Modeling language acquisition, change and variation

Willem Zuidema
Language Evolution and Computation Research Unit
School of Philosophy, Psychology and Language Sciences
and Institute of Animal, Cell and Population Biology
University of Edinburgh
40, George Square
Edinburgh EH8 9LL, United Kingdom
jelle@ling.ed.ac.uk

http://www.ling.ac.uk/ ∼ jelle

Abstract

The relation between Language Acquisition, Language Change and Language Typology is a fascinating topic, but also one that is difficult to model. I focus in this paper on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and language change is parameter change. I review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach that is based on "Explicit Induction" algorithms for grammatical formalisms. I discuss which approach is most useful for which problems.

1 Language acquisition, change and typology

Every healthy human infant is capable of acquiring any one of a dazzling variety of human languages. This simple fact poses two fundamental challenges for linguistics: (1) understanding how children are so extremely successful at this apparently complex task, and (2) understanding how, although all humans have such similar linguistic abilities, such a wide variety of languages has emerged. These challenges are intricately linked: the languages that we observe today, are the result of thousands of years of cultural transmission, where every generation has acquired its language from the observed use by previous generations. That makes the acquisition of language a rather unique learning problem for learning theory, because what is being learned is itself the result of a learning process. Conversely, the structure of a language (say modern English) at any one time (say, 2003) is the result of perhaps millions of individuals learning from examples from a language with a very similar structure (say, the English of the 1960s).

This so-called *circular causality* (Steels, 1999) makes the relation between Language Acquisition, Language Change and Language Typology a fascinating topic, but also one that is difficult to model. I will focus here on the question how theories of language acquisition constrain theories of language change and typology. In the generative tradition and "Learnability Theory" this problem is approached by assuming that all linguistic variation can be described in terms of a relatively small number of parameters of a universal, innate core, the Universal Grammar. In this view, language acquisition is parameter setting, and

language change is parameter change. In the following I will review some simple acquisition models and their consequences for language change, and discuss some problems with this approach. I will then discuss an alternative approach from the emerging field of computational modeling of the evolution of language (Kirby, 2002b), that is based on "Explicit Induction" algorithms for grammatical formalisms. I will argue that the differences between the two approaches have been exaggerated, and will discuss for which sort of problems which sort of approach is most useful.

2 Parameter models

The "Parameter change" approach to this problem is based on *parameterizing* linguistic structure, such that we can characterize all differences between possible human languages by a vector of a small number of parameters. E.g., in the Principles and Parameters approach (Chomsky, 1981; Bertolo, 2001), language acquisition is described in terms of parameter settings for a universal core, the Universal Grammar. With such a description of language in hand, we can reformulate the challenges as follows: (1) how can learning, given primary linguistic data that conforms to any particular set of parameters, find that set of parameters? (2) given a set of learning procedures that are capable of finding the correct parameters, which ones predict the type of language change and statistical distributions (universals tendencies, Kirby 1999) that we can actually observe?

2.1 Parameter setting

In the "parameter setting" models of language acquisition, one assumes a finite number N of possible grammars. If all variation can be described by n different, Boolean and independent parameters, such that the total number of possible grammars is $N = 2^n$. Such parameters determine, for instance, whether or not an object precedes the main verb in a sentence, or whether or not the subject can be left out. Typically, although the number of parameters is estimated at around 30, concrete examples are only worked out for the 2 or 3 least controversial proposed parameters. A lot of work in parameter setting works with rather simplified models that can be studied analytically, and that depend only on the finiteness of N. Examples of such models are "memory-less learning", "batch learning" (e.g. Nowak *et al.*, 2001) and "learning by enumeration" (Gold, 1967). It is useful to look in a bit more detail at these models.

Memory-less learning (Niyogi, 1998) is arguably the simplest language acquisition model. The algorithm works by choosing a random grammar from the set of possible grammars each time the input data shows that the present hypothesis is wrong. The algorithm obviously is not very efficient, because it can arrive at hypotheses it has already rejected before; i.e. each time it randomly chooses a new grammar, it forgets what it has learned from all data it has received before. This algorithm is only of interest because it is simple and provides a lower bound on the performance of any reasonable algorithm (Nowak *et al.*, 2001).

The *batch learner*, in contrast, memorizes all received sentences and finds all grammars from the set of possible ones that are consistent with these sentences. Equivalently, it keeps track of all possible grammars that are still consistent with the received data. In any case, for any reasonably large set of possible grammars, the batch learner has monstrous memory and processing requirements. Its value lies in the fact that it is simple, and provides an upper bound on the performance of any reasonable learning algorithm, as long as there is no a-priori reason to prefer one grammar that is consistent with the data over another.

As exemplified by appendix A, we can, with a bit of effort, derive explicit formulas that describe the probability of success q as a function of the number of input sentences for both the memory-less and the batch learner. Under the assumption that every wrong grammar is equally similar to the right grammar (described with a similarity parameter a), we can in fact give a complete transition matrix T, where all

diagonal values are $q_{memoryless}$ and all off-diagonal values are $(1 - q_{memoryless})/(N - 1)$. This transition matrix plays an important role in models of language change described in the next section.

It is important to realize that these algorithms only work because a finite (and in fact, relatively small) number of possible grammars is assumed. Moreover, calculations such as in appendix A are relatively easy due to some important assumptions: (1) that the algorithms are not biased at all to favor certain possible grammars over others; (2) that (in the case of the memory-less learner) the probability of jumping to a wrong or right grammar remains constant throughout the learning process; and (3) that all grammars are equally similar to each other. Without these assumptions, similar calculations quickly get rather complex.

For instance, *learning by enumeration* (Gold, 1967), as the name suggests, proceeds by enumerating one at a time, and in prespecified order all possible grammars. Only if a grammar is inconsistent with incoming data ("text"), does the algorithm move on to the next grammar. The procedure is of interest, because it can be used as a criterion for learnability (Gold, 1967)[1]. Calculating q is more difficult than before, because the probably of changing to a wrong grammar *decreases* over time.

The *trigger learning algorithm* (Wexler & Culicover, 1980) is a popular model that is of (slightly) more practical interest. Rather than picking a random new grammar, as the memory-less learner does, or enumerating grammars in a random order, as in learning by enumeration, it changes a random parameter when it finds an input sentence that is inconsistent with the present hypothesis. If with the new parameter setting the sentence can be parsed, the change is kept, otherwise it is reverted. The trigger learning algorithm thus implements a kind of hill-climbing (gradient ascent), by keeping parameters that do well and only making a small change when it improves performance. The probability of the trigger algorithm to give the right grammar after b sentences is even more tricky to calculate, because the probability to reject a wrong hypotheses *decreases* as more and more parameters get correctly set.

Many other parameter setting models exists. E.g. Briscoe (2002a) develops a variant of the trigger learning algorithm, where parameters are no longer independent, but fall into linguistically motivated inheritance hierarchies. Further, rather than choosing a single parameter at random and changing it, as in the TLA, Briscoe's algorithm selects several random parameters and keeps track of their most likely setting in a Bayesian, statistical fashion. Yang (2000) argues that language acquisition is best viewed as a selectionist process, where many different parameter sets are considered in parallel. Niyogi & Berwick (1995) and Yang (2000) consider the further complication that children learn from input sentences that are drawn from different languages, and explore the expectations on what grammar settings they will end up with. In all these models, calculating the probabilities of the outcome of learning gets very complex and results are typically obtained by using computer simulations.

2.2 Parameter change

Niyogi & Berwick (1995), as well as neural network modelers Hare & Elman (1995), argue that a theory of language acquisition – and the mistakes children make when confronted with insufficient or ambiguous input – implies a theory of language change. Similarly, Kirby (1999) explores the idea that a theory on language use and processing – which alter the primary linguistic data – leads to specific expectations on language change and the resulting linguistic variation. Hence, by working out the consequences for language change and comparing them to empirical data, theories on language use, processing and acquisition can be

[1] Learning by enumeration can, within finite time, find the target grammar from a class of grammars if the following conditions hold: (1) the class of grammars is finite (enumerable), (2) for every two grammars in the class, there exists a sentence that distinguishes between two grammars (i.e. that is grammatical according to one, and ungrammatical according to the other), and (3) the distinguishing sentence will occur within a finite amount of time in the the text, generated by the target grammar. It follows that the class of grammars is then learnable from text. It can be shown that superfinite classes of grammars, such as the context-free or context-sensitive grammars, are not learnable in this sense (Gold, 1967). Principles & Parameters-models, in contrast, are learnable (Wexler & Culicover, 1980) and so are many other classes (Angluin, 1980).

tested. Formally, a class of grammars $\mathcal{G}$, a learning algorithm $\mathcal{A}$ and a model of the primary linguistic data (a probability distribution $\mathcal{P}_i$ over the possible sentences of language i) together constitute the main ingredients of a dynamical system that describes the change in numbers of speakers of each language[2].

Several general results have been obtained. For instance, Niyogi & Berwick (1995) and Yang (2000) find that with different choices for $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$, the change in the number of speakers of a particular language tends to follow an S-shaped curve, consistent with observed patterns in historical data. More interestingly, Nowak *et al.* (2001) derive a *coherence threshold*. In their model, natural selection selecting for more frequent grammars, helps a population to converge on a specific grammar. Mistakes in learning, on the other hand, lead to divergence, because it essentially randomizes the choice of grammars. Nowak et al. find that if the accuracy in learning is below a precise threshold, all coherence in the population is lost and all languages are spoken with equal probability[3].

Niyogi and Berwick apply their methodology to a number of case studies. For instance, they look at a simple 3-parameter system where the parameters determine whether or not specifiers (1) and complements (2) come before the head of a phrase, and whether or not the verb is obligatorily in second position (3). In this system, there are 8 different possible grammars (languages). By making assumptions on the frequency with which triggers for each of the parameters are available to the child, they can estimate the probability a specific learning algorithm can learn each language. They numerically determine the probabilities of transitions between each of the 8 language over 30 generations with 128 triggers per generation. They find that languages with the third parameter set to "0" ($V2-$) are extremely unstable and that the $V2+$ parameter therefore quickly gets fixed in all simulations. This observation is contrary to observed trends in historical data, where $V2+$ is typically lost. Niyogi and Berwick argue that this falsifies their preliminary model, and thus illustrates the feasibility of testing the diachronic accuracy of the assumptions on $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$.

2.3 Some features of parameter change models

Several other parameter change models have been studied. They have in common the emphasis on the uniformity of languages, i.e. all possible languages (grammars) are of equal quality. Hence, children acquiring a language do not go from a simple grammar to a more complex one, but rather jump from one grammar to an equally complex alternative. Not the quality of the language, but the uncertainty about which is the correct one changes over time.

Moreover, in all these models the acquisition of syntax is studied independently from the acquisition of phonology, semantics, pragmatics and the lexicon, and, usually, independent from the particularities of the child's parsing algorithm. The training data are "triggers", i.e. strings of grammatical categories. The problems of learning the syntactic categories of words and their meaning, and learning to recognize the phonological form and the boundaries between words are all ignored.

Further, the models fit into a tradition that is much mathematically oriented. Although many results are obtained through numerical simulations, the models are formulated at a rather abstract level. Generations are typically discrete, the number of parameters small (2, 3, 5), number of training samples and the number of individuals in a population very small or, alternatively, infinite.

The models are valuable, because they give a *general* insight in how linguistic conventions can change and spread in a population. However, the problem with this approach is that its potential for explaining *specific* aspects of language acquisition and language typology depends completely on the successful parametrization of linguistic descriptions. That dependence has advantages, because it makes the relation with other linguistic theories very clear, but it has some major disadvantages as well.

[2] In addition to the triple $\{\mathcal{G}, \mathcal{A}, \mathcal{P}\}$ (Niyogi & Berwick, 1995), one needs assumptions on population and generation structure and the number of training sentences the algorithm receives.

[3] Presumably, a similar mechanism explains the lack of coherence in the simulations of Niyogi & Berwick (1995).

First, there is, as for now, no such parametrization available. If efficient parametrization (i.e. with 20 or 30 parameters) turns out to be impossible, models that depend on them will be inadequate. Second, even if it is possible in principle, without a complete theory available on what each parameter means, solutions in terms of these parameters give little insight on why children learn certain things with more ease than others, or why languages tend to show certain patterns more often than others. Finally, parameter-models might give an adequate description of the variation in languages in a quasi-stable state, but that does not necessarily mean that they also give an adequate description of language variety when languages are changing. In particular, observed trends in language change regarding the interaction between phonology, syntax, semantics and pragmatics seem hard to capture in available parameter models.

3 Explicit Induction

3.1 Grammar Induction: impossible and irrelevant?

Grammar Induction algorithms are usually based on the intuition that the frequency of occurrence of substring in the training sentences, and the contexts in which they appear, contain information on what the underlying constituents and the rules of combination of the target grammar are. E.g. Zellig Harris, in describing the methods linguists use to infer the grammar of an unknown language, defines the crucial concept of "substitutability" as follows: "If our informant accepts DA'F as a repetition of DEF, and if we are similarly able to obtain E'BC as equivalent to ABC, then we say that A and E are mutually substitutable" (Zellig Harris, 1951, quoted in van Zaanen 2001).

It is possible to design induction algorithms that, just like Harris's linguist, use observed patterns in training sentences to induce the underlying grammar. However, due to initial negative results on the theoretical possibility of learning a grammar from positive data (Gold, 1967) and developments in linguistic theory (e.g. Chomsky, 1965), the *induction* of grammar has been widely viewed as both impossible and irrelevant.

The supposed impossibility of grammar induction is based on a widespread misinterpretation of negative learnability results. Gold (1967) showed that e.g. the class of context-sensitive languages is not *identifiable in the limit*. Even we if accept identification in the limit as the appropriate criterion for learnability, Gold's results mean nothing more than, in his own words:

> *"The class of possible natural languages is much smaller than one would expect from our present models of syntax. That is, even if English is context-sensitive, it is not true that any context-sensitive language can occur naturally. Equivalently, we may say that the child starts out with more information than that the language it will be presented is context-sensitive. In particular, the results on learnability from text imply the following: The class of possible natural languages if it contains languages of infinite cardinality, cannot contain all languages of finite cardinality." (Gold, 1967)*

In other words, a class of context-sensitive grammars needs to be constrained to make it learnable. Angluin (1980) has shown that very non-trivial classes of formal languages are learnable. Nothing in the formal results, however, proves that the necessary restrictions are due to an extensive, innate, language-specific Universal Grammar; they could be simply generic properties of the human brain[4].

The supposed irrelevance of grammar induction algorithms is based on the fact that the dominant linguistic theories of the last decades assume extensive innate knowledge. If children don't do grammar induction, why design computer programs that do? Evidence for this view comes – in addition to the learnability

[4] Although it is of course true that learnability is a valid test for judging the validity of a (grammatical) theory, and that few proposed theories other than those from the nativist tradition pass it. However, one can argue that nativist theories, rather than solving the learnability problem, simply shift it to the domains of evolutionary theory and cognitive neuroscience.

results – from empirical observations in child language acquisition. Typically, such arguments have the form: the child correctly uses construction X very early in life, even though the primary linguistic data it has received up to that point does not provide enough evidence to choose between X and several alternative logical possibilities. Thus, it is concluded, the child must have prior (innate) knowledge of X.

More and more it is now recognized that this "knowledge of X" might be an emergent result of the interaction between not necessarily language-specific cognitive and learning abilities, and the structure, meaning and pragmatics of the linguistic data the child received (MacWhinney, 1999). Consequently, the need to postulate language-specific adaptations might be limited (Jackendoff, 2002; Hauser *et al.*, 2002).

3.2 Induction Algorithms

Wolff (1982), and similarly Stolcke (1994), Langley & Stromsten (2000) and Zuidema (2003), presents a model based on the idea that a grammar is a compressed representation of a possibly infinite language (string set). These algorithms all use context-free grammars as the grammar formalism, learn from text and run through three phases that can be termed "incorporation", "compression" and "generalization". I will refer to these algorithms as "compression-based induction".

In the incorporation phase, input sentences s are stored as idiosyncratic rewrite rules $S \mapsto s$. In the compression phase (or "syntagmatic merging"), the most frequent substrings z in the right-hand sides of the stored rules are replaced by a unique non-terminal symbol N. Rules of the form $N \mapsto z$ are added to the grammar. In the generalization phase (or "paradigmatic merging"), two nonterminals N and N' are considered *substitutable* if they occur in the same context; all occurrences of N' are then replaced by N. Different variants of the basic algorithm differ in how *greedy* they are, and in whether or not they are *incremental*. Kirby (2000), and later papers, uses a algorithm were the context-free grammars are enriched with a predicate-logic based semantics.

A related framework based on substitutability is developed by van Zaanen (2001) and termed "Alignment Based Learning" (ABL). Van Zaanen develops a number of algorithms for the two phases of the ABL framework: Alignment learning and selection learning. In the alignment learning phase input sentences are compared, aligned and common substrings are identified. The *unequal* parts z and z' of the two sentences are labeled with a non-terminal. The non-terminal is unique if neither z nor z' was labeled already, but the algorithm reuses the existing label if available, and equates the two non-terminals if both z and z' were labeled already. In the latter two conditions a form of generalization occurs. Each labeling is a hypothesis on a possible constituent of the target language, and very many such hypotheses are generated.

In the selection learning phase, a subset of the generated hypotheses is selected. That subset is chosen such that it is concise (each hypothesis can be used to analyze many sentences), and that it is internally consistent (hypotheses do not overlap). The ABL algorithm yields a tree-bank: an annotated version of the input corpus (it thus implements automated tagging). From the tree-bank, context-free grammars can be trivially induced.

3.3 Language Evolution

In the "Explicit Induction" approach to modeling language change and evolution, language change is studied based on similar induction algorithms, i.e. learning algorithms that produce an explicit grammar based on training sentences (see Hurford, 2002, for a review). Such an approach avoids the problems of parameter models, because they can incorporate any available linguistic formalism. However, they have two major disadvantages as well: (1) language induction is very challenging problem that is far from solved, even for simplified and well understood grammar formalisms; (2) models that incorporate a full-blown linguistic formalism, including procedures for language production and interpretation, quickly get very complex.

Two recent models by Kirby (2002a) and Batali (2002) show that there is reason for optimism for progress on bl problems. Kirby presents a model that is very clear in its set-up. It uses first-order predicate logic with a small set of entities and predicates to represent semantics, and a extension of context-free grammars to represent syntax and the syntax-semantics mapping. The model thus uses well-understood and conventional linguistic formalisms and a simple learning procedure. However, by using the output of one learning cycle as input for the next Kirby was able to get some unconventional results: the spontaneous emergence of a recursive, infinite but learnable language. However, the learning algorithm used is very brittle, and it's difficult to extend the model to domains with more diverse semantics and a more heterogeneous syntax.

In contrast, Batali's model is very difficult to understand. It also uses a form of predicate logic to represent semantics, but it uses "exemplars" as the basic representation of the grammar, and "argument maps" to guide the combination of exemplars into meaningful sentences. The results show the emergence of a complex language, with properties similar to case marking and subordinate clause marking in natural languages. The emergent languages are essentially infinite but nevertheless learnable (from meaning–form pairs). The learning algorithm is successful and robust in this complex domain presumably because of the redundancy it allows.

3.4 Some features of explicit induction models

Several other explicit induction models have been studied. They have in common that no uniformity of languages are assumed. Typically, individuals in these models start with an empty grammar and empty lexicon, and gradually add new rules and lexical items based on the received sentences and observed patterns. Individuals are, however, equipped with an invention procedure, such that they can generate new sentences when required.

Further, in these models learning is typically from form–meaning pairs and a lexicon is built-up in parallel with the grammar. The recognition of phonemes and the pragmatics of dialogs are built-in as assumptions of the models.

The models are all implemented as computer programs. Typically, the models are rather concrete: they consist of a population of individuals, with procedures for production, invention, interpretation and induction, and a set of possible message to communicate. The languages studied in these models are still relatively simple, and exhibit just some basic word orders or morphological markers for the semantic roles of agents, patients and action. Empirical data from historical linguistics has so far played no role in these studies.

4 Discussion

I have reviewed some models of language acquisition and language change from two different traditions. The crucial question – which approach is best? – is still largely open to discussion. The following issues are important in comparing both approaches:

Learnability - Theoretical arguments. From the field of learnability theory it has sometimes been argued that grammar induction is impossible. In section 3.1 I have argued that this position is based on a misunderstanding of the negative learnability results. Learnability, however, is an important test for the validity of a grammar formalisms and induction algorithms. The challenge is to find a combination of a formalism that is as expressive as human languages are (i.e. mildly context-sensitive), and a learning algorithm that can induce it from the available primary linguistic data. In my view, parameter setting

models meet this challenge, but only by making unsatisfactory assumptions on the prior knowledge the algorithms start with. Explicit induction models, on the other hand, present considerable progress (i.e. most work with context-free grammars), but more work still needs to be done.

Learnability - Empirical arguments. From the field of psycholinguistics it has been argued that children have prior knowledge of syntactic constructions, because they choose, from apparently many logical possibilities that are consistent with the received evidence, the correct, seemingly arbitrary option. Grammar induction models, in this view, are – if not impossible – irrelevant, because children do not do induction. I believe that explicit induction algorithms have already shown that the logic of this argument is false. There is no need for assuming explicit prior knowledge, because the outcome of the interaction between learning biases and training data is subtle and often unexpected. Moreover, because languages are transmitted culturally from generation to generation, seeming arbitrary choices are likely to be the correct ones, because previous generations have used the same arbitrary learning algorithm to learn their language (Deacon, 1997; Kirby, 2000; Briscoe, 2002a; Zuidema, 2003).

Equivalence More subtly, it has been suggested that explicit induction models might in some sense be equivalent to parameter setting models. If the space of grammars that induction algorithms explore is finite, then that space could in principle be parametrized and hence described by a finite number of parameters. The induction algorithm can then be described, albeit possibly in a clumsy and complicated way, as a parameter setting procedure. If this is true – and it presumably is for the context-free grammar and finite-state machine inducers – the crucial issue is parsimony and clarity. Presumably, for some purposes the representation in terms of parameters is more useful, but for comparison with psycholinguistic, neurological and historical data the explicit grammar representation seems more appropriate. Further, the parameterized representation leads naturally to the uniformity assumptions, whereas the explicit grammar representation leads naturally to the view that grammars grows over time. Finally, stochastic grammar formalisms can not be parametrized in the concise way that parameter setting models usually assume. Worse, lexicalized, exemplar-based models can not be parametrized because there are infinitely many probability distributions that can be assigned to the string set (Bod, 1998).

In conclusion, the two approaches to modeling of language change are rooted in different theoretical positions on the nature of language and language acquisition. If one adopts the Principles and Parameters framework, the parameter change approach is the appropriate way to conceptualize language change. However, this approach requires more work to make explicit how each parameter is to be interpreted, which triggers for each parameter are available, how the child learns her lexicon and recognizes syntactic categories in the sentences it receives, how parameters depend on each other, etc. Moreover, it requires a satisfactory explanation for the evolution and development of the Universal Grammar in the child's brain. However, some Explicit Induction models might, even if one adopts this approach, still be useful as an equivalent representations that can be more easily compared to empirical data.

If one rejects the Uniformity Hypothesis and conceptualizes grammar acquisition as the gradual built-up of a grammar in the mind of the child, explicit induction models are the appropriate approach. Parameter change models are still useful as simple, but mathematically sophisticated models of how conventions spread in a population.

References

ANGLUIN, D. (1980). Inductive inference of formal languages from positive data. *Information and Control* **21**, 46–62.

BATALI, J. (2002). The negotiation and acquisition of recursive grammars as a result of competition among exemplars. In: Briscoe (2002b).

BERTOLO, S., ed. (2001). *Language Acquisition and Learnability*. Cambridge University Press.

BOD, R. (1998). *Beyond Grammar: An experience-based theory of language*. Stanford, CA: CSLI.

BRISCOE, T. (2002a). Grammatical acquisition and linguistic selection. In: Briscoe (2002b).

BRISCOE, T., ed. (2002b). *Linguistic evolution through language acquisition: formal and computational models*. Cambridge University Press.

CHOMSKY, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

CHOMSKY, N. (1981). *Lectures on Government and Binding*. Dordrecht: Foris.

DEACON, T. (1997). *Symbolic species, the co-evolution of language and the human brain*. The Penguin Press.

GOLD, E. M. (1967). Language identification in the limit. *Information and Control (now Information and Computation)* **10**, 447–474.

HARE, M. & ELMAN, J. (1995). Learning and morphological change. *Cognition* **56**, 61–98.

HAUSER, M., CHOMSKY, N. & FITCH, W. (2002). The faculty of language: what is it, who has it, and how did it evolve? *Science* **298**, 1569–1579.

HURFORD, J. R. (2002). Expression / induction models of language. In: Briscoe (2002b).

JACKENDOFF, R. (2002). *Foundations of Language*. Oxford, UK: Oxford University Press.

KIRBY, S. (1999). *Function, selection and innateness: The emergence of language universals*. Oxford University Press.

KIRBY, S. (2000). Syntax without natural selection: How compositionality emerges from vocabulary in a population of learners. In: *The Evolutionary Emergence of Language: Social function and the origins of linguistic form* (Knight, C., Hurford, J. & Studdert-Kennedy, M., eds.). Cambridge, UK: Cambridge University Press.

KIRBY, S. (2002a). Learning, bottlenecks and the evolution of recursive syntax. In: Briscoe (2002b).

KIRBY, S. (2002b). Natural language from artificial life. *Artificial Life* **8**, 185–215.

KOMAROVA, N., NIYOGI, P. & NOWAK, M. (2001). The evolutionary dynamics of grammar acquisition. *J. Theor. Biology* **209**, 43–59.

LANGLEY, P. & STROMSTEN, S. (2000). Learning context-free grammars with a simplicity bias. In: *Proceedings of the Eleventh European Conference on Machine Learning*, pp. 220–228. Barcelona: Springer-Verlag.

MACWHINNEY, B., ed. (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates.

NIYOGI, P. (1998). *The informational complexity of learning*. Boston, MA: Kluwer.

NIYOGI, P. & BERWICK, R. C. (1995). The logical problem of language change. Tech. rep., M.I.T.

NOWAK, M. A., KOMAROVA, N. & NIYOGI, P. (2001). Evolution of universal grammar. *Science* **291**, 114–118.

STEELS, L. (1999). The puzzle of language evolution. *Kognitionswissenschaft* **8**.

STOLCKE, A. (1994). *Bayesian Learning of Probabilistic Language Models*. Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, University of California at Berkeley.

WEXLER, K. & CULICOVER, P. (1980). *Formal principles of language acquisition*. Cambridge MA: MIT Press.

WOLFF, J. G. (1982). Language acquisition, data compression and generalization. *Language & Communication* **2**, 57–89.

YANG, C. D. (2000). Internal and external forces in language change. *Language Variation and Change* **12**, 231–250.

VAN ZAANEN, M. (2001). *Bootstrapping Structure into Language: Alignment-Based Learning*. Ph.D. thesis, School of Computing, University of Leeds.

ZUIDEMA, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In: *Advances in Neural Information Processing Systems 15 (Proceedings of NIPS'02)* (Becker, S., Thrun, S. & Obermayer, K., eds.). Cambridge, MA: MIT Press. (forthcoming).

A Memory-less learner and batch learner

To estimate the probability that memory-less learning finds the correct grammar after a certain number (b) of sample sentences, we need to consider the inverse: the probability that the algorithm still has a wrong hypothesis after b sample sentence.

$$P(right\ grammar\ after\ b\ samples) = 1 - P(wrong\ grammar\ after\ b\ samples) \tag{1}$$

The probability that the learner still has the wrong hypothesis, depends on the probability that it initially chose the wrong hypothesis (simply $(N-1)/N$) times the probability that it remained for all b sentences at a wrong hypothesis. If it makes no essential difference which wrong grammar is the present hypothesis and how long it has held it as the hypothesis[5], the probability that the algorithm remain for b sentences at a wrong hypothesis is simply $P(remain)^b$. Hence,

$$P(wrong\ grammar\ after\ b\ samples) = \frac{(N-1)}{N}\left(P(remain)\right)^b \tag{2}$$

The probability to remain at a wrong grammar for each random input sentence is given by the probability that that input sentence happens to be consistent with the present (wrong) grammar, plus the probability that the algorithm jumps to another wrong grammar:

$$P(remain) = P(consistent) + P(another\ wrong\ grammar) \tag{3}$$

The probability that a sentence is consistent with a wrong grammar is simply the similarity parameter a in Nowak *et al.* (2001). The probability that the algorithm jumps to another wrong grammar is given by the probability that the input sentence is inconsistent $(1-a)$ times the fraction of other wrong grammars $((N-2)/N)$.

Putting all this together, the probability (q) that the memory-less learner has found the correct grammar after b input sentences is given by (Komarova *et al.*, 2001)[6]:

$$\begin{aligned} q_{memoryless} &= 1 - \frac{(N-1)}{N}\left(a + \frac{(N-2)(1-a)}{N-1}\right)^b \\ &= 1 - \frac{(N-1)}{N}\left(1 - \frac{(1-a)}{N-1}\right)^b \end{aligned} \tag{4}$$

The probability that the batch learner has found the correct grammar after b input sentences is found by Nowak *et al.* (2001) to be

$$q_{batch} = \frac{\left(1 - \left(1 - a^b\right)^N\right)}{(Na^b)} \tag{5}$$

[5]That is the case, for the memory-less learner, under the assumption of Nowak *et al.* (2001) that all grammars are equally similar to each other. In contrast, in a Principles & Parameters model, we can calculate the expected similarity based on estimates of how many parameters are revealed in a single sentence. Under the assumption that every sentence reveals m parameters, that all parameters are Boolean and that all parameters are revealed with equal probability: $a \approx \left(\frac{1}{2}\right)^m$. $a \approx \left(\frac{1}{2}\right)^m$. a is then an expected value rather than a constant, and equation (2) needs to be adapted. For simplicity, we will here follow the assumption of Nowak et al.

[6]Note that there is an error in this equation in Nowak *et al.* (2001) that is corrected in Komarova *et al.* (2001)