Internal and External Factors Affecting Language Change: A computational model

Robert A. J. Clark

A thesis submitted in fulfilment of the requirements
for the degree of MSc Speech and Language Processing
to the
University of Edinburgh
1996

Declaration

This thesis has been composed by myself and it has not been submitted in any previous application for a degree. The work reported within was executed by myself, unless otherwise stated.

September 1996

Acknowledgements

Thanks go to many people: To Simon Kirby, for suggesting that there was some interesting stuff to look at in this field, for his supervision, and his willingness to talk computers in a sometimes hostile environment, and to Jim Hurford for his additional supervision and support. To Bob Berwick for providing papers and encouragement. To those that keep the department running including: Ethel Jack, Irene McLeod, Morag Brown, Cedric Macmartin and Norman Dryden. To friends in the department, including: Cassie, Laurence, Shane, Catriona, Miriam, Diane and Dave, whose presence helps maintain the sanity-insanity balance.

And finally to Michael Brooke at Bath University who sparked my interest in Speech and Language with his enthusiasm.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This project concerns language and *glossogenetic* language change. That is the diachronic change that a particular language goes through over a number of successive generations. Many examples of such changes can be seen historically. For example, one of the syntactic changes that has occurred in French (Clark & Roberts 1993) is that null subjects have been lost. Compare:

(1.1) *Ainsi s'amusaient bien cette nuit. (Modern French)
thus (they) had fun that night.

(1.2) Si firent pro grant joie la nuit. (Old French)
thus (they) made great joy the night.

In modern French the subject must be explicitly given, making the above Modern French example ungrammatical.

The aim here is to take a computational modelling viewpoint, and look at some of the issues involved in accounting for language change. The theories concerning language change break into two distinct approaches.

One approach to the subject is carried out within a principles and parameters framework (Chomsky 1981), where an underlying set of innate Universal Grammar principles and parameters exist. The principles are fixed and must always be adhered to, while the parameters are variable and are set by a child learning their first language. The parameters may be binary valued (having values *on* or *off* signifying if a particular attribute in the grammar is allowed/not–allowed, exists/does–not–exist) or they may be n-ary valued (having more than two values). The parameters may be independent of each other, or may be arranged hierarchically—where a parameter is only used if other parameters have already been set. How setting, unsetting and resetting is carried out depends on the particular theory in question. Some theories may allow a parameter's value to be altered many times during the acquisition process, others may only allow it to be altered once. Chomsky's argument for the existence of such a framework concerns the complex structure underlying the surface of the linguistic data that a learner is subjected to, and the necessity of a learning mechanism to cope with this.

The properties of any given natural language are regarded as a direct result of a particular configuration of these parameters. Language acquisition concerns the setting of these parameters from an initial configuration (again what this configuration actually is, be it random or all parameters unset, depends on the particular theory in question) to a configuration that best represents the perceived language that the learner in question is subjected to. Diachronic change is then attributed to the misconvergence that occurs in the process of setting these parameters at acquisition.

Another approach dwells on the communicative function of language, where it is processing constraints imposing pressures on language which account for the the particular properties that a language has (e.g. Hawkins (1994)). Processing constraints are factors like 'effort to convey a message'. This is optimised in such a way so that least effort is put in by the speaker, but enough effort is put in so that the listener fully comprehends the intended message. This manifests itself in that language constructions that are less complex and hence more *fit* for their environment are more likely to occur than equivalent more complex ones.

Within this framework, language is thought to be acquired by more generalised learning procedures (i.e. the learning processes that the brain uses to learn other things like walking, and playing) with no mechanism specifically adapted for learning language. It is the way in which language is used within the imposed constraints that determines the form that it takes. The imposed constraints shape the language and effectively determine the underlying grammar that a child acquires. Diachronic change can be attributed to the way in which a population adapts its language to the optimum fitness; a learner at each generation will tend to acquire the fittest language that it can that accounts for what it is hearing. If the resulting acquired language differs slightly from what the population actually speaks, there can be a knock-on effect in successive generations.

This project examines and evaluates some of the current proposals for modelling language and language change using the above approaches, and tries to incorporate some of the ideas about parsing complexity and fitness into a formal parameter–based framework where it is usually unaccounted for. The aim is to show that by combining these theories and incorporating both an innate set of parameters with a linguistically orientated learning mechanism, and a mechanism which evaluates parsing complexity into the same model, it is not only a plausible model, able to produce some interesting results, but it is also a realistic model, able to overcome some of the problems posed by an approach that takes only one factor into consideration. Furthermore it will be seen that the interaction of the elements from each component of the model is such that an adequate account of language acquisition and language change requires both viewpoints to be taken into consideration, as the interactions associated by a particular component will be missing where there is nothing to interact with.

Chapter 2 looks in more detail at some of the current work which is relevant. Chapter 3 replicates work based around an existing parameter–based model, and Chapter 4 develops the idea of incorporating parsing complexity issues into this framework, by proposing and demonstrating some possible alterations to the parameter–based model.

Chapter 2

Background

It is appropriate to look in some detail at the work that currently exists in the field of modelling language change. The work of Niyogi & Berwick (Niyogi & Berwick 1995, Niyogi & Berwick 1996*a*, Niyogi & Berwick 1996*b*) provides the initial framework for the work done here, along with the work on triggers by Gibson & Wexler (1994) which Niyogi & Berwick's work in turn is based upon. The parsing complexity issues are looked at through the work of Kirby (1996) and the work of Hawkins (1994) on which some of Kirby's work is based. Clark & Roberts's (1993) work is of interest because it also tries to incorporate ideas about acceptability within a given parameterisation along with issues of fitness and complexity.

2.1 Triggers

In many respects the work of Niyogi & Berwick is an extension and formalisation of work done by Gibson & Wexler (1994). Gibson & Wexler investigate parameter setting with the basic assumption that *triggering data* is required by a learner of a first language, namely data from a target language, which can only be analysed if the parameter that that data is a trigger for is correctly set.

Gibson & Wexler (1994) formalise the notion of two types of trigger:

1. A *global trigger* for value v of parameter P_i, $P_i(v)$ is a sentence S from the target grammar L such that S is grammatical if and only if the value for P_i is v, no matter what the values for parameters other than P_i are.
2. Given values for all parameters but one, parameter P_i, a *local trigger* for value v of parameter P_i, $P_i(v)$ is a sentence S from the target

> grammar L such that S is grammatical if and only if the value for P_i is v.

One of the difficulties associated with the above definitions is the consequence of *subset conditions* (Berwick 1985) which can occur. Suppose that two languages L_1 and L_2 differ by only one parameter value, and all of the sentences that are part of L_1 are also part of L_2, L_2 however has sentences which are not part of L_1. In these circumstances L_1 is a subset of L_2 as everything in L_1 is contained within L_2. The point is, that with the above trigger definitions, there cannot exist a trigger to set the parameter that distinguishes these two languages to the value which defines L_1 as such an S would also be grammatical in L_2. Negative evidence would be required to reject the hypothesis L_2, and this is generally regarded as not available to the learner.

Gibson & Wexler imply that it is generally accepted in the (psycho)linguistic literature that a learning algorithm which relies on the existence of triggers is assumed. They give the Trigger Learning Algorithm[1] (TLA) as being a natural choice:

> The Trigger learning algorithm. Given a set of initial values for n binary valued parameters, the learner attempts to syntactically analyse an incoming sentence S. If S can be successfully analysed, then the learner's hypothesis regarding the target grammar is left unchanged. If, however the learner cannot analyse S, then the learner uniformly selects a parameter P (with probability $1/n$ for each parameter), changes the value associated with P, and tries to re-process S using the new parameter value. If analysis is now possible, then the parameter value change is adopted. Otherwise the original parameter value is retained. (Gibson & Wexler 1994)

The constraint that makes the TLA conservative, in that it only considers grammars that differ from from the current hypothesis by one parameter, is called the Single Value Constraint (Clark 1990). And the constraint that requires the new hypothesis to give an analysis of the current sentence is referred to as the Greediness Constraint (Clark 1990). The algorithm is greedy in the sense that it is only prepared to change its current hypothesis if it can gain something from it (namely an analysis for the current sentence). Gibson & Wexler prove that this algorithm with the above constraints guarantees that a learner can converge to a target grammar in the limit (i.e. given enough data) with input that consists of positive examples from this grammar when every grammatical hypothesis has a local trigger associated with it.

[1] see also Sections 2.2 & 3.2.

Gibson & Wexler note that the way in which the algorithm selects a parameter to change by random choice, is possibly a simplification of what a real learner would do. The learner may be able to use partial analysis to determine which parameters are viable candidates to alter, or may have sufficient linguistic knowledge to weight the distribution with which parameters are selected to be altered. However, the resulting outcome of such changes would still allow the convergence to occur.

The obvious question to ask is "how realistic is it to suppose the existence of local triggers for every grammatical hypothesis, as required for convergence?"

The simple answer seems to be: "not very realistic". Gibson & Wexler show that a simple three–parameter system (also used by Niyogi & Berwick, and which will be used for further development), can have grammatical hypotheses other than the target grammar with no local triggers at all. That is, if a learner acquires such a hypothesis at any stage of learning then they will not be able to eventually acquire the correct target grammar as their current hypothesis, because there exist no triggers in the sentences they hear which will allow the correct resetting of wrong parameters. States representing these no-return hypotheses are referred to as local maxima. This problem can be overcome by various means: by setting the initial hypothesis appropriately, or by imposing an ordering on the parameters themselves, removing one of the above mentioned constraints. It is however unclear exactly how appropriate this is with our current knowledge, and the problem could be with the TLA framework itself not reflecting Universal Grammar properties suitably well.

2.2 Dynamical

Niyogi & Berwick (1995) propose a model that shows language change to be a dynamical system. Their model of language change is built around a parameter–based theory of language acquisition.

The basic scenario they assume is as follows: A language learner sets out to acquire the target grammar of his peers. He is exposed to primary linguistic data, the utterances of his peers. What this PLD consists of will be dependent upon the grammar(s) of his peers. It is assumed initially that all these peers produce utterances according to a particular target grammar, and the distribution of utterances heard by the learner reflects this. After processing a finite number of utterances the learner will accurately acquire the target grammar of his peer group. If a whole generation of learners is considered and not just one in isolation, then ideally all of these learners would accurately acquire the grammar of the peer group. These learners would then become mature and then themselves produce utterances that would form part of the PLD for the next generation of learners.

In this ideal situation the distribution of the PLD does not change from one generation to the next, as all of the population produce utterances from the same target grammar which all the learners acquire. However, if a small proportion of the learners misconverge and acquire a slightly different grammar from the original target, then the distribution of the PLD for the next generation of learners is altered slightly to reflect this. Niyogi & Berwick show how this affects successive generations of learners, and produces diachronic change in the language of the population as a whole. Looking at this kind of scenario on a discusive level is nothing new. However, what Niyogi & Berwick have done is to show the mathematical consequences of such an approach.

Logistic S-shaped curves are a talking point in the field of evolution and change, in both the contexts of Biology and Linguistics. Some models of language evolution, specifically (Kroch 1989), have shown these curves in their results, reflecting the data found historically. Niyogi & Berwick also show this behaviour. However, Niyogi & Berwick claim that they *do not* impose this behaviour on their system like other models do. That is, they claim that other models incorporate the logistic nature directly into the model, usually in the form of a probability distribution, and then show that it can be used to model language change. This is specifically true in the case of the Kroch (1989) who proposes a logistic model based on the analysis historical data. Niyogi & Berwick however show that a mathematical acquisition based model of language change can exhibit logistic behaviour due to the underlying nature of the dynamical system. It is also shown that the system need not necessarily show such behaviour, and they show that the resulting change need not even be monotonic.

Niyogi & Berwick assume that language change is a logical consequence of specific assumptions about:

1. the grammar hypothesis space

2. the language acquisition device

3. the primary linguistic data

The formal acquisition framework that they propose is as follows:

G is taken to be a set of possible target grammars and $g_t \in G$ is the target grammar in question.

P is taken to be the probability distribution defining the PLD, that is P determines the likelihood with which any particular grammatical construction will be next heard by the learner. If the population is homogeneous and all speak the target language governed by the target grammar g_t then all constructions not found in $L(g_t)$, the language defined by g_t, will have probability zero. If the population is non-homogeneous then P will reflect this, and be the composite distribution of the individual distributions that represent the individual languages, weighted accordingly to the proportion of the population that speaks them.

A is taken to be a learning algorithm that maps D_n a sequence of n utterances, drawn from the population with distribution P, onto a hypothesis grammar $h_i \in G$ with probability p_i

The resulting dynamical system is then defined as follows:

A state space S defines the set of all possible linguistic compositions (i.e. the different possible combinations of proportions of the population speaking each language) of the population. $s \in S$ is defined by a distribution P_{pop} on G which describes the proportion of the population that speaks each of the languages $g_i \in G$

An update rule f is then derived from the learning algorithm A and the PLD distributions P, to take the system from one state $s_t \in S$ to the next state s_{t+1}.

If the state of the population at generation t is $P_{pop,t}$, that is the linguistic composition of the t-th generation of speakers is defined by the distribution $P_{pop,t}$, then the probability that the next utterance that a learner will hear is the utterances w is:

$$P(w) = \sum_i P_i(w).P_{pop,t}(i)$$

That is, the sum over all languages of the probability of the utterance being produced by a speaker of a particular language multiplied by the proportion of the population speaking that language.

The update rule is defined by calculating the probability

$$p_n(h_i) = Prob[A(d_n) = h_i]$$

(The probability that a learner develops an arbitrary hypothesis h_i after n examples. The actual details of the calculation being dependent on the particular learning algorithm in question.) Assuming that a learner matures after n examples and becomes part of the next generation of speakers, then a proportion $p_n(h_i)$ of the population in next generation will have grammar h_i.

The update rule

$$f : P_{pop,t} \longrightarrow^{A} P_{pop,t+1} \qquad \text{is then}$$

$$P_{pop,t+1}(h_i) = p_n(h_i) \qquad \forall h_i \in G$$

All of the above is true for any choice of G, A and P_i for which the outlined constraints hold. That is the above argument does not assume any particular linguistic theory, learning algorithm or distribution of sentences to produce a dynamical system.

Niyogi & Berwick then demonstrate this theory using a parameterised grammatical theory, a uniform distribution of unembedded sentences generated by g_i, and the TLA (Trigger Learning Algorithm).

The TLA starts with the hypothesis grammar's parameters set randomly, and on receiving an example utterance, it attempts to parse it with its current hypothesis. If this fails it is allowed to flip one parameter setting at random, but only if the new hypothesis after the flip results in a new grammar which can parse the utterance. This procedure then repeats until maturity (after n utterances have been processed).

Algorithms of this type can be modelled exactly by a Markov chain, where each possible grammar is represented by a state in the model. The model changes state by following transitions, which have probabilities associated with them. There are transition probabilities for transitions from one state to another and from a state to itself. The transition from a state to itself corresponds to a utterance being parsed by the current hypothesis. Take two possible grammars g_1 and g_2, represented in the model by states s_1 and s_2 respectively. If the parameter settings of g_1 and g_2 differ by more than binary value then the transition probability for the transition from s_1 to s_2 and for the transition from s_2 to s_1 will be zero, corresponding to only being allowed to flip one parameter setting. If all the utterances that g_1 produces can also be produced by g_2, then the transition from s_2 to s_1 will be zero because if g_2 is the current hypothesis, there are no utterances that can be parsed by g_1 that cannot be parsed by the current grammar. Other transition probability values depend on the distribution of the utterances produced by the population. In fact, the specifics of the model are completely reliant on the parameterisation of the grammar space that is employed and the resulting distribution of the utterances that are produced, this being dependent upon the current population mix.

The Markov chain technology enables the probability of being in a particular state at a particular time to be calculated. Whether or not the model is convergent to a particular state, representing convergence to a target grammar, can also be calculated. This provides the probabilities needed to calculate the update rule stated above. More details of the TLA including the use of matrices of transition probabilities are given in Section 3.2.

The logistic behaviour talked about earlier can be seen in changes over time in the proportion of the mature population with a particular grammar or with a grammar that includes certain constructions, i.e. speaking one of a specified number of grammars. The logistic behaviour is most notable when a particular language trait takes over from another in the majority of the population.

Niyogi & Berwick present various examples of the model with differing numbers of parameters, slight alterations to the TLA and different maturing times. The fact that the technique does not depend on any particular set of parameters, means that it is difficult to evaluate its linguistic accuracy, and possibly wrong to. However, what it does allow is the ability to show how a particular set of parameters behaves, or how different aspects of different sets of parameters behave under the same circumstances. This provides a very strong framework for further development.

2.3 Complexity

A closer look will now be taken at some of the current issues from the viewpoint of parsing complexity.

To gain some background information, we first take a look at Hawkins' (1994) idea of Early Immediate Constituent Recognition. The human parser's preference to gain as much constituency information in the shortest possible time is expressed as 'Early Immediate Constituent' (EIC) recognition by Hawkins (1994). That is mother nodes are built above syntactic categories as soon as the node's presence is guaranteed by the input (Kirby 1996) (i.e. the mother node is constructed at the earliest opportunity, and the parser does not wait until all of the constituents below it have been fully constructed). Syntactic categories which allow the mother node to be built are referred to as *Mother node constructing categories* (MNCCs).

Mother node construction occurs as soon as a MNCC is encountered. Other constituents are then attached to the mother node as soon as possible (including those dominated by the mother node, that were encountered before the mother node could be constructed). See Hawkins (1994) for a more detailed account of MNCCs.

Hawkins' EIC metric allows a measure of parsing difficulty to be calculated.

The Constituent Recognition Domain (CRD) for a mother node M consists of the sequence of nodes (terminal and non-terminal) that must be parsed in order to recognise M and all of its ICs, starting with the first terminal of the first IC on the left, to the first terminal of the last IC on the right.

For example, the CRD for the mother node VP in Example 2.1. is highlighted (Hawkins 1994)

(2.1) I $\underbrace{\text{gave the valuable book that was extremely difficult to find to}}$ Mary

the terminal 'gave' constructs the mother node for the VP, 'the' constructs the NP IC, and only when 'to' is parsed seven words later can the PP IC be constructed.

The idea of the EIC metric is to provide a measure to quantify the preference for shorter CRDs, and to give CRDs with earlier constituency information, in the left to right parse, preference over ones with later constituency information when their length is equal.

The EIC metric uses the IC–to–non-IC ratios. The IC–to–non-IC ratio for a CRD is just the number of the ICs in that domain, divided by the number of non-ICs (words). For a whole sentence, then, the aggregate of the ratios for all the CRDs contained within that sentence is the measure for that sentence.

From this metric Hawkins defines the principle of Early Immediate Constituents:

"The human parser prefers linear orders that maximise the IC–to–non-IC ratios of the CRDs."

This principle is reflected in the the frequency of left or right branching languages spoken in the world compared to those with inconsistent branching properties. It is worth noting that the existence of such real data does not necessarily show that the assumption of the EIC principle is correct, as any languages in a sample of languages looked at are not necessarily independent of each other, as they have common roots. For the current discussion and work we do however make the assumption that the EIC principle, or something akin to it, holds.

Of specific interest to this project is the phenomenon that there is a skewing in the world's languages of subject and object order, (we will use this observation later as a simple example complexity metric). There is a definite bias towards subject appearing before objects, irrespective of verb position. Hawkins (1994) gives the following statistics (Tomlin 1986). From a sample of 402 languages, 168 (42%) have SVO order, 180 (45%) have SOV order, 12 (3%) have VOS, 5 (1.2%) have OVS and the remaining 37 (9%) have VSO. That is that over 95% of the sample show word order where S precedes O. Hurford (1990, pp. 328–339) accounts for this by the EIC's predictions based on the average lengths of the S, V and O, and shows that constructions where S precedes O do indeed result in better EIC scores.

Kirby (1996) then asks the question of how such a principle inherent to the human parser can affect the distribution of languages found in the world. His answer is that languages can adapt, and that this adaptation is effected by linguistic selection.

Kirby's (1996) argument takes on board the notion that PLD is filtered in such a way to produce triggers. That is that not all linguistic data that is heard by a child learner is actually used by the child as a trigger in their acquisition of language (Lightfoot 1991), and that parsing preference itself plays a role in the cycle of language acquisition an use.

Kroch (1989), as mentioned earlier, discusses historical data showing with two competing linguistic forms, how one can replace the other, proposing a simple logistic model to account for his findings:

$$(2.2) \qquad p_{f_1}(t) = \frac{e^{k+st}}{1+e^{k+st}}$$

p_{f_1} = frequency of f_1.
k = initial frequency.
s = slope parameter.
t = time.

This shows that replacement through competition can indeed bring about syntactic change.

Kirby (1996) then ties in this notion of competing variants with a performance theory, such as Hawkins' by proposing a model and demonstrating it by computer simulation. Kirby is attempting to use processing pressures to explain cross–linguistic distributions and the existence of language Universals. The model consists of the following elements:

- Simple types or features found in the language, (e.g. SVO).
- Arena of Use(Hurford 1990): A pool from which utterances are drawn.
- Speakers: These speak with a grammar[2] and produce utterances which contribute to the pool.
- Acquirers: Speakers who have not yet acquired a grammar, and take input from the pool.

Two dynamic processes take place within the framework of the model.

1. Production: Speakers contribute from their grammar to the Arena of Use
2. Parsing/Acquisition
    - The linguistic data forming the input of each learner is taken from the Arena of Use
    - Triggers are produced through a filtering process. The probability that an utterance forms a trigger is related to its frequency and its pre-defined parsing complexity.
    - The triggers are added to an individual acquirer's grammar

[2] In this context a grammar is just a collection of utterances which are regarded to be produced by an unspecified underlying set of rules

The simulation is run with the same number of speakers and acquirers. Once the acquirers have acquired, the speakers are discarded and the acquirers become the speakers for the next generation.

An example is given with competing prepositions and postpositions, and basic VO order. The EIC metric favours prepositions in such a scenario because with postpositions the CRD for the VP stretches across the NP.

$$_{VP}[\underbrace{V_{PP}[NP\quad P]}] \qquad vs \qquad _{VP}[\underbrace{V_{PP}[P}\quad NP]]$$

Triggers are selected by the following equations:

$$\begin{aligned} p(prep) &= \tfrac{1 \cdot n_{prep}}{1 \cdot n_{prep} + 0.79 \cdot n_{postp}}, \\ p(postp) &= \tfrac{0.79 \cdot n_{postp}}{0.79 \cdot n_{postp} + 1 \cdot n_{prep}}. \end{aligned} \tag{2.3}$$

where $p(f)$ is the probability that form f will become a trigger, and n_f is the frequency of f in the Arena of Use. The values 1 and 0.79 are the associated complexity values.

Kirby gives an example run of this simulation showing logistic behaviour. It can however also be shown mathematically to behave logistically, by formulating $p_f(t)$ the proportion of form f found at time t. We can derive for the more general case, where $0 < c_f < 1$ is the complexity associated with form f and $0 < n_f < 1$ is the initial proportion of form f, that $p_f(t)$ is of the form given in Equation 2.2. (See Appendix A. for proof.)

The framework is then extended to incorporate two competing factors, adposition order and VO order. Here the fitness of a particular form of one type (e.g. adposition) is made dependent on the frequency of the complementary form of the other type (e.g. VO order), and vice versa. Complementary in this context means that the fitness of these two forms together is maximal, (i.e. they are an optimal pairing with respect to fitness). Kirby shows how the average fitness of the population increases over time as the fitness landscape is climbed. The complexity introduced by the interdependence between types of form unfortunately means that mathematical analysis of this model is beyond the scope of this text.

2.4 Genetic

Clark & Roberts (1993) discuss parametric change from a genetic algorithm point of view. They 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. (Clark & Roberts 1993)

So when a learner hears data that is compatible with differing, conflicting parameter settings, its hypothesis will be evaluated using other criteria such as subset conditions with respect to the grammar space and elegance of actual derivations using different grammars. The linguistic environment then under-determines the learner's appropriate choice of grammar (this is the orthodox Chomskyan viewpoint), and the various factors of fitness are combined in deciding which hypothesis is appropriate.

The subset condition (Berwick 1985), provides that a learner must be able to know the minimal language compatible with a given input sequence of linguistic data, such that the problem of becoming stuck with an incorrect parameter settings where one language is a proper subset of another, (i.e. the acquired grammar will over generate) can be overcome. Clark (1990) observes a further complication in the *shifting–property* where the language subsets can occur due to parameter groupings, (i.e. where subset like conditions occur, though not as the result of one particular parameter, but as the result of the interaction between parameters). These problems are predominantly due to the fact that only positive linguistic data is received by the learner, and an incorrectly set parameter that does not contradict constructions of the target grammar is unable to be unset unless negative data is presented to contradict its setting.

A genetic algorithm (Clark 1990) is defined as having the following properties:

- A string representation of each hypothesis.
- Operators to combine existing hypotheses to produce new hypotheses
  - a cross over mechanism to combine two hypotheses to produce a new hypothesis.
  - a mutation operator that introduces a minimal random element into the combination.
- A fitness measure to judge hypothesis performance in an environment.

The fitness measure that Clark & Roberts suggest consists of three elements. The first is based around the ability to parse a given input. The parser is considered to be made up of modules (case, binding and X-bar theory are given as examples) and a count is made of the number of modules that are violated. A module is considered violated if its principles[3] are not adhered to by the input, i.e. when the input can only be parsed by going against the principles of that module. A count of violated modules contributes to the first component in the fitness measure.

The second element specifies that a subset hypothesis should be preferred over a superset hypothesis, where both are adequate to parse the input.

The third component considers the 'elegance' of the hypothesis, and that more compact representations will be preferred over less compact ones, all else being equal.

The mathematics for this goes something like:

Let n be the number of parsing devices and v_i be the number of violations of the i-th parsing device caused by the input.

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

represents the statistical fitness of a particular parsing device when compared to all others.

Let s_m represent the number of superset setting in hypothesis h_m [4] and let $0 < b < 1$ be a appropriate weighting factor. And let e_i be a measure of the general elegance of the parser P_i, this is a count of the nodes in the parse tree produced, and $0 < c < 1$ be another appropriate weighting factor, then adding three similar terms together gives the fitness metric:

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

Any particular sentence that a learner hears will require some, but not all of the parameters available to be set to specific values, for a well formed representation to be assigned by the parser. If a sentence σ requires parameter p_i to be set a particular way, then σ is said to express p_i. This is used as a definition for the notion of a 'trigger':

> A sentence σ is considers a trigger for parameter p_i if σ expresses p_i.

Clark & Roberts go on to show how parameter changes that have occurred in French, can be accounted for by such a model. One such change the discuss is the loss of null subjects mentioned in Chapter 1.

[3] Clark & Roberts use the term principles when referring to the rules that would be defined by a set of Chomskyan parameters and not in reference the the invariant Chomskyan principles

[4] It is not made how the learner comes by this kind of information, presumably it is regarded as innate

2.5 Overview

S-shape curves are an issue in the computer modelling of language change because the *logistic* behaviour that they represent has been shown to exist historically in the changes that languages undergo (Kroch 1989). As little is understood about the processes behind language change, producing behaviour that reflects historical data shows that a model may accurately reflect some property of language change, although it would be wrong to jump to the conclusion that a model is completely accurate just because it reflects this behaviour.

The main problem is that it is difficult to say whether a model incorporates such behaviour or just displays it. This is because, if it does display it, it then surely must incorporate it at some level—even if it is not apparently obvious from the formulation of the model, the behaviour is still there. A possible distinction that could be made is that Kroch's (1989) and Kirby's (1996)[5] have an explicit logistically-derived function, which contributes to the step calculating the next generation of the population. Niyogi & Berwick's model has no such step, and that is the claim that they are making. It is arguable that in the formulation used by Kirby, the model is not defined logistically, and it is only a mathematical consequence that the logistic function arises. This contrasts Kroch's model which is an explicit logistic model based on his study of historical data. A more complete mathematical treatment of Niyogi & Berwick's model may show that logistic behaviour is inherent to this model too–this is unfortunately well beyond the scope of this text. The outcome is that it is difficult to know if Niyogi & Berwick's claim is a valid one.

Within the rigorous mathematical framework of Niyogi and Berwick a fully specified grammatical hypothesis is *mutated* randomly only when linguistic data fails to be accounted for by this hypothesis. Clark and Roberts' framework however allows an under-specified hypothesis to be more fully specified by external factors influencing the fitness of constructions. Which of these approaches is more realistic is unclear.

The Niyogi and Berwick framework is difficult to fault because of its generality. They state clearly some of the deficiencies that filling in the gaps in the framework with particular theories introduces. The influence of factors not directly related to the parameterisation of the grammar are not considered by Niyogi & Berwick; in fact they are trying to showing that language change can come about because of the inherent internal properties of the model, but a theory where other factors affect the way in which language changes could be introduced into a particular learning algorithm (like

[5] as discussed in Section 2.3

the TLA).

One point of complication arises within the Niyogi and Berwick framework when the population is (or becomes) non-homogeneous. Here the population does not all have the same internal grammar, and there is no longer a single target grammar for a learner to acquire. Niyogi and Berwick do not make it clear how a learner's hypothesis will behave once this situation arises, specifically they do not clearly detail how the convergence behaviour of a learners hypothesis behaves with utterances from a collection of grammars.

Clark & Roberts' proposed fitness measure looks very promising but few details are put forward on how the formula that they put forward are used. It seems somewhat unreasonable that all of the information that the formulation requires is readily available to the learner, or could be efficiently used even it if were. The methodology suggests that the learner has a complete database of information about possible consequences of various parameter that it has, and that it tries to parse a given input with a selection of possible grammars. It is not made clear how large the set of hypotheses is that any particular hypothesis is compared to. It would be unrealistic to have a database of all possible hypotheses, but if only a few are present, they do not make it clear how reliable the method would be.

Clark & Roberts also give no details about how the subset component is calculated, but do comment that it would not account for the *shifting–property* across grammars where parameter settings combine to form subsets. This seems to imply that the model would fail in respect of avoiding subsets anyway. The elegance component is almost glossed over altogether, with a mention of counting the nodes in the derived parse trees produced by a particular parsing device. If and how the particular structure of a given tree is evaluated is not made clear.

The remainder of this dissertation considers the Niyogi & Berwick model in more detail. The next chapter concerns replicating the model as it stands and showing how it performs, and the following chapter looks at developing the model to incorporate a mechanism that considers parsing complexity.

Chapter 3

Replication

3.1 Parameters

The three parameter grammar discussed by Gibson & Wexler (1994) and used by Niyogi & Berwick and also adopted later, is based on a two parameter grammar which defines base word order (i.e the underlying word order in an utterance) and an additional parameter which causes verb movement when set to give a surface structure which differs from the original base structure.

The first two parameters define the base word order. We take the the following X-bar (see e.g. Haegeman (1991) for a review) schemata:

$$(X1) \qquad X' \rightarrow X, Complement,$$

$$(X2) \qquad XP \rightarrow X', Specifier.$$

Where rule $(X1)$ states that X(head) and $Complement$ are sisters, but says nothing about ordering. Similarly, rule $(X2)$ states that X' and $Specifier$ are sisters. Two binary-valued parameters are then employed to set the ordering. The first parameter *spec-first* when set to 1 defines a specifier-first language and when set to 0 a specifier-final language. Similarly the second parameter *comp-first* defines a complement-first language when set to 1 and a complement-final language when set to 0.

Figure 3.1. shows extracts of the affected nodes in the parse tree for each parameter setting, and Figure 3.2. shows the base word order SVO obtained by the parameter settings *spec-first*, *comp-final*.

A third parameter is now introduced. This parameter when set forces:

> a finite verb to move from its base position to the second position in root declarative clauses. (Gibson & Wexler 1994)

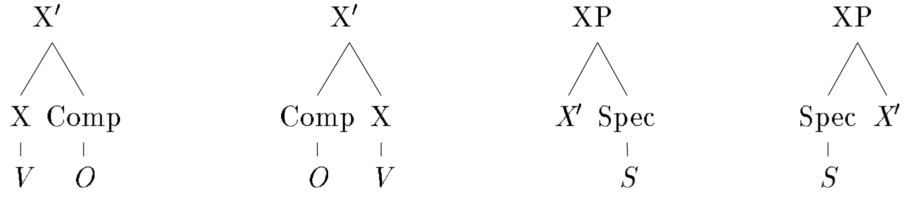
spec	*comp*	*verb mov.*	resulting grammar
0	0	0	VOS -V2
0	0	1	VOS +V2
0	1	0	OVS -V2
0	1	1	OVS +V2
1	0	0	VSO -V2
1	0	1	VSO +V2
1	1	0	SOV -V2
1	1	1	SOV +V2

Table 3.1. Parameter settings and their resulting grammars

This means that although the underlying base word order is unchanged the surface structure may be re-ordered. Theories about why verb movement occur are varied, (see Gibson & Wexler (1994) for more details), but the mechanism assumed here is that the verb moves to the C position. Figure 3.3. shows the movement involved. The movement of the verb requires that the CP/Spec position be filled, for theory internal reasons. Figure 3.3. shows this being filled by O, but it could equally be filled by S.

The grammar space defined by these parameters consists of 8 grammars made from 4 base word orders, each with or without verb movement. These grammars can be referred to by their base word ordering, followed by either $+V2$ or $-V2$ to signify the presence of verb movement. Table 3.1. lists the parameter settings along with its given name and Appendix B. shows the actual surface constructions which are associated with each grammar.

It is worth commenting that the assumption that only unembedded sentences form the triggers that are used by a learner in acquiring language is taken here. This issue is not discussed as it is an assumed part of the framework being employed. For a full detailed discussion and evidence to support this assumption see Lightfoot (1991).



comp-first = 0 *comp-first* = 1 *spec-first* = 0 *spec-first* = 1

Figure 3.1. Parts of the parse trees affected by the two base word order parameter

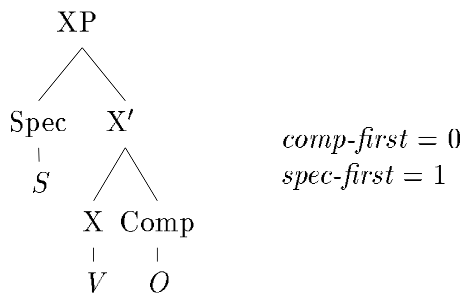


Figure 3.2. Example tree structure for the parameter setting producing the base word order *SVO*

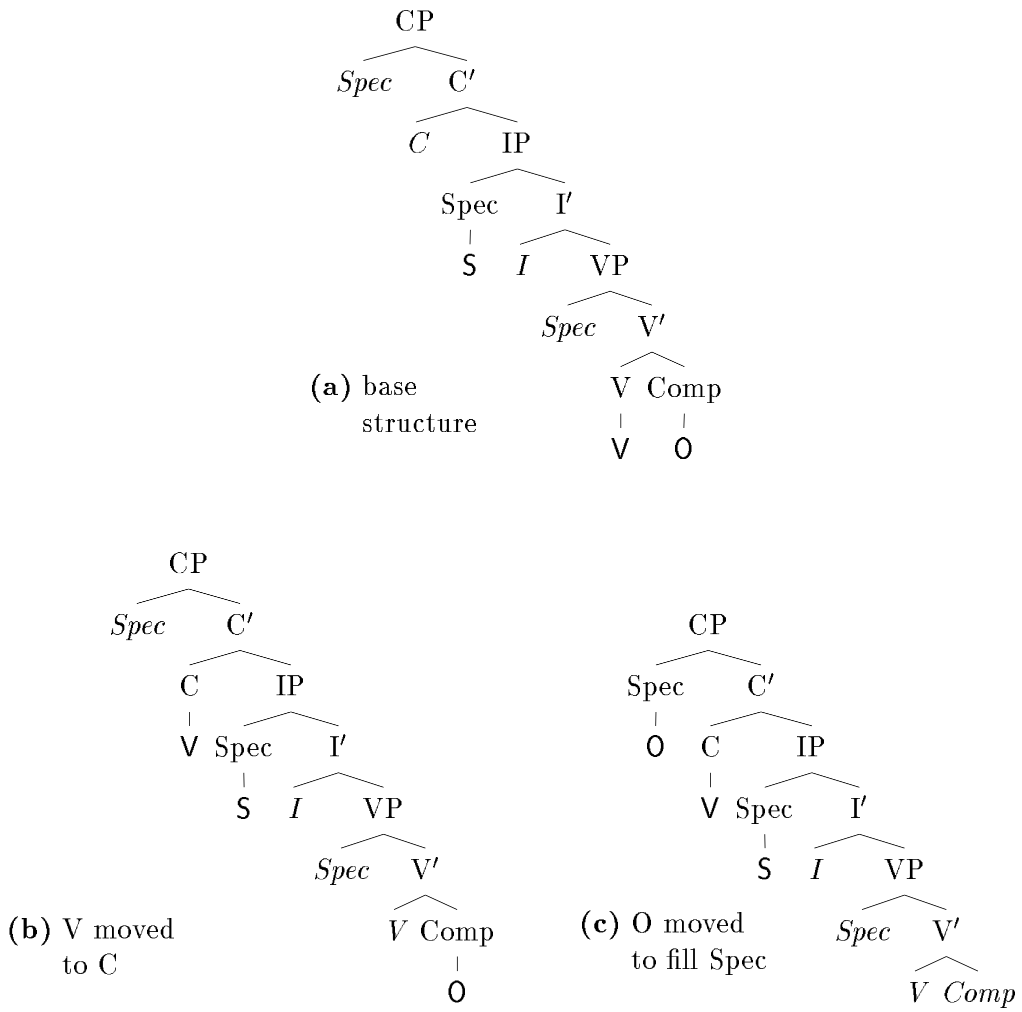


Figure 3.3. Tree structures showing the stages of verb movement: (a) shows the underlying base tree structure, (b) shows V moved to the C position, (c) shows O having moved to fill the CP/Spec position.

3.2 Methodology

To further evaluate the model proposed by Niyogi & Berwick and to provide the groundwork for further development, a computational replication of the model was produced. This was done mathematically and probabilistically by working with the transition matrices of the Markov process modelling acquisition, and the probabilistic distributions of the population each of which are outlined in Section 2.2. This method models the theoretical outcome of each step in the model in terms of proportions, instead of individually and randomly simulating a large number of learners hearing individual utterances. This both saves computational time, as it is effectively only modelling one *average* learner, as it is modelling the gross behaviour of the system, and hence not a large number of individual learners. Also this method removes the need to run a particular model many times to produce statistically significant results as the method produces the statistical outcome directly. This basic procedure follows:

1. Decide on parameters being used and define resulting languages. The three parameter grammar discussed in Section 3.1. was used here.

2. Set up the distribution for the initial population of speakers: Here each language (or parameter configuration) is assigned a value [0–1] which represents the proportion of the population which speaks that language. The main sets of initial conditions that were experimented with included: Homogeneous populations of each language, a mixed population speaking an equal proportion of each language and a mixed population speaking each of the -V2 languages.

3. Set up the primary linguistic data: Knowing the linguistic composition of the population, (from the previous step) each utterance that is producible within the defined grammars is allocated a value [0–1] to represent the probability that it could be the next utterance produced by that population.

4. Calculate the transition matrix: The transition matrix represents the Markov process modelling acquisition. The effect that hearing an utterance has on a learner is stored in the matrix as the probability of shifting from one hypothesis to another.

5. Calculate product matrix: The transition matrix is then raised to the power m (multiplied by itself m times[1]), where m is the number of utterances after which

[1] Matrix multiplication is defined as follows: If A,B and C are $n \times n$ matrices, and a_{ij} is the element in the ith row and jth column of A, then if $C = AB$, $c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$

a learner matures. This matrix then provides the information to show what proportion of a population of learners will acquire each language in the language space.

6. Generate the new population of speakers: The product matrix is then used to generate a new population of speakers for the next generation.

7. Return to Step 3

Transition matrix entries of non-neighbouring states are zero, otherwise entries are calculated by the following:

$$\begin{aligned} P[s \to k] &= \sum_{s_j \notin L_s, s_j \in L_k} \frac{1}{n} P(s_j), \\ P[s \to s] &= 1 - \sum_{\substack{k \text{ a neighbouring} \\ \text{state of } s}} P[s \to k]. \end{aligned} \tag{3.2}$$

L_s and L_k are the languages defined by the grammars s and k respectively. The transition from s to k is made up of a components relating to each utterance's frequency in the PLD ($P(s_j)$).

The three parameter grammar was used because these parameters provide interaction in that a surface utterance can be the result of the underlying word order or the result of verb movement. These parameters have also been fairly well discussed (Gibson & Wexler 1994, Niyogi & Berwick 1995, Niyogi & Berwick 1996*b*, Niyogi & Berwick 1996*a*). It is noted that they are too simplistic or completely inappropriate when compared to the real parameter system present in the human brain (if indeed such parameters do exist at all), but as the aim here is to look at the effect of parsing complexity issues within this type of parameterised framework, in comparison to systems which neglect it, their simplicity serves well. Unrealistic behaviour is also to be expected. The complete realism of such a simulation is not the primary issue here. As insufficient is known about the learning processes being modelled, it would be very difficult to be certain what completely realistic behaviour comprises anyway. What is being focused on here are the effects that including or excluding parsing complexity and fitness issues have on the outcome of the model.

The eight hypotheses in the three parameter scenario can be regarded as the corners of a cube. The greediness constraint in the TLA restricts the single step movement, from one hypothesis to the next, (that which may occur due to the processing of one utterance), to the transitions which follow along the edges of the cube. Figure 3.4. illustrates this. Assuming[2] that hypothesis 000 representing L_0 (VOS –V2), is on the front most face of the cube, then each face of the cube represents a particular parameter setting, (e.g. the front face represents the hypotheses which are *comp-final* and the top face represents the +V2 languages). This illustration clearly shows which transitions are possible due to the greediness constraint and which are not.

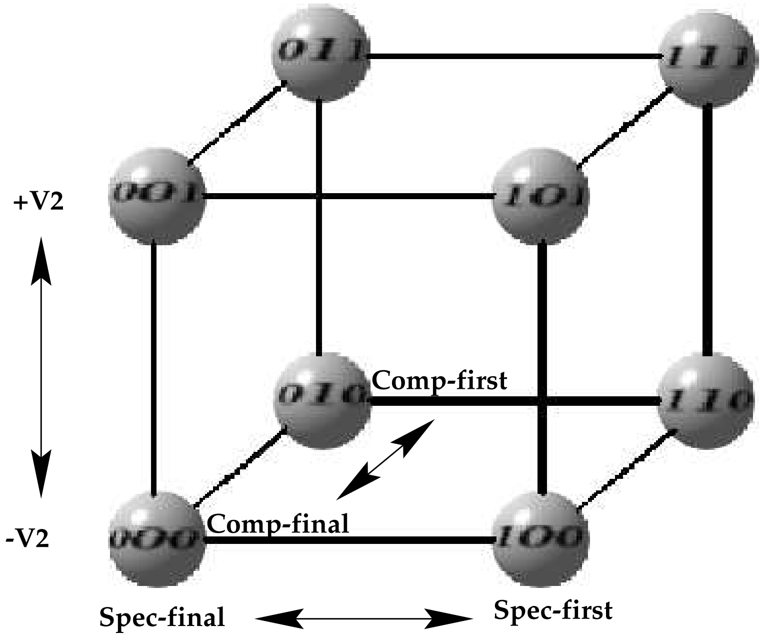


Figure 3.4. Hypotheses represented as the corners of a cube

An example single step transition matrix is shown in Figure 3.2. (Niyogi & Berwick 1996*b*) is that for a learner exposed to a homogeneous population speaking SVO –V2. The top row of this matrix holds the transition probabilities associated with a transition from hypothesis L_0 (VOS -V2, corner 000 on the cube) after hearing one utterance. The cube shows that the only possible transitions are to L_1 (001) by setting +V2, L_2 (010) by setting *comp-first*, L_4 (100) by setting *spec-first*, or this hypothesis can be kept and no transition taken. The top row of Figure 3.2. shows non-zero probabilities only for transition to L_1 (VOS +V2), and L_4 (SVO –V2), and of course remaining in L_0. This means that there are no triggers produced by the SVO -V2 speaking population to set the *comp-first* parameter. This is hardly surprising as SVO -V2 is a

[2] optical effects when viewing such diagrams and all...

comp-final grammar without verb movement. The other rows of the matrix represent the transitions from the other hypotheses. The row representing transitions from L_4, shows that there are no transitions away from L_4; this is because all the utterances that a learner is exposed to can be accounted for by this grammar, and hence once a learner has this hypothesis it is never rejected.

Figure 3.3 shows the resulting matrix when the single-step transition matrix in Figure 3.2 is multiplied by itself sufficient times for the values to converge; this is equivalent to a learner being exposed to sufficient utterances that hearing any more would make no difference to his hypothesis. Again the first row represents the transition probabilities from L_0. The row shows that if a learner starts with hypothesis L_0, after hearing many utterances he/she will have hypothesis L_1 with probability 1/3 and L_4 with probability 2/3. Equivalently, if a large number of learners started off with hypothesis L_0 then 1/3 would end up with L_1 and 2/3 with L_4. As the learners in the model start with a random hypothesis, there will be an equal proportion of learners with each initial hypothesis. So if we sum the values in a particular column and divide by the number of rows, we can see what proportion of the population would acquire a particular hypothesis. In the example there are only values in two columns (L_1 and L_4). This means that all learners exposed to enough L_4 utterances will either take L_1 or L_4 to be their final hypothesis. The probability of taking L_1 is $(1/3+1+3/5+1)/8 = 11/30$ and the probability of taking L_2 is $(2/3+2/5+1+1+1+1)/8 = 19/30$, so approximately 2/3 of the learners will acquire the target grammar L_4.

Transition to

Transition from

$$\begin{array}{c} \\ L_0 \\ L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \\ L_6 \\ L_7 \end{array}
\begin{array}{c}
\begin{array}{cccccccc} L_0 & L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_7 \end{array} \\
\left(\begin{array}{cccccccc}
\frac{1}{2} & \frac{1}{6} & & & \frac{1}{3} & & & \\
 & 1 & & & & & & \\
 & & \frac{31}{36} & \frac{1}{12} & & & \frac{1}{18} & \\
 & \frac{1}{12} & & \frac{11}{12} & & & & \\
 & & & & 1 & & & \\
 & & & & \frac{1}{6} & \frac{5}{6} & & \\
 & & & & \frac{5}{18} & & \frac{2}{3} & \frac{1}{18} \\
 & & & & & \frac{1}{12} & \frac{1}{36} & \frac{8}{9}
\end{array}\right)
\end{array}$$

Table 3.2. Transition matrix for a learner, assuming a population of speakers of SOV –V2. This matrix shows the probabilities relating to the processing of one utterance.

3.3 Outcome

The simulation was then run with a selection of initial conditions. Initially those discussed by Niyogi & Berwick were used to check that the simulation was in fact producing the expected results and then a wider range of initial conditions was looked at.

The effect of maturation time affects language change quite considerably. A very low maturation time (i.e. after only two or thee utterances) results in approximately equal proportions of all languages being spoken in the next generation, with only a slight bias from the distribution of speakers in the current population. This results from an almost random choice of final hypothesis from an individual learner, this is what you would intuitively expect as a learner's initial hypothesis is chosen at random and very little data is being presented to support or oppose this, so it, or something close to it ends up as the final hypothesis, and there is no time allowed for convergence to a particular grammar to occur.

Niyogi & Berwick only briefly comment on how the model behaves over successive generations, and show that under certain conditions the S-shape logistic curves arise.

A more detailed study of the behaviour of the model using these three particular parameters shows that the *clear cut* logistic behaviour which Niyogi & Berwick show is in fact a rarity and exponential behaviour is more often the outcome observed as shown below.

Figures 3.5–3.7. show how maturation time has an effect from various initial conditions. Figure 3.5. shows the situation where initially the whole population speaks VOS -V2. The result is that eventually VOS +V2 takes over (Niyogi & Berwick 1995).

$$\begin{array}{ccc} & & \textbf{Transition to} \\ & & \begin{matrix} L_0 & L_1 & L_2 & L_3 & L_4 & L_5 & L_6 & L_7 \end{matrix} \\ \textbf{Transition from} & \begin{matrix} L_0 \\ L_1 \\ L_2 \\ L_3 \\ L_4 \\ L_5 \\ L_6 \\ L_7 \end{matrix} & \begin{pmatrix} & \frac{1}{3} & & & \frac{2}{3} & & & \\ & 1 & & & & & & \\ & \frac{3}{5} & & & \frac{2}{5} & & & \\ & 1 & & & & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \\ & & & & 1 & & & \end{pmatrix} \end{array}$$

Table 3.3. Transition matrix for a learner, assuming a population of speakers of SOV –V2. This matrix shows the probabilities relating to the processing of enough utterances for convergence to occur.

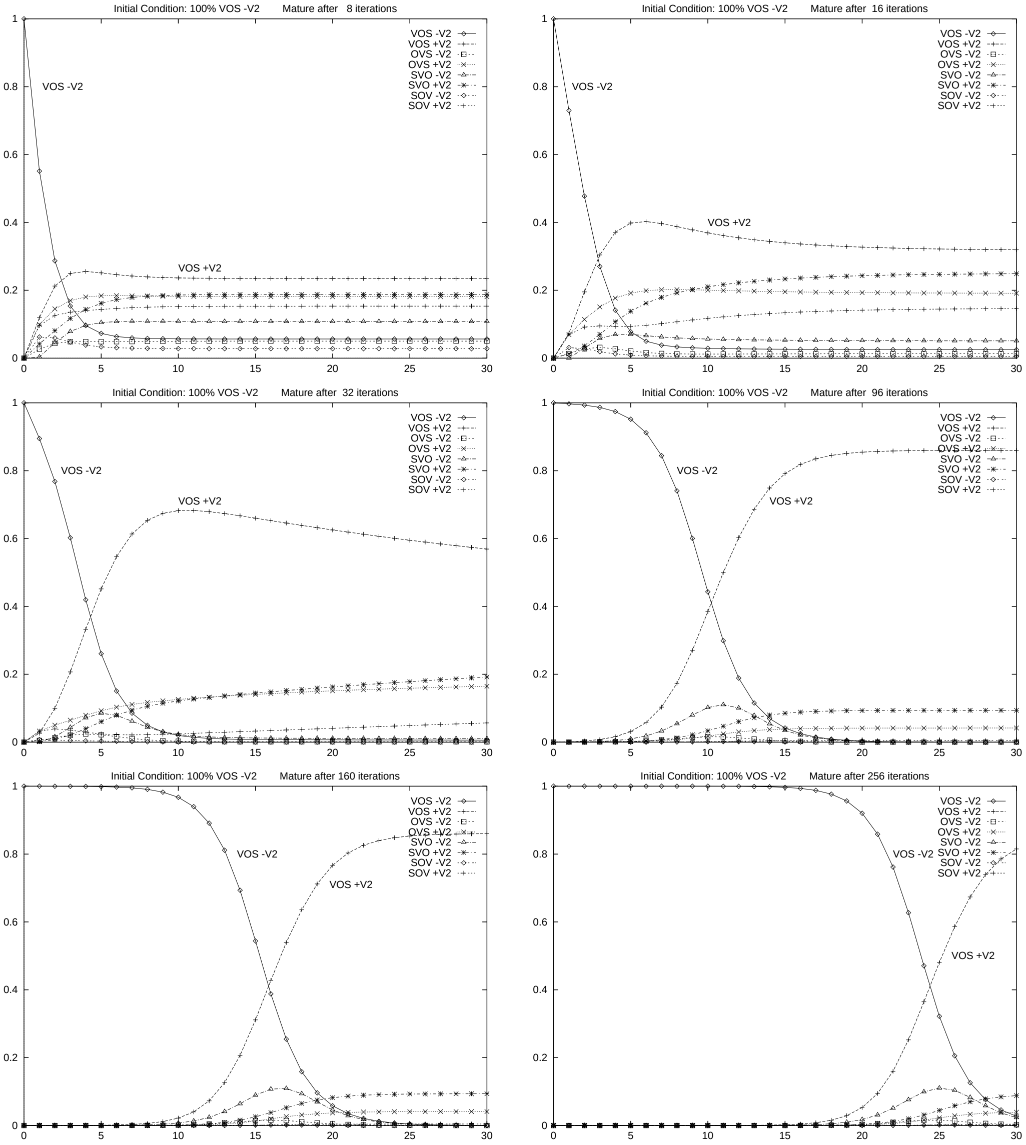


Figure 3.5. Simulation plots showing the effect of altering maturation time from 8 to 256 utterances over 30 generations. Initially population is speaking a VOS –V2 language. Plots show how +V2 language takes over.

The rate at which it does so is dependent on the maturation time, a longer maturation time means at each generation a larger percentage acquire VOS -V2 successfully, that with a shorter maturation time, but a small proportion will still acquire VOS +V2, and this will eventually take over as the prominent language. The later plots clearly show the logistic S-shape curve as –V2 becomes taken over by +V2. In contrast Figure 3.6. shows the behaviour when the initial population speaks only VOS +V2. Here the more random acquisition effect can be seen, when learners do not get the chance to hear enough utterances to converge to any particular grammar, and the population all successfully acquire the +V2 language when the maturation time is appropriately high. It is interesting to note that there is a hierarchical effect—there is a definite order to proportions of the population speaking each language. Compare the graph showing maturity after eight utterances in Figure 3.6. with the graph showing maturity after thirty-two utterances in Figure 3.7. The ordering which is arising in the latter case is

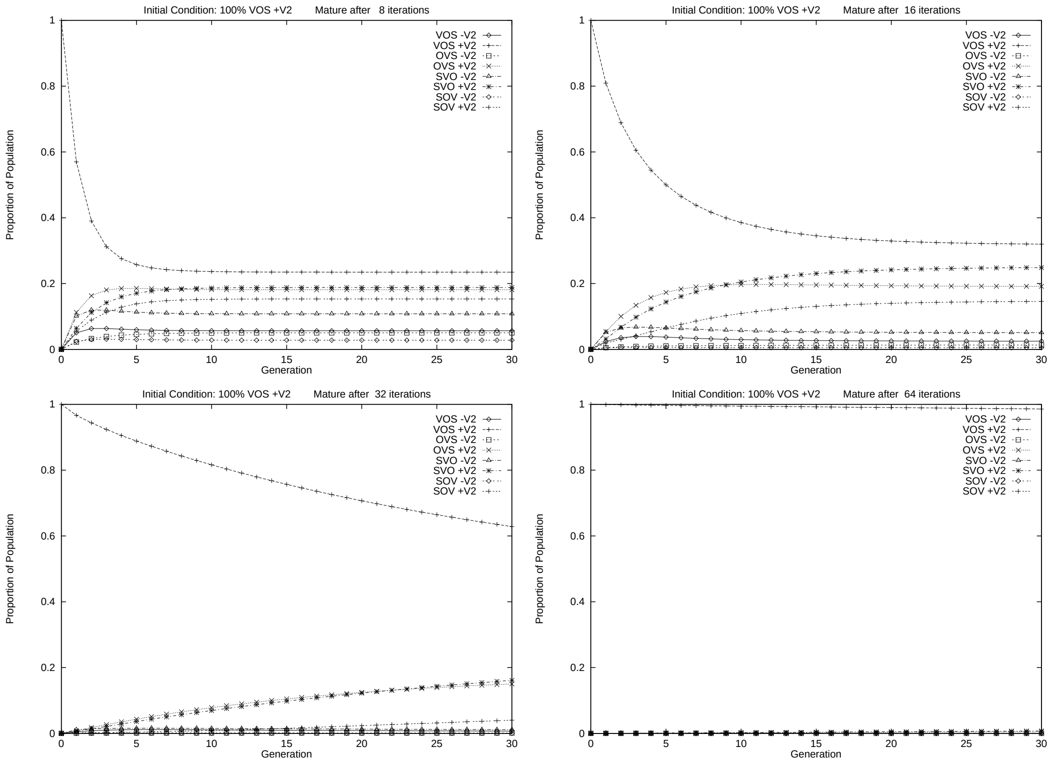


Figure 3.6. Simulation plots showing the effect of altering maturation time from 8 to 64 utterances over 30 generations. Initially population is speaking a VOS +V2 language. Plots show how +V2 language stays prominent

the same as that which is initially present in the former. This suggests that a frequency ordering is imposed on the grammars due to the way different grammars can account for particular utterances.

One clear observation is that these particular parameters favour +V2 languages. The +V2 languages in this parameterisation have more constructions than −V2 languages and the organisation of the utterances is such that it is easier to set the verb movement parameter than it is to unset it, that is on average there are more triggers to set the verb-movement parameter than to unset it. This phenomenon could be a direct result of the parameters being used and the fact these parameters do not reflect the real world, or that other factors not accounted for by the model, such as the distribution of utterances which are produced within any one language, meaning that this model alone is too simple to reflect the real world.

Figure 3.7. where the initial population is composed of a equal mix of each −V2

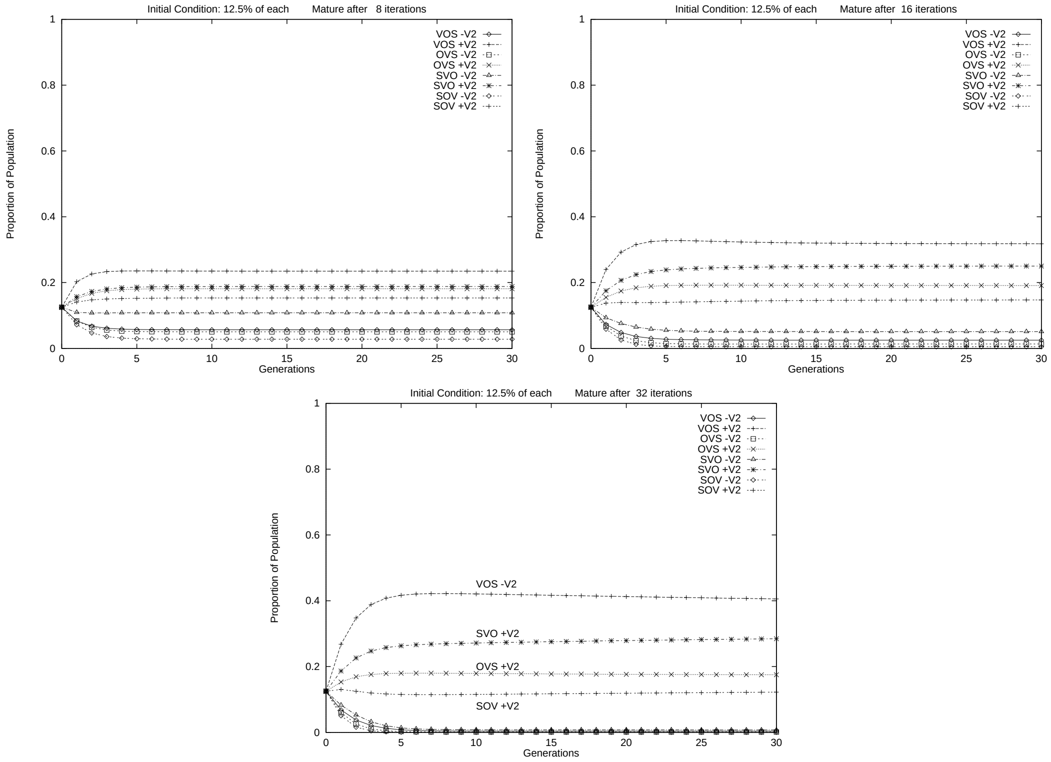


Figure 3.7. Simulation plots showing the effect of altering maturation time from 8 to 32 utterances over 30 generations. Initially population is speaking all 8 languages. Plots show how +V2 languages become prominent

language shows the difficulty for a –V2 language to survive at all, even when given a reasonable maturation time. This can also be shown with simulations with homogeneous initial populations for each –V2 language. This evidence is by no means exhaustive but the general behaviour of +V2 prominence does seem to be the rule rather than an exception.

3.4 Assumptions

There are many factors which could be affecting the results of this simulation, some related to the modelling framework, some to the implementation, and some due to the initial conditions and distributions chosen for the simulations:

1. Utterance production within a language is uniform. The model assumes that the utterances produced by a speaker of a particular language, are uniformly distributed across the utterances which are part of that language. That is if a and b are both constructions of language L, that a speaker of L is equally likely to produce the construct a as he is the construct b. It is unlikely that all the constructions in a language are used to the same extent.

2. No overlap in generations. In the scenario presented here and in that presented by Niyogi & Berwick, a generation of speakers is replaced by the new generation of learners with their newly acquired languages. This is a very simplistic view of real life where there is no such generation distinction, and new learners are born across a continuum and not in specific batches representing a particular generation.

3. An interesting methodological point is that of the accuracy of the calculations. Many of the rational fractions which contribute to the calculations cannot be represented as finite binary fractions by the computer running the simulation, and the values stored are in fact only accurate to about 16 decimal places. The implications of this can be demonstrated by asking the computer to print the value 1/3 to more than a few decimal places. The result as a double-precision floating point value is 0.333333333333333314829616256247390992939472198486328125 which is clearly accurate to only 16 decimal places. Worse still the computer equates $(1/3)^{11}$ to zero, which is the sort of calculation which could occur with a learner processing 11 utterances. However, as $(1/3)^{11} = 0.00000564502926\ldots$ which is close to 0, this is not as bad as it first seems.

Unfortunately both the calculating of the error bounds in the simulation, which would show the limit to which the results are correct, or implementing the simulation with arbitrary precision arithmetic—to eliminate the problem altogether, are beyond the scope of this work. The outcome of this situation is that the transition matrix probabilities could be inaccurate in that rounding towards 0 or 1 could occur.

In some sense this type of quantisation error could be regarded as realistic, where a population would consist of a finite number of people, and the proportion of learners acquiring a particular language would be rounded to the nearest whole person. It should be noted however that the possible current inaccuracies in the model should not be considered the right way to go about modelling this quantisation effect.

4. Initial parameter settings for learner's are random. It is assumed that a learner's parameters have no default values. If they did it may be possible that some cases where convergence to a target grammar fails could be overcome.

5. Spatial Organisation. No account is taken in this model of spatial organisation. That is each learner is exposed to the whole population. A more realistic situation would where the population is arranged spatially in two dimensions, and each learner only hears utterances that are produced nearby. Or more mathematically a 2-dimensional Gaussian distribution can be used to weight the utterances produced, so that the closer a speaker is to a learner, the more his/her utterances contribute to that learner's acquisition. Kirby (1996) and Oliphant (1996) illustrate the profound differences that spatial organisation can make on a model.

3.5 Summary

In duplicating the work done by Niyogi & Berwick some of the details of the TLA and an overall perspective on the model are seen. The randomness present in acquisition with a very low maturation level is clearly visible, and it can be seen that maturation levels over 50 only delay the eventual outcome of the system and do not change it.

The advantage that the system gives to +V2 languages also stands out. A possible explanation for this is that parsing utterances with a +V2 grammar is more complex than parsing the same utterance with a −V2 grammar. The model however does not take this into consideration, and as the +V2 languages account for a wider range of utterances, they are more likely to survive than the −V2 languages. This clearly illustrates one disadvantage of a model which only accounts for one linguistic approach.

Chapter 4

Development

The next step is to show how issues such as parsing complexity can affect the behaviour of such a system and to see if they can be of help in overcoming some of the problems with the previous model. In the spirit of the original model an approach is taken that will not insist on a specific set of parameters, but provide the framework for any appropriate set of parameters to be used.

4.1 A first Attempt

With the model as it stands the current hypothesis is only rejected when it cannot account for the current utterance. What is required to take issues such as parsing complexity into account is a mechanism that allows the current hypothesis to be rejected if an acceptable parse of the current utterance is deemed unfit (i.e. too complex). The ability to account for the current utterance should not be considered enough to keep a hypothesis, the hypothesis should be able to account for it efficiently.

To do this some additional framework is required: Let F_k be a fitness function on the strings of language k, such that:

$$F_k \colon L_k \longrightarrow [0,1]$$

That is, each utterance of a language is assigned a complexity score, 0 meaning not at all complex up to 1 meaning most complex[1], with intermediate values representing intermediate complexities. It should be noted that a construction that appears in more than one language, (i.e. can be generated by different parameter settings) may have a

[1] This formulation is in terms of complexity, and complexity will be minimised, which is equivalent to climbing a fitness landscape

different complexity with respect to each language. That is, an utterance's complexity may be dependent upon its underlying grammatical construction. For example, both grammars L_1 (VOS +V2) and L_4 (SVO −V2) can account for the utterance S V O1 O2, but each grammar may have a different complexity score associated with it. For instance: $F_1(\mathsf{S\ V\ O1\ O2}) = 0.2$, $F_4(\mathsf{S\ V\ O1\ O2}) = 0.6$.

These complexity scores can then be used to measure the probability with which a hypothesis should be rejected if the current utterance being processed can be accounted for, (i.e. utterances with a complexity score greater that zero with respect to the current hypothesis may result in the current hypothesis being rejected). The higher the complexity score is the more chance there is of it causing the current hypothesis to be changed. A complexity value of 1 will force the current hypothesis to be rejected on receiving this utterance. Formally, the transition probabilities associated with a state s are altered in the following way:

Previously, for a n parameter system, the transition matrix was calculated from Equation 3.2. Adding a such rejection factor relating to parsing complexity gives the equations:

$$\begin{aligned} P[s \rightarrow k] &= \frac{1}{n}\left(\sum_{s_j \in L_s} F_s(s_j)P(s_j) + \sum_{s_j \notin L_s, s_j \in L_k} P(s_j)\right), \\ P[s \rightarrow s] &= 1 - \sum_{\substack{k \text{ a neighbouring} \\ \text{state of } s}} P[s \rightarrow k]. \end{aligned} \tag{4.1}$$

Here the transition from state s to k is additionally taken when the current utterance can be accounted for L_s with probability $F_s(s_j)$. The behaviour resulting from adding this mechanism can now be demonstrated with the three parameter system used earlier and a very simplified complexity factor. This heuristic metric states utterances with SO ordering are less complex than OS ordered utterances. A much more detailed study of SO and OS orderings and complexity, and why SO is preferable is carried out by Hawkins (1994), as the outcome of using the EIC metric discussed in Section 2.3.

The following results are for when:

$$F_k(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is SO ordered} \\ 1 & \text{if } \omega \text{ is OS ordered} \end{cases}$$

The outcome is that +V2 languages indirectly lose their advantage over the −V2 ones, and an overall steady population state is reached more quickly than without the complexity factor. As we will see it is not because the +V2 languages are more complex that they lose their advantage, it is because acquisition is becoming more random in nature.

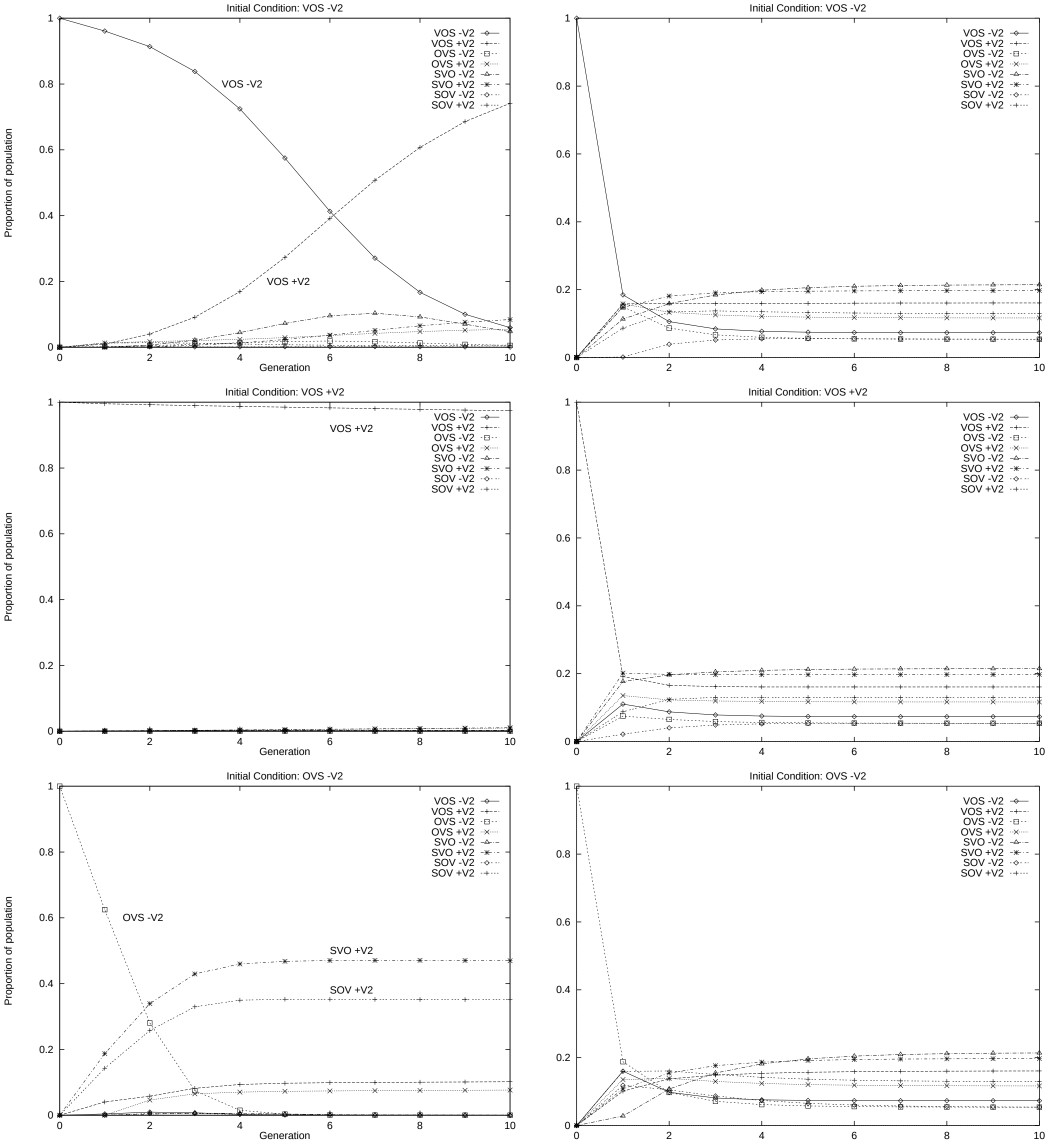


Figure 4.1. Graphs showing the effect of a complexity factor weighting SO ordering in utterances. Graphs on the left are with no complexity factor. Graphs on the right are with complexity factor. Maturation time is 50 utterances for all plots.

Figures 4.1 & 4.2. show the outcome of introducing this complexity factor, compared to the same initial conditions in the basic model.

The result of adding the chosen complexity factor is that after five generations the population mix reaches an equilibrium and no more than twenty percent of the population ever acquires a particular language; although some languages are spoken by three times the proportion of the population as others. The complexity factor in this form is re-introducing the some of the random nature that occurs when the maturation level is very low by causing the rejection of hypotheses on the processing of specific utterances. The factor is effectively providing a handicap system which means that any strong advantage that a language inherently has over others is eliminated. Any advantage that the metric gives to SO ordering is dwarfed by the randomness introduced. This complexity factor seems inappropriate in that often there is rapid change in the language

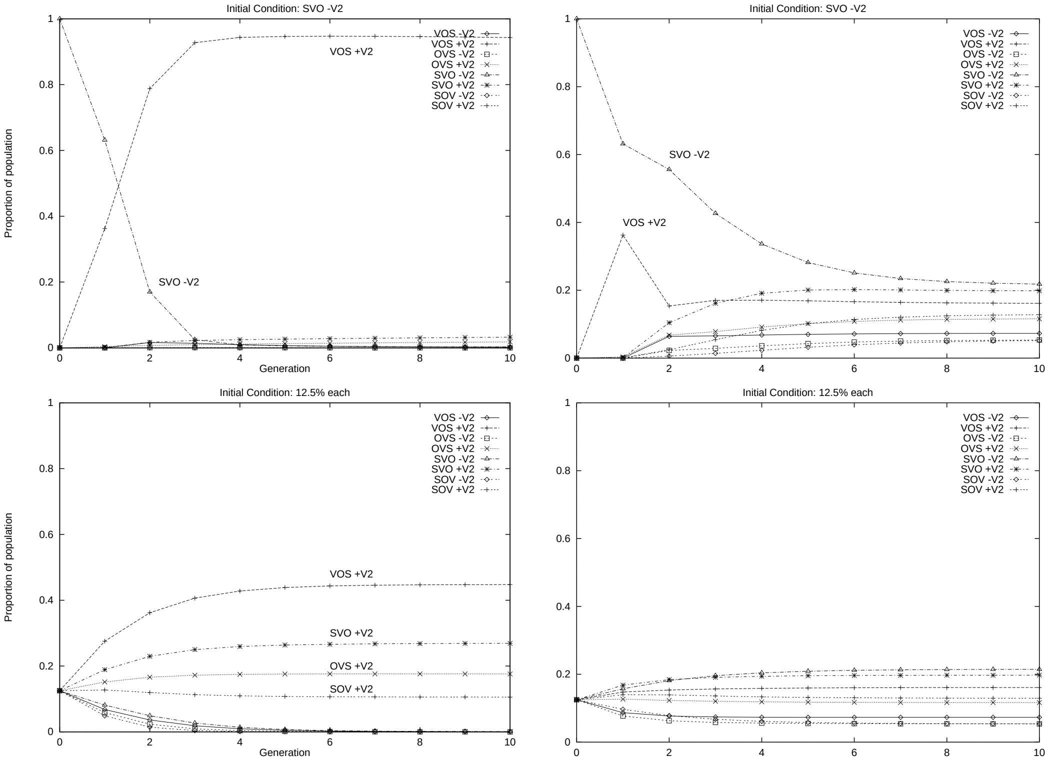


Figure 4.2. More graphs showing the effect of a complexity factor weighting SO ordering in utterances. Graphs on the left are with no complexity factor. Graphs on the right are with complexity factor. Maturation time is 50 utterances for all plots.

distribution from the initial conditions to the second generation of speakers. The languages that learners in a generation acquire is much more dependent on the structure of the model than on the distribution of utterances that they encounter. This would prove particularly unrealistic in a spatially distributed model, where learners only hear other speakers that are in their spatial locality, as spatial clustering of languages would be eliminated. The shapes of the curves produced here differ greatly to those in the basic model. Almost all logistic and exponential behaviour has been removed, and replaced by angular jumps to the equilibrium state.

The outcome can altered by varying the complexity weightings.

$$F_k(\omega) = \begin{cases} 0 & \text{if } \omega \text{ is SO ordered} \\ v & \text{if } \omega \text{ is OS ordered} \end{cases}$$

Figure 4.3. shows the effect of varying the weighting of the complexity factor with different values of v.

Here v is varied from 0 to 0.8. Figure 4.3. clearly shows the way increasing the weighting on the complexity factor smoothly reduces the inherent advantage that any language has over others. A weighting of 0.2 is sufficient to bring all the proportions to less than 0.3 and increasing the weighting further shifts the order of prominence of the languages, leaving SVO −V2 as the most prominent.

Interesting results occur when $0 < v < 0.1$. Here the initial behaviour is like that without the complexity factor present, but the longterm behaviour is more subtle and a steady state is reached much later. For example when $v = 0.05$ an equilibrium may not be reached until after 150 generations. This long term smoothing also reduces the rapid change found between the first and second generations.

The above approach to parsing complexity pays no attention to the new hypothesis (that which the current hypothesis is being rejected for), and its ability to account for the current utterance. This goes against the Greediness Constraint mentioned above, in that by rejecting the current hypothesis, there is no guarantee of ending up with a better hypothesis (with respect to the current utterance), in fact the new hypothesis may not be able to account for the current utterance at all. It would also be more in the spirit of the TLA framework to have the current hypothesis either kept or rejected outright given any particular utterance, and not be rejected probabilistically (the probabilistic nature in the model should be seen as a surface characteristic, where the underlying internal workings are discrete decision making processes). There is also no real sense of competition in the above approach. The whole point of the complexity issue is that two forms compete with each other and the one which is deemed fittest for its environment proliferates.

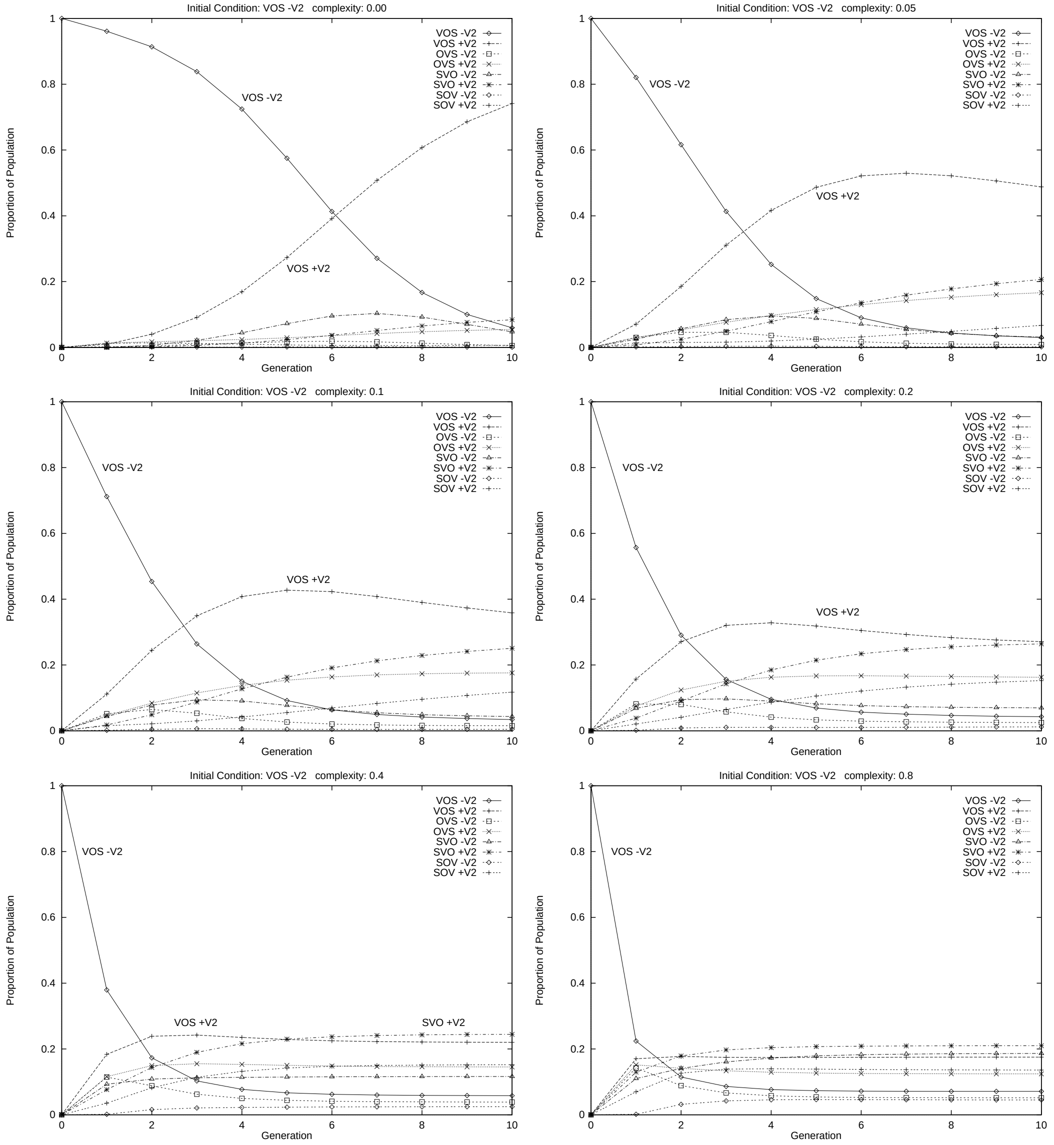


Figure 4.3. Graphs showing the effect of a complexity factor weighting SO ordering in utterances. Complexity weightings from 0 to 0.8 are shown. Maturation time is 50 utterances for all plots.

4.2 Competition

Some of the problems and difficulties discussed above can be remedied by the following formulation:

$$
\begin{aligned}
P[s \rightarrow k] &= \frac{1}{n}\left(\sum_{\substack{s_j \, \in \, L_s, \, s_j \, \in \, L_k \\ F_s(s_j) \, > \, F_k(s_j)}} P(s_j) + \sum_{s_j \notin L_s, s_j \in L_k} P(s_j) \right), \\
P[s \rightarrow s] &= 1 - \sum_{\substack{k \text{ a neighbouring} \\ \text{state of } s}} P[s \rightarrow k].
\end{aligned} \tag{4.2}
$$

Here, the transition from s to k occurs if the current utterance, s_j, is accountable by the current hypothesis, $(s_j \in L_s)$ and the randomly chosen new hypothesis, $(s_j \in L_k)$, if the parse with L_s is more complex that the parse with L_k, $(F_s(s_j) > F_k(s_j))$. That is, a current hypothesis that can account for the current utterance will only be rejected in favour of a new hypothesis which not only accounts for the current hypothesis, but accounts for it is a less complex way. Simply this means that the algorithm will only reject a hypothesis that works for a better, more fit one.

Allowing the current hypothesis to be rejected on processing an acceptable utterance probabilistically has been also been dropped under this formalism in accordance with the above criticisms.

It should be apparent that for this formalism to reject an acceptable current hypotheses, utterances must have different complexity values with respect to different grammars. If the complexity of an utterance is independent of the grammar that produced it, and hence is the same across all grammars, then an acceptable current hypothesis will never be rejected on the grounds that no other acceptable hypothesis is any better than the current one. This means that the complexity measure of surface SO order used in the previous formulation will have no effect under this formalisation and the resulting model will behave exactly as the basic model does. This point is important because is shows that it is not just getting the framework to account for complexity right, but the metric used to define complexity must work well within this framework.

The verb movement to second position in the surface structure of utterance is an obvious candidate to use to measure complexity where complexity values for the same utterance must differ with respect to different grammars. To demonstrate the above formalism the following example metric can be used:

$$F_k(\omega) = \begin{cases} 1 & \text{if } k \text{ is a } +V2 \text{ language} \quad \forall\omega \in F_k \\ 0 & \text{if } k \text{ is a } -V2 \text{ language} \quad \forall\omega \in F_k \end{cases}$$

This basically ranks all utterances which are parsed with verb movement as being more complex than those which are parsed without. That is, any utterance parsed by a $+V2$ grammar is complex, and any utterance parsed by a $-V2$ grammar is simple (e.g. the construction SVO is complex when parsed by a $+V2$ grammar and simple when parsed by a $-V2$ grammar, and the utterance VSO is always considered simple as it only occurs in $-V2$ languages).

Figures 4.4 & 4.5. show a selection of results using this metric. The results are interesting for a number of reasons.

1. The same steady state is reached after about 30 generations, in that each of the languages SVO $-V2$, VOS $-V2$ and VOS $+V2$ end up spoken by approximately 30% of the population and all other languages are spoken by less than 7% of the population, none of them being eliminated completely. The fact that languages can survive in very small proportions and not be completely wiped out is of interest because there exist in the world language types which survive in very small proportions. This contrasts the results from the basic model where the behaviour often all or nothing.

2. Intuitively you might well expect that by marking all $+V2$ utterances as complex, effectively making them less fit, that a strong advantage is being given to $-V2$ languages and they would dominate over $+V2$ languages. This is not what the results show. Out of the three languages that dominate one of them is a $+V2$ language.

3. Logistic change is now much more widespread, (possibly reflecting the competition between rival forms that is present), and if this is how real language change progresses over time, then it is possible that parsing complexity, or some similar filtering mechanism, should be an issue in any such model of language change.

4. There is only a slight internal change to the model. The result of adding V2 complexity is that some of the internal transitions in the model originally from a +V2 state to itself, on presentation of a particular utterance, are altered to being from the +V2 state to the -V2 state. In total there are 15 such altered transitions. Initially 250 of the 576 transitions (1 from each state for each utterance) result in the model changing state. So the change from the basic formulation of the

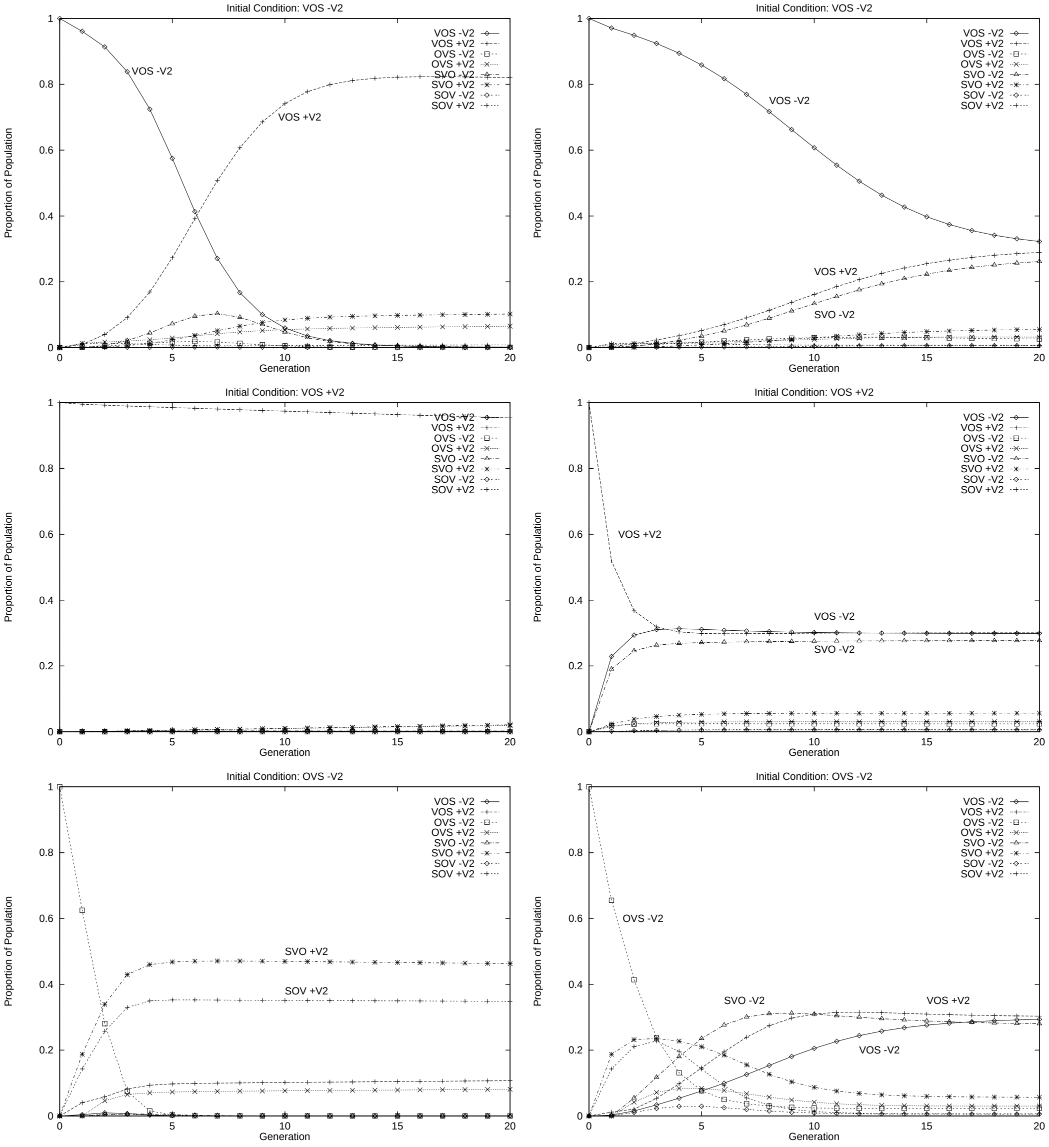


Figure 4.4. Graphs showing the effect of a complexity factor weighting V2 movement in parsing of utterances. Graphs on the left are with no complexity factor. Graphs on the right are with complexity factor. Maturation time is 50 utterances for all plots.

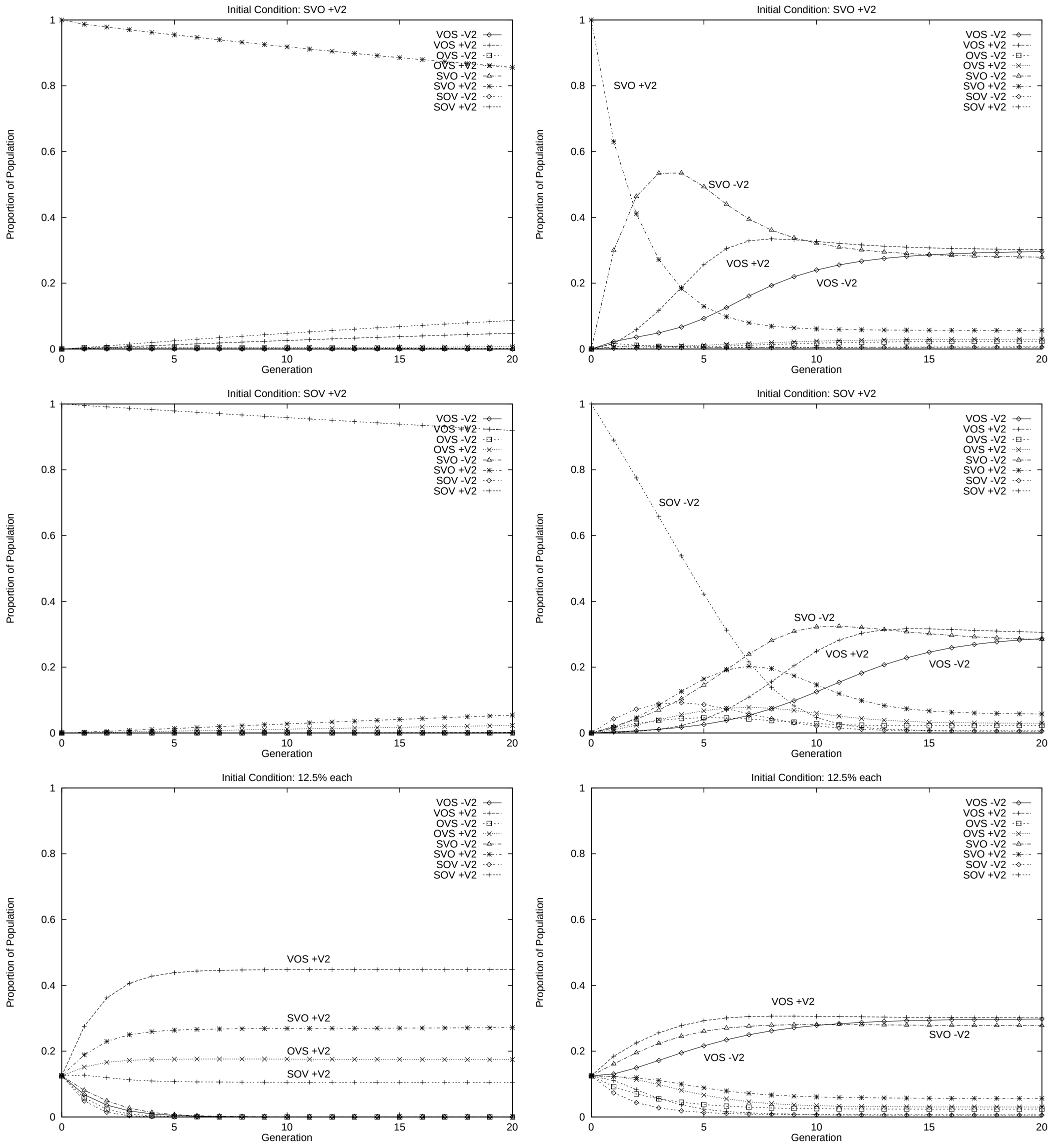


Figure 4.5. Graphs showing the effect of a complexity factor weighting V2 movement in parsing of utterances. Graphs on the left are with no complexity factor. Graphs on the right are with complexity factor. Maturation time is 50 utterances for all plots.

> model is caused by increasing the number of set transitions (i.e. not from a state to itself) by 15 (6%), which is equivalent to altering 2.6% of the total number of transitions.

With respect to the steady state reached, it is difficult to compare this result to historic change, as issues such as growing population sizes and spatial distributions are not considered by the model. The difference between the basic model and the model proposed here though is that different steady states are reached by the basic model depending on the initial conditions, were as in the extended model the same steady state is always reached.

In an attempt to explain the above phenomena we take a closer look at the structure of the particular model we are using as defined by the parameters that have been chosen.

		Transition to								
		L_0	$\mathbf{L_1}$	L_2	$\mathbf{L_3}$	L_4	$\mathbf{L_5}$	L_6	$\mathbf{L_7}$	total
Transition from	L_0		12	10		12				34
	$\mathbf{L_1}$	6			8		10			**24**
	L_2	10			16			12		38
	$\mathbf{L_3}$		8	10					10	**28**
	L_4	12					12	10		34
	$\mathbf{L_5}$		10			6			8	**24**
	L_6			12		10			17	39
	$\mathbf{L_7}$				10		8	11		**29**
	total	28	**30**	32	**34**	28	**30**	33	**35**	

Table 4.1. Number of utterances which will cause transitions in the basic model (+V2 languages in bold font).

Table 4.1. shows the number of utterances which cause a transition from one hypothesis to another. The row totals show that for each language pair, (a –V2 and its corresponding +V2 language), there are always 10 more utterances to cause a transition away from the –V2 language than there are to cause a transition away from the corresponding +V2 language. The column totals show a similar result in the there are always 2 more utterances to cause a transition to a +V2 language then there are to cause a transition to the corresponding –V2 language. The gross behaviour seen here is that on average (assuming that each utterance is equally likely, which is not strictly true), more transitions occur from –V2 languages to +V2 language that the other way around. This shows that the grammar space used in the model does indeed favour +V2 language.

Table 4.2. shows similar data but includes the extra transitions associated with parsing complexity where +V2 language are marked as more complex than –V2 ones. Here the differences in the row sums are such that the difference of 10 is reduced somewhat, but there is still a difference. The situation with the column sums is different in that now the number of utterances to a –V2 language is greater or equal to that of its corresponding +V2 language.

The ratios between transitions into a state and transitions out of a state are shown in Table 4.3. Values greater than 1 reflect there being more possible transitions to a state that away from it. The consequence of this is that if presented with completely random utterances, statistically on average the model should end up in states with rations greater than 1 more often that it does in states with ratios less than 1. These figures show that from the grammar structure alone, in the basic model the +V2 languages do have a definite advantage over the –V2 ones, as the +V2 language are the only ones with values greater than 1. The figures for the complexity model show that language L_0, one of the –V2 languages which arises strongly has a ratio of 1.00 and the L_4 which is the other *strong* –V2 language has a corrected ratio of 0.972 which are much higher than in the basic model, but still less than 1 and less than all of the +V2 languages. These show that the structure of the grammars alone does not fully account for the outcome of the complexity model.

Looking at the compatability of the utterances in the three resulting languages with respect to each other, (i.e. seeing which other grammars also account for an utterance) as shown in Table 4.4., shows that 12/18 (66%) of the utterances in L_1 the

		Transition to								
		L_0	$\mathbf{L}_1$	L_2	$\mathbf{L}_3$	L_4	$\mathbf{L}_5$	L_6	$\mathbf{L}_7$	total
	L_0		12	10		12				34
	$\mathbf{L}_1$	12			8		10			**30**
	L_2	10			16			12		38
Transition from	$\mathbf{L}_3$		8	12					10	**30**
	L_4	12					12	10		34
	$\mathbf{L}_5$		10			10			8	**28**
	L_6			12		10			17	39
	$\mathbf{L}_7$				10		8	12		**30**
	total	34	**30**	34	**34**	32	**30**	34	**35**	

Table 4.2. Number of utterances which cause transitions in the model with parsing complexity (V2 complexity version), +V2 language in bold font.

	Uncorrected		Corrected	
	Basic	Complexity	Basic	Complexity
Language	model	model	model	model
0	0.823	1.000	(0.916)	(1.000)
1	1.250	1.000	(1.083)	(1.000)
2	0.842	0.894	(0.916)	(0.944)
3	1.214	1.133	(1.083)	(1.055)
4	0.823	0.941	(0.916)	(0.972)
5	1.250	1.071	(1.083)	(1.072)
6	0.846	0.871	(0.916)	(0.930)
7	1.206	1.166	(1.083)	(1.069)

Table 4.3. *To–from* transition ratios for basic and complexity models. Uncorrected ratios are calculated from the values given in Tables 4.1 & 4.2. Corrected ratios also include transitions from a state to itself.

+V2 grammar are accounted for by either L_0 or L_4. Conversely 12/24 (50%) of the utterances that would be produced by a population speaking L_0 and L_4 are compatible with L_1. Although there is a high level of compatibility between these three languages, the question of complexity not ruling out the +V2 language in favour of the –V2 still

Utterance	L_0	L_1	L_4	Utterance	L_0	L_1	L_4
ADV AUX V O S	•	•		O AUX V S		•	
ADV AUX V O1 O2 S	•	•		O V S		•	
ADV AUX V S	•	•		O1 AUX V O2 S		•	
ADV S AUX V			•	O1 V O2 S		•	
ADV S AUX V O			•	O2 AUX V O1 S		•	
ADV S AUX V O1 O2			•	O2 V O1 S		•	
ADV S V			•	S AUX V		•	•
ADV S V O			•	S AUX V O		•	•
ADV S V O1 O2			•	S AUX V O1 O2		•	•
ADV V O S	•	•		S V		•	•
ADV V O1 O2 S	•	•		S V O		•	•
ADV V S	•	•		S V O1 O2		•	•
AUX V O S	•			V O S	•		
AUX V O1 O2 S	•			V O1 O2 S	•		
AUX V S	•			V S	•		

Table 4.4. Compatability of the utterances in language L_0, L_1 & L_4 with respect to each other. A '•' signifies that an utterance is accounted for by a particular grammar.

needs to be accounted for. What appears to be happening is that a learner with hypothesis L_1, on hearing utterances from L_4 does not reject L_1 as its hypothesis in favour of L_0 as none of the utterances of L_4 are also in L_0, and no transition is possible to L_4 itself because it it requires 2 parameters to be changed. This means that it is because the two −V2 language that proliferate are complementary to each other, in that no utterance occurs in both languages, that a complex +V2 language can be supported. It is also likely that the pair L_0 & L_4 do so well because they are complementary in their utterances, so one is not rejected in favour of the other, and the support for the +V2 language is a mutual one in that the +V2 language also supports the −V2 languages.

Chapter 5

Conclusions

It has been shown that by adding a complexity factor into the basic TLA definition, the behaviour of the TLA can be changed quite considerably. A first attempt by weighting surface word ordering of subject and object resulted in the model losing its characteristic behaviour almost completely. By dampening the effect of parsing complexity under this formulation, it was shown that the parsing complexity factor in the model took much longer to take effect, but it did eventually take hold of the model and determine the resulting language distribution of the population. The resulting distribution however was not of particular interest as all languages ended up being spoken by an approximately equal proportion of the population. This means that learners will basically end up acquiring a random grammar after a given number of generations.

The outcome is that this result seems less realistic than that produced by the unaltered TLA, and hence the incorporation of parsing complexity in this manner is considered unsuccessful.

The difference in results is itself useful however, because it shows how sensitive the formulation of the TLA is to slight changes. This is important when considering how well it reflects real learning because it shows that the TLA or a similar learning algorithm can produce greatly differing behaviour with only marginal differences in its definition. There are four probable reasons why the results produced are as they are:

1. The metric used is inappropriate. The metric reflects the consequences of the EIC metric, it may be that its simplicity is its downfall, in that there other basic consequences of the EIC metric which are being ignored by using this simplified form.

2. The way the metric was used within the algorithm is unrealistic. As noted earlier the implementation does seem unrealistic in that there is no real competition between forms, but only absolute complexities taken in isolation.

3. The issue of parsing complexity is inappropriate in this context. In other words processing issues play no part in language acquisition.

4. The framework is an unrealistic model of language acquisition. The behaviour of the TLA with the parameters used here obviously has its deficiencies as its behaviour does not generally reflect historical data. But, it is not possible with our current understanding to know exactly where these deficiencies lie.

More favourable results were produced by using verb movement as the complexity measure, and a more competitive decision making process under which the greediness constraint is adhered to. Sample runs of such a model show that smooth logistic change is widespread. This suggests that the model may reflect historical change well. Showing some historic data which the model with the added parsing complexity functionality can account for is unfortunately beyond the scope of the work here, but it would be interesting to see if there are historic changes that it reflects.

It is unclear to the author exactly how useful it is to show a small extract from the model's general behaviour matching a small extract of historical data, especially if the widespread behaviour of the two differ. It is obviously useful in showing that something about the model is right, but focusing on where the more general behaviour of the model does not so closely match what is seen historically, and finding out why this is so, would be more useful as an analytical tool in understanding the processes behind language acquisition and language change.

What has been clearly shown is that parsing complexity is definitely an issue that should be considered in a model of language change, and that the formulation of it is important, as slight variations in formulation can cause large differences in results–a common trait of complex dynamical systems. It has also been shown how a mechanism involving parsing complexity can be effectively incorporated into the TLA framework. The results shown by using verb movement as the complexity factor do match some of the behaviour seen historically

The clear advantages seen in the results of the complexity model include:

- Logistic behaviour is more widespread

- Language structures surviving in small proportions.

- Less +V2 (cross–linguistically rare) languages

- More clearly defined competition between two opposing forms.

Having a model which is solely based upon complexity presents similar problems in that, the consequences of underlying parameterisation are ignored. For example in Kirby's model a grammar is represented just by the set of utterances it can produce, and the consequences of any inherent grammatical structure are lost. Both the effects of parsing complexity and underlying parameterisation need to be accounted for by a model.

There are still however a lot of areas that need further work and a lot of issues that are unclear. The V2 complexity model proposed here only assigns binary values (1 and 0) splitting the grammars into two subsets: one composed of languages of fit utterances and one composed of languages of unfit utterances. A metric over the whole interval would reflect current theories (e.g. those of Hawkins (1994) and Kirby (1996)) more accurately. There is also no interaction within the current metric; that is, fitness is the result of only one factor. How interaction in the metric would effect the model, (c.f. the way surface word order is the result of the interaction between base word order and verb movement) needs to be looked at. It would also be interesting to know exactly what extent the behaviour of the model depends on a particular complexity metric employed and the parameterisation of the grammars. For instance, the extent to which the resulting stable solution is dependent upon the parameter interaction or upon the parsing complexity is unclear. Are other weak forms (like the complex +V2 language that is found to survive), able to survive under different parameterisations or different complexity metrics?

What is clear is the need for further research. As things stand, explanations of language change and language acquisition are split into the functionalist and the innateness approaches, and both present valid issues that need to be considered. However, it seems unlikely that one approach is right and one is wrong, and the need to find middle ground where aspects of both approaches are taken into consideration is needed. The model developed here has shown how such theories can be successfully combined, and that the properties of each part of the combined theory interface well with each other.

Appendix A

Proposition 1 *The model proposed by Kirby in Section 2.3. is equivalent to the Kroch model defined by Equation 2.2.*

If we take the general form with complexities c_f and c'_f and define the functions of time. Without loss of generality $p_f(t)$ is explicitly shown, and $p_{f'}(t)$ is assumed to be of similar form.

$$p_f(t) = \frac{c_f \cdot n_f}{c_f \cdot n_f + c_{f'} \cdot n_{f'}}, \qquad \textit{where } t = 1. \tag{A.1}$$

n_f and $n_{f'}$ are the frequencies at time, $t = 0$, so we now show, by induction, that:

$$p_f(t) = \frac{{c_f}^t \cdot p_f(0)}{{c_f}^t \cdot p_f(0) + {c_{f'}}^t \cdot p_{f'}(0)}, \qquad \forall t > 0. \tag{A.2}$$

The base step is true from its definition and Equation (A.1), so assuming (A.2):

$$\begin{aligned}
p_f(t+1) &= \frac{c_f \cdot p_f(t)}{c_f \cdot p_f(t) + c_{f'} \cdot p_{f'}(t)}, \\
&= \frac{c_f \cdot \frac{{c_f}^t \cdot p_f(0)}{{c_f}^t \cdot p_f(0) + {c_{f'}}^t \cdot p_{f'}(0)}}{c_f \cdot \frac{{c_f}^t \cdot p_f(0)}{{c_f}^t \cdot p_f(0) + {c_{f'}}^t \cdot p_{f'}(0)} + c_{f'} \cdot \frac{{c_{f'}}^t \cdot p_{f'}(0)}{{c_{f'}}^t \cdot p_{f'}(0) + {c_f}^t \cdot p_f(0)}}, \\
&= \frac{\frac{{c_f}^{t+1} \cdot p_f(0)}{{c_f}^t \cdot p_f(0) + {c_{f'}}^t \cdot p_{f'}(0)}}{\frac{{c_f}^{t+1} \cdot p_f(0) + {c_{f'}}^{t+1} \cdot p_{f'}(0)}{{c_f}^t \cdot p_f(0) + {c_{f'}}^t \cdot p_{f'}(0)}}, \\
&= \frac{{c_f}^{t+1} \cdot p_f(0)}{{c_f}^{t+1} \cdot p_f(0) + {c_{f'}}^{t+1} \cdot p_{f'}(0)} \qquad \text{as required.} \qquad \text{(A.3)}
\end{aligned}$$

Now, $p_f(0) = e^{f_0}$ where $f_0 = \ln p_f(0)$ and $p_{f'}(0) = e^{f'_0}$ where $f'_0 = \ln p_{f'}(0)$. Similarly, ${c_f}^t = e^{kt}$, where $k = \ln c_f$ and ${c_{f'}}^t = e^{k't}$, where $k' = \ln c_{f'}$

Substituting, we now have:

$$\begin{aligned} p_f(t) &= \frac{e^{kt+f_0}}{e^{kt+f_0} + e^{k't+f'_0}}, \\ &= \frac{e^{a+bt}}{e^{a+bt}+1}, \qquad \textit{where } a = f_0 - f'_0,\ b = k - k' \end{aligned} \tag{A.4}$$

as required.

Appendix B

The following table shows the surface constructs for each Grammar with in the three parameter framework (Gibson & Wexler 1994).

VOS -V2
ADV AUX V O S
ADV AUX V O1 O2 S
ADV AUX V S
ADV V O S
ADV V O1 O2 S
ADV V S
AUX V O S
AUX V O1 O2 S
AUX V S
V O S
V O1 O2 S
V S

VOS +V2
ADV AUX V O S
ADV AUX V O1 O2 S
ADV AUX V S
ADV V O S
ADV V O1 O2 S
ADV V S
O AUX V S
O V S
O1 AUX V O2 S
O1 V O2 S
O2 AUX V O1 S
O2 V O1 S
S AUX V
S AUX V O
S AUX V O1 O2
S V
S V O
S V O1 O2

OVS -V2
ADV O V AUX S
ADV O V S
ADV O2 O1 V AUX S
ADV O2 O1 V S
ADV V AUX S
ADV V S
O V AUX S
O V S
O2 O1 V AUX S
O2 O1 V S
V AUX S
V S

OVS +V2
ADV AUX O V S
ADV AUX O2 O1 V S
ADV AUX V S
ADV V O S
ADV V O2 O1 S
ADV V S
O AUX V S
O V S
O1 AUX O2 V S
O1 V O2 S
O2 AUX O1 V S
O2 V O1 S
S AUX O V
S AUX O2 O1 V
S AUX V
S V
S V O
S V O2 O1

VSO -V2
ADV S AUX V
ADV S AUX V O
ADV S AUX V O1 O2
ADV S V
ADV S V O
ADV S V O1 O2
S AUX V
S AUX V O
S AUX V O1 O2
S V
S V O
S V O1 O2

VSO +V2
ADV AUX S V
ADV AUX S V O
ADV AUX S V O1 O2
ADV V S
ADV V S O
ADV V S O1 O2
AUX S V
O V S
O1 AUX S V O2
O1 V S O2
O2 AUX S V O1
O2 V S 01
S AUX V
S AUX V O
S AUX V O1 O2
S V
S V O
S V O1 O2

SOV -V2
ADV S O V
ADV S O V AUX
ADV S O2 O1 V
ADV S O2 O1 V AUX
ADV S V
ADV S V AUX
S O V
S O V AUX
S O2 O1 V
S O2 O1 V AUX
S V
S V AUX

SOV +V2
ADV AUX S O V
ADV AUX S O2 O1 V
ADV AUX S V
ADV V S
ADV V S O
ADV V S O2 O1
AUX S V
O V S
O1 AUX S O2 V
O1 V S O2
O2 AUX S O1 V
O2 V S 01
S AUX O V
S AUX O2 O1 V
S AUX V
S V
S V O
S V O2 O1

Bibliography

Berwick, R. (1985), *The Acquisition of Syntactic Knowledge*, Cambridge, Mass.: MIT Press.

Chomsky, N. (1981), *Lectures on Government Binding*, Dordrecht–Holland: Foris Publications.

Clark, R. (1990), *Papers on Learnability and Natural Selection*, Technical Reports in Computational Linguistics, No. 1.

Clark, R. & Roberts, I. (1993), 'A computational model of language learnability and language change', *Linguistic Inquiry* **242**, 299–345.

Gibson, E. & Wexler, K. (1994), 'Triggers', *Linguistic Inquiry* **25**(3), 407–454.

Haegeman, L. (1991), *Introduction to Government and Binding Theory*, Oxford: Blackwell.

Hawkins, J. A. (1994), *A Performance Theory of Order and Constituency*, Cambridge Studies in Linguistics, Cambridge, UK.: Cambridge University Press.

Hurford, J. R. (1990), Nativist and functional explainations in language acquisition, *in* I. M. Roca, ed., 'Logical Issues in Language Acquisition', Dordrecht–Holland: Foris Publications, pp. 85–136.

Kirby, S. (1996), Function, Selection and Innateness The Emergence of Language Universals, PhD thesis, Department of Linguistics, University of Edinburgh.

Kroch, A. S. (1989), Function and grammar in the history of English: Periphrastic "do.", *in* R. Fasold, ed., 'Language Change and Variation', Amsterdam: Benjamins.

Lightfoot, D. (1991), *How To Set Parameters: Arguments from Language Change*, Cambridge, Mass.: MIT Press.

Niyogi, P. & Berwick, R. C. (1995), 'The logical problem of language change', C.B.C.L Paper No. 115, MIT AI Lab.

Niyogi, P. & Berwick, R. C. (1996*a*), 'A dynamical systems model for language change'.

Niyogi, P. & Berwick, R. C. (1996*b*), 'A language learning model for finite parameter spaces', *Cognition* **nn**(n), yyy–zzz.

Oliphant, M. (1996), 'The dilemma of saussurean communication', *BioSystems* **37**, 31–38.

Tomlin, R. S. (1986), *Basic Word Order: Functional Principles*, London: Routledge (Croom Helm).