A Simulation of Lardil Language Change
Guanjun Bella Feng
University of Southern California

Introduction

This paper reports on a project simulating language change in Lardil over the past 30 years. Lardil is a Pama-Nyungan language spoken on Mornington Island in Australia. Like other Australian languages, Lardil used to have a very rich morphology. Richards (1997) gives the following table to show the richness of Old Lardil (Old Lardil refers to the Lardil spoken until about thirty years ago, as documented by Ngakulmungan Kangka Leman (1997).

<table>
<thead>
<tr>
<th>Base</th>
<th>kirdikirdi “moon”</th>
<th>wangalk “boomerang”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominative</td>
<td>kirdikir</td>
<td>wangal</td>
</tr>
<tr>
<td>Objective</td>
<td>kirdikirdi-n</td>
<td>wangalk-in</td>
</tr>
<tr>
<td>Future</td>
<td>kirdikirdi-wur</td>
<td>wangalk-ur</td>
</tr>
<tr>
<td>Marked non-future</td>
<td>kirdikirdi-ngarr</td>
<td>wangalk-arr</td>
</tr>
<tr>
<td>Locative</td>
<td>kirdikirdi-i</td>
<td>wangalk-e</td>
</tr>
<tr>
<td>Genitive</td>
<td>kirdikirdi-kan</td>
<td>wangalk-kan</td>
</tr>
<tr>
<td>Intransitive allative</td>
<td>kirdikirdi-ya</td>
<td>wangalk-inya</td>
</tr>
<tr>
<td>Transitive allative</td>
<td>kirdikirdi-mari</td>
<td>wangalk-mari</td>
</tr>
<tr>
<td>Intransitive ablative</td>
<td>kirdikirdi-burri</td>
<td>wangalk-burri</td>
</tr>
<tr>
<td>Transitive ablative</td>
<td>kirdikirdi-burri</td>
<td>wangalk-burri</td>
</tr>
<tr>
<td>Comitative</td>
<td>kirdikirdi-ngun</td>
<td>wangalk-ingun</td>
</tr>
<tr>
<td>Prorptive</td>
<td>kirdikirdi-wur</td>
<td>wangalk-ur</td>
</tr>
<tr>
<td>Instrumental</td>
<td>kirdikirdi-wurr</td>
<td>wangalk-werr</td>
</tr>
</tbody>
</table>

Table 1. Old Lardil Nominal Morphology

The richness of the morphology is also reflected in the free word order of Old Lardil sentences, as in the following sentences. These sentences also illustrate the use of the future case, which occurs on nouns when the verb is marked for future:

(1) a. Ngada latha diini libani
     I spear this-OBJ pumpkinhead-OBJ
     ‘I speared/am spearing this pumpkinhead’

b. Ngada lathu diinku libanku
   I spear-FUT this-FUT pumpkinhead-FUT
   ‘I will spear this pumpkinhead’

c. Diinku libanku lathu ngada.
   this-FUT pumpkinhead-FUT spear-FUT I
   ‘I will spear this pumpkinhead’

During the past thirty years, Lardil has undergone some substantial changes. The major difference between Old and New Lardil is that New Lardil has comparatively impoverished nominal morphology. Thirty years ago, only a small set of Old Lardil speakers did not mark the future and objective for the nominals, especially for vowel final bases. In New Lardil, there are even more speakers that do not mark these cases. The reason for the change, according to Richards (1997) is the scarce input for the
language learners. The program presented below simulates this change and claims that the reason for the change is that the limited input of Old Lardil prevents the learners from generalizing the complex phonological rules deriving the nominative form. The learners then obtain two base forms for one nominal: the real base and the nominative form. These two forms compete to be learned as the base form. Because the "pseudo base form"—the nominative forms introduces more vowel-final forms into the base form, the learner produces more objective and future forms without the case marking.

I claim that because of the low frequency of Old Lardil input, learners fail to learn the phonological rules. This is in line with the study by Bybee (1994)’s finding that frequency of use is an important factor in the transmission of phonological and morphological processes.

I. The idea of I-language and E-language

Language exists in two different domains (Chomsky 1986; Hurford 1987; Kirby 1999c): I-language, the internal language as represented in the brains of the population, is the language user’s knowledge of language; E-language is the external language that exists as utterances in the arena of use (Hurford 1987). For a language to survive from generation to generation, it must be learned by individuals through observing behaviors of other individuals. The process of language transmission is done through two processes: induction and production. Induction is the process of the learner acquiring his own I-language by generalizing what he has heard in the form of E-language. At the same time, the learner produces his own E-language utterances—this is production—which serve as learning data for others. Looking at the language at large, the processes of language induction and language production are the processes that impose endogenous selection pressures on the language.

These two processes are what Kirby (1999b) called bottlenecks on the persistence of linguistic variation. In the process of going through the bottlenecks, a language with high compositionality is more easily transmitted than a language with lower compositionality. Compositionality is defined as the property whereby an expression’s meaning is a function of the meanings of parts of that expression and the way they are put together. In other words, compositionality is the regular pattern of a language. The more regular a language pattern it is, the easier it is for the learner to induce the faithful I-language from the E-languages that he heard. The bottleneck is thus less severe.

II. Application of the Iterative Learning Model

The idea of I-language and E-language is applied in computational model of Iterative Learning Model (ILM) by Kirby (1999a, 1999b, in press). It has four components: a meaning space, a signal space, one or more learning agents and one or more adult agents. Each iteration of the ILM begins with the adult agents being given a set of randomly chosen meanings for which to produce signals. The resulting meaning-signal pairs become the training data (E-language) for the learning agent. Based on the training data, the learning agents come up with their own grammar (I-language) and then produce utterances (E-language) according to their acquired I-language. This process iterates a fixed number of times until the learner becomes an adult agent and another new learner of the language replaces him.
The agents in the ILM need at least four components: 1) an internal representation of language that specifies the ways in which signals can be produced for particular meanings; 2) an algorithm for inducing this representation given examples of meanings and signals; 3) some means of generating signals for meanings that the induced language representation does not include; The algorithm in 2) is the learner’s induction algorithm, which includes two steps: incorporation and generalization. In the step of incorporation, the meaning-string pairs are added to the grammar as unitary, non-compositional rules, if the learning agent could not already parse the signal component of the pair with its existing set of rules. Generalization is the process of integrating the new rule into the existing grammar by searching for possible generalizations. For example, for the two rules: C:(a, b)->cde and C:(a, f)->cdg that have a common meaning component “a” in the left side of the rule and a common substring “cd” on the right side of the rule, the learner will create a more generalized rule: C:(a, x)->cd D:x and two subrules: D:b->e and D:f->g.

One finding of Kirby (in press) is that the induction pressure (also called learning bottleneck) becomes more severe for low-frequency meanings since these will have less likelihood of being expressed in the learner’s training set. The low frequency forms need therefore to belong to regular paradigms in order to be learned by the learner. With a restricted amount of input, low-frequency forms can only be learned if they are compositional, because they don't need actually need to be heard; the learner can produce them merely by generalizing. In this study, we will simulate the effect of low frequency on a language change process. By restricting the amount of input that the learner gets, we predict that it will be more difficult for the learner to learn partial meaning-signal pairs when phonological rules apply to them, particularly when the learner can make more than one guess about the underlying form and rule pair that generated an observed form.

III. Background of Lardil language change

In 1960 and 1967 Hale did an extensive study on Lardil. At that time, the language was spoken by many adults on the island, though the younger generation consisted primarily of monolingual English speakers. In 1996, Hale went to the island again. By then, the language was spoken by considerably fewer people, the youngest of whom were in their early fifties. During the thirty years in between Hale’s two visits, the language of Lardil has undergone some systematic changes. The comparison between Old and New Lardil shows that New Lardil has comparatively impoverished nominal morphology and a more fixed word order than Old Lardil. There might be two reasons for these changes: the decline in common everyday use of Lardil, and the contact with English.

Richards (1997) claims that although the two factors are sociologically related, they are linguistically distinct. To distinguish the two, the theory that holds that the changes in Lardil are entirely due to the English influence is called the "English influence theory" and the theory that attributes the change to the scarce data that New Lardil learners are exposed to is referred to as the "scarce data theory". Richards holds that the scarce data theory is closer to the truth and better accounts for why New Lardil differs from Old Lardil in the way that it does. Other studies have noted that language change appears to accelerate in situations in which children are learning the language on the basis
of scarce data, such as Dorian (1981), Schmidt (1985), and Pensalfini and Richards (2000). I will not go through Richards' arguments in detail, but they center on the claim that in the process of language learning, New Lardil learners reanalyzed the nominals. This reanalysis originally comes from an opaque relationship between the nominative form and the base.

Old Lardil nominal morphology is added to the base (underlying form of the stem), which is distinct from the surface nominative (citation) form. The nominative form is predictable from the base via the following five rules:

a. final high vowels become non-high. /nguku/ --> nguka "water"

b. trisyllabic or longer bases are shortened. /kirdikirdi/ --> kirdikir "moon"

c. monomoraic bases are lengthened. /yak/ --> yaka "fish"

d. final clusters are simplified. /wangalk/ --> wangal "boomerang"

e. certain final consonants (including all bilabials and velars) are deleted. /kurkang/ --> kurka "panja" (data from Richards, 1997)

In Old Lardil, the objective and future cases—the ones I will focus on here, and which are included in the simulation—are marked with the suffixes -in and -ur respectively. Even for many Old Lardil speakers, however, the objective and future markings are often absent, especially with vowel-final bases. Richards (1997) proposes that New Lardil speakers have generalized this absence of inflection. New Lardil speakers failed to realize, from the small Lardil sample from which they were working, that the relevant factor determining whether inflection appears or not has to do with the presence or absence of a base-final vowel. This is in part because the phonological rules listed above made it difficult for them to learn which bases were vowel-final. In other words, New Lardil differs from Old Lardil in that the null alternate of certain inflectional suffixes may appear freely, rather than being phonologically conditioned.

Several of the listed rules above (c, e) have the effect of creating vowel-final nominative forms out of consonant-final bases. Richards proposes that, in the process of the evolution, New Lardil speakers have reanalyzed these nominals, making the bases identical to the nominative forms. New Lardil would then have considerably more vowel-final bases than Old Lardil, and consequently more cases in which objective and future endings would have a null realization, even following the Old Lardil rules.

Following Richards’ line of thinking, this study implements a simulation of the language change that occurred in Lardil using the Iterative Learning Model. The purpose is to see how the number of utterances by the speaker affects the grammar generalized by the learner. In particular, I examine whether restricting the number of utterances can explain why the new Lardil speaker starts reanalyzing the nominals and why he identifies the base form with the nominative form. This study tries to give a plausible reason for this process.

IV. Details of the program

The details of the program differ from those of Kirby’s in some respects for practical reasons. Instead of having a speaker without any grammar of the language, this simulation endows the first speaker with a full, Old-Lardil-like grammar. This is because the language change to be simulated here starts with a fully developed language, whereas Kirby’s experiment simulates the emergence of a language from nothing. In addition,
only one generation is simulated. The model also differs substantively from Kirby’s in
that it includes phonological rules.

Ten words and the three inflections to be simulated (objective, future and
nominative) are chosen as the meaning set for the speaker. These give thirty meaning
pairs. An example is (obj, fish), which denotes the objective form of fish.

Items from among the thirty meaning pairs are randomly chosen for the speaker to
express. To produce the objective and future forms, the speaker adds “-in” and “-ur”
respectively to the base form. To produce the correct nominative forms, the speaker goes
through the phonological rules listed in Section III. The speaker then passes the utterance
to the learner in the form of, for example, S:(obj, fish)->yakin, which says that the
Objective form of ‘fish’ is “yakin”.

The input for the learner thus takes the form of: C:(A, B)->S. “C” refers to the
category of the rule. A rule with category of “D” means the rule is a starting point for a
parsing process. Category of “N” means the rule is a subrule for generating the base of
the noun. The “A” part of the rule refers to the inflection of the noun, which includes
“OBJ”(Objective), “FUT”(Future) and “NOM”(nominative). If it is an “N” subrule, the
“A” part will be “0”. The “B” part refers to the meaning of the nominal which is
represented by the English gloss. The “S” part of the rule refers to the utterance produced
by the speaker. The non-compositional rule internalized by the learner takes almost the
same form (if the learner has learned it correctly, of course), except that the learner’s rule
has a frequency counter that records how many times he has heard this input.

If the learner’s current grammar is unable to parse the input, the input to the
learner is added to the learner’s grammar as a non-compositional rule. The learner then
generalizes this new rule and tries to see if it has anything in common with an old rule If
an existing rule can parse the input, the existing rule gets one more frequency point

After a certain number of inputs (this number is decided by the user of the
program; it is the independent variable in the simulation experiments performed here),
the learner will form a set of rules and make utterances accordingly.

Here the simulation is different from Kirby’s again. What we are interested in the
program is whether the speaker’s utterance will reflect the language change, namely,
whether he is more likely to come up with the non-inflected objective and future nominal
forms or the fully inflected ones. So a counter is set in the program to see the ratio
between the two. The number is the dependent variable in the simulation experiments.

The remainder of this section gives more detail about the meaning set. The
program chooses from the objective, future and nominative forms of the following ten
words as meaning inputs to the speaker. The ten words in their base and nominative
forms are as follows:

<table>
<thead>
<tr>
<th>Base</th>
<th>Gloss</th>
<th>Nominative</th>
<th>Future</th>
<th>Objective</th>
</tr>
</thead>
<tbody>
<tr>
<td>kang</td>
<td>speech</td>
<td>kangka</td>
<td>kangur</td>
<td>kangin</td>
</tr>
<tr>
<td>nguka</td>
<td>water</td>
<td>nguka</td>
<td>ngukur</td>
<td>ngukin</td>
</tr>
<tr>
<td>kurkang</td>
<td>panja</td>
<td>kurka</td>
<td>kurkangur</td>
<td>kurkangin</td>
</tr>
<tr>
<td>jul</td>
<td>hair</td>
<td>julda</td>
<td>julur</td>
<td>julin</td>
</tr>
<tr>
<td>yak</td>
<td>fish</td>
<td>yaka</td>
<td>yakur</td>
<td>yakin</td>
</tr>
<tr>
<td>kerndi</td>
<td>wife</td>
<td>kernde</td>
<td>kerndur</td>
<td>kerndin</td>
</tr>
<tr>
<td>wun</td>
<td>rain</td>
<td>wunda</td>
<td>wunur</td>
<td>wunin</td>
</tr>
<tr>
<td>kirdikirdi</td>
<td>moon</td>
<td>kirdikir</td>
<td>kirdikurdur</td>
<td>kirdikirdin</td>
</tr>
</tbody>
</table>
Among Old Lardil speakers, there is variation in whether the inflections appear. According to Richards (1997), a certain set of Old Lardil speakers that does not inflect nominals for objective and future cases, especially with vowel-final bases. Unfortunately, the exact number of these speakers is not available (p.c. with Norvin Richards). This is implemented in the program by having the original speaker’s utterances fail to show objective and future inflections on the nominals 1/5 of the time. If the base is vowel final, 2/5 of the speaker’s objective and future utterances will not show the inflections.

If the input is in the form of S:(OBJ, fish)->yakin or the form of S:(OBJ, fish)->yak, the learner, by comparing with his existing grammar, is going to deduce that the general rule is S:(OBJ, N)->Nin or S:(OBJ, N)->N and at the same time, he will have a subrule saying: N(0, fish)->yak. If this subrule already exists, the existing subrule gets one more frequency point. Future forms are processed analogously. If the learner “hears” enough Objective and Future forms of other nouns, the learner will easily obtain the general rules for the Objective and the Future form: either “add -in/-ur to the base” or “use the base form of the noun”. However, when the input a nominative, like S:(NOM, fish)->yaka, the learner will have a hard time generalizing the rules for the form because the derivations for different base forms are different. The phonological rules, repeated here, are quite complicated:

a. final high vowels become non-high. /nguku/-->nguka "water"
b. trisyllabic or longer bases are shortened. /kirdikirdi/-->kirdikir "moon"
c. monomoraic bases are lengthened. /yak/--yaka "fish"
d. final clusters are simplified. /wangalk/--wangal "boomerang"
e. certain final consonants (including all bilabials and velars) are deleted. /kurkang/--kurka "panja"

There is an opaque relation between the nominative form and the base to which nominal morphology is added. For example:

/wun/→ wunda ‘rain’ (undergoes rule(3c))
/wunda/→ wunda ‘stingray sp.’ (no change)

With limited input as in the case of New Lardil, the learner is assumed not to be able to generalize the correct phonological rules. He can only learn the noncompositional rule, without generalizing from it, or guess that the nominative form is equivalent to the underlying form. For example, when the input to the learner is S:(NOM, fish)->yaka, he could do two things. One is to posit S:(NOM, fish)->yaka as a new rule. At the same time, he could guess that, as in the case of future and objective inflections, the speaker has just used the base form of the noun in the nominative form. The general rule he supposes the speaker has used is: S:(NOM, N)->N. And the subrule he would have is: N:(0, fish)->yaka. The result now is, the learner has two base forms for fish, N:(0, fish)->yak from the Future and Objective forms the learner has heard and N:(0, fish)->yaka. All these rules the learner acquires are put in a separate file as output of the program.

When it comes to production, the learner will choose between the two base forms he learned previously. The chances of choosing either one are dependent on how many times the learner has encountered that base. To take our previous example, whether the learner will choose “yak” or “yaka” depends on (i) how many times he has heard rules
related to “yak”, like S:(OBJ, fish)->yakin, S:(FUT, fish)->yakur, S:(OBJ, fish)->yak or S:(FUT, fish)->yak, and S:(NOM, fish)->yaka when correctly identified as the augmented form of /yak/, and (ii) how many times he has heard the nominative rule related to “yaka”: S:(NOM, fish)->yaka and failed to identify it as a case of the augmentation rule applying. The frequency counter in the learner’s rule now comes into use. If he has heard the former T1 times and the latter T2 times, he will choose “yak” as the base form T1/(T1+T2) percent of the time, and choose “yaka” T2/(T1+T2) percent of the time.

Another counter is set for both the speaker’s and the learner’s productions, counting how many times the speaker uses the base form, and how many times the speaker uses the “-in” and “-ur” forms to mark future and objective. The ratio between the two is computed.

When the input increases to a certain point, the learner is allowed to learn the phonological rules, and stops considering the “pseudo” base form (nominative) as a possible underlying form.

V. Result of running the program

The program, written in C++ and given in the Appendix, was run in two different versions: in one version, the learner never achieves the phonological rule; in the other, he achieves the phonological rule for a given base after hearing it in the nominative ten times, which happens at about the time he has received a total input of 200-300 rule occurrences.

Because a random number function is used in the program, each run of the program will not give the same result. So the program is run ten times for each number of inputs. The variation of the results depends on the frequency of input. When the time of input is set at a smaller number, the variation among the ten result is quite big. When the input times increases, this variation decreases. The variation is shown by the error bar in Chart I.

Putting the average of the ten results for each number of input, we have the following chart:
It can be seen from above that the ratio of unmarked to marked forms for the learner without phonological rules using the base form for the future and objective forms is consistently higher than when the speaker speaks. This replicates the language change seen in Lardil over the past thirty years. That is, nominal inflections are less frequently marked after thirty years. By obtaining the result without any interference of English data, the simulation supports Richards’ claim that the change of Lardil is due not to the influence of English but to the scarce input that new Lardil learners get from the previous generation. The complexity of the phonological rules for the nominative form prevents the learner with limited input from generalizing the rules. This also confirms Hawkins’ (1998) prediction that morphemes and allomorphs exist in proportion to their potential for use.

For the learner who learned the phonological rules after 10 inputs of the rule occurrences, the ratio of use of the unmarked form is very close to that of the adult speaker. As the number of inputs gets bigger, the two get closer and closer. If the New Lardil learner had had adequate input, his grammar would have been very similar to that of the Old Lardil speaker. This supports the claim that the main factor for Lardil language change is the acquisition of the phonological rules.

The interesting part of the language change in Lardil is that the complexity of derivation of one form (nominative form) has an impact on the others (future and objective). Richards (1997) claims that the change is caused by a reanalysis of the nominals by the learners, but doesn’t offer an explanation for why the reanalysis occurs.
and why the learner identifies base with nominative forms in the process of reanalysis. Through the computational model, we see that the reason the learner takes the nominative form as the base is that some speakers of old Lardil do not inflect future and objective forms, giving the learner a reason to guess that the nominative forms he hears may also be a non-inflected form, that is the base which is in fact the uninflected nominative form. Actually, the learner doesn’t exclusively identify the base form with the nominative form, but rather has an uncertainty about the base form which is reasonable when there is limited input to make a choice. The learner incorporates the surface form of the nominative as one of the base forms. Because the phonological rules that apply to the nominative often create vowel final forms from the base, there are more “bases” (including the real base and the learner’s guess) ending with a vowel. This, in turn, creates more non-inflected forms for Objective and Future forms since for vowel final bases, speakers tend to use more non-inflected forms. The whole process starts from the fact that there is limited input to the learner who, therefore, couldn’t generalize the phonological rules.

In conclusion, this study simulates the language change of Lardil with Kirby’s ILM model and it clarifies the reason for the change. At the same time, it shows that the ILM is a good model for simulating language change. There remains the question as to how phonological rules can be generalized by the learner and to what degree the complexity of a rule and its learnability are related (Hayes, 1999).

Future research could clarify the role of frequency, both across froms of nouns and across applications of phonological rule, by investigating the corpus of Old and New Lardil which will make the program more realistic. Instead of having an arbitrary ratio between the nominative, future and objective forms, investigation on the corpus will give a frequency ratio among the three based on real language contact.
Reference:

    Historische Germanische und deutsche Syntax, Sonderdruck.
Hayes, Bruce. (1999). An Automated Learner for Phonology and Morphology. ms.
Kirby, Simon. in press. Spontaneous Evolution of Linguistic Structure: an iterated learning model of the emergence of regularity and irregularity.
    Philadelphia: University of Philadelphia Press

Contact Information:
Guanjun Bella Feng
Department of Linguistics
University of Southern California
University Park, Los Angeles, CA90089
Email: gfeng@usc.edu