedback. But there is an alternative explanation, which is entirely compatible with a modular view: Subjects' decisions concerning which phoneme was heard is influenced by both phonological and lexical representations of the stimulus. According to this view, the lexical level directly influences the subject's decision, without any top-down influence on the phoneme detection process itself.

Experimentally disentangling these two explanations is extremely difficult. But Elman and McClelland (1988) noticed a prediction of their interactive TRACE model, which appeared to suggest an appropriate crucial experiment. In natural speech, the pronunciation of a phoneme will to some extent be altered by the phonemes that surround it, in part for articulatory reasons: This phenomenon is known as coarticulation. This means that listeners should adjust their category boundaries depending on the phonemic context. Experiments confirm that people do indeed exhibit this "compensation for coarticulation" (Mann & Repp, 1980). For example, given a series of synthetically produced tokens between /t/ and /k/, listeners move the category boundary towards the /t/ following a /s/ and towards the /k/ following a /sh/.

This phonemic phenomenon suggests a way of detecting whether lexical information really does feed back to the phoneme level. Elman and McClelland (1988) considered the case where compensation for coarticulation occurs across word boundaries, for example, a word-final /s/ influencing a word-initial /t/ as in *Christmas tapes*. If lexical-level representations feed back on to phoneme-level representations, the compensation of the /t/ should still occur when the /s/ relies on lexically driven phoneme restoration for its identity (i.e. in an experimental condition in which the identity of /s/ in *Christmas* is obscured, the /s/ should be restored and thus compensation for coarticulation proceeds as normal). Elman and McClelland noticed that their interactive TRACE model does indeed produce this prediction, whereas it is difficult to see how a modular, noninteractive model could account for this effect. They therefore decided to conduct the crucial experiment.

Subjects heard pairs of words such as *Christmas tapes* or *foolish capes*, where the last segment of *Christmas* or *foolish* was replaced by a synthetic segment midway between /s/ and /sh/. The first segment of *tapes*/*capes* was a synthetic segment drawn between /t/ and /k/. Subjects were required to report the identity of the second word. The results indicated that compensation for coarticulation across the word boundary occurred just as if the final phoneme of the first word had been unambiguous (i.e. restored), suggesting that the lexically restored final phoneme was able to trigger compensation for coarticulation at the phonemic level.

Advocates of bottom-up connectionist models have argued that Elman and McClelland's (1988) results do not demonstrate top-down lexical influences on phoneme identification. Norris (1993) trained a recurrent connectionist network with no lexicon (and therefore no possibility of top-down lexical effects) on a small artificial data set, and observed compensation for coarticulation similar to that observed experimentally by Elman and McClelland. However, the crucial question is whether a network trained on natural speech, rather than artificial data will model the Elman and McClelland's results?

Shillcock *et al.* (1992) constructed such a network and trained it on the phonologically transcribed version of the London–Lund corpus. A recurrent network was trained on the corpus of phonologically transcribed conversational English, with inputs and outputs at the level of phonetic features. As in Norris's (1993) simulations, there was no lexical level of representation from which top-down information could flow. None the less, phoneme restoration follows the pattern that Elman and McClelland (1988) explain in terms of lexical influence.

Why is it that in the simulation purely bottom-up processes appear to mimic lexical effects? Restoration occurs because the network has picked up distributional regularities at the phonemic level, rather than because of lexical influence. It just happens that the lexical items that Elman and McClelland (1988) used experimentally are more statistically regular at the phonemic level than the nonwords with which they are contrasted. This is confirmed by a purely statistical analysis of the corpus of speech on which the network is trained. By carefully choosing stimulus items for which statistical regularities at the phonemic level have the opposite bias to that which would be provided by lexical status, it may be possible to distinguish experimentally between the interactive and bottom-up connectionist accounts. This experimental test is yet to be conducted, however.

Note that a finding in favour of the bottom-up connectionist model has strong developmental implications; it would suggest that the distributional properties to which the simple recurrent network is sensitive could be utilised by human infants in learning to identify phonemes.

The debate between interactive and bottom-up models of speech perception that we have just described is a good illustration of the way in which using distributional methods, on real data, has led to unexpected theoretical predictions being derived (e.g. that bottom-up models can account for apparently lexically based phoneme restoration), as well as indicating new directions for empirical research.

Summary

To conclude, in the case of learning phonological and prosodic structure, there appears to be little doubt that distributional information is of considerable significance, and research with lexicons and corpora suggests that this information can be extracted relatively easily. The complexity of the speech signal makes the distributional information used in the speech-to-phoneme mapping difficult to study. But processes involved in learning phonological regularities subsequent to the acquisition of phonemes appear to be much more tractable to distributional methods. Understanding the distributional information that may be drawn on in the acquisition of phonological and prosodic structure is not only a tractable project, but it may be important as a foundation for studying acquisition of higher-level linguistic information. As we shall see, phonological and prosodic cues have been widely proposed as important sources of distributional information in speech segmentation, and the acquisition of word classes and syntactic structure.

Segmentation

The Problem

As well as discovering the sound patterns (phonology) of their native language, the child must also discover how to *segment* appropriately these sound patterns into words. This is a difficult problem because in conversational speech there are generally no "gaps" or other obvious acoustic markers to signal the boundaries between words (Cole, 1980).

Theories of adult speech processing and segmentation, such as the Cohort model (Marslen-Wilson & Welsh, 1978), propose that possible segmentations are licensed, or constrained, by the lexicon. However, in the case of the child, this points to a chicken and egg problem; segmentation into words appears to require knowing what the words of the language are. But the child cannot know what the words of the language are until segmentation can be successfully effected. Somehow, the child must "bootstrap" the ability to segment and learn the lexicon of the natural language.

Development of Segmentation

The child's development of segmentation can be divided into two stages (Plunkett, 1986, 1990). Between the ends of the first and second years (i.e. from the child's first productions to the onset of the vocabulary spurt)

children appear to be able to use and identify some words appropriately, but their speech also contains units which are either parts of words (*idiosyncratic expressions*) or sequences of words (*formulaic expressions*), which are used as lexical items. Thus, the child shows a tendency to under- and over-shoot in finding the appropriate solution to the segmentation problem.

The infant's tendency to over-segment prior to the vocabulary spurt has been observed to vary between children (e.g. Bates, Bretherton, & Snyder, 1988), and to correlate with aspects of their speech style; formulaic expressions are rare in the productions of infants with an "analytic/referential" speech style and relatively high use of concrete nouns, whereas formulaic expressions are more common in infants with a "social/expressive" speech style, and relatively high pronoun use (Nelson, 1973).

As well as individual differences in segmentation strategies, there is evidence that infants may switch between, or explore, different segmentation strategies during development. Plunkett (1986, 1990), in a longitudinal study of two Danish children, observed that, between 13 and 16 months, the proportion of formulaic expressions in Jens' Vocabulary consistently decreased, while the proportion of idiosyncratic expressions increased, in contrast to Jens' previously high proportion of formulaic expressions. Anne showed the opposite pattern, with an initially high proportion of idiosyncratic expressions, and then an increasing use of formulaic expressions (and decreasing use of idiosyncratic expressions) from 16 months onwards.

After the vocabulary spurt (at around 18–24 months), the proportion of idiosyncratic and formulaic expressions decreases rapidly, and the proportion of actual words increases. This appears to coincide with the achievement of a majority of correctly identified lexical items in the child's vocabulary, and hence the finding of an appropriate solution to the segmentation problem.

Possible Sources of Information

A range of possible sources of information that the child may use in segmentation have been suggested. One of the simplest is that children first learn words heard in isolation and then identify these words in fluent speech (Suomi, 1993). However, various problems with this account suggest that additional sources of information are required. First, the speech stream is highly locally ambiguous, so that even if a sequence of phonemes has been observed in isolation, it may be difficult to decide whether an occurrence of that sequence in fluent speech corresponds to a word. For example, the string of phonemes for *but* could occur in *rebuttal*, *butter*, *abut*, and so on. Second, some words (e.g. function words such as *in*, *as*, *and*, etc.) seldom or never occur in isolation, and moreover single word utterances are relatively

rare in conversation.[8] Third, without some further cue, the child cannot know which short utterances are single words rather than, for example, short phrases. Finally, the acoustic properties of words spoken in isolation typically differ considerably from the acoustic properties of the same words in fluent speech (although see Jusczyk & Aslin, 1995, for evidence that eight-month-old infants can recognise previously heard single words in fluent speech). Despite these difficulties, this remains an interesting source of information, and makes some nice predictions (for instance, the rarity of function words in single utterances would predict that these words would be learnt later rather than sooner, as is indeed the case; McCarthy, 1954).

Another source of information is the "flagging" of new words by the speaker by, for example, placing them at the end of the utterance (Woodward & Aslin, 1990, cited in Jusczyk, 1993).

Lehiste (1971) has proposed that although there are no obvious signals of word boundaries in the acoustic input, there may be subtle acoustic/phonetic juncture markers. Many potential markers are identified by Peters (1985), in a comprehensive survey of "Operating Principles" (Slobin, 1973) for speech segmentation. As well as the possibilities mentioned earlier, Peters considers intonation contours, melody, rhythm, stress, and distribution as potential cues, although she provides only anecdotal evidence in support of these proposals.

A number of other authors (e.g. Cutler & Norris, 1988; Gleitman, Gleitman, Landau, & Wanner, 1988) have also proposed that prosodic cues such as stress or vowel lengthening may serve to identify word boundaries, or initial segments. Saffran, Newport, and Aslin (1996) cite the observation that mothers often vary their pitch so as to highlight topical words (Fernald & Mazzie, 1991) and the occurrence of stressed and/or word-final syllables amongst children's early (and often idiosyncratic) productions as evidence for the use of prosodic cues. However, Saffran, Newport, and Aslin also note that the variation in prosody across languages requires that the infant possesses some means of identifying the relevant cues for their native tongue.

All of these cues are obvious candidates for exploitation by distributional learning mechanisms. A further cue, suggested by early work in distributional linguistics (e.g. Harris, 1955), and utilised in the models described later, is the redundancy, or predictability of the speech stream. The phonotactic constraints of a language dictate that the predictability of the speech stream will be higher at word and phrase boundaries than within words and phrases. Saffran, Newport, and Aslin (1996) and Saffran, Aslin,

[8]For instance, in the CHILDES corpus (MacWhinney & Snow, 1985), only approximately 15% of the adult utterances are single-word. While not fatal for the use of single-word utterances as a cue to segmentation, it does place a bound on the utility of this cue.

and Newport (1996) have recently shown that both adults and eight-month-old infants can exploit the difference between transitional probabilities within and between words of an artificial language, as shown by their ability to identify the words of the language (in a recognition test or preferential listening paradigm). Furthermore, both adults and infants were also able to exploit additional prosodic cues in order to improve their performance still further.

A language external source of information for segmentation may be the rough correlation between words and the world. Strings of phonemes which correspond to words will presumably correlate reliably with aspects of the environment (e.g. the presence of a particular object, or performance of a particular action), while strings which do not correspond to words will presumably not have reliable environmental correlates. This also suggests another reason why content words are acquired earlier than function words; content words will often correspond to salient aspects of the environment (e.g. *dog*), whereas function words (which correspond to abstract relationships) would plausibly seem to be much less perceptually salient.

Case Studies

A simple distributional strategy for segmentation, which was initially proposed as playing a role in adult performance, is the Metrical Segmentation Strategy (MSS; Cutler, 1993; Cutler & Norris, 1988). This strategy assumes that prosodic markings correlate with the beginning of a new word. The specific prosodic cues may vary between languages. In English, the cue is that strong syllables typically mark the beginning of a new word. How might this strategy be involved in the acquisition of segmentation? The idea is that infants have a pre-existing sensitivity to rhythm, which picks up periodic prosodic regularities (Cutler & Butterfield, 1992; Otake, Hatano, Cutler, & Mehler, 1993). Note that this rhythmic regularity is not based on any simple physical feature of the speech input, but is defined in terms of "strong" and "weak" vowels, categories that the child is assumed already to have acquired. This assumption is supported by experiments which appear to show that nine-month-old infants are sensitive to prosodic features of English words (Jusczyk, Cutler, & Redanz, 1993). The child is also assumed to be able to segment the input into syllables before the MSS can be applied.

A simple distributional analysis of stress patterns in the English lexicon confirmed that the overwhelming majority of content words have strong initial stress (Cutler & Carter, 1987). Thus, in English at least, the MSS can provide useful information for segmentation. Whether such regularities are also present in other languages is presently undetermined.

The most well-known distributional approach to segmentation has been developed by Wolff (1975, 1977, 1988; see also Redlich, 1993, and Servan-Schreiber, 1992, for similar approaches). Wolff's model begins with a dictionary of atomic symbols (e.g. phonemes or letters). The co-occurrence statistics for these symbols are collected from the input, and the most frequently co-occurring pair are added to the dictionary as a new symbol. An iterative application of this strategy leads to progressively larger groupings. When run on artificial or simple natural language texts, the boundaries between units tend to respect word boundaries and so provide a constraint for segmentation (see Fig. 3).

However, although the general approach taken by Wolff is very interesting, the method has a number of limitations. For instance, there is no clear-cut way to assess the "goodness" of the results (e.g. the amount of information with respect to word boundaries), although Chi-squared tests reveal that the method does do significantly better than chance (Wolff, 1977). With large (millions of words) corpora the model quickly grinds to a halt, although it produces good results for small (thousands of words) corpora (although a parallel implementation might be able to cope with very large corpora). Another problem with Wolff's account is that it predicts that shorter words should be acquired first. Amongst the shortest words in

FIG. 3. The output of Wolff's MK10 model (Wolff, 1977) applied to a 10,000 letter sample from a book of the Ladybird Reading Series. The tree markers indicate the "chunks" developed by the model and their internal structure. Note that this example was not picked at random, although generally the model does respect word boundaries to a reasonably high degree.

English are the function words (such as *as*, *of*, *or*, etc.), which are known to be acquired (or at least used productively) relatively late. Wolff (1988) suggests that function words may be difficult to learn because having no concrete or perceptually salient referents, their meanings are difficult for the child to acquire. However, he claims that his theory predicts that children possess knowledge of function words from an early stage, even though they do not know what they mean, or how to use them appropriately (see Gerken, Landau, & Remez, 1990, for evidence that this is the case).

Underlying Wolff's distributional method is the general concept that the cognitive system aims to provide a *compressed* representation of the input. The general aim of compression requires that the cognitive system finds and exploits as much of the structure of the input as possible.[9] Recoding the corpus in terms of the units constructed by the model allows it to be represented more efficiently. Brent and Cartwright's (1997) work on segmentation reintroduces these ideas in the form of the minimum description length approach (Rissanen, 1989), and overcomes some of the resource limitations and problems with quantitative assessment of Wolff's model.

A different approach to segmentation is to use predictability as a guide to word boundaries. The assumption here is that regularities within words will be stronger than regularities between words, and hence that the difficulty of predicting the next phoneme should be greatest across word boundaries. This approach involves using distributional information to attempt to predict the next phoneme on the basis of previous phonemes, and using prediction error as an index of the probability that a word boundary has been encountered. This can be done either with a connectionist network or using simple co-occurrence statistics (Cairns, Shillcock, Chater, & Levy, 1994; see also Harrington, Watson, & Cooper, 1988, and also Christiansen, Allen, & Seidenberg, this issue). Both of these approaches demonstrate empirically that this method provides a useful, but only very partial, source of information about word boundaries. Additional cues would be required to provide even an approximate solution to the segmentation problem.

---

[9]The idea that the goal of compression can drive the search for underlying structure in data forms the basis for an important approach to inductive inference, based on Kolmogorov complexity theory (e.g. Li & Vitanyi, 1993), which has been used in statistics (e.g. Rissanen, 1989), machine learning (e.g. Wallace & Boulton, 1968), and psychology (Chater, 1996). This approach is used in some of the work described here, including Ellison's (1992) methods for learning phonological structure, Brent's (1993) work on finding morphemes, and Grünwald's (in press) approach to learning phrase structure. Furthermore, there is a close relationship between the goal of compression and the goal of optimal Bayesian inference, which is also a widely used learning method (e.g. Li & Vitanyi, 1995). Wolff uses an informal notion of compression, although integration with these formal ideas would seem to be feasible. A step in this direction is taken by Redlich (1993), who provides an information-theoretic version of Wolff's technique.

Summary

In summary, the means by which children solve the segmentation problem remain far from clear. For English at least, a number of sources of information are present in the child's input, including single word utterances, prosody, and redundancy (Wolff's approach) or predictability (Cairns et al.'s, 1994 approach). The studies of Saffran and colleagues (Saffran, Newport, & Aslin, 1996; Saffran, Aslin, & Newport, 1996) suggest that both infants and adults can exploit such cues. Assessing the value of distributional cues across languages, and whether and how such cues can be usefully combined to improve segmentation performance remains as future projects (but see Christiansen, Allen, & Seidenberg, this issue). There is also great scope for the relating corpus-based work in this area to the known variation in infants' segmentation strategies, both between individuals, and over time.

Morphology

The Problem

The first problem of acquiring morphology is the identification of the relevant morphological processes in the language. Across languages, these processes are very diverse, including suffixes, prefixes, infixes, circumfixes, ablaut/umlaut, vowel-tier morphemes, tonal morphemes, metatheses, and truncations (Anderson, 1992).

The second problem is relating these morphological processes to semantics, so that the child understands that adding /-ed/ means that the verb takes the past tense, and learns how to apply these processes (in both their regular and irregular forms) to the relevant stems to form new meanings. Presumably, these two problems are not independent.

Development of Morphology

The acquisition of morphology was classically viewed as a three-stage process (Ervin, 1964). Lexical items were held to be initially rote-learned with no distinction between regular and irregular forms, and correct production of both. The second stage was characterised as involving the identification of morphological structure (the first problem described previously) and the occurrence of strong morphological regularities (e.g. the signalling of English past tense by the addition of /-ed/ to the verb stem), the second problem described. This stage was held to be signalled by the child's ability to use morphology productively (as in Berko's, 1958 wug test), in combination with the "over-regularisation" of previously correctly produced forms. For example, the child may say *goed* instead of *went*. The third stage consisted of the correct application of regular and irregular

morphology (Bybee & Slobin, 1982; Kuczaj, 1977). These three stages appear to illustrate U-shaped development (Bever, 1982; Bowerman, 1982): The past tense of a common irregular verb such as *to go* may initially be produced as *went*, then the incorrect *goed*, then finally *went*.

More recent studies of the development of morphology (e.g. Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992) reveal that, rather than going through a second "stage" of predominant over-regularisation, children tend to make over-regularisation errors at a very low (median 2.5% of irregular verbs), but constant rate, between two years and school-age. The frequency of parental production of irregular forms, and the occurrence of similar sounding irregulars lead to lower rates of irregularisation of particular verbs (although similar sounding regulars do not lead to higher rates of over-regularisation).

Possible Sources of Information

The problem of identifying the language's morphemes is closely analogous to the problem of segmenting the speech stream into words,[10] and therefore the range of prosodic, phonetic, acoustic, and other distributional cues that bear on the speech segmentation problem may also be informative here. Syntax is another potential cue to morphology, as morphology predicts a word's syntactic category, and vice versa. A similar two-way relationship holds between morphology and language external factors, such as semantics; meaning is a good predictor of morphology, and may be exploited in its acquisition, just as morphology is used in comprehension to alter meaning, and might serve a useful role in the development of language comprehension. However, the computational research on this problem discussed below has mainly concentrated on distributional cues.

Case Studies

Wolff's (1975, 1977, 1988) approach to the segmenting streams of phonemes, outlined above, also provides a mechanism for finding morphological structure. Much as later units in Wolff's model correspond to words, so the earlier units of which they are constructed may correlate to some degree with morphemes. For instance, units corresponding to common stems, such as *stand*, will develop, as will units for common suffixes, such as /-ing/ and /-s/. In practice, as for segmentation, the degree to which this is actually true (that unit boundaries reflect morphemic boundaries) is difficult to assess. Additionally, in Wolff's model word boundaries are not assumed, and therefore suffixes (especially suffixes such as /-s/) tend to unite with

[10]Although this analogy does not extend to, for instance, tonal languages, such as Mandarin Chinese.

function words, with which they frequently co-occur, rather than the words that they modify; for example, see how the model treats *thinks* in Fig. 3.

Brent (1993) develops a similar approach, based on the minimum description length (MDL) principle in statistics (Rissanen, 1989). MDL provides a means of comparing different accounts of the structure in a data set (in this case, the set of strings of phonemes corresponding to English words). The "goodness" of a particular account is measured by the length of description required to encode the data set via that account. This description has two parts: (1) The description of the postulated structure (here, the inventory of postulated morphemes); and (2) the description of the data in terms of that structure (the original lexicon, encoding in terms of these morphemes).

At a general level, the MDL principle offers a simple way of trading off these two parts of the description, recommending that the postulated structure with the minimum overall description (summing the lengths of the two parts of the description) should be preferred. Intuitively, the rationale for this approach is that any postulated structure should capture regularities in the data set, and thus allow the data to be encoded concisely.

Specifically, Brent's (1993) system commences with an input lexicon of English words, and assumes that each lexical entry consists of a stem and suffix (note that the use of a lexicon assumes a solution to the segmentation problem). The system attempts to find the shortest overall description of the lexicon (the data) in terms of these stems and suffixes (the postulated structure).

A variation on this approach is to include the syntactic category of each word as a property of the lexicon, and to use the correspondence between a word's syntactic category (which is assumed to be known from the lexicon) and its possible suffixes as a constraint on identifying the suffixes; as each description of a suffix must be indexed with a single syntactic category, to which it can be applied (i.e. the /-s/ suffix must be represented once for verbs, and once for nouns) Brent's algorithm favours suffixes which apply to many members of a syntactic category, as real regular suffixes do (in English at least).

The performance achieved by Brent's approach is impressive. Using a 1000 word lexicon of highly frequent words occurring in newspaper text (from the *Wall Street Journal*), the best description of the lexicon utilised only the following suffixes: /-age/, /-al/, /-ed/, /-ing/, /-ion/, /-ity/, /-ly/, /-ment/, /-nce/, and /-s/. On the basis of form alone, the system managed to identify many English morphemic suffixes, while avoiding nonmorphemes, such as word final /-sk/ and /-ld/, which have a relatively high frequency in English. Generally, with lexicons of 1000 words or more, a minimum of 70% of the identified suffixes were "perfect" morphemes such as /-ed/, and approximately 80–95% of the identified suffixes were linguistically

meaningful, such as /-mental/ (as in *governmental*). This finding demonstrates that this single source of information (which Brent refers to as "form") can be highly informative concerning English regular morphology. Although the frequent words in the *Wall Street Journal* are not typical of either written or spoken language, it seems likely that this method will enjoy a similar level of success with more realistic input, although this has not been empirically confirmed. However, one problem that the model does not address is the identification of irregular morphology. It is unclear whether this can be accounted for on the basis of form and/or syntactic category membership.

We now turn to the problem of learning how to apply and appropriately generalise (as opposed to identifying) both regular and irregular morphological mappings. Connectionist research in this area has focused on the English past tense, the extent to which connectionist models can learn both regular and irregular mappings, and how their over-regularisation errors coincide with pattern observed in children, on the basis of phonemic and frequency information. In order to permit such investigations, some syntax and semantics are taken as given: Models are specific to particular syntactic classes and parts of speech and particular transformations (e.g. the past tense).

Traditionally, the developmental data on over-regularisation have been taken as evidence for a "dual route" model of past tense formation, with over-regularisation being explained in terms of a default rule-based route for dealing with regular morphological forms. In adults, the use of this rule-based route is supposedly blocked for irregulars, with an associative memory supplying the correct irregular form.

Rumelhart and McClelland (1986) considered the morphological mapping between the present and past tenses of English verbs. Both present and past tenses were represented using a complex representation based on phonetic "Wickelfeatures." A single layer of learnable connections was trained to map the present tense to the past tense from a sample of English verbs. Rumelhart and McClelland found that this single route connectionist network appeared to be able to deal both regular and irregular forms, and under some conditions appeared to demonstrate U-shaped learning. Thus, it appears that a single-route connectionist model can explain both the developmental profile and adult performance in English past tense formation.

However, there has been a vigorous debate concerning the merits and limitations of the rule-based and connectionist accounts of past tense acquisition. Here we focus on factors relating to the distributional properties of language.

Although Rumelhart and McClelland's (1986) model *can* demonstrate U-shaped learning, this appears to require a careful manipulation of the

input at each of the three developmental stages. In particular, the number of verbs, and proportion of regular and irregular verbs in the training set was altered significantly. From a psychological point of view, this manipulation is only valid if it corresponds to the input received by the child. Pinker and Prince (1988) pointed out that these manipulations are not supported by the empirical evidence concerning the language received by the child—they constitute unwarranted distributional assumptions.

Plunkett and Marchman (1993; see also Plunkett & Marchman, 1991) trained a connectionist network with hidden units (i.e. with two layers of learnable connections, rather than one) with a gradually increasing vocabulary of up to 500 verbs. Although the phonology of these verbs was artificial (but consistent with the phonotactics of English), to avoid the complexities associated with Wickelfeature representations, they were chosen to be representative of type and token frequences of English verbs.[11]

With this more realistic input, the network exhibited a similar pattern of performance to children. A period when all verbs were correctly marked for tense, followed by a period of very low, roughly constant rate of over-regularisation. The kinds of error made by the network were similar to those made by children. For example, the network was sensitive to subregularities within the irregular verbs (e.g. *sleep* → *slept*, *keep* → *kept*) and showed some tendency to overgeneralise these subregularities.

Prasada and Pinker (1993) have argued that the success of connectionist models of the acquisition of English past tense morphology may rely on the particular distributional statistics of English. In English, there are many regular /-ed/ verbs, which individually have low token frequences, allowing a connectionist network to find a very general default category. In the case of irregular verbs, type frequency is relatively low, but token frequency is typically high, allowing the network to override this general default. Prasada and Pinker argued that languages where the default regular mapping has both low type and token frequency cannot be learned by a connectionist network. The Arabic plural system (Forrester & Plunkett, 1994) and the putative default /-s/ inflection of plural nouns in German (Clahsen, Rothweiler, Woest, & Marcus, 1993) appear to provide examples of such "minority default mappings".

Forrester and Plunkett (1994) showed that under some circumstances minority default mappings can be learned by connectionist networks. In their artificial training set, regular items were scattered throughout the input space. By contrast, irregular items were grouped into subregions of the

[11]In this context, *type* frequency refers to the number of different verbs within in each class: The regular forms, where past tense is constructed by adding a suffix, and irregulars whose past tenses are formed by a vowel change, by the identity mapping, or by an arbitrary mapping. Token frequency refers to the frequency of individual lexical items.

space, each corresponding to a distinct subregularity. The network was able to take advantage of the distribution of regular and irregular items in the input space to appropriately capture the patterns of the irregular items, while correctly using the minority regular rule as a default. Forrester and Plunkett (1994) argued that the distributional structure of their training set mirrored the distribution in the Arabic plural system. Here, irregular plurals follow a range of subregularities, determined by whether they match particular phonological templates, whereas the regular applies to diverse phonological forms. Detailed analysis of the lexicon of Arabic is required to assess the extent to which Forrester and Plunkett's idealisation is valid.

Nakisa and Hahn (1996) have analysed the distributional properties of the German plural system. Recent dual-route accounts of morphology have suggested that the minority German regular plural form /-s/ can be captured by the joint operation of a default rule and a pattern associator (Marcus, Brinkmann, Clahsen, Weise, & Pinker, 1995). Specifically, the default "add /-s/" rule is held to be applied whenever a particular word cannot be found in the lexicon of irregulars. This lexicon is held to include a phonologically based associative memory, which allows the model to account for irregularisation of novel irregulars observed in German (and Arabic).

Nakisa and Hahn (1996) investigated whether simple single-route associative models (the nearest neighbour algorithm, the Generalised Context Model—Nosofsky, 1990; and a simple feed-forward connectionist net, with one hidden layer) could learn the German plural system, and generalise appropriately to novel regular and irregular nouns. The associative models' task was to predict to which of 15 different plural types the input stem belonged. The inputs to the learning mechanisms were phonetic representations of approximately 4000 German nouns taken from the CELEX database (token frequency was ignored). The three associative models scored, respectively, 71%, 75% and 84% of classifications correct on a test set of 4000 previously unseen test nouns.

Nakisa and Hahn (1996) also simulated the Marcus et al. (1995) model, by assuming that any test word that is not close to a training word, according to the associative model (i.e. for which the lexical memory fails) will be dealt with by a default "add /-s/" rule. The associative models were trained on the irregular nouns, and the models were tested as before. Nakisa and Hahn found that, in all three cases, the presence of the rule led to a decrement in performance. Generally, the higher the threshold for memory failure (the more similar a test item had to be to a training item to be irregularised via the associative memory), the greater the decrement in performance.

The use of a default rule could only have improved performance if regular nouns occupied very sparsely populated regions of phonemic space. In real German, Nakisa and Hahn's (1996) findings demonstrate that this is not the

case. The extension of Nakisa and Hahn's findings to the production of the plural form (instead of merely indicating the plural type), and to more realistic input (e.g. taking account of token frequency) remains to be performed.

This is an excellent illustration of both the value of distributional models as accounts of acquisition, and of how such models can be used to illuminate the potential role, or lack of role, of proposed pieces of innate knowledge, such as the provision of apparatus for learning "default" rules.

Summary

Morphology has proved to be a fruitful ground for the application of distributional learning methods, both connectionist and statistical, and this research has led to important findings concerning both the distributional properties of language, and the abilities of connectionist learning methods. The development or morphology in other languages, the potential interaction between morphological development and the acquisition of word classes, and other aspects of language are all promising areas for future research.

Word Classes

The Problem

The acquisition of syntactic categories such as *noun* and *verb* is a central problem in language acquisition. There are two related problems: (a) Discovering that there are different classes, or categories of words; (b) discovering which words are members of which syntactic category. These problems are likely to be highly interrelated, although the existence of pre-existing innate categories is often proposed as a solution to the first (e.g. Pinker, 1984).

Even for theorists who assume that the child possesses an innate universal grammar, and innate categories, identifying the syntactic category of words must primarily be a matter of *learning*, because the phonological strings associated with words of the language are clearly not universal. The universal grammatical features of language can only be mapped on to the specific surface appearance of a particular natural language once the identification of words with syntactic categories has been made. Of course, once some identifications have been successfully made, it may be possible to use prior grammatical knowledge to facilitate further identifications. To solve the problem of establishing initial linguistic categories, however, it seems that the contribution of innate knowledge must be relatively slight.

Development of Word Classes

Assessing the child's knowledge of word classes is difficult (see Pine & Martindale, 1996, for discussion), as appropriate usages of a particular word do not necessarily indicate that the infant possesses word-class knowledge (for instance a word may be used within a rote-learned utterance), and similarly, incorrect usages do not necessarily indicate that the infant lacks word-class knowledge (for instance, inappropriate inflections of a particular word may be due to a lack of morphological knowledge).

By the age of three or four, children's spontaneous productive use of morphological inflection (Ervin, 1964) and verb argument structure (Bowerman, 1982), and ability to correctly inflect nonsense words in a laboratory setting (Berko, 1958), indicate that they can both comprehend and use adult word classes. However, in part for the reasons outlined earlier, relatively little concrete data is available for the development of word classes prior to this age.

Although infants' earliest multiword utterances (just prior to 18 months) do show some syntactic consistency, generally consisting of a nominal and action words or modifiers (Bloom, 1970), this may be due to the infant's word order preferences (Braine, 1976; Tomasello, 1992), or to constraints imposed by semantics or communicative intent (Sachs, 1976), rather than knowledge of grammatical categories per se. Attributions of early or innate knowledge of syntactic categories (e.g. Pinker, 1984; Valian, 1986) have been challenged on the grounds that early syntactic knowledge is in fact highly restricted to specific lexical items (e.g. Braine, 1988; Pine & Martindale, 1996).

The most recent and detailed picture of the time course of syntactic category acquisition (for English at least) is provided by Tomasello's (1992) diary study. Tomasello found evidence of an early (prior to 18 months) noun category, with productive use of the past tense and possessives, pronoun substitution, and flexible use of newly learned nouns. However, a full verb category did not develop during the period of development that Tomasello studied (up to 24 months). This finding is supported by laboratory studies. A majority of two-year-old children exposed to novel nouns used them productively in novel argument roles and with plural morphology (Tomasello & Olguin, 1993). However, children of similar age exposed to novel verbs did not show any generalisation of verb arguments or morphology (Olguin & Tomasello, 1993).

Possible Sources of Information

Both language external and language internal sources of information have been proposed to influence the learning of syntactic categories. One language external approach, "semantic bootstrapping" (see Grimshaw,

1981; Pinker, 1984; Schlesinger, 1981, 1988, for a variety of different approaches), assumes that there is a correlation between linguistic categories (in particular *noun* and *verb*) and the child's perception of the environment (in terms of objects and actions). It is proposed that this provides the learner's initial means of "breaking in" to the system of syntactic categories. A somewhat different extralinguistic approach is the "social/interaction" model (see Bruner, 1975; Nelson, 1977; Snow, 1972, 1988, for a range of views). Here, pragmatic factors such as *force of request*, *object under consideration*, and *location of object* are assumed to be correlated with syntactic categories such as *verb* and *noun* and *preposition* (e.g. Ninio & Snow, 1988).

While extralinguistic factors may be very important, it is difficult to quantify the strength of the correlations on which they rely, for the reasons already outlined. Additionally, the developmental evidence is not consistent with the sole use of semantic or pragmatic relationships in syntactic category acquisition. For example, children's application of morphological inflection (e.g. the English -ed past tense suffix) is not initially restricted to verbs which denote actions (Maratsos, 1988; Maratsos & Chalkley, 1980), and children also appear to be able to acquire distinctions that have no semantic basis at a very early age (e.g. linguistic gender in French, Karmiloff-Smith, 1979; Hebrew, Levy, 1983; and Spanish, Perez-Pereira, 1991). Clearly, whatever role language external sources of information play in the acquisition of word classes, they are not the whole story.

There are also a range of language internal factors, four of which have been discussed in the literature. First, distributional analysis of morphological variations across lexical items. Maratsos (1988; Maratsos & Chalkley, 1980) notes that in English, word roots that take the suffix /-ed/ typically take the suffix /-s/ and are verbs. Words that take the suffix /-s/, but not the suffix /-ed/ are typically count-nouns. Patterns of correlation between simple properties of word roots might therefore potentially be used to infer proto-word classes which can later be refined to the adult word classes. Empirical assessment of the usefulness of these relationships in finding word classes from real corpora appears to be tractable, although, to our knowledge, has not yet been attempted.

A second source of language internal information is the regularities between the phonology of words and their syntactic categories (Kelly, 1992). In English disyllabic words, for example, nouns tend have stress on the initial syllable, whereas verbs have final syllable stress (e.g. Liberman & Prince, 1977); English polysyllabic words are predominantly nouns (Cassidy & Kelly, 1991); English open-class words are generally stressed more strongly than closed-class words (Gleitman, et al., 1988). These and many other cues, both in English and across languages have been subject to very little empirical investigation.

The third source of information is prosody. Morgan and Newport (1981) and Hirsh-Pasek, Kemler-Nelson, Jusczyk, Wright, and Druss (1987) propose that learners exploit the mutual predictability between the syntactic phrasing of a sentence, and the way it is said (i.e. its *prosodic phrasing*). Consequently, if the child takes note of how something is said, he or she has information about the "hidden" syntactic phrasing of the sentence. This information might provide clues about the syntactic properties of words in the input, and thereby constraints on their possible syntactic categories. The strength of this prosody–syntax relationship has not been assessed using corpora of natural speech. As mentioned earlier, large corpora marked for prosody are not currently available.

Finally, the fourth source of information, which we consider later, is distributional analysis at the level of lexical items.

Case Studies

Various authors have proposed that a variant upon one of the main legacies of structural linguistics provides a valuable clue to syntactic category (e.g. Brill, Magerman, Marcus, & Santorini, 1990; Finch & Chater, 1991, 1993; Grünwald, 1996; Hughes & Atwell, 1994; Kiss, 1973; Marcus, 1991; Rosenfeld, Huang, & Schneider, 1969; Scholtes, 1991a, 1991b; Schütze, 1993). The "distributional test" (e.g. Radford, 1988) is based on the observation that if all occurrences of word A can be replaced by word B, without loss of syntactic well-formedness, then they share the same syntactic category. For example, *dog* can be substituted freely for *cat*, in phrases such as *the cat sat on the mat*, *nine out of ten cats say* . . ., indicating that these items have the same category. By contrast, *purr* cannot be substituted in the vast majority of phrases containing *cat*, without giving rise to ungrammatical phrases such as **the purr sat on the mat*, **nine out of ten purrs say* . . . .

In linguistics, the distributional test involves generating possible contexts for words and consulting native speakers concerning whether the words of interest can legitimately occur in these contexts. But, in investigating the potential contribution of distributional learning mechanisms in language acquisition, we should, in the first instance, explore methods that rely purely on exposure to a corpus of language.

The most widely used approach is to collect co-occurrence statistics concerning the context words adjacent to the "target" word of interest, throughout a corpus, such as the statistics for the corpus *to be or not to be* described earlier. As words in the same syntactic category will tend to occur adjacent to the same words, then co-occurrence statistics contain a potentially useful clue to syntactic category. Syntactic categories can be postulated by grouping together words with similar co-occurrence statistics.

We have explored one implementation of these ideas (Finch & Chater, 1991, 1992, 1994; Finch, Chater, & Redington, 1995; Redington, Chater, & Finch, 1993, in press) in which the context was defined as the two words before and after each target word. For each target word, vectors representing the co-occurrence statistics for these positions were constructed. The similarity of distribution between target words was calculated using Spearman's rank correlation, and words were grouped together by hierarchical cluster analysis. Typically the most frequent 1000 or 2000 words in the corpus might be used as the target words (these words will account for the bulk of the corpus, typically 75–90%).

This approach does not partition words into distinct groups corresponding to the syntactic categories. Rather, it produces a hierarchical tree, or dendrogram, which to some extent reflects the syntactic categories of words. Figure 4 shows the high-level structure of such a dendrogram, which was the result of analysing approximately 2.5 million words of transcribed adult speech taken from the CHILDES corpus (MacWhinney & Snow, 1985). The speech in the corpus was not guaranteed to be child directed, but was largely recorded in North American domestic settings, in the presence of young

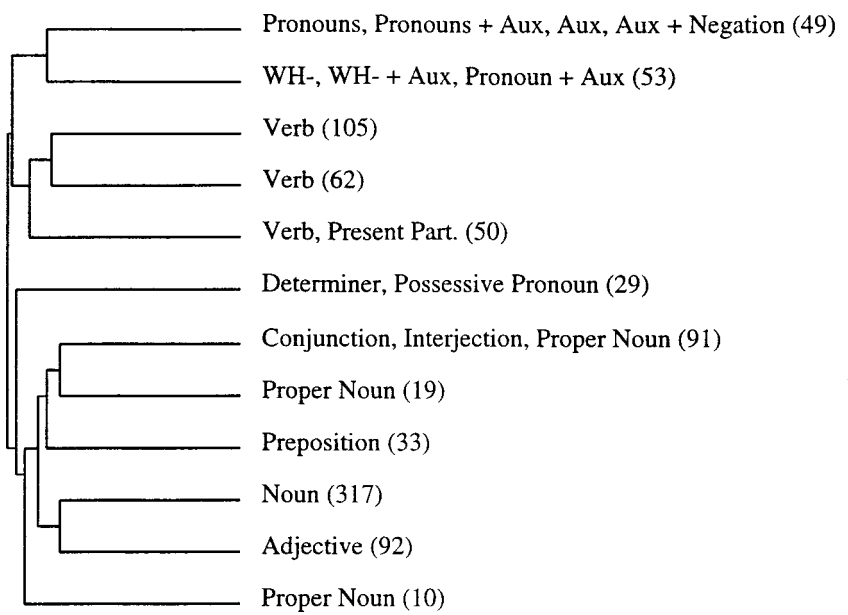

FIG. 4. A dendrogram resulting from a word level analysis of the distributed statistics of the CHILDES corpus. The dendrogram has been truncated at a chosen level of similarity, and the resulting discrete clusters labelled by hand with the syntactic categories to which they correspond. The number of items in each cluster is shown in parentheses. Only clusters with 10 or more members are shown here.

infants, and forms a good approximation to the language to which a child might be exposed.

Figures 5 and 6 show examples of the structure of the dendrogram, and its relation to syntactic category at a very fine level.

The goodness of the classification can be quantitatively assessed in terms of accuracy and completeness. Accuracy is defined as the proportion of pairs of words that are grouped together by the method, and that belong to the same syntactic category. Completeness is defined as the proportion of words that belong to the same syntactic category, and that are grouped together by the method. A canonical syntactic category for each word was derived from its most common usage according to the CELEX database. At all levels of similarity (i.e. from the gross scale of Fig. 4 to the fine scale of Figs. 5 and 6), the method provided more information about syntactic category relationships than would be expected by chance. At the level corresponding to Fig. 4, the method's accuracy was 0.72, and completeness was 0.47. The mean accuracy and completeness of 1000 random simulations (where the number and size of clusters was held constant, and words were randomly assigned to clusters) were 0.27 and 0.17 respectively.

Similar analyses of written corpora have demonstrated that this method can provide information concerning syntactic category membership for a range of languages; French, German, and Mandarin Chinese (Redington, Chater, Huang, Chang, Finch, & Chen, 1995). This work provides a feasibility proof for the validity of distributional information in the acquisition of syntactic categories.

This method can also be implemented in a connectionist network (Finch & Chater, 1992). The architecture and functioning of the network are highly similar to the simple example shown in Fig. 2. Figure 7 shows how this architecture can be extended to collect statistics over many context positions. As before, each unit in the "current word" and "context word" layers represents a single word, and is activated when that word occurs in the appropriate position in the input (i.e. as the current word, or in one of the context positions). The weights between these two layers are strengthened by a simple Hebbian learning rule, and represent the co-occurrence statistics between the current and context words. In order to obtain a classification, the units of the "current word" layer are activated in turn, and the resulting pattern of activation on the "context word" layer serves as the input to a process of Kohonen clustering (Kohonen, 1982). This implies the use of Euclidean distance (rather than rank correlation) as an effective measure of similarity between the patterns of activation for each word. This measure of similarity effectively "comes for free". Rank correlation (which is generally more effective with this method for noisy natural language corpora) cannot be implemented straightforwardly in a connectionist net. Therefore, the connectionist architecture places a constraint on the statistical mechanisms

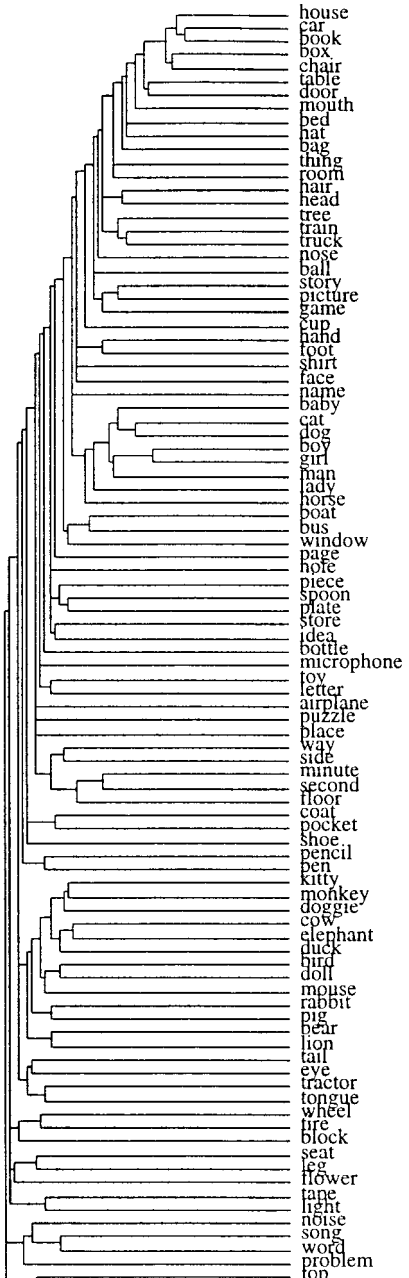

FIG. 5. A noun branch, from the CHILDES analysis.

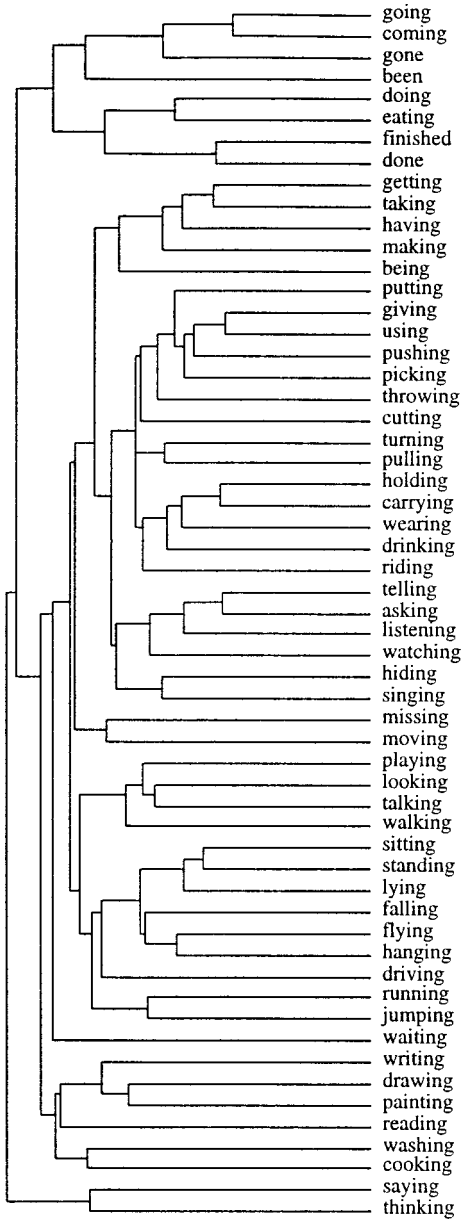

FIG. 6. A verb cluster (featuring the progressive participle form) from the CHILDES analysis.

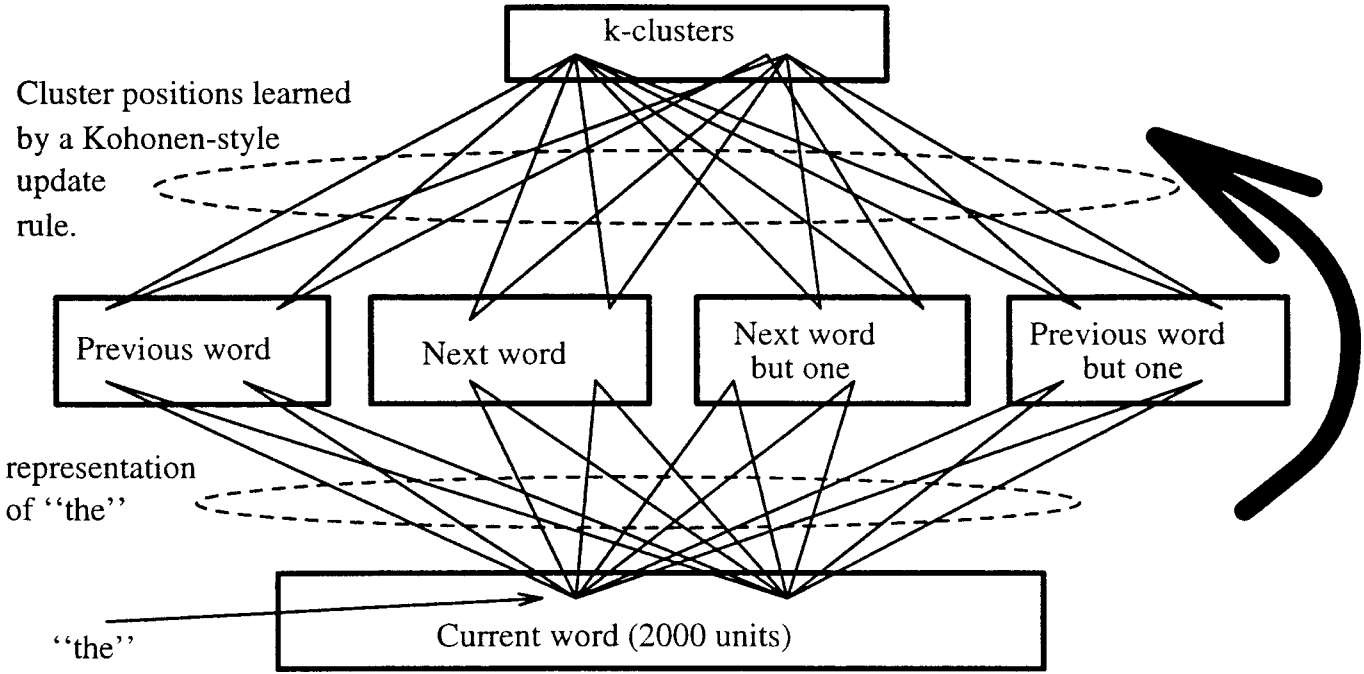

FIG. 7. The architecture of the connectionist implementation of Finch and Chater's (1992) method. For clarity, only a few sample connections are shown, although there is in fact full connectivity between layers.

that are available. None the less, the discrete classifications produced by the net reflect the syntactic relationships amongst the target words to an impressively high degree.

An important feature of this network is that it can cope with very large amounts of data (e.g. corpora of 40 million words). This is in contrast to many connectionist models, which are not able to scale up from small artificial domains. A second notable feature of this network is that it is an example of a network which was specifically designed to implement a particular statistical learning mechanism, or to approximate it as closely as possible (e.g. the substitution of Euclidean distance for rank correlation). This contrasts against the usual procedure in connectionist research, where the network is first designed and trained, and the statistical structure which it exploits is then determined by a post hoc analysis.

There are a number of aspects of syntactic categories which this approach does not address. The first of these is the phenomenon of syntactic ambiguity: Many words are members of more then one syntactic category. For example, *fire* can be a noun or a verb, and, as part of compound nouns (e.g. *fire engine*), can play an adjective-like role. Because this method averages contexts over all occurrences of a word, a word cannot receive more than one classification. There are some indications that the method categorises ambiguous noun/verbs together to some extent, but this obviously depends on their usage in the corpus. In general, the method is unable to deal with syntactic ambiguity. However, this is not necessarily a problem for distributional methods in general. It may be possible to identify words with more than one syntactic category by observing that the distribution of contexts in which they occur is bimodal, or by forming a rough initial classification (using the method described earlier), and then refining this model to include multiple classifications.

The second problem is that the measure of syntactic similarity used by this approach is continuous, whereas syntactic categories are generally conceived of as discrete. This is not a fatal problem, as the usefulness of this information (i.e. its association with canonical discrete categories) has been shown. However, an account of how this information is combined with other sources to derive discrete, adult categories would be required by a complete account of syntactic category development.

A further problem concerns the application of the method to so-called free word order languages such as Turkish. It seems probable that the method will be less effective with these languages. However, this is a matter for empirical investigation; although such languages have no prescribed word order constraints, it may well be the case that some word orders are more common than others, and that these regularities can be exploited by this method.

Summary

To summarise, distributional information at the word level has been shown to be highly informative of syntactic category, for corpora both of written text, and of speech, contrary to claims such as those of Pinker (1984). Indeed, in this case distributional relationships appear to provide such a powerful cue that it would be odd if children did not exploit them in some form. Many other sources of distributional information, such as prosody and phonology, and especially morphology, have not as yet been empirically investigated. The interaction of distributional information at the word level, and morphological information, and how this constrains the development of both morphology and the identification of syntactic categories, appears to be a particularly fruitful field for further research.

Phrase Structure

The Problem

One of the most controversial problems in language acquisition concerns the acquisition of phrase structure grammar. Given only positive examples, some of which are incorrect, the child must acquire a grammar, which accepts as grammatical all, and only grammatical sentences of the language. This has been held up as a paradigm case for the dominance of innate knowledge; according to nativist accounts, the child's input serves only to "trigger" the selection of a particular variant of the child's innate universal grammar (Chomsky, 1980).

Development of Phrase Structure

At around 18 months, children move from single-word utterances and rote-learned phrases to the production of combinations of words in two-word phrases. This development coincides with the naming explosion (Bates et al., 1988). This "telegraphic speech" has a number of properties. Typical two-word phrases generally express simple semantic relationships, often consisting of a nominal, together with an action word or modifier (Bloom, 1970). Function words (such as *the* and *and*) and morphemes (such as /-ed/) are generally absent, although word order errors are very rare (Bloom, 1970; Brown & Bellugi, 1964).

Children do not pass through a three- or four-word stage. Rather, from around 2½ years, the average length of their utterances gradually increases. Although these longer sentences can initially be characterised as simple expansions of children's two-word phrases (Bloom, 1970), after this point there is a consistent increase in the appropriate use of function words and more complex syntactic constructions, such as questions and relative clauses.

Possible Sources of Information

There are many possible sources of information that might aid in the development of grammar. Language external sources include semantics and pragmatics. For instance, the "actor" referred to by an utterance will usually occur in a noun phrase, and there will (hopefully) be many other associations between syntax and semantics. Additionally, the full range of language internal sources may also be exploited. As has already been discussed, the mutual predictability between an utterance's "prosodic contours" and its grammatical structure might be exploited to provide information about the latter (Hirsh-Pasek et al., 1987; Morgan & Newport, 1981). Similarly, just as phonology and morphology may provide constraints on a word's syntactic category, so they may be informative about an utterance's syntactic structure. Finally, there is distributional information at the level of words and word classes, which we consider later.

Case Studies

Within the fields of computational linguistics and machine learning, there is a vast array of research on the problem of discovering a grammar, from a corpus of the language (e.g. Langley, 1987; Solomonoff, 1964a, 1964b). However, very little of this work is psychologically motivated.

The example that we give here has much more modest aims than the discovery of a complete grammar, from scratch. It is an extension of Finch and Chater's (1991, 1992) model for word classes, but rather than using distributional information to constrain the classification of words into categories such as *noun* or *verb*, it attempts to constrain the classification of short sequences of words into categories such a *noun phrase*, or *verb phrase*.

Rather than classify sequences of words (such as *the brown dog*) the method considers sequences of words in terms of their categories, using the results of the word-level analysis already described (so that, for instance *the brown dog* and *a black cat* might be equivalent when described in terms of the categories). Context is also measured in terms of these categories, rather than the actual words themselves.

The results here are from an analysis where sequences of one to three words were considered, from an analysis of 40 million words of English USENET news. Using a corpus of this size ensures that the syntactic relevance of the method's output is obviously apparent (as opposed to the 2.5 million word CHILDES corpus).

Figure 8 shows a few sample clusters, specifically chosen for illustrative purposes. The clusters derived from the word-class analysis have been hand-labelled with the syntactic category to which they correspond. It should be clear that our labelling of the clusters of phrases as noun, verb, and prepositional phrases does to some extent reflect aspects of linguistic

Noun Phrase
Det Noun, Det Adjective Noun, Det Noun Noun, Determiner Verb/Noun, Det Adjective Verb/Noun, Det Inf, Det Verb/Noun Noun, Det Noun Verb/Noun, Det Inf Noun, Det ing Noun, Det PastPpl Noun, Det Det Noun, Det Adjective Noun, Det Adjective Inf, Det Adjective Verb/Noun, Det ing, Det Noun Adjective, Det Place Noun, Det Adjective QuantProNP

Verb Phrase
Inf ProObj, Inf ProObj Noun, Inf Det Noun, Inf Det Verb/Noun, Inf Det Inf, Verb/Noun Det Noun, Verb/Noun ProObj, Inf ProObj Prep/Adv, Inf Quant/NP, Inf QuantProNP, Inf ProObj Adjective, Inf Countries, Inf Noun, Inf Adjective Noun, Inf Noun Noun, Inf PastPpl, PastPpl PastPpl, PastPpl Adjective

Prepositional Phrase
Prep Noun, Prep Det Noun, Prep Adjective Noun, Prep Det Verb/Noun, Prep Inf, Prep Det Inf, Prep Adjective Noun, Prep Verb/Noun, Prep Adjective, Prep QuantProNP, Prep ProObj Noun, Prep Conj &WH Noun, Prep Noun Noun, Prep QuantProNP Noun

FIG. 8. "Noun Phrase", "Verb Phrase", and "Prepositional Phrase" clusters from the longer sequences.

structure. Note that an ambiguous "verb/noun" cluster was present in the word-level analysis. Within the "noun phrase" cluster in Fig. 8, both the ambiguous class and nonfinite verbs act in a similar manner to nouns when preceded by a determiner. Within the "verb phrase" cluster, the ambiguous class acts in a similar manner to nonfinite verbs. This suggests that distributional information at the phrase level might be used overcome the problem of syntactic ambiguity.

Having formed classes of short sequences, these can be used as the basis for the classification of longer sequences. In the analysis described next, pairs of short sequences were considered (i.e. from two to six words in length), and a similar process was performed. The results shown are from an analysis of 10 million words of USENET news, and again clearly illustrate that the method captures linguistically relevant structure.

Here, we simply show examples randomly drawn from a few sample clusters, expressed terms of phrases in the original corpus, rather than in terms of sequences of syntactic labels. Figure 9 shows sequences from a cluster corresponding to "proto-sentences"—phrases that could reasonably be thought to be a candidate sentence if parsed out of context, such as *the*

Proto-sentences
what is a context, it might be a good idea, that's a different story, you see a problem, you will also receive a copy, that there was an error, the world isn't perfect, you start out, it really was lost, it does have a german title, we are looking, the government won't let them, it would be a good idea, i did notice it, i can get the book, you can actually see it, you need more information, this information is available, they were picked up, we could hold some events, you carry them, i just received my copy, you were found out, i just don't want it

Prepositional Phrases
out, out of this state, into a form, to those questions, of language and information, in the appropriate box, to a function, in school french, it out, by the way, of the terms, on its argument structure, of a variable, with this, of program performance, on the basis, of the file, in other words, to the development, in general, of such a news group, for this, on usenet, in areas of political rights, on the basis of religious law, up, to such a rule

Noun Phrases
the reason, such questions, a moral law, the problem with it, a more accurate memory, the real number system, the article, it, many cases the option, a discussion on this, the child of a woman, a problem here, a gun, the six day war, his behaviour during his life, his ideas about the rights, the four letter name, it for no reason, a piece of paper, someone at the post, some sources for your last statement

FIG. 9. Random selections of sequences of words from categories corresponding to "Proto-sentences", "Prepositional phrases", and "Noun phrases".

man ate would be a proto-sentence, even if it occurred in the string *the man ate the apple*. Additionally, because of ellipsis, NP movement, and the like, many sequences may be analysed as sentences that do not themselves stand alone as sentences.

Figure 9 also shows examples drawn from a cluster that largely corresponds to prepositional phrases, and to the first part of the prepositional phrase, including the head noun of the rest of prepositional phrase. Also shown is a cluster corresponding to noun phrases including the short noun phrases described before, and more complicated constructions such as Det $\overline{\text{N}}$ PP, as in *the child of a woman* or *a piece of paper*, but not

sequences such as *the man who I saw yesterday*, possibly because these are typically too long to be considered.

There are some major problems with this approach. Primarily, it is very difficult to assess the results empirically, as this would require a canonical classification of short sequences, which is not available in practice. Additionally, the method also appears to capture some structure that is not as linguistically relevant as, for instance, being a noun phrase, but nevertheless has some reasonable linguistic interpretation (e.g. the proto-sentences shown in Fig. 9). It is unclear how to make a principled assessment of the value of such structure.

Without a quantitative measure of success, it is difficult to assess the results of this method with the typically smaller corpora of transcribed speech, or to compare its applicability across languages. Nevertheless, even on the basis of a post hoc, qualitative assessment, it seems clear that this method is capable of capturing relevant linguistic structure. Developing measures of the success of such methods, and relating these results to the developmental phenomenon remain as problems for future research.

Summary

The acquisition of grammatical structure is an area where the potential contribution of distributional information is essentially unknown. Initial psychological work in this area, and more technically sophisticated work from computational linguistics, suggests that distributional learning mechanisms can provide some useful information regarding grammatical structure. Particularly important research goals here include finding methods for qualitatively assessing the success of such methods, and investigating how information from multiple sources (including multiple distributional sources) can be integrated successfully (as individual cues are likely to be of relatively limited value in this complex domain).

Lexical Semantics

The Problem

Acquiring lexical semantics involves identifying the meanings of particular words. Even for concrete nouns, this problem is complicated by the difficulty of detecting which part of the physical environment a speaker is referring to. Even given that this can be ascertained, it may still remain unclear whether the term used by the speaker refers to a particular object, a part of that object, or a class of objects (as illustrated by Quine's, 1960 famous discussion of "Gavagai"). For abstract nouns, and other words which have no concrete referents, these difficulties are compounded further.

Development of Lexical Semantics

As with many language acquisition processes, the acquisition of word meanings appears to follow a stage-like progression, with the vocabulary spurt at 18 months marking the boundary between stages. When acquiring their first words, children appear to interpret word meaning in a highly restrictive fashion (that is, they underextend the possible meanings). For instance, Allison initially used *car* only when observing cars from a particular location (Bloom, 1973). Another common "error" at this stage is overextension, when children appear to associate a word with many instances; for instance, *clock* might be used to refer to clocks, dials, timers, bracelets, etc. (Rescorla, 1980). This pattern of under- and overextension is classically interpreted as evidence that first word meanings are acquired on the basis of association between a word and the context, or particular features of the context in which it was learnt (e.g. Bloom, 1973; Schlesinger, 1982).

By the time of the vocabulary spurt, under- and overextension have diminished considerably, and children's interpretation of word meanings appears to be much less bound to specific experiences, and more in line with the concepts and categories used by adult speakers (e.g. Bates, Benigni, Bretherton, Camaioni, & Volterra, 1979; Bloom, 1973).

Possible Sources of Information

The primary sources of information for the development of lexical semantics are presumably language-external. Relationships between the physical, and especially the social, environment of the child are likely to play a major role in the development of lexical semantic knowledge.

However, it also seems plausible that language-internal information can be used to constrain the identification possible meaning of words. For instance, just as semantics might constrain the identity of a word's syntactic category (words referring to concrete objects are likely to be nouns) so knowing a word's syntactic category provides some constraint on its meaning; knowing that a word is a noun, perhaps because it occurs in a particular set of local contexts, generally implies that it will refer to a concrete object or an abstract concept, rather than an action or process (Brown, 1957).

Because there are potentially informative relationships between aspects of language at all levels, this means that even relatively low level properties of language, such as morphology and phonology, might provide some constraints on lexical semantics.

Within the language acquisition literature, Gleitman (1990) has proposed that syntax is a potentially powerful cue for the acquisition of meaning. Gleitman notes a number of problems for the proposal that word meanings

can be acquired by mapping from words to the world. For instance, blind children interpret *look* and *see* as referring to manual exploration, even though their mothers often use *see* when the object referred to is out of the child's reach (Landau & Gleitman, 1985); verbs such as *chase* and *flee* and *win* and *beat* represent both events and the perspective taken by the speaker, and it is not clear how an observer could determine the latter; words such as *think*, *guess*, *know*, etc., have no clearly observable physical properties or correlates.

Gleitman (1990) proposes that the syntactic context in which a word appears can provide information about its meaning. For example, Landau and Gleitman (1985) show that blind infants' mothers typically use *look* and *see* within a restricted set of syntactic constructions. Detailed analysis revealed that different constructions were used when the object referred to was within the child's reach from when no object was referred to or nearby (e.g. *Look what you're doing*). According to Gleitman, in order to exploit these correspondences between syntax and the world, the child must possess the ability to parse utterances, and must also possess some notion of the semantic value of a particular syntactic construction (e.g. the relationship between verb argument structure and verb semantics). The child must either acquire the requisite abilities and knowledge at a very early stage, or they must form part of the child's innate language learning apparatus.

We will now consider a number of methods based on the much simpler property of co-occurrence statistics. These methods require no a priori knowledge of syntax or of the correspondence between syntax and semantics. Although it might seem far-fetched that such simple properties could constrain language external aspects of linguistic knowledge, we will see that in practice they may be highly informative.

Case Studies

The first method we describe for revealing words' semantic properties is simply the method described earlier for identifying a word's syntactic category. As already noted, within the dendrogram produced by this method, words of the same syntactic category tend to be clustered closer together than expected by chance. As we have already discussed, knowing a words' syntactic category places a (very limited) degree of constraint on its meaning or referent.

However, close examination of such dendrograms reveals that *within* clusters of words sharing the same syntactic category, the clustering of words sometimes appears to reflect semantic relationships. Figure 10 shows an example of such a subcluster, taken from the analysis of the adult speech in the CHILDES corpus. This is the clearest and best example of semantic relatedness from that analysis.

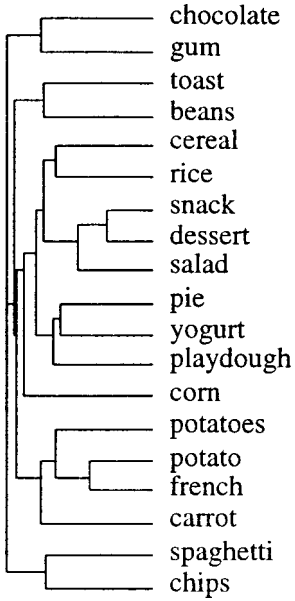

FIG. 10. A subcluster of highly semantically related nouns from the CHILDES analysis.

This demonstrates that distributional methods can provide some degree of additional constraint on a word's semantics. However, the degree of semantic relatedness in the dendrogram is relatively small and haphazard.

A much more effective method (for deriving semantic relationships) is presented by Burgess and Lund (e.g. 1997a, 1997b; Lund & Burgess, 1996; see Schütze, 1992, 1993, for a similar, less psychologically motivated approach). Burgess and Lund's "Hyperspace Analogue to Language" (HAL) constructs semantic representations using a similar principle to Finch and Chater's (1991, 1992) method for word classes. Words that are semantically related will tend to occur close together within a particular corpus.

HAL constructs semantic representations by collecting "collocation statistics" for words occurring in a very large corpus (e.g. 160 million words of USENET news). In this context, collocation refers to the co-occurrence of two words within a short stretch (e.g. a 10-word window) of the corpus, with the measure being weighted according to the number of intervening words. The output of this process is a matrix representing the extent to which a set of context words occurred within the same window as the target word. Lund and Burgess (1996) used the most frequent 70,000 words as both target and context words, their aim being to construct semantic representations for as large a set of words as possible, in order to model data from adult studies using a range of stimuli, rather than modelling lexical development. For each

target/context word, the row and column of the matrix indexed by that word which indicate, respectively, words that followed and preceded the target word within the 10-word window) were then concatenated to produce a 140,000 element "semantic vector".[12] Burgess and Lund's claim is that these semantic vectors, derived purely from a corpus-based analysis, captured aspects of the semantic relationships between the target words.

Figure 11, from Burgess and Lund (1997a), shows the spatial relationships between vectors representing words from the categories of animal names, body parts, and geographical locations. Multidimensional scaling was used to represent the distance relationships within the high-dimensional space of the semantic vectors in two dimensions. Clearly the semantic vectors do capture aspects of the semantic distinctions between these categories: Distributional statistics do carry information about semantic relationships.

Although Burgess and Lund's method is not a model of acquisition per se, they have shown that the semantic relationships captured by the HAL system are related to psychological phenomena in the adult literature. For example, the distance between words in HAL's high dimensional semantic space is reliably correlated with semantic priming effects in lexical decision tasks (Lund & Burgess, 1996). More detailed work has used HAL to model cerebral asymmetries in the time course of semantic priming of multiple meanings: Ambiguous words (such as *bank*) presented centrally or to the left visual field (LVF) prime both their dominant (*money*) and subordinate (*river*) meanings in a lexical decision task with a short (35 ms) delay. After a 70 ms delay, only the dominant meaning is primed, with the subordinate meaning showing inhibition. In contrast, ambiguous words presented to the right visual field prime only the dominant meaning after 35 ms, but prime both dominant and subordinate meanings after 70 ms. Burgess and Lund (1997a) modelled this pattern of performance in terms of activation spreading through a region of HAL's semantic space, and decaying over time. Their model was able to account for the cerebral asymmetry in terms of processing differences (differing levels of initial activation and rates of spread and decay) between the hemispheres. Thus, these factors, rather than representational differences, or modulation of information by the corpus callosum may underlie observed cerebral asymmetries. This and other work (e.g. on using HAL representations as constraints on parsing; Burgess & Lund, 1997b) shows a close tie between the properties of the representations acquired by HAL, and human psychological effects.

[12] The very large amount of data and the massive number of target words used should not be construed as being crucial for the basic ideas embodied by Burgess and Lund's method. The effects reported by Burgess and Lund degrade gracefully as elements of the vectors are removed, and there is only a small difference (in terms of effects) between vectors of 140,000 and vectors of 1000 elements.

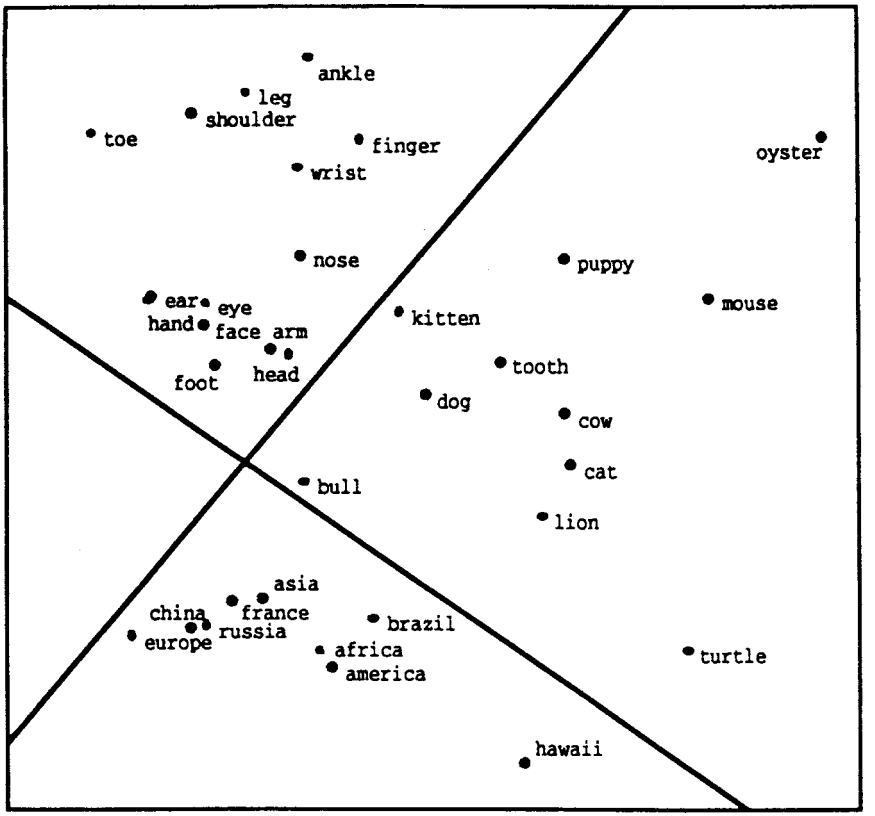

FIG. 11. Results from Burgess and Lund (1997a). The distance relationships between semantic vectors for words belonging to three categories (animals, locations, and body parts) are shown here in two dimensions (via multidimensional scaling). ©1997, Lawrence Erlbaum Associates Inc. Reproduced with permission from Burgess and Lund (1997a).

As with phrasal categories, the quantitative assessment of the informativeness of the relationships between semantic vectors such as those constructed by Lund and Burgess (1996) and Shütze (1992) is problematic. Some indication of the success of the method is indicated by Schütze's (1992) experiments on word sense disambiguation, using semantic vectors derived from this method. Schütze was able to distinguish between meaning such as *factory/living being* for the word *plant*, with success rates of approximately 90%, based on the context in which they occurred.

Summary

As for the acquisition of grammatical structure, the study of distributional information and semantics from a psychological perspective is in its infancy. However, even this early research suggests that connectionist and statistical learnings mechanisms can provide useful constraints on semantics, without requiring either the ability to parse parental utterances, or some initial knowledge of the syntax–semantics relationship.

CONCLUSION

We have argued that common objections to the utility of distributional information for language acquisition are flawed, and these objections are weakened even further by the existence of numerous studies across the range of language acquisition phenomena, in which distributional properties of language have been shown unequivocally to provide information about linguistic properties. These studies demonstrate that distributional information cannot be dismissed—it must be empirically assessed, for particular aspects of language, and for particular learning mechanisms. Although it is unfortunate that the study of distributional learning mechanisms within developmental psychology has been neglected for so long, as a consequence the scope for future work is very wide.

For many aspects of language there remain many potential cues whose value is currently unknown (e.g. the use of single word utterances to break into segmentation, or of stress markings for determining a word's syntactic category). Similarly, most work has been performed using English corpora; the applicability of distributional techniques across languages is likely to be a major focus for future research.

Additionally, although the relevance of distributional information to developmental research should now be clear, there is clearly a need for stronger ties between the predictions of statistical and connectionist learning mechanisms and developmental evidence. This relationship is likely to be bidirectional, with developmental evidence constraining possible learning mechanisms, and with modelling work suggesting new lines of experimental enquiry. As well as its developmental relevance, such inquiry is likely to have implications for theories of adult language processing as well (e.g. the case of phonology and spoken word recognition discussed earlier).

Future research is also likely to be increasingly concerned with the problem of how particular cues (such as morphology and syntactic category) might be combined or interact (MacWhinney, Leinbach, Taraban, & McDonald, 1989; Christiansen, Allen, & Seidenberg, this issue, provide examples of this kind of investigation). Such studies are also likely to illuminate the role of innate knowledge, either in declarative form, or in terms of the initial structure of the learning mechanism.

In short, we have argued that distributional information is a potentially valuable source of information for many aspects of language acquisition, and that the study of both connectionist and statistical learning mechanisms is a valuable research strategy, complementing existing lines of research in developmental psychology.

REFERENCES

Abu-Bakar, M., & Chater, N. (1993). Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 191–196). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Anderson, S.R. (1992). *A-morphous morphology.* New York: Cambridge University Press.

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

Bates, E., Benigni, L., Bretherton, I., Camaioni, L., & Volterra, V. (1979). *The emergence of symbols: Communications and cognition in infancy*. New York: Academic Press.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge, UK: Cambridge Universtity Press.

Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150–177.

Best, C.T., McRoberts, G.W., & Sithole, N.M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 345–460.

Bever, T.G. (Ed.). (1982). *Regressions in mental development: Basic phenomena and theories.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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

Bloom, L. (1973). *One word at a time: The use of single word utterances before syntax.* The Hague, Netherlands: Mouton.

Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In E. Wanner & L. Gleitman (Eds), *Language acquisition: The state of the art.* Cambridge, UK: Cambridge University Press.

Bowerman, M. (1987). Commentary: Mechanisms of language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 443-466). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Bowerman, M. (1988). The "no negative evidence" problem: How do children avoid constructing an overly general grammar? In J.A. Hawkins (Ed.), *Explaining language universals* (pp. 73–101). Oxford, UK: Blackwell.

Braine, M.D.S. (1971). On two types of models of the internationalization of grammar. In D.I. Slobin (Ed.), *The ontogenesis of grammar* (pp. 153–186). New York: Academic Press.

Braine, M.D.S. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development*, *41*(1), Serial no. 164.

Braine, M.D.S. (1988). Review of "Language learnability and language development" by S. Pinker. *Journal of Child Language*, *15*, 189–199.

Brent, M. (1993). Minimal generative explanations: A middle ground between neurons and triggers. In *Proceedings of the 15th annual conference of the Cognitive Science Society* (pp. 28–36). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Brent, M.R., & Cartwright, T.A. (1997). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *63*, 121–170.
Brill, E., Magerman, D., Marcus, M., & Santorini, B. (1990). Deducing linguistic structure from the statistics of large corpora. *DARPA Speech and Natural Language Workshop.* Hidden Valley, PA: Morgan Kaufmann.
Brown, R. (1957). Linguistic determination and the part of speech. *Journal of Abnormal and Social Psychology*, *55*, 1–5.
Brown, R., & Bellugi, U. (1964). Three processes in the child's acquisition of syntax. In E.H. Lenneberg (Ed.), *New directions in the study of language.* Cambridge, MA: MIT Press.
Bruner, J. (1975). The ontogenesis of speech acts. *Journal of Child Language*, *2*, 1–19.
Burgess, C., & Lund, K. (1997a). *Modeling cerebral asymmetries in high-dimensional semantic space.* In M. Beeman & C. Chiarello (Eds), *Right hemisphere language comprehension: Perspectives from cognitive science*, Mahwah, NJ: Lawrence Erlbaum Associates Inc.
Burgess, C., & Lund, K. (1997b). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes*, *12*, 177–210.
Bybee, J., & Slobin, D. (1982). Rules and schemas in the development and use of the English past tense. *Language*, *58*, 265–289.
Cairns, P., Shillcock, R.C., Chater, N., & Levy, J. (1994). Lexical segmentation: The role of sequential statistics in supervised and unsupervised models. In A. Ram & K. Eiselt (Eds), *Proceedings of the 16th annual conference of the Cognitive Science Society* (pp. 136–141). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Cairns, P., Shillcock, R.C., Chater, N., & Levy, J. (1995). Bottom-up connectionist modelling of speech. In J. Levy, D. Bairaktaris, J.A. Bullinaria, & P. Cairns (Eds), *Proceedings of the connectionist models of memory and language* (pp. 289–310). London: UCL Press.
Cassidy, K.W., & Kelly, M.H. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language*, *30*, 348–369.
Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review*, *103*, 566–591.
Chomsky, N. (1959). A review of B.F. Skinner's verbal behavior. *Language*, *35*, 26–58.
Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.
Chomsky, N. (1980). *Rules and representations.* Oxford, UK: Blackwell.
Christiansen, M., & Chater, N. (1994). Generalization and connectionist language learning. *Mind and Language*, *9*, 273–287.
Clahsen, H., Rothweiler, M., Woest, A., & Marcus, G.F. (1993). Regular and irregular inflection in the acquisition of German plural nouns. *Cognition*, *45*, 225–255.
Cole, R.A. (1980). *Perception and production of fluent speech.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Cottrell, G.W., Nguyen, M., & Tsung, F. (1993). Tau net: The way to do is to be. *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 365–370). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Cutler, A. (1993). Phonological cues to open- and closed-class words in the processing of spoken sentences. *Journal of Psycholinguistic Research*, *22*, 109–131.
Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation—evidence from juncture misperception. *Journal of Memory and Language*, *31*, 218–236.
Cutler, A., & Carter, D.M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*, 133–142.
Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 113–121.
Ellison, T.M. (1992). The machine learning of phonological structure. Unpublished doctoral dissertation, Department of Computer Science, University of Western Australia, Perth, Australia.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.
Elman, J.L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.
Elman, J.L., & McClelland, J.L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language*, *27*, 143–165.
Ervin, S. (1964). Imitation and structural change in children's language. In E. Lenneberg (Ed.), *New directions in the study of language.* Cambridge, MA: MIT Press.
Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to adults and infants. *Developmental Psychology*, *27*, 209–221.
Finch, S.P., & Chater, N. (1991). A hybrid approach to the automatic learning of linguistic categories. *Artificial Intelligence and Simulated Behaviour Quarterly*, *78*, 16–24.
Finch, S.P., & Chater, N. (1992). Bootstrapping syntactic categories. *Proceedings of the 14th annual conference of the Cognitive Science Society of America* (pp. 820–825). Bloomington, IN: Cognitive Science Society.
Finch, S.P., & Chater, N. (1993). Learning syntactic categories: A statistical approach. In M. Oaksford & G.D.A. Brown (Eds), *Neurodynamics and psychology.* London: Academic Press.
Finch, S.P., & Chater, N. (1994). Distributional bootstrapping: From word class to proto-sentence. In A. Ram & K. Eiselt (Eds), *Proceedings of the 16th annual meeting of the Cognitive Science Society* (pp. 301–306). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Finch, S.P., Chater, N., & Redington, M. (1995). Acquiring syntactic information from distributional statistics. In J. Levy, D. Bairaktaris, J.A. Bullinaria, & P. Cairns (Eds), *Connectionist models of memory and language* (pp. 229–242). London: UCL Press.
Fodor, J. (1981). *Representations: Philosophical essays on the foundation of cognitive science* (pp. 257–316). Brighton, UK: Harvester Press.
Forrester, N., & Plunkett, K. (1994). Learning the Arabic plural: The case for minority default mappings in connectionist networks. In A. Ram & K. Eiselt (Eds), *Proceedings of the 16th annual conference of the Cognitive Science Society* (pp. 319–323). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Gerken, L.A., Landau, B., & Remez, R. (1990). Function morphemes in young children's speech perception and production. *Developmental Psychology*, *26*, 204–216.
Gibson, J.J. (1979). *The ecological approach to vision.* Boston, MA: Houghton Mifflin.
Gleitman, H. (1991). *Psychology* (3rd edn.). New York: Norton.
Gleitman, L.R. (1990). The structural sources of word meaning. *Language Acquisition*, *1*, 3–55.
Gleitman, L.R., Gleitman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representations for language learning. In F.J. Newmeyer (Ed.), *Linguistics: The Cambridge survey, Vol. 3* (pp. 150–193). Cambridge, UK: Cambridge University Press.
Gold, E.M. (1967). Language identification in the limit. *Information and control*, *10*, 447–474.
Grimshaw, J. (1981). Form, function, and the language acquisition device. In C.L. Baker & J. McCarthy (Eds), *The logical problem of language acquisition.* Cambridge, MA: MIT Press.
Grünwald, P. (1996). A minimum description length approach to grammar inference. In S. Wermter, E. Riloff, & G. Scheler (Eds), *Symbolic, connectionist, and statistical approaches to learning for natural language processing. Springer Lecture Notes in Artificial Intelligence 1040* (pp. 203–216). Berlin, Germany: Springer-Verlag.
Harrington, J., Watson, G., & Cooper, M. (1988). Word boundary identification from phoneme sequence constraints in automatic continuous speech recognition. *Proceedings of the 12th international conference on Computational Linguistics* (pp. 225–230).
Harris, Z.S. (1955). From phoneme to morpheme, *Language*, *31*, 190–222.
Hirsh-Pasek, K., Kemler-Nelson, D.G., Jusczyk, P.K., Wright, K., & Druss, B. (1987). Clauses are perceptual units for prelinguistic infants. *Cognition*, *26*, 269–286.

Horn, B.K.P. (1975). Obtaining shape from shading information. In P.H. Winston (Ed.), *The psychology of computer vision* (pp. 115–155). New York: McGraw-Hill.

Hughes, J., & Atwell, E. (1994). the automated evaluation of inferred word classifications. In T. Cohn (Ed.), *Proceedings of the European conference on Artificial Intelligence (ECAI)* (pp. 535–539). Chichester, UK: John Wiley.

Jusczyk, P.W. (1993). Discovering sound patterns in the native language. In *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 49–60). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Jusczyk, P.W., & Aslin, R.N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1–23.

Jusczyk, P.W., Cutler, A., & Redanz, N.J. (1993). Infants' preference for the predominant stress patterns of English words. *Child Development*, *64*, 657–687.

Jusczyk, P.W., Friederici, A.D., Wessels, J.M.I., Svenkerud, V.Y., & Jusczyk, A.M. (1993). Infants' sensitivity to the sound patterns of native language words. *Journal of Memory and Language*, *32*, 402–420.

Jusczyk, P.W., Luce, P.A., & Luce, J.C. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language*, *33*, 630–645.

Karmiloff-Smith, A. (1979). *A functional approach to child language: A study of determiners and reference.* Cambridge, UK: Cambridge University Press.

Kelly, M.H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*, 349–364.

Kirsh, D. (1991). PDP learnability and innate knowledge of language. *Center for Research in Language Newsletter*, *6*, December, 3–17.

Kiss, G.R. (1973). Grammatical word classes: A learning process and its simulation. *Psychology of Learning and Motivation*, *7*, 1–41.

Kohonen, T. (1982). Self organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Kuczaj, S. (1977). The acquisition of regular and irregular past forms. *Journal of Verbal Learning and Verbal Behavior*, *16*, 589–600.

Landau, B., & Gleitman, L. (1985). *Language and experience: Evidence from the blind child.* Cambridge, MA: Harvard University Press.

Langley, P. (1987). Machine learning and grammar induction. *Machine Learning*, *2*, 5–8.

Lehiste, I. (1971). The timing of utterances and linguistic boundaries. *Journal of the Acoustical Society of America*, *51*, 2018–2024.

Levy, Y. (1983). It's frogs all the way down. *Cognition*, *15*, 75–93.

Li, M., & Vitanyi, P. (1993). *An introduction to Kolmogorov complexity and its applications.* Berlin, Germany: Springer Verlag.

Li, M., & Vitanyi, P. (1995). *Computational machine learning: Theory of praxis.* NeuroCOLT Tech. Report No. NC-TR-95-052. Department of Computer Science, University of London, UK.

Liberman, M., & Prince, A.S. (1977). On the stress and linguistic rhythm. *Linguistic Inquiry*, *8*, 249–336.

Lund, K., & Burgess, C. (1986). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, and Computers*, *28*, 203–208.

MacKay, D.J.C. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, *4*, 448–472.

MacWhinney, B., Leinbach, J., Taraban, R., & McDonald, J. (1989). Language learning: Cues or rules? *Journal of Memory and Language*, *28*, 255–277.

MacWhinney, B., & Snow, C. (1985). The child language data exchange system. *Journal of Child Language*, *12*, 271–295.

Mann, V.A., & Repp, B.H. (1980). Influence of vocalic context on perception of the [s]–[d] distinction. *Perception and Psychophysics*, *28*, 213–228.

Maratsos, M. (1988). The acquisition of formal word classes. In Y. Levy, I.M. Schlesinger, & M.D.S. Braine (Eds), *Categories and processes in language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Maratsos, M., & Chalkley, M. (1980). The internal language of children's syntax: The ontogenesis and representation of syntactic categories. In K. Nelson (Ed.), *Children's language, Vol. 2.* New York: Gardner Press.

Marcus, G.F. (1993). Negative evidence in language acquisition. *Cognition*, *46*, 53–85.

Marcus, G.F., Brinkmann, U., Clahsen, H., Weise, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*, 189–256.

Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*, 5–165.

Marcus, M. (1991). The automatic acquisition of linguistic structure from large corpora. In D. Powers (Ed.), *Proceedings of the 1991 Spring symposium on the Machine Learning of Natural Language and Ontology*. Stanford, CA.

Mareschal, D., Plunkett, K., & Harris, P. (1995). Developing object permanence: A connectionist model. In J.D. Moore & J.F. Lehmann (Eds), *Proceedings of the 17th annual conference of the Cognitive Science Society* (pp. 170–175). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Marslen-Wilson, W., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*, 29–63.

Marr, D. (1982). *Vision.* San Francisco: W.H. Freeman.

Marr, D., & Hildreth, E. (1980). Theory of edge detection. *Proceedings of the Royal Society of London*, *B207*, 187–217.

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London*, *B204*, 301–328.

McCarthy, D. (1954). Language development in children. In L. Carmichael (Ed.), *Manual of child psychology.* New York: Wiley.

Mehler, J., Jusczyk, P.W., Lambertz, G., Halstead, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*, 144–178.

Morgan, J., & Newport, E. (1981). The role of constituent structure in the induction of an artificial language. *Journal of Verbal Learning and Verbal Behavior*, *20*, 67–85.

Nakisa, R.C., & Hahn, U. (1996). Where defaults don't help: the case of the German Plural System. In G.W. Cottrell (Ed.), *Proceedings of the 18th annual conference of the Cognitive Science Society* (pp. 177–182). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Neisser, U. (1967). *Cognitive psychology.* New York: Blackwell.

Nelson, K. (1973). Structure and strategy in learning to talk. *Monographs of the Society for Research in Child Development*, *38*, (1–2), Serial No. 149.

Nelson, K. (1977). Facilitating children's syntax acquisition. *Developmental Psychology*, *13*, 101–107.

Newport, E. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Ninio, A., & Snow, C.E. (1988). Language acquisition through language use: The functional sources of children's early utterances. In Y. Levy, I.M. Schlesinger, & M.D.S. Braine (Eds) *Categories and processes in language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Norris, D.G. (1993). Bottom-up connectionist models of "interaction". In G. Altmann & R. Shillcock (Eds.). *Cognitive models of speech processing* (pp. 211–234). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Nosofsky, R.M. (1990). Relations between exemplar similarity and likelihood models of classification. *Journal of Mathematical Psychology*, *34*, 393–418.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month old children do not have a grammatical category of verb. *Cognitive Development*, *8*, 245–272.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*, 258–278.

Perez-Pereira, M. (1991). The acquisition of gender: What Spanish children tell us. *Journal of Child Language*, *18*, 571–590.

Peters, A.M. (1985). Language segmentation: Operating principles for the perception and analysis of language. In D.I. Slobin (Ed.), *The crosslinguistic study of language acquisition: Vol.2. Theoretical issues* (pp. 1029–1067). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Pine, J.M., & Martindale, H. (1996). Syntactic categories in the speech of very young children: The case of the determiner. *Journal of Child Language*, *23*, 369–395.

Pinker, S. (1984). *Language learnability and language development.* Cambridge, MA: Harvard University Press.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cogition*, *28*, 73–193.

Plunkett, K. (1986). Learning strategies in two Danish children's language development. *Scandinavian Journal of Psychology*, *27*, 64–73.

Plunkett, K. (1990). The segmentation problem in early language acquisition. *Center for Research in Language Newsletter*, *5*, November; 1–17.

Plunkett, K. (1996). Development in a connectionist framework: Rethinking the nature–nurture debate. *Center for Research in Language Newsletter*, *10*, February, 3–14.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43–102.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, *48*, 1–49.

Plunkett, K., Sinha, C. Møller, C.F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in a connectionist net. *Connection Science*, *4*, 293–312.

Prasada, S., & Pinker, S. (1993). Similarity-based and rule-based generalizations in inflectional morphology. *Language and Cognitive Processes*, *8*, 1–56.

Quine, W.V.O. (1953). Two dogmas of empiricism. In W.V.O. Quine (Ed.), *From a logical point of view.* Harper Torchbooks.

Quine, W.V.O. (1960). *Word and object.* Cambridge, MA: MIT Press.

Radford, A. (1988). *Transformational grammar* (2nd edn.). Cambridge, UK: Cambridge University Press.

Redington, F.M., Chater, N., & Finch, S. (1993). Distributional information and the acquisition of linguistic categories: A statistical approach. *Proceedings of the 15th annual meeting of the Cognitive Science Society* (pp. 848–853). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Redington, M., Chater, N., & Finch, S. (in press). The potential contribution of distributional information to early syntactic category acquisition. *Cognitive Science*.

Redington, M., Chater, N., Huang, C., Chang, L., Finch, S., & Chen, K. (1995). The universality of simple distributional methods: Identifying syntactic categories in Chinese. In *Proceedings of the Cognitive Science of Natural Language Processing.* Dublin City University: Dublin.

Redlich, A.N. (1993). Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, *5*, 289–304.

Rescorla, L. (1980). Overextension in early language development. *Journal of Child Language*, *7*, 321–335.

Richards, W. (1988). *Natural computation.* Cambridge, MA: MIT Press.

Rissanen, J. (1989). *Stochastic complexity and statistical enquiry.* Singapore: World Scientific Publishers.

Rosenfeld, A., Huang, H.K., & Schneider, V.B. (1969). An application of cluster detection to text and picture processing. *IEEE Transactions on Information theory*, *15*, 672–681.

Rumelhart, D., & McClelland, J. (1986). On learning the past tenses of English verbs: Implicit rules or parallel distributed processing. In J. McClelland, D. Rumelhart, & the PDP Research Group (Eds), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2* (pp. 216–271). Cambridge, MA: MIT Press.

Sachs, J. (1976). The development of speech. In E.C. Carterette & M.P. Friedman (Eds), *Handbook of perception: Vol. 7. Language and speech* (pp. 145–172). New York: Academic Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In G.W. Cottrell (Ed.), *Proceedings of the 18th annual conference of the Cognitive Science Society* (pp. 376–380). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Saffran, J.R., Newport, E.L., & Aslin, R.N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Schlesinger, I.M. (1981). Semantic assimilation in the acquisition of relational categories. In W. Deutsch (Ed.), *The child's construction of language.* New York: Academic Press.

Schlesinger, I.M (1982). *Steps to language: Toward a theory of native language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Schlesinger, I.M. (1988). The origin of relational categories. In Y. Levy, I.M. Schlesinger & M.D.S. Braine (Eds), *Categories and processes in language acquisition.* Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Scholtes, J.C. (1991a). Kohonen's self-organising map applied towards natural language processing. *Proceedings of the CUNY conference on Human Sentence Processing.*

Scholtes, J.C. (1991b). Using extended feature maps in a language acquisition model. *Proceedings of the 2nd Australian conference on Neural Networks.*

Schütze, H. (1992). Dimensions of meaning. In *Proceedings of Supercomputing.*

Schütze, H. (1993). Word space. In S.J. Hanson, J.D. Cowan, & C.L. Giles (Eds), *Advances in neural information processing systems, vol. 5.* San Mateo, CA: Morgan Kaufmann.

Seidenberg, M., & McClelland, J.L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Servan-Schreiber, E. (1992). Chunking processes and context effects in letter perception. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 78–83). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Shillcock, R., Hicks, J., Cairns, P., Levy, J., & Chater, N. (in press). A statistical analysis of an idealised phonological transcription of the London–Lund corpus. *Computer Speech and Language.*

Shillcock, R., Lindsey, G., Levy, J., & Chater, N. (1992). A phonologically motivated input representation for the modelling of auditory word perception in continuous speech. In *Proceedings of the 14th annual conference of the Cognitive Science Society* (pp. 408–413). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Slobin, D.I. (1973). Cognitive prerequisites for the development of grammar. In C.A. Ferguson & D.I. Slobin (Eds), *Studies of language development.* New York: Holt, Rinehart & Winston.

Snow, C.E. (1972). Mother's speech to children learning language. *Child Development*, *43*, 549–565.

Snow, C.E. (1988). The last word: Questions about the emergence of words. In M. Smith & J. Locke (Eds), *The emergent lexicon.* New York: Academic Press.

Solomonoff, R.J. (1964a). A formal theory of inductive inference, Pt. 1. *Information and Control*, *11*, 1–22.

Solomonoff, R.J. (1964b). A formal theory of inductive inference, Pt. 2. *Information and Control*, *11*, 224–254.

Suomi, K. (1993). An outline of a developmental model of adult phonological organization and behavior. *Journal of Phonetics*, *21*, 29–60.

Svartvik, J., & Quirk, R. (1980). *A corpus of English conversation.* Lund, Sweden: LiberLaromedel Lund.

Tomasello, A. (1992). *First verbs: A case study of grammatical development.* Cambridge, UK: Cambridge University Press.

Tomasello, M., & Olguin, R. (1983). Twenty-three-month old children have a grammatical category of noun. *Cognitive Development*, *8*, 451–464.

Valian, V. (1986). Syntactic categories in the speech of young children. *Developmental Psychology*, *22*, 562–579.

Wallace, C., & Boulton, D.M. (1968). An information measure for classification. *Computer Journal*, *11*, 185–195.

Waltz, D.L. (1975). Understanding line drawings of scenes with shadows. In P.H. Winston (Ed.), *The psychology of computer vision* (pp. 19–91). New York: McGraw-Hill.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual re-organisation during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Wolff, J.G. (1975). An algorithm for the segmentation of an artificial language analogue. *British Journal of Psychology*, *66*, 79–90.

Wolff, J.G. (1977). The discovery of segmentation in natural language. *British Journal of Psychology*, *68*, 97–106.

Wolff, J.G. (1988). Learning syntax and meanings through optimization and distributional analysis. In Y. Levy, I.M. Schlesinger, & M.D.S. Braine (Eds), *Categories and processes in language acquisition* (pp. 179–215). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

Woodward, J.Z., & Aslin, R.N. (1990). Segmentation cues in maternal cues to infants. Paper presented at the 7th Biennial Meeting of the international conference on Infant Studies. Montreal, Canada.

Zemel, R.S. (1993). A minimum description length framework for unsupervised learning. Unpublished doctoral dissertation, Department of Computer Science, University of Toronto, Canada.