J. Linguistics **44** (2008), 205–228. © 2008 Cambridge University Press
doi:10.1017/S0022226707004999 Printed in the United Kingdom

REVIEW ARTICLE

The great number crunch[1]

CHARLES YANG

University of Pennsylvania

(Received 16 March 2007; revised 4 July 2007)

Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*. Cambridge, MA: MIT Press, 2003. Pp. xii + 451.

1. Probability and linguistics

A hard look in the mirror, as they say, is good for fitness and vitality. The time seems ripe, then, fifty years after the birth of modern linguistics, to re-examine its foundations. Or rather, the rubble, as the editors of *Probabilistic linguistics* suggest: corpus statistics, Markov chains, information theory, and the very notion of probability that were supposedly buried by the Chomskyan landslide.

> One of the foundations of modern linguistics is the maxim of categoricity: language is categorical. Numbers play no role, or, where they do, they are artifacts of nonlinguistic performance factors. (1)

This sets the volume on the wrong footing, and on too narrow a ground. After all, linguistics in the past half-century has had the good fortune of witnessing not one revolution, but two. The very essence of William Labov's groundbreaking work is that the individual's knowledge of language is inherently variable, and this is a field where probability and statistics reign supreme. But linguistic probability isn't inherently incompatible with linguistic categoricity. Variationist analysis, as Sankoff (1988: 986) explains, concerns the DISTRIBUTION of DISCRETE linguistic choices:

> Whenever a choice among two (or more) discrete alternatives can be perceived as having been made in the course of linguistic performance, and where this choice may have been influenced by factors such as features in the phonological environment, the syntactic context, discursive function of the utterance, topic, style, interactional situation or personal

[1] I thank Bob Berwick, Abby Cohn, Julie Legate, my colleagues at the University of Pennsylvania, and two anonymous *JL* referees for helpful comments and suggestions. I alone am responsible for the views expressed here.

or sociodemographic characteristics of the speaker or other participants, then it is appropriate to invoke the statistical notions and methods known to students of linguistic variation as *variable rules*.

Much more on variationist analysis later. But even if the focus is on generative grammar, the charge that probability had been systematically excluded appears to be a misreading of the historical and intellectual context. Dusting off *The logical structure of linguistic theory* (*LSLT*; Chomsky 1955/1975), widely regarded as the founding document of generative grammar, one finds little sign of grave injustice to probability (Legate & Yang 2005).[2] For instance, the *LSLT* program explicitly advocates a probabilistic approach to words and categories 'through the analysis of clustering … the *distribution* of a word as the set of contexts of the corpus in which it occurs, and the distributional *distance* between two words' (*LSLT*: section 34.5). The conception of syntactic analysis has a direct information-theoretic interpretation: 'defined the best analysis as the one that minimizes information per word in the generated language of grammatical discourses' (*LSLT*: section 35.4). Grammars are evaluated such that 'any simplification along these lines is immediately reflected in the length of the grammar' (*LSLT*: section 26): this was later termed the Minimum Description Length (MDL) principle (Rissanen 1989), now widely used in statistical models of language. On the nature of grammaticality, it 'might turn out to be the case that statistical considerations are relevant to establishing, e.g., the absolute, non-statistical distinction between G and $\overline{G}$ [grammatical vs. ungrammatical] … Note that there is no question being raised here as to the legitimacy of a probabilistic study of language' (*LSLT*: section 36.3). Perhaps best known to the current audience is the suggestion that word boundaries might be defined via transitional probabilities over successive syllables, an idea implemented experimentally in a widely influential study by Saffran, Newport & Aslin (1996), which helped popularize the probabilistic approach to language and cognition.[3] Probabilistic considerations would seem to sit squarely at the center of linguistic theorizing, rather than on the margins as 'artifacts of nonlinguistic performance factors', as the editors contend. Indeed, given the intellectual environment in which modern linguistics planted its roots, it is inconceivable that a maxim of categoricity could have been established.

So much for the intellectual history (for now): as we shall see, the linguist's reading list no longer seems to feature the classics. Let us turn to the present. What is the proper interpretation of probabilistic effects in language, which

[2] The section numbers herein refer to the 1975 edition.

[3] Though it is still useful to recall the caution that 'the problem is whether this [word segmentation via transitional probability] can be done on the basis of a corpus of a reasonable size' (*LSLT*: section 45 footnote). It cannot (Yang 2004).

are amply represented in the present volume? How much of 'categorical linguistics', to paraphrase the editors, can be maintained and how much of it has run its course? These questions are taken up in section 3 below, focusing on the role of probability in linguistic models and in the broader context of cognitive studies. In section 4, we turn to the potential benefits as well as difficulties of incorporating a probabilistic component into language, with specific attention to the computational issues in models of language processing and learning. First, an overview of the evidence.

2. A THUMBNAIL SUMMARY

Probabilistic linguistics is a collection of papers originally presented at an LSA Workshop on the same topic. It starts with a short introduction by the editors (Rens Bod, Jennifer Hay & Stefanie Jannedy), which lays out the goals of the volume together with an overview of how the chapters individually and collectively contribute to the theme of probabilistic linguistics.

Perceiving, accurately, that probability theory does not feature in every linguist's curriculum, Rens Bod ('Introduction to elementary probability theory and formal stochastic language theory') presents the basics of frequencist and Bayesian probability and probabilistic grammar formalism–standard fare from introductory computational linguistics courses. Then, rather oddly, we are treated to a tutorial on the author's own framework 'Data-oriented Parsing' (DOP) and how DOP – actually DOP1, one of the DOP installments – compares with better-known formalisms of (probabilistic) grammar. What's troubling, though, is that neither this nor later chapters that refer approvingly to DOP make any mention of the formal and empirical problems with that work, which are well known in the computational linguistics community (Goodman 1996, Collins 1999, Johnson 2001).

Dan Jurafsky's chapter ('Probabilistic modeling in psycholinguistics: Linguistic comprehension and production') is an excellent summary of probabilistic effects and models in psycholinguistics. The topics span from lexical frequency to collocation effects, from syntactic subcategorization to sentence processing. Some of these topics are covered by other authors in the volume, but this chapter alone makes a convincing case for probability in language processing and use. Toward the end of the chapter, Jurafsky also deals with several potential objections to probabilistic models of psycholinguistics.

Norma Mendoza-Denton, Jennifer Hay & Stefanie Jannedy ('Probabilistic sociolinguistics: Beyond variable rules') provide a review of quantitative methods in sociolinguistics and introduce a new statistical tool that makes inferences about dependent variables in linguistic behavior. As a case study, they analyze the distribution of monophthongization in the speech of Oprah Winfrey, the noted media personality. It is found that Winfrey's use of the variable is subtly attuned to her conversational partner's

ethnic identity, frequencies of word usage, and other factors. The demonstration is convincing, but in light of Labov's pioneering effort, I do not see in what sense this chapter is 'beyond variable rules', as its title advertises. Perhaps 'beyond VARBRUL' would be more appropriate, with reference to the well-known statistical package for sociolinguistic analysis, now that a new tool kit is available. We will return to variationist analysis in section 3.2.

Kie Zuraw ('Probability in language change') starts out with an introduction to some statistical methods in the inference of historical relations among languages, before turning to recent work that uses probabilistic techniques to investigate language change. That language change often takes place gradually is a traditional observation, and the need for probabilistic methods is underscored by the analysis of historical corpora within the generative framework (Kroch 1989). Unfortunately, Zuraw's chapter is marred by the omission of several prominent lines of work. [Full disclosure: my own research on language change did receive several pages.] There is an increasing body of work that applies phylogenetic methods to the natural history of languages, much of which was available at the time of writing (McMahon & McMahon 1995, Grey & Jordan 2000, Ringe et al. 2002). Perhaps the most glaring omission, and one that would presumably make the strongest case for a probabilistic view of language change, is the research on lexical diffusion by Wang and his colleagues (Wang 1969, Chen & Wang 1975, etc.). The responses to lexical diffusion, which maintain the classical view of phonological change (Labov 1981, 1994; Kiparsky 1995), would also have made a welcome addition.

Janet Pierrehumbert's contribution ('Probabilistic phonology: Discrimination and robustness') centers on the probabilistic aspects of sound patterns. Taking the famous plot of American English vowel variation (Peterson & Barney 1952) as her point of departure, she presents several strands of evidence that language users are sensitive to minute details of phonetic and phonological information, which can be modeled in a probabilistic framework. Pierrehumbert, more so than the other contributors, makes a strong appeal to distinct linguistic levels: acoustic, phonetic, phonological, morphological, among others. Statistical generalizations at lower levels are used to form primitives at higher levels. She also provides a summary of several probabilistic models of phonology, including cluster analysis of the phonetic space (Kornai 1998), the Gradual Learning Algorithm (Boersma & Hayes 2001), and the exemplar approach of which Pierrehumbert is a leading proponent. We will return to a brief assessment of these in section 4.1.

Much of Harald Baayen's chapter ('Probabilistic approaches to morphology') focuses on the issue of morphological productivity and the tension between storage vs. computation in the mental lexicon. At times it appears that Baayen is advocating the probabilistic approach not against the categorical approach in general. Rather, the target is the specifics of the dual-route model of morphology (Clahsen 1999, Pinker 1999), which maintains a

sharp separation between memorized and rule-based forms.[4] Baayen presents evidence that even regularly inflected words, which are assumed to be formed compositionally in traditional theories (as well in the dual-route model), can be stored holistically, especially in the high frequency region: the dichotomy is thus less clear.[5] The issue of morphological productivity, which also receives substantial treatment in this chapter, is discussed in a similar vein. Baayen claims that the productivity of many morphological classes, particularly those in the derivational domain, falls on a gradient scale. These issues will be picked up in section 3.1 and section 4.1.

The chapter by Christopher Manning ('Probabilistic syntax') is arguably the broadest in scope. It begins with a discussion of the relationship between linguistic corpora and linguistic theories. A review of probabilistic models of grammar follows, including probabilistic context-free grammar and stochastic Optimality Theory, among others. Manning also suggests that a probabilistic treatment of the grammar eases the task of learning, a claim highlighted in the introductory chapter by the editors as well. But these statements need to be qualified; we will do so in section 4.

The last chapter, by Ariel Cohen ('Probabilistic approaches to semantics'), surveys recent research on probabilistic methods in semantics. Some of the topics he reviews, such as the treatment of frequency adverbs ('always', 'often', 'sometimes'), readily lend themselves to a probabilistic analysis, while topics such as conditionals, vagueness, and discourse inference are less developed. It seems to me that a probabilistic foundation for semantics has not yet emerged to the point of consensus, though Cohen's understated but effective chapter highlights some potential benefits in that direction.

3. Linguistic probability and linguistic theory

In this section, we consider how probabilistic effects in language may be reconciled with – or in fact, strengthen – the categorical approach. Again, we do so with an eye to the history of probability in linguistics. At the minimum, probabilistic effects do not automatically constitute a rebuttal of categorical linguistics, as some of the contributors to the volume contend; careful dissection of the probabilistic facts is required. And once we do so, it may turn out that numbers really ARE 'performance factors': probabilistic effects of the cognitive system at large which interact – maybe even in crucial

[4] That's not the only game in town; see Yang (2002), Embick & Marantz (2005), Stockall & Marantz (2006), which augment and directly place the generative approach to morpho-phonology in a psycholinguistic setting.

[5] As far as I can see, even the dual-route model could easily accommodate Baayen's findings by augmenting the model with a clause: 'store high-frequency items as well', though the threshold for storage is a separate, and empirical, question. We should also point out that there has been no report of storage effects in *auditory* presentations of regularly inflected items (Pinker 1999). This leaves open the possibility that Baayen's results may be an artifact of more familiar orthography.

ways – with the fundamentally categorical system of linguistic knowledge. But first, a discussion of the facts is in order.

3.1 *Probabilistic facts*

Much of this volume focuses on probabilistic factors in language use and their implications. Some of these claims, however, seem empirically questionable; it's best to get these out of the way.

One source for concern has to do with the methodologies whereby probabilistic effects are established. Baayen's contribution makes reference to various gradient phenomena in morphological complexity and productivity uncovered by Hay (2000; but see Poplack 2001 for some counterexamples). Just as one needs to be critical about the grammaticality data in categorical linguistics, we must also scrutinize the protocols with which gradient effects are harnessed. It is well known in experimental psychology that categorical tasks are likely to elicit categorical results, and gradient tasks – such as rating, which has been gaining currency in the probabilistic approach to language – are likely to result in, alas, gradient results, as the subject is more inclined to spread responses over multiple choices (Parducci & Perrett 1971). For instance, in a classic study, Armstrong et al. (1983) find that gradient judgment can be obtained even for uncontroversially categorical concepts such as 'even numbers' and 'females'. I will not discuss the specifics of Baayen's and Hay's claims but only direct the reader to Schütze's critique (2005). Schütze notes in Hay's studies, the instructions for the judgment tasks were sometimes inconsistent and biased against specific reactions. Moreover, a significant number of subjects seem to have misunderstood the task, and the results are often barely distinguishable from chance-level performance.

Here I will focus on the empirical basis for frequency effects in language change. Zuraw, with reference to Bybee's usage-based model (2001), asserts that 'frequent lexical items are the first to adopt automatic, phonetic rules' (157). Statements of this nature have featured prominently in the probabilistic linguistic literature and already form an integral part of probabilistic models such as the exemplar theory (Pierrehumbert 2002). Unfortunately, these claims have not been substantiated in the quantitative study of sound patterns.

An oft-cited example in the probabilistic linguistic literature is Bybee's assertion that frequent lexical items more often undergo t/d deletion in word-final consonant clusters (see also Jurafsky's contribution). But as Abramowicz (2006) points out, t/d deletion is in fact an instance of stable variation (Guy 1991), not a change in progress. This is further evidenced by Roberts's study of language acquisition (1996), which shows that young children match the probabilistic phonetic and grammatical patterns of adult speech nearly perfectly.

Recent work on sound change in progress, such as the Telsur survey of American English (Labov et al. 2006), has generated the volume of data necessary for quantitative findings in language change to be established and evaluated. In a study of the fronting of the diphthong nuclei /uw/, /ow/, and /aw/ in American English, Labov (2006) finds that high-frequency words are neither more nor less advanced in the sound change than low-frequency ones, and that virtually all variation can be accounted for by purely phonetic factors, a result recently corroborated by the work of Dinkin (2007). Dinkin evaluates the front/backness of short vowels in Northern United States English, which are involved in the ongoing Northern Cities Vowel Shift. A large amount of data – about 13,000 measurements of vowels – is examined with multiple regression, and there is no evidence that frequency affects the progression of change. Dinkin does find, however, that high frequency more readily leads to LENITION, as observed by Phillips (1984). The connection between frequency and change may well be genuine, but we need a far more careful look at the data before slippery facts harden into the foundations of probabilistic linguistics.

3.2 *Variation and grammar*

As remarked by several contributors, the case for a probabilistic linguistics is built on the 'gradient middle', and this challenges the categorical linguist's focus on the endpoints. But what is this gradient middle any way? If the grey area is indeed grey, then the categorical linguist has a problem, for s/he would have failed to identify one of the ways in which a linguistic phenomenon could vary. However, if 'grey' is just $(p \times [\text{black}] + (1-p) \times [\text{white}])$, i.e., if the language learner/user probabilistically accesses alternative forms that fall in the realm of possible variation, then there is nothing wrong if one chooses to only look at black and white.

Indeed, I know of no grammatical model that PROHIBITS the association with probabilities. Many contributors here do just that: adding probabilities to grammars developed in the categorical tradition. For example, Jurafsky considers probabilistic context-free grammar to capture frequency effects in sentence processing, Pierrehumbert and Manning use stochastic Optimality Theory to describe, respectively, phonetic variation and syntactic alternations, and Manning augments subcategorization frames with probabilities to encode lexical selectional tendencies, as follows (Manning, 303):

$$P\,(NP\,[_{\text{SUBJ}}] \mid V = \textit{retire}) = 1.0$$
$$P\,(NP\,[_{\text{OBJ}}] \mid V = \textit{retire}) = .52$$
$$P\,(PP\,[_{\text{from}}] \mid V = \textit{retire}) = .05$$
$$P\,(PP\,[_{\text{as}}] \mid V = \textit{retire}) = .05$$

The categorical linguist's response is straightforward: these are NOT genuinely grey areas, just add your numbers to my rules. To be sure, the

dependence of probabilistic linguistics on entities and generalizations from categorical linguistics is recognized by some of the contributors. Pierrehumbert, for instance, points out that '[t]his conception of phonology as a formal grammar (with abstract variables) is often assumed to stand in opposition to the idea that phonology involves statistical knowledge. However, this opposition is spurious, because probability theory requires use to assign probability distributions to variables' (p. 178). Yet this volume is far more content to trumpet the probabilistic crisis that generative grammar is mired in (e.g., Manning) and, in the words of Joan Bresnan on the book jacket, 'calls for a revolution in the models and theories that linguists commonly use'.

Except that the revolution has already been underway for four decades. In a celebrated paper (1969), Labov significantly extends the range of linguistic data into new empirical and social dimensions and introduced quantitative tools into linguistic analysis. And statistical evidence provides support for both the broad and the specific claims of linguistic theories (e.g., Chomsky 1965, Chomsky & Halle 1968). Labov's remarks are worth quoting at length (1969: 761):

> I do not regard these methods or this formal treatment as radical revisions of generative grammar and phonology. On the contrary, I believe that our findings give independent confirmation of the value of generative techniques in several ways. First, I do not know of any other approach which would allow us to work out this complex series of ordered rules, in which both grammatical and phonological constraints appear. Secondly, the stress assignment rules of Chomsky & Halle seem to yield precisely the right conditions for vowel reduction and the contraction rule. Since the contraction rule has never been presented before in detail, we must consider this independent confirmation on the basis of discrete data, clearer evidence than we can obtain from the continuous dimensions of stress or vowel reduction. We also find independent confirmation of the position and role of the tense marker, even where it takes a zero form. Third, we find abundant confirmation of Chomsky's general position that dialects of English are likely to differ from each other far more in their surface representation than in their underlying structure.

The variationist perspective invites at least two lines of discussion, both of which are sorely missing in this collection. First, what is the proper place for probabilistic linguistics in the history of linguistic thought? The contributors are treading over VERY familiar terrain; it is crucial to bear in mind that the object of Labov's study is the language faculty in its entirety – both statistical and categorical – not just its social and cultural components. In light of this, the volume has nothing to say, aside from a few technical remarks, about the conceptual and methodological issues surrounding variationist analysis; perhaps some useful lessons can be learned from the earlier controversy

(Cedergren & Sankoff 1974, Kay & McDaniel 1979, Sankoff & Labov 1979). Second, how do we (re)assess the continuity between probabilistic and categorical linguistics observed in Labov's classic paper? Numbers do not produce analysis all by themselves; statistical tools verify, but do not produce, empirical hypotheses. After all, the subcategorization choices to which Manning assigns probabilities come straight out of the categorical linguistics literature. There is no need to disparage the categorical linguist's interest in the endpoints; they provide the very units of distribution that the probabilistic linguist works with.

3.3 *The locus of linguistic probability*

The controversy over probabilistic effects in language stems, at least in part, from the border war between the grammar and its use, which in turn touches on the familiar competence vs. performance issue.[6] My concerns here are narrower: probabilistic effects are not a homogeneous lot, and their accommodation may be afforded by existing models of language and cognition.

First, the issue of linguistic levels. For example, the classic conception of sound patterns takes the phonological system to be categorical and the phonetic system to be continuous. The gradient effects reported in Pierrehumbert's chapter indicate that the boundary between the levels may not be clear cut, but that does not mean that such effects aren't the joint product of two distinct domains (see Cohn 2006 for an extensive discussion of facts and theories in gradient phonology).

Second, the language faculty is embedded in and interacts with the cognitive/perceptual system at large, which may carry its own numbers; a probabilistic effect in the composite output does not automatically pick out the factor that causes the effect. The problem is not restricted to linguistic research. I recognize my son's face more rapidly than my cousin's – likely a consequence of frequency – but that alone tells us nothing about how faces are represented and perceived, or how I can recognize a face, any face, in the first place. Indeed, frequency effects alone in no way undermine the claim that there may be a domain-specific module associated with face recognition, as recent work suggests (Kanwisher 2000, though see Tarr & Cheng 2003).

With these complications in mind, consider the interpretation of frequency effects in lexical processing.[7] Surely, a statistical/associative process could

[6] For a recent exchange, see Newmeyer (2003) and the responses it has engendered (Clark 2005, Guy 2005).

[7] These effects appear stronger in comprehension than production. In the latter case, a more diverse range of interpretations have been offered (Guion 1995, Lavoie 2002), where a central concern is whether the findings are consistent with a more abstract conception of lexical structures or with the exemplar approach; see also the discussion of t/d deletion in section 3.1. Moreover, apparent frequency effects may sometimes be factored into

replicate the outcome of repeated exposure, but it is worth noting that one of the earliest models of lexical frequency effects was one which made no reference to probability at all (Forster 1976). According to what has come to be known as the bin search model, words are represented in an orthographically/phonetically defined bin, and ranked by frequencies: the higher up a word resides, the faster it will be accessed. And there exist several classes of online algorithms that manipulate such a list in a computationally trivial way while retaining near-optimal efficiency (Forster 1992, Yang 2005). This simple and plausible model has the additional property of predicting that non-existing words will generally take longer to recognize/reject in lexical decision tasks, for all the known words will have to be scanned first. Without going into the finer-grained predictions of this model (Murray & Forster 2004), it suffices to say that when equipped with adaptive online algorithms for processing, discrete linguistic structures are capable of producing stochastic properties – no floating points attached.

Yet another way to account for probabilistic effects in language is to develop precise models of perception, memory, learning, etc. which may allow us to tease apart the interacting components in language use. Consider Hale's probabilistic model of sentence processing (2001), which is reviewed in Jurafsky's chapter. This model is designed to handle the asymmetry in processing time between subject and object relative clauses – such as *The man who saw you saw me* vs. *The man who you saw saw me* (Gibson 1998). It does so by essentially keeping track of the frequencies of subject and object relative clauses, with the former considerably greater than the latter.[8] Putting aside the formal problems with this and related models, to which we return in section 4.1, the asymmetry is plausibly accounted for by models that integrate the independently motivated working memory system into language processing (Pritchett 1992, Just & Carpenter 1992, Gibson 1998). Of particular interest is the work of Vasishth & Lewis (2006), who appeal to temporal aspects of a precise model of working memory (Anderson & Schooler 1991) to explain the apparent probabilistic aspects of language processing.

Taken together, the gradient middle may well be the converging ground of multiple cognitive systems interacting with each other. Its very existence is no disconfirmation of the categorical conception of linguistic structures. Quite the opposite may be true, as we illustrate presently.

---

completely predictable structural conditions (Jurafsky et al. 2002). I thank Abby Cohn for discussion of these matters.

[8] Associating probability with processing cost may be too strong: Is 'Is that …' easier for the sentence processor than 'Is the …', just because Aux [$_{NP}$ Pronoun] is more frequent than Aux [$_{NP}$ Det …]? Would extremely rare syntactic constructions (e.g., parasitic gaps) entail extremely slow processing time?

3.4 *Probabilistic evidence for categorical linguistics*

In the natural sciences, statistical variation often provides the most compelling evidence for a discrete system. Perhaps the best example comes from biology, where the discrete genetic basis is apparently at odds with the continuous distribution found among the individual organisms. Indeed, a great crisis arose for the theory of evolution at a time when genetic materials were believed to mirror the continuous variables at the phenotypic level: the so-called 'blending inheritance' would bleach out variation and ground evolution to a halt. Only the rediscovery of Mendel's work saved the day. The discrete Mendelian principles of inheritance, as is well known, were deduced entirely from the statistical distribution of phenotypes (smooth vs. wrinkled seeds, purple or white flowers). Similarly, the probabilistic variation in the phenotype of language – use, learning, and change – may be the reflex of an underlying system of discrete linguistic units; this is again the line of reasoning in Labov's (1969) classic study.

Take a prima facie case of variation: the rise of periphrastic *do* in the history of English (Ellegård 1953, Kroch 1989; see Zuraw's chapter). From the probabilistic perspective, this particular case of change is unremarkable: every student of historical linguistics knows that language change typically takes place gradually, characterized by a mixture of linguistic forms whose distribution is in fluctuation. From the categorical perspective, however, the emergence of *do* is an extremely important discovery. Kroch provides statistical evidence that the uses of *do* in several seemingly unrelated constructions follow the same trajectory of change. Unless the correlation turns out to be a complete accident, the grammatical change must be attributed to the change in a SINGLE syntactic parameter: that of V-to-T movement, independently recognized in the Principles and Parameters framework – categorical linguistics par excellence. The key point is not just the fact of variation, but the unifying force of the parameter behind it.

Take another well-known case of probabilistic phenomena in language: language acquisition. There is a good deal of evidence against a categorical perspective on learning, according to which the child learner is to hop from one grammatical hypothesis to another as the input is processed (e.g., 'triggering'). On the one hand, quantitative analysis of child language frequently shows variation that cannot be attributed to any single potential grammar. On the other, it is well known that child language does not change overnight; variation is eliminated gradually as the child approaches adult-like grasp of the grammar. But one needn't, and shouldn't, abandon the categorical theory of GRAMMAR; we only need to call upon a probabilistic model of LEARNING (Bush & Mosteller 1951; Yang 2002, 2006; see Roeper 2000, Crain & Pietroski 2002, Rizzi 2005 for similar views). Consequently, the variation in child language can be interpreted as a statistical ensemble of possible grammars whose distribution changes over time. The pool of

grammars with specific properties – identified in categorical linguistics – appear to be just the right kind of filter to mediate the mismatch between the distributional regularities in the input (i.e., adult language) and the output (i.e., child language): language acquisition is full of cases where the development of children's grammar does not correlate with the statistical distributions seen in adult language (Hyams 1986, Wexler 1994). Moreover, the probabilistic model of learning makes it possible to relate the distributional patterns in the input data to the longitudinal trends of grammar acquisition (Yang 2004, Legate & Yang 2007, Yang to appear).

The probabilistic aspects of language learning, use, and change do raise a challenge – to categorical models of learning, use, and change, but not to the categorical view of language itself. As Mehler et al. remark in a recent paper (2006), the 'soul' of language does not use statistics. This volume has given us no reason to disagree.

4. Data, model, and inference

My final set of comments will have to be registered with a sense of conflict. As a computer scientist reared in the statistical approach to machine learning, and one who has made probabilistic models a main theme of my work on language acquisition and change, I agree with the editors on the necessity of integrating probability into a comprehensive theory of language and cognition.

But probability is no silver bullet. Regrettably, the present volume as a whole does not recognize the full scale of challenges that linguistic models face, nor does it present a realistic assessment of the powers – and limitations – of probabilistic methods. Let me make it clear before we get to the specifics: in voicing my concerns, I in no way imply that the non-probabilistic way provides a better treatment. The matters are genuinely difficult, and it is misleading to suggest that the probabilistic approach holds a better hand.

4.1 *Probability, reality, and computation*

As Jurafsky notes, it is often premature to ask questions of psychological reality when dealing with abstract models of language and cognition: 'Surely you don't believe that people have little symbolic Bayesian equations in their heads?' (89). That, however, does not completely absolve the linguist from the responsibility of empirical justification, especially when the model is intended to capture aspects of linguistic performance. After all, a probabilistic model of language is a model of reality, and it needs to be measured against and constrained by independent findings regarding what the human language user can or cannot do, on the one hand, and regarding complexity, scalability, and other computational issues on the other.

https://doi.org/10.1017/S0022226707004999 Published online by Cambridge University Press

At times, the book's presentation of probabilistic models shows an unfortunate disconnection from the state of the art in computational linguistics. (The case of DOP models was already mentioned in section 2.) Both Manning and Pierrehumbert give an extensive discussion of how the Gradual Learning Algorithm (GLA; Boersma & Hayes 2001) can be applied to model variation and ambiguity, but neither author mentions Keller & Asudeh's criticism (2002) of that work. Keller & Asudeh point out that the GLA has not been subject to the rigorous evaluations that are standard in natural language processing. On the one hand, the GLA model is trained and tested on the same data where the standard practice is to use separate sets. On the other, virtually nothing is known about its formal properties, such as learnability, convergence time, etc., and there are datasets that are unlearnable under the GLA model. These problems may not be Manning or Pierrehumbert's direct concerns, but without addressing them, their appeal to the GLA is not convincing.

Another troubling example is the discussion of distributional learning in language acquisition. Taking up the challenge posed by Peterson & Barney's (1952) findings, while drawing contextual support from automatic statistical learning by infants (Maye et al. 2002), Pierrehumbert cites a result of Kornai's (1998) 'unsupervised cluster analysis on the vowel formants data … The clusters identified through this analysis are extremely close to the mean values for the 10 vowels of American English' (p. 187). This is an astonishing claim: pattern recognition of exactly this kind has tormented statisticians and computer scientists to no end. But it is also an overstatement: according to Kornai's original paper, the model actually employs the so-called K-means algorithm, which requires the scientist to *specify* exactly how many partitions of the data are to be expected. That, of course, has the effect of whispering into the baby's ear: 'Your language has ten vowels; now go get them'. Kornai's result is not without interest but it certainly cannot be taken as a demonstration that phonetic categories would emerge through distributional learning.

My biggest concern, however, lies not with specific algorithms but with probabilistic models in general. In computational linguistics, the perennial difficulty is the *sparse data* problem: as the statistical model gets richer, the number of parameters to be valued tends to shoot up exponentially. The recent progress in statistical natural language processing can largely be attributed to assumptions about the language model that simplifiy and reduce the interactions among the parameters. The most basic kind is that of *independence*, the Markovian assumption that the components of a linguistic expression (phonemes in words, words/phrases in sentences, sentences in discourse) are independent of each other. This allows for the multiplication of the components' individual probabilities, which are relatively easy to estimate due to their smaller size, in order to obtain the probability of the composite. Linguists will no doubt find this move dubious. In general,

$P(AB) = P(A) \times P(B)$ only if A and B are independent; but in a linguistic expression, hardly any two items are ever independent. Several contributors to this volume make this assumption of independence, though only Manning (303) explicitly alerts the reader that it may be problematic. But even this does not get around the sparse data problem: a trigram model of grammar may greatly overwhelm the available data (Jelinek 1998).[9]

Consider again the model of Hale (2001) in light of the sparse data problem. Hale notes that the difficulty (e.g., reading time) in sentence processing can be modeled in an information-theoretic way: specifically, how surprised the reader is when encountering the next word (w_i) given preceding words ($w_1 w_2 \ldots w_{i-1}$ or w_1^{i-1}). Technically, this notion (called *surprisal* of word w_i) can be formalized as:

$$h(w_i) = -\log P(w_i|w_1^{i-1}) = -\log \frac{P(w_1^i)}{P(w_1^{i-1})} = -\log \frac{\alpha_i}{\alpha_{i-1}}$$

This formulation provides a simple, incremental, and theory-independent measure of cognitive load in processing. However, scaling it up to a realistic setting may run into practical problems. The probability of a string under a language model is actually the sum of the probabilities of all possible structural analyses, which requires a commitment to the parallel parsing strategy; in principle, the number of possible parses that need to be tallied up may grow exponentially as well. Hale's implementation thus only uses a small handcrafted context-free grammar that can only deal with several examples from the sentence processing literature. It seems that the scalability issue needs to be addressed before probabilistic models can make a true impact on empirical research.

Hale in fact makes the interesting suggestion that one may take α to represent the probability of syntactic structures rather than of strings (Gibson & Perlmutter 1998). The computational benefits of this move may be significant: more abstract representations give rise to fewer parameters, which may in turn alleviate the sparse data problem (though no one knows by how much). Ironically, then, more help from categorical linguistics may be the right way to cash out a probabilistic model of linguistics. If recent trends in computational linguistics can serve as a guide, that turns out to be exactly the case (Gabbard et al. 2006).

Once again, I am not advocating some categorical mechanism of learning and processing (as opposed to the GLA and the surprisal model). Nor am I

[9] A trigram model is a probabilistic model that takes the product of the probabilities of successive triples in a string of linguistic units to approximate the probability of that string, or $P(w_1 w_2 \ldots w_n) = \Pi_{i=1}^{n}\ p(w_i|w_{i-2}w_{i-1})$. Obviously, this involves an independence assumption that does not reflect linguistic reality; but even so, it runs into the sparse data problem as the possible triples (i.e., three-word combinations) are numerous and cannot all be sampled in any reasonable corpus.

https://doi.org/10.1017/S0022226707004999 Published online by Cambridge University Press

suggesting that the acquisition of phonetic categories requires an innate feature system (though see Dresher 2003).[10] But I do not share the enthusiasm for recent advances in computational linguistics and machine learning, as Manning expresses here and as is echoed by other commentators (e.g., Lappin & Shieber 2007); see also section 4.3. Whatever progress statistical models have made in applied natural language processing, they do not directly translate into success or even utility in the understanding of language and cognition. Anyone can write a computer program to do something, but it is an altogether different matter whether the model reflects the psychological reality of the human language user. For instance, current statistical models of language induction have reported useful progress in parsing, which is typically measured by the accuracy of finding appropriate bracketings of words into phrases or constituents. But even putting aside assumptions about the training data and the learning algorithms, which presently lack psychological backing, I fear that parents would run screaming to the nearest neurologist if their child got one out of ten phrase structures wrong – which is about the state of the art in computational linguistics these days. Overall, a more comprehensive treatment of probabilistic models would have been more helpful to the reader who does not come from a computational linguistics background, and more convincing to one who does.

4.2 *The case of missing data*

Related to the sparse data problem is the problem of what to do with the data at hand. Corpus usage is featured prominently in this volume. The corpus is no doubt a valuable depository of linguistic insights; how best to mine it, however, is a complex issue and ought not to be trivialized to 'If Google can find it, it is linguistics'. Perhaps two linguists who are known for their attention to data put it best: 'Not every regularity in the use of language is a matter of grammar' (Zwicky & Pullum 1987: 330).

No such caution for Baayen, however. After finding on the web several uses of *-th* as a nominalizing suffix, he proceeds to claim 'the residual degree of productivity of *-th* and the graded, scalar nature of productivity' (236), while criticizing Bauer (2001) for assuming that *-th* is an unproductive affix. But it is worth noting that all instances of *-th* in Baayen's search[11] were, and perhaps still are, in use, albeit with very low frequencies. Finding them on the

[10] A promising approach to acquiring phonetic categories may well be probabilistic. The recent work of Coen (2006) develops a novel framework of cross-modal clustering and is able to extract the vowel categories of English without supervision, though with additional assumptions that resemble some version of the Motor Theory of Speech Perception.

[11] These are *gloomth*, *greenth*, and *coolth*; Bauer (2001; cited by Baayen, page 234) already points out that the first two are frozen forms.

ever-expanding web is thus hardly a surprise. However, this has nothing to do with productivity, which is generally taken as a measure of how a particular form generalizes to *novel* items. Following Baayen's logic, if we find a few instances of *work-wrought* and *cleave-clove*, then the strong verbs must be making a 'graded' and 'scalar' return.

This is not to deny the utility of the corpus to linguistic analysis. Keeping to the topic of productivity, one of the most informative sources of data comes from the most error-prone stage of language use, i.e., child language acquisition (CHILDES; MacWhinney 1995). A most robust finding in morphological learning is that the misuse of unproductive morphology is vanishingly rare (Xu & Pinker 1995). By contrast, the overapplication of productive morphemes is abundant (Pinker 1999). This asymmetry has been repeatedly observed across languages, in fact by the use of corpus statistics (Guasti 2002). It would follow, then, that a categorical distinction of productive vs. unproductive morphology shouldn't be dismissed so quickly.

The flip side of the data–theory problem is more troubling; let's paraphrase it as 'if Google can't find it, it's not linguistics'. Manning, in his discussion of the theoretical implications of corpus data, voices his concern that generative grammar 'is increasingly failing because its hypotheses are disconnected from verifiable linguistic data' (296). There are several problems with this view. First, why does only the corpus count as 'verifiable linguistic data'? Surely grammaticality judgments – even if we DO complain about them – can be confirmed or rejected as well. By contrast, corpus statistics are known to be highly sensitive to genre and style. Regularities from the Wall Street Journal hardly generalize to CHILDES, so the corpus does not necessarily constitute a bias-proof fount of linguistic knowledge. Second, and more important, a fundamental goal of linguistic theory is to establish the bounds of possible and impossible linguistic forms. The impossible forms, by definition, are not uttered for they are unutterable and may never be found anywhere. Even the possible forms, at least the type most revealing of the limits of grammar and language (think of island constraints, the wug test, and center embedding), may not be readily available from the corpus either. While we can all agree on the need for better methods for getting better data, it does not seem necessary to pay special reverence to corpora.

Once we consider the cognitive status of the corpus, there are even more reasons for concern. Consider a specific kind of corpus, the primary linguistic data that a child receives during the course of language acquisition. Such a corpus is by definition finite – nobody learns forever – yet the child is able to attain a system of knowledge that is capable of generating an infinite set of grammatical linguistic expressions (and not generating an infinite set of ungrammatical ones). This fact alone seems to support the contention that 'Absence of evidence is not evidence of absence', as former U.S. Secretary of Defense Donald Rumsfeld once infamously, but logically, quipped. However, Pierrehumbert seems to be suggesting just

that: 'Statistical underrepresentation must do the job of negative evidence' (196). Her discussion makes no mention of the empirical research on negative evidence, but merely states, without reference, that 'recent studies indeed show that children are quite sensitive to the statistics of sound patterns'. Presumably the child can compute statistics over certain linguistic units (Saffran et al. 1996), but I don't see exactly how learning something in the data tells the learner what is not in the data. Even if the Rumsfeldian logic were too strict – is that possible? – it is still imperative that the learner's hypothesis space be properly bounded; after all, evidence, positive or negative, is defined relative to hypotheses, and the problem of proper generalization over data does not go away. The learnability puzzles concerning negative evidence are challenging (Lasnik 1989), and the field of language acquisition has gone over this issue several times already; the reader is directed to Marcus (1993) for a useful review.

4.3 *Probabilistic learning and language learning*

Finally, a note on the potential benefits and difficulties of incorporating probability into a theory of language learning. Probability can and does help. In the case of parameter setting discussed above, a probabilistic model has provably superior formal properties as compared to its counterparts, with implications for language development as well (Yang 2002). In computer science, we often find problems where the discrete version is extremely difficult but where relaxation of discreteness makes for far greater tractability. A classic example is the optimization problem of LINEAR PROGRAMMING, where the goal is to maximize (or minimize) the value of a linear function under various constraints:

$$f(x_1, x_2, \ldots, x_n) = c_1x_1 + c_2x_2 + \ldots + c_nx_n$$

If the unknown variables must be integers, then optimization can only be solved by means of infeasible computational resources. By comparison, if the variables are allowed to be real numbers, then the problem can be solved efficiently – by a straightforward subroutine call in Microsoft Excel.

But there is no mathematical evidence that probability ALWAYS helps. The editors remark, citing the results of Horning (1969) and Manning's contribution, that 'unlike categorical grammars, probabilistic grammars *are* learnable from positive evidence' (6) and that 'if the language faculty is probabilistic, the learning task is considerably more achievable' (7).[12]

[12] It is useful to note that this is (at least) the second time that probability has made a hopeful return to learnability. In particular, the probabilistic learning models of Suppes (1969, 1970) generated a good deal of interest in the 1970s. A careful reading of the literature at the time (Arbib 1969, Batchelder & Wexler 1979, Pinker 1979) would still prove useful today.

This is simply a mistake. At the risk of straying too far afield, let me provide some background on mathematical learning theory.

Pertinent to our discussion are two related but distinct frameworks of learning, both of which have developed a very large technical literature. The 'categorical' framework of Gold (1967) requires exact identification of the target hypothesis, whereas the Probably Approximately Correct (PAC) framework (Valiant 1984) only requires the learner to get arbitrarily close to the target though s/he must do so within reasonable bounds of computational resources. Both frameworks are broad enough to allow modifications of the assumptions about the learner, the presentation of the data, the criterion for convergence, etc.; for example, Horning (1969) presents a probabilistic instantiation of the Gold learning paradigm (see Pitt 1989 for a general formulation). But it is not accurate to claim that probabilistic learning is 'more achievable' than exact identification in the Gold framework. It is actually difficult to compare learnability results from these frameworks, which operate under different assumptions (see Nowak, Komarova & Niyogi 2002 and Niyogi 2006 for insightful reviews). For example, the PAC learner has access to both positive and negative data; if the Gold learner is given negative data, then all recursively enumerable sets are learnable. Moreover, the Gold learner can take an arbitrarily long time – as long as it's finite – to convergence, a luxury the PAC learner cannot afford. Finally, it is not the case that PAC learning admits a larger class of languages: for instance, finite languages, which are learnable under the Gold framework with positive evidence alone, are not learnable under the PAC framework even with both positive and negative evidence.

Both the Gold and the PAC frameworks aim to derive learnability results in the 'distribution-free' sense – that is, no prior assumptions are made about the distribution from which the learning sample is drawn. This requirement produces results of the greatest generalizability (and thus interest) but it can be relaxed as well. It has been shown (Angluin 1988, among others) that if one has certain information about the distribution from which the sample is drawn, then the class of learnable languages is considerably enlarged. But this is a very strong assumption to make, as the estimation of the distribution of a function is generally harder than the approximation of the function itself – and it's the function itself the child is trying to identify during the course of language acquisition: the child is to learn how to say 'I am hungry', not how often 'I am hungry' is said.

Viewed in this light, Horning's (1969) result, which has often been cited as an argument for probabilistic models of language (e.g., Abney 1996), is a special case and a very weak result. It is well known that under the Gold framework, context-free languages are not learnable. However, once the context-free grammar rules are associated with probabilities, the distribution of sentences becomes very favorable to learning in the sense that Angluin (1988) describes. In a probabilistic context-free grammar, the probability of a

https://doi.org/10.1017/S0022226707004999 Published online by Cambridge University Press

sentence is the product of the probabilities of rules that lead to its expansion – and thus we encounter the independence assumption again, see section 4.1. It follows, then, that longer sentences are vanishingly unlikely. Horning's learner can, in effect, ignore these long sentences without affecting the overall approximation to the target. Now the grammar is, in effect, FINITE – a position that I think few linguists would find appealing. Finite languages, however, ARE learnable, as Gold had already shown.

Another problem is that Horning's learning algorithm actually works by enumeration, i.e., searching through the entire space of all probabilistic context-free grammars. The computational cost of this is prohibitive, as Horning himself notes. Once the computational complexity of learning is taken into account (i.e., the PAC framework), most language classes of linguistic interest – finite-state languages, context-free languages, etc. – are NOT learnable (Kearns & Valiant 1994). Horning's result is obtainable only under strong and unrealistic assumptions about grammar and learning; the conclusion that probability makes language learning more tractable is a misreading of the mathematical results.

Overwhelmingly, learnability results in both the Gold and the PAC frameworks are negative. These results are generally obtained irrespective of the learning algorithm; in other words, no salvation can be found in semantic information, social learning, cultural cues, or distributional regularities. Once again caution is needed to find interpretations that are relevant and specific to the study of language acquisition. One conclusion does emerge convincingly from both frameworks: learning is not possible unless the hypothesis space is tightly constrained by prior knowledge, which can be broadly identified as Universal Grammar. Of particular interest to linguistics is the fact that, if there is a finite number of hypotheses, then learnability is in principle ensured. In this sense, grammatical theories that postulate a finite range of variation, such as the Principles and Parameter framework, Optimality Theory, and others, are FORMALLY learnable.[13] But this does not mean that the learnability problem is resolved as an empirical problem – and hence the extended effort to see how parameter setting and constraint ranking actually work. As noted by Chomsky (1965: 61), the crucial question is how the grammatical hypotheses are 'scattered' such that they can be distinguished by data in a computationally tractable way. Another challenge, of course, is whether computational models of language learning pass the acid test of psychological plausibility and match the findings of language development (Yang 2002). In any case, the fundamental problem in language acquisition remains empirical and linguistic, and I don't

[13] That is not to say that only finite hypothesis spaces are learnable. An infinite hypothesis space, if bounded a priori in some appropriate way, may also be learnable, though this involves a different notion of finiteness (Vapnik 1995).

see any obvious reason to believe that the solution lies in the learning model, be it probabilistic or otherwise.

5. Conclusion

The present volume is an accessible collection that assembles in one place research from a diverse range of linguistic sources. Probability is here to stay, and may indeed have an important role to play in the study of language and cognition. For these achievements, the editors are to be applauded.

But the proof of the pudding should have been left in the eating. Though several contributors make measured and effective assessments of probability in language, the volume as a whole comes across as an indictment of previous achievements in linguistics and a sermon that probability will somehow fix everything. The case is further undermined by a lack of historical context regarding the role of probability in linguistics: that we have had to quote BOTH Chomsky AND Labov to set the record straight is a sad commentary in itself. Moreover, the volume is based on a rather narrow set of goals and methods of linguistic research, and in some cases, misleading and inaccurate readings of the technical materials. The Maxim of Categoricity – if there ever was such a thing – seems to emerge largely unscathed. But the Maxim of Probability, so far as I can see, isn't quite forthcoming.

REFERENCES

Abney, Steven. 1996. Statistical methods and linguistics. In Phillip Resnick & Judith Klavins (eds.), *The balancing act*, 1–26. Cambridge, MA: MIT Press.

Abramowicz, Lukasz. 2006. Sociolinguistics meets exemplar theory: Frequency and recency effects in (ing). Presented at the 35th New Ways of Analyzing Variation (NWAV) Conference, Columbus, OH. [To appear in the proceedings.]

Anderson, John & Lael Schooler. 1991. Reflections of the environment in memory. *Psychological Science* 2, 396–408.

Angluin, Dana. 1988. *Identifying languages from stochastic examples* (Technical Report 614). New Haven, CT: Yale University.

Arbib, Michael. 1969. Memory limitations of stimulus–response theory. *Psychological Review* 76, 507–510.

Armstrong, Susan, Lila Gleitman & Henry Gleitman. 1983. What some concepts might not be. *Cognition* 13, 263–308.

Batchelder, William H. & Kenneth Wexler. 1979. Suppes' work in the foundations of psychology. In Radu Bogdan R. (ed.), *Profiles – An international series on contemporary philosophers and logicians: Patrick Suppes*, 149–186. Dordrecht: Reidel.

Bauer, Laurie. 2001. *Morphological productivity*. Cambridge: Cambridge University Press.

Boersma, Paul & Bruce Hayes. 2001. Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* 32, 45–86.

Bush, Robert & Frederick Mosteller. 1951. A mathematical model for simple learning. *Psychological Review* 68, 313–323.

Bybee, Joan. 2001. *Phonology and language use*. Cambridge: Cambridge University Press.

Cedergren, Henrietta & David Sankoff. 1974. Variable rules: Performance as a statistical reflection of competence. *Language* 50, 333–355.

Chen, Matthew Y. & William S.-Y. Wang. 1975. Sound change: Actuation and implementation. *Language* 51, 255–281.

Chomsky, Noam. 1955/1975. *The logical structure of linguistic theory*. Ms., Harvard University & MIT. [Published 1975. New York: Plenum; available http://alpha-leonis.lids.mit.edu/chomsky/ (19 September 2007).]
Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, Noam & Morris Halle. 1968. *The sound pattern of English*. New York: Harper & Row.
Clahsen, Harald. 1999. Lexical entries and rules of language: A multi-disciplinary investigation. *Behavioral and Brain Sciences* 22, 991–1060.
Clark, Brady. 2005. On stochastic grammar. *Language* 81, 207–217.
Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* 2, 83–149.
Coen, Michael. 2006. Multimodal dynamics: Self-supervised learning in perceptual and motor systems. Ph.D. dissertation, Department of Electrical Engineering and Computer Science, MIT.
Cohn, Aabigail. 2006. Is there gradient phonology? In Gisbert Fanselow, Caroline Frey, Matthias Schlesewsky & Ralf Vogel (eds.), *Gradience in grammar: Generative perspectives*, 25–44. Oxford: Oxford University Press.
Collins, Michael. 1999. Review of *Beyond grammar: An experience-based theory of language* by Rens Bod. *Computational Linguistics*, 25(3), 440–444.
Crain, Stephen & Paul Pietroski. 2002. Why language acquisition is a snap. *Linguistic Review* 19, 163–183.
Dinkin, Aaron. 2007. The real effect of word frequency on phonetic variation. Presented at the 31st Penn Linguistics Colloquium, Philadelphia, PA. [To appear in the proceedings.]
Dresher, Elan. 2003. On the acquisition of phonological contrasts. In Jacqueline van Kampen & Sergio Baauw (eds.), *The generative approaches to language acquisition*. Utrecht: 27–46.
Ellegård, Alvar. 1953. *The auxiliary* do*: The establishment and regulation of its use in English* (*Gothenburg Studies in English*). Stockholm: Almqvist and Wiksell.
Embick, David & Alex Marantz. 2005. Cognitive neuroscience and the English past tense: Comments on the paper by Ullman et al. *Brain and Language* 93(2), 243–247.
Forster, Kenneth. 1976. Accessing the mental lexicon. In Roger J. Wales & Edward C. T. Walker (eds.), *New approaches to language mechanisms*, 257–287. Amsterdam: North-Holland.
Forster, Kenneth. 1992. Memory-addressing mechanisms and lexical access. In Ram Frost & Leonard Katz (eds.), *Orthography, phonology, morphology and meaning*, 413–434. Amsterdam: Elsevier.
Gabbard, Ryan, Mitch Marcus & Seth Kulick. 2006. Fully parsing the Penn Treebank. *The Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 184–191. New York.
Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68, 1–76.
Gibson, Edward & Neal J. Perlmutter. 1998. Constraints on sentence comprehension. *Trends in Cognitive Sciences* 2(7), 262–268.
Gold, Mark. 1967. Language identification in the limit. *Information and Control* 10, 447–474.
Goodman, Joshua. 1996. Efficient algorithms for parsing the DOP model. *The Conference on Empirical Methods in Natural Language Processing*, 143–152.
Grey, Russell & Fiona Jordan. 2000. Language trees support the express-train sequence of Austronesian expansion. *Nature* 6790, 1052–1054.
Guasti, Maria T. 2002. *Language development: The growth of grammar*. Cambridge, MA: MIT Press.
Guion, Susan. 1995. Frequency effects among homonyms. *Texas Linguistic Forum* 35, 103–116.
Gussennhoven, Carlos & Natasha Warner (eds.). 2002. *Laboratory phonology 7*. Berlin: Mouton de Gruyter.
Guy, Gregory. 1991. Explanationin variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3, 1–22.
Guy, Gregogy. 2005. Grammar and usage: A variationist response. *Language* 81, 561–563.
Hale, John. 2001. A probabilistic Earley parser as a psycholinguistic model. *The 2nd North American Chapter of the Association for Computational Linguistics*, 1–8. Philadelphia, PA.
Hay, Jennifer. 2000. Causes and consequences of word structures. Ph.D. dissertation, Department of Linguistics, Northwestern University.
Horning, James. 1969. A study of grammatical inference. Ph.D. dissertation, Department of Computer Science, Stanford University.

Hyams, Nina. 1986. *Language acquisition and the theory of parameters*. Dordrecht: Reidel.
Jelinek, Frederick. 1998. *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.
Johnson, Mark. 2001. The DOP estimation method is biased and inconsistent. *Computational Linguistics* 28(1), 71–76.
Jurafsky, Daniel, Alan Bell & Cynthia Girand. 2002. The role of lemma in form variation. In Gussenhoven & Warner (eds.), 3–34.
Just, Marcel A. & Patricia A. Carpenter. 1992. A capacity theory of comprehension: Individual differences in working memory. *Psychological Review* 98, 122–149.
Kanwisher, Nancy. 2000. Domain specificity in face perception. *Nature Neuroscience* 3, 759–763.
Kay, Paul & Chad McDaniel. 1979. On the logic of variable rules. *Language in Society* 8, 151–187.
Kearns, Michael & Leslie Valiant. 1994. Cryptographic limitations on learning Boolean formulae and finite automata. *Journal of the ACM* [Association for Computing Machinery] 41, 67–95.
Keller, Frank & Ash Asudeh. 2002. Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry* 33, 225–244.
Kiparsky, Paul. 1995. The phonological basis of sound change. In John Goldsmith (ed.), *Handbook of phonological theory*, 640–670. Oxford: Blackwell.
Kornai, Andras. 1998. Analytic models in phonology. In Jacques Durand & Bernard Laks (eds.), *The organization of phonology: Constraints, levels, and representations*, 395–418. Oxford: Oxford University Press.
Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1, 199–244.
Labov, William. 1969. Contraction, deletion, and inherent variability of the English copula. *Language* 45, 715–762.
Labov, William. 1981. Resolving the neogrammarian controversy. *Language* 57, 267–308.
Labov, William. 1994. *Principles of language change: Internal factors*. Oxford: Blackwell.
Labov, William. 2006. A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34, 500–515.
Labov, William, Sharon Ash & Charles Boberg. 2006. *The Atlas of North American English: Phonetics, phonology, and sound change*. Berlin: Mouton de Gruyter.
Lappin, Shalom & Stuart M. Shieber. 2007. Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics* 43(2), 393–427.
Lasnik, Howard. 1989. On certain substitutes for negative data. In Robert J. Matthews & William Demopoulos (eds.), *Learnability and linguistic theory*, 89–105. Dordrecht: Reidel.
Lavoie, Lisa. 2002. Some influences on the realization of *for* and *four* in American English. *Journal of the International Phonetic Association* 32, 175–202.
Legate, Julie A. & Charles Yang. 2005. The richness of the poverty of the stimulus. Presented at the Golden Anniversary of *The Logical Structure of Linguistic Theory*, 16 July 2005. Cambridge, MA.
Legate, Julie A. & Charles Yang. 2007. Morphosyntactic learning and the development of tense. *Language Acquisition* 14, 315–344.
MacWhinney, Brian. 1995. *The CHILDES project*. Mahwah, NJ: Lawrence Erlbaum.
Marcus, Gary. 1993. Negative evidence in language acquisition. *Cognition* 46, 53–85.
Maye, Jessica, Janet Werker & LouAnn Gerken. 2002. Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* 82(3), B101–B111.
McMahon, April & Robert McMahon. 1995. Linguistics, genetics and archaeology: Internal and external evidence in the Amerind controversy. *Transactions of the Philological Society* 93, 125–225.
Mehler, Jacques, Marcela Peña, Marina Nespor & Luca Bonatti. 2006. The 'soul' of language does not use statistics: Reflections on vowels and consonants. *Cortex* 42, 846–854.
Murray, Wayne & Kenneth Forster. 2004. Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review* 111, 721–756.
Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Language* 79, 682–707.
Niyogi, Partha. 2006. *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.
Nowak, Martin, Natalia Komarova & Partha Niyogi. 2002. Computational and evolutionary aspects of language. *Nature* 417, 611–617.

Parducci, Allen & Linda F. Perrett. 1971. Category rating scales: Effects of relative frequency of stimulus values. *Journal of Experimental Psychology* 89(2), 427–452.
Peterson, Gordon & Harold Barney. 1952. Control methods used in a study of the vowels. *Journal of the Acoustical Society of America* 24, 175–184.
Phillips, Betty. 1984. Word frequency and the actuation of sound change. *Language* 60, 320–342.
Pierrehumbert, Janet B. 1994. Syllable structure and word structure. In Patricia A. Keating (ed.), *Papers in laboratory phonology* 3, 168–188. Cambridge: Cambridge University Press.
Pierrehumbert, Janet B. 2002. Word-specific phonetics. In Gussenhoven & Warner (eds.), 101–139.
Pinker, Steven. 1979. Formal models of language learning. *Cognition* 7, 217–283.
Pinker, Steven. 1999. *Words and rules*. New York: Basic Books.
Pitt, Leonard. 1989. Probabilistic inductive inference. *Journal of the ACM* [Association for Computing Machinery] 36, 383–433.
Poplack, Shana. 2001. Variability, frequency and productivity in the irrealis domain of French. In Joan Bybee & Paul Hopper (eds.), *Frequency effects and emergent grammar*, 405–428. Amsterdam: Benjamins.
Pritchett, Bradley. 1992. *Grammatical competence and parsing performance*. Chicago, IL: University of Chicago Press.
Ringe, Donald, Tandy Warnow & Ann Taylor. 2002. Indo-European and computational cladistics. *Transactions of the Philological Society* 100(1), 59–130.
Rissanen, Jorma. 1989. *Stochastic complexity in statistical inquiry*. Singapore: World Scientific.
Rizzi, Luigi. 2005. Grammatically based target inconsistencies in child language. In Kamil Ud Deen, Jun Nomujra, Barbara Schultz & Bonnie Schwartz (eds.), *The Inaugural Conference on Generative Approaches to Language Acquisition – North America* (*GALANA*) (MIT Working Papers in Linguistics), 19–49. Cambridge, MA: MIT.
Roberts, Julie. 1996. Acquisition of variable rules: A study of (-t, -d) deletion in preschool children. *Journal of Child Language* 24, 351–372.
Roeper, Tom. 2000. Universal bilingualism. *Bilingualism: Language and Cognition* 2, 169–185.
Saffran, Jenny, Elissa Newport & Richard Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language* 35, 606–621.
Sankoff, David. 1988. Variable rules. In Ulrich Ammon, Norbert Dittmar & Klaus Mattheier (eds.), *Sociolinguistics: An international handbook of the science of language and society*, 984–997. Berlin: Walter de Gruyter.
Sankoff, David & William Labov. 1979. On the uses of variable rules. *Language in Society* 8, 189–222.
Schütze, Carson T. 2005. Thinking about what we are asking speakers to do. In Stephan Kepser & Marga Reis (eds.), *Linguistic evidence: Theoretical, empirical, and computational perspectives*, 457–485. Berlin: Mouton de Gruyter.
Stockall, Linnaea & Alec Marantz. 2006. A single route, full decomposition model of morphological complexity: MEG evidence. *The Mental Lexicon* 1, 85–123.
Suppes, Patrick. 1969. Stimulus–response theory of finite automata. *Journal of Mathematical Psychology* 6, 327–355.
Suppes, Patrick. 1970. Probabilistic grammars for natural languages. *Synthese* 22, 95–116.
Tarr, Michael J. & Yi D. Cheng. 2003. Learning to see faces and objects. *Trends in Cognitive Sciences* 7(1), 23–30.
Ullman, Michael, Roumyana Pancheva, Tracy Love, Eiling Yee, David Swinney & Gregory Hickok. 2005. Neural correlates of lexicon and grammar: Evidence from the production, reading, and judgment of inflection in aphasia. *Brain and Language* 93(2), 185–238.
Valiant, Leslie G. 1984. A theory of the learnable. *Communications of the ACM* [Association for Computing Machinery] 27, 1134–1142.
Vapnik, Vapnik. 1995. *The nature of statistical learning theory*. Berlin: SpringerVerlag.
Vasishth, Shravan & Richard Lewis. 2006. Argument–head distance and processing complexity: Explaining both locality and anti-locality effects. *Language* 82, 767–794.
Wang, William S.-Y. 1969. Competing changes as a cause of residue. *Language* 45, 9–25.
Wexler, Kenneth. 1994. Optional infinitives, head movement, and the economy of derivation in child language. In David W. Lightfoot & Norbert Hornstein (eds.), *Verb movement*, 305–350. Cambridge: Cambridge University Press.
Xu, Fei & Steven Pinker. 1995. Weird past tense forms. *Journal of Child Language* 22, 531–556.

Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford: Oxford University Press.
Yang, Charles. 2004. Universal grammar, statistics, or both. *Trends in Cognitive Sciences* 8(10), 451–456.
Yang, Charles. 2005. On productivity. *Yearbook of Language Variation* 5, 333–370.
Yang, Charles. 2006. *The infinite gift: How children learn and unlearn languages of the world.* New York: Scribner.
Yang, Charles. To appear. Counting grammars. In Insa Guzlow & Natalia Gagarina (eds.), *Frequency effects in language acquisition*. Berlin: Mouton de Gruyter.
Zwicky, Arnold & Geoffrey K. Pullum. 1987. Plain morphology and expressive morphology. *The Thirteenth Annual Meeting, General Session and Parasession on Grammar and Cognition*, 330–340. Berkeley, CA: Berkeley Linguistics Society.

Author's address: *Department of Linguistics and Computer Science, 608 Williams Hall, University of Pennsylvania, Philadelphia, PA 19104, U.S.A.*
E-mail: charles.yang@ling.upenn.edu