1

Introduction

Ted Briscoe
Natural Language and Information
Processing Group, Computer Laboratory,
University of Cambridge

1.1 Linguistic theory and evolutionary theory

Taking an evolutionary perspective on the origins and development of human language, and on linguistic variation and change, is becoming more and more common, as the papers in Hurford *et al.*(1998) attest. The term 'evolution' now crops up regularly in work emerging from the broadly generative tradition in linguistic theory (e.g. Jackendoff, 1997; Steedman, 2000). The latter development is probably a more or less direct consequence of several influential attempts to reconcile Chomskyan nativism with evolutionary theory, primarily in terms of a gradualist and adaptionist account of the origins and development of the language faculty (e.g. Hurford, 1989; Newmeyer, 1991; Pinker and Bloom, 1990). However, most of the contributions to this book owe more to the complementary but very different insight (e.g. Hurford, 1987, 1999) that not only the language faculty *per se*, but also the origins and subsequent development of *languages themselves* can be fruitfully addressed within the framework of evolutionary theory. Under this view, languages are evolving, not metaphorically but literally, via cultural rather than biological transmission on a historical rather than genetic timescale. This represents a very distinct and quite narrow theme within the broader program of integrating linguistic theory and evolutionary theory, and it is this theme which is primarily addressed by the contributors to this volume.

Evolutionary ideas have had a rather checkered history within linguistic theory despite their close mutual influence in the nineteenth century. McMahon (1994:ch12) provides a brief account of this history and also discusses linguistic work influenced by evolutionary theory during the

fifties and sixties. However, the insight that languages *per se* can be studied as (culturally) evolving systems, post the modern synthesis in biology and post mathematical and computational work on dynamical systems, does not seem to have reemerged until the eighties when Lindblom (1986) in phonology, Keller (1984, (1994)) in historical linguistics, and Hurford (1987) in syntactic theory independently articulated this view (using somewhat different terminology). The idea is an instance of the 'universal Darwinist' claim (Dawkins, 1983; Dennett, 1995:343f) that the methodology of evolutionary theory is applicable whenever any dynamical system exhibits (random) variation, selection amongst variants, and thus differential inheritance. In the nineties, this perspective on languages has been espoused enthusiastically and persuasively by non-linguists (e.g. Cziko, 1995; Deacon, 1997). However, it has not had significant impact in mainstream linguistic theory as yet, perhaps partly because work has only recently begun to address questions seen as central to (generative) linguistic theory.

The contributions to this volume are less concerned with questions of linguistic origins or the development of a broad evolutionary account of human language, than with why and how specific syntactic universals evolved (Kirby, Batali, Briscoe), why homonymy and synonymy are present and maintained in vocabulary systems (Steels and Kaplan), the nature of (E-) language syntactic change (Niyogi, Briscoe), the kind of language learning mechanism required to not only acquire an existing linguistic system accurately but also impose further structure on an emerging system (Oliphant, Kirby, Worden), and the (co)evolution of language(s) and this learning mechanism (Turkel, Briscoe). A second and equally important way in which the contributions here represent a tightly circumscribed theme within evolutionary linguistic work is that all utilize a methodology of computational implementation and simulation of (more or less explicit) formal models. For this reason too, there is a close connection with formal generative linguistic theory. Mathematical modeling and/or computational simulation help ensure that theories constructed are complete and precise, and also help with their evaluation by making the assumptions on which they rest fully explicit. This is particularly critical in the context of largely speculative accounts of the prehistoric development of human languages, as without such a methodology there is little to constrain such speculation.

The rest of this introduction describes the key ideas and techniques which underlie the contributions to this book, and, more broadly, the

evolutionary approach to linguistic variation, change and development, relating them to current linguistic theory and discussing critical methodological issues. The contribution by Hurford contains a thorough and insightful analysis and comparison of five different computational models of linguistic evolution, two of which are described here (Batali, Kirby), as well as developing a more general framework for such comparisons that could, in principle, be applied to all the work presented here. Therefore, I limit myself here to additional, I hope complementary, remarks and refer the reader to Hurford's contribution for a much more detailed exposition of the general structure of many of the models.

1.2 The formal framework

1.2.1 Generative linguistics

Chomsky (1965) defined grammatical competence in terms of the language of (i.e. stringset generated by) an ideal speaker-hearer at a single instant in time, abstracting away from working memory limitations, errors of performance, and so forth. The generative research program has been very successful, but, one legacy of the idealization to a single speaker at a single instant has been the relative sidelining of language variation, change and development. More recently, Chomsky (1986) has argued that generative linguistics can offer a precise characterization of I-language, the internalized language or grammar of an individual speaker, but has little to say about E-language, 'external' language, which is an epiphenomenon of the I-languages of the individual speakers who comprise a speech community.

Consequently, the study of language change within the generative tradition has largely focused on 'I-language change'; that is, the differences between I-languages or their corresponding grammars internalized by child language learners across generations. And within I-language change on the (parametric) properties of internalized grammars (e.g. Lightfoot, 1979, 1999). The generative approach to language change treats (major) grammatical change as a consequence of children acquiring different grammars from those predominant amongst the adults in the population, perhaps as a consequence of variation in the internalized grammars of these adults. However, theories of language variation, change and development will (minimally) require an account of how the E-language(s) of an adult population can be defined in terms of the aggregate output of these (changing) individuals.

1.2.2 Language agents

A language agent is a idealized model of just what is essential to understanding an individual's linguistic behavior. I use the term 'agent', in common with several contributors to this volume and with (one) current usage in computer science and artificial intelligence, to emphasize that agents are artificial, autonomous, rational and volitional, and that agents are embedded in a decentralized, distributed system, i.e. a speech community.

A language agent must minimally be able to learn, produce and interpret a language, usually defined as a well-formed set of strings with an associated representation of meaning, by acquiring and using linguistic knowledge according to precisely specified procedures. Beyond this, the models of language agents deployed by the contributors differ substantially, depending on theoretical orientation and the precise questions being addressed. Oliphant, and Steels and Kaplan define linguistic knowledge entirely in terms of word–meaning associations in a lexicon, reflecting their focus on the acquisition of vocabulary. Niyogi, Turkel and Briscoe focus on the acquisition of parametrically-defined generative grammars and thus define linguistic knowledge primarily in terms of (sets of) parameter settings. Batali, Kirby and Worden all develop broadly lexicalist models of linguistic knowledge, in which the acquisition of lexical and grammatical knowledge is closely integrated.

All the models provide some account of the acquisition, comprehension and production of (I-) language. Again the details vary considerably depending on the theoretical orientation and questions being addressed. For example Niyogi and Turkel largely assume very idealized, simple accounts of parameter setting in order to focus on the dynamics of E-language change and the genetic assimilation of grammatical information, respectively. The other contributors concentrate on specifying acquisition procedures in some detail, since properties of the acquisition procedure are at the heart of linguistic inheritance and selection. As acquisition is closely bound up with comprehension, most of these contributors also develop detailed accounts of aspects of the comprehension, or at least parsing, of linguistic input. However, none really provide a detailed account of language production, beyond the minimal assumption that linguistic utterances are generated randomly from a usually uniform distribution over the strings licensed by an agent's grammar and/or lexicon.

Additionally, language agents can have further properties, such as the

ability to invent elements of language, the ability to reproduce further language agents, an age determining the learning period and/or their 'death', and so forth. For example, the contributors on the development of language or emergence of new traits, often endow their language agents with the ability to 'invent' language in the form of new utterance–meaning pairs, where the utterance can either be essentially an unanalysed atom ('word') or a string with grammatical structure ('sentence'). Invention is again modeled very minimally as a (rare) random process within a predefined space of possibilities, and is one method of providing the variation essential to an evolutionary model of linguistic change and/or development.

1.2.3 Languages as dynamical systems

E-languages are the aggregate output of a population of language users. Such a population constitutes a speech community if the internalized grammars of the users are 'close' enough to support mutual comprehension most of the time. Membership of the population/speech community changes over time as people are born, die or migrate.

Perhaps the simplest model which approximates this scenario is one in which the population initially consists of a fixed number of 'adult' language agents with predefined internalized grammars, and their output constitutes the data from which the next generation of 'child' language learning agents acquires new internalized grammars. Once the learning agents have acquired grammars, this new generation replaces the previous one and becomes the adult generation defining the input for the next generation of learners, and so on. We can define a dynamical model of this form quite straightforwardly. A dynamical system is just a system which changes over time. We represent it by a sequence of states where each state encodes the system properties at each time step and an update rule defines how state s^{t+1} can be derived from state s^t:

$$s^{t+1} = \mathit{Update}(s^t)$$

Time steps in this model correspond to successive non-overlapping generations in the population. Minimally, states must represent the E-language(s) of the current generation of language agents, defining the input for the next generation of learners. The *Update* rule must specify how the internalized grammars of the learners are derived from the E-language input.

Niyogi and Berwick (1997) develop a deterministic version of this model in which each state is defined by a probability distribution over triggers, a finite subset of unembedded sentences from each language defined by each internalized grammar present in the population. The deterministic update rule defines a new probability distribution on triggers by calculating the proportions of the population which will acquire the internalized grammars exemplified in the input data. In this volume, Niyogi describes this model in detail and develops it by exploring the predictions of deterministic update rules which assume that different learners will receive different input depending on their parents or on their geographical location. Niyogi shows how this model makes predictions about the direction and timecourse of E-language change dependent on the learning algorithm and the precise form of the update rule. Throughout, E-language change is modeled as a consequence of a number of 'instantaneous' I-language changes across generations, in common with standard generative assumptions about major grammatical change. However, the population-level modeling demonstrates that the consequent predictions about the trajectory and direction of change are often surprising, very varied, and always sufficiently complex that mathematical modeling and/or computational simulation are essential tools in deriving them.

Niyogi's use of deterministic update rules assumes that random individual differences in the learners' input are an insignificant factor in language change. In his model, learners are exposed to a finite number of triggers randomly drawn according to a probability distribution defined by the current adult population. Sampling variation may well mean that learners will or will not see triggers exemplifying particular internalized grammars present in the adult population. If the number of triggers sampled and/or the size of the population is large, then this variation is likely to be insignificant in defining the overall trajectory and timecourse of E-language change. Therefore, Niyogi models the behavior of an *average* learner in the population. In the limit, the behavior of the overall model will be identical to one in which the behavior of individuals is modeled directly but the population is infinite. The great advantage of this approach is that it is possible to analytically derive fixed points of the resulting dynamical models, and thus prove that certain qualitative results are guaranteed given the model assumptions.

The models utilized by the other contributors are all stochastic in the sense that they model the behavior of individual agents directly

and deploy stochastic or random agent interactions. Therefore, there may be sampling variation in learner input. Time steps of the resulting dynamical models are defined in a more fine-grained way in terms of individual agent interactions or sets of such interactions. For example, Batali, Kirby, Oliphant, and Steels and Kaplan all take individual linguistic interactions as the basic time step, so the update rule in their simulations is defined (implicitly) in terms of the effect on E-language of any change in the linguistic knowledge of two interacting agents. In these and most of the other models, language acquisition is no longer viewed as an 'instantaneous' event. Rather agents interact according to their (partial) knowledge of the E-language(s) exemplified in the environment and continue to update this knowledge for some subset of the total interactions allotted to them. Turkel uses a standard (stochastic) genetic algorithm architecture with fitness-based generational replacement of agents so that time steps in his system correspond to non-overlapping generations. However, the fitness of each agent is computed individually based on 10 learning trials between it and another randomly chosen agent in the current population. Briscoe defines time steps in terms of interaction cycles consisting of a set number of interactions proportional to the current population size. Agents interact randomly and a proportion of interactions will involve learners. Once a stochastic model of this type is adopted it is also easy to introduce overlapping generations in which learners as well as adults may contribute utterances to E-language. The stochastic approach provides greater flexibility and potential realism but relies even more heavily on computational simulation, as analytic mathematical techniques are only easily applicable to the simplest such systems. For this reason, it is important that the results of simulation runs are shown to be statistically reliable and that the stochastic factors in the simulation are not dominating its behavior.

Interestingly, though Kirby derives his results via a stochastic simulation of a single speaker providing finite input to a single learner, the critical time steps of his model are generation changes, in which the learner becomes the new adult speaker, and a new learner is introduced. Therefore, it would appear that the analytic model developed by Niyogi and Berwick could, in principle, be applied to Kirby's simulation. The effect of such an application would be to factor out sampling variation in learner input. It should then be possible to prove that the qualitative results observed are guaranteed in any run of such a simulation. Indeed, what we might expect is that, over the predefined meaning space,

a single *optimal* grammar, relative to the subsumption based grammar compression algorithm employed, is the sole fixed point of the dynamical system.

1.2.4 Languages as adaptive systems

Niyogi and Berwick (1997) argue that their model of E-language does not need or utilize a notion of linguistic selection between linguistic variants. However, the specific learning algorithm they utilize is selective, in the sense that it is parametric. They examine, in detail, the predictions made by the Trigger Learning Algorithm (TLA, Gibson and Wexler, 1994) embedded in their dynamical model. The TLA is a parameter setting algorithm based on the principles and parameters account of grammatical acquisition (Chomsky, 1981). The TLA selects one grammar from the finite space of possible grammars defined by the settings of a finite number of finite valued parameters. Thus, when faced with variation exemplifying conflicting parameter settings in the input, the TLA selects between the variants by assigning all parameters a unique value. So, selection between variants is a direct consequence of the learning procedure.

It is possible to imagine a learning procedure which when faced with variation simply incorporated all variants into the grammatical system acquired. Briscoe (2000a) describes one such algorithm in some detail. In order to claim that no selection between linguistic variants is happening in dynamical models of the type introduced in the previous section, we would need to demonstrate that the specific learning procedure being deployed by agents in the system was not itself selective in this sense. However, such a learning procedure seems implausible as a model of human language learning because it predicts that the dynamic of language change would always involve integration of variation and construction of larger and larger 'covering' grammars of learner input. Loss of constructions, competition between variants, and the very existence of different grammatical systems would all be problematic under such an account.

Once we adopt an account of language learning which is at least partially selective, then it is more accurate to characterize linguistic dynamical systems as *adaptive* systems; that is, as dynamical systems which have evolved in response to environmental pressure. In this case, to be learnable with respect to the learning algorithm deployed by child language learners (whatever this is). The nature of the pressure depends

on properties of the learning procedure and need not be 'functional' in the conventional linguistic sense. For example, the TLA selects between variants by either selecting the parameter setting dictated by the last unambiguous trigger (with respect to the relevant parameter) in the input before the end of the learning period or by making an unbiased random guess. Therefore, the relative frequency with which variants are exemplified in learner input is the main determinant of which variants are culturally transmitted through successive generations of language learning agents. However, most of the learning procedures developed by other contributors exhibit various kinds of inductive bias which interact with the relative frequency of variant input to create additional pressures on learnability.

It is striking that with the exception of Turkel's quite idealized account of learning (which is not intended as a serious model of parameter setting), the other contributors all develop learning algorithms which, unlike the TLA, incorporate Ockham's Razor in some form; that is, a broad preference for the *smallest* grammar and/or lexicon ('compatible' with the input). In addition, most of the models remain selective, in the sense defined above with respect to the TLA, in that they bias learning towards acquisition of *unambiguous* word-meaning associations and/or syntactic means of realizing non-atomic meaning representations. Indeed the latter bias is a direct consequence of the former, as alternative encodings of the mapping from meaning to form result in larger descriptions. All the models impose hard constraints in the form of representational assumptions about the kind of grammars and/or lexicons which can be acquired; that is, assumptions about the form of universal grammar. It is in terms of such representational assumptions which incorporate hard inviolable constraints on what can be learnt that the soft, violable constraints or inductive bias in favour of small unambiguous mappings can be stated. As these representational assumptions vary a good deal between the contributions, the precise effect of the bias will also vary. Nevertheless, very broadly, Ockham's Razor creates an additional selection pressure for regularity in linguistic systems, over and above the requirement for frequent enough exemplification in learner input.

One might argue that the incorporation of such inductive biases into these models is no more than a method of ensuring that the simulations deliver the desired results. However, Ockham's Razor has been a central tenet of learning theory for centuries, and in the theory of informational complexity has been formally proved to provide a universally

accurate prior or inductive bias over a universal representation language (Rissanen, 1989). In the framework of Bayesian learning, the minimum description length principle, over a given representation language or class of grammars/models, provides a concrete, practical instantiation of Ockham's Razor, which has been used to develop learnability proofs for non-finite classes of grammar (e.g. Muggleton, 1996) and to develop theoretical and computational models of lexical and grammatical acquisition (e.g. Brent and Cartwright, 1996; de Marcken, 1996; Rissanen and Ristad, 1994; Osborne and Briscoe, 1997). Therefore, the learning procedures developed here, which incorporate this principle in some form, are not in any way unusual, controversial or surprising. Indeed, inductive bias has been argued to be essential to successful learning (Mitchell, 1990, 1997), this insight is central to the Bayesian framework, and within the space of possible inductive biases, Ockham's Razor remains the single most powerful and general principle, which under the idealized conditions of a universal representation language has been shown to subsume all other forms of bias (e.g. Rissanen, 1989).

Kirby (this volume, 1998, 2000) extends this insight in several ways arguing that the bias for smaller grammars is tantamount to the assumption that learners generalize from data and will, therefore, be a component of any language learning procedure. He argues that the syntactic systems which emerge in his simulations would emerge given many other possible learning procedures. Oliphant, in the context of word learning, similarly argues that the only kind of learning procedure which will *impose* order on random, inconsistent vocabulary systems is one which prefers unambiguous word-meaning mappings. However, as we have seen above, at root this follows from Ockham's Razor, since this is equivalent to saying that a learner prefers to retain the smallest number of word–meaning associations.

The picture which emerges then, is that languages have adapted to the human language learning procedure, in the sense that this procedure incorporates inductive bias – itself virtually definitional of the concept of learning. Inductive bias creates linguistic selection for more learnable linguistic variants relative to this bias and thus as languages are culturally transmitted from generation to generation via successive child language learners, linguistic systems will evolve that fit, or are adapted to, these biases. However, this picture cannot be the whole truth, for if it were we would predict that all languages should eventually converge to a single optimal system, that change should always be unidirectional, and