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. 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. 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. 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.
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. 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

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.

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, exploiting subtle acoustic/phonetic boundary markers in the speech signal, prosody (including pauses, segmental lengthening, metrical patterns, intonation contours and stress patterns) and phonotactic constraints (sequential regularities between phonemes).

Probabilistic computational models have focussed primarily on lexical stress and phonotactic constraints. 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 trained a simple recurrent network (SRN) (see Fig. 1) to predict the next input from a representation of previous speech signal, prosody (including pauses, segmental lengthening, metrical patterns, intonation contours and stress patterns) and phonotactic constraints (sequential regularities between phonemes).

Probabilistic computational models have focussed primarily on lexical stress and phonotactic constraints. 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

Performing segmentation
Segmentation is the process of dividing continuous speech into discrete units. The boundary unit code is set for utterance boundaries in the input, and is a useful predictor of word boundaries in the output. Reproduced, with permission, from Ref. 1b.

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. 1b.

References
methods are capable of even better performance: a state-of-the-art specialized statistical method proposed by Brent and Cartwright\textsuperscript{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).

**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\textsuperscript{5}. 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' \text{\rightarrow} 'went').

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

---

**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.

---

\begin{align*}
X: & 10011101011000111010101 \\
Y: & 0110100110110001110001 \\
Z: & 111110100110100111101100011101100
\end{align*}

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.
route may handle both cases. However, Prasada and Pinker 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 appears to provide an example of such a 'minority default mapping'. Marcus et al. 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 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.

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
learning. Universal grammatical features can only be
recognized 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 ap-
pear 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
categories. Here, context is defined in terms of co-occurrence
statistics (see Box 3).

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 contin-
gency table, as in Fig. A.

Co-occurrence statistics can also be captured easily by a con-
nectionist 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 cor-
pus 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 showed how a single layer connec-
tionist network can map from present to past tense for both
regular and irregular English verbs, using this kind of distribu-
tional information. The problem of optimally training a single-
layer neural network is directly analogous to a conventional sta-
tistical 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 implement-
ing general statistical principles, such as maximizing the prob-
ability of the weights chosen according to Bayesian principles;
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,
b Mackay, D.J.C. (1993) A practical Bayesian framework for back
propagation networks Neural Comput. 4, 448-472

Fig. A Contingency table. In this case, each cell of the
table indexes the number of times that word was fol-
lowed immediately by word...
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.

Finch, Chater, and Redington 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.

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.

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 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, Schutze and Landauer and Dumais.

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 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. The distance between vectors has also been shown to correlate reliably with psychological phenomena such as semantic priming effects in lexical decision tasks.

Burgess and Lund 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.

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 callous.

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...
word meaning, because learning word meaning requires relating words to the world, to which distributional methods have no access. Nonetheless, language-internal distributional information is a potentially valuable cue to word meaning. 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.

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. 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. 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 R000429342400 and N. Chater was supported, in part, by Research Grant SPG0209590 from the Joint Councils Initiative in Cognitive ScienceHCI. Both authors are supported currently by ESRC grant R000236214.

References

2 Harris, Z. (1955) Methods in Structural Linguistics, University of Chicago
6 Marslen-Wilson, W. and Welsh, A. (1978) Processing interactions and lexical access during word recognition in continuous speech Cognit. Psychol. 10, 29–63
23Clausen, H. et al. (1993) Regular and irregular inflection in the acquisition of German plural nouns Cognition 49, 235–255
24Marcus, G. et al. (1995) German inflection: The exception that proves the rule Cognit. Psychol. 29, 189–256
Redington and Chater – Statistical learning mechanism and language acquisition

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

TICS Editorial Policy

Trends in Cognitive Sciences is indispensable reading for all interested in this diverse field. The journal carries a core of authoritative Reviews, written in an easily accessible style by leading authors, summarizing the exciting developments of the subjects you need to know about. Accompanying the Reviews are a range of other types of articles. The Comment section offers short updates and discussions of one or two ground-breaking and important primary papers. The authors of the primary papers have the opportunity of replying to the Comments, thus extending the dialogues on their work. Monitor pieces summarize in 100-200 words recently published papers. Meeting reports convey the excitement of recent conferences or seminars, focusing on recent developments and the discussions surrounding them. Opinion articles are reviews of subjects written from a personal slant, and therefore highlight some of the controversies in cognition. Books etcetera features stimulating essay-reviews of recent publications, whether books, software, CD-ROMs, films, etcetera. And a list of books received for review will be published regularly. Letters stimulated by any of the articles published in TICS are welcome. Authors will be offered the chance to reply to any points raised.