Review TRENDS in Cognitive Sciences Vol.10 No.7 July 2006

Special Issue: Probabilistic models of cognition

Probabilistic models of language processing and acquisition

Nick Chater[1] and Christopher D. Manning[2]

[1]Department of Psychology, University College London, Gower Street, London, WC1E 6BT, UK
[2]Stanford 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). (Another interesting line of work that we do not review assumes instead that language processing is based on memory for

Corresponding authors: Chater, N. (n.chater@ucl.ac.uk); Manning, C.D. (manning@cs.stanford.edu).

Available online 19 June 2006

 1364-6613/$ - see front matter doi:10.1016/j.tics.2006.05.006

Table 1. Applications of probability in language

Type of explanation	Probabilistic perspective	Examples	Non-probabilistic alternative
Probabilistic linguistics	*Complementary*: Describing language variability	Phonetic variation. [61] Corpus counts of different syntactic structures. Sociolinguistic variation [62].	Proper scope of linguistics is competence; assign probability to performance [1]
	Revisionist: Probabilistic versus rigid linguistic rules	Status of rules / subrules / exceptions in morphology [7,14] Gradedness of grammaticality judgements [11,12]	To restrict linguistics to core competence grammar, where intuitions are clear [35].
Probabilistic models of cognitive processes	Language processing	Stochastic phrase-structure grammars and related methods [29] Connectionist models [42]	Assume that structural principles guide processing, e.g. minimal attachment [18]
	Language acquisition	Probabilistic algorithms for grammar learning [46,47] Theoretical learnability results [38,39] Bayesian word learning [17]	Trigger-based acquisition models [54] Identification in the limit [36]

past instances and not via the construction of a model of the language [15]). Moreover, for reasons of space, we shall focus mainly on parsing and learning *grammar*, rather than, for example, exploring probabilistic models of how words are recognized [16] or learned [17]. We will see that a probabilistic perspective adds to, but also substantially modifies, current theories of the rules, representations and processes underlying language.

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:

$$\Pr_m(t|s) = \frac{\Pr_m(t,s)}{\sum_{t'} \Pr_m(t',s)}$$

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

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

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

[a]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.

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

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.

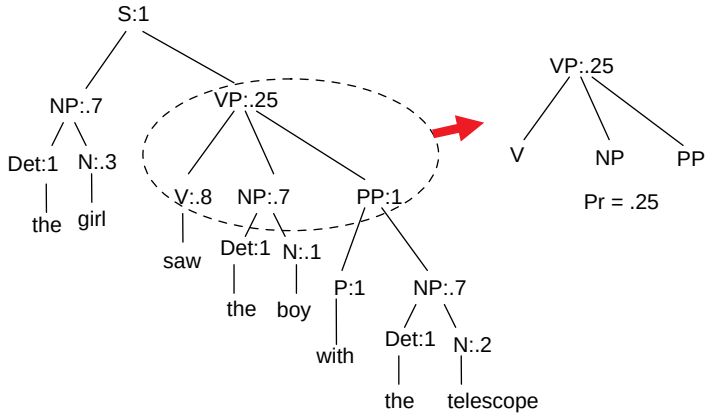
Computational linguistics has typically taken a fairly cavalier approach to existing linguistic theory. The explanatory goals of linguistics, attempting to account for linguistic patterns across languages, with speaker judgments as primary data, has yielded complex representations and principles, which 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.

(a)

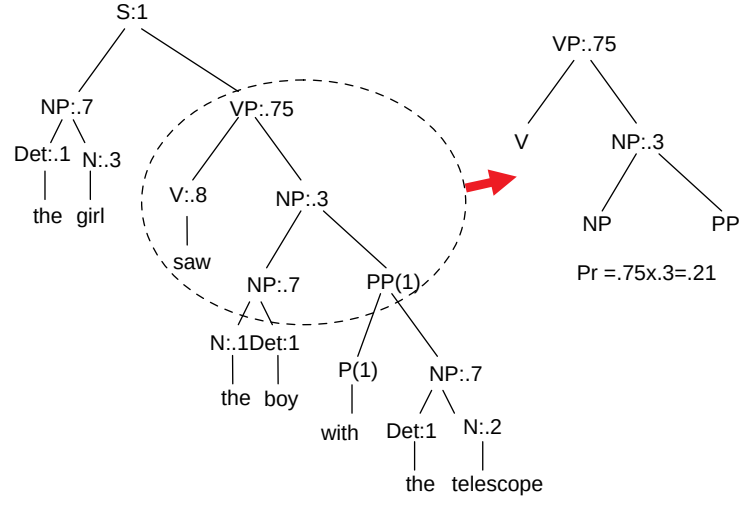
S→ NP VP	(1)	V→ saw	(.8)	N→ cat	(.1)
VP → V NP	(.75)	V → prodded	(.2	Det → the	(1)
VP → V NP PP	(.25)	N → telescope	(.2)	P → with	(1)
NP → Det Noun	(.7)	N → stick	(.3)		
NP → NP PP	(.3)	N → girl	(.3)		
PP → P NP	(1)	N → boy	(.1)		

(b)



Pr(tree) = 1 x .7 x 1 x .3 x .25 x .8 x .7 x 1 x .1 x 1 x 1 x .7 x 1 x .2 ≈ 0.00041

(c)



Pr(tree) = 1 x .7 x 1 x .3 x .75 x .8 x .3 x .7 x 1 x .1 x 1 x 1 x .7 x 1 x .2 ≈ 0.00037

TRENDS in Cognitive Sciences

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 parses 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.

Table 3. Probabilistic methods applied across a wide range of domains in the cognitive science of language

	Theoretical framework	Sub-topics	Empirical data
Speech processing and word recognition	Connectionist models [70] Probabilistic phonetics [71]	Feature integration Incremental on-line word recognition	'soft' integration of features analysis by synthesis [78]
Probabilistic phonology	Stochastic optimality theory [72] *N*-grams+finite state models [64] Exemplar models	Stochastic optimality theory Probabilistic phonotactics	Graded linguistic judgements
Morphology	Connectionism [14] Exemplar models [66]	Regularities/subregularities/exceptions Level of morphological generalizations	Data on acceptability Linguistic data
Syntax	Probabilistic parsing [28] Identifying linguistic classes [44] Connectionism [42]	Integration of information resolving local ambiguity Recursion	Graded linguistic judgements Eye-tracking data Reading times [30,73]
Lexical semantics	Distributional analysis [57] Bayesian networks [45]	Finding word classes from corpora Relating words to 'world'	Acquisition data Semantic priming
Acquisition	Learnability VC dimension; Minimum description length [17,55,74]	Learning parameters, grammar, word meanings	Corpus data; experimental data; linguistic data

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 off-line 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 of that linguistic input, given the parse, together with the prior probability of each possible parse (see Figure I).

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.

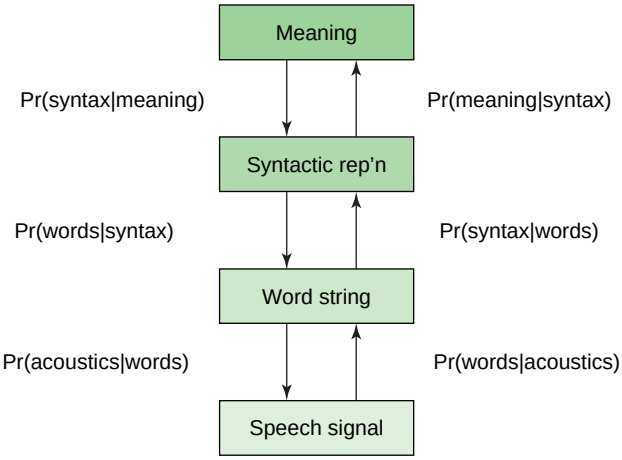


Figure I. The reversibility of language processing. See text for explanation.

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

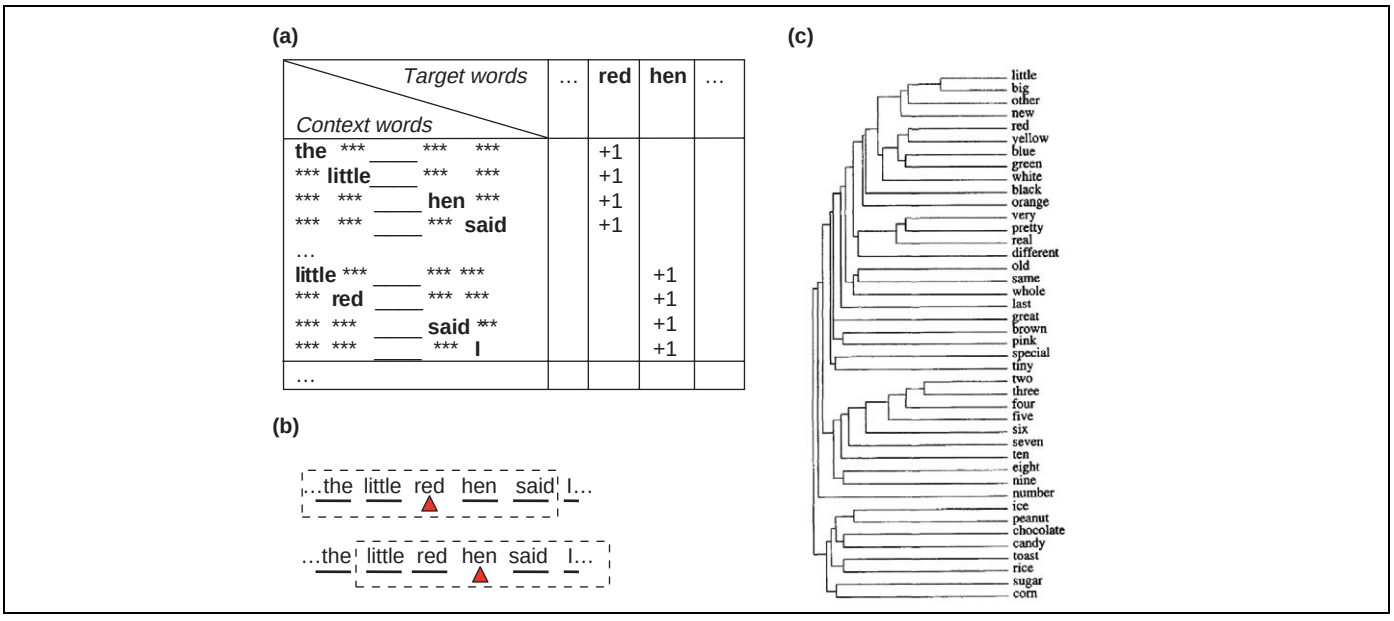


Figure 2. Clustering words into syntactic classes by context. (a) shows the modification to a table of co-occurrences as a moving window **(b)** passes over the text, centred on a target word. Here, separate counts are made for context words in four different locations – two slots before and after the 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 theses 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.

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

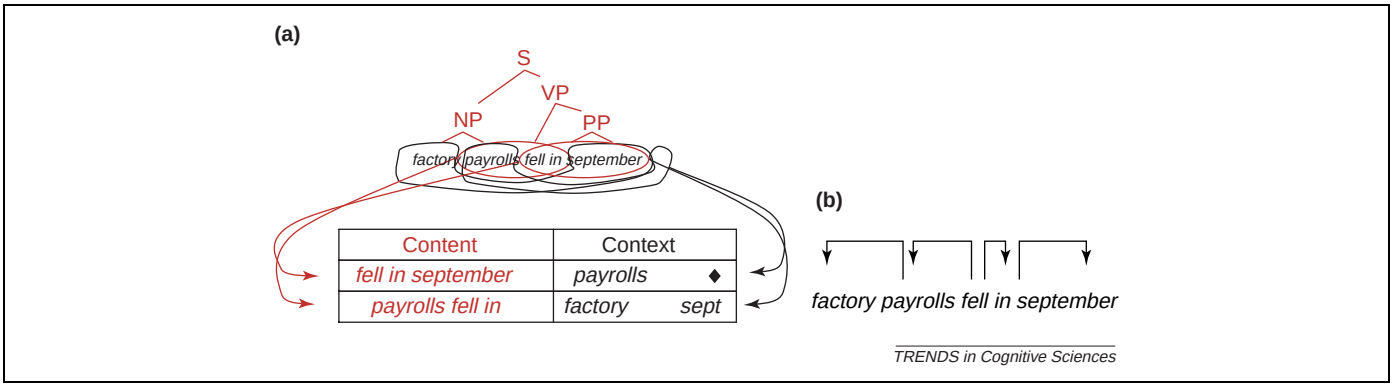


Figure 3. Unsupervised grammar induction. The task of grammar induction can be thought of as two correlated tasks: learning the constituents in text and learning modification or dependency relationships between words. Klein and Manning's grammar induction system [47] exploits two representations focussed 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 Chomskyan 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

1 Chomsky, N. (1965) *Aspects of the Theory of Syntax*, MIT Press
2 Fodor, J.A. *et al*. (1974) *The Psychology of Language*, McGraw-Hill
3 Shannon, C.E. (1951) Prediction and entropy of printed English. *Bell System Technical Journal* 30, 50–64
4 Marcus, G.F. *et al*. (1999) Rule learning by seven-month-old infants. *Science* 283, 77–80
5 Pinker, S. (1999) *Words and rules: The Ingredients of Language*, Basic Books
6 Seidenberg, M.S. (1997) Language acquisition and use: learning and applying probabilistic constraints. *Science* 275, 1599–1603
7 Seidenberg, M.S. and Elman, J.L. (1999) Do infants learn grammar with algebra or statistics? *Science* 284, 434–435
8 Klavans, J. and Resnik, P., eds. (1996) *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, MIT Press
9 Manning, C. (2003) Probabilistic Syntax. In *Probabilistic Linguistics* (Bod, R. *et al*., eds), pp. 289–341, MIT Press
10 Bod, R. *et al.*, eds. (2003) *Probabilistic Linguistics*, MIT Press
11 Fanselow, G. *et al*. eds., *Gradience in Grammar: Generative Perspectives*, Oxford University Press (in press)
12 Hay, J. and Baayen, H. (2005) Shifting paradigms: gradient structure in morphology. *Trends Cogn. Sci.* 9, 342–348
13 Smolensky, P. and Legendre, G. (2006) *The Harmonic Mind* (*2 Vols*), MIT Press
14 Hahn, U. and Nakisa, R. (2000) German inflection: Single route or dual route? *Cogn. Psychol.* 41, 313–360
15 Daelemans, W. and van den Bosch, A. (2005) *Memory-based Language Processing*, Cambridge University Press
16 Norris, D. The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychol. Rev.* (in press)
17 Xu, F. and Tenenbaum, J.B. (2005) Word learning as Bayesian inference: Evidence from preschoolers. In *Proc. 27th Annu. Conf. Cogn. Sci. Soc.* Erlbaum
18 Frazier, L. and Fodor, J.D. (1978) The sausage machine: A new two-stage parsing model. *Cognition* 13, 187–222
19 Desmet, T. *et al*. Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Lang. Cogn. Process.* (in press)
20 Desmet, T. and Gibson, E. (2003) Disambiguation preferences and corpus frequencies in noun phrase conjunction. *J. Mem. Lang.* 49, 353–374
21 Charniak, E. (1997) Statistical parsing with a context-free grammar and word statistics. In *Proc. 14th Natl. Conf. Artif. Intell.* pp. 598–603, AAAI Press
22 Collins, M. (2003) Head-driven statistical models for natural language parsing. *Comput. Linguist.* 29, 589–637
23 Oaksford, M. and Chater, N. (1998) *Rationality in an Uncertain World*, Psychology Press
24 Tenenbaum, J.B. *et al*. (2006) Theory-based Bayesian models of inductive learning and reasoning. *Trends Cogn. Sci.* DOI:10.1016/j.tics.2006.05.009
25 Jurafsky, D. (2005) Pragmatics and computational linguistics. In *Handbook of Pragmatics* (Horn, L.R. and Ward, G., eds), pp. 578–604, Blackwell
26 Tanenhaus, M.K. *et al*. (1995) Integration of visual and linguistic information in spoken language comprehension. *Science* 268, 632–634
27 McDonald, S.A. and Shillcock, R.C. (2003) Eye movements reveal the on-line computation of lexical probabilities. *Psychol. Sci.* 14, 648–652
28 Jurafsky, D. (1996) A probabilistic model of lexical and syntactic access and disambiguation. *Cogn. Sci.* 20, 137–194
29 Narayanan, S. and Jurafsky, D. (2002) A Bayesian model predicts human parse preference and reading time in sentence processing. In *Advances in Neural Information Processing Systems Vol. 14* (Dietterich, T.G. *et al*., eds), pp. 59–65, MIT Press
30 Hale, J. (2003) The information conveyed by words in sentences. *J. Psycholinguist. Res.* 32, 101–123
31 Grodner, D. and Gibson, E. (2005) Consequences of the serial nature of linguistic input. *Cogn. Sci.* 29, 261–291
32 Chater, N. *et al*. (1998) The rational analysis of inquiry: The case of parsing. In *Rational Models of Cognition* (Oaksford, M. and Chater, N., eds), pp. 441–469, Oxford University Press
33 Pickering, M.J. *et al*. (2000) Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *J. Mem. Lang.* 43, 447–475
34 Crocker, M.W. and Brants, T. (2000) Wide-coverage probabilistic sentence processing. *J. Psycholinguist. Res.* 29, 647–669
35 Chomsky, N. (1981) *Lectures on Government and Binding*, Foris, Dordrecht
36 Gold, E.M. (1967) Language identification in the limit. *Information and Control* 10, 447–474
37 Pinker, S. (1979) Formal models of language learning. *Cognition* 7, 217–283
38 Horning, J. (1971) A procedure for grammatical inference. In *Proc. IFIP Congress 71*, pp. 519–523, North Holland
39 Chater, N. (2004) What can be learned from positive data? Insights from an 'ideal learner'. *J. Child Lang.* 31, 915–918
40 Lari, K. and Young, S.Y. (1990) The estimation of stochastic context-free grammars using the inside-outside algorithm. *Comput. Speech Lang.* 4, 35–56
41 Elman, J.L. (1990) Finding structure in time. *Cogn. Sci.* 14, 179–211
42 Christiansen, M.H. and Chater, N., eds (2001) *Connectionist Psycholinguistics*, Ablex
43 Schütze, H. (1995) Distributional part-of-speech tagging. In *Proc. 7th Conf. Eur. Chapter Assoc. Comput. Linguist.,* pp. 141–148
44 Redington, M. *et al*. (1998) Distributional information: A powerful cue for acquiring syntactic categories. *Cogn. Sci.* 22, 425–469

45 Griffiths, T.L. *et al*. (2005) Integrating topics and syntax. In *Advances in Neural Information Processing Systems Vol. 17* (Saul, L.K. *et al*., eds), pp. 537–544, MIT Press
46 Klein, D. and Manning, C. (2002) A generative constituent-context model for improved grammar induction. In *Proc. 40th Annual Meeting Assoc. Comput. Linguist.*, pp. 128–135
47 Klein, D. and Manning, C. (2004) Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. 42nd Annual Meeting Assoc. Comput. Linguist.*, pp. 479–486
48 Johnson, M. and Riezler, S. (2002) Statistical models of language learning and use. *Cogn. Sci.* 26, 239–253
49 Toutanova, K. *et al*. (2005) Stochastic HPSG parse disambiguation using the Redwoods corpus. *Res. Lang. Comput.* 3, 83–105
50 Zettlemoyer, L.S. and Collins, M. (2005) Learning to map sentences to logical form: Structured classification with probabilistic categorical grammars. In *Proc. 21st Conf. Uncertainty Artif. Intell.* (UAI-05), pp. 658–666
51 Swier, R. and Stevenson, S. (2005) Exploiting a verb lexicon in automatic semantic role labelling. In *Proc. Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (HLT/EMNLP-05), pp. 883–890
52 Pullum, G. and Scholz, B. (2002) Empirical assessment of stimulus poverty arguments. *Linguist. Rev.* 19, 9–50
53 Legate, J.A. and Yang, C.D. (2002) Empirical re-assessment of stimulus poverty arguments. *Linguist. Rev.* 19, 151–162
54 Gibson, E. and Wexler, K. (1994) *Triggers. Linguist. Inq.* 25, 407–454
55 Niyogi, P. (2006) *The Computational Nature of Language Learning and Evolution*, MIT Press
56 Brent, M. and Cartwright, T. (1996) Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition* 61, 93–125
57 Landauer, T.K. and Dumais, S.T. (1997) A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–140
58 Culicover, P.W. (1999) *Syntactic Nuts*, Oxford University Press
59 Tomasello, M. (2003) *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press
60 Baayen, R.H. and Moscoso del Prado, M.F. (2005) Semantic density and past-tense formation in three Germanic languages. *Language* 81, 666–698
61 Pierrehumbert, J. (2001) Stochastic phonology. *Glot Intl.* 5, 1–13
62 Sankoff, D. (1988) Variable rules. In *Sociolinguistics: An International Handbook of the Science of Language and Society Vol. 2* (Ammon, U. *et al*., eds), pp. 984–997, Walter de Gruyter
63 Rabiner, L.R. (1989) A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* 77, 257–286
64 Goldsmith, J. (2002) Probabilistic models of grammar: Phonology as information minimization. *Phonol. Stud.* 5, 21–46
65 Goldsmith, J. (2001) Unsupervised learning of the morphology of a natural language. *Comput. Linguist.* 27, 153–198
66 Hayes, B. and Albright, A. (2003) Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition* 90, 119–161
67 Clark, A. (2003) Combining distributional and morphological information for part of speech induction. In *Proc. 7th Conf. Eur. Chapter Assoc. Comput. Linguist.*, pp. 59–66
68 Griffiths, T. and Steyvers, M. (2004) Finding scientific topics. *Proc. Natl. Acad. Sci. U. S. A.* 101(Suppl. 1), 5228–5235
69 Resnik, P. (1996) Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61, 127–159
70 McClelland, J.L. and Elman, J.L. (1986) The TRACE model of speech perception. *Cogn. Psychol.* 18, 1–86
71 Liberman, A.M. and Mattingly, I.G. (1985) The motor theory of speech perception revised. *Cognition* 21, 1–36
72 Boersma, P. and Hayes, B. (2001) Empirical tests of the gradual learning algorithm. *Linguist. Inq.* 32, 45–86
73 Gibson, E. (2006) The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *J. Mem. Lang.* (in press)
74 Seidenberg, M.S. and MacDonald, M.E. (1999) A probabilistic constraints approach to language acquisition and processing. *Cogn. Sci.* 23, 569–588
75 Manning, C.D. and Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press
76 Newport, E.L. and Aslin, R.N. (2004) Learning at a distance: I. Statistical learning of non-adjacent dependencies. *Cogn. Psychol.* 48, 127–162
77 Yuille, A. and Kersten, D. (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* DOI:10.1016/j.tics.2006.05.002
78 Körding, K.P. and Wolpert, D.M. (2006) Bayesian decision theory in sensorimotor control. *Trends Cogn. Sci.* DOI:10.1016/j.tics.2006.05.003
79 Norris, D. *et al*. (2000) Merging information in speech recognition: Feedback is never necessary. *Behav. Brain Sci.* 23, 299–370
80 Pickering, M. and Garrod, S. (2004) Toward a mechanistic psychology of dialogue. *Behav. Brain Sci.* 27, 169–190
81 Levelt, W.J.M. (2001) Relations between speech production and speech perception. In *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler* (Dupoux, E., ed.), pp. 241–256, MIT Press