Probabilistic models of language processing and acquisition

Nick Chater¹ and Christopher D. Manning²

¹Department of Psychology, University College London, Gower Street, London, WC1E 6BT, UK
²Stanford University, Depts of Linguistics and Computer Science, Gates Building 1A, 353 Serra Mall, Stanford, California, 94305-9010, USA

Probabilistic methods are providing new explanatory approaches to fundamental cognitive science questions of how humans structure, process and acquire language. This review examines probabilistic models defined over traditional symbolic structures. Language comprehension and production involve probabilistic inference in such models; and acquisition involves choosing the best model, given innate constraints and linguistic and other input. Probabilistic models can account for the learning and processing of language, while maintaining the sophistication of symbolic models. A recent burgeoning of theoretical developments and online corpus creation has enabled large models to be tested, revealing probabilistic constraints in processing, undermining acquisition arguments based on a perceived poverty of the stimulus, and suggesting fruitful links with probabilistic theories of categorization and ambiguity resolution in perception.

Probability in language

The processing and acquisition of language is a central topic in cognitive science. Yet, perhaps surprisingly from the perspective of this Special Issue (see also Conceptual Foundations Editorial), the first steps towards a cognitive science of language involved driving out, rather than building on, probability. Whereas structural linguistics focussed on finding regularities in language corpora, the Chomskyan revolution focussed on the abstract rules governing linguistic ‘competence’, based on judgements of linguistic acceptability [1]. Whereas behaviourism viewed language as a stochastic process determined by principles of reinforcement between stimuli and responses, the new psycholinguistics viewed language processing as governed by internally represented linguistic rules [2]. And interest in statistical and information-theoretic properties of language [3] was replaced by the mathematical machinery of formal grammar.

Thus, probability has suffered a bad press in the cognitive science of language. The focus on complex linguistic representations (feature matrices, trees, logical representations), and rules defined over them, has crowded out probabilistic notions. And the impression that probabilistic ideas are incompatible with the Chomskyan approach to linguistics has been reinforced by debates that appear to pitch probabilistic and related quantitative/connectionist approaches against the symbolic approach to language [4–7].

The development of sophisticated probabilistic models, such as described in this Special Issue, casts these issues in a different light. Such probabilistic models may be specified in terms of symbolic rules and representations, rather than being in opposition to them. Thus, grammatical rules may be associated with probabilities of use, capturing what is linguistically likely, not just what is linguistically possible. From this viewpoint, probabilistic ideas augment symbolic models of language [8,9].

Yet this complementarity does not imply that probabilistic methods merely add to symbolic work, without modification. On the contrary, the ‘probabilistic turn’, broadly characterized, has led to some radical re-thinking in the cognitive science of language, on several levels (see Table 1).

In linguistics, there has been renewed interest in phenomena that seem inherently graded and/or stochastic, from phonology to syntax [10–12] – this linguistic work is complementary to the focus of Chomskyan linguistics (Table 1, first row). There have also been ‘revisionist’ perspectives on the strict symbolic rules thought to underlie language (Table 1, second row). Although inspired by a type of probabilistic connectionist network, standard optimality theory attempts to define a middle ground of ranked, violable linguistic constraints, used particularly to explain phonological regularities [13]. However, it has also been extended into increasingly rich probabilistic variants. And in morphology, there is debate over whether ‘rule + exception’ regularities (e.g. English past tense, German plural) are better explained by a single stochastic process [14].

Although it touches on these issues, this review explores a narrower perspective: the idea that language is represented by a probabilistic model [9], that language processing involves generating or interpreting using this model, and that language acquisition involves learning probabilistic models (Table 1, rows 3 and 4).
From grammar to probabilistic models

To see the contribution of probability, let us begin without it. According to early Chomskyan linguistics, language is internally represented as a grammar: a system of rules that specifies all and only allowable sentences. Thus, parsing is viewed as the problem of inferring an underlying linguistic tree, \( t \in T \), from the observed strings of words, \( s \in S \). Yet natural language is notoriously ambiguous – there are many ways in which local chunks can be parsed, and exponentially many ways in which these parses can be stitched together to produce a global parse. Searching these possibilities is hugely challenging; and there are often many globally possible parses (many \( t \), for a single \( s \)). The problem gets dramatically easier if the cognitive system knows that the bracketing [the old [man]] is much more likely than [[the old] man] (although this latter reading is possible, as in the old man the boats). This helps locally prune the search space; and helps decide between interpretations for globally ambiguous sentences. In particular, Bayesian methods specify a framework showing how information about the probability of generating different grammatical structures, and their associated word strings, can be used to infer grammatical structure from a string of words. This Bayesian framework is analogous to probabilistic models of vision, inference and learning; what is distinctive is the specific structures (e.g. trees, dependency diagrams) relevant for language.

In computational linguistics, the practical challenge of parsing and interpreting corpora of real language (typically text, sometimes speech) has led to a strong focus on probabilistic methods (Table 2). However, computational linguistics often parts company from standard linguistic theory, which focuses on much more complex grammatical frameworks, where probabilistic and other computational methods cannot readily be applied (see Box 1 for discussion). But computational linguistics does, we suggest, provide a valuable source of hypotheses for the cognitive science of language.

Formally, probabilistic parsing involves estimating \( \Pr_m(t|s) \) – estimating the likelihood of different trees, \( t \), given a sentence, \( s \), and given a probabilistic model \( \Pr_m \) of the language (see the online article by Griffiths and Yuille for Technical Introduction: Supplementary material online). This quantity can be evaluated by using Bayes’ theorem:

\[
\Pr_m(t|s) = \frac{\Pr_m(t,s)}{\sum_i \Pr_m(t',s)}
\]

The probabilistic model can take as many forms as there are linguistic theories (and linguistic structures, \( t \), may equally be trees, attribute-value matrices, dependency diagrams, etc.). For simplicity, suppose that our grammar is a context-free phrase-structure grammar, defined by rules such as those in Figure 1a. The bracketed numbers indicate the probabilities of expanding each node using a given rule. The product of probabilities in a derivation gives the overall probability of that tree (Figures 1b and 1c).

This grammar fragment encodes a syntactic ambiguity concerning prepositional phrase attachment that has been much studied in psycholinguistics. The parser has to decide: does the prepositional phrase (e.g. ‘with the telescope’) modify the verb phrase describing the girl’s action (i.e. she saw-with-a-telescope the boy); or the noun phrase the boy (i.e. she saw the-boy-with-a-telescope)? This question is a useful starting point for discussing the role of probability in the cognitive science of language.

Principles, probability and plausibility in parsing

Classical proposals in psycholinguistics assumed that disambiguation occurs using structural features of the trees. For example, the principle of minimal attachment
would prefer the first reading, because it has one less node [18]. The spirit of this proposal could, however, be recast probabilistically: the probability of a tree is the product of the probabilities at each node; and hence, other things being equal, fewer nodes imply higher probability. This is illustrated using the (arbitrary) probabilities in Figure 1: the key structural difference is highlighted to the right of the trees – all other structure, and its probability, is shared.

Structural principles in parsing have come under threat from the variety of parsing preferences observed within and across languages. But a stochastic grammar can capture parsing-preference variation across languages, because the probability of different structures may differ across languages. A structure with fewer nodes, but using highly improbable rules (estimated from a corpus) will be dispreferred. Psycholinguists are increasingly exploring corpus statistics across languages, and parsing preferences seem to fit the probabilities evident in each language [19,20].

A second problem for structural parsing principles is the influence of lexical information. Thus, the preference for the structurally analogous ‘the girl saw the boy with a book’ appears to reverse, because books, unlike telescopes, are not aids to sight. The pattern flips back with a change of verb: ‘the girl hit the boy with a book’, because books can be aids to hitting. The probabilistic approach seems useful here because it is important to integrate the constraint that ‘seeing-with-telescopes’ is much more likely than ‘seeing-with-books’. But our particular stochastic grammar above does not help, because each node is expanded independently – the grammar is ‘context free’.

One way to capture these constraints aims to capture statistical (or even rigid) regularities between head words of phrases. For example, ‘lexicalized’ grammars, which

<table>
<thead>
<tr>
<th>Speech recognition</th>
<th>Representation</th>
<th>Model</th>
<th>Primary objective</th>
<th>Learning method</th>
</tr>
</thead>
<tbody>
<tr>
<td>[63]</td>
<td>Phonemes</td>
<td>Hidden Markov Models, Finite state models, with multiple levels</td>
<td>Describing phonological principles across languages; phonotactics</td>
<td>EM algorithm</td>
</tr>
<tr>
<td>Computational phonology [64]</td>
<td>Series of phonemes; Levels of autosegmental phonology</td>
<td>Bigrams; Language as a sequence of letter strings</td>
<td>Learning morphological structure from lists of words; relevance across languages</td>
<td>Simulated annealing search; minimum description length</td>
</tr>
<tr>
<td>Morphology [56,65,66]</td>
<td>Letter strings</td>
<td>Context-free phrase-structure grammar, and variants; n-gram based models</td>
<td>Broad coverage parsing; syntactic tagging; basis for machine translation, semantic analysis etc; automated discovery of syntactic categories</td>
<td>EM algorithm; correlating context 'vectors' and clustering</td>
</tr>
<tr>
<td>Syntax [22,43,47,67]</td>
<td>Syntactic categories for words; either flat or hierarchical syntactic structure</td>
<td>Bayesian mixture</td>
<td>Automated discovery of semantic relations</td>
<td>Markov Chain Monte Carlo; Singular value decomposition</td>
</tr>
<tr>
<td>Corpus based lexical semantics [57,68,69]</td>
<td>Word and 'bag' of surrounding words</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Computational models of language using probabilistic and statistical methods

Recent work has especially favoured the use of statistical methods for which a clear Bayesian analysis can be given, i.e. the inferential assumptions are specified by an explicit probabilistic model, and inference involves Bayesian updating over the model. Connectionist models of psycholinguistic phenomena (see [42]) have many features in common with probabilistic models, although the probabilistic assumptions they impose are not explicit.

Box 1. Linguistics, computational linguistics and cognitive science

The driving force in the development of many of the probabilistic methods discussed in this article has been the creation of practical computational systems for language processing – for recognizing speech, analysing or retrieving information in texts, question-answering, and machine translation. The goal here is getting systems to work, rather than modelling human language processing.

Computational linguistics has typically taken a fairly cavalier approach to existing linguistic theory. The explanatory goals of linguistics, attempting to account for linguistic patterns across languages, with speaker judgments as primary data, has yielded complex representations and principles, which are difficult to work with computationally. Computational linguists have instead focussed on simpler language models, based on finite state, or phrase-structure grammars and variants. Computationally, the emphasis on simple formalisms is guided by the need to parse, produce, learn and construct semantic representations robustly on real corpora. ‘Broad coverage’ grammars have tried to cope with real language use, while of necessity riding rough-shod over many linguistic subtleties. Yet the need to tackle ‘real language’ has also led to insights that might transfer more naturally to models of cognitive processes for robustly dealing with language, than do insights from traditional linguistic theory.

Early computational linguistic methods focussed primarily on capturing rigid linguistic constraints, but recent probabilistic methods have had a revolutionary impact [75]. In parsing, for example, probability helps resolve the massive local syntactic ambiguity of natural language, by focussing on the relatively small number of potential parses with significant probability (given what is known about the frequency of different structures across a corpus). Similarly, probabilistic methods in learning dramatically narrow the infinite set of grammatical rules that could generate a given set of sentences or structures. Probabilistic methods are increasingly widespread in the psychology of language acquisition and processing, and in linguistics [10,74], and human abilities to pick up probabilistic constraints have been extensively studied experimentally [76]. The practical success of probabilistic methods in computational linguistics suggests that human processing and acquisition might also exploit probabilistic information.
Figure 1. Ambiguity resolution in probabilistic parsing. (a) A simple stochastic phrase-structure grammar fragment – note that each symbol (e.g. NP) expands into one or more symbol sequences (Det Noun; NP PP) whose probabilities sum to 1. From a start symbol, here S, the application of a sequence of rules replaces the initial S with a sequence of words, and in doing so, generates a tree, such as those shown in (b) and (c). The probability of a tree is just the product of the probabilities of the rules required to generate that tree. Syntactic ambiguity arises because different trees can generate the same string of words, as (b) and (c) illustrate. According to a probabilistic approach to ambiguity resolution, the processor should prefer the parse with the highest probability. The alternative parse of the girl saw the boy with the telescope in (b) and (c) differ in whether the prepositional phrase (with a telescope) attaches to the verb phrase (the seeing is done with a telescope), or the object noun phrase (the boy has the telescope). The points at which the trees differ are shown to the right of the trees. Notice that the flatter structure for the first reading, which contains one less node (and hence one less syntactic rule), and has a higher probability.
carry information about what material co-occurs with specific words, substantially improve computational parsing performance [21,22].

**Plausibility and statistics**

Statistical constraints between words provide, however, a crude estimate of which sentences are plausible. In an offline judgement task, we use world knowledge, understanding of the social and environmental context, pragmatic principles, and much more, to determine what people might plausibly say or mean. Determining whether a statement is plausible may involve determining how likely it is to be true; but also whether, given the present context, it might plausibly be said. The first issue requires a probabilistic model of general knowledge [23] and Tenenbaum et al., this issue [24]. The second issue requires engaging ‘theory of mind’ (inferring the other’s mental states), and invoking principles of pragmatics. Computational models of these processes, probabilistic or otherwise are very preliminary [25].

A fundamental theoretical debate is whether plausibility is used on-line in parsing decisions. Are statistical dependencies between words used as a computationally cheap surrogate for plausibility? Or are both statistics and plausibility deployed on-line, perhaps in separate mechanisms? Eye-tracking paradigms [26,27] have been used to suggest that both factors are used on-line, although the interpretation of the data is controversial. Recent work indicates that probabilistic grammar models often predict the time course of processing [28–30], although parsing preferences also appear to be influenced by additional factors, including the linear distance between the incoming word and the prior words to which it has a dependency relation [31].

**Is the most likely parse favoured?**

In the probabilistic framework, it is typically assumed that on-line ambiguity resolution favours the most probable parse. Yet Chater, Crocker and Pickering [32] suggest that, for a serial parser, whose chance of ‘recovery’ is highest if the ‘mistake’ is discovered soon, this is oversimple. In particular, they suggest that because parsing decisions are made on-line [26], there should be a bias to choose interpretations which make specific predictions, that might rapidly be falsified. For example, after ‘John realized his...’ the more probable interpretation is that realized introduces a reduced relative clause (i.e. ‘John realized (that) his...’). On this interpretation, the rest of the noun phrase after his is unconstrained. By contrast, the less probable transitive reading (‘John realized his goals/potential objectives’) places very strong constraints on the subsequent noun phrase. Perhaps, then, the parser should favour the more specific reading, because if wrong, it may rapidly and successfully be corrected. Chater et al. [32] provide a Bayesian analysis of ‘optimal ambiguity resolution’ capturing such cases. The empirical issue of whether the human parser follows this analysis [33], and even the correct probabilistic analysis of sentences of this type [34], is not fully resolved.

**Beyond parsing**

We have here focussed on parsing. But the ‘probabilistic turn’ applies across language processing, from modelling lexical semantics to modelling processing difficulty (see Table 3). Note, though, that integrating these diverse approaches into a unified model of language is extremely challenging; and many of the theoretical issues that have traditionally concerned psycholinguists are re-framed rather than resolved by a probabilistic approach (e.g. the relation between understanding and production becomes: how far are the relevant probabilistic models shared? (see Box 2); the issue of the degree of modularity between separate processes becomes: how far are cognitive models of different levels of linguistic analysis probabilistically independent?). Probability might prove important as a unifying theoretical framework for understanding how
Box 2. Probabilistic models, Bayes and the ‘reversibility’ of language processing

If the cognitive system uses a probabilistic model in language processing, then it can infer the probability of a word (or parse/interpretation) from speech input. It does this from the reverse probability: the probability that linguistic input, given the parse, together with the prior probability of each possible parse (see Figure 1).

This pattern is an instance of the more general principle that Bayesian approaches to recognition typically involve analysis-by-synthesis (see Yuille and Kersten, this issue) [77]. That is, the mapping from low- to high-level representation (e.g. from acoustic to word-level) is computed using the reverse mapping, from high- to low-level representation. This pattern is standard in Bayesian models of perception, but it also has the interesting additional feature that the structure being modelled (the production of speech, rather than the production of natural acoustic or visual stimuli) is typically part of a person’s cognitive equipment. Indeed, not only do people produce speech, but as with other motor outputs, it is likely that they can compute a ‘forward model’ for predicting the acoustic consequences of their own speech, before the motor output is given. This forward model is presumed to be useful in feedforward control of the speech apparatus (see Körding and Wolpert, this issue [78], for a discussion of the general motor control case); and the phenomenology of ‘inner voices’, whether in normal imagery or mental illness, might arise from its functioning. This perspective is a return to the motor theory of speech perception. Analysis-by-synthesis also opens up a possible mechanism for top-down influences on speech perception, although empirical evidence that such effects occur on-line is mixed [79].

Details aside, the Bayesian approach raises the possibility that there may be substantial sharing of information between producing and understanding speech. Indeed, there is substantial behavioural and neuropsychological evidence that the levels of processing in comprehension and production are intricately linked (e.g. [80]). For example, despite superficial asymmetries between reception and production of language, it seems that people are roughly able to understand the linguistic forms they can generate. The apparent asymmetry is explicable because ‘guessing’ using background knowledge can successfully recover meaning, but guessing is unlikely to yield linguistically correct output (although see [81]). In summary, we see that what might be a deep inter-relationship between language understanding and production is, at a more general level, a natural consequence of the more general idea that the cognitive system constructs a probabilistic model of the language.

The cognitive system makes the uncertain inference from speech signal to message, and vice versa. As we now see, it may also help understand how, and to what extent, learners infer language structure from linguistic input.

Probabilistic perspectives on language acquisition

Probabilistic language processing presupposes a probabilistic model of the language; and uses that model to infer, for example, how sentences should be parsed, or ambiguous words interpreted. But how is such a model, or for that matter a traditional non-probabilistic grammar, acquired? Chomsky [1] frames the problem as follows: the child has a hypothesis-space of candidate grammars; and must choose, on the basis of (primarily linguistic) experience one of these grammars. From a Bayesian standpoint, each candidate grammar is associated with a prior probability; and these probabilities will be modified by experience using Bayesian updating (see Griffiths and Yuille Technical Introduction: Supplementary material online). The learner will presumably choose a language with high, and perhaps the highest, posterior probability.

The poverty of the stimulus?

Chomsky [1] influentially argued that the learning problem is unsolvable without strong prior constraints on the language, given the ‘poverty’ (i.e. partiality and errorfulness) of the linguistic stimulus. Indeed, Chomsky [35] argued that almost all syntactic structure, aside from a finite number of binary parameters, must be innate. Separate mathematical work by Gold [36] indicated that, under certain assumptions, learners provably cannot converge on a language even ‘in the limit’ as the corpus becomes indefinitely large (see [37], for discussion).

A probabilistic standpoint yields more positive learnability results. For example, Horning [38] proved that phrase-structure grammars are learnable (with high probability) to within a statistical tolerance, if sentences are sampled as independent, identically distributed data. Chater and Vitányi generalize to a language that is generated by any computable process (i.e. sentences can be interdependent, and generated by any computable grammar; see [39] for a brief summary), and show that prediction, grammaticality and semantics are learnable, to a statistical tolerance. These results are ‘ideal’ however; that is, they consider what would be learned if the learner could find the shortest representation of linguistic data. In practice, the learner will find a short code, not the shortest, and theoretical results are not available for this case. Nonetheless, from a probabilistic standpoint, learning looks more tractable – partly because learning need only succeed with high probability; and to an approximation (speakers might learn slightly different idiolects).

Computational models of language learning

Yet the question of learnability, and the potential need for innate constraints, remains. Machine learning methods have successfully learned small artificial context-free languages (e.g., [40]), but profound difficulties in extending these results to real language corpora have led
computational linguists to focus on learning from parsed trees [21,22] – presumably not available to the child. Connectionism is no panacea here; indeed, connectionist simulations of language learning typically use small artificial languages [41,42], and, despite having considerable psychological interest, they often scale poorly.

By contrast, many simple but important aspects of language structure have successfully been learned from linguistic corpora by distributional methods. For example, good approximations to syntactic categories and semantic classes have been learned by clustering words based on their linear distributional contexts (e.g. the distribution over the word that precedes and follows each token of a type) or broad topical contexts (e.g. [43,44]) (see Figure 2).

One can even simultaneously cluster words exploiting local syntactic and topical similarity [45].

Recently, however, Klein and Manning [46,47] have made significant progress in solving the problem of learning syntactic constituency from corpora of unparsed sentences. Klein and Manning [46] extended the success of distributional clustering methods for learning word classes by using the left and right word context of a putative constituent and its content as the basis of similarity calculations. Such a model better realizes ideas from traditional linguistic constituency tests which emphasize (i) the external context of a phrase (‘something is a noun phrase if it appears in noun phrase contexts’) at least as much as its internal structure, and (ii) proform

**Figure 2. Clustering words into syntactic classes by context.** (a) shows the modification to a table of co-occurrences as a moving window (b) passes over the text, centred on a target word. Here, separate counts are made for context words in four different locations – two slots before and after the target word. Each target word is then associated with a ‘context vector’ consisting of the counts for an entire corpus, corresponding to columns in the table in (a). Target words are then clustered based on the similarity of these vectors, leading to an overall clustering into syntactic categories, and a rich fine-grained structure showing a mixture of syntactic and semantic factors. An adjective subcluster is illustrated in (c). This method is used in [44], from which (c) is reproduced with permission.

One can even simultaneously cluster words exploiting local syntactic and topical similarity [45].

**Figure 3. Unsupervised grammar induction.** The task of grammar induction can be thought of as two correlated tasks: learning the constituents in text and learning modification or dependency relationships between words. Klein and Manning’s grammar induction system [47] exploits two representations focussed on these tasks. (a) Distributional word clustering techniques are extended to phrasal constituents by performing clustering over a representation that focuses on the content and context of both putative constituents (the upper example is a constituent, but the lower example is not). (b) The model over word dependency structures. The model includes the directionality, distance and count of dependents. Both these models are learnt using the expectation-maximization algorithm, and are then combined to give a unified probabilistic model of grammar induction.
tests (testing replacing a large constituent with a single word member of the same category). Klein and Manning [47] extended this work by combining such a distributional phrase clustering model with a dependency-grammar-based model (see Figure 3). The dependency model uses data on word co-occurrence to bootstrap word-word dependency probabilities, but the work crucially shows that more is needed than simply a model based on word co-occurrence. One appears to need two types of prior constraint: one making dependencies more likely between nearby words than far away words, and the other making it more likely for a word to have few rather than many dependents. Both of Klein and Manning’s models capture a few core features of language structure, while still being simple enough to support learning. The resulting combined model is better than either model individually, suggesting a certain complementarity of knowledge sources. Klein and Manning show that high-quality parses can be learned from surprisingly little text, from a range of languages, with no labeled examples and no language-specific biases. The resulting model provides good results, building binary trees which are correct on over 80% of the constituency decisions in hand-parsed English text.

This work is a promising demonstration of empirical language learning, but most linguistic theories use richer structures than surface phrase structure trees; and a particularly important objective is finding models that map to meaning representations. This remains very much an area of ongoing research, but *inter alia*, there is work on probabilistic parsing with richer formalized grammar models based on learning from parsed data [48,49], some work on mapping to meaning representations of simple datasets [50], and work on unsupervised learning of a mapping from surface text to semantic role representations [51].

**Poverty of the stimulus, again…**

The status of Chomsky’s poverty of the stimulus argument remains unclear, beginning with the question of whether children really do face a poverty of linguistic data (see the debate between [52] and [53]). Perhaps no large and complex grammar can be learned from the child’s input; or perhaps certain specific linguistic patterns (e.g. those encoded in an innate universal grammar) are in principle unlearnable. Probabilistic methods provide a potential way of assessing such questions. Oversimplifying somewhat, suppose that a learner wonders whether to include constraint C in her grammar. C happens, perhaps coincidentally, to fit all the data so far encountered. If the learner does not assume C, the probability that each sentence will happen to fit C by chance is p. Thus, each sentence obeying C is 1/p times more probable, if the constraint is true than if it is not (if we simply rescale the probability of all sentences obeying the constraint). Thus, after n sentences, the probability of the corpus, is 1/p^n greater, if the constraint is included. Yet, a more complex grammar will typically have a lower prior probability. If the ratio of priors for grammars with/without the constraint is greater than 1/p^n, then, by Bayes’ theorem, the constraint is unlearnable in n items.

Presently, theorists using probabilistic methods diverge widely on the severity of the prior ‘innate’ constraints they assume. Some theorists focus on applying probability to learning parameters of Chomsky’s Universal Grammar [54,55]; others focus on learning relatively simple aspects of language, such as syntactic or semantic categories, or approximate morphological decomposition, with relatively weak prior assumptions [44,56,57]. Probabilistic methods should be viewed as a framework for building and evaluating theories of language acquisition, and for concretely formulating questions concerning the poverty of the stimulus, rather than as embodying any particular theoretical viewpoint. This point arises throughout cognition; although probability provides natural models of learning, it is an open question whether initial structure is crucial in facilitating such learning. For example, Tenenbaum et al. [24] argue that prior structure over Bayesian networks is crucial to support learning.

**Language acquisition and language structure**

How far do probabilistic perspectives on language structure and language acquisition interact? Some theorists argue that language should not best be described as rules and exceptions, but as a system of graded ‘quasi-regular’ mappings (this is ‘revisionist’ probabilistic linguistics; Table 1). Notable examples of such mappings including the English past-tense, the German plural, and spelling-to-sound correspondences in English; but a closely related viewpoint has been advocated for syntax [58,59] and aspects of semantics [60]. Some theorists argue [13] that such mappings are better learned using statistical or connectionist methods, which learn according to probabilistic principles. By contrast, traditional rule-and-exception views are typically associated with non-probabilistic hypothesis generation and testing. Nonetheless, we see no necessary connection between these debates on the structure of language and models of acquisition.

---

### Box 3. Open questions

- Are the same probabilistic model and computational processes used in language comprehension and production? (see also Box 2). How does the picture change for comprehension based on pragmatics, world knowledge and ‘theory of mind’?
- Is local ambiguity handled by using a single underspecified representation; or by pursuing distinct parses in parallel or in sequence?
- Over what levels of representation (words, word classes, structures) is frequency information represented by the language processor?
- How far is speech and language optimized for communication? What features of language (e.g. the brevity of common words; nature of local ambiguity) might such optimization explain?
- How are convergent sources of linguistic information exploited in learning and processing?
- How can non-linguistic cues from the social and physical environment be exploited by the child?
- Can specific features of language be proved to be unlearnable from the input available to the child, using the probabilistic arguments discussed here, or other methods.
Conclusion
Understanding and producing language involves complex patterns of uncertain inference, from processing noisy and partial speech input to lexical identification, syntactic and semantic analysis, to language interpretation in context. Acquiring language involves uncertain inference from linguistic and other data, to infer language structure. These uncertain inferences are naturally framed using probability theory: the calculus of uncertainty. Historically, probabilistic approaches to language are associated with simple models of language structure (e.g. local dependencies between words), but, across the cognitive sciences, as described in this special issue, technical advances have reduced this type of limitation. Probabilistic methods are also often associated with empiricist views of language acquisition. But the framework is equally compatible with nativism – that there are prior constraints on the class of language models. Indeed, as we have seen, probabilistic analysis can provide one line of attack (alongside the empirical investigation of child language) in assessing the relative contributions of innate constraints and corpus input in language acquisition. Overall, we view probabilistic methods as providing a rich framework for theorizing about language structure, processing and acquisition, which may prove valuable in developing, and contrasting between, a wide range of theoretical perspectives (see also Box 3, and Editorial ‘Where next?’ in this issue).

Acknowledgements
We would like to thank Ted Gibson, Harald Baayen and an anonymous reviewer for comments on this paper. Nick Chater was partially supported by ESRC grant RES-000–22–1120 and the Human Frontiers Science Program. Christopher Manning was supported in part by the Advanced Research and Development Activity (ARDA)’s AQUAINT Program and by an IBM Faculty Partnership Award.

Supplementary data
Supplementary data associated with this article can be found at doi:10.1016/j.tics.2006.05.006

References
11 Fanselow, G. et al., eds., Gradience in Grammar: Generative Perspectives, Oxford University Press (in press)
16 Norris, D. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. Psychol. Rev. (in press)


www.sciencedirect.com

Reuse of Current Opinion and Trends journal figures in multimedia presentations

It’s easy to incorporate figures published in Trends or Current Opinion journals into your PowerPoint presentations or other image-display programs. Simply follow the steps below to augment your presentations or teaching materials with our fine figures!

1. Locate the article with the required figure in the Science Direct journal collection
2. Click on the ‘Full text + links’ hyperlink
3. Scroll down to the thumbnail of the required figure
4. Click on the ‘Full text + links’ hyperlink
5. On a PC, right-click over the expanded image and select ‘Copy’ from pull-down menu (Mac users: hold left button down and then select the ‘Copy image’ option)
6. Open a blank slide in PowerPoint or other image-display program
7. Right-click over the slide and select ‘paste’ (Mac users hit ‘Apple-V’ or select the ‘Edit-Paste’ pull-down option).

Permission of the publisher, Elsevier, is required to re-use any materials in Trends or Current Opinion journals or any other works published by Elsevier. Elsevier authors can obtain permission by completing the online form available through the Copyright Information section of Elsevier’s Author Gateway at http://authors.elsevier.com/. Alternatively, readers can access the request form through Elsevier’s main web site at http://www.elsevier.com/locate/permissions.