The MIT Press

A Computational Model of Language Learnability and Language Change
Author(s): Robin Clark and Ian Roberts
Source: *Linguistic Inquiry*, Vol. 24, No. 2 (Spring, 1993), pp. 299-345
Published by: The MIT Press
Stable URL: http://www.jstor.org/stable/4178813
Accessed: 26-05-2016 14:19 UTC

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at
http://about.jstor.org/terms

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

The MIT Press is collaborating with JSTOR to digitize, preserve and extend access to *Linguistic Inquiry*

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Robin Clark
Ian Roberts

A Computational Model of Language Learnability and Language Change

1 Introduction

Darwin's (1859) theory of natural selection had an important influence on the Neogrammarians. Like Darwin, they believed that diachronic change was the result of selective pressures on organisms from the environment operating on random variation within a population (see Haldane 1990 for a classic exposition of natural selection as the motive force underlying evolution). Darwin proposed that natural selection was accounted for by the greater reproduction rates of fitter organisms; in the linguistic realm, Paul (1920) proposed that language change is driven by restructuring of the target grammar that may take place during language acquisition. If the input to language acquisition is taken to be the environment and if language acquisition is taken to be the linguistic correlate of biological reproduction, a clear parallelism between Darwin's view of natural selection and Paul's view of the selection of grammars emerges. Despite the appeal of this notion, no successful evolutionary theory of the relationship between language acquisition and language change has been developed in the 130 years since Darwin's *On the Origin of Species*. The purpose of this article is to relate natural selection, language acquisition, and language change in light of current computational models of learning.

The basic problem for the hypothesis that language change is driven by acquisition concerns the relationship between the adult input, which is generated by one grammar, and the learner's hypotheses, which may differ at certain points from the adult grammar. We have grown accustomed to thinking of acquisition as a relation between linguistic experience and a target grammar; the learner must converge to a single target grammar in order for learning to be considered successful (see Gold 1967, Osherson, Stob, and Weinstein 1986). Although this idealization has proven useful in the study of the logical problem of language acquisition, it renders opaque the relationship between language acquisition and language change. If each generation converges successfully to the adult

The first author received support from grant 11-25362.88 from the Fonds national suisse pour la recherche scientifique and from a grant from the Fondation Ernst et Lucie Schmidheiny. This article has greatly benefited from comments made by two anonymous reviewers for *Linguistic Inquiry*.

Linguistic Inquiry, Volume 24, Number 2, Spring 1993
299–345
© 1993 by The Massachusetts Institute of Technology

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

grammar, how can languages ever change? One would expect them to remain forever fixed since change entails that there must be at least one generation whose grammar differs from its parents' grammar; yet, by definition, this generation would have misconverged. We can easily state the problem in terms of parameter setting. Acquisition is a process of accurately fixing parametric values. That is, the learner sets parameter p_n to the value v_i in response to some property, c_i, of the input text; the usual idealization states that the learner has successfully converged to the value v_i for the parameter p_n if the target grammar has p_n set to v_i. Language change, on the other hand, presupposes that a population must converge on a value v_i for at least one parameter, p, where the adult grammar has $p(v_j)$ and $v_i \neq v_j$. Strictly speaking, the learner has failed to learn. More puzzling still, the property c_i of the input text that allowed adults to induce p_n (v_i) when they were learning the language should be present in the speech that they, in turn, address to children. How is it that, for one generation, property c_i causes learners to hypothesize $p_n(v_i)$ whereas in a succeeding generation it loses its causal force?

We will argue that the question of how parametric change can take place given reasonable constraints on learnability is fundamental both for understanding of language acquisition and for understanding of language change. Indeed, the logical problem of language change cannot be separated from the logical problem of language acquisition; one of the claims of this article is that the former problem is a subcase of the latter (see Lightfoot and Hornstein 1981) in that the answer reduces to the relation between property c_i, the structure of the learner, and p_i (the same point has been made by Lightfoot (1991)). We will formalize this problem in light of current thinking on language learnability; doing this elucidates both the processes that underlie diachronic change and those that drive learning. The result is of importance for an understanding both of language acquisition and of diachronic change.[1] A central problem for acquisition theory is that of characterizing how the learner formulates and retracts hypotheses in light of its linguistic environment. Equally, one of the central problems for language change concerns how a population of learners can converge on a grammar that is systematically different from the adult grammar in the sense defined above. In both cases, hypothesis formation and retraction by learners appear to be the crucial mechanisms.

We will adopt the *genetic algorithm* approach to learnability developed in Clark 1990, 1992.[2] This approach treats learning as a special case of natural selection. In what follows, we will show how to encode the learner's hypotheses about the target sequence of parameter settings as "bit strings"—that is, strings of *0*s and *1*s—that serve to enumerate not only hypotheses but also, by extension, grammars and parsing devices. These bit strings, then, can be treated like genetic material that specifies grammatical "pheno-

[1] The first person to formulate this problem in terms of generative syntax was Lightfoot (1979).

[2] Genetic algorithms were developed by Holland; see particularly Holland 1975. Goldberg 1989 provides a comprehensive overview of the technique; see also Booker, Goldberg, and Holland 1990. Clark 1990 develops a model of parameter setting in terms of genetic algorithms as an approach to demonstrating the learnability property. See also Clark 1992 for a comprehensive theoretical treatment.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

types'' that may be expressed by parsing devices. These parsing devices are then run against an input text, and their relative fitness is measured by a simple metric. Those hypotheses that are judged most fit are then combined via a special mating operation; in other words, we will literally allow hypotheses to mate and thereby produce ''offspring'' hypotheses that share genetic material (subsequences of bit strings) of both parents. Since the mating operation prefers the most fit hypotheses, this technique allows the learner to search the hypothesis space efficiently while optimizing the learner's computational resources.

The genetic algorithm technique presupposes that the input text expresses each parameter with sufficient frequency that the learner's hypotheses are placed under pressure to bear that parameter setting. Hypotheses that carry a parameter value corresponding to a parameter setting frequently expressed in the input text will be strongly selected for by the fitness metric. As a result, hypotheses containing ''favorable'' parameter settings will tend to reproduce more frequently, whereas the ''unfavorable'' setting will disappear from the population, where *favorable* simply means 'better able to parse the input'. If, on the other hand, a parameter is not expressed frequently in the input text, the learner will be under less pressure to set that parameter in accordance with the target setting. In this case, the fitness metric will not be decisive in driving the learner toward the target setting, so that either the correct setting or the incorrect setting can survive in the linguistic environment. The fitness metric, which we will describe in detail below, plays a crucial role in mediating between the learner and the input text. Implicit to this discussion is the notion that relative fitness determines convergence; the learner converges to the most fit hypothesis relative to the input text even if this grammar differs from the adult state for the values of some parameters.

We will propose that parametric change occurs when the target of acquisition contains parameter values that cannot be uniquely determined on the basis of the linguistic environment. This can occur when the evidence presented to the learner is formally compatible with a number of different, and conflicting, parameter settings. In these cases the learner must evaluate its hypotheses using criteria that are not purely a response to the external environment; in particular, the learner must consider factors like the Subset Condition (Berwick 1985) and elegance of derivations (the least effort strategy; Chomsky 1991). Thus, the consideration of language change from a learnability perspective gives us access to how learners evaluate the relative merit of their hypotheses. Our goal here will be to characterize, in a precise manner, the conditions under which a learner arrives at a grammar distinct from the target, thus fueling diachronic change. Moreover, this approach reduces the logical problem of language change to the logical problem of language acquisition by relating both to the question of how learners set parameters to particular values.

Intuitively, our argument will be that, because of various factors, the input data do not put pressure on the learner to set certain parameters to a definite value; several alternative grammars can adequately account for the input stream; the appropriate choice

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

of grammar is underdetermined by the linguistic environment, even given the learner's rich internal structure. Since external pressures do not force the learner to select a particular grammar, it will turn in on itself, abandoning external pressure, and rely on its own internal structure to select from the alternatives at hand. If this is correct, then diachronic change can provide crucial information on those factors that learners rely on to select hypotheses. Since the external environment is not decisive in these cases, diachronic change reflects pure learnability considerations. Thus, diachronic change reflects what is, in a sense, "pathological" learning, and so a careful study of its properties can reveal a great deal about how learning transpires in nonpathological cases (a similar idea is developed for phonological change by Kiparsky (1982)).

We will argue that parametric change can involve a variety of factors. Change in one component—for example, the phonology—can obscure syntactic parameter expression. The resulting text will not uniquely drive the learner toward the target. At this point the learner appeals to the fitness metric to select an appropriate parameter setting, and factors such as the Subset Condition or general economy of representations come into play rather than pure selective pressure from the input text. This type of change is exemplified by the introduction of subject clitics in 15th-century French. A second important factor is instability due to independent parametric changes within a component; change in one parameter setting can trigger a number of changes to other parameter settings. As we will show, parametric change in 16th-century French provides a case study on how parametric change can cascade through a system (see Roberts 1992b). During this period, French ceased to be both a null subject language and a verb-second (V2) language. We will show that, because of innovations in the 15th century, the system became unstable, and deep parametric change was forced on the learner via the fitness metric.

Fundamental to this analysis is the formalization of the notion of *stability* relative to a particular parameter setting: A parameter setting is stable to the degree that its expression in the input data is unambiguous. Following Clark (1990, 1992), we will say that a parameter value, $p(v_j)$, is expressed by an input sentence, s_i, just in case a grammar must have p set to value v_j in order to assign a well-formed representation to s_i (see section 2.4). We should note that this does not mean that the parameter is set by raw data; rather, parameter expression defines a class of representations that are compatible with the current input sentence and the parameter values that those representations entail. An unstable parameter setting, then, is one whose expression is ambiguous. We will show that, through a variety of independent changes, 16th-century French became highly unstable, resulting in the loss of null subjects and V2 phenomena.

The article is organized as follows. In section 2 we discuss the formal and conceptual underpinnings of the learning theory. In section 3 we apply the learning theory to a particular case of change. Finally, in section 4 we discuss some of the consequences of the current approach for the theory of learning and change.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

2 Genetic Algorithms and Language Learnability

The basic problem faced by a language learner is to discover a target grammar based on a plausible input text.[3] A principles-and-parameters (see Chomsky 1981) approach to grammar provides a powerful way of limiting the problem of discovering the appropriate target grammar given the impoverished nature of the input data. Parameters can be viewed as finite vectors along which natural languages may vary; the learner is faced with the problem of searching a finite space of possible grammars rather than the more difficult problem of inducing a set of rules that lies at an undetermined point in an infinite hypothesis space.

Learning theory must provide an account of how the learner's search through the set of possible combinations of parameter values takes place, and of how certain values are chosen over others. We believe, with Lightfoot (1979), that such an account should give a solution to the logical problem of language change. In this section we will describe in detail our account of how the learner searches through the available parameters and fixes their values. The approach is based on the notion of a genetic algorithm (Holland 1975, Goldberg 1989, Clark 1990, 1992). Genetic algorithms model the basic process of natural selection in the biological world: how certain patterns of genetic material are more adapted to their environment (i.e., fitter) than others, and hence tend to reproduce at the expense of the others. Our account of language learning is analogous: the input text is the analogue of the environment, and so "fitness" means consistency with this; parameter settings correspond to the genetic material of the biological world (and so a whole grammar would be a genome). Successful combinations of parameter settings "reproduce" (i.e., contribute to the formation of new hypotheses about the target grammar) at the expense of others. In this way, the learning mechanism gradually eliminates "unfit" hypotheses (those that are not consistent with the input text) and arrives at a single fittest grammar. Since nothing in the approach requires this grammar to be consistent with the one that underlies the input text, learners may arrive at final-state systems that differ from those of their parents; this, in essence, is our solution to the logical problem of language change.

2.1 The Nature of the Learning Problem

It is possible to see the learner as a relation between input data and a sequence of parameter values (see Clark 1990, 1992). More precisely, we can view the learner as a function from input texts to parameter values, as in (1).

[3] We will assume, with many researchers in developmental psycholinguistics, that an input text consists of short, simple, grammatical sentences. Little in the present discussion hinges on the precise nature of the text, so long as the basic constructions of the language are adequately exemplified. For further discussion (and debate) on the nature of the input evidence, see Wexler and Culicover 1980, Lightfoot 1989, and the discussion of the latter work. For a formal characterization of the input evidence and its relation to learning, see Osherson, Stob, and Weinstein 1986.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

(1) $\varphi(\sigma_i) = \langle x_1, x_2, \ldots, x_n \rangle$

Here the learner is the function φ that applies to an arbitrarily selected text, σ_i, and gives a sequence of n parameter values, $\langle x_1, x_2, \ldots, x_n \rangle$. Given a sequence of parameter values, we can imagine that a special compiling function, ϕ_n, maps the sequences of parameter values onto a grammar, G_i, for the input text σ_i. We can further define a function, γ, which, given a grammar G_i, returns a parsing device P_m for the grammar G_i. Thus, we can view learning as a relation between inputs and parsing devices. This is important, since the notion of fitness with respect to input texts is most naturally defined in terms of the number of failed or successful parses of those texts. We will discuss how this is done below. Putting the above together, the learning situation is as described in (2).

(2) $\gamma[\phi_n(\varphi(\sigma_i))] = P_m$

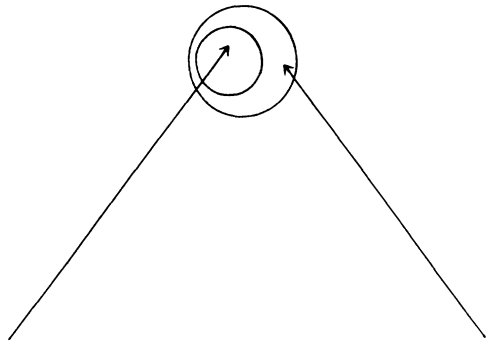
In considering the learning problem, it is important to recall that the learner is computationally bounded. In other words, the learner has finite resources in terms of time and memory. It cannot take indefinite periods of time before converging to the target grammar, nor does it have a perfect memory for past sequences in the input text or past (unsuccessful) hypotheses. Furthermore, the learner is given little information about the proper analysis to be accorded to the input data. It has only limited information about the proper structural analysis for any given datum, and little to no access to input that is ill formed with respect to the target.

The claim that the hypothesis space, under a principles-and-parameters approach, is finite is not, in itself, sufficient to guarantee that the learner can converge in a reasonable amount of time. Finite problems can be sufficiently large that their solution might take an impractical amount of time to compute. Suppose, for example, that the hypothesis space is determined by 30 binary parameters. In this case there are 2^{30}, or 1,073,741,824, possible grammars. If the learner could test each of these grammars at the rate of one per second, it might in the worst case take the learner over 34 years to converge on the target. Clearly, the learner must be capable of searching the hypothesis space in a more efficient manner.

Beyond efficiency considerations, it is clear that the learner cannot use a brute-force search technique to converge on the target since certain parameters may fall into subset relations; allowing *0* to stand for the negative value of a parameter and *1* to stand for the positive value, we can indicate as in (3) that the language that results when a certain parameter, p_x, is set to 0 is a proper subset of the language that results when p_x is set to 1. All the sentences that are grammatical in the subset language will also be grammatical in the superset language. If the learner guesses the superset language, then no further evidence will contradict its hypothesis. Thus, the learner will never have grounds to retract this (incorrect) hypothesis. Thus, the learner must guess the minimal language compatible with the input sequence σ_i. Given that the learner has no reliable access to negative evidence, it appears that the learner must guess the smallest possible

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

(3) $L[p_1,\ldots, p_{x-1}, p_x(0), p_{x+1},\ldots, p_z] \subset L[p_1,\ldots, p_{x-1}, p_x(1), p_{x+1},\ldots, p_z]$



language compatible with the input at each step of the learning procedure. This is, in essence, the Subset Condition proposed by Berwick (1985), which is intended to circumvent the sort of trap posed by subset parameters.

A further possibility arises if we consider that sets of parameters might interact in such a way as to generate superset languages. That is, when considered individually, the parameters in question may not necessarily generate superset languages, but when they act in a group, they do generate a superset language. This is the *shifting* relation observed by Clark (1990):[4]

(4) *Shifting*

Two parameters, x_i and x_j, cause a *shift* at values $x_i(1)$ and $x_j(1)$ just in case:

a. $L[\phi_n(x_1,\ldots, x_i(1),\ldots, x_j(0),\ldots, x_n)] \not\subset L[\phi_n(x_1,\ldots, x_i(0),\ldots, x_j(1),\ldots, x_n)]$

b. $L[\phi_n(x_1,\ldots, x_i(0),\ldots, x_j(1),\ldots, x_n)] \not\subset L[\phi_n(x_1,\ldots, x_i(1),\ldots, x_j(0),\ldots, x_n)]$

c. $L[\phi_n(x_1,\ldots, x_i(1),\ldots, x_j(0),\ldots, x_n)] \subset L[\phi_n(x_1,\ldots, x_i(1),\ldots, x_j(1),\ldots, x_n)]$

d. $L[\phi_n(x_1,\ldots, x_i(0),\ldots, x_j(1),\ldots, x_n)] \subset L[\phi_n(x_1,\ldots, x_i(1),\ldots, x_j(1),\ldots, x_n)]$

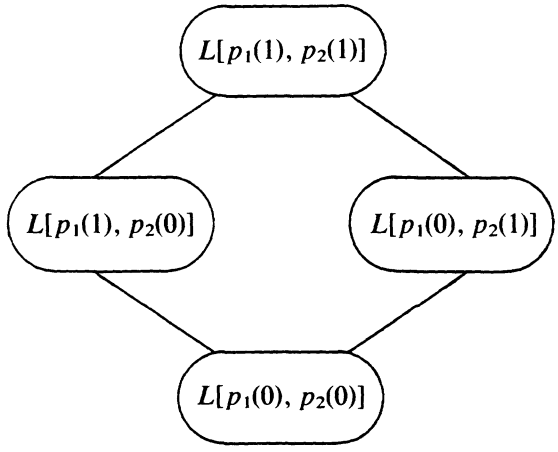
In other words, a shift occurs given two parameters that generate superset languages when they are both set to some particular value. Notice, crucially, that if the language generated by setting x_i to 0 is a subset of the language generated by setting x_i to 1, this relationship is preserved in the shifted language. In brief, a learner could obey the Subset Condition on the microscopic level (with respect to a single parameter) while violating it on the macroscopic level (due to shifting interactions between parameters). In order for the learner to avoid these higher-level violations of the Subset Condition, it would have to calculate interactions between parameter settings. But this would become increasingly difficult as the number of parameters that could "conspire" to generate a shifted language increased; given n parameters, the learner may have to consider $n!$ possible interactions.

[4] As we will show, shifting is more than a logical possibility and serves to force parametric change over time.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

The graph in (5) illustrates a case of shifting that involves superset parameters. In this example we have two parameters p_1 and p_2 that interact to generate a shifted language, $L[p_1(1),p_2(1)]$. In (5) dominance indicates the subset/superset relation.

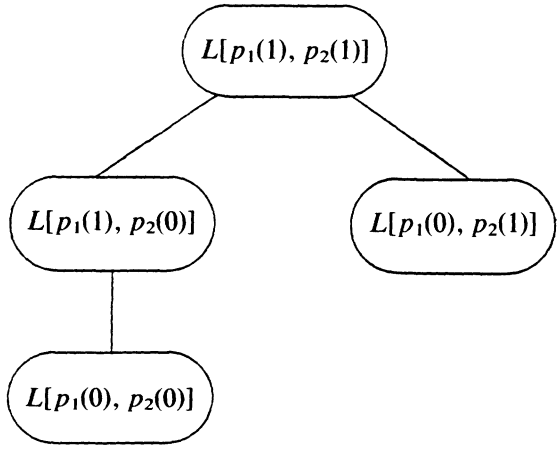
(5)



In this case both p_1 and p_2 are superset parameters; any language with p_1 set to 0 is a subset of a language with p_1 set to 1, and any language with p_2 set to 0 is a subset of a language with p_2 set to 1. Note that $L[p_1(1),p_2(0)]$ and $L[p_1(0),p_2(1)]$ are not in the superset relation with each other. The language $L[p_1(1),p_2(1)]$, however, properly contains the other three possible options. As we will show, the learner will be reluctant to posit the language $L[p_1(1),p_2(1)]$ and will only do so if faced with a significant amount of empirical prodding in the form of failed parses.

A more difficult case is illustrated in (6).

(6)



In this case only one of the parameters, p_1, is a superset parameter. One might imagine that p_1 regulates the option of having left-dislocation of a constituent. The parameter p_2

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

does not generate languages in the superset relation. For example, one might take p_2 to be a parameter that regulates V2 phenomena in matrix clauses. Suppose that p_1 and p_2 interact in such a way that, when both are set to 1, the language allows left-dislocation of a constituent over the V2 structure of the root clause; the resulting language has all of the normal V2 orders plus clauses with an additional constituent left-dislocated before the normal V2 order. Such a language would be a shifted language.

Take the case where the target language is V2 without left-dislocation. Suppose that the learner, during an early phase of the learning cycle, erroneously sets p_1 to 1, allowing left-dislocation of an NP (or DP) in response to the presence of nonsubject NPs/DPs in clause-initial position. This hypothesis, however, is inadequate to account for all the root V2 orders that the learner encounters—for example, those with initial adverbials and also possibly those with initial NPs/DPs without a resumptive pronoun. In response, the learner sets p_2 to 1, allowing for the possibility of V2, but does not reset p_1 to 0. In this case the learner has now entered a shifted language; because of the interaction between p_1 and p_2, all the target orders will be consistent with the learner's hypothesis, which, nevertheless, overgenerates. We will show that such a hypothesis will be selected against in such a way that the learner can retract its overgeneral hypothesis without access to direct negative evidence. Such a shifted language, although a possibility empirically, will tend to be unstable diachronically, with one of the two superset possibilities, V2 or left-dislocation, being quickly lost. Notice that a learner will have two analyses available for "V3" structures (structures with two constituents before the tensed verb); either such a structure involves left-dislocation with a standard V2 structure, as in (7a), or it involves simple left-dislocation, as in (7b).

(7) a. $[_{CP}$ DP $[_{CP}$ DP $[_{C'}$ V $[_{IP}$ . . .]]]]
b. $[_{CP}$ DP $[_{IP}$ DP V . . .]]

We will argue that (7b) involves a simpler syntactic analysis, with shorter chains, than (7a). Thus, the learner will tend to prefer the hypothesis that allows (7b) over one that requires the analysis in (7a). We will discuss further how this factor of "elegance" influences parameter setting below. In our discussion of the data from French in section 3, we will present some cases of this type.

2.2 Genetic Algorithms

Clark (1990, 1992) proposes that *genetic algorithms* provide a computational model of learning for a principles-and-parameters theory that circumvents the problems discussed in section 2.1 while accounting for the relationship between input evidence and parameter setting. Genetic algorithms mimic natural selection by representing hypotheses about a problem in a way that is similar to the way in which genetic material is represented. Hypotheses are then tested against the problem space, with the most fit hypotheses contributing to the formation of new hypotheses via reproduction (the combination of

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

preexisting hypotheses to form new hypotheses in a way that is similar to the biological recombination of DNA present in mating). By "breeding" the most fit hypotheses, testing them against the problem space, and pruning the least fit, a genetic algorithm can efficiently search large spaces and find optimal solutions.[5] More precisely, a genetic algorithm defines a number of automatic mechanisms for combining hypotheses that are, in some sense to be defined below, "fit." These mechanisms, which simulate breeding or reproduction, produce new hypotheses that are likely to replicate the advantageous properties of existing hypotheses while eliminating those properties that are ill adapted to the environment (in our case, the sequence of input sentences that the learner encounters). By repeating this process over successive "generations" of hypotheses, the learner is able to approximate the target sequence of parameter settings.[6]

A genetic algorithm consists of the following components:

- A *representation* of hypotheses in terms of *strings*, similar in structure to genetic material. In our case we will encode sequences of parameter values as strings of binary numbers.
- A set of *reproduction* operators that combine or alter existing "parent" hypotheses in order to produce new "offspring" hypotheses. Reproduction will be based on the performance of the hypotheses relative to the input stream; those that perform best will reproduce most prolifically. Furthermore, since reproduction is based on existing hypotheses, the search of the hypothesis space is highly constrained and not random (see Holland 1975 and Goldberg 1989 for careful discussions of these points):
  1. A *crossover mechanism*. This mechanism combines two hypotheses and produces a new hypothesis by combining parts of each of the parents' genetic material.
  2. A *mutation operator*. This mechanism randomly alters an offspring's genotype to produce a new hypothesis close to, but not identical with, the parents' genetic endowment.
- A *measure of fitness* of hypotheses in terms of their performance in an environment. The fitness metric defines how well adapted hypotheses are to their en-

[5] Space prevents a comprehensive discussion of this class of algorithms; see Goldberg 1989 for a general introduction and Clark 1990, 1992, for an application to the learnability problem for natural languages.

[6] Genetic algorithms are part of a class of algorithms that approximate some desired optimum but are not absolutely guaranteed to return the optimum. Other such algorithms include the "simulated annealing" found in some applications using neural networks, as pointed out by an anonymous reviewer. The property of returning a result that is "probably approximately correct" (PAC) is important for our purposes since such approximations are the fuel for language change (see the discussion of PAC learning in Natarajan 1991). We have selected genetic algorithms from the class of PAC algorithms because genetic algorithms incorporate a notion of relative fitness and for the formal clarity of the resulting model of parameter setting. We will argue below that this notion of fitness provides some insight into the nature of the learner and how properties of the learner govern diachronic change.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

vironment. In our case the fitness metric mainly measures success in parsing the input text (although it does contain other factors, as we will show).

Most crucial for our purposes are the representation of hypotheses in terms of strings and the notion of a fitness metric. Let us first turn to a more careful consideration of the representation of hypotheses. It is common to think of parameters as variables in Universal Grammar that range over a limited set of values. The bounding nodes for classical Subjacency (Chomsky 1977) provide a good example of such a parameter. Here subjacency is taken as an invariant property of natural languages whereas the bounding nodes may be contingently selected from a restricted set:

(8) *Subjacency*

No rule may involve X and Y in the configuration:
$\ldots X \ldots [_{\alpha} \ldots [_{\beta} \ldots Y \ldots \beta] \ldots \alpha] \ldots$
(order irrelevant)
where α and β are bounding nodes;
$\alpha, \beta \in \{NP, IP, CP\}$.

Parameters can equally be viewed as variant properties of natural languages; in other words, a parameter can be thought of as a descriptive statement that may be either true or false of a given grammatical system. From this perspective, we could rewrite the parameter for the bounding nodes as a series of three statements:

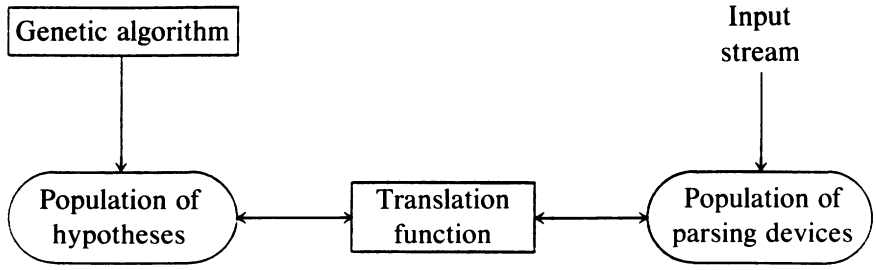
(9) a. IP is a bounding node for Subjacency.
b. CP is a bounding node for Subjacency.
c. NP is a bounding node for Subjacency.

The learner's task would be to scan the input data and attempt to assign truth-values, 1 for *true* and 0 for *false,* to each of the above propositions. The learner's hypotheses could then be taken as strings of *0*s and *1*s corresponding to the truth-value associated with each parameter. For example, the string *100* could correspond to the hypothesis that IP is a bounding node for Subjacency, but neither CP nor NP is. Thus, it is relatively natural to represent parameter settings in terms of strings. Notice that this binary representation of sequences of parameter values serves both to encode grammars as binary numbers and to enumerate the set of possible natural languages (see Clark 1992).

Crucially, given the above method of encoding parameter sequences, we must be capable of recovering the grammars and parsing devices that these encodings represent. This is crucial because fitness will be measured in terms of the performance of parsers relative to a stream of input data; the actual algorithm, however, will operate on the string representation of the hypotheses. We must, then, have a translation function that relates our hypotheses (strings) to the parsers that they represent, as shown in (10).

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

(10)



In fact, we have already defined all the machinery needed to accomplish the above. We conceive of the learner, φ, as operating on strings of parameter settings; thus, φ is the set of reproduction operators in the genetic algorithm. The translation function in (10) then maps the learner's hypothesis strings onto parsing devices; in other words, the translation function is comparable to the functions ϕ_n, which maps sequences of parameter settings onto grammars, and γ, which maps grammars onto parsers. In a sense, the hypothesis strings represent genotypes for parsing devices whereas the translation function (ϕ_n and γ) maps genotypes onto phenotypes. Overlying all of this is the fitness metric, which guides the learner's application of the reproduction operators.

The crossover operator combines two hypothesis strings to create new hypotheses. For example, suppose that the two hypotheses in (11) have been selected for reproduction.

(11) a. 000111
b. 101000

Now suppose we "cut" both strings after the third position in the bit string:

(12) a. 000——111
b. 101——000

The first part of string (12a) is then recombined with the second part of string (12b), and the first part of string (12b) is recombined with the second part of string (12a):

(13) a. 000——000
b. 101——111

And thus two new "offspring" hypotheses that have inherited genetic material (hypotheses about settings of particular parameters) from each parent are created. It should be noted that fitness interacts in a crucial way with the crossover operation. Highly fit hypotheses are more likely to be selected to take part in crossover and therefore are more likely to pass the parameter settings that made them fit on to new generations of hypotheses.

The mutation operator similarly creates new hypotheses on the basis of existing ones. In essence, it must slightly alter a hypothesis string in order to create a new, but

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

"nearby," hypothesis. We can do this by flipping a randomly selected bit position in a hypothesis string by the following rules:

(14) a. $0 \rightarrow 1$
b. $1 \rightarrow 0$

Thus, selecting the second position of the following hypothesis for mutation would yield a "mutant" that is nearly identical to its parent structure:

(15) $000111 \rightarrow 010111$

The mutation operation can be viewed as a means of searching the immediate hypothesis space surrounding a parameter string. Thus, the learner can, in a sense, experiment with near-optimal hypotheses that approximate, but do not correspond to, the target.

In terms of an actual parsing framework, there would be a fixed central algorithm, corresponding to UG. Within this algorithm would be various flags, indicating points where code must be inserted for the parser to function. The *0*s and *1*s in the hypothesis strings could be interpreted as pointers to the parameterized code. Upon receiving a hypothesis string, the machine would look up the various pieces of code indicated by the *0*s and *1*s and systematically substitute the code it finds for the flags in the main algorithm. The result would be a special parsing device designed to analyze the language enumerated by the hypothesis string. Thus, a "self-constructing" parser would be the ensemble of the core algorithm, the parameterized code, and a learning device that would select the appropriate hypothesis string in response to the input text. We then have a straightforward model of the *translation function* required by the genetic algorithm to relate hypothesis strings to parsing devices. Recall that this translation function, itself, corresponded to the functions γ and ϕ_n in the formalization of the learning problem, above.

2.3 Fitness

Having shown how hypotheses can be represented in terms of strings and how these can be combined systematically to form new hypotheses, we still face the problem of defining the relative fitness of a hypothesis with respect to a linguistic environment. Ultimately, we want the learner to become better able to represent the input data. In other words, the learner should change its hypothesis on the basis of evidence from the external environment, and its new hypothesis must be better able to account for this evidence. In some sense, new hypotheses must be an improvement over the old hypotheses.

Clark (1990) provides a crude definition of improvement based on the ability to parse input sentences in terms of failed parses. We will modify his treatment by supposing that the crucial property of a failed parse is that it violates at least one principle of core

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

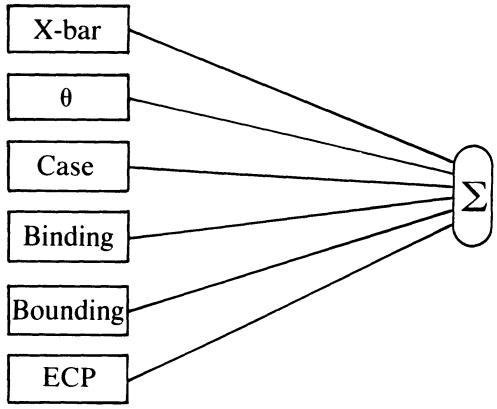
grammar.[7] In particular, we will suppose that a parser consists of a number of modules (Case, binding, X-bar theory, and so on) that operate in tandem to produce a full syntactic representation. When a principle in one of these modules is violated, when the current grammar cannot assign a well-formed representation to some input, the offended component will signal a violation. With this in mind, we adopt the following notion of *improvement* of one hypothesis with respect to another:

(16) A hypothesis A is an improvement over a hypothesis B if, given an input datum, s_i, A signals m violations of core grammatical principles while B signals n violations and $m < n$.

Intuitively, a parser that signals 3 violations on a parse is rather better than one that signals 4 violations, and a parser that signals 2 violations is superior to one that signals 3. Crucially, parsers need not perform perfectly in order for the performance to be distinguished.

We will suppose, then, that the various modules of the parser are connected to a summation function, $\sum$, as shown in (17).

(17)



Each module can signal a violation to the function $\sum$, which then sums up the number of violations and passes the number on to the learner. Notice that the learner has no access to which grammatical principles have been violated; it only receives a number representing the sum of the violations for each parse.

As noted above, the learner must be able to distinguish between hypotheses that generate a superset language and those that do not. If a superset hypothesis and a subset hypothesis can both account for an input datum, then, all things being equal, the learner should prefer the latter to the former. Thus, any fitness metric should be such that it generally rates a subset hypothesis more highly than a superset hypothesis just so long as the subset hypothesis is empirically adequate (does not fail to parse the input data).

[7] See Clark 1992 for an extensive formal discussion of fitness and reproduction and of their influence on convergence. Here, we will mainly be concerned with the intuitions that underlie the formalism.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Finally, we will assume that the learner can take into account the overall "elegance" of its hypotheses. That is, the learner will, all else being equal, prefer hypotheses that lead to more compact representations. Compactness, here, can be defined in terms of such factors as the number of nodes required to cover the input string, the length of the chains associated with arguments and operators, or both. For the moment we will assume that the measure of elegance is a raw node count from each parse.

With these factors in mind, we suggest the following as a fitness metric, defined over a population of parsing devices relative to an input sentence (see Clark 1990 for an earlier version of this metric). It should be noted that hypotheses are judged indirectly by means of the parsing devices that they determine, just as a genotype is judged through its expression as a phenotype. In particular, the learner has no information about *why* certain hypotheses perform better than others, only that certain hypotheses do, in fact, perform better. In assessing the performance of hypotheses, the fitness metric will consider a number of different factors. Above all, it will consider raw success or failure to parse; other factors, like subset relations and elegance of representation, are also taken into account, although their contribution is weighted so that they influence the learner slightly less than actual success or failure to parse.

Let the number of parsing devices be n. We then need a way to count up the number of violations incurred by a given parser P and, since we are defining *relative* fitness, to relate this to the number of violations signaled by all the parsing devices together. We indicate the total number of violations of all parsing devices by $\sum_{j=1}^{n} v_j$; this operation simply sums the number of violations in the entire population of parsing devices. We indicate the number of violations incurred by P as v_i. To relate v_i to the number of violations incurred by other parsers, we follow a standard statistical technique and subtract the number of violations incurred by P from the total and divide that figure by the total:

$$(18) \quad \frac{\sum_{j=1}^{n} v_j - v_i}{\sum_{j=1}^{n} v_j}$$

Thus, if the total number of violations is 1,000 and P_i produces 10 violations, the metric will give $\frac{1000-10}{1000} = 0.99$. Where P_j produces 100 violations, the metric gives $\frac{1000-100}{1000} = 0.9$. P_i is thus more highly valued than P_j.

For complete precision, we must prevent the parser in question from being compared with itself, so we exclude it from the population as follows:

$$(19) \quad \frac{\sum_{j=1}^{n} v_j - v_i}{(n-1)\sum_{j=1}^{n} v_j}$$

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

As noted earlier, we also want to evaluate whether a given hypothesis gives rise to a superset grammar. We can do this by proceeding in the same way as above: If we allow s_m to represent the number of superset settings in the hypothesis h_m, then $\sum_{j=1}^{n} s_j$ is the number of parameters set to superset values in the population.[8] We now introduce a "superset penalty," the constant $b < 1$, and multiply the count of superset settings by b. In this way, Subset Condition violations are scaled so that they will have less weight in the overall metric than a simple failure to parse a sentence. The product of b and the superset count for a single parsing device is evaluated relative to the population of parsers in the same way as above. Combining the superset factor with the parsing factor, then, produces the metric in (20).

$$(20)\quad \frac{(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j) - (v_i + bs_i)}{(n-1)(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j)}$$

Finally, we need to weigh in the relative elegance of parses as a factor. Again we proceed in the same fashion: $\sum_{j=1}^{n} e_j$ is the measure of the general elegance of the analyses in the entire population of parsers (which we continue to take to be a simple tally of the number of nodes) and e_i is the measure for parser P_i. Analogous to the superset penalty, we introduce the constant c, which is a scaling factor for the elegance of the representation. Here again, this means that elegance is a less important factor than failure to parse. If we include the elegance factor in the equation, we arrive at the fitness metric:

(21) *The fitness metric*

$$\frac{(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j) - (v_i + bs_i + ce_i)}{(n-1)(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j)}$$

We will leave the question of the exact values of the constants b and c open, assuming only that $1 > b, c > 0$ (preliminary calculations suggest that both of these constants are in fact very small, in the region of 0.00002; see Clark 1990). It is worth emphasizing that the fitness metric takes these factors into consideration, but that they are weighted so that they always count less than straightforward failure to parse. Notice, though, that they become crucial in distinguishing between successful parses. This will play an important role in our discussion of language change in section 3. Finally, the fitness metric in (21) blurs the reasons for success or failure of a hypothesis relative to a population; the learner has no way of knowing why a given hypothesis succeeds or fails.

It is perhaps useful to consider the contribution of each of the above factors, using

[8] For simplicity, we assume that the learner has access to a table that tells it which settings are superset settings; this is much simpler than forcing the learner to calculate whether or not a given parameter value generates a superset language. Note that shifting relations will not be included on the table. These will be selected against by the fitness metric in an indirect way.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

some hypothetical examples. Let us turn first to the way in which the fitness metric treats grammatical violations. For the population, this is the term $\sum_{j=1}^{n} v_j$ in the fitness metric; for the individual parsing device, it is the term v_i. Suppose we have the three parsing devices p_1, p_2, and p_3. Running these on an input sentence yields the following results:

(22) a. p_1 returns 1 violation, covering the input with 15 nodes.
b. p_2 returns 2 violations, covering the input with 15 nodes.
c. p_3 returns 3 violations, covering the input with 15 nodes.

Running the above results through the fitness metric gives the following results, with $b = 0.02$ and $c = 0.05$ (we ignore, here, the contribution of the subset factor by assuming that none of the hypotheses underlying the parser contain superset settings):

(23) a. p_1 receives a fitness rating of 0.393939.
b. p_2 receives a fitness rating of 0.333333.
c. p_3 receives a fitness rating of 0.272727.

Thus, parser p_1 is judged the most fit, p_2 the next most fit, and p_3 the least fit. Notice that the learner does not receive information about which grammatical principles are violated. It has no need of such information in order to distinguish between the hypotheses at hand. Instead, it need only observe the performance of its hypotheses in an external manner, without information about their inner workings. The learner will base its new hypotheses on those old ones that are relatively more fit, thus passing on the parameter settings that made those hypotheses fit to future generations. Those parameter settings that avoid grammatical violations relative to the input text will be preserved, and those that tend to generate violations will gradually disappear.

Let us turn, now, to the contribution of the superset penalty, the term $\sum_{j=1}^{n} s_j$ for the entire population and the term s_i for a single parsing device. Suppose that p_1 and p_2 both signal no violations of any grammatical principles and both cover the input in 20 nodes. Suppose further that p_2 contains a superset setting for one parameter and p_1 contains no superset settings. The fitness metric will then return the following results:

(24) a. p_1 receives a fitness rating of 0.50495.
b. p_2 receives a fitness rating of 0.49505.

Notice that the "smallest hypothesis," in this case the one underlying p_1, is judged more fit than the one that violates the Superset Condition. Thus, the fitness metric can distinguish both between hypotheses that are unequal in their parsing powers and between hypotheses that are equal in parsing power but differ with respect to the Subset Condition.

We turn, finally, to the contribution of the "elegance" factor; this is the term $\sum_{j=1}^{n} e_j$ for the entire population and e_i for individual parsing devices. Consider two hypotheses, p_1 and p_2, which both return no violations and contain no superset settings

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

but cover the input with trees of different elegance. Suppose that p_1 is able to cover the input with 17 nodes whereas p_2 covers the input with 18 nodes. The results of the fitness metric are then as follows:

(25) a. p_1 receives a fitness rating of 0.514286.
b. p_2 receives a fitness rating of 0.485714.

The first hypothesis is preferred by the fitness metric since it is able to span the input in a more elegant way than the second hypothesis.

In order to see the importance of this factor, consider the case where the target is SVO. Suppose that hypothesis h_1 treats the subject as being in the Spec of IP at S-Structure whereas hypothesis h_2 treats the subject as having moved to the Spec of CP, attracting the main verb with it. For a simple clause, h_1 and h_2 will return the following structures:

(26) a. h_1: $[_{IP}$ DP $[_{I'}$ I VP]]
b. h_2: $[_{CP}$ DP_i $[_{C'}$ V_j $[_{IP}$ t_i $[_{I'}$ t_j VP]]]]

By assumption, both h_1 and h_2 can account for the input stream. Notice, however, that h_2 involves systematically longer chains than h_1 since the former always involves movement of the subject to the Spec of CP, with subsequent attraction of the verb to C^0, whereas the latter does not. The representations returned by h_1 are simpler than those returned by h_2. Since the learner, via the fitness metric, can take into account the general elegance of representations, it can successfully distinguish between h_1 and h_2. Notice, however, that elegance is defined quite simply as a count of the nodes in the tree covering an input item plus the lengths of the chains in the representation.

The fitness metric can be considered to work as follows. The population of parsing devices specified by the learner's hypothesis strings is run against each input item. The term $\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j$ yields the total number of violations, the total number of superset settings, and the total elegance of representations of the entire population, with the various factors weighted appropriately by the constants b and c. Dividing this term by n, the size of the population, would give the average number of undesirable properties for the entire population. Next consider the term $v_i + bs_i + ce_i$. This yields the number of unhealthy properties each individual parsing device carries. As this term grows in relation to the population average, the relative fitness of the parsing device decreases. If this term decreases with respect to the population average, then the parsing device is judged relatively more fit.[9]

The opportunity to reproduce (that is, be selected for the crossover operation and

[9] The results discussed here receive a more formal discussion in Clark 1992, where proofs of certain theorems entailed by the fitness metric are given. For present purposes, the important point is that, relative to an input text, the fitness metric drives the learner toward a hypothesis that minimizes the number of violations and the number of superset settings and that generates the most elegant syntactic representations possible, given that grammatical violations are avoided.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

mutation) is a direct function of relative fitness. The simulation developed in Clark 1990 assumes that the fitness associated with a hypothesis corresponds transparently to its proportion of the general population. In an environment with random mating, then, those hypotheses with a high proportion in the population are more likely to meet and reproduce. The fitness ratings are used to simulate a weighted roulette wheel, the results of which undergo the crossover and mutation operations. In other words, successful hypotheses will receive a high fitness rating. The fitness rating corresponds to the probability that the hypothesis will get to reproduce. Thus, the fittest hypotheses will reproduce more frequently and pass on their parameter settings to new hypotheses. Cumulatively, then, the population will tend toward the optimal set of parameter settings for the target.

Crucially, the most fit hypotheses are the most likely to contribute to the formation of new hypotheses. These hypotheses have the greatest opportunity to pass on the parameter settings that made them fit to new hypotheses. Because weak hypotheses are pruned at random intervals, these are ultimately prevented from contributing their inferior parameter settings to the general pool. Thus, fit parameter settings tend to take over while unfit parameter settings are purged. By iterating the process of parsing, judging fitness, reproduction, and "death," the learner is able to incrementally approach the target grammar.

2.4 P-Encodings

Before we turn to the diachronic data, two other definitions are required. Consider a simple example like (27).

(27) John loves Mary.

Notice that certain parameters must be set in a particular way if the sentence is to be parsed. Both *John* and *Mary* must receive θ-roles and Case, the verb *love* must be capable of picking up its inflectional affix, and so on. Any parsing device that can successfully account for these features of the sentence in (27) will return a well-formed representation. Other parameters (e.g., bounding nodes and those that regulate conditions on anaphora) are irrelevant to the representation of this sentence. It will not matter what values for these parameters the parsing device presupposes. This suggests that any given input sentence *expresses* certain parameters and that a set of distinct parsing devices can account for (27):

(28) *Parameter expression*

A sentence σ expresses a parameter p_i just in case a grammar must have p_i set to some definite value in order to assign a well-formed representation to σ.

When a given datum expresses some parameter value, the learner will be under pressure to set that parameter to the value expressed by the datum. This is because the fitness

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

metric will prefer hypotheses with the correct setting to those without it. This provides a simple definition of the intuitive notion of *triggering datum:*

(29) *Trigger*

A sentence σ is a *trigger* for a parameter p_j if σ expresses p_j.

Given the above interpretation of the input data, we can imagine a method of encoding the data in string form. Suppose we have a function ψ that maps a sentence onto the set of sequences of parameter settings that are compatible with that sentence. For example, a given input sentence, s_m, can be accounted for by grammars with the second and third parameters set to 0 and the fifth parameter set to 1. Applying ψ to s_m would give the following set of parameter strings:

(30) $\psi(s_m) = \{00001, 10001, 00011, 10011\}$

Using "*" as a variable to range over 0 and 1, we could replace the above set of strings with a cover term:

(31) $\{00001, 10001, 00011, 10011\} = [\ *\ 0\ 0\ *\ 1\ ]$

We will refer to the sequence [* 0 0 * 1] as the *p-encoding* for s_m; the p-encoding of a sentence may be thought of as a "pure" representation of the parameters expressed by the sentence.[10] Notice that, in principle, one could replace the sentences in an input text with their p-encodings and, so, study the frequency of expression for various parameters and the overall structure of the text relative to parameter expression.

There is an important relationship between parameter expression and the fitness metric. Ultimately, the fitness associated with a hypothesis governs its probability of being selected for reproduction. The more fit a hypothesis is, the more likely it is to pass on those parameter settings that made it fit. Now consider parameter expression. When a parameter is expressed, those hypotheses that have the correct value for that parameter will be judged more fit than those that lack the proper value. If a parameter is expressed robustly by several different construction types (and, hence, has a higher probability of occurring in the input text), then those hypotheses bearing the correct value will have more opportunity to be selected for reproduction and the appropriate parameter setting will tend to dominate in the population. Furthermore, those hypotheses bearing the incorrect value will have a lower fitness rating and will tend to reproduce less so that the parameter values that made them unfit are washed from the population. Thus, parameter settings that are expressed robustly will tend to be set quickly and

[10] The notion of p-encoding defined here is essentially isomorphic to that of "schemata," which has been widely discussed in the genetic algorithm literature (see, in particular, Goldberg 1989 and the references cited there). There is one important difference, however; schemata are usually taken as ranging over empirical generalizations whereas p-encodings represent the ambiguities inherent in the input stream. The two are similar in that p-encodings represent the set of grammars that can, in principle, assign a well-formed representation to a given string.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

efficiently by the learner. Parameters that are not expressed robustly, however, will tend not to affect the fitness of a hypothesis in the same way. The learner will have correspondingly less stake in setting the parameter correctly and will not converge so readily to the parameter value.

Now consider the case where parameters are ambiguously expressed. In our terms, there might be several contradictory p-encodings associated with a class of data, for example. Here the learner has several possible solutions available that can account for the input without generating grammatical violations. In this case frequency of parameter expression will not aid the learner in distinguishing between its hypotheses. Instead, the learner will have to rely on the structure of the hypotheses themselves, and not their empirical coverage, in order to select a winning hypothesis. These internal factors are the overall elegance of representations and the number of superset settings in each hypothesis, both of which are factors in the fitness metric. We argue, here, that it is this sort of case that provides the fuel for core diachronic change in a parameter setting. In the next section we will turn to a case where learners were faced with just such an ambiguity.

3 A Case Study in Diachronic Change

We believe that applying genetic algorithms in the form outlined above to the acquisition of natural languages is not only possible but desirable. It is desirable in part because it avoids the problems discussed in the previous section: it allows convergence over a finite but large hypothesis space, and it can be defined such that superset traps can be avoided (which the version of the fitness metric given in (21) in fact does). Our main contention here, however, is that the genetic algorithm approach provides a solution to the logical problem of language change. We will now turn to an application of the genetic algorithm approach to learning and show how it can model diachronic change as well.

3.1 The History of French

Roberts (1992b) analyzes three major syntactic changes in the history of French as reflexes of a single underlying parametric change. The three changes are as follows (here and elsewhere, unless otherwise noted, OF and MidF data are from Roberts 1992b):

(32) *Loss of "simple inversion" in interrogatives*

a. *A Jean pris le livre? ModF
has Jean taken the book

b. Comment fu ceste lettre faitte? OF
how was this letter made

(33) *Loss of null subjects*

a. *Ainsi s'amusaient bien cette nuit. ModF
thus (they) had fun that night

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

b. Si firent pro grant joie la nuit. OF
thus (they) made great joy the night

(34) *Loss of V2*

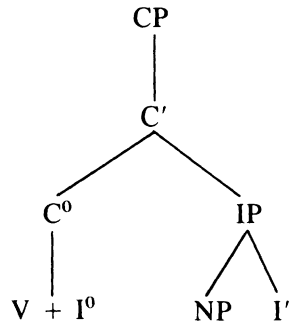
a. *Puis entendirent-ils un coup de tonnerre. ModF
then they heard a clap of thunder

b. Lors oïrent ils venir un escoiz de tonoire. OF
then they heard come a clap of thunder

As Roberts shows in some detail, each of these constructions was lost in the early 16th century. Roberts argues that these changes reflect an underlying change in the value of the parameter determining nominative Case assignment proposed by Koopman and Sportiche (1991): nominative Case may be assigned (by I) either under government, or under agreement, or under both. The central idea of this account of the history of French is that OF allowed nominative Case assignment under government, whereas ModF does not.

More precisely, all of the OF constructions depend on the possibility of the inflected verb, V + I, assigning nominative Case to the subject in Spec of I′ from C, as shown in (35).

(35)



This situation was allowed in the grammar of OF (and is still allowed in, for example, the contemporary Germanic languages). In a grammar where this configuration of Case assignment is not allowed, no lexical NP can survive in subject position in inversion contexts; this is the situation in ModF, where (32) thus violates the Case Filter. Following Kayne (1983) and Rizzi and Roberts (1989), we assume that clitics can survive in subject position in this context since they are able to pass the Case Filter in other ways (also see Baker 1988, Everett 1986).

Adopting Rizzi's (1986a) proposal that the necessary condition on formal licensing of pro is that it occupy a Case-marked position, Roberts accounts for the change illustrated in (33) by extending the nominative Case parameter to the pro module; it is well known that OF null subjects were licensed only in contexts of inversion (see Thurneysen 1892, Price 1971, Einhorn 1974, Foulet 1982, Vanelli, Renzi, and Benincà 1986, Adams 1987a,b), and so a natural interpretation of this is that null subjects could only be licensed

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

where nominative Case was assigned under government, that is, in the configuration (35). This in turn accounts for why null subjects were lost when nominative Case assignment under government was lost.[11] Regarding (34), V2 also depends on the capacity of I to assign nominative Case to the subject under government after being raised to C with the verb. Note that nominative-Case-under-government is a necessary, not a sufficient, condition for V2. Hence, a system without this possibility cannot have V2. However, a system with this possibility need not have V2 (Modern English is probably such a system). In fact, as we will illustrate, obligatory V2 was already eroding in MidF—this was a crucial factor in the instability that led to the change in the nominative Case assignment parameter.

The principal trigger for the change in the possibilities of nominative Case assignment was the introduction of new word orders that did not strictly conform to V2, notably XSVO (where "X" could be a topic or an adverb). This innovation was probably caused by the development of a series of subject clitics in MidF (see below). The cumulative effect of the new word orders was to destabilize the system in such a way that setting the parameter for nominative Case assignment under government positively became impossible by about 1500, and learners converged on a grammar lacking this property. The result was the elimination of the structures in (32)–(34) in 16th-century texts—a major change in the grammar of French. Note that we do not consider the null subject parameter or the V2 parameter as in any sense subsumed by the nominative Case parameter; however, the particular circumstances of French at the time the change took place were such that the loss of nominative Case assignment under government entailed the loss of null subjects and the elimination of V2. Our proposal is that the initial weakening of V2 combined with the development of a series of subject clitics created a system that ultimately eliminated V2, and in doing so eliminated null subjects and simple inversion. In particular, the weakening of V2 had the effect that hypotheses that allowed an input datum to be analyzed as a V2 structure became more costly relative to the fitness metric; thus, the learner was under pressure from fitness to eliminate the V2 hypothesis.

Although we concentrate exclusively on French here, there is also evidence (see in particular Vanelli, Renzi, and Benincà 1986) that many of the Gallo-Italian dialects of northern Italy have undergone the same parametric change, since in their recorded history, simple inversion, V2, and, arguably, null subjects have been lost (although the contemporary dialects in fact have a kind of "disguised" null subject system that prob-

[11] Vance (1989) in fact shows that 15th-century MidF null subjects could be licensed under agreement as well as government. Nevertheless, both null subjects licensed under government and null subjects licensed under agreement are lost with simple inversion in the 16th century. Roberts (1992b:sec. 2.4.3) proposes that the loss of null subjects where they were licensed under government also entailed their loss throughout the system on the basis of the idea that, for null subjects to be licensed only under agreement, a very rich "pronominal" morphology is required. This type of morphology is found in Italian and Spanish but not in MidF or ModF. Hence, the "poverty" of French agreement, combined with the change in the nominative Case parameter, led to the loss of null subjects everywhere. We will discuss Vance's data further below.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

ably represents an independent diachronic innovation; see Poletto 1990, Renzi and Vanelli 1983, Rizzi 1986b). Moreover, Renzi (1983) argues that Modern Standard Italian has undergone the same changes as French regarding inversion while retaining null subjects.

In all, five parameters are relevant to our account of the historical development of French. These are given in (36).

(36) a. Nominative Case is assigned (by I) under agreement.
{1,0}
b. Nominative Case is assigned (by I) under government.
{1,0}
c. Clitic nominative pronouns.
{1,0}
d. Null subjects licensed canonically (Case-dependently).
{1,0}
e. Obligatory V-movement to C in matrix declaratives (V2).
{1,0}

Note that we split Koopman and Sportiche's parameter for nominative Case assignment into two separate parameters in order to preserve a basically binary vocabulary for parameters (see the discussion of Subjacency and bounding nodes in section 2). We take it that (36a) has been constant at 1 throughout the entire period (but see section 4). As just mentioned, (36b,d,e) changed together in the 16th century. The shift in (36d) and (36e) was forced by the change in the value of (36b). This is presumably quite a standard situation with parametric change: changes in parameter values interact. Moreover, parameter values can be affected by nonsyntactic factors, notably phonological changes. This is the case with (36c); properties connected to the stress system may cause a class of pronouns to cliticize and thereby trigger a shift in the value of this parameter.

We now review the relevant data from the different periods of French and show how the data trigger parameter settings. To illustrate the general technique, we will first consider Modern French. Then we will consider Old French and finally the period of greatest "structural instability" (and, hence, of greatest interest), Middle French.

3.2 Learning Modern French

Before we consider the earlier periods of French, let us first look at the situation in the contemporary language. What are the parameter values for ModF? It is clear that nominative Case is assigned by I to its Spec position; hence, the first position in the string must be set to 1. On the other hand, Rizzi and Roberts (1989) argue that ModF does not allow nominative Case assignment under government; this is what leads to the restriction to clitics in contexts where the inflected verb, a complex head that contains I, moves to C (e.g., in interrogatives or conditionals; cf. also (32)):

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

(37) a. Ont-ils/*les enfants vu ce film?
have they/the children seen this film
b. Aurait-elle/*Marie fait cela . . .
had she/Marie done this

Once moved to C, I must Case-mark the subject position under government; the ungrammaticality of (37a–b) with a nonclitic subject shows this to be impossible. In terms of this analysis, I does not assign nominative Case under government in ModF, and we therefore set the second position in the string to 0.[12] It is well known that ModF has a class of clitic nominative pronouns (see Kayne 1975, Rizzi 1986b); the contrasts in (37) in fact illustrate that these elements interact with Case theory in a manner distinct from nonclitics. Rizzi and Roberts (1989) propose that clitics can satisfy Case theory by incorporating with the verb in C (see Baker 1988, Everett 1986, 1989). Thus, we take it that in ModF parameter (36c) is set to 1. Both parameters (36d) and (36e) are set to 0: ModF is neither a null subject nor a V2 language, as comparisons with contemporary Italian and German show, respectively.[13]

These remarks on the grammar of ModF (which we of course cannot fully substantiate here; see the references cited for further arguments) lead to the following conclusion regarding the representation of the parameters in (36) as a string of binary units:

(38) The "target string" for ModF is 10100.
Nominative Case is assigned under agreement and subject clitics are allowed.

Let us now consider how the parameter values for ModF are expressed in the input text. Recall that a sentence S expresses a parameter P_i iff a grammar must have P_i set to a particular value in order to assign a well-formed representation to S. In such a situation, S is a trigger for P_i. The following examples illustrate a significant part of the trigger for the parameter values of ModF:

(39) a. Jean aime Marie.
Jean loves Marie
b. Hier Jean est parti.
yesterday Jean left

[12] In our presentation, we abstract away from the "split Infl" hypothesis of Pollock (1989), restricting ourselves to projections of I. To fully account for the facts of ModF inversion, however, it is necessary to split I into at least Agr and T (and their projections). In terms of the "Agr over T" system proposed by Belletti (1990), our nominative Case parameter refers to Agr. To account for stylistic inversion, we probably need to say that T can assign nominative Case to a postverbal subject under government (see Rizzi 1990). (Also see footnote 13.)

[13] Literary ModF allows strings that appear to be V2—for example, *Dans cette maison vécut Racine* 'In this house lived Racine'. However, such examples should be treated as instances of stylistic inversion. Stylistic inversion differs from V2 and subject-clitic inversion in that the subject appears in a position following the entire verbal complex in a compound tense and is not sensitive to the root/embedded distinction, unlike true V2. (See Kayne and Pollock 1978 and Pollock 1986.) In fact, Pollock (1986) suggests that stylistic inversion may involve a nonreferential null subject in Spec of I′. If so, ModF allows at least some highly restricted occurrences of null subjects and (36d) should therefore be reformulated to refer to referential null subjects.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

c. Où est-il allé?
where did he go

Recall that the conditions of acquisition are such that starred examples like the (a) cases of (32)–(34), which can be used by the linguist to justify a particular analysis, are not available. Moreover, many sentences are amenable to differing structural analyses that can affect their status as triggers. This last point is crucial to understanding how change takes place, as we will show.

Consider first (39a), a simple declarative sentence with canonical SVO order. In terms of the usual analysis of ModF, the relevant parts of this sentence are as follows:

(40) $[_{IP}$ Jean $[_{I'}$ aime . . .]]

Parsed in this way, (39a) triggers nominative Case assignment under agreement and indicates that V-movement to C is not required in matrix declaratives—in other words, that ModF is not V2. Thus, (40) is associated with the following p-encoding:

(41) [1 * * * 0]
Nominative Case is assigned under agreement, and V-movement to C is *not* allowed in matrix declaratives.

(41) indicates that (39a) tells the learner that nominative Case is assigned under agreement, and that French is not V2; but it does not say anything about whether nominative Case is assigned under government, whether subject clitics are allowed, or whether null subjects are allowed.

However, strings exactly equivalent to (39a) are grammatical in the Germanic V2 languages. In these languages the relevant parts of the structure are as follows:

(42) $[_{CP}$ Jean $[_{C'}$ aime $[_{IP}$ t $[_{I'}$ t. . .]

Call this the "V2 parse" of an SVO sentence. Here I assigns nominative Case to the Spec of I′ (i.e., the position occupied by the trace of the subject) under government; we will refine this analysis in section 4. Hence, the p-encoding for this parse is as follows:

(43) [* 1 * * 1]
The parser must have nominative Case assignment under government, and V-movement to C is obligatory in matrix declaratives.

As (43) shows, (39a) remains silent regarding subject clitics and null subjects.

To sum up, SVO declaratives in ModF have the following p-encodings:

(44) SVO declaratives p-encode
a. [1 * * * 0]
Nominative Case is assigned under agreement, and V remains in I in matrix declaratives.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

b. [* 1 * * 1]
Nominative Case is assigned under government, and V moves to C in matrix declaratives.

SVO sentences are thus associated with different p-encodings depending on the parse they are given. We can characterize this situation in terms of the following notion of p-ambiguity:

(45) A sentence S is *p-ambiguous* with respect to some parameter P_i just in case S has the set of well-formed representations ($R_1 \ldots R_n$) and P_i must be set to some definite value v_1 in order to assign R_i to S (i.e., R_i triggers a $P_i(v_1)$), whereas P_i does not need to be set to v_1 in order to assign $R_j \neq R_i$ to S.

ModF SVO sentences are p-ambiguous, as (44) shows. As will be discussed in section 3.3, however, the representation where V is in C is disfavored since it involves a more complex structure than the representation where V is in I.

Now consider (39b). In V2 languages generally, orders of this type are impossible (see Schwartz and Vikner 1989). This can be interpreted in terms of a ban on adjunction to CP. Supposing that this is so, this example must be parsed with the adverb attached to IP, V in I, and the subject in Spec of I′. In other words, the relevant parts of the structure are like the parse of (39a) given in (40), and the triggering properties of the sentence are the same. More generally, we can conclude the following:

(46) XSV p-encodes [1 * * * 0]
Nominative Case is assigned under agreement, and movement of V to C is not allowed in matrix declaratives.

Now consider the interrogative in (39c). (39c) provides evidence for the subject clitic (this evidence is probably morphological, given the existence of a separate paradigm of clitic pronouns) and therefore, given that clitic pronouns do not obey the Case Filter in the same way as nonclitic NPs, provides no evidence for either Case assignment parameter. We take it that interrogative sentences by their nature provide no evidence regarding V2 in declaratives, and the null subject parameter is not determined either. We therefore arrive at the p-encoding in (47) (where *s* indicates a subject clitic in the schematic word order).

(47) *wh*VsO p-encodes [* * 1 * *]
Subject clitics are possible.

If the subject clitic is not recognized as such, but treated as a full NP, this sentence would p-encode (48).

(48) [* 1 * * *]
Nominative Case is assigned under government.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

We assume, however, that phonological and morphological evidence disfavors this possibility.

When we put the p-encodings in (43)–(47) together (and disregard the one in (48)), the following picture emerges:

(49) a. SVO, XSV: [1 * * * 0]
Nominative Case is assigned under agreement; no V2 is possible in declaratives.

b. SVO: [* 1 * * 1]
Nominative Case is assigned under government; V2 is possible in declaratives.

c. *wh*VsO: [* * 1 * *]
Subject clitics are possible.

The two parameters that are not positively set are nominative Case assignment under government and null subjects. These are both set to 0 in the optimal case. Let us consider why.

The two parameters determining nominative Case assignment by I, (36a–b), are in a shifting relation. Although neither parameter directly determines a superset relation (a grammar that allows nominative Case assignment under agreement generates a language that intersects with one that does not; similarly for nominative Case assignment under government), if both parameters are set to 1, they together generate a language that is the superset of the one that results from setting either parameter to 0. This is a classic case of shifting (of the type seen in section 2). Now, as we have shown, (36a) is unambiguously expressed in the input for ModF and thus is set to 1. In order to avoid shifting, a positive value for (36b) is strongly disfavored. Since there is no unambiguous evidence for nominative Case assignment under government, the pressure against shifting is decisive and the parameter is set to 0 in the optimal grammar.

It should be noted that the only evidence for nominative Case assignment under government consists of sentences with the order SVO, with a V2 parse, which can also be analyzed more compactly under the assumption that nominative Case is assigned under agreement. In particular, the V2 parse for the SVO order must involve movement of the subject to the Spec of CP and thus entails a longer chain than would occur under the competing analysis. Thus, the non-V2 parse is again favored and the V2 parse is disfavored by the fitness metric. This provides the learner with further evidence in favor of setting the V2 parameter to 0, as well as disfavoring nominative Case assignment under government.

For the null subject parameter, we could follow Berwick's (1985) reasoning and invoke the Subset Condition. If null subject languages are a superset of non–null subject languages, the lack of a trigger for a positive value of the null subject parameter will guarantee that (36d) is set to 0. Alternatively, we could appeal to morphological conditions and say that, although the syntactic evidence does not determine a value for

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

(36d), the "poverty" of French verbal inflection determines a negative value. We will leave this question open here.

The above paragraphs demonstrate how the various factors we are concerned with work. On the basis of simple, plausible, positive evidence, the learner can converge on the correct parameter settings for Modern French. In what follows we will show how these same factors led to a major parametric change in French, circa 1500.

3.3 Old French

As mentioned earlier, OF allowed nominative Case assignment under government (see (34a–b)). We assume that nominative Case could also be assigned under agreement, although we will return to this point in section 4. (34b) shows that OF allowed null subjects, although it is well known that these were possible only in contexts of inversion. Another well-known and much-discussed difference between OF and ModF is that the OF nominative pronouns *je, tu, il,* etc., were potentially tonic elements, unlike their ModF counterparts (see Kayne 1975 on ModF; Adams 1987a,b, Roberts 1992b:sec. 2.2, and below on OF). These facts about OF syntax lead to the following parameter settings, in terms of (36):

(50) The target string for OF is 11011.
Nominative Case assignment was possible both under agreement and under government; null subjects were possible; V2 was obligatory in matrix declaratives.

As in the previous section, we now show how this string could be determined on the basis of simple, positive evidence.[14]

The following kinds of sentence were available as evidence, where (*S*) indicates a null subject:

(51) a. XVS
(Et) lors demande Galaad ses armes.
(and) then asks Galahad (for) his arms
b. SVO
Aucassins ala par le forest.
Aucassin went through the forest

[14] In the case of OF, as in the case of all languages now without native speakers, negative evidence in the form of grammaticality judgments is unavailable. Linguists working on such languages are in a situation almost analogous to that of children acquiring their native language, although in fact the linguists' situation is worse since they have no access to UG and their data are seriously degenerate owing to dialect mixture, scribal error, and so on. Unlike children, however, linguists have no access to a regular input text. Children are surrounded by native speakers producing grammatical utterances. Linguists obviously are not, since all the native speakers are dead.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

c. XV(S)O
Si firent grant joie la nuit.
so (they) made great joy the night

(52) a. *wh*VSO
(Mais) ou fu cele espee prise. . . ?
(but) where was that sword taken

b. *wh*VSO
Ne nos connoissez vos mie?
NEG us know you not

(51a) is a V2 declarative (as in modern Germanic languages, conjunctions like 'and' and 'but' do not count in the computation of V2; these elements can be external to CP when they conjoin CPs). The relevant parts of the structure of this sentence are as follows:

(53) $[_{CP}$ Lors $[_{C'}$ demande $[_{IP}$ Galaad . . .]]]

Here the inflected verb in C assigns nominative Case to the subject NP, *Galaad,* under government. Of the five parameters in (36), this example then positively triggers nominative Case assignment under government and V2. More generally, this word order has the following p-encoding:

(54) XVS p-encodes [* 1 * * 1]
Nominative Case is assigned under government, and V2 is obligatory in matrix declaratives.

OF also allowed SVO sentences like (51b). As in the case of the ModF SVO order, this kind of sentence is p-ambiguous in the following way:

(55) SVO p-encodes either [* 1 * * 1]
Nominative Case under government, V2
or [1 * * * 0]
Nominative Case under agreement, no V2

We will return to this point below.

As noted earlier and illustrated in (51c), OF allowed null subjects in V2 contexts. Such examples are also p-ambiguous from the point of view of the learner: if V is in C, then the null subject is licensed under government in Spec of I; if V is in I, then the null subject is licensed under agreement in Spec of I. In the former situation, nominative Case under government and V2 are triggered; in the latter, nominative Case under agreement is triggered along with a negative value for V2. In both cases, the null subject parameter is positively triggered. The following p-ambiguity arises:

(56) XV(s)O p-encodes either [* 1 * 1 1]
As above, with null subject
or [1 * * 1 0]
As above, with null subject

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Now consider the interrogatives in (52). (52a) has the same trigger properties as a V2 declarative, except that by assumption interrogatives cannot trigger the V2 parameter. On the assumption that the nominative pronouns were tonic,[15] (52b) involves nominative Case assignment under government to the clitic, just as with any other NP subject. These examples, then, have the following p-encoding:

(57) *wh*VSO p-encodes [* 1 * * *]
Nominative under government

Putting the above p-encodings together, we arrive at (58).

(58) a. [* 1 * * 1]
Nominative under government and V2
b. [1 * * * 0]
Nominative under agreement and no V2
c. [* 1 * 1 1]
Nominative under government, null subject, and V2
d. [1 * * 1 0]
Nominative under agreement, null subject, and no V2
e. [* 1 * * *]
Nominative under government

Both nominative Case parameters are triggered positively. (Notice that the positive evidence overrides the fact that these two parameters are in a shifting relationship; we return to this in section 4.) The null subject parameter is also positively triggered. V2 is also triggered if we take it that the positive evidence for the more complex trigger weighs more heavily than the pressure in favor of the simpler structure in the p-ambiguous cases; this is a matter that can be captured by the fitness metric. Finally, as mentioned in footnote 11, there is no morphological evidence in favor of subject clitics, in that there was only one series of subject pronouns at this time. Phonological evidence presumably militates against treating the nominative pronouns as obligatory clitics; for example, these pronouns could be stressed in OF, as their occurrence in topicalized position indicates:

(59) Je, que sai?
me what do I know

[15] In fact, there are reasons to think that in the position immediately following the inflected verb, as in (52b), these pronouns did cliticize in OF (see Dupuis 1989:119f., Roberts 1992b:sec. 2.2.2, and Vance 1989: 70ff.). However, Roberts argues that the crucial step in the development of the system of subject pronouns in French was the emergence of complementary distribution between the *je*-series and the *moi*-series. This happened because the cliticization of the *je*-pronouns became obligatory in MidF. What the OF evidence shows is that these pronouns were optionally clitics in that they cliticized only in certain contexts. In other contexts, such as those in (59) and (60), these pronouns were clearly tonic. It may be, then, that the correct formulation of the parameter in (36c) should refer to obligatory cliticization of nominative pronouns, or, more likely, to the existence of a special series of clitic pronouns. Note that in the latter case the trigger for the parameter is morphological: the learner must recognize two paradigms of subject pronouns.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Moreover, subject pronouns, unlike object pronouns, could appear first in V2 declaratives. This indicates that they "counted" just like other XPs for the determination of V2; object pronouns did not "count," however:

(60) a. Tu es or riche et ge sui po proisié.
you are now rich and I am little valued
b. Toutes ces choses te presta Nostre Sires.
all these things to you lent our Lord

On the basis of evidence of this kind, the subject clitic parameter was set to 0.

Thus, we have demonstrated how simple, positive data could trigger the parameter settings for OF. Indeed, this discussion of the OF data brings out one important point: clear, positive evidence overrides all other considerations. We showed this in two cases. First, OF had a shifted system with respect to the nominative Case parameter, but learners nevertheless converged on this system since there was clear, positive evidence for it. Second, the p-ambiguities of SVO and V2/null subject examples are resolved by the unambiguous V2 cases, and moreover this resolution is in the direction of the more complex structure. In other words, clear, positive evidence can override both subset/shifting considerations and the pressure toward the simplest possible structure. In terms of our assumptions and definitions, "clear, positive evidence" means non-p-ambiguous evidence. Since the only non-p-ambiguous evidence for V2 is the XVS order, this type of sentence clearly played a crucial role. This order was very frequent in OF matrix declaratives. Roberts (1992b:sec. 2.3.1) gives the following percentages for (X)VS and SV(X) order (based on the first 100 matrix declaratives with overt subjects in six representative texts):

(61) (X)VS = 58%
SV(X) = 34%

Although a more sophisticated and exhaustive quantitative analysis is needed in order to fully demonstrate the point, we can conclude that (X)VS orders were sufficiently frequent to trigger a positive setting of the V2 parameter. This in turn means that SVO sentences could be analyzed as V2, unlike in ModF. Thus, a shifted system is allowed because there is clear evidence for it; the situation is quite different in ModF, where the only evidence for the shifted system is p-ambiguous and is therefore disregarded.

In section 1 we introduced the notion of stability of parameter setting, proposing that a parameter setting is stable to the degree that its expression in the input data is unambiguous. Was the V2 parameter stable in OF? The only non-p-ambiguous trigger for V2 is provided by XVS orders. The frequency of these orders positively sets the nominative-Case-under-government parameter and thereby makes the V2 parse available for the p-ambiguous SVO and null subject structures. The potential instability created

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

by the "non-V2 parses" of these examples is eliminated in the optimal grammar of OF. Nevertheless, it is likely that the non-V2 parse for SVO and null subject sentences was a close rival for the V2 parse, even in (later) OF, especially since elegance considerations always favor a non-V2 parse over a V2 parse where there is a choice. More explicitly, in terms of the fitness metric, the existence and frequency of an unambiguous trigger for V2 was sufficient to establish a positive setting for the V2 parameter. Recall that the relative elegance of a parse plays a less crucial role in judging fitness than real grammatical violations. This is because the elegance factor is scaled down by the constant c of the fitness metric, whereas violations are not scaled down. Thus, a hypothesis that leads to slightly more inelegant representations without generating grammatical violations will ultimately drive a hypothesis that generates elegant violations out of the population.

In the next section we will show how the MidF situation contrasts with what we have just described for OF. In particular, we will show that, in part because of the introduction of new word orders and in part because of the diminishing frequency of XVS, XVS orders were no longer able to trigger a positive value for the V2 parameter. As a result, the V2 parameter became maximally unstable. The instability was resolved by a parametric change that led to the loss of the constructions in (32)–(34).

3.4 Middle French

In MidF, XSV was introduced, and SVO and V1 became more frequent. These facts are standardly described in histories of French (see Harris 1978, Marchello-Nizia 1979, Vance 1989, and, for a detailed treatment in terms of the parameters under discussion here, Roberts 1992b). Together, they meant that the V2 constraint was less rigorously respected than it had been in OF (although V2 orders were still possible throughout this period, unlike ModF). Also, a separate series of nominative clitics emerged. For now, we will take the introduction of the new word orders as given, although we discuss possible causes for this change in section 4 (also see Adams 1987a,b, Roberts 1992b). We treat the cliticization of nominative pronouns as a phonologically driven change. Otherwise, MidF was like OF and different from ModF, in particular with respect to nominative Case assignment under government and null subjects. We do not present a target string for MidF, however, since we precisely wish to show how indeterminacy in one parameter (V2) created indeterminacy elsewhere (nominative Case assignment under government and the possibility of null subjects).

Let us consider the types of evidence available in MidF. As in OF, the following kinds of declaratives were found:

(62) a. XVS
Or avoit nostre curé priez des aultres prebtres.
now had our priest asked the other priests

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

b. SVO
Les Anglais veulent un roi guerrier.
the English want a warrior king
c. XV(S)O
Or ai eu plusseurs fois grant imagination.
now have (I) had several times great imagination

Also as in OF, these constructions have the following p-encodings, corresponding to (62a), (62b), and (62c), respectively:

(63) a. XVS: [* 1 * * 1]
Nominative under government and obligatory V2 in matrix declaratives
b. SVO: [* 1 * * 1]
As above
or [1 * * * 0]
Nominative under agreement and no V2
c. XV(S)O: [* 1 * 1 1]
As in (a) with null subjects
or [1 * * 1 0]
As in (b) with null subjects

The changes that took place in MidF created further possibilities, however. Consider the following examples (where *s* indicates a subject clitic):

(64) a. XVs
Or ai je proposé ensi que . . .
now have I proposed thus that
b. XsV
Et ce conseil nous vous donnons.
and this advice we to you give

Taking these examples to positively trigger the subject clitic parameter, we propose that they have the p-encodings in (65a) and (65b), respectively.

(65) a. XVs: [* * 1 * 1]
Subject clitics and V2
b. XsV: [* * 1 * 0]
Subject clitics and no V2

Since clitics can receive Case in ways unavailable to other nominal elements, sentences containing subject clitics provide no information about either nominative Case assignment parameter. The order verb-clitic in (64a) triggers a positive setting for the V2 parameter. On the other hand, since French subject clitics (then as now) do not attach to a verb and move with it (unlike object clitics), the order clitic-verb in (64b) triggers

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

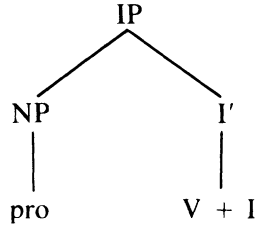
a negative value for the same parameter (but see below for further discussion of this kind of case).

As mentioned earlier, MidF allowed, with growing frequency, other word orders that were not found in OF:

(66) a. XSV
Lors la royne fist Saintré appeller.
then the queen had Saintré called
b. (S)VY
Se appensa de faire ung amy.
(he) to himself thought to make a friend

(66a), combined with the greater frequency of SVO orders in MidF as compared to OF, shows that V2 began to "erode" at this period. Sentences like (66b) illustrate another phenomenon, noticed and analyzed by Vance (1989): the fact that null subjects increase their distribution in this period, no longer being licensed only in inversion contexts. Roberts (1992b:sec. 2.3.5) analyzes this situation in terms of the idea that null subjects could be licensed under agreement as well as under government in MidF, whereas in OF they were licensed only under government. So MidF allowed a null subject in the following configuration:

(67)



The p-encodings for these orders are as follows:

(68) a. XSV: [1 * * * 0]
Nominative under agreement and no V2
b. (S)VY: [1 * * 1 0]
As in (a) with null subject
or [* 1 * 1 1]
Nominative under government, null subject, and V2

In interrogatives the same general situation holds as in declaratives. On the one hand, the same kinds of examples are found as in OF:

(69) a. *wh*VSO
Que voelt ceste parolle dire?
what wants this word to say
'What does this word mean?'

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

b. *wh*VsO
A qui estes vous?
whose are you

(69a) has the same p-encodings as its OF counterpart:

(70) *wh*VSO: [* 1 * * *]
Nominative under government

(69b), on the other hand, no longer encodes nominative Case under government, since the subject has cliticized:

(71) *wh*VsO: [* * 1 * *]
Subject clitics

Let us now put together the MidF p-encodings:

(72) a. [* 1 * * 1]
Nominative under government and V2
b. [1 * * * 0]
Nominative under agreement and no V2
c. [* 1 * 1 1]
Nominative under government, null subject, and V2
d. [1 * * 1 0]
Nominative under agreement, null subject, and no V2
e. [* * 1 * 1]
Subject clitics and V2
f. [* * 1 * 0]
Subject clitics and no V2
g. [* 1 * * *]
Nominative under government
h. [* * 1 * *]
Subject clitics

In terms of p-encodings alone, the V2 parameter appears to be no more or less unstable than it was in OF. However, two factors distinguish the MidF situation from the OF one. First, the unambiguous trigger for V2—XVS order—was much less frequent in MidF than in OF. According to Marchello-Nizia (1979), the mean orders for three texts from the late 15th century are as follows:

(73) (X)VS = 10%
SV(X) = 60%

This is a significant difference in frequency as compared to OF (see (61)). The second factor concerns the status of SVO clauses. As shown earlier, the "V2 parse" for these clauses is disregarded in ModF, yet it was favored in OF. In MidF there is total inde-

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

terminacy on this point: there is (infrequent) evidence for V2 in the form of XVS order, and there is evidence against V2 in the form of XSV. Any parsing device with a positive setting for V2 would engender a violation on this word order and would be disfavored by the fitness metric. Another factor that adds to the instability of V2 at this point is the development of left-dislocation with a resumptive pronoun (Priestley 1955, Kroch 1989). This is illustrated by the following example from Priestley:

(74) Les autres arts et sciences, Alexandre les honoroit bien.
the other arts and sciences Alexandre them-honored well

The development of this type of construction led to shifting of the type described in (5) and (6). That is, the interaction between left-dislocation and V2 further obscured the latter due to surface "V3" orders. Kroch (1989:215) shows that there is a real correlation between the rise of the construction in (74) and the loss of V2. The correlation results from the action of the fitness metric, which will judge a system of this type as relatively unfit.

Late MidF V2 provides an instance of the situation described in section 1: learners are unable to converge on a single value for a parameter. In other words, the V2 parameter is maximally unstable. This case therefore exemplifies the "pathological" situation for acquisition. Since the available data cannot decide between two parametric values, other aspects of the fitness metric come into play: the Subset Condition and the elegance criterion.

As noted earlier, a language with both V2 and left-dislocation will be disfavored by the Subset Condition, since it is a case of shifting. Another factor that can decide between competing parses, and therefore competing p-encodings and triggers, is the criterion of elegance. It is reasonable to suppose that learners follow a least effort strategy in that they try to assign the simplest possible parse to the input string.[16] This idea can be instantiated in terms of counting nodes, traces, or chain positions. We will not attempt to choose between those possibilities (Roberts (1992b) opts for chain positions; for a formal statement of this, see his chapter 2, note 26); what is important here is that any parse that represents the inflected verb as being moved to C is more costly in terms of the least effort strategy than one that represents the verb as being moved only to I (by any of the above criteria).

Suppose, then, that the least effort strategy plays a crucial role in resolving the instability in the data, by penalizing all p-encodings that depend on V-movement to C where there is a choice between this and V-movement to I. More technically, suppose that hypothesis h_1 is identical to hypothesis h_2 except that h_2 allows for V2 in matrix

[16] This idea is discussed at length in the context of syntactic change by Roberts (1992b), who notes the close resemblance between this idea and the Transparency Principle proposed by Lightfoot (1979). Also see de Vincenzi's (1989) proposal that something of this kind is a general parsing strategy, not limited to language learners. Note that the least effort strategy as conceived here is not a principle of grammar; in this we differ from Chomsky (1991).

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

declaratives whereas h_1 does not. That is, h_1 and h_2 admit the same sentences and contain the same number of superset settings to parameters, differing only in the value for the V2 parameter. Hypothesis h_2, then, systematically includes more structure in its representation than h_1 since h_2 will represent the verb as having moved to C (as well as movement of the subject in SVO). In other words, if h_1 returns k nodes on a structure, h_2 will return $k + n$ nodes. Letting m represent the number of superset settings in each hypothesis, running each of the above through the fitness metric will yield the following ratings:

(75) a. h_1: $1 - \dfrac{m + ck}{2m + c(2k + n)}$

b. h_2: $1 - \dfrac{m + c(k + n)}{2m + c(2k + n)}$

Since $1 - \dfrac{m + ck}{2m + c(2k + n)}$ is greater than $1 - \dfrac{m + c(k + n)}{2m + c(2k + n)}$, the fitness metric prefers h_1 over h_2 and the learner is under pressure to select h_1. This, then, effectively sets the V2 parameter to 0.

Like OF, MidF had one order where the V2 parameter was unambiguously p-encoded as 1: namely, XVS orders, which unambiguously p-encode [* 1 * * 1]. In the situation of instability that reigned in MidF, the fitness metric, formulated to take account of the way in which the least effort criterion resolves p-ambiguities, will lead to convergence on a grammar where such experience is simply disregarded (i.e., not parsed where no alternative analysis can be found).[17] Thus, this case shows how an unambiguous trigger for a given property can be disregarded when the system is maximally unstable, even if the instability is located in another area of the grammar.

If the hypotheses where the V2 parameter has a positive value are penalized, the only remaining triggers for nominative Case assignment under government are *wh*VSO orders. This order, too, is only weakly triggered in 15th-century French. The difference between MidF and OF in this regard was that several new constructions were available in MidF, notably complex inversion (as in *Où Jean est-il allé?* '(lit.) where Jean is-he gone') and (*qu'*)*est-ce que* '(what) is-it that'.[18] Whereas nominative Case assignment under agreement received strong support from the input data, nominative Case assignment under government received very little. Since the two parameters are in a shifting relationship, there was some pressure (built into the fitness metric, as shown in section 2) not to set them both to 1. In this situation, the fact that nominative Case assignment

[17] An alternative analysis is often available. Roberts (1992b:sec. 2.4.1) shows that many cases of V2 could be treated as "free inversion."

[18] For a synchronic analysis of the former construction, see Rizzi and Roberts 1989, and for a discussion of its diachronic development, Roberts 1992b:sec. 2.3.4. Concerning the development of the latter construction as a nonemphatic interrogative, see Foulet 1921.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

under government was only weakly triggered led to a change in the value of this parameter.

The change to a system with nominative Case assignment under agreement entailed a change in the null subject parameter (already only weakly triggered, as (74) shows) for theory-internal reasons. Under the assumption that null subjects can only be licensed in positions where Case is assigned (Rizzi 1986a), once nominative Case could no longer be assigned under government, null subjects could no longer be licensed under government. In this way, French lost null subjects with no significant change in the verbal inflectional morphology. There is a complication here, however—namely, that MidF, unlike OF, also allowed null subjects to be licensed in configurations of agreement. Why were these null subjects lost along with those licensed in government configurations? Roberts (1992b) answers this question in terms of a postulate concerning the identification of null subjects that we can phrase as follows:

(76) Where null subjects are licensed only in configurations of agreement, they require a "pronominal" Agr for identification.

A "pronominal" Agr is an Agr that morphologically distinguishes at least five persons—that is, an Agr of the kind found in languages like Spanish and Italian. French Agr is not pronominal in this sense, and indeed has not been since early in the OF period. The intuition behind (76) is that a system where null subjects are licensed under government requires less inflectional morphology to recover the content of those null subjects than one where the only licensing configuration is agreement, since government is a closer syntactic relation than agreement. A system that licenses null subjects both under government and under agreement, like MidF, tolerates a relatively poorer agreement morphology. Therefore, once null subjects could no longer be licensed under government in French, the relative "poverty" of the verbal morphology became crucial, and null subjects were also lost in contexts where they had been licensed under agreement. As Roberts shows, the parallel development of Gallo-Italian dialects, in particular Veneto, supports the postulation of (76).

Thus, at the beginning of the MidF period (ca. 1300) the relevant parameter settings were those in (77a); by the end of this period (ca. 1500) they had become those in (77b).

(77) a. 11011 (= OF)
b. 10100 (= ModF)

It is clear that the crucial element of instability was created by the gradual erosion of V2 as a rigid constraint on word order in matrix declaratives. In particular, the introduction and spread of XSV orders brought about a situation that eliminated a crucial trigger for nominative Case assignment under government—XVS order. The previous discussion shows how the genetic algorithm approach to learnability, and in particular the fitness metric, can shed light on this. What seems to have happened is that V2 was mildly unstable in, say, 1300 (recall the discussion at the end of section 3.3) in the sense

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

that non-V2 parses for certain types of sentence (e.g., SVO) were close competitors for V2. These competitors generated "mutant" word orders, notably XSV, which were highly successful. The critical point was reached in the late 15th century, when V2 was eliminated. For completely contingent reasons (which concern the overall organization of the MidF grammatical system), the loss of V2 led to the loss of nominative Case assignment under government. And for reasons having to do with the organization of UG, this entailed the loss of null subjects. Moreover, Roberts (1992a) argues that this in turn led to the loss of clitic climbing (also see Kayne 1989). This account of syntactic changes in the history of French illustrates how syntactic change can be internally driven: change in one parameter can destabilize another. We will provide another example of this in section 4.

However, we now find ourselves up against the problem posed by innovations: How were XSV orders introduced into a V2 system? Since these orders are ungrammatical in modern V2 Germanic languages, their introduction into a V2 system requires some comment. If we say that the weakening of V2 was a condition for this development, we risk falling into an unproductive regress. It was in part for this reason that we avoided the issue earlier and simply took this innovation as given. However, there are good reasons to think that the introduction of XSV order is related to the cliticization of subject pronouns. Adams (1987b) points out that the overwhelming majority of early cases of XSV involved a pronominal subject. As Adams suggests, it is possible that XSV originates from cases of V2 where the clitic subject pronoun is not counted in determining V2.[19] If Adams's idea is correct, then the initial stimulus for the erosion of V2 comes from a morphophonological change in the subject pronouns. As is frequently the case, syntactic change can be traced back to extrasyntactic factors, although the relationship between the extrasyntactic factors and the syntactic changes they cause can be extremely indirect. This is because instability, once introduced, can propagate through a grammatical system.

4 Some Concluding Remarks

Here we wish to address some of the wider issues raised by our case study of language change. These concern the shifting relationship between the nominative Case parameters in section 3.1 with respect to the OF data and what our approach has to say about the classic questions for diachronic linguistics concerning the nature of innovation and loss.

How is it that a massively unstable system of parameter settings, like the one in MidF, can come into being in the first place? Of course, factors external to the syntax can destabilize a syntactic system, but we believe that instability can propagate within

[19] We do not want to propose that preverbal subject pronouns in MidF or ModF are syntactic clitics; rather, following Kayne (1983), we believe that these pronouns cliticize only in PF. However, the ultimately unsuccessful hypothesis that these pronouns were indeed syntactic clitics could nevertheless have given rise to XSV orders at the time when the subject-pronoun system was undergoing change. See Roberts 1992b for a more elaborated approach.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

a syntactic system and that exactly this has happened in the history of French. Consider again the p-encodings for the OF data:

(78) a. [* 1 * * 1]
b. [1 * * * 0]
c. [* 1 * 1 1]
d. [1 * * 1 0]
e. [* 1 * * *]

Bearing in mind that the correct grammar for OF did not contain non-V2 parses (i.e., the p-encodings in (78b) and (78d) are discarded in the correct grammar), it seems that nominative Case assignment under agreement had a quite precarious status in OF. There was another trigger for nominative Case assignment under agreement, however: the fact that subordinate clauses regularly had SVO order (assuming, contra Lightfoot (1989), that subordinate word order can trigger parameter settings). Thus, it is the fact that OF had a root/embedded asymmetry with respect to V2 order that is crucial for triggering nominative Case assignment under agreement. Now, there is evidence that early OF (prior to ca. 1200) allowed embedded V2 (Cardinaletti and Roberts, to appear, Dupuis 1989, Hirschbuhler 1990). This means that nominative Case assignment under agreement was an OF innovation, emerging in subordinate clauses as V2 became a uniquely root phenomenon. This innovation started the chain of changes leading to the MidF innovations that were crucial to our account in section 3.4 (and hence to the later changes discussed there).

Assume that an archaic stage of OF did not allow nominative Case assignment under agreement. How can Case assignment under agreement arise? Notice that when such assignment comes into the grammar, a shifted system is introduced on the basis of a nonshifted one. Following an idea originally due to Cardinaletti (1990), let us suppose that expletive elements can never topicalize. In a V2 system, however, Spec of C′ is a topic position: it is an Ā-position and a position that does not receive Case. Cardinaletti proposes that when an expletive occupies this position, as frequently happens in the V2 Germanic languages, the position is able to count as an A-position in that (nominative) Case can be assigned there. Thus, we can attribute the introduction of nominative Case assignment under agreement to the introduction of a lexical expletive capable of occupying Spec of CP in matrix declaratives. OF had a lexical expletive *il* that appeared in Spec of C′ in examples like (79) (from Einhorn 1974:123).

(79) Il ne me chaut.
it not to me matters
'It doesn't matter to me.'

Supposing that this construction emerged in archaic OF, we can then say that nominative Case assignment under agreement was triggered by this kind of example.

Finally, let us briefly consider what implications our proposals may have for tra-

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

ditional preoccupations of diachronic syntax: the nature of innovation and the nature of loss. Of course, it should be immediately clear that the conception of how grammatical systems differ from one another that lies at the heart of the principles-and-parameters approach means that parameters themselves never change.[20] What changes over time are parametric values.

Nevertheless, at the level of constructions (e.g., available word order types) it is clear that possibilities are both innovated and lost. In our terms, innovation may arise from one of two sources: either internally, when a parametric change makes new constructions available, or externally, when phonological or morphological change weakens evidence for certain hypotheses. The second type of innovation is likely to lead to instability at the level of parameter settings, as in the case of the introduction of XSV orders triggered by the cliticization of subject pronouns in MidF.[21]

Concerning loss, it seems that only parametric change can truly eliminate a construction in the sense that construction C is accepted by native speakers of language L at time T and rejected at T' ($T > T'$). This has been the fate of simple inversion, V2, and null subjects in French. In terms of the standard view of language acquisition, this situation seems problematic. Put very simplistically, Why is one generation's trigger experience the next generation's fossil? This is the logical problem of language acquisition again. Various solutions have been proposed, but we believe we have discovered a new and interesting one.

An approach to learnability based on a genetic algorithm including a version of the fitness metric makes it possible to see how a data point can be disregarded in a situation of instability (where instability can be formalized); this was what happened in the case of XVS orders in 15th-century French. Although relatively infrequent and often parsable as some other construction, XVS was certainly found in 1500, and so, given the standard assumption that parameters can be set on the basis of quite impoverished experience, an account of loss based on frequency considerations alone will not answer the fundamental question. The fitness metric, properly formulated so that frequency and other considerations are taken into account, seems able to resolve this tension between standard views of acquisition and the fact that structures are lost in the course of language change, since it can be seen why one class of input strings may be rendered unparsable. This can happen even where, as in the case of XVS orders, the input in question is intrinsically simple and structurally "transparent"; here we see a major difference between our account and the approach to language change based exclusively on something like Lightfoot's (1979) Transparency Principle, although we believe our approach retains

[20] Except perhaps at the higher diachronic level of phylogenetic change; it is a reasonable assumption that the set of parameters available to modern *Homo sapiens* is not the same as the set that was available to the first hominids with a language faculty. Of course, we are concerned in the text with changes in the recorded history of languages that by assumption fall within the set of human languages, so this question does not arise.

[21] There is at least a metaphorical sense in which cases like XSV are successful rogue hypotheses, where success is determined by the least effort criterion. This is mutation at the level of constructions, not at the level of parameters, so the mutation operator of section 2 is presumably not relevant.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

the basic insight behind the Transparency Principle in the elegance part of the fitness metric.[22]

Another important consideration that emerges from our discussion is that exactly the same string S_i can successfully trigger a parameter setting $P(v_1)$ in one grammatical system G_i, but fail to trigger $P(v_1)$ in system $G_j \neq G_i$. French XVS order is a case of exactly this sort, where G_i is the grammar of OF and G_j that of late MidF. In terms of the genetic algorithm, S_i can trigger a successful hypothesis or an unsuccessful one. As in the biological world, successful propagation depends as much on the external environment as on internal properties, so that little can be predicted purely on the basis of internal structural criteria. It is this aspect of the genetic algorithm that makes possible a deeper understanding of language change and demonstrates how successive generations may treat the "same" trigger experience differently. Note also that in these terms, language change refers not only to the "limit cases" of innovation and loss, but also to the varying success of strings in encoding viable parameter settings.

Our approach also has implications for the theory of markedness. It is part of the classical concept of markedness that marked properties are both diachronically unstable and "difficult" in terms of acquisition. A shifted system of parameter settings can be thought of as a marked system. It is clear from our discussion that a shifted system is diachronically unstable. Consider again the shifted system discussed in section 2, which featured both V2 and left-dislocation. Neither V2 nor left-dislocation is marked on its own (note the stability of Germanic V2 and the fact that all periods of French feature left-dislocation of one kind or another); however, their combined presence in a system leads to markedness—witness the instability of MidF.[23] So we suggest that in general markedness, rather than being an inherent property of certain parameter values, is a property that derives from the interaction of parameters in a given grammatical system, relative to the fitness metric. This in turn implies that a given parameter value can be marked in one grammatical system (or at one period) and unmarked in another system (e.g., at another period).

Diachronic studies of the type discussed here also have important implications for the study of language learnability and language acquisition. As discussed briefly above, diachronic change represents a type of "pathological" learning, where learners systematically arrive at the wrong grammar for the target language. Strictly speaking, these are cases where learners fail. We would argue that learners fail for reasons that reveal something important about their internal structure. Parametric change is the result of an input text that places indifferent pressure on the learner's hypotheses; several different

[22] More recently, Lightfoot (1989) has proposed a new approach to change based on "Degree-0 learnability." A detailed comparison of that approach with the one developed here is beyond the scope of this article (though see Clark, in preparation).

[23] Modern German also has left-dislocation, but with a tonic resumptive pronoun. On the other hand, MidF left-dislocation featured atonic resumptive pronouns. This was yet another way in which the clitic nature of pronouns in MidF created instability.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

grammars can provide an acceptable account for the input text. We have shown that other factors, always related to the learner's internal fitness metric, come into play to distinguish between the competing hypotheses. These factors involve the Subset Condition and a measure of elegance. Let us return to the fitness metric, repeated here as (80).

(80) *The fitness metric*

$$\frac{(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j) - (v_i + bs_i + ce_i)}{(n-1)(\sum_{j=1}^{n} v_j + b\sum_{j=1}^{n} s_j + c\sum_{j=1}^{n} e_j)}$$

Our study of diachronic change reveals certain facts about the scaling constants b and c. We assume that empirical coverage of the input text is the learner's central interest; thus, violations (calculated by $\sum_{j=1}^{n} v_j$ for the population and by v_i for the individual) are the single most important factor in the equation. Both superset settings and elegance are scaled down by the constants b and c, respectively.

Let us now consider what the relative magnitudes of b and c are. At a certain point, French was a V2 language that allowed for left-dislocation (the latter associated with atonic pronouns), and it was a shifted language that would be selected against by the fitness metric. Furthermore, the relative frequency of structures that would have required both V2 and left-dislocation was relatively low, placing little pressure on the learner in terms of violations. All else being equal, learners could have preferred either a language with matrix V2 and no left-dislocation or a language with left-dislocation and no matrix V2. Notice that left-dislocation is a superset parameter; a language that allows left-dislocation in addition to its basic word order is a superset of a language that allows only the basic word order. We argued, on the other hand, that matrix V2 led to more complex representations, relative to the input text, than a grammar without matrix V2.

Now, the changes we have illustrated in French involve the abandonment of matrix V2, a nonsuperset parameter, and the persistence of left-dislocation, a superset parameter. Given our premises, then, the fitness metric must have preferred a grammar that generated an elegant set of representations and a superset language over a grammar that generated inelegant representations and a subset language. Thus, learners appear to consider elegance a more important factor than superset settings when evaluating hypotheses:

(81) $c > b$

Thus, our study of diachronic change has enabled us to make a concrete hypothesis about how learners evaluate parameter settings. We can now test this hypothesis against actual child grammars, perhaps by attempting to characterize successive developmental stages in child language. In general, we should see children avoiding grammars that create inelegant representations. More to the point, we should find children resisting

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

grammars that force longer chains to the point of, temporarily at least, preferring grammars with superset settings if these grammars can approximate the target.

We have shown how a theory of language learning based on a genetic algorithm affords a novel and insightful account of language change, taking as our case study of language change the development of word order and null subjects in French. We believe that our account sheds light both on the mechanisms of language change and on those of language acquisition, and goes some way toward building a bridge between these two domains; in this respect, our work is conceptually very close to work by Lightfoot (1989, 1991). Moreover, we have shown that it is possible to characterize the markedness of systems and to clearly see the role played by such factors as elegance and frequency of input, and the interactions between these factors. We know of no other approach to language learnability and language change that achieves these results.

References

Adams, Marianne. 1987a. From Old French to the theory of pro-drop. *Natural Language and Linguistic Theory* 5:1–32.

Adams, Marianne. 1987b. Old French, null subjects and verb second phenomena. Doctoral dissertation, UCLA, Los Angeles, Calif.

Baker, Mark. 1988. *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.

Belletti, Adriana. 1990. *Generalized verb movement: Aspects of verb syntax*. Turin: Rosenberg and Sellier.

Berwick, Robert. 1985. *The acquisition of syntactic knowledge*. Cambridge, Mass.: MIT Press.

Booker, L. B., David E. Goldberg, and John H. Holland. 1990. Classifier systems and genetic algorithms. In *Machine learning: Paradigms and methods,* ed. Jaime Carbonell, 235–282. Cambridge, Mass.: MIT Press.

Cardinaletti, Anna. 1990. *Pronomi nulli e pleonastici nelle lingue germaniche e romanze: Saggio di sintassi comparata*. Dottorato di ricerca in linguistica, Università di Padova.

Cardinaletti, Anna, and Ian Roberts. To appear. Clause structure and X-second. In *Levels, principles and processes: The structure of grammatical representations,* ed. W. Chao and G. Horrocks. Berlin: Foris/de Gruyter.

Chomsky, Noam. 1977. On *wh*-movement. In *Formal syntax,* ed. Peter Culicover, Thomas Wasow, and Adrian Akmajian, 71–132. New York: Academic Press.

Chomsky, Noam. 1981. *Lectures on government and binding*. Dordrecht: Foris.

Chomsky, Noam. 1991. Some notes on economy of derivation and representation. In *Principles and parameters in comparative grammar,* ed. Robert Freidin, 417–454. Cambridge, Mass.: MIT Press.

Clark, Robin. 1990. *Papers on learnability and natural selection*. Technical Reports in Formal and Computational Linguistics, No. 1. Université de Genève.

Clark, Robin. 1991. A computational model of parameter setting. Paper presented at the American Association for Artificial Intelligence Spring Symposium on Machine Learning, Natural Language, and Ontology. Stanford, Calif.

Clark, Robin. 1992. The selection of syntactic knowledge. *Language Acquisition* 2:85–149.

Clark, Robin. In preparation. Finitude, boundedness, and approximate learning of natural languages. Ms., Université de Genève.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Darwin, Charles. 1859. *On the origin of species*. London: John Murray.
Dupuis, Fernande. 1989. L'expression du sujet dans les subordonnées en ancien français. Thèse de Ph.D., Université de Montréal, Montréal, Québec.
Einhorn, Einar. 1974. *Old French: A concise handbook*. Cambridge: Cambridge University Press.
Everett, Daniel. 1986. Pirahã clitic doubling and the parametrization of nominal clitics. In *MIT working papers in linguistics* 8, 85–127. Department of Linguistics and Philosophy, MIT, Cambridge, Mass.
Everett, Daniel. 1989. Clitic doubling, reflexives and word order alternations in Yagua. *Language* 65:339–372.
Foulet, Lucien. 1921. Comment ont évolué les formes de l'interrogation? *Romania* 47:243–348.
Foulet, Lucien. 1982. *Petite syntaxe de l'ancien français*. 3d ed. Paris: Editions Champion.
Gold, E. M. 1967. Language identification in the limit. *Information and Control* 16:447–474.
Goldberg, David. 1989. *Genetic algorithms in search, optimization, and machine learning*. Reading, Mass.: Addison-Wesley.
Haldane, J. B. S. 1990. *The causes of evolution*. Princeton, N.J.: Princeton University Press.
Harris, Martin B. 1978. *The development of French syntax: A comparative approach*. London: Longmans.
Hirschbuhler, Paul. 1990. La légitimation de la construction V1 à sujet nul dans la prose et le vers en ancien français. *Revue québecoise de linguistique* 19:32–55.
Holland, John. 1975. *Adaptation in natural and artificial systems*. Ann Arbor, Mich.: University of Michigan Press.
Kayne, Richard. 1975. *French syntax*. Cambridge, Mass.: MIT Press.
Kayne, Richard. 1983. Chains, categories external to S, and French complex inversion. *Natural Language and Linguistic Theory* 1:109–137.
Kayne, Richard. 1989. Null subjects and clitic climbing. In *The null subject parameter*, ed. Osvaldo Jaeggli and Ken Safir, 239–261. Dordrecht: Kluwer.
Kayne, Richard, and Jean-Yves Pollock. 1978. Stylistic inversion, successive cyclicity, and Move NP in French. *Linguistic Inquiry* 9:595–621.
Kiparsky, Paul. 1982. *Explanation in phonology*. Dordrecht: Foris.
Koopman, Hilda, and Dominique Sportiche. 1991. The position of subjects. In *The syntax of verb-initial languages*, ed. James McCloskey, 211–258. Amsterdam: Elsevier. [Special issue of *Lingua* 85.]
Kroch, Anthony. 1989. Reflexes of grammar in patterns of language change. *Language Variation and Change* 1:199–244.
Lightfoot, David. 1979. *Principles of diachronic syntax*. Cambridge: Cambridge University Press.
Lightfoot, David. 1989. The child's trigger experience: Degree-0 learnability. *Behavioral and Brain Sciences* 12:321–334; commentary 334–375.
Lightfoot, David. 1991. *How to set parameters*. Cambridge, Mass.: MIT Press.
Lightfoot, David, and Norbert Hornstein. 1981. *Explanation in linguistics*. London: Longmans.
Marchello-Nizia, C. 1979. *Histoire de la langue française aux XIVe et XVe siècles*. Paris: Bordas.
Natarajan, Balas. 1991. *Machine learning: A theoretical approach*. Palo Alto, Calif.: Morgan Kaufmann.
Osherson, Daniel, Michael Stob, and Scott Weinstein. 1986. *Systems that learn: An introduction to learning theory for cognitive and computer scientists*. Cambridge, Mass.: MIT Press.
Paul, Hermann. 1920. *Prinzipien der Sprachgeschichte*. 5th ed. Halle: Niemeyer.
Poletto, Cecilia. 1990. Diachronic development of subject clitics. Talk given at the Crucial Languages Workshop, Université de Genève.

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms

Pollock, Jean-Yves. 1986. Sur la syntaxe de *en* et le paramètre du sujet nul. In *La grammaire modulaire,* ed. Mitsou Ronat and Daniel Couquaux, 211–246. Paris: Editions de Minuit.
Pollock, Jean-Yves. 1989. Verb movement, Universal Grammar, and the structure of IP. *Linguistic Inquiry* 20:365–424.
Price, Glanville. 1971. *The French language: Present and past.* London: Edward Arnold.
Priestley, Lawrence. 1955. Reprise constructions in French. *Archivum Linguisticum* 7:1–28.
Renzi, Lorenzo. 1983. Fiorentino e italiano: Storia dei pronomi personali soggetto. In *Italia linguistica: Idee, storia, struttura,* ed. F. Albano Leoni et al., 223–239. Bologna.
Renzi, Lorenzo, and Laura Vanelli. 1983. I pronomi soggetto in alcune varietà romanze. In *Scritti linguistici in onore di Giovan Battista Pellegrini,* 121–145. Pisa.
Rizzi, Luigi. 1986a. Null objects in Italian and the theory of *pro*. *Linguistic Inquiry* 17:501–557.
Rizzi, Luigi. 1986b. On the status of subject clitics in Romance. In *Studies in Romance syntax,* ed. Osvaldo Jaeggli and Carmen Silva-Corvalàn, 391–419. Dordrecht: Foris.
Rizzi, Luigi. 1990. *Relativized Minimality.* Cambridge, Mass.: MIT Press.
Rizzi, Luigi, and Ian Roberts. 1989. Complex inversion in French. *Probus* 1:1–30.
Roberts, Ian. 1992a. Two types of head-movement in Romance. Ms., Université de Genève/ University College of North Wales, Bangor.
Roberts, Ian. 1992b. *Verbs and diachronic syntax.* Dordrecht: Kluwer.
Schwartz, Bonnie, and Sten Vikner. 1989. All verb second clauses are CPs. In *Working papers in Scandinavian syntax* 43, 27–49. Department of Linguistics, University of Trondheim.
Thurneysen, Robert. 1892. Die Stellung des Verbums im Altfranzösischen. *Zeitschrift für Romanische Philologie* 16:289–371.
Vance, Barbara. 1989. Null subjects and syntactic change in medieval French. Doctoral dissertation, Cornell University, Ithaca, N.Y.
Vanelli, Laura, Lorenzo Renzi, and Paola Benincà. 1986. Typologie des pronoms sujets dans les langues romanes. In *Actes du XIIe Congrès de Linguistique et Philologie Romanes.* Aix-en-Provence.
de Vincenzi, Maria. 1989. Syntactic parsing strategies in a null subject language. Doctoral dissertation, University of Massachusetts, Amherst.
Wexler, Kenneth, and Peter Culicover. 1980. *Formal principles of language acquisition.* Cambridge, Mass.: MIT Press.

(Clark)
Department of Linguistics
619 Williams Hall
University of Pennsylvania
Philadelphia, Pennsylvania 19104-6305

rclark@babel.ling.upenn.edu

(Roberts)
School of English and Linguistics
University College of North Wales
Bangor, Gwynedd LL57 2DG
Wales

ELS011@bangor.ac.uk

This content downloaded from 165.123.34.86 on Thu, 26 May 2016 14:19:47 UTC
All use subject to http://about.jstor.org/terms