Probabilistic and distributional approaches to language acquisition

Martin Redington and Nick Chater

Recent computational research on natural language corpora has revealed that relatively simple statistical learning mechanisms can make an important contribution to certain aspects of language acquisition. For example, statistical and connectionist methods can provide valuable cues to word segmentation and to the acquisition of inflectional morphology, syntactic classes and aspects of word meaning. In each case, these cues are partial, and must be integrated with additional information, whether from other environmental cues or innate knowledge, to provide a complete solution to the acquisition problem. The success of these methods with real natural language corpora demonstrates their feasibility as part of the language acquisition mechanism, an area where previously most research has been limited to highly idealized artificial input or to a priori considerations regarding the feasibility of acquisition mechanisms. Exploring probabilistic learning mechanisms with natural language input provides both an empirical basis for assessing how innate constraints interact with information derived from the environment, and a source of hypotheses for experimental testing.

Theoretical accounts of language acquisition have emphasized the role of innate linguistic knowledge, with the influence of the child's environment playing a relatively minor role[1]. However, 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 (such as vocabulary) must be learnt. Moreover, even putatively innate knowledge must be tuned (for instance, by 'parameter setting') to the specific properties of the language to be learned.

Recently, it has become possible to study the contribution of learning in language acquisition from a new perspective. Potential models can be coded as computer programs, and exposed to (some approximation of) natural language input. This work explores the utility of important classes of language-internal, or distributional information, derived from the relationships between linguistic units such as phonemes, morphemes, words and phrases. Distributional information can be extracted readily by a range of probabilistic learning mechanisms, including connectionist networks and conventional statistics, which collectively we shall term distributional learning mechanisms. This approach is inspired by and builds on work in structural linguistics, where distributional methods were used as a methodology for deriving linguistic theories, rather than as models of acquisition[2]. The research we review shows that distributional information provides valuable cues to many aspects of language, which potentially may be exploited by the child.

Distributional information contrasts with the extra-linguistic sources of information that infants might exploit, including features of the physical and social environment or the meaning of an utterance. Undoubtedly, extra-linguistic information plays a major role in the acquisition of language, but its utility is difficult to evaluate computationally, because the child's representation of the environment is unknown; even if the resources to compile 'corpora' relating language and environment were available, it would still be unclear how the environment should be encoded. This provides a methodological reason to focus on language-internal aspects of environmental input, although this approach is consistent with the possibility that distributional information may only be relevant to some aspects of language acquisition (see Box 1), and is compatible with the innateness of both domain-specific language learning mechanisms and knowledge of many universal properties of language[3,4]. By evaluating probabilistic learning mechanisms empirically with natural language input, it may be possible to assess how language-external factors and innate constraints interact with distributional information.

M. Reding[…] the Depar[…] Psychology, University […] London, UK WC1E[…]

tel: +44 171 380 7[…]
fax: +44 171 436 4276
e-mail: m.redington@ucl.ac.uk

N. Chater is at the Department of Psychology, University of Warwick, Coventry, UK CV4 7AL.

tel: +44 1203 523096
fax: +44 1203 524225
e-mail: nick.chater@warwick.ac.uk

 1364-6613/97/$17.00 PII: S1364-6613(97)01081-4

Box 1. Can distributional methods account for all language acquisition?

Two extreme views concerning the utility of distributional methods are possible. One is that distributional methods can learn all aspects of language unaided. The other is that distributional methods can provide no useful information about any aspect of language. Debates concerning specific distributional methods often adopt implicitly, or are (mis)interpreted as advocating, one of these extreme positions. Our position is that distributional learning methods are valuable in a number of domains, such as those outlined in this article. But there are many aspects of language (such as syntax and compositional semantics) that exhibit highly complex and structured regularities. It has been argued that these are intractable to any learning method, including distributional methods and, hence, require the existence of innate symbolic linguistic knowledge[a]. Whatever the strengths of these arguments, they are not in any way undermined by the success of the distributional methods described in this paper. Indeed, it might be suggested that distributional methods are useful in learning to encode the aspects of the language that are specific to particular languages, so that innate language universal knowledge can be brought to bear. More generally, this might suggest a possible division of labour between distributional methods and traditional formal learning theory[b]. Nonetheless, we believe that the success of distributional methods in the limited aspects of language so far attacked does show that empirical research may produce better results than may be expected from considerations of linguistic theory. Therefore, we believe that pushing distributional methods as far as possible is an important enterprise, which is likely to illuminate both the value of distributional information and the nature of innate constraints.

References

a Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press

b Osherson, D.N., Stob, M. and Weinstein, S. (1985) *Systems that Learn*, MIT Press

Below, we outline the application of probabilistic methods to four important aspects of language acquisition, outlining a case study of recent research for each.

Segmentation

A problem faced early in language acquisition is segmenting the continuous speech stream into discrete words. This is difficult because conversational speech contains no 'gaps' or obvious acoustic markers to signal word boundaries[5].

Theories of adult segmentation (for example, the 'cohort model'; see Ref. 6) propose that the lexicon crucially constrains segmentation. But the child, possessing no initial lexicon, faces a chicken and egg problem. Somehow, the child must 'bootstrap' the ability to segment and learn the lexicon of the natural language.

Many possible language-internal cues that the child may use in segmentation have been suggested, including bootstrapping a vocabulary from single word utterances[7], exploiting subtle acoustic/phonetic boundary markers in the speech signal[8], prosody (including pauses, segmental lengthening, metrical patterns, intonation contours[9] and stress patterns[10]) and phonotactic constraints (sequential regularities between phonemes)[11].

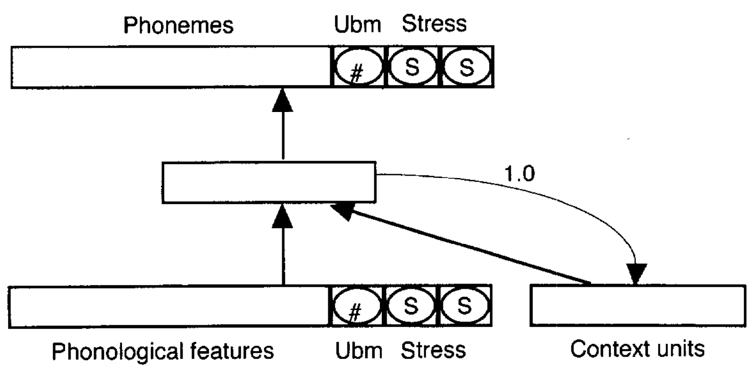

Fig. 1 The simple recurrent network model used by Christiansen *et al.* The current phoneme is represented as a set of features on the input layer. At the output layer, individual units represent each phoneme. The activation of the output units represents the network's prediction of the next phoneme in the sequence. At each time step, the hidden unit activations are copied back on to the context units, allowing the network to maintain prediction-relevant information. The 'Ubm' unit codes for utterance boundaries in the input, and is a useful predictor of word boundaries in the output. Reproduced, with permission, from Ref. 16.

Probabilistic computational models have focussed primarily on lexical stress[12] and phonotactic constraints[13–15]. The work we describe here shows how these cues can be combined within a single learning mechanism. Studying combinations of cues is important, because no single cue is likely to produce a complete solution to any problem in language acquisition. Christiansen, Allen and Seidenberg[16] trained a simple recurrent network (SRN) (see Fig. 1) to predict the next input from a representation of previous inputs [where inputs are coded in terms of phonological features, utterances boundaries (but not word boundaries) and stress] using a corpus of maternal speech to preverbal infants[17]. Stress patterns for words were obtained from a standard database (the MRC Psycholinguistic Database). Initially, the network's connection strengths are random. During training, it learns to exploit phonotactic regularities (for example, in English, /a/ is rarely followed by another /a/, but quite frequently by /b/) in the input. Because certain combinations of phonemes are more likely to occur at the beginning and end of words, these regularities provide a potentially useful cue for word segmentation.

Although the only boundary information that the net received concerned utterance boundaries, there was a good correlation between the SRN's prediction of boundaries (the activation of the output boundary unit) and the occurrence of word boundaries in the corpus. Figure 2 shows the activation of the network's boundary output unit over a short stretch of the corpus. Although the model has no lexicon, over the entire corpus, 43% of words were segmented correctly and over 45% of segmented units correspond to words. Performance dropped marginally when stress was ignored and dropped significantly if phonology was ignored. Distributional methods are capable of even better performance: a state-of-the-art specialized statistical method proposed by Brent and Cartwright[14] segments 72% of words correctly and 65% of segmented units correspond to words. Christiansen *et al.* argue that their model is more psychologically plausible because it uses a general purpose sequential learning, which can combine different cues to segmentation, whereas Brent and Cartwright's model is cast at a relatively abstract level.

The work of Christiansen *et al.* illustrates how a simple and general learning method can find a considerable amount of information about the structure of language, even though that structure (discrete words) is not marked overtly in the input. Moreover, it illustrates how computational analyses of corpora can shed light on how the informational value of different cues can interact (see Box 2).

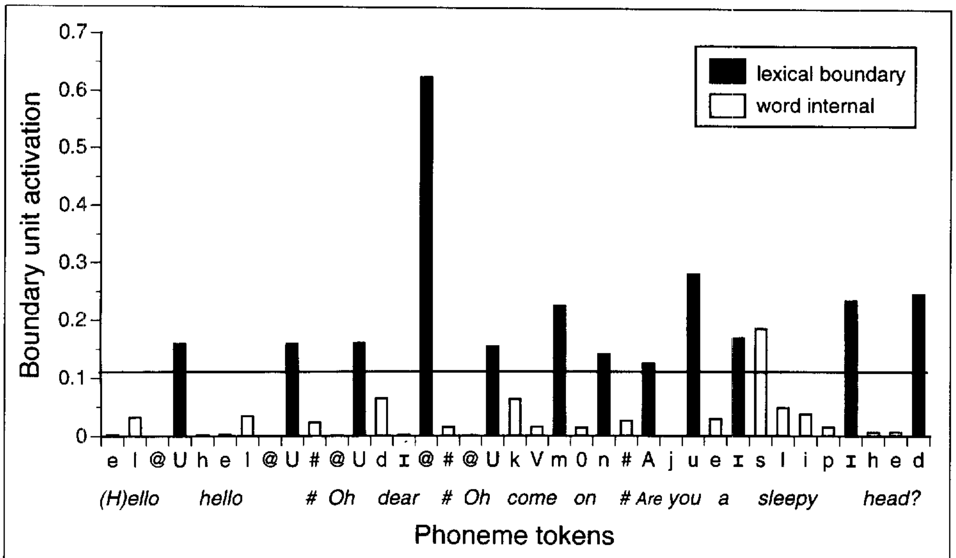

Fig. 2 The activation of the output boundary unit in the Christiansen *et al.* network over a short stretch of the training corpus. Activation of the boundary unit at a particular position corresponds to the network's hypothesis that a boundary follows this phoneme. Black bars indicate the activation at lexical boundaries, whereas the white bars correspond to activation at word-internal positions. The horizontal line indicates the mean boundary unit activation across the whole corpus. A gloss of the input utterances is found beneath the input phoneme tokens (with # denoting an utterance boundary). Reproduced, with permission, from Ref. 16.

Inflectional morphology

Acquiring morphology involves identifying the 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[18]. We focus here on how computational analysis has addressed a key theoretical question: whether inflectional morphology requires two 'routes', one to handle regular morphology (for example, add /-ed/) and one to handle irregulars (for example, 'go' → 'went').

Connectionist studies with idealized languages patterned on English past tense morphology suggest that a single

Box 2. Interaction of cues

Many problems in language acquisition are difficult because no single feature of the input correlates with the relevant aspects of language structure. Although it is a natural starting point for computational and empirical research to study cues in isolation, it may be that the problem of acquisition is easier when multiple cues are taken into account. Figure A shows how three constraints A, B and C, represented by regions of the hypothesis space, are insufficient to identify the correct hypothesis when considered in isolation. It is only by combining these cues that the hypothesis space can substantially be narrowed down. Thus, as the number of cues that the learner considers increases, the difficulty of the learning problem may decrease. This suggests that the cognitive system may aim to exploit as many sources of information as possible in language acquisition.

Moreover, it is possible that cues are only useful when considered together. For example, in the sequences in Fig. B, each cue X and Y seems completely random with respect to the target Z; but when considered together, X and Y perfectly determine Z (specifically, Z has the value 1 exactly when either X or Y have the value 1). Considering cues in isolation assumes implicitly that there is a simple additive relation between cues. The relationship between cues, and the extent to which multiple cues to particular aspects of language reinforce (or potentially, conflict with) each other, is a matter for empirical investigation.

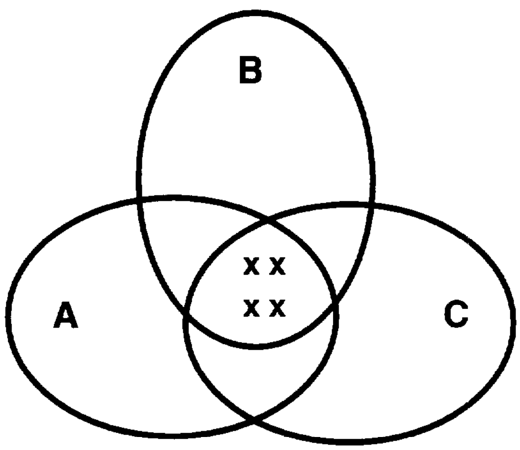

Fig. A A conceptual illustration of three hypothesis spaces given the information provided by the cues A, B, and C. The x symbols correspond to hypotheses that are consistent with all three cues. Reproduced, with permission, from Ref. 16.

X: 1 0 0 1 1 1 0 1 0 1 1 0 0 0 1 1 0 1 0 1 0 1
Y: 0 1 1 0 1 0 0 1 1 0 1 1 0 0 0 0 1 1 1 0 0 1
Z: 1 1 1 1 0 1 0 0 1 1 0 1 0 0 1 1 1 0 1 1 0 0

Fig. B Cues may be uninformative in isolation, but highly informative when combined. The Z sequence is independent of the value of X, and independent of the value of Y, but can be predicted exactly (via XOR) from X and Y.

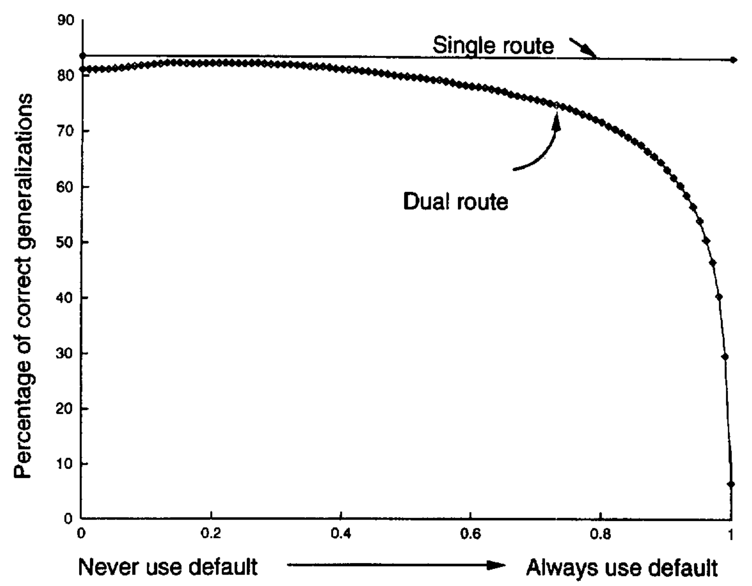

Fig. 3 The percentage of German nouns correctly pluralized. Pluralization by a single rate neural network model, and by a dual route model, with a default 'add /-s/' rule taking over when the output activation of the network is low. The x-axis corresponds to a measure of output activation, with dual route performance depending on the criterion for using the default rule, but always being lower than the single route model. Adapted, with permission, from Ref. 25.

route may handle both cases[19–21]. However, Prasada and Pinker[22] argued that the success of these models results from the distributional statistics of English. Many regular English /-ed/ verbs have low token frequencies, which a connectionist model can handle by learning to add /-ed/ as a default. For irregular verbs, token frequency is typically high, allowing the network to override the default. Prasada and Pinker argued that a default regular mapping with both low type and token frequency could not be learned by a connectionist network. The putative default /-s/ inflection of plural nouns in German[23] appears to provide an example of such a 'minority default mapping'. Marcus *et al.*[24] proposed that the German plural system must be modelled by two routes: a pattern associator which memorizes specific cases (both irregular and regular) and a default rule (add /-s/) which applies when the pattern associator fails.

Nakisa and Hahn[25] asked whether single route associative models (the nearest neighbour algorithm, the 'generalized context model'[26] and a simple feed-forward connectionist net with one hidden layer) could learn the German plural system, and generalize appropriately to novel regular and irregular nouns. The associative model's task was to predict which of 15 different plural types the input stem belonged to. The inputs to the learning mechanisms were phonetic representations of ~4000 German nouns taken from the CELEX database (token frequency was ignored). The three simple associative models scored 71%, 75%, and 84% correct classifications, respectively, on a test set of 4000 previously unseen test nouns.

Nakisa and Hahn also simulated the Marcus *et al.* model, by assuming that any test word which is not close to a training word, according to the associative model (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 for all three models, the presence of the rule led to a decrement in performance. In general, the higher the threshold for memory failure (the more similar a test item had to be to a training item to be irregularized via the associative memory), the greater the decrement in performance (see Fig. 3).

The use of a default rule could only have improved performance for regular nouns occupying regions of phonemic space surrounding clusters of irregulars (see Fig. 4). In real German, Nakisa and Hahn's findings demonstrate that very few regular nouns occur in these regions. 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 (for instance, taking account of token frequency), remains to be performed. Further work might also focus on the extent to which different single and dual route models are able to capture changes in detailed error patterns of under- and over-regularization during development, as well as considering overall levels of performance.

This is an excellent illustration of how distributional analysis of the statistical structure of real language is crucial in assessing the feasibility of psychological proposals, such as whether default rules are involved in learning inflectional morphology.

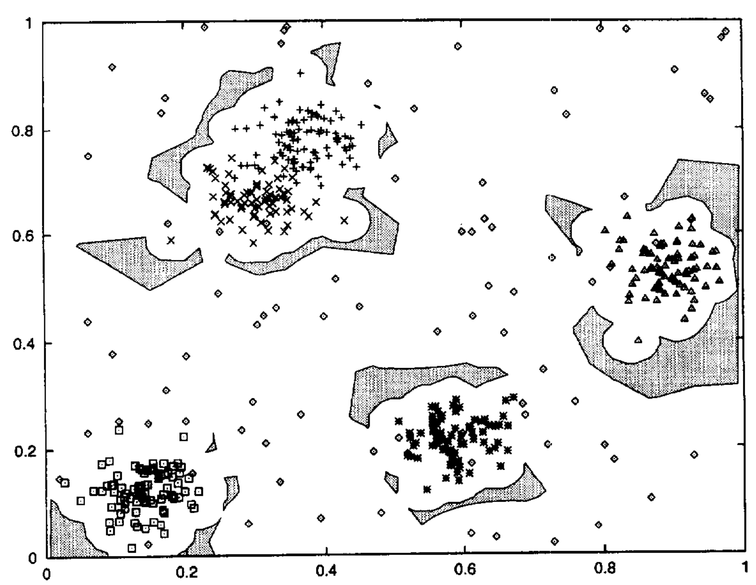

Fig. 4 Regions of phonological space for which single and dual rate models differ. In this artificially generated data, diamonds correspond to regular nouns, whereas the other symbols represent irregular nouns, which are clustered together in phonological space. Single and dual route model classifications differ for the shaded regions surrounding the clusters of irregulars. Regular nouns in these areas are classified correctly by the dual route model, via the default rule, and classified incorrectly as irregulars by the single route model. However, Nakisa and Hahn's results suggest that in real German, as in this artificial data, very few regular nouns are found in these regions of phonological space. Reproduced, with permission, from Ref. 25.

Word classes

A central problem in language acquisition is the acquisition of syntactic categories such as noun and verb. This encompasses both discovering that there are different classes and ascertaining which words belong to each class. Even for theorists who assume that the child innately possesses a universal grammar and syntactic categories, identifying the category of particular words must primarily be a matter of

Box 3. Distributional methods, statistics and connectionism

Many distributional methods exploit simple properties such as 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. Such statistics can be represented easily in a contingency table, as in Fig. A

Co-occurrence statistics can also be captured easily by a connectionist network. In the network shown in Fig. B, units in the first layer are activated 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 (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[a] showed 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 information. 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 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 as maximizing the probability of the weights chosen according to Bayesian principles[b], or minimizing description length (R.S. Zemel, 1993, PhD thesis, Department of Computer Science, University of Toronto). It is remarkable that such simple statistics, which ignore so much important language structure are, nonetheless, so informative about word boundaries, word classes and lexical semantics.

References

a Rumelhart, D. and McClelland, J. (1986) On learning the past tenses of English verbs. Implicit rules or parallel distributed processing, in *Parallel Distributed Processing* (Vol. 2) (McClelland, J. and Rumelhart, D., eds), pp. 216–271, MIT Press

b MacKay, D.J.C. (1992) A practical Bayesian framework for back propagation networks *Neural Comput.* 4, 448–472

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

Fig. A Contingency table. In this case, each cell of the table indexes the number of times that word was followed immediately by word_{n+1}.

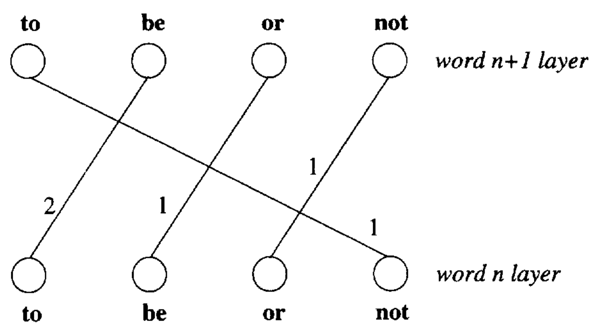

Fig. B Network with a Hebbian learning rule. The weights of the trained network reflect the same statistics as the contingency table shown in Fig. A. For clarity, only non-zero weights are shown.

learning. Universal grammatical features 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. Although once some identifications have been made, it may be possible to use prior grammatical knowledge to facilitate further identifications, the contribution of innate knowledge to initial linguistic categories must be relatively slight.

Both language-external and -internal cues may be relevant to learning syntactic categories. One language external approach[27], 'semantic bootstrapping', exploits the putative correlation between linguistic categories (in particular, noun and verb) and the child's perception of the environment (in terms of objects and actions). This may provide a means of 'breaking in' to the system of syntactic categories. Also, there may be many relevant language-internal factors: regularities between phonology and syntactic categories[28], prosody (relations between intonation and syntactic structure)[29] and distributional analysis, both over morphological variations between lexical items (for example, affixes such as '-ed' are correlated with syntactic category)[30], and at the word level. We focus on this last approach which has a long history[31–33], although such approaches to finding word classes have often been dismissed on a priori grounds within the language learning literature[27].

The 'distributional test' in linguistics[34] 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 prefer...', indicating that these items have the same category. The distributional test is not a foolproof method of grouping words by their syntactic category, because distribution is a function of many factors other than syntactic category (such as word meaning). Thus, for example, cat and barnacle might appear in very different contexts in some corpora, although they have the same word class. Nevertheless, it may be possible to exploit the general principle underlying the distributional test to obtain useful information about word classes. The method described here records the contexts in which the words to be classified appear in a corpus of language, and groups together words with similar distributions of contexts. Here, context is defined in terms of co-occurrence statistics (see Box 3).

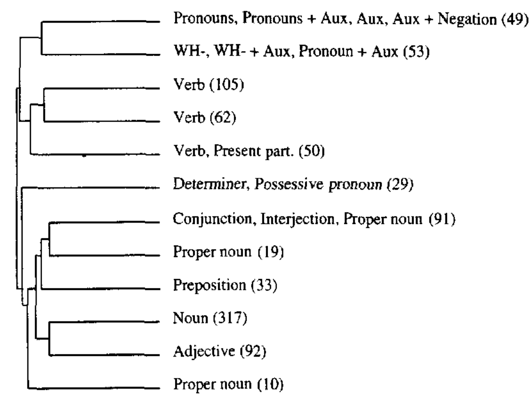

Fig. 5 A dendrogram resulting from a word level analysis of the distributional statistics of the CHILDES corpus. The dendrogram has been truncated at a chosen level of similarity, and the resulting discrete clusters labelled by hand with the syntactic categories to which they correspond. The number of items in each cluster is shown in parentheses. Only clusters with ten or more members are shown.

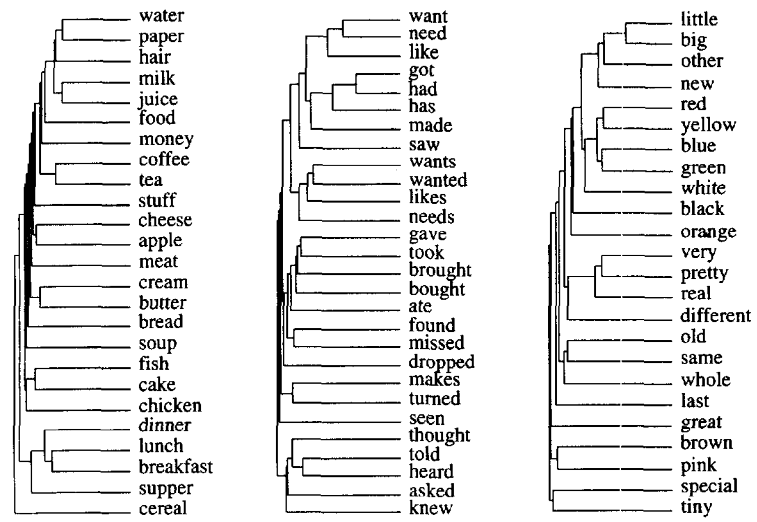

Fig. 6 Low-level clusters of nouns, verbs and adverbs taken from the dendrogram presented in summary form in Fig. 5.

Finch, Chater, and Redington[35–37] used the two words before and after each target word as context. Vectors (rows of a contingency table; see Box 3) representing the co-occurrence statistics for these positions were constructed from a 2.5 million word corpus of transcribed adult speech taken from the CHILDES corpus (much of which was child-directed). The vectors for each position were concatenated to form a single vector for each of 1000 target words. The similarity of distribution between the vectors was calculated using Spearman's rank correlation, and hierarchical cluster analysis was used to group similar words together.

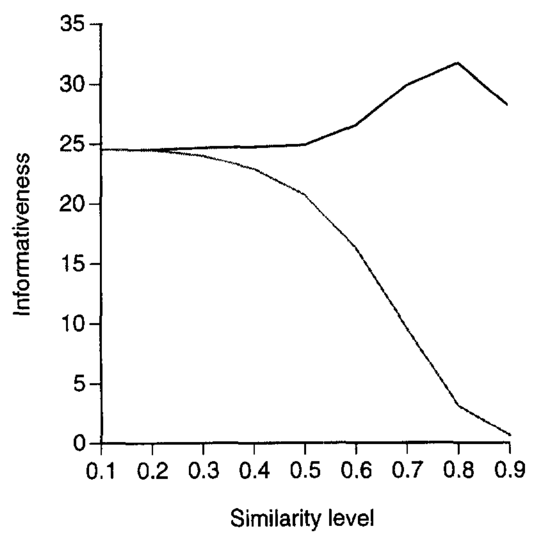

Fig. 7 The informativeness of the hierarchical classification of the target words with respect to the most common syntactic categories of those words. Informativeness is an information theoretical measure of the degree to which words belonging to the same syntactic category are grouped together, and words belonging to different syntactic categories are separated in the dendrogram. The lower line is a baseline value for informativeness, where words were clustered together randomly. The plot shows that the distributional analysis provides information about syntactic categories at all levels of the dendrogram. The most informative level (0.8) is the level at which the summary dendrogram shown in Fig. 5 was cut.

This approach does not partition words into distinct groups corresponding to the syntactic categories, but produces a hierarchical tree, or dendrogram, whose structure reflects to some extent the syntactic relationships between words. Figure 5 shows the high-level structure of the dendrogram resulting from the above analysis. Figure 6 shows examples of the structure of the dendrogram, and its relation to syntactic category at a very fine level.

A quantitative analysis (see Fig. 7) of the mutual information between the structure of the dendrogram, and a canonical syntactic classification of the target words (defined as their most common syntactic usage in English) as a percentage of the joint information in both the derived and canonical classifications, revealed that at all levels of similarity, the dendrogram conveyed useful information about the syntactic structure of English. Words which were clustered together tended to belong to the same syntactic category, and words that were clustered apart tended to belong to different syntactic categories. Thus, computational analysis of real language corpora shows that distributional information at the word level is highly informative about syntactic category, despite a priori objections to its utility.

Lexical semantics

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 if 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. For abstract nouns, and other words which have no concrete referents, these difficulties are compounded further.

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

However, it also seems plausible that language-internal information might be used to constrain the identification of the 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), knowing a word's syntactic category provides some constraint on its meaning; in general, knowing that a word is a noun, perhaps because it occurs in a particular set of local contexts, implies that it will refer to a concrete object or an abstract concept, rather than an action or process[38].

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. Gleitman[39] has proposed that syntax is a potentially powerful cue for the acquisition of meaning. Gleitman assumes that the child possesses a relatively high degree of syntactic knowledge. However, an examination of Fig. 6 shows that the distributional method used above to provide information about syntactic categories also captures some degree of semantic relatedness, without any knowledge of syntax proper. More effective methods for deriving semantic relationships have been discussed by Burgess and Lund[40,41], Schutze[42] and Landauer and Dumais[43].

We focus here on Burgess and Lund's work. Semantic representations are constructed by collecting 'collocation' statistics, capturing the co-occurrence of target and context words within a ten word window of the input corpus [typically a large (160 million) corpus of USENET news], weighted according to the separation of the two words within this window. 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. The row and column of the matrix corresponding to each word are concatenated to form a 'semantic vector'. The claim is that the similarity between semantic vectors for different words captures aspects of the semantic relationships between these words.

Figure 8 shows the spatial relationships between vectors representing words from the categories of animal names, body parts and geographical locations. Multidimensional scaling was used to rerepresent 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. The distance between vectors has also been shown to correlate reliably with psychological phenomena such as semantic priming effects in lexical decision tasks[44].

Burgess and Lund[41] have also shown that a model of spreading activation through the space of semantic vectors is able to account for cerebral asymmetries in the time course of semantic priming of multiple meanings; ambiguous words (such as bank) presented to the left visual field prime both meanings initially (within 35 ms), but only the dominant meaning after a 70 ms delay. Ambiguous words presented to the left visual field prime only the dominant meaning initially, but both meanings after a 70 ms delay.

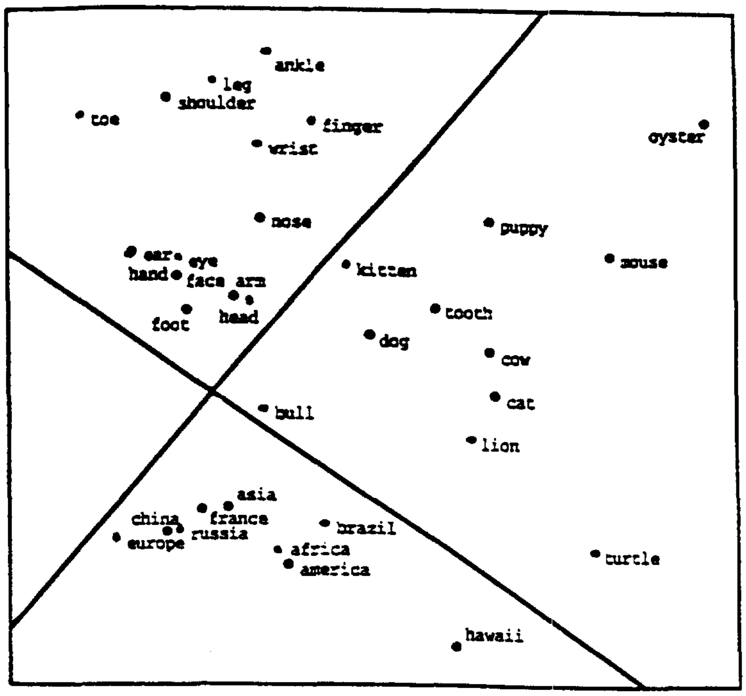

Fig. 8 The semantic distance relationship between animals, locations and body parts, according to Burgess and Lund's (1997) analysis. The distance relationships between semantic vectors for words belonging to three categories (animals, locations and body parts) are shown here in two dimensions (via multidimensional scaling). Clearly, these distance relationships capture something of the semantic relationships between these words. Reproduced, with permission, from Ref. 41.

Using semantic representations derived from HAL, Burgess and Lund were able to model this difference in terms of differing initial activation, and differing rates of spread and decay between the hemispheres, without appealing to representational differences or modulation of information by the corpus callosum.

We have seen that distributional methods are informative about semantic relatedness but, clearly, language-internal information alone cannot be the basis for the acquisition of

Outstanding questions

- How do language learners identify and resolve syntactic and semantic ambiguities? For example, words such as fire can be either nouns or verbs, and many words (such as bank) have multiple meanings.
- How can multiple cues from disparate sources be integrated effectively? For example, phonetics, morphology and word-level information can all contribute towards identifying a word's syntactic class and/or meaning. Although Christiansen *et al.* have made some progress in integrating multiple cues in the context of segmentation, the generality of their approach to other aspects of language acquisition is unclear.
- Distributional relationships, while often reflecting underlying linguistic structure, are noisy and sometimes will be misleading. How serious a problem is this, and how can it be minimized?
- To what extent is distributional information useful across languages? Are different kinds of distributional information present in different languages?
- Can distributional methods be applied successfully to more accurate representations of the input, such as raw speech signals? To what extent do processes occurring at lower levels (speech perception and segmentation) influence higher level processes (for example, the identification of the syntactic classes of words)?
- How can we determine empirically whether infants exploit particular sources of distributional information?

Box 4. How good are distributional methods?

In all the examples in this paper, the distributional methods are shown to provide useful information, but do not approach human levels of linguistic knowledge or performance. Indeed, human level performance would not be expected if distributional methods are, as we suggest, only one among many sources of information involved in language acquisition. If human level performance is too ambitious a standard, how can we assess how good distributional methods are? One approach is to compare them against random benchmarks. Thus, Christiansen *et al.* show that their segmentation model greatly exceeds the performance obtained by randomly assigning word boundaries to respect mean word length, and Fig. 7 shows a comparison of the Redington and Chater syntactic classification against a random classification. Although this shows that the method is finding some useful information, a better comparison is between different algorithms and/or sources of information. Of course, this requires that competing proposals are computationally explicit and applied to appropriate corpora. Currently, most non-distributional proposals in language acquisition are described in purely conceptual terms, which makes comparison difficult. Only when a variety of sources of information and/or algorithms can be compared directly will it be possible to assess accurately the potential contribution of distributional methods. More importantly, it may then be possible to study how different sources of information can be combined to obtain something close to human level performance.

word meaning, because learning word meaning requires relating words to the world, to which distributional methods have no access. Nonetheless, language-internal distributional information about semantic relatedness may be important in helping the child constrain hypotheses about word meaning.

Conclusion

Computational studies using natural language corpora show that distributional information is a potentially valuable cue for many aspects of language acquisition (see Box 4). Does the child use these sources of information? As with all theories of language acquisition, empirical evidence regarding distributional methods is difficult to obtain and interpret[45]. It is encouraging that recent experimental evidence in both children and adults shows that the cognitive system is sensitive to features of the input (for example, co-occurrence statistics) which underlie the mechanisms described here[46–48]. It seems a reasonable working assumption that, given the immense difficulty of the language acquisition problem, the cognitive system is likely to exploit such simple and useful sources of information.

Acknowledgements

M. Redington was supported by the UK Economic and Social Research Council (ESRC) Research Studentship R00429234268 and N. Chater was supported, in part, by Research Grant SPG9029590 from the Joint Councils Initiative in Cognitive Science/HCI. Both authors are supported currently by ESRC grant R000236214.

References

1 Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press

2 Harris, Z. (1955) *Methods in Structural Linguistics*, University of Chicago

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

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

5 Cole, R.A. (1980) *Perception and Production of Fluent Speech*, LEA

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

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

8 Lehiste, I. (1971) The timing of utterances and inguistic boundaries *J. Acoust. Soc. Am.* 51, 2018–2024

9 Gleitman, L. *et al.* (1988) Where learning begins: Initial representations for language learning, in *Linguistics: The Cambridge Survey* (Vol. 3) (Newmeyer, F.J., ed.), pp. 150–193, Cambridge University Press

10 Cutler, A. and Mehler, J. (1993) The periodicity bias *J. Phonetics* 21, 103–108

11 Jusczyk, P.W. (1993) Discovering sound patterns in the native language, in *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 49–60, Erlbaum

12 Cutler, A. and Carter, D.M. (1987) The predominance of strong initial syllables in the English vocabulary *Comp. Speech Lang.* 2, 133–142

13 Wolff, J.G. (1988) Learning syntax and meanings through optimization and distributional analysis, in *Categories and Processes in Language Acquisition* (Levy, Y., Schlesinger, I.M. and Braine, M.D.S., eds), pp. 179–215, LEA

14 Brent, M.R. and Cartwright, T.A. (1996) Distributional regularity and phonotactic constraints are useful for segmentation *Cognition* 61, 93–125

15 Cairns, P. *et al.* (1994) Modelling the acquisition of lexical segmentation, in *Proceedings of the Child Language Research Forum*, University of Chicago Press

16 Christiansen, M.H., Allen, J., and Seidenberg, M.S. Learning to segment speech using multiple cues: A connectionist model *Lang. Cogn. Proc.* (in press)

17 Korman, M. (1984) Adaptive aspects of maternal vocalizations in differing contexts at ten weeks *First Lang.* 5, 44–45

18 Anderson. S.R. (1992) *A-morphous Morphology*, Cambridge University Press

19 Rumelhart, D. and McClelland, J. (1986) On learning the past tenses of English verbs. Implicit rules or parallel distributed processing, in *Parallel Distributed Processing* (Vol. 2) (McClelland, J., and Rumelhart, D., eds), pp. 216–271, MIT Press

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

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

22 Prasada, S. and Pinker, S. (1993) Similarity-based and rule-based generalizations in inflectional morphology *Lang. Cogn. Proc.* 8, 1–56

23 Clahsen, H. *et al.* (1993) Regular and irregular inflection in the acquisition of German plural nouns *Cognition* 45, 225–255

24 Marcus, G. *et al.* (1995) German inflection: The exception that proves the rule *Cognit. Psychol.* 29, 189–256

25 Nakisa, R. C. and Hahn, U. (1996) Where defaults don't help: the case of the German plural system, in *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (Cottrell, G.W., ed.), pp. 177–182, Erlbaum

26 Nosofsky, R.M. (1990) Relations between exemplar similarity and likelihood models of classification *J. Math. Psychol.* 34, 393–418

27 Pinker, S. (1984) *Language Learnability and Language Development*, Harvard University Press

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

29 Morgan, J. and Newport, E. (1981) The role of constituent structure in the induction of an artificial language *J. Verb. Learn. Verb. Behav.* 20, 67–85

30 Maratsos, M. and Chalkley, M. (1980) The internal language of children's syntax: The ontogenesis and representation of syntactic categories, in *Children's Language* (Vol. 2) (L. Nelson, ed.), pp. 127–214, Gardner Press

31 Brill, E. *et. al.* (1990) Deducing linguistic structure from the statistics of large corpora *DARPA Speech and Natural Language Workshop*, Morgan Kaufmann

32 Kiss, G.R. (1973) Grammatical word classes: A learning process and its simulation *Psychol. Learn. Motiv.* 7, 1–41

33 Rosenfeld, A., Huang, H.K. and Schneider, V.B. (1969) An application of cluster detection to text and picture processing *IEEE Trans. Info. Theory* 15, 672–681

34 Radford, A. (1988) *Transformational Grammar* (2nd edn), Cambridge University Press.

35 Finch, S.P. and Chater, N. (1991) A hybrid approach to the automatic learning of linguistic categories *Artif. Intell. Simul. Behav. Qtrly* 78, 16–24

36 Finch, S.P., Chater, N. and Redington, M. (1995) Acquiring syntactic information from distributional statistics, in *Connectionist Models of Memory and Language* (Levy, J. *et al.*, eds), pp. 229–242, University of London Press

37 Redington, M. and Chater, N. Connectionist and statistical approaches to language acquisition: A distributional perspective *Lang. Cogn. Proc.* (in press)

38 Brown, R. (1954) Linguistic determinism and the part of speech *J. Abn. Soc. Psychol.* 55, 1–5

39 Gleitman, L.R. (1990) The structural sources of word meaning *Lang. Acquis.* 1, 3–55

40 Lund, K. and Burgess, C. (1996) Producing high-dimensional semantic spaces from lexical co-occurrence *Behav. Res. Methods Instrument. Comput.* 28, 203–208

41 Burgess, C. and Lund, K. (1997) Modeling cerebral asymmetries of semantic memory using high-dimensional semantic space, in *Right Hemisphere Language Comprehension: Perspectives from Cognitive Neuroscience* (Beeman, M. and Chiarello, C., eds), Erlbaum

42 Schutze, H. (1993) Word space in *Neural Information Processing Systems 5* (Hanson, S.J., Cowan, J.D. and Giles, C.L., eds), pp. 895–902, Morgan Kaufmann

43 Landauer, T.K. and Dumais, S.T. (1997) A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge *Psychol. Rev.* 104, 211–240

44 Burgess, C. and Lund, K. (1997) Modeling parsing constraints with high-dimensional context space *Lang. Cogn. Proc.* 12, 177–210

45 Ninio, A. and Snow, C.E. (1988) Language acquisition through language use: The functional sources of children's early utterances in, *Categories and Processes in Language Acquisition* (Levy, Y., Schlesinger, I.M. and Braine, M.D.S., eds), pp. 11–30, Erlbaum

46 Jusczyk, P.W. (1997) *The Discovery of Spoken Language*, MIT Press

47 Saffran, J.R., Aslin, R.N. and Newport, E.L. (1996) Statistical cues in language acquisition: Word segmentation by infants, in *Proceedings of the 18th Annual Conference of the Cognitive Science Society* (Cottrell, G.W., ed.), pp. 376–380, Erlbaum

48 Saffran, J.R., Newport, E.L. and Aslin, R.N. (1996) Word segmentation: The role of distributional cues *J. Mem. Lang.* 35, 606–621