Computerized tools for reconstruction in Tibeto-Burman

John Brandon Lowe, Martine Mazaudon

to appear in BLS 15 (1989)

Proceedings of the Fifteenth Annual Meeting of the Berkeley Linguistics Society, Feb 18-20 1989: 367-378

Computerized Tools for Reconstruction in Tibeto-Burman

John Lowe & Martine Mazaudon
*UC Berkeley & CNRS, Paris**

The methodology of reconstruction involves a series of steps, repeated cyclically, starting with the gathering of data from modern languages, their comparison, and the establishment of a list of potential cognates on the basis of similarity of form and meaning. This is analyzed to propose a set of sound correspondences in the modern languages, and a first hypothesis on the identity of the proto-phonemes which could be posited for each correspondence. Hypothetical proto-morphemes are then reconstructed. At this point it is advisable to verify that the modern forms can actually be regularly derived from the posited protoforms through the supposed sound-laws, in order to check the internal consistency of the model of proto-language and evolution hypothesized on the basis of the limited amount of data first analyzed. There follows a series of revisions of cognate matching, modern correspondences, reconstructed segments and reconstructed morphemes, the extent of which depends on the degree of closeness of the languages, and the intuition and experience of the linguist who made the first set of hypotheses. The cycle is repeated several times with more data.

Automating part of the process has two advantages. First, the internal consistency of the proposed model of proto-language and evolution can be checked more thoroughly. One might hope that it would also be done more easily and more quickly, but anyone with some experience of automation will know that this is not necessarily so. The process will be slower at first, but the result will be better, because the machine tolerates no inconsistencies, although uncertainty can be allowed for, as we will see below. In the later stages of the research, when more data is integrated and the model is already refined, some gain in the speed of analysis may be realized. The second benefit of automation arises after obvious cognates have been noted with the naked eye; at this point the computer can generate additional comparisons, based on form only, between words which have diverged in meaning. The acceptance or rejection of these suggestions lies entirely with the linguist at this stage, as no automation of semantic analysis has been developed. Figure 1 outlines our conception of this methodology and the computer objects with which we have implemented it.

This paper is a report on work in progress, and is thus tentative. Indeed, the project is by nature constantly evolving. Since we are developing a *research* tool, rather than a demonstration or teaching tool, all parts of the model, including the programs and the data files, are constantly revised, as they are progressively improved by being made more consistent with each other. Ultimately, when the model reaches perfection (in an ideal world), it ceases to be useful, as its flawless functioning would signify that it had come to constitute an adequate and complete description of the phonological evolution of the languages considered, and the study would be over! The protoforms at that point would be perfectly computed from the reflexes.

Methodological steps in reconstruction	Computer data files
• gathering of data from modern languages	• Dictionaries (in machine-readable form)
• comparison and establishment of a list of potential cognates on the basis of similarity of form and meaning	• Cognate file
• sound correspondences between the modern languages and potential proto-phonemes	• Table of Correspondences
• list of hypothetical proto-morphemes	• Cognate file (enriched with proto-morphs) or Proto-language file

Methodological steps in reconstruction	Computer programs	
• deriving modern forms from the proto-forms through the correspondences in a systematic way	• input: output:	proto-language file modern forms
• addition-of-data-cycle (more cognates by form and meaning)	• input: output:	dictionaries potential cognates by form
• refining reconstructions of individual etyma and proposing new ones	• input: output:	cognate file proto-forms

Figure 1: "Traditional" methods of reconstruction and their computer analogs

1.0 ENVIRONMENT

1.1 LINGUISTIC ENVIRONMENT

We began developing our model using a subgroup of Tibeto-Burman, the Tamang group of the Bodic Division of Sino-Tibetan in Shafer's classification, about which we know enough to be able to supply potential sound correspondences and reconstructions for most of the phonological system, and where there remain enough unsolved problems to provide an incentive to try new methods, and more seriously to give a real-life research situation where everything is not already known. [1]

We have taken full advantage of the basic monosyllabicity of the Sino-Tibetan morpheme, and initially restricted the data treated to monosyllabic forms. This allows the inclusion of almost all the verbs in the Tamang subgroup and about half the nouns, which is a sufficient amount of data to get significant results. We have also taken advantage of the more or less "isolating" structure of the subgroup, which means that morphemes combined in words remain basically unchanged (in particular we have no vowel harmony between root and suffix, as occurs in Eastern Himalayish languages).

On the other hand, we have had to accommodate some structural difficulties that might not have arisen in other language families (like Oceanic languages); in particular the sound laws are so sensitive to context that most authors writing on Sino-Tibetan diachronic phonology have adopted statements in terms of syllabic constituents instead of phonemes. [2]

1.2 PROGRAMMING ENVIRONMENT

The current implementation of the checking program, which we regard as a prototype, operates in the MS-DOS environment. The checking program is written in PC-SPITBOL, a version of SNOBOL 4. The Duke Language Toolkit provides the ability to display special characters; BLS and PRINT3 were used to produce hardcopy. Some of the input data was initially entered into HyperCard stacks on Apple Macintoshes. All data was converted into a common format for processing under MS-DOS. The format used is compatible with Lexware, a software package which assists in the processing of lexicographic data. Some of the data preparation tasks were accomplished using Lexware.

The initial design and development of the program was done on a Macintosh using HyperCard. However, to speed development we switched to SPITBOL, a programming language which provides extraordinary pattern matching capability and greatly reduced programming time. Also, as SPITBOL is a compiled language it has a substantial performance edge over HyperCard. [3]

2.0 THE MODEL: ASSUMPTIONS AND STRUCTURES

Structurally, our computer model is composed of two sets of data files (protoforms at one end and modern reflexes -- both computer generated and attested -- at other end), and an active component, which incorporates two pieces of linguistic information, the Table of Correspondences and the Syllable Canon.

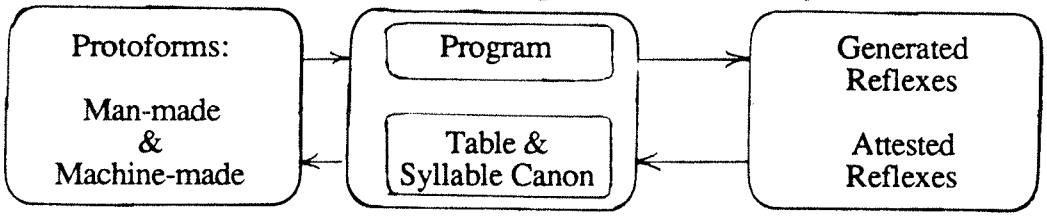

Figure 2: Components of the Model

2.1 TABLE OF CORRESPONDENCES

The Table of Correspondences reflects the linguist's conception of the etymological relationship between phonotactically significant portions of words in the languages compared.

The Table (see Figure 3) contains two types of information. First, on the right side, the correspondences observed between a number of related languages are stated in terms of phonemes, features, or groups of phonemes. Each line of the Table we will call a correspondence.[4] Secondly, on the left side of the Table, we list the reconstruction proposed for each correspondence [column 3], with the particular syllabic and phonological contexts where the correspondence can be observed [in columns 2 and 4 respectively]. The letter in the second column, which codes the phonotactic characteristics of the correspondence, links the correspondence to the Syllable Canon (which is described below in section 2.2). At the extreme left is a line number, which is used by the algorithm to identify the correspondence.

A question mark in a language column or two reflexes separated by a comma indicate an unknown reflex or unexplained variation.

Proto-Environment				Observed Correspondences in Modern Languages							
1	2	3	4	5	6	7	8	9	10	11	12
#	Slot	TGTM	Context*	Ris	Sahu	Tag	Tuk	Mar	Syang	Gur	Man
3	T	B	/_ U	z	z	z	z	z	z	z	z
5	T	B	/_ W	4	4	4	4	4	4	4	4
31	R,V	a		a	a	a	ə	ə	ə	a	ɤ
35	R,V	wa		wa	wa	wa	o	o	o	o	ɔ
36	R,V	wa	/t^h _	o	?	o	o	o	o	u	a
37	R,V	wa	/s_	wa	a	a,wa	ə	ə	o	a	ɤ
172	V	a	/ _p	a	a	a	ə	o	o	a	e
183	I,G	w		w	w	w	w	w	w	w	w
99	O	gr	/A _	kr	kr	w,h	ʈ	k	g	kr	kr
100	O	gr	/B _e	kr	?	h	?	?	?	?	?
101	O	gr	/B _i	k	k	k	ʈ	k	g	gr	hř,k^hr
102	O	gr	/B _	kr	kr	kr	ʈ	k	g	kr	hř=r^h
186	O	gr	/B _ w	k	ø	h	ʈ	k,ʈ	k,ʈ	kr	hř
90	I	g	/A _	k	k	k	k	k,g	k,g	k	k
91	I	g	/B _	k	k	k	k	k	g	k,g	k^h
131	L	r	/P _	r	r	r	r	r	r	r	r
159	I	r		r	r	r	r	r	r	r	r

* U = [-voiced], W = [+voiced], P = [+labial], A B 1 2 3 4 are tones

Figure 3: Excerpt from the Table of Correspondences

2.2 SYLLABLE CANON

The syllable canon is a statement of the phonotactics of monosyllabic morphemes in the proto-language.[5] It provides a means for us to analyze syllables into distributionally defined constituents. We call these "slots" and the members of a slot "fillers." Note that fillers need not be individual phonemes but may be comprised of sequences of phonemes. "Slots" may be suprasegmental (e.g. tone). The occurrence of a particular slot in a syllable may be optional, in which case the slot is parenthesized. Square brackets are used to indicate that one of the slots contained within must occur. However, a null element is possible, and this is indicated by a comma followed by nothing.

The "slots" which we have presently retained as potentially significant for the evolution of the Tamang group of languages are the following:

O = onset	L = liquid {r,l}	R = rhyme	F = final consonant
I = initial consonant	V = vowel	G = glide {j,w}	T = tone

and the Proto-Tamang Syllable Canon has been established as in Figure 4: [6]

[T,][O(G),I(L)(G),][R,VF]

Figure 4: Proto-Tamang Syllable Canon

To interpret, syllables may or may not have tone (depending on whether they are root or suffix, and on whether we *want* to consider tone or not at a given stage in the study); a syllable is minimally composed of either a Rhyme or the sequence of a Vowel and a Final consonant; this can be preceded by an Onset with optional Glide, or an Initial consonant with optional Liquid or Glide, or nothing.

A morpheme may have more than one analysis according to the canon for reasons we will clarify in section 3.1 below. For example, the parsing of the word *Bgrwat 'hawk,eagle' could be realized in several ways, as shown in Figure 5:

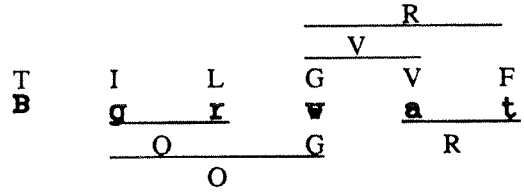

Figure 5: Parsing according to the Syllable Canon

2.3 THE ALGORITHM

The algorithm we use to propose forms is "bidirectional". That is, it can either:

- given a reconstruction (a "protoform"), compute the expected or regular reflexes (outcomes) in daughter languages.
- given a form in one of the modern languages, propose reconstructions which would produce the form as a regular reflex.

Computing reflexes from a protoform

In both cases, the same procedures are followed, at least conceptually. We shall consider the generation of reflexes from a reconstructed form first. As an example consider the parsing of the proto-form *Bgrwa 'wheat' according to the Table of Correspondences and Syllable Canon stated above.

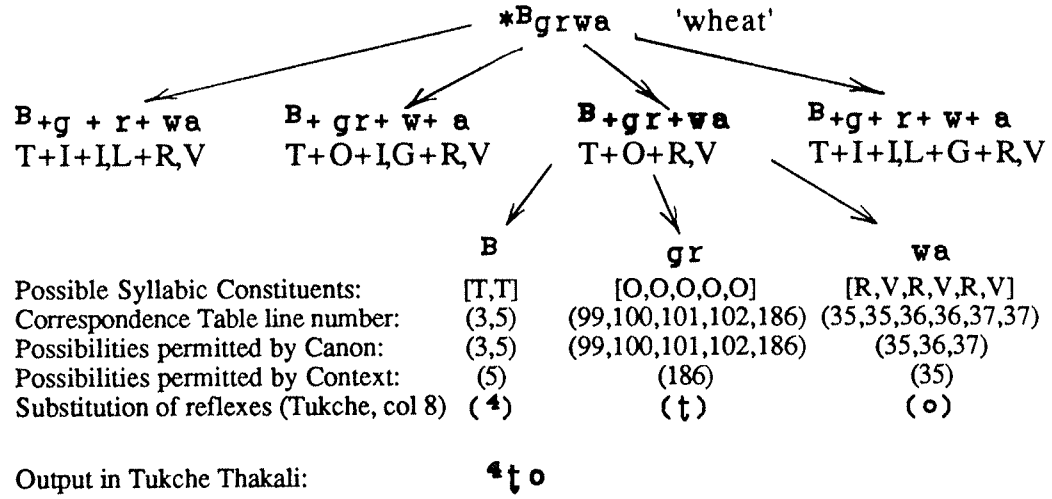

Figure 6: "Tokenization" of a protoform

Starting at the left of the form, the computer selects successively longer initial substrings of the given protoform and attempts to match them with one or more entries in the "TGTM column" of the Correspondence Table. When a match occurs, the matching correspondence sets (i.e. rows of the Table) are recorded by their row numbers and the substring is removed from the form. The matching process is repeated recursively with the remaining portion of the form. In this way all parsings which are permissible according to the table are produced. The parsing of the protoform results in zero or more "tokenized" versions of the original form. Each token in a tokenized form is composed of pointers to one or more rows of the Correspondence Table (represented in Figure 6 as parenthesized lists of row numbers); that is, a segment of the protoform may have different reflexes in daughter languages depending on the syllabic and phonological environment. Next, each element of each token is combined with the elements of the other tokens comprising the tokenized form. Combinations which would result in a violation of the Syllable Canon or the phonological context for the element are eliminated. Note that permitted phonological environments and permitted syllable types are stated exclusively in terms of the proto-environment. To produce reflexes in daughter languages, the row numbers are replaced with the corresponding phonological elements from each language's column.

Comparing generated forms with attested forms

In this way, for each reconstructed etymon, we derive one or more reflexes for each daughter language. This output is compared to the attested forms, and those sections of the Table and Syllable Canon which permit the generation of forms contradicted by the data are corrected or eliminated.

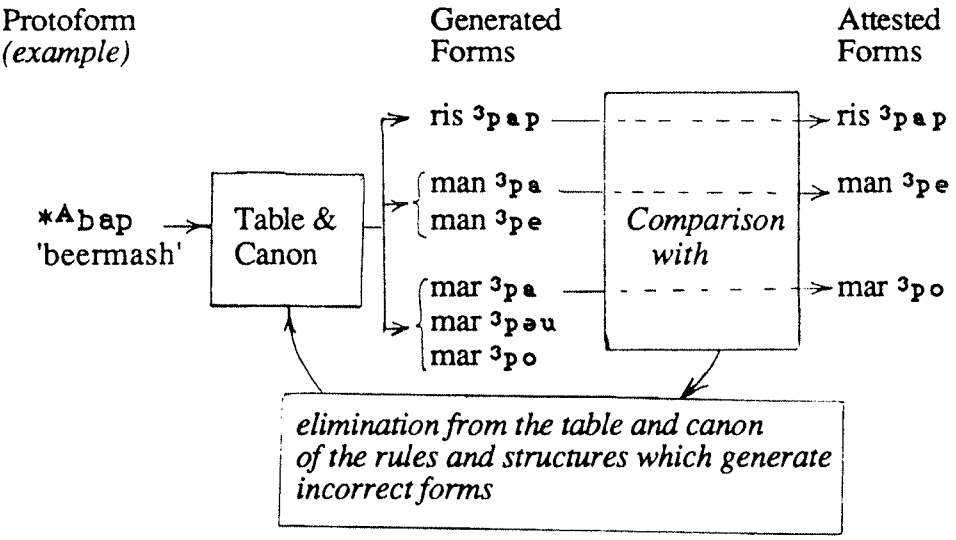

Figure 7: Computing reflexes from a protoform as a means of checking the model

After a number of repetitions of the cycle illustrated in Figure 7, the algorithm should generate all and only the attested forms.

Computing protoforms from reflexes

Conversely, the program can also "reverse engineer" protoforms on the basis of one or more reflexes in modern languages. In this case, the computer

recursively tokenizes the modern form according to the reflexes listed in that language's column of the correspondence table. In this way, lists of protosegments from which the daughter reflexes may have derived are accumulated. By combining the elements of the tokenized forms as described above, a set of protoforms is generated. Once again, only those forms which meet syllabic and contextual constraints are retained.

The program generates and stores a separate set of possible proto-forms for each daughter language reflex. Then, if this process is applied to a set of presumed cognates in several languages, the program can compute the "set intersection" of the sets of possible protoforms and determine which protoforms (if any) produce all the reflexes through regular correspondences (see Figure 8).

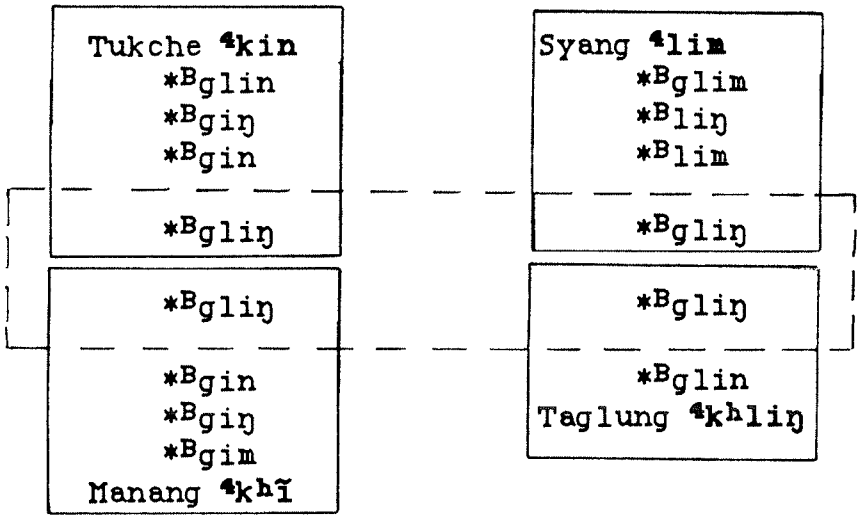

*Figure 8: "Triangulating" on protoforms: *Bgliŋ 'snow' in Proto-Tamang*

The problem of computational complexity:

The drastic reduction in the complexity of TB monosyllables in general has an unfortunate consequence in the computation of protoforms: since modern reflexes may derive from any of a large number of protosegments, it may be necessary to generate a large number of reconstructions. For example, Gurung `pli` requires that over 200 protoforms be tested. Of course, syllabic and contextual constraints reduce the number of valid reconstructions in some cases. Ideally, no matter how many forms are evaluated, only the possible ones according to the model should be retained. This, however, assumes that the Table and algorithms are nearly perfect. Thus, this combinatorial explosion poses a significant problem both for the computer, which must have the processing and memory capacity to test all the alternatives, and for the linguist, who must deal with this "noise" in the output data. The performance problems are purely technical and can be solved if necessary with supercomputers. The "noise" problem may require the development of other heuristics, such as global rules, to reduce the volume of output.

2.4 THE DATA

The program has both an interactive mode and a batch mode: input can come either from the keyboard or from data files. A limited amount of data preparation is required for batch mode operation.

• The *transcription* of each language input file has to be consistent with the one used in that language column in the Table and must be linearized in a way that is consistent with the Syllable Canon.

• The proliferation of generated forms (mentioned above and illustrated in Figure 7) makes it advisable to reduce *redundancy* in the input data as much as possible. In semantically arranged word lists, commonly used in comparative work, polysemous items may be listed multiply, without being distinguished from true homonyms (which have a different etymology). The preferred data format for input in our algorithm is thus the classical dictionary format, where each entry reflects the linguist's analysis of the item as a single lexeme. [7]

• We cannot expect to have perfect data on all the languages involved before starting reconstruction, so some incompleteness and imprecision should be tolerated. But the *imprecision should be precisely coded.* For instance tones ¹ and ² are not distinguished in the source on Tukche Thakali for monosyllabic nouns and are both transcribed as tone ¹. We preprocessed this file, replacing tone ¹ in that context by tone ᴴ (i.e indistinct "high" tone).

Of course, this type of pre-treatment of the data implies a careful philological evaluation of the sources. It should be emphasized also that using the output of the programs in a meaningful way requires a good familiarity with the linguistic data treated.

3.0 LIMITATIONS AND IMPROVEMENTS

The computer tools we have developed so far only claim to check the internal consistency of a linguistic model which states 1) that a set of modern languages are genetically related, 2) that they can reconstruct at the level of a common ancestor in the guise of a given list of protoforms, and 3) that *one way* that this relationship can be described is via the Table of Correspondences and Syllable Canon.

While the development of a checking tool remains our primary goal, we believe that with our model we can test the efficacy of a variety of theoretical approaches to the data with a minimum of apparatus. It is clear to us that the ways in which correspondences are coded in the Table, for example, reflect choices made on the basis of some sort of theoretical framework. We would like that framework to mirror linguistic reality insofar as we understand what that reality is.

3.1 ANALYSIS OF THE TABLE STRUCTURE

3.1.1. Phonemes, clusters, features

In the Table, we have used two different types of statements simultaneously: correspondences on the basis of individual phonemes and correspondences in terms of "slot-fillers", which allow clusters of phonemes. One of our aims is to discover if there is any motivation for choosing one representation (or analysis) over the other. Analysis on the basis of phonemes reflects the "classical" segmental approach favored in Indo-European linguistics; the Onset + Rhyme analysis is typically used by Asianists. The "Asianist" approach reflects an underlying position against the strict localization of features on individual phonemes. It may be of interest to note that "prosodic" theory (in the Firthian sense) has received its best support from Asian languages.

A rigorous analysis of the internal structure and economy of the Table would help discover whether or not the differences of approach are motivated by differences in the typology of the languages themselves. Our research on this point

could make a modest contribution towards realizing the goal, mentioned by Prof. Emeneau in his contribution to this volume, of making an "explicit study of the typology of rules".

Figure 9 presents the "classical" representation of the evolution of PTam *^{B}grwa by means of four segmental rules.

Proto-Environment				Observed Correspondences in Modern Languages							
1	2	3	4	5	6	7	8	9	10	11	12
#	Slot	TGTM	Context	Ris	Sahu	Tag	Tuk	Mar	Syang	Gur	Man
91a	I	g	/^{B} _r	k	ø	h	t̪	k	k	k	h
131a	L	r	/k,k^{h},g _w	ø	ø	ø	ø	ø	ø	r	ř
31a	R,V	a	/w _	a	a	a	o	o	o	o	ɔ
183a	G	w	/_a	w	w	w	ø	ø	ø	ø	ø

*Figure 9: Alternative statement of the correspondences for *grwa*

Compare this with the synthetic "cluster" versions presented in correspondences 186 and 35 cited in Figure 3. In this case no generalization is gained by the "analytical" coding, since the two interacting rules in each pair are so specific that the input for one is the context for the other and vice versa.

At another level of analysis, considering the example in Figure 10 below, we may observe that if we were able to represent correspondences in terms of *features*, instead of phonemes, we could state 121, 126 and 91 as a single rule describing the devoicing of initial stops in Gurung. It would remain to be decided whether or not such features should be strictly linearized, as in the "classical" model.

3.1.2 Exceptions

Including or excluding a correspondence from the Table reflects another theoretical decision: how specific can a rule be and still count as a sound-change rather than a single word history? Or to say it differently, when do you have to move out of internal causality (inside the phonological system) and look for external causality (analogy, language contact)? Simple statistics will not answer this question. A genuine sound-change sequence may be exemplified only once because data is insufficient, especially in the case of a complex sequence of multiply conditioned changes. An example would be Gur ^{1}p^{h}rɛ̃ 'young man', regularly derived from Proto-Tamang *^{B}bjon by such a long list of context-sensitive rules that the only convenient way to code it in the Table would be as a whole syllable!

3.2 COMPLEXIFICATION OF THE MODEL

The development of PTam *dut into Gur ^{4}ti 'to pick up' (Figure 10) illustrates another flaw of our model: it assumes that sound changes occur in one step. This unrealistic view leads to the postulation of *ad hoc* unnatural rules like 121c to ensure proper output from the reconstructed input.

Certainly, rules 121c and 77 do give the correct result (as illustrated in the "one-step" model), but 121c is totally *ad hoc*. Why should the devoicing of initial stops (rules 120, and 121a) [paralleled by similar rules for labials and velars],

which is blocked before a vowel under proto-tone *B (rule 121b) [also paralleled by similar rules for labials and velars], fail to be blocked only for the dental stop, and only in the context of the proto-rhyme *ut (rule 121c)? The chronological succession of changes symbolized in (b) is the only way to make sense of this development. Note that (17) and (75) are independently required and context-free.

Line #	Slot	TGTM	Context *	Ris	Sahu	...	Gur	Man
17	R,V	wi					i	
75	R	ut					wi	
77	**R**	**ut**	**/t,d _**				**i**	
120	I	d	/A _				t	
121	I	d	/B _				t,d	
121a	I	d	/B _ G,L				t	
121b	I	d	/B _ Vowel				d	
121c	**I**	**d**	**/B _ ut**				**t**	
126	I	b	/B _				p,b	
91	I	g	/B _				k,g	

* G = glide, L= liquid

(a) *Bdut $\xrightarrow{(77 + 121c)}$ 4ti 'one-step' model *vs* (b) *Bdut $\xrightarrow{(75)}$ *Bdwi $\xrightarrow{(121a)}$ *4twi $\xrightarrow{(17)}$ 4ti 'real-life progressive change'

*Figure 10: The chronology of sound change: TGTM *Bdut > Gur 4ti*

The modifications required to incorporate this reality into the model could be symbolically represented by a sequence of Tables, which would parallel the chronology of changes.

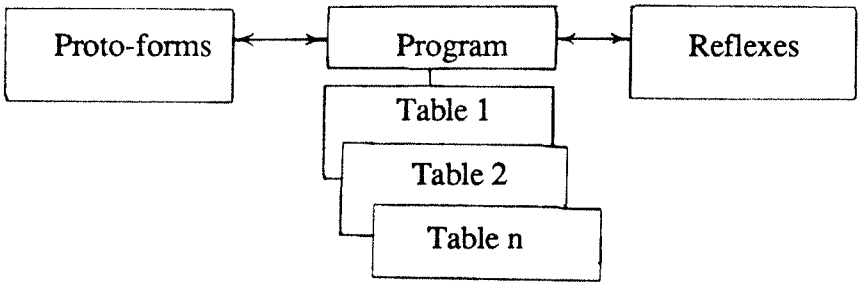

Figure 11: Complexification of the model

4. CONCLUSION

While it may seem at first that turning to automation to help in linguistic research implies that we expect linguistic processes to follow exceptionless, rigid principles, we hope we have shown that quite the opposite is true, and that, in our view, the computer is most useful as a means of keeping track of a large number of possibly divergent hypotheses and observations, permitting us to delay making generalizations until such time as they are justified by an adequate analysis of a sufficient amount of data.

Notes

* The authors gratefully acknowledge the advice and assistance given by Robert W. Hsu and Boyd Michailovsky. Some of this material is based upon work done at the Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) Project, supported by the National Science Foundation under Grant No. BNS-867726 and by the Division of Research Programs of the National Endowment for the Humanities, an independent federal agency, under Grant No. RT-20789-87. Additional support was also provided by the Centre National de la Recherche Scientifique (CNRS), Paris, through its Laboratoire des Langues et Civilisations à Tradition Orale (LACITO).

After the completion of this paper and its presentation at BLS 15, it came to our attention that a similar project had been conducted on Algonquian languages by John Hewson, at the Memorial University of Newfoundland (Hewson 1974).

1. Shafer refers to this group as the"Gurung Branch" of the Bodish section of the Bodic division. We have established that it covers four ethnic groups, Tamang, Gurung, Thakali, and Manangba or Manangke, whence the short-hand denomination of the language group as TGTM in the Table below. Dialects are designated by an abbreviation of the name of the village where they are spoken: ris. Risiangku, sahu. Sahugaon, tag.Taglung, tuk.Tukche, mar. Marpha, Syang (unabridged), gur. Gurung (Ghachok village), man. Manang. For more details see Mazaudon 1978.
2. See Matisoff's description of the canonic form of ST morphemes as "bulging monosyllables".(Matisoff 1989)
3. MS-DOS is a registered trademark of Microsoft Corporation. PC-SPITBOL is a licensed product of Realia Corporation, copyright R.K.B. Dewar 1983, 1984. Apple, HyperCard, and Macintosh are registered by or trademarks of Apple Computer, Inc. Lexware was written by Robert W. Hsu at the University of Hawaii.
4. The transcription follows the principles of the IPA. The phonetic font used in this paper was designed by Stephen P. Baron for STEDT.
5. When the model is expanded, a canonical form for polysyllabic morphemes will be established, which will probably not be a sequence of syllabic canons.
6. This is not yet completely accurate, as a structure OGG has to be allowed in limited cases, unless a new composite filler for the G slot (G = jw) is introduced in the Table.
7. The algorithm pretreats redundant data by compacting homophonous entries (whether polysemous, or genuine homonyms) into single entries with a list of different meanings. This reduces computing time, but does not help with the "noise" generated by the lack of prior lexical analysis. Thus the usefulness of the ouput *is* dependent on the quality of the input (mass still does not compensate for quality ...)

References

Emeneau, Murray, 1989, Phonetic Laws and Grammatical Categories, *BLS* 15.

Hewson, John, 1974, Comparative Reconstruction on the Computer, in J.M. Anderson and C. Jones, eds, *Proceedings of the First International Conference on Historical Linguistics,* Edinburgh 2-7 September 1973, Amsterdam: North-Holland.

Matisoff, James A., 1989, Bulging monosyllables: Areal tendencies in Southeast Asian Diachrony, *BLS* 15.

Mazaudon, Martine, 1978, Consonantal mutation and tonal split in the Tamang sub-family of Tibeto-Burman, *Kailash* 6:3, 157-180, Kathmandu.

_____, 1988, "The influence of tone and affrication on manner: some irregular manner correspondences in the Tamang group", paper presented at the 22nd Conference on Sino-Tibetan languages and linguistics, Lund.

Shafer, Robert, 1955, "Classification of the Sino-Tibetan languages", *Word* 11, 94-111.