AUTOMATIC SIMULATION OF HISTORICAL CHANGE

Raoul N. Smith
Northwestern University

0.0 Purpose. One of the principal reasons for studying the history of a language has been to explain the system of its modern reflex, the contemporary language. This has been especially true in attempting to deal with certain anomalies in the modern language. But the role, if appropriate, of utilizing information concerning diachronic processes in a synchronic description is not at all clear. Recent studies[1] describing contemporary languages, based purely on synchronically motivated grounds, suggest a much more intimate relation between a synchronic grammar and what has been previously posited as a diachronic description of that language.[2]

The two major problems involved in historical studies have been the statement of the sound change (or, as this has been reinterpreted, the grammar change) and the relation of this change to other diachronic changes, that is, its relative chronology. A great deal of attention has been paid to the former but very little to the latter, whose significance has greatly increased

---

[1] See, for example, Lightner, 1965 and Schachter and Fromkin, 1968.

[2] Since a naive native speaker of a language can not be expected to know the history of his language, the reason for this relation may lie in the manner in which the rules were added to the grammars of his ancestors. It is hoped that the future results of this study will help to shed light on this relation by comparing them and their associated grammars with synchronic descriptions.

due to recent results in generative grammar. One of the reasons for the lack of rigor in stating the relative chronology has probably been the large amounts of data required for input to the set of rules and the very large number of stages/rules which must be accounted for and the many permutations of these rules which should be tested. This lack of rigor, in turn, has made it very difficult to discuss coherently the historical development of a language.

The purpose of this paper is to discuss certain limited aspects of historical language change and suggest the possible use of the computer in approaching their solution. The types of problems discussed are only phonological and include only those changes conditioned by phonetic environment and which do not require syntactic information (for example, the change of the Old Russian unstressed infinitive ending /t'i/ to /t'/).

1.0 Language change. The discovery or postulation of the history of a language has been approached by two rather well-known methods. These are the reconstruction of the parent language of a set of genetically related languages and the reconstruction of an older stage of a language given a later stage. Both assume that languages evolve, that is, change (either gradually or abruptly) and that the relation of one language stage to the other is that one preceded the other. By comparing these stages in the development of the same or related sister languages one can

therefore reconstruct or recover the parent or proto language from which it or its sister languages evolved.

The two problems of reconstruction -- the reconstruction of a parent language of related sister languages and the reconstruction of an earlier form of a single language -- have been approached by separate methods.

The problem of reconstructing the parent of a set of related sister languages has been formalized by the so-called comparative method. The comparative method assumes, among other things and in a simplified version, that by comparing sets of sounds occurring in the same positions of the same words in the sister languages one can reconstruct the sound from which these sister sounds evolved. ("Same position" and "same word" may be difficult to define in a particular case.) For example, the word for "three" in various Indo-European languages is

Sanskrit	tráyaḥ
Greek	τρεῖς
Latin	trēs
Gothic	þreis
Lithuanian	trỹs
Old Church Slavonic	trьje

Since five of the six forms have some kind of voiceless, dental stop in word initial position and that the fricative in Gothic can be accounted for by a change particular to Gothic, the assumption is made that the parent language of these, Proto-Indo-European, had

a voiceless, dental stop symbolized by *t, and this fact is highlighted by arranging these correspondences in tabular form:

PIE	Skr	Gr	Lat	Go	Lith	OCS
*t	t	τ	t	þ	t	t

Once the proto language has been reconstructed,[1] the correspondences used in the reconstruction can be reinterpreted as the results of phonological changes, for example, Proto-Indo-European *t became t in Old Church Slavonic, or, as it is usually expressed

PIE *t > OCS t

The problem of discovering an older stage of a language given only a later stage of that language is approached by examining certain alternations in the language at the later stage and from these postulate a proto form from which it could have evolved. The alternations in the later stage which are usually chosen are in the form of morphophonemic alternations, that is, phonemically differently shaped forms of the same morpheme. The assumption is made that these irregularities in the shape of one and the same morpheme must have been conditioned regularly so that by postulating one proto form and accounting for the change by a general rule, we have successfully reconstructed that form of the earlier language.

---

[1] The forms of the proto language reconstructed by the comparative method can be interpreted as a statement of the state of the art in reconstruction for that language family.

For example, in Modern Russian the first person singular present of the verb "to be able" is /mogu/ and the second person singular is /možiš/. The first singular of "to read" is /čitaju/ and the second singular is /čitajiš/. From these and other sets of verbs we would conclude that the first singular ending is /u/ and that of the second singular is /iš/. Therefore, the stem morpheme alternants must be $\{mog \sim mož\}$ and $\{čitaj \sim čitaj\}$. The first of these two sets exhibits a morphophonemic alternation of g/ž. This same alternation in the first person singular occurs in other sets of verbs when the stem ends in a velar. From this (and other corroborative forms) we postulate that an earlier stage of Russian had one form for this verb stem, namely /mog/, and that before the vowel /i/ /g/ later became /ž/.

The proposal in this paper is to reverse the bottom-to-top model of the comparative method and that of internal reconstruction into a top-to-bottom generative model where the input forms are reconstructed lexical items and the rules are the set of postulated sound changes for the language. But there are two major difficulties in reversing the older models. One, the documented changes have often been incorrectly or incompletely stated and, two, the relative chronology of various rules has not been adequately described. We hope to show how the computer can be used at least to test the accuracy of the rules and secondly to test or, hopefully, to help discover the relative chronology of the rules exhibited by their ordering.

The model proposed here is one where the proto language is a set of reconstructed forms (chosen, for example, from a standard reference work). The rules describing the phonological changes in that language are then described and ordered. As the program operates on these forms, the output from each rule represents a synchronic stage in the development of that language. As final output one hopes to get the modern language. If any of the output is incorrect, then it is assumed to be from one of four possible sources: an incorrectly formulated rule (including analogical formation), a non-existent rule, an incorrectly ordered rule, or an incorrectly reconstructed form. Being able to differentiate which of these is the actual cause for the incorrect output is simplest only in the case where all of the output was the result of the application of only one rule.

2.0 A sketch of the phonological history of Russian. The rules which were tested were an abridged version of a set presented by the author in a recent paper.[1] The rules attempt to account for certain aspects of the development of the phonological system of Contemporary Standard Russian from a late form of Proto-Indo-European. These rules were:

---

[1] Kantor, Marvin and R.N. Smith, "A sketch of the major develppments in Russian historical phonology" (to appear). The original formulation was in terms of distinctive features; however, for this programmatic study a segmental notation has been used for ease of statement, etc.

1. $\begin{bmatrix} k^h \\ g^h \\ k\overset{u}{\smile}^h \\ g\overset{u}{\smile}^h \end{bmatrix} > \begin{bmatrix} k \\ g \\ k\overset{u}{\smile} \\ g\overset{u}{\smile} \end{bmatrix}$

2. $\begin{bmatrix} k\overset{u}{\smile} \\ g\overset{u}{\smile} \end{bmatrix} > \begin{bmatrix} k \\ g \end{bmatrix}$

3. $s > x \; / \; \begin{Bmatrix} i \\ u \\ r \\ k \end{Bmatrix} \; \_\_\_\_ \; \text{non-obstruer}$

4. $\begin{bmatrix} \overset{\frown}{k} \\ \overset{\frown}{g} \end{bmatrix} > \begin{bmatrix} s \\ z \end{bmatrix}$

5. $\begin{bmatrix} \left\{\begin{smallmatrix} o\breve{} \\ a\breve{} \end{smallmatrix}\right\} \\ \left\{\begin{smallmatrix} \bar{o} \\ \bar{a} \end{smallmatrix}\right\} \end{bmatrix} > \begin{bmatrix} o \\ a \end{bmatrix}$

6. $\begin{bmatrix} k \\ g \\ x \end{bmatrix} > \begin{bmatrix} \check{c} \\ \check{z} \\ \check{s} \end{bmatrix} / \_\_\_\_ \begin{Bmatrix} f \\ v \\ j \end{Bmatrix}$

7. $\begin{bmatrix} t \\ d \\ s \\ z \end{bmatrix} > \begin{bmatrix} \check{c} \\ \check{z} \\ \check{s} \\ \check{z} \end{bmatrix} / \_\_\_\_ \; j$

8. $j > l' / \left\{ \begin{matrix} p \\ b \\ m \\ w \end{matrix} \right\} \_\_$

9. $\left[\begin{matrix} i \\ e \\ u \end{matrix} \right] > \left[\begin{matrix} i \\ y \\ e \end{matrix} \right]$

10. $j > ' / \left\{ \begin{matrix} l \\ n \\ r \end{matrix} \right\} \_\_$

11. $\left[\begin{matrix} f \\ v \end{matrix} \right] > \left[\begin{matrix} e \\ o \end{matrix} \right] / \_\_ \; N \left\{ \begin{matrix} C \\ \# \end{matrix} \right\}$

12. $\left[\begin{matrix} e \\ a \\ o \end{matrix} \right] + \left[\begin{matrix} j \\ w \end{matrix} \right] > \left[\left[\begin{matrix} i \\ e \end{matrix} \right] \left[\begin{matrix} u \\ u \end{matrix} \right] \right] / \_\_ \left\{ \begin{matrix} C \\ \# \end{matrix} \right\}$

13. $\left[\begin{matrix} k \\ g \\ x \end{matrix} \right] > \left[\begin{matrix} c \\ z \\ s \end{matrix} \right] / \_\_ \left\{ \begin{matrix} i \\ e \end{matrix} \right\}$

14. $\begin{bmatrix} k \\ g \\ x \end{bmatrix} > \begin{bmatrix} c \\ z \\ s \end{bmatrix} / \text{f} \_\_\_ \text{þ}$

15. $\check{e} > a / \left\{ \begin{matrix} s\check{} \\ z\check{} \\ c\check{} \\ j \end{matrix} \right\} \_\_\_$

16. $\begin{bmatrix} or \\ ol \\ er \\ el \end{bmatrix} > \begin{bmatrix} oro \\ olo \\ ere \\ olo \end{bmatrix} / C \_\_\_ C$

17. $\begin{bmatrix} d \\ t \end{bmatrix} > \begin{bmatrix} \ \end{bmatrix} / \_\_\_ l$

18. $w > v$

19. $c > c' / \_\_\_ \text{f}$

20. $\begin{bmatrix} ę \\ ǫ \end{bmatrix} > \begin{bmatrix} a \\ u \end{bmatrix}$

21. $\check{e} > e$

Some of these rules appear in a less formal and explicit way in all standard texts of the history of the Russian language but none of them has all of these rules and in most cases there is little or no discussion of ordering.

3.0 Description of test. Approximately five hundred reconstructed Proto-Indo-European forms were chosen from Walde and Pokorny (1932). These were punched onto cards along with their English glosses. A separate Russian gloss was typed onto a print-out of the PIE forms. The transcription of Walde and Pokorny for the PIE lexical items was maintained as closely as possible, including such notations as subscript e and o. The only criterion imposed on choice of words was that they be as long as possible, so as to have a variety of environments.

The program was written in SNOBOL4 for the CDC 6400. Each rule set was numbered so as to coincide with the set of rules listed in section 2, with a zero appended to each rule number so as to allow for later insertions. Changing a rule consists at the moment of simple removal and replacement of cards. The history of a word or set of words can be gotten by allowing it to be processed with accompanying output generated by each rule set. Similarly, the lexicon for a particular stage can be generated by allowing the input to be processed up through the rule covering that stage and, if wanted, suppressing output from intermediate stages. With the availability of larger storage capacity the output from each stage can be generated once and stored in such a way that it can be referenced simply and thereby eliminate regeneration of input forms when the need for a rule change arises.

Frequency counters will be added in ofder to measure the functional load of a rule, at least in terms of dictionary

frequency. How this can be incorporated meaningfully into a theory of language change is not clear at this time.

The effect of borrowing can be simulated by introduction of lexical items just prior to a specific stage. There are too many variables involved in this case and the predictions have been poor. The effects of loss of original PIE are even more obvious but will require much further study.

4.0 Discussion of results. The program is obviously language dependent but the basic conception is of general applicability. The set of rules described in section 2.0 has been programmed and has successfully predicted the Modern Russian form from the PIE input in many cases including the following:

PIE	Mod. Russian
*$b^h el\hat{g}^h$-	boloz-
*$med^h i$-	mež-
*sloχg-	slug-
*ang$^{\mu}$^{h}i-	už
*$\hat{k}$orm-	sram
*$\hat{g}$omb^h-	zub
*g$^{\mu}$rūg^h-	griz-

The number of forms of PIE which were not related to Modern Russian was much greater. The main reason is probably due to loss (again assuming a uniform parent PIE as the sole source of the lexicon). The differences between generated

and actual Modern Russian could be accounted for in a few instances, for example *apsa did not become Mod. R 'osina' in part because there are no explicit rules in the program for the simplification of consonant clusters. But other incorrect outputs can not be accounted for in many instances by any easily accessible, documented rule. For example, there is no rule to handle the two disparate outputs from similar input forms with respect to the initial cluster 'sp-' in the forms

PIE	Mod. Russian
*sperg-	pr'ad-
*splen$\hat{g}^h$-	selez-

Similarly, there is a rule eu > u, for example, to account for *leut- becoming l'ud-, and others, but there are at least two exceptions to this rule:

*b^{h}reu$\hat{k}$-	bros-
*b^{h}leu-	bles-

Whether the original rule has too general an environmental condition or whether u was generated and later underwent some other changes, heretofore unpostulated, is unknown. It is possible that these forms should not have been considered as correspondences.

No case of an error in rule ordering has been found yet for this sample.

Examining the output is at times a forbidding task. It may help to simplify the discovery of causes of error by generating the output of each rule in the form of a KWIC index where the occurrences of each phone would be grouped together and the

environments then made quite clear. When input and output are then compared one might find more easily why a rule was not applied or why it should be generalized, etc.

5.0 Conclusions. This paper is necessarily meant to be only a preliminary progress report and as such has raised many other questions in addition to its concrete results. Some of these questions are very basic, in particular, since many of the output forms could not be accounted for, should one really attempt to generate forms of a modern language from reconstructed lexical items if the rules used are not those postulated during the process of reconstruction since the former should be a record of the latter. Given, therefore, a set of reconstructed forms and a separate set of rules, it becomes very difficult to account for the source of errors in the output. Also, the assumption of a uniform, single-stage proto language may require many restrictions.

The computer can under ideal conditions be successfully used in testing hypothesized changes in the history of a language, given certain simplifying assumptions. It can be expected to operate best when the rules and reconstructed proto forms are established by the same investigator, working within the bounds established by a general theory of historical change.

Bibliography

Kantor, Marvin and R. N. Smith, "A sketch of the major developments in Russian historical phonology" (to appear).

Lightner, Theodore M., Segmental phonology of modern standard Russian. MIT dissertation, 1965.

Schachter, Paul and V. Fromkin, A phonology of Akan: Akuapem, Asante and Fante. (Working Papers in Phonetics No. 9) UCLA, 1968.

Walde, A. and J. Pokorny, Vergleichendes Worterbuch der indo-germanischen Sprachen. Berlin, 1932.