# Cross-linguistic Lexicographic Databases for Etymological Research, with Examples from Sino-Tibetan and Bantu Languages

by

John Brandon Lowe

A.B. (Yale College) 1977

M.L.I.S. (University of California at Berkeley) 1987

M.A. (University of California at Berkeley) 1989

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Linguistics
in the
GRADUATE DIVISION
of the
UNIVERSITY of CALIFORNIA at BERKELEY

Committee in charge:

Professor James A. Matisoff, Chair
Professor Gary B. Holland
Professor Bruce R. Pray

1995

The dissertation of John Brandon Lowe is approved:

James A. Matisoff 5/18/95

James A. Matisoff, Chair Date

Gary B. Holland 5/18/95

Gary B. Holland Date

Bruce R. Pray 5/18/95

Bruce R. Pray Date

University of California at Berkeley

1995

# Cross-linguistic Lexicographic Databases for Etymological Research, with Examples from Sino-Tibetan and Bantu Languages

by

John Brandon Lowe

To

my mother

Thelma Larson Lowe

## List of Symbols and Abbreviations

| | |
|---|---|
| ≍ | separates 'co-allofams' |
| § | Chapter or section |
| *LB (cf. *LB) | Proto-Lolo-Burmese, a protolanguage |
| *TGTM | Proto Tamang-Gurung-Thakali-Manang |
| ahi | Ahi |
| ak | Akha |
| ASCII | A standard code for representing roman characters, numbers, punctuation, etc. |
| bi | Bisu |
| CBOLD | Comparative Bantu Online Dictionary (project) |
| CL | Classical Latin |
| Codepoint | Numeric value assigned to a glyph |
| Env | Environment (phonological) |
| gha | Ghachok |
| H | High tone, usually the high-stopped tone |
| IE | Indo-European, a language group |
| L = *L | Proto-Loloish, a protolanguage |
| L | Loloish, a protolanguage |
| L | Low tone, usually the low-stopped tone |
| LB (cf. *LB) | Lolo-Burmese, a language group |
| lc | Luqüan |
| lh | Lahu |
| li | Lipho |
| mar | Marpha |
| na | Nasu |
| Ø | Linguistic zero |
| OALD | Oxford Advanced Learner's Dictionary |
| OCR | Optical Character Recognition |
| OE | Old English |
| OS | Operating system |
| PAF | Pan-allofamic formula, a way of show alternations in the shape of etyma |
| PGmc | Proto-Germanic |

| | |
|---|---|
| PIE | Proto-Indo-European, a protolanguage |
| pra | Prakaa |
| PST | Proto-Sino-Tibetan, a protolanguage |
| PTB = *TB | Proto-Tibeto-Burman, a protolanguage |
| RE | Reconstruction Engine |
| ris | Risiangku |
| sa | Sani |
| sahu | Sahu |
| SGML | Standard Generalized Markup Language |
| ST | Sino-Tibetan, a language group |
| STC | Sino-Tibetan: a conspectus (Benedict 1972) |
| STEDT | Sino-Tibetan Etymological Dictionary and Thesaurus (project) |
| syang | Syang |
| tag | Taglung |
| TB (cf. *TB) | Tibeto-Burman, a language group |
| TEI | Text Encoding Initiative |
| tuk | Tukche |
| Unicode | A standard for encoding scripts and characters |
| WB | Written Burmese |
| WGmc | West Germanic |
| wo | Woni |
| WT | Written Tibetan |
| WWW | World Wide Web, a part of the internet |
| ZMYYC | *Zang-Mian yuzu yuyin cihui [A Tibeto-Burman lexicon].* Dai Qingxia 1992. |

## Acknowledgments

Many people have helped me in many different ways to reach this point in my research. There are far too many to include here; I wish specifically to thanks Jim Matisoff, Gary Holland, and Bruce Pray for their sympathetic support and encouragment.

I also wish to thank Martine Mazaudon; without her most of the software and the thought that went into it would not have occurred. Boyd Michailovsky, linguist and programmer, provided much good advice. I am also indebted to David Bradley George van Driem, Larry Hyman, for teaching me what I know about Bantu (so far!), Inga-Lill Hansson, David Solnit, Sun Hongkai, Sun "Jackson" Tianshin, Dai Qingxia (who provided me with many insights into the Sinospheric side of Tibeto-Burman); Chen Kang (whose meticulously recorded data has been a wonderful asset in my research). My collegues and friends at STEDT have been great to work with; I wish to thank Randy LaPolla, ex-STEDTnik and friend; several other STEDTniks have given me aid and advice over the years, Zev Handel, Jonathan Evans, and Leela Bilmes.

I was fortunate to have suberb editorial help from Orin Gensler, Nancy Urban, Sara Gesuato, Françoise Peters (CNRS/LACITO, Paris), and Richard Schaedel.

The computer side of the research benefited greatly from discussions with and programming assisitance from Dan Jurafsky, Mike Brodhead, Gene Gragg, Robert Nicolaï, and Joel Brogniart.

Several of my friends, in particular, Alex Hewitt and Gordon van Kessel have seen me through the ups and downs of the process, providing moral support and all the intangbles that make it possible to get some work done,

I wish also to thank the following institutions for their support: the STEDT project, LACITO/CNRS, CBOLD, DYLAN, and the Luce Foundation.

Some of the material included here was prepared with the support of grants from the National Science Foundation under Grants No. SBR-9319415, BNS-867726, BNS-90-911918, and FD-92-09841; and from the Division of Research Programs of the National Endowment for the Humanities, in independent federal agency under Grant Nos. RT-20789-87, RT-21203-90, and RT-21420.

# 1. THE COMPARATIVE METHOD AND CROSS-LINGUISTIC LEXICOGRAPHIC DATABASES

## 1.1. The comparative method: aıı endangered methodology?

In recent years a debate about the efficacy and even validity of the comparative method has flourished intermittently in the pages of journals and more recently on the email lists of the internet. The goal of the research presented here is to show how computational techniques, and in particular database techniques, can be used to refute or at least clarify some of these challenges.

Challenges to the method have come principally from three directions:

First, the theoretical underpinnings of the method have been questioned by linguists seeking a simpler and more unified motivation for phonological change (Wang 1969, Wang and Cheng 1977, etc.). These linguists have proposed that the aggregate of partial theories which underlie traditional explanations of diachronic changes (i.e. phonetically motivated sound change, borrowing, and analogy) be replaced with a theory based on a single stochastic process (lexical diffusion). These claims are examined in detail in §1.6.7.

Secondly, the necessity for the methodological rigor normally employed has been challenged. Asserting that an abundance of data can overcome the need for precise correspondences and reconstructions, some linguists (Greenberg et al.) have classified the languages of Africa and the Americas using the method of 'mass comparison' (also called *megalocomparison*Matisoff 1990a); see Greenberg (1949), Greenberg (1957), Greenberg (1960), Greenberg (1987). The likelihood of several languages having similar words purely by chance is higher than one might think (Ringe 1992:1). These claims are examined in §1.6.1, §1.6.2, and §1.6.3. It has always been difficult to make a principled judgment between competing hypotheses; experienced linguists can sometimes discern the patterns of relationship between language groups long before the evidence and research are complete enough to support their guesses. Thus, Benedict's method of *teleoreconstruction* (Benedict 1976, Benedict 1973) contrasts methodologically with methods of *microreconstruction* (discussed in detail here with regard to the Tamang and Loloish subgroups of Tibeto-Burman, §6) and *macroreconstruction.* The accuracy and application of these different approaches are discussed in §1.6.9.

Finally, there have been attempts to improve the utility and accuracy of the results obtained by applying the comparative method. The lexicostatistical experiments of Swadesh and Dyen may be interpreted in this way inasmuch as they assume the results of traditional reconstruction as the input to their statistical model of genetic affiliation (Dyen 1969). It is unfortunate that their other assumptions were so faulty as to compromise their method as a whole. The relation of traditional comparative linguistics to lexicostatistics is dealt with briefly in §2.1.

This thesis is devoted to improving the comparative method and to refurbishing its image as a scientific method. I wish therefore to make explicit the theoretical assumptions and methodological principles which support the traditional method and, secondly, to show how computer technology might improve its ability to withstand recent challenges. In the process I will show how and why these challenges arose, and how they might be directly addressed.

The discussion will proceed as follows:

**Chapter One** outlines the basic elements of the comparative method. It recounts briefly the development of the comparative method as a tool in historical linguistics and rehearses arguments dealing with philosophical, theoretical, and methodological issues, including the notion of scientific proof, the structure of theories of language change, the neogrammarian regularity hypothesis, and issues relating to the stages or levels of reconstruction (i.e. mesolanguages). It also suggests how *cross-linguistic lexicographic databases* might contribute to answering some of the current questions about comparative reconstruction and linguistic affiliations.

**Chapter Two** reviews past and present computational efforts to perform computer-aided diachronic research. These efforts range from speculations about how to implement facets of historical research on a computer to full-fledged database development efforts.

**Chapter Three** discusses issues involved in *creating cross-linguistic lexicographic databases* for etymological research, especially issues related to converting linguistic data into a machine-readable format suitable for comparative work. Included here are problems of transcription, extracting useful information from extant sources, comparison of data across languages and across sources, and strategies for organizing and cross-referencing the data in useful ways.

**Chapter Four** exemplifies some of the principles presented in the previous chapter using as examples several efforts at converting lexicographic sources of Sino-Tibetan and Bantu languages.

**Chapter Five** discusses strategies implementing some of the principles of the comparative method to *create hypotheses from data* contained in cross-linguistic lexicographic databases. These should not be understood as discovery tools, but as computational heuristics which help organize the possibly large amount of data in a way suitable for evaluation.

**Chapter Six** discusses strategies for *testing* arrangements of data in cross-linguistic lexicographic databases. Treated here are techniques for organizing data into semantic, phonological, etymological, and genetic categories. Exemplification is given from the Sino-Tibetan Etymological

Dictionary and Thesaurus project (STEDT), the Comparative Bantu Online Dictionary project (CBOLD), and the Reconstruction Engine (RE) development project.

**Chapter Seven** sketches a design for a *sound law database* operating with a cross-linguistic lexicographic database which could assist in carrying out the types of research discussed in the earlier sections.

**Chapter Eight** concludes by speculating on the prospects and consequences of converting all or most of the available lexicographic sources in the world into machine-readable form. This is not as daunting a task as it seems, and at any rate, only a portion of any languages vocabulary is germane to the search for etymologies.

## 1.2. The comparative method as an arrangement of data

The exegesis of the comparative method in most elementary linguistic texts usually begins by supplying sets of cognates from two or more languages as in (1), (2), and (3) below. The sets usually 1) demonstrate the relationship between languages on the basis of 'core vocabulary' items; 2) illustrate interesting features of the hypothesized ancestor or of the common development of the daughter languages; and/or 3) consist of multiple 'parallel forms' from related languages, in which each set contains at least one item exemplifying a given proto-phoneme or class of phonemes. These concepts may be illustrated by the following tables of cognates taken from well-known linguistics texts:

*(1) Core vocabulary : Romance numerals (Meillet 1967:15)*

| *Gloss* | *French* | *Italian* | *Spanish* |
|---|---|---|---|
| 1. | un, une | uno, una | uno, una |
| 2. | deux | due | dos |
| 3. | trois | tre | tres |
| 4. | quatre | quattro | cuatro |
| 5. | cinq | cinque | cinco |
| 6. | six | sei | seis |
| 7. | sept | sette | siete |
| 8. | huit | otto | ocho |
| 9. | neuf | nove | nueve |
| 10. | dix | dieci | dies |
| 20. | vingt | venti | veinte |
| 30. | trente | trenta | treinta |
| 40. | quarante | quaranta | cuarenta |
| 100. | cent | cento | ciento |

*(2) Examples of Germanic obstruents (after Hock 1986:37)*

| *PIE* | *Gothic* | *Old English* | *gloss* |
|---|---|---|---|
| *pətḗr | fadar | fæder/fader | father |
| *tréyes | þreis | þrī | three |
| *ḱm̥tóm | hund | hund(rāþ) | hundred |
| *déḱm̥(t) | taihun | tēon | ten |
| *ǵews- | kiusan | cēosan | choose |
| *bher- | bairan | beoran | bear |
| *dhē- | (ga-)dēps | dǣd | deed |
| *ǵhew- | giutan | gēotan | pour |

(3) *Correspondences between Germanic obstruents, supporting Grimm's Law (after Hock 1986:37)*

| *PIE* | | | | *PGmc.* | | |
|---|---|---|---|---|---|---|
| p | t | k | | f | þ | x/h |
| (b) | d | g | ⇨ | (p) | t | k |
| bh | dh | gh | | b/β | d/ð | g/ɣ |

(4) *Parallel forms (illustrating the set of proto-medial consonants) (after Bloomfield 1933:*310*)*

(a) List of cognate sets

| | *gloss* | *Tagalog* | *Javanese* | *Batak* | *Primitive Indonesian* | |
|---|---|---|---|---|---|---|
| (1) | choose | 'pi:li? | pilik | pili | *pilik | |
| (2) | lack | 'ku:laŋ | kuraŋ | huraŋ | *kuLaŋ | |
| (2) | nose | i'luŋ | iruŋ | iguŋ | *iguŋ | |
| (4) | desire | 'hi:lam | iDam | idam | *hiDam | 1 |
| (5) | point out | 'tu:ru? | tuduk | tudu | *tuduk | |
| (6) | spur | 'ta:ri? | tadi | tadi | *tadi | |
| (7) | sago | 'sa:gu | sagu | sagu | *tagu | 2 |
| (8) | addled | bu'guk | vu? | buruk | *buɣuk | |

(b) Correspondences for the medial consonant in the forms above[3]

| *Tagalog* | *Javanese* | *Batak* | *Primitive Indonesian* |
|---|---|---|---|
| l | l | l | l |
| l | r | r | L |
| l | r | g | g |
| l | D | d | D |
| r | d | d | d |
| r | d | d | d |
| g | g | g | g |
| g | ? | r | γ |

By induction, various sound correspondences are extracted from the data. Once correspondences are in hand, an argument of some sort is presented for reconstructing one or another protophoneme for each correspondence set (usually based on the notion of 'majority rules,' 'phonetic plausibility,' or typological considerations such as 'system congruence').

Finally, reconstructed word-forms or morpheme-forms (etyma) are built up from the reconstructed protophonemes (though in some cases the etyma are provided 'up front').

Two important features of expository presentations of this type should be noted.

The first is the implicit assumption of the neogrammarian theory of exceptionless sound change, or some version of it. Without this assumption there is no justification for extracting correspondences or recreating unique ancestor forms. This theory of phonetically based sound change is discussed in more detail in §1.5.6 below.

The second feature of such procedures is that they present the comparative method simply as an *arrangement*. That is, the presentation does not detail or explain the processes of selecting particular forms for comparison and of extracting the *comparanda* (the exact material to be compared) from compound, inflected, or derived forms; nor usually are the fine details of phonetic transcription which permit or obscure the cross-linguistic phonological comparison discussed as they are assumed to be part of another discussion.[4]

Now these are not insignificant omissions; after all, '[t]he first law of comparative grammar is that you've got to know what to compare' (Watkins 1976:312). But such presentations often start from the final reconstructed result and work backwards, introducing the apparatus of analysis and comparison in the process of motivating and justifying that result.

Such expository procedures typically take up and work through (usually in this order) a number of interlocking pieces of evidence. I describe these below in a definitional way (shown in SMALL CAPS); the intention is not to introduce new jargon but to use the conventional terms in a precise way, especially as a number of them can be ambiguous.

- COGNATE SETS; here the term cognate set has a precise meaning as a list of modern or attested forms presumed to be COGNATE, i.e. descended from a common ancestor. The set may or may not be headed by a reconstruction (defined below). To avoid confusion, a set of cognate sets will be called a COGNATE SET LIST. Cognate lexical items from modern or attested languages will be called REFLEXES.

- CORRESPONDENCES; here the term correspondence is used to mean an observed or proposed equivalence between pieces of attested forms. The pieces may be features, segments, or larger units: corresponding elements between languages can be of almost any size or type. The pieces will be called OUTCOMES. Correspondences are often conceived of as pairs of elements (a:b) as this is supposed to represent the simplest case of comparison; here, however, the 'n-ary' case in which several languages are compared simultaneously is the default and will be referred to as CORRESPONDENCE ROWS. A group of correspondence rows which treats the complete phonologies of the ancestor and daughter languages (i.e. has ancestors for each phonological element in each daughter language and daughters for each ancestor) will be called a TABLE OF CORRESPONDENCES.

- RECONSTRUCTIONS; here the term reconstruction and etymon will be used more or less interchangeably for the formula or entity used to represent the ancestor of a group of words. ETYMON will generally be preferred (this avoids confusing 'reconstruction' the *item* with 'reconstruction' the *process* by which cognate sets, correspondences, and etyma are created); I should note also that 'reconstruction' can also stand for the *entire result* of the process (i.e. all the bulleted items defined or described here). The constituent pieces of an etymon (i.e. the protolanguage analogue of an outcome) will be called a PROTOCONSTITUENT.[5]

- EXPLANATIONS of semantics, morphology, and phonology.

- LISTS OF EXCEPTIONS, with apologias (based on sporadic processes); the notion of exception is a complex one. In general it will be used to refer to those forms which fail to be accounted for by the explanatory apparatus in use, and therefore stand as potential counterexamples or theory-breakers.[6] Note that ISOLATES, forms which may be regular reflexes of some ancestor form but have no sisters should not be regarded as exceptions, though this is a troublesome methodological point, as it can be difficult to tell the difference.

Abstracting away from such presentations, we can see how the data structures act as constraints on each other: as the reconstruction progresses the degree of freedom decreases as the number of correspondences and cognate sets increases. We will consider in detail how this arrangement is arrived at.

First, a DATA MATRIX of comparable forms is created. Note that this structure is merely a tabular presentation of cognate sets.

(5) *An abstract version of the data matrix, one way of presenting cognate sets*

| Meaning | Set # | Reconstruction | $Lg_1$ | $Lg_2$ | $Lg_3$ | ... |
|---|---|---|---|---|---|---|
| $m_1$ | $c_1$ | $r_1$ | $w_{11}$ | $w_{21}$ | $w_{31}$ | |
| $m_2$ | $c_2$ | $r_2$ | $w_{12}$ | $w_{22}$ | $w_{32}$ | |
| $m_3$ | $c_3$ | $r_3$ | $(w_{13})$ | $w_{23}$ | $w_{33}$ | |
| $m_4$ | $c_4$ | $r_4$ | $w_{14}$ | $w_{24}$ | $(w_{34})$ | |
| | | | . | . | . | |
| | | | . | . | . | |
| | | | . | . | . | |
| $m_n$ | $c_n$ | $r_n$ | $w_{1n}$ | $w_{2n}$ | $w_{3n}$ | |

where:

$w_{ij}$ = word forms for i languages, over j possible meanings; these (or some part of these) are the actual comparanda.

$m_i$ = i different meanings, usually expressed as a gloss in some metalanguage.

$c_i$ = (optional) identifier of the cognate set, used to differentiate synonyms in the protolanguage.

$r_i$ = (less optional) reconstructed form = reconstruction = etymon (see §1.3).

$(w_{34})$ = forms which should not be included in the analysis (i.e. they are presumed to be replacements or variations of the expected regular form). Often these forms are omitted completely (though in fact they may have an important function as exceptions or potential counterexamples).

Next, the phonological constituents of comparable forms are aligned and put forward as correspondences. That is, suppose:

*(6)*

$w_{11}$ = abcd
$w_{21}$ = efgh
$w_{31}$ = ijk;

the comparativist must then align the constituents of each form to decide which sounds correspond. This is not an uncomplicated process, as has been observed by many linguists (Anttila 1989, Fox 1995).[7] However, absent any other phonological or phonotactic considerations, a reasonable alignment for the schematic data above would be that shown in (7) below.

*(7)*

| | | | | |
|---|---|---|---|---|
| $w_1$, $Lg_1$ | a | b | c | d |
| | \| | \| | \| | \| |
| $w_1$, $Lg_2$ | e | f | g | h |
| | \| | \| | \| | \| |
| $w_1$, $Lg_3$ | i | j | k | Ø |

The last alignment, for *ijk*, is problematic, for several equally valid alignments exist. The problem is treated in some generality in §5.2. From this alignment actual 'n-ary' correspondences can be extracted, that is, multilateral correspondences among several languages, as shown in (8) below.

*(8)*

a:e:i
b:f:j
c:g:k
d:h:Ø

Of course, the alignment of forms is not usually such a simple matter. A number of possible alignments and resulting correspondences are possible given the data in (6) above. Other putative cognate sets would have to be evaluated to see whether the correspondences proposed can be supported; indeed it is precisely this validation of correspondences across sets that protects the comparativist from false conclusions. Computational methods for performing this validation have been proposed (cf. for example Kay 1966 and Veatch 1993, discussed in §2.3 below, and in §6, where this process is applied to languages in the Himalayish and Loloish branches of the Tibeto-Burman family).

The 'strength' of a given correspondence is based on the frequency with which it is observed in the data set. Thus, each correspondence row is weighted according to the size of its list of supporting cognate sets. The 'strength' of the set of correspondence rows (i.e. of the table of correspondences) as a whole is based on the overall parsimony and phonetic plausibility of the table of correspondences, the proportion of the data matrix which is characterized (explained) by the set of correspondences, and other less tangible values like degree of semantic fit and typological properties of the modern and reconstructed languages.

These qualities are hard to quantify, making it difficult to distinguish good reconstructions from bad ones on a statistical basis.

Potential reconstructors are cautioned by standard textbooks:

- about the limits of relying on reconstructed lexical items in establishing genetic affiliation (as opposed to other linguistic structures, such as morphological and grammatical categorial correspondences, which at least in the case of Indo-European are deemed more reliable).

- about the need for precision in correspondence as a control against borrowing, analogy, and other sporadic processes which might give a false picture of the relationship. This precision is usually stated in terms of requirements for *regularity* and *exhaustiveness* (discussed in §1.7.5).

### 1.3. Interlocking parts of a reconstruction as constituting a proof

In this procedural statement the correspondence sets and the data matrix form a closed and interlocking structure: changing one perforce changes the other. Each form in each language in the data matrix has some relation to the overall structure, either as a regular reflex or as an exception. A completed reconstruction, one in which all known forms are completely accounted for, constitutes a proof of genetic relation. The validity of the proof rests on the completeness and plausibility of its treatment of the lexicons and phonologies of the languages; the data and its arrangement thus form a closed set. This 'closed-catalog' model is discussed in more detail in §1.7.5.

Clearly, a real reconstruction involves a rather large amount of data. Lexicons of languages contain thousands of words.[8] Given the depth and breadth of such a proof (see §1.7.3 on the characteristics of large proofs), it is not unusual for some of the relevant evidence to be missing, especially at the early stages of reconstruction. However, as the pieces are filled in the specificity and precision of the reconstruction increases. Indeed, as noted above, the specific relationships which link the comparanda express constraints on the result, providing a 'lock' on the proof:

- Correspondences cannot be changed without changing the cognate sets;
- Cognate sets cannot be changed without changing the correspondences;
- New data (new language data sets or new lexical items for existing languages) must either

1) fit smoothly into the scheme provided by the pre-existing cognate sets and correspondences,

2) be added (with further apologia) to the list of exceptions, or

3) be accommodated into the scheme by modifying the correspondences.

—or some combination of these.

## 1.4. The pieces of the lock

A great deal of this dissertation will be devoted to the problems (computational and otherwise) of creating and testing the comparative arrangements of large data sets; I must therefore take apart and inventory the components of the arrangement, showing how they fit together and what their function is in the final result. The arrangement revolves around the participants in the cognate set structure and the correspondences. Below (9) I give an abstract representation of both, showing which elements they have in common.

(9) *Linked data structure supporting a reconstruction*

*(a) Cognate set structure*

| | | | | | |
|---|---|---|---|---|---|
| Set number | sss. | | | | |
| possible reconstruction | *PLg abc [c1.c2.c3] | | | | |
| possible reconstruction | *PLg defg [c4.c5] | | | | |
| protogloss | *ppp* | | | | |
| | (1) lg1 | | <rrr1> | -mmm | *g1* |
| supporting forms | (2) lg2 | | <rrr2> | | *g2* |
| (with the cognate portion | (3) lg3 | | <rrr3> | -s1 | *g3* |
| distinguished and glosses | (4) lg3 | p1- | <rrr3> | | *g4* |
| given for individual forms) | (5) lg4 | p2- | <rrr4> | -s2 | *g5* |
| Annotations | cite1, cite2, cite3 | | | | |

*(b) Correspondence row structure (for n-ary case)*

| | | | |
|---|---|---|---|
| Correspondence row number | | c1 | |
| ancestor & environment | *a | /_*b | > |
| Outcomes in daughter | | lg1 | x1 |
| languages | | lg2 | x2 |
| | | ... | |
| | | lgn | xn |

*Supporting sets (i.e. CORRESPONDENCE INDEX):*

(consisting of at least the set number, etymon, formula, and perhaps supporting forms as well.):

| | | |
|---|---|---|
| sss | *abc | <c1.c2.c3> (NB: *defg <c4.c5> also possible) |
| nnn | *bxzc | <c1.c29> |

**Legend:**

sss, nnn number identifying this set in the list of cognate sets

*PLg Name of reconstructed language (PROTOLANGUAGE)

abc, defg reconstructed forms (etyma)

lg1 - lg4 language names

rrr1-rrr4 cognate portion of individual forms (comparanda); note that the 'cognate portion' of the form is bracketed off

[c1.c2.c3] the etymon stated as a formula (i.e. a composition of correspondences, indicated by their row numbers)

p1-, p2-, -s1, -s2 prefixes, suffixes, and other morphemes, grammatical and otherwise

mmm other morphemes, perhaps etymologized in another set

g1 - g5 glosses for individual words (perhaps in several different glossing metalanguages)

c1 - c5 references to the correspondences which this set exemplifies

ppp reconstructed meaning

Some elements function as deictics for locating or coindexing elements across and within the two data structures. For example, the language names (lg1-lgn) coindex the language forms in the cognate set with the outcomes in the correspondence row; the correspondence row number (c1) links the correspondence row to a specific ancestor of the reconstruction in the cognate set; and so on. These links are discussed in more detail below, when this abstract structure is fleshed out with real data in (10).

The structure accommodates a certain amount of precisely stated imprecision; for example, it allows for the inclusion of several possible etyma. There are several good reasons (discussed in §6.5.1) why more than one ancestor form may be possible for a given cognate set. Also, it is not unusual to find the same morpheme in one of the daughter languages exemplified synchronically in several different forms (as shown in reflexes (3) and (4) in (9) above). Indexing multiple occurrences of the same morpheme in this way may improve the quality of the reconstruction, but we would want to coindex all forms here at any rate to ensure that they are accounted for in the analysis.

Some of the structure shown in (9) is redundant: there is, for example, no requirement that the entire cognate set with supporting forms be listed again for each correspondence row (as is shown in the correspondence index). It can be very helpful as research progresses, however, to have large portions of coindexed structure available for easy reference.

Many details of these related structures remain to be discussed, but their significance will be much clearer if the abstract structure above is made more tangible. I have therefore provided a concrete example (10) of a cognate set and correspondence row based on a computer-assisted verification of an existing reconstruction of Proto-Lolo-Burmese (*LB) (Matisoff 1972), hereinafter called TSR (i.e. *[the Loloish] Tonal Split Revisited*).

*(10) Concrete exemplification of linked data structure shown above:*

*(a) A cognate set from* Matisoff 1972[9]

| | | | | |
|---|---|---|---|---|
| Set number | 185 | | | |
| possible reconstruction | *LB $^{H}$sə-<$^{H}$wat> [1.93.104] | | | |
| suggested protogloss | *FLOWER* | | | |
| | (1) na | | <$^{32S}$vi> | *flower* |
| supporting forms | (2) sa | | <$^{44}$vi> | *flower* |
| (with the cognate portion | (3) ahi | | <$^{44}$vi> | *flower* |
| distinguished and glosses | (4) ak | $^{H}$a- | <$^{HS}$yeh> | *flower* |
| given for individual forms) | (5) lh | $^{21}$ɔ- | <$^{54}$veʔ> | *flower* |
| | (6) lh | $^{35}$šɨ- | <$^{54}$veʔ> | *flower.* |
| Annotations | *LB *-at: *see also* alive [TSR 1], deer [TSR 10], run [TSR 18], bite [TSR 24], vomit [TSR 38], break [TSR 40(a)], break [TSR 40(b)], pluck[2] [TSR 57], pour [TSR 114], kill [TSR 124], hungry [TSR 132], spirit [TSR 136(a)], leech [TSR 167], etc. | | | |

*(b) A line from the correspondence index. Each correspondence row is annotated with the list of cognate sets which exemplify the correspondences. (created on the basis of data from* Matisoff 1972. *Partial annotation shown here in matrix format.*

| *#* | *Type* | **LB* | *Env* | *lh* | *ak* | *wo* | *sa* | *ahi* | *na* | *li* | *lc* | *bi* | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 104 | R | *-at | > | eʔ | eh | i | i.ɯ.ɛ | o.ɹ | ɤ.ɹ | e | e | ɛ | |

| Set | Gloss | *LB | lh | ak | wo | sa | ahi | na | li | lc | bi ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | alive | *dat$^{L}$ | tèʔ | deh$^{LS}$ | | | | | | | |
| 10 | deer | *tsat$^{H}$ | | tseh$^{HS}$ | | tˢhɯ$^{44}$ | | | htsye$^{2}$ | | tshɛ̄ |
| 18 | run | *kyat$^{H}$ | | ceh$^{HS}$ | | cɛ$^{44}$ | | | hchye$^{2}$ | | |
| 124 | bite | *{C}-tsat$^{L}$ | chèʔ | tseh$^{LS}$ | ši$^{33}$ | | tʂho$^{44S}$ | tš·ɤ$^{55}$ | | | |
| ... | | | | | | | | | | | |
| 185 | flower | *sə-wat$^{H}$ | ɔ̀-vêʔ<br>šɨ́-vêʔ | a-yeh$^{H-HS}$ | | vi$^{44}$ | vi$^{44}$ | vi$^{32S}$ | si$^{2}$-vé$^{3}$ | e$^{55}$ | wɛ̄ |

Besides the structural links mentioned above, there are several important points to be made about the linked cognate set/correspondence representation given here.

The etymon is given as a bimorphemic form, only the second morpheme of which is supported by the reflexes in the cognate set. The first morpheme is supported by another set of reflexes in another cognate set, which would include, for example, the first morpheme of the Lahu form šɨ́-vê?.[10] In both the etymon and the supporting forms the cognate portion is indicated by angle brackets.

In the correspondence line (104) shown, there are several different outcomes given for Sani (abbrev. 'sa'), Ahi ('ahi'), and Nasu ('na'). This is because the supporting forms for these languages (in bold font in sets 1, 10, 18, 124, and 185) reflect several possible outcomes for *LB *-at. These represent unfinished business in the analysis: the environment (in either the protolanguage or the modern language) which conditions these outcomes should eventually be supplied, or else these outcomes should be explained by other diachronic processes. The forms may not be cognate despite their resemblance, in which case (using the conventions given above on p. 12) we should indicate that they are exceptional by putting them in parentheses. For the time being it is sufficient to list these forms, perhaps with footnotes, and to include their outcomes in the table with the most common (i.e. most regular) outcome first.

Examining the complete list of cognate sets supporting the *-at rhyme given in TSR (shown in (11) below), one other important feature should be noted. Two other sets which might be adduced as support for this rhyme have irregular outcomes in Lahu. These forms, found in sets #57 PLUCK and #132 HUNGRY, seem to point to a different protorhyme.

(11) *All sets for *LB rhyme -at from TSR.*

| Set | Protogloss | *LB Reconstructions and Reflexes |
|---|---|---|
| 1 | alive | *dat$^{L}$. Ak deh$^{LS}$. Lh tè? |
| 10 | deer | *tsat$^{H}$. Ak tseh$^{HS}$. Bi tshɛ̄. Ha[HT] tse. LC ts·i$^{22S}$. Li htsye$^{2}$. Sa tshɯ$^{44}$ |
| 18 | run | *kyat$^{H}$. Ak ceh$^{HS}$. Li hchye$^{2}$. Sa cɛ$^{44}$ |
| 24 | bite | *{C}-tsat$^{L}$. Ahi tʂho$^{44S}$. Ak tseh$^{LS}$. Lh chè?. Mo ts·a$^{55}$. Na tš·ɤ$^{55}$. Wancho tsat |
| 38 | vomit | *C-pat$^{L}$. -Na p·i$^{213}$. Ahi phi$^{44S}$. Ak peh$^{LS}$. Ch pha. Ha[K] phə$^{21}$. LC p·i$^{55c}$. Lh phè?. Li hpē$^{6}$. Sa phi$^{22S}$ |
| 40(a) | break | *tsat$^{H}$. Ak tseh$^{HS}$. Ha[HT] tse. Lh chê?. Sa tshɨ$^{44}$ |
| 40(b) | break | *{C}-tsat$^{L}$. Li hchē$^{6}$ |
| 57 | pluck[2] | *ʔcwat$^{H}$. Ak ci$^{HS}$. Ak ci$^{HS}$. Lh (cɨ?). Sa tši$^{44}$ |
| 114 | pour | *šat$^{H}$. Ak sheh$^{HS}$. -Bi šɛ̀t. Lh šê?. Sa xɤ$^{44}$ |
| 124 | kill | *C-sat$^{L}$. -Ahi xo$^{21}$. Ak seh$^{LS}$. Bi sɛ$^{21}$. Ch tʂhu. Ha [HT] se$^{21c}$. -Ha [K] še$^{33}$. Li sye$^{6}$. Na si$^{55}$. -Sa xa$^{11}$. Wo ši$^{33}$ |
| 132 | hungry | *mwat$^{L}$ ~ ŋwat$^{L}$. -Ha [K] mie$^{33}$. Ahi ni$^{44S}$. Ak meh$^{LS}$. Bi bɛ̀. Ha [HT] me$^{21c}$. Lh (mə̀?). Li mrghe$^{6}$. Na ñi$^{55}$. Sa ŋ̥$^{22}$s. Wo me$^{33}$ |
| 136(a) | spirit | *nat$^{L}$. Ak neh$^{LS}$. Ha [HT] ne$^{21c}$. Na nɛ$^{55}$. Wo ni$^{33}$ |
| 167 | leech | *k-r-wat$^{L}$. Ak yeh$^{LS}$. Lh vè?. Li vé$^{6}$ |
| 185 | flower | *sə-wat$^{H}$. Ahi vi$^{44}$. Ak a-yeh$^{H}$-$^{HS}$. Bi wɛ̄. Ha [HT] je. Ha [K] βæ$^{33}$. -LC e$^{55}$. Lh ɔ̀-vê?. Lh šɨ-vê?. Li [F] si$^{?}$-vɛ̃$^{3}$. Li [J] ve$^{33}$. Na vi$^{32S}$. Sa vi$^{44}$ |

The etyma for these sets *mwat$^L$ and *ʔcwat$^H$ point the way: these are in fact not *-at rhymes, but *-wat rhymes. This type of phonological distinction (called 'open mouth' (kāikǒu) versus 'closed mouth' (hékǒu) articulation in traditional Chinese grammatical terms) is not at all uncommon in Tibeto-Burman languages. The w functions not as an initial (as it does in the regular but prefixed *k-r-wat$^L$ (#167) and *sə-wat$^H$ (#185)), but as a glide. A different correspondence row annotated with these sets (and perhaps others) is thus required to account for these variations.

It is easy to see that a complete reconstruction based even on a small number of words in a few languages requires a large amount of work.

### 1.4.1. Creating the arrangement

Data does not come pre-arranged in the convenient format just presented; the set of related data structures described above must be extracted and homogenized from a variety of sources. The usually tedious, iterative labor required to create these tightly-knit representations is backgrounded in standard presentations of the comparative method. In fact, complete reconstructions, those with many cognate sets and involving a large set of languages, take many years to prepare.

Presentations of the comparative method typically provide no description of how the items compared are chosen. There is a significant amount of fairly well understood semantic and phonological apparatus required to support the creation of sets of cognates and sets of sound correspondences, the 'principal step' in the comparative method (cf. Hoenigswald 1966). In most explications of the method, however, the problem of matching items semantically and phonologically is implicitly (and invisibly) taken to have already been solved. The true starting situation is more like that represented in (12) below.

(12)

| Lexicon of $Lg_1$ : | | | Lexicon of $Lg_2$ : | | | Lexicon of $Lg_3$ : | | |
|---|---|---|---|---|---|---|---|---|
| $w_{11}$ | $m_{11}$ | $c_1$ | $w_{21}$ | $m_{21}$ | $c_1$ | $w_{31}$ | $m_{31}$ | $c_1$ |
| $w_{12}$ | $m_{12}$ | $c_2$ | $w_{22}$ | $m_{22}$ | $c_2$ | $w_{32}$ | $m_{32}$ | $c_2$ |
| $w_{13}$ | $m_{13}$ | $c_3$ | $w_{23}$ | $m_{23}$ | $c_3$ | $w_{33}$ | $m_{33}$ | $c_3$ |
| $w_{14}$ | $m_{14}$ | $c_4$ | $w_{24}$ | $m_{24}$ | $c_4$ | $w_{34}$ | $m_{34}$ | $c_4$ |
| . | | . | | | . | | | |
| . | | . | | | . | | | |
| . | | . | | | . | | | |
| $w_{1n}$ | $m_{1n}$ | $c_n$ | $w_{2n}$ | $m_{2n}$ | $c_n$ | $w_{3n}$ | $m_{3n}$ | $c_n$ |

In this case, several analyses must be carried out in order to create an arrangement which demonstrates the etymological relationship. In fact, we can identify at least two different types of approaches to the problem of organizing data for comparison:

- A 'semantics-first' approach, in which words having related meanings are brought together and examined for principled phonological similarities. Matisoff's 'organic-semantic method' (Matisoff 1978; Matisoff

1980; Matisoff 1985}:421) is an example of such a method. Computer-aided techniques applying this approach are discussed in §5.1.7.

- A 'phonology-first' approach, in which words which exhibit similar phonological patterns are grouped together into sets. This approach is exemplified by, for example, the Reconstruction Engine program (Lowe and Mazaudon 1989; Lowe and Mazaudon 1994).

### 1.4.1.1. The semantic side of the equation

On the one hand, some equation must be made between the different meanings of words across languages (and sources). That is, it must be determined when $m_{ij} = m_{ab}$ for the various values of i, j, a, and b. This is a trivial task in some cases (e.g. for 'core vocabulary' such as *eye, father,* and so on). However, since the $m_{ij}$ are in fact represented by glosses in varying metalanguages, conceptual difficulties arise straightaway. For example, should a word glossed (in French) as *voler* in one dictionary be compared with a word meaning *fly* in a source whose glosses are given in English, or with *steal*? (cf. Benveniste 1971) Sometimes homographs in the glossing metalanguage are not distinguished (e.g. *tear (n.)* and *tear (v.)*). Orthographic conventions and idiosyncratic characteristics of the glossing language may prevent an easy equation of glosses (e.g. trivially *brain* may be used for *brains,* preventing any fully automatic, morphology-insensitive matching of glosses). The problems associated with identifying semantic equivalence for the purposes of diachronic analysis are discussed in §5.1.7.

There are limits on how far the semantic links can be stretched, but these are in principle not simple to state. An example is Greenberg's 'Proto-Sapiens' reconstruction meaning 'hand; finger; to point (with the finger); one' which is supported by lower-level reconstructions such as Proto-Sino-Tibetan (*ST) *tik 'one', Proto-Indo-European (PIE) *deik 'point', and Amerindian forms like Karok ti:k 'hand;finger' as well as others. (Greenberg 1987:62), cited by Matisoff 1990a:112)

A serious problem in establishing equivalence in meaning is that there may be several forms to choose from. The linguist accordingly must have access to the lexicons of the various target languages in some depth, lest possible cognates be missed. Lexicostatistical studies have been criticized on this account: when counting cognates between languages, missing a cognate may skew the results. Consider table (13) below, from Dyen 1992:

(13) *German and Spanish pairs used to compute 'lexicostatistic' percentage (after Dyen 1992:96)*

| Meaning | German | Spanish | Cognation |
|---|---|---|---|
| all | alle | todo | no |
| and | und | y | no |
| animal | Tier | animal | no |
| ashes | Asche | ceniza | no |
| ... | ... | ... | ... |
| fat | fett | grasa | no |
| father | Vater | padre | yes |
| to fear | fürchten | temer | no |

Comparison of all words in the 200 word Swadesh list gives a 25.3% rate of 'cognation' between German and English. However, as Dyen notes, complications occur when a speech variety uses two or more forms for a single meaning, and when cognation is judged to be indeterminate because the evidence does not support a clear decision. (Dyen 1992:96). So, for example, the following data from English and German are problematic.

*(14)*

| English | German | Possibly cognate? |
|---|---|---|
| cat | Katze | yes |
| **dog** | **Hund** | **no** |

The English word *hound* has no place in this scheme as it is not properly core vocabulary. One has the sense that decisions of this sort are being made with one eye closed; the comparativist using the traditional comparative method on the other hand is relatively free to choose from the available vocabulary. The arbitrariness of signifier and signified and the requirements for regularity protect against the possibility of chance similarity being mistaken for cognacy.

Nevertheless, the received wisdom is that it is best if the lexemes compared are drawn from 'basic' or 'core' vocabulary. The idea is that such data are the least subject to arbitrary replacement and reshaping through time. Restricting comparanda to basic vocabulary also reduces the amount of semantic variation which the comparativist has to deal with. Meillet (Meillet 1967:48-49} on the one hand notes that vocabulary is 'the most unstable element of all in language,' but that 'in spite of this frequent instability ... , it is the agreements in vocabulary which are the most striking when languages are compared with each other.' In this regard, Matisoff 1978 and others have noted that the notion of what constitutes 'basic vocabulary' is culturally conditioned.[11]

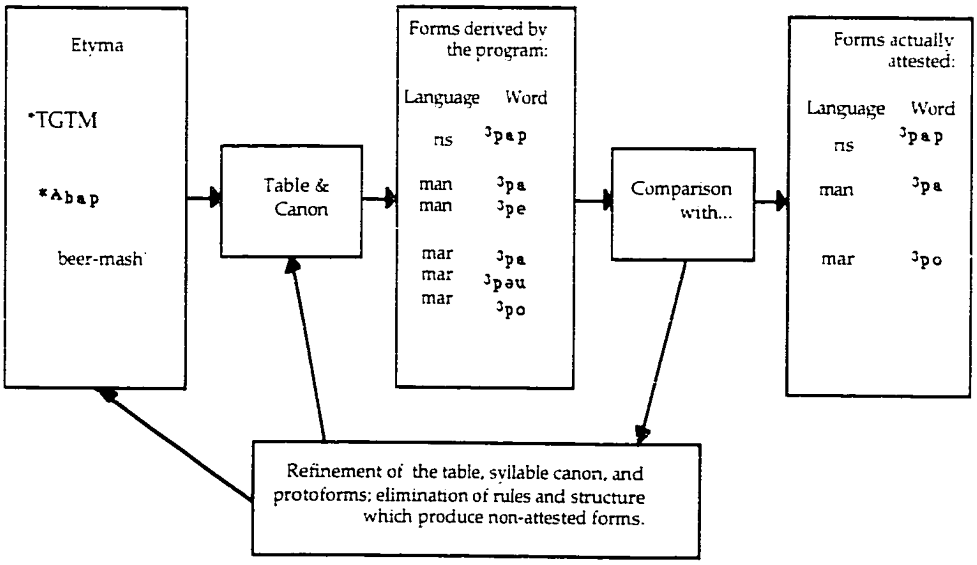
### 1.4.1.2. The phonological side of the equation

An equally serious problem is the necessity of establishing some equation between the different sounds of words across languages (and sources). That is, it must be determined when $w_{ij} \approx w_{ab}$ for the various values of i, j, a, and b. '≈' is used here is the sense of *regularly corresponds*. It is possible and indeed likely that the transcription of the forms will vary; Meillet, in his example above (1), cited the data in its conventional orthography though of course he knew quite well that the comparison of forms must be carried out on the basis of sounds. It is presumed that the comparison of elements is between immediate and continuous constituents; to do otherwise would be to complicate the analysis immensely. As noted above, the alignment of phonological material for comparison is far from automatic. Here, however, I only broach the problem; the complex relationship between the representation of data and their analysis is treated in several chapters below.

### 1.4.2. Expanding the scope of the arrangement

The process of creating the arrangement is iterative: the arrangement is constantly in flux, and indeed remains so until the last form is in place and the last gap in a correspondence row filled in. After some correspondences are noted, the lexicons are searched again for more supporting items. These in turn may provide evidence for other correspondences. Many iterations and revisions are required before a complete analysis can be offered. The method can be seen in this light to consist of alternating cycles of deduction and induction.

(15) *Iterative improvement in the reconstruction process (data from a computational study of the Tamang languages of Nepal, discussed in detail in §6)*



### 1.4.2.1. A short aside on drudgery

A final note about this time-consuming iterative process: several authors who have written on the subject of automating parts of this process have claimed that one of the expected benefits of automation is a reduction in drudgery and the amount of time it takes to complete a reconstruction project. The author of an early effort relates the following challenge and her reply:

> A noted linguist has asked me what I have accomplished that had not been previously accomplished with the application of a lot of elbow grease. The question seems to me rhetorical. Eliminating for future scholars the elbow grease now required for etymological studies ... has reduced work which could have been a lifetime's undertaking to perhaps a sabbatical project, or a year's work for a graduate seminar, or a brief funded research project. With standardization, each language history which is programmed can be added to the files available to scholars. and each addition increases greatly the probability that a time will come when the entire family of [in this case] Indo-European languages will be instantly available for comparative, anthropological, historical and etymological studies. (Burton-Hunter 1976:219)

Twenty years have passed and this Holy Grail of computational historical research has not been realized. I shall conclude this section by speculating briefly on why this is so. First, it is the experience of most automation specialists that automation does not in general save labor, at least not on the order of magnitude suggested by the optimistic linguist quoted above. The primary effect of automation is to improve the consistency of the work, and to make the results of an effort available for further research. Many factors affect whether automation really produces this effect; certainly most people who have undergone the horror of automating their work have at first experience a sense of tremendous loss of time and effort. It is only after some time has passed and some experience gained that the benefits are realized. Second, and finally, the process of 'programming the history of a language' is not a task to be compared with programming a payroll system. In a research environment, such ends are rarely attained, since it is rare that research activity is 'operationalized' to the point that the automated activities become routine. Historical reconstruction will never be, probably can never be, as routine an activity as processing ATM transactions at a bank. It would be unwise to expect to feed the dictionary of a previously unknown language into a computer and have a reasonable diachronic analysis come out the other end.[12] Many well-entrenched, well-understood and difficult problems stand in the way of a single solution. The sheer amount of data which must be sifted through has limited the research:

> The greatness of Bloomfield's treatment of Algonkian can be ascribed to a variety of reasons. A very important point of methodology lies in his LIMITATION OF THE PROBLEM to the comparison of four languages for which he had adequate (though not abundant) descriptive materials....To have attempted to use all available materials on the dozens of Algonkian languages for which some kind of information was available would have rendered the task so *unwieldy and unmanageable* that he would not have been able to complete it in a lifetime.' (Haas 1969:25; italics mine)

The amount of data has deterred linguists from carrying out various kinds of analytic tasks:

> In the historical study of written languages it is commonplace to make use of evidence from loanwords, outloans as well as inloans. This potentially valuable source of information has been almost totally neglected in the study of unwritten languages, not because such valuable information is not available, but because the *drudgery* of patient sifting of many sources has all too often been neglected. (Haas 1969:48; italics mine)

These and other challenges to thorough comparison have hampered the reconstruction of the world's languages. The long and colorful history of comparative linguistics testifies to the resourcefulness and determination required to discover and explain the facts underlying the present state of a language despite the difficulties. A short consideration of some of the realizations that contributed to meeting these challenges is in order before some possible improvements are proposed.

## 1.5. A brief history of the comparative method

The reader will forgive me if I do not recount in detail the early events which led to the development of the comparative method. These are recounted by better and more thorough story-tellers than I, and on whom I have relied, among them Davies 1987, Fox 1995, Hoenigswald 1990, Lehmann 1967, Percival 1987.

### 1.5.1. Before the comparative method: pre-romantic and romantic views of language change

Early researchers in etymology had as their espoused goals the discovery of the primordial meanings of words, a search that was imbued with a mystical or religious significance; Plato, for example, asserts in *Cratylus* that the source of a word is not merely the historically earliest ascertainable form, but the actual *etymos logos*, that is, its *true* meaning. (Percival 1987:11) Such knowledge, could, for example, provide speakers with 'correct' definitions. The notion that words derived from some primordial vocabulary persisted until (and somewhat beyond) the first 'scientific' approaches to etymology were used in the beginning of the nineteenth century. Bopp, for example, believed in the notion of *prōta onomata*, primordial words endowed with a rationality lost in subsequent periods of decay and degeneration. The 'organic' metaphor slowly came to dominate descriptions of language change; the terminology of grammar reinforced it (cf. for example Latin *radix* 'root') until in Bopp's review of Grimm's work we find:

> Languages must be taken as organic natural bodies which form themselves according to definite laws, develop carrying in themselves an internal life principle, and gradually die... (cited from (Davies 1987:84)

We will return to the biological metaphor later.

Early comparativists felt that the development of languages was guided by divine principles. The interpretation of Grimm's (really Rask's) Law as a cycle illustrated its mystical significance. Discovery of principled subregularities such as Verner's Law reinforced the idea of a divine regularity, intricately ordered and assembled like a watch, and set the stage for the strict Neogrammarian view of a linguistic universe governed, like the physical universe, by exceptionless laws.

### 1.5.2. Early attempts at comparison

At first, examination of the data supported most intuitions about relatedness. As time went on, however, it became clear that additional controls on comparison were needed to arrive at a consistent result.

The history of Indo-European studies provides no support for claims of the efficacy of superficial lexical comparison (the importance of this point will be made quite clear later on). Indeed, this history provides important object lessons on how to establish genetic affiliation, and how easy it is to go astray.[13] Among the first to observe that several Indo-European languages might be related was Sir William Jones, known to most linguists from the famous passage below in which he proposed the nucleus of the Indo-European language family (Jones 1798:422-423).

> The Sanscrit language, whatever be its antiquity, is of a wonderful structure; more perfect than the Greek, more copious than the Latin, and more exquisitely refined than either; yet bearing to both of them a stronger affinity, both in the roots of verbs, and in the forms of grammar, than could possibly have been produced by accident; so strong, indeed, that no philologer could examine them all three without believing them to have sprung from some common source, which perhaps no longer exists. There is a similar reason, though not quite so forcible, for supposing that both the Gothick and the Celtick, though blended with a very different idiom, had the same origin with the Sanscrit; and the old Persian might be added to the same family, if this were the place for discussing any question concerning the antiquities of Persia.

However, no one has seriously claimed that Jones was a strict adherent to the (as yet unformulated) comparative method and its principal tenets; his published work provides only the skimpiest evidence as to his methods, for he generally gave only his conclusions, not detailed arguments and data. Though Jones was aware of the possibility of borrowing, and that borrowing is especially likely in cultural and technological vocabulary, he (like Greenberg and some other modern scholars) did not recognize that massive borrowing was possible, or that even relatively basic vocabulary can be borrowed (Jones 1799a:54-55):

> I close this head with observing, that no supposition of a mere political or commercial intercourse between the different nations, will account for the Sanscrit and Chaldaic words, which we find in the old Persian tongues; because they are, in the first place, too numerous to have been introduced by such means; and secondly, are not the names of exotic animals, commodities, or arts, but those of material elements, parts of the body, natural objects and relations, affections of the mind, and other ideas common to the whole race of man.

As a result, he was ready to postulate genetic affiliation on the basis of large numbers of similar words.

Jones was also aware that grammatical correspondences provide stronger evidence of genetic affiliation than lexical correspondences (Jones 1799c:4):

> That the written Abyssinian language, which we call Ethiopick, is a dialect of old Chaldean, and sister of Arabick and Hebrew; we know with certainty, not only from the great multitude of identical words, but (which is a far stronger proof) from the similar grammatical arrangement of the several idioms.

Jones was not careful about excluding loans in spite of his recognition of the problem, and, since he did not establish phonological correspondences, his methods led him astray in many cases. A particularly striking case is his misidentification of Pahlavi, an Indo-European language of the Iranian branch, as Semitic. (Jones 1799a:52):

> This examination gave me perfect conviction, that the Pahlavi was a dialect of the Chaldaic; Chaldaic refers to the Semitic family, especially to Aramaic ...and of this curious fact I will exhibit a short proof. By the nature of the Chaldean tongue most words ended in the first long vowel, like shemia, heaven; and that very word, unaltered in a single letter, we find in the Pazend, together with lailia, night; meyd, water; nira, fire; matra, rain; and a multitude of others, all Arabic or Hebrew, with a Chaldean termination; so zamar, by a beautiful metaphor, from pruning trees, means in Hebrew to compose verses, and thence, by an easy transition, to sing them; and in Pahlavi we see the verb zamruniten, to sing, with its forms zamrunemi, I sing, and zamrunid, he sang; the verbal terminations of the Persian being added to the Chaldaic root. Now

> all those words are integral parts of the language, not adventitious to it like the Arabic nouns and verbals engrafted on modern Persian; and this distinction convinces me, that the dialect of the Gabrs, which they pretend to be that of Zeratusht, and of which Bahman gave me a variety of written specimens, is a late invention of their priests, or subsequent at least to the Muselman invasion. [14]

Not only did Jones also mistakenly classify other Iranian languages as Semitic (Jones 1799c:7-8), but he included Malay in this family as well (Jones 1799c:10). In fairness, Jones labored under another misconception which would have made it difficult for him to make a correct identification of Malay: perhaps influenced by Bopp, who had tried to prove that the Malayo-Polynesian languages were Indo-European, Jones mistakenly regarded the Austronesian languages as Indo-European, specifically Indic (Jones 1799c:12). Deceived it appears by the influence of Indic culture on Tibet, Jones wrongly grouped Tibetan in the Indo-European family showing heavy influence from Chinese, rather than as Sinitic showing heavy influence from Indic (Jones 1799c13):

> For, although it [Tibetan] was anciently Sanscrit, and polysyllabick, it seems at present, from the influence of Chinese manners, to consist of monosyllables, to form which, with some regard to grammatical derivation, it has become necessary to suppress in common discourse many letters, which we see in their books, and thus we are enabled to trace in their writing a number of Sanscrit

> words and phrases, which, in their spoken dialect are quite undistinguishable.

Another case in which Jones failed to recognize a real relationship is that of Hindi, which he denied could be related to Sanskrit on the grounds that its grammar was typologically so different (Robins 1990:93).

Neither Grimm nor Bopp required exhaustive regular correspondences in establishing cognacy. Grimm claimed that 'the sound shift takes place in the mass, but never neatly in individual items; words remain in the relationship of the old arrangement — the stream of innovation has passed them by' (cited from Lehmann 1967: 57).

### 1.5.3. Classification, *Stammbäume*, and the biological metaphor

Towards the middle of the nineteenth century, the flowering of scientific thought inspired in part by the earlier romantic trends bore fruit: a number of correlated theories about natural systems took root and began to flourish. Perhaps the most significant of these, Darwin's theory of evolution, found immediate parallels in describing the evolution of languages (the association is linked most closely to Schleicher 1873). The mystical insights of Grimm and others came to be perceived as scientific truth exposed by the discovery of the organic basis of language. The extension of the biological metaphor, and in particular the notion of family trees, to language evolution provided an explanation for certain historical artifacts in language. The sequentialization of Verner's Law with respect to Grimm's Law showed that it was necessary to reconstruct stages in language change.[15] These changes could in turn be related to a tree structure, providing a natural means for organizing the sequence of events. The interpretation of the branching structure of the trees as points of divergence (or more precisely, as the last points of unity) gave linguists a natural means to organize their hypotheses about language change.

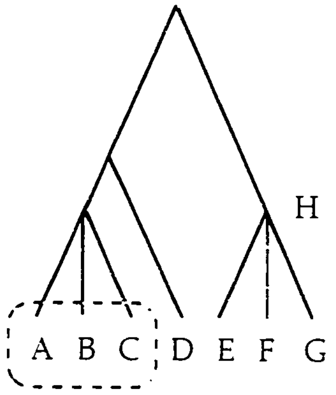
Now, it has always been eminently reasonable to hypothesize that 'between the initial *common language* reconstructed by comparison and a language attested in fact, one or more intermediate *common languages* [may] be interposed...These stages much facilitate the explanation [of the historical development]' (Meillet 1967:29). In fact, without such stages, as Meillet notes, the explanation of the facts would 'remain singularly incomplete, most often impossible'. The identification of the branchings of the tree with intermediate common languages is a natural choice. In allowing for intermediate common languages, however, Meillet did not mean that meso-languages should be freely constructed simply for the sake of facilitating explanation. He notes that these intermediate common languages (mesolanguages) 'reflect periods of linguistic unity' (Meillet 1967:30) and that each common language 'must express a type of civilization.' The linguistic tree model was in this sense constrained by reality; it was not to be interpreted merely as a classificatory device.

Some notion of descent was clearly needed, but the notion provided by the seductive metaphor from biology was found inadequate to the task. Languages, unlike organisms, permit and even encourage horizontal transmission. In biology, such transmission would be the equivalent of members of one species acquiring features of some other species in the course of a single generation simply by exposure.[16] Over the course of time, one species can indeed adopt another's features (butterflies which assume the appearance of a predator); languages, less constrained by the physical limits of DNA, can exhibit this type of convergence as well.[17]

Therefore, the tree model was challenged almost immediately as linguists noted that other modes of propagation of changes existed. Schmidt's wave model Schmidt 1872, for example, provided an intuitive and very reasonable alternative to the tree model of transmission. It rapidly became clear that both notions (at least) were needed for an adequate description of language development.

The tree model had implications for carrying out reconstructions. Comparing distantly related languages would (in theory) be more difficult because the number of branchings would be greater. Comparison, then, should made between languages which are presumed to be related, and in the best of all possible worlds, closely related (most textbook presentations of the comparative method, for example, pick examples from a particular subgroup; it is rare to see forms from modern Romance languages compared directly with those of modern Germanic languages!). The time of separation of closely related languages is relatively recent; hence other extralinguistic processes have not had much time to cloud regular phonological developments. The problem of relative chronology of sound changes is thereby reduced (though of course not eliminated). Thus, in the tree metaphor of genetic affiliation, the linguist aims *first* at comparing languages which are clearly related to each other. In principle, the clearest relation (or perhaps merely the greatest similarity) should appear between languages which are in adjacent branches of the tree:

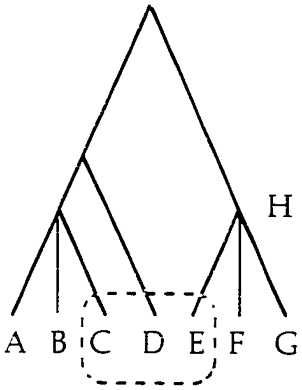
(16) *Compare closely related languages at a small time depth first*



Of course, it may not be clear at the outset that the comparanda are in fact related in any such direct way; the comparative method (in its most direct application) can only provide this single step in the evolution from parent to child. Attempting comparison across levels and involving more distant relatives inevitably causes problems, as the chain of phonetic events becomes more convoluted. Haas notes that protolanguages can (and indeed should) be compared with other protolanguages (Haas 1969: 49). Like is compared with like, at the same level.

When it proves difficult to develop plausible and regular correspondences within a group of modern languages, it may be the result of a situation such as that illustrated in (17) below. Languages D and E are rather distantly related (this fact could certainly be indicated by flipping the branches of the tree around); the chain of 'bifurcations' between these languages and their shared ancestor is long. Hoenigswald (Hoenigswald 1963:4) (among others) has shown how to create and arrange trees according to the results of binary comparisons of correspondences between each of the possible pairs of languages. This clever procedure allows the linguist to rank the languages according to their shared innovations and retentions and therefore present a plausible chronology.

*(17) Other choices for comparison may be problematic*



Hoenigswald (Hoenigswald 1963:7) has taken great pains to indicate how the 'precious irreversibility' of phonological mergers, etc. is to be interpreted in a stepwise fashion within a hypothesis of relatedness. Some have interpreted the requirement for selecting adjacent branches of a family tree even more rigidly, requiring strictly binary trees (Ruvolo 1991:213); but this must be regarded as fashionable wishful thinking.[18]

This is not to say that a reconstruction based on a set of languages like those in (17) above would not work: it would, however, be a macro-reconstruction, or even a megalo-reconstruction (cf. §1.6.9 below).

### 1.5.4. Mesolanguages, or, what is a protolanguage anyway?

Clearly the branchings of the tree have significance, and Meillet, among others, interpreted this to mean that the nodes reflected some period of proto-unity. What could be said about these stages? That is, what is a protolanguage? Haas, in answer to this question, states that 'quite simply any language is an actual or potential protolanguage' (Haas 1969:31). In this view, a protolanguage has both unity and concrete reality. Other linguists have taken a much more cautious view. While there may have been at some point *a* language that is reflected in the reconstructed protolanguage arrived at by application of the comparative method, that protolanguage is an entirely *artificial construct*, a model of a lost reality useful only in this particular sort of scientific investigation. Concerning the hypothesis of a unitary protolanguage, Trubetskoy (Trubetskoy 1939) notes that 'its first, purely linguistic nature has been forgotten. Prehistorical archeology, anthropology, and ethnology have been drawn in in an unjustifiable manner.' Meillet, for example, scolded Schleicher for composing a fable in PIE:

> It was a daring feat of genius for Schleicher to *reconstruct* Indo-European with the aid of historically attested languages of the family; but it was a grave error on his part to compose a text in this reconstructed language. Comparison brings a system of equations on which the history of a family of languages can be based; it does not furnish a real language with all the means of expression which this language had. (Meillet 1967:28-29})

This, however, is not the end of the saga; other linguists (Lehmann and Hirt) have provided similar reconstructions of connected texts, and it seems clear that reconstruction should *not* stop at the level of lexicon and morphology (or syntax) unless there are insurmountable methodological obstacles. If protolanguages are languages, then we are bound to try to recover as much of their character as possible; we cannot be satisfied with reconstructed vocabulary and basic elements of grammar. Certainly the amount of proto-syntax reliably reconstructed to date appears rather small. The reconstruction of syntax by the comparative method has had some obvious successes. (Holland 1992) As the methodology of lexical and morphological reconstruction are better understood we may be able to advance in syntactic reconstruction as well, despite substantial differences in method required. Machine-readable cross-linguistic text corpora text corpora are increasingly available for this kind of research: for example, machine-readable texts exist for Vedic Sanskrit (van Nooten and

Holland 1994), the Pali Canon, Classical Greek (the TLG of Berkowitz 1992), Old English, and several other older languages.

### 1.5.5. The neogrammarian principle

The fundamental sub-theory underlying the comparative method remains the principle that sound change, when properly formulated (i.e. taking account of conditioning environments, etc.), operates without exception. 'Sound change', as opposed to a simple 'change in sound', is defined as those phonetic changes whose reflexes are found recurrently in the lexicon. Three things are asserted about sound change: 1) it exists, 2) it occurs through progressive unconscious changes, and 3) it lies at the root of 'changes in sound'. Exceptions to regular sound change have to be explained by other factors. One weakness of the theory is that it does not provide a detailed account of how regular change is implemented across generations in a homogeneous community, and in what way the interaction with the other forces (stylistic restoration, analogy, etc.) actually takes place.

It should be noted that proponents of the regularity hypothesis did not equate regularity with simplicity: regular rules could be quite complicated. Any incompleteness or inaccuracy in the statement of rules therefore understates the degree of regularity in the language.

> Karl Verner, who provided the most striking evidence [that language is governed by rules], wrote in a letter of 1872 that the

> accepted proposition 'no rule without exceptions' should be rephrased as 'no exception without a rule'. And in his greatly admired article three years later he put it: 'there must be a rule for irregularity; the problem is to find it.' [Lehmann 1991:12]

Regular sound laws were the backdrop against which irregularities were highlighted. Exceptions indeed *proved* (i.e. tested) the rules.

### 1.5.6. The classical approach

With the parallels between the study of language and the study of other natural systems now firmly entrenched, linguists began to demand that theories and conclusions in linguistics adhere to the same standards as those in other sciences. 'Like the sciences, linguistics requires exact observation, thereupon description, and conclusions (explanations) based on exact observations' (Schleicher 1873:1) cited in Lehmann 1991:13). The comparative method had evolved into a powerful tool for drawing conclusions. And the 'exact observations' are the words occurring in the languages compared—exact, it must be noted, to the limit of available descriptive tools.

The theory of phonological change which underlies the classical approach (and indeed the work of most historical linguists to date) is that first expressed by the neogrammarians, and modified in accordance with the structuralist and functionalist point of view as developed by the Prague School.[19] It is true that this set of hypotheses, which I am calling the 'classical approach', has not been greatly modified recently, and it is also true that it is still incomplete. We will see later that Henning Andersen's detailed account of a particularly vexing problem for the classical approach goes a long way toward filling one of its gaps (Andersen 1974:773).

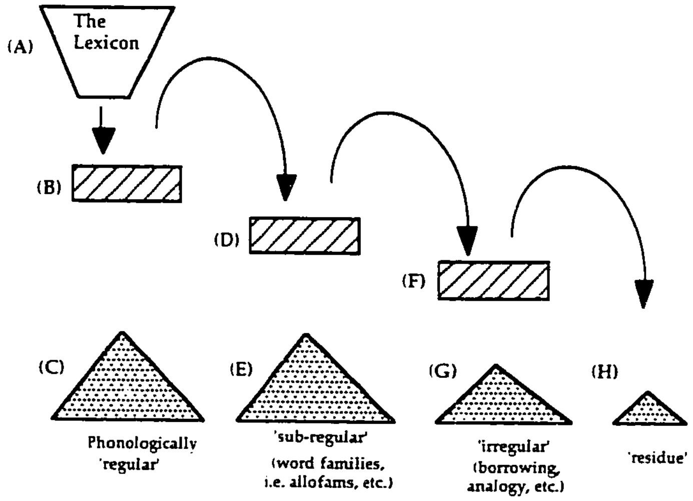
The classical approach posits a set of concomitant forces that influence the development of languages. The resultant of these forces, each pulling in its own direction, is the actual history of the language. This is probably not the way a historian of linguistics would formulate the neogrammarians' theoretical outlook, but it states their views in a more sympathetic, less dogmatic light.

Such a 'global theory' calls for several sub-theories to account for the different factors or 'forces', and for a theory of their interaction. We are very far from possessing such a complete set of theoretical tools. Clearly, as long as we do not possess this complete overall theory, any part of it can be considered as unfalsifiable, since a counterexample can always be ascribed to the domain of a different part of the theory which is not yet fully explicit. For instance, the theory of contact and borrowing can be invoked as an easy 'escape hatch' for apparent exceptions to the principle of the regularity of sound change. The interaction between the theory of contact and the classical theory of sound change is discussed (albeit briefly) in §1.6.8 below.

The major subtheories which comprise the global theory are illustrated in (18) below as a set of 'sieves' and the 'piles' into which they apportion the lexicon. It must be emphasized that the relative sizes of these piles are completely arbitrary; the proportions will vary from family to family.[20] It would not be unusual for the list of problem cases (H) to be the largest, especially for less-studied language groups. Especially in cases where languages are in close contact or are only distantly related, the regular component of the lexicon may be expected to be quite small. Some dialects of Bai, for example, a Tibeto-Burman language of as yet undetermined affiliation, have had 75% of their vocabulary replaced by Chinese loans. Such massive influence (from the point of view of reconstruction we might say contamination) makes it difficult to arrive at a satisfactory result through the application of the comparative method.

It is not clear from this diagram where isolates (such as English *pour*) should settle. Isolates, the 'only children' of some ancestor form or perhaps early nonce forms, should logically be classified with the regular component in the absence of evidence to the contrary. However, from a methodological point of view this is difficult since the regularity of isolates cannot be determined except by comparison with other sets which exemplify the same correspondences. A similar situation arises when a form can only be reconstructed in one subgroup of a language family (e.g., '< PGmc *xyz (unattested elsewhere).' There is no reason to challenge the authenticity of an etymology of this sort; which enough support they are independently justifiable, though they are still isolates in terms of the higher-order comparison.

*(18) The comparative method is a collection of subtheories*


(A)
The Lexicon
(B)
(D)
(F)
(C)
(E)
(G)
(H)
Phonologically 'regular'
'sub-regular' (word families, i.e. allofams, etc.)
'irregular' (borrowing, analogy, etc.)
'residue'

Key:

A. The complete lexicon.
B. Regular sound change.
C. Regular, 'expected' reflexes of the ancestor forms.
D. Domain of 'protovariation', perhaps due to morphological/derivational processes.
E. Sub-regularities emerging from a relaxation of strict regularity constraints (word families, etc.)
F. Sociolinguistic explanation. Domain of lexical diffusion and other sporadic processes.
G. Borrowings, analogized forms, hypercorrections, prestige pronunciations, etc.
H. The 'mystery pile': counterexamples and other troublesome words.

Sub-theories are not to be understood as defined by different points of view on the facts, but by different subsets of facts, each calling for a different explanation. Most historical linguists are in agreement as to which sub-theories are called for, at least in broad categories; where they differ is mostly on the weight to be assigned to each one. Indeed, some of the sub-theories we will need belong less to the domain of linguistics proper than to sociology or history.

### 1.5.7. American Structuralists and the comparative method

The comparative method came to be accepted as the scientific method for establishing linguistic affiliations. As such, linguists regarded it as the solid mainstay of linguistics; it was pointed to as a lasting intellectual accomplishment in the study of language.

> [...T]he comparative method is one of the most powerful theories about human language that has ever been proposed — and the one most consistently validated and verified over the longest period of time. (Watkins 1989783)

To study linguistics was first and foremost to study the classical languages and the reconstruction of Indo-European. Comparative historical linguistics has been the jumping-off place for many linguists. As Watkins notes (quoting Newmeyer):

Chomsky attributes his early interest in explaining linguistic phenomena, as opposed to simply describing them, to his early childhood exposure to historical linguistics. In a period when leading theorists tended to look upon the desire for explanations as a sort of infantile aberration, historians of language like his father, either ignorant of or indifferent to the contemporary 'scientific' wisdom in the field, clung to the 19th-century desire to explain why a particular distribution of forms existed at a particular point in time.

Indeed, until rather recently there was only praise and adulation for the comparative method, especially from American structuralists:

> ... a technique which is *more nearly perfect than that of any other science* dealing with man's institutions. (Sapir 1929, cited by Haas 1969:15)

> ... one of the *triumphs of nineteenth-century science.* (Bloomfield 1933:2)

> ... *beyond any doubt the most important historical tool ever devised* in linguistics. (Haas 1969:102; emphasis mine in all cases)[21]

The situation has certainly changed! Watkins laments:

> It is possible to get a Ph.D. degree in linguistics at a number of fine and distinguished American universities without ever taking a

> course in historical linguistics, and there are good linguists teaching in my own department who have never had such a course. The result is that there are a lot of people out there who have very little idea of what it is that we do, of what historical linguistics has accomplished. That is a pity. (Watkins 1989;784)

It is a little odd that a research method as well-established as the comparative method should be surrounded by misunderstanding and controversy both inside and outside the field. Both the method's reliance on the principle of regular sound change and the application of the results of the method for subgrouping have been criticized. These issues 'have been extensively discussed in the literature, both general and Indo-Europeanist' (Hoenigswald 1963:1),[22] but for many reasons these controversies, both specifically and generally, can in no way be called 'resolved.'[23] Some these issues are summarized in the next sections.

## 1.6. Recent challenges to and enrichments of the comparative method

### 1.6.1. Looking across disciplines for external validation

Recently the results of certain kinds of comparative linguistic research have been adduced as external validation of hypotheses in other fields, notably in the areas of archaeology and population genetics (Greenberg 1987:113, Renfrew 1990, Cavalli-Sforza 1991). Since the linguistic result (based in large part on Greenberg's method of mass comparison, discussed below) used in these cases has itself been challenged, it is obviously premature to offer it as support for other hypotheses. While marshaling independent evidence to verify related results across disciplines is a powerful validation technique, the strength of the validation is only as good as the result in each particular discipline: using controversial conclusions about human genetic data to validate controversial linguistic classifications or vice versa may result in compounded errors. As Matisoff notes, the idea of errors 'canceling each other out' evokes the image of two drunks supporting each other (Matisoff 1990a:110). Especially in historical disciplines where it is not possible to entertain the idea of reproducible results, it is tempting to try to prop up one's conclusions with evidence from other historical fields. As will be pointed out below, one result of this is that researchers in other fields are now retracing the methodological footsteps of linguists, and may be making some of the same missteps linguists have made in the past.[24]

### 1.6.2. Greenberg's mass comparison

One element of the recent controversy over historical methodology has been whether it is possible to classify languages on the basis of superficial lexical resemblances, with no attempt to establish phonological correspondences and no evidence from submerged morphology. *Mass comparison*, also called the *inspection method*, is based on the premise that a comparison of a large numbers of forms from a large number of languages will not likely be wrong because another principle, the arbitrariness of the relationship between signifier and signified, will prevent chance similarities from ever gaining the upper hand to the point of giving a false impression. Proponents of this methodology have used mass comparison to classify the American Indian languages. They argue (Greenberg 1949, Greenberg 1987, Greenberg 1990, Greenberg 1991, Ruhlen 1987) that this has in fact been the methodology used to establish the Indo-European language family, and that the success of these methods in the Indo-European case shows them to be reliable. For example, Greenberg claims that Delbrück is on his side, quoting Delbrück's remark that given that 'the formation of the inflectional forms of the verbs, nouns, and pronouns ... and ... an extraordinary number of inflected and uninflected words agree in their lexical parts, the assumption of chance agreement must appear absurd.' (cited by Greenberg 1991:128)

Certainly today the comparison of an outcome with one expected by chance is conventionally understood to be a *statistical* claim. In passing, the famous remark of Jones about Indo-European also carries this same implication, that the relatedness noted is the result of a statistical inference. Finally, Meillet, in presenting the comparative method, gave as his first example the first few numerals of three Romance languages (French, Spanish, and Italian) (shown in (1) above), with which he could assume everyone was familiar, noting: 'The relationship cannot be due to chance ' (Meillet 1967). It is clear from the rest of their texts, however, that none of these authors intended this to be their or the reader's conclusion. It is important here to distinguish the claim that a relationship exists (which can be supported at least in some cases by 'inspection') and any claims about the specific nature of that relationship (for example subgrouping). Jones compared languages in widely separated branches of the Indo-European family; the flaws in his approach have already been pointed out (§1.5.2).

Most of Greenberg's evidence for Amerind consists of lists of words taken to be similar in form and meaning, with no attempt to establish phonological correspondences. He also presents what he calls 'grammatical evidence', which is not, however, the sort of submerged morphology that other scholars consider probative. 'Grammatical' evidence is provided only in a minority of cases. Poser and Campbell Campbell and Poser 1992 note that very few grammatical equations span subgroups and that more than half are in fact restricted to a single subgroup. The significance of these facts is discussed below.

Greenbergian mass comparison has gained ground rather rapidly; indeed, the method can (and perhaps must) be applied to large numbers of languages quickly, since all that is required is the ability to 'grok'[25] the relationships based on one's (presumably informed) intuitions about relationship (Matisoff 1990a:106). On the other hand, adherents of traditional methods are in many cases not yet ready to do battle: few language groups have been reconstructed in enough detail to provide a basis for larger-scale comparisons. Indeed, the methodology for comparing intermediate levels of reconstructed languages (called *mesolanguages* or *common languages* (French *langue commune* ) (Meillet 1966) is not yet well understood, since so few mesolanguages and middle-level comparisons have been completed and accepted.

### 1.6.3. Statistical properties of reconstructions and lexicons

Some, perhaps most, of the uncertainty about how to compare and classify languages results from the sketchy and sometimes incomplete evidence which is available for comparison of any sort. While compendious lexicographic and grammatical resources exist for the Indo-European languages, most of the other language families of the world have only partial treatments. Inventorying a few extant etymological resources for several language families and subfamilies shows the varied nature of etymological sources (shown in (19 on the next page).

Even though some of these efforts have taken many years to complete, the authors themselves often note that they are merely scratching the surface in terms of what a complete treatment would require. Any judgments about genetic affiliations based on such comparative tools must perforce be somewhat tentative.

*(19) Summary statistics for a number of etymological sources (both at the level of families and subgroups)*[26]

| (1) | (2) | (3) | (4) | (5) | (6) | (7) |
|---|---|---|---|---|---|---|
| Family | Source | Lgs | Sets | Forms | Yrs | |
| Austronesian | Tryon 1995 | 80 | 1200 | 96,000 | | 27 |
| Dravidian | Burrow and Emeneau 1961 | 63 | 5557 | 46,000 | 12 | 28 |
| Indo-European | Buck 1949 | 32 | 1,300 | 41,600 | 20 | 29 |
| | Pokorny 1969 | 80 | 2,044 | 50,000 | >30 | 30 |
| | Walde and Pokorny 1930 | 80 | 2,100 | 41,000 | | 31 |
| Indo-Aryan | Turner 1962-69 | 260 | 14,845 | 143,700 | | 32 |
| Karen | Jones 1961 | 6 | 859 | 4,900 | | |
| Loloish | Matisoff 1972 | 13 | 300 | 1,800 | | 33 |
| Naga | Marrison 1967 | 38 | 917 | 6,480 | | |
| Sino-Tibetan | Benedict 1972 | 5+80 | 500 | 6000 | | 34 |
| | Matisoff forthcoming | 250 | 1300 | 41,500 | >8 | 35 |

**Legend:**

(1) Language family
(2) Source of data
(3) Languages
(4) Number of cognate sets given
(5) Total number of supporting forms in all cognate sets
(6) Years of work (if given or deducible)
(7) Notes

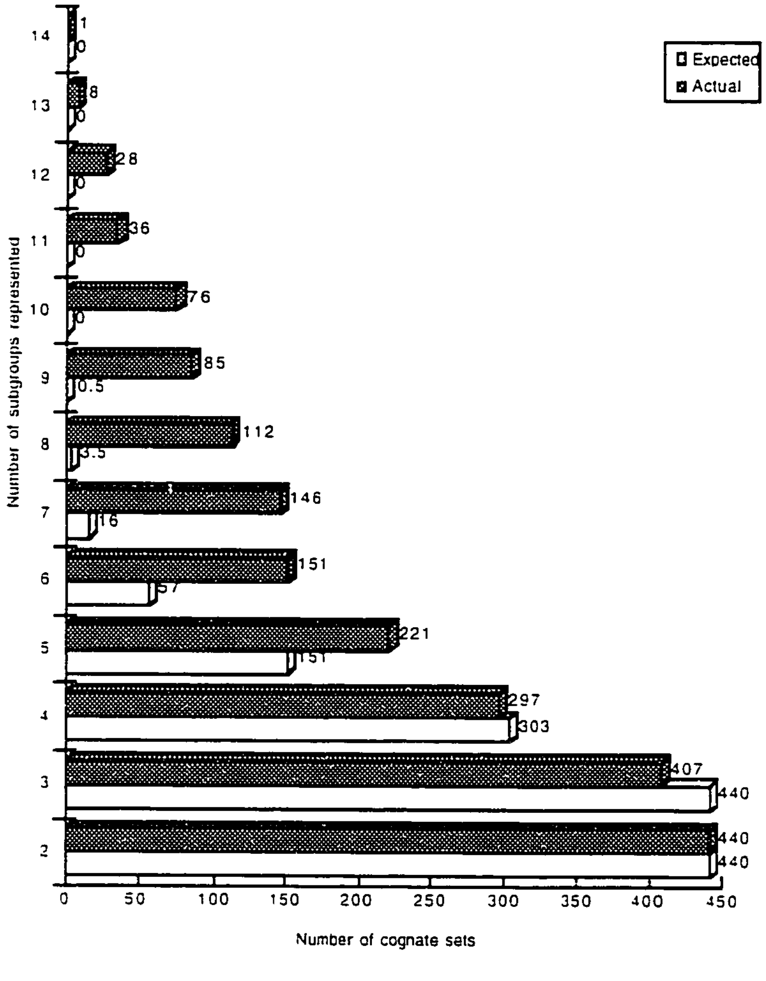
Many important properties (statistical and otherwise) of reconstructions are as yet unknown. Still, estimates of these properties based on one or two accepted reconstructions are used as evidence corroborating others. Ringe (Ringe 1992) criticized Greenberg's Amerind reconstruction by noting that most of Greenberg's reconstructions contain reflexes from only three or four of the eleven Amerind subgroups. Greenberg responded by saying that 'such a situation is quite normal, and is true for Indo-European as well.' His proof is interesting:

> I examined the first etymologies on pages 10, 20 ... 100 in the standard comparative dictionary of Pokorny [Pokorny 1969]. Of these ten items, two (pages 50, 70) were found in only two branches, four (pages 10, 20, 60, 90) in three branches, three (pages 30, 40 100) in four branches, and one (page 80) in seven branches. Indo-European may be regarded as having eleven well-documented branches, thus matching the eleven branches of Amerind. (Greenberg 1995:82)

One might be tempted to believe that a consideration of ten sets (out of a possible two thousand) would not yield a significant result. However, if the question is re-phrased to ask how many sets have no more than four members, then the answer on the basis of this small sample (90%) is indeed statistically significant. Were the distributions of accepted reconstructions known (and themselves accepted), it might be possible to compare them with proposed reconstructions to determine whether the proposed reconstruction was a good one. This has been attempted to some extent. Ringe (Ringe forthcoming) has calculated that the distribution of cognates by subgroup in the cognate sets listed in the Indo-European dictionary of Pokorny is unlikely to have arisen by chance. He first computes the expected distribution of subgroups in cognate sets would be if the probability of a subgroup having a supporting form in a cognate set were one in five. Normalizing this curve to the size of Pokorny's reconstruction (n = 2044), it is clear that the two curves (shown in (20) below) are quite different.

(20) *Representation of Indo-European subgroups in cognates sets*

*(based on* Pokorny 1969, *cited from* Ringe forthcoming:*16 and* Bird 1982*)*[36]



Ringe then shows that the list of Nostratic reconstructions of Illich-Svitych 1971 exhibit a pattern of distribution across subgroups which is not different from chance.

The argumentation for both these points is based on the premise that reconstructions have consistent properties across languages, and the methodological problems with the data will not skew the results. While this may be true, we are a long way from having statistical summaries of reconstructions which would facilitate reaching these conclusions. For those languages for which decent (i.e. relatively exhaustive, complete) reconstructions have been published, a substantial conversion and analysis process is still required to extract this information. And most families are not yet this far along.

### 1.6.4. Data must be published!

In this regard, Greenberg defends himself against Ringe's reproof that the data on which is classification is based is not published by noting that the original data, contained in the famous twenty-three handwritten notebooks, is indeed available via inter-library loan. An extensive quotation of Greenberg's response here is merited:

> In these notebooks, the source of every individual item is given. To have published the exact source of each form cited in the book would have added enormously to the length and cost of the book. Moreover, in no etymological dictionary of which I am aware, for example, in those of the Indo-European and Uralic language families, are the sources of each form cited.

> I have been assuming that Ringe is simply asking for the sources of the several thousand forms cited in the etymological portions of my work. A closer reading of Ringe suggests, however, that in order for him to test the validity of Amerind by the statistical methods he advocates, he requires the publication of my entire database [sic]. If such is indeed the case, he is asking me to publish upwards of a quarter of a million words already available from my notebooks. (Greenberg 1995:83)

From a scientific point of view, there is no question but that any claim should stand or fall on the evidence it is based on and that that evidence should be easily available. In most other scientific fields the sharing of such data is routine and failure to do so is grounds for censure. We should therefore urge Prof. Greenberg to make such a publication in order to facilitate the investigation of his claims.[37] The creation and dissemination of such a store of lexical information is exactly what is needed as a preliminary step towards settling questions of reconstruction and affiliation. Certainly paper publication of such a word-hoard would be expensive; but the costs of electronic publication (even in image form initially) would be quite modest. I am certain, though, (and will show in subsequent chapters) that such a distribution would only be the beginning of a long process of evaluation, correction, and interpretation.

It should also be noted that such gross statistical comparisons must miss some significant points about the reconstruction process. For example, etymologies which are found only in a single subgroup, exemplified by the set of forms from Burmish languages below, may not be useful in large scale comparisons of the sort discussed above, but with enough evidence can provide solid support for subgrouping.

(21) *An etymon meaning* shoulder *apparently found only in a single *TB subgroup (Burmish)*[38]

| Language | Form | Source | |
|---|---|---|---|
| Zaiwa=Atsi | koʔ$^{21}$san$^{51}$ | ZMYYC | 250 |
| Achang Lianghe | lɑ$^{31}$ sɑn$^{55}$ | JZ-AClh | |
| Achang Luxi | laʔ$^{31}$ san$^{35}$ phu$^{55}$ | JZ-AClx | |
| Achang Xiandao | lɔʔ$^{31}$ san$^{55}$ kʰʐ̯oʔ$^{55}$ | DQ-Xiandao | 114 |
| Bola | laʔ$^{31}$ sɛ̃$^{55}$ | DQ-Bola | 114 |
| Langsu (=Maru) | lɔʔ$^{31}$saŋ$^{31}$ | ZMYYC | 250 |
| Lashi | lɔʔ$^{31}$ sa̱n$^{33}$ | DQ-Lashi | 5.3 |

Also, statistical estimates must give each form an equal weight in the result, and most historical linguists would claim that some words are worth more than others in providing evidence. Of course, statistical studies could eventually take this into account, but estimating weights of this sort would be a difficult and methodologically suspect process.

Much of the debate is anticipatory: since a classification of e.g. the Bantu or Sino-Tibetan languages according to traditional means (i.e. detailed reconstruction of common languages followed by typological and other comparison of mesolanguages for grouping into larger phyla) has not yet been accomplished, no challenge can currently be raised to the efficacy of the traditional methods: there is no result to dispute! On the other hand, alternative methodologies like mass comparison resist refutation except on methodological grounds precisely because so little actual reconstruction has been accomplished. In these cases, proponents of the comparative method point out methodological flaws in such alternative methodologies, noting, for example, that a classification of a language based on six words is unlikely to be reliable (Poser 1995:Linguist List electronic newsletter). However, there is no accepted metric for deciding how many words are necessary to reliably establish an affiliation, whether the comparative method is the tool used or not.

### 1.6.5. Does the comparative method apply across languages?

It is probably true that the comparative method actually does not apply in one and the same way across different language families, as some groups seem to yield easily to traditional techniques whereas others require broad knowledge of the linguistic milieu, sophisticated analysis, and an intuitive knack for diachronic phenomena (i.e. what has been called *Protosprachgefühl* (Matisoff 1982)). The former have been dubbed 'reconstruction-friendly', the latter 'providing less joy to the linguist' (Hoenigswald 1990). The typological properties of the language family as well as the typological properties of the reconstruction process itself affect the results. As Gamkrelidze notes, 'comparative reconstruction must go hand-in-hand with typology and language universals' (Gamkrelidze 1989:118, Garrett 1991).

### 1.6.6. Opposition to the notion of exceptionlessness

Opponents of the neogrammarian principle of the regularity of sound change have from the beginning found their strongest argument in the existence of a non-negligible proportion of exceptions to the postulated regular laws. After the vociferous quarrel of the years 1870-1880 between proponents and opponents of the principle (with Hugo Schuchardt as the most visible figure of opposition), for most of this century it seemed that peace had been established in the shape of a not fully explicit synthesis, weighted in favor of regular sound change but allowing for other processes as well. In recent years, however, this equilibrium has been brought into question, mostly following the work of William Wang on lexical diffusion (§1.6.7 below), to the point not only of diminishing the part played by regular sound change, but of removing it from the core of the theory, and even denying it any place whatsoever.

Exceptions, real or apparent, are a major methodological problem for the historical linguist. Determining what constitutes a genuine exception and what does not is of fundamental importance to the further development of the theory of sound change.

### 1.6.7. Wang's version: the 'radical lexical diffusion' theory

Wang's theory, though concerned with lexical diffusion, is a radical departure from 'classical' lexical diffusion, and in accordance with previous work of mine and others, I will refer to it here as the 'radical lexical diffusion' theory (or RLD). Mazaudon and Lowe forthcoming 'Lexical diffusion' as classically conceived is a process of change which has been recognized for a long time as a particular case of linguistic change. Everybody agrees, and has agreed for decades, that those changes in sound which result from analogy, restoration, hypercorrection, imitation of socially prestigious groups, and the like, are often (though not always) propagated through the lexicon one or a few lexical items at a time. In the course of history, some linguists have given more weight than others to these 'external' causes of change.

The parade example of 'lexical diffusion' according to the classical theory is probably Bloomfield's presentation of the evolution of the pair —'mouse'/'house' in Dutch dialects (Bloomfield 1933: 329-331). Here we see early Germanic [*u:] developing into a variety of phonetic realizations [u:. aw. y. y:. øɥ. ø:]. all respecting the structural identity of the vowel in the two members of the pair — except in one border area between the [u:] and the [y:] zone, where we find [y:] for 'house' but [u:] for 'mouse'. Bloomfield explains the different reflexes by the fact that [y:] was a prestige pronunciation imported from Flanders; a word like 'house' is more often used in conversation with outsiders, where one is using elegant pronunciation, than a homely word like 'mouse', whence [hy:s] but [mu:s]. Conversely, on the east of the zone, the North German Hanseatic cities constituted another source of cultural influence, tending to maintain [u:].

'Our isoglosses of *mouse* and *house*,' says Bloomfield, 'are the result of the varying balance of these two *cultural forces*' [emphasis added].

If Wang were only underscoring the omnipresence of diglossia if not bilingualism, and emphasizing the importance of studying these factors of language change to uncover the actual development of a language—as opposed to the selective, overly abstract account that the exclusive study of regular sound change would provide—then this would not be a new theory. Under this interpretation, Wang's methodological position would simply be advocating the same kind of study that sociolinguists like William Labov are conducting, and which no historical linguist perceives as resting on a theory which contradicts his own, but only on a sub-theory which complements his own sub-theory. This is the way Martinet, in a conciliatory note of 1987 (Martinet 1987) understood Wang's position, and so also Haudricourt and Hagège in their *Phonologie panchronique* (Hagège and Haudricourt 1978:53). Such an interpretation of course makes Wang's approach acceptable to historical phonologists, but it also empties it of all new substance. This is not what Wang is claiming.

Labov's extensive article Labov 1981 assesses the main point of the theory correctly: Wang asserts that sound change even *in a homogeneous community* proceeds *from its very inception* by lexical diffusion. The theory was originally formulated to account for irregularities which, says Wang, remain after all efforts at treatment with the tools of the classical theory—a thorough analysis of phonetic context, analogical influence, and possible foreign sources for the vocabulary—have been exhaustively tried and have failed. Faced with such irreducible irregularity, Wang was led to propose that we should establish irregular change at the very root of phonological evolution, and stop considering it (as the classical theory had done) as a sort of accident. [39] When Wang and his collaborators talk of 'competing sound changes', they are not referring to competing cultural influences entering the language from the outside, but rather to competing and coexistent language-internal evolutions—or at least they do not think it necessary to distinguish between the two apparently very different sociolinguistic situations (see for example Chen 1986a). The 'radical lexical diffusion' theory is meant to replace the theory of regular sound change as a global theory, not to complement it as a sub-theory. In later articles, Wang and his collaborators have given lip-service to 'neogrammarian'-type sound change, either in passing remarks or in general conclusions (Wang 1979).

It should be mentioned that clear cases of lexical diffusion are much less numerous than has been claimed. Mazaudon and Lowe forthcoming The founding example of the theory, the apparently unmotivated split of a tonal category in the Chinese dialect of Chao-zhou (Wang and Cheng 1977) has been convincingly reanalyzed as a case of language contact by Søren Egerod (Egerod 1982). Egerod shows that there exists in that dialect a colloquial level of speech and a literary level; the latter constitutes a different dialect, with its own regular development from Middle Chinese, distinct from that of the colloquial level.

Nevertheless, there remain cases where actual language contact cannot be shown and where change seems to be diffusing from word to word in a monolingual situation. Andersen 1973 handles one of these troublesome situations. Without going into the details of the Czech case, we may follow Andersen in distinguishing two kinds of 'evolutive changes', defined as changes which are 'entirely explainable in terms of the linguistic system that gave rise to [them]': 1) changes which occur without stylistic variation being established and 2) changes where a stylistic variation gets established. In this second case, Andersen proposes schematically the following process: 1) A regular sound change occurs between one 'generation' and the next. 2) A restoration to the old form occurs for some lexical items under the influence of the older generation. This second phase is a typical 'contact' situation, the limiting case of 'borrowing', and it occurs, like all such events, on an item by item basis, or perhaps one word class at a time. 3) The double reflex of the proto-phoneme is acquired as a stylistic variation by the third generation, which is exposed simultaneously to the speech of both of the earlier generations, with the newer form as the unmarked term. Stylistic variants typically apply on an item by item basis, and unless hypercorrection occurs, they eventually die out. *What is lexically diffused here is not the change, but the* *restoration.* The change has occurred regularly, at the inception of evolution, and in that sense it is fundamental; it is at the source.

With this scenario, which Andersen expresses as internalized rules of the grammar, and which we have retold in the psychological, or cognitive terms that the rules represent, it becomes possible to understand the fact that although most changes are most regular at their beginning and acquire irregularities as time passes, some changes seem to regularize themselves over a long period of time. It also becomes possible to reconcile the findings of dialect geography with those of classical comparative phonology. It might be important to note here that stylistic variation is not as likely to be established in all types of social settings. If we want to study 'internally motivated changes' as a clue to general phonological processes, we might want to select for study communities where stylistic variation is either small or strictly codified.

From a methodological point of view it could be considered that the Wang scenario and the Andersen scenario are equivalent: they end up the same. But from the point of view of the theory of linguistic processes of change they are very different.

The theory of regular sound change is presented by Wang as a theory which has, so to speak, been shown wrong with honor. He professes great respect for the accomplishments of that theory, and claims it was the best theory for its time, as it was methodologically very productive; but even a wrong theory, as everyone knows, can continue to be useful and be held on to for some time after it has been shown to be 'wrong' (Mazaudon and Lowe forthcoming).

### 1.6.8. Mischsprachen

The challenges to the comparative method resulted in the perception that historical linguists had (and perhaps have) an overweening interest in finding regularity. The last ten years have seen a resurgence of interest in contact phenomena (cf. for example Thomason and Kaufman 1988). The comparative method is neutral with respect to 'the case where one would have to take account of two initial systems and of their reactions to each other' (Meillet 1967:102); the method cannot be applied. Yet this is exactly the situation which gives rise to pidgins.

Thus it makes little sense to try to reconstruct the ancestor of such 'mixed languages.' The well-known example of Japanese, which has affiliations with both the Altaic and Austro-Tai families, is a case in point. Researchers have claimed that the language is one or the other with later 'strata' borrowed from the other family. (Miller 1991,Benedict 1975)

The dominance of the neogrammarian view and the political unpopularity of the idea of 'mixed languages' tarnished its reputation until the 1960s, when fieldwork on creoles and pidgins resumed. Some researchers (e.g., Bickerton 1984) claimed that creoles provided insight into the workings of universal grammar. The wealth of new theory about processes of creolization may shed new light on historical processes, and in fact, such a rapprochement is overdue (McWhorter 1995).

### 1.6.9. 'Scaling up' the analysis: micro-, macro-, and megalo-reconstruction

Another methodological problem confronting the comparative linguist has to do with the 'scale' of the comparison and the standard of proof required by the choice of scale. The most productive and acclaimed protolanguage reconstructions have typically been based principally on a few carefully chosen languages.[40] To include a larger set of languages complicates the reconstruction task exponentially; however it is important to test the reconstruction with new data. This will surely complicate the analysis and lead to the discovery many more non-cognates and difficult cases. The result, however, is still a better result, more reliable and general. As a practical matter, the comparativist must keep track of the many correspondences discovered and the forms in which they are manifested. This is a laborious task even for a small number of languages. More data does not necessarily (or even typically) make the situation clearer, just more realistic. Adding more data adds more irregularity as well: while some cognate sets become better supported, others remain unchanged when the cognate form is not retained, giving the impression that the status quo might not be enough. A cognate set which has two forms looks fine if there are only two language in the comparison; if there are fifteen it is not so impressive. This simple arithmetic fact is often overlooked.

The initial steps of comparison must necessarily be followed by further attempts to fill in the gaps. The ideal goal is a complete explication of all the words in the lexicons of all languages in the subgroup. While not all words are entitled to an etymology (Meillet 1967:58), those that are must find their way into a niche, i.e. a cognate set or a borrowed form from which the word derives. As more languages and more words are brought in to test some hypothesis of a particular relationship, the hypothesis will inevitably bend and creak, perhaps break, to be replaced with another. Additional evidence sometimes contradicts earlier hypotheses, or at least may appear to, as this quote from Ruhlen illustrates:

> How can it be that the dramatic increase in both the quantity and quality of Sino-Tibetan sources during this century seems to have had the effect of eroding confidence in the Sino-Tibetan family? Have these new materials really called into question the validity of Sino-Tibetan? Of course not. What has happened in the study of Sino-Tibetan parallels what happened in African classification: the standard of proof of genetic affinity has been modified to such an extent by the requirement of regular sound correspondences that even such obvious groupings as Sino-Tibetan fail to meet the new requirements. (Ruhlen 1987 )

Whose confidence, we might ask, has been eroded we may ask? A fitting response would be that while the quantity and quality of field work in these languages has increased manyfold, the analysis and integration of these new data into the Sino-Tibetan scheme has not kept pace. The amount of time and effort required for comparison and reconstruction is on a par with that required to gather the data in the first place (to wit, many years).[41] The process is always somewhat lumpy and fitful, as the research must sometimes pause to await new data and revision, only to rush forward to the next obstacle. In the course of this laborious process one is occasionally led astray, loses track of the state of things, or may begin to feel nagging doubts about significant details. The standard of proof in Sino-Tibetan *has* been modified, and soon we may expect a fuller and more precise explication of the relationships between the Sino-Tibetan languages, based on systematic and precise comparison of the new wealth of data.

As noted above, it is at the level of microcomparison that the most precise statement of variation is required. It is at this level that sound change can be observed 'in progress.' Precise records of phonetic detail are required, as the differences between dialects may be quite small. Strict semantic and phonetic criteria for accepting cognates must be applied. Since by definition comparison at this level involves languages or dialects in relatively close contact, it is of crucial importance to distinguish inheritance from influence.

'Macroreconstruction,' that is, reconstruction between subgroups of the same family, works with much greater divergence between languages. At this point, the comparison is between reconstructions resulting from internal reconstruction and/or comparison within each participating subgroup. The decision of whether or not to accept a certain semantic fit becomes more problematic. Differences which have developed in phonological and morphological structures may make it difficult to find appropriate units for comparison. Moreover, there are usually many languages to choose for comparison at this level. As noted above, it is important initially to select a representative few, lest the forest not be seen for the trees.

Megaloreconstruction is the most perilous form of reconstruction. It may also be the most fun, as the limits on what is acceptable are essentially set by oneself, and there is an almost limitless range of data to choose from. Language universals may influence the perception of relatedness: most languages of the world contain monosyllabic or reduplicated monosyllabic forms beginning with voiced bilabial or dental nasals as referents for one's parents, but reconstructing such words in the megaloprotolanguage is not very persuasive. Protolanguages are fairly imprecise objects to start with, in terms of their chronology, phonology, and semantics; *proto*protolanguages can only be less precise. Given that sounds change and words are replaced, there must be some limit to the depth at which the comparative method could possibly apply. This limit, even under the most generous of interpretations, must be much later than the time when *homo sapiens* actually began speaking. The implication is that many languages and language families have come and gone before the particular developments we have access to today. It is difficult to believe that the linguistic past we dimly perceive today represents the beginnings of human speech, or manifests some past unity over and beyond that imposed by putative universals of human cognition and biology.

### 1.6.10. There are few hard-and-fast rules

Because so little actual reconstruction has been done (I am speaking on a global scale), and because so little of it is available in a coherent testable form, it is difficult to say just what the characteristics of a 'good' reconstruction are or should be, and to compare reconstructed languages across and even within language families. Many Indo-European scholars have explained at length the principles used for reconstruction and for judging reconstructions (Meillet, Hoenigswald, etc.). However, the metrics they provide are necessarily not hard-and-fast rules, or countable relations.[42] Each language is situated in its own social, geographical, and historical milieu which determines how accurately it can testify to the nature of its ancestors. An example where revision in standards has resulted in a dramatic revision of supposed relationship is Thai and Chinese (Maspero 1911; Maspero 1952). The original belief that Thai descended from Chinese was based on the extensive amount of Chinese vocabulary found in Thai. Benedict's revision Benedict 1942, showing that Thai was a member of a distinct family (Tai-Kadai), is based in part on a lexicostatistical study. It is likely that as time goes on and these methods are refined, further adjustment of our understanding will be required.

### 1.6.11. A predilection for iconoclasm

Besides the problem of incomplete data sets coupled with incomplete analysis, uncertainty can also result from methodological problems. The perception of disarray (*anomaly* in the Kuhnian sense) gives alternative theories a shot at the status of paradigm. (Kuhn 1962:77ff). Modern linguists, especially American linguists, seem quite ready to gives up the dominant paradigm. And perhaps with good reason. The dominant paradigm requires a lot of work, often years of work to come up with a result. Also, Americans place a premium on novelty, especially in science, further increasing the incentive to do something different whether it is an improvement or not.

In this case the dominant paradigm is (or is represented by) the comparative method. The comparative method as properly applied is a tool for verification of relationship. However, it holds the promise of answering other questions as well. The process of reconstruction cannot be divorced from questions of typology, subgrouping, and language universals. While the use of the comparative method for purposes of reconstructing lexicons has gradually evolved into a process which can be carried out fairly routinely with scientific precision, the interpretation of its results for other purposes is still very much an art. Nevertheless, progress will not wait; researchers have many other questions which need answering, questions which the comparative method is not currently (and perhaps will never be) able to answer in a straightforward way.

### 1.7. How cross-linguistic lexicographic databases might help

Until a substantial body of well-arranged and easily accessible evidence is shared among linguists, including a means of expressing claims of affiliation and comparing such claims, it is unlikely (in my opinion) that the several bones of contention discussed above can be resolved. Whatever approach to establishing language relationships, the results can only be improved by as Matisoff notes, 'looking at well-recorded and well-analyzed forms from as many languages as possible.' (Matisoff 1990b:118). Providing such a classified data set presents a number of challenges:

- The data set would have to be large enough to contain a representative sample, perhaps even the complete inventory, of the linguistic structures found in each of the languages used for exemplification. Such structures have traditionally been lexical items and morphological features, but should include syntactic and semantic facts as well.

- It would have to contain a large enough set of languages to address issues of statistical validity and coverage. For each language it would have to contain enough forms and information about the forms to satisfy the need for completeness.

- The content and structure of the data set itself would have to be precise and accurate enough to be relied on as a reference source. If too many shortcuts, inconsistencies, or out-and-out errors are found the data set becomes useless. This may seem obvious, but given the range of data which have to be included, finding techniques and representations which work in all or even most cases is difficult.

Clearly a data set of such depth and breadth would require a very powerful computational structure. To be useful it would have to be presented in a way that would allow scholars to examine in minute detail its contents and the relationships embodied therein. Restricting the discussion now to the lexicon, the traditional domain of reconstruction, the ideal data set would contain the entire lexicon for all languages of the world.[43] These lexicons would necessarily be linked to grammatical descriptions, since words themselves in isolation are of limited use to the linguist. The lexicons would be have to be adjusted to permit cross-linguistic comparison. Despite advances in description in the past hundred years, substantial problems exist in comparing forms across languages. Since all elements of a language are specifiable only contrastively within the language itself. comparing individual elements across languages must also take structure into account somehow. The implication is that the lexicons would also have to be linked to phonological descriptions as well. This is a well-understood constraint on interpretation of transcribed data: Pokorny's *Etymologisches Wörterbuch*, for example, sorts each language in his *Register* in the traditional lexicographic sort order for the language and gives that order at the head of each language's list. The recently published *Comparative Austronesian Dictionary* devoted nearly two entire volumes of the five-volume set to descriptions of the eighty languages treated (Tryon 1995:Part I, fascicles 1 and 2)).

Until such a body of evidence is presented (language family by language family), with concomitant statements of the principles around which the evidence is organized, both the methodological and the theoretical debates will continue: until testable hypotheses are created and repeatedly tested, no firm result can be claimed. And even then, in the strictest sense, nothing is proven.

### 1.7.1. Logical positivism

I will argue here for an approach to historical reconstruction based on a strict logical positivist program for evidence and argumentation. The model for reconstruction presented, based on rigorous computational verification of correspondences against extensive data sets, would satisfy the requirements of even the most ardent neogrammarian. The Popperian view of the scientific method is assumed: all hypotheses must perforce remain unproven, and once counterexamples surface, the hypothesis must be abandoned, modified, or clung to in defiance of the evidence.[44]

It is reasonable to ask why such an effort should be made, and to question the likelihood of its success. After all, real language is rather messy and is unlikely to conform to the needs of a neat theory. The answer to this is that we wish to improve the empirical basis for answering the questions of method and interpretation which continue to challenge the discipline of historical linguistics.

'Proofs' of genetic relationship have many properties of which make their validity difficult to settle once and for all. These properties have been discussed above, but I wish to emphasize a few of them again:

- their size: vocabularies and grammars of many languages must be compared, and therefore huge numbers of items must be examined.

- their complexity: such proofs often rely on several different explanatory mechanisms (regular sound change, analogy, borrowing, etc.) which interact with each other.

- their heterogeneity: sometimes the types of evidence for relationship in one group cannot be used in another. And the evidence may be better at one time depth than another. For example, for reconstructing early Indo-European, both lexical roots and morphology can be reconstructed with certainty given the ample evidence from classical languages, later intermediate stages rely more on lexical evidence; for Proto-Bantu (and Proto-Sino-Tibetan), lexical roots are most secure at a great time depth; in later stage and in some subgroups, some morphology can be reliably reconstructed (e.g., in the Himalayish subgroup of Tibeto-Burman (Michailovsky 1976)); for Tai-Kadai (which may not be reconstructible due to the extent of borrowing and contact), shared features of grammar (Benedict 1942)

Providing a means for examining such proofs in detail would provide a service to the field: a scheme for cataloguing the data, a set of specific claims made about the data, and tools for arranging the data according to alternative hypotheses would go a long way towards resolving differences. It would also be an extraordinary pedagogical tool in many other areas of linguistics.

### 1.7.2. The Enormous Theorem

Proofs of the type just discussed fall into a larger class. Linguists are not alone in having such problems. Mathematicians have recently had to confront them as well: the famed Four Color theorem was finally proved by brute-force exhaustive enumeration of the possibilities by supercomputer. The process took several months of computer time. Many mathematicians are understandably dissatisfied with a proof which cannot be 'synoptically' understood by a human being.

There is another proof of this type in mathematics which bears an even greater resemblance to genetic proofs in linguistics. This is the proof of the so-called Enormous Theorem,

> [...]one of their [mathematicians] most important theorems, describing the taxonomy of the mathematical objects known as simple groups, has a proof that runs an estimated 15,000 pages, spread over upwards of a thousand separate papers written in widely varying styles by hundreds of researchers. Cipra 1995:795)

Part of the reason the original proof turned out to be so long is that the four (major topological) categories have widely varying properties, so that unifying concepts are hard to come by. The proof is the result of a gigantic long-term collaboration between mathematicians world-wide who have '[...]chipped away at the [...] problem for 30 years[...]'. One of these mathematicians (Ron Solomon, of Ohio State) has remarked that if 'the generation of people who worked on the proof were to vanish, it would be very hard for future generations to reconstruct the proof out of the literature' (Cipra 1995).

'To some mathematicians, it's worrisome that they can't check the theorem on their own.' Noting that 'everyone would be happier with a shorter, more readable proof,' several researchers have embarked on an effort to redo the proof, streamlining it as they go. Nevertheless, 'the second generation proof is still going to be gigantic, with estimates ranging between three and five thousand pages' (Cipra 1995).

### 1.7.3. Similarities and differences between the Enormous Theorem and 'proofs' of linguistic affiliation

Several striking similarities between the challenges confronting both mathematicians and linguists immediately come to mind. Consider, for example, the case of Indo-European, a linguistic unit whose genetic unity is universally considered 'proven'. Just what is it? How big is it? Where is it? What would you point to as the best place to look to find out details about the proof?

#### Similarities between linguistic and mathematical proofs

Like the four color theorem and the Enormous Theorem, proofs of affiliation using the comparative method are proofs by exhaustive enumeration; the number and complexity of the enumeration make the proof too large for one person to be 'sure of' (how many articles and pages does the proof of Proto-Indo-European take up?). They also, in consequence of their size, require substantial storage capacity (of one sort or another); One specialist cannot hold all of the relevant information in his or her head (but should in principle know where to look for it).

Like the Enormous Theorem, linguistic proofs are often heterogeneous proofs. The entities involved are of disparate types; that is, different argumentation and types of evidence are needed, and the strength of the different sub-parts of the proof is not the same.

### Differences between linguistic and mathematical proofs

Mathematicians believe that theorems are disproved if even a small facet is disproved. As in the case of a computer program, a misplaced comma or a typo can bring the whole thing crashing down. In historical linguistics (and indeed in linguistics and perhaps science generally), however, the proper role of counterexamples and missing data remains unresolved. Since counterexamples, or at least difficult examples, abound in historical linguistics, this problem can make it difficult to determine whether to reject a given proof of genetic relationship. Counterexamples have a special status in linguistics and their own history and literature. (cf. Inkelas 1993:556, Mazaudon and Lowe forthcoming)

### 1.7.4. One possible answer to the problem of large proofs

A possible answer to the problem of 'large, spread-out, heterogeneous proofs' is to provide the entire proof in such a way that any part of it can be quickly retrieved and examined. A computer is in many ways the perfect tool for making such a catalog of data and claims available. However, as will be discussed in detail below, the complexity of linguistic data and of their interpretation constitutes significant obstacles to making this possible. The model of 'computer-assisted proof' offered here is based on the 'closed-catalog' model of grammar proposed by (among others) Vennemann.

Inasmuch as no such closed-catalog model of a language family exists (even if Indo-European arguably has such a model, it is not explicit in a single work), the work of reconstruction (in any family) must always be regarded as tentative and provisional.

### 1.7.5. The closed-catalog model of a grammar

Before treating the closed-catalog model, a few words about the notions of *exhaustivity* and *exceptionlessness* in linguistics are in order. The certitude with which two words can be said to be cognate is proportional to the precision with which the sound correspondences between them can be stated. It is worth quoting Meillet at some length about this:

> When it is a matter of words really going back to the 'common language,' it is necessary to reconstruct a word of this language in *every respect,* and not to be content with comparing small root elements. And as the risks of error are great, it is necessary to assure oneself *precisely* that the agreements observed are not fortuitous.
>
> The first point, on which [every]one is in agreement in fact if not in principle, is that an etymology is valid only if the rules of phonological correspondences are applied in an *exact* way, or in case a divergence is accepted, if this divergence is explained by special circumstances rigorously defined. ...

> The agreement in meaning should be as *exact* and as *precise* as the agreement in phonological form (according to the rules of correspondence). [Meillet 1967:51] [italics mine]

There is clearly here no notion that the action of sound laws and semantic shifts is in any way 'fuzzy' or 'statistical'. The best analysis which *exhausts* all the data, both in terms of the range of forms treated as well as in its treatment of each form, which has the fewest unsupportable *exceptions*. I interpret Meillet's injunction as requiring the principal metric of the reconstruction process to weight these two criteria heavily in judging reconstructions. This is an ideal, of course, those words which have such precise agreements in form and meaning are to be treasured. Many times the data cannot support such a high degree of certainty (which is one reason why reconstruction is at least as much art as science). Not all words, even those that are cognate, are of equal value in reconstruction.

To satisfy the exhaustivity and exceptionlessness criteria imposed by accepting a strict interpretation of the comparative method, I assume here a model of data and accompanying explanation proposed by Theo Vennemann. Vennemann's view of the responsibilities of linguists is worth quoting:

> The central task of our discipline, linguistics, is to produce a theory (or theories, if you are a pluralist) of all human languages, no more, no less. Since this task cannot be accomplished with one

> blow, we break it down into many smaller tasks and produce theories covering only sections of the total object: partial theories. Partial theories may be partial in either or both of two senses: they may be partial general theories, i.e. theories covering all languages but only certain of their properties [or they may] be partial by being language-specific, viz. by being intended to hold for only a limited group of languages[...] (Vennemann 1983:6)

Vennemann continues in his (by his own admission) unoriginal elaboration of the philosophical underpinnings of linguistic theory, examining the role of external evidence in linguistic explanation and laying the groundwork for what has become a comprehensive theory of language structure based on preference laws. But it is his discussion about what constitutes a theory of language change which interests us here. Dismissing the assertion that language change cannot be explained, he asserts that 'there can be such theories, theories of precisely the Popperian kind.' According to Vennemann, 'a list, a 'taxonomy' as it used to be called, or a 'catalog' as I prefer to call it in order to avoid the negative implications of the term taxonomy with the additional stipulation of *closure*, constitutes such a falsifiable theory' (Vennemann 1983:18). A list with a closure stipulation is simply a finite list; a 'closed catalog', then, is a Popperian theory because it is 'capable of falsification: any *bona fide* language change not subsumable under one of the types of language change specified in the closed catalog falsifies the closed catalog.'[45]

Falsification of a closed catalog may not be very damaging, however, for, as Vennemann notes, 'either we have not considered a type of language change, in which case we may simply open the catalog, let the type in, and close it again; or we have formulated too tight a constraint on one type of change, in which case we need only modify or remove the constraint' (Vennemann 1983:19).

Constructing a closed-catalog theory of language change is by no means a trivial matter, as Vennemann in an understatement observes (Vennemann 1983:19). Creating such a catalog even just for Indo-European, a family in which the laws of sound change are relatively well-known and described, would entail a substantial bibliographic and analytic research effort, the result of which might be some improvement in data access and argumentation (cf. for example Collinge 1985) and better pedagogy. It is unlikely that such an effort would provide many new insights into Indo-European languages. In Tibeto-Burman, on the other hand, and other less-reconstructed language families, where (to use the words of Matisoff (Matisoff 1991)) 'we still do not know the sound laws or *Lautgesetze* of [... the ...] various languages', such a catalog would be more useful.

The catalog would have to include at least the following:

- The data on which the reconstruction depends, including not only the forms which support the reconstruction, but the counterexamples as well.

The system for representing the data should provide for retrieving and reconciling several instances of the same datapoint (e.g. the same word may be provided from different sources, transcribed in different ways).

- A list of the sound changes, with 'metainformation' about the changes, such as the source of information, the forms in which the change is observed, etc.

- A means for cross-checking and searching the data and proposed changes.

A catalog of sound changes is likely to undergo a great deal of revision, especially in the early stages of its development. Representing the catalog on a computer eases the task of revision and allows for a number of improvements in the method as discussed below. The Reconstruction Engine (discussed in §6), is an example of one part of such a catalog. Some problems and possibilities for designing more general-purpose data structures and algorithms, called in this case a *sound law database* , are discussed in §7.

But there are a number of issues that need to be treated before building a sound law database: a sound law database would, after all, rest on a body of lexical evidence (entered into the computer); software would have to be provided to make the intricate links between the bits of evidence. There is some prior work (though not much) in this area which I review in the next chapter.

# 2. PRIOR ART

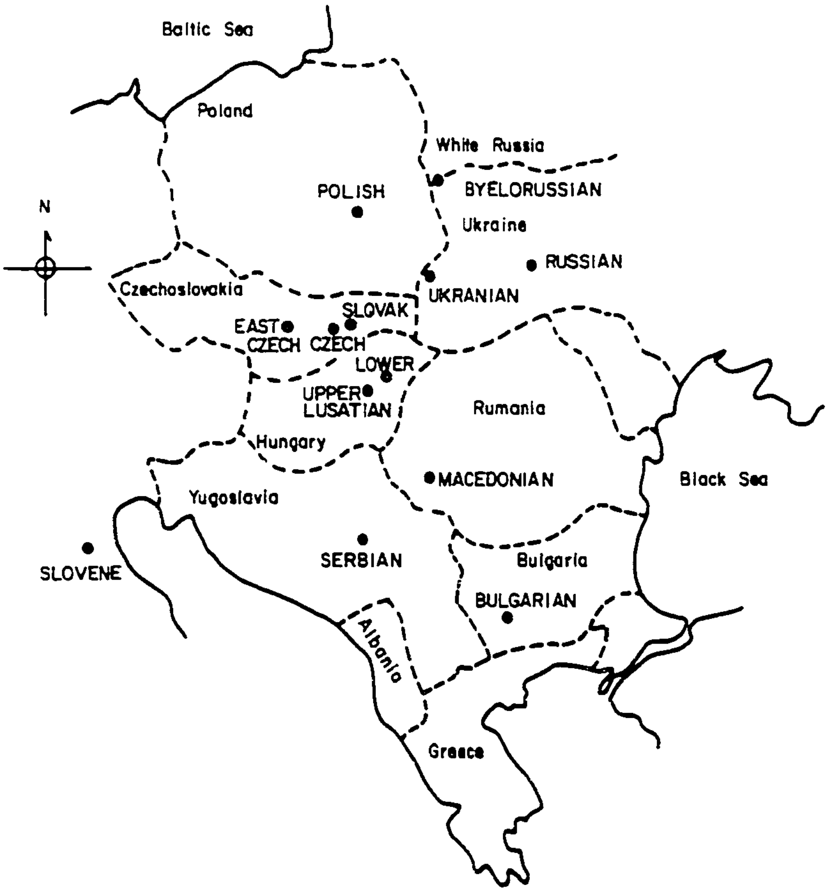
Linguists and programmers have taken a variety of approaches to make computers do useful work in historical linguistics. Prior attempts have met with varying degrees of success and have been carried out with varying degrees of commitment; some have a long history, while others were essentially experiments, brief forays into the area of computer applications in linguistics.

Computer applications in historical linguistics fall into two distinct categories: those based on numerical techniques, usually relying on methods of statistical inference; and those based on combinatorial techniques, usually implementing some type of rule-driven apparatus specifying the possible diachronic development of language forms. The major features of a few of these programs, and mainly those of the rule-driven variety, are reviewed briefly below. The projects reviewed below do not exhaust the field of computational historical linguistics, especially if lexicostatistical approaches are included. Indeed, lexicostatistical approaches dominate the computational historical linguistic literature. Here, however, I will eschew most discussion of this work in favor of one particular such approach since it is this approach that is the focus of this dissertation. The criteria for selecting the particular set of projects is that they have been described in the literature and elsewhere sufficiently for an evaluation. The literature in this subfield of computational historical linguistics is fragmented; starting in the 1960s and 70s a sizable literature on the lexicostatistic properties of language change developed in the wake of Swadesh's earlier glottochronological studies (for example Swadesh 1950) and later (Dyen 1969; Dyen 1970; Dyen 1973; Dyen 1975; Dyen 1992). On the other hand, only a handful of attempts to produce and evaluate software of the rule-application type (for use in historical linguistics) can be found in the literature (Becker 1982; Brandon 1984; Durham and Rogers 1971; Frantz 1970). In general, such computer programs seem to have been abandoned after a certain amount of experimentation. Certainly the problem of articulating a set of algorithms and associated data sets which completely describe the regular sound changes evinced by a group of languages is a daunting task.

## 2.1. Lexicostatistical approaches

To the first class belong lexicostatistic models of language change. While the approach and its results have been well documented, not much off-the-shelf software is available to do the work of producing the statistics. The COMPASS module of the WORDSURV program described below belongs to this class (cf. Wimbish 1989). It measures degree of affiliation using a distance metric based on the degree of similarity between corresponding phonemes in different languages. Also to this class belong applications which measure genetic affiliation as a function of the number of shared words in a selected vocabulary set, such as Guy's COGNATE, which implements a somewhat more sophisticated algorithm which compares the frequencies of segments in words from pairs of languages. Any method which depends on counting 'shared words,' we note, assumes the existence and prior application of a means of determining which forms are cognate; such estimates of the relatedness of languages clearly are only as good as means used for determining cognacy. Only a very general criticism of these approaches is offered here: to the extent that the methods rely on the previous application of the traditional comparative method, they answer a question which is already answered. Also, to given a single numeric value as a measure of 'distance' between languages is to take a drastically oversimplified view of the nature of linguistic relationships, genetic or otherwise. Even if such distance measures correlate with the geographical distances between languages (as shown below in (22) for example), we are left with the problem of explaining how such a precise relationship can exist without explicitly including the variables of geography and culture.

(22) *Slavic pseudomap superimposed on a geographical map (Dyen 1992:75)*



The 'pseudomap' (also known as a configuration) is an arrangement of points (each designating a language) such that the physical distances between the points is proportional to the computed lexicostatistic distances. Dyen claims that this graphical technique, based on the method of 'multidimensional scaling' of (Black 1976), works in some cases where lexicostatistical dendrograms do not. It provides a 'nonhierarchical approach,' suitable for cases where 'wave or diffusion effects ... suggest ...

some sort of spatially oriented classification to supplement the hierarchical classification' (Dyen 1992:71). Of course, the points can be arranged in a space with any number of dimensions, and Dyen notes that the choice of number of dimensions is strongly influenced by the data; miraculously or implausibly depending on one's point of view, it turns out that for linguistic pseudomaps, 'two dimensions (n = 2) turns out to be appropriate in every case'.

One last criticism of these methods: the lexicon provides only one incomplete perspective on the degree of relatedness; even if a comparison of lexical items from a set of languages produced a map which exactly overlaid a geographic one, this would not be completely persuasive. A complete picture should take into account a broader spectrum of linguistic structure, including morphology, syntax, and semantics (cf. for example (Nichols 1992, Nichols 1994) which treat the issue of relatedness without reference to specific lexical items).

## 2.2. Combinatorial approaches

To the second class belong programs which model sound change as sets of rules applied to derive later forms from earlier forms. Examples of programs of this sort are VARBRUL (by Susan Pintzuk), based on a rule-processing subsystem called BETA (not critiqued here), used to analyze Old English; several programs used to analyze Romance languages: PHONO Hartman 1993, Iberochange (Eastlack 1977, both applied to Latin-to-Spanish data; three programs only slightly described, one for Indo-

European to Early Latin (Maniet 1980; Maniet 1983) one for Classical Latin to Old French (Burton-Hunter 1976); and one for Balto-Finnic (Remmel 1979).

## 2.3. Program, projects, and databases

The projects are described in chronological order of first publication; there are certainly other taxonomies for arranging them, but the historical view affords an automatic orientation and some perspective.

### 2.3.1. Kay's algorithm

Martin Kay (Kay 1964) outlines an approach that, given matched pairs of cognate words, works out the correspondences according to an evaluation metric based on the parsimony of the set of correspondences. Noting that 'the 'comparative method' ...is not the well-defined technique that the name suggests', Kay presents a 'formalization in terms of elementary propositional logic of one of the most crucial steps in the comparative method, that is, [the step in] which modern derivatives of prehistoric phonemes are recognized.' (Kay 1964:v) Kay summarizes the result of applying the comparative method as follows:

> For each associated set of forms which are judged to be related, an artificial form is constructed which fills the role of their common ancestor within the model. The letters in these reconstructed forms stand for the phonemes of the extinct language. The aim is to make the reconstructions in such a way that the history of each form does

> not have to be written separately. Instead, a history is written for each phoneme in the original language, and from these the history of the forms can be inferred. (Kay 1964)

Kay confines himself to the problem of finding correspondences between forms from pairs of modern languages, noting that his method can be extended to any number of languages with 'only trivial modifications' (Kay 1964:5).

Applying the terminology of reconstruction in a particular formal sense useful for his exegesis, Kay calls a *correspondence* 'an ordered pair of strings where the first member is taken from one extant languages and the second from another [and written] ... separated by a stroke, e.g., 'abcd/xyz" (Kay 1964:6). He considers all possible *decompositions* of each string into substrings, which are then associated with the corresponding substrings from the other member of the pair:

(23)

| | | | |
|---|---|---|---|
| (i) | a/x | bcd/yz | |
| (ii) | a/xy | bcd/z | |
| (iii) | ab/x | cd/yz | |
| (iv) | ab/xy | cd/z | |
| (v) | abc/x | d/yz | |
| (vi) | abc/xy | d/z | |
| (vii) | a/x | b/y | cd/z |
| (viii) | a/x | bc/y | d/z |
| (ix) | ab/x | c/y | d/z |

These decompositions, together with the initial correspondence, represent all the possible decompositions of 'abcd/xyz' into matching sets of substrings. Noting that in general most of these theoretically possible decompositions have 'no significance for reconstruction', Kay turns to the problem of discovering which of the decompositions do represent valid correspondences. He gives an illustrative pair of items from English and German:

*(24)*

that/dass

and notes that of the twenty possible decompositions, '...[o]nly one of these has a correspondence for each Indo-European phoneme' (Kay 1964:8).

*(25)*

th/d a/a t/ss

Using an algorithm for creating a gigantic logical disjunction of the possible correspondences over a set of data, Kay proceeds to show how the unfruitful decompositions can be eliminated, retaining only the smallest (i.e. most parsimonious) set of decompositions for which 'every correspondence represents a phoneme of the language being reconstructed' (Kay 1964:12). In a sense, the algorithm is presented as an exercise in predicate calculus.

Certain modifications of these algorithms, Kay notes, would make it possible to handle certain troublesome cases. Metathesis could be handled by starting with 'a list in which the forms in one language were paired with all permutations of their equivalents in the other'. And for cases of loss (where 'an ancient phoneme is without issue in some of the daughter languages') he proposes to insert a 'zero' at the beginning and end of each word and between each pair of phonemes.[46] Kay notes that this solution, while straightforward, 'results in a possibly unacceptable increase in the amount of computation to implement the theory.'

Kay concludes with a section on implementing the theory, noting that

> the possibility of applying the method mechanically is open, but barely so.[47] The author estimates that it would take some four or five hours of computer time to analyse a list of a hundred pairs of forms. Where the connection between a pair of languages is remote, this may well be worthwhile, for the amount of human labor that is put into such problems is often prodigious, and it is inefficiently spread out over a long period of time. (Kay 1964:181)

In the course of implementing his algorithm, Kay actually tried it on the following set of words, which, he says, is 'as small a set of data as the method can be applied to and produce a non-trivial result':

(26) *A small set of cognates for mechanical comparison*[48]

| English | German |
|---|---|
| on | an |
| nut | nuß |
| that | daß |
| bath | bad |

It would be altogether out of the question, Kay notes, to apply the method even to such a corpus as this without machine aid. While 'conceptually trivial,' the computation required 'rapidly becomes prohibitive as the number of variables increases. The belligerently incredulous are urged to try the example for themselves' (Kay 1964:18).

Computers have come a long way since 1964; however, the complexity of many of the computations associated with computer implementations of the comparative method has not changed. As noted below in the section on the Reconstruction Engine (§6), some of these problems (which are NP-hard[49]) could challenge even the limits of modern supercomputers.

### 2.3.2. Hewson's Proto-Algonkian experiment (the 'Electronic Neogrammarian'

The first description I have found of a computer program which uses correspondences and modern forms to create cognate sets is John Hewson's long-term experiment in reconstructing Algonkian (eventually dubbed *the electronic neogrammarian* (Hewson 1973)). John Hewson and others at the Memorial University of Newfoundland pioneered these 'proto-projection' techniques, which were later (and independently) used by the Reconstruction Engine (Hewson 1973; Hewson 1974; Hewson 1993). The strategy is in some ways quite transparent; as Hewson notes, he and his team also decided to 'follow the basic logic used by the linguist in the comparative method' (Hewson 1974:193). The results of this research have recently been published in the form of an etymological dictionary of Proto-Algonkian (Hewson 1993).

The program as first envisioned was to operate on 'consonant-only' transcriptions of polysyllabic morphemes from four Amerindian languages (as shown in (27) below). The program would take a modern form, 'project it backwards' into one or more proto-projections, then project these proto-projections forward into the next daughter language, deriving the expected regular reflexes. The lexicon for this language would be checked for these predicted reflexes; if found, the program would repeat the projection process, zig-zagging back and forth in time until all reflexes were found. For example, given Fox /poohkešamwa/ *he cuts it open*, the program would match the correct Cree form, as indicated in (27).

(27) *Potential Proto-Algonkian cognates (after Hewson 1974:193-94)*

| Language | C1 | C2 | C3 | C4 | Reflex | Gloss |
|---|---|---|---|---|---|---|
| Fox | p | hk | š | m | poohkešamwa | 'he cuts it open' |
| Cree | p | sk | s | m | pooskosam | 'he cuts it open' |
| Menomini | - | - | - | - | | |
| Ojibwa | p | šk | š | n | paškošaan | 'he cuts it down |
| Ojibwa | p | kk | š | n | pakkweešaan | 'he slices off a part' |

There were problems with this approach. In cases where no reflex could be found, (as shown in (27) above where no Menomini cognates for this form existed in the database) the process would grind to a halt though other cognate forms in other languages remained to be identified. Recognizing that 'the end result of such a programme would be almost nil' (Hewson 1973:266), the team developed another approach in which the program generated *all possible* proto-projections for the 3,403 modern forms. These 74,049 reconstructions were sorted together, and 'only those that showed identical proto-projections in another language' (some 1,305 items) were retained for further examination. At this point Hewson claimed that he and his colleagues were then able to quickly identify some 250 new cognate sets. (Hewson 1974:195). The vowels were added back into the forms, and from this a final dictionary file suitable 'as input to an automated typesetting machine' was created. A cognate set from this file, consisting of a reconstruction and two supporting forms, is reproduced in (28) below.

*(28) Proto-Algonkian cognate set (after Hewson 1973:273)*

| Language | | Form | Gloss | Protomorpheme |
|---|---|---|---|---|
| * | (ProtoAlq.) | PEQTAAXKWIHCINWA | BUMP | (*-AAXKW) |
| M | (Menomini) | P3QTAAHKIHSEN | HE BUMPS INTO A TREE OR SOLID... | |
| O | (Ojibwa) | PATTAKKOCCIN | BUMP/KNOCK AGAINST...[STHG] | |

## 2.3.3. Iberochange

One of the first rule-based applications for historical linguistics, Iberochange modeled the

> 'derivational etymological' technique, which works primarily for a given 'mother' language to a specific 'daughter' dialect ... It assumes that both the etyma and the corresponding 'modern' forms are known, e.g., Classical Latin LUPUM, VĪTAM, SACRĀTUM, modern Spanish lobo, vida, sagrado. (Eastlack 1977:82).

The creators of Iberochange made the following linguistic assumptions:

1) linguistic change has two components — systematic sound change and various types of non-systematic change such as analogical change, sporadic sound change, dialect leveling, etc.

2) systematic sound change can be described in terms of a fully explicit ordered set of rules ...;

3) at some point in time words differ from ... related forms at some earlier time only in having undergone the ... systematic set of sound changes specified by the set of rules mentioned in ... 2). (Eastlack 1977:81).

Since many symbols needed were not available in the 'computer alphabet,' the developers devised a system of symbolization by which it would be possible to transcribe input items without ambiguity. Specifically, syllable and morpheme boundaries were encoded so that the program could avail itself of them and complex segments such as /ts/ and

/dz/ were retranscribed with single letters (/C/ and /Z/. Examples are shown below:

(29)

| Orthography | 'Symbolization' |
|---|---|
| fortiam | #FOR TI AM# |
| caecum | #KAE KUM# |
| bracchium | #BRA:K KI UM# |

Forty two ordered rules (with subrules) describe the development of Latin to Spanish. Rule 3, which is used in the example in (30) below, says 'word-final M becomes N following a stressed vowel; elsewhere it is deleted.' Several step-by-step derivations are given, such as:

(30) *Derivation of Spanish* cabeça *from Latin* capitiam, *cf. CL* capitem[50]

| | |
|---|---|
| Starting form | #KA PI TI AM# |
| Rule 3 | #KA PI TI A# |
| Rule 7(a) | #KA P'I TI A# |
| Rule 7(d) | #KA P'I TIA# |
| Rule 7(g) | #KA P'E TIA# |
| Rule 15 | #KA P'ET TIA# |
| Rule 16(a) | #KA P'EC CA# |
| Rule 22 | #KA B'EC CA# |
| Rule 23 | #KA B'E CA# |
| Rule 27(a) | #KA B'E C;A# |

According to the author, the program provides 'rather conclusive evidence in support of the theory of language change propounded in

King's (1969) discussion of *Historical Linguistics and Generative Grammar.*' (Eastlack 1977:84); this says more about the generality and durability of the program and its algorithms than I could. Like the Reconstruction Engine, Iberochange is written in SNOBOL4, a powerful text-oriented programming language.

### 2.3.4. COMPASS

COMPASS Frantz 1970 applies both combinatoric and statistical techniques to the problem of comparative reconstruction. It is based on a relatively traditional model of reconstruction involving correspondences and cognate sets. Frantz is clearly aware of the problem surface similarities pose for reconstruction:

> The linguist who suspects a genetic relationship between two languages first compares lexical items of similar meaning in the two languages. Should he, in so doing, find a number of word pairs that are phonetically similar, he would be aware that this shows him little or nothing about the possibility of genetic relationship (Frantz 1970:353).

Noting that such similarities may be coincidental, Frantz goes on to elaborate how the COMPASS program attempts to 'weed out' the bulk of such accidental correspondences by statistical techniques. Correspondences which are frequently observed are more likely to be the result of genuine inheritance than those which are unique or of low frequency. Frantz uses 'hypothetical data' (reproduced in (31), his *Table I*, below) in supporting the explication of the operation of COMPASS.

*(31) Hypothetical data from 'Table I' (after Frantz 1970:354)*

| | Language | | |
|---|---|---|---|
| | *A* | *B* | *gloss(es)* |
| (1) | pakol | phogor | hand |
| (2) | feku | fögu | water/liquid |
| (3) | likel | riger | woman |
| (4) | pano | phono | tree/wood |
| (5) | kene | khene | mother |
| (6) | xipo | xöbo | uncle/elder |
| (7) | pepo | phöbo | stone |
| (8) | xana | xana | gourd |
| (9) | fapa | faba | good |
| (10) | kitu | khödu | tomorrow |
| (11) | kito | khotu | red |

Note that COMPASS requires the investigator to 'arrange the data for input so that the program compares the characters that he assumes would have to correspond if the members of each pair are cognate' (Frantz 1970:354):

*(32)*

```
p  a k o l
ph o g o r
p  a n o
ph o n o
```

The investigator must leave blank spaces in the appropriate places to account for lack of one-to-one correspondence in number of characters (the constituent size problem raising its ugly head again) (Frantz 1970:354).

The program can then compute the frequency of occurrence of each correspondence (only the correspondence p:p is illustrated below). The program lists the segmental correspondences, their frequencies, and an indication of the word pair which contains each correspondence.

(33) *Correspondence with count and list of supporting forms*

| | | | | | |
|---|---|---|---|---|---|
| p:p | tokens: 3 | | | | |
| | (1) | p | akol | phogor | hand |
| | (4) | p | ano | phono | tree/wood |
| | (7) | p | epo | phöbo | stone |

The program computes a *correspondence value*, essentially an estimate of how well attested it is, using the following formula:

$$\text{correspondence value} = \frac{\sum_{i=0}^{n}(v_i)}{n}$$

where: n = number of comparable segments
and: $v_i$ is the frequency of the correspondence at position i

This value is used to rank the correspondences with respect to each other. Frantz notes that his program is 'merely a tool; it is no substitute for the ingenuity and experience of the investigator. Rather it is a partial remedy for the limitations placed upon the investigator by the time-consuming nature of data-manipulation' (Frantz 1970:353)

COMPASS, according to Frantz, has been used for Proto-Algonkian, Cheyenne and Arapaho, 'with each other and with Bloomfield's PA'. He notes a difference in the comparison of two languages at a time as opposed to three, and that this poses a problem for his algorithms:

> The output that resulted from the simultaneous comparison of PA, Cheyenne, and Arapaho, while useful, is not nearly so useful as the output of pairs [i.e. pairwise comparisons]. In many sets only two of the words, but not the third, are cognate; the result is that there is a disproportionately large number of correspondences listed which are not regular. The three-language program would be more useful after work with the three pairwise combinations of the languages enables the investigator to remove sets which contain a member which is probably not cognate. (Frantz 1970:356)

The probative value of additional data and the problems associated with it have been discussed in §1.6.9.

### 2.3.5. Guy's COGNATE

Jacques B.M. Guy's own appellation for his program says a lot: he calls it 'an apparently wonderfully useless program implementing an algorithm'. COGNATE, according to Guy, implements a prototype algorithm for identifying related words across languages. Guy's purpose was to 'take a first step towards solving a far more interesting, and difficult, problem of automatic machine translation: given a bilingual text, find the rules for translating from either language into the other' (Guy 1992).

According to Guy, COGNATE operates as follows:

> given the same [sic] list of words in two different languages, COGNATE will determine which words are likely to be regularly derivable from each other, and which are not. The longer the list, or the more closely related the two languages are, the better the performance of COGNATE. For instance, suppose that you have typed into a file 200 words in English (one per line), and in another file the same 200 words, in the same order, in German (again one per line). English and German are fairly close languages. Given these two files, and no other information whatsoever, COGNATE will be able to tell for instance that English 'TWENTY' and German 'ZWANZIG' are almost certainly derivable from each other, and so are English 'HONEY' and German 'HONIG'; but it will also tell you that English 'HORSE' and German 'PFERD' are not so related. COGNATE will also tell you, when comparing 'TWENTY' with 'ZWANZIG', that English 'T' corresponds to German 'Z'. ' (Guy 1992)

Guy notes that because of the 'very nature' of the algorithm, the program is not sensitive to the actual scheme used for encoding the data: the program would work just as well if the letters were shifted using a simple-substitution code.

> For instance, if you have encoded the English data by shifting one letter forward (so that 'TWENTY' becomes 'UXFOUZ') and the German data by shifting one letter backward (so that 'ZWANZIG' becomes 'YVZMYHF'), COGNATE will still able to tell that 'UXFOUZ' and 'YVZMYHF' are related, and that 'IPSTF' ('HORSE') and 'OEDQC' ('PFERD') are not. (Guy 1992)[51]

COGNATE is supplied with three sample files of 200 words each, English, German, and Dutch. Like many of the historical linguistic applications described here, COGNATE has a checkered development history. It was first implemented around 1978 in Simula 67 on a DEC KL10. Then, as a 'self-inflicted challenge which I did not expect to win', Guy translated it into Turbo Pascal, to run on his Kaypro II. It is now available over the Internet at a number of FTP sites.

### 2.3.6. DOC: Chinese Dialect Dictionary on Computer

DOC is one of the earliest projects to attempt a comprehensive treatment of the lexicons of a group of related languages. DOC was developed 'for certain problems [in which] the linguist finds it necessary to organize large amounts of data, or to perform rather involved logical tasks — such as checking out a body of rules with intricate ordering relations' (Wang 1970:57). The original data design, implemented in punch cards, organized each dialect entry into a '22-byte word'. A sample of a few of the approximately 70,000 Middle Chinese and dialect records (in one of the original formats) is illustrated in (34) below. Note that as in WORDSURV, the data is pre-segmented according to a universal phonotactic description (in this case the Chinese syllable canon) which the program and data structures are built to handle. The one-segment-one-constituent restriction does not exist, though the (maximum) size of constituents is fixed within the data structure.

(34) *A Dialect record in DOC (cited from Fig. 7 in Wang 1970)*

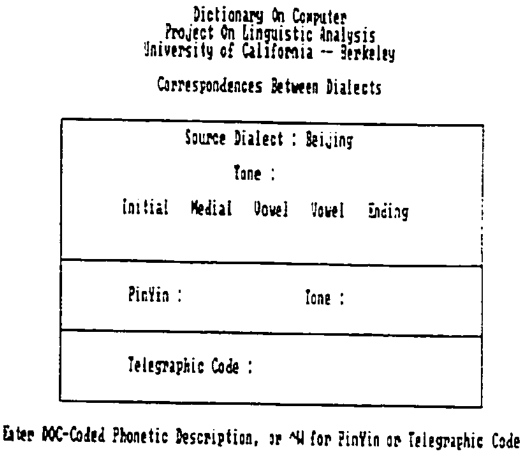
| Dialect | Tone | Initial | Medial | Nucleus | Ending |
|---|---|---|---|---|---|
| 0052 192- | 3 | L | H1 | WN | 7 |
| PEKING | 3 | L | U | A | N |
| XI-AN | 3 | L | U | A | Z |
| TAI-YUAN | 3 | L | U | A | Z |
| HAN-KOU | 3 | L | U | AE | Z |
| CHENG-DU | 3 | L |  | A | N |
| YANG-SHOU | 3 | L | U | O | Z |
| WEN-ZHO | 3B | L | U | O3 |  |
| CHANG-SHA | 3B | N |  | O | Z |

The data in (34) above is to be interpreted as follows: the line beginning '0052' (the so-called telegraphic code[52]) records the Middle Chinese form (with cross reference 192- to another source, the *Qiè-Yùn*); 'H1', 'WN', and '7' are coded representations of phonetic characters.

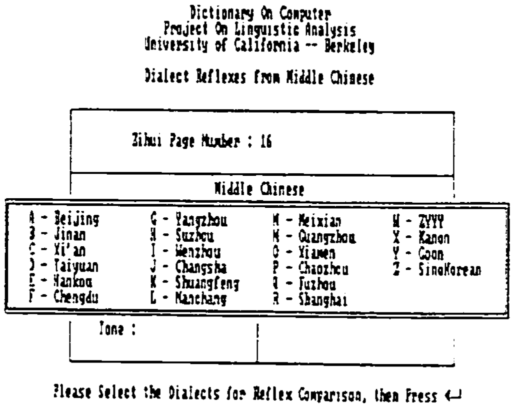
Following this are eight dialect records, giving reflexes of this word in modern Chinese dialects, also in a coded phonetic form.

At least four versions of this database and associated software were produced (described in Lyovin 1968; Streeter 1972; Cheng 1993:13). Originally processed as a punched-card file on a LINC-8, the program underwent several metamorphoses. An intelligent front-end was developed in Clipper (a microcomputer-based database management system) which allows the user to perform faceted queries (i.e. multiple keyterm searches) against the database and also contains the actual Chinese characters. (Yaruss 1990) So, as shown in the upper screen in (35), for example, the user could select a particular dialect (in this case Beijing), and search for words containing particular phonological constituents (Initial, Medial, etc.). The lower screen shows to retrieve the reflexes of a particular MC form (from page 16 of the Hànyǔ Fānyán Zìhuì Anonymous 1962 in this case).

*(35) User interface to the Clipper version of DOC (Yaruss 1990:215)*



```
              Dictionary On Computer
           Project On Linguistic Analysis
         University of California -- Berkeley

          Correspondences Between Dialects

            Source Dialect : Beijing

                    Tone :

      Initial   Medial   Vowel   Vowel   Ending

      PinYin :                Tone :

      Telegraphic Code :

Enter DOC-Coded Phonetic Description, or ^W for PinYin or Telegraphic Code
```



```
              Dictionary On Computer
           Project On Linguistic Analysis
         University of California -- Berkeley

        Dialect Reflexes from Middle Chinese

        Zihui Page Number : 16

                  Middle Chinese
 A - Beijing    G - Yangzhou    M - Meixian     W - ZYYY
 B - Jinan      H - Suzhou      N - Guangzhou   X - Kanon
 C - Xi'an      I - Wenzhou     O - Xiamen      Y - Goon
 D - Taiyuan    J - Changsha    P - Chaozhou    Z - SinoKorean
 E - Hankou     K - Shuangfeng  Q - Fuzhou
 F - Chengdu    L - Nanchang    R - Shanghai
        Tone :

 Please Select the Dialects for Reflex Comparison, then Press ↵
```

The database is also available as a text file (slightly over a megabyte) containing forms in 17 dialects for some 2,961 Chinese characters (Cheng 1993:12). DOC has no 'active' or rule-application component: it is a database of phonologically analyzed lexemes organized for effective retrieval.

### 2.3.7. Programs for CARP: Computer Aided (Historical Linguistic) Reconstruction in Phonology

Veatch (Veatch 1993) has drafted a Unix-based software suite which can extract correspondences from appropriately formatted data. Veatch's made-up examples clearly illustrate the bare-bones logic of the comparative process usually followed implicitly by linguists in creating correspondence sets. The starting point of Veatch's program is a list of potential cognates in the language group under study, aligned in columns. Veatch's discussion starts with the creation of a set of source data:[53]

> Use your favorite editor to create a list of cognates, in the following format. Put each language's cognates in a column; and each of the cognates of a single proto-form in a line. Separate the columns by a tab or spaces.

Veatch gives some contrived forms, illustrated below in (36).

(36) *Contrived cognates for use with CARP*

```
cognate1   COGNATE1   KAGNET_1
cognate2   COGNATE2   KAGNET_2
cognate3   COGNATE3   KAGNET_3
cognate4   COGNATE4   KAGNET_4
cognate5   COGNATE5   KAGNET_5
cognate6   COGNATE6   KAGNET_6"
```

The coding of the source data is rather strict, notes Veatch:

> Notice that where segments are deleted, i.e., in one language there is a segment present but in another it is missing, [...] an underscore is inserted in the location in the cognate which lost the segment. This is so that corresponding characters in the cognates actually correspond, and where a character corresponds to a deleted segment, the underscore gives it something explicit to correspond to. [...][W]hen you have edited the cognate file, all the cognates on one line have the same number of characters, so that corresponding characters actually correspond in the cognate. Thus, extra morphemes in one language must be deleted, to make the correspondences right.

Of course, in reality, considerable insight into the phonologies of the language studied, whether human or machine, is required to provide a correct alignment (this fact was noted above in §1.4, and will be discussed in more detail in §5.3).

Having created a set of putative cognates, Veatch's program next proceeds to match the 'corresponding' segments of each form, producing a list of 'merged proto-forms' [my term] as shown in (37) below. The resulting proto-forms have a rather 'strange-looking form'. For example, the above cognates would result in the following list:

*(37)*

```
cCK oOA gGG nNN aAE tTT eE_ 111
cCK oOA gGG nNN aAE tTT eE_ 222
cCK oOA gGG nNN aAE tTT eE_ 333
cCK oOA gGG nNN aAE tTT eE_ 444
cCK oOA gGG nNN aAE tTT eE_ 555
cCK oOA gGG nNN aAE tTT eE_ 666
```

Veatch explains:

> [...] Each phoneme [...][of each] proto-word is represented by a correspondence-set, so that an 8-segment proto-form will look like 8 queer words, each formed from the concatenation of corresponding segments in the cognates. So each line is really the proto-word which the cognates are related to. It's just that instead of writing it in letters, it is written in correspondence-sets.

Next another program picks out all the unique correspondence sets and makes a list of them (shown in column (1) of (38) below). The linguist adds a column (as shown in column (2) in (38) below) for the reconstructed ancestor to allow the computer to recode the 'queer words' of (37) into protoforms as shown in (39) below.

> we need to make a list of all the correspondence sets, and specify what characters to use to represent each correspondence set in the proto-allophone forms. So for example, we want a list like this:

(38) *The user supplies an ancestor for each 'proto-allophone'*

| (1)<br>Proto-allophone | (2)<br>Reconstructed ancestor (added by user) |
|---|---|
| cCK | k |
| oOA | o |
| gGG | g |
| nNN | n |
| aAE | a |
| tTT | t |
| eE_ | e |
| 111 | 1 |
| 222 | 2 |
| ... | ... |
| etc. | |

A final program does the actual retranscription:

(39) *Retranscription of the 'queer words' according to the 'proto-allophones'*

*kognate1
*kognate2
*kognate3
*kognate4
*kognate5
*kognate6

Veatch goes on to point out that the allophones file (exemplified in (38) above) can be further refined to identify 'proto-allophones'.

(40)

> Look at the distribution of the proto-allophones, and determine which ones may be collapsed into one category, using complementary distribution and phonetic similarity as criteria. When you find proto-allophones that may be collapsed, then go

> back and edit the allophones file, which specifies the proto-sounds for each correspondence set, and specify the same symbol for each collapsed pair of allophones. Then you can redo codecorr and ccon,[54] if necessary, to see if any further reconstruction is possible, and re-edit the allophone list, and so on, until nothing else can be collapsed.

Veatch does not say whether this program has ever been used on real data, and he provides no such examples. A similar procedure of aligning input data according to semantic and phonological criteria is required by the WORDSURV and COMPASS program described elsewhere in this chapter.

### 2.3.8. PHONO: a program for testing models of sound change

PHONO (Hartman 1981, Hartman 1993) is an MS-DOS program which applies ordered sets of phonological rules to input forms for the purpose of 'developing and testing models of regular historical sound change.' (Hartman 1994) The rules are expressed by the user in a notation composed of if- and then- clauses that refer to feature values and locations in the word. The feature specification is communicated to the program via an *Alphabet*. The Alphabet is a list of symbols and associated feature set expressed as a matrix of characters. PHONO converts input strings (words in the ancestor language) into their equivalent feature matrices using this table of alphabetic characters and feature values. The program then manipulates the feature matrices according to the rules, converting the matrices back into strings for output. Hartman has developed a detailed set of rules which derive Spanish from Proto-Romance. Besides allowing the expression of diachronic rules in terms of features, facilities are included to handle metathesis. PHONO has both an interactive mode for trying out individual forms from the keyboard as well as a batch mode, in which lists of forms are read from an external file. PHONO is available over the Internet via FTP.

### 2.3.9. WORDSURV

The Summer Institute of Linguistics (SIL), a prodigious developer of software for the translating and field linguist which is headquartered in Dallas, Texas, provides a variety of integrated tools for linguistic analysis. One of these tools, the COMPASS module of WORDSURV, allows linguists to compare and analyze word lists from different languages and to perform phonostatistic analysis. To do so, the linguist first enters 'survey data' into the program; reflexes are arranged together by gloss, as illustrated in the reproduction in (41).

(41) *'Properly aligned word forms' in WORDSURV (Wimbish 1989:43)*

| (1) | (2) | (3) | (4) |
|---|---|---|---|
| Group | Reflex | [metathesis] | Language Abbreviation |
| 0 | -- no entry -- R | | |
| A | faDer | | E |
| A | fater | | G |
| A | padre | >4 | S |
| B | ara | | iT |
| C | bapa -- | | MPB |
| C | bapak-- | | I |
| C | bapa da | | h |
| D | tataN | | wm |
| D | tatay | | ab |

In addition to the *a priori* semantic grouping of reflexes by gloss, the linguist must also re-transcribe the data in such a way that each constituent of a reflex is a single character, that is, 'no digraphs are allowed. Single unique characters must be used to represent what might normally be represented by digraphs ... e.g. N for ng' (Wimbish 1989:43). The program also requires that part of the diachronic analysis be carried out before entering the data into the computer in order to incorporate that analysis into the data. For example, when the linguist hypothesizes that 'a process of sound change has caused a phone to be lost (or inserted), a space must be inserted to hold its place in the forms in which it has been deleted (or not been inserted)' (Wimbish 1989:43). That is, the zero constituent must be represented *in the data itself.* The program also contains a 'provision for metathesis. ...Enter the symbols >n (where n is a one or two digit number) after a word to inform WORDSURV that metathesis has occurred with the nth character and the one to its right' (Wimbish 1989:43). An example of this may be seen in column 3 of (41). This provision is clearly intended to allow linguists to 'correct' (my term) for sporadic changes such as metathesis, bringing the proper elements into juxtaposition for comparison.

To represent tone, the author notes that 'there are at least two solutions. The first is to use a number for each tone (for example 1ma3na). The second solution is to use one of the vowel characters with an accent. ... The two methods will produce different results' when the analysis is performed (Wimbish 1989:44). While the last statement may surprise some strict empiricists (after all, the same data should give the same results under an identical analysis), it should come as no surprise to linguists who recognize that the selection of unit size, the type of constituency, and other problems of representation may have a dramatic effect on conclusions.[55] Two requirements of this program 1) that forms be grouped *a priori* by gloss and 2) that segments be aligned according to their supposed correspondences are fraught with methodological difficulty: these requirements force the linguist to decide *a priori* which forms might be related semantically and also to supply a singular phonological analysis (both synchronic and diachronic. The phonological inventory is thus limited to segments. In passing, the lexicostatistics which are computed are based on the 'Manhattan distance' (in a universal feature matrix) between corresponding phonemes from different languages as a measure of their affiliation. The validity of this measure for establishing genetic affiliation is suspect: corresponding phonemes may be quite different in terms of their phonological features without altering the strength of the correspondence or the closeness of the genetic affiliation. Also, the metrics of features spaces are notoriously hard to quantify, so any distance measures are themselves likely to be unreliable.

### 2.3.10. MARIAMA

MARIAMA (Nicolai 1993), is a *manager of [linguistic] hypotheses*. It was conceived as computational aid in comparative research and lexicography. It applies the 'classical' database functions (searching, sorting, and import/export) to linguistic data; but it also incorporates a number of functions particular to 'sa finalité propre.' This are described briefly below. MARIAMA was developed using Nilo-Saharan languages. It is being distributed now and applied to other language groups. The description below is based on draft documentation Nicolai 1991 and conference presentations Nicolai 1993 on the program, and the reader is advised that the program is still being developed; its description and evaluation here is tentative.

From the point of view of MARIAMA, a *hypothesis* is a *relationship established between several records* of the same data type. This relationship is defined via a system for marking pertinent data (*un système de pertinence particulier*). In the most basic sense, the linguist working on a comparative project — ultimately to establish genetic affiliations, to perform phonological, dialectological, morphological, and semantic analyses — proposes one or more hypotheses concerning the data under comparison. So, for example, given the three forms in (42) below:

*(42) Connections established between forms by MARIAMA*

Exemple : le rapprochement particulier effectué entre les trois formes suivantes :

| **'Dialect'**[56] | **Entry** | **Meaning** |
|---|---|---|
| kaado | debe | damer, tasser en frappant |
| touareg | atAbbi | taper avec la paume de la main |
| bozo | tEbE | piétiner |

The program permits the user to:

1. establish a hypothesis about the relationship exhibited by the forms (by linking them together according to a key)

and

2. define the nature of the relationship by giving the set of linked forms a name. The name used can define a semantic relation (i.e. some sort of cover gloss for the forms) or a phonological relation (i.e. a tentative reconstruction).

The hierarchical structure of MARIAMA provides for three 'levels' of representation of the data; the second level provides five 'sublevels' to record relational hypotheses about the relationship between words. These levels provide a means for unifying data from different sources into a consistent description and can become more abstract as the levels go up.

Level 0 is the 'reference level' *('niveau de référence');* contains the raw data. This level is provided to allow the research to record the source data just as it is found in the source.

Level 1 is the 'work level': this is the level at which the user can 'homogenize' the data for his or her own purposes, retranscribing the data and normalizing glosses. In (43)(a) and (b) below Nicolai supplies examples of various transcribed forms which can be homogenized into a single transcription (not shown) using the program (the homogenization, as noted above, is implement by the user, not by the program). (43)(a) illustrates the equation of forms from two source (abbreviated RN and OY) which differ in their transcription. (43)(b) illustrates a similar equation with forms from three sources. In (43)(c) he gives four glosses from different sources which can be unified into a single set. The process of homogenization is carried out 'by hand:' that is, the linguist marks each term in the equations himself.

*(43) Examples of 'Homogenization':*

- of phonetic transcriptions:

(a) RN : saayi = OY : sa:ji
(b) FD : touri = RN : tuuri = OY : tu:ri ; etc.

- of semantic distinctions:

(c) RN : 'palmier-doum' = XX : 'espèce d'arbre' = YY : 'arbre dont le fruit est utilisé pour [...] et les feuilles[...]' = A.P. : 'Balanites thebaica' ; etc.

Level 2 is the 'research level' ('niveau des hypothèses'). It pertains to the various types of linguistic hypotheses (discussed above) which may be applied to the data. Level 2 is the point at which data is grouped into sets for further research; to this end, five 'sub-levels' are provided, amounting to five *plans of analysis* (described below).

Level 3 is an additional hierarchical level. Basically it provides another level of data grouping above level 2 for consolidation of hypotheses stated at lower levels used, for example, for bringing together 'etyma' which are supported by several 'roots' (*radicaux*, illustrated in (44) below) already reconstructed. It provides two optional analysis plans.

As noted above (in discussing Level 2, the research level), MARIAMA provides five *plans of analysis* over the data set. These five plans correspond to five different sets of hypotheses, each with its own particular attributes:

- Le niveau 2 des 'références'
- Le niveau 2 des 'reconstructions'
- Le niveau 2 des 'clefs'
- Le niveau 2 des 'classificateurs'
- Le niveau 2 de 'travail'

- *Le niveau 2 des 'références'* : the level of intuitions and hypothesis advanced without necessarily having a solid justification. Preliminary research level.

- *Le niveau 2 des 'reconstructions'* : Represented here are sets of data supported by linguistic reconstruction in the 'strict sense'. Alternative forms may be specified along with justifications concerning phonetic regularities in the form of features, rules, and rule changes.

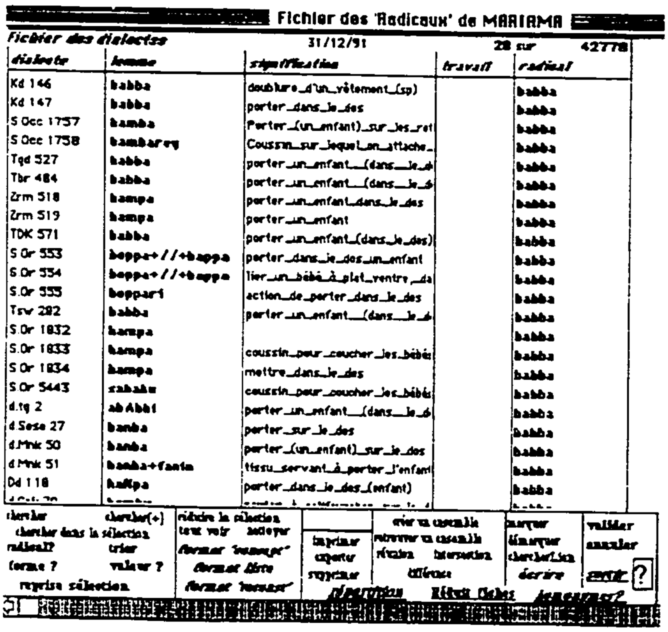
- *Le niveau 2 des 'clefs'* : this plan provides a means to 'index' the data according to a classificatory grid. The units of the grid may be semantic, ethnographic, linguistic, or otherwise. A multivalued classification is permitted so that the same form may receive several classifications; consequently, this plan permits working with forms in terms of 'features' (*traits*), or in terms of 'matrices.' The use of the word *clef* (key) is probably intended to evoke the notion of a limited set of possible values used in classifying the data.

- *Le niveau 2 des 'classificateurs'* : this plan also permits an 'indexation' of the forms, but at organized according to 'classification types.' I am not sure how to interpret the function of this plan.

- *Le niveau 2 de 'travail'* : this plan allows the specification of tentative hypotheses based on alternative organization of data created at other levels. It is for experimentation and intermediate groupings.

MARIAMA is a complicated program with a large number of features, many as yet only partially documented. The flavor of the program, which is written in 4D for the Macintosh, can be glimpsed in (44) below, which shows a list of forms from different languages which have the same root (*babba*), meaning something like 'carry on the back.'

(44) *Display of words for *babba 'carry on the back' in MARIAMA*



Fichier des 'Radicaux' de MARIAMA

Fichier des dialectes 31/12/91 28 sur 42778

| dialecte | lemme | signification | travail | radical |
|---|---|---|---|---|
| Kd 146 | babba | doublure_d'un_vêtement_(sp) | | babba |
| Kd 147 | babba | porter_dans_le_dos | | babba |
| S Occ 1757 | bamba | Porter_(un_enfant)_sur_les_rei | | babba |
| S Occ 1758 | bambareŋ | Coussin_sur_lequel_on_attache_ | | babba |
| Tgd 527 | babba | porter_un_enfant__(dans__le_d | | babba |
| Tbr 484 | babba | porter_un_enfant__(dans__le_d | | babba |
| Zrm 518 | bampa | porter_un_enfant_dans_le_dos | | babba |
| Zrm 519 | bampa | porter_un_enfant | | babba |
| TDK 571 | babba | porter_un_enfant_(dans_le_dos) | | babba |
| S Or 553 | boppa+//+bappa | porter_dans_le_dos_un_enfant | | babba |
| S Or 554 | boppa+//+bappa | lier_un_bébé_à_plat_ventre_da | | babba |
| S.Or 555 | boppari | action_de_porter_dans_le_dos | | babba |
| Tsw 282 | babba | porter_un_enfant__(dans__le_d | | babba |
| S.Or 1832 | bampa | | | babba |
| S.Or 1833 | bampa | coussin_pour_coucher_les_bébés | | babba |
| S Or 1834 | bampa | mettre_dans_le_dos | | babba |
| S.Or 5443 | sabahu | coussin_pour_coucher_les_bébés | | babba |
| d.tg 2 | ab Abbi | porter_un_enfant__(dans__le_d | | babba |
| d.Sese 27 | bamba | porter_sur_le_dos | | babba |
| d.Mnk 50 | bamba | porter_(un_enfant)_sur_le_dos | | babba |
| d Mnk 51 | bamba+fanta | tissu_servant_à_porter_l'enfant | | babba |
| Dd 118 | baKpa | porter_dans_le_dos_(enfant) | | babba |

## 2.3.11. CUSHLEX

CUSHLEX, for 'Cushitic Lexicon', is being developed by Gene Gragg at the Oriental Institute of the University of Chicago. As of then end of 1994, CUSHLEX contains some Cushitic lexical items (20,000 words in some 70 Cushitic languages, plus 5,000 or so from some 230 other Afroasiatic), an index of the cognate sets that have been proposed for them (including extra-Cushitic Afroasiatic where they exist), and a set of tools for registering, maintaining, and cross-referencing correspondence sets and rules. It is an attempt to provide a 'useful tool for historical linguistic research developed with off-the-shelf DBMS software, and on platforms readily available to the working historical linguist.' (Gragg 1994) Development went through several phases: a first attempt in dBASE IV was unacceptably slow, and suffered from an archaic interface. The current implementation is in FoxPro for Windows, a powerful multi-platform DBMS. CUSHLEX is a sophisticated tool providing a means of creating cognate sets (by hand) and making correspondences sets linked to sets and reconstructions (again by hand). Its interface and functionality will likely make it a model for future development of cross-linguistic etymological databases; combined with a phonological component like the Reconstruction Engine (described below) it could have a broad appeal to working comparativists.

### 2.3.12. The Reconstruction Engine (RE)

This program is treated in some detail in §6 and so only mentioned here. The programs models the classical view of the comparative method for establishing genetic affiliation among a group of languages via sound correspondences and cognate sets (as described in §1). The program is a research tool designed to aid the linguist in evaluating specific hypotheses, by calculating the consequences of a set of postulated sound changes (proposed by the linguist) on complete lexicons of several languages. It divides the lexicons into a phonologically regular part, and a part which deviates from the sound laws. The Reconstruction Engine is bi-directional: given words in modern languages, it can propose cognate sets (with reconstructions); given reconstructions, it can project the modern forms which would result from regular changes. The Reconstruction Engine operates either interactively, allowing word-by-word evaluation of hypothesized sound changes and semantic shifts, or in a 'batch' mode, processing entire multilingual lexicons *en masse.* (Lowe and Mazaudon 1994)

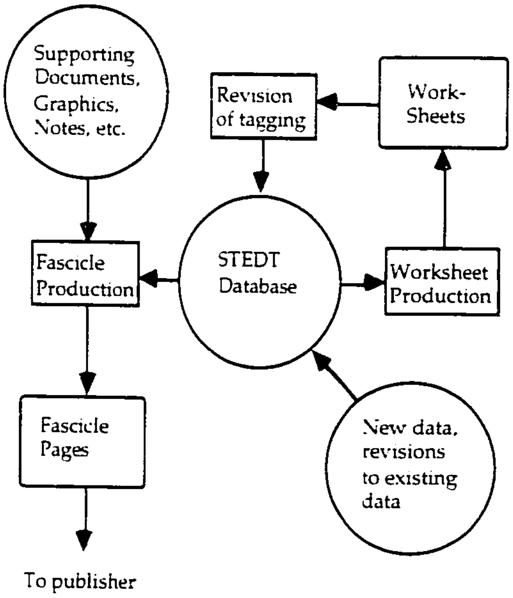
### 2.3.13. Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT)

The STEDT project, begun in 1987, is creating an etymological dictionary-thesaurus of Proto-Sino-Tibetan, the reconstructed ancestor of many of the languages spoken in China, NE India, and peninsular SE Asia. The goal of the project is to publish a multi-volume work, each on a different semantic area.

To support the creation of the dictionary thesaurus, a sizable lexicographic database has been created from a variety of disparate sources including monolingual and bilingual dictionaries, word lists, published and unpublished articles and manuscripts, and other linguistic databases and files. As of April 1995, the database contains approximately 232,000 language forms from 250 Sino-Tibetan languages and dialects. A portion of these have been grouped into some 2,000 cognate sets for eventual publication in the dictionary-thesaurus. Problems concerning transcription and representation, semantic relationships, etymologization, and indexing are discussed elsewhere in this dissertation.

The production of the printed thesaurus requires the integration of a wide variety of documents, including graphics, a large variety of marginalia and notes, and lexicographic and etymological information. The essence of the process is illustrated in (45) below.

(45) *Components of the STEDT document production system*



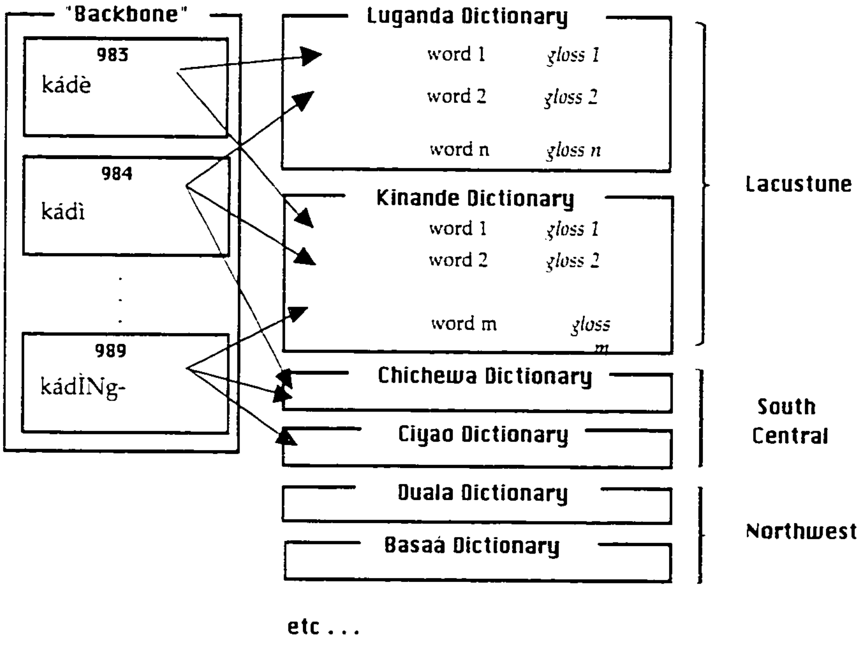
The database is an essential sideline to the production of the published work, though it may turn out to be one of the most useful results when it is made available to the research community. (Lowe 1993, Matisoff 1991)

### 2.3.14. The Comparative Bantu Online Dictionary (CBOLD)

The CBOLD project is an international effort to establish a computerized database consisting of approximately 4,000+ Proto-Bantu roots as well as reflexes of these and additional regional roots for an initial 50 or so of the 500+ daughter languages. The major concrete goals of the project are 1) to set up a unified database for lexical research in Bantu languages, and to input data from as many languages as possible into the database; and 2) to establish a means for sharing the data as widely as possible among scholars around the world. As April 1995, CBOLD had received or converted data from over 30 sources (institutions and contributing scholars), representing approximately 230,000 words in 118 Bantu languages (including data for reconstructed languages).

The CBOLD database consists of a number of parallel bilingual dictionaries with an overlaid semantic and phonological analysis. Functioning as the 'backbone' of the database are existing reconstructions of Bantu, notably those of Guthrie and Meussen (Guthrie 1967; Meussen 1967). The lemmata in these dictionaries are 'aligned' etymologically as illustrated in (46).

*(46) File design for CBOLD (N.B. only six of the set of 'core' dictionaries are depicted)*



To the extent possible, dictionaries and other lexicographic data is acquired through scanning and OCR (optical character recognition). Software to parse these texts into 'fields' which can be loaded into the database is being developed. The data acquisition and preparation process is discussed in §3. Tools for searching the database on the basis of morphological and phonological structure are planned. The database and the tools for using it are being developed in an environment which will allow researchers access on most of the popular computing platforms, at least Apple Macintosh and IBM compatibles. Several types of tools will be developed:

Besides producing an up-to-date, revised, and expanded etymological dictionary of Proto-Bantu, other types of documents which might be useful are synonym lists, phonological inventories with supporting forms, thesauruses, and multi-lingual dictionaries. In general, these tools will operate on the database as a whole and produce sizable documents. Thus, the users of CBOLD will be able to support a certain amount of 'demand publishing', providing interim and final versions of their analyses in a timely fashion.

Queries of high complexity need to be answered by the database. The queries may refer to specific segments or broad classes of segments. They may refer to adjacency or boundary conditions defined by morphological or phonological criteria. Some approaches to providing these types of facilities are discussed in §5.

### 2.4. A critical evaluation

There seems to be in general an inverse relationship in the programs and projects above between the amount of effort devoted to gathering the data and the amount of effort devoted to analyzing it. Indeed, the goal of some of the projects is primarily to gather the data (with some idea of how it will eventually be used) while other projects are applied to small data sets with the idea that someday they will be applied to larger sets.

Several of the programs are explicitly 'corpus-based:' (CBOLD, CUSHLEX, DOC, the Electronic Neogrammarian, MARIAMA, the Reconstruction Engine, STEDT, and WORDSURV) and provide some sort of functions for diachronic analysis. Others provide only apparatus for analyzing the data, and sometimes include some sample data, either real (as in COGNATE, Iberochange, PHONO, and Kay's test program) or made-up (COMPASS, CARP). They differ substantially in the portion and proportion of the lexicon treated. Veatch has not said if his programs have actually been used on real data; Kay's experiment used only eight forms in two languages, and he said that that was a lot given his approach.[57] Others are large database projects incorporating tens and sometimes hundreds of thousands of forms in a myriad languages.

Some programs require that the data be initially arranged in some fashion, either by semantic relationship (WORDSURV) or cognacy (DOC).

Some programs require prior segmentation according to some abstract constituent structure prior to computer analysis (COMPASS, DOC, the Electronic Neogrammarian, Iberochange, and WORDSURV). Some provide no means to do this at all, representing such segmentation implicitly (MARIAMA, CUSHLEX). Others, like PHONO and the Reconstruction Engine, can perform phonological analysis on constituents of a variety of sizes, including features, segments, or larger constituents. I should point out that to be general, the phonology and phonotactics should be parameters to the extent possible, so one can test different structural and phonological hypotheses without having to recode the data.

WORDSURV counts correspondences in order to arrive at a statistical measure of their 'strength.' Hewson's program finesse their actual representation as a distinct computer object; indeed he notes that 'the method does not use the correspondences in order to predict possible cognates, but the reflexes.' Clearly however, the 'proto projections' used to bring cognates together rely on the notion of correspondence. Although the columns of a dialect record in DOC might reflect some kind of genuine correspondence, it is not clear what status the authors of the program believe them to have, inasmuch as DOC is meant to test the hypothesis of lexical diffusion. Recently however, the data in DOC has been used to test the notion of regular correspondence among Chinese dialects.

Of these programs only the Reconstruction Engine actually generates cognate sets with complete reconstructions on the basis of correspondences and semantic information. The Hewson program produced pieces of cognates: consonants in the languages treated show 'greater regularity and simplicity' (Hewson 1974:192) than vowels; vowels were consequently ignored. Software being developed on the CBOLD and STEDT projects will take the some of the ideas incorporated in these programs a step further, implementing the cognate sets and reconstruction generating algorithms developed for the Reconstruction Engine, the database approaches of MARIAMA and CUSHLEX, as well other features borrowed from these programs.

# 3. DATA PREPARATION AND CONVERSION

## 3.1. Number-crunchers versus letter-crunchers

Given the amount of discussion about the positive contribution of computers in many (scientific) domains, it is remarkable how little use linguists have made of them for anything beyond word processing and email.

Recently, with the increasing popularity of the 'corpus-based approach', linguists have begun to use computational techniques in their research. Techniques such as simulation, statistical analysis, and database searching have all found a niche in historical linguistics. It is this last area that I will focus on here. Linguistic databases (such as text corpora and machine-readable dictionaries) have become increasingly important parts of computational linguistic research which previously focussed on building tools to parse individual sentences and solve known problems in syntax.

Many barriers to the general use of database tools in linguistics have yet to be overcome, and this is especially true in cross-linguistic and diachronic studies. As a pre-eminent tool for research in fields such as physics and geology, computers crunch *numbers* efficiently. *Letters*, on the other hand, still pose a problem. In part this is due to the fact that computers, even when processing text data, are in fact processing underlying numerical representations. The relationship between these numerical representations and the linguistic phenomena they symbolize is still unclear from a computational point of view. Both linguistic and computational difficulties are involved, as will be shown in subsequent discussion. Another problem is size. Some linguistic research requires that a large amount of data be processed. Corpora for syntactic and speech recognition purposes may contain millions of words, taxing the resources of even the largest computers. As I will show below, the resources required for many historical linguistic problems are also likely to require large databases and powerful computers. I will first address the problems of creating such corpora below, starting from the point at which the data is gathered and proceeding through several typical levels of analysis; later chapters will deal with techniques for using them.

### 3.2. Form, meaning, and provenance

The comparative method is based on equations of *form* and *meaning* involving lexical items from different languages. The recurrent intersection of these two criteria across languages constitutes the basis for reconstruction.

Besides these two data elements (form and meaning), the comparativist working with data from numerous languages needs to retain the *provenance* of each of the lexical items used in his research—a term which has at least two senses here.

First, I use it in the sense of *source* from which the data is drawn. Reconstruction requires reference to data from several languages, and almost inevitably the researcher must rely on data gathered by others, usually in the form of written works. The good use of such 'second-hand' data depends on the accurate interpretation of the source documents. Any description, no matter how thorough, has limitations, some of which have long been the subject of debate in linguistics. For instance, though for some purposes a narrow phonetic transcription may be useful, for comparative work a well-thought out and highly phonologized transcription is usually easier to work with.

Second, provenance has a sociolinguistic and ethnolinguistic interpretation, which I will call *language variety*: each lexical item represents a speech form drawn from a particular speaker or community of speakers; it is located at a particular moment in history and place in the world. Sometimes it is sufficient to know that a word is *German*; at other times it is necessary to know that it is from a particular *dialect* of German. Further, the same language variety may go under different names depending on who is referring to it, as I will discuss in some detail below. These important facts should be borne in mind, at least implicitly, during the process of reconstruction.

The bipartite notion of provenance given here is significant: identical data from different sources may differ just as much in representation as different data sets from the same source (examining (88) on page 212 below may make this clearer).

I identify therefore four basic, indeed essential, dimensions of the data which must have a systematic representation for comparative work:

- The *form* of the lexical items. The usual starting point for the representation of form is a transcription of the pronunciation. Phonemic analysis lays bare the phonological structure, which provides a basis for further work. For purposes of reconstruction, however, several more layers of linguistic complexity must normally be undone to render the data suitable for comparison. Besides recovering phonological structure from surface phonetic form, morphological processes must be 'reversed' to retrieve the root morphemes which are the normal basis for comparison; where relevant, internal reconstruction must also be carried out. In some cases, only written records are available, which provide an imperfect mirror for the spoken language.

- The *meaning* of the lexical items. In everyday life, the semantic burden borne by lexical items varies from use to use, and it can be helpful if the full range of meanings is available to the comparativist. In practical comparative work, however, the meaning of lexical items is normally represented only by a short gloss. Sometimes very precise meaning nuances are required to understand the semantic link between cognate forms in different languages; in other cases a broad categorization is required to make the link. An example of the need for a precise meaning would be understanding how a word etymologically meaning 'sit' can mean 'goodbye': the admonition 'remain sitting!' is conventionalized;[58] an example of the need of a broad meaning would be the Proto-Tibeto-Burman reconstruction *r-kliŋ (STC 126) meaning both 'brain' and 'marrow'.

- The *source* of the lexical items, that is, a reference to the works from which data are drawn. From a linguistic point of view, this dimension is relatively unproblematic, since few databases draw data from more than a few hundred sources. As noted below, however, there are a few twists in the road which need to be considered.

- The *language* of the lexical items, by which I mean a naming of the set of lexical items according to its areal, temporal, and genetic context. We need to be able to refer to a set of data using some referent which reminds us of important facts about where the words were spoken and by whom. 'Tibetan' is so named because it is from a region we in English call Tibet. Tibetans, of course, call their language something else, and this name depends on which group of Tibetans is being asked.[59] The referent used should be as evocative as possible, but should also consider the usage and custom of the speakers and their neighbors. This is not, I will show, always a straightforward task for a human or a machine.

### 3.2.1. Representing form

#### 3.2.1.1. Symbols and segmentation

Before there can be any hope of having the computer do something useful, it must be given some representation of the linguistic forms. Computers are notoriously poor at handling variation of any kind, and stories abound of satellites misplaced or lost for want of a comma in a piece of controlling software code. I must therefore digress somewhat to discuss the implications of certain kinds of linguistic representations. I will first point out features of writing systems which present challenges to computer processing, and then treat more typically linguistic issues which must be confronted for cross-linguistic comparison. The initial examples I will use deal mainly with the type of difficulties involved in transcribing Asian and particularly Tibeto-Burman languages, but it will be clear that the problems are not restricted to this family.

First, at a basic level there is a need to segment the continuous and one-dimensional stream of symbols used to represent speech into those smaller repeating units which are useful for analysis. Many languages do not divide the continuous stream of text into smaller units, requiring the reader to do so on the basis of their knowledge of the language. Burmese, for example, usually only breaks sequences of symbols at phrasal boundaries (Okell 1971). Breaks in Sanskrit texts are determined by the rules of saṃdhi and the smaller units thus created are not 'words.' The larger units of text do not contain repeating units (except for formulas and standard collocations which repeat relatively infrequently); they are, however, hierarchically composed of smaller units such as morphemes, which in turn are divided into smaller repeating entities, which may or may not be simply segments.[60] One such set of repeating units finds its way into dictionaries in the form of *headwords*. For purposes of the preparation of lexicographic databases, I will assume that some linguistic analysis has already extracted or identified the significant units and perhaps even clothed them in a suitable transcription. In short, I will assume that we have a dictionary, wordlist, or other lexicographic source whose contents we wish to make usable for further etymological research. We wish then, to give a machine-usable interpretation to the *words* of the source.

A 'word' is internally represented by a computer simply as a string of bytes. A 'word' in a text, for example, is conventionally defined (by computer specialists especially) as a string of characters found between some delimiting character, such as a blank. In Unix, for example, the typical delimiter for a word in text data is the occurrence of one or more blanks, tabs or 'new line', called 'white space'.[61] This is clear enough; however, the relationship between the word we perceive as a displayed image (called a glyph) and its 'underlying' or internal representation in the computer, called *rendering*, can be quite complex.

I will call the graphic image generated from a single underlying (i.e. machine-internal) unit a *glyph*. This image may be shown on a display device (such as a computer screen) or on a printer, and it is often the case that the same representations will not work for both. A complete symbol for some phonetic element (usually a segment) created using one or more glyphs will be called a *grapheme*. Some examples of graphemes are:

(47) *Examples of graphemes used in transcribing TB languages*

| | |
|---|---|
| g | grapheme composed of a single glyph |
| g̈ | digraph composed of two glyphs: ̈ . and g |
| š | digraph composed of two glyphs: ˇand s or ˇand s [62] |
| ṳ̈ | trigraph composed of three glyphs: ̈ . u. and ̣.[63] |

Even assuming that the division of the stream of symbols into 'word-sized' units has been accomplished, problems of constituent size and order persist. In Devanagari scripts, for example, the order of writing (i.e. surface form) and the order of phonetic transcription (underlying form) differ, requiring the rendering process to break down and reorder glyphs.

(48)

| 'Phonetic Order' | | | | | 'Writing Order' |
|---|---|---|---|---|---|
| ह् | ि | न् | द् | ी | हिन्दी |
| h(a) | i | n | d(a) | ī | i h n d i |

(Apple Computer 1991:section 14-33)

Devanagari also has diacritical glyphs which must be rendered above or below (or above and below simultaneously) other glyphs. Burmese is written using the Mon script, a descendent of the Nagari script, and illustrates even more rendering problems which not only encompass 'order' in the sense of left and right (or before and after), but of up, down, and 'surround' as well.

(49) *Written Burmese (WB) uses graphemes which require that glyphs be rendered above, below, and around other glyphs*[64]

ကြွတ် WB krwat 'leech'

(Consortium 1992)

But most of these conceptual and technical problems have been solved, at least in principle, with the advent of standards for representations of scripts (such as the Unicode Standard). The lag between solution and practice, however, means that most of us are still dealing with computer whose character encodings are very limited, forcing us to write phonetic characters using sequences of the existing ASCII codes as a stopgap measure. (Neuhaus 1986:160) The TeX system of Knuth is a good example of the kinds of conventions required to communicate complex rendering commands to the computer in cases where only the usual ASCII characters are available. (Knuth 1979). I will not attempt to deal with these kinds of problems, which will disappear in the next few years, and devote my attention here to the kind of rendering problems which are intrinsic to most phonological representations and to phonology itself and therefore are unlikely to go out of date.

In spite of the fact that several glyphs may be used to create one displayed image, no phonotactic relationship necessarily exists between them. Most people feel that '*e* plus *accent aigu*' is two glyphs combined and that the result is a vowel, but in a computer (internally) the symbol *é* on the screen may be represented by one glyph or two (with the order dependent on the implementation). This diacritic, when it appears over a consonant, as in the letter ś, must be treated as consonantal. One (or one's software) must be sensitive to these differences, if a correct result is to be obtained.

Just as a rendering algorithm is required to produce the appropriate visual image for some internal representation (as in the Hindi example (48) above), algorithms are required to provide a phonological interpretation for internal representations (for example to find all instances of *e*, regardless of the diacritical marks they bear). Transcription techniques will of course vary from language to language and author to author. The first major barrier to comparing forms by computer is unifying differing phonetic transcriptions. Several theoretical problems immediately crop up in this regard. It can be difficult to tell when different sequences of graphemes mark similar phonetic material, and (conversely) when a similar grapheme sequence marks different phonetic material. If in our data we encounter a sequence of graphemes /ch/ in one language and a /ch/ grapheme in another we would be remiss in assuming automatically that these indicate the same sound.[65] Only through an understanding of the system within which the symbols are used can they be compared. So for example, the sequence /ch/ is used in Fraser's orthography for Lisu to represent an unaspirated affricate, which /hch/ is used for the aspirate (this transcription is discussed in §3.2.1.2 below). In some circumstances it can be useful to ignore fine phonetic detail and perhaps even distinctive phonological oppositions in order to be able to make comparisons across languages. In other cases, ignoring such distinctions could cause important features to be overlooked. Cross-linguistic unification of transcription requires careful analysis and coding of transcriptions from different sources. In many if not all cases it can only be accomplished as an approximation. Given the appropriate instructions by the transcriber, or other extralinguistic information, we may often be able to equate /ʃ/ with /š/, /ph/ with /pʰ/, and /ñ/ with /ɲ/ in specific circumstances. Such instructions permit us to retranscribe the data of different languages with one set of graphemes, provided that the total grapheme set contains enough distinct elements to represent all the phonetic distinctions we may require. Doing so does not, of course, mean that we have a universal transcription! An example of two of each of these cases will be informative.

### 3.2.1.2. When the same symbols are used with different meanings

It is not unusual when comparing data from several sources to have the cases where the use of a symbol in one is quite different than that of another. Even when conventions and common sense converge to provide the basic scheme upon which both works are based, details contrive to make a straightforward comparison impossible. The *meaning* of a symbol in this case has the Prague School interpretation of designating a place in the structure of distinctions which are required to represent the sounds of the language.

The phonetic system of Proto-Bantu is conventionally reconstructed with seven vowels. In some Bantu languages this distinction is maintained, especially as the two highest reconstructed vowels are supposed to have been unstable and prone to merging with their neighbors. For languages which retain seven vowels, there are several transcriptions in use (as there are for five vowel systems), though only one type is shown in (50) below:

*(50) Different transcriptions for five- and seven-vowel systems in Bantu languages*

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | i | | u | i | | u | i(:) | | u(:) |
| 2 | ɪ | | ʊ | e | | o | | | |
| 3 | e | | ɔ | ɛ | | ɔ | e | | o |
| 4 | | a | | | a | | | a(:) | |
| | | 7 vowels | | | 7 vowels | | | 5 vowels | |
| | | Londo | | | Lingala | | | Laadi | |
| | | (Kuperus 1985) | | | (Dzokanga 1979:209) | | | (Jacquot 1982:17) | |

If we number the four vowel heights as shown in (50), we can see that while /i/ represents the highest vowel in all cases, /e/ represents vowels of at least two different heights. I say *at least* because /e/ in Laadi, which has a three-way vowel height distinction, is structurally different from the /e/ in the other two languages. We might say that /e/ in Laadi is 'in the middle' (i.e., it is *both* one feature from the top *and* one from the bottom), whereas in the other two languages it is *either* one from the top *or* one from the bottom. As we code this data for machine processing, we need to consider the eventual use to which we will put it: if we wish to find all words in Bantu languages which use the symbol /e/ in their transcription, we need think no further. If, on the other hand, we wish to retrieve all words which have a front vowel of 'aperture 3' (as numbered in (50) above), we will have to look for forms with /e/ in Londo and /ɛ/ in Lingala; and we will have to decide if the /e/ Laadi should qualify (as depicted in (50) /e/ has an aperture between 2 and 3, which we may wish to designate 2.5). The implication is that we need a representation of the symbols which describes for each language how it functions phonologically. This would probably be a description in terms of phonological features, in which case we would also need to explicitly relate the features used for each language.

Note that there are other transcriptions in use for 5- and 7- vowel systems in Bantu languages, in which some of these symbols would overlap in different ways; sometimes, for example, /i/ and /u/ are used for the second aperture of a seven vowels system, and /i̦/ and /u̦/ are used for the highest 'super-close' pair.[66]

Phonological inventories such as those above display the symbols in an array indicating the commonality of features. While such arrays are conceptually 'n-dimensional,' they must be flattened onto the printed page. Any computer representation should be able to translate between these representations. Consider for example the arrays given in (51) below:

(51) *Lisu Initial Consonants according to various sources*

Fraser 1922

| | | | |
|---|---|---|---|
| p | ch | r | |
| b | hch | ng | |
| hp | dz | sh | |
| d | ts | w | |
| t | hts | y | |
| ht | m | h | |
| g | n | h° | 'Nasal h's |
| k | l | hh | 'Guttural h' |
| hk | s | v | 'sometimes resembles ü' |
| j 'as in English' | | | |

Hope (in prep.)

| | | | | |
|---|---|---|---|---|
| p | t | ts | k | ʔ |
| ph | th | tsh | kh | |
| b | d | dz | g | |
| f | s | | x | h |
| v | z | | ɣ | |
| m | n | | ŋ | |
| | l | | | |

Burling 1967

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| p | py | t | ty | ts | c | k | kw | | |
| ph | phy | th | | tsh | ch | kh | khw | | |
| b | by | d | | dz | j | g | gw | | |
| f | | s | | | š | x | | h | hy |
| | | ř | | | y | ɣ | | | |
| m | my | n | ny | | | | ŋw | | |
| w | | l | ly | | | | | | |

Xu Lin 1986

| | | | | | |
|---|---|---|---|---|---|
| p | | t | | k | |
| ph | | th | | kh | |
| b | | d | | g | |
| | ts | | tʃ | | |
| | tsh | | tʃh | | |
| | dz | | dʒ | | |
| m | | n | ȵ | ŋ | |
| | | l | | | |
| f | s | | ʃ | x | h |
| v | z | | ʒ | ɣ | |

For some time the only data on Lisu was that of Fraser (1922).[67] More recently data transcribed according to modern descriptive principles has been published. Nevertheless, though the symbols used, e.g., /ts/ in Hope, Burling, and Xu all pattern differently (though the exact feature descriptions, i.e. the labels for the columns, are not given). From a diachronic point of view, Burling's is the most sophisticated, since it Hope's and Burling's transcriptions are the most similar, and may even be featurally equivalent if the handling of medials /y/ and /w/ are accounted for.

Some of these are not strictly a data preparation issues. However, they should be considered before data preparation in order that any solution be compatible with the use of symbols in the larger data cross-linguistic data set.

#### 3.2.1.3. And when different symbols should have the same meaning

There will be other cases in which we wish to systematically bring together sets of segments on the basis of shared traits, and we should consider how we should provide for this eventuality before entering the data. For example, Ringe notes that when comparing English and German initial consonants that both the matching of English /s/ and German /z/as well as English /s/ and German /s/ reflect a Proto-Germanic *s. Lexicostatistics computed on the basis of modern forms would compute different frequencies of occurrence, giving a skewed impression of the distribution of this sound in the comparative lexicon. Therefore, Ringe algorithmically 'groups together as a 'single consonant' all the consonants of a language that might have resulted from such a phonemic split' (Ringe 1992:67). He does this by retranscribing these symbols into cover symbols (as shown in (52) below) and computing the statistics on the basis of their distribution.

| | | |
|---|---|---|
| (52) | f,p,b | as 'P' |
| | z,s,ʒ,c,c' | as 'S' |
| | g,k | as 'K' |
| | etc. | |

(Ringe 1992:68)

Note that normally this type of retranscription or recoding takes place after the data has been prepared for computer use. Sometimes it may be easier to prepare the data in a retranscribed version first. Indeed, for many years, computer users had no choice but to use often arcane key combinations and sequences to transcribe characters not found on typical keyboards. This technique is applied to diachronic data in Tibeto-Burman in §5.1.4.

#### 3.2.1.4. Conventions for marking features in transcriptions

Representations also have to be sensitive to the *absence* of symbols. A transcription may allow 'default' or common features to be unmarked. However, for purposes of search and retrieval, this implied marking must be made explicit. Consider the following example from Written Burmese and Lahu:

*(53) from the Introduction to the Written Burmese Rhyming Dictionary (Benedict 1976vii)*

| | | |
|---|---|---|
| kyaŋ | filth | level tone |
| kyâŋ | trench | "heavy" tone |
| kyaŋ' | practice | creaky tone |

(54) *after Figure 2, Lahu Tones (Matisoff 1973:22, Matisoff 1979:35)*

| Name | Approx. Value | Symbol | Example | Gloss | Provenience[68] |
|---|---|---|---|---|---|
| Mid | $^{33}$ | (unmarked) | ca | 'look for' | < *1 -V, G, S; *3 |
| Low-falling | $^{21}$ | /` / | cà | 'fierce' | < *1 P |
| High-falling | $^{54}$ | / ˆ / | câ | 'eat' | < *2 V,-V |
| Very-low | $^{11}$ ~ $^{112}$ | / ˉ / | cā | 'feed' | < *2 G, S |
| High-rising | $^{45}$ | / ´ / | cá | 'boil' | < *G , S _ (p,t,k) |
| High-checked | $^{54}$ | / ˆ?/ | câ? | 'string' | < *-V _ (p,t,k) |
| Low-checked | $^{21}$ | /` ?/ | cà? | 'machine' | < *V _ (p,t,k) |

In this case, a 'canonical' transcription (which marks all significant features for searching, sorting, and retrieval) would have to note the fact that the level tone is unmarked. Furthermore, it would be best (i.e. simplest from the point of view of computer-assisted comparison) if tones were marked in a uniform way across the languages under consideration (i.e. if the tones occupied the same 'slot' in each case). As it stands, each of the Burmese tones is marked in a different fashion: level tone is unmarked, heavy tone is marked medially (on the vowel), and creaky tone is marked finally. In the case of the Lahu tones, the mid tone is unmarked, and the other tones are marked with diacritics, or a combination of diacritics and final ?. Note that the provenience of the unmarked tones is different (but overlapping) in Burmese and Lahu. When the time comes to compare forms from different languages transcribed in these ways, a great deal of retranscription and 'rectification' may be required. Consider the following data and possible 'canonical representations' (data from Matisoff 1979):

(55) *Canonical forms of various transcriptions*

|  |  |  |  |  |  |  |  |  |  | "Canonical"<br>form |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Language | Cited<br>form |  |  | Glyphs |  |  |  | I | G | V | F | T |
| WB | kok | k | o | k |  |  |  | k |  | o | k | $^{1}$ |
| Maru | kyuk | k | y | u | k |  |  | k | y | u | k |  |
|  |  |  |  |  |  |  |  | ky |  | u | k |  |
| Lahu | g̱̈ɔ̂?~vɔ̂? | ¨ | g̱ | ^ | ɔ | ? |  | ɣ |  | ɔ |  | $^{3}$ |
| Akha | g̱'oˆ | g̱ | ' | o | ^ |  |  | g̱h |  | o |  | $^{1}$ |
| Lisu | g̱aw3 | g̱ | a | w | 3 |  |  | g̱ |  | a | w | $^{3}$ |

**Legend:**

Columns headed by capital letters indicate syllable positions ('slots') as follows:

| | |
|---|---|
| I = initial | F = final |
| G = glide | T = tone |
| V= vowel | |

At some point, especially for etymological work, it will become necessary to compare the *constituents* of these forms; if so, depending on choices made earlier, it may be necessary to rearrange and recode the data. In the case of the Written Burmese data above, for example, an (arbitrary) tone number (1) has been supplied to mark the tone of the otherwise 'unmarked' case. In some instances it may be better to use a representation which would allow comparison of the tone values themselves, in which case a notational system such as the Chao system could be used. This system uses superscripted 'tone letters' (e.g., ˥˧ ˧ ˆ) which correspond to the perceived or measured F0 track. 'Tone numbers' can be used in the same way. When recording relative pitch or contour, by convention five distinctions, numbered $^{1}$ to $^{5}$ are used, with two numbers, e.g., high-falling might be $^{51}$, mid $^{33}$, and so on. Complex contour tones may require a longer sequence. Single tone numbers, or more accurate, a ordinal sequence of numbers are used if one is merely numbering the tones of a language. In the case of Lahu, the disjunct graphemic sequence /ˆ?/ as a whole has been replaced by a tone number. In some cases a one-glyph grapheme has been substituted for a two-glyph variety (e.g., ɣ for g̈). Note that the segmentation of the forms into initial (I), glide (G) vowel (V) , final (F), and tone (T) is in part arbitrary at this stage. The Maru form, for example, is ambiguous with regard to this segmentation; more information about the nature of the constituents in this language is needed to resolve the matter. In fact, this type of segmentation problem is ubiquitous in Tibeto-Burman languages, especially when seen from a diachronic and comparative perspective.

Many computer programs for storing and retrieving cross linguistic data (including several discussed in §2) have required that the data be entered initially in a canonical form, either to avoid the need to implement complicated language-specific program modules to create this form when it is needed or to avoid the inevitable ambiguities which arise when trying to segment the forms. However, as a result of making decisions like these, the data is now frozen in a single invariant and perhaps in hindsight unsuitable structure.

The problem of creating machine-internal phonetic and phonological representations is far from solved; the IPA is not universally accepted as a means of transcribing speech, and even if it were, numerous phonological issues would remain (cf. for example Chao 1966 on the non-uniqueness of phonemicization, or the literature on the 'proper' use of the IPA (Ladefoged forthcoming)).

The specification and implementation of those language-specific processes that are required to make computer software operate flexibly in a sensible way for different languages is called *localization.* Localization must have linguistic and cultural inputs to work properly; information about direction of writing, hyphenation (if any), representations of numbers, times, dates, currencies, and so on must all be accounted for; that these types of information are the targets of localization reflect the commercial orientation of most computer users. The process is still in its infancy for most computer systems. Some of the major problems which linguists must be aware of when transcribing their data for computer processing are discussed below.

### 3.2.1.5. Sorting and searching

The problems associated with making claims about the identity of segments (phonetic or phonemic) within and across languages constituted a significant focus of debate among American structuralists. The same problems exist for computer implementations, except that here even the most obvious relationships must be explicitly and carefully stated. Once a well-defined method of transcription is established, considerable work is still required to provide a means of manipulating data in that transcription. For example, the ordering of lexical items with respect to each other is not always a simple task. For purposes of *machine* retrieval the ordering of lexical items is relatively trivial: a default appeal to the conventionally assigned alphanumeric values of the machine-internal representations (normally as sequences of bytes) will work. However, the assignments suitable for keyboarding and display may conflict with the assignments which would facilitate sorting. For example, when sorting words in a Norwegian dictionary all variants of a segment should sort together; thus in (56) below, all the versions of *A* should sort together, as indicated in (a). However, since numeric codes assigned internally to the conventional (undiacriticized) letters have preempted the language-specific ones, there is a gap in the values between the related letters (as indicated in (b)). Since sorting occurs on these internal numeric representations, the letters will not sort properly unless specific steps are taken (i.e. the collating sequence of the characters must be specified for the language being sorted). The < symbol shows how these symbols should collate.

*(56) 'All of the forms of A must precede all of the forms of B'*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (a) Letter | A | < | Å | < | a | < | å |
| (b) Mac OS Codepoint | 65 | | 129 | | 97 | | 140 |

| | | | |
|---|---|---|---|
| (a) Letter | B | < | b |
| (b) Mac OS Codepoint | 66 | | 98 |

(after Apple Computer 1991:14-27)

Many searching and sorting applications, however, require linguistically sensitive orderings.

> Sorting or comparing strings can be an extremely intricate operation. Subtle issues like expansion, contraction, ignorable characters, and exceptional words may [have] to be taken into account. Sorting cannot be done properly by a simple table look-up, even for such straightforward cases as English. Sorting depends not just on the script, but on the individual language. (Apple Computer 199114.26)

Of course, many of these problems are obvious to anyone who has tried to use a telephone book, but when the issues are treated as problems in many languages requiring a single solution, they become more subtle, which is computerese for 'intractable or requiring a lot of programming.' Some issues which need to be addressed are given in (57) below. In German, for example, *bäk* and *baek* are lexicographically equivalent, though here the version with *ä* is shown a sorting before *ae*. In Spanish, *ch* must sort as a unit, and so all words beginning *ch* should sort after *czar*. Punctuation must be accounted for, usually, but not always, ignored.

(57) *Examples of 'subtle issues' sorting Roman characters*

| | | |
|---|---|---|
| a single character sorts as a sequence | bäk < baek < bäks | German |
| a sequence sorts as a single character | czar < char < dar | Spanish |
| characters must be ignored in sorting | blackbird < black-birds | English |

From a linguistic point of view, schemata for lexicographic ordering break down into three groups:

- Otherwise unmotivated orderings based on a traditional sequence, such as the ordering of the Roman alphabet.[69]

- Orderings based on similarities in the appearance of the symbols used. An example is Pullum and Ladusaw's phonetic symbol guide, in which the various characters are ordered according to their appearance (Pullum and Ladusaw 1986:xxiv). Thus the following characters appear together in sequence in their list:

*(58) Examples of 'm-like' graphemes*

| (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|
| | | | Usage in | |
| Grphm | Keystroke | Name (Pullum and Ladusaw) | transcriptions | CodePoint |
| [m] | m | Lower-case m | bilabial nasal | 109 |
| [ɱ] | opt-m | m with leftward tail at right | labiodental nasal | 181 |
| [ɯ] | opt-w | Turned m | high back unrounded vowel | 183 |
| [M] | sh-m | Upper-case m | | 77 |
| [ᴹ] | sh-opt-n | Raised upper-case m | mid-tone category | 247 |

The ordering of Chinese characters based on radicals is another example of this type of shape-based ordering.

- Orderings based on the phonetic values of the symbols used. Examples of this are the Devanāgarī writing systems, where the consonants appear in a sequence determined by their position and manner of articulation. While this is also a 'traditional' ordering, it is a motivated one. Alphabets of this type have been adopted for all the great literary languages of Southeast Asia except Vietnamese (viz. Mon, Khmer, Cham, Old Javanese, Burmese, Tibetan, Thai, Lao), as well as for several minority languages with much more recently devised writing systems (e.g. Karen, Shan, Newari, Manipuri). There is also a tradition going back to the 19th century of arranging bilingual dictionaries of SE Asian languages in a Devanāgarī-style order (as in (59) below) even when the words in fact appear in a Romanized orthography (e.g. Mainwaring 1898, Srinuan 1976, Grüssner 1978, Matisoff 1988 exemplified below). The five 'cardinal' vowels also appear in a conventional order *(a i u e o)* in Indic-influenced alphabets, with additional or modified symbols intercalated or added to these basic five according to the needs of the particular language (Matisoff 1991).

(59) *Lexicographic collating order of the Sanskrit alphabet, based on manner and place of articulation. The table is to be read left to right and top to bottom. (after Whitney 1889:26) and (Macdonell 1929:x)*[70]

**Vowels**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| a | ā | | i | ī | | u | ū |
| ri | (- ṛ) | | rī | (- ṝ) | | li | (- ḷ) |
| | e | ai (- āi) | | | o | au (- āu) | |

**Consonants**

| | | | | | |
|---|---|---|---|---|---|
| k | kh | g | gh | ṅ ( = ŋ) | |
| *k* (- c) | *kh* (- ch) | *g* (- j) | *gh* (- jh) | (- ñ) | |
| *t* (- ṭ) | *th* (- ṭh) | *d* (- ḍ) | *dh* (- ḍh) | *n* | |
| t | th | d | dh | n | |
| p | ph | b | bh | m | |
| | y | r | l | | |
| | *s* (- ś) | sh (- ṣ) | s | | h |

A concrete example of an ordering sequence of this sort is seen at the bottom of each page in Matisoff's Lahu Dictionary:

(60)

a á â à ā â? à? aiueoɛɔɨə q qh k kh g ŋ c ch j t th d n p ph b m h ğ š y f v l

Significantly, most Tibeto-Burman lexicographers nowadays adopt the conventional ABC order of romanization—undoubtedly in some cases to facilitate machine-sorting of the data.

Dictionaries of monosyllabic tonal languages must be alphabetized on another dimension as well, namely by *tone*. The many syllables which are homophonous with respect to consonants and vowels, but which differ in tone, must be presented in a consistent order. Tonal contrasts may be indicated using several different symbolic systems:

- by diacritics (acute, grave, circumflex, etc.). The Dictionary of Lahu (discussed in §4.2.5 below) exemplifies this method.

- by a superscripted numeral or sequence of numerals (Chao numbers or letters, exemplified above).

- by arbitrary syllable-final consonants (this last option works particularly well in languages with no 'real' final consonants). Bradley's Lisu dictionary ((Bradley 1994) discussed below in §4.2.4), and a Hani transcription, (described in §6.6.4) both use this method.

Lately some Chinese-English dictionaries have abandoned the traditional arrangement in terms of graphic elements ('radicals' and number of strokes) in favor of a listing based on the alphabetical order of transcription in the excellent *'pinyin'* romanization of Mandarin (used in the People's Republic of China).

### 3.2.1.6. Morphology, compounding and homophony

Sorting applications may have to be cognizant of morphological boundaries:

> Sometimes prefixes pose lexicographical problems. The nominalizing or bulk-providing Lahu prefix ɔ̀- occurs as the first syllable in 6% (86 pages) of the entries in Matisoff 1988 (1414 pages). Similarly 11.8% of Judson's 1061-page Burmese-English Dictionary begins with the functionally identical prefix ʔə-. Even at the cost of the extra space, it is a good idea to list each of these prefixed morphemes twice, both in prefixed and unprefixed form, since there is often semantic differentiation and their syntactic properties are different. (Matisoff 1991)

Lemmatization may affect sorting decisions. For example, the relatively complex morphophonemics of some 'Indospheric'[71] TB languages may require the listing of several variants in an entry—e.g. the four 'principal parts' of Tibetan verbs, the independent vs. subjunctive forms of Kuki-Chin Naga verbs, or the pairs of Newari verb-forms both with and without a thematic stem-final consonant. These different forms must either be kept together as part of a dictionary article or the sorting algorithm must be sensitive to particular morphological facts of the language in order to bring the entries into correct sequence.

The generally monosyllabic Asian languages compensate for the homophony problem endemic to languages of this type by creating polysyllabic compounds. By and large each such syllable has a relatively clear meaning of its own; Matisoff has coined the term *morphan* (i.e. orphan morph) to designate semantically obscure syllables which only occur in one or two collocations.) A recurrent morph should be listed as a separate head entry even if it does not occur initially in compounds. By the same token it is often a good idea to list morphans separately as well, especially for comparative purposes—a bound morpheme in one language may well be cognate to a free morpheme in another.

Matisoff deplores the widespread practice of 'pernicious interalphabetization,' where homophonous monosyllables are listed *en bloc*, after which all the collocations beginning with that same phonological syllable appear in strict alphabetical order according to the initial of their second syllable, regardless of what the morphemic identity of their homophonous first syllable may be. Unless the meanings of the first syllables are quite disparate, it is often unclear which collocations relate to which homophone. Folk etymologies and conflations can go unrecognized (Matisoff 1987).

### 3.2.1.7. Phonotactics and parsing

For some applications it is necessary to indicate (or to be able to find algorithmically) various phonological and morphological boundaries. Languages usually have a range of phonotactic structures and, as pointed out above (p. 160), a morpheme may fit into more than one structure at a time (i.e. its parsing is ambiguous). From a computer programming point of view, the problem can be stated as follows: given a form in a particular language or proto-language, a suitable specification of the syllable canon (or other phonotactic description) and its constituents, is it possible to uniquely decompose the form into its morphological and syllabic constituents and if not, what are the consequences? For one of the simplest situations, in which we attempt to divide a monosyllable into initial (i.e. a sequence of [+consonantal] graphemes) and rhyme (i.e. a sequence of [+vocalic] followed by [+consonantal]), specifying the algorithm is not too hard. If the vocalic elements are all transcribed by single graphemes and no grapheme serves more than one purpose, the syllable may be divided by examining each grapheme in a lexeme from left to right until the first non-consonantal grapheme is found, or equivalently by proceeding from right to left until the leftmost vocalic element (i.e. the last grapheme marked [+vocalic] in the grapheme inventory) is found. For example, consider the following data set:

*(61)*

| | | |
|---|---|---|
| 1. kok | k- | ok |
| 2. kyuk | ky- | uk |
| 3. g̈ɔ̂? | g̈- | ɔ̂? |
| 4. g'oˆ | g'- | oˆ |
| 5. gawʒ | g- | awʒ |

Correct syllabification in these cases depends on several chance features of the transcription. First, it is assumed that the diacritic marks (i.e. glyphs used to compose complex graphemes) for vowels and for consonants are distinct. Thus, if the umlaut was a separate glyph, the umlaut over /g̈/ would have to differ (in its internal representation) from the umlaut used to mark vowels (no example of this use is given above, however). Similarly, the circumflex used to mark tone in (3) must be regarded for purposes of syllabification as a vocalic feature. If the glyphs composing vowels and consonants are not distinct, then the syllabification algorithm will have the added problem of determining the class membership of the grapheme from context. The order of glyphs in a grapheme may also have to be taken into account. Furthermore, since the glyphs making up the graphemic representation of some phonological unit may not be spatially adjacent, some means of grouping or resequencing graphemic subsequences is needed in order to assign units to the correct phonotactic slots. These considerations (and others) make the dividing of transcribed forms into their phonotactic constituents a non-trivial programming problem, and can make whatever solutions are found difficult to apply to new data or to implement in another programming environment. An application of this technique to the Loloish languages is given in §5.2.

### 3.2.1.8. Prosodic structure and the 'chaining' of syllable canons

The preceding discussion opens another area which needs to be considered in preparing diachronic lexical data for computational analysis, and which is especially important in the study of TB languages: how to handle the interaction of syllabic and morphological phenomena. As the quote below indicates, traditional analyses of Tibeto-Burman languages have had to accommodate both types in a single structure:

> Tibeto-Burman, as reconstructed, can be described in general terms as a relatively isolating language with roots of simple monosyllabic

type, normally prefixing but occasionally suffixing. (Benedict 1972:92)

Benedict's general description has been recast in more concise notation by Matisoff 1979. This restatement allows us to define a data structure which can make the Tibeto-Burman syllable 'processable' by computer. Given the elements, we first build a simplified version of the *TB syllable canon as follows:

*(62)*

| | | | |
|---|---|---|---|
| let | S | = | the syllable |
| | P | = | prefix |
| | C | = | core, i.e. $(C_i)V(:)(C_f)$ |
| | F | = | suffix ('final') |
| | () | = | optional element |
| then | S | = | (P)C(F) |

Each of the constituents of the syllable canon is manifested as a composition of elements selected from a constrained set of possibilities. For example, prefixes may consist of one of the following types:

1. More-or-less meaningless segments, including 'formatives,' and 'pre-initials,' which provide 'phonological bulk' (Matisoff 1978:23).

2. Elements with a concrete lexical content, used to classify or specify.[72]

3. Elements with an abstract grammatical function.[73]

The syllable is being viewed here as composed of an ordered set of sometimes optional phonemes drawn from restricted sets. Additionally, tone (and perhaps register) must be included suprasegmentally over the entire syllable or at least those portions of the syllable for which it is relevant. Matisoff has specified the types of phonetic material that can occur in the various elements of the *TB syllable, thereby 'fleshing out' the definition of the canon. In his terms,

(63) *Constituents of the *TB syllable*

let: S = the syllable, as above.

P = prefix; one or two obligatory elements selected from a specified set of syllabic and non-syllabic elements; i.e. P = $((P_1)(P_2))$.

R = composed of an obligatory 'initial', which may be a 'simple' consonant $C_i$, or a cluster (consonant plus glide: $C_iG$); a 'vowel' V (including diphthongs, along with other prosodic features such as length, creakiness, etc.) ; and an optional 'final consonant' ($C_f$).

F = post-final; like prefixes, an element selected from a specified set of segments; in this case, however, only one element can occur, i.e. $(F_1)$. The only segment that can fill this slot is /s/.

T = tone, associated with the entire syllable, no matter where or how it is marked.

then:

(64) *TB syllable canon: S = $((P_1)(P_2))C_i(G)V(:)(C_f)(F_1)^{(T)}$

(Matisoff 1987)

Traditionally, the portion of the syllable beginning with the vowel and ending with the final consonant ($C_f$)) and/or ($F_1$) is called the 'rhyme', and the preceding material is called the 'initial:'[74]

| (65) | $((P_1)(P_2))C_i(G)$ | $V(:)(C_f)(F_1)^{(T)}$ |
|---|---|---|
| | 'Initial' | 'Rhyme' |

Rhymes are important in TB historical linguistics inasmuch as phonological developments often occur with respect to these sets of segments as a whole; at least, the changes are described in these terms.

Note that this syllable canon admits a maximum of 24 different syllables. Since the structure above is panchronic, we cannot expect to find any maximal instances of it. Tibetan, however, whose syllable structure is (not by accident) quite similar to that of the *TB protolanguage, exemplifies forms which are nearly maximal:

(66) Written Tibetan (WT) *brgyad '8'*

| | | | | | |
|---|---|---|---|---|---|
| *b* | *r* | *g* | *y* | *a* | *d* |
| $P_1$ | $P_2$ | $C_i$ | G | V | $C_f$ |

Written Tibetan *bsnyigs 'sediment'*[75]

| | | | | | | |
|---|---|---|---|---|---|---|
| *b* | *s* | *n* | *y* | *i* | *g* | *s* |
| $P_1$ | $P_2$ | $C_i$ | G | V | $C_f$ | F |

It may not be possible to unambiguously parse any particular form with respect to this canon, since glides may be interpreted either as consonants or vocalic elements. Ambiguity in decomposing a syllable according to a given syllable canon is problematic for language speakers in much the same way as it is problematic for computers. Consider the Proto-Lolo-Burmese reconstruction /krak/ with regard to the *LB syllable canon. Since a synchronic sequence /kr/ can reflect either a *prefix plus root-initial or an *initial plus glide, and since /r/ can occur as either an initial or a medial, the decomposition of this form is ambiguous as the examples in (67) below illustrate. When the syllabification process is performed by human speakers, such ambiguity can result in 'reanalysis' or 'metanalysis.'[76] Similarly, the second to last form in (67) could be analyzed as either /t+wak/ or /tw+ak/. Given the evidence from Lahu và? pig, it is clear that the /t+wak/ analysis is more appropriate.

(67) *Etyma showing metanalysis of initial*

| Gloss | *LB | WB | Lahu |
|---|---|---|---|
| chicken | *k-rak | krak | ğ̱â? |
| crossbow | *krak | - | khâ? |
| weave | *rak | rak | ğ̱à? |
| | | | |
| emerge | *?twak | thwak | tɔ̂? |
| pig | *wak | wak | và? |

In all Sino-Tibetan reconstruction the decomposition of lexemes into these types of syllabic constituents is an absolutely crucial step. While many modern Sino-Tibetan languages have simple syllable structures, historically the reconstructed Proto-Sino-Tibetan syllable was quite complex:

> 'The basic phonological unit in Sino-Tibetan was the syllable. The following syllable canon is posited[...}:
>
> $C(C)(M_1)(M_2)(M_3)V(V)C(V)^{T(?)}$
> C = Consonant
> V = Vowel
> T = Tone
> M = Medial semivowel:
> Class $M_1$ (high front)= -j- -y-
> Class $M_2$ (liquid) = -r- -'r- -l- -'l-
> Class $M_3$ (labial) = -w-
>
> [There follows a 'system of sounds' specifying which phonemes may occur as initial consonant, vowel, and final consonant segments of the syllable.]
> In the system presented here more than one medial class may be represented in a syllable containing members of all three classes, but no syllable containing members of all three classes has so far been reconstructed. Only one member of each medial class may occur in a syllable.' (Coblin 1986:13)[77]

Inasmuch as modern syllable structure is a product of changes in an earlier structure, the syllable canons of each intermediate chronological level must be specified and related. This can be accomplished by specifying the chain of structural mergers by protoconstituents combined into more modern constituents; of course, there may be several possibilities (Matisoff 1987). Below are given the Proto Tibeto-Burman syllable canon, that of two of its 'daughter proto-languages,' Proto Lolo-

Burmese (*LB) and Proto-Loloish (*L), and that of a particular modern Lolo-Burmese language, Lahu:

*(68)*

| | |
|---|---|
| *TB: | $((P_1)(P_2))C_i(G)V(:)(C_f)(s)$ |
| *LB: | $((P_1)(P_2))C_i(G)V(:)(C_f)^{(T)}$ |
| *L: | $(P)C_i(G)V(:)(C_f)^{(T)}$ |
| Lahu: | $(C_i)V^T$ |

Here:

| | |
|---|---|
| P | = Prefix |
| $C_i$ | = Initial consonant |
| G | = Glide (only w,y,r,l in these cases) |
| V | = Vowel (including diphthongs) |
| : | = Vowel length |
| $C_f$ | = Final consonant |
| T | = Tone[78] |

Viewing the syllable in terms of the traditional break between 'initial' and 'rhyme' highlights the simplification of structure which is supposed to have occurred. As noted above (p. 171), sound changes in these language families are described in terms of these larger constituents. The implication therefore is that computer software must be able to not only to manage rather complex arrangements of glyphs and graphemes, but also to take into account the larger linguistic structures in which these elements play a part. Given the different assumptions made by different researchers, attempting to relate these structures to each other is difficult to do in any precise and deterministic way. The following represents something like the desired result:

*(69)*

| Language | 'Initial' | 'Rhyme' |
|---|---|---|
| *ST | $C(C)(M_1)(M_2)(M_3)$ | $V(V)C(V)^{(T)(?)}$ |
| *TB | $((P_1)(P_2))C_i(G)$ | $V(:)(C_f)(s)^{(T)}$ |
| *LB | $(P)C_i(G)$ | $V(:)(C_f)^T$ |
| Lahu | $(C_i)$ | $V^T$ |

As in the problem of disambiguating transcriptions, the strings of characters themselves give little information about which graphemes (=segments here) are to be assigned to which slots.[79] Indeed, in some cases the glide (G in the TB canons, M in the Proto-Sino-Tibetan canon) may be regarded as part of the rhyme rather than the initial. Any algorithm which specifies how to decompose a lexeme according to the syllable canon must therefore at times avail itself of some non-phonetic information to operate correctly.

It is important to emphasize the quantitative and qualitative complexity of the syllable canon in Sino-Tibetan. Assume that the number of 'valid' or 'possible' syllables is explicitly limited by:

1. The order of the syllabic subelements (which I have been calling 'slots').

2. The phonetic material (usually segments) which can occur in a given slot.

3. Cooccurrence limitations on different slots.

In this analysis, Lahu, according to its syllable canon, has a maximum of 1,575 different syllables:

*(70) The number of possible Lahu syllables*

$$(24\,C_i + 1\,\emptyset) \times (9\,V) \times 7^T \qquad = \qquad 1{,}575$$

In fact, the *LB syllable is more constrained than the version given in (71) above shows: the syllable structure is a conjunction of two types, an open syllable and a closed syllable, each with its own tonal system. It has at any rate a larger inventory of positional slots, and a larger inventory of fillers than the Lahu syllable. This syllable canon is discussed in more detail in §6.4.3. Here, however, I give it as an illustration of the differences in complexity of various structures:

*(71) The *LB syllable canon*

$$\left\{ \begin{array}{l} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \begin{pmatrix} P \\ \emptyset \end{pmatrix} \begin{pmatrix} I \\ \emptyset \end{pmatrix} \begin{pmatrix} G \\ L \end{pmatrix} \begin{pmatrix} VN \\ V \end{pmatrix} \\ \\ \begin{pmatrix} HS \\ LS \end{pmatrix} \begin{pmatrix} P \\ \emptyset \end{pmatrix} \begin{pmatrix} I \\ \emptyset \end{pmatrix} \begin{pmatrix} G \\ L \end{pmatrix} (VS) \end{array} \right\}$$

These two different descriptions 'generate' different syllables, of course. Another level of constraint is understood to exist (and would have to be encoded in any computer representations) which limits which elements can cooccur. Stating the vowel+final combinations as rhymes is one way of indicating such constraints. For example, e. ö. and o occur only in closed syllables, and the distribution of medial -w- is at least somewhat restricted.[80] These limitations are empirical; while 'violations' of these constraints are articulatorily possible (and indeed exploited in other languages), they do not occur here. Other limitations, both phonetic and structural, may be assumed which would further reduce the number of syllables; it would be quite absurd, for instance, to assert that *LB would have tolerated a syllable like /l-l-l-ö-:-l/ just because the syllable canon permits it. A computer implementing the *LB syllable canon, however, would not be able to reject such a reconstruction unless all cooccurrence limitations were explicitly incorporated.

In fact, only about two dozen stop syllable rhymes are possible in *LB (Matisoff 1979:18). Even if a high range of consonantal variation were tolerated, therefore, the number of possible *LB syllables would only be in the thousands.

These facts about the complexity of syllable canons are significant when considering the design of algorithms and data structures which provide intelligent decomposition of syllables. If it were possible simply to list and store all the possible syllables for a particular language, for example, the problem of syllable parsing could be reduced to the problem of searching a 'syllable database' which contained, for each possible orthographic syllable, the syllable shapes it might represent.

### 3.2.1.9. Writing etyma

Structural considerations of the type discussed above are immediately evident when the linguist considers the problem of representing reconstructed forms. The need to record various types of variation in form arises when representing modern forms of course, but comes into sharpest focus when comparing reconstructions. So, when converting or comparing existing etymological sources, some method for indicating the range of phonological shapes and the semantic centers of gravity must be developed. Here I discuss how the variations arise and analyze the implications for computer treatment of this type of data.

Reconstructions, as phonological entities, have a number of unusual properties, some of which have already been noted. Many linguists regard them merely as formulas for the correspondences observed in their reflexes (cf. for example, Meillet), while others assert that they are our best estimates of actual spoken words. Morphological processes in the protolanguage leave traces which obscure the exact form of the original words. And at any rate, we can see them only indistinctly, limited as we are by the haphazard nature of the evidence preserved and the instruments of internal and comparative reconstruction.

> ...[E]tyma are not invariant in shape, but form clusters of morphophonemically related sub-roots, which have traditionally been referred to as *word families*. The 'allofams' of a word family

> may differ from each other by their prefixes, by the voicing or voicelessness of their initial consonant, by their nuclear vowel, by their final consonant, and/or by their tone. These patterns of variation are not random, but fall into certain well-defined classes of phenomena.[81] Great care needs to be exercised in distinguishing genuine, well-attested patterns of allofamic variation[82] from accidental similarities between non-related forms—and in attributing particular forms in modern languages to the particular proto-allofam from which they descend. (Matisoff 1994:51)

Matisoff identifies two major categories of such allofamic variants (Matisoff 1994:52):

(a) *Language-internal morphological processes.* In Sino-Tibetan these are typically derivational, not inflectional—hence idiosyncratic and sporadic, and highly susceptible to analogical pressure.[83]

(b) *Borrowing between related languages.* An extreme example is the huge Chinese component in the lexicon of Bai. (Matisoff 1994:52)

Matisoff exemplifies these two types of allofamic variation in English with the Indo-European example **wed-* 'water, wet' (< (Watkins 1985:73):

*(72) (a)* *Inherited Germanic material*

1. *wod-ōr [suffixed o-grade]
> PGmc *watar > OE wætar > ***water***

2. *wēd-o- [suffixed lengthened grade]
> PGmc *wēd- > OE wǣt, wēt > ***wet***

3. *wod- [o-grade]
> PGmc *wat-skan > OE wæscan, wacsan > ***wash***

4. *we-n-d- [form with nasal infix]
> PGmc *wintruz 'wet season' > OE winter > ***winter***

5. *ud-ro-, *ud-rā [suffixed zero-grade]
> PGmc *otraz 'water animal' > OE otor > ***otter***

*(72) (b)* *Borrowings from other Indo-European languages*

6. *ud-ōr [suffixed zero-grade]
> Greek hudōr 'water' > HYDRO- (incl. ***clepsydra, dropsy***)

7. *u-n-d-ā [suffixed zero-grade with nasal infix]
> Latin unda 'wave' > ***undulate, inundate, abound,***
***redundant, surround***

8. *ud-skio [suffixed zero-grade]
> Scot. and Ir. Gaelic uisge 'water' > ***uisquebaugh, whiskey***

9. *wod-ā- [suffixed o-grade]
> Russ. voda 'water', with -ka 'diminutive' > ***vodka***

(After Matisoff 1994:52, Matisoff 1992:160)

To these I would add two more less common but nevertheless important categories:

(c) *Differential preservation of dialect variations which existed in the protolanguage.*

In some cases, especially at the microlinguistic level, it is possible to reconstruct variations which must reflect dialect differences in the protolanguage which are differentially preserved by the modern dialects. Where otherwise well-supported correspondences produce two different reconstructions for subsets of clearly related forms, we may wish to retain two proto-variants to avoid unwarranted 'stuffing' of the reconstruction.[84]

An excellent example of this type of variation from the Tani subgroup of Tibeto-Burman (located in Assam and Arunachal Pradesh, see Appendix 2) is given by Sun Tianshin in his reconstruction of Proto-Tani. Noting that

> [...] variations, both on the phonological and semantic level, must be taken account of in historical reconstruction (Matisoff 1978a). One of the implications of this principle is that not every observed synchronic correspondence goes back to a uniform proto-entity. (Sun 1994),

Sun presents the cognate sets below as an illustration:

(73) **Tani 'tail'*

| **GROUP A** | | **GROUP B** | | |
|---|---|---|---|---|
| PT *me | | PT *mjo | | |
| Apatani S | a-mi | Bengni S | ñu-bjuŋ | |
| Padam L | (t)a-me | Bokar OY | e-mño | (<-mjo) |
| Damu OY | me-čuŋ | Bori M | ño-buŋ | |
| Milang T | ta-mi | Gallong W | ¯ ño-bu | |
| Nyisu H | ta-mi | Hill Miri S | añ-ño | |
| Yano B | me-uŋ | Mising L | ta-mño | (<-mjo) |
| Tagin B | a-me | Tagin DG | ña-buŋ | |

As Sun explains, Apatani S -i and Padam-Mising L -e here exemplify a regular correspondence pattern **-e* (cf. Apatani S si-bi, Padam-Mising L si-be, Bokar OY sə-be, Gallong W ˆ ho-be 'monkey' < PT *be:), while the forms in Group B exemplify **-jo* (cf. Bengni S rju:; Bokar OY o-jo; Bori M a-jo; Gallong W ˆa-jo; Mising L a-jo; Tagin DG rju 'tongue' < PT *rjo). Rejecting 'the reductionist view that historical reconstruction should always reduce synchronic variation to earlier invariance (for discussion please see (Hock 1986:18.7)', Sun decides to stop after the two alternating reconstructions have been worked out on the basis of modern forms in Group A and Group B,[85] thus claiming that there already existed two competing variant proto-forms, *me and *mjo, at the Proto-Tani stage.

(d) *Allofams which are vestiges of the data collection and reconstruction process.*

In this case, the variant etyma betoken nothing more than uncertainty about the reconstruction. However, this information can be important, especially at early stages of the reconstruction process, and so it is prudent to provide a means to indicate it. When the reconstruction of a particular single segment is uncertain, this could be marked using typographic styles of some sort (bold or italic). However, many other types of uncertainty exist; conventions are needed to indicate that an element may or may not be present, or that the element may be one of a list of possibilities.

An example from another reconstructed language, Proto-Tamang-Gurung-Tukche=Marpha (*TGTM) (Mazaudon 1973; Mazaudon 1994), illustrates how the data may require the systematic recording in the reconstruction of variation of this type. Here several cognate sets appear to have both $*^{A}$ and $*^{B}$ variants; this can occur either where some of the daughter languages do not give good evidence for the tone (i.e. tone not recorded or incorrectly recorded) or where evidence from several dialects is contradictory or ambiguous:[86]

(74) *Apparent co-allofams which result from inconclusive data (Mazaudon 1994)*

22. $^{A}$kam [1.100.59]

| | | |
|---|---|---|
| ris | $^{1}$kam | *fiel, bile* |
| tag | $^{X}$kam | *bile* |
| tuk | $^{H}$kəm << kɔm | *bile* |
| mar | $^{51}$kʌm | *bile* |
| mar | $^{55'44}$kʌm-mu | *amer* |
| syang | $^{54}$kʌm | *bile* |
| gha | $^{1}$kã: << kã: | *bile* |
| gha | $^{1}$kã:- << kã:ba:q | *bitter.* |

23. $^{B}$kam [2.100.59}

| | | |
|---|---|---|
| sahu | $^{2}$kam << $^{2}$kam=pa << 'kampa | *bitter* |
| tag | $^{X}$kam | *bile* |
| tuk | $^{H}$kəm << kɔm | *bile* |
| tuk | $^{2}$kəm =pə << kɔm-pɔ | *bitter* |
| mar | $^{51}$kʌm | *bile* |
| mar | $^{55'44}$kʌm-mu | *amer* |
| syang | $^{54}$kʌm | *bile.* |

*NB: << is used to indicate the chain of retranscription from the original source to the form used in the automated comparison.*

Here Sahu $^{2}$kam reflects an ancestor with tone $*^{B}$, whereas Risiangku ('ris') $^{1}$kam can only have descended from tone $*^{A}$. Since tone was not recorded for the Taglung ('tag') form, it may match either an $*^{A}$ or $*^{B}$ reconstruction; data from the other languages is likewise either ambiguous of contradictory with respect to the tone. Thus two different reconstructions predicated on highly overlapping sets are quite possible (see §6.5 for details about the constituency of cognate sets). I should make it quite clear that the idea here is not that there was definitely a tonal alternation in the protolanguage, but that we cannot avoid suspecting such an alternation given the data available to date. Three further such examples are given below to show that the pattern of ambiguous and contradictory data for tone is not restricted to particular dialect groups, so, like the example in (c) above (73), the effect cannot be the result of modern dialect variation.

(75) *Three *TGTM cognate sets exhibiting tonal variation*

188. $^{A}$gla: [3.125.34]

| | | |
|---|---|---|
| tuk | $^{3}$kja << kjah | *place* |
| syang | $^{11}$lja | *place* |
| gha | $^{3}$ɬo | *place.* |

189. $^{B}$gla: [4.126.34]

| | | |
|---|---|---|
| sahu | $^{4}$kla: << $^{4}$kla: | *place* |
| mar | $^{54}$lja | *place* |
| syang | $^{11}$lja | *place* |
| pra | $^{4}$k$^{h}$ja | *place.* |

315. $^{A}$tsja: [1.134.34]

| | | |
|---|---|---|
| ris | $^{1}$tsja: | *regarder/ examiner* |
| sahu | $^{1}$tsja: << $^{1}$tsja:=pa << cya:pa | *to see/ to look at* |
| tag | $^{1}$tsja:-ba | *to look* |
| tag | $^{X}$tsja:-dzi | *to look* |
| mar | $^{55'44}$tsja-wa | *chercher / regarder, examiner* |
| syang | $^{55'44}$tsja•-go | *chercher.* |

316. $^{B}$tsja: [2.134.34]

| | | |
|---|---|---|
| tag | $^{X}$tsja:-dzi | *to look* |
| tuk | $^{2}$tsja =lə << cja-lɔ | *search for* |
| mar | $^{55'44}$tsja-wa | *chercher / regarder, examiner* |
| syang | $^{55'44}$tsja•-go | *chercher.* |

317. $^{A}$tsjaŋ [1.134.53]

| | | |
|---|---|---|
| ris | $^{1}$tsjaŋ | *être petit, un peu, y en avoir peu* |
| mar | $^{55'33}$tsjaŋ-ba | *petit* |
| syang | $^{55'44}$tsjaŋ-ba | *petit.* |

318. $^{B}$tsjaŋ [2.134.53]

| | | |
|---|---|---|
| tuk | $^{2}$tsjaŋ =pə << cjaŋ-pɔ | *small* |
| mar | $^{55'33}$tsjaŋ-ba | *petit* |
| syang | $^{55'44}$tsjaŋ-ba | *petit.* |

### 3.2.1.10. Pan-allofamic formulas

The need for some convention to record such variations in a succinct and useful way is a significant desideratum in comparative work. Indo-Europeanists normally use the e-grade (also called 'normal grade') of the root as the citation form (for sorting and reference purposes) and list the co-allofams under this form. Thus, the nine allofams given in

(72)(a)(b) above may be found in Watkins' etymological appendix to the American Heritage Dictionary under *wed-. An analogous convention, however, is not (yet) available for other, less-studied families. Matisoff instead proposes using a notation which 'displays all well-attested variants simultaneously' (Matisoff 1994). Using this notational device, called a 'pan-allofamic formula' or PAF, the nine Indo-European forms in (72) and (73) above could be represented as:

*(76)*

$$(\mathrm{w})\underset{\mathrm{u}}{\begin{pmatrix}\mathrm{o}\\ \bar{\mathrm{e}}\\ \mathrm{e}\end{pmatrix}}(\mathrm{n})\mathrm{d}\text{-}\begin{pmatrix}\bar{\mathrm{a}}\text{-}\\ \bar{\mathrm{o}}\mathrm{r}\\ \mathrm{o}\\ \mathrm{r}\begin{pmatrix}\mathrm{o}\\ \bar{\mathrm{a}}\end{pmatrix}\\ \mathrm{skio}\end{pmatrix}$$

I stress that this is a notational device; options like 'wu', which seem to be possible given the structure above, are excluded as a matter of principle. Its primary purpose as presented by Matisoff is as a means for bringing together and listing a variety of known co-allofams. An example of such a PAF is shown below in (77).

*(77) Pan-allofamic formula for* tongue/lick *in Tibeto-Burman (Matisoff 1994:54)*

$$\begin{pmatrix} \mathrm{m} \\ \mathrm{s} \end{pmatrix} \mathrm{l}\,(\mathrm{y})\,\mathrm{a}\,(:) \begin{pmatrix} \mathrm{w} \\ \mathrm{y} \\ \mathrm{m} \\ \mathrm{k} \end{pmatrix}$$

Matisoff has used the formalism to group 'some 30 putatively distinct proto-roots (most of them [at the time] new)' having to do with limbs of the body. (Matisoff 1985:422). The lion's share of these roots falls under the first PAF he cites, given in (78)(a) below. Note that this PAF is a disjunction of two similar PAFs, and that the two terms of the disjunction in fact differ only in the arrangement (paradigmatic or syntagmatic) of the initial element. The individual allofams (shown in (78)(b) are then listed and discussed separately.

*(78)(a)*

$$\begin{bmatrix} \mathrm{d-} \\ \mathrm{g-} \\ \mathrm{p-} \end{bmatrix} \begin{matrix} \mathrm{l} \\ \mathrm{y} \end{matrix}\, \mathrm{a\,k} \qquad \underline{\mathrm{or}} \qquad \begin{bmatrix} \mathrm{d-} \\ \mathrm{g-} \\ \mathrm{p-} \end{bmatrix} \mathrm{l}^{\mathrm{y}}\, \mathrm{a\,k}$$

*(78)(b)*

1.1 simple (unprefixed) sonorant initial.
1.11 *lak
1.12 *yak > *zak
1.121 *yak

1.2 with dental prefix.
1.21 *d-lak
1.22 *d-yak
1.23 *d-[]ak ('prefix preemption' of the initial)

1.3 with velar prefix.
1.31 *g-lak
1.32 *g-yak and *g-ya

1.4 with fused (affricated) initial.
1.41 *džak < $^{**}\begin{bmatrix}\text{d-}\\ \text{g-}\end{bmatrix}$y a k

1.5 forms from languages in which *hand* = *foot* + -k.[87]
1.6 with labial prefix *p-yak.

(Matisoff 1985:423-430)

In this case, more than half the possible forms 'licensed' by the PAF occur.

The same type of variation occurs in Bantu languages; an examination of Guthrie's reconstructions reveals several roots which should be brought together into a single word family, such as the list supplied in (79). The fact that most of these are numbered sequentially indicates that Guthrie certainly noticed that these are related, but he gave no formalism for showing how they can be combined. In fact, several of these may merely be transcriptional variants. Even so, somehow their differences must be reconciled, and as efficiently as possible since there are nearly 4,000 of them.

*(79)* *Several Bantu etyma (after Guthrie 1967)*

| Etymon | Protogloss | Source | Source ID |
|---|---|---|---|
| -kín- | dance; gambol | G1967.CB | 1063 |
| kén | play | G1967.CB | ps.291 |
| -kínà | dance | G1967.CB | 1064 |
| kìnd | dance | G1967.CB | 1065 1/2 |

For purposes of etymologization, especially at the early stages, it may be unclear which precise reconstruction is the ancestor of a particular form. At the CBOLD project, the reconciliation of differing existing reconstructions is being carried out in a similar way. The several forms above, which could be represented by the following PAF:

*(80)* *A possible PAF for the Bantu reconstructions above*

$$*k\begin{Bmatrix}\varepsilon\\ e\end{Bmatrix}n\begin{Bmatrix}a\\ d\end{Bmatrix}$$

are grouped using a phonological and semantic *handle* or key. The handle is a bipartite key composed of a standardized gloss and a simplified version of the reconstruction. Handle-making processes were discussed in §5.1.4, for example; the handle for each of these reconstructions is

*(81)* DANCE.ken

The handle has several uses besides acting as a temporary catch-all during the reconstruction process. It can be used to sort the reconstructions and associated etyma and supporting forms into semantic sets (e.g., to bring together all the etyma meaning something like *dance*), or as a means to bring together reconstructions with a similar shape (etyma that resemble *ken*). But how would we determine that *gambol, play,* and *dance* might all go together? This problem is addressed in the next section.

### 3.2.2. Representing meaning

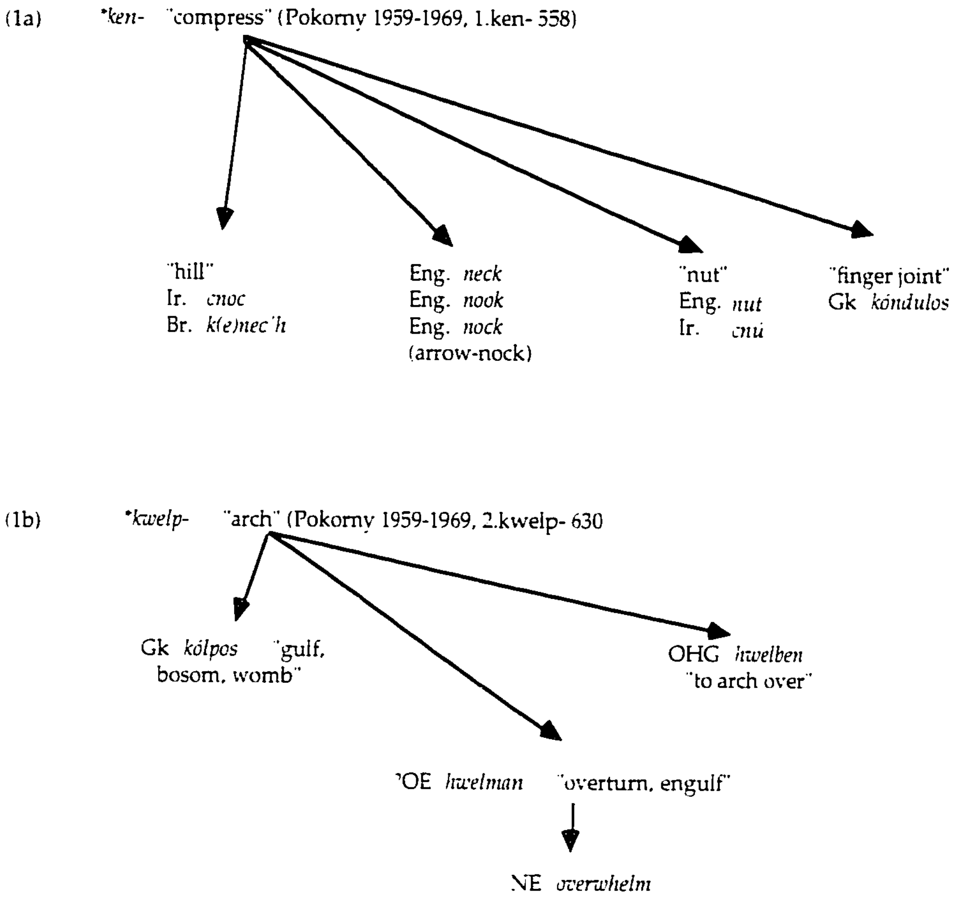
Before any computational analysis of the meanings of linguistic forms can be carried out, some representation of that meaning must be chosen. From an information-theoretical point of view, the particular representation chosen may not be so terribly important: the symbols used in a representation convey a *message*. And 'frequently the messages have meaning: that is they refer to or are correlated according to some system with physical or conceptual entities.' (Shannon 1949:31) As long as the correlation of the symbols with elements and subelements in the language studied conveys the correct meaning to the researchers (users), any notation will do. Whether the message (and associated meaning) is signaled by a single character (as in ideographic scripts), sequences of simpler characters (as in various Roman scripts), or in the form of on/off sequences (as in Morse code and computer-to-computer communications) is not important as long as the conventions of interpretation are understood. From both a linguistic and computational point of view, however, representations have consequences. And when data from many sources is compared, problems of divergent representations become more important. The complexity of comparison and the possibility of error increase rapidly as the number of different representations and the details they encode increase. In dictionaries and other lexicographic works meanings are expressed by words and phrases ('strings' in computer parlance) called glosses or definitions. I will tend to use the word gloss here to convey the more concise (and multilingual) notion of a precise meaning which is attached to a particular form. Glosses are constrained by the syntax and semantics of the glossing languages. But before taking up the issue of glosses themselves as data (i.e. as messages which should convey a particular meaning), I must say something about the concept of *meaning* and its use in historical reconstruction.

A notorious problem in reconstruction is setting limits on the degree and nature of semantic similarity required to establish cognacy. A related problem is establishing the precise meanings of reconstructed forms. Scholars have noted that often it is difficult if not impossible to assign a precise meaning to a reconstruction. Sweetser (1990) analyzes this problem in terms of semantic features.

> Semantics is limited only by our capacity for meaning, i.e. by our cognitive capacity, which is dauntingly ill-understood in comparison to the physical limits of the vocal tract. And yet lexical semantics and semantic change have frequently been analyzed as

> based on groupings of features, the semantic analogues of phonological distinctive features. Semantic feature-analyses and feature-based etymologies such as those in [82a] and [82b][reproduced below] abound in the literature. In these etymologies, the supposed common semantic feature of the descendent words is the compressed state or arched shape; this feature is viewed as being retained by the descendent lexemes, while other features are added or dropped. [...] The resulting proto-meaning thus becomes a sort of 'lowest common denominator' of the descendent meanings. (Sweetser 1990:24)

*(82) Two 'feature-based' etymologies (after Sweetser 1990:24)*


(1a)
*ken- "compress" (Pokorny 1959-1969, 1.ken- 558)
"hill"
Ir. *cnoc*
Br. *k(e)nec'h*
Eng. *neck*
Eng. *nook*
Eng. *nock*
(arrow-nock)
"nut"
Eng. *nut*
Ir. *cnú*
"finger joint"
Gk *kóndulos*
(1b)
*kwelp- "arch" (Pokorny 1959-1969, 2.kwelp- 630
Gk *kólpos* "gulf, bosom, womb"
OHG *hwelben* "to arch over"
?OE *hwelman* "overturn, engulf"
NE *overwhelm*

While noting that 'there has been some excellent work in historical semantics, often by researchers whose thorough knowledge of the older Indo-European languages and good 'feel' for word usage have enabled them to establish intuitively satisfying etymologies in cases where the descendent words would never have had a common denominator of feature,' Sweetser nevertheless questions the 'whole corpus of received etymological research' because, she says, 'we have little or no idea of what constitutes a reasonable semantic reconstruction, and are only starting to be aware of what regularities may be generally observable in semantic change' (Sweetser 1990:26).

This is a valid criticism, and until our understanding of semantics, especially diachronic semantics, improves, we are probably stuck with what we have. Computers are unlikely to help. And so, despite the above caveats, the method of semantic decomposition criticized above still provides, *as a heuristic*, an initial basis for grouping words together by meaning and justifying their inclusion in cognate sets, and it is thus likely to be used for years and perhaps centuries to come. Furthermore, the method is amenable to computational applications and especially to certain problems of dealing with large corpora, some of which are discussed below. And, as I will show, some of these techniques will at least help us reasonably defer the decision as to whether, for example, the 'historical prior' sense of a particular root (**deru)* meant 'tree/oak' or 'strong/trustworthy' until we can learn enough to make a good choice. (Sweetser 1990:26)

### 3.2.2.1. Using synonym sets in diachronic research

A common first step in comparative reconstruction is the arrangement of forms from different languages into *synonym sets* (cf. §5.1.7 above and (83) below). Scanning the set, it is easy to find forms which have a similar shape and are therefore likely to be related (either by borrowing or inheritance). The apparent correspondence can be extracted and, by comparing these correspondences with those found in other synonym sets, it is possible to begin to establish possible cognate sets. Several factors contribute to reducing the efficacy of this approach.

*(83) BACK (from ZMYYC set 258, Dai Qingxia 1992)*

| | | | |
|---|---|---|---|
| Written Tibetan | sgal pa | Bla-brang Tibet | hga wa |
| Sde-dge Tibetan | ge$^{13}$ba$^{53}$ | Cuona Monpa | phuŋ$^{53}$ |
| Zeku Tibetan (Amdo) | rga wa | Mawo Qiang | ɹa sta |
| Motuo Monpa | tshiŋ | Taoba Pumi | do$^{55}$ |
| Taoping Qiang | de$^{241}$ | rGyarong | tɐ zgən |
| Jinghua Pumi | do$^{13}$ | Ergong | qo mn̥o |
| Muya (=Minyak) | khø$^{35}$dyi$^{53}$ | Queyu | gõ$^{35}$ |
| Guiqiong | gi$^{35}$kuɔ̃$^{53}$ | Ersu | ga$^{33}$ma$^{55}$ |
| Namuyi | dʐu$^{33}$tsɛ$^{33}$tsɛ$^{33}$ | Shixing | ʁua$^{55}$sɿ$^{33}$pũ$^{35}$ |
| Xide Yi | kɯ$^{21}$tɯ$^{21}$ | Dafang Yi | bɯ$^{21}$gɯ$^{33}$ |
| Nanjian Yi | kɑ$^{21}$tu$^{33}$ | Nanhua Yi | gɯ$^{33}$di$^{21}$mo$^{33}$ |
| Mile Yi (=Axi) | khʌ$^{33}$ko$^{33}$dɯ$^{21}$ | Mojiang Yi | gɯ$^{21}$ɣɯ$^{33}$ |
| Lisu | kɑ$^{44}$tɛ$^{44}$ | Lijiang Naxi | gɯ$^{33}$tsɯ$^{33}$ |
| Yongning Naxi | gv$^{33}$dv$^{33}$ | Caiyuan Hani | ɔ$^{31}$mu$^{55}$ |
| Dazhai Hani | dɔ$^{55}$the$^{55}$ | Shuikui Hani | tu$^{31}$xu$^{55}$ |
| Lahu | tsɔ$^{21}$gɔ$^{53}$ | Jinuo | tə$^{42}$tshə$^{42}$ |
| Dali Bai | to$^{31}$ko$^{31}$ɕi$^{35}$ | Jianchuan Bai | to$^{42}$ko$^{42}$tɕi$^{55}$phiɛ$^{55}$ |
| Bijiang Bai | do$^{42}$pẽ$^{42}$ | Tujia | pɯi$^{55}$ |
| Written Burmese | kjɔ$^{3}$ | Spoken Rangoon | tɕɔ$^{55}$ |
| Achang | xa$^{31}$luŋ$^{35}$ | Zaiwa (=Atsi) | nuŋ$^{51}$kuŋ$^{51}$ |
| Langsu (=Maru) | kauŋ$^{31}$tɔ̃$^{55}$ | Anong | ɖɛ$^{31}$guŋ$^{31}$ʈhɑŋ$^{55}$ |
| Nusu | * | Dulong | gɔŋ$^{55}$ɹiʔ$^{55}$ |
| Jingpo | ʃin$^{33}$ma$^{33}$ | Geman (=Kaman) | glɑ̆u$^{53}$ |
| Darang (=Taraon) | plɯn$^{53}$ | Idu | plɯn$^{53}$ |
| Bokar Adi | lam ko | Sulong (=Sulung) | kə$^{33}$tse$^{53}$ |
| Lhasa Tibetan | kɛ$^{15}$pa$^{53}$ | | |

Several different etyma may be represented in a synonym set. Some of these distinct etyma may in fact be related (members of 'word families', or, to use Matisoff's terminology 'co-allofams'). (In the set given in (83) for example, reflexes of *TB *gyo⪤dzo (STC 216), *s-nuŋ ⪤ s-nuk (STC 203), *ko

(STC 211) may be found.) The correspondence patterns exhibited may therefore be of the 'partially regular' variety (noted in §6.6.2).

Due to *semantic shift,* (discussed in §5) an expected (and existing) regular reflex may not be included in the synonym set. The inclusion of a particular word in a particular synonym set often has as much to do with idiosyncrasies of the glossing language(s) as it does with the semantics of the languages being grouped.

Certainly any obvious correspondences will stand out. However, phonologically unusual correspondences (e.g., PIE **dwo* and Armenian *erku* 'two') will not and may be overlooked.

Borrowings from neighboring or prestige languages may (initially) convey a falsely uniform impression of regular correspondence that is not borne out by comparison with other sets—as in example (84) below, where *nari* is a borrowing from Indo-European. Where languages share the an *identical or nearly identical* form for a word, we may suspect borrowing. While this is a useful fact for the etymologist to identify, it does little to help find the language's actual cognate for the given etymon, which may have 'migrated' to another synonym set or become hidden in more obscure forms.

*(84) Forms meaning* wrist, pulse, *and* heart *in several Himalayish languages*

| | | | | |
|---|---|---|---|---|
| Chantyal | nari | *wrist* | NPB-CHANQ 6.1.6 | [127671] |
| Chepang (Eastern) | nari | *pulse* | RC-CHEPQ 9.3.2 | [128082] |
| Kham | nari | *wrist* | DNW-KHAM1 | [12860] |
| Newari | nari | *forearm / lower arm* | SH-KNw 6.1.5 | [62242] |
| Newari | nari | *pulse* | SH-KNw 9.3.2 | [62347] |
| Newari | nari | *wrist* | SH-KNw 6.1.6 | [62243] |
| Chantyal | nari bɦətki-wa | *pulse* | NPB-CHANQ 9.3.2 | [127771] |
| Limbu | nari∘bet | *heart* | JAM-VSTB | [144561] |

The notion of 'synonymy', moreover, is a subtle and complex one. Often synonym sets overlap considerably. For example, in (84) above the words given for *wrist* and *pulse* are identical, perhaps because of the way the forms were elicited (the idea of the pulse being in the wrist is not a semantic universal). Without other evidence, it is impossible to establish whether:

- in Newari there is truly no distinction between the two; or

- the elicitation failed to retrieve the distinction.[88]

This is a general methodological problem having to do with the use of 'second-hand' sources (i.e. sources not specifically prepared by the researcher), and certainly not unique to a computerized environment. However, the fact of using a large computerized database virtually guarantees that some of the data will be second-hand, thus exacerbating the problem considerably.

When creating synonym sets from different sources, the linguist must decide which words from the target languages are to be included together in a set. This may not be immediately evident from the glosses in the various sources. When aggregating words into a synonym set, the linguist is thus making an explicit equation between the glosses of words in different sources. Unless the precise glosses are supplied for the forms listed (and encyclopedic listings of synonyms sets usually do not), the metaglosses (i.e. meanings assigned to the synonym sets) convey an equivalence of meaning that may be quite spurious. Thus, words meaning *moon* or *month* in TB languages (shown in (85) below) are usually related, just as the words are in English. However, words meaning *cloud* or *cloudy* arc hardly so closely related, since many TB languages express the notion cloudy with words meaning *sky-dark*, *sunless*, etc.

*(85) Four Synonym Sets from the ZMYYC (Dai Qingxia 1992)*

| Language | CLOUD (6) | CLOUDY (836) | MOON (3) | MONTH (74) |
|---|---|---|---|---|
| Written Tibetan | sprin pa | thibs po | zla ba | zla ba |
| Sde-dge Tibetan | tʂin$^{55}$ | si$^{55}$xu$^{53}$ | da$^{13}$wa$^{53}$ | da$^{13}$wa$^{53}$ |
| Zeku Tibetan | ʂən | ɣnam rəp | rda wa | rdza |
| Motuo Monpa | muk pa | * | la ɳ̥i | la ɳ̥i: da wa |
| Taoping Qiang | χde$^{33}$ | də$^{33}$ | ɕy$^{33}$ɕya$^{55}$ | ʂʅ$^{33}$ |
| Jinghua Pumi | sdĩ$^{55}$ | khə$^{13}$sdĩ$^{55}$ | ɬi$^{55}$ | ʒ̥i$^{13}$ |
| Muya (=Minyak) | ndɯ$^{33}$ʐe$^{35}$ | ŋgɯ$^{55}$di$^{35}$ | lɛ$^{35}$nɯ̱$^{35}$ | dɛ$^{55}$wɛ$^{55}$ |
| Guiqiong | ʐɔ̃$^{35}$kuɛ̃$^{35}$ | xɔ̃$^{35}$ | li$^{35}$mo$^{33}$ | li$^{53}$ |
| Namuyi | tʂu$^{33}$ | nɛ$^{33}$ | ɬi$^{55}$mi$^{55}$ | ɬi$^{55}$ |
| Xide Yi | m̩(u)$^{33}$ti$^{33}$ | m(u)$^{33}$ŋo$^{33}$ | ɬo$^{21}$bo$^{21}$ | m̩(u)$^{33}$ɬɯ$^{33}$ |
| Nanjian Yi | ɑ$^{55}$m̩(u̱)$^{21}$ti$^{55}$ | ti$^{55}$ | xɑ$^{33}$bɑ$^{33}$ | xɑ$^{33}$bɑ$^{33}$ |
| Mile Yi (=Axi) | tɛ$^{33}$ | mu$^{21}$dɯ̱$^{33}$ | ɬo$^{33}$bo$^{33}$ | ɬo$^{33}$ |
| Lisu | mu$^{44}$ku$^{55}$ | nɛ$^{44}$mɯ$^{35}$ | hɑ$^{33}$bɑ$^{33}$ | hɑ$^{33}$ |
| Yongning Naxi | tɕi$^{33}$ | mv$^{33}$tɕi$^{55}$ | ɬe$^{33}$mi$^{33}$ | ɬe$^{33}$ |
| Dazhai Hani | dzo$^{31}$xø$^{31}$ | tshø$^{31}$ | ba$^{33}$la$^{33}$ | ba$^{33}$la$^{33}$ |
| Lahu | mo$^{31}$ | nᴀ$^{54}$ | xᴀ$^{33}$pᴀ$^{33}$ | xᴀ$^{33}$pᴀ$^{33}$ |
| Dali Bai | ŋv$^{21}$ | vu$^{42}$ | mi$^{55}$uɑ$^{44}$ | ʊɑ$^{44}$ |
| Bijiang Bai | mɯ$^{21}$ko$^{42}$ | ue$^{44}$ | ɳ̥u$^{55}$ŋu$^{55}$ | ŋʊɑ$^{44}$ |
| Written Burmese | tim$^{2}$ | mo$^{3}$uṁ$^{1}$ | lɑ$^{1}$ | lɑ$^{1}$ |
| Achang | xaŋ$^{31}$tɕin$^{31}$ | tʂau$^{35}$ | phă$^{31}$lɔʔ$^{31}$ | pau$^{51}$l̥ɔ$^{35}$ |
| Langsu (=Maru) | tʃa̱m$^{31}$thɔi$^{35}$ | tap$^{55}$ | lɔ̱$^{55}$ | lɔ̱$^{55}$ |
| Nusu | tʂhuẽ$^{31}$mɔ$^{55}$ | tɕy$^{53}$ɑ$^{31}$ | ɬɑ$^{31}$ | ɬɑ$^{31}$ |
| Jingpo | să$^{33}$mui$^{33}$ | muŋ$^{33}$ | ʃă$^{33}$ta̱$^{33}$ | ʃă$^{33}$ta̱$^{33}$ |
| Darang (=Taraon) | ɑ$^{31}$m̩$^{55}$ | bɯm$^{55}$ | xa$^{55}$lo$^{55}$ | xɑ$^{55}$lo$^{55}$ |
| Bokar Adi | dɑŋ muk | * | poŋ lo | poŋ lo |
| Lhasa Tibetan | tʂĩ$^{55}$pa$^{53}$ | (nam$^{55}$)thip$^{53}$ | ta$^{13}$wa$^{13}$ | ta$^{13}$wa$^{13}$ |
| Bla-brang Tibet | ʂən | hman thəp | hda wa | hda |
| Cuona Monpa | sᴀ$^{55}$cᴀʔ$^{53}$ | lup$^{13}$ | lɛ:$^{55}$thøn$^{55}$ | le$^{53}$ |
| Mawo Qiang | zdɤm | mə mɑ ʂqɑ | tʃhə ʂa | ʂə |
| Taoba Pumi | zə$^{55}$rẽ$^{55}$ | mə$^{55}$dʐɐ$^{35}$mə$^{53}$ | ɬi$^{55}$ | zi$^{35}$ |
| rGyarong | zdᴇm | kə jam kə khi | tsə la | tsə la |
| Ergong | zdo mɛ | zdo ʐɛ | ɬɯ va | ɬɯ |
| Queyu | xu$^{55}$pa$^{53}$ | kə$^{35}$mu$^{53}$ | ɬo$^{55}$ɳ̥u$^{53}$ | lø$^{53}$ |
| Ersu | tsɛ$^{55}$ | mɛ$^{33}$ŋa$^{55}$ | ɬɑ$^{55}$phɛ$^{55}$ | ɬɑ$^{55}$ |
| Shixing | tɕi$^{55}$rɑ̃$^{33}$ | mɐ$^{33}$zɑ$^{55}$ | ɬi$^{33}$mi$^{55}$ | ɬɯ$^{55}$ |
| Dafang Yi | tie$^{33}$ | dɯ$^{21}$ | ho$^{21}$bo$^{21}$ | ho$^{21}$ |

| | | | | |
|---|---|---|---|---|
| Nanhua Yi | ti$^{33}$ | dɿ$^{21}$ | ɕo$^{33}$bo$^{33}$ | ɕo$^{33}$ |
| Mojiang Yi | tɛ$^{55}$ | dɯ$^{33}$ | xo$^{21}$bo$^{21}$ | xo$^{21}$ |
| Lijiang Naxi | tɕi$^{31}$ | mɯ$^{33}$lv$^{33}$lv$^{33}$ | xe$^{33}$me$^{33}$ | xe$^{33}$ |
| Caiyuan Hani | nɿ$^{31}$tshi$^{31}$ | na̱$^{33}$ | pɔ$^{33}$lɔ$^{33}$ | lɔ$^{33}$ |
| Shuikui Hani | u$^{31}$tu$^{55}$ | na$^{33}$ | pɔ$^{33}$l̥ɔ$^{33}$ | pɔ$^{33}$l̥ɔ$^{33}$ |
| Jinuo | mɯ$^{33}$tjɐ$^{33}$ | xo$^{42}$ | pu$^{33}$ɬa$^{44}$ | ɬɔ$^{33}$ |
| Jianchuan Bai | ŋv$^{21}$ | ŋv$^{55}$ | mi$^{55}$ŋuɑ$^{44}$ | ŋʊɑ$^{44}$ |
| Tujia | mɯe$^{35}$la$^{55}$ɣoŋ$^{21}$ | jin$^{55}$ | su$^{21}$su$^{21}$ | jie$^{55}$ |
| Spoken Rangoon | tẽ$^{22}$ | mo$^{55}$õ$^{53}$ | lɑ$^{53}$ | lɑ$^{53}$ |
| Zaiwa (=Atsi) | mṵt$^{55}$mau$^{55}$ | tsau$^{55}$ | lɔ̆$^{55}$mo$^{55}$ | lɔ̆$^{55}$mo$^{55}$ |
| Anong | io$^{31}$mɯn$^{55}$ | muʔ$^{53}$ɖɯʔ$^{53}$ | sɿ$^{31}$lɑ$^{55}$ | sɿ$^{31}$lɑ$^{55}$ |
| Dulong | ɹɯ$^{31}$mɯ̆t$^{55}$ | nɑ̆m$^{53}$dɯ$^{55}$ | sɯ$^{31}$lɑ$^{55}$ | sɯ$^{31}$lɑ$^{55}$ |
| Geman (=Kaman) | kɑ$^{55}$mɑ̆i$^{35}$ | ȵɑuŋ$^{31}$nɯm$^{53}$ | lɑi$^{53}$ | lɑi$^{53}$ |
| Idu | ɑ$^{55}$mu$^{55}$ | boŋ$^{35}$ | e$^{55}$lɑ$^{55}$ | e$^{55}$lɑ$^{55}$ |
| Sulong (=Sulung) | kə$^{33}$tɯ$^{33}$ | lɯ$^{33}$ | aŋ$^{33}$bo$^{33}$ | aŋ$^{33}$bo$^{33}$ |

When researching the data contained in such sets, it will be useful to recognize that in the glossing metalanguage the gloss given for a synonym set can be related, morphologically or otherwise, to other glosses; it is important to remember caveats above and to the extent possible control for chance synonymy (or the lack of it) affecting the comparison. Several techniques for analyzing and representing synchronic and diachronic semantic relationships are presented in §5.

### 3.2.3. Representing data sources

This important link in the chain of scholarly research need not concern us much here; maintaining a link between the data as it undergoes analysis and its origin is mainly a problem of recording with each extracted data element a pointer to the original work and a pointer to the exact spot in the original work. It is only a little different from procedures for citing works in published papers, but there are a few points worth noting.

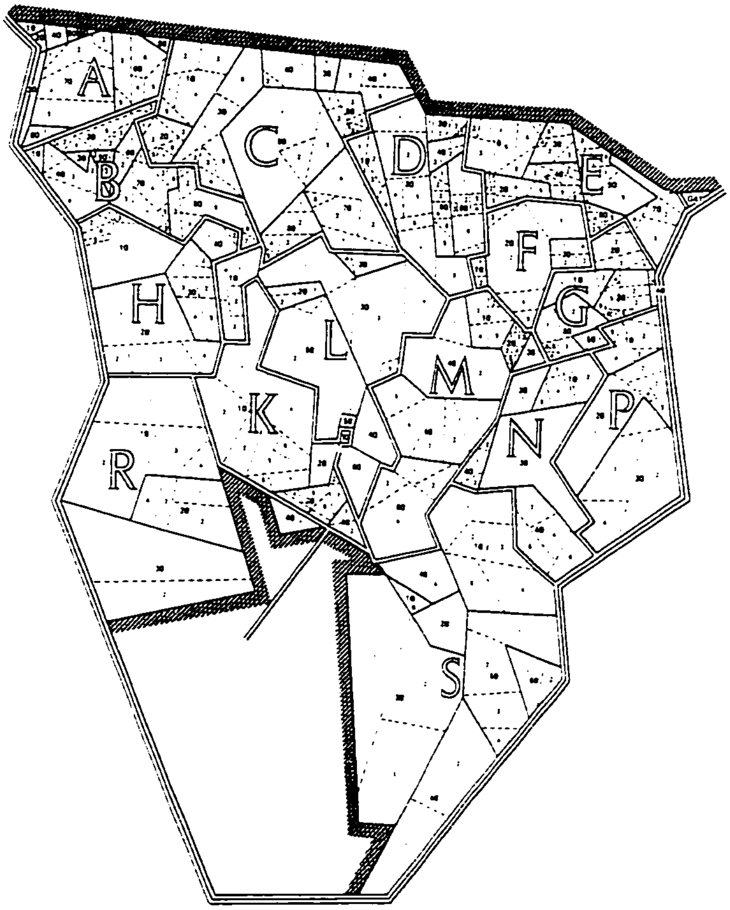
It is a good idea to have a short and mnemonic pointer to the source; while the pointer could simply be a number, in practice this is not a good idea because human users respond to and remember a sequence of letters better than a sequence of numbers. A key like that used for online bibliographic systems works well. A key built from the first letter or letters of the author's names combined with a key built from the first few letters of the first few words of the title of the work and the last digit or two of the year of publication serves to uniquely identify almost all the books in the Library of Congress. There are many more letters than numbers so a short alphabetic key contains many more possibilities than a numeric key of the same length. Very common or important works deserve their own short keys though their construction violates the conventions in use. Thus, at STEDT, Paul K. Benedict's seminal work, *Sino-Tibetan: a conspectus* is known almost exclusively by the abbreviation *STC*, a shortening of the expected key *PKB-STC*.

Since forms cited in one work may in fact be originally drawn from other works, it is a good idea to consider how to represent the chain of use of particular forms, as it is not uncommon for the transcription and meaning to change as it is cited. Exigencies of typography and space may result in changes with unforeseen consequences. As an example, the STC cited in the last paragraph, cites its Burmese forms from Judson 1921, 1986 (using an idiosyncratic system for retranscribing the tone) and its Tibetan forms from Jäschke 1881, two of the standard Western works for these languages. Since eventually the ultimate sources may be processed and compared, it is important to know when two forms, ostensibly from different sources, are really one and the same. In a large cross-linguistic database it is not unusual for certain 'touchstone' forms and parade examples to be cited many times in different works.

### 3.2.4. Representing language varieties

It is practically axiomatic in linguistics that the distinction between language and dialect is a matter of social or political convenience.[89] For comparative work, the linguist needs sometimes to mask small dialect differences and sometimes to highlight them. The well-known processes of 'lumping' and 'splitting' must therefore be incorporated into the computer representation as well. When a large number of languages and dialec:s are gathered together, the problem of identifying and distinguishing particular varieties can become especially challenging. Several conventions are available for making the appropriate distinctions. Some linguistic distinctions are best visualized cartographically. The *linguistic atlas* provides one means of graphically and directly representing the distribution of language varieties.

*(86) The distribution of Bantu languages according to (Guthrie 1967)*



The cartographic model breaks down, however, in a number of cases:

- where linguistic areas overlap, or where speakers of different languages inhabit the same region.

- where the population is sparse or there are isolated 'pockets' of speakers, for example in individual villages. This is the case in Eastern Nepal, for example, where a village or group of villages has its own specific linguistic identity distinct from others in the area. This gives the distribution of 'hill' languages a speckled appearance compared to languages spoken in more populated areas. The map of the Tamang speaking area in Appendix 4.2 illustrates the 'point source' interpretation of a language area which contrast with the more areal interpretation given for African languages above.

Another way of denoting distinctions in language varieties involves a traditional hierarchical classification similar to the Linnaean system of biological classification. Language varieties are named first of all according to their genetic or areal affiliation in a linguistic phylum (e.g. Tibetan, English), and then subcategorized according to location, ethnolinguistic group, or degree of conventionality (Written Tibetan, RP English, Cockney). Especially when reporting the results of field work or discussing less-described language varieties, it is important to encode such details.

Substantial ethnolinguistic and descriptive insight is required to identify and code the important distinctions among languages, and a good deal of ink may be required to make the distinctions clear. Consider, for example, Marrison's description of the names for Naga languages:

> The nomenclature of the Naga tribes is complex. The tribes themselves are much sub-divided; but apart from this, in many

> cases there are alternative names, as well as alternative spellings of the same name. When the Nagas were first described, it was usually an outsider's name for a particular tribe which was used; the tribe's own name for itself often was not known till later.
>
> In reference to language, especially in the reports made in the 19th century, it is often the name of the village, rather than that of the tribe or sub-tribe, which is given. This arose from a need to provide some means of identification; but it may be justified by the fact that nearly every village has its own variety of speech.
>
> Different names have been applied to the same tribes or other groups at different times.' (Marrison 1967377)

In his *Directory of Tibeto-Burman languages* (Matisoff 1986), Matisoff systematizes the 'nomenclatural complexity' of languages, and his system is summarized here, as it will be used later when discussing the forms of language names. ETHNONYMS (people-names), Matisoff notes, are often used as GLOSSONYMS (language-names). *Laitong*, for example, refers both to a tribe of Tripura and to the dialect of Tripuri that they speak (Karapurkar 1972). But sometimes the correspondence is not one-to-one. The name *Kham* refers to a language of Nepal spoken by Magars of the Bhuda, Gharti, Pun, and Rokha subtribes (Watters 1975). On the other hand, *Kheja*, *Kheza*, and *Khezha* are merely alternative spellings (CO-ALLOGRAMS) of a single name referring to a certain Naga group.

Matisoff distinguishes genuinely different names for the same people/language—ALLONYMS—and merely different spellings or pronunciations of the same name—ALLOGRAMS. *Hsi-hsia* (the Wade-Giles transcription; the pinyin, another allogram, is *Xixia*) is the name given by Chinese and Japanese scholars to a certain extinct Tibeto-Burman language,[90] while Russian writers use a different allonym, *Tangut*. Allonyms can be further subdivided into autonyms and exonyms; the AUTONYM (self-name) for a given group is apt to be totally different from the EXONYM[91] (outsiders' name) that others use for them. Thus, the tribes of Tripura refer to their language as *Kok-borok* , literally 'speech of men.'[92] Such egocentric names are hardly likely to be adopted exonymically by neighboring groups. Exonyms are not infrequently pejorative. The Greeks referred to non-Greeks as *barbaroi* ('foreign, rude', originally 'stutterers'), and Sanskrit speakers referred to outsiders as *mleccha*, 'any person who does not speak Sanskrit' (Monier-Williams :837). The term *Lolo,* used for the Yi languages of China and Thailand, means *barbarian*. 'Yi' is itself pejorative in this way, but has be ameliorated by the use of a different Chinese homophonous character meaning 'sacrificial vessel'. (Bradley 1979, Matisoff 1988)

A great many ethno- and glosso-nyms are primarily or originally names of places (TOPONYMS ). Matisoff uses the term LOCONYM to refer specifically to 'a place-name that has been extended to serve as the name of a language or dialect.' Names of rivers are sometimes applied to ethnic/linguistic groups in Southeast Asia, e.g. the Hka-hku or 'up-river' Jinghpo of the upper Irrawaddy Valley; we may call such names POTAMONYMS . (See also Glover's Maiwa River branch of the Limbu subfamily (Glover 1974:11]).)

These nomenclatural conventions may be compounded: the Phom used to call themselves Chingmengnu (PALEOAUTONYM), while others called them Tamlu (PALEOEXONYM). Matisoff exemplifies some of the possibilities as in (87) below.

(87) *Part of the hierarchy of names*

| | | | |
|---|---|---|---|
| people | auto- | autoethnonym | Memi |
| village | auto- | autoloconym | Sopvoma |
| language | auto- | autoglossonym | Memi |
| people | exo- | exoethnonym | Mao |
| village | exo- | exoloconym | Mao |
| language | exo- | exoglossonym | Mao-Sopvoma. |

Some, perhaps all, of this information must be incorporated into the data structures and algorithms for handling the data.

Several major functions of naming conventions are discussed below.

First, the naming system should be responsive to different transcriptions of the same language variety, assigning (in some sense) a different 'language name' to each differently transcribed version. Because phonetically identical forms may be written in different transcriptions, the result can be an appearance of variation which may or may not be real (or significant), but which the nomenclature system should faithfully record.

(88) *Three transcriptions of Black Lahu (Lahu Na) forms from three sources (and from three different locales)*

| | Thailand | | Burma | | China | |
|---|---|---|---|---|---|---|
| Gloss | Matisoff 1972 | | Luce 1981 | | Dai Qingxia 1981 | |
| hand | làʔ-šɛ | 166 | la‸sheh: | U.33 | lᴀ$^{21}$sɛ$^{33}$ | 251 |
| eye | mɛ̂ʔ-šī | 145 | mehˆshi | U.31 | mɛ$^{54}$si$^{11}$ | 238 |
| pig | vàʔ | 168 | | | vᴀ$^{21}$ | 115 |

While in the above example we may recognize all the words as forms of a language called *Lahu Na,* we must distinguish them from each other based on the source of data, which was gathered at different times and places, and from different informants.

Another function of naming conventions is to equate different names for the same language variety. This 'lumping' function serves as a means of cross-referencing related languages. Different names may have been used for the same language variety.

The lumping may bring together languages in different ways. It can make it possible to group language varieties according to genetic or areal criteria. All dialects of a language group or subgroup, for example, could be listed under one heading as is shown in (89) below, where all Hani dialects in the STEDT database are listed together.

The names, abbreviations, and codes used in published works cannot simply be taken at face value. Dictionaries and thesauri are among the most telegraphic of texts. Selected information from the front matter of the work must therefore be incorporated into the database scheme to prevent confusion. So, for example, among languages in the STEDT database, the letter *K* is variously used in published sources to abbreviate the languages *Jingpho* (a.k.a. *Kachin*), *Karen*, and *Khàtú*; *T* may stand for *Tamang*, *Tibetan*, or *Tangkhul Naga*. Retaining the conventions used in the source document is useful for proofreading and maintaining fidelity to the original. However, to prevent confusion the designation used in the original is linked to three additional variant designations to facilitate searching and sorting. The four language designators in use at STEDT are:

(1) The LANGUAGE ABBREVIATION, which is the designation for the language used in the original document. Usually, but not always, this is an abbreviation. Sometimes (rarely) it was necessary to create such a designation when the source document failed to provide one.

(2) The STANDARDIZED LANGUAGE NAME, a designation created on the basis of orthographic conventions and aimed at providing a simple, unified, and precise way of referring to language varieties. It rectifies the transcription of language names so that they may be conveniently published in a single font (it is not usual for sources to refer to the name of the language in phonetic transcription or even in the native orthography).

(3) The LANGUAGE SORT KEY, a designation used for grouping different language varieties under a common rubric. This provides one type of 'lumping' designator.

(4) The LANGUAGE GROUP IDENTIFIER, a multipart numerical designator which locates the language variety in a hierarchical scheme for purposes of subgrouping.

These are described in more detail below.

The Standardized Language Name gives, for each language variety, a regularized form which conforms to predetermined conventions. These conventions specify the internal syntax of names and how they should be subcategorized. Given the variety of nomenclature in use, it is impossible (or at least unwise) to provide a strict syntax for names which would explicitly mark all distinctions in an explicit way. The language spoken in the city of Rangoon (now perhaps Yangoun) differs substantially from the early language, though the writing preserves the earlier pronunciations. To indicate this, it is common to refer to the written form as Written Burmese and to call the modern version by another name. Here the commonality ends: some authors call the language Modern Burmese (Rangoon dialect assumed as the standard); others call it Rangoon Burmese; and so on. To capture these distinctions, we might wish to indicate that the languages is 1) modern, 2) spoken, 3) from Rangoon, and 4) that it is a variety of Burmese. If all four features were to be indicated in the name for all languages in this database, a great deal of redundancy and complexity would be introduced, especially since some languages would be essentially 'underspecified' for these features, and for others this specification would be insufficient to distinguish them from other varieties (e.g., it does not include an indication of the register, for example). Instead, the format of this designator varies according to the need for subdividing language varieties and the need to maintain a more or less conventional way of referring to them. The internal syntax is given below in (89).

*(89) The internal syntax of Standardized Language Names at STEDT*

| *FORMAT* | *EXAMPLE* | *REMARKS* |
|---|---|---|
| LgName (Loconym) | Hani (Dazhai) | |
| | Burmese (Rangoon) | |
| LgName (Ethnonym) | Naga (Konyak) | |
| LgName (Orthography) | Burmese (Written) | not *Written Burmese* (sorts under B) |
| LgName (Variety/Dialect) | Gyarong (Eastern) | |
| | Nung (dialect) | dialect unspecified |
| LgName (Author's Name) | Lisu (Fraser) | |
| | Hani (Hu T'an) | |
| LgName (Title or Work) | Hani (Wordlist) | distinguishes transcription |
| LgName (= Allonym) | Hruso (=Aka) | used only to distinguish common allonyms |
| | Haka (=Laizo) | |

**Legend:**

LgName normally a major designation of language variety taken from 'shortlist'[93]

(Nym) subcategorizing element to distinguish this particular variety from other varieties.

Another designation, the Language Sort Key, is used to implement the lumping function. The name in this case is a shortened form of the most common designation for a whole group of language varieties. This designation is called the Language Sort Key since it allows the language varieties to be sorted together by major categories. An example of a number of language varieties sorted together under this rubric is shown in (90) below. Designators are usually drawn from the 'shortlist' mentioned above.

*(90) Versions of* HANI: *in STEDT database*

| SORT | STANDARDIZED | SOURCE OF | LANGUAGE |
|---|---|---|---|
| KEY | LANGUAGE NAME | DATA | ABBREVIATION |
| LGSORT | LANGUAGE NAME | SRCABBR | LGABBR |
| HANI | Hani | DQ-Hani | Hani |
| HANI | Hani | HU/DAI1964 | Hani |
| HANI | Hani | JAM-Ety | Hani |
| HANI | Hani | JAM-GSTC | Hani |
| HANI | Hani | KH-YHyc | Hani |
| HANI | Hani | PC | Hani |

| | | | |
|---|---|---|---|
| HANI | Hani (Caiyuan) | JZ-HNcy | HNcy |
| HANI | Hani (Caiyuan) | ZMYYC | Hani.Caiyuan30 |
| HANI | Hani (Dazai) | JZ-HNdz | HNdz |
| HANI | Hani (Dazai) | ZMYYC | Hani.Dazhai31 |
| HANI | Hani (Gao Huanian) | JAM-GSTC | Hani (Gao Huanian) |
| HANI | Hani (Gelanghe) | JZ-HNgl | HNgl |
| HANI | Hani (Hu T'an) | JAM-TSR | Ha [HT] |
| HANI | Hani (Kao Hua-Nien) | JAM-TSR | -Ha [K] |
| HANI | Hani (Kao Hua-Nien) | JAM-TSR | Ha [K] |
| HANI | Hani (Lüchun) | IH-PL3 | HaL |
| HANI | Hani (Shuikui) | DQ-Haoni | Haoni |
| HANI | Hani (Shuikui) | IH-PL1 | Hao |
| HANI | Hani (Shuikui) | ILH | Haoni |
| HANI | Hani (Shuikui) | JZ-HNsk | HNsk |
| HANI | Hani (Shuikui) | ZMYYC | Hani.Shuikui32 |
| HANI | Hani (Wordlist) | IH-PL3 | HaW |
| HANI | Haoni | see | HANI (SHUIKUI) |
| HANI | Pijo | IH-PL1 | P |
| HANI | Pijo | ZMYYC | Hani Pijo |

Note that several somewhat arbitrary choices have been made in choosing these language names: e.g. the name *Pijo* has been retained instead of *Hani Pijo*; *Haoni*, an exonym for the dialect of Hani spoken in the Shuikui region, is not used, although it is fairly common.

At STEDT, these names are stored in a separate database file, distinct from the lexical file. The two files are linked via a database relation based on the 'Source-Language binome', an identifier composed of the data source designator and the language abbreviation which uniquely identifies each language variety in the database.[94] Thus it is a simple matter to globally change the name and grouping of a particular language data set.

I hope to have demonstrated that converting one's data into machine-readable form requires some forethought, and in some cases, a great deal of planning. As the size of the data set gets larger, the number of exceptional cases increases. What seem like small problems or deficiencies in software and planning can easily turn into a substantial headache. Imagine coding all your data so that it can be represented in a particular font, and later finding that new data cannot be transcribed using it. Another problem which obsesses computer people but is rarely considered by linguists planning databases is performance. Many techniques which work well with small amounts of data fail when confront with a 'real' data set. It is well-known in computer circles that problems in 'scaling-up' must be explicitly accounted for in the design of any software system. In the next chapter I will show a few examples of conversion of printed sources into database-ready formats.

# 4. CONVERTING EXISTING SOURCES

## 4.1. Data gathering

Entry of machine-readable data may be accomplished by:

- typing it in oneself, which normally ensures that the representations will be useful to the typist/creator. This is usually a very slow process.

- having someone else type it in or accepting a machine-readable version provided to you by someone else, which implies accepting the coding and representational decisions of the creator(s).

- scanning and recognizing text that has already been printed, in which case one must ensure that the newly scanned representation and the original printed version resemble each other near-exactly. A second step will be the extraction or conversion of the data into a more usable form.

The following section discusses the ramifications of these different methods of data entry.

### 4.1.1. Scanning and converting existing sources

One must not assume that the conversion of a published work into a usable machine-readable form is ever a trivial task.[95] Unless the amount of data involved is small, the application of heuristics and a substantial amount of quality control (proofreading, etc.) will be required to ensure an accurate representation. The traditional paradigm for text data in the computer world is still ASCII, a small set of characters suitable for printing bank statements and telephone bills. More complex documents still present a maze of potential difficulties to computer users.

This fact seems to be lost on some computational linguists, who assume that someday (perhaps soon) machine techniques will surely be able to resolve many such problems routinely. The conversion process involves a number of steps, at each of which errors may be introduce or content lost. The prototypical case of conversion is the transformation of a printed page into an intelligent machine-readable form via scanning and optical character recognition. At least five steps are required:

- Scanning the document, which is the converts the page into an *image* of the page in the computer's memory. This representation (which is a machine-readable form of a sort) is useless for any linguistic purpose: the characters are not recognize as characters, but merely as bits indicating darker or lighter spots on the page.

- Optical character recognition (OCR), which converts the image representation into text. This is a difficult pattern recognition process and (as will be discussed below), and computers are still not very good at it. Especially if the characters are not members of the prim and proper ASCII set in a conventional typeface, the process is likely to be rife with errors.

Scanning and OCRing of linguistically interesting texts (i.e. texts which contain phonetic transcriptions, incorporate non-alphabetic scripts such as Chinese or Hittite, even subscripted letters and numerals) often results in an error-ridden version which requires a lot of editing.

- The correction process, in which the text from the recognized page is corrected (usually by hand) to match as closely as possible the content and format of the original. This is usually the most time-consuming part of the conversion process; though some heuristics can applied on a case-by-case basis to make the process more efficient, it normally requires a lot of human intervention. At this point, conversion needs diverge according to the ultimate use of the text. Some users do not need to retain the formatting information (such as indications of which characters were in italic or bold): it is sufficient if the string of characters is correct. Others, such as converters of dictionaries, need this information as it encodes important features of the work.

- The markup process, in which the text is transformed into a representation suitable for the analyses required. This process interprets implicit characteristics of the text (such as the significance of bold type or the fact that items are in a certain order) into explicit representations which can be used in the research. Again, depending on the exact application and the nature of the data, this step can be either trivially simple, or (and this is the more likely alternative), it will take almost as much time as the correction process to get it right.

Consider the following introduction to a recently published dissertation on computational phonology:

> Imagine that you are an archeologist working in the central plains of Asia. One day, while rooting around for fireplaces and flint shards in a cave, you happen across an ornately carved box. Inside the box, you find a collection of paper leaves bound between flat wooden plates.
>
> [Back] in your camp laboratory, you separate the leaves of the book, laying each of them on a white plate, then digitising them with your portable scanner. When you have them all recorded, you begin the analysis of their contents. First, you train your optical character recognition software to recognise the symbols on the digitised pages. All of the pages are then parsed into a binary symbolic representation. Given the small number of symbol types, it seems likely that the representation is phonemic, and broken up into words.
>
> The data is now in the right format for your linguistic analysis software. First you key in the instructions to initiate the phonology discovery program. Satisfied that you have specified the correct search controls, you leave the computer running while you go and eat your dinner.

> Some time later, you return to the computer and discover that the program has terminated. On the screen you find details of the computer's analysis of the language of the text. The analysis details the segmental categories which are significant for word structure, dividing the segments into two classes. It specifies the syllable structure in terms of a sonority hierarchy of six distinctive levels, describing the restrictions on segments occurring in the syllable codas. It informs you that the seven vowels participate in a harmony system with a single transparent vowel. It tells you... (Ellison 1994)

Ellison proposes to '[take] this vision of the machine learning of phonological structure some of the way from fantasy to reality' (Ellison 1994:3). However, some of the most fundamental and difficult problems are ignored completely. Consider the problem of identifying the characters in the text based on their visual appearance, the process called optical character recognition (described above). Ellison hypothesizes that the scanned text transparently resolves itself into phonemic chunks. However, if we look at the real world, the possibility of this happening (especially in Central Asia!) is rather small.

Even granting that the text represents some 'roman' type, a number of problems must be overcome before the output of the OCR process is usable as input to other processes. If the text, for example, were a photocopy of a printed piece of technical prose in modern English, even then the problem would not be simple:

(91) *Uncorrected version of a scanned and OCRed text (taken from a poor photocopy of Ellison's dissertation):*

*Orlthc othcr harld, liriguists rarely agree orl featurcl decompositions of segments. Couchirig data in a particular decornposition makes it uninteresting to those who disagrcc with the decorripositiclll, and ally lesults learned froril it, irrelevant So cipher-independence in the data leads mearis that the results of learning will be ltiore widely acceptablc, arld therefore more objective.*

The text is barely readable in this form; however, only a few substitutions are required to render it into legible English, shown in (92) below.[96]

(92) *The corrected text*

*On the other hand, linguists rarely agree on featural decompositions of segments. Couching data in a particular decomposition makes it uninteresting to those who disagree with the decomposition, and any results learned from it, irrelevant. So cipher-independence in the data leads means that the results of learning will be more widely acceptable, and therefore more objective.*

The following few corrections must be made to render the text into English:

*(93)*

e (and in one case a) are sometimes scanned as c
n is mis-scanned as orthographically similary sequences rl, ri, and ll
m is mis-scanned as rn, ril, lti, rti, and sometimes ris
periods and occasional blanks are missing
decomposition is mis-scanned as decorripositiclll

Note that these changes must be made in the correct places only, or new errors will result (for example, *m* cannot be substituted for *rn* in the word *learning*, which is already correct). Changes must thus be sensitive to lexical and morphological context. Once these changes are made, the text becomes quite clear. Note that in the case of a monolingual text in a *known* language, it is possible to apply heuristics such as spell-checking to the correction process. If the text is multilingual, however the efficacy of these heuristics is greatly reduced, and their application greatly complicated. In fairness, Ellison's claim is perhaps best portrayed as an attempt at computer-aided decipherment of an unknown language in an unknown script. Such decipherments are known to be the most intractable and difficult.

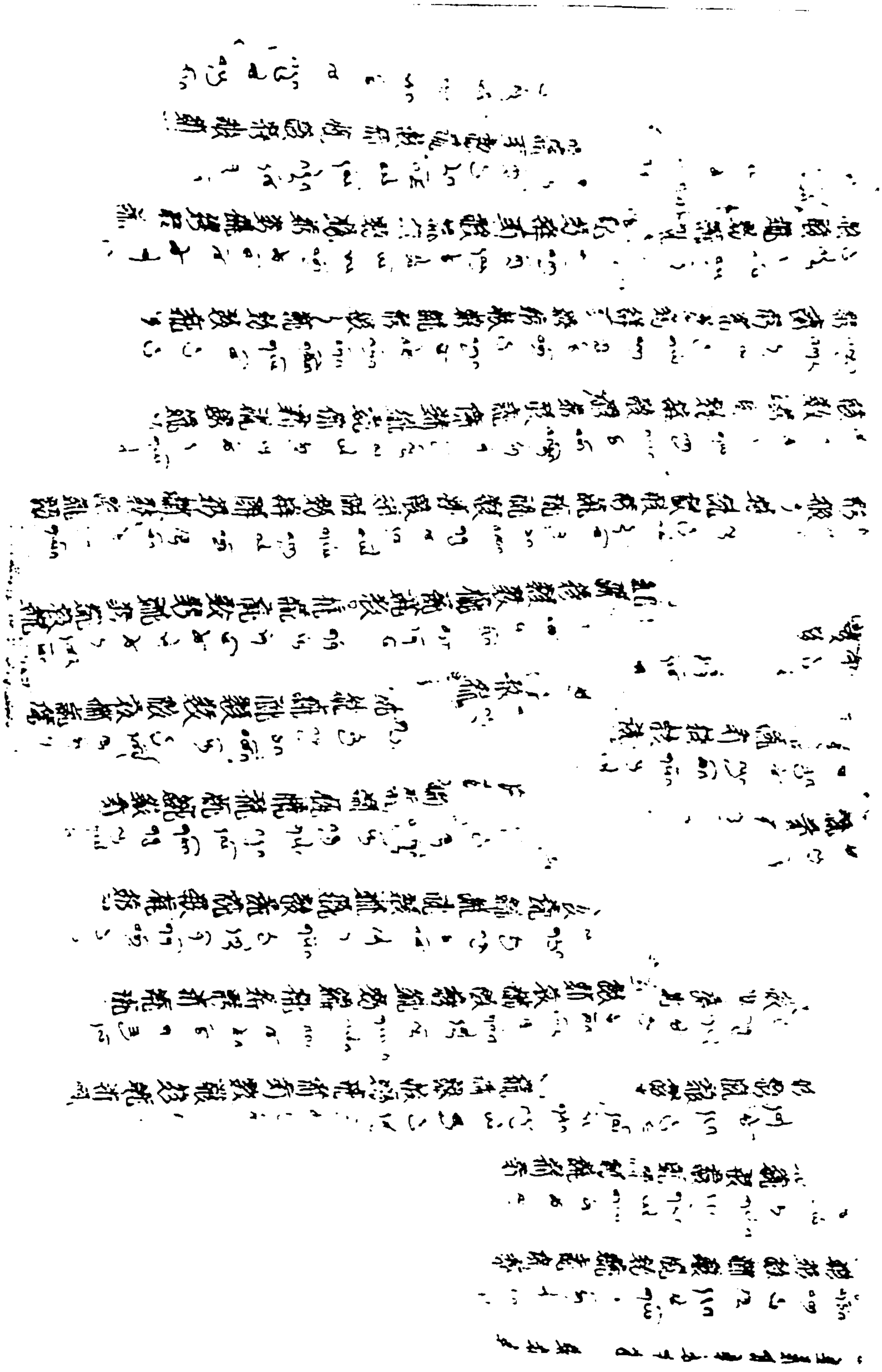
### 4.1.2. Tangut-Tibetan interlinear texts as a test of 'mechanical' scanning

Ellison specifies two goals for his research: 'Firstly, to show that algorithms for learning phonology, like the ones in the scenario, are possible, and not just a pipe dream. Secondly, to present a methodology for building algorithms that learn phonology' (Ellison 1994). Let us take Ellison's words at face value and examine the feasibility of his claims. His description of the hypothetical linguist's discovery (above) bears a curious resemblance to an actual event:

> ... in 1892 in St. Petersburg, [a Russian explorer] reported the existence of ruins of a Tangut city at a site named Khara Khoto, although [he] had never set eyes on these ruins himself. ... [In 1908] an expedition of the Imperial Russian Geographic Society into Inner Mongo*l*ia [discovered] Khara Khoto. Hidden inside a stūpa, an entire library of Tangut books and much Tangut art were unearthed and taken back to St. Petersburg. Today the Tangut manuscripts and wood-block prints are kept in the Manuscript Department of the Institute of Oriental Studies of the Russian Academy of Sciences. ... Sir Aurel Stein led a British expedition to the Khara Khoto ruins in 1914 and salvaged a smaller but very valuable collection of Tangut manuscripts, now kept at the Oriental and India Office Collections of the British Library on Blackfriars Road in London. (Driem 1991)

One of these texts found find Stein is reproduced in (94).

(94) *Tangut-Tibetan interlinear text (from Stein 1928:)*



### 4.1.3. Notes about Tangut-Tibetan text

The reader should note a few features of this text which would defeat any such attempt to perform the type of process Ellison proposes. First, two different languages are 'conflated.' These languages are only distantly related in the Tibeto-Burman family (Tangut is perhaps related to Qiangic, but this has yet to be established).[97] The scripts are of two completely unrelated types: Tangut is logographic (even 'hieroglyphic' (Laufer 1916), Tibetan is phonetic (with the graphically complex Nagari type of orthography presented in §3.2.1). Since it should fit the 'phonemic' paradigm Ellison presupposes, it at least might be optically recognizable. But would the scanner and OCR software be able to distinguish the Tangut from the Tibetan?

The content of the text is rather unusual: the Tangut characters transcribe syllable for syllable the Tibetan; that is, each Tangut character approximates the sound of the syllable in the base Tibetan text. What computer-based recognition scheme would be able to interpret this unusual relationship?

Finally, the characters are handwritten. Handwritten text is the final frontier in OCR and is still virtually unscannable (especially without *a priori* training). Scanning ideographic scripts (such as Chinese or Tangut) is also virtually impossible: pattern recognition algorithms can made a good guess about which character out of, say, a hundred of so should be used to represent a particular blob on the page. The problem posed by selecting one out of a set of 5,000 is of a different character altogether.

### 4.2. Converting dictionaries into database format

For many more conventional works, scanning and OCR works quite well. Dictionaries in conventional roman orthography scan and OCR relatively well, and given that dictionary format are relatively consistent, it is possible to find heuristics to help with the markup as well.

The field of computational lexicography is young but maturing rapidly. Researchers in the fields of natural language processing and artificial intelligence have recently recognized the importance of having access to detailed machine-readable versions of the lexicon of the languages being processed. Until recently, however, most researchers in historical linguistics had to content themselves with using printed sources.

I turn now to the practical problem of making the words in printed dictionaries and so on useful for research, and to the problem of using existing converted dictionaries. Several possible forms are useful, and given the difficulties involved it is important to pick the correct target representation and to avoid doing more work than necessary.

The conversion may be done with two goals in mind: either publication or research. These are two very different ends which may at time meet. Preparing data for publication (merely) means rendering the information in a form suitable for transfer to the printed page, and computers are a great boon to this process as they make the revision of the printed page a matter of changing exactly and only what needs changing, while retaining what was already correct. Preparing data for computer-aided research, however, usually implies a much more complete analysis of the content and structure of the information. Many works originally prepared in machine-readable form for convenience of publication have found their way into research settings. On the other hand, the ultimate goal of the research is publication; if the material to be published is already converted and in use, so much the better.

There is a relatively well-accepted typology of lexicographic sources. The cline starts with relatively unstructured data and moves towards highly articulated structures.

- Text base (TxB): the most basic version of a machine-processable lexicon is simply a verbatim text version of the original document (prototypically a printed dictionary). With such a document one can, for example, easily search for particular strings of characters, making it possible to locate information that would otherwise be virtually inaccessible.

- Machine-readable dictionary (MRD): when the component parts of dictionary entries (e.g. headword, part-of-speech, definition, etc.) are distinguished from each other by marks within the entries it becomes possible to use the dictionary for more sophisticated analysis employing (with some limitations) conventional database strategies.

- Knowledge base (KB): at this high level of analysis most of the implied structure in lexical entries has been made explicit and a good deal of real-world knowledge necessary to interpret the meanings and relationships between entries is encoded. A KB is therefore closer to an encyclopedia.

### 4.2.1. SGML markup of dictionaries

The *de facto* (and soon to be *de jure*) standard for marking the distinctions in content required for any of these types of sources is SGML, Standard Generalized Markup Language, discussed in some detail below. Recently, standards for computer encoding of a wide variety of document types, including dictionaries, have been established as part of the Text Encoding Initiative (TEI).[98] The section on Print Dictionaries is particuiarly interesting:

> Both typographically and structurally, dictionaries are extremely complex. [...] the many general problems of encoding text are particularly pronounced here, and more compromises and alternatives within the encoding scheme may be required. Two problems are particularly prominent.
>
> First, because the structure of dictionary entries varies widely both among and within dictionaries, the simplest way for an encoding

> scheme to accommodate the entire range of structures is to allow virtually any element to appear virtually anywhere in a dictionary entry. It is clear, however, that strong and consistent structural principles do govern the vast majority of conventional dictionaries, as well as many or most entries even in more 'exotic' dictionaries. [...] (TEIP3 v1}:321)

To accommodate the wide variety of coding styles, the TEI therefore provides two types of entries: a 'strictly typed' entry (<entry>) and a more lenient type (<entryFree>).

Next, the TEI notes that space considerations normally dictate a highly abbreviated format for entries in dictionaries, in which much of the information is available only to the human reader who is able to interpolate the missing or implied information.

> Second, since so much of the information in printed dictionaries is implicit or highly compressed, their encoding requires clear thought about whether [...] the goal of encoding [...] is to capture the precise typographic form of the source text or the underlying structure of the information [...] they present.

What is meant here by this somewhat telegraphic statement is that dictionaries rely on a number of conventions to abbreviate entries, including:

- Copious front matter describing typographic conventions, abbreviations, etc. The reader of an article must mentally interpolate the symbols and conventions in the front matter into the entry. Human beings, I note, can do this effortlessly, reading dictionary entries as either prose or poetry.

- Iconic relations between parts of entries (e.g. the first definiendum is the primary one, indentation of subparts can indicate a hierarchical relation of derivation.

- Repeated letters may be replaced by a shorter string if it is clear (or clearly implied) where the segmentation is being made, e.g., the use of ~ to indicate the headword in examples: **zugrunde** ... *adv.* ... ~ *gehen fig.* go to ruin, perish ... (Messinger 1973:631)

- Typographic variations, including size, style (bold, italic, small caps), font (serif, sans-serif), to indicate the type of content.

In light of our earlier discussions (§3.2.1) on problems of representation, typography, and formatting, it is interesting to note that the authors of TEIP3, in order to simplify their electronic presentation on encoding dictionaries, elected to avoid transcriptional problems in their own exegesis:

> 'to simplify the electronic presentation of this document on systems with limited character sets, many of the pronunciations are

presented using the transliteration found in the electronic edition of the OALD[...]'

Notably, this transcription uses only the standard ASCII characters — no fancy phonetic characters — and this is the only place I found in the two-volume, 1,290-page work that such shortcuts have been taken!

The principles of markup are best illustrated by example. Consider the dictionary entry below:

*(95) A monolingual dictionary article (TEIP3 v1}:330-331)*

**disproof** (dIs"pru:f) n. 1. facts that disprove something. 2. the act of disproving.

[Sinclair 1994]

When SGML tagging is introduced into this entry, each functional item in the entry is distinguished:

*(96) SGML markup of the entry given in (95) above*

```
<entry>
  <form>
        <orth>disproof</orth>
        <pron>dIs"pru:f</>
        <gramGrp><pos>n</></gramGrp>
        <sense n = 1><def>facts that disprove something.</def></sense>
        <sense n = 2><def>the act of disproving.</def></sense>
  </form>
</entry>
```

Just a few details of the tagging conventions will be mentioned here; the interested reader is referred to the source. The tags (text within

<>) are essentially labeled brackets which indicate the type of data enclosed. A bracket can be closed without naming which type of bracket is being closed by using </>; the convention is that this closes the most recently opened bracket. In some rare cases tags do not have to be closed: these tags indicate the start of some conventions within the text which hold until they are changed. Entities (this terms has a precise meaning in SGML which irrelevant here) can be imbedded inside other entities. There are restrictions on this embedding enforced by an external set of conventions which define the structure of the document; this set of conventions, called a document type definition (DTD) must be specified for each type of document. DTD have been created for most types of documents, and it is the Print Dictionary DTD which is being used and discussed here.

Of course, there are many additional types of information that one would like to be able to distinguish in dictionaries. The TEI standard discusses how to handle more complicated entries, entries from bilingual dictionaries, and so on. Of particular interest here is the treatment of etymological information. Here the standard is adequate for marking most of the conventional pieces of etymologies found in dictionaries, but not for including the level of detail required for ongoing comparative linguistic research. The subentry <etym> in the TEI provides only a few 'content-specific' tags:

*(97) Content-specific tags for etymological information (345-346)*

| | |
|---|---|
| <etym> | encloses etymological information in dictionary entry |
| <lang> | language name mentioned in etymological or other linguistic discussion |
| <date> | date in any format |
| <mentioned> | mentioned, not used |
| <gloss> | phrase or word used to provide a gloss or definition |
| <pron> | pronunciation |
| <usg> | usage information |
| <lbl> | in dictionaries, contains a label for a form, e.g. 'literally', 'abbreviation for', etc. |

Thus, for example, marking up the following entry:

*(98) A dictionary article containing an etymology*

> **neume** \'n(y)üm\ n [F, fr. ML pneuma, neuma, fr. Gk pneuma breath — more at **pneumatic**]: any of various symbols used in the notation of Gregorian chant...
>
> [Webster's 1975]

we would see something like this:

*(99) The same dicionary article in SGML with an <etym> subentity*

```
<!-- ... -->
<etym>
<lang>F</>
    fr. <lang> ML</><mentioned> pneuma</><mentioned> neuma</>,
    fr. <lang>Gk</><mentioned> pneuma </><gloss>breath </>
    <xr type=etym>more at <ptr target ='pneumatic'></xr>
</etym>
<def>any of various symbols used in the notation of Gregorian chant...
<!-- ... -->
```

N.B. <!-- ... --> indicates elided text in SGML document

Note that there is no provision here for explicitly representing some of the important details of etymology. Thus, considering the example above (99), the form *neume* is a borrowing from French, which borrowed it from Latin, which in turn borrowed it from Greek; but the only indication of the nature of the relation is the abbreviation 'fr.', which is external to the markup. The issue of distinguishing between genetic inheritance and borrowing is not explicitly treated in the set of etymological markup tags, though there are *ad hoc* methods for making the distinction in SGML. Using markup conventions beyond the standard may, however, also require more programming to be done down the line: standard software packages usually handle the specified document types, extensions are extra.

### 4.2.2. Views

The TEI recognizes that dictionaries might be encoded with several different ends (called *views*) in mind. Three possible views of dictionaries include (359-363):

- the typographic view, concerned with the two-dimensional printed page;
- the editorial view, concerned with the one-dimensional string of tokens which can be seen as the input to the typesetting process;

> - the lexical view, which includes the underlying information represented in a dictionary, without concern for its exact textual form.
>
> It is sometimes desirable to retain both the lexical and editorial view, in which case a potential for conflict exists between the two. (TEIP3 :364).

Given the complexity of representations used in different dictionaries, it is likely that etymological research would require the retention of both the typographic and the lexical views: the typographic view in order to allow researchers to verify the source data, the lexical view to provide a means of for computer software to get at the information implicit in the front matter and iconic representation.

### 4.2.3. Snoxall's Luganda dictionary

The results of scanning and marking up a dictionary of a Bantu language (the Luganda dictionary of Snoxall 1967, part of the CBOLD project discussed below) are exemplified in (100) below. This version of the dictionary is merely a machine-readable text version of the original printed work.

*(100) Source text from Luganda dictionary (p. 21)*

**bìkudumo, è** *n.* VIII dregs of *mùbisì.*
**bìkujjujju, è** *n.* VIII knots; roughness: *`Olùgoyè lùnò òk`ubaàko èbikujjujju, si kìrungì* for this cloth to be knotty is not right.
**bìkukujju, è** *n.* VIII moss; lichen. Also **kàkukujju** *n.* Ia.
**bìkwakwàya, è** *n.* VIII dust (of tobacco and grasses).
**bìkwêra, è** *n.* VIII scaly rash in syphilis.
**bìkyâ, è** *n.* VIII 1. tendons of neck. 2. neck (Buddu). *cf.* **lùkyâ,ò-**.
*kù*-**bimba** *v.i.* foam; froth; effervesce.

To convert this version of the dictionary into a machine-readable dictionary (i.e. an MRD), a one-shot program had to be written to parse the text and insert the tagging. The results of this process are depicted in (101).

*(101) (100) as encoded in SGML (N.B.: tagging is non-standard)*

```
<entry id=476>
    <lex>bìkudumo</>,
                <prf> è</><pos>n</>.<cls>VIII</>
                <def>dregs of <lex>mùbisì</>.</>

<entry id=477>
    <lex>bìkujjujju</>,
            <prf> è</><pos>n</>.<cls>VIII</>
            <def>knots; roughness</>:
    <ex>`Olùgoyè lùnò òk`ubaàko èbikujjujju, si kìrungì
    <tr>for this cloth to be knotty is not right</>.</>

<entry id=478>
    <lex>bìkukujju</>,
            <prf> è</><pos>n</>.<cls>VIII</>
            <def>moss; lichen</>.
    <xref>
    <x>Also</>
                <xref n=> <lex>kàkukujju</>
                <pos>n</>.
            <cls>Ia</>.</></>
```

```
...

<entry id=481>
            <lex>bìkyâ</>,
            <prf> è</><pos>n</>.<cls>VIII</>
            <def n=1>tendons of neck</>.
            <def n=2>neck (Buddu)</>.
<cf>lùkyâ, ò-</>.
    <prf> è</><pos>n</>.<cls>VIII</>.</>

<entry id=482>
    <pfx>kù</>-<lex>bimba</>
    <pos>v.i</>.
    <def>foam; froth; effervesce</>.</>
```

Some features to note about the markup:

- All individual Luganda forms except examples (<ex>) are marked up as lexical items <lex>.

- Punctuation is retained outside of the markup (hence the markup approximates the so-called lexical view).

- Morphological distinctions (i.e. separation of prefixes) are retained in the markup.

From the marked up version of the text, information can be extracted for use in a database. Data from various tags can be extracted to make database entries as shown in (102) below. Note that, for example, lexical items from within the entry are extractable (cf. *kàkukujju*, in #478).

*(102) Data extracted and sorted from SGML markup (given in (101) above)*

| DataType | Entry # | Contents |
|---|---|---|
| Defn | 467 | dregs of *mùbisì* |
| Defn | 479 | dust (of tobacco and grasses) |
| Defn | 482 | effervesce |
| Defn | 482 | foam |
| Defn | 482 | froth |
| Defn | 477 | knots |
| Defn | 478 | lichen |
| Defn | 478 | moss |
| Defn | 481 | neck (Buddu) |
| Defn | 477 | roughness |
| Defn | 480 | scaly rash in syphilis |
| Defn | 481 | tendons of neck |
| | | |
| LexItem | 476 | bìkudumo |
| LexItem | 478 | bìkujjujju |
| LexItem | 477 | bìkukujju |
| LexItem | 479 | bìkwakwàya |
| LexItem | 480 | bìkwêra |
| LexItem | 481 | bìkyâ |
| LexItem | 482 | bimba |
| LexItem | 478 | kàkukujju |
| ... | | |
| etc. | | |

One use of the marked up text, for example, is to create a 'reverse dictionary' of Luganda using the contents of the <def> and <lex> tags could be sorted and used:

(103) *English-Luganda'Reverse dictionary'*

| English Gloss | Luganda |
|---|---|
| dregs of mùbisì | bìkudumo |
| dust (of tobacco and grasses) | bìkwakwàya |
| ... | |
| effervesce | kù-bimba |
| ... | |
| foam | kù-bimba |
| froth | kù-bimba |
| ... | |
| knots | bìkujjujju |
| ... | |
| lichen | bìkukujju |
| ... | |
| moss | bìkukujju |
| ... | |
| neck (Buddu) | bìkyâ |
| ... | |
| roughness | bìkujjujju |
| ... | |
| scaly rash in syphilis | bìkwêra |
| ... | |
| tendons of neck | bìkyâ |

Of course, making a reverse dictionary is merely *aided* by having a computerized copy; the process of selecting headwords and lemmatization, both of which are only moderately automatable, still need to be done.

### 4.2.4. David Bradley's Lisu Dictionary

The recently published *Dictionary of the Northern Dialect of Lisu (China and Southeast Asia)* (Bradley 1994) illustrates a number of issues commonly encountered in converting dictionary data into machine-readable form. The source is a translation of a Lisu-Chinese original into English, and forms as it stands a generally consistent and coherent whole, several steps must be taken to make its contents accessible to the non-specialist and available for further computer processing. The steps are described in detail in Handel forthcoming.

The figure below illustrates briefly the original form of the dictionary, the converted form, and the file (generated in the same pass) for loading into the STEDT lexical database.

*(104) Original Lisu dictionary entries (after Bradley 1994:15)*

| | |
|---|---|
| bbaithaint | *N* termite |
| bbaithaint alnat | white ants |
| bbaithaintlox | 'black-charcoal-stick-fungus' (fungus which grows on the nests of termites) |
| bbaithaintm_u | *collybia albuminosa* (fungus, speciality in Yunnan) |
| bbaithaintnai | flying ant |
| bbaithaintnai | *collybia albuminosa* (big ones) |

To the extent possible, of course, one should preserve the original orthography of the transcribed data. Since transcription systems vary, this will necessitate providing an explanation of the transcription system in a separate document (as mention before in the discussion of *phonological inventories* §3.2.1.2). In the present case, Bradley's Lisu transcription (based primarily on the transcription system developed in the People's Republic of China in the 1950s) requires a good deal of interpretation by the researcher. Among other things, the system uses final consonants to indicate tones. This leads to unwieldy and misleading forms like bbaithaint 'termite' (bæ$^{21}$ hæ̃$^{21}$ in IPA). Retranscribing the data into IPA according to the description given in the front matter reduces the possibility of misinterpretation at a later stage when the data is compared with data from other languages.

A computer program was written to effect this retranscription.[99] Difficulties surfaced almost immediately. For one thing, a simple attempt to specify a list of simple unconditioned conversion rules (such as bb- > b-; -ai- > -æ-; -t > $^{21}$ [a tone]) revealed some ambiguities in the highly phonologized transcription system, which required that the phonological environment of the symbols be accounted for . For example, Bradley notes that '*ea* and *eo* represent /ɣa/and /ɣɔ/, and *ei* represents /ji/ [if syllable initial], otherwise consonant plus -e is used to represent a back unrounded vowel...'. (Bradley 1994:ix) Thus the symbol *e*, may variously represent /ɣ/, /j/, or /ɤ/ depending on context.

One of the interesting side effects of running the program was to identify ambiguous or incorrect transcriptions. Any word that could not be parsed by the computer program (which encoded the essence of the phonotactics of Lisu as represented in Bradley's transcription) was automatically suspect. In this way a number of typographical errors that might have escaped notice were identified almost instantaneously. It also became clear that, although the orthography included a conventional symbol (the apostrophe) to indicate otherwise ambiguous syllable boundaries, this symbol was underutilized in the dictionary. To someone familiar with the Lisu language these ambiguities would be resolved by context, but to the non-specialist using the dictionary as a reference work, this kind of ambiguity could lead to misreadings.

Many words in the dictionary are Chinese borrowings, and are marked as such. However, these are phonologically distinct from the native forms and could not be parsed with the same algorithm. Additional coding was required to generate an appropriate retranscription in these cases. The results of the retranscription and database reformatting are shown in (105) and (106) below. Note that the numbering of the sub-entries in the database format refer back to the head entry, permitting database users to find the superordinate term regardless of how the data are actually sorted.

(105) *Lisu data after retranscription (after Bradley1994:15)*

| | |
|---|---|
| bæ$^{21}$hæ̃$^{21}$ | N termite |
| bæ$^{21}$hæ̃$^{21}$ a$^{55}$na$^{21}$ | white ants |
| bæ$^{21}$hæ̃$^{21}$lɔ$^{44}$ | 'black-charcoal-stick-fungus' (fungus which grows on the nests of termites) |
| bæ$^{21}$hæ̃$^{21}$my$^{33}$ | collybia albuminosa (fungus. speciality in Yunnan) |
| bæ$^{21}$hæ̃$^{21}$næ$^{33}$ | flying ant |
| bæ$^{21}$hæ̃$^{21}$næ$^{33}$ | collybia albuminosa (big ones) |

(106) *Data above 'massaged' into a typical database-ready format*

| reflex | gloss | POS | # |
|---|---|---|---|
| bæ$^{21}$hæ̃$^{21}$ | termite | N | #97 |
| bæ$^{21}$hæ̃$^{21}$ a$^{55}$na$^{21}$ | white ants | N | #97.1 |
| bæ$^{21}$hæ̃$^{21}$lɔ$^{44}$ | 'black-charcoal-stick-fungus' (fungus which grows on the nests of termites) | N | #97.2 |
| bæ$^{21}$hæ̃$^{21}$my$^{33}$ | collybia albuminosa (fungus. speciality in Yunnan) | N | #97.3 |
| bæ$^{21}$hæ̃$^{21}$næ$^{33}$ | flying ant | N | #97.4 |
| bæ$^{21}$hæ̃$^{21}$næ$^{33}$ | collybia albuminosa (big ones) | N | #97.5 |

### 4.2.5. James Matisoff's Lahu Dictionary

*The Dictionary of Lahu* (DL) (Matisoff 1988) provides another example of the challenges of database conversion. This work was printed using special software developed during the 1980s, when phonetic fonts and large text files required lots of custom software and careful machine and human manipulation. As a result, however, the basic text was available in machine-readable form.

The data contained in these files would be quite useful for both cross-linguistic and language-internal research if it were available in a conventional data-processing format. For example, 'reversing' the dictionary to create an English-Lahu index would be possible. To do this, the carefully formatted textbase version of the dictionary was converted into a 'markup format' (in this case Lexware).[100] (Handel and Lowe 1995) Like SGML, Lexware permits the tagging of particular parts of the text according to their function. In particular, the Lahu headwords and the English glosses could be distinguished, allowing the software to extract elements for other purposes, such as inclusion in databases.

A relatively simple side effect of this conversion process was to generate, along with the reverse dictionary, a number of appendices containing lists of Lahu plant and animal names, grammatical particles, dialectal terms, kinship terms, and other categories of lexical items not immediately accessible through Lahu or English headwords but constituting useful and interesting categories for the linguist.

The first step was to 'parse' the dictionary articles in DL into the markup format (i.e. Lexware). Along the way, a retranscription of the ancient character encoding into a more tractable encoding for the current generation of computers was performed. Below is the rendering of a dictionary article in its original format, and as converted into the Lexware markup. Note that the hierarchical relationship between the head and sub- entry indicated in the source document by indentation is reflected in the marked up version by distinct tags (*.hw* means headword, *..hw* means subheadword in this markup language):

(107) *Source article from DL*

| | | |
|---|---|---|
| bə̂ʔ-bə̂ʔ | (AE) | [*RL*] gushing; flowing out constantly and copiously |
| šɨ̆ʔ-cɨ̂ bə̂ʔ-bə̂ʔ qay ve | (S + AE + V) | the sap gushed out |

(108) *As marked up in Lexware format*

| | |
|---|---|
| .hw | bə̂ʔ-bə̂ʔ |
| gfn | AE |
| dfe | [*RL*] gushing; flowing out constantly and copiously |
| ..hw | šɨ̆ʔ-cɨ̂ bə̂ʔ-bə̂ʔ qay ve |
| gfn | S + AE + V |
| dfe | the sap gushed out |

To create a proper reverse index, the markup text must be reviewed and corrected to ensure that the proper words are marked for reverse indexing. This procedure consists (in the Lexware model) of inserting asterisks and other markup characters in front of words that should be indexed and inserting the appropriate tagging to reflect morphological and syntactic boundaries.

The markup conventions are used in Lexware 'finderlists' (i.e. reverse indexes) are summarized below, as they exemplify simple and effective conventions for the task of identifying headwords and lemmatizing the elements for the reversing process. The results are shown below in (109).

*(109) Lexware markup conventions for reverse indexing*

a. If a word is 'starred,' it will be indexed up to the next 'white space' character. The asterisk may be imbedded inside a word as well:

be *angry; be annoyed with, *mad at
be un*clear; blurry; dull in color

b. If only part of a word is to be indexed, place a vertical bar where the index is to break off:

*gush | ing; *flow | ing out constantly and copiously

c. If a phrase is to be indexed, use a tilde between each word after the asterisk:

A number of cases must be treated specially. For example, how are subentries to be handled? In general, one would bother to create an index entry from a subheadword only when the gloss of the subheadword is significantly different from the gloss of its superordinate headword.

(110) *An article from DL in Lexware 'Invert' markup, and as converted for database use.*

| | |
|---|---|
| .hw | bí ɛ̀ |
| gfn | $AE_{stat}$ |
| dfe | __*swollen__ |
| ..hw | mɛ̂? bí ɛ̀ qay/te ve |
| gfn | $N_{spec}$ + AE + V |
| dfe | be nice and __*plump__; have a pleasantly round face |

| Form | gfn | Gloss |
|---|---|---|
| bí ɛ̀ | $AE_{stat}$ | swollen |
| bí ɛ̀ | $AE_{stat}$ | plump |
| mɛ̂? bí ɛ̀ qay/te ve | $N_{spec}$ + AE + V | be nice and plump; have a pleasantly round face |

Sample of Lexware format articles from DL, with automatic markup:

(111) *Preliminary version of one piece of the marked up database on which the reverse dictionary excerpt is based:*

| | |
|---|---|
| .hw | à-thò?=ma |
| pos | Nintg |
| gl | *what? |
| note | GL 3.323: SYNS. à-thò? (Nintg). à-ma (Nintg) |
| ex | chi à-thò?=ma le |
| exgl | What is this? |
| ..hw | à-thò?=ma te le |
| pos | N + V + Punf |
| gl | *why? ("doing/having done what?") |
| note | cf. Thai thammaj and Jse. dooshite 'why' ("doing how") |
| ..hw | à-thò?=ma + V + kà?/thɔ̂ |
| pos | Clnf |
| gl | *whatever one V's: *anything that is V'd |
| ex | à-thò?=ma pî kà? yɔ̂ mâ hə̂? |
| exgl | He didn't accept anything they offered: Whatever they offered he didn't accept. |

*Reverse dictionary (including some of the indexed items above)*

| | |
|---|---|
| anything | à-ma thɔ̂ < à-ma. |
| anything at all | à-thòʔ=ma ɔ̀-cə̀ thɔ̂ < à-thòʔ=ma. |
| anything that is V'd | à-thòʔ=ma + V + kàʔ/thɔ̂ < à-thòʔ=ma. |
| approximate | |
| approximately | à-là. |
| aunt | |
| [Chin. Lh.] father's older sister | á-ku=ma. |
| [Chin. Lh.] father's sister | á-šū=ma < á-šū. |
| [Chin. Lh.] father's younger sister | á-ni-a. |
| [Chin. Lh.] mother's older sister | á-yèʔ=ma. |
| uncle or aunt | á-šū. |
| wife of father's brother | á-šū=ma < á-šū. |
| awe | |
| be awed | â ~ ân. â-na ~ â-nà ~ á-nà ~ âʔ-nà. |
| bamboo | |
| section of bamboo | á-dɛ́-qō. |
| banana | á-pɔ̂. |

Another substantial advantage of MRDs is that they provide a means to easily gather statistical information about their contents. Such statistics can be useful in supporting typological arguments about both the documents they represent (in this case dictionaries) and the underlying content (in this case, information about the language). Any such statistics must be used with great caution inasmuch as the properties of the document may or may not reflect properties of the underlying content. If we consider a statistical summary of this dictionary's contents (shown in (112) below) we would be quite wrong to conclude that because there are about 28,000 headwords and subheadwords that there are 28,000 words in the language. However, the conclusion that the average word in Lahu requires around twenty symbols to transcribe is probably justified. Information of this type is particularly useful in database design.

*(112) Dictionary of Lahu contents: statistical summary*[101]

| Description | band (=tag) | N | Total Size<br>(bytes) | Average<br>(bytes) |
|---|---|---|---|---|
| headword | .hw | 5301 | 108863 | 20 |
| subheadword | ..hw | 22710 | 555684 | 24 |
| Definition | dfe | 27917 | 939227 | 33 |
| Examples | ex | - | - | - |
| Part of speech | gfn | 28011 | 223221 | 7 |
| Loan indicator | loan | 1285 | 5140 | 4 |
| Notes | note | 11465 | 893000 | 77 |
| Translations | tr | 4861 | 265885 | 54 |
| Total | | 101,550 | 2,991,020 | |

The conversion process, long and fraught with problems and the potential for error, is best carried out following the maxim: 'follow copy!' (Schneider 1974) At least then one can be assured that what one sees in the database bears a reasonable resemble to the original document. However, to make use of the data converted, further steps are required; one must often take several steps away from the original in order to achieve a cross-linguistic perspective.

# 5. HYPOTHESIS CREATION

## 5.1. 'Projections' from the source data

The words and forms in the database are of course not indivisible atomic units, nor are they necessarily useful as they stand without analysis, decomposition and transformation. From a database point of view, these latter elements should be stored as separate entities from the source data (which for obvious reasons of verification and attribution should remain inviolate). Some analyses take the form of queries against existing data structures, but many require different parts, versions, or re-codings of the source data (some of these recodings have already been discussed). I find it useful to view these other versions as *projections*, since the analytic process may take one source item and produce several analytic objects, or vice versa; I think of these as different from the processes described in the two previous chapters in which the object is to create machine-processable entities which resemble the original to the greatest extent possible. Here I will treat the kinds of manipulations which create new and different entities and arrangements.

### 5.1.1. Compounds and etymology

One of the tasks confronting the etymologist is to unravel and undo the processes of compounding and derivation which have interacted to produce the attested forms. Often, such research yields insights into semantic connections between words and things (cf. the example below, where the meanings of the words *foot* and *eye* are shown to be conceptually related (in TB languages) to words meaning *ankle*. (Matisoff 1978)).

In particular, it is useful to know which words in a language are compounds of other words, and to attempt to explain why these compound words mean what they do. Processes which create such compounds (hereinafter called *polynyms*) have both a synchronic and a diachronic component: sometimes the process is synchronically productive, or at least transparent (e.g. English *bookkeeping*); at other times the compounds are composed of bound morphemes which have etymological significance and apparent lexical independence but no longer have a recognized distinct meaning (e.g. *fro* in English *to and fro*, etc.)[102]

The process of identifying and analyzing such forms can be automated to a certain degree. This process is outlined below.

### 5.1.2. Synchronic decomposition of compound lexemes into constituents

#### 5.1.2.1. Computer-aided lemmatization in Tibeto-Burman

First, each word in each language is compared with all other words in the given language data set to find those in which it constitutes a part. Next, this list is sorted by the extracted 'complex' forms, in order to bring identical constituents together. Finally, the compound forms are matched with their constituents to provide a constituent analysis. The process is illustrated in (113).

*(113) Procedure to identify and parse compounds*

*Step 1: Search each word (possibly compound) as a constituent of other words (by language)*

| In language, | the word: | contains the word: |
|---|---|---|
| Limbu | lāŋmik 'ankle' | 'eye' mik |
| Lushai | ke mit 'ankle' | 'eye' mit |
| . | | |
| . | | |
| Lalung | iathong mi 'ankle' | 'foot' iathong |
| Limbu | lāŋmik 'ankle' | 'foot' lāŋ |
| . | | |

*Step 2: Sort by compound form and gloss to bring identical constituents together*

| Language | Word | Constituent |
|---|---|---|
| Lalung | iathong mi | 'foot' iathong |
| Limbu | lāŋmik | 'eye' mik |
| Limbu | lāŋmik | 'foot' lāŋ |
| Lushai | ke mit | 'eye' mit |

*Step 3: Consolidate Entries for identical forms*

| Language | Word | Decompostion |
|---|---|---|
| Lalung | iathong mi | 'foot' iathong + '?' mi |
| Limbu | lāŋmik | 'foot' lāŋ + 'eye' mik |
| Lushai | ke mit | '?' ke + 'eye' mit |

Now the Lalung word for *eye* is not mi but mik, and the Lushai word for *foot* is not ke but ket. Native speakers may or may not be aware of the connections however obvious (cf. English *elbow*, which is not transparently an 'L' to English speakers). These morphemes, however, cannot be transparently (i.e. synchronically) connected to the corresponding unbound form. That step is described next in section §5.1.3. Of course, there is nothing new in this procedure or its results *per se*.[103] However, the implementation as a search heuristic for resolving compounds in multilingual databases is my idea. The result of this process is a set of partially and sometimes completely decomposed lexemes, as is illustrated by the results for reflexes meaning ANKLE in (9). (The data cited here reflects, of course, only a small part of the computer-analyzed corpus.)

(114) *Partial list of constituents of the word for ANKLE in several TB languages*

| | Language | Word | Constituents |
|---|---|---|---|
| 1. | Bisu | làkhɨ̀khɨ̀kjú~khakjú | ['foot' làkhɨ̀]khɨ̀kjú~khakjú |
| 2. | Lahu | khɨ-mɛ̂ʔ-ši | ['foot' khɨ]-['eye' mɛ̂ʔ-ši] |
| 3. | Lakher | phei-lo-byu | ['leg' phei]-lo-byu |
| 4. | Lalung | iathong mi | ['foot' iathong] mi |
| 5. | Limbu | lāŋmik | ['foot' lāŋ]['eye' mik] |
| 6. | Lushai | ke mit | ke ['eye' mit] |
| 7. | Meithei | khumit | khu['eye' mit] |
| 8. | Miji | lai-cuŋ-kʰu | ['leg, foot' lai]-cuŋ-kʰu |
| 9. | Nasu | ʂɿ$^{21}$ na̠$^{21}$ tsɿ̠$^{55}$ | ʂɿ$^{21}$ na̠$^{21}$ ['joint' tsɿ̠$^{55}$] |
| 10. | Nesu | gɤ$^{21}$ ɳɹ̠$^{33}$ ʂe$^{33}$ | ['foot' gɤ$^{21}$] ['eye' ɳɹ̠$^{33}$] ʂe$^{33}$ |
| 11. | Tangkhul | phei mik ra | phei ['eye' mik] ra |
| 12. | WBurmese | khre myak ci' | ['foot' khre] ['eye' myak] ci' |

The idea is to automatically or semiautomatically analyze the morphological and lexical structure of compound expressions. This effort has both synchronic and diachronic components: some compounds are transparent to speakers because their constituents are lexical items in their own right (e.g. *somebody* contains the word *body*). Note that in many cases, the semantics of a compound are opaque because semantic or phonological change has obscured the origin. Performing this process on these forms is unlikely to be enlightening. The forms for *rainbow* ((115) below) in Tibeto-Burman languages are often opaque in this way. Only a few of the forms are shown here; the complete list may be found in Appendix 7. The extensive list is shown here to illustrate that even in related languages very different types of compounds, composed of unrelated morphemes are used; while one morpheme (reconstructed by Bradley as *L ʃi$^{3}$) is given, most of the forms are four to six unetymologizable syllables.

(115) *Lolo-Burmese forms meaning rainbow from the STEDT database*

| | | |
|---|---|---|
| *Loloish | ʃi$^{3}$ | DB-PLOLO 322 |
| Achang | xɔŋ$^{51}$tɕin$^{31}$nam$^{31}$ | ZMYYC 16 |
| Burmese (Rangoon) | tθɛʔ$^{44}$tɑ̃$^{53}$ | ZMYYC 16 |
| Burmese (Written) | cui | PKB-WBRD |
| Burmese (Written) | sak-taṁ | PKB-WBRD |
| Burmese (Written) | saktan | GEM-CNL |
| Burmese (Written) | thɑk$^{3}$tɑm$^{1}$ | ZMYYC 16 |
| Hani (Caiyuan =Biyue) | tshɤ$^{55}$thɤ$^{55}$lɤ$^{55}$khy̠$^{33}$ | ZMYYC 16 |
| Jinuo | ja$^{33}$mɔ$^{33}$ko$^{44}$ta$^{55}$ko$^{44}$tɕhø$^{33}$ | ZMYYC 16 |
| Lahu (Na) | ᴀ$^{33}$mu$^{53}$lᴀ$^{31}$si$^{35}$dzɔ$^{33}$ | ZMYYC 16 |
| Langsu (=Maru) | ɣɔʔ$^{31}$kəŋ$^{35}$saŋ$^{31}$ŋjɔ̃$^{31}$ | ZMYYC 16 |
| Lijiang Naxi | mɯ$^{33}$lɯ$^{55}$xɯ$^{55}$dzi$^{31}$ | ZMYYC 16 |
| Lisu | ɑ$^{55}$mu$^{31}$ʃɿ$^{44}$kho$^{55}$ | ZMYYC 16 |
| Lisu (Northern) | a$^{55}$mɔ$^{21}$ ʃʐ̩$^{44}$khɔ$^{55}$ | DB-LISU |
| Lisu (Northern) | a$^{55}$mɔ$^{21}$ʃʐ̩$^{44}$ | DB-LISU |
| Lushai | chhimbâl | GEM-CNL |
| Nanhua Yi | mɯ$^{21}$ɕi$^{33}$dᴀ$^{33}$ | ZMYYC 16 |
| Nanjian Yi | ɑ$^{55}$m̩(u̠)$^{21}$tshɑ$^{55}$du$^{55}$ro$^{21}$ | ZMYYC 16 |
| Yi (Xide) = Gazhuo | si$^{33}$si$^{33}$ | ZMYYC 16 |
| Yongning Naxi (=Moshang) | mv$^{33}$ɕi$^{55}$ | ZMYYC 16 |
| Zaiwa=Atsi | voʔ$^{21}$kan$^{21}$ʃiŋ$^{55}$jaŋ$^{21}$ | ZMYYC 16 |

### 5.1.2.2. Abstracting away from surface forms in Bantu

The units of diachronic analysis are rarely surface forms, especially in languages which have a rich morphological life. In order to compare elements of Bantu languages historically it is necessary to peel away several layers of morphological and phonological material into order to extract the comparanda. In the case of nouns, the Bantu concord markers must be accounted for. In the case of verbs, there are a number of affixal processes which must be undone.

Dictionaries of Bantu languages often cite full undifferentiated words (e.g. *kulima* 'to cultivate', *mulimi* 'farmer'[104]) as headwords. The lexicographic sort sequence may or may not take account of prefixes for either nouns or verbs. Adverbs and other parts of speech, though diachronically and even synchronically analyzable, are often cited as indeclinables. Other dictionaries provide entries in stem form, e.g. *-lima* 'cultivate', or even by root, e.g. *-lim-*. (Hyman 1994) Particularly in the case of verbs there are least four distinct domains (for which the standard Bantu terminology is indicated below) which must be controlled for in diachronic research:

*(116) The morphological structure of the Bantu verb*

```
                 WORD
                /    \
        prefixes      STEM
                     /    \
                 BASE      final vowel
                /    \
            ROOT      extensions
```

e.g. ku-lim-il-an-a 'to cultivate for each other'

```
[ ku-  [ [ [ -lim- ] root  -il-an-  ] base  -a ] stem  ] word
       inf      'cultivate'              app-recip      fv
```

According to Hyman, several of these morphological domains are 'obligatory' (and indicated above in SMALL CAPS). As shown, the word divides up into a stem and from zero to several prefixes. The stem in turn consists of a base and a final vowel. The base consists of an obligatory root and zero to several extensions. Both the extensions and the final vowel are suffixes. Within verb stems, extensions are generally derivational (marking e.g. causative, applicative, passive, reciprocal), while the final vowel is inflectional (marking e.g. indicative, subjunctive, perfective). The distinction between an extension and a final vowel can also be applied to derived nouns, e.g. the -i of *mu-lim-i* 'farmer' is a final vowel deriving agentive nouns from verbs.

And while many sources do provide full word entries they differ in the degree to which they segment individual morphemes. While some provide no segmentation (hence *mulimi, kulima, kulimilana*), others separate prefixes from stems (mu*limi*, ku*lima*, ku*limilana*). Few provide full words with morpheme breaks within stems (mu-*lim-i*, ku-*lim-a*, ku-*lim-il-an-a*). Sources that attempt to provide extensive morpheme divisions typically cite verbs in stem form (*lim-a, lim-il-an-a*).

Some sources analyze the data even further, providing abstract morphophonemic representations and sometimes examples of surface phonetic transcriptions as well. In such morphophonemic representations morpheme breaks are generally indicated, and the consonants, vowels and tones are cited in appropriately abstract, underspecified forms. In the following (117) from the Shi lexicon, for example, both surface and underlying representations are given.

*(117) Two types of representations of words in Shi*

| Morphophonemic<br>representations | Surface form | Gloss (Fr.) |
|---|---|---|
| *-buid-id-I-* | óokubwiirizaa | enseigner, apprendre à qn., souffler (ex. aux examens) |
| *-bùnb-à* 5 | íibuùmba | argile à poterie |

(Polak-Bynon 1975)

In this source, both underlying (morphophonemic) and broad phonetic transcriptions are indicated for all forms. As indicated in bold, alphabetization here is by underlying representation (UR): Nouns are cited in the left column by stem, while verbs are cited by base (i.e. exclusive of the predictable FV -a). Morpheme breaks are indicated only in the UR's. Phonological rules are responsible for the discrepancies seen between the UR's and the surface forms, e.g. /d/ becomes [z] before /I/ (< *į), otherwise [r] unless preceded by a nasal, the sequence /Cui/ is realized [Cwii], and the nasal of /nb/ undergoes homorganic nasal assimilation to produce [mb]. (The final lengthening in óokubwiirizaa is also from gliding, /dI-a/ → /zI-a/ → /zyaa/.) The tones are also marked differently in the two columns. Thus, the Shi lexicon provides two entries for each form which differ in three ways: (i) by domain (stem/base vs. word); (ii) by morpheme segmentation (has vs. has not); and (iii) by level of representation (UR vs. phonetic). Most sources do not provide two representations, as in this case, but instead give only one, which is either phonetic or (quasi-) phonemic. (Hyman 1994)

In order to compare data from different sources some means must be developed to create or project forms of these different types into a unified structure, similar to the 'canonical forms' discussed in §3.2.1.7.

Even given workable URs, some effort must be devoted to creating structures encoding the various domains. For example, we might wish (as in the Tibeto-Burman examples) to find all occurrences of a given morpheme (in this case, *roots* abstracted or segmented away from their surface representations).

First, we need to break words up into their morphological constituents according to the scheme given (118) above. This would enable us to index the various parts of the forms separately:

*(118)*

| *Surface form* | *Gloss* | *Morphologically segmented form* |
|---|---|---|
| *kulima* | 'to cultivate' | P(ku)R(lim)F(a) |
| *mulimi* | 'farmer' | P(mu)R(lim)F(i) |
| *meeso* | 'eyes' | P(ma)SR(iso) |

**Legend:**

| | | | |
|---|---|---|---|
| P = | Prefix | R = | Root |
| S = | Stem | F = | Final |

We could then identify all the forms in which the root *-lim-* occurred. Note that in *meeso* we should 'undo' the morphophonological rule which combines vowel-initial stems. To do this properly we must either create the parsed expressions by hand or implement some parser into the process. Since there are thousands of forms we should try to use the parser! Also note that the root *iso* in 'eyes' should also be marked as a stem (i.e. there is no final vowel or extension, so this is 'obligatorily' both stem and root).

Consider the Nyamwezi data below. Given this kind of representation we might want to search for forms which meet certain structural criteria, for example, to find forms which have a high tone in the second syllable and a complex onset. The second form below meets this criteria.

*(119)Nyamwezi forms in surface and 'UR'*

| Surface | UR | UR | Class | English |
|---|---|---|---|---|
| Surface | (segments) | (Tone) | Class | Gloss |
| pé | pe | H | I | *all* |
| dwií | dwii | LH | I | *all* |

i.e.

dwií =

| | |
|---|---|
| dwi | i |
| L | H |

A way of representing the data in such a way that this query could be met is with labeled brackets (which could of course be, and elsewhere have been, implemented as SGML tags). Creating (again algorithmically, one hopes) bracketed representations like those shown below would make some queries possible.

Recognizing that each morpheme has a structure we wish to be able to refer to in developing hypotheses, we should further divide up the roots (and perhaps other lexemes) into their phonological constituents in a similar way. There may be several such divisions, as shown in (120) below, and we may wish to generate and retain all of them, depending on the needs of our research.

*(120) Various bracketings of the Nyamwezi forms in (119) into phonological constituents*

| | |
|---|---|
| C(**p**)V(**e**)T(**H**) | Codes each elements in brackets |
| C(**d**)G(**w**)V1(**i**)V2(**i**)T(**LH**) | Codes and numbers each segment |
| L[C(**d**)G(**w**)V1(**i**)] H[V2(**i**)] | Tonal domain indicated by brackets |
| C1(**d**)G(**w**)V1(**i**)C2(**Ø**)V2(**i**) | explicit Zero initial coded in 2nd syllable |
| S(**dwi**)S(**i**) | Syllable domain bracket |

**Legend:**

| | | |
|---|---|---|
| P | = | prefix |
| S, $S_n$ | = | syllable domain |
| V, $V_n$ | = | vowel |
| C, $C_n$ | = | consonant |
| L, H, T | = | tonal domains |

Having multiple parsings according to different analyses would make it possible to perform a variety of complex queries based on morphological and phonological criteria.

Many cases are more difficult and may require a relatively sophisticated parser. Yaka *lusóngáni* (class 11 singular) 'red forest ant' has transparently a prefix *lu-* and stem *sóngáni.* The plural, *tsóngáni* (class 10) incorporates a prefix *n-*. In order to give the correct UR shown in (121), the algorithm creating the bracketed form would either have to be sensitive to the morphophonemic rules required to analyze the form or the analysis would have to be done by a human being.

(121) *Yaka* tsóngáni *(class 10 plural)*, *'red forest ant'*

P(**n**) S1(**son**) S2(**ga**) S3(**ni**) T(**-HH-**)

**Legend:**

| | | |
|---|---|---|
| P | = | prefix |
| $S_n$ | = | syllable domain |
| $V_n$ | = | vowel |
| $C_n$ | = | consonant |

### 5.1.3. Diachronic decomposition of lexemes into constituent morphemes

As noted in (113) above, some part of the form is often *synchronically* unanalyzable (e.g., somewhat trivially mi in Lalung iathong mi 'ankle' and khu in Meithei khumit). Here we must seek an explanation in the historical development of the languages. Matisoff lays out the historical process concisely:

> Each TB language, when faced with the necessity of forming polysyllabic compounds, dips into its inheritance from the proto-treasury in its own unpredictable way. Some languages are satisfied with a single root morpheme, and prepose a prefix to it for 'phonological bulk' (WT, Bisu, Lahu Shi). Several select two root morphemes (Akha, Lisu, Burmese, Black Lahu, Maru, Lashi). Red Lahu has a trisyllabic form... (Matisoff 1978:63)

These options are illustrated in the following table:

*(122) Alternative selections from the proto-treasury of morphemes in TB compound-formation (Matisoff 1978:63-64)*

(Adapted from Nishida 1967: p.68)
*Compounds meaning HEAD*

| Language | Form | Prefix | *bu ( > *wu) | *du(k) | *s-ko(ŋ) | *l(yam) ≭ *lum |
|---|---|---|---|---|---|---|
| Written Tibetan | d-bu | d- | bu | | | |
| Bisu | ?aŋ̀-tù | ?aŋ- | | tù | | |
| Akha | ?ù-dù | | ?ù | dù | | |
| Written Burmese | ?û-khôŋ | | ?û- | | khôŋ | |
| Lisu | wu$^1$-dū$^3$ | | wu$^1$- | dū$^3$ | | |
| Black Lahu | ó-q̄o | | ó- | | q̄o | |
| Red Lahu | a-tù-kù | ?a- | | -tù- | kù | |
| Yellow Lahu | ?ùdù | | ?ù | dù | | |
| Maru | ?àu-làm | | ?àu- | | | làm |
| Lashi | ?ù-lèm | | ?ù- | | | lèm |
| Atsi | u-lum | | u- | | | lum |

Here each language makes its own idiosyncratic and unpredictable selection of inherited root-morphemes and/or prefixes to create compounds: *bu and *du(k) are root-morphemes meaning 'head' (also 'neck' in the case of *du(k)), *s-ko(ŋ) basically means 'hollow object', while *l(y)am ⪤ *lum means 'round object'. Both semantically and etymologically the structures of these compounds thus represent many heterogeneous possibilities.

Using software developed for the STEDT database, it is possible to retrieve any and all such compounds (previously marked by hand) using the fundamental tool of *etymological tagging* . An example of the results of such tagging, involving some compounds meaning *corpse,* are displayed in the KWIC (keyword in context) list shown in (123) below:

*(123) A KWIC (keyword in context) list*

| Tagging | Reflex | ID# | Language Name | Gloss |
|---|---|---|---|---|
| 27 | ɕɛ$^{53}$ | 131008 | Shixing | *die/be dead* |
| 27,1 | ɕi$^{55}$ gɯ$^{55}$ | 40109 | Dulong Dulonghe | *corpse* |
| 27,2 | ɕi$^{55}$ gɯ$^{55}$ | 40293 | Dulong Nujiang | *corpse* |
| 27,17 | shi- mang | 114867 | Abor-Miri | *carcass/corpse* |
| 27,17 | si: mhɔ | 62102 | Newari | *corpse / dead body* |
| p, 27,17 i$^{33}$ | ʂi$^{33}$ mu$^{33}$ | 47295 | Ahi | *carcass / dead animal* |
| 27,17,s | ʂi$^{33}$ mu$^{33}$( mɒ$^{33}$) | 131896 | Sani | *corpse* |
| 27,17 | ʂi$^{22}$ mu$^{33}$ | 47294 | Ahi | *corpse / dead body* |
| 27,17 | ʂʅ$^{33}$ ṃ$^{33}$ | 16450 | Nyiq | *corpse / dead body* |
| 27,17,s | ʂʅ$^{33}$ ṃ$^{33}$ mɑ$^{33}$ | 17254 | Sani (Wu) | *corpse* |
| 27,17 | ʂɿ$^{33}$ mɒ$^{33}$ | 15692 | Lolopho | *carcass / dead animal* |
| 28, 27,17 | tshɒ$^{33}$ ʂɿ$^{33}$ mɒ$^{33}$ | 15691 | Lolopho | *corpse / dead body* |
| 27,17 | ʂɿ$^{55}$ mʐaŋ$^{55}$ | 38400 | Achang Longchuan | *corpse* |
| 27,17 | ʃi$^{44}$ mʌ$^{44}$ | 14275 | Jinuo (A) | *corpse* |
| 28,19, 27,17tsʰə$^{42}$ zɔ$^{44}$ | ʃi$^{44}$ mə$^{44}$ | 41078 | Jinuo (Youle) | *corpse* |
| 27,25, 27,17 ɕi$^{33}$ ŋɯ̃$^{21}$ | ɕi$^{33}$ mæ$^{33}$ | 80552 | Bai | *carcass/dead animal* |

| | | | | |
|---|---|---|---|---|
| 27,17 | ɕi$^{33}$ mɔ$^{33}$ | 9475 | Nasu | *corpse / dead body* |
| 27,25,27,17 | ɕi$^{33}$ ŋɯ̰̃$^{21}$ ɕi$^{33}$ mæ$^{33}$ | 80552 | Bai | *carcass/dead animal* |
| 28,17 | tsʰo$^{33}$- mo$^{33}$ | 10262 | Yi (Xide) | *corpse* |
| 28.17 | tsʰọ$^{44}$ mo$^{33}$ | 41871 | Lisu | *corpse* |
| 28,17 | tʂha$^{21}$ ma$^{55}$ | 9691 | Nesu | *corpse / dead body* |
| 28,17 | tʂu$^{55}$ mɔŋ$^{55}$ | 17447 | Achang Xiandao | *corpse* |
| 28,19,27,17 | tsʰə$^{42}$ zɔ$^{44}$ ʃi$^{44}$ mə$^{44}$ | 41078 | Jinuo (Youle) | *corpse* |
| 28,27,17 | tshɒ$^{33}$ ʂɿ$^{33}$ mɒ$^{33}$ | 15691 | Lolopho | *corpse / dead body* |
| 28,27 | tsho$^{55}$ si$^{55}$ | 144705 | Hani | *corpse* |
| 28,27 | tshy̩$^{55}$ sɿ$^{55}$ | 144706 | Hani | *corpse* |

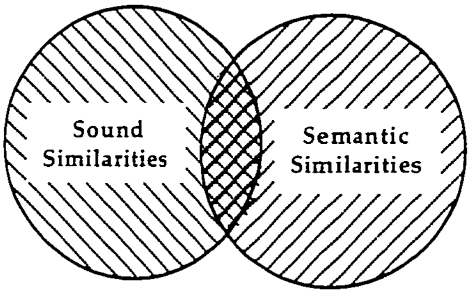
**N.B.** Each letter or number in the *Tagging* identifies with each morpheme of each language form in the associated *Reflex* column an etymon (in the case of numbers) or a morphological category (p = prefix, s=suffix, m (not shown) = as yet unetymologized morpheme).

This excerpt from the KWIC index of polynyms shows compounds built on reflexes of *TB *səy DIE (etymon #27 in STEDT database) and *tsu DEAD, SPIRIT OF THE DEAD (#28); most of the compounds also incorporate a reflex of *TB *s-maŋ (#17) BODY, CORPSE.

### 5.1.4. Computer-aided tagging based on existing reconstructions

Those cases in which form and meaning are most similar present themselves first to the eye as potential cognates; through trial and error the linguist distinguishes cases in which this similarity is the result of a regular correspondence from cases of borrowing, analogy, etc. Large numbers of 'surface similarities' in phonological form *may* indicate some form of 'relatedness', the explanation of which is one of the goals of the comparative method:

(124) *Both borrowed and inherited forms may both be found in the intersection*



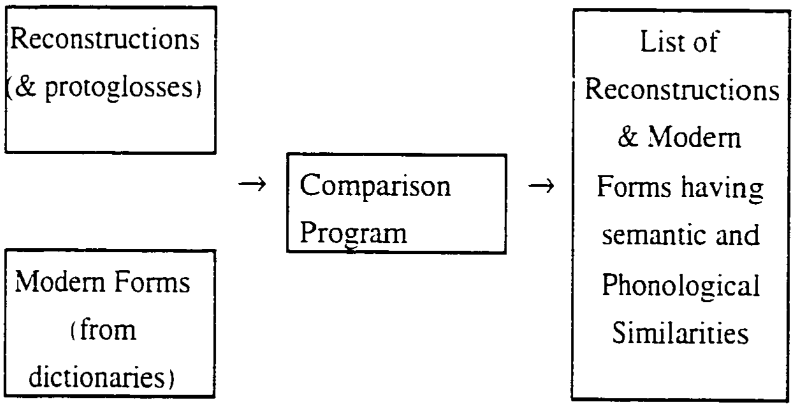
The intersection of sound similarities and meaning contains both cognates (especially in languages without much divergence) and borrowings. Data organized in this way, however, is the starting point for more detailed research and so it behooves us to identify this situation as quickly as possible and get on with the rest of the work of comparison.

Once the concept of 'similarity' in the semantic and phonological dimensions is specified in a way which can be programmed, a computer can make this 'first cut' by bringing together forms which have features in common. Linguists have traditionally made extensive use of 'synonym lists' as a means of bringing together potential cognates (see discussion above in §3.2.2.1). Within a synonym list, the linguist attempts to analyze the forms (i.e. to separate morphemes and undo morphological processes) in order to bring together just those constituents which are suitable for comparison.

Two algorithms are described here. The first, which matches modern forms to a list of possible etyma created beforehand, has been applied to data at the STEDT project in an attempt to assist linguists making the 'first pass' in the reconstruction process[105]; The second version of the algorithm proposes highly simplified etyma on its own.

The first algorithm attempts to match an *existing* list of reconstructions (and protoglosses) with a list of modern forms (and their glosses); the basic dataflow of the program is illustrated in Figure 1.

(125) *Data flow diagram for automated cognate matching*


Reconstructions (& protoglosses)
Modern Forms (from dictionaries)
Comparison Program
List of Reconstructions & Modern Forms having semantic and Phonological Similarities

### 5.1.4.1. Semantic similarity

For programming purposes, a reconstruction and a modern form are *semantically similar* when one or more words used in the protogloss field of a reconstruction matches one or more words in the gloss field of a modern form. In both cases, the glosses are 'normalized', that is, they are translated to uppercase, special characters are translated or removed, leading and trailing blanks removed, etc. Thus the protogloss 'BODY/CORPSE' will match any of the following glosses from modern forms:

(126) *Glosses of modern words containing the strings BODY or CORPSE*

*corpse*
*body*
*carcass/corpse*
*body hair*
*dead body*
*ashes of a dead body*
*carcass (dead body)*

Of course, this is only one of many possible algorithms for judging semantic similarity.

### 5.1.4.2. Phonological similarity

A reconstruction and a modern form are *phonologically similar* when the normalized modern form matches the initial characters of a normalized reconstruction. Here extensive caveats are in order.

The reconstruction may itself reflect several possible phonological forms, as in a PAF or 'pan-allofamic formula' (see 3.2.1.10 above). PAFs are reconstructions which contain optional elements or specify alternative constituents for a particular constituent of the reconstruction. Thus, the reconstruction

(127) *TB *s-maŋ

stands for two possible forms, with and without prefix (*maŋ or *smaŋ). Similarly, parenthesized elements indicate possible alternatives, as in (3):

*(128) Expansion of a PAF*

*(r/l)wa(ŋ/k)

> *rwaŋ
*lwaŋ
*lwak

Note that not all the possible combinations appear here (*rwak is missing). Phonotactic constraints in the protolanguage may disallow some of the combinations implied by the PAF (cf. §3.2.1.10 above). This is purely a 'shorthand' for representing an etymon with several variants. Note also that such proto-variation is not necessarily reflected synchronically in a given language.

Of course, if protoforms are compared directly with modern forms, there would be very few cases in which identity would be found. The segments of the reconstruction evolve into the segments of the daughter forms through a variety of diachronic processes: loss of features, assimilation, dissimilation, etc. The syllable structure may also be affected in the course of time. A comparison of phonological forms must take these processes into account in a systematic way. Here the program implements an algorithm which makes a rather universalist assumption: that the modern forms generally reflect the loss of features and segments.[106] Thus, the program reduces the number of distinctions between segments, and allows segments to be lost in some cases from the protoform. It does this by conflating some segments and by eliminating others. It is worth noting that the 'reduction' described, inasmuch as the elements may be grouped across some phonological oppositions, may also be viewed as *creating* other distinctions.

The details of the normalization performed are documented elsewhere. Here it will suffice to give a couple of examples. First, the program conflates constituents based on the membership in some *a priori* universal class. In general, voicing distinctions are eliminated, and place and manner features are reduced in number. Thus, the [s.z.ç.sh.ʃ.ɕ.ʂ.ʐ.ʒ] are all replaced by /S/[107]. Some segments which consist of several graphemes are rewritten; [ts.dz.tʃ.tɕ.tʂ.dʒ], for example, are all replaced by /C/. Most diacritics and tone marks are removed. Nazalization, represented by a tilde, is replaced by /N/ following the vowel which was nasalized. A number of other violent eviscerations are performed on the data as well which are not described here to spare the reader's sensibilities.

(129) below exemplifies this procedure. Column 1 indicates the language, in this case forms from three dialects of Jinuo, a Lolo-Burmese language, and from Chepang and Newari, which are Himalayish languages[108]; columns 2 and 3 give the form and gloss as cited in the source, and the column 4 shows the result of normalization.

*(129) Normalization of modern forms for comparison with etyma*

| (1) | (2) | (3) | (4) |
|---|---|---|---|
| Language | Form | Gloss | 'Normalized Form' |
| Jinuo (A) | a$^{33}$ mʌ$^{44}$ | *body* | **A MA** |
| | mɯ$^{44}$ tʃ$^{h}$a$^{55}$ | *body hair* | **MA CA** |
| Jinuo (B) | a$^{33}$ mʌ$^{44}$ | *body* | **A MA** |
| | tʃ$^{h}$a$^{55}$ ṃɯ$^{44}$ | *body hair* | **CA MA** |
| Jinuo (Youle) | mə$^{42}$ to$^{44}$ | *body* | **MA TO** |
| | ts$^{h}$ə$^{42}$ zɔ$^{44}$ ʃi$^{44}$ mə$^{44}$ | *body hair* | **CA SO SI MA** |
| Chepang | hmaŋ? | *corpse* | **MAN** |
| Newari | hma | *corpse* | **MA** |
| Newari | si: mɔ | *corpse* | **SI MO** |

An identical normalization procedure is performed on the reconstructions.

### 5.1.4.3. Identification of the cognate morphemes

The program, having first brought together reconstructions and modern forms which have *semantic* similarity, then makes a phonological comparison. It does the comparison for each morpheme (defined simple-mindedly as a blank-delimited string of characters in the normalized form) using a *soundex* function.[109] The soundex system permits forms which differ by a few letters to 'match', especially if the differences are towards the end of the word. Thus two morphemes which differ by a single letter towards the end of the word will often have the same soundex value. As the program proceeds through the morphemes, it creates a parallel list of tokens for the morphemes which match a reconstruction. The token is either the tagging number (identifier) for the reconstruction, or a single letter: *m* (morpheme) in cases where no match is found, *p* (prefix) for single letter initial 'morphemes', and *s* (suffix) for single letter 'suffixes' and the string *BA*$^4$ . (The conventions for tagging are those used for marking etymologies in the STEDT database.) The results of 'tagging' the data in (129) with the reconstruction *s-maŋ would be as in (130).

*(130) Results of 'tagging' modern forms with etyma numbers*

*TB *s-maŋ tag number **123**

Normalized forms: SMAN
MAN

| Form | Normalized Version | Tagging |
|---|---|---|
| 1. a$^{33}$ mʌ$^{44}$ | **A MA** | p.123 |
| 2. mɯ$^{44}$ tʃ$^{h}$a$^{55}$ | **MA CA** | 123.m |
| 3. a$^{33}$ mʌ$^{44}$ | **A MA** | p.123 |
| 4. tʃ$^{h}$a$^{55}$ ṃɯ$^{44}$ | **CA MA** | 123.m |
| 5. mə$^{42}$ to$^{44}$ | **MA TO** | 123.m |
| 6. ts$^{h}$ə$^{42}$ zɔ$^{44}$ ʃi$^{44}$ mə$^{44}$ | ***CA SO SI MA*** | *m.m.99.123* |
| 7. hmaŋ? | **MAN** | 123 |
| 8. hma | **MA** | 123 |
| 9. si: mɔ | **SI MO** | m.m |

Note that none of the forms here matched the s- prefixed form of the reconstruction though the initial aspirates of the Chepang and Newari forms hmaŋ? and hma may be evidence of it. Note also that the final Newari form did not match because the normalized form of the vowel in its second syllable did not quite 'merge' with the other. Both these problems are the result of the insensitive way in which structural and phonological contrasts are treated by the program.

It is a relatively simple matter now to list all the forms for a particular reconstruction. This process is repeated for all reconstructions/

glosses/modern forms in the database. Each of the morphemes can be tagged individually, as illustrated in the italicized Youle form in line 6 of Figure 9 above. In the case where a modern form matches more than one reconstruction the program should record all possible taggings. Handling these possibilities requires some complexification of the tagging structure.

The algorithm described above is very rudimentary. It implements a type of mass comparison: bringing together (large) numbers of semantically and phonologically similar forms. It may be a useful tool for initially organizing the data, especially new data not before analyzed, in a way that it can be more easily used by the linguist. It is likely to be able to link some forms from languages which are closely related and have not diverged much, but is likely to fail with more distantly related languages. It will not distinguish borrowings from true cognates. It is as likely to make incorrect assignments as it is to make correct assignments of forms to reconstructions. Since the list of reconstructions to which the modern forms are assigned is a closed set no new possible cognate sets will be created. Except for sound symbolic forms, words from different languages do not have both similar form and meaning unless there is a historical relationship of *some* sort; after all, the relationship between the signifier and the signified is arbitrary. The task of the comparative method is to apply controls to this process.

### 5.1.5. Proposing a 'simplified reconstruction'

If no list of reconstructions were available beforehand, the machine could propose the 'reduced' forms as possible reconstructions. In this case, the machine would simply number all the 'reduced' morphemes, assigning the same value to those that were identical or 'appropriately' similar. It would be a bit of a challenge to make a 'reduced' protoform whose soundex value matched a set of possible cognates (as illustrated in the 'automatic' cognate set below).

*(131) 'Automatic' Protoform generation*

| (1) | (2) | (3) | (4) | (5) |
|---|---|---|---|---|
| Language | Form | Gloss | Normalized Form | Tagged Form |
| Jinuo (A) | a$^{33}$ mʌ$^{44}$ | *body* | **A MA** | p.1 |
| | mɯ$^{44}$ tʃʰa$^{55}$ | *body hair* | **MA CA** | 1.2 |
| Jinuo (B) | a$^{33}$ mʌ$^{44}$ | *body* | **A MA** | p.1 |
| | tʃʰa$^{55}$ ṃɯ$^{44}$ | *body hair* | **CA MA** | 2.1 |
| Jinuo (Youle) | mə$^{42}$ to$^{44}$ | *body* | **MA TO** | 1.3 |
| | tsʰə$^{42}$ zɔ$^{44}$ ʃi$^{44}$ mə$^{44}$ | *body hair* | **CA SO SI MA** | 2.4.5.1 |
| Chepang | hmaŋ? | *corpse* | **MAN** | 1a |
| Newari | hma | *corpse* | **MA** | 1 |
| Newari | si: mɔ | *corpse* | **SI MO** | 5.1b |

Column 1 indicates the language, in this case forms from three dialects of Jinuo, a Lolo-Burmese language, and from Chepang and Newari, which are Himalayish languages[110]; columns 2 and 3 give the form and gloss as cited in the source, and the column 4 shows the result of normalization. Column 5 'tokenizes' each 'simplified reconstruction' with a number for reference. Further relaxation of matching constraints (i.e. 'fuzzy' vowels, alternation of finals) may be allowed.

(132) *An 'Automatic' cognate set*

| | |
|---|---|
| ***M[A/O]N** | **BODY/CORPSE** |
| | |
| **1.** | **MA** |
| **1a.** | **MAN** |
| **1b.** | **MO** |

| | | | |
|---|---|---|---|
| Jinuo (A) | a$^{33}$ | mʌ$^{44}$ | *body* |
| | | mɯ$^{44}$ tʃʰa$^{55}$ | *body hair* |
| Jinuo (B) | a$^{33}$ | mʌ$^{44}$ | *body* |
| | tʃʰa$^{55}$ | ṃɯ$^{44}$ | *body hair* |
| Jinuo (Youle) | | mə$^{42}$ to$^{44}$ | *body* |
| | tsʰə$^{42}$ zɔ$^{44}$ ʃi$^{44}$ | mə$^{44}$ | *body hair* |
| Chepang | | hmaŋʔ | *corpse* |
| Newari | | hma | *corpse* |
| Newari | si: | mɔ | *corpse* |

While this approach is appealing it has a number of problems. For example, it is possible that the reduction in distinctions is so severe that chance matching in form and meaning will obscure the result: there are a number of words meaning 'body hair' in TB languages with bilabial nasal initials and occluded rhymes which are reconstructed as *mul = (s-)mul ~

(s-)mil ~ (r-)mul [STC 2, Benedict 1972:15]. These might be erroneously included in the above set. It is likely that the other morpheme in the Jinuo forms descend from *tsam 'hair (head)' [STC 73, Benedict 1972:]. If this is the case, these compounds would be formed of two morphemes meaning hair. This is not unreasonable, of course. Examination of other compounds in the language and in other related languages would shed more light on the matter. The point is that 'surface similarities' may have led the computer program astray here. Precise correspondences are needed to prevent embarrassment.

A number of refinements might be proposed for the program. Most of these refinements would be in the form of heuristics to make the matching more 'realistic' given what is known about the semantic and phonological qualities of the languages being studied. So far, it has not been possible to predict the direction of sound change and semantic shift. Until it is possible to do so, any program attempting to make such matches will be severely limited. As a means of breaking down the complexity of representations, it has a place in the repertoire of algorithms for historical research.

### 5.1.6. Decomposition of morphemes into phonological constituents

Once the database entries have been analyzed into morphological components, a detailed phonological analysis of each morpheme is possible. Of course, such analysis is neither trivial nor transparent; some procedure, whether carried out by a computer or by a human analyst, must be applied to each element to supply the appropriate analysis.

The bracketing conventions and bracketed forms in §5.1.2.2 above should be seen not as independent entities but as part of an integrated scheme of analysis which allows access to both the content and the structure of representations. The example below (from Guthrie's (1967) reconstruction of Proto-Bantu) illustrates how a complex lexical entry can be broken down into hierarchical, fine-grained morphological and phonological elements of the types described above (in §5.1.2) for further analysis. Note that the only part of the description which is from Guthrie is in 'Level 0' (to use the MARIAMA terminology); the other levels are *projections* from this data element based on (hypothetical) automated morphophonological analysis.

(133) *After Guthrie 1967: #600a dioŋodioŋo* giddiness *(Class 7, D:5; NW East)*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | L660a (=dioŋodioŋo) | | | | | | | | | |
| 1 | M226 (= dioŋo) | | | | | M226 (= dioŋo) | | | | |
| 2a | S120 (= dio) | | | S213 (= ŋo) | | S120 (= dio) | | | S213 (= ŋo) | |
| 2b | H | | | L | | H | | | L | |
| 3a | C1 | V11 | V12 | C2 | V21 | C1 | V11 | V12 | C2 | V21 |
| 3b | d | i | o | Ng | o | d | i | o | Ng | o |
| 4a | +stop | | | +stop | | +stop | | | +stop | |
| 4b | +dent | | | +vel | | +dent | | | +vel | |
| 4c | +vcd | +vcd | +vcd | +vcd | +vcd | +vcd | +vcd | +vcd | +vcd | • |
| 4d | | +high | +mid | | +mid | | +high | +mid | | • |

**Legend**

N.B.: since this is a reduplicated syllable, the features of the last syllable in tiers 4c and 4d have been elided to save space.

'L660a' refers to the original lexical entry in the dictionary, and identifies a full lexeme, here a reduplicated reconstructed form.

'M226' means morpheme #226; i.e. each morpheme is individually identified and distinguished. Morphemes are subcategorized by part of speech, semantic field, noun or verb class, and so on in other parts of the database record.

'S120' and 'S213' are numbered syllables. Like morphemes, each syllable in each language is uniquely identified and indexed (elsewhere in the database).

H = High; L = Low

$C_1$, $V_{12}$, etc. are syllable slot identifiers; '$V_{12}$' means 'second vowel of 1st syllable'. Additionally, elements like '$C_i$' (= initial consonant), '$C_f$' (=final consonant), and 'C' (= any consonant) can serve to define 'metaslots.'

**'Tier' descriptions**

| | | |
|---|---|---|
| 1 | 'Morphology' | |
| 2 | 'Syllable' | 2a 'Foot', or 'segmental tier' 2b Tone |
| 3 | 'Segmental' | 3a Syllable Slot, 3b Grapheme |
| 4 | 'Feature' | 4a Manner, 4b Place, 4c Voicing; 4d Height |

Such a representation would allow the instant identification of words matching some particular phonological description via regular expressions which refer to the elements of various levels of representation, e.g.:

*(134)*

| To Find: | Search for: |
|---|---|
| 'Cases of reduplication' | $M_i = M_j$ for all i, for all j = i + 1 |
| 'Initial /d/ followed immediately by /i/' | $C_i$ = /d/ ^ $V_{11}$ = /i/ |
| 'lexemes containing dental stops' | Tier 4a = [+stop] ^<br>Tier 4b = [+dental] |
| 'a labial followed by a velar' | Tier 4b of $C_j$ = [labial] ^ Tier 4b of Ci+i = [velar], for all i |

Searching of the 'segmental and lower' tiers (3 and 4 in (133) above) could be accomplished using algorithms such as those developed for 'feature analysis' in the Reconstruction Engine (Lowe and Mazaudon 1994) which describe a scheme which permits the coding of feature representations for the individual constituents, which in turn are propagated to any morphemes containing those constituents. This would reduce the work and storage requirements for searching at levels 3 and 4.

### 5.1.7. Semantic classification and 'synsets'

The glosses of forms in lexicographic databases are hardly sufficient for use as a means to bring forms together for diachronic analysis. This presents us with a conundrum as we consider how to recode the meanings of our data for searching: we will want to retain the specific meanings of words for some purposes, but also to be able to lump words with similar meanings together to search for shifted meanings. One way to do this is to perform a classification of the data into semantic fields. The prototypical instance of such a classified lexicon is the thesaurus built on semantic principles. There are a number of these, for no one pretends that semantic classification is something that can be done once and for all and forgotten.

Buck, when talking about the process of organizing a semantically-based thesaurus, discusses the approaches taken by an number of his predecessors and observes:

> Of all these [works], no two, apart from direct imitations, will agree in the order of classification. For example, Pollux thought proper to begin with the gods (so in many lists), passing to man (with the parts of the body), relationship, science and art, hunting, meals, trades, law, town organizations, utensils. Aelfric began with agricultural tools, passing to men (by office or graft), diseases, law, insects, vessels, drinks, birds, plants, trees, arms, winds, cereals, clothes, physical world, parts of the body, colors. But actually, all sorts of miscellaneous items are mixed in. In Roget's Thesaurus the parallelism of opposites and some minor subdivisions may be convenient. But the main groups and larger subdivisions are so

> comprehensive as to have no obvious coherence. What may one not find under Motion (e.g. eat, food) or Volition (e.g. clean)! The fact is, of course, that relations are too complex to admit any truly scientific and complete classification (cf. the remark of Jespersen 1924) (Buck 1949:xiii)

Bringing a number of meanings (or more precisely lexical items expressing meanings) into a machine-processable set is a technique developed in the artificial intelligence and natural language processing world as a means of encoding lexical semantic relations (Miller 1990). The WordNet approach uses (among other structures) a (monolingual) list of synonyms called a SYNSET to instantiate the common semantic link between dictionary headwords. Synsets must not be confused with synonym sets, such as those mentioned in §3, which are lists of words with the 'same' meaning in different languages. A synset in the WordNet sense from which the idea is drawn is a 'set of words which is substitutable in some context.' The following (135) are synsets from the WordNet database. They are (and will be) represented between parentheses as a convention.

*(135) Synsets instantiate the synchronic notion of synonymy between dictionary headwords; the context is not given here*

(fawn,crawl,creep,cringe,cower,grovel)

(dart,dash,scoot,scud,flash,shoot)

(smash,dash,break into pieces)

Here, the definition will be extended for cross-linguistic comparison as follows. A synset will be:

- a synset more or less in the sense given above; that is, a list of words in a single language which either 1) refer to the same (tangible or intangible) thing, to the extent that this is possible or 2) are substitutable for each other; or,

- a set of words in the metaglossing language or languages which share a core meaning diachronically, or

- a set of words in the metaglossing language or languages used to gloss lexical items in the target languages which are sometimes etymologically related.

The formalism is borrowed and introduced here as a means of representing *diachronic* semantic relations. I stress that it is principally the *formalism*, especially as a machine-processable representation that is being borrowed. It is also used as an *ad hoc* facility for controlling homonymy in reconstruction in the Reconstruction Engine, discussed in §6.6.1 below. Some examples of the diachronic specification of synsets is shown in (136) below. Note that specification is based on the premise that the metaglossing language is English and the target languages are Tibeto-Burman.

*(136) Possible diachronic synsets*

| | Synset | Linking concept/etymon |
|---|---|---|
| (a) | (EYE, EYEBROW, EYELID, DIE, ANKLE ...) | 'eye' |
| (b) | (EYEBROW, HAIR, BEARD, FUR, ...) | 'hair' |
| (c) | (SIT, GOODBYE, LEAVE, ...) | 'departure' (cf. §3.1) |

In (a) above, it is obvious that words glossed EYEBROW and EYELID in the metaglossing language (English) might contain morphemes in the target languages meaning 'eye'. ANKLE is included for reasons discussed below (§5.1.7.2). DIE is included here because the concept is expressed in some TB languages with a compound analyzable as 'close eyes;' in (b) words glossed *hair* or *fur* (and *feathers* and so on) contain a common etymon.

The synchronic and diachronic dimensions of the semantic distinctions being made here are not specified, nor are the accidents of the semantics and derivational morphology encoded in the glosses distinguished. One can imagine ways to do this, but for the practical purposes of grouping lexical items together for analysis it is not necessary. The composition of such synsets may at some point be an interesting area of study.

#### 5.1.7.1. Semantic classification

Of course, it would be possible to create such synsets beforehand, based on one's knowledge of the languages and their history. However, there are several reasons why this is not a good idea:

- the words used in the synsets would have to be coordinated with the words used to gloss the lexical data sets being treated, otherwise the classification and the glosses would not match

- many relations, both synchronic and diachronic, would be missed if not prodded by noticing the fact that certain words or collocations are used in glosses

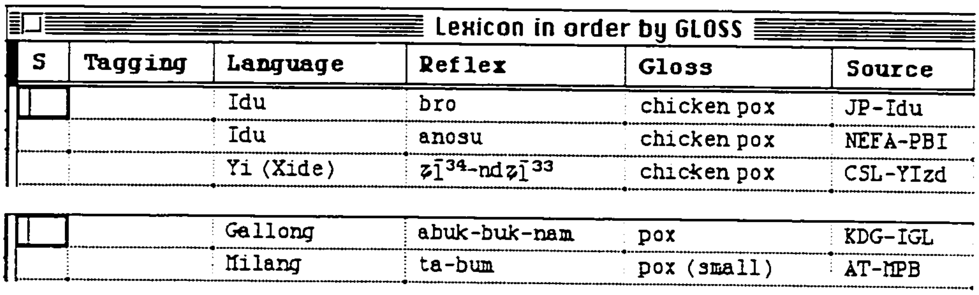
Therefore, the classification should be done on the basis of words and collocations in the data sets being treated.

This is being done for the STEDT database, and the method and results are discussed in Appendix 8.

On one hand, it is useful to have a universal classification of all meanings into a consistent hierarchical structure, and on the other one needs to record the detailed semantic facts of relationship, such as details about words meaning ANKLE in the target languages. The implication is that, for practical purposes at least, a hybrid structure of explicit classification (à la Buck) and implicit linking (as in the synset formalism) is needed. I will first treat the problem of creating a hierarchical semantic over a large data set.

The problem is simple to state. We wish to classify all forms in the data set according to our *a priori* classification scheme (see Appendix 8 or the table of contents of Buck (1949) or Roget (Mawson 1911) for an example of such a scheme). It is sufficient if we assign to each form in the database a number or code representing its place in the semantic classification. So for example, we would wish to classify the following entries (137) into the *Diseases* category, perhaps subdivided into *Diseases (communicable)*. We could do this by hand, assigning each form to the appropriate category. But in a large database, this is time-consuming and error-prone: the first three words sort under C, while the others sort under P. Would we be sure to assign them to the same category, especially if there are a number of possible places or levels to assign them in the hierarchy?

*(137) 5 lexical entries from STEDT Database*



Lexicon in order by GLOSS

| S | Tagging | Language | Reflex | Gloss | Source |
|---|---|---|---|---|---|
| □ | | Idu | bro | chicken pox | JP-Idu |
| | | Idu | anosu | chicken pox | NEFA-PBI |
| | | Yi (Xide) | ʑɿ̄$^{34}$-ndʑɿ̄$^{33}$ | chicken pox | CSL-YIzd |
| □ | | Gallong | abuk-buk-nam | pox | KDG-IGL |
| | | Milang | ta-bum | pox (small) | AT-MPB |

Note that in the five lexical entries above (for five different lexemes in four languages), there are three glosses:

*(138)*

CHICKEN POX
POX
POX (SMALL)

and three words used in these glosses:

*(139)*

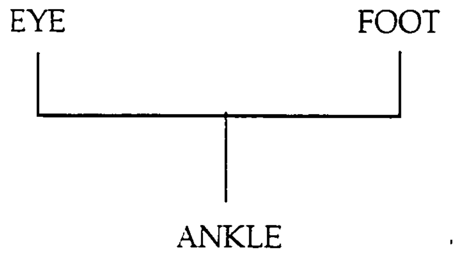
CHICKEN
POX
SMALL

In the database as a whole, the gloss *chicken pox* (as a two-word phrase) occurs three times (and these are shown in (137)); elsewhere the word *chicken* alone occurs 66 times, and *chickenpox* (as one word) occurs once. Using the words themselves (as described in Appendix 8) we can approximately classify all the glosses, though of course there will be ambiguity and overlap. Nevertheless, as a first cut, this approach is useful.

### 5.1.7.2. Metastatic flowcharts

Sets of glosses listed in synsets like those given in (136) above bring together words which fall into the same semantic field without having to specify exactly what that field is or what its limits are. Synsets can be linked by linking members of synsets (an 'implicational linking') or by Using Matisoff's graphical formalism, which he calls *metastatic flowcharts* (Matisoff 1978). The relationship between both synsets and glosses in synsets can be named, and the nature of the relationship specified. For example, morphemes meaning *eye* and morphemes meaning *foot* are often combined in Tibeto-Burman languages into a compound meaning ankle (i.e. the ankle is the eye (round protruding object) of the foot). Matisoff expresses this relations graphically in a metastatic flowchart:

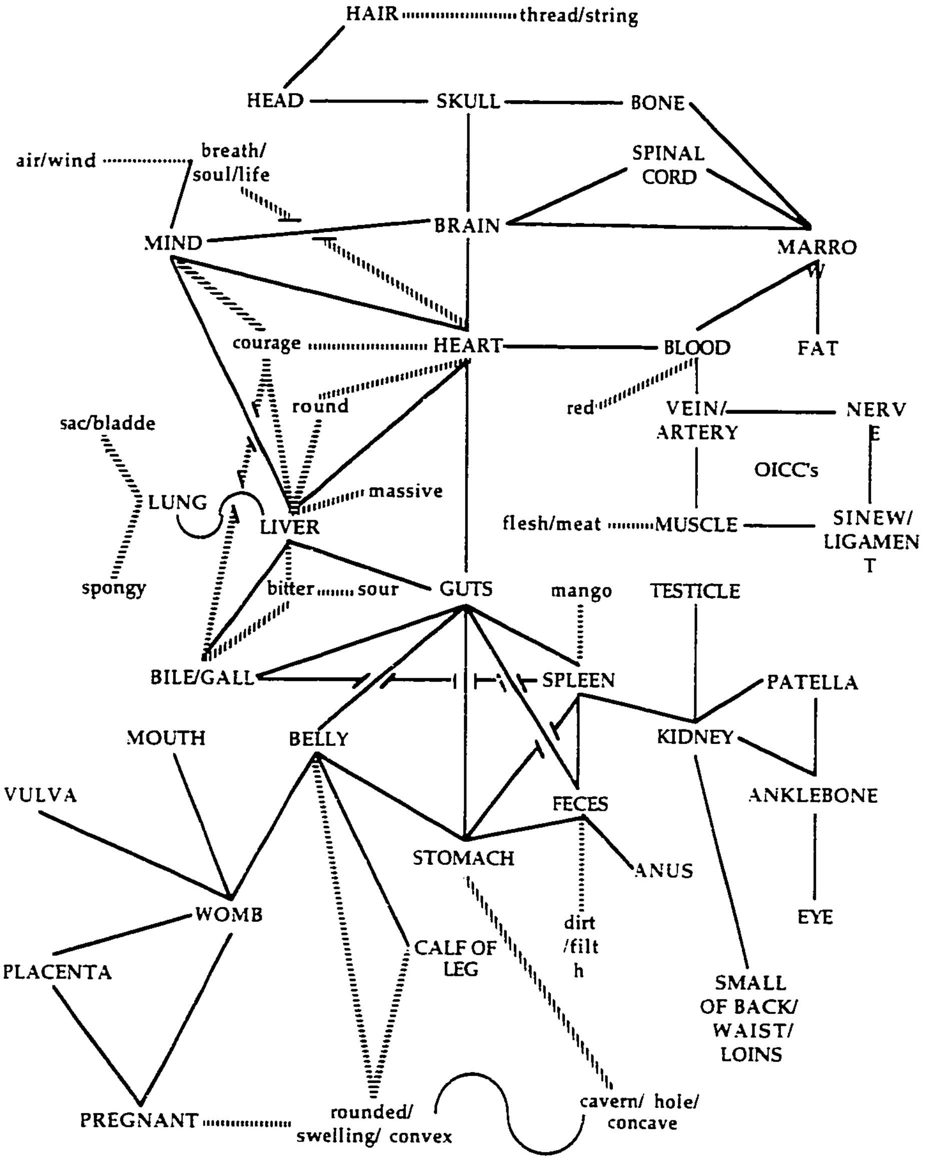
*(140) A simple metastatic flowchart: 'ankle'*



This connection is the basis for research into compounds of the sort discussed in §5.1.2 above.

Different types of links (usually metaphorical extensions) can be expressed by different styles and shapes of linking lines, as shown below in (141). In this diagram, for example, the relation of antonymy is encoded using a curved connector, as shown between ROUNDED/SWELLING/CONVEX and CAVERN/HOLD/CONCAVE.

*(141) A more complicated metastatic flowchart: the Sino-Tibetan Body*



There is a direct relationship between synsets and their graphical representation as metastatic flowcharts; the synset formalism represents a graph structure like the flowcharts. The flowchart metaphor is easy for people to grasp conceptually and would be the obvious choice for an interface to study and manipulate the underlying semantic relations. The possibility is discussed further in the database design chapter §7.

### 5.1.7.3. Problems with semantic classification

There are a number of problems with this scheme in some cases requiring quite *ad hoc* solutions. Some of these problems are structural, and cannot be corrected by improved or refined implementations. Others are accidents of the process and can be corrected by hand or by the use of heuristics.

1) Words which are high in frequency and low in information (and which would thus normally occur on a stop list) must be handled. For some, their information content varies depending on their context so they cannot be suppressed unconditionally. In the semantic classification of the STEDT database, they were assigned (temporarily) to a special 'dead end' category; such words would not be used in further semantic processing.

2) The categorization of gloss-words into the major grammatical categories of the glossing metalanguage (i.e. English) can be somewhat problematic for the target languages, in this case Tibeto-Burman languages. For example, the distinction between adjective and verb is often irrelevant (*cold* and *be cold* are often the same lexical item).

4) If a word is clearly homonymous or homographic in English (e.g., 'tear'), several distinct entries will eventually be needed.

5) For sets like BEAT/BEATEN/BEATING, where a single lexeme occurs in several inflected forms straddling several parts of speech, one semantic classification is chosen (e.g., V act). (Other examples: BOIL/BOILED/BOILING: V act; BRIGHT/BRIGHTNESS: V adj.). These forms also tend to clog synsets with uninformative repetition; some means should be developed to morphologically collapse forms in the metaglossing languages.

## 5.2. Creating Tables of correspondences

One useful function which might be performed automatically by the computer is the *initial* positing of sound laws. After all, the comparative method relies in part on the numerical strength of correspondences for its efficacy and this can be estimated by computer.

### 5.2.1. Synopsis of approach

The research scheme presented here is:

- Creation of tables compatible with the Reconstruction Engine tables of correspondences for the three data sets described in Appendix 3.2. The tables were provided in advance for ILH (and are analyzed here), computer-generated for CK, and merely conjectured for TSR, about which only a few remarks are offered here (no analysis is presented).

- Processing of the data sets with the Reconstruction Engine and comparison of the reconstructions produced thereby with those in the literature (for sets which already have a proposed reconstruction).

### 5.3. Characteristics of the data set used

The data set for Northern Loloish languages provided in (Chen 1986a) has several distinct characteristics. First, the data constitute a carefully analyzed set: in almost all cases the synonym sets contain witnesses from all eight Yi dialects. Further, in each case any 'extraneous' morphemes have been stripped away, leaving only the most likely candidate for cognacy. This 'pre-treatment' of the data is clear when the data in Chen's comparative set is compared with the fuller data he and others have contributed to the STEDT database:

(142) *Comparative set for* EYE *in eight Yi languages (Chen 1986a)*

| | |
|---|---|
| sani | ne$^{44}$ |
| axi | nḛ$^{33}$ |
| nesu | nḛ$^{33}$ |
| lalo | ʔmɪ̠$^{33}$ |
| li | mḛ$^{33}$ |
| nasu | na̠$^{2}$ |
| neisu | na$^{33}$ |
| nosu | ɳo̠$^{4}$ |

(143) *Words for* EYE *in five Yi languages (from the STEDT database (Matisoff forthcoming))*

| | | | |
|---|---|---|---|
| Sani (Maa)=Nyi | ne$^{44}$ sɿ$^{11}$ | DQ-SANIMA 320.3 | [16891] |
| Ahi | nḛ$^{33}$ sɑ$^{21}$ | LMZ-AY 3.4 | [47313] |
| Nesu | ɳɹ̠$^{33}$ du$^{33}$ | CK-NESU 3.4 | [9704] |
| Lalo | ʔmɪ̠$^{33}$ tse$^{21}$ | CK-LALO 3.4 | [9309] |
| Nasu | na̠$^{21}$ du$^{33}$ | CK-NASU 3.4 | [9489] |

Chen's data are rather narrowly transcribed and incompletely phonemicized; and in fact the data appear insufficient to permit a proper phonemicization. So in Nasu , for example, it is fairly clear that [ȵ] is the allophone of /n/ occurring before /i/ and perhaps /e/ and /ɛ/ as well:

*(144)*

| *N* | *Gloss* | *sani* | *nasu* | *neisu* |
|---|---|---|---|---|
| 4. | buckwheat | qa$^{21}$ | ngho$^{33}$ | ngu$^{33}$ |
| 7. | chew | ga$^{21}$ | ngho$^{33}$ | ngu$^{33}$ |
| 39. | ill | na$^{33}$ | no$^{21}$ | no$^{21}$ |
| 29. | ask | na$^{33}$ | no$^{33}$ | no$^{13}$ |
| 54. | stop | na$^{55}$ | no$^{33}$ | nu$^{33}$ |
| 112. | soft | no$^{21}$ | nu$^{33}$ | nu$^{33}$ |
| 131. | cry | ŋɯ$^{33}$ | ŋɯ$^{33}$ | ny$^{13}$ |
| 192. | press | ȵɤ$^{44}$ | ȵi$^{2}$ | ȵi$^{33}$ |
| 121. | day | ni$^{33}$ | ȵi$^{21}$ | ȵi$^{21}$ |
| 23. | two | ni$^{21}$ | ȵi$^{55}$ | ȵi$^{55}$ |
| 138. | frost | ŋɪ$^{33}$ | ȵe$^{33}$ | ȵi$^{13}$ |
| 89. | brain | no$^{2}$ | nǫ$^{55}$ | ne$^{13}$ |
| 69. | bean | no$^{44}$ | nǫ$^{2}$ | ȵe$^{33}$ |
| 13. | near | næ$^{21}$ | nɔ$^{33}$ | nɛ$^{33}$ |
| 147. | smell | nɤ$^{21}$ | ne$^{33}$ | ȵɛ$^{33}$ |

However, this is not a strong enough basis to justify a full-blown retranscription of the entire data set. I have therefore elected to use the data just as supplied, and to account for all variation via a diachronic analysis using the Reconstruction Engine.

The Northern Loloish data used in this study came pre-arranged by gloss. Indeed, it is often the case that researchers pre-arrange their data into *synonym sets* for comparison, and often members of such a set are actually cognates. I propose an algorithm here which takes advantage of this fact to build correspondence sets. The algorithm is similar to ones used or proposed in §2; it has several advantages over these previous approaches:

1) no alignment of segments is required, and

2) the constituent inventory is not limited to segments.

It does, however, place other restrictions on the user/programmer:

1) An algorithm must be devised which divides each morpheme in each language into comparable constituents. Of course, such an algorithm encodes diachronic and synchronic phonological structures known beforehand to the linguist. This is exactly what most other programs have the linguist do by hand in the data preparation step.

Consider the following data (synonym sets from Chen 1986a):

(145)

| set# | gloss | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 972. | *eye* | ne$^{44}$ | nę$^{33}$ | nę$^{33}$ | ʔmɿ$^{33}$ | mę$^{33}$ | na̱$^{2}$ | na$^{33}$ | ɲǫ$^{4}$ |
| 30. | *black* | ne$^{44}$ | nę$^{33}$ | nę$^{33}$ | nɿ$^{33}$ | nę$^{33}$ | na̱$^{2}$ | na$^{33}$ | nǫ$^{4}$ |
| 339. | *chicken* | ze$^{44}$ | zę$^{33}$ | zę$^{33}$ | zɿ$^{33}$ | zę$^{33}$ | ɣa̱$^{2}$ | ɣa$^{33}$ | va̱$^{4}$ |
| 196. | *pig* | ve$^{2}$ | vę$^{2}$ | vę$^{2}$ | vɿ$^{2}$ | vę$^{2}$ | va̱$^{55}$ | va$^{13}$ | vǫ$^{55}$ |

Obviously a number of constituents occur here 'in parallel' and almost undoubtedly cognate. A human being could list the correspondences straightaway. To have this done by a computer, however, we must provide a algorithm to divide the forms up into the appropriate constituents for comparison; then the correspondence sets could be extracted directly, as shown in (8) below:

(146) *words in CK set 972 'eye' divided according to a subgroup-specific syllable canon I+R+T*

972. *eye*

| Initial | Rhyme | Tone |
|---|---|---|
| n | e | $^{44}$ |
| n | e̱ | $^{33}$ |
| n | e̱ | $^{33}$ |
| ʔm | i̱ | $^{33}$ |
| m | e̱ | $^{33}$ |
| n | a̱ | $^{2}$ |
| n | a | $^{33}$ |
| ɳ | o̱ | $^{4}$ |

In fact, on a language group by language group basis, it is not difficult in general to specify criteria for dividing forms into their constituent parts. If the symbols used in representing the form have only one use in the transcription, and there are no discontinuous constituents or glyphs which serve multiple functions, the algorithm is simple: look for adjacent symbols of a particular class and group them into a constituent, a typical application for a finite-state transducer. This is illustrated in the following table:

(147) *Dividing linearly contiguous representations into constituents (Lalo* ?mi̠$^{33}$*'eye' )*

| Form | | | | | | Level |
|---|---|---|---|---|---|---|
| ? | m | _ | i | $^{3}$ | $^{3}$ | Glyphs |
| I | I | R | R | T | T | Class |
| ?m | | i̠ | | $^{33}$ | | Constituents |
| I | | R | | T | | Slot |

| | |
|---|---|
| Initial Class glyphs: | I = n.m.?.ɲ̥.h, etc. |
| Rhyme Class glyphs: | R = i.e.o._. etc. |
| Tone Class glyphs: | T = $^{1}.^{2}.^{3}.^{4}.^{5}$ etc. |

Having defined the process, we can run through hundreds or thousands of forms looking for this pattern; if this process is carried out for each form in each row in (145) above, and the corresponding elements extracted, we can then compare rows which have the same or similar constituents, as shown in (148):

*(148) Northern Loloish cognate sets divided and exploded into correspondence sets (already cited in (145) above).*

*eye*

| set# | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 972. | I | n | n | n | ʔm | m | n | n | ɳ |
| 972. | R | e | e̱ | e̱ | i̱ | e̱ | a̱ | a | o̱ |
| 972. | T | 44 | 33 | 33 | 33 | 33 | 2 | 33 | 1 |

*black*

| set# | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 30. | I | n | n | n | n | n | n | n | n |
| 30. | R | e | e̱ | e̱ | i̱ | e̱ | a̱ | a | o̱ |
| 30. | T | 44 | 33 | 33 | 33 | 33 | 2 | 33 | 1 |

*chicken*

| set# | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 339. | I | z | z | z | z | z | ɣ | ɣ | v |
| 339. | R | e | e̱ | e̱ | i̱ | e̱ | a̱ | a | a̱ |
| 339. | T | 44 | 33 | 33 | 33 | 33 | 2 | 33 | 1 |

*pig*

| set# | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 196. | I | v | v | v | v | v | v | v | v |
| 196. | R | e | e̱ | e̱ | i̱ | e̱ | a̱ | a | o̱ |
| 196. | T | 2 | 2 | 2 | 2 | 2 | 55 | 13 | 55 |

The rows thus generated can now be compared. Duplicate rows are consolidated (i.e. these represent correspondence sets supported by several cognate sets); a count of the number of supporting sets is kept, and the sets supporting a particular correspondence are linked together.

The real world is rarely as neat as theory would suggest. In order to eliminate some of the noise which results from small variations between rows a row comparison algorithm brings similar rows (and their supporting sets) together, as shown in (149) below. Rows which have mostly similar outcomes are listed together.

*(149) CK Initial correspondence #20*

| # | N | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 20a | 2 | ph | ph | ph | ph | ph | ph | ph | - |
| 20b | 1 | ph | ph | ph | f | ph | ph | ph | ph |
| 20c | 1 | ph | ph | ph | ph | ph | ph | ph | ph |

*(150) Cognate sets supporting the correspondences above*

| set# | Gloss | Corrs. | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 115. | price | 20.71.75. | phu$^{21}$ | phu$^{21}$ | phu$^{33}$ | fi$^{21}$ | phɯ$^{21}$ | phu$^{33}$ | phi$^{33}$ | phu$^{33}$ |
| 99. | shell | 20.65.78. | phɤ$^{2}$ | pho̜$^{2}$ | phɣ̩$^{2}$ | phy$^{2}$ | phɣ$^{2}$ | phɣ̩$^{55}$ | phi$^{13}$ | ? |
| 186. | swell | 20.65.78. | phɤ$^{2}$ | pho̜$^{2}$ | phɣ̩$^{2}$ | phy$^{2}$ | phɣ$^{2}$ | phɣ̩$^{55}$ | phi$^{13}$ | ? |
| 95. | vomit | 20.64.78. | phɪ$^{2}$ | phi̠$^{2}$ | phi̠$^{2}$ | phi̠$^{2}$ | phi̠$^{2}$ | phi̠$^{55}$ | phi$^{13}$ | phɪ̠$^{55}$ |

In fact, it is possible to suggest some simple and not-too-dangerous heuristics for combining (at least for the time being) rows which differ only slightly. So rows were merged according to the following criteria:

- If two rows were identical except for missing data they were merged. Thus 20a and 20c in (149) above merge into a single row.

- Two rows differing from each other by only one constituent, are merged into one. Thus 20b in (149) above merges with the previous merged 20a,c; the outcomes in Lalo are merged into a single cell.

The three correspondence lines above (151) are therefore merged:

*(151) Merged correspondence row for data in (149) with reconstruction*

| # | N | *NL | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 20 | 1 | *p | ph | ph | ph | ph.f | ph | ph | ph | ph |

It is possible for two cognate sets to dovetail and fill in each other's empty cells. The two sets below (152) have the same rhyme and tone correspondence according to this algorithm.

*(152)*

| set# | gloss | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 64 | SHY | to$^{33}$ | tu$^{33}$ | ta$^{55}$ | tu$^{55}$ | ta$^{33}$ |  | to$^{13}$ | to$^{33}$ |
| 104 | TURN | tṣo$^{33}$ |  | tsa$^{55}$ | tṣu$^{55}$ | tṣa$^{33}$ | tṣɔ$^{33}$ | tṣo$^{13}$ | tɕo$^{33}$ |

Such a conflation need not be justified on a diachronic basis; rows may be merged in this way when the variation observed is due to some synchronic factor and should not be carried back into the protolanguage. For example, the synchronic alternation of /n/ before high vowels (giving /ɳ/ as shown in (144) above) results in a large number of correspondence rows. A conflation algorithm which operates only on rows having a single difference does poorly here. Besides, evidence from other reconstructions of Loloish show that at least two different *initials are mixed up here (153).

(153) *CK Initial correspondence #16*

| N | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|
| 1 | n | n | n | m | m | ɳ | ɳ | ȵ |
| 3 | n | n | n | n | n | n | n | n |
| 1 | n | n | n | n | n | n | n | ȵ |
| 2 | n | n | n | n | n | n | ȵ | n |
| 1 | n | n | n | ?m | m | n | n | ȵ |
| 1 | n | n | n | ?n | n | n |  | ȵ |
| 1 | n | n | n | ?n | n | n | n | ṇ |
| 2 | n | n | n | ?n | n | n | n | n |
| 1 | n | ȵ | ȵ |  | ȵ | n | n |  |
| 1 | ȵ | n | n |  | n | ȵ | ȵ | ȵ |
| 1 | ŋ | n | n |  | n | ȵ | ȵ |  |

(154) *Cognate sets supporting the correspondences above*

| set# | Gloss | Corrs. | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| ***L** | ***my** |  |  |  |  |  |  |  |  |  |
| 161. | monkey | 16.68.78. | no$^{2}$ | nu̱$^{2}$ | nu̱$^{2}$ | mo̱$^{2}$ | mu̱$^{2}$ | ɳo̱$^{55}$ | ɳe$^{13}$ | ȵu̱$^{55}$ |
| 149. | eye | 16.62.75. | ne$^{44}$ | ne̱$^{33}$ | ne̱$^{33}$ | ?mi̱$^{33}$ | me̱$^{33}$ | na̱$^{2}$ | na$^{33}$ | ȵo̱$^{4}$ |
| ***L** | ***n** |  |  |  |  |  |  |  |  |  |
| 39. | ill | 16.58.76. | na$^{33}$ | no$^{2}$ | no$^{21}$ | na$^{55}$ | no$^{33}$ | no$^{21}$ | no$^{21}$ | na$^{33}$ |
| 66. | black | 16.62.75. | ne$^{44}$ | ne̱$^{33}$ | ne̱$^{33}$ | ni̱$^{33}$ | ne̱$^{33}$ | na̱$^{2}$ | na$^{33}$ | no̱$^{4}$ |
| 112. | soft | 16.70.75. | no$^{21}$ | no$^{21}$ | nu$^{33}$ | nu̱$^{21}$ | nu$^{21}$ | nu$^{33}$ | nu$^{33}$ | nu$^{33}$ |
| 13. | near | 16.55.75. | næ$^{21}$ | na$^{21}$ | nɛ$^{33}$ | ne$^{21}$ | nɛ$^{21}$ | nɔ$^{33}$ | nɛ$^{33}$ | ȵi$^{33}$ |
| 69. | bean | 16.68.75. | no$^{44}$ | nu̱$^{33}$ | nu̱$^{33}$ | no̱$^{33}$ | nu̱$^{33}$ | no̱$^{2}$ | ȵe$^{33}$ | nu̱$^{4}$ |
| 147. | smell | 16.66.75. | nɤ$^{21}$ | nɯ$^{21}$ | nɤ$^{33}$ | ny$^{21}$ | nɤ$^{21}$ | ne$^{33}$ | ȵɛ$^{33}$ | ni$^{33}$ |
| 164. | paste | 16.59.78. | na$^{2}$ | na̱$^{2}$ | na̱$^{2}$ | ?na̱$^{2}$ | na̱$^{2}$ | nɚ$^{55}$ | ? | ȵo̱$^{55}$ |
| 29. | ask | 16.58.75. | na$^{33}$ | no$^{33}$ | no$^{55}$ | ?na$^{55}$ | no$^{33}$ | no$^{33}$ | no$^{13}$ | ṇa$^{33}$ |
| 54. | stop | 16.58.75. | na$^{55}$ | no$^{55}$ | no$^{33}$ | ?na$^{21}$ | no$^{55}$ | no$^{33}$ | nu$^{33}$ | nɯ$^{33}$ |
| 89. | brain | 16.68.78. | no$^{2}$ | nu̱$^{2}$ | nu̱$^{2}$ | ?no̱$^{2}$ | nu̱$^{2}$ | no̱$^{55}$ | ne$^{13}$ | no̱$^{55}$ |
| 174. | heart | 16.64.75. | ni$^{44}$ | ȵi̱$^{33}$ | ȵi̱$^{33}$ | ? | ȵi̱$^{33}$ | ni̱$^{2}$ | nɪ$^{33}$ | ? |
| 192. | press... | 16.65.75. | ȵɣ$^{44}$ | no̱$^{33}$ | nɣ$^{33}$ | ? | nɣ$^{33}$ | ȵi$^{2}$ | ȵi$^{33}$ | ȵi̱$^{4}$ |
| 138. | frost | 16.61.75. | ŋɪ$^{33}$ | ni$^{33}$ | ne$^{55}$ | ? | ne$^{33}$ | ȵe$^{33}$ | ȵi$^{13}$ | ? |

Allophones may be inserted into the outcome columns of the table separated by an equals sign, as shown in (155) below. Thus the cells in the table of correspondences for those languages in which this alternation is observed contain both segments as possible reflexes of *L n:

*(155)*

| *#* | *T* | *PL* | *PNL* | *sani* | *axi* | *nesu* | *lalo* | *li* | *nasu* | *neisu* | *nosu* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 44 | I | *n | *n | ȵ=n | ȵ=n | ȵ=n | n,ʔn | ȵ,ɲ | ȵ=n | ȵ=n | ȵ,ɲ,ɲ̣ |

This certainly gives the appearance of unfinished business in the diachronic analysis. The coding conventions distinguishes the use of the equals sign (=) to distinguish synchronic variants present in the data (and so not handled by separate correspondence lines) and the comma (,) used to designate as-yet unexplained variance. The program treats these two delimiters identically. Using the Reconstruction Engine (§7) in the *downstream* direction, such compound entries produce reflexes containing all possible outcomes; in the *upstream* direction, a modern form matches any one of the variants.

The results of applying these two criteria for merging (and assigning a reconstruction, discussed below) are illustrated in correspondence rows 2 and 5 in (1) below.

*(156) Correspondence sets after sorting and some data reduction*[111]

| Corr | Type | *L | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | I | s-n | n | n | n | n | n | n | n | n |
| 2. | I | s-my | n | n | n | ʔm.m | m | n | n | ȵ |
| 3. | I | w | v | v | v | v | v | v | v | v |
| 4. | I | k-r | ʑ | ʑ | ʑ | ʑ | ʑ | ɣ | ɣ | v |
| 5. | R | ak | e | e̠ | e̠ | i̠ | e̠ | a̠ | a | o̠.a̠ |
| 6. | T | H | 44 | 33 | 33 | 33 | 33 | 2 | 33 | 4 |
| 7. | T | L | 2 | 2 | 2 | 2 | 2 | 55 | 13 | 55 |

As the figure shows, each row in the resulting consolidated table of correspondences has been assigned (by hand on the comparison with other Loloish reconstructions) a reconstruction (third column, *L). If there were no other evidence, the reconstruction would have to be based on the evidence of the rows themselves, either on the basis of 'majority rules' or arguments based on typological, phonological, or phonetic grounds. However, in the present case, we can 'peek' at other reconstructions for the broader Loloish group, and also at Written Burmese, which tends to conserve many archaic features. I have listed the reconstructions for the four words in question (taken from the literature) in (157). These I adduce as support for assigning the reconstructions in the correspondence rows in (156) above.

*(157) Reconstructions from the literature for the sets under study*

| Gloss | TSR | Bradley 1979 | STC (*TB) |
|---|---|---|---|
| *eye* | s-myak$^{H}$ | (C)-myak$^{H}$ | mik ~ myak |
| *black* | (s-)nak$^{H}$ | C-nak$^{H}$ | nak |
| *chicken* | k-rak$^{H}$ | k-rak$^{H}$ | rak ('fowl') |
| *pig* | wak$^{L}$ | wak$^{L}$ | p-wak |

## 5.4. GENERATING COGNATE SETS FROM THESE TABLES

The next step in the process is to verify that the correspondences and reconstructions proposed in the various sources are consistent among themselves and are supported by the data. The Reconstruction Engine is being used as a means to organize and test the correspondences already given by authors of the original sources and as a framework for adding extending the depth and breadth of the comparison of Loloish languages. The data from the ZMYYC are fairly new (or at least are not well-known in the West) and have not previously been compared with other reconstructions. I have compared the forms in these languages with those in the other languages for which correspondences have already been noted and am attempting to fit these new data into the existing schema.

There are a number of problems, however, with making the reconstruction of *NL jibe with other *L and *LB reconstructions. Consider the cognate sets supporting the *NL correspondence row 163, shown in (158) below. Clearly, these correspondence sets all have the same proto-rhyme at the level of Proto-Northern Loloish. (It may be argued that the Nosu reflexes should not be conflated, but this conflation has no effect on the point being made here).

(158) *A few of Chen Kang's cognate sets, with possible correspondence rows*

| Corr | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| **163** | **R** | a | a̠ | a̠ | a̠ | a̠ | ɚ̠ | ɪ | o̠.ɹ̠ |

| Set | Gloss | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 76. | shoot | ba$^{44}$ | ba̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | mbhɚ̠$^{2}$ | mbɪ$^{33}$ | bɹ̠$^{4}$ |
| 77. | light | dla$^{44}$ | ɖa̠$^{33}$ | da̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | ɖɚ̠$^{2}$ | ɖɪ$^{33}$ | dɹ̠$^{4}$ |
| 92. | rub | va$^{2}$ | va̠$^{2}$ | va̠$^{2}$ | va̠$^{2}$ | va̠$^{2}$ | vɚ̠$^{55}$ | vɪ$^{13}$ | vo̠$^{55}$ |
| 93. | lick | ɖa$^{2}$ | ɖa̠$^{2}$ | la̠$^{2}$ | la̠$^{2}$ | la̠$^{2}$ | lɚ̠$^{55}$ | lɪ$^{13}$ | zo̠$^{55}$ |
| 167. | lack | qha$^{2}$ | kha̠$^{2}$ | kha̠$^{2}$ | kha̠$^{2}$ | kha̠$^{2}$ | khɚ̠$^{55}$ | khɪ$^{13}$ | tɕho̠$^{55}$ |
| 168. | paste | na$^{2}$ | na̠$^{2}$ | na̠$^{2}$ | ʔna̠$^{2}$ | na̠$^{2}$ | nɚ̠$^{55}$ | ? | ȵo̠$^{55}$ |
| 169. | bandit | dza$^{44}$ | dza̠$^{33}$ | dza̠$^{33}$ | ? | dza̠$^{33}$ | dzɚ̠$^{2}$ | dzɪ$^{33}$ | dzo̠$^{4}$ |

| Corr | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 164 | R | a | a̠ | a̠ | a̠ | a̠ | ɚ̠ | a | a̠ |

| Set | Gloss | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 75. | hold in arms | va$^{44}$ | va̠$^{33}$ | va̠$^{33}$ | va̠$^{33}$ | va̠$^{33}$ | vɚ̠$^{2}$ | va$^{33}$ | va̠$^{4}$ |

| Corr | Type | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 165 | R | a | a̠ | a̠ | a̠ | a̠ | ɚ̠ | e | i̠ |

| Set | Gloss | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|
| 170. | know | sa$^{2}$ | sa̠$^{2}$ | sa̠$^{2}$ | sa̠$^{2}$ | sa̠$^{2}$ | sɚ̠$^{55}$ | se$^{13}$ | si̠$^{55}$ |

However, when we go up a level from Northern Loloish to Loloish and compare the *L protoforms as given in various sources (as shown in (159)), the words shows a different *rhyme. What, then, should the reconstruction of *NL be in this case?

*(159)*

| | | | |
|---|---|---|---|
| 76. | shoot | Npök$^{H}$/?pök$^{H}$ or ?bök$^{H}$ | TSR 108 |
| 77. | light | | ? |
| 92. | rub | | ? |
| 93. | lick | m-lyak ≍ ?-lyak | TSR 179a |
| 167. | lack | | ? |
| 168. | paste | nay$^{2}$ | GSTC 113 (='sticky, cohesive') |
| 169. | bandit | | ? |

We have several choices. Either the items in (158) represent distinct roots in *NL, unrelated to those reconstructed for the other branches, or we must conclude that the distinct protorhymes in (159) had merged by the time of *NL—there is no evidence in *NL to support distinguishing them. At least for sets 76 *shoot* and 93 *lick* , the *NL forms do appear to be cognate to the *L forms given in (159), so we are justified in concluding (perhaps once a few more sets are compared):

*(160)*

| | | | | |
|---|---|---|---|---|
| *L | -ök | > | *NL | -Vk or perhaps even -V̲ |
| | -yak | | | |

Being able to chronologically order some of the steps in the evolution of protorhymes may allow us to make more accurate judgments about other changes.

Because the synonym matrix contains a complete set of forms in nearly all cases it is possible to uniquely determine which correspondence set each synonym set is aligned with despite the massive merger of initials and rhymes. Thus, though the dialects are very closely related and differences between correspondence lines often rather small, there is usually only one possible path through the table of correspondences and therefore only one reconstruction generated. An example of the result of the automated verification by the Reconstruction Engine is shown in (161). In the next chapter I will show how the Reconstruction Engine builds these sets from the correspondence rows and synonym sets.

*(161)*

52. myak$^{H}$ [45.87.1] / *EYE*

| | | |
|---|---|---|
| [1] sani | ne$^{44}$ | *eye* |
| [2] axi | ne̜$^{33}$ | *eye* |
| [3] nesu | ne̜$^{33}$ | *eye* |
| [4] lalo | ʔmʅ$^{33}$ | *eye* |
| [5] li | me̜$^{33}$ | *eye* |
| [6] nasu | na̠$^{2}$ | *eye* |
| [7] neisu | na$^{33}$ | *eye* |
| [8] nosu | ȵo̜$^{4}$ | *eye.* |

# 6. HYPOTHESIS TESTING

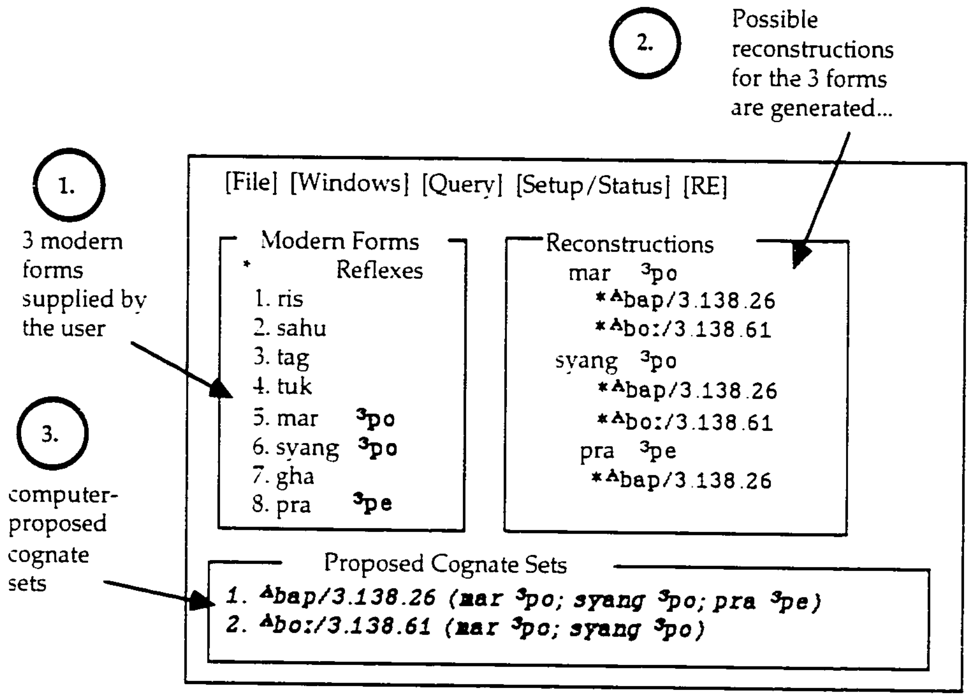
This chapter is devoted almost exclusively to the Reconstruction Engine. It is a portion of a reworked version of the paper I published with Martine Mazaudon in Computational Linguistics 20:30 (1994). I received permission to include it in this dissertation as it is a logical next step in the development of cross-linguistic lexicographic databases as I have presented it so far.

## 6.1. Tools for phonological reconstruction : The reconstruction engine

The Reconstruction Engine is a prototype computational tool which automates a crucial portion of the comparative method: the process of creating cognate sets and proposing reconstructions on the basis of observed correspondences between modern languages.[112] The Reconstruction Engine functions as a 'checker' of hypotheses proposed by the linguist. It uses a body of rules, but has no inferential component, at least not in the sense usual used with regard to expert systems. (Charniak and McDermott 1985) Its role is to verify the internal consistency of a set of phonological correspondences, created beforehand by the linguist, against the lexicons of an ensemble of putatively related languages, and to gauge the extent to which those data are consistent with the given phonological and phonotactic descriptions (i.e. correspondences and syllable canon).

The essence of the Reconstruction Engine's processing is illustrated in (162) and (163) below, which shows how the Reconstruction Engine in *interactive mode* can 1) propose cognate sets on the basis of modern forms and 2) generate the expected regular reflexes for a given reconstruction. The illustrations are meant to lay out the essence of the program's operation; the many details are discussed in later sections.

*(162) Interactive generation of reconstructions (an 'upstream' computation; transcription and languages exemplified are described in Appendix 4.1)*



2. Possible reconstructions for the 3 forms are generated...

1. 3 modern forms supplied by the user

3. computer-proposed cognate sets

[File] [Windows] [Query] [Setup/Status] [RE]

Modern Forms
* Reflexes
1. ris
2. sahu
3. tag
4. tuk
5. mar $^{3}$po
6. syang $^{3}$po
7. gha
8. pra $^{3}$pe

Reconstructions
mar $^{3}$po
*$^{A}$bap/3.138.26
*$^{A}$bo:/3.138.61
syang $^{3}$po
*$^{A}$bap/3.138.26
*$^{A}$bo:/3.138.61
pra $^{3}$pe
*$^{A}$bap/3.138.26

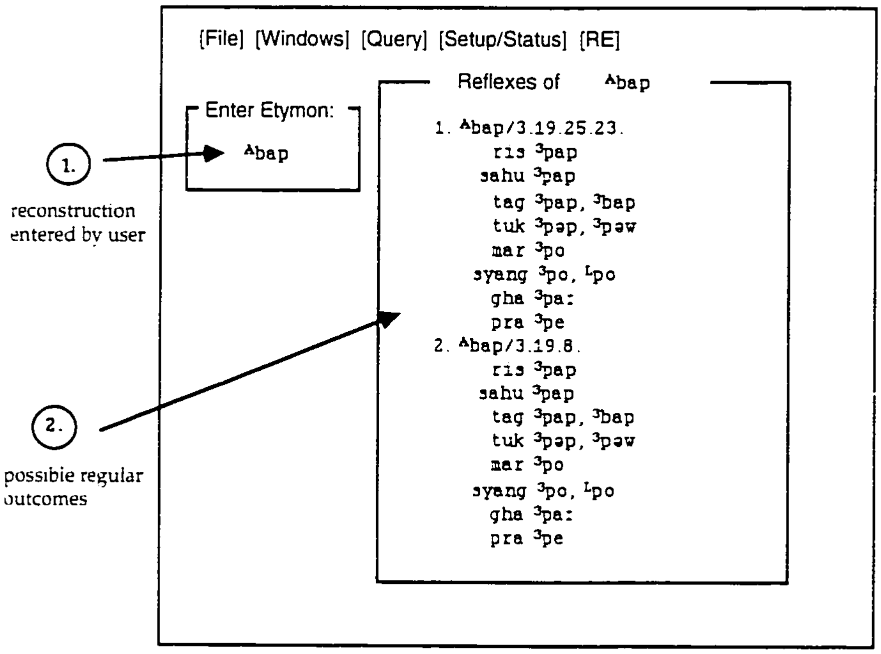
Proposed Cognate Sets
*1. $^{A}$bap/3.138.26 (mar $^{3}$po; syang $^{3}$po; pra $^{3}$pe)*
*2. $^{A}$bo:/3.138.61 (mar $^{3}$po; syang $^{3}$po)*

*(163)* is a representation of the contents of the computer screen after the user has entered three modern words (1). The program has generated the set of possible reconstructions from which these forms might be derived

(2). The list of numbers (called the *analysis* and discussed below §6.4.1) following each reconstruction refers to the row numbers in the table of correspondences which were used by the program in generating the reconstructions. In two cases (Marpha (mar) and Syang (syang)) reflexes have more than one possible ancestor; here the program has proposed the two possible cognate sets which result from computing the set intersection of the possible ancestors (3). The proposed sets are listed in descending order by population of supporting forms.[113]

Conversely, given a protoform, the Reconstruction Engine will predict (actually 'postdict') the regular reflexes in each of the daughter languages. (163) reproduces the results on the computer screen of performing such a 'downstream' calculation.

*(163) The expected outcomes of* *$^{A}$bap *(a 'downstream' computation)*


[File] [Windows] [Query] [Setup/Status] [RE]
Enter Etymon:
$^{A}$bap
Reflexes of $^{A}$bap
1.
reconstruction entered by user
2.
possible regular outcomes

```
1. ^Abap/3.19.25.23.
     ris $^{3}$pap
    sahu $^{3}$pap
     tag $^{3}$pap, $^{3}$bap
     tuk $^{3}$pəp, $^{3}$pəw
     mar $^{3}$po
   syang $^{3}$po, $^{L}$po
     gha $^{3}$pa:
     pra $^{3}$pe
2. ^Abap/3.19.8.
     ris $^{3}$pap
    sahu $^{3}$pap
     tag $^{3}$pap, $^{3}$bap
     tuk $^{3}$pəp, $^{3}$pəw
     mar $^{3}$po
   syang $^{3}$po, $^{L}$po
     gha $^{3}$pa:
     pra $^{3}$pe
```

Here the etymon entered by the user (1) produced reflexes (2) through two different syllabic analyses (shown by the chain of numbers follows each reconstruction): $^{A}$bap as initial /b-/ plus vowel /-a-/ plus final /-p/, and as initial /b-/ followed by rhyme /-ap/. In this case, the results are the same, but this is not always so. The algorithms used in this process are described in §6.4.

Using the Reconstruction Engine interactively as shown above is a great aid in testing the details of a hypothesized set of sound changes. The primary use of the Reconstruction Engine, however, is to check whole lexicons containing thousands of words; this *batch process* is illustrated below in §6.4.4.

The Reconstruction Engine has several features which represent a significant advance in the automated handling of diachronic data. It provides *part* of the 'closed catalog' (discussed in §1.7.5) required (it has been claimed) for a complete and verifiable hypothesis of diachronic development. In particular, it provides a closed set of rules describing a closed multilingual data set.

The first and most important feature of the Reconstruction Engine is that it provides an exhaustive treatment of the data in several dimensions:

- It processes complete lexicons of modern languages. Every modern form is evaluated by the program in a consistent and complete way.

- Each form is completely analyzed. Modern forms which are not *completely regular* (according to the rule set and phonotactic structure, described in §6.3 below) are not included in cognate sets.

- The two major data structures of the program (*correspondences* and *syllable canon* constitute a complete and unified statement of the diachronic phonology of the languages treated.

Second, the Reconstruction Engine contains a number of features which make it flexible in handling the kinds of data realistically encountered in historical research.

- Provisions exist for allowing several different transcriptions to be used in representing the data.

- There are no requirements that the data be organized beforehand by gloss, semantic field, phonological shape, or any other criterion.

- The size and type of constituents used in the analysis are not limited by the program. There is no requirement, for example, that a segmental analysis be used (as opposed to, for example, the initial-plus-rhyme-plus-tone analysis commonly used for many Asian languages). However, the program does not provide for non-linear representations or discontinuous constituents: the 'absolute slicing hypothesis' is assumed. Also, the linearization of constituents must be the same for all the language data used by the program. For example, the tone numbers used in the languages cited in this dissertation, which might equally well be ordered before as after the segmental strings to which they apply, are uniformly written at the beginning. Problems associated with non-linear constituents were discussed in §3.2.1.4.

- Several competing analyses of the same data can be managed and compared simultaneously.

## 6.2. INTERNALS OF THE RECONSTRUCTION ENGINE

The Reconstruction Engine implements 1) a set of algorithms which generate possible reconstructions on the basis of word forms in modern languages (and vice-versa as well) and 2) a set of algorithms which arrange the input modern forms into possible cognate sets based on those reconstructions. The first set uses a simple bottom-up parser; the second automates database management chores, such as reading multiple input files and sorting, merging, and indexing the parser's output.

The core functions of The Reconstruction Engine compute all possible ancestors for a given set of putative cognate forms (using a TABLE OF CORRESPONDENCES and a phonotactic description, a SYLLABLE CANON, both described in §6.3) and makes sets of those modern forms which share the same reconstructions. The initial creation of possible cognate sets is done on a purely phonological basis, without reference to semantics. Other software tools can then be used to further divide the computer-proposed cognate sets on the basis on semantic distinctions. The linguist (that is, the user) collects and inputs the source data, prepares the table of correspondences and phonotactic description (syllable canon), and verifies the semantics of the output of the phonologically based reconstruction process. The Reconstruction Engine, *qua* 'linguistic bookkeeper,' makes the projections and keeps track of several competing hypotheses as the research progresses. Specifically, the linguist provides as input to the program:

(a) Word forms from several modern languages, with glosses.

(b) Parameters which control the operation of the program and the interpretation of the input data (mostly not described here).

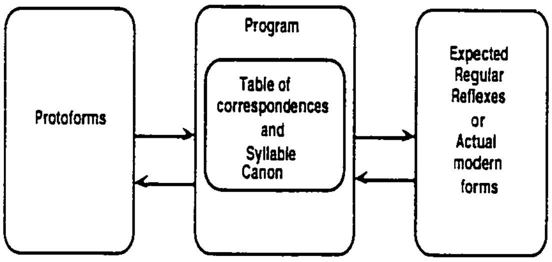
(c) A file containing the table of correspondences, detailed below.

(d) The syllable canon, described below.

(e) Semantic information for disambiguating modern and reconstructed homophones, described below.

Though I noted this point above, I wish to reiterate that the parsing algorithm implemented in the Reconstruction Engine is bidirectional: the 'upstream'[114] process involves projecting each modern form *backward* (cf. Hock 1986) in time and merging the sets of possible ancestors generated thereby to see which, if any, are identical. Conversely, in the 'downstream direction', the program computes the expected regular reflexes in the daughter languages on the basis of given protoforms. In this way the Reconstruction Engine follows a long research tradition in natural language processing concerned with reversible grammars. Reversing grammatical rules has a number of profound computational and linguistic implications which I only mention here. Some issue which arise are ambiguity in the output of reversed grammars, algorithmic obstacles to reversing any given grammar, and so on. Many of these problems are relatively well understood and a substantial literature has grown up around these issues. (cf. for example Kaplan September 1994 for a summary).

(164) *Input-Output diagram of the Reconstruction Engine's basic projection functions*


Protoforms
Program
Table of correspondences and Syllable Canon
Expected Regular Reflexes or Actual modern forms

## 6.3. PRINCIPAL DATA STRUCTURES: THE CORRESPONDENCE TABLE AND THE SYLLABLE CANON

Two data structures (internal to the program) are relevant to the phonological reconstruction, and these are passed as arguments to the Reconstruction Engine. The table of correspondences represents the linguist's hypothesis about the development of the languages being treated. The columns of the table (already identified in §1.2, but repeated briefly here for convenience) are (1) a correspondence set number, uniquely identifying the correspondence; (2) the distribution of the correspondence within the syllable structure (i.e. the type of syllable constituent: in this case, Tone, Initial, Liquid, Glide, Onset, Rhyme, Vowel, or Final); (3) the PROTOCONSTITUENT itself; (4) the phonological context (if any) in which the correspondence is found; (5-12) the OUTCOME or reflex of the protoconstituent in the daughter languages.[115] The CONSTITUENT TYPES (T, I, F, L, etc. in column 2) are specifiable by the user. So, for example, C and V could be chosen if no other types of constituents need to be recognized for the research. Note that the table allows for several different outcomes depending on context; the absence of context indicates either an unconditioned sound change or the Elsewhere case of a set of related rules (as discussed below).

*(165) Excerpt from the Table of Correspondences*

| (1)<br>N | (2)<br>ConT | (3)<br>* | (4)<br>context | (5)<br>ris | (6)<br>sahu | (7)<br>tag | (8)<br>tuk | (9)<br>mar | (10)<br>syang | (11)<br>gha | (12)<br>pra |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | T | $^{A}$ | $\_C_{VL}$ | $^{1}$X.$^{H}$ | : | :$^{H}$ | :$^{H}$ | : | :$^{H}$ | : | : |
| 3 | T | $^{A}$ | $\_C_{VD}$ | $^{3}$X.$^{L}$ | : | : | : | : | :$^{L}$ | : | : |
| 93 | I | k | k | k | k | k | k | k | k | k | |
| 94 | I | k | _w | k | Ø | h | k | k | k | k | k |
| 181 | F | k | | k | : | Ø.k | Ø.k | Ø | Ø | Ø | Ø |
| 142 | L | r | p.p$^{h}$.b_ | r | r | r | r | r | r | r | r=ŕ |
| 169 | O | r | | r | r | r | r | r | r | r | r |
| 102 | O | kr | _e: | k | k | k | ʈ | k | k | kr | kr |
| 103 | O | kr | _u | kr | kr | ɦ | ʈ | k | k | kr | kr |
| 104 | O | kr | _a | kr | kr | hw | ʈ | kj | kj | kr | kr |
| 105 | O | kr | _at | kr | kr | h | ʈ | kj | kj | kr | kr |
| 106 | O | kr | | kr | kr | h | ʈ | k | k | kr | kr |
| 31 | R.V | a | | a | a | a | ə | ə | ə | a | ɤ |
| 186 | V | a | _p | a | a | a | ə | o | o | a | ɐ |

*Column headings (discussed in the text below):*

(1) Row number uniquely identifying this correspondence set

(2) Slot type; shows phonotactic position of this correspondence (according to syllable canon)

(3) Reconstruction

(4) Phonological context (in the protolanguage)

(5-12) Outcomes in daughter languages; actual observed correspondences[116]

The SYLLABLE CANON provides a template for building monosyllables. It specifies how the constituents of the table of correspondences may be combined based on the (syllable) constituent types (column (2) of the table). Thus, the outcomes for a final /k/ (correspondence 181) and an initial /k/ (correspondence 93) are never confused by the program. The program takes the syllable canon as an argument expressing the adjacency constraints on the constituents found in the Table. For example, the canon for *TGTM, illustrated in (166) below, has three SLOTS, each having its own substructure: first, a tone (optional, as indicated by the possibility of a zero element); followed by an (also optional) initial element consisting of various combinations of Onset, Glide, Initial (consonant), and Liquid; and terminating with either a Rhyme element or a Vowel plus Final consonant.[117] A syllable is composed of zero or more elements from each of these SLOTS. Picking the longest possible combination from each slot produces the maximal syllable permitted by the canon — six constituents (TILGVF), one from each constituent type except O and R. Similarly, the minimal syllable has only one constituent (R). Parentheses indicate optional elements within a slot, brackets separate the sequential slots in the syllable structure.

*(166) Syllable Canon in Proto-Tamang*

[T,Ø]  [O(G),I(L)(G),Ø]  [R,VF]

| | | | |
|---|---|---|---|
| T | =Tone | L | = Liquid |
| O | = Onset (complex initial) | R | = Rhyme |
| G | = Glide | V | = Vowel |
| I | = Initial Consonant | F | = Final consonant |
| Ø | = Zero | | |
| () | encloses optional element | | |
| [] | encloses alternative possible slot fillers | | |

*(167) Brace notation for Proto-Tamang Syllable structure*

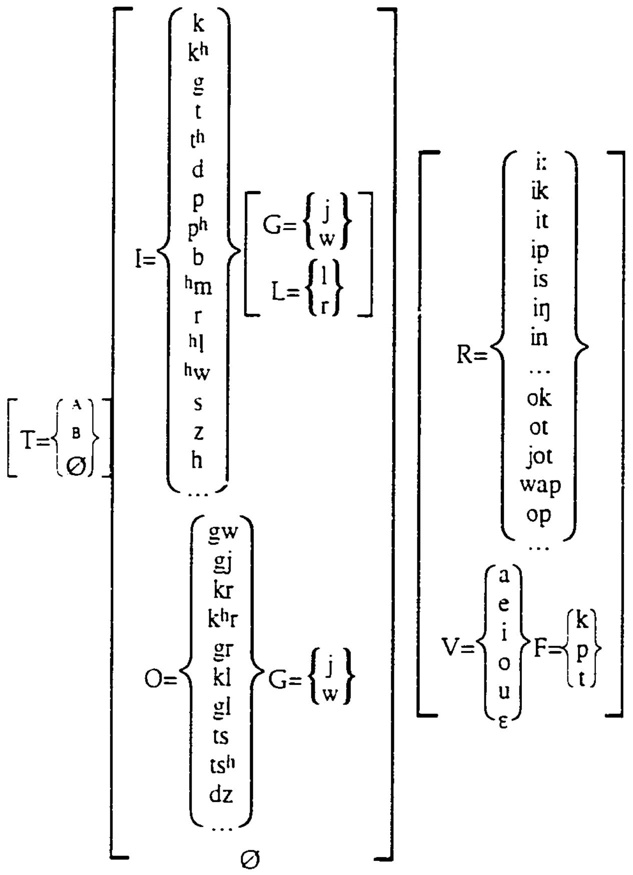
$$\begin{Bmatrix}\mathrm{T}\\ \varnothing\end{Bmatrix}\begin{Bmatrix}\mathrm{O(G)}\\ \mathrm{I(L)(G)}\\ \varnothing\end{Bmatrix}\begin{Bmatrix}\mathrm{R}\\ \mathrm{VF}\end{Bmatrix}$$

Another way to conceive of the relationship between the table and the canon is to imagine each slot of the canon (optionally) filled with the elements from the table which are licensed for that slot. (168) below gives for each slot the list of protoconstituents which can fill that slot.

*(168) Possible (proto-)fillers for syllable canon slots*

| Count | Slot | | Set of elements which can fill specified slot |
|---|---|---|---|
| 17 | I | = | k. k$^h$. g. t. t$^h$. d. p. p$^h$. b. $^h$m. r. $^h$l. $^h$w. s. z. h.... |
| 27 | O | = | gw. gj. kr. k$^h$r. gr. kl. gl. ts. ts$^h$. dz.... |
| 2 | G | = | j. w |
| 2 | L | = | l. r |
| 6 | V | = | a. e. i. o. u. ε |
| 3 | F | = | k. p. t |
| 74 | R | = | iː. ik. it. ip. is. iŋ. in. ... ok. ot. jot. wap. op.... |
| 2 | T | = | $^{A}$. $^{B}$ |

*(169) Possible fillers for Proto-Tamang Syllable structure given in (168) above*



The syllable canon is thus a type of regular expression, providing a shorthand device for expressing 28 possible syllable structures as illustrated below in (170).

*(170) The 28 possible *TGTM syllable structures, with two illustrations*

| | | |
|---|---|---|
| IGR | TIGR | |
| IGVF | TIGVF | |
| ILGR | TILGR | |
| ILGVF | TILGVF | $^{B}$grwat 'hawk' |
| ILR | TILR | |
| ILVF | TILVF | |
| IR | TIR | |

| | | |
|---|---|---|
| IVF | | TIVF |
| OGR | | TOGR |
| OGVF | | TOGVF |
| OR | $^{A}$kra 'hair' | TOR |
| OVF | | TOVF |
| R | | TR |
| VF | | TVF |

Note that in the illustrations supplied in (170) more than one analysis according to the canon may be possible. These two protoforms (among others) are discussed below. The program computes *all possible segmentations* of a given protoform. The possible segmentations of the *TGTM etymon *$^{B}$grwat. for example, are given below in (179).

The Proto-Tamang syllable canon is quite complex because several hypotheses about syllable structure are encoded in it. For other languages, in which only consonant and vowel need be distinguished in describing syllable structure, a simpler canon (e.g. CV(C)) might suffice. Polysyllabic syllable canons can be expressed and used by the canon in two ways:

- Explicitly; for example,

  [CV(C)]CV,Ø]

  specifies a bisyllabic canon in which the minimal form is CV and the maximal is CVCCV.

- As a recursive application of a single syllable. This is done via a software toggle which allows the canon structure to be repeatedly mapped over an input form. For example, if the polysyllable toggle is turned on, the canon

[(C)V(C)]

would match structures of the form V, CVC, CVCCV,CVCVCVV, etc.

## 6.4. ALGORITHMS

### 6.4.1. Generating proto-projections

Three steps are required to project or transform a given modern form into a set of possible reconstructions or vice versa:

- tokenizing the given form into a list of row numbers in the table of correspondences (i.e. column 1 of (165)); this is a recursive process;
- filtering the tokenized forms according to syllabic and phonological constraints; and
- substituting the actual outcomes in the table of correspondences for the tokens, also a recursive process.

#### 6.4.1.1. Tokenization

On its first recursive pass the Reconstruction Engine generates (recursively from the left of the input form) all possible segmentations of the form. That is, starting from the left, the program divides the form in two, and then repeats the process on the right hand part until the end of the form is reached. Essentially, this algorithm implements a standard solution to a standard problem, that of finding all parses of an input form with respect to the given regular expression (encoded in the syllable canon and table). As the segmentation tree is created, the program checks to see that the node being built is actually specified as an element of the table of correspondences; it thereby avoids having to build branches of the tree which cannot produce outcomes (according to the table of correspondences). The pseudocode below (171) outlines the algorithm.

*(171) Pseudocode for tokenizing forms into table constituents*

```
/* GENERATE:   STEP ONE: Tokenize input form   */
Tokenize(InputString,TokenList)
          /* base case */
    if InputString is null then return(TokenList)
          /* recursive step */
    for i = 1 to the length of InputString
          leftside = leftmost i characters of InputString
          rest     = the rest of InputString
          lookup leftside in list of constituents for this table column
          if found then
               add tokens (i.e. TofC row numbers)  for this constituent to
                                                              TokenList
          otherwise
               /* abandon this parse */
          end if
    Tokenize(rest,TokenList)
    end for
end Tokenize
```

Consider for example the segmentations of:

(172) *$^{A}$kra *head hair*

There are eight ways to segment this protoform:

| | | | | |
|---|---|---|---|---|
| (173) | $^{A}$kra | $^{A}$-kra | $^{A}$k-ra | $^{A}$kr-a |
| | | $^{A}$-k-ra | $^{A}$k-r-a | |
| | | $^{A}$-k-r-a | | |
| | | $^{A}$-kr-a | | |

Of these eight segmentations, only two ($^{A}$-k-r-a and $^{A}$-kr-a)are composed entirely of elements which occur in the protoconstituent column (3) of the table given in (165) above. For each of these valid segmentations, the Reconstruction Engine constructs a tokenized version of the form, wherein each element of the segmented form is replaced with the correspondence or list of correspondences for that constituent in the table. *k, for example, has three possible outcomes (given by rows 93, 94, and 181 of the table of correspondences) depending on its syllabic position and environment. These segmented and tokenized versions are listed in the following two tables.

(174) *Segmentations whose elements are ALL constituents of the table:*

| | | | | | |
|---|---|---|---|---|---|
| (a) | Segmentation: | ^ | k | r | a |
| | Tokenized form: | (1,3) | (93,94,181) | (142,169) | (31,186) |

| | | | | |
|---|---|---|---|---|
| (b) | Segmentation: | ^ | kr | a |
| | Tokenized form: | (1,3)(102,103,104,105,106) | | (31,186) |

(175) *Segmentations which contain one or more elements NOT found in the table (i.e. which are only partially parsable; these are only listed here, the parser stops building when an unknown element is seen):*

| | Segmentation | Tokenized form |
|---|---|---|
| (c) | ^kr-a | (?)(31,186) |
| (d) | ^kra | (?) |
| (e) | ^-k-ra | (1,3)(93,94,181)(?) |
| (f) | ^k-ra | (?)(?) |
| (g) | ^k-r-a | (?)(142,169)(31,186) |
| (h) | ^-kra | (1,3)(?) |

### 6.4.1.2. Filtering

Having created and tokenized a list of all valid segmentations, the algorithm traverses each tokenized form, looking up (in the table) each correspondence row number of each segment and substituting the outcome of that row from the appropriate column of the table. As the output form is being built, the phonological and phonotactic contexts are checked to eliminate disallowed structures, as illustrated in the pseudocode given in (176).

(176) *Pseudocode for filtering possible projections and substituting regular outcomes*

```
/* GENERATE: STEP TWO: Convert tokenized form into possible
outcomes*/
FilterAndSubstitute(TokenList,ListofPossibleForms)
    /*  base case */
    if TokenList is NULL then return(ListofPossibleForms)
    /* recursive step */
    for each RowNumber of first segmented element in TokenList
        if phonological and syllabic context constraints are met then
            for each language in the table
                add outcomes for this RowNumber to each output form in
                        ListofPossibleForms for this language
            end for
        otherwise
            * do not use this token in building output forms *
        end if
    end for
    remove first segmented element from TokenList
    FilterAndSubstitute(TokenList.ListofPossibleForms)
end FilterAndSubstitute
```

If we turn again to (174)(b) above, then were it not for syllabic structure constraints and phonological context constraints expressed in the table, the segmentation (restated below)

(177) ʌ-kr-a (1,3)(102,103,104,105,106)(31,186)

would produce 20 (= 2 x 5 x 2) different outcomes based on the different ordered combinations of its tokens :

(178) 1.102.31
1.103.31
1.104.31
1.105.31
1.106.31
1.102.186
1.103.186
1.104.186
1.105.186
1.106.186
3.102.31
3.103.31
3.104.31
3.105.31
3.106.31
3.102.186
3.103.186
3.104.186
3.105.186
3.106.186

*With* these constraints, however, only one combination is licensed: 1.104.31, because:

- only the tone correspondence for row 1 applies since it specifies the outcome of prototone A for voiceless initials;

- only outcomes of row 104 for *kr- are generated since this is the most specific rule that applies (see discussion below); and because

- row 186 is eliminated as a possibility for *-a in this case since these outcomes only occur when *-a is followed by *-p.

Some conditions on the application of the rules should be noted here. The program *does* apply Pāṇini's principle, also known as the Elsewhere Condition ((Kiparsky 1973), Kiparsky 1982). Thus, of all possible *kr- correspondences, only the most specific is selected. For example, though the context in line 104 *-a is a substring (or subcontext) of line 105 *-at, only one or the other is selected for any particular segmentation of a protoform ending in *-at (i.e. 104 for *-a- + *-t versus 105 for *-at).[118] If the 'specificity' of several applicable contexts is the same, all are used by the program in generating the forms.[119] Also, note that since the context is stated in terms of proto-elements, when computing backwards (*upstream*) the program must tokenize *and* substitute ahead (i.e. look ahead) to determine if the context of a correspondence applies. The other possible segmentation, $^{\text{A}}$-k-r-a, though a valid segmentation of the input form into table constituents, would fail to produce any reflexes because the phonological context criterion is not met.

### 6.4.1.3. Substitution

In the final step, the program substitutes the outcomes for each correspondence row in each of the language columns of the table and outputs the expected reflexes. The expected outcome of *just the segmentation* $^{\text{A}}$-kr-a in Tukche, for example, ((165) above column 8 tuk) is either $^{\text{l}}$ʈə or $^{\text{H}}$ʈə.[120]

This process is performed for each language column in the table, resulting in a list of all the modern reflexes of the input protoform. This assumes, of course, that the reconstructed forms are correct, the rules are correct, and no external influences have come into play. By comparing these computer-generated modern forms with the forms actually attested in the living languages we can check the adequacy of the proposed analysis, and make improvements and extensions as required.

### 6.4.2. Combinatorial explosion and syllable structure

The example given above has only a few possible segmentations. Consider, however, the set of all eight possible *valid* segmentations of the *TGTM form *$^{B}$grwat *hawk, eagle,* schematized in (179). There is, of course, a substantially larger number of *invalid* segmentations. Each token of a segmentation may have a sizable list of possible outcomes. One can see that even relatively uncomplicated monosyllables may generate massive ambiguity in structural interpretation. Indeed, some of the monosyllabic forms in the Tamang database generate nearly a hundred reconstructions, even given the limitations of syllabic and phonological context. Uncontextualized correspondence rows describing languages in which many mergers have occurred may contain syllables which produce thousands of reconstructions.

(179) *Segmentation of* *$^{B}$grwat *'hawk, eagle'*

| | Type of syllable constituent | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | T | I | L | O | G | R | V | F | Structure |
| (1) | $^{B}$ | | | gr | w | | a | t | T O G V F |
| (2) | $^{B}$ | | | gr | w | at | | | T O G R |
| (3) | $^{B}$ | | | gr | | | wa | t | T O V F |
| (4) | $^{B}$ | | | gr | | wat | | | T O R |
| (5) | $^{B}$ | g | r | | w | | a | t | T I L G V F |
| (6) | $^{B}$ | g | r | | w | at | | | T I L G R |
| (7) | $^{B}$ | g | r | | | | wa | t | T I L V F |
| (8) | $^{B}$ | g | r | | | wat | | | T I L R |

### 6.4.3. Computing upstream and creating a set of cognates

The preceding discussion shows how the table of correspondences can be read from ancestor to daughter (left to right), *downstream* in the sense of history. It can also be read from daughter to ancestor (right to left), *upstream*, revealing all the possible ancestors of a given modern segment of a particular language. For example, we can see from the excerpt in (165) that Syang k (in column 10 of the table) could be derived from either *k- or *kr- (to be read from the column of protoconstituents (column 3) of the table).

By combining, according to the syllable canon, all the possible permutations of *initial, *tone and *rhyme for the initial, tone and rhyme of a given modern word, the computer can, using exactly the same procedures as described in §6.4.1, create a list of its possible reconstructions. If the possible reconstructions of a set of *cognate* words are compared, it must be true that one or more of the reconstructions is the same for all words in the set (assuming, of course, that the words are related via regular sound changes). This process is illustrated below using data from Northern Yi languages of China. (See Appendix 3.2 for details of the data set.)

The table of correspondences (excerpted in (187) below) has only a few of the most significant facts about phonological context specified (as yet), and only a few of the correspondences are exemplified. Note that the table expresses several well-known facts about the reconstruction of Loloish:

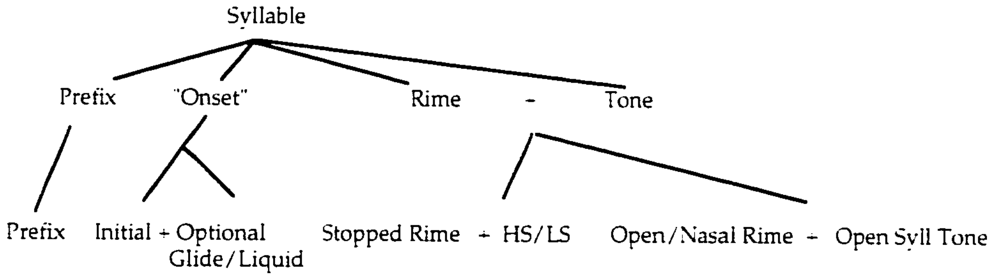
- The tonal split in stopped syllables (Matisoff 1972, Bradley 1979) is encoded in the tone correspondences of the table (187) and in the syllable canon. Since the rhyme types are indicated in the Syllable Constituent column (column 2 of (187)), the Reconstruction Engine, using the syllable canon in (185) below, will produce only those reconstructions which have the 'correct' relationship between rhyme and tone: stopped tones with stopped rhymes, open tones with open rhymes.

> In most Loloish languages, there is a basic two-term syllable-type system. One syllable-type, the vowel/nasal syllable-type, is characterized by final nasals or fully voiced vowels. The other syllable-type, the stop syllable type, is characterized by final stops, by constriction or laryngealization of the vowel, several or all of the above (Bradley 1979:206).

Therefore, and in accord with every other reconstruction of *LB and *L based on data with tonal recordings, the reconstruction of two separate *tonal systems for Proto-Loloish is encoded in the syllable canon shown below in (180, 185).

The alternative would have been to state all rhyme+tone combinations in the table as unitary constituents. This would certainly work, but it would be cumbersome and would fail to capture succinctly the important generalization that the tonal outcomes of prototones *$^{L}$ and *$^{H}$ are conditioned by the coda of the rhyme. Furthermore, the two values of the *tone category itself are conditioned by the voicing features of the prefix and initials, as is discussed below.

*(180) ***Loloish Syllable Structure***


Syllable
Prefix
"Onset"
Rime
-
Tone
Prefix
Initial + Optional Glide/Liquid
Stopped Rime + HS/LS
Open/Nasal Rime + Open Syll Tone

The disjunction of the two syllable types may be express formulaically as:

*(181)*

$$(P)(I(L,G))R_S T_S \quad \wedge \quad (P)(I(L,G))R_O T_O$$

where:

P = Prefix  
O = Onset, either a simple segment or an initial plus glide/liquid cluster, i.e. O = I or I(G) or I(L)  
$R_S$ = Stopped rhyme  
$R_O$ = Open/nasal rhyme  
$T_S$ = Tone for stopped syllable (= $^{\text{HS,LS}}$)  
$T_O$ = Tone for open/nasal syllable (= $^{1,2,3}$)

() indicate optionality

The disjunction of syllable structures is much clearer when the syllable canon is viewed in bracket notation (182) below:

*(182)*

$$\left[\begin{array}{l} \begin{Bmatrix}1\\2\\3\end{Bmatrix}\begin{Bmatrix}P\\ \emptyset\end{Bmatrix}\begin{Bmatrix}I\\ \emptyset\end{Bmatrix}\begin{Bmatrix}G\\ L\end{Bmatrix}\begin{Bmatrix}VN\\ V\end{Bmatrix} \\ \\ \begin{Bmatrix}HS\\ LS\end{Bmatrix}\begin{Bmatrix}P\\ \emptyset\end{Bmatrix}\begin{Bmatrix}I\\ \emptyset\end{Bmatrix}\begin{Bmatrix}G\\ L\end{Bmatrix}(VS) \end{array}\right]$$

This statement of the syllable canon fails to capture a few features of Loloish syllable structure. For example, there are in fact only fourteen syllable types in *L:

*(183) The 14 possible *Loloish syllable structures*

$T_O IGR_O$ $\qquad$ $T_S IGR_S$

| | |
|---|---|
| $T_OILR_O$ | $T_SILR_S$ |
| $T_OIR_O$ | $T_SIR_S$ |
| $T_OPIGR_O$ | $T_SPIGR_S$ |
| $T_OPILR_O$ | $T_SPILR_S$ |
| $T_OPIR_O$ | $T_SPIR_S$ |
| $T_OR_O$ | $T_SR_S$ |

Some possible of the possible structures licensed by the above description are not actually included in this list because other conditions, which would complexify the expression of the canon, have not yet been accounted for. These are:

- prefixes are only possible before initials; that is, Ø-initial reconstructions do not have prefixes. Thus there is no structure $T_SPR_S$.

- The Glide and Liquid slots are also limited in their distribution; a glide or a liquid can only occur between a (non-zero) initial and a rhyme. Thus there is no structure $T_SGR_S$ or $T_SLR_S$; perhaps it would be better to say that Glides and Liquids (as segments) can occupy the Initial slot when there is no initial. A few such cognate sets and reconstructions from previous research make this clear:

*(184)*

| | | |
|---|---|---|
| 57 | pluck[2] | *ʔcwat$^{H}$. Ak ci$^{HS}$. Ak ci$^{HS}$. Lh cɨʔ. Sa tši$^{44}$ |
| 132 | hungry | *mwat ~ ŋwat. -Ha [K] mie$^{33}$. Ahi ni$^{44S}$. Ak meh$^{LS}$. Bi bɛ̀. Ha [HT] me$^{21c}$. Lh mə̀ʔ. Li mrghe$^{o}$. Na ñi$^{55}$. Sa ŋ̩$^{22}$s. Wo me$^{33}$ |

| | | |
|---|---|---|
| 167 | leech | *k-r-wat. Ak yeh$^{LS}$. Lh vè?. Li vé$^{o}$ |
| 185 | flower | *sə-wat. Ahi vi$^{44}$. Ak a-yeh$^{H-HS}$. Bi wɛ̃. Ha [HT] je. Ha [K] βæ$^{33}$. -LC e$^{55}$. Lh ɔ̀-vê?. Lh šɨ-vê?. Li [F] si$^{2}$-vé$^{3}$. Li [J] ve$^{33}$. Na vi$^{325}$. Sa vi$^{44}$ |

- Unlike *TGTM, Glides and Liquids cannot cooccur; they are in paradigmatic alternation. Thus there is no structure T$_S$IGLR$_S$.

These constraints are, however, implemented in the syllable canon used by the Reconstruction Engine. In the Reconstruction Engine, this structure is expressed slightly differently by the syllable canon:

*(185)*

[I,PI,Ø][L,G,Ø][OT,NT,SC]

That is, the syllable is composed of two possible optional constituents (simple initial (=I) or prefix plus initial (=PI)), followed by an optional liquid or glide, and terminating with either an open rhyme and an open syllable tone (OT), a nasal rhyme and open syllable tone (NT), or a stopped rhyme and stopped tone (SC). Expressing the prefix+initial as a single constituent in the table is required since prefixes are hypothesized to have affected the voicing of the initial (as well as other features) and therefore the tonal outcome (Matisoff 1972:14-22; Matisoff 1970). At the present point in the research, the distinction between open and nasal syllables is indicated simply as optional; none of the results of the program depend on it. It is coded in this way in the event that the distinction is needed in the future.

Note too that the context for the tonal split (i.e. voicing of the initial or onset (= prefix + initial) is also stated in the table (187 below) : *$^{H}$ results from voiceless initials, and *$^{L}$ from voiced. Stating context here disambiguates a number of possible reconstructions where the initial voicing distinction has been lost (that is, transphonologized) in the daughter languages (examples are shown in Chen Kang's Northern Loloish data set in Appendix 3.3).

The extensive merger of both initials and rhymes is captured. Evidence external to Loloish indicates that *-ut and *-ok, for example have merged in many of these languages, so that in many of these languages words like 'brain' (*TB *nuk (> PL *nok) STC 483) and 'blow' (*TB (s)-mut STC p.75) have the same rhymes:

*(186)*

| | | |
|---|---|---|
| s-mut$^{H}$ (TSR 143, DB-Plolo 690) | > | Axi mu̠$^{33}$ 'blow' |
| ʔnok$^{L}$ ≭ nok$^{L}$ (TSR 156(a.b)). C-nok$^{L}$ (DB-Plolo 140) | > | Axi nu̠$^{2}$ 'brain' |

The table (187) reflects this merger (and other mergers as well).

*(187) Partial table of correspondences for Northern Loloish (based on data from Chen Kang 1986)*

| (1) | (2) | (3) | | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # | T | PNL | | sani | axi | nesu | lalo | li | nasu | neisu | nosu |
| 1 | C | *L | $I_{VL-}$ | 44 | 33 | 33 | 33 | 33 | 2 | 33 | 4 |
| 2 | C | *H | $I_{VD-}$ | 2 | 2 | 2 | 2 | 2 | 55 | 13 | 55 |
| | | | | | | | | | | | |
| 6 | T | *1 | | 33 | 2 | 21 | 55 | 33 | 21 | 21 | 33 |
| 7 | T | *2 | | 55 | 55 | 2 | 2 | 55 | 55 | 55 | 55 |
| 8 | T | *3 | | 44 | 33 | 33 | 21 | 33 | 2 | 33 | 33 |
| | | | | | | | | | | | |
| 41 | I | *m | | m | m | m | m.ʔm | m | m | m | m.ṃ |
| 43 | I | *mb | | p | p | b | v.b | b | mb.bh | mb | b.mb |
| 44 | I | *n | | n | n | n | n.ʔn | n | ȵ.n | ȵ.n | ȵ.n.ṇ |
| 45 | I | *my | | n | n | n | m.ʔm | m | ɳ | ɳ | ȵ |
| 46 | I | *ny | | ŋ | n | n.ʔn | n.ny | n | ȵ | ȵ | n |
| | | | | | | | | | | | |
| 82 | N | *aŋ | | o.ı.u | o.u | u.o | i.u̠.u | u | u | u | u |
| 86 | N | *aŋ | | o.u | o.u | a | u | a.o | ɔ | o | o |
| | | | | | | | | | | | |
| 78 | O | *u | | u | u | u | u | u | u | u | ɯ |
| 80 | O | *uw | | u | o.u | u | i | ɯ | u | i | u |
| 79 | O | *uw | | ɯ.o | ɯ | ɯ | u̠.u.ɯ | ɯ.u | ɯ | y | o.u |
| | | | | | | | | | | | |
| 87 | S | *ak | | e | ę.ɛ | ę | i̠.i | ę | a.a̠ | a | o̠.a̠ |
| 96 | S | *ok | | ɤ.o | u̠.o̠ | u̠.u | o.o̠ | u̠ | o̠ | e | u̠.o̠ |
| 97 | S | *ök | | a | a̠ | a̠ | a̠ | a̠ | ɚ | a.ɪ.e | a̠.o̠.ɪ.i̠ |
| 98 | S | *uk | | ɯ | ɯ̠ | ɯ̠ | i̠ | ɯ̠ | ɯ̠ | ɤ | i̠.ɯ̠ |
| 99 | S | *ut | | u | u̠ | u̠ | u̠ | ɯ̠ | u̠ | u | u |

To generate the possible reconstructed forms which might be the ancestor of a given daughter form, the Reconstruction Engine divides the given form and looks up all possible segmentations of it in the table of correspondences. Consider, for example, a form from the Axi language:

(188) Axi nu̱$^2$ brain

There are in fact four ways to segment this simple monosyllable (assuming that u̱ is in fact a counts as a single glyph (§3.2.1.1):

(189) nu̱$^2$ n+u̱+$^2$ nu̱+$^2$ n+u̱$^2$

Of these four segmentations, only one (n+u̱+$^2$) is composed completely of elements which occur in the Axi column (5) of the table of *NL correspondences (i.e. of elements which according to the creator of the table to have some statable phonological significance).

(190) *Segmentations whose elements are ALL constituents of the table:*

| | | | |
|---|---|---|---|
| Table constituent: | n- | -u̱- | -$^2$ |
| Found in rows: | (44,45,46) | (96,99) | (2,6) |
| Slot type: | I | S | C,T |

(191) *Segmentations which contain elements NOT found in the table:*

| | Segmentation | Tokenized form |
|---|---|---|
| (c) | nu̱+$^2$ | (?)(2,6) |
| (d) | nu̱$^2$ | (?) |
| (e) | n+u̱$^2$ | (44,45,46)(?) |

For this segmentation, the Reconstruction Engine constructs a tokenized version of the form, in which each element of the segmented form is replaced with the correspondence or list of correspondences for that constituent in the table. Axi n-, for example, has three possible ancestors (*n-, my-, and *ny-, given by rows 44, 45, and 46 of the table of correspondences) when it occurs as an initial (i.e. syllable slot type I, column (2) of the table).

Were it not for syllabic structure constraints (expressed in column (2) of the table given in (187) the segmentation in (190) above, to wit

(192) n+u̲+$^{2}$ (44,45,46)(96,99)(2,6)

would produce 12 (= 3 x 2 x 2) different ancestors based on the different ordered combinations of its tokens:

(193)

| (1) | (2)<br>Row<br>Numbers | (3)<br>Syllable<br>Structure | (4)<br><br>Reconstruction |
|---|---|---|---|
| | 44.96.2 | I+S+C | nok$^{H}$ |
| * | 44.96.6 | I+S+T | |
| | 44.99.2 | I+S+C | nut$^{H}$ |
| * | 44.99.6 | I+S+T | |
| | 45.96.2 | I+S+C | myok$^{H}$ |
| * | 45.96.6 | I+S+T | |
| | 45.99.2 | I+S+C | myut$^{H}$ |
| * | 45.99.6 | I+S+T | |
| | 46.96.2 | I+S+C | nyok$^{H}$ |
| * | 46.96.6 | I+S+T | |
| | 46.99.2 | I+S+C | nyut$^{H}$ |
| * | 46.99.6 | I+S+T | |

*With* these constraints, however, the number of possible combinations that are possible is reduced to the six listed in column (4) in (193) above because the stopped tone (slot type C for 'closed') only cooccurs with stopped syllables (designated S for 'stopped'). All combinations in which a stopped tone ancestor would be generated with an open syllable are ruled out (indicated with * in column (1) of (193) above).

This situation is still not very satisfying: there are still six possible reconstructions as a result of the hypothesized mergers in initials and rhymes in this language. If we perform the same process with another form from another language, however, and compare the results, the situation improves considerably. Thus with Sani:

*(194)*

(a) Sani no$^{2}$ brain

| (b) | | | | |
|---|---|---|---|---|
| | Table constituent: | n | o | $^{2}$ |
| | Found in rows: | (44.45) | (79.82.86.96) | (2) |
| | Slot type: | I | O.S.N | C |

| (c) | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| | | Row | Syllable | |
| | | Numbers | Structure | Reconstruction |
| | * | 44.79.2 | I+O+C | |
| | * | 44.82.2 | I+N+C | |
| | * | 44.86.2 | I+N+C | nak$^{H}$ |
| | | 44.96.2 | I+S+C | nok$^{H}$ |
| | * | 45.79.2 | I+O+C | |
| | * | 45.82.2 | I+N+C | |
| | * | 45.86.2 | I+N+C | myak$^{H}$ |
| | | 45.96.2 | I+S+C | myok$^{H}$ |

Note that in the two sets of possible reconstructions for these two languages (Sani and Axi) only two are the same (indeed, these are also the two possible reconstructions in Sani):

| (195) | | | |
|---|---|---|---|
| | 44.96.2 | I+S+C | nok$^{H}$ |
| | 45.96.2 | I+S+C | myok$^{H}$ |

When this process is performed in parallel for the eight Northern Yi languages given in this data set, only *one* of the thirteen reconstructions which are possible for the entire set of forms can be (according to the table and canon) their common ancestor, as is shown in (196) below. Other possible reconstructions for various proper subsets of the daughter languages are listed (in order of number of supporting members) under the 'best' reconstruction.

(196)

| Possible | | Languages for which the reconstruction is possible | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Reconstructions | | sa | ax | ne | la | li | na | ne | no |
| nok$^{L}$ | 44.96.2 | x | x | x | x | x | x | x | x |
| nyok$^{L}$ | 46.96.2 | | x | x | | x | | | x |
| nok$^{L}$ | 44.96.3 | | | x | x | | x | | x |
| myok$^{L}$ | 45.96.2 | x | x | x | | | | | |
| nyok$^{L}$ | 46.96.3 | | | x | | | | | x |
| nyut$^{L}$ | 46.99.2 | | x | x | | | | | |
| mwok$^{L}$ | 42.96.2 | x | x | | | | | | |
| myok$^{H}$ | 44.96.4 | | x | | | | | x | |
| myok$^{L}$ | 44.95.2 | | | | x | | x | | |
| myok$^{L}$ | 44.95.3 | | | | x | | x | | |
| myut$^{L}$ | 44.99.2 | | x | x | | | | | |
| myut$^{L}$ | 45.99.2 | | x | x | | | | | |
| myök$^{L}$ | 44.97.2 | | | | | | | x | x |

The Reconstruction Engine generates and preserves this matrix of possible reconstructions in list form in the cognate set:

*(197)*

| *NL Set 55 | BRAIN | |
|---|---|---|
| nok$^{L}$ | 44.96.2. | [178.576.975.1374.1763.2155.2557.2950] |
| nyok$^{L}$ | 46.96.2. | [576.975.1763.2950] |
| nok$^{L}$ | 44.96.3. | [975.1374.2155.2950] |
| myok$^{L}$ | 45.96.2. | [178.576.975] |
| nyok$^{L}$ | 46.96.3. | [975.2950] |
| nyut$^{L}$ | 46.99.2. | [576.975] |
| mwok$^{L}$ | 42.96.2. | [178.576] |
| myok$^{H}$ | 44.96.4. | [576.2557] |
| myok$^{L}$ | 44.95.2. | [1374.2155] |
| myok$^{L}$ | 44.95.3. | [1374.2155] |
| myut$^{L}$ | 44.99.2. | [576.975] |
| myut$^{L}$ | 45.99.2. | [576.975] |
| myök$^{L}$ | 44.97.2. | [2557.2950] |

| | | | |
|---|---|---|---|
| 0178 | sani | no$^{2}$ | *brain* |
| 0576 | axi | nu̯$^{2}$ | *brain* |
| 0975 | nesu | nu̯$^{2}$ | *brain* |
| 1374 | ialo | ʔno̯$^{2}$ | *brain* |
| 1763 | li | nu̯$^{2}$ | *brain* |
| 2155 | nasu | no̯$^{55}$ | *brain* |
| 2557 | neisu | ne$^{13}$ | *brain* |
| 2950 | nosu | no̯$^{55}$ | *brain* |

Another set, this one for *eye*, is shown below in the internal Lexware format used by the Reconstruction Engine; there are numerous possible reconstructions for various subsets; it is the 'triangulation' process on the full set which disambiguates the 'correct' reconstruction:

*(198)*

```
.mrc myak$^{H}$/45.87.1. 297.694.1095.1493.1885.2278.2679.3070.]
rc nak$^{H}$/44.87.1. 297.694.1095.2278.2679.3070.]
rc mwak$^{H}$/42.87.1. 297.694.1885.]
rc mak$^{H}$/41.87.1. 1493.1885.]
rc nak$^{H}$/46.87.1. 694.1095.]
rc nak$^{L}$/44.87.3. 1095.2679.]
rc ŋak$^{H}$/47.87.1. 1885.2679.]
rc mwak$^{H}$/42.87.4. 694.1885.]
rc myak$^{H}$/45.87.4. 694.1885.]
rc myak$^{L}$/45.87.3. 1095.2679.]
rc nök$^{H}$/44.97.1. 2679.3070.]
rc myök$^{H}$/45.97.1. 2679.3070.]
pg              EYE
set             50
0297/sani       <ne$^{44}$>         eye
0694/axi        <nḛ$^{33}$>         eye
1095/nesu       <nḛ$^{33}$>         eye
1493/lalo       <?mḭ$^{33}$>        eye
1885/li         <mḛ$^{33}$>         eye
2278/nasu       <na̰$^{2}$>          eye
2679/neisu      <na$^{33}$>         eye
3070/nosu       <ɳo̰$^{4}$>          eye
```

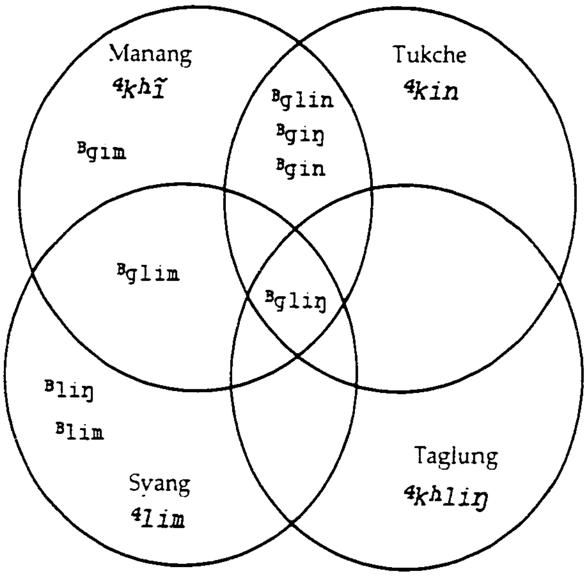
In the example above, each reconstruction is followed by a list of numbers which identifies the reflexes that support that reconstruction. In the final analysis, only the 'winning' reconstruction would be kept, of course. However, while research is still in process (that is, while the data sets are incomplete and the table of correspondences is in flux) it is useful to retain the variant reconstructions.

Another example of the 'triangulation' performed by the Reconstruction Engine, this time with forms for the word *snow* taken from four languages in the Tamang (TGTM) group, is exemplified in (199). Each column contains the possible reconstructions for the modern reflex that heads the column. A comparison of the columns (or an examination of the Venn diagram below) shows that only one reconstructed form, *$^{B}$gliŋ (in row 1), is supported by all the members of the cognate set; and we see that these four languages provide sufficient data to rule out some of the other reconstructions proposed on the basis of one language alone.[121]

*(199) Selecting the 'best' reconstruction from the list of possible reconstructions*

*SNOW*

| | Tukche | Manang | Syang | Taglung |
|---|---|---|---|---|
| | $^{4}$kin | $^{4}$k$^{h}$ĩ | $^{4}$lim | $^{4}$k$^{h}$liŋ |
| 1. | $^{B}$gliŋ | $^{B}$gliŋ | $^{B}$gliŋ | $^{B}$gliŋ |
| 2. | $^{B}$glin | $^{B}$glin | | |
| 3. | | $^{B}$glim | $^{B}$glim | |
| 4. | $^{B}$giŋ | $^{B}$giŋ | | |
| 5. | $^{B}$gin | $^{B}$gin | | |
| 6. | | $^{B}$gim | | |
| 7. | | | $^{B}$liŋ | |
| 8. | | | $^{B}$lim | |



$^{B}$gliŋ *snow* in Proto-Tamang
(8 possible protoforms produced from 4 reflexes)

### 6.4.4 The database management side of historical reconstruction

Using the interactive mode of the Reconstruction Engine described above is a good way to 'debug' the table of correspondences and canon. However, the Reconstruction Engine is most useful as a means of analyzing complete lexicons. The four steps involved in creating reasonable cognate sets from a set of lexicons of modern forms are schematized in (200) below. They are:

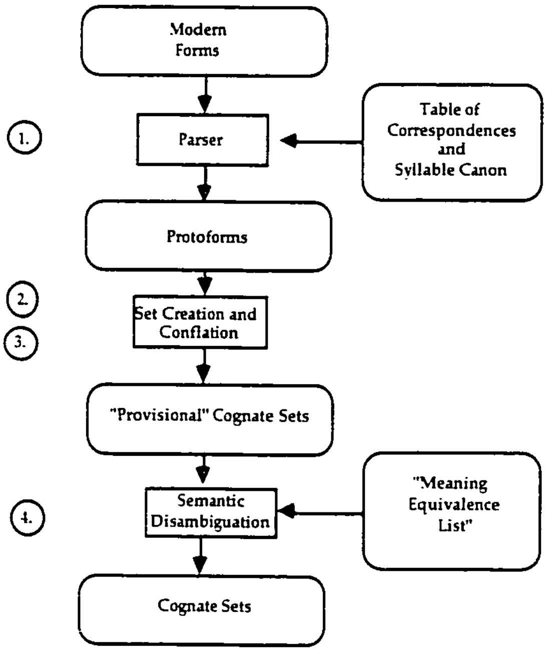
1. segmentation of lexemes and generation of proto-projections,

2. comparison of proto-projections and creation of tentative cognate sets,

3. merging (conflation) of subsets in the list of tentative cognate sets, and

4. conflict resolution within and between cognate sets of homophonous reflexes and homophonous reconstructions (via the application of semantic information).

(200) *Input-Output diagram of the Reconstruction Engine's basic batch functions*


Modern Forms
1.
Parser
Table of Correspondences and Syllable Canon
Protoforms
2.
3.
Set Creation and Conflation
"Provisional" Cognate Sets
4.
Semantic Disambiguation
"Meaning Equivalence List"
Cognate Sets

The algorithms for three of these four processes are outlined in the pseudocode in (201) through (203) below (the forth is not exemplified, for reasons to be explained). First, the Tokenize and FilterAndSubstitute procedures are performed for each form in each source dictionary.

(201) *Pseudocode for the Reconstruction Engine's basic batch functions - first create reconstructions*

```
/* STEP ONE: Backward projection of modern forms */
setup tables for the language data file to be processed
   get appropriate columns from TofC
   set language codes, etc. for output
   Initialize list of reconstructions
end setup
for each language_dictionary
   for each modern_form from language_dictionary
      Initialize TokenList
      Tokenize(modern_form)   /* see pseudocode for this function above */
      Initialize ListofPossibleForms
      FilterAndSubstitute(TokenList,ListofPossibleForms) /* upstream! */
      Apply Panini's Principle (Elsewhere condition) to select "allowed"
                                                        reconstructions
      check each reconstruction generated against list of reconstructions:
          if the reconstruction already exists in the list,
               link modern_form to existing reconstruction
          otherwise
               add reconstruction and link to modern_form into list
          end if
   end for
end for
```

Next, the list of reconstructions generated is examined and those reconstructions which fail to have sufficient support are eliminated. The remaining reconstructions are retained.

(202) *Pseudocode for the Reconstruction Engine's basic batch functions - next, create first group of cognate sets*

```
/* STEP TWO: create "pseudo" cognate sets */
for each reconstruction in list of reconstructions
     if reconstruction is supported by data from two or more languages then
           output reconstruction
           output supporting forms
     end if
end for
```

Third, each set is compared with each other set to get rid of those which are subsets of other sets (a type of 'set covering problem', discussed in §6.5.1 below). This is primarily a data reduction process, and not interesting algorithmically; we have therefore not provided the pseudocode describing it. It is, however, NP-hard, and therefore takes a lot of time for a data set of any size. The NP-hardness of this particular problem is treated in Appendix 5.

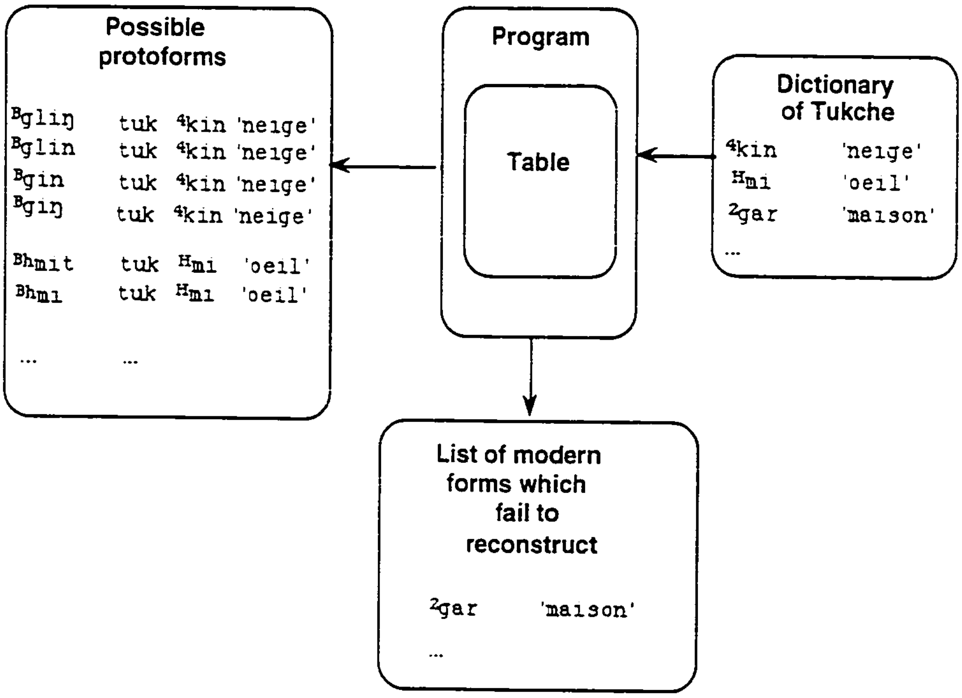
Finally, if the linguist is able (on the basis of analysis of previous runs) to provide semantic criteria for distinguishing homophones, the program can further partition the data into sets which contain only semantically compatible reflexes. The method for accomplishing this fourth step is described in §6.6.1 below.

(203) *Pseudocode for the Reconstruction Engine's basic batch functions - semantic component*

```
/* STEP FOUR: semantic processing: splitting and remerging of sets based on
semantics */
for each cognate set
  divide the set in two based on list of glosses selected
  for each of the newly created divided sets
      if (it is supported by data from at least two different languages)      and
              (it is not now a subset of some other existing cognate set)   then
              retain this half of the divided set
      otherwise
              delete this half of the divided set
      end if
  end for
end for

      /* check the division in the rest of the sets */
for all other sets containing any subset of these glosses
      if (there are words in this set with semantically incompatible
glosses) then
              divide the set (as was done above)
      end if
end for
output cognate sets
```

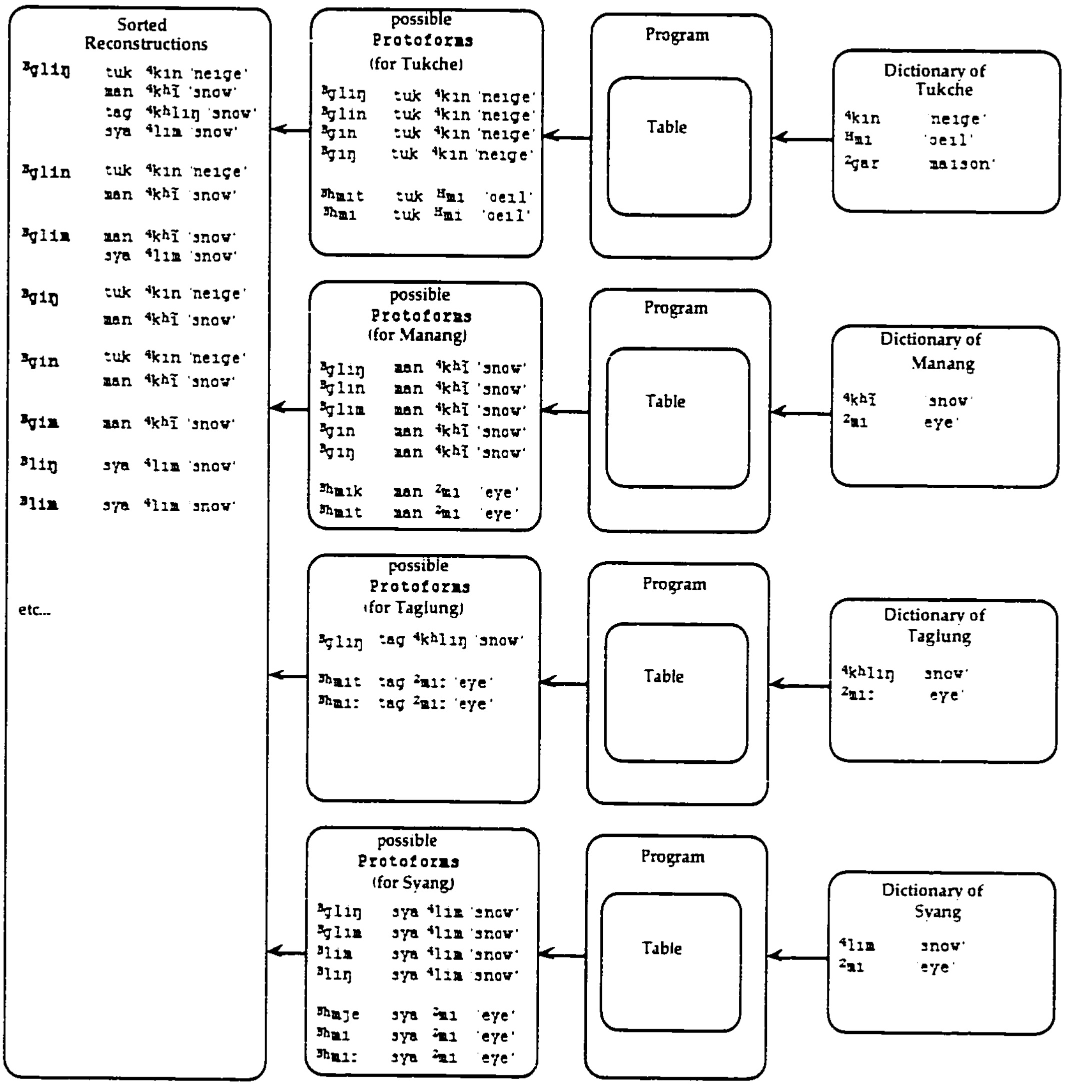
The first step, creating the list of proto-projections, is merely a matter of iteratively applying the reconstruction-generating procedures (already described in §6.4.1) to all the forms in the files. The list of protoforms obtained by running all the entries of a modern dictionary backward through the program is saved for later combination with reconstructions generated by words from other languages. The process is illustrated in (204). Note that forms which fail to produce any reconstructions are saved in a residue file for further analysis. In the example in (204) below, we see that a Nepali loan word, Tukche $^{2}$gar *house*, failed to produce any reconstructions in the proto-language, because there exists a phonological subsystem (substratum) for Nepali loans in Tukche which does not conform to the phonology of native words (i.e. to the phonology described by the table of correspondences).[122] In particular, no inherited voiced initials have survived in Tukche even as allophonic variants. In other cases, forms collected in the residue files may indicate a mistake in the table of correspondences, which will then need to be corrected to allow the words to reconstruct successfully. Note that the Tukche words in this example are glossed in French (*neige* 'snow', *oeil* 'eye' *maison* 'house'), as they are taken from a Tukche-French dictionary. This is a significant fact, as will be explained below in §6.6.1.

*(204) Proposition of protoforms and the residue ('check') file*



Combining the lists of reconstructions for several languages into a single sequence and sorting by the proposed reconstructions brings together all reflexes which descend from a particular reconstruction (205).

*(205) Upstream computation in batch*


Sorted Reconstructions
$^{3}$gliŋ tuk $^{4}$kin 'neige' man $^{4}$kʰĩ 'snow' tag $^{4}$kʰliŋ 'snow' sya $^{4}$lim 'snow'
$^{3}$glin tuk $^{4}$kin 'neige' man $^{4}$kʰĩ 'snow'
$^{3}$glim man $^{4}$kʰĩ 'snow' sya $^{4}$lim 'snow'
$^{3}$giŋ tuk $^{4}$kin 'neige' man $^{4}$kʰĩ 'snow'
$^{3}$gin tuk $^{4}$kin 'neige' man $^{4}$kʰĩ 'snow'
$^{3}$gim man $^{4}$kʰĩ 'snow'
$^{3}$liŋ sya $^{4}$lim 'snow'
$^{3}$lim sya $^{4}$lim 'snow'
etc...
possible Protoforms (for Tukche)
$^{3}$gliŋ tuk $^{4}$kin 'neige'
$^{3}$glin tuk $^{4}$kin 'neige'
$^{3}$gin tuk $^{4}$kin 'neige'
$^{3}$giŋ tuk $^{4}$kin 'neige'
$^{3}$hmit tuk $^{H}$mi 'oeil'
$^{3}$hmi tuk $^{H}$mi 'oeil'
Program
Table
Dictionary of Tukche
$^{4}$kin 'neige'
$^{H}$mi 'oeil'
$^{2}$gar 'maison'
possible Protoforms (for Manang)
$^{3}$gliŋ man $^{4}$kʰĩ 'snow'
$^{3}$glin man $^{4}$kʰĩ 'snow'
$^{3}$glim man $^{4}$kʰĩ 'snow'
$^{3}$gin man $^{4}$kʰĩ 'snow'
$^{3}$giŋ man $^{4}$kʰĩ 'snow'
$^{3}$hmik man $^{2}$mi 'eye'
$^{3}$hmit man $^{2}$mi 'eye'
Program
Table
Dictionary of Manang
$^{4}$kʰĩ 'snow'
$^{2}$mi 'eye'
possible Protoforms (for Taglung)
$^{3}$gliŋ tag $^{4}$kʰliŋ 'snow'
$^{3}$hmit tag $^{2}$mi: 'eye'
$^{3}$hmi: tag $^{2}$mi: 'eye'
Program
Table
Dictionary of Taglung
$^{4}$kʰliŋ 'snow'
$^{2}$mi: 'eye'
possible Protoforms (for Syang)
$^{3}$gliŋ sya $^{4}$lim 'snow'
$^{3}$glim sya $^{4}$lim 'snow'
$^{3}$lim sya $^{4}$lim 'snow'
$^{3}$liŋ sya $^{4}$lim 'snow'
$^{3}$hmje sya $^{2}$mi 'eye'
$^{3}$hmi sya $^{2}$mi 'eye'
$^{3}$hmi: sya $^{2}$mi 'eye'
Program
Table
Dictionary of Syang
$^{4}$lim 'snow'
$^{2}$mi 'eye'

From this sorted reconstruction list, the Reconstruction Engine extracts matching reconstructions, along with their supporting forms, and proposes them as potential cognate sets (206). Ideally, the rules would be sufficiently precise for the program to propose only valid sets, and yet sufficiently broad not to exclude legitimate possibilities. However, there is a certain amount of redundancy and uncertainty in the rules which tends to result in several possible reconstructions for the same cognate set. On the other hand, some forms which do produce possible reconstructions cannot be included in a cognate set because their reconstruction does not match that of any word in the other languages. These isolates (not illustrated) are collected by the program during the set creation process and maintained as a separate list.

The evaluation measure initially hypothesized for establishing the validity of a cognate set was the number of supporting forms. The program would retain a cognate set when the number of supporting forms from different languages reached a certain preset threshold value. However, many reasonable cognate sets had forms from only a few languages. The handling of this problem and other problems having to do with the composition of the proposed cognate sets is dealt with in more detail in §6.6 below.

*(206) A 'Cognate Set' generated by the Reconstruction Engine*

| | | |
|---|---|---|
| *TGTM $^{A}$ba: | 3.136.32. | |
| gha | $^{3}$po | leaf |
| man | $^{3}$pa: | leaf |
| ris | $^{3}$pa: | feuille d'arbre |
| sahu | $^{3}$pa: | leaf |
| tuk | $^{3}$pa | leaf |

## 6.5. THE CONSTITUENCY OF COGNATE SETS

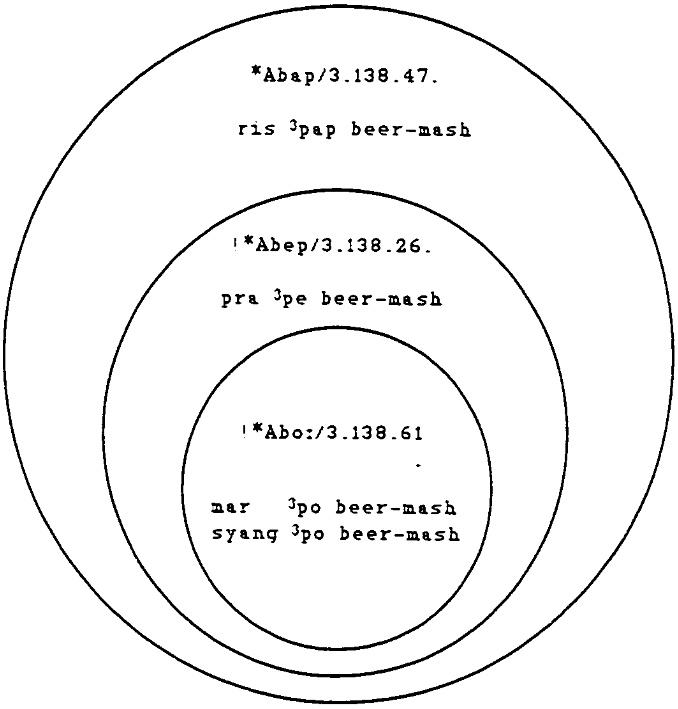
If sound change proceeded in such a way as to perfectly maintain semantic and phonological contrasts through time, the diachronic situation would be quite simple. Phonemes would change into other phonemes but not merge or split; words would mutate in phonological shape, but would remain distinct from other words in both form and meaning. Making cognate sets in such a situation would be quite straightforward. In reality, of course, neither semantic and phonological distinctions are maintained over time. We will now examine some of the implications of this situation.

### 6.5.1. Many reconstructions may be possible for a given set of cognates.

The process of 'triangulation' (discussed above in §6.5.1 and in some detail in Lowe and Mazaudon 1989, Lowe and Mazaudon 1990, Lowe and Mazaudon 1994) provides a means for selecting the best reconstruction out of several candidate reconstructions. If it were possible *a priori* to determine which reflexes might fall together based on the correspondences, one could preserve just those reconstructions from which all the reflexes might descend (and discard the other reconstructions). This situation is the one illustrated with the words for *snow* illustrated in (199) above. However, when entire lexicons are processed, it is not necessarily possible or even desirable to attempt to partition the lexemes into comparable sets to begin with. It is necessary to generate all possibilities and later to eliminate the undesirable ones through a sort of competition. Simply using the number of supporting forms in a cognate set as the sole or primary criterion for keeping the set, however, was found to be an inadequate heuristic: some perfectly good sets have only three members, while others with more members may be shaky and in some cases simply wrong.

An example of this competition is illustrated in (207). Here, as in the $^{B}$gliŋ *snow* example (199), several different cognate sets composed of the same reflexes but having different reconstructions have been generated. All reflexes clearly fall together into the same overall cognate set, but the reconstruction of several of the forms is non-unique: the reconstruction *$^{A}$bo: is supported by forms from Marpha and Syang, $^{A}$bep by three of the languages, while only *$^{A}$bap is supported by all four languages. This cognate set, then, reflects a merger of several smaller sets, as indicated in the Venn diagram. It is the Risiangku form that disambiguates the reconstructions, showing that *$^{A}$bap should be recognized as the winning reconstruction (it alone is marked with the *; other reconstructions supported by various subsets of the reflexes are marked with !*). As an aside, this process of set conflation can only be accomplished once all the words in all the lexicons have been processed. The problem then presented is a *set covering problem* (one of the various kinds of NP-complete problems for which no fast or easy solution exists, discussed in Appendix 5). In the case of the Tamang database, in which about 7,000 modern forms yield about 8,000 different reconstructions having supporting forms from at least two languages, set conflation takes several hours on the 486-based machine used for these analyses.

*(207) Nested cognate sets*



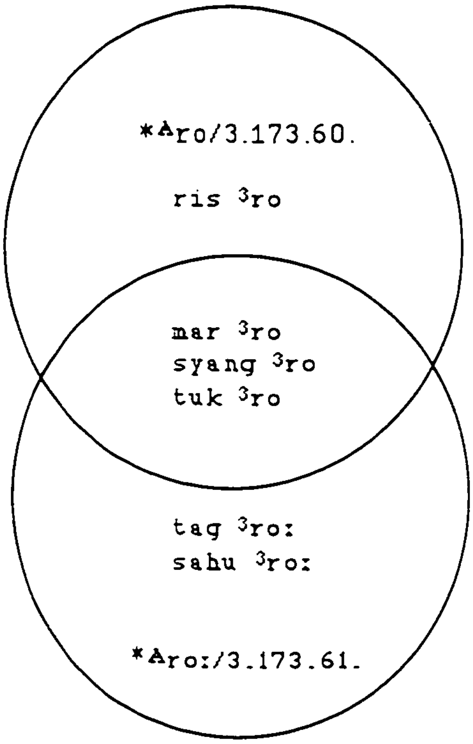
| | | |
|---|---|---|
| *$^{A}$bap | | **3.138.47.** |
| !*$^{A}$bep | | **3.138.26.** |
| !*$^{A}$bo: | | **3.138.61.** |
| mar | $^3$po | *beer-mash* |
| pra | $^3$pe | *beer-mash* |
| ris | $^3$pap | *beer-mash* |
| syang | $^3$po | *beer-mash* |

### 6.5.2. More complex kinds of competition for reflexes

A number of other types of overlapping can occur. In general, supporting forms for competing reconstructions do not nest as neatly as shown in the previous example. It is often the case that cognate sets may merely show partial overlap to varying extents. Such cases fall into several classes:

First, there exist some overlapping sets in which neither set is a proper subset of the other; the reflexes fit semantically, so the problem is one of phonologically reconciling the reconstructions. (208) illustrates the problems which arise when semantically compatible reflexes support different reconstructions.

*(208)* *$^{A}$ro(:) *'friend'*

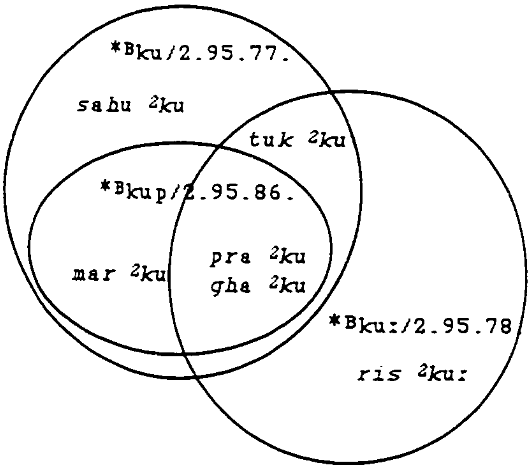


| | | | | |
|---|---|---|---|---|
| *$^{A}$ro | **3.173.60.** | | | |
| | | mar | $^{3}$ro | *friend* |
| | | ris | $^{3}$ro | *friend* |
| | | syang | $^{3}$ro | *friend* |
| | | tuk | $^{3}$ro | *friend* |
| *$^{A}$ro: | **3.173.61.** | | | |
| | | mar | $^{3}$ro | *friend* |
| | | sahu | $^{3}$ro: | *friend* |
| | | syang | $^{3}$ro | *friend* |
| | | tag | $^{3}$ro: | *friend* |
| | | tuk | $^{3}$ro | *friend* |

This particular situation results from an uncompensated merger (of length) which is now in progress: the length distinction appears to be on its way out in the Risiangku ('ris') dialect. This explanation is based on knowledge of the language and an internal analysis of its phonology. At this time it is not clear what algorithm (if any) could be used to sort out cases like these.

A similar but more complicated situation can be seen in (209) below. Here not only does the free variation in length in Risiangku generate an additional possibility, but those dialects where final consonants have been lost permit the reconstruction of a variant with a final stop. However, the form from the Risiangku dialect, which does preserve finals, cannot reconstruct (according to the table) to either a short vowel or a stopped rhyme, and so another cognate set supporting a long-vowel reconstruction is created. This result illustrates how crucially the fidelity and sharpness of the reconstruction is dependent on the accuracy of the modern forms (cf. also §3.2.1.10 on the theoretical significance of this phenomenon).

*(209)* *$^{B}$ku(:) *'nine'*



| | | |
|---|---|---|
| *$^{B}$ku | **2.95.77.** | |
| sahu | $^{2}$ku | *nine* |
| tuk | $^{2}$ku | *nine* |
| !*$^{B}$kup | **2.95.86.** | |
| gha | $^{2}$ku | *nine* |
| mar | $^{2}$ku | *nine* |
| pra | $^{2}$ku | *nine* |
| *$^{B}$ku: | **2.95.78.** | |
| gha | $^{2}$ku | *nine* |
| pra | $^{2}$ku | *nine* |
| ris | $^{2}$ku: | *nine* |
| tuk | $^{2}$ku | *nine* |

## 6.6. EXTENSIONS TO THE RECONSTRUCTION ENGINE

The processing described above produces cognate sets which are often found wanting upon closer inspection. For example, homophones may be conflated in the same set even in cases where they can (and indeed should) be derived from different etyma. Also, small irregularities in the data, perhaps only partially understood by the linguist, may cause forms to fail to reconstruct into the set in which they plausibly belong. We discuss below extensions to the Reconstruction Engine which deal with some of these problems.

### 6.6.1. CONSTRAINING COGNATE SETS: INCORPORATING SEMANTICS

Note that we have nowhere mentioned semantics in the backward computation process: in its search for new cognate sets, the algorithm works strictly on form, not on meaning. Indeed, one of the strengths of the Reconstruction Engine is that it initially ignores the glosses altogether, operating strictly on the forms themselves. This approach has the advantage of allowing a set like *$^{A}$ba: (shown in (206) above) to be created in spite of the fact that two different metalanguages (French and English, as illustrated in some of the earlier examples) are used in the data files for the glosses; meaning and glossing are simply not being taken into consideration in creating the set to start with. However, in cases of homophony in the protolexicon (like *$^{A}$bam. shown in (210)) it is plainly undesirable to conflate all the reflexes which might appear to support a given reconstruction. Here we can see that words belonging to at least two different etyma (*shoulder/thigh* vs. *soak /wet*) have been mixed into a single proposed etymon.

(210) *Sample of comparative set proposed by the Reconstruction Engine (without semantic component)*[123]

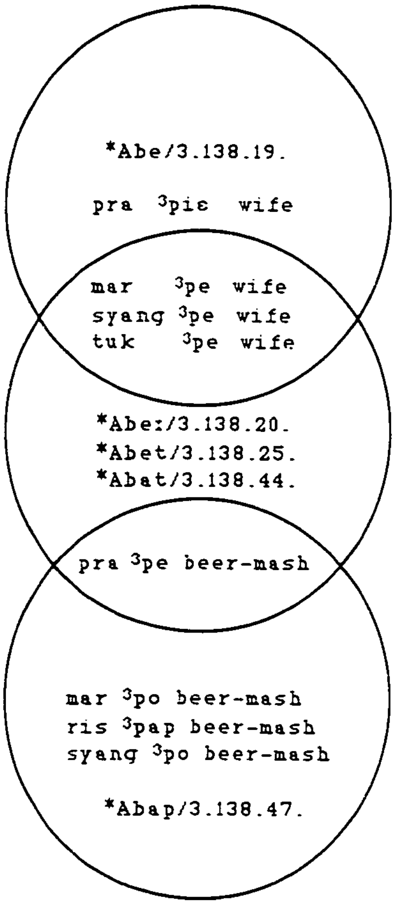
| TGTM *$^{A}$bam | 3.136.53. | |
|---|---|---|
| gha | $^{3}$pā: | thigh |
| tag | Xbam-ba | to soak |
| ris | $^{3}$pam | épaule |
| ris | $^{3}$pam | mouiller, tremper |
| sahu | $^{3}$pam | shoulder |
| tuk | $^{3}$pom | wet (v.i.) |

Besides allowing incompatible forms into an established cognate set, a lack of semantic differentiation permits the creation of some spurious cognate sets, as is illustrated in the Venn diagrams in (211). The set supporting the three reconstructions $^{A}$be:, $^{A}$bet, and $^{A}$bat should be eliminated and the reflexes permitted to migrate to the cognate set to which they are semantically related. Were some means available to specify *in this data set only* that no set could contain reflexes meaning both *wife* and *beer-mash*, the superfluous middle set and the overlap it causes would be eliminated (and the reconstructions would be listed under the *wife* set of which their supporting reflexes form a proper subset in the same way as depicted in (207}.[124]

(211) *$^A$be *and* *$^A$bap *'compete' for reflexes*

| | | | |
|---|---|---|---|
| *$^A$be | **3.138.19.** | | |
| | mar | $^3$pe | *wife* |
| | pra | $^3$piɛ | *wife* |
| | syang | $^3$pe | *wife* |
| | tuk | $^3$pe | *wife* |
| *$^A$be: | **3.138.20.** | | |
| *$^A$bet | **3.138.25.** | | |
| *$^A$bat | **3.138.44.** | | |
| | mar | $^3$pe | *wife* |
| | pra | $^3$pe | *beer-mash* |
| | syang | $^3$pe | *wife* |
| | tuk | $^3$pe | *wife* |

*$^A$bap (only the largest subset from (207) above is used here)



General theories of semantics and semantic shifts are not yet sufficiently developed to be used as an *a priori* reference framework to constrain the search for cognates.[125] Moreover, using such a framework would preclude the possibility of discovering any new semantic relationships which might be specific to the linguistic group or the linguistic area under study. So we do not wish to constrain the program at all in its first pass through the data in search of cognate sets. On subsequent passes though it would be convenient not to repeatedly encounter putative cognate sets which contain semantic discrepancies. It should be noted in passing that semantic discrepancies have nothing to do with semantic distance.

In order to separate incompatible etymological sets, we devised an *ad hoc* system of 'semantic tagging' using a structure we call an 'exclusion list.' These are bracketed lists of glosses specifying which glosses are semantically compatible and so might be found glossing reflexes in the same cognate set. The formalism used is as follows:

(212) $G_1, G_2, \ldots G_n]$

'Extract sets which contain any gloss $G_1$ to $G_n$ and eliminate from those sets reflexes with any other gloss. Eliminate any sets which as a result have too few members or become subsets of other sets.'[126]

(213) $G_1, G_2, \ldots G_m] \; G_{m+1}, G_{m+2}, \ldots G_n] \ldots G_{p+1}, G_{p+2}, \ldots G_q]$

'Divide any set which contains any of the glosses $G_1$ to $G_p$ into sets each of which contains reflexes which

- contain glosses only from one of the subsets $G_1, G_2 \ldots G_m]$, $G_{m+1}, G_{m+2} \ldots G_n]$, etc.; but
- retain any reflexes which are NOT specified in any of the subsets.

Eliminate any sets which as a result of the division have too few members or become subsets of other sets.'

Some examples of such exclusion lists are:

(214) SHOULDER, THIGH, ÉPAULE] MOUILLER, SOAK, TO SOAK, TREMPER, WET (V.I.)]

(215) BEER-MASH] WIFE]

(216) ANSWER, DIRE, REPLY, SAY, TELL, TO SAY]

(217) FEUILLE D 'ARBRE, LEAF, SMALL LEAF]
etc...

Specifically, these lists identify words *in the specific language data sets being processed* which are homophonous or might have homophonous reconstructions. They also bring together glosses from different languages which should be equated for purposes of diachronic comparison.

The exclusion lists are created based on examination of the initial sets proposed by the program. On subsequent passes through the data, the Reconstruction Engine can be instructed to take semantics into account and (by processing a file containing these lists) separate sets of potential cognates according to the exclusion lists.[127] The result is that reflexes which would otherwise fall together into one semantically incompatible cognate set can now be differentiated on the basis of meaning and separate sets can be created (see Figures 24 & 25 below). The set conflation procedure described in §6.5.1 can be applied after this differentiation to see if either of the resulting sets have become subsets of other sets. The result is a more reasonable list of cognates.

(218) *Sample of comparative cards proposed by the Reconstruction Engine (with semantic component) No separation (all glosses found in the same exclusion list, (12) above)*

| | | |
|---|---|---|
| *$^{A}$ba: | 3.136.32. | |
| gha | $^{3}$po | leaf |
| man | $^{3}$pa: | leaf |
| ris | $^{3}$pa: | feuille d'arbre |
| sahu | $^{3}$pa: | leaf |
| tuk | $^{3}$pa | leaf |

(219) *Sample of comparative cards proposed by the Reconstruction Engine (with semantic component) A set (exemplified in (210)) divided using exclusion list (9) above*

| | | |
|---|---|---|
| *$^{A}$bam | 3.136.53. | |
| gha | $^{3}$pā: | thigh |
| ris | $^{3}$pam | épaule |
| sahu | $^{3}$pam | shoulder |
| *$^{A}$bam | 3.136.53. | |
| ris | $^{3}$pam | mouiller, tremper |
| tag | Xbam-ba | to soak |
| tuk | $^{3}$pom | soak |
| tuk | $^{3}$pom | wet (v.i.) |

This device is presently conceived of as a simple tool to reduce noise in the output, but the exclusion lists might be studied later for an analysis of semantic shift in the particular group of languages. Note that the gloss lists are created from and therefore specific to the particular language data sets and glossing metalanguages used.

### 6.6.2. EXTENSION TO ALLOFAMIC[128] FAMILIES AND SYSTEMATIC IMPRECISION IN THE TABLE OF CORRESPONDENCES

As has often been noted and deplored, irregular sets of quasi-cognates, or simple groups of look-alikes are a necessary evil of comparative linguistics.[129] If we discarded immediately all sets that are not absolutely regular according to the rules of phonological change already uncovered for the language group, we would stand no chance of ever improving our understanding of the facts. Using a computer to mechanically apply a set of rules to the data implies that we believe to a large extent in the regularity of sound change. But the comparative method itself implies such an assumption. This does not mean that we cannot also admit that 'each word has its own history', when we take into account competing trends or influences on the languages. This flexibility towards irregularity has been incorporated everywhere in the computing mechanism.

One example of this flexible approach is evident in examining the table of correspondences. Turning back to the table excerpt (Figure 9a) notice the presence of multiple outcomes separated by commas in the columns of modern outcomes (columns 5-12). These mean that we do not know yet what the regular outcome of a given change is. Question marks are also allowed, in which case the program can (with the proper switch settings) borrow the outcomes from adjacent languages. Neither of these conventions should remain in the table of correspondences when the analysis of the group of languages is completed. But, as a working tool, the table of correspondences tolerates them, and they do not hamper the functioning of the Reconstruction Engine.

### 6.6.3. 'FUZZY' MATCHING IN THE TABLE OF CORRESPONDENCES

We can also reduce some of the specificity of the rules in the table of correspondences in a controlled way in order to produce 'allofam' sets or 'irregular cognate' sets. These cognate sets are composed of reflexes which are irregular only by one or two features specified as parameters by the linguist. This is accomplished using a 'fuzzy file' in which elements of the table of correspondences are conflated, allowing the linguist to systematically relax distinctions between segments. In (220), distinctions in tone and the mode of articulation at the proto-language level (indicated by the code TGTM) are being ignored in order to concentrate on patterns of correspondences between rhymes and between initial points of articulation.

*(220) A 'Fuzzy file'*

TGTM X=A.B
TGTM G=k.$k^h$.g
TGTM GR=kr.$k^h$r.gr
TGTM GL=kl.$k^h$l.gl
TGTM GW=kw.$k^h$w.gw
TGTM GJ=kj.$k^h$j.gj
...
TGTM P=p.$p^h$.b
TGTM PR=pr.$p^h$r.br
TGTM PL=pl.$p^h$l.bl
TGTM PW=pw.$p^h$w.bw

The 'fuzzy file', which states that TGTM proto tones $^{A}$ and $^{B}$ should be equated to the 'cover symbol' X , and that TGTM *k, *$k^h$, and *g should be conflated into *G, a velar stop, unspecified for aspiration and voicing. Similar conflations are stated for *p, *$p^h$, *b, and so on.

The result of conflating certain segments is to bring together certain reflexes which would otherwise fail to be included in cognate sets. (221) and (222) below illustrate how this procedure brings reflexes which are tonally irregular into the appropriate cognate sets for further study. In (221), an 'irregular correspondence' (that is, a lack of a correspondence where one might be expected to exist) results in forms from Tukche and Ghachok being left out of the cognate set. With a 'fuzzy' value for the tone (illustrated in part B of (221)), the two aberrant forms fall into place.

*(221)* $*^{A}$gla: ˜ $*^{B}$gla:

**A. Without 'Fuzzy' Constituents**

| | | |
|---|---|---|
| recon | $^{B}$gla: | |
| analysis | 4.117.32. | |
| | | |
| mar | $^{4}$lja | place |
| pra | $^{4}$k$^{h}$ja | place |
| sahu | $^{4}$kla: | place |
| syang | $^{4}$lja | place |
| tag | $^{4}$k$^{h}$la: | place |
| | | **gha** |
| | | **tuk** |

**B. With 'Fuzzy' Constituents**

| | | |
|---|---|---|
| recon | XGLa: | |
| analysis | 4.117.32. | |
| **analysis** | **3.116.32.** | |
| mar | $^{4}$lja | place |
| pra | $^{4}$k$^{h}$ja | place |
| sahu | $^{4}$kla: | place |
| syang | $^{4}$lja | place |
| tag | $^{4}$k$^{h}$la: | place |
| $^{3}$ɬo | **place** | |
| $^{3}$kja | **place** | |

In (222), the features conflated are tone and voicing.

*(222)* $*^{B}$pap ˜ $*^{A}$bap

**A. Without 'Fuzzy' Constituents**

| | | |
|---|---|---|
| recon | $^{A}$bap | |
| analysis | 3.136.46. | |
| | | |
| mar | $^{3}$po | beer-mash |
| pra | $^{3}$pe | beer-mash |
| ris | $^{3}$pap | beer-mash |
| syang | $^{3}$po | beer-mash |

**B. With 'Fuzzy' Constituents**

| | | |
|---|---|---|
| recon | XPap | |
| analysis | 3.136.46. | |
| **analysis** | **2.135.46.** | |
| mar | $^{3}$po | beer-mash |
| pra | $^{3}$pe | beer-mash |
| ris | $^{3}$pap | beer-mash |
| syang | $^{3}$po | beer-mash |
| **gha** | $^{-}$pa: | **beer-mash** |

### 6.6.4. Amplifying support for existing data sets

Like the data set for *NL, the data supplied by Hansson (1989) for *SL is a fairly uniform data set. The sets given are true cognate sets, and in many cases reconstructions (mainly from Bradley 1979) are supplied. Some caveats about the transcription are in order, however. First, in two of the languages (Akha and Hani), aspiration of initials is predictable:

voiceless initials are aspirated in open syllables, and unaspirated in checked ones.

The transcription of the Hani data from the Wordlist (see appendix 3.2 section C), described in Hansson 1982 used otherwise unused roman letters to indicate tone. The orthography is similar to the one used in the Lisu dictionary of Bradley's described in §4.2.4 above.

*(223)*

| | | |
|---|---|---|
| v | = | H |
| l | = | L |
| | = | mid |

Doubled voiced consonants are voiced, single voiceless consonants are voiceless aspirated, and /q/ indicated a stopped syllable (glottal stop in Hani):

*(224)*

| | | |
|---|---|---|
| p | = | ph |
| b | = | p |
| bb | = | b |
| t | = | th |
| d | = | t |
| dd | = | d |
| q | = | ʔ |
| etc. | | |

So, for example,

*(225)*

| | | | | |
|---|---|---|---|---|
| bbiavq | = | bja$^{HS}$ | = | |
| tul | = | thu$^{H}$ | = | |

Hansson supplies correspondences with the cognate sets she gives, and these have been used with the Reconstruction Engine to evaluate the reconstruction of *L. Based on the analysis I carried out with the Reconstruction Engine, I am able to offer a few improvements to her analysis, and indeed to use the same data set to further buttress the analysis she offers.

For example, Hansson offers four sets to support the initial correspondence for *L b:

(226) *Hansson's sets supporting *L b-*

| Set No | 498 | 53 | 267 | 488 |
|---|---|---|---|---|
| *Gloss* | *Rotten* | *Thin* | *Deaf* | *Give* |
| *PL | m-bup$^{L}$ | $^{2}$ba | $^{2}$baŋ | $^{2}$be |
| WB | $^{M}$pup | $^{M}$pâ | $^{M}$pâŋ | $^{M}$pê |
| Akha | $^{L}$buq | $^{L}$ba | $^{L}$bɔ | $^{L}$biq |
| Khatu | $^{L}$po | $^{L}$po | $^{L}$pu | $^{L}$pi |
| Pijo | - | $^{L}$pɔ | $^{L}$pu | $^{L}$pi? |
| Haoni | - | - | - | - |
| Mpi | $^{1}$pu? | $^{1}$po | $^{1}$poŋ | $^{5}$pe |
| HaniW | $^{MS}$bbuv | $^{MS}$bba | $^{MS}$bbo | $^{MS}$bbiv |
| HaniL | $^{L}$bu | $^{L}$ba | $^{L}$bo | $^{MS}$bi |

These are 'parade' example of this correspondence. In fact, the Reconstruction Engine can supply several additional sets which reflect this same correspondence (some of which also fall into the category 'parade example' it must be admitted):[130]

(227) *Additional sets proposed by the Reconstruction Engine for *L b*

75. $^{2}$ba [2.17.49] ⋇ $^{2}$ba [2.17.48] / CHEEK

| | | |
|---|---|---|
| [1] psl | $^{2}$ba | *Cheek* |
| [2] akha | $^{11}$ba | *Cheek* |
| [3] hanil | $^{11}$ba | *Cheek* |
| [4] khatu | $^{11}$pɔ | *Cheek* |
| [5] bi | $^{2}$po $^{4}$po | *Cheek.* |
| [1] wb | $^{2}$pa | *Cheek* |

299. $^{H}$bet [4.17.77] NUMB

| | | |
|---|---|---|
| [1] akha | $^{33}$by̯q | *Numb* |
| [2] khatu | $^{33}$pi | *Numb* |
| [3] pijo | $^{33}$pi | *Numb.* |

124. $^{1}$bin [3.17.81] ⋇ $^{H}$bap [4.17.71] / DIKE

| | | |
|---|---|---|
| [1] akha | $^{55}$baŋ | *Dike* |
| [2] hanil | $^{55}$bɔ | *Dike* |
| [3] khatu | $^{55}$py̯ | *Dike.* |
| [2] pijo | $^{55}$pu | *Dike.* |

284. $^{1}$ba [3.17.49] ⋇ $^{1}$ban [3.17.87] / MUCUS

| | | |
|---|---|---|
| [1] akha | $^{55}$bɛ | *Mucus* |
| [2] hanil | $^{55}$bɛ | *Mucus* |
| [3] khatu | $^{55}$pi | *Mucus* |
| [4] pijo | $^{55}$pi | *Mucus.* |

In the first two cases the reconstructions already existed in Hansson's data set. In the last four, the reconstructions are 'new': proposed by the Reconstruction Engine on the basis of correspondences found in other cognate sets supplied by Hansson. The Reconstruction Engine is useful in 'fleshing out' support for correspondence sets, relieving the linguist of a little of the tedium of identifying and annotating a correspondence table with all possible supporting sets. In most cases Hansson cautiously chose not to use those cognate sets for which a published *L reconstruction did not exist in exemplifying her correspondences or when some irregularity made the set incomplete. Here a more risky tack has been taken and some of the gaps filled in. The Reconstruction Engine will, of course, not fill in gaps with irregular forms, but if the set has enough support, the regular subset will be proposed.

In the case of *L open rhymes, Hansson's correspondences imply that the outcomes are at least partially conditioned by the initials. However, the degree of overlap between the different correspondence sets makes it difficult to verify that the differentiation is anything but observational. Consider the six correspondence sets for *L -i in different environments:

*(228)*

| *#* | *Type* | *PL* | *Context* | *WB* | *Akl* | *HaW* | *HaL* | *K* | *P* | *Hao* |
|---|---|---|---|---|---|---|---|---|---|---|
| 8 | R | *i | /L_ | i | i | i | i | i | i | |
| 9 | R | *i | /Z_ | i | i.y.e | i.y | i.y | y.ə.e | y.ə.e | i |
| 10 | R | *i | /C_ | i=e | i | y | y | y | y | |
| 11 | R | *i | /V_ | i | i.y | i.y | i.y | y.i | i | |
| 12 | R | *i | /F_ | i | i | y | y | y.i | y.i | |
| 13 | R | *i | /N_ | i | ö.y | ö.y | ö.y.ü | i.y | i.y | ɯ |

**Legend:**

L =
Z =
C =

V =
F =
N =

Given the twenty-one supporting cognate sets (seven are used in the main text; all are shown in (229) below), it is possible to collapse all possible outcomes in all environments into one correspondence row, without loss of specificity (in essence, the correspondence set for *-i after dental and alveolar stops is a superset of all the other outcomes).

(229) *Twenty one cognate sets supporting *L -i*

| *PL -i | *PL | WB | Akha | Hani W/L | Khàtú | Pîjɔ̀ | Háoní | Mpi |
|---|---|---|---|---|---|---|---|---|
| 125 Chilli | pi$^{2}$ | - | phí | pil/phí | phí | phí | - | phi$^{6}$ |
| 127 Close | pi$^{2}$ | - | phì | piq/phì | phì | phì | - | phi$^{1}$ |
| 129 Tear (eye) | - | - | bí | bbil/ - | pí | pí | - | m$^{4}$pi$^{6}$ |
| 130 Cat | mi$^{1}$ | - | mí | mil/mí | mí | njí | - | - |
| 21 Fire | C-mi$^{2}$ | mî | mí | miq/mì | mì | mì | - | mi$^{2}$ |
| 248 Tail | ʔ-mri$^{2}$ | mrî | mì | miq/mì | mì | mì | - | m$^{2}$pɑ$^{4}$ |
| 123 One | t/di$^{2}$ | thî | thì/tìq | qiq/tjhì | thỳ | khỳ | - | thɯʔ$^{2}$/tho$^{2}$ |
| 124 Hit | m-di$^{2}$ | tî | dì | ddiq/dì | tỳ | tỳ | ti$^{31}$ | tɯ$^{1}$ |
| 307 Heart | ni$^{3}$ | - | ny | nee/ny | nə | nə | - | no$^{4}$ |
| 185 Red | ʔni$^{1}$ | ni | né | nil/ní | ný | ný | ɳi$^{55}$ | nə$^{5}$ |
| 122 Two | s-ni(k)$^{2}$/$^{L}$ | hnac | njì/njìq | niq/njì | njè | njè | - | ɲi$^{2}$/ɲiʔ$^{2}$ |
| 112 Ride | dzi$^{2}$ | cî | dzì | zziq/dzỳ | tsỳ | tsỳ | - | tɯ$^{1}$ |
| 117 Liquor | ji$^{1}$ | se | djí | zzil/dzý | tsý | tjý | - | - |
| 115 Paddyhouse | ʔ-gyi$^{1}$ | kyi | djí | jjil/djí | tjhí khɔ́ | tjhítsh aŋkhu | - | - |
| 113 Lift | kyi$^{2}$ | khyî | tjhì | qiq/tjhì | - | - | - | tɕhi$^{1}$ |
| 303 Nest | - | - | gý | ggeel/gý | khý/tjhı | śjhí | - | khɯ$^{6}$ |
| 137 Sharpen | si$^{2}$ | - | shì | siiq/shỳ | shỳ | shỳ | - | sɯ$^{1}$ |
| 140 Fruit | si$^{2}$ | si | shì | siiq/shỳ | shì | shì | - | ʔɑ$^{2}$sɯ$^{2}$ |
| 366 Mo Br | ʔəri$^{1}$ | rî | ɣö | hhyu/ɣö | tji | ki | - | - |
| 319 Big | k/ʔ-ri$^{2}$ | krî | hỳ | heeq/xhỳ | xhỳ | xhỳ | xɯ$^{31}$ | hɯ$^{1}$ |
| 338 Old | ʔ-li$^{1}$ | - | ö | yul/ü | tjí | kí | - | li$^{5}$ |

Hansson gives only a single example of the /ö/ outcome after resonants in Akha and Hani, in 366 *mother's brother* and 338 *old.* I am regarding these as irregular and not including these as reflexes of *L -i.

Since the Reconstruction Engine generates all possible ancestors for a form, each of these correspondence lines results in at least one protoform being generated. There are now only 9 possible rhymes in Khatu and Pijo; however, from a purely computational point of view the rhyme /y/ in Pijo has 19 possible ancestors in *L (as shown in (230) below), and Khatu has 14 possible ancestors. Of course, in some cases /y/ is the only outcome of a certain protorhyme, in other cases, /y/ is one of the infrequent possibilities. There is currently no way of weighting the possibilities in the software, though clearly it would be useful to know which of the possibilities is the best-attested.

*(230) Possible ancestors of /y/ in Khatu and Pijo*

| # | T | PL | Khatu | Pijo |
|---|---|---|---|---|
| 88 | S | *ik | i | y |
| 62 | O | *u | u | y |
| 63 | O | *u | u.ɯ.ʔ | y |
| 52 | O | *ay | y | y |
| 67 | O | *wa | y | y |
| 68 | S | *ak | a.o.ɔ | y.a.aq.ɔ |
| 51 | O | *ay | u.ɹ | y.ɹ |
| 65 | O | *we | y | y.i |
| 54 | O | *e | y.i | y.ɹ |
| 58 | O | *i | y.ɹ | y.ɹ |
| 66 | O | *we | y.ɹ | y.i |
| 81 | N | *in | y.ɹ | y.ɹ |
| 53 | O | *ay | y.u.ɹ | y.ɹ |
| 60 | O | *o | y.ɹ.e.o | y.ɹ.e.o |

| | | | | |
|---|---|---|---|---|
| 77 | S | *et | y.əq.ə.ɹ | y.ɹ.yq.əq |
| 57 | O | *i | y.ɹ.ö.ü | y.ɹ.ə.e |
| 90 | N | *in | y.ɔ | y.ɔ |
| 71 | S | *ap | y.ɛq | y.u.ɛq |
| 59 | O | *i | y.ə.e | y.ə.e |
| 87 | N | *an | y.ɹ | i |

Analysis of Hansson's data using the Reconstruction Engine has turned up a number of interesting methodological points. A number of Hansson's sets are 'irregular' in the sense that some constituents of some forms fail to follow the regular pattern. Thus, set 233, *C-la 'ashes', is adduced to support the reconstruction of initial *C-l- but not to support the rhyme correspondence *-a, since the forms in the set do not follow this pattern. In fact, the words in these sets support correspondence sets for a nasal rhyme like *an or *wan much better. At any rate, the forms given are not regular reflexes of the cited reconstruction via the correspondence rows supplied. Whether this is an artifact of the data (as discussed in §3.2.1.10) or a genuine alternation reconstructible in the protolanguages remains to be shown.

(231) Set 233 *Ashes* from *Hansson 1989*; tentatively reconstructed Proto-Loloish *C-la$^{1}$

| | |
|---|---|
| Gloss | ashes |
| Akha(Thai) | xhà-lɛ́ |
| Akha(Yunn) | xhà lɛ́ |
| Hani (W) | haqleil |
| Hani (L) | xhà lɛ́ |
| Khàtú | mì sjhí. khà lí |
| Pîjɔ̀ | sjhí pho |
| *Mpi* | kho˥lo$^{6}$ |

(232) *The Reconstruction Engine cognate set for ASH based on data from Hansson 1989*

*$^{1}$lan [3.13.87] / *ASHES*

| | |
|---|---|
| [1] akha | $^{55}$lɛ |
| [2] hanil | $^{55}$lɛ |
| [3] khatu | $^{55}$sjhi |
| [4] pijo | $^{55}$li |
| [x] mpi | $^{2}$kho $^{6}$lo |

## 6.7. Conclusion

While the prototype software implementing the functions above has been demonstrated to be useful in a number of ways, it is awkward to use and cannot itself answer questions about the contents of the input or output files. It produces summary statistics of the results of its processing, but other software is required to search the files to evaluate the results in more depth. While the sophisticated computer user is not challenged by the need to develop other tools and routines to sort, search, and print the files, the average user is likely to need a tool with more 'bells and whistles,' to use the appropriate computer jargon. What would a more suitable tool look like? In the next chapter, I suggest some ways of combining the phonological and semantic apparatus of the Reconstruction Engine with a true database model of the input and output data structures.

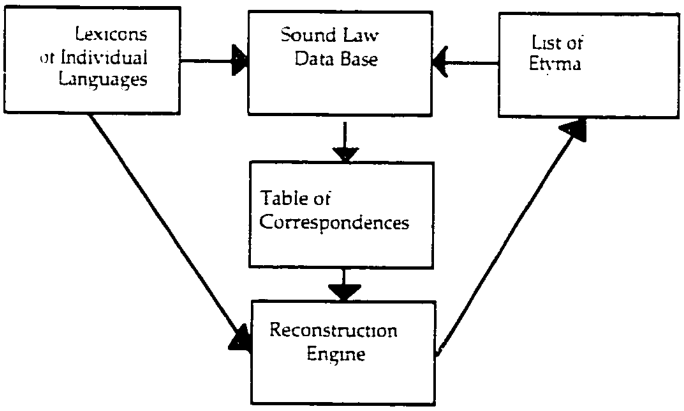
# 7. DESIDERATA FOR A SOUND LAW DATABASE

## 7.1. On the utility of a Sound Law Database ('SLDB')

For the hypothesis creation and testing apparatus described in Chapters 4 and 5 to be useful it must be connected to machine-readable data sets of the sort described in Chapter 3. None of the programs described in Chapter 2 provides anything like the full range of facilities needed. Indeed, no off-the-shelf software exists today which can manipulate large and complex sets of linguistic data in the ways required to apply the traditional methods of comparison, and developing such software is a challenging task. The Reconstruction Engine, described in Chapter 6 above, is a start on providing such a tool, and it addresses (and solves) many of the computational issues concerning the reconstruction process. And the Lexware program provides a flexible and general-purpose means for managing lexicographic data files used by the Reconstruction Engine. Both programs have a number of shortcomings, however, as they are based on an antiquated input-output model of processing: they read a set of input files, process them, and produce a set of output files. While this model is adequate for many purposes, it is an inconvenient way to carry out one's research, and represents in fact an earlier generation of thinking about software design. Today it is obvious that an online, interactive version based on a database model is the appropriate approach. However, it is only recently that software and hardware advances have made it possible to consider creating such a utility without a long and expensive development process. Programs such as MARIAMA and CUSHLEX (described in Chapter 2) are a start on implementing this approach; however, they lack many features that a complete historical linguistics software suite would have. This chapter is a speculation on how to build such a program, though I will try to keep the speculation within the realms of the feasible-with-current-technology.

A simple block diagram of the basic elements of a sound law database is shown in (233) below. This structure illustrates only the most crucial portion of the comparative process, that portion concerned with managing the relationship between modern lexicons and a reconstructed ancestor via a set of correspondences. Many other software components would have to be included in a full-blown system.

(233)



The sound law database indicated above is, therefore, merely a repository of relationships, of links between database elements: it links the individual forms in the modern lexicons, their reconstructions (if any), and the correspondences. It presumes that a number of other functions are taken care of by software components not shown. The components would handle data management tasks associated with the individual lexicons and other data structures, the treatment of semantics, manually entered annotations, sorting, output, and so on.

In fact, the component which handles that portion of the research concerned with diachrony sits on top (in the software design sense) of the machinery for data preparation. The sound law component operates for the most part on refined data, semantic and phonological abstractions prepared specifically for the purpose of applying the comparative method.

A sound law database of this type would serve three primary functions:

- as a *ready reference collection* of proposed, hypothesized, and published correspondences between a language and its putative ancestor. The contents of the catalog in this case would be based on the results of published research, and would be a bibliographic and encyclopedic reference tool.

- as a *research tool* for linguists working with the languages cataloged. An easy-to-use program with a simple and uniform format for importing and exporting data using a consistent representation would allow researchers to share results and verify each other's claims. Such uniformity of representations is common, indeed indispensable, in other research disciplines.

- as a convenient *repository* of information used by other programs and for other purposes. Data from a sound law database might be used in making comparative, monolingual, or bilingual dictionaries, or provide formatted examples to be incorporated into papers.

In these roles, the sound law database would act as a linguistic bookkeeping tool. All analysis and hypothesis creation would be done by the researcher. The sound law database would serve some of the research-assistant functions called for by scholars involved in the day-to-day work of historical reconstruction:

> Once there have been archived enough facts about the regular phonological developments in particular languages, the user of the [sound law] database will be able to test new etymologies by seeing if parallel examples of phonological correspondences have already been attested. Suppose, e.g., that a word for 'pig' of the shape phɛ is found in a certain TB language. Since the accepted *TB reconstruction of the most widespread etymon for 'pig' is *p-wak, the user will quickly be able to check the sound-law database to ascertain whether or not -ɛ is the 'regular reflex' of the rhyme *-wak in the language in question.
>
> (Matisoff 1993, NSF grant proposal)

The utility of such an electronic bookkeeper should not be belittled: gathering and organizing lexical evidence for a sizable number of languages or dialects is a time-consuming and painstaking task. The process is not a matter of creating a static structure for retrieval from a read-only source (except perhaps at the moment of initial loading), but of maintaining a constantly change and sometimes conflicting set of ideas about a number of hypotheses (à la MARIAMA §2.3.10).

Sound law database software would have to be highly customizable in order to be of use to other researchers for use in creating their own sound law databases. As previous chapters have made clear, the range of language types and data sources is enormous.

### 7.2. The interface

Interface design reconciles two opposing force: the need to have the computer perform certain chores under the control of the user and the limits of the software to carry out those chores. I will describe some of the ways the user may wish to interact to get what he needs from the sound law database, and at the same time indicate how they might be accomplished. The interface should be as natural and intuitive as possible, so that the user does not have to spend time and energy imagining how something might look or work. Thus, sound correspondences should look like sound correspondences as seen in books and articles; entries for words should look like dictionaries, and etymologies should look like they do in the standard sources, but better, of course. Some of these I need not detail here as they are easy to imagine or implemented elsewhere. I will, however, investigate how the concept of 'sound law' should be represented.

### 7.2.1. Several conceptions of the notion 'sound law'

There are at least three common (and related) visualizations of the notions 'sound law:'

- as a correspondences pair between phonological constituents in two languages. A correspondence usually on gains the status of 'sound law' when it has unassailable support and a bit of history behind it.

- as sets of correspondences between several languages (the matrix view); here

- as phonological rules describing the transformation of individual proto-elements into modern elements.

### 7.2.2. Correspondence pairs

In its most elemental version, a sound law is a correspondence between sounds in two languages which, and expressed as a proportion *a:b*. This formalism and the interpretation of the historical circumstances has been masterfully elaborated by a number of linguists, for example Hoenigswald 1960 and Anttila 1995. Extending the formalism to handle the case of several languages and several correspondences (i.e. sets of 'n-ary correspondences') results in the familiar matrix or tabular view of correspondences (a version of which has been discussed in detaii above with respect to the Reconstruction Engine in §6). The interface should include this conception of relation as a means of viewing the database contents.

### 7.2.3. The tables metaphor

In this approximation or view, sound laws appear as cells in a two-dimensional matrix, the columns of which indicate the languages in which a correspondence is observed, the rows the outcomes of the various protosegments (or vice versa).

The example on the following page (234) shows some of the complications to a strict '2-D matrix idea.' In encodes:

- some orthographic variants (e.g. WB *pl* (ins.), 'dkv' in Nakhi),
- some synchronic conditioning (ts before i in Lahu),
- uncertainty, in the form of empty cells (perhaps interpreted as correspondences as yet unattested) or question marks (perhaps interpreted as unstatable or difficult),
- unconditioned alternates in modern language (of two types 'or' and x/x/x)

*(234) Lolo-Burmese positions )f articulations (Matisoff 1979:31)*

| PLB | WB | Lahu | Lisu | Akha | Sani | Bisu | LC | Nakhi |
|---|---|---|---|---|---|---|---|---|
| *p | p | p | p | p | p | p | p | p |
| *pr | pr | p | p | p | tɬ | p | *pw>p<br>*pr>ṭ | p<br>*mr>ž |
| *pl | pl(insc)<br>>pr/py | p<br>(pw?) | p | py | tɬ | pl | ṭ | p |
| *py | py | p | py | py | tɬ | py | ṭ | p |
| *t | t | t | t | t | tl | t | t | t |
| *ts | ts or<br>tš | tš<br>~ts/_ɨ | ts | ts | tl | ts | ts | ts/tš |
| *tš | ts or<br>tš | tš<br>~ts/_ɨ | tš | tš | ts/tš/tṣ | ts | ts | tš/tš |
| *ky | ky | tš<br>~ts/_ɨ | tš | tš | ts/tš/tṣ | tš/ky | tṣ | ιš/ky |
| *kl | kl(insc)<br>>kr/ky | tš | ts | $t^{š}$ | $t^{š}$ | kl | tṣ | "dky<br>~ dty" |
| *kr | kr | k | $t^{š}$ | k | ts/tš/k | ky/k | tš | tš/k<br>"t'ky" |
| *kw | kw | p | ?? | k | ts | k | tš | k |
| *k | k | q | k | x | q | k | k | k |

It is a useful way to lay out correspondences for comparison and is easy to process, but it is merely a shorthand, and lacks several important features:

- The structure assumes that all the languages are directly reconstructible at the level of a single common ancestor. Thus, to handle reconstruction at the subgroup (mesolanguage) level, the structure must be repeated or otherwise amplified (perhaps as suggested in (236).

- Even at the level of a single subgroup, the structure is limited in its ability to record the differing reflexes which may result from context-dependent change (i.e. Cells would have to have a complex structure (as indicated, for example in the outcomes of *ts, *tš, and *ky in Lahu) or the columns would have to be repeated).

- Other important information relevant to the correspondence cannot be recorded; in particular, data such as the source of information (if any), list of the forms in which the correspondence is observed, list of counterexamples, and notes might all be linked to a cell. In the text accompanying (234), for example, Matisoff notes that the merger of *ts to tl in Sani is well established, being supported by 'such key roots as *fill, bee, flee, flat*, and *to fly*.' (Matisoff 1979:30)

### 7.2.4. Sound laws as phonological rules

Specifying an ancestor for an observed correspondence set gives a diachronic orientation to the formalism: a correspondence set can be interpreted as a set of simultaneous 'rules' in which the ancestor 'changes' into the daughter. Of course, the trivial case in which ancestor and daughter remain identical is as significant as the case in which a change is observed.

Converting algorithmically from correspondence to rules is a trivial process, the results of which are shown below in (235).

(235) *Correspondences vs. 'Sound Laws' in *TGTM (after Mazaudon and Lowe 1991)*

| N | Type | TGTM | Context | ris | sahu | tag | tuk | mar | syang | gha | pra |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 47 | R | *ap | | ap | ap | ap | əp.əu | o | o | a: | e |
| 94 | I | *k | /_w | k | Ø | h | k | k | k | k | k |

| Rule (Corr) | Type | TGTM | | Outcome | Context | Language(s) |
|---|---|---|---|---|---|---|
| 181(0047) | R | *ap | > | ap | | ris.sahu.tag |
| 182(0047) | R | *ap | > | a: | | gha |
| 183(0047) | R | *ap | > | e | | pra |
| 184(0047) | R | *ap | > | o | | mar.syang |
| 185(0047) | R | *ap | > | əp.əu | | tuk |
| 365(0094) | I | *k | > | h | /_w | tag |
| 366(0094) | I | *k | > | k | /_w | ris.tuk.mar.syang.gha.pra |
| 367(0094) | I | *k | > | Ø | /_w | sahu |

This view essentially turns the correspondences 'sideways' giving a diachronic view of the laws as a chain of changes (as opposed to the tabular view, which emphasizes the observed variations).

### 7.2.5. Ordered rules and tables

A sound law database must be able to provide a means to view or manipulate the sound laws according to accepted linguistic principles. Several levels of genetic affiliation may be required to explain the observed distribution of forms. The tree model of genetic affiliation can be used to accomplish this, linking correspondence tables together as illustrated below (236). The 'chain' of tables derives modern Loloish forms from *LB according in three stages of development: from *LB to *L, from *L into one of the three subgroups hypothesized for *L (Northern, Central, and Southern), and from each of these subgroups into the modern attested daughter. The Burmish subgroup, including Written Burmese (WB) is shown diverging from the *LB parent.

(236) *'Chained' tables of correspondence*[131]

| **LB* | **L* | **B* |
|---|---|---|
| *d | *d | d |
| *g | *g | g |
| *k | *k | g |
| *l | *l | l |

| Proto-Loloish | 'Nasoid' | 'Lahoid' | 'Mosoid' |
|---|---|---|---|
| **L* | **NL* | **CL* | **SL* |
| *d | *d | *d | *d |
| *g | *k | *k | *g |
| *k | *g | *g | *k |
| *l | *ɬ | *l | *l |

| Burmish | |
|---|---|
| **B* | *WB...* |
| *d | d |
| *g | k |
| *k | k |
| *l | l |

| **NL* | *sani* | *axi* | *nesu ...* | **CL* | *ak* | *lh* | *lc ...* | **SL* | *Akha* | *Khatu* | *Pijo ...* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *d | t | t | d | *b | b | p | p | *d | d.th | d.tj.t | t.tj |
| *k | kh | kh | kh | *d | d | t | ʔ | *g | g.ɣ | k.tj | k |
| *k | k.qh | k | k | *dž | j.c | c | dẓ | *k | k | kh.tjh | kh |
| *g | q | g.k | g | *g | x | q | g | *l | l | xh.sjh.ɹ | l |
| *l | ɬ | l | ɬ | *k | k· | qh.q | k· | | | | |

Based on Chen Kang 1986 Based on TSR Based on Hansson 1989

This sort of structure would also be able to accommodate a number of types of diachronic changes in syllable and morpheme structure, as each table (representing each intermediate mesolanguage) would have its own phonotactic constraints (syllable canon) attached.

Computationally, such a model could become quite large and complex. While it is not clear how many cells would be required for each stage of development (i.e. each individual table of correspondences) even a few stages would probably have several thousand cells.[132] So far, based on the experiments with Proto-Tamang and Loloish described above, it appears that each stage requires a table with approximately two hundred rows. This may be an artifact of the type of analysis chosen, since a phonological description based on initials and rhymes tends to have more elements that one based on segments. However, it is a good starting point. If we make a few more assumptions:

- that a maximum of five or six levels are reconstructible (this figures appears to be a reasonable approximation for both Indo-European and Sino-Tibetan );[133]

- that each subgroup table has around eight to ten languages (this assumption made exclusively to permit assigning a breadth to the table);

then a language family of five hundred languages would have breadth of between 50 and 63 tables at the bottom of the tree. These tables would in turn link upwards to five or six other tables (i.e. relating the first stage mesolanguages together). The tree would obviously not be dense like this, but this approximation indicates that there would be between 50 and 100 total tables, each with, as estimated before, about 200 rows. These tables would have around a million cells.

### 7.2.6. Some statistics worth keeping track of

Each data point in the database, whether it is an attested form in a cognate sets, an etymon, or a counterexample contributes (or detracts) a certain amount to the solidity of the reconstruction as a whole. The different elements do not contribute equally. This is one of the biggest weaknesses of the computational approach to historical linguistics: each element has attached to it some measure of its value in a particular circumstance, and this is difficult if not impossible to capture in a database structure. It would be possible, for example to assign on some basis a 'grade' to each reconstruction. (Matisoff 1990b) Certainly some reconstructions are better than others by virtue of the number of supporting forms, the regularity and plausibility of the correspondences they are based on, the goodness of the semantic fit, and so on. Matisoff (1990) notes that there is a typology of irregularities which must be considered when judging possible cognates. In this typology for Loloish reconstructions, forms can be The grade assigned is thus a multivariate measure which cannot be accurately reflected in a single value. Nevertheless, some evaluation is better than no evaluation, and we can hope to improve the metrics as time goes on.

The sound law database should provide for both computer-generated metrics as well as human intuition. The software can, for example, count the number of supporting forms for a cognate set or correspondence and provide this to the user. We can imagine a number of interesting situations where statistics of this type might be useful. If a pair of correspondence lines of the form shown below were to appear in a table of correspondences, we might wish to investigate further.

| Corr. | Frequency | | Correspondences & N of forms | | | |
|---|---|---|---|---|---|---|
| No. | (# of sets) | *Lg | Lg1 | Lg2 | Lg3 | Lg4 |
| 97 | 10 | *A | a | b | c | d |
| | | | 10 | 10 | 1 | 1 |
| 98 | 10 | *B | e | f | g | h |
| | | | 1 | 1 | 10 | 10 |

The numeric contribution of Languages Lg3 and Lg4 (in terms of number of supporting forms) to correspondence row 97 is negligible, as is the contribution of languages Lg1 and Lg2 to reconstructions exemplifying row 98. In fact, there is a rather complex arithmetic involved here which depends on the distribution of supporting forms among the sets and correspondence rows (imagine that there were no examples of correspondence row 97 in Lg3 and Lg4; the numbers of supporting forms in both Lg1 and Lg2 would both have to be 10 in order to justify having the row at all). Analysis of existing and reconstructions along these lines might produce interesting results, and a sound law database could provide this capability.

A sound law database could provide another useful 'grade' for etyma and their associated cognate sets: if the strength of a correspondence is based (in part) on the number of times it is observed, and if solidity of a reconstruction is a reflection of the number of reflexes it has, then it follows that reconstructions with many reflexes exemplifying commonly occurring correspondences should receive a high grade. If a reconstruction is only moderately supported, but the correspondences of which it is composed are very robust, this should count for something. This is easily quantifiable. Consider two reconstructions, both having 10 supporting forms, and sharing a correspondence (in this case they have the same rhyme). The initial *l* in the first is not as well attested as the *p* in the second. Should we conclude that the second reconstruction is better supported than the first? Probably, but not by much.

| | Initial | Rhyme | Number of supporting forms in set |
|---|---|---|---|
| | *l- | *-ak | 10 |
| Overall f | 20 | 20 | (i.e. number of sets using the corr.) |
| | *p- | *-ak | 10 |
| Overall f | 50 | 20 | " |

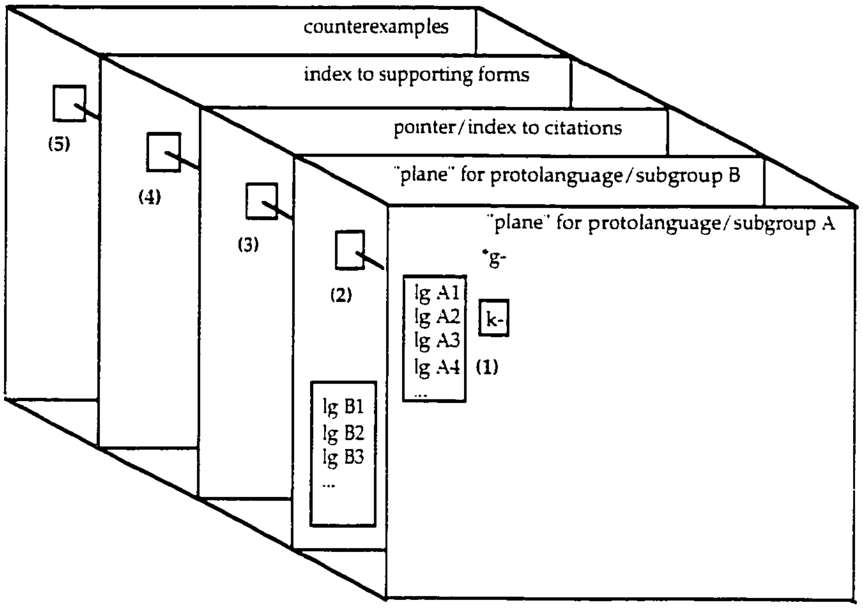
It is easy to propose such measures, but only empirical testing will show whether they have any value. Nevertheless, it seems likely that a good sound law database would have a substantial statistical component that is able to provide this type of information; without it, many of these questions will remain unanswered.

## 7.3. Integrating these conceptualizations

### 7.3.1. The n-dimensional space of sound laws

An extended data model should not therefore be two-dimensional but n-dimensional, each dimension handling one of the salient kinds of information relevant to the given correspondence. Any particular correspondence would thus have associated with it a number of linked data elements. Extending the 'table' metaphor, it should be possible to 'look behind' each table entry to see the other tables in which the entry participates, as well as to pull up a variety of other information:

(237)


counterexamples
index to supporting forms
pointer/index to citations
"plane" for protolanguage/subgroup B
"plane" for protolanguage/subgroup A
*g-
lg A1
lg A2
lg A3
lg A4
...
k-
(1)
(2)
(3)
(4)
(5)
lg B1
lg B2
lg B3
...

Actually, each correspondence *x:x would be a 'record' (in the database sense) in its own right. Even a short list of desiderata for inclusion as fields in such a record shows that it would be a relatively complex structure:

language
ancestor
protolanguage/subgroup
links to other sound law records for which this one is an ancestor or a descendant
time depth/chronological or order of application
context (phonological)
type of constituent (rhyme, initial, etc.)
source citation(s)
estimate of reliability/'grade'
list of firm supporting forms
list of possible supporting forms
other statistics
counterexamples and problems

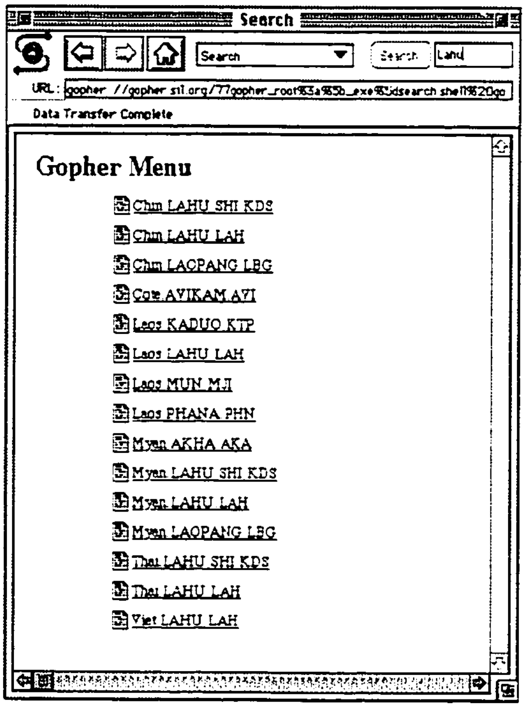
notes and comments
links to phonological component: statements in terms of features,
position in phonological inventory, etc.

A 'spreadsheet' or 'matrix' view of the data would be one of the possible 'user views' into the sound law database. It could be created 'on the fly', based on queries from the researcher. Creating such a structure would involve access to a great deal of other information in principle already available in the database: lexicons of modern languages, bibliographies, synchronic phonological analyses and inventories, and subgroup classifications. This brings up another function of the sound law database, that of reference and pedagogical tool.

### 7.3.2. Encyclopedic functions of a sound law database

A basic function of the sound law database is to provide access to what is already known about the languages in the database, both in the sense of providing dictionary and thesaurus access, as well as more encyclopedic information. Some of this functionality is already available, and the access to it is worth describing. The Ethnologue World Wide Web site provides an example of how such an encyclopedic facility would work and of the type of information it would provide. Searching the Ethnologue may be done using a Gopher interface, as shown below in (238). The term *Lahu* was entered (in the Search box at upper right), and a number of Lahu dialects (subcategorized by country loconym) are returned as possible answers.

*(238) Gopher interface to Ethnologue*



Search: Land

URL: gopher //gopher sil.org/77gopher_root%3a%5b_exe%5dsearch shell%20go

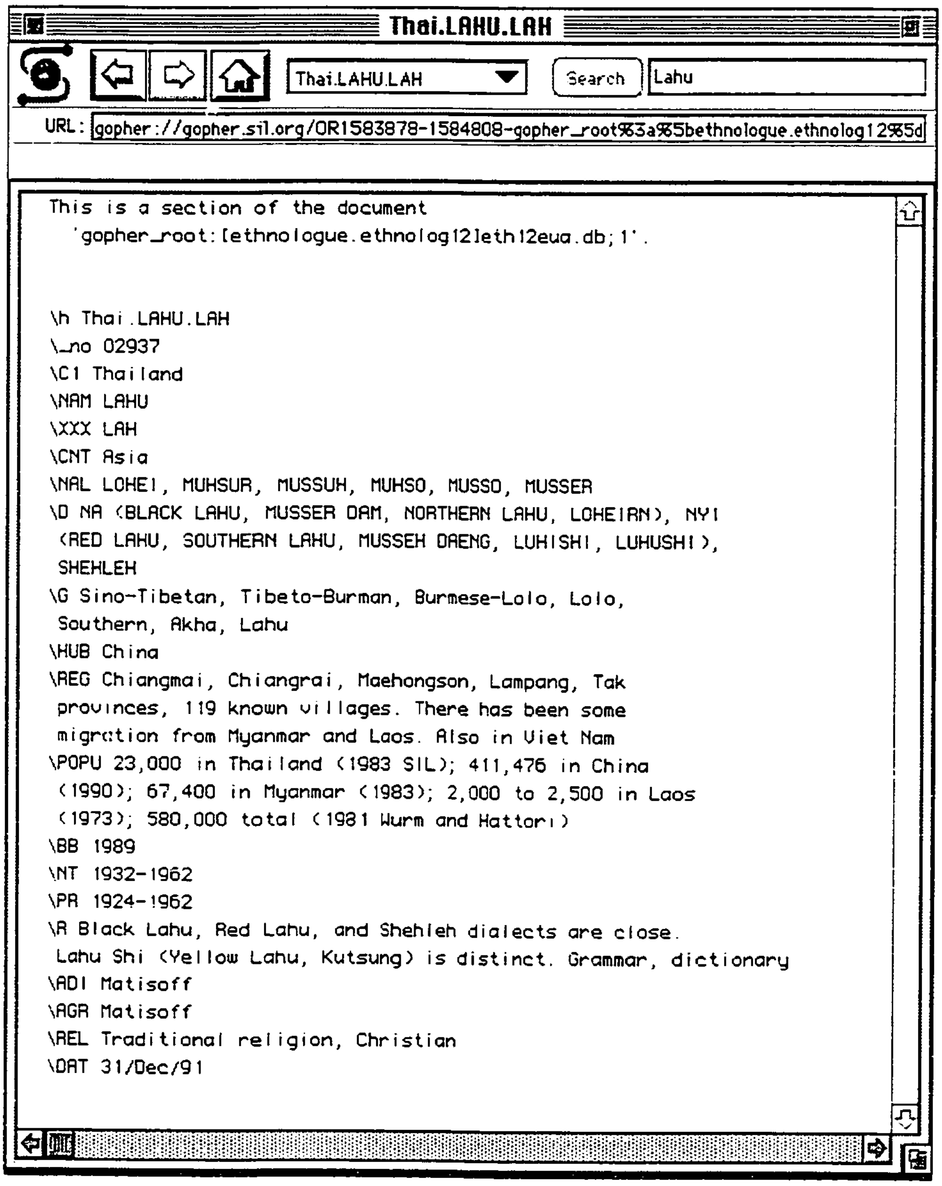
Data Transfer Complete

**Gopher Menu**

- Chm LAHU SHI KDS
- Chm LAHU LAH
- Chm LAOPANG LBG
- Cote AVIKAM AVI
- Laos KADUO KTP
- Laos LAHU LAH
- Laos MUN MJI
- Laos PHANA PHN
- Myan AKHA AKA
- Myan LAHU SHI KDS
- Myan LAHU LAH
- Myan LAOPANG LBG
- Thai LAHU SHI KDS
- Thai LAHU LAH
- Viet LAHU LAH

Selecting a particular dialect (by double clicking) retrieves the specific entry. Note that a variety of information is included, such as distribution, subgrouping, allonyms, and so on.

*(239) A language entry from the WWW interface to Ethnologue*



URL: gopher://gopher.sil.org/OR1583878-1584808-gopher_root%3a%5bethnologue.ethnolog12%5d

```
This is a section of the document
  'gopher_root:[ethnologue.ethnolog12]eth12eua.db;1'.



\h Thai.LAHU.LAH
\_no 02937
\C1 Thailand
\NAM LAHU
\XXX LAH
\CNT Asia
\NAL LOHEI, MUHSUR, MUSSUH, MUHSO, MUSSO, MUSSER
\D NA (BLACK LAHU, MUSSER DAM, NORTHERN LAHU, LOHEIRN), NYI
 (RED LAHU, SOUTHERN LAHU, MUSSEH DAENG, LUHISHI, LUHUSHI),
 SHEHLEH
\G Sino-Tibetan, Tibeto-Burman, Burmese-Lolo, Lolo,
 Southern, Akha, Lahu
\HUB China
\REG Chiangmai, Chiangrai, Maehongson, Lampang, Tak
 provinces, 119 known villages. There has been some
 migration from Myanmar and Laos. Also in Viet Nam
\POPU 23,000 in Thailand (1983 SIL); 411,476 in China
 (1990); 67,400 in Myanmar (1983); 2,000 to 2,500 in Laos
 (1973); 580,000 total (1981 Wurm and Hattori)
\BB 1989
\NT 1932-1962
\PR 1924-1962
\R Black Lahu, Red Lahu, and Shehleh dialects are close.
 Lahu Shi (Yellow Lahu, Kutsung) is distinct. Grammar, dictionary
\ADI Matisoff
\AGR Matisoff
\REL Traditional religion, Christian
\DAT 31/Dec/91
```

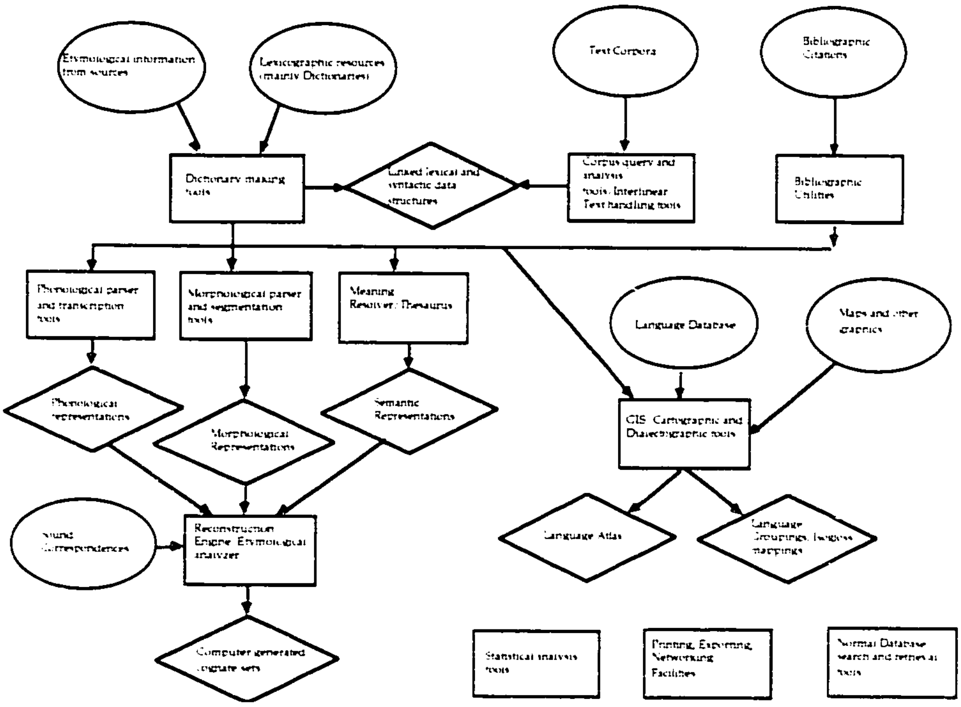
Another way to access this type of information is, of course, through the agency of a 'clickable map.' Besides depicting boundaries of speech communities and locating them in their geopolitical context, the map can portray isoglosses and other distributions. Of course, providing a facility for *making* such maps on the basis of the contents of a database is an advance programming task, and doing so in a user-friendly, interactive way is even more challenging.

## 7.4. A glimpse under the hood: internals of the sound law database

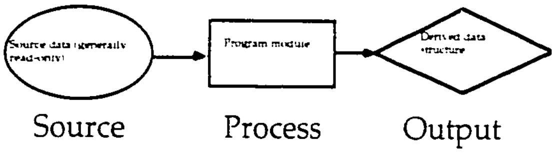
The software would have to provide a wide range of facilities. Embedded in such a software suite would be a number of subsystems, each of which has already been implemented in one program or another. These different subsystems would have to be integrated, that is, the output of one process would have to be usable as input to another. Software integration based on standardized representations is becoming more common; many programs can import or export a variety of types of data. The software would provide for handling complex lexicographic data structures; for handling text corpora and linking corpus entries to lexicographic entries; for carrying out phonological and morphological analysis such as retranscription, segmentation, parsing, and tagging; for managing simultaneously several types of research projects and hypotheses, each an incrementally-created set of relations and data over some fixed data set; for detecting and correcting errors and propagating the corrections to all affected subcorpora (with appropriate audit trails and quality assurance provisions). Of course, the software and data would have to be sharable, probably over networks such as the internet. It should support the publication of data and results in both printed and electronic forms, and so would have a variety of formats for data extraction and reporting.

A more complete diagram showing the additional components which would be required for a full-fledged research tool is shown in (240) below.

(240) *Data flow diagram of sound law database*


Etymological information from sources
Lexicographic resources (mainly Dictionaries)
Text Corpora
Bibliographic Citations
Dictionary making tools
Linked lexical and syntactic data structures
Corpus query and analysis tools / Interlinear Text handling tools
Bibliographic Utilities
Phonological parser and transcription tools
Morphological parser and segmentation tools
Meaning Resolver / Thesaurus
Language Database
Maps and other graphics
Phonological representations
Morphological Representations
Semantic Representations
GIS Cartographic and Dialectographic tools
Sound Correspondences
Reconstruction Engine Etymological analyzer
Language Atlas
Language Groupings, Isogloss mappings
Computer generated cognate sets
Statistical analysis tools
Printing, Exporting, Networking Facilities
Normal Database search and retrieval tools

**Legend:**



This design incorporates seven types of source data, eleven types of tools for processing that data, and seven types of derived or intermediate data structures. Note that while the diagram represents source and derived data as separate objects these entities may in fact share data files or other program entities. Outputs of one process of course can be inputs to another.

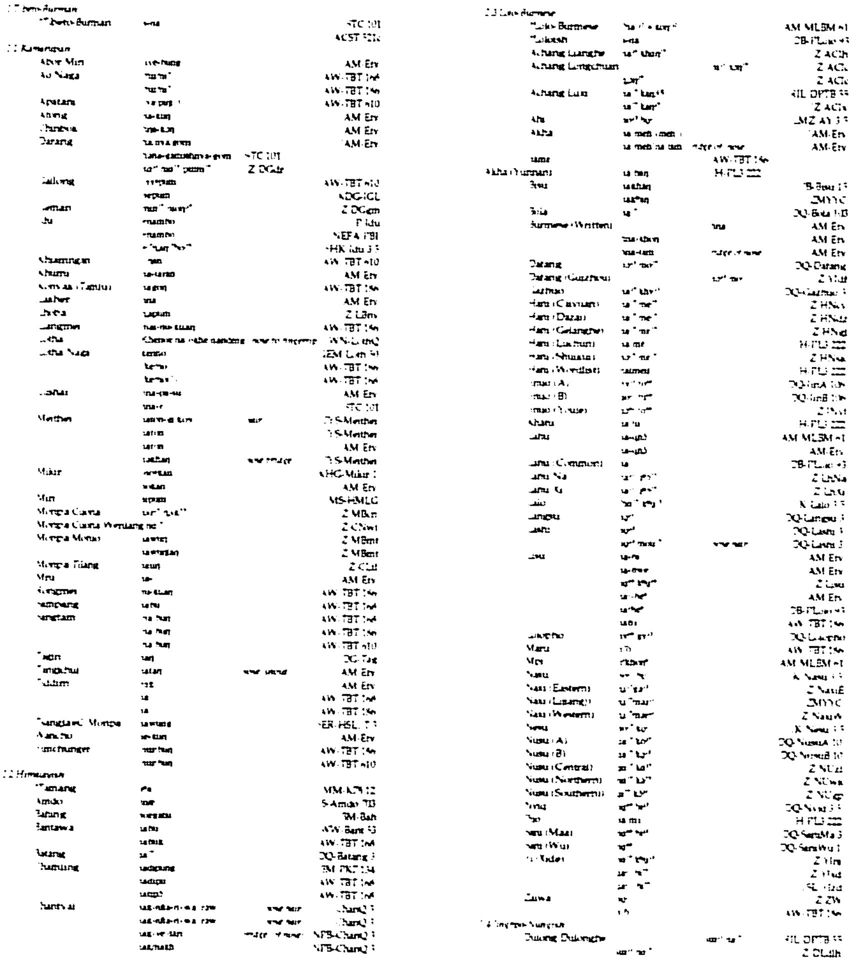
The specification of such a software tool, done in detail, would probably be nearly as long as this dissertation. It would be of interest to the programmers and documenters, but not to others. I will describe only the most general functionality and interface needs here, leaving the bulk of the internals for another publication. I will first outline the data structures, then describe the processes that use them.

### 7.4.1. Data Design

The data structures parallel the four basic elements of diachronic lexical research: form, meaning, source, and provenience (described in §3). These are the basic building blocks of a sound law database. Each of these elements has unique requirements in terms of database design. For example, language names in the a cross-linguistic database form a relative small closed set of items. Glosses, on the other hand, form a relatively large and open set. Some data elements (for example the list of names of language subgroups) have few enough elements that they can be implemented as menus and select lists; others have hundreds of thousands of values requiring different presentational techniques. The numbers may change in the course of research. For example, a cognate set may contain as few as two forms (there are some reasons why we might wish to keep 'cognate sets' with one form around (i.e. isolates), discussed in §6.5.1 and §1.5.6 above); or it may have hundreds (as shown in (241) below) or, in extreme cases, thousands of forms.

(241) *A 'Largish' Cognate Set (from the Sino-Tibetan Dictionary and Thesaurus (forthcoming); this version is not meant to be read, but reflected upon. It is reproduced in larger type in Appendix 6.*

3.00 *s-na *NOSE*



This cognate set contains 145 forms from 95 Tibeto-Burman languages drawn from four subgroups. In contrast, the equivalent entry in Pokorny 1969 (p. 775) cites 27 roots in 14 languages across 6 subgroups, while Buck's synonym dictionary cites only 31 forms (not all cognate), with 4 discussion notes (Buck 1949). Of course, we would expect that the synonym dictionary, based on semantic principles, would list more forms in a set like this that would the etymological dictionary, based on reconstructions.

Implementing a design with these sorts of requirements demands a development platform which is fairly robust and flexible. The requirement for 'free-form' data structures will exclude most conventional database programs from consideration as suitable development environments. As discussed above the structure of the source data as well as any projections made from the source data demand that the software accommodate several types of variation, including:

- variable length fields of unspecified length but in some cases very long;

- repeating fields, both indexed (ordered with respect to each other) and non-indexed;

- several data types beyond the usual (the usual being text, numeric, date, and so forth). Such data types would include 'word-indexed text' (i.e. a field in which the individual words are indexed and searchable, not just the initial strings); 'morphologically analyzed string' in which the components of underlying representations of surface forms is searchable.

- potentially huge numbers of records, in the thousands at least and probably in the hundreds of thousands or millions ultimately.

- a highly relational structure.

In the discussion below, the following conventions: names of program items, both procedures and data structures, will be represented in a distinct font, have no blanks, and be significantly capitalized (LikeSo).[134]

### 7.4.2. Source data

Data from source documents is generally read-only: once the data is correctly entered, there should be no reason to modify it. These data resemble the original document as much as possible given the constraints of the hardware and software. The specific files (in alphabetical order) called for are:

*(242) Source files for a sound law database*

| | |
|---|---|
| LanguageList | Data on languages used in Lexicon file |
| Lexicon | Lexical forms from various languages; essentially a set of merged dictionaries and other data sources |
| Maps | A catchall for information represented in geographical or spatial terms |
| SoundCorrs | Tables of correspondences used in reconstruction subsystem |
| SourceEtyma | Existing reconstructions from source documents, linked to the lexical data which support them |
| Sources | Bibliographic citations for data sources used in Lexicon, Etyma, and other files |
| SubgroupList | Multivariate groupings of languages based on derived and source information |
| TextCorpora | Full texts used to support lexicographic functions |

### 7.4.3. Derived data

It is likely that the data structures used to represent the relations between the source elements, their projections, and other machine-created elements will be just as complex and space-consuming as the source data.

Most of this structure will be submerged: the user will not see it directly, but only its results.

(243) *Program-internal files for a sound law database*

| | |
|---|---|
| CognateSets | Index to the Lexicon file, specifies which forms are cognate, based on data published sources as well as machine-created cognate sets. |
| DerivedEtyma | Computed reconstructions based on comparison tools |
| PhonoReps | Phonological representations |
| MorphoReps | Morphological representations |
| SemReps | Semantic representations |
| LanguageAtlas | Annotated versions of maps and graphic representations |
| LanguageGroups | Structure imposed on source references language names, used to distinguish and unifying differing conventions |

## 7.5. Conclusion

Whether such an effort is worth the substantial design and development time required is an open question. A number of important questions remain to be answered before an effort to create a general-purpose SLDB is started: Who would use such a program and for what? How developed and flexible would the formalism (and its implementation) have to be before it would be accepted by linguists, especially given the range of applications of the comparative method? How should the SLDB be interfaced to the lexicons of modern languages; that is, how much should the SLDB have to know about the morphology and internal reconstruction of the languages? What range of data should the program be able to accept as input? What is the best way to arrange the user interface, and on what computers should it be developed?

# 8. CONCLUSION

A striking fact about etymological research is its boundedness. There are a finite number of words in a finite number of languages which can shed light on the past; these are the words and languages that exist at this instant. Of course, recent work observing sound change in progress may make it possible to make better use of the evidence provided by these words; in fact, the detailed dialectological observations of Labov and others on dialect change now span generations of speakers (i.e. about 30 years) and have already provided many insights in language change, especially those driven by sociolinguistic forces. However, the time scale is still rather short, and we will have to wait several lifetimes before a complete picture based on direct observations is possible. Assuming that humanity survives a few more centuries, our descendants will be in the position of being able to study before-and-after spectrograms of pronunciation and usage at different periods; they will be able to avail themselves of tremendous corpora recording various aspects of language activity. Admittedly, their task will be greatly complicated by the extensive language contact in the world: the language isolate is virtually a thing of the past, and all evidence points to the eventual development of a 'climax community' of languages, in which a few hardy survivors have staked out their niches. This statement is merely an observation based on the rate of extinction of the world's languages.

I have already estimated (and very roughly so) the number of words which exist which might be useful in etymologies; the number is not that large. The number of different arrangements is much larger, but only one corresponds to the historical reality of their relationships. A word is either a descendant of an earlier form of the same language, or it has been borrowed from a neighbor. There are, of course, neologisms, and these must form a stratum of some sort at each point in time; their contribution to the whole lexicon is probably small.

We may expect that someday all the words that are still available will be recorded and entered in some datafile. For Sino-Tibetan, there are nearly a quarter of a million words entered now; the same (or nearly so) for Bantu. Databases containing smaller numbers of forms (in the 'mere' hundred thousand range) for Mon-Khmer, Austronesian, Nilo-Saharan, and so on are in the computers of various researchers; Tools for conversion are getting better and we may expect the rate of doubling (in size!) to increase rapidly. Of course, there is a distinct limit to all this. Soon the rate of 'capture' of linguistic data will meet the rate of extinction of languages globally. Unfortunately, it is likely that this meeting will occur mainly as a result of the rate of extinction overtaking the rate of capture: fieldwork takes much longer than database preparation and a sizable portion of the world's languages are moribund. Much of the data will not be captured before it becomes 'unavailable.' There is, of course, the substantial backlog of texts and dictionaries to be analyzed, converted, and added to the store, but it is a finite quantity and will never increase.

# NOTES

[1] 'Javanese [D] is a domal stop... The Tagalog form means "pain, smart." The Batak form ... is from the Dairi dialect.' (Bloomfield 1933:310). Bloomfield does not indicate what the italicization of segments means but we must assume that the difference is distinctive.

[2] 'The Tagalog form means "exudation"; in poetic use also "sap."' (Bloomfield 1933:310)

[3] N.B.: these sets could also just as well be used to illustrate vowel correspondences such as *a, *i and *u.

[4] This is true of the three examples I give, though clearly all the presenters have their reasons for not including them. Bloomfield's exposition is the most curious for its attention to but non-explication of these details.

[5] It would be nice to be able to use the term 'protophoneme' (following (Anttila 1989:279) for example. However, as will become clear in later discussion, there is no need to assume that the protoelement involved is necessarily a phoneme.

[6] *Exception*, to some linguists, also has the meaning *peripheral*; that is, in this view, exceptions may not have to be handled at all by the explanatory apparatus. For discussion see for example Inkelas 1993.

[7] Indeed, Anttila notes, deciding 'which part corresponds to what is an enormous operation logically.' (Anttila 1989:230)

[8] Of course, there is quite a range in the number of words available or needed, especially since some words are more useful than others for

the purpose of comparison. The idea that some languages have more words than others, especially the notion that 'primitive' languages have smaller lexicons than more 'developed' languages is preposterous: the size of a language s lexicon is primarily based on the willingness (and resources!) of researchers gathering the data.

[9] For convenience, the abbreviations are interpreted here: na (= Nasu), sa (= Sani), ahi (= Ahi), ak (= Akha), lh (= Lahu), wo (= Woni), bi (= Bisu), and lc (= Lüquan Lolo) are the name of Loloish languages. *LB is Proto-Lolo-Burmese, the name of their reconstructed ancestor language. This data set and the Loloish languages are discussed in more detail in §5.2 and §6.4.

The annotations given are other cognate sets which support the reconstruction of this set with an **-at* rhyme. Of course, it might be good to include a list of other sets which support the reconstruction of the *$^H$ tone or the **w* initial as well.

The first four columns of the correspondence row are the row number (#), the syllable constituent type (Type: R means Rhyme, as opposed to Initial or Tone), protoconstituent (in this case label *LB for Proto Lolo-Burmese), phonological environment (Env), here unspecified (i.e. unconditioned change); the rest of the columns identify outcomes in the Loloish languages.

[10] This morpheme in the Lahu form means 'tree;' presumably the $^H$sə- in the etymon is a reduction of the full morpheme *LB sik, 'tree, wood.'

[11] The problem was first noted by critics of glottochronology. Matisoff, for example, has provided a culturally-sensitive 'Swadesh-type' 200-word list specifically tailored to Southeast Asia.

[12] Though it appears that at least one computer scientist seems to think some version of this statement is not true! (see §4.1.2 about scanning Central Asian texts).

[13] The quotes and the essence of my remarks on Jones paraphrase views published most notably by Campbell and Poser 1992.

[14] The 'beautiful metaphor' of pruning trees into composing verses reminds one of the Biblical verse (Isaiah 2:4) about beating swords into plowshares and pruning hooks.

[15] I used the word sequentialization here as a convenient way of expressing the relationship between the two laws. Verner's Law can be seen in several different lights, for example as a condition on the application Grimm's Law or as as applying in sequence after Grimm's Law.

[16] I.e., Lamarckian evolution.

[17] As an aside, biologists have had an interest in linguistics corresponding to linguists' interest in biology. 'Evolutionists have always viewed linguistic change as a fertile field for meaningful analogies. Charles Darwin, advocating an evolutionary interpretation for such vestigial structures as the human appendix and the embryonic teeth of whalebone whales, wrote: "Rudimentary organs may be compared

with the letters in a word, still retained in the spelling but become useless in the pronunciation, but which serve as a clue in seeking its derivation." Both organisms and languages evolve.' (Gould 1980:27). Schleicher was himself devoted to the biological metaphor and Darwinian philosophy. See Percival 1987 for a fascinating discussion of these connections.

[18]The problem of cladistics (a method of hierarchical classification based on the observation that the distribution of characteristics among organisms can be perceived as exhibiting a hierarchical pattern(Crane 1987:140)) is quite topical in evolutionary biology these days. Phenotypical and genotypical classifications have come into conflict, and the Linnaean system of classification is rapidly becoming outmoded. See for example Hoenigswald and Wiener 1987 for a number of insightful essays on this subject from the point of view of both biology and linguistics.

[19] Kiparsky (Kiparsky 1988, Kiparsky 1993) seems to believe that the currently received theory, which he refers to as the 'neogrammarian/structuralist' approach, is a direct heir of the neogrammarian view in its most extreme form, which Kiparsky states as holding that '[...] sound changes originate through gradual articulatory shifts *which operate blindly without regard for the linguistic system*' [emphasis added], to which he opposes his own view that sound change is phonological rather than phonetic. This discussion is

very surprising to historical linguists, who have believed, for a good 75 years, that when they used the traditional expression 'phonetic change' (*changements phonétiques*) or 'sound change' it was understood as a technical term meaning precisely 'phonological change'. Perhaps the subtitle of such a non-recent work as Martinet's 1955 *Economie des changements phonétiques: traité de phonologie diachronique* would suffice to indicate that systemic pressure on language evolution is not a discovery of generative phonology. How else could one understand the ever-present insistence on the role of '*cas vides*' as triggers of change in the works of practitioners of the 'neogrammarian/structuralist' approach?

[20] Though as Malkiel notes, it makes quite a difference if it is 90% one way or 90% the other (Malkiel 1967).

[21] The comparative method certainly fit quite neatly into the intellectual program of American structural linguistics; this may explain the praise bordering on adoration heaped on the method by American linguists. I have not found similar adulation in the works of European linguists; perhaps they had no reason to remark on such a familar and accepted tool.

[22] In a footnote to this remark, Hoenigswald faults specifically Greenberg and Haas for misunderstanding the use of the comparative method in language classification. 'It is true that it is easier to prove relationship than subrelationship. But the criteria of subgrouping have constituted

one of the most frequently discussed topics in comparative linguistics from the middle of the last century to the present day ' Hoenigswald 1963:1)

[23] The controversies referred to here are dealt with in details below; they include for example the controversy about the mechanism of sound change treated in (but not resolved by) Labov's classic article *Resolving the Neogrammarian controversy* (Labov 1981)

[24] I refer here to recent assertions about hypotheses like the uniformity of the so-called molecular clock (Wilson and Cann 1992), the agricultural basis for the spread of Indo-European languages (Renfrew 1990), and so on.

[25] The term, due to Robert A. Heinlein in *Stranger in a Strange Land,* etymologically in Martian means *drink;* by extension 'to understand so thoroughly that the observer becomes part of the observed — to merge, to blend, intermarry, lose identity in group experience. It means almost everything we mean by religion, philosophy, and science — and it means as little to us as color means to a blind man.' (Heinlein 1961:213-214)

[26] It has been pointed out to me that even such a 'philological' language family as Semitic still has no comparative dictionary on the scale of Pokorny 1959! (Orin Gensler p.c. 1995)

[27] Statistics are given in the front matter of Part 1: Fascicle 1 (p.1).

[28] In the front matter '...two dozen, more or less,' languages are mentioned (vii), but 63 language names (29 Dravidian and 34 others) are listed therein as well, and many of these are cited in the cognate sets; started in 1949 and first published in 1961. The count is based on the size of the index: 90% of 3 columns x 66 items x 256 pages.

[29]Between announcement (1929) and publication (1949) Buck 1929

[30] '50,000 Wörter' cited in preface (v. 2, p. 8). The set count comes from (Ringe forthcoming:15) who cites (Bird 1982:11). This work is a revision of the previous work.

[31]These statistics are extrapolations based on a sample of pages: the Indo-European index contains 35 pages with about 61 reconstructions per page; the Register of modern forms contains about 66 forms per column, 3 columns per page, and comprises 211 pages.

[32]Estimates based on a sample; the number of languages is taken from the front matter; clearly some of the 'languages' given there merely identify the particular literary work from which the form is cited and so this number is somewhat, but only somewhat, inflated.

[33]NB: Three dialects of Lisu are included in this total; 192 numbered cognate sets are presented, but a large portion of these have several subsets given.

[34]Five major languages are cited, but occasional forms from approximately eighty other languages also appear.

[35] The STEDT project was started in 1987. The statistics here reflect the number of sets and forms analyzed as of May, 1995. Most of these reconstructions are words for body parts.

[36] For comparison the distribution of subgroups among cognate sets in the STEDT database (Matisoff forthcoming) is given below. Note that most of these sets are from a particular semantic area (body parts), and are drawn from a 232,000 word database containing data from 250 languages.

| #groups in set | # of Sets |
|---|---|
| 1 | 207 |
| 2 | 283 |
| 3 | 237 |
| 4 | 166 |
| 5 | 112 |
| 6 | 83 |
| 7 | 49 |
| 8 | 31 |
| 9 | 32 |
| 10 | 12 |
| 11 | 10 |

[37] Availability of the notebooks of course is the beginning of this process, and I certainly do not mean to impugn Prof. Greenberg's integrity in this matter. The undertaking of such data preparation and publication is a serious matter.

[38] From the STEDT database, Tag 3348. Most of the forms are compounds composed of a morpheme meaning ARM/HAND (< *TB lak) plus this etymon (perhaps *Burmish san). This cognate set and its significance was pointed out to me by Jim Matisoff.

[39] 'But what about the residual forms which remain, after we have taken into account the phonetic and morphological factors and the multilayered structure of the vocabulary? What explanation can we give for sounds changing differently *under completely comparable conditions?*' [emphasis added] (Wang 1969:10).

[40] Bloomfield's Algonquin: (Fox, Cree, Menomini, and Ojibwa) (cited from Haas 1969); Benedict's Tibeto-Burman (Garo, Lushei, Kachin=Jingpho, Written Tibetan, Written Burmese) (cited from Benedict 1972); Bopp's Indo-European (Greek, Latin, Sanskrit, Avestan, and Germanic) (cited from Bloomfield 1933:14); Dempwolff's Austronesian Dempwolff 1934-1938 (Indonesian, Javanese, Toba-Batak, and Tagalog) (cited from Haas 1969).

[41] This author has not yet been fortunate enough to have the opportunity to do 'real' field work, and so his assertions about the relative time commitments are based purely on his own conjectures.

[42] For example, there is a clearly implied expectation that reconstruction at a shallow time depth can and should be supported by a larger number of forms than those for a more distant relationship.

[43] If there are 5,000 languages in the world, and each language contributed 10,000 forms (the number in a medium-sized dictionary), this database would have 50 million entries. If we assume each entry (including gloss, part of speech, etc.) is 100 characters long (a rather minimal entry), this database would occupy at least 5 gigabytes, about 10 CD-ROMs worth.

[44] This last alternative is chosen more often than one might think. Einstein, when asked what he would do if Eddington's eclipse expedition to Brazil in 1929? yielded evidence that the theory of relativity was wrong, said, 'but the theory is so beautiful — it must be true!' (Clark 1984). 'Se non è vero, è ben trovato.'

[45] As Vennemann notes, the 'closed catalog' method of linguistic explanation has a rather long history, cf. for example Bredsdorff (1821) cited by H. Andersen (1975), Enkvist (1979).

[46] This is how the reconstruction of lost constituents is done by the Reconstruction Engine. The increase in computation is quite substantial.

[47] I believe that Kay means here that it might be possible to implement the method as he has described it directly given the computational power available at the time.

[48] Kay himself has spelled *daß* in two different ways in this article; his orthography is retained here.

[49] NP-hardness as it applies to this particular problem in reconstruction is given in Appendix 5.

[50] 'If the etymon is not identical to the Classical Latin (CL) form the latter is also listed.'

[51] 'Single letters shifts' have a special significance to programmers. For example, bit-shift instructions are commonly used in assembly language programs; 'right-shifting' the letters used to designate the evil HAL 9000 series computer in the cult classic *2001: a space odyssey* spell the initials of a rather well known computer company.

[52] The telegraphic code is a four-digit number assigned to each Chinese character to make it possible to send Chinese over telegraph lines.

[53] The description of the program given here is taken from an electronic manuscript provided to me by Dr. Veatch. No pagination is given.

[54] These are the names of two of the programs in the suite. I have tried to avoid mentioning the names of the programs as these details do not add to the description.

[55] Without going into too much detail, the difference in results is due to the fact that if tonal and segmental representations are conflated (e.g., by using accented vowels), the frequency distributions of the symbols will be different than for the tone number transcription; this in turn will give a different measure of relatedness.

[56] The term dialect may be misleading: these are apparently different languages.

[57] I should reiterate that Kay's project is really one of the earliest attempts to look inside forms and do real comparison, and that my criticism is intended to be quite mild: the average desktop computer today is perhaps two orders of magnitude more powerful that the machines Kay had to work with.

[58] This example is from Mazaudon's Proto-Tamang reconstruction (descibed in §5.1 and Appendix 4.1);

#504. $^{A}$dan [3.145.55]

| | | |
|---|---|---|
| tag | $^{3}$tan | *way of sitting* |
| pra | $^{3}$tē $^{1}$c$^{h}$atse $^{3}$mo: | *good day.* |

[59] In this context, I note that some people might find it a bit odd that Americans persist in calling the language they speak *English*, given the facts surrounding the unfriendly separation of the two groups at the end of the eighteenth century.

[60]
So, naturalists observe, a flea
Hath smaller fleas that on him prey;
And thse have smaller still to bite 'em;
And so proceed ad infinitum!

Swift 1733

[61] This definition is from Kernighan and Ritchie's classic work on the C programming language (Kernighan and Ritchie 1978:20). This definition of white space has some well-known tender spots: the treatment of punctuation, the English possessive and contraction *'s*, other contractions, hyphenated words, exceptional cases like *O'Malley*.

[62] Order may be significant in the case of digraphs; some computer implementations 'backspace' to place the diacritic over (or under) a character, other provide 'zero-width' characters which allow the following character to be placed underneath (or over) the diacritic.

[63] The characters given were generated using the 'STEDT' phonetic font created by Stephen P. Baron, updated by this author and Zev Handel. The exact combinations of graphemes used to compose a graph vary from system to system. Some systems may, for example, represent all graphs by one grapheme. In these cases, u, ü, and ú have nothing in common (as far as the computer is concerned).

[64] In passing, I note that not all 'root' characters in the Burmese script are the same width, so that different versions of these diacritical marks are required.

[65] Attempts to provide a universalist transcription system have met with mixed results. The IPA, for example, was to provide the basis for such a system of universal representations. However, it also explicitly allows for the symbols to be assigned according to the needs of the situation. Ladefoged and others have noted that it might be possible to arrive at an adequate representation of articulatory and acoustic features to support universal transcription. At the level of phonology this cannot be possible, because the symbols must reference just those distinctions which are significant in the language. Any symbols must therefore be accompanied by their (idiosyncratic) interpretation.

[66] Most of the preceeding was first pointed out to me by Larry Hyman.

[67] This phonological description contains the famous (within Tibeto-Burman) description of the vowel /ɤ/, transcribed by Fraser as /rgh/ as 'a plain guttural vowel sound, difficult to describe. Approximated in involuntary retching.'

[68] The provenience of the tones reflects the changes proposed in Matisoff 1979. The capital letters refer to the type of initial, which together with the type of rhyme condition the tonal outcomes of the three proto-tones, numbered *1, *2, and *3: P = *voiced, G = glottalized initial, V = prenasalized, -V =* voiceless, S = voiceless spirantal.

[69] Which is based on the Greek, which in turn is based on the Phoenician.

[70] The devanāgarī order is quite rational until it is rendered in roman letters, where, as is shown, sometimes two letters are required to represent one segment, or special versions of roman characters are required leading to graphic differences (Whitney, Monier-Williams, and most Sanskritists use the forms in parentheses, while MacDonell, to make work a bit easier for his publisher, used italics).

[71] This term is due to Matisoff; it means 'under Indic influence' and should be juxtaposed with its twin sister *Sinospheric*, ('under Chinese influence'). It designates the two of the major *Sprachbünde* in Asia.

[72] For example, Lushai sa- 'animal' in sa-va *'bird'*; sa-vom *'bear'*, sa-hya *'fish'*.

[73] For example, *s*-'causative marker' and *m*- 'stative marker.'

[74] Note that tone is included here only because the reconstruction of *TB as a toneless language is not completely uncontroversial.

[75] Though perhaps in this case, the /ny/ should be regarded as part of the $C_i$.

[76] A number of these processes as they apply to the interaction of intials are described in Matisoff (1979:24). For example, 'prefix-preemption,' the process by which a "powerful' prefix drives out the root initial consonant altogether,' is shown by *LB s-nit > Lahu šɨ̄.

[77] The liquid medials -'r- and -'l- are written in Coblin as barred-r and barred-l, without comment. The question mark indicates that he is uncertain about the reconstruction of tone in the ST protolanguage.

[78] (Matisoff 1979:11ff). Note that there is a slight discrepancy (in the representation of vowel length) between the canon as cited here and as given in (64) above.

[79] Lexicostatistic techniques based on n-gram frequency distributions of glyphs may be able do a bit of this, but it is beyond the scope of this dissertation to discuss them in detail.

[80] {Matisoff, 1979 #520} notes that it 'must be set up after velars, ... dentals, and possibly labials.' However, he has also noted that it occurs after palatal affricates as well (cf. ʔcwat 'pluck'). (Matisoff 1995 p.c.)

[81] For extended discussion see Matisoff 1978.

[82] New patterns of this sort continue to be discovered, e.g. alternations within a word-family between the open rhyme **-a** on the one hand and the diphthongal rhyme **-ay** on the other. See Matisoff 1985, 1989.

[83] It is actually not always easy to distinguish between inflection and derivation in TB. Some good candidates for the label 'inflection' are Kiranti-type agreement systems, Kuki-Chin independent vs. subordinate verb forms, and Tibetan verbal ablaut. But what about, e.g., simplex vs. causative verbs in Burmese? Are we to assume that the more regular and transparent a morphological process is, the less ancient it is? Or have the truly ancient processes become more transparent because of analogical leveling?

[84] 'Protoform stuffing' (Matisoff 1980) is the result of trying to accommodate all protovariants in a single form, resulting in phonologically implausible monstrosities such as *mrgsla.

[85] These are by no means to be interpreted as dialect groups. The lexical variation discussed here happens to cut across major dialect boundaries, as can be seen in the forms from the two closely related Eastern Tani languages Padam L (ta-me, a-me < PT *me) and Mising L (ta-mño < PT *mjo).

[86] The provenience of the data used here is described in detail in Appendix 4.1.

[87] e.g., Miri əlak 'hand', əla 'foot'; Tableng (Konyak) yak 'hand', ya 'foot'; etc.

[88] I am reminded of an apparently apocryphal etymology of the word *kangaroo*, which according to this etymology means *pointing finger* in an Aboriginal language: on reaching Australia, one of Captain Cook's crew pointed to a kangaroo and ask the language consultant 'what's that?,' to which the consultant replied, 'That's a pointing finger.'

[89] As Jespersen says: a language is a dialect with an army and a navy.

[90] Though of course the Japanese pronounce it *Sei-ka*. *Seika* and *Hsi-hsia* are 'allograms' in Matisoff's terminology.

[91] This term is due to Gérard Diffloth.

[92] An exonymic name for the Kok-borok dialects is *Tripuri*, which is of course a *loconym*.

[93] The 'shortlist' currently contains the names of about 200 TB languages.

[94] I am indebted to Boyd Michailovsky for pointing out to me the importance of maintaining this invariant link between the source of data and how it is referred to in documents generated from the database.

[95] Machine-readable, a word which has been and will be used heavily, is a familiar buzzword; it has no technical sense in computer science (just as the word *word* has no technical sense in linguistics).

[96] Of course, if this were merely a simple substitution cipher it would be relatively easy to decode; however, there are both many-to-one and one-to-many substitutions here, as well as just plain garbling.

[97] See ST Stammbaum in Appendix 1.

[98] The Text Encoding Initiative is a collaboration of a number of national institutions to provide standards for encoding documents, tools to use them, and the documents themselves.

[99] Program written in MaxSpitbol (a computer language with extensive text-handling features) by Zev Handel at STEDT. I am indebted to Mr. Handel for his help with this example.

[100] Lexware was developed by Robert Hsu of the University of Hawaii. (Hsu 1970s) It appears to antedate SGML by a few years.

[101] Statistics on the number of example sentences in the dictionary because the operation of the parsing program, a 'one-shot', is suspect. The statistics are provisional and may be revised as the analysis of this machine-readable version progresses.

[102] This is not exactly a *compound* in the conventional sense as it is used in English, since there are still spaces between the words.

[103] Matisoff (1978:198, footnote 247), for example, gives this analysis of ANKLE, and notes that the same kind of composition is apparently attested in Quiché Maya.

[104] I use italics to indicate how a form is alphabetized within the different sources. Thus, *kulima* 'to cultivate' indicates that the form is alphabetized under the infinitive prefix *ku-*, while ku*lima* indicates that it is alphabetized under the stem *lima*.

[105] This procedure has been useful in making the cognate connection in the most obvious cases.

[106] This assumption will produce several wrong results; it is not the best such algorithm, merely one of the simplest. A more sophisticated one would be able to compensate for the *gain* of features, such as occurs in epenthesis.

[107] Braces and slashes are used in an informal sense, to distinguish the graphemes found in the forms as transcribed in the database, and the result as 'normalized' for comparison.

[108]The data cited here is from questionnaires submitted to STEDT.

[109] Soundex is a language-specific algorithm for comparing strings for phonetic similarity. The soundex system experimented with here is the one available in the FoxBase microcomputer database management system and is rather primitive: for example, the soundex requires that at least the first letters of the compared strings be the same in order to match. It is also an English language soundex. A more appropriate version would use a soundex function suited to Tibeto-Burman.

[‡] *BA* is a rather ubiquitous verbal suffix in Tibeto-Burman. Rather than attempt to have the computer distinguish the *BA* suffix from other *BA*, word-final BA is treated specially.

[110]The data cited here is from questionnaires submitted to STEDT.

[111]As an aside, the tonal correspondences 6 and 7 here show evidence of so-called 'tonal flip-flop,' also evidenced and discussed in (Matisoff 1972).

[112] The word *modern* here is meant to contrast with *reconstructed;* that is, the languages treated need not be modern, but merely attested, or perhaps even intermediate common languages. At any rate the idea to to distinguish the (usually) *attested* forms supplied as input to the program from the *unattested* forms (reconstructions) which it generates.

[113] In fact, the situation is slightly more complicated than is shown here: there are two other possible reconstructions and another possible cognate set which are not shown because of space considerations. This example is discussed in more detail in §6.5.1 below.

[114] Upstream in the sense of time. Originally the temporal directions of the program were described as backward and forward. The opposition of upstream and downstream, was suggested by Professor John Hewson, one of the developers of the first 'Electronic Neogrammarian,' Hewson 1973 is much more intuitive.

[115] The term REFLEX will be reserved for describing a complete modern form which is the regular descendant of some protoform. OUTCOME will be used for the regular descendent of a protoconstituent.

[116] The languages are discussed in Appendix 4.1.

[117] The use of *initial* in two senses here (specifying either a class of constituent, or a syllable position) is unfortunate, but seems to be unavoidable; it is tolerated in most discussions of this type and is clear from context.

[118] Another switch in the program determines whether to allow matching on substrings. This is significant inasmuch as some substrings might not be stated as constituents in the table, which would permit context conditions to be met by a constituent whose neighbors (immediate constituents) are not reconstructable via the table. The selection of this row would block the selection of other correspondence rows, perhaps producing undesired results.

[119] There is a great deal more to say about specificity and the complexity of the environmental constraints, so much so that a separate and rather lengthy discussion of it is merited. As currently implemented in the Reconstruction Engine, context must be stated in terms of immediately adjacent constituents (remote context cannot be used). Also, the context must be stated in terms of constituents (i.e. *atoms*), or lists of constituents: regular expressions and other possible definitions are not supported. Specificity is measured in a straightforward way: correspondence rules with no context have low specificity (specificity = 0). Rules with a 'one-sided' context (X_ or _Y) have specificity 1. Rules with a contextualizing element on both sides (X_Y) have specificity 2. Only integer specificities are supported. Cover symbols (such as '$C_{VL}$' for voiceless consonants in correspondence 1 in (165) above) and lists of constituents (e.g. p,b,m_) are supported but do not alter the specificity.

[120] The cover symbol $^{H}$ is used to permit the upstream reconstruction of Tukche forms in which the tone of the modern form is not precisely known. In the downstream direction, however, it licenses the generation of two possible reflexes.

[121] While the Taglung form itself is sufficient to determine the 'proper' reconstruction in this case, and if the Syang form were not available, the Taglung form would break the tie between the other competing reconstructions ($^{B}$glin. $^{B}$giŋ, and $^{B}$gin), it is usually difficult to pick out such decisive lexical items from a list of words.

[122] In essence, tone 2 in Tukche (< *TGTM B before ) does not occur with voiced initials in native words.

[123] The X which occurs in the Taglung form is a cover symbol meaning 'unspecified tone') and is used when the tone of the form is unknown. This allows the Reconstruction Engine to reconstruct the form under any of the tones. If this cover symbol were left out, the Reconstruction Engine would reconstruct this form without a prototone (permitted by the canon), and the form would fail to form a set with other forms which do have the tone specified.

[124] As discussed in §5.1.7 above, such a distinction may or may not reflect a cross-linguistic or even subgroup-wide truth. So, for example, by a chain of semantic shifts, the term *mother* is used to describe a moldy by-product of the process of making wine and vinegar. (Matisoff

1991:298) It is a small step to imagine that in other language families the distance between *wife* and *beer-mash* might not be great.

[125]It might be possible to apply the results of some recent research in the area, for example Wordnet (Miller 1990), to part of the problem. Indeed, the 'exclusion lists' developed for the Reconstruction Engine are similar structurally and conceptually to the 'synsets' of Wordnet. Ultimately, any solution would have to be sensitive not only to synchronic relationships in a single language (like Wordnet) but also to semantic shifts (both universal and language-specific) and the possibility of several different glossing metalanguages (in this case both French and English are used).

[126]In cases where a set is eliminated as a result of becoming a subset of another set, the reconstructions of the set being eliminated may have to be merged into the larger set.

[127]Using exclusion lists such as those defined in (7), for example, creates precise cognate sets which are composed only of reflexes which are assured to be semantically compatible (though some likely candidates might be eliminated when the exclusion list is incomplete). Using exclusion lists such as those defined in (8) would remove semantically incompatible reflexes, but leave those which for which semantic compatibility is unspecified.

[128]Allofamy, the relationship between words in a word-family, is described in more detail in §3.2.1.

[129] For example English *have* and Latin *habēre.*

[130] These sets have been simplified somewhat from those actually generated: I have eliminated some of the 'competing' reconstructions which are still permitted by the table at this stages. I have also conflated two overlapping sets, indicated by the allofam alternation between the possible reconstructions.

[131] The correspondences shown here are for purposes of exemplification only! While the correspondences in the lower level tables (for *NL, *SL, *CL) are at least tentatively established, the correspondences at the mesolevels are completely unverified (and so are shown as "identify correspondences" of the sort *k > *k.

[132] The STEDT database, for example, contains data on several hundred languages and dialects, of which perhaps twenty-five or thirty might be regarded as 'major languages'. Currently there are at least nine subgroups proposed for Tibeto-Burman.

[133] Compare, say, PIE > Gmc > WGmc > OE > Modern English with *PST > *TB > *LB > *L > *NL > Akha.

[134] This is becoming the *de facto* standard for naming variables in computer programs, especially in the more 'object-oriented languages' such as C++, contrasted with the now unfashionable 'underscore' approach (e.g., Language_List). (Winder 1991:18)

# BIBLIOGRAPHY

Adelaar, Willem F. H. 1989. Review of Language in the Americas. *Lingua* 78.249-255.

Andersen, Henning. 1974. Towards a typology of change: bifurcating changes and binary relations. Proceedings of the first international conference on historical linguistics II, Edinburgh 2-7 September 1973 J.M Anderson and C. Jones (eds.) 17-60. Amsterdam: North-Holland Publishing Company.

Andersen, Henning. 1990. The Structure of Drift. Historical Linguistics 1987: Papers from the 8th International Conference on Historical Linguistics, Lille 66. Henning Andersen and Konrad Koerner (eds.) 1-20. Amsterdam: John Benjamins.

Andersen, Henning, and E. F. K. Koerner. 1987. *Historical Linguistics, 1987: papers from the 8th international conference on historical linguistics.* International Conference on Historical Linguistics Lille: John Benjamins.

Anderson, J.M, and C. Jones. 1973. Historical linguistics II : Theory and description in phonology. First international conference on historical linguistics. S.C Dik and J.G Kooij (eds.). North-Holland Publishing Company.

Anonymous. 1962. Hànyǔ Fāngyán Zìhuì. Peking: Peking University.

Anttila, Raimo. 1989. Historical and comparative ling·:istics. 2nd ed. Amsterdam studies in the theory and history of lir.guistic science 6. Amsterdam: John Benjamins.

Anttila, Raimo. 1995. Historical explanation and historical linguistics. Explanation in historical linguistics 84. Garry W. Davis and Gregory K. Iverson (eds.) 17-40. Amsterdam: John Benjamins.

Apple Computer, Inc. 1991. *Inside Macintosh.* Apple technical library VI. Reading, MA: Addison-Wesley.

Baldi, Philip ed. 1990. *Linguistic change and reconstrucion methodology.* Trends in linguistics: studies and monographs. New York: Mouton de Gruyter.

Baron, Stephen P. 1990. Glottal retention and the transphonologization of stopped rimes in Loloish and Chinese. *paper presented at the International Conference on Sino-Tibetan Languages and Linguistics 23*

Baxter, William H. 1992. *A handbook of Old Chinese phonology.* Trends in linguistics. studies and monographs 34. Berlin: Mouton de Gruyter.

Becker, D.A. 1982. Teaching Lautgesetze in the computer age. *Yearbook of the seminar for Germanic philology* 5.8-37.

Benedict, Paul K. 1942. Thai, Kadai, and Indonesian: a new alignment in Southeastern Asia. *American Anthropologist* 44.576-601.

Benedict, Paul K. 1943. Studies in Thai kinship terminology. *Journal of the American Oriental Society* 63.168-75.

Benedict, Paul K. 1972. *Sino-Tibetan: a Conspectus.* Cambridge: Cambridge University Press.

Benedict, Paul K. 1973. Tibeto-Burman tones, with a note on teleoreconstruction. *Acta Orientalia* 35.127-38.

Benedict, Paul K. 1975. *Austro-Thai Language and Culture, with a glossary of roots.* New Haven: Human Relations Area Files Press.

Benedict, Paul K. 1975. A note on Proto-Burmese-Lolo prefixation. Linguistics of the Tibeto-Burman Area2.289-91.

Benedict, Paul K. 1976. Rhyming dictionary of Written Burmese. *Linguistics of the Tibeto-Burman Area*3.1.

Benedict, Paul K. 1976. Sino-Tibetan: another look. *Journal of the American Oriental Society* 96.167-97.

Benveniste, Emile. 1971. *Problems in general linguistics.* Miami linguistics series 8. Trans. Mary Elizabeth Meek. Coral Gables, Fla.: University of Miami Press.

Berkowitz, Luci. 1992. *Thesaurus Linguae Graecae. CD ROM #D.* Irvine, CA: Thesaurus Linguae Graecae Project.

Bickerton, Derek. 1984. The language bioprogram hypothesis. *Behavioral and Brain Sciences* 7.173-221.

Bird, Norman. 1982. *The distribution of Indo-European root morphemes*. Wiesbaden: Otto Harrassowitz.

Black, P. 1976. Multidimensional scaling applied to linguistic relationships. Lexicostatistics in genetic linguistics II 3.5-6. I. Dyen and G. Jucquois (eds.) 43-92. Montreal: Cahiers de l'Institut de Linguistique de Louvain.

Bloomfield, Leonard. 1933. *Language*. New York: H. Holt and Company.

Bloomfield, Leonard. 1966. A set of postulates for the science of language. Readings in linguistics I Martin Joos (ed.) 26-31. Chicago, London: University of Chicago Press.

Bradley, David. 1975. Nahsi and Proto-Burmese-Lolo. *Linguistics of the Tibeto-Burman Area*2.93-150.

Bradley, David. 1979. *Proto-Loloish*. Scandinavian Institute of Asian Studies Monograph Series #39. Copenhagen/London.

Bradley, David. 1994. *A Dictionary of the Northern Dialect of Lisu (China and Southeast Asia)*. Pacific Linguistic: Series C 126. Canberra, Australia: Dept. of Linguistics, Research School of Pacific and Asian Studies, Australian National University.

Brandon, Frank R. 1984. A phonological rule interpreter for microcomputers. Computers in literary and lingustics computing 1. J Hamesse and A Zampolli (eds.) 47-62. Louvain-la-Neuve: Université Catholique de Louvain.

Buck, Carl Darling. 1929. Announcement of IE Synonyms. *Language* 5.215-227.

Buck, Carl Darling. 1949. *A Dictionary of Selected Synonyms in the Principal Indo-European Languages: a contribution to the history of ideas*. Chicago: University of Chicago Press.

Burling, Robbins. 1959. Proto-Bodo. *Language* 35.433-453.

Burling, Robbins. 1967. Proto-Lolo-Burmese. *International Journal of American linguistics* 33.101.

Burrow, Thomas, and Murray B. Emeneau. 1961. *A Dravidian etymological dictionary*. Oxford: Clarendon Press.

Burrow, Thomas, and Murray B. Emeneau. 1984. *A Dravidian etymological dictionary*. 2nd New York: Clarendon Press.

Burton-Hunter, Sarah K. 1976. Romance Etymology: a computerized model. *Computers and the humanities* 10.217-220.

Campbell, Lyle. 1988. Review of Language in the Americas. Language 64.591-615.

Campbeli, Lyle. 1992. Inside the American language classification debate. ms.

Campbell, Lyle, and William Poser. 1992. Indo-European Practice and Historical Methodology. *Proceedings of the Berkeley Linguistics Society* 18

Cannon, Garland. 1990. *The Life and Mind of Oriental Jones: Sir William Jones, the Father of Modern Linguistics.* Cambridge: Cambridge University Press.

Cannon, Garland. 1991. Jones's "Sprung from some common source": 1786-1986. Sprung from some common source: Investigations into the prehistory of languages Sydney Lamb and e. Douglas Mitchell (eds.) 23-50. Stanford: Stanford University Press.

Cavalli-Sforza, Luigi Luca, and et al. 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences* 85.6002-06.

Cavalli-Sforza, Luigi Luca. 1991. Genes, peoples and languages. *Scientific American* 265.104-110.

Chao, Yuen-Ren. 1966. The non-uniqueness of phonemic solutions of phonetic systems. Readings in Linguistics 1. Martin Joos (ed.) 38-54. Chicago, London: University of Chicago Press.

Charniak, Eugene, and Drew McDermott. 1985. *Artificial Intelligence.* Reading, Mass.: Addison-Wesley.

Chen, Kang. 1986a. *A Comparative Study of the Vowels among the Dialects of the Yi Languages.* Paper presented at the 2nd International Workshop on the Lolo (Yi) Languages, Lund, Sweden. ms. [published in Chinese as "Study of the rhyme correspondence in Yi" *Yuyan Yanjiu* 1987:2(1987)]

Chen, Kang. 1986b. *The Proto-Yi Tone System (summary).* Lecture delivered in Lund, Sweden, 1986. ms. [published in Chinese as "Tonal correspondences in Yi" *Minzu Yuwen* 1986:5]

Cheng, Chin-Chuan. 1993. DOC: Its birth and life. Linguistic Essays in honor of William S-Y Wang Matthew Chen and Ovid Tzeng (ed.) (forthcoming).

Cipra, Barry. 1995. Enormous theorem eclipses Fermat. *Science* 267.794-797.

Clark, Ronald William. 1984. *Einstein : the life and times : an illustrated biography.* New York: H.N. Abrams.

Coblin, W.S. 1986. *A Sinologist's Handlist of Sino-Tibetan Lexical Comparisons.* Monumenta Serica Monograph Series 18. Nettetal: Steyler Verlag.

Collinge, N. E. 1985. *Laws of Indo-European.* Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory v. 35 Amsterdam: J. Benjamins.

*Collins English dictionary.* 1975. Glasgow: Harper Collins.

Comrie, Bernard. 1981. *Language universals and linguistic typology.* Chicago: University of Chicago Press.

Consortium, Unicode. 1992. The Unicode standard. 1-2

Crane, Peter R. and Christopher R. Hill. 1987. Cladistic and Paleobotanical Approaches to Plant Phylogery. Biological

Metaphor and Cladistic Classification: an Interdisciplinary Perspective Hoenigswald (ed.) 139-155. Philadelphia: University of Pennsylvania Press.

Dai Qingxia, et al. 1981. *Chinese-Jingpho Dictionary.* Kunming: Yunnan People's Press (In Chinese).

Dai Qingxia, et al. 1992. *Zang-Mian yuzu yuyin cihui [A Tibeto-Burman lexicon].* Beijing: Central Institute of Minorities Press.

Davies, Anna Morpurgo. 1987. "Organic" and "organism" in Franz Bopp. Biological metaphor and cladistic classification Henry M Hoenigswald and Linda F. Wiener (eds.) Philadelphia: University of Pennsylvania Press.

Dempwolff, Otto. 1934-1938. *Vergleichende Lautlehre des austronesischen Wortschatzes.* Zeitschrift für Eingeborenen-Sprachen 15, 17, 19. Berlin: D. Reimer.

Diamond, Jared. 1990. The talk of the Americas. *Nature* 344.589-90.

Driem, George van, and Ksenija Borisovna Kepping. 1991. The Tibetan transcriptions of Tangut (Hsi-hsia) ideograms. *Linguistics of the Tibeto-Burman Area.*117-128.

Durham, Stanton P, and David Ellis Rogers. 1971. An applicaton of computer programming to the reconstruction of a proto-language. *ITL* 5.70-81.

Dyen, Isadore. 1969. Reconstruction, the comparative method, and the protolanguage uniformity assumption. *Language* 43.150-171.

Dyen, Isidore. 1970. Background 'noise' or 'evidence in comparative linguistics: the case of the Austronesian-Indo-European hypothesis. Indo-European and Indo-Europeans George Cardona, Henry M Hoenigswald and Alfred Senn (eds.) 1-10. Philadelphia: University of Pennsylvania Press.

Dyen, Isadore. 1973. The impact of lexicostatistics on comparative linguistics. Lexicostatistics in genetic linguistics I Dyen (ed.) 11-29. The Hague: Mouton.

Dyen, Isidore. 1975. *Linguistic subgroupings and lexicostatistics.* The Hague: Mouton.

Dyen, Isadore. 1992. An Indo-European classification: a lexicostatistical experiment. *Transactions of the American Philosophical Society* 82.1-132.

Dzokanga, Adolphe. 1979. *Dictionnaire lingala-français: suivi d'une grammaire lingala.* Leipzig: VEB Verlag Enzyklopädie.

Eastlack, Charles L. 1977. Iberochange: a program to simulate systematic sound change in Ibero-Romance. *Computers and the humanities* 11.81-88.

Egerod, Søren. 1982. How not to split tones: the Chaozhou case. *Fangyan* 3.169-173.

Ellison, T. Mark. 1994. Machine learning of phonological structure. Ph. D. Dissertation. University of Western Australia.

Fodor, István. 1965. *The rate of linguistic change.* London: Mouton & Co.

Fox, Anthony. 1995. *Linguistic Reconstruction.* Oxford: Oxford University Press.

Frantz, Donald G. 1970. A PL/1 program to assist the comparative linguist. *Communications of the ACM* 13.353-356.

Fraser, J.O. 1922. *Handbook of the Lisu (Yawyin) Language.* Rangoon, Burma: Superintendent, Govt. print.

Gamkrelidze, Thomas. 1989. Language typology and Indo-European reconstruction. The new sound of Indo-European Theo Vennemann (ed.) 117-21. Berlin: Mouton de Gruyter.

Garrett, Andrew. 1991. Indo-European reconstruction and historical methodologies. *Language* 67.790-803.

Glover, Warren W. 1974. *Sememic and Grammatical Structures in Gurung (Nepal).* SIL Publications Kathmandu: Tribhuvan University Press.

Gould, Stephen Jay. 1980. *The Panda's Thumb: More Reflections in Natural History.* New York: W.W. Norton & Company.

Gragg, Gene. 1994. CUSHLEX. (computer program).

Greenberg, Joseph H. 1949. Studies in African Linguistic Classification: I. The Niger-Congo Family. Southwestern Journal of Anthropology 5. 79-100.

Greenberg, Joseph H. 1957. Genetic Relationship Among Languages. Essays in Linguistics 35-45. Chicago: University of Chicago Press.

Greenberg, Joseph H. 1960. The general classification of Central and South American languages. Men and Cultures: Selected Papers of the 5th International Congress of Anthropological and Ethnological Sciences Anthony F. C. Wallace (ed.) 791-94. Philadelphia: University of Pennsylvania Press.

Greenberg, Joseph H. 1966. *Language universals.* The Hague, Paris: Mouton & Co.

Greenberg, Joseph H. ed. 1978. *Universals of Human Language Volume I: Method & Theory.* Stanford: Stanford University Press.

Greenberg, Joseph H., Turner, Christy G., II, & Steven L. Zegura. 1986. The settlement of the Americas: A comparison of the linguistic, dental, and genetic evidence. Current Anthropology 27.477-497.

Greenberg, Joseph H., and respondents. 1987. A CA book review: Language in the Americas. *Current Anthropology* 27.477-97.

Greenberg, Joseph H. 1987. *Language in the Americas.* Stanford: Stanford University Press.

Greenberg, Joseph H. 1987. *Language in the Americas.* Stanford: Stanford University Press.

Greenberg, Joseph H. 1990. The American Indian language controversy. *The Review of Archaeology* 11.5-14.

Greenberg, Joseph H. 1990a. Indo-European Practice and American Indianist Theory in Linguistic Classification. paper presented at the Boulder conference

Greenberg, Joseph H. 1990b. Correction. *Language* 66.3.660

Greenberg, Joseph H. 1991. Some problems of Indo-European in historical perspective. Sprung from Some Common Source Sidney M. Lamb & E. Douglas Mitchell (ed.) 125-140. Stanford: Stanford University Press.

Greenberg, Joseph H. 1995. Observations concerning Ringe's 'Calculating the factor of chance in language comparison'. *Proceedings of the American Philosophical Society* 137.79-90.

Grierson, Sir George A., ed. 1903-08 [reprinted 1967]. *Linguistic Survey of India.* III, Pts.1-3: Tibeto-Burman Family. Delhi, Varanasi, Patna: Motilal Banarsidass.

Griswold, R. E, J. F Poage, and I. P. Polonsky. 1971. *The SNOBOL4 programming language.* Englewood Cliffs, NJ: Prentice-Hall.

Grüssner, Karl-Heinz. 1978. *Arleng Alam: die Sprache der Mikir.* Wiesbaden: Franz Steiner Verlag.

Guthrie, Malcolm. 1967. *Comparative Bantu: An Introduction to the Comparative Linguistics and Prehistory of the Bantu Languages.* 1. London: Greggs.

Guy, Jaques. 1992. Notes on COGNATE. x.xxx ed. Linguist List.

Haas, Mary R. 1969. *The Prehistory of Languages.* Studia memoriae Nicolai van Wijk dedicata: series minor 57. The Hague, Paris: Mouton.

Hagège, Claude, and André-Georges Haudricourt. 1978. *La phonologie panchronique : comment les sons changent dans les langues.* Le Linguiste 20. Paris: Presses Universitaires de France.

Handel, Zev. forthcoming. Review of Bradley's Northern Lisu Dictionary. To appear in *Linguistics of the Tibeto-Burman Area.*

Handel, Zev, and John B. Lowe. 1995. Analysis of Matisoff's Dictionary of Lahu. *Linguistics of the Tibeto-Burman Area.*

Hansson, I.-L. (in prep.). *Akha-English Dictionary.*

Hansson, Inga-Lill. 1982. A phonological comparison of Akha and Hani. *Linguistics of the Tibeto-Burman Area* 7.63-115.

Hansson, I.-L. 1989. A comparison of Akha, Hani, Khatu and Pijo. *Linguistics of the Tibeto-Burman Area* 12.6-91.

Hartman, Steven Lee. 1981. A universal alphabet for experiments in comparative phonology. *Computers and the Humanities* 15.75-82.

Hartman, Steven Lee. 1993. Writing rules for a computer model of sound change. *Southern Illinois Working Papers in Linguistics and Language Teaching* 2.31-39.

Hartman, Steven Lee. 1994. PHONO. v3.1 (computer program).

Heinlein, Robert A. 1961. *Stranger in a strange land.* Ace. Berkeley: Berkeley Publishing Co.

Hewson, John. 1973. Reconstructing prehistoric languages on the computer: the triumph of the electronic neogrammarian. Proceedings of the international conference on computational linguistics. eds.) A. Zampolli and N. Calzolari. Linguistica. Pisa: Leo S. Olschki.

Hewson, John. 1974. Comparative reconstruction on the computer. Proceedings of the first international conference on historical linguistics. eds.) J.M. Anderson and C. Jones. North Holland Linguistic Series 12a. Edinburgh 2-7 September 1973: North-Holland Publishing Company.

Hewson, John. 1989. Computer-Aided Research in Comparative and Historical Linguistics. Computational Linguistics István S Bátori, Winfried Lenders and Wolfgang Putschke (eds.) 576-580. Berlin, New-York: Walter de Gruyter.

Hewson, John. 1993. *A Computer-Generated Dictionary of Proto-Algonquian.* Mercury Series Canadian Museum of Civilization.

Hock, Hans Heinrich. 1986. *Principles of historical linguistics.* Trends in linguistics : Studies and Monographs 34. New York: Mouton de Gruyter.

Hockett, Charles F. 1955. *A manual of phonology*. International Journal of American Linguistics 21. Memoir 11. Baltimore: Waverly Press.

Hockett, Charles F. 1967. *Language, mathmatics, and linguistics*. The Hague,Paris: Mouton & Co.

Hoenigswald, Henry. 1963. On the history of the comparative method. *Anthropological Linguistics* 5:1-11.

Hoenigswald, Henry. 1990. Is the comparative method general or family-specific? Linguistic change and reconstruction methodology 45. Baldi (ed.) Berlin: Mouton de Gruyter.

Hoenigswald, Henry M. 1960. *Language change and linguistic reconstruction*. Chicago: University of Chicago Press.

Hoenigswald, Henry M. 1966. Criteria for the subgrouping of languages. Ancient Indo-European Dialects: Proceedings of the conference on Indo-European linguistics Henrik Birnbaum and Jaan Puhvel (eds.) 1-12. Berkeley Los Angeles: University of California Press.

Hoenigswald, Henry M. 1966. The principal step in comparative grammar. Readings in Linguistics 1. 298-302. Martin Joos (ed.) Chicago, London: University of Chicago Press.

Hoenigswald, Henry M. 1966. Sound change and linguistic structure. Readings in linguistics I. 139-141. Martin Joos (ed.) Chicago, London: University of Chicago Press.

Hoenigswald, H.M. 1974. Internal reconstruction and context. Proceedings of the first international conference on historical linguistics. eds.) J.M. Anderson and C. Jones. North-Holland Linguistic Series 12a. Edinburgh 2-7 September 1973: North-Holland Publishing Company.

Hoenigswald, Henry M. 1990. Descent, perfection and the comparative method since Leibniz. Leibniz, Humboldt, and the Origins of Comparativism. 119-132. Tullio de Mauro and Lia Formigari (ed.). Amsterdam: John Benjamins.

Hoenigswald, Henry M, and Linda F. Wiener eds. 1987. *Biological metaphor and cladistic classification: an interdisciplinary perspective.* Philadelphia: University of Pennsylvania Press.

Holland, Gary B. 1992. Syntactic reconstruction. International encyclopedia of linguistics 4. William Bright (ed.) 115-117. New York, Oxford: Oxford University Press.

Hope, Edward R. (in prep.). *Lisu-English Dictionary.*

Hope, Edward R. 1974. The deep syntax of Lisu sentences : a transformational case grammar. Canberra: Dept. of Linguistics.

Hsu, Robert. 1970s. *Lexware manual, draft ms.* Honolulu: University of Hawaii Computer Center.

Hyman, Larry. 1994. *CBOLD phonological representation.* ms.

Illič-Svityč, V.M. 1971. *Opyt sravnenija nostraticeskix jazykov: Vvedenie; sravnitel'nyj slovar' (b-K).* Moscow: "Nauka".

Inkelas, Sharon. 1993. Inalterability as prespecification. *Language* 69.529-573.

Jacquot, A. 1982. *Lexique laadi (koongo).* Oralité-documents 3. Paris: Société d'études linguistiques et anthropologiques de France: Office de la recherche scientifique et technique outre-mer.

Jäschke, H. A. 1881. *A Tibetan-English Dictionary.* London: Routledge and Kegan Paul.

Jespersen. 1924. *Philosophy of Grammar.* New York: H. Holt and company.

Jones, Robert B. 1961. *Karen Linguistic Studies: description, comparison, and texts.* University of California Publications in Linguistics 25. Berkeley and Los Angeles: University of California Press.

Jones, Sir William. 1798. Third Anniversary Discourse: On the Hindus. Asiatick Researches 1. 415-431.

Jones, Sir William. 1799a. Sixth Anniversary Discourse: On the Persians. Asiatick Researches 2. 43-66.

Jones, Sir William. 1799c. Eighth Anniversary Discourse: On the Borderers, Mountaineers, and Islanders of Asia. Asiatick Researches 3. 1-20.

Judson, A. 1921, 1986. *Burmese-English Dictionary.* Centeanry Edition, 2nd Printing Revised and enlarged by R. C. Stevenson. Rangoon, Baptist Board of Publications.

Kaplan, Ronald M. and Martin Kay. September 1994. Regular models of phonological rule systems. *Computational Linguistics* 20.331-379.

Karapurkar, Pushpa. 1972. *Tripuri Phonetic Reader.* CIIL Phonetic Reader Series 5. Mysore: Central Institute of Indian Languages.

Kay, Martin. 1964. The logic of cognate recognition in historical linguistics. The Rand Corporation, Santa Monica, CA.

Kay, Martin. 1966. xxx.

Kernighan, Brian W., and Dennis M. Ritchie. 1978. *The C programming language.* Englewood Cliffs: Prentice Hall, Inc.

King, Robert D. 1969. *Historical linguistics and generative grammar.* New Jersey: Prentice-Hall.

Kiparsky, Paul. 1968. Linguistic universals and linguistic change. Univerals in linguistic theory. 170-202. Emmon Bach and Robert T Harms (eds.) New York: Holt, Rinehart and Winston.

Kiparsky, Paul. 1973. Elsewhere in phonology. in *A festschrift for Morris Halle,* ed. by Stephen R. Anderson and Paul Kiparsky, 93-106. New York: Rinehart & Winston.

Kiparsky, Paul. 1974. Remarks on analogical change. Proceedings of the first international conference on historical linguistics. eds.) J.M Anderson and C Jones. North Holland Linguistic Series 12a. Edinburgh 2-7 September 1973: North-Holland Publishing Company.

Kiparsky, Paul. 1982. *Explanation in phonology.* Dordrecht: Foris Publications.

Kiparsky, Paul. 1988. Phonological Change. Linguistic Theory: Foundations 1. F.J. Newmeyer (ed.) 363-415. Cambridge: Cambridge University Press.

Kiparsky, Paul. 1993. The phonological basis of sound change. *Stanford Workshop on Sound Change*

Knuth, Donald Ervin. 1979. *TEX and METAFONT : new directions in typesetting.* Providence, R.I.: American Mathematical Society.

Kuhn, Thomas S. 1962. *The structure of scientific revolutions.* 2nd (1970) Chicago: University of Chicago Press.

Kuperus, J. 1985. *The Londo word: its phonological and morphological structure.* Annalen. Menselijke wetenschappen 119. Tervuren, Belgie: Musée royal de l'Afrique centrale.

La Polla, R. J., Lowe, J. B. 1989. *Bibliography of the International Conferences on Sino-Tibetan Languages and Linguistics, I - XXI. STEDT Monograph Series,.* 1. Institute of International Studies. Berkeley: University of California Press.

Labov, William. 1981. Resolving the neogrammarian controversy. *Language* 57.267-308.

Ladefoged, Peter. forthcoming. Some reflections on the IPA. *Language*

Laufer, B. 1916. *The Nichols Mo-so manuscript. Notes sur quelques populations au nord de l'Indo-Chine.* The Geographical Review 1, 274-85. Lefévre-Pontalis, P. 1892.

Lehmann, Winfred P. ed. 1967. *A reader in nineteenth-century historical Indo-European linguistics.* Indiana University studies in history and theory of linguistics. Bloomington and London: Indiana University Press.

Lehmann, Winfred P. 1991. The process of linguistics. Sprung from some common source: Investigations into the prehistory of languages. 11-22. Sydney Lamb and E. Douglas Mitchell (eds.) Stanford: Stanford University Press.

Lewis, Paul. 1968. *Akha-English Dictionary.* 70, Southeast Asia Program. Ithaca, NY: Data paper

Cornell Univ.

Lewis, P. 1986. *Lahu-English-Thai Dictionary.* Thailand Lahu Baptist Convention. Bangkok: Darnsutha Press.

Li, Fang kuei. 1977. A Handbook of Comparative Tai. Oceanic Linguistics Special Publication 15. 389. Honolulu: University Press of Hawaii.

Li, Wang. 1982. *A Dictionary of Word Families.* Beijing: Commercial Press.

Lowe, John B. 1993. The methodology of the Sino-Tibetan Etymological Dictionary and Thesaurus. *SEALS* 3

Lowe, John B., and Martine Mazaudon. 1989. Computerized tools for reconstruction for Tibeto-Burman. Proceedings of the Annual Meeting of the Berkeley Linguistics Society, 18-20 February1989 367-378.

Lowe, John B, and Martine Mazaudon. 1990. Phonological change in the Tamang languages of Nepal: Problems and prospects of a computerized study. 23rd Conference on Sino-Tibetan Languages and Linguistics. Arlington, Texas:

Lowe, John B., and Martine Mazaudon. 1994. The Reconstruction Engine: a computer implementation of the comparative method. *Computational Linguistics* 20.381-418.

Luce, G. H. 1981. *A Comparative Wordlist of Old Burmese, Chinese, and Tibetan.* London: School of Oriental and African Studies.

Lyovin, Anatole. 1968. A Chinese dialect dictionary on computer: progress report. *POLA* 2.c1-c43.

Macdonell, Arthur Anthony. 1929. *A practical Sanskrit dictionary : with translation, accentuation, and etymological analysis throughout.* London: Oxford University Press.

Mainwaring, G.B. A. Gruenwedel. 1898. *Dictionary of the Lepcha Language.* Berlin: Unger Bros.

Malkiel, Yakov. 1967. Every word has its own history. *Glossa* 1.137-49.

Malkiel, Yakov. 1967. Linguistics as a genetic science. 43.223-245.

Maniet, Albert. 1980. Recherche par ordinateur sur la phonologie diachronique du latin. 16th International Congress of Romance Linguistics. Palma de Mallorca:

Maniet, Albert. 1983. Justification of the formulation and position of phonological rules in a algorithmic series generating Early Latin from "Indo-European".

Marrison, Geoffrey E. 1967. The Classification of the Naga Languages of North-east India. Doctoral dissertation. School of Oriental and African Studies, Univ. London.

Martinet, André. 1987. Note sur les 'changements phonétiques'. *La Linguistique* 23.43-46.

Maspero, Henri. 1911. Contribution à l'étude du système phonétique des langues thai. *Bulletin de l'École Française d'Extrême Orient* 11.153-69.

Maspero, Henri. 1952. Les langues thai. Les Langues du Monde A. and Marcel Cohen Meillet (ed.) Paris: E. Champion.

Matisoff, James. 1986. Hearts and minds in South-east Asian languages and English: an essay in the comparative lexical semantics of psycho-collocations. *C. L. A. O* XV.5-57.

Matisoff, J.A. 1970. Glottal dissimilation and the Lahu high-rising tone: a tonogenetic case-study. *Journal of the American Oriental Society* 90.13-44.

Matisoff, James A. 1972. *The Loloish Tonal Split Revisited.* Research Monograph 7. Berkeley: Center for South and Southeast Asia Studies.

Matisoff, James A. 1972. Tangkhul Naga and comparative Tibeto-Burman. *Tonan Azia Kenkyu [Southeast Asian Studies]* (Kyoto) 10.2, 1-13.

Matisoff, James A. 1973. *The Grammar of Lahu.* University of California Publications in Linguistics 75. Berkeley and Los Angeles.: University of California Press.

Matisoff, James A. 1978. *Mpi and Lolo-Burmese microlinguistics.* Monumenta Serindica 4. Tokyo: Institute for the Study of Languages and Cultures of Asia and Africa.

Matisoff, James A. 1978. *Variational Semantics in Tibeto-Burman: the 'organic' approach to linguistic comparison.* Philadelphia: Publication of the Institute for the Study of Human Issues.

Matisoff, James A. 1979. Problems and progress in Lolo-Burmese: Quo Vadimus? *Linguistics of the Tibeto-Burman Area* 4.11-43.

Matisoff, James A. 1980. Stars, moon, and spirits: bright beings of the night in Sino-Tibetan. *Gengo Kenkyu* (Tokyo) 77.1-45.

Matisoff, James A. 1982. Proto-languages and proto-Sprachgefühl. *Linguistics of the Tibeto-Burman Area* 6.1-64.

Matisoff, James A. 1983. God and the Sino-Tibetan copula with some good news concerning selected Sino-Tibetan rhymes. Sixteenth

International Conference on Sino-Tibentan languages and linguistics, September 15-18, 1983 Seatt!e:

Matisoff, James A. 1985. A new Sino-Tibetan root *d-yu-k: BELONG/TRUST/DEPEND/ACCEPT/TAKE and A note of caution to megaloreconstructionists. *Prosodic Analysis and Asian Linguistics: to honour R. K. Sprigg* David Bradley, Eugénie J.A. Henderson and Martine Mazaudon (eds.) 265-269. Canberra: Pacific Linguistics.

Matisoff, James A. 1985. Out on a Limb: Arm, Hand and Wing in Sino-Tibetan. *Linguistics of the Sino-Tibetan Area: the State of the Art: Papers Presented to Paul K. Benedict for his 71st Birthday* James Matisoff, Graham Thurgood and David Bradley (eds.) 421-450. Canberra: Pacific Linguistics.

Matisoff, James A. 1986. Languages and Dialects of Tibeto-Burma. Contributions to Sino-Tibetan studies John McCoy and Timothy Light (eds.) 477. Leiden: E. J. Brill.

Matisoff, James A. 1987. Bulging monosyllables: areal tendencies in Southeast Asian diachrony. Proceedings of the Berkeley Lingustics Society. 15. Kira Hall (ed.) 543-59.

Matisoff, J. A. 1987. Review of D. Bernot, Dictionnaire birman-français. *Bulletin of the School of Oriental and African Studies* 50.191-5.

Matisoff, J. A. 1988. *The Dictionary of Lahu.* University of California Publications in Linguistics 111. Berkeley, Los Angeles, London: University of California Press.

Matisoff, James A. 1988. Universal semantics and allofamic identification: two Sino-Tibetan case-studies - 'straight/flat/full' and 'property/livestock/talent'. Languages and history in east Asia. Festschrift for Tatsuo Nishida on the occasion of his 60th birthday Kyoto: Shokado.

Matisoff, James A. 1991. Areal and universal dimensions of grammatization in Lahu. 383-453. Elizabeth C. Traugott and Bernd Heine, eds., *Approaches to Grammaticalization,* Vol. II. Amsterdam: John Benjamins.

Matisoff, James A. 1990a. Review article: On megalocomparison. *Language* 60.106-20.

Matisoff, James A. 1990b. Cognate grading and other desiderata for Lolo-Burmese studies. Paper presented at the International Conference on Sino-Tibetan Languages and Linguistics 23

Matisoff, James A. 1991. Lexicography of other Tibeto-Burman languages. *Dictionaries: an International Encyclopedia of Lexicography* 3. F. J. Hausmann, et al. (eds.) 2555-60. Berlin and New York: Walter de Gruyter and Co.

Matisoff, James A. 1991. The mother of all morphemes: augmentatives and diminutives in areal and universal perspective. Papers from the first annual meeting of the Southeast Asian Linguistics Society Martha Ratliff and Eric Shiller (eds.) 293-349. Tucson: Arizona State University.

Matisoff, James A. 1991. Sino-Tibetan linguistics: present state and future prospects. *Annual Review of Anthropology* 20.469-504.

Matisoff, James A. 1992. Following the marrow: two parallel Sino-Tibetan etymologies. *Fourth Spring Workshop on Theory and Method in Linguistic Reconstruction,* March 27-29, 1992. University of Pittsburg.

Matisoff, James A. 1994. Regularity and variation in Sino-Tibetan. Current issues in Sino-Tibetan linguistics: Kitamura, Nishida and Nagano (eds.) 36-58. Osaka: Organizing committee of the 26th ICSTLL.

Matisoff, James A. forthcoming. *The Sino-Tibetan etymological dictionary and thesaurus.* Berkeley: University of California Press.

Mawson, C. O. Sylvester ed. 1911. *Roget's Thesaurus of English words and phrases classified and arranged so as to facilitate the expression of ideas and assist in literary composition.* New York: Thomas Y. Crowell company.

Mazaudon, Martine. 1973. *Phonologie Tamang (Népal).* Paris: Société d'études Linguistiques et Anthropologiques de la France (SELAF).

Mazaudon, Martine. 1994. Problèmes de comparatisme et de reconstruction dans quelques langues de la famille tibéto-birmane. Thèse d'état. Université de la Sorbonne Nouvelle: Paris III.

Mazaudon, Martine, and John B. Lowe. 1991. Du bon usage de l'informatique en linguistique historique. *Bulletin de la Société de Linguistique de Paris* 86.49-87.

Mazaudon, Martine, and John B. Lowe. forthcoming. Regularity and exceptions in sound change. 1993 Annual meeting of the Belgian Linguistics Society.

McWhorter, John. Forthcoming. Renewing our vows: creole studies and historical linguistics. *BLS* 21

Meillet, Antoine. 1966. *La méthode comparative en linguistique historique.* Paris: Librairie ancienne Honoré Champion.

Meillet, Antoine. 1967. *The comparative method in historical linguistics.* Trans. Gordon B. Ford Jr. Paris: Librairie Honoré Champion.

Messinger, Heinz. 1973. *New college German dictionary.* Berlin and Münich: Langenscheidt.

Meussen, A. E. 1967. Bantu grammatical reconstructions. Annalen van het Koninklijk Museum voor Midden-Afrika 61. 79-21. Tervuren: Musée Royal de l'Afrique Centrale.

Michailovsky, Boyd. 1976. Notes on the Kiranti verb (East Nepal). *Linguistics of the Tibeto-Burman Area* 2.183-218.

Michailovsky, Boyd. 1981. *La langue Hayu: phonologie, morphologie, syntaxe.* Thèse de Doctorat Troisième Cycle, Université de Paris III.

Miller, G.A. et al. 1990. Wordnet: an on-line lexical database. *International Journal of Lexicography* 3

Miller, Roy Andrew. 1991. Genetic connections among the Altaic languages. Sprung from some common source: Investigations into the prehistory of languages Sydney Lamb and E. Douglas Mitchell (eds.) 293-327. Stanford: Stanford University Press.

Monier-Williams, Monier. 1899. *A Sanskrit-English dictionary etymologically and philologically arranged with special reference to cognate Indo-European languages.* Oxford: Clarendon Press.

Neuhaus. 1986. Phonetic character sets: printing, sorting, and computing. *Literary and linguistic computing* 1.166-167.

Nichols, Johanna. 1992. *Linguistic diversity in space and time.* Chicago: University of Chicago Press.

Nichols, Johanna. 1994. AAAS article.

Nicolai, Robert. 1991. *MARIAMA: base de données dialectologiques et lexicales sur la zone sahélo-saharienne et gestionanaire d'hypothèses.* Nice: URA 1235 du CNRS.

Nicolai, Robert. 1993. MARIAMA. (Paper presented at the first workshop on tools in computational historical linguistics, Brussels)

Ohala, J. J. 1981. The listener as a source of sound change. Papers from the parasession on language and behavior. eds.) C. S Masek, R. A Hendreck and M. F Miller. Chicago: Chicago Linguistic Society.

Okell, John. 1971. *A guide to the romanization of Burmese.* James G. Forlong Fund publications XXVII. London: The Royal Asiatic Society of Great Britain and Ireland.

Percival. 1987. Biological analogy before comparative grammar. Biological metaphor and classification Henry M. Hoenigswald and Linda F. Wiener (eds.) Philadelphia: University of Pennsylvania Press.

Pokorny. 1969. *Indogermanisches Etymologisches Wörterbuch.* Bern: A. Francke AG Verlag.

Polak-Bynon, Louise. 1975. *A Shi grammar: surface structures and generative phonology of a Bantu language.* Annalen Tervuren: Musee royal de l'Afrique centrale.

Popper, Sir Karl. 1938. *Scientific method.*

Pullum, Geoffrey K., and William A. Ladusaw. 1986. *Phonetic symbol guide.* Chicago: University of Chicago Press.

Ralston, Anthony and Edwin D. Reilly. 1993. *Encyclopedia of Computer Science.* 3rd Van Nostrand Reinhold.

Remmel, Mart. 1979. Computer techniques in Balto-Slavic historical phonetics. *Eesti NSV Teaduste Akadeemia Preprint KKI-11.*

*(Paper presented to the Symposium of Balto-Finnic Philology, Petrozavodzk).*

Renfrew, Collin. 1990. Spread of IE. *Sci Am?*

Ringe, Donald. forthcoming. 'Nostratic' and the factor of chance. *MS* [to appear in *Diachronica*].

Ringe, Donald A., Jr. 1992. On calculating the factor of chance in language comparison. *Transactions of the American Philosophical Society* 82.1-109.

Robins, Robert H. 1990. Leibniz and Wilhelm von Humboldt and the History of Comparative Linguistics. Leibniz, Humboldt, and the Origins of Comparativism. 49. Tullio de Mauro and Lia Formigari (ed.) 85-102. Amsterdam, Philadelphia: John Benjamins.

Ruhlen, Merritt. 1987. *A guide to the world's languages*. 1. Stanford, CA: Stanford University Press.

Ruvolo, Maryellen. 1991. Reconstructing genetic and linguistic trees: phenetic and cladistic approaches. Biological metaphor and cladistic classification. 193-216. Henry M. Hoenigswald and Linda F. Wiener (eds.) Philadelphia: University of Pennsylvania Press.

Schleicher. 1873. *Die Darwinsche Theorie und die Sprachwissenschaft*. 2nd Weimar: H. Bölau.

Schmidt, Johannes. 1872. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Weimar: H. Böhlau.

Schneider, Ben Ross. 1974. *Travels in computerland : or, Incompatabilities and interfaces : a full and true account of the implementation of the London stage information bank.* Reading, Mass.: Addison-Wesley Pub. Co.

Shannon, Claude, and Warren Weaver. 1949. *The mathematical theory of communication.* Urbana: University of Illinois Press.

Snoxall, R.A. 1967. *Luganda-English Dictionary.* Oxford: Oxford University Press.

Srinuan, D. 1976. *An Mpi Dictionary.* Bangkok: Indigenous Languages of Thailand Research Project, Central Institute of English Language, Office of State Universities: Ed. W. Pantupong.

Stein, Sir Aurel. 1928. *Innermost Asia: a detailed report of explorations in Central Asia, Kan-SU, and Eastern Iran.* reprinted 1981 2. New Delhi: Cosmos.

Streeter, Mary L. 1972. DOC: 1971. *Computers and the humanities* 6.259-70.

Sun, Tianshin. 1994. Proto-Tani. Ph.D. U.C. Berkeley.

Swadesh, Morris. 1950. Salish Internal Relationships. IJAL 16:157-67.

Sweetser, Eve. 1990. *From etymology to pragmatics.* Cambridge studies in linguistics 54. Cambridge: Cambridge University Press.

Swift, Jonathan. 1733. *On poetry: a rhapsody.* Dublin, and re-printed at London: J. Huggonson, next to Kent's Coffee-house, near

Serjeant's-inn, in Chancery-lane; [and] at the bookseller's and pamphletshops.

TEI P3. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Text Encoding Initiative 1. Chicago: The Association for Computers and the Humanities.

Thomason, Sarah, and Terrence Kaufman. 1988. *Language contact, creolization, and genetic classification.* Berkeley and Los Angeles: University of California Press.

Trubetskoy, N. S. 1939. Gedanken über das Indogermanenproblem. *Acta Linguistica* 1.81-89. [translated by Gary Holland]

Trubetzkoy, N. S. 1968. *Introduction to principles of phonological descriptions.* Trans. Murray, L. A. The Hague: Martinus Nijhoff.

Tryon, Darrell T. ed. 1995. *Comparative Austronesian dictionary.* Trends in Linguistics 10. Berlin, New York: Mouton de Gruyter.

Turner, Sir Ralph Lilley. 1962-69. *A comparative dictionary of the Indo-Aryan languages.* London, New York: Oxford University Press.

van Nooten, Barend A., and Gary B. Holland. 1994. *Rig Veda : a metrically restored text with an introduction and notes.* Harvard oriental series 50 Cambridge, Mass: Dept. of Sanskrit and Indian Studies, Harvard University.

Veatch, Thomas. 1993. Programs for CARP: Computer Aided (Historical Linguistic) Reconstruction in Phonology. (computer program).

Vennemann, Theo. 1983. Causality in language change : theories for linguistic preferences as a basis for linguistics explanations. *Folia Linguistica Historica* 6.5-26.

Walde, Alois, and Julius Pokorny. 1930. *Vergleichendes Wörterbuch der indogermanischen Sprachen.* Berlin, Leipzig: W. de Gruyter & Co.

Wang, William, S.-Y. 1979. Language change - a lexical perspective. *Ann. Rev. Anthropol* 8.353-71.

Wang, William S.-Y. 1969. Competing changes as a cause of residue. *Language* 45.9-25.

Wang, William S.-Y. 1970. Project DOC: its methodological basis. *Journal of the American Oriental Society* 90.57-66.

Wang, William S.-Y., and Chin-Chuan Cheng. 1977. Implementation of phonological change : the shuang-feng Chinese case. 148-159. The lexicon in phonological change 5. William S.-Y. Wang (ed.) The Hague: Mouton.

Watkins, Calvert. 1976. Towards Proto-Indo-European syntax: problems and pseudo-problems. Papers from the parasession on diachronic syntax Steever, Walker and Mufwene (eds.) 305-326. Chicago: Chicago Linguistic Society.

Watkins, Calvert. 1985. *The American heritage dictionary of Indo-European roots.* Boston: Houghton Mifflin.

Watkins, Calvert. 1989. New parameters in historical linguistics, philology, and culture history. *Language* 65.783-811.

*Webster's new collegiate dictionary..* 1975. Springfield, Mass.: G. & C. Merriam Co.

Wells, Rulon. 1979. Linguistics as a science: the case of the comparative method. The European Background of American Linguistics Henry M Hoenigswald (ed.) 23-61. Dordrecht: Foris.

Whitney, William Dwight. 1889. *A Sanskrit grammar including both the classical language, and the older dialects, of Veda and Brahmana.* 2nd Cambridge: Harvard University Press.

Whorf, Benjamin. 1956. *Language, thought and reality. Selected writings of B. L. Whorf.* Cambridge: MIT Press.

Wilson, Allen C, and Rebeca L. Cann. 1992. The recent African genesis of humans. *Scientific American* 266.68-73.

Wimbish, John S. 1989. *WORDSURV: A program for analyzing language survey word lists.* Occasional publications in academic computing 13. Dallas: Summer Institute of Linguistics.

Winder, Russel. 1991. *Developing C++ Software.* 2nd Chicester: John Wiley and Sons, Inc.

Wolfenden, Stuart N. 1929. *Outlines of Tibeto-Burman Linguistic Morphology*. Royal Asiatic Society, London.

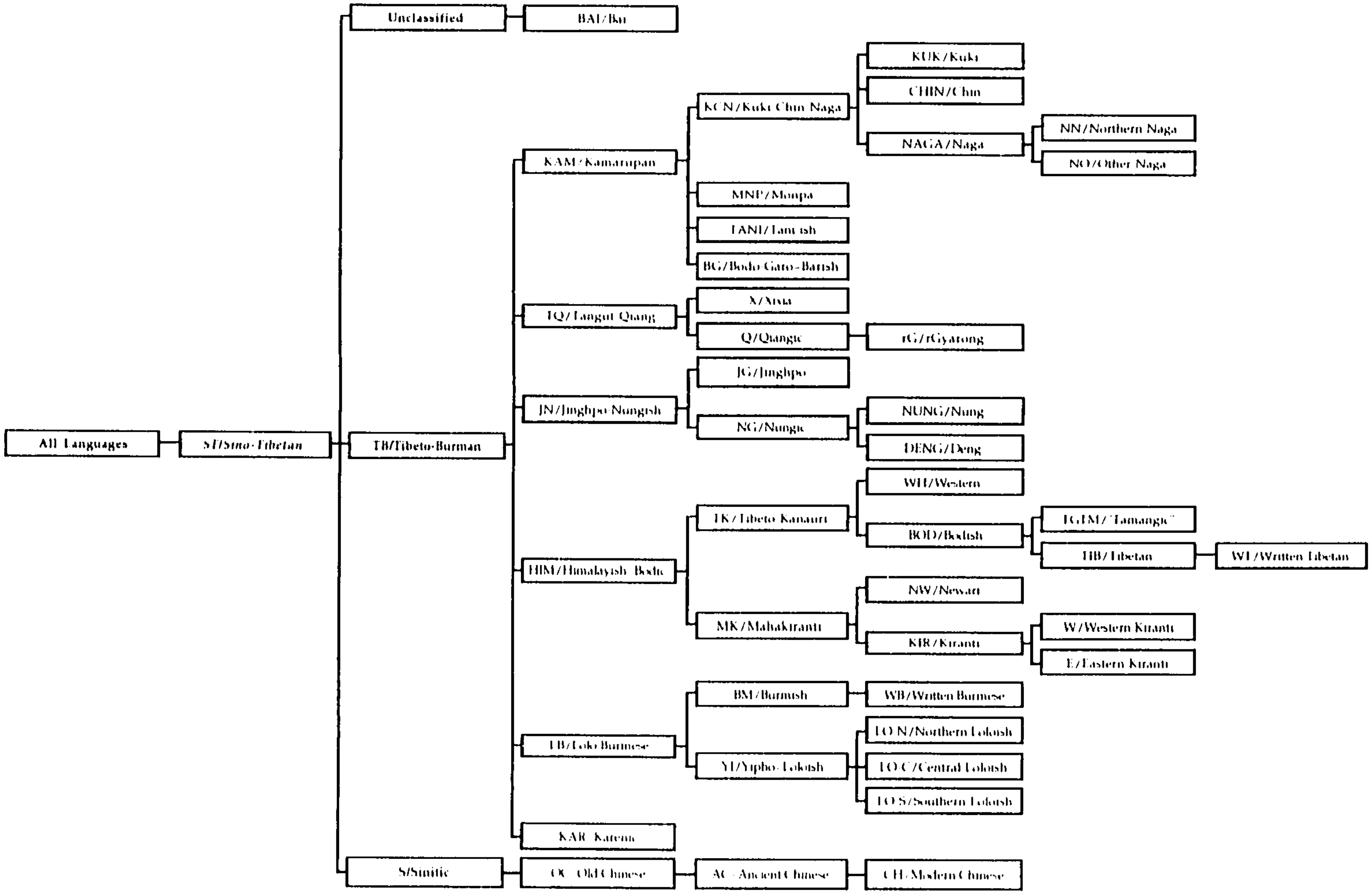
Xu Lin, Mu Yuzhang, Gai Xingzhi, eds. 1986. *Lisuyu jianzhi: a brief description of the Lisu language*. Beijing: Nationalities Press.

Yaruss, Jonathan Scott. 1990. DOC 1988: the modernization of a Chinese dialect dictionary on computer. *Computers and the humanities* 24.207-219.

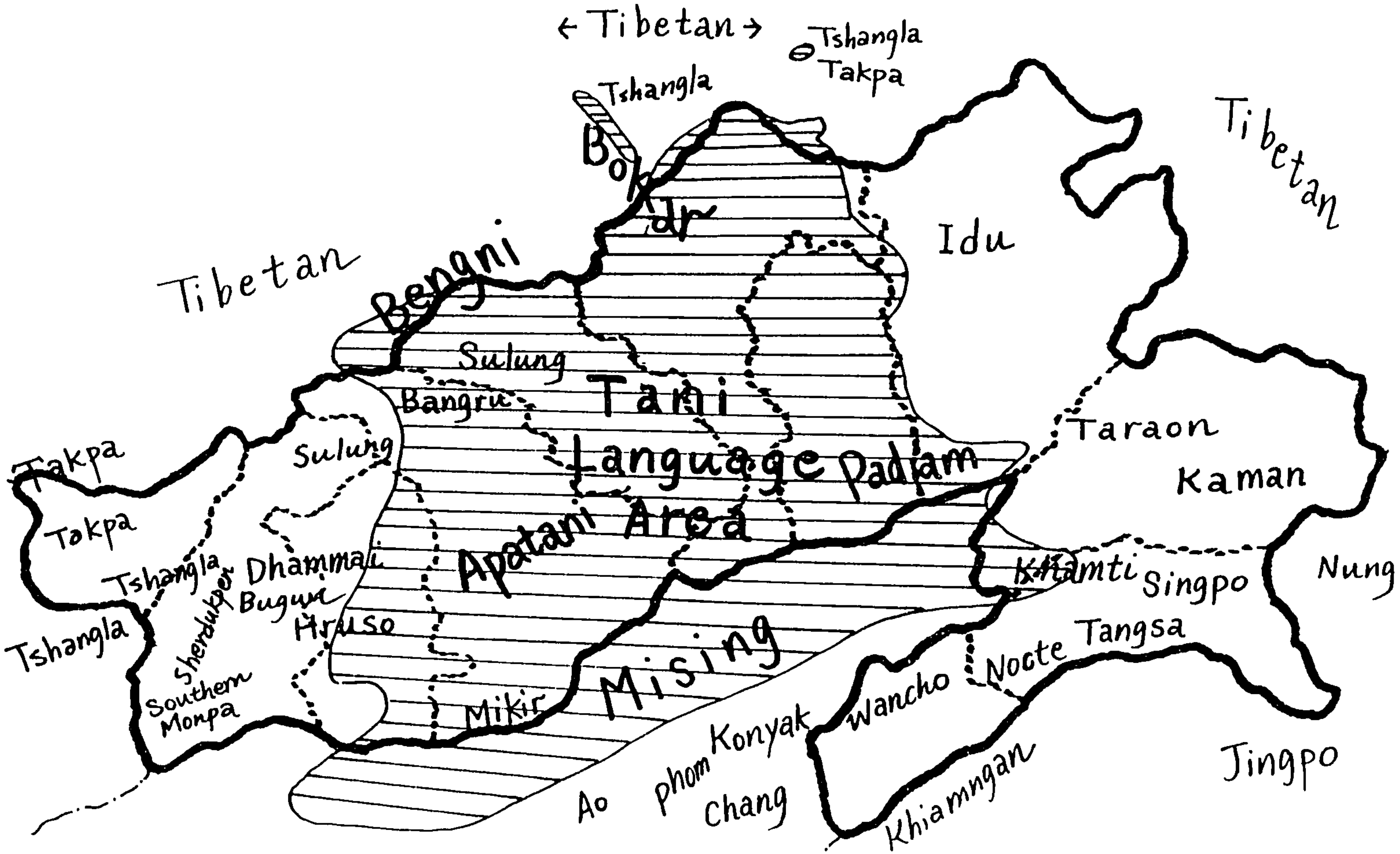
Zipf, G. K. 1939. The pyschobiology of language: An introduction to dynamic philology. London: G. Routledge & sons.

# Appendices

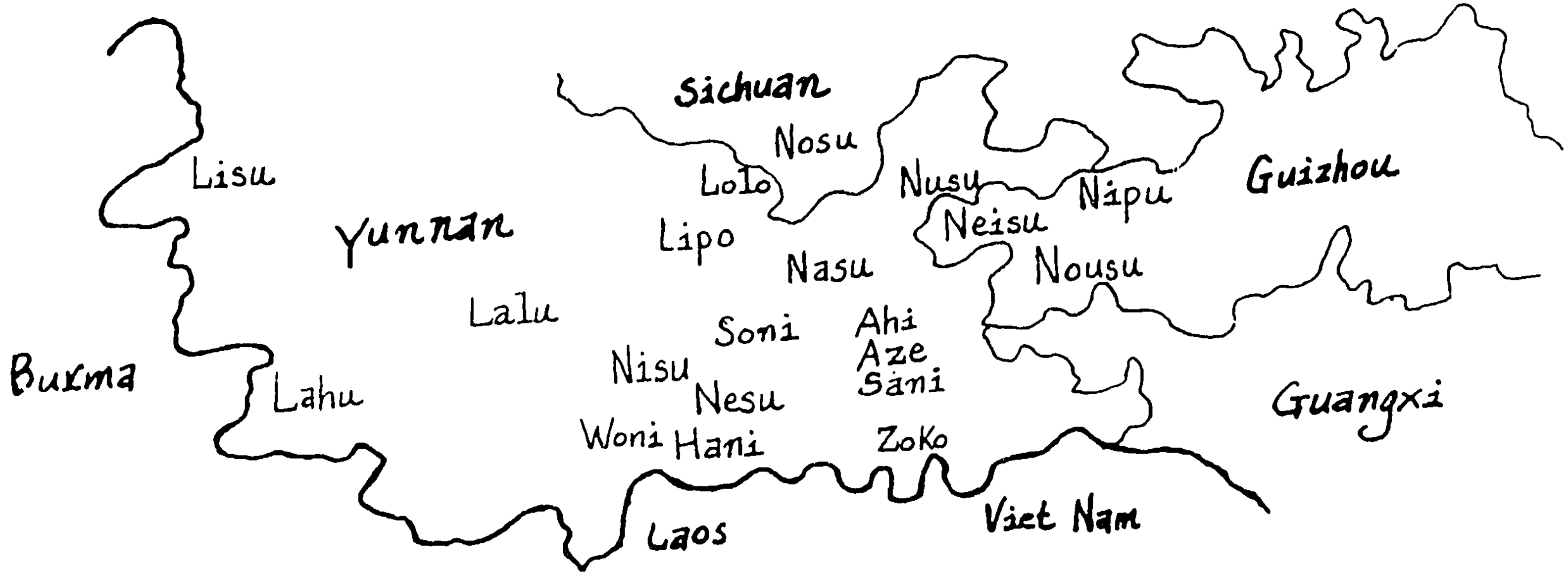
## Appendix 1: Preliminary Stambaum of Sino-Tibetan languages, based on data and research at the STEDT project



Appendix 2: The Tani Area (from Sun 1994)



Appendix 3.1: Northern Loloish languages (after Chen Kang 1986)



## Appendix 3.2: A brief description of the Loloish language data cited

The following summarizes some statistical and linguistic and linguistic properties of the three principal Loloish data sets used in exemplification in the main part of the dissertation. Since the analysis of these datasets is not complete, the presentation here is sketchy in places.

### A. Statistical Summary of the three Loloish datasets

The data from Appendices 1 and 2 of Inga-Lill Hansson's LTBA article ({Hansson, 1989 #852}) (hereinafter "ILH"). 306 cognate sets containing 1,934 forms in 8 languages are given (WB and the *L reconstructions are counted, since they are included in the Table of Correspondences and have been processed with RE). The Hani forms cited from the Chinese Wordlist are not, however, included. Including these forms would bring the reflex count for this dataset to about 2400.

The data from the three papers presented by Chen Kang at the Sino-Tibetan Conferences in 1986 and 1987 ({Chen 1986; 1987}) and also published in Chinese in *Minzu Yuwen* (hereinafter "CK"). 202 synonym sets containing 1,608 forms from 8 languages are given. A Proto-Yi reconstruction is offered, but it is not related to the reconstructions proposed by other linguists, and is restricted to Northern Loloish.

The data in numbered sets in Jim Matisoff's "The Loloish Tonal Split Revisited" ({Matisoff, 1972 #515}) (hereinafter TSR). I am using an annotated version of the original which contains about 1,868 forms in both Loloish and Burmish languages, organized into 192 cognates sets (with some sets further subdivided). Of these, I have focused only on the 1,199 forms from the nineteen Loloish languages and Burmese.

These three sets comprise about 700 cognate sets altogether, containing about 4,740 reflexes in 22 dialects (see (2) below). For comparison, {Bradley, 1979 #741} has slightly over 1,000 sets (the highest numbered set is 866, but many sets are broken down into smaller subsets), and supporting forms in seven languages are given. This implies that Bradley's dataset contains between 6,000 and 7,000 forms, though the number is probably somewhat less since not all cells in the matrix are filled. (The actual size of this dataset will be known soon, when conversion into computer form is completed.)

## Appendix 3.2: A brief description of the Loloish language data cited

### B. Statistical Summary of the Contents of the Three Loloish Datasets used here

|  | CK |  | ILH |  | TSR |  |
|---|---|---|---|---|---|---|
| Language | LgAbbr | N | LgAbbr | N | LgAbbr | N |
| **Northern Loloish (9 dialects/sources)** |  |  |  |  |  |  |
| Sani | sani | 202 |  |  | sa |  |
| Axi | axi | 202 |  |  | ahi |  |
| Nesu | nesu | 202 |  |  |  |  |
| Lalupo | lalo | 202 |  |  |  |  |
| Lipho | li | 202 |  |  |  |  |
| Nasu | nasu | 202 |  |  | na |  |
| Neisu | neisu | 202 |  |  |  |  |
| Nosu | nosu | 202 |  |  |  |  |
| Luquan |  |  |  |  | le |  |
| **Central Loloish (3 dialects/sources)** |  |  |  |  |  |  |
| Lahu |  |  |  |  | lh |  |
| Lisu |  |  |  |  | li |  |
| Woni |  |  |  |  | wo |  |

## Appendix 3.2: A brief description of the Loloish language data cited

### B. Statistical Summary of the Contents of the Three Loloish Datasets used here

|  | CK |  | ILH |  | TSR |  |
|---|---|---|---|---|---|---|
| Language | LgAbbr | N | LgAbbr | N | LgAbbr | N |
| **Southern Loloish (7 dialects/sources)** |  |  |  |  |  |  |
| Akha |  |  | akha | 311 | ak |  |
| Hani (L) |  |  | hanil | 311 |  |  |
| Haoni |  |  | haoni | 78 |  |  |
| Khatu |  |  | khatu | 261 |  |  |
| Pijo |  |  | pijo | 268 |  |  |
| Bisu |  |  |  |  | bi |  |
| Mpi |  |  | mpi | 254 | mpi |  |
| **Other (3 languages)** |  |  |  |  |  |  |
| Moso |  |  |  |  | mo |  |
| Proto-Loloish |  |  | pl | 246 | PL |  |
| Written Burmese |  |  | wb | 205 | WB |  |
| Totals |  | 1616 |  | 1934 |  | 1199? |

## Appendix 3.2: A brief description of the Loloish language data cited

### C. Sources Citations

The citations here are from Hansson 1989, except where noted by *.

Benedict, Paul. *Sino-Tibetan: A Conspectus*. Contributing Editor: James A. Matisoff, Cambridge University Press, 1972.

Benedict, Paul. *Rhyming dictionary of written Burmese*, LTBA 3:1, 1976.

Bradley, David: *Proto-Loloish*, Scandinavian Institute of Asian Studies, Monograph Series No. 39, London and Malmö, 1979.

Bradley, David: *The Háoní 'Dialect' of Hani*, 1982 (?), 22 pp.

*Chen Kang. *A Comparative Study of the Vowels among the Dialects of the Yi Languages*. Paper presented at the 2nd International Workshop on the Lolo (Yi) Languages, Lund, Sweden, 1986. [published in Chinese as "Study of the rhyme correspondence in Yi" *Yuyan Yanjiu* 1987:2(1987)]

*Chen Kang. *The Proto-Yi Tone System (summary)*. MS. (lecture delivered in Lund, Sweden, 1986). [published in Chinese as "Tonal correspondences in Yi" *Minzu Yuwen* 1986:5]

Duanghom, Srinuan: *Mpi-Thai-English Dictionary*, Bangkok, 1976.

*Dai Qingxia: *Zangmianyu Shiwuzhong* [Fifteen Tibeto-Burman languages]. Beijing: Yanshan Chubanshe, 1991.

Egerod, Søren, and Inga-Lill Hansson: "An Akha conversation on death and funeral," *AO* 36 (1974), 225-284.

*Haqniq pyulniul soqmiav niq pyu hu Zangmianyu Yuyin. Han Cihuiaol e nilgevnei soqhhavq/Ha-Hàn duìzhào xiao cíhuì*, Kunming, 1959, 108 p. (*A Short Hani-Chinese Vocabulary*).

## Appendix 3.2: A brief description of the Loloish language data cited

### C. Sources Citations (continued)

*Hansson, Inga-Lill: A Comparison of Akha, Hani, Khàtú, and Pîjɔ̀. *Linguistics of the Tibeto-Burman Area* 12.1. (1989), 6-91.

Hansson, Inga-Lill: Sound Changes in Akha. Paper presented at 12th Sino-Tibetan Conference, Paris, 1979, 25 pp.

Hansson, Inga- Lill: "A Phonological Comparison of Akha and Hani," *Linguistics of the Tibeto-Burman Area*7.1 (1982), 63-115.

Hú Tan and Dài Qìng-xià: "Haní yu yuán yin song jin," *Zhongguó yuwén* 1 (1964), 76-87 (Vowels with and without stricture in the Hani Language).

Li Yong-suì: *Ha-ní yu gàikuàng*, *Mínzú yuwén* 2 (1979), 134-151 (A Brief Descrip-tion of the Hani Language).

Matisoff, James A.: *The Loloish Tonal Split Revisited*, Research Monograph No. 7, Center for South and Southeast Asia Studies, University of California, Berkeley, 1972.

Matisoff, James A.: "Mpi and Lolo-Burmese Microlinguistics," *Monumenta Serindica* No. 4 (1978), 36 p.

Wáng Er-song: Cóng fangyán bijiào kàn Háoní huà de yuyin tèzheng, in: *Mínzú yuwén lùn jí*, Beijing, 1981, p. 495-504. (Phonological characteristics of Hani as seen from dialect comparisons.)

**Zangmianyu Yuyin. Han Cihui* [Tibeto-Burman. Phonology and Lexicon]. Beijing: Chinese Social Sciences Press, 1991. 1420 pp.

## Appendix 3.2: A brief description of the Loloish language data cited

### D. Languages Cited

### D1 Languages cited by Hansson (quoted almost verbatim from Hansson 1989 pp.6-7)

1. **Akha** (Thailand). Based on Hansson's files from fieldwork among the djə-ɣò àkhà in northern Thailand. Described e. g. in Hansson 1982.

2. **Akha** (Yunnan). All 581 words taped by Hansson in Kunming with two (part of it with three) male informants, aged 17, 18 and 32 respectively, who had spent 1-2 years in Kunming. All of them come from Mĕnglà county, in the Xīshuāngbănnà Dăi Autonomous District. Apart from the Mĕng-là county, the Akha mainly live in the Mĕnghăi and Jĭnghóng counties, all in the same district. According to Lĭ Yŏngsuì they amount to around 100,000 people there. About 30,000 Akha live in the Lăncāng Lāhù Autonomous County and in the Mĕnglián Dăi-Lāhù-Wă Autonomous County. They are reported as calling themselves ja$^{21}$ ɳi$^{21}$, written in Chinese as or (Li 1979). The informants Hansson worked with called themselves àkhà, which they claimed was true for the whole group in Xīshuāng-bănnà. This was also confirmed to Hansson by some Akha farmers, whom she happened to meet in Jĭnghóng, and who also were of the same clan, djə-ɣò, as the main group in Thailand. Hansson believes that the language is the same as the one spoken by the Akha in Thailand.

3. **Hani Wordlist** (Hani W). This is the language described in Hansson 1982, based on the Hani-Chinese Wordlist, 1959.

4. **Xhànì**, in Chinese Hāní, here written Hani Lüchūn (Hani L). All 581 words were recorded by Hansson with one male informant in Kunming, aged 17, resident in Kunming since the previous year. He is from Lüchūn county. There are said to be around 500,000 Hani, mainly living in the Hónghé Hāní Yí Autonomous District.

## Appendix 3.2: A brief description of the Loloish language data cited

5. **Khàtú**, in Chinese Kăduō. Glosses 1-508 on tape, the rest only in Hansson's notes. Recorded in Kunming with one male informant, aged 16, resident in Kunming for a year. He is from Mòjiāng county.

6. **Pîjɔ̀**, in Chinese Bìyuē. Glosses 1-451 on tape, the rest from Hansson's notes. Recorded in Kunming with one male informant, aged 29, resident in Kunming for three years. He is from Mòjiāng county.

7. **Xɔ$^{31}$ni$^{31}$**, in Chinese Háoní. Available words (84) given in Bradley (1985) based on Wáng (1981) are quoted for comparison. Khàtú, Pîjɔ̀, and Háoní are mainly spoken in the Mòjiāng, Jiāngchéng, and Yuánjiāng counties by around 300,000 speakers.

8. **Mpi**. Cognates according to Duanghom, *Mpi-Thai-English Dictionary*, 1976. Some words added by Bradley (1979).

9. **Proto-Loloish**. Reconstructions based in part on Bradley (1979) and Matisoff (1972).

10. **Written Burmese**. Cited by Hansson from Benedict (1976).

D2 **Languages cited by Chen Kang (1986). Included here are data from languages cited either in his paper delivered at the Lolo (Yi) language workshop or in the lecture given on Proto-Yi tones.**

11. Yunnan Lùnán Sani (nɪ$^{21}$: Southeastern Dialect, Sani Variety)

12. Yunnan Mílè Axi (a$^{21}$ɕi$^{33}$p$^{h}$o$^{21}$: Southeastern Dialect, Axi Variety)

13. Yunnan Shípíng Nesu (ne̱$^{33}$su$^{55}$: Southern Dialect, Shípíng Variety)

14. Yunnan Wéishān Lalupa (la̱$^{2}$lu$^{33}$pa̱$^{2}$: Western Dialect)

15. Yunann Dàyáo Lipo (li$^{33}$p$^{h}$o$^{21}$: Central Dialect, Lipo Variety)

## Appendix 3.2: A brief description of the Loloish language data cited

16. Yunnan Lùqüan Nasupo (nạ$^{55}$su$^{33}$p$^{h}$o$^{55}$: Eastern Dialect, Northeastern Yunnan Subdialect, Black Yi variety)

17. Guizhou Wēiníng Nɤsu (nɤ$^{55}$su$^{13}$: Eastern Dialect, Northwestern Subdialect, Wūsǎ Variety)

18. Sichuan Xǐdé Nosu (nọ$^{33}$su$^{33}$: Northern Dialect, Northern Subdialect, Shèngzuò Variety)

## D3 Languages cited by Matisoff (1972)

Not included here as they are only treated in passing.

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 1. | ʒa$^{2}$ | *son* | za$^{21}$ | zo$^{21}$ | zo$^{33}$ | za$^{21}$ | zo$^{21}$ | zo$^{33}$ | zu$^{33}$ | zɯ$^{33}$ |
| 2. | tsa$^{2}$ | *salt* | tsha$^{21}$ | tsho$^{21}$ | tsho$^{33}$ | tsha$^{21}$ | tsho$^{21}$ | tsho$^{33}$ | tshu$^{33}$ | tshɯ$^{33}$ |
| 3. | xa$^{2}$ | *meat* | xa$^{21}$ | xo$^{21}$ | xo$^{33}$ | xa$^{21}$ | xo$^{21}$ | xo$^{33}$ | xu$^{33}$ | ʂɯ$^{33}$ |
| 4. |  | *buckwheat* | qa$^{21}$ | go$^{21}$ | go$^{33}$ | ɣa$^{21}$ | go$^{21}$ | ngho$^{33}$ | ngu$^{33}$ | ngɯ$^{33}$ |
| 5. | dza$^{2}$ | *eat* | dza$^{21}$ | dzo$^{21}$ | dzo$^{33}$ | dza$^{21}$ | dzo$^{21}$ | dzo$^{33}$ | dzu$^{33}$ | dzɯ$^{33}$ |
| 6. | k-ra$^{2}$ | *strength* | ɣa$^{21}$ | ɣo$^{21}$ | ɣo$^{33}$ | ɣa$^{21}$ | ɣo$^{21}$ | ɣo$^{33}$ | ɣu$^{33}$ | ɣɯ$^{33}$ |
| 7. | (g)wa$^{2}$ | *chew* | ga$^{21}$ | go$^{21}$ | go$^{33}$ | ga$^{21}$ | go$^{21}$ | ngho$^{33}$ | ngu$^{33}$ | ngɯ$^{33}$ |
| 8. | < Ch. | *dragon* | lu$^{21}$ | lo$^{21}$ | lo$^{33}$ | lu$^{21}$ | lu$^{21}$ | lu$^{33}$ | lu$^{33}$ | lu$^{33}$ |
| 9. | baŋ$^{2}$ | *deaf* | bu$^{21}$ | bo$^{21}$ | ba$^{33}$ | bu$^{21}$ | ba$^{21}$ | bɔ$^{33}$ | bo$^{33}$ | bo$^{33}$ |
| 10. |  | *under* | qho$^{21}$ | ko$^{21}$ | ka$^{33}$ | ku$^{21}$ | ka$^{21}$ | kɔ$^{33}$ | ko$^{33}$ | ko$^{33}$ |
| 11. | ʔ-maŋ$^{2}$ | *pine* | tho$^{21}$ | tho$^{21}$ | tha$^{33}$ | thu$^{21}$ | tha$^{21}$ | thɔ$^{33}$ | tho$^{33}$ | tho$^{33}$ |
| 12. | tsiy$^{2}$ | *wash* | tʂhi$^{21}$ | tshi$^{21}$ | tɕhi$^{33}$ | tshi$^{21}$ | tshi$^{21}$ | tshi$^{33}$ | tshi$^{33}$ | tshi$^{33}$ |
| 13. | b-ni$^{2}$ | *near* | næ$^{21}$ | na$^{21}$ | nɛ$^{33}$ | ne$^{21}$ | nɛ$^{21}$ | nɔ$^{33}$ | nɛ$^{33}$ | ɳi$^{33}$ |
| 14. | C-mi$^{2}$ | *daughter* | mæ$^{21}$ | mɛ$^{21}$ | mɛ$^{33}$ | me$^{21}$ | mɛ$^{21}$ | mɔ$^{33}$ | mɛ$^{33}$ | mi$^{33}$ |
| 15. | zum$^{2}$ | *use* | zɤ$^{21}$ | zi$^{21}$ | zɤ$^{33}$ | zy$^{21}$ | zɤ$^{21}$ | zɤ$^{33}$ | zɤ$^{33}$ | zɤ$^{33}$ |
| 16. | ko$^{2}$ | *steal* | khɯ$^{21}$ | khɯ$^{21}$ | khɯ$^{33}$ | khu$^{21}$ | khɯ$^{21}$ | khɯ$^{33}$ | tɕhy$^{33}$ | khu$^{33}$ |
| 17. | ɣo$^{2}$ | *bone* | ɣɯ$^{21}$ | ɣɯ$^{21}$ | ɣɯ$^{33}$ | ʔvu$^{21}$ | ɣɯ$^{21}$ | vɯ$^{33}$ | ʑy$^{33}$ | vu$^{33}$ |
| 18. | ǰim$^{2}$ | *raw* | dzi$^{21}$ | dzɛ$^{21}$ | dzɛ$^{33}$ | dzi$^{21}$ | dzi$^{21}$ | dzɿ$^{33}$ | dzɛ$^{33}$ | dzɿ$^{33}$ |
| 19. | ʔgyak | *vegetable* | ɣo$^{21}$ | vu$^{21}$ | ɣa$^{21}$ | ɣu$^{21}$ | va$^{21}$ | ɣɔ$^{55}$ | ɣo$^{55}$ | vo$^{21}$ |
| 20. | maŋ$^{2}$ | *old* | mo$^{21}$ | mu$^{21}$ | ma$^{21}$ | mu$^{21}$ | ma$^{21}$ | mɔ$^{55}$ | mo$^{55}$ | mo$^{21}$ |
| 21. | gaŋ$^{2}$ | *empty* | qo$^{21}$ | gu$^{21}$ | ga$^{21}$ | gu$^{21}$ | ga$^{21}$ | gɔ$^{55}$ | go$^{55}$ | go$^{21}$ |
| 22. | (k)-rwaŋ$^{2}$ | *sell* | vu$^{21}$ | vu$^{21}$ | ɣu$^{21}$ | vu$^{21}$ | vu$^{21}$ | vu$^{55}$ | ɣu$^{55}$ | vɯ$^{21}$ |
| 23. | ʔnit$^{1}$ | *two* | ni$^{21}$ | ni$^{21}$ | ɳi$^{21}$ | ni$^{21}$ | ɳi$^{21}$ | ɳi$^{55}$ | ɳi$^{55}$ | ɳi$^{21}$ |
| 24. | ʔ-kri$^{2}$ | *rot* | tʂhi$^{21}$ | tʂhi$^{21}$ | tshi$^{21}$ | ? | tʂhi$^{21}$ | tʂhi$^{55}$ | tʂhy$^{55}$ | tʂhi$^{21}$ |
| 25. | N-šit$^{1}$ | *seven* | ʂi$^{21}$ | ʂi$^{21}$ | ɕi$^{21}$ | xɯ$^{21}$ | ʂi$^{21}$ | ɕi$^{55}$ | ɕi$^{55}$ | ʂi$^{21}$ |
| 26. |  | *let go* | ɬɤ$^{21}$ | ɬhɯ$^{21}$ | thɤ$^{21}$ | phy$^{21}$ | phɤ$^{21}$ | ɬhɤ$^{55}$ | ɬhɤ$^{55}$ | thi$^{21}$ |
| 27. | ʔ-l(y)a$^{1}$ | *tongue* | ɬa$^{33}$ | lo$^{33}$ | ɬo$^{55}$ | ʔla$^{55}$ | lo$^{33}$ | ɬo$^{33}$ | ɬu$^{33}$ | ha$^{33}$ |
| 28. | ma$^{1}$ | *bamboo* | ma$^{33}$ | mo$^{33}$ | mo$^{55}$ | ma$^{55}$ | mo$^{33}$ | mo$^{33}$ | mu$^{33}$ | ma$^{33}$ |
| 29. | ʔ-na$^{1}$ | *ask* | na$^{33}$ | no$^{33}$ | no$^{55}$ | ʔna$^{55}$ | no$^{33}$ | no$^{33}$ | no$^{33}$ | ṇa$^{33}$ |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 30. | ʔ/g-raw$^{1}$ | *fry* | ɬu$^{33}$ | lo$^{33}$ | ɬu$^{55}$ | ʔlṳ$^{55}$ | lu$^{33}$ | ɬu$^{33}$ | ɬu$^{13}$ | ɬu$^{33}$ |
| 31. | | *bake* | ko$^{33}$ | ku$^{33}$ | ka$^{55}$ | ku$^{55}$ | ka$^{33}$ | kɔ$^{33}$ | ko$^{13}$ | ko$^{33}$ |
| 32. | | *turn* | tʂo$^{33}$ | tʂu$^{33}$ | tsa$^{55}$ | tʂu$^{55}$ | tʂa$^{33}$ | tʂɔ$^{33}$ | tʂo$^{13}$ | tɕo$^{33}$ |
| 33. | | *shy* | to$^{33}$ | tu$^{33}$ | ta$^{55}$ | tu$^{55}$ | ta$^{33}$ | tɔ$^{33}$ | to$^{13}$ | to$^{33}$ |
| 34. | tsu$^{1}$ | *fat* | tshi$^{33}$ | tsho$^{33}$ | tshu$^{55}$ | tshi$^{55}$ | tshu$^{33}$ | tshu$^{33}$ | tshu$^{13}$ | tshu$^{33}$ |
| 35. | C-gray$^{1}$ | *star* | tɕæ$^{21}$ | tʂa$^{21}$ | tsɛ$^{33}$ | tɕe$^{21}$ | kɛ$^{21}$ | tɕɔ$^{33}$ | tɕɛ$^{33}$ | tɕi$^{33}$ |
| 36. | | *grow* | tæ$^{21}$ | ta$^{21}$ | tɛ$^{33}$ | te$^{21}$ | tɛ$^{21}$ | tɔ$^{33}$ | tɛ$^{33}$ | tsi$^{33}$ |
| 37. | dza-n | *rice (cooked)* | tsa$^{33}$ | tso$^{2}$ | dzo$^{21}$ | dza$^{55}$ | dzo$^{33}$ | dzo$^{21}$ | dzo$^{21}$ | dza$^{33}$ |
| 38. | C-ŋa$^{1}$ | *I* | ŋa$^{33}$ | ŋo$^{2}$ | ŋo$^{21}$ | ŋa$^{55}$ | ŋo$^{33}$ | ŋo$^{21}$ | ŋo$^{21}$ | ŋa$^{33}$ |
| 39. | C-na$^{1}$ | *ill* | na$^{33}$ | no$^{2}$ | no$^{21}$ | na$^{55}$ | no$^{33}$ | no$^{21}$ | no$^{21}$ | na$^{33}$ |
| 40. | C-ra$^{1}$ | *hundred* | ha$^{33}$ | xo$^{2}$ | xo$^{21}$ | ha$^{55}$ | xo$^{33}$ | ho$^{21}$ | ho$^{21}$ | ha$^{33}$ |
| 41. | | *intestines* | vu$^{33}$ | ɣo$^{2}$ | vu$^{21}$ | vṳ$^{55}$ | vu$^{33}$ | ɣu$^{21}$ | ɣu$^{21}$ | vu$^{33}$ |
| 42. | | *make* | mu$^{33}$ | mo$^{2}$ | mu$^{21}$ | mṳ$^{55}$ | mu$^{33}$ | mu$^{21}$ | mu$^{21}$ | mu$^{33}$ |
| 43. | doŋ$^{1}$ | *wing* | tu$^{33}$ | to$^{2}$ | do$^{21}$ | dṳ$^{55}$ | du$^{33}$ | du$^{21}$ | du$^{21}$ | du$^{33}$ |
| 44. | | *human* | tsho$^{33}$ | tshu$^{2}$ | tsha$^{21}$ | tshu$^{55}$ | tsha$^{33}$ | tshɔ$^{21}$ | tsho$^{21}$ | tsho$^{33}$ |
| 45. | m-gaŋ$^{1}$ | *pull* | qo$^{33}$ | ku$^{2}$ | ga$^{21}$ | ɣu$^{55}$ | ga$^{33}$ | gɔ$^{21}$ | go$^{21}$ | go$^{33}$ |
| 46. | ʒo$^{1}$ | *sheep* | ʐo$^{33}$ | ʐu$^{2}$ | xa$^{21}$ | ʐu$^{55}$ | ʐa$^{33}$ | hɔ$^{21}$ | ho$^{21}$ | ʐo$^{33}$ |
| 47. | ti(y) | *water* | ʐi$^{33}$ | ʐi$^{2}$ | ʐi$^{21}$ | ɣɯ$^{55}$ | ʐi$^{33}$ | ʐi$^{21}$ | ʐi$^{21}$ | ʐi$^{33}$ |
| 48. | ʔkuk/ʔguk | *skin* | tʂi$^{33}$ | tɕi$^{2}$ | dzi$^{21}$ | gɯ$^{55}$ | dzi$^{33}$ | ndzi$^{21}$ | ndzi$^{21}$ | ndzi$^{33}$ |
| 49. | siy$^{1}$ | *die* | ʂi$^{33}$ | ʂi$^{2}$ | ɕi$^{21}$ | xɯ$^{55}$ | si$^{33}$ | ɕi$^{21}$ | ɕi$^{21}$ | si$^{33}$ |
| 50. | | *melt* | ? | tɕi$^{2}$ | dzi$^{21}$ | gɯ$^{55}$ | dzi$^{33}$ | dzi$^{21}$ | dzi$^{21}$ | dzi$^{33}$ |
| 51. | | *wine* | tʂi$^{33}$ | tɕi$^{2}$ | dʐi$^{21}$ | dʐi$^{55}$ | dzi$^{33}$ | ɳɖʐi$^{21}$ | ɳɖʐi$^{21}$ | ɳɖʐi$^{33}$ |
| 52. | way$^{1}$ | *buy* | væ$^{33}$ | va$^{2}$ | vɛ$^{21}$ | ve$^{55}$ | vɛ$^{33}$ | vɔ$^{21}$ | vɛ$^{21}$ | vi$^{33}$ |
| 53. | s-ŋa$^{2}$ | *borrow* | ŋa$^{55}$ | ŋo$^{55}$ | ŋo$^{33}$ | a$^{21}$ | ŋo$^{55}$ | ŋo$^{33}$ | ŋu$^{33}$ | hɯ$^{33}$ |
| 54. | | *stop* | na$^{55}$ | no$^{55}$ | no$^{33}$ | ʔna$^{21}$ | no$^{55}$ | no$^{33}$ | nu$^{33}$ | nɯ$^{33}$ |
| 55. | | *lay up* | ta$^{55}$ | to$^{55}$ | to$^{33}$ | ta$^{21}$ | to$^{55}$ | to$^{33}$ | tu$^{33}$ | ta$^{33}$ |
| 56. | m-ga$^{2}$ | *want* | ŋo$^{55}$ | ŋu$^{55}$ | ŋa$^{33}$ | ? | ŋo$^{55}$ | ŋɔ$^{33}$ | ŋo$^{33}$ | ŋo$^{33}$ |
| 57. | ʔduk | *kneel* | kɯ$^{55}$ | kɯ$^{55}$ | kɯ$^{33}$ | gɯ$^{21}$ | kɯ$^{55}$ | kɯ$^{33}$ | tɕy$^{33}$ | ku$^{33}$ |
| 58. | (k-)la | *tiger* | la$^{55}$ | lo$^{55}$ | lo$^{2}$ | la$^{2}$ | lo$^{55}$ | lo$^{55}$ | lu$^{55}$ | la$^{55}$ |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 59. | | *trousers* | ɬa$^{55}$ | lo$^{55}$ | ɬo$^{2}$ | ʔla$^{2}$ | ɬu$^{55}$ | ɬo$^{55}$ | ɬu$^{55}$ | ɬa$^{55}$ |
| 60. | s-ma$^{2}$ | *teach* | mo$^{55}$ | mu$^{55}$ | mo$^{2}$ | ʔma$^{2}$ | mu$^{55}$ | mo$^{55}$ | mu$^{55}$ | m̩a$^{55}$ |
| 61. | ʔ-krwe$^{2}$ | *sweat* | tçæ$^{55}$ | tʂa$^{55}$ | tʂa$^{2}$ | tçe$^{2}$ | kɛ$^{55}$ | tçɔ$^{55}$ | tçɛ$^{55}$ | ? |
| 62. | | *congeal* | tɤ$^{55}$ | tɯ$^{55}$ | tɤ$^{2}$ | ty$^{2}$ | tɤ$^{55}$ | te$^{55}$ | tɤ$^{55}$ | ti$^{55}$ |
| 63. | k-rak$^{H}$ | *chicken* | ze$^{44}$ | ze̠$^{33}$ | ze̠$^{33}$ | zi̠$^{33}$ | ze̠$^{33}$ | ɣa̠$^{2}$ | ɣa$^{33}$ | va̠$^{4}$ |
| 64. | ʔtak$^{H}$ | *embrace* | te$^{44}$ | te̠$^{33}$ | te̠$^{33}$ | ti̠$^{33}$ | te̠$^{33}$ | ta̠$^{2}$ | ta$^{33}$ | to̠$^{4}$ |
| 65. | | *harrow* | tçe$^{44}$ | tçe̠$^{33}$ | tçe̠$^{33}$ | tçi̠$^{33}$ | tçe̠$^{33}$ | tça̠$^{2}$ | tça$^{33}$ | go̠$^{4}$ |
| 66. | s-nak$^{H}$ | *black* | ne$^{44}$ | ne̠$^{33}$ | ne̠$^{33}$ | ni̠$^{33}$ | ne̠$^{33}$ | na̠$^{2}$ | na$^{33}$ | no̠$^{4}$ |
| 67. | kywan$^{1/2/3}$ | *sharpen* | the$^{44}$ | the̠$^{33}$ | the̠$^{33}$ | thi̠$^{33}$ | the̠$^{33}$ | tha̠$^{2}$ | tha$^{33}$ | tho̠$^{4}$ |
| 68. | | *go up* | de$^{44}$ | de̠$^{33}$ | de̠$^{33}$ | di$^{33}$ | de̠$^{33}$ | da̠$^{2}$ | da$^{33}$ | do̠$^{4}$ |
| 69. | s-nök$^{H}$ | *bean* | no$^{44}$ | nu̠$^{33}$ | nu̠$^{33}$ | no̠$^{33}$ | nu̠$^{33}$ | no̠$^{2}$ | ɳe$^{33}$ | nu̠$^{4}$ |
| 70. | m-bliŋ$^{3}$ | *be full* | bo$^{44}$ | bu̠$^{33}$ | bu̠$^{33}$ | bo̠$^{33}$ | bu̠$^{33}$ | mbho̠$^{2}$ | mbe$^{33}$ | mbu̠$^{4}$ |
| 71. | ʔdwak$^{L}$ | *go out* | do$^{44}$ | du̠$^{33}$ | du̠$^{33}$ | do̠$^{33}$ | du̠$^{33}$ | do̠$^{2}$ | de$^{33}$ | du̠$^{4}$ |
| 72. | | *lean on* | to$^{44}$ | tu̠$^{33}$ | tu̠$^{33}$ | to̠$^{33}$ | tu̠$^{33}$ | to̠$^{2}$ | te$^{33}$ | tu̠$^{4}$ |
| 73. | Nkrok$^{H}$ | *fear* | <go$^{44}$> | dʑu̠$^{33}$ | dʑu̠$^{33}$ | go̠$^{33}$ | dʑu̠$^{33}$ | dzo̠$^{2}$ | dze$^{33}$ | dzu̠$^{4}$ |
| 74. | k-lok$^{L}$ | *stone* | lo$^{44}$ | lo̠$^{33}$ | lu̠$^{33}$ | lo̠$^{33}$ | lu̠$^{33}$ | lo̠$^{2}$ | le$^{33}$ | lu̠$^{4}$ |
| 75. | | *hold in arms* | va$^{44}$ | va̠$^{33}$ | va̠$^{33}$ | va̠$^{33}$ | va̠$^{33}$ | vɚ̠$^{2}$ | va$^{33}$ | va̠$^{4}$ |
| 76. | N-pök $^{H}$ | *shoot* | ba$^{44}$ | ba̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | mbhɚ̠$^{2}$ | mbɪ$^{33}$ | <bɪ̠$^{4}$> |
| 77. | | *light* | dla$^{44}$ | ɖa̠$^{33}$ | da̠$^{33}$ | ba̠$^{33}$ | ba̠$^{33}$ | ɖɚ̠$^{2}$ | ɖɪ$^{33}$ | dɪ̠$^{4}$ |
| 78. | sik | *tree* | si$^{44}$ | si̠$^{33}$ | çi̠$^{33}$ | si̠$^{33}$ | si̠$^{33}$ | si̠$^{2}$ | sɪ$^{33}$ | si̠$^{4}$ |
| 79. | | *blossom* | vɪ$^{44}$ | vi̠$^{33}$ | vi̠$^{33}$ | vi̠$^{33}$ | vi̠$^{33}$ | vi̠$^{2}$ | vɪ$^{33}$ | vi̠$^{4}$ |
| 80. | | *bubble* | ti$^{44}$ | ti̠$^{33}$ | ti̠$^{33}$ | ti̠$^{33}$ | ti̠$^{33}$ | ti̠$^{2}$ | tɪ$^{33}$ | tsi̠$^{4}$ |
| 81. | tsat$^{H}$ | *break* | tshi$^{44}$ | tsi̠$^{33}$ | tçhi̠$^{33}$ | tçhi̠$^{33}$ | tshi̠$^{33}$ | tshi̠$^{2}$ | gɪ$^{33}$ | tçhi̠$^{4}$ |
| 82. | lak$^{L}$ | *hand* | le$^{2}$ | le̠$^{2}$ | le̠$^{2}$ | ʔli̠$^{2}$ | le̠$^{2}$ | la̠$^{55}$ | la$^{13}$ | lo̠$^{55}$ |
| 83. | wak$^{L}$ | *pig* | ve$^{2}$ | ve̠$^{2}$ | ve̠$^{2}$ | vi̠$^{2}$ | ve̠$^{2}$ | va̠$^{55}$ | va$^{13}$ | vo̠$^{55}$ |
| 84. | mak$^{L}$ | *soldier* | me$^{2}$ | me̠$^{2}$ | me̠$^{2}$ | ʔmi̠$^{2}$ | me̠$^{2}$ | ma̠$^{55}$ | ma$^{13}$ | mo̠$^{55}$ |
| 85. | rak$^{L}$ | *weave* | ze$^{2}$ | ze̠$^{2}$ | ze̠$^{2}$ | ? | ze̠$^{2}$ | ɣa$^{55}$ | ɣa$^{13}$ | dʑo̠$^{55}$ |
| 86. | C-kok$^{L}$ | *year* | kho$^{2}$ | khu̠$^{2}$ | khu̠$^{2}$ | kho̠$^{2}$ | khu̠$^{2}$ | kho̠$^{55}$ | tçhe$^{13}$ | khu̠$^{55}$ |
| 87. | lok$^{L}$ | *enough* | lo$^{2}$ | lu̠$^{2}$ | lu̠$^{2}$ | ʔlo̠$^{2}$ | lu̠$^{2}$ | lo̠$^{55}$ | le$^{13}$ | lu̠$^{55}$ |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 88. | C-krok$^{1\cdot}$ | *six* | kho$^{2}$ | tʂhu̱$^{2}$ | tʂhu̱$^{2}$ | kho̱$^{2}$ | tʂhu̱$^{2}$ | tʂho̱$^{55}$ | tɕhe$^{13}$ | fu̱$^{55}$ |
| 89. | ?nok$^{1\cdot}$ | *brain* | no$^{2}$ | nu̱$^{2}$ | nu̱$^{2}$ | ?no̱$^{2}$ | nu̱$^{2}$ | no̱$^{55}$ | ne$^{13}$ | no̱$^{55}$ |
| 90. | džok$^{1\cdot}$ | *waist* | dʑo$^{2}$ | dʑu̱$^{2}$ | dzu$^{2}$ | dʑo̱$^{2}$ | dʑu̱$^{2}$ | dʑo̱$^{55}$ | dʑe$^{13}$ | dzu̱$^{55}$ |
| 91. | ?dok$^{1\cdot}$ | *poison* | do$^{2}$ | du̱$^{2}$ | du̱$^{2}$ | do̱$^{2}$ | du̱$^{2}$ | <to̱$^{55}$> | <te$^{13}$> | du̱$^{55}$ |
| 92. | sap$^{11}$ | *rub* | va$^{2}$ | va̱$^{2}$ | va̱$^{2}$ | va̱$^{2}$ | va̱$^{2}$ | və˞$^{55}$ | vɪ$^{13}$ | vo̱$^{55}$ |
| 93. | m-lyak$^{1\cdot}$ | *lick* | ɗa$^{2}$ | ɗa̱$^{2}$ | la̱$^{2}$ | la̱$^{2}$ | la̱$^{2}$ | lə˞$^{55}$ | lɪ$^{13}$ | zo̱$^{55}$ |
| 94. | zik$^{1\cdot}$ | *leopard* | zi$^{2}$ | zɿ$^{2}$ | zɿ$^{2}$ | zɿ$^{2}$ | zɿ$^{2}$ | zɿ$^{55}$ | zi$^{13}$ | zɿ$^{55}$ |
| 95. | C-pat$^{1\cdot}$ | *spit vomit* | phɪ$^{2}$ | phɿ$^{2}$ | phɿ$^{2}$ | phɿ$^{2}$ | phɿ$^{2}$ | phɿ$^{55}$ | phi$^{13}$ | phɿ$^{55}$ |
| 96. | k-r-wat$^{1}$ | *leech* | vɪ$^{2}$ | vɿ$^{2}$ | vɿ$^{2}$ | vɿ$^{2}$ | bɿ$^{2}$ | vɿ$^{55}$ | vi$^{13}$ | mbɿ$^{55}$ |
| 97. | g-rap$^{1\cdot}$ | *needle* | ɣɤ$^{2}$ | ɣo̱$^{2}$ | ɣɤ̱$^{2}$ | zy$^{2}$ | vɤ̱$^{2}$ | ɣɤ̱$^{55}$ | ɣi$^{13}$ | zɿ$^{55}$ |
| 98. |  | *hoof* | bɤ$^{2}$ | bo̱$^{2}$ | zɤ̱$^{2}$ | by$^{2}$ | bɤ̱$^{2}$ | bɤ̱$^{55}$ | bi$^{13}$ | bɿ$^{55}$ |
| 99. |  | *shell* | phɤ$^{2}$ | pho̱$^{2}$ | phɤ̱$^{2}$ | phy$^{2}$ | phɤ̱$^{2}$ | phɤ̱$^{55}$ | phi$^{13}$ | ? |
| 100. |  | *put on* | dɤ$^{2}$ | do̱$^{2}$ | dɤ̱$^{2}$ | dy$^{2}$ | dɤ̱$^{2}$ | de̱$^{55}$ | di$^{13}$ | ndɿ$^{55}$ |
| 101. | ?rap$^{1\cdot}$ | *stand* | hɤ$^{2}$ | xo̱$^{2}$ | xɤ̱$^{2}$ | hy$^{2}$ | xɤ̱$^{2}$ | he̱$^{55}$ | hi$^{13}$ | hɿ$^{55}$ |
| 102. | s-mi$^{1}$ | *catch* | da$^{21}$ | do$^{21}$ | do$^{21}$ | da$^{21}$ | do$^{21}$ | do$^{55}$ | du$^{55}$ | ? |
| 193. | ba$^{2}$ | *thin* | ba$^{21}$ | bo$^{21}$ | bo$^{33}$ | ba$^{21}$ | bo$^{21}$ | bo$^{33}$ | bu$^{33}$ | bo$^{33}$ |
| 104. |  | *snow* | va$^{21}$ | ɣo$^{21}$ | ɣo$^{33}$ | va$^{21}$ | ɣo$^{21}$ | vo$^{33}$ | ɣu$^{33}$ | vo$^{33}$ |
| 105. |  | *plough* | ŋa$^{21}$ | ŋo$^{21}$ | ŋo$^{33}$ | ma$^{21}$ | mo$^{21}$ | ŋo$^{33}$ | ? | mo$^{33}$ |
| 106. | m-ba$^{3}$ | *bright* | ba$^{33}$ | bo$^{2}$ | bo$^{21}$ | ba$^{55}$ | bo$^{33}$ | bo$^{21}$ | bo$^{21}$ | bo$^{33}$ |
| 107. |  | *nephew* | tu$^{33}$ | to$^{33}$ | du$^{55}$ | du̱$^{55}$ | du$^{33}$ | ndu$^{33}$ | ndu$^{13}$ | ndu$^{33}$ |
| 108. | ?plu$^{1}$ | *porcupine* | pu$^{33}$ | po$^{33}$ | pu$^{55}$ | ku̱$^{55}$ | pu$^{33}$ | pu$^{33}$ | pu$^{13}$ | pu$^{33}$ |
| 109. | tu$^{1}$ | *thick* | thu$^{33}$ | tho$^{2}$ | thu$^{21}$ | thu̱$^{55}$ | thu$^{33}$ | thu$^{21}$ | thu$^{21}$ | tu$^{33}$ |
| 110. | < Ch. | *chopsticks* | dʑi$^{33}$ | dʑo$^{2}$ | dzu$^{21}$ | dzu̱$^{55}$ | dʑu$^{33}$ | dʑu$^{21}$ | dʑu$^{21}$ | dʑu$^{33}$ |
| 111. | mraŋ$^{2}$ | *horse* | mu$^{21}$ | mo$^{21}$ | mo$^{33}$ | mu̱$^{21}$ | mu$^{21}$ | mu$^{33}$ | mu$^{33}$ | mu$^{33}$ |
| 112. | C-nu$^{2}$ | *soft* | no$^{21}$ | no$^{21}$ | nu$^{33}$ | nu̱$^{21}$ | nu$^{21}$ | nu$^{33}$ | nu$^{33}$ | nu$^{33}$ |
| 113. | mo$^{2}$ | *sky* | mu$^{21}$ | mu$^{21}$ | mu$^{33}$ | mi$^{21}$ | mɯ$^{21}$ | mu$^{33}$ | mi$^{33}$ | mu$^{33}$ |
| 114. | bi/bo$^{2}$ | *insect* | bu$^{21}$ | bu$^{21}$ | bu$^{33}$ | <vi$^{21}$> | bu$^{21}$ | bu$^{33}$ | bi$^{33}$ | bu$^{33}$ |
| 115. | po$^{2}$ | *price* | phu$^{21}$ | phu$^{21}$ | phu$^{33}$ | fi$^{21}$ | phɯ$^{21}$ | phu$^{33}$ | phi$^{33}$ | phu$^{33}$ |
| 116. | po$^{2}$ | *carry (on back)* | bu$^{21}$ | bu$^{21}$ | bu$^{21}$ | <vi$^{21}$> | bɯ$^{21}$ | bu$^{55}$ | bi$^{55}$ | <pu$^{21}$> |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 117. | s-mo$^{1}$ | *mushroom* | mu$^{33}$ | mo$^{33}$ | mu$^{55}$ | mi$^{55}$ | mɯ$^{33}$ | mu$^{33}$ | mi$^{13}$ | m̩u$^{33}$ |
| 118. | ak$^{1}$/a$^{3}$ | *crow* | pu$^{33}$ | pu$^{2}$ | bu$^{21}$ | vi$^{55}$ | bɯ$^{33}$ | mbu$^{21}$ | mbi$^{21}$ | ku$^{33}$ |
| 119. | | *excrement* | ɬi$^{21}$ | ʈhi$^{21}$ | thi$^{33}$ | tɕhi$^{21}$ | si$^{21}$ | ɬi$^{33}$ | ʈhi$^{33}$ | tɕhi$^{33}$ |
| 120. | ʃi$^{2}$ | *urine* | zi$^{21}$ | zi$^{21}$ | ɕi$^{33}$ | ʐi$^{21}$ | zi$^{21}$ | zi$^{33}$ | zi$^{33}$ | zi$^{33}$ |
| 121. | (ʔ)ne$^{3}$ | *day* | ni$^{33}$ | ni$^{2}$ | ȵi$^{21}$ | ʔni$^{55}$ | ȵi$^{33}$ | ȵi$^{21}$ | ȵi$^{21}$ | ȵi$^{33}$ |
| 122. | gre$^{2}$ | *copper* | dʐi$^{21}$ | dʐi$^{21}$ | dʐi$^{33}$ | gɯ$^{21}$ | <dzi$^{21}$> | dʑi$^{33}$ | dʑi$^{33}$ | dʑi$^{33}$ |
| 123. | | *dung* | tʂhi$^{21}$ | tɕhi$^{21}$ | tɕhi$^{33}$ | khɯ$^{21}$ | tshi$^{21}$ | tɕhi$^{33}$ | tɕhi$^{33}$ | ɕi$^{33}$ |
| 124. | kyo$^{1}$ | *sweet* | tʂhi$^{33}$ | tʂhi$^{2}$ | tshi$^{21}$ | tʂhi$^{55}$ | tʂhɿ$^{33}$ | tʂhɿ$^{21}$ | tʂhy$^{21}$ | tɕhi$^{33}$ |
| 125. | swe$^{2}$ | *blood* | si$^{21}$ | si$^{21}$ | si$^{33}$ | si$^{21}$ | sɿ$^{21}$ | sɯ$^{33}$ | sy$^{33}$ | si$^{33}$ |
| 126. | ʔnip$^{1}$ | *squeeze* | tshi$^{21}$ | tshi$^{21}$ | tshi$^{21}$ | tshi$^{21}$ | tshɿ$^{21}$ | tshɯ$^{55}$ | tshy$^{55}$ | tshi$^{21}$ |
| 127. | dzi$^{2}$ | *ride* | ? | dza$^{21}$ | dzɛ$^{33}$ | dze$^{21}$ | dzɛ$^{21}$ | dzɔ$^{33}$ | dzɛ$^{33}$ | dzi$^{33}$ |
| 128. | s-rwe$^{1}$ | *yellow* | ʂi$^{33}$ | ʂa$^{33}$ | ɕɛ$^{55}$ | xe$^{55}$ | sɛ$^{33}$ | ʂɔ$^{33}$ | ʂɛ$^{13}$ | ʂi$^{33}$ |
| 129. | tsi$^{1}$ | *grease* | tshæ$^{33}$ | tsha$^{2}$ | tshɛ$^{21}$ | tshe$^{55}$ | tshɛ$^{33}$ | tshɔ$^{21}$ | tshɛ$^{21}$ | tshi$^{33}$ |
| 130. | ko$^{2}$ | *smoke* | khɯ$^{21}$ | khɯ$^{21}$ | khɯ$^{33}$ | khụ$^{21}$ | khɯ$^{21}$ | khɯ$^{33}$ | tɕhy$^{33}$ | ku$^{33}$ |
| 131. | ŋo$^{1}$ | *cry* | ŋɯ$^{33}$ | ŋɯ$^{33}$ | ŋɯ$^{55}$ | ŋụ$^{55}$ | ŋɯ$^{33}$ | ŋɯ$^{33}$ | ny$^{13}$ | ŋo$^{33}$ |
| 132. | go$^{2}$ | *nine* | kɯ$^{55}$ | kɯ$^{55}$ | kɯ$^{33}$ | kụ$^{21}$ | kɯ$^{55}$ | kɯ$^{33}$ | tɕy$^{33}$ | gu$^{33}$ |
| 133. | can$^{1}$ | *paddy* | tɕhɪ$^{33}$ | tɕhi$^{2}$ | tɕhe$^{21}$ | tɕhi$^{55}$ | tɕhe$^{33}$ | tʂhe$^{21}$ | tʂhi$^{21}$ | tʂhɯ$^{33}$ |
| 134. | san$^{1}$/$^{2}$ | *louse* | ɕɪ$^{33}$ | ɕi$^{2}$ | ɕe$^{21}$ | ɕi$^{55}$ | ɕe$^{33}$ | ɕe$^{21}$ | ɕi$^{21}$ | ʂɯ$^{33}$ |
| 135. | | *livestock* | dʑɪ$^{21}$ | dʑi$^{21}$ | dʑe$^{33}$ | dʑi$^{21}$ | dze$^{21}$ | dʐe$^{33}$ | dʐɤ$^{33}$ | dʐɯ$^{33}$ |
| 136. | | *be left* | tsɪ$^{33}$ | tsi$^{2}$ | dze$^{21}$ | dzi$^{55}$ | dze$^{33}$ | dze$^{21}$ | dzi$^{21}$ | dzɯ$^{33}$ |
| 137. | ʔ-kyin$^{1}$ | *sour* | tɕɪ$^{33}$ | tɕi$^{33}$ | tɕe$^{55}$ | tɕi$^{55}$ | tse$^{33}$ | tʂe$^{33}$ | tʂi$^{13}$ | tɕɪ$^{33}$ |
| 138. | | *frost* | ŋɪ$^{33}$ | ni$^{33}$ | ne$^{55}$ | ? | ne$^{33}$ | ȵe$^{33}$ | ȵi$^{13}$ | ? |
| 139. | s-r-way | *lead* | si$^{33}$ | sɛ$^{2}$ | sɛ$^{21}$ | si$^{55}$ | si$^{33}$ | sɿ$^{21}$ | sɛ$^{21}$ | sɪ$^{33}$ |
| 140. | r-miŋ$^{1}$ | *name* | mæ$^{33}$ | mɛ$^{33}$ | mɛ$^{55}$ | ʔmɯ$^{55}$ | mi$^{33}$ | mɚ$^{33}$ | mɛ$^{13}$ | m̩ɪ$^{33}$ |
| 141. | C-dim$^{1}$ | *cloud* | tæ$^{33}$ | tɛ$^{33}$ | tɛ$^{55}$ | tɯ$^{55}$ | ti$^{33}$ | tɚ$^{33}$ | tɛ$^{13}$ | tɪ$^{33}$ |
| 142. | s-miŋ$^{3}$ | *ripe* | mæ$^{33}$ | mɛ$^{2}$ | mɛ$^{21}$ | ʔmɯ$^{55}$ | mi$^{33}$ | mɚ$^{21}$ | mɛ$^{21}$ | m̩ɪ$^{33}$ |
| 143. | s/m-riŋ$^{1}$ | *long* | ɕæ$^{33}$ | ? | sɛ$^{55}$ | ʂɯ$^{55}$ | ʂi$^{33}$ | ʂɚ$^{33}$ | ʂɛ$^{13}$ | ʂo$^{33}$ |
| 144. | g-rap$^{1}$ | *thread* | tɕhæ$^{33}$ | tʂhɛ$^{2}$ | tshɛ$^{21}$ | khɯ$^{55}$ | tshi$^{33}$ | tɕɚ$^{21}$ | khɛ$^{21}$ | ɕɪ$^{33}$ |
| 145. | | *wake up* | tɕhæ$^{21}$ | ? | tshɛ$^{33}$ | gɯ$^{21}$ | tshi$^{21}$ | tɕhɚ$^{33}$ | gɛ$^{33}$ | dʐɪ$^{33}$ |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 146. | dzam$^{1}$ | *bridge* | tsɤ$^{33}$ | tsi$^{2}$ | dzɤ$^{21}$ | dzy$^{55}$ | dzɤ$^{33}$ | ndzhe$^{21}$ | dzɛ$^{21}$ | dzɿ$^{33}$ |
| 147. | ʔ-nam$^{2}$ | *smell* | nɤ$^{21}$ | mɯ$^{21}$ | nɤ$^{33}$ | ny$^{21}$ | nɤ$^{21}$ | ne$^{33}$ | ȵɛ$^{33}$ | ni$^{33}$ |
| 148. | xam$^{1}$ | *iron* | xɤ$^{33}$ | xɯ$^{2}$ | ɕe$^{21}$ | ɕy$^{55}$ | xɤ$^{33}$ | ɕe$^{21}$ | ɕɛ$^{21}$ | ʂi$^{33}$ |
| 149. | (s-)myak | *eye* | ne$^{44}$ | ne̱$^{33}$ | ne̱$^{33}$ | ʔmi̱$^{33}$ | me̱$^{33}$ | na̱$^{2}$ | na$^{33}$ | ȵo̱$^{4}$ |
| 150. | C-sak$^{1}$ | *air* | se$^{2}$ | se̱$^{2}$ | ɕe̱$^{2}$ | ɕi̱$^{2}$ | se̱$^{2}$ | sa̱$^{55}$ | sa$^{13}$ | so̱$^{55}$ |
| 151. |  | *link up* | tse$^{2}$ | tse̱$^{2}$ | tɕe̱$^{2}$ | tɕi̱$^{2}$ | ? | tsa̱$^{55}$ | tsa$^{13}$ | tso̱$^{55}$ |
| 152. | Ntsak | *drip* | dze$^{44}$ | dzɛ̱$^{33}$ | the̱$^{33}$ | dzi̱$^{33}$ | dze̱$^{33}$ | ndzha̱$^{2}$ | ndza$^{33}$ | ? |
| 153. | dak | *cockspur* | de$^{2}$ | de̱$^{2}$ | de̱$^{2}$ | di̱$^{2}$ | de̱$^{2}$ | da̱$^{55}$ | da$^{13}$ | do̱$^{55}$ |
| 154. |  | *lock (v.)* | dʑo$^{44}$ | dʑu̱$^{33}$ | dzu̱$^{33}$ | ? | dʑu̱$^{33}$ | ɳdʑo̱$^{2}$ | ɳdʑe$^{33}$ | ɳdʑu̱$^{4}$ |
| 155. | tok | *peck* | tho$^{44}$ | thu̱$^{33}$ | thu̱$^{33}$ | tho̱$^{33}$ | thu̱$^{33}$ | tho̱$^{2}$ | ? | thu̱$^{4}$ |
| 156. | pup$^{11}$ | *turn over* | po$^{44}$ | pu̱$^{33}$ | pu̱$^{33}$ | pho̱$^{33}$ | ? | pho̱$^{2}$ | pe$^{33}$ | phu̱$^{4}$ |
| 157. | ʔkrok | *frighten* | ? | tʂu̱$^{33}$ | tsu̱$^{33}$ | ko$^{33}$ | ku̱$^{33}$ | tɕo̱$^{2}$ | tɕe$^{33}$ | ku̱$^{4}$ |
| 158. | sök | *scrape* | ? | tʂu̱$^{33}$ | tsu̱$^{33}$ | <ko$^{33}$> | tʂu̱$^{33}$ | tʂo̱$^{2}$ | tʂe$^{33}$ | <ku̱$^{4}$> |
| 159. |  | *herd* | ɬo$^{2}$ | lu̱$^{2}$ | ɬu̱$^{2}$ | ʔlo̱$^{2}$ | lu̱$^{2}$ | ɬo̱$^{55}$ | ɬe$^{13}$ | ɬu̱$^{55}$ |
| 160. | $^{1}$duk | *kindle/light (v)* | tɤ$^{2}$ | to̱$^{2}$ | tu̱$^{2}$ | to̱$^{2}$ | tu̱$^{2}$ | to̱$^{55}$ | te$^{13}$ | tu̱$^{55}$ |
| 161. | myok$^{1}$ | *monkey* | no$^{2}$ | nu̱$^{2}$ | nu̱$^{2}$ | mo̱$^{2}$ | mu̱$^{2}$ | ɳo̱$^{55}$ | ɳe$^{13}$ | ȵu̱$^{55}$ |
| 162. | C-mruk$^{1}$ | *weeding* | ŋo$^{2}$ | ŋu̱$^{2}$ | ŋu̱$^{2}$ | mo̱$^{2}$ | ? | ŋo̱$^{55}$ | ? | mu̱$^{55}$ |
| 163. |  | *lack* | qha$^{2}$ | kha̱$^{2}$ | kha̱$^{2}$ | kha̱$^{2}$ | kha̱$^{2}$ | khə˞$^{55}$ | khɿ$^{13}$ | tɕho̱$^{55}$ |
| 164. |  | *paste* | na$^{2}$ | na̱$^{2}$ | na̱$^{2}$ | ʔna̱$^{2}$ | na̱$^{2}$ | nə˞$^{55}$ | ? | ȵo̱$^{55}$ |
| 165. |  | *bandit* | dza$^{44}$ | dza̱$^{33}$ | dza̱$^{33}$ | ? | dza̱$^{33}$ | dzə˞$^{2}$ | dzɿ$^{33}$ | dzo̱$^{4}$ |
| 166. | si$^{2}$ | *know* | sa$^{2}$ | sa̱$^{2}$ | sa̱$^{2}$ | sa̱$^{2}$ | sa̱$^{2}$ | sə˞$^{55}$ | se$^{13}$ | si̱$^{55}$ |
| 167. | ʔyip$^{1}$ | *sleep* | zi$^{2}$ | zi̱$^{2}$ | zi̱$^{2}$ | ʔi̱$^{2}$ | zi̱$^{2}$ | zi$^{55}$ | zi$^{13}$ | i̱$^{55}$ |
| 168. | {C}-tsit$^{11}$ | *pinch* | tshi$^{2}$ | tshi̱$^{2}$ | tshi̱$^{2}$ | tɕhi̱$^{2}$ | tshi̱$^{2}$ | tshi̱$^{55}$ | tshi$^{13}$ | tshi̱$^{55}$ |
| 169. | C-mwat$^{1}$ | *hungry* | ni$^{2}$ | mi̱$^{2}$ | ȵi̱$^{2}$ | mi̱$^{2}$ | mi$^{2}$ | ȵi̱$^{55}$ | ŋi$^{13}$ | mi̱$^{55}$ |
| 170. | C-sip$^{1}$ | *thirsty* | si$^{2}$ | si̱$^{2}$ | si̱$^{2}$ | ɕi̱$^{2}$ | si̱$^{2}$ | si̱$^{55}$ | si$^{13}$ | si̱$^{55}$ |
| 171. | C-šik$^{1}$ | *new* | ɕi$^{2}$ | ɕi̱$^{2}$ | ɕi̱$^{2}$ | xɯ̱$^{2}$ | xi̱$^{2}$ | ɕi̱$^{55}$ | xi$^{13}$ | ʂi̱$^{55}$ |
| 172. | Ntsit$^{1}$ | *split* | bɿ$^{2}$ | bi̱$^{2}$ | bi̱$^{2}$ | bi̱$^{2}$ | bi̱$^{2}$ | bi̱$^{55}$ | bi$^{13}$ | bɿ̱$^{55}$ |
| 173. | ʔwik$^{1}$ | *stomach* | hɿ$^{2}$ | xi̱$^{2}$ | xi̱$^{2}$ | hi̱$^{2}$ | xi̱$^{2}$ | hi̱$^{55}$ | hi$^{13}$ | hi̱$^{55}$ |
| 174. | s-nik $^{1}$ | *heart* | ni$^{44}$ | ȵi̱$^{33}$ | ȵi̱$^{33}$ | ? | ȵi̱$^{33}$ | ni̱$^{2}$ | nɿ$^{33}$ | ? |

## Appendix 3.3 Data from Eight Northern Loloish (Yi) dialects (Chen Kang 1986)

| No. | Etymon | Gloss | sani | axi | nesu | lalo | lipho | nasu | neisu | nosu |
|---|---|---|---|---|---|---|---|---|---|---|
| 175. | s-mit | *blink* | tshi$^{44}$ | tshʅ$^{33}$ | tɕhʅ$^{33}$ | tshʅ$^{33}$ | ? | tshʅ$^{2}$ | tshi$^{33}$ | tshʅ$^{4}$ |
| 176. | V-ci:t$^{L}$ | *goat* | tɕhi$^{2}$ | tɕhʅ$^{2}$ | tɕhʅ$^{2}$ | tʂhʅ$^{2}$ | tʂhɿ$^{2}$ | tʂhɿ$^{55}$ | tʂhi$^{13}$ | tʂhʅ$^{55}$ |
| 177. | ʔ(c/j)up$^{L}$ | *suck* | ? | tʂʅ$^{2}$ | tsʅ$^{2}$ | tʂʅ$^{2}$ | tʂɿ$^{2}$ | tʂɿ$^{55}$ | tʂɤ$^{13}$ | tɕʅ$^{55}$ |
| 178. | s-ga | *saddle* | ɣa$^{44}$ | ɣo̱$^{33}$ | ɣo̱$^{33}$ | ho̱$^{33}$ | xo̱$^{33}$ | ɣo̱$^{2}$ | ɣo̱$^{33}$ | ɣa̱$^{4}$ |
| 179. | C-ma$^{3}$ | *female* | ma$^{44}$ | mo̱$^{33}$ | mo̱$^{33}$ | mo̱$^{33}$ | mo̱$^{33}$ | mo̱$^{2}$ | mo̱$^{33}$ | ma̱$^{4}$ |
| 180. | ʔpa$^{1}$≍ʔpa$^{3}$ | *exchange* | pa$^{44}$ | po̱$^{33}$ | po̱$^{33}$ | po̱$^{33}$ | po̱$^{33}$ | ɬo̱$^{2}$ | ɬo̱$^{33}$ | pa̱$^{4}$ |
| 181. |  | *obtain* | ɣa$^{44}$ | ɣo̱$^{33}$ | ɣo̱$^{33}$ | ɣo̱$^{33}$ | ɣo̱$^{33}$ | ɣo̱$^{2}$ | ɣo̱$^{33}$ | ɣɯ̱$^{4}$ |
| 182. | loŋ$^{2/1}$ | *rabbit* | ɬa$^{44}$ | lo̱$^{33}$ | ɬo̱$^{33}$ | lo̱$^{33}$ | lo̱$^{33}$ | ɬo̱$^{2}$ | ɬo̱$^{33}$ | lɯ̱$^{4}$ |
| 183. |  | *brood* | tsa$^{44}$ | tso̱$^{33}$ | tso̱$^{33}$ | tso̱$^{33}$ | tso̱$^{33}$ | tso̱$^{2}$ | ? | ? |
| 184. |  | *shine upon* | ɬɤ$^{2}$ | li$^{2}$ | ɬɤ̱$^{2}$ | ʔly$^{2}$ | lɤ̱$^{2}$ | ɬɤ̱$^{55}$ | ɬi$^{13}$ | ɬʅ$^{55}$ |
| 185. | {C}-tsat$^{L}$ | *bite* | qhɤ$^{2}$ | tʂho̱$^{2}$ | tshɤ̱$^{2}$ | kho$^{2}$ | khɤ̱$^{2}$ | khɤ̱$^{55}$ | khi$^{13}$ | ɕʅ$^{55}$ |
| 186. | {C}-pwap$^{L}$ | *swell* | phɤ$^{2}$ | pho̱$^{2}$ | phɤ̱$^{2}$ | phy$^{2}$ | phɤ̱$^{2}$ | phɤ̱$^{55}$ | phi$^{13}$ | ? |
| 187. |  | *prickly ash (tree)* | dzɤ$^{2}$ | dzo̱$^{2}$ | dzɤ̱$^{2}$ | dzy$^{2}$ | dzɤ̱$^{2}$ | dze̱$^{55}$ | dzi$^{13}$ | dzʅ$^{55}$ |
| 188. |  | *drill* | lɤ$^{2}$ | ɬi$^{2}$ | lɤ̱$^{2}$ | ʔly$^{2}$ | lɤ̱$^{2}$ | lɤ̱$^{55}$ | ? | ? |
| 189. |  | *burn grass on waste land* | dɤ$^{2}$ | do̱$^{2}$ | dɤ̱$^{2}$ | <ty$^{2}$> | dɤ̱$^{2}$ | de̱$^{55}$ | ? | ndi$^{55}$ |
| 190. |  | *bear (fruit)* | dɤ$^{2}$ | do̱$^{2}$ | dɤ̱$^{2}$ | ? | ? | de$^{55}$ | di$^{13}$ | ndi$^{55}$ |
| 191. |  | *family* | ɣɤ$^{2}$ | ɣo̱$^{2}$ | ɣɤ̱$^{2}$ | ? | vɤ̱$^{2}$ | ɣɤ̱$^{55}$ | zi$^{13}$ | zɿ$^{55}$ |
| 192. | C-ɲap$^{L}$ | *press from both sides* | ȵɤ̱$^{44}$ | no̱$^{33}$ | nɤ̱$^{33}$ | ? | nɤ̱$^{33}$ | ȵi$^{2}$ | ȵi$^{33}$ | ȵʅ$^{4}$ |
| 193. |  | *come apart* | ɬɤ$^{44}$ | lo̱$^{33}$ | lɤ̱$^{33}$ | ? | lɤ̱$^{33}$ | ɬɤ̱$^{2}$ | ɬi$^{33}$ | lʅ$^{4}$ |
| 194. | ʔtap | *scoop up* | tɤ$^{44}$ | to̱$^{33}$ | tɤ̱$^{33}$ | ty$^{33}$ | tɤ̱$^{33}$ | tɤ̱$^{2}$ | ti$^{33}$ | ? |
| 195. | kyap | *stick in* | tʂhɤ$^{44}$ | tʂho̱$^{33}$ | tshɤ̱$^{33}$ | ? | tʂhɤ̱$^{33}$ | tʂhɤ̱$^{2}$ | tʂhi$^{33}$ | ? |
| 196. |  | *layer* | tɤ$^{44}$ | to̱$^{33}$ | tɤ̱$^{33}$ | ? | ? | tɤ̱$^{2}$ | ti$^{33}$ | tʅ$^{4}$ |
| 197. | ra$^{3}$ | *can* | kɯ$^{2}$ | kɯ̱$^{2}$ | kɯ̱$^{2}$ | ʔʅ$^{2}$ | kɯ̱$^{2}$ | kɯ̱$^{55}$ | kɤ$^{13}$ | kɯ̱$^{55}$ |
| 198. |  | *hold (between fingers)* | tshi$^{44}$ | tshʅ$^{33}$ | tɯ̱$^{33}$ | tshʅ$^{33}$ | tshʅ$^{33}$ | tshɯ̱$^{2}$ | tshɤ$^{33}$ | tshi$^{4}$ |
| 199. |  | *take off* | ɬɯ$^{2}$ | lɯ̱$^{2}$ | ɬɯ̱$^{2}$ | ʔlʅ$^{2}$ | lɯ̱$^{2}$ | lɯ̱$^{55}$ | lɤ$^{13}$ | lʅ$^{55}$ |
| 200. | s-mut$^{H}$ | *blow* | mu$^{44}$ | mu̱$^{33}$ | mu̱$^{33}$ | ʔmu̱$^{33}$ | mɯ̱$^{33}$ | mu̱$^{2}$ | mu$^{33}$ | m̥u$^{4}$ |
| 201. |  | *Buddha* | bu$^{44}$ | bu̱$^{33}$ | bu̱$^{33}$ | ʔbu̱$^{33}$ | bɯ̱$^{33}$ | bu̱$^{2}$ | bu$^{33}$ | bu$^{4}$ |
| 202. |  | *jar* | bu$^{2}$ | bu̱$^{2}$ | bu̱$^{2}$ | bu̱$^{2}$ | bɯ̱$^{2}$ | bu̱$^{55}$ | ? | ? |

## Appendix 3.4 Sources of *Loloish reconstructions cited in previous table

| No. | Etymon | Source | No. | Etymon | Source |
|---|---|---|---|---|---|
| 1. | ʒa$^2$ | DB 206 | 31. | | |
| 2. | tsa$^2$ | JAM-TIL 9; DB 408 | 32. | pup | JAM-TSR 19 |
| 3. | xa$^2$ | DB 135 | 33. | | |
| 4. | | | 34. | tsu$^1$ | JAM-MLBM 31 |
| 5. | dza$^2$ | DB 629; JAM-MLBM 47 | 35. | C-gray$^1$ | DB 319 |
| 6. | k-ra$^2$ | JAM-DL p.1116 | 36. | | |
| 7. | (g)wa$^2$ | DB 635A | 37. | dza-n | TIL 8 (eat/food) |
| 8. | | | 38. | C-ŋa$^1$ | DB 438 |
| 9. | baŋ$^2$ | DB 573 | 39. | C-na$^1$ | DB 763S |
| 10. | | | 40. | C-ra$^1$ | DB 488 |
| 11. | ʔ-maŋ$^2$ | DB 311C | 41. | | |
| 12. | tsiy$^2$ | JAM-MLBM 34 | 42. | | |
| 13. | b-ni$^2$ | DB 751; JAM-GSTC 55 | 43. | doŋ$^1$ | DB 83 |
| 14. | C-mi$^2$ | DB 207 | 44. | | |
| 15. | zum$^2$ | STEDT 2810 | 45. | m-gaŋ$^1$ | DB 728A |
| 16. | ko$^2$ | DB 615 | 46. | ʒo$^1$ | DB 5 |
| 17. | ɣo$^2$ | AW-TBT 173 | 47. | ti(y) | STC 55 |
| 18. | jim$^2$ | JAM-MLBM 45 | 48. | ʔkuk/ʔguk | JAM-TSR 71 |
| 19. | ʔgyak | JAM-TSR 49 | 49. | siy$^1$ | JAM-DL p.1229 |
| 20. | maŋ$^2$ | DB 535 | 50. | | |
| 21. | gaŋ$^2$ | DB 546A | 51. | | |
| 22. | (k)-rwaŋ$^2$ | DB 604 | 52. | way$^1$ | DB 603 |
| 23. | ʔnit$^1$ | JAM-TSR 160(c) | 53. | s-ŋa$^2$ | DB 600 |
| 24. | ʔ-kri$^2$ | DB 552B | 54. | | |
| 25. | N-šit$^1$ | JAM-TSR 128(b) | 55. | | |
| 26. | | | 56. | m-ga$^2$ | DB 827A |
| 27. | ʔ-l(y)a$^1$ | DB 95 JAM-GSTC 56 | 57. | ʔduk | JAM-TSR 45a |
| 28. | ma$^1$ | DB 295B | 58. | (k-)la | STEDT 2389 |
| 29. | ʔ-na$^1$ | DB 667 | 59. | | |
| 30. | ʔ/g-raw$^1$ | DB 640 | 60. | s-ma$^2$ | DB 721 |

## Appendix 3.4 Sources of *Loloish reconstructions cited in previous table

| No. | Etymon | Source | No. | Etymon | Source |
|---|---|---|---|---|---|
| 61. | ʔ-krwe$^{2}$ | DB 151 | 91. | ʔdok$^{L}$ | JAM-TSR 113(b) |
| 62. | | | 92. | sap$^{H}$ | JAM-TSR 116 |
| 63. | k-rak$^{H}$ | JAM-TSR 184 | 93. | m-lyak$^{L}$ | JAM-TSR 179(a) |
| 64. | ʔtak$^{H}$ | JAM-TSR 54 | 94. | zik$^{L}$ | JAM-TSR 122 |
| 65. | | | 95. | C-pat$^{L}$ | JAM-TSR 38 |
| 66. | s-nak$^{H}$ | JAM-TSR 142; JAM-MLBM 79 | 96. | k-r-wat$^{L}$ | JAM-TSR 167 |
| 67. | kywan$^{1}$/$^{2}$/$^{3}$ | JAM-GSTC 9 | 97. | g-rap$^{L}$ | DB 382; TSR 191(a) |
| 68. | | | 98. | | |
| 69. | s-nök$^{H}$ | JAM-TSR 140 | 99. | (kroy) | JAM-GSTC 88 |
| 70. | m-bliŋ$^{3}$ | DB 547 | 100. | | |
| 71. | ʔdwak$^{L}$ | TSR 102(c) | 101. | ʔrap$^{L}$ | JAM-TSR 175 |
| 72. | | | 102. | s-mi$^{1}$ | DB 701 |
| 73. | Nkrok$^{H}$ | JAM-TSR 104(a) | 103. | ba$^{2}$ | DB 533A |
| 74. | k-lok$^{L}$ | DB 337A; JAM-TSR 190(a) | 104. | | |
| 75. | | | 105. | | |
| 76. | N-pök $^{H}$ | JAM-TSR 108(b) | 106. | m-ba$^{3}$ | DB 556 |
| 77. | | | 107. | | |
| 78. | sik | JAM-TSR 118(a) | 108. | ʔplu$^{1}$ | JAM-MLBM 25 |
| 79. | | | 109. | tu$^{1}$ | DB 531 |
| 80. | | | 110. | | |
| 81. | tsat$^{H}$ | JAM-TSR 40(a) | 111. | mraŋ$^{2}$ | DB 6 |
| 82. | lak$^{L}$ | JAM-TSR 166 | 112. | C-nu$^{2}$ | DB 528 |
| 83. | wak$^{1}$ | JAM-TSR 168 | 113. | mo$^{2}$ | DB 321 |
| 84. | mak$^{1}$ | JAM-TSR 135 | 114. | bi/bo$^{2}$ | DB 71 |
| 85. | rak$^{1}$ | JAM-TSR 192(a) | 115. | po$^{2}$ | DB 421 |
| 86. | C-kok$^{L}$ | DB 477B | 116. | po$^{2}$ | DB 661A |
| 87. | lok$^{L}$ | JAM-TSR 164 | 117. | s-mo$^{1}$ | DB 288 |
| 88. | C-krok$^{1}$ | JAM-MLBM 17 | 118. | ak$^{L}$/a$^{3}$ | DB 52 |
| 89. | ʔnok$^{L}$ | JAM-TSR 156(b) | 119. | | |
| 90. | džok$^{L}$ | JAM-TSR 6 | 120. | ʃi$^{2}$ | DB 149B |

## Appendix 3.4 Sources of *Loloish reconstructions cited in previous table

| No. | Etymon | Source |
|---|---|---|
| 121. | (ʔ)ne$^{3}$ | DB 461 |
| 122. | gre$^{2}$ | DB 404 M 18 |
| 123. | | |
| 124. | kyo$^{1}$ | DB 551 |
| 125. | swe$^{2}$ | DB 147 |
| 126. | ʔnip$^{1}$ | JAM-TSR 159(b) |
| 127. | dzi$^{2}$ | DB 651; JAM-MLBM 36 |
| 128. | s-rwe$^{1}$ | DB 506 |
| 129. | tsi$^{1}$ | DB 384 (fat) |
| 130. | ko$^{2}$ | DB 333; JAM-MLBM 11 |
| 131. | ŋo$^{1}$ | DB 670 |
| 132. | go$^{2}$ | DB 486 |
| 133. | can$^{1}$ | DB 280 |
| 134. | san$^{1}$/$^{2}$ | JAM-GSTC 7 |
| 135. | | |
| 136. | | |
| 137. | ʔ-kyin$^{1}$ | DB 549 |
| 138. | | |
| 139. | s-r-way | JAM-GSTC 121 |
| 140. | r-miŋ$^{1}$ | JAM-TIL 25 |
| 141. | C-dim$^{1}$ | DB 320-2 |
| 142. | s-miŋ$^{3}$ | DB 764B |
| 143. | s/m-riŋ$^{1}$ | DB 754 |
| 144. | g-rap$^{L}$ | JAM-TSR 191(b) (needle) |
| 145. | | |
| 146. | dzam$^{1}$ | DB 393 |
| 147. | ʔ-nam$^{2}$ | DB 513 |
| 148. | xam$^{1}$ | DB 403 |
| 149. | (s-)myak | JAM-TSR 145 |
| 150. | C-sak$^{1}$ | JAM-TSR 126 |
| 151. | | |
| 152. | Ntsak | JAM-TSR 82 |
| 153. | dak | STEDT 2003 |
| 154. | | |
| 155. | tok$^{H}$ | JAM-TSR 15 |
| 156. | pup$^{H}$ | JAM-TSR 19 |
| 157. | ʔkrok | JAM-TSR 54a |
| 158. | sök | JAM-TSR 117 |
| 159. | | |
| 160. | $^{L}$duk | JAM-TSR 62(a) |
| 161. | myok$^{L}$ | JAM-TSR 133 |
| 162. | C-mruk$^{L}$ | DB 621; TSR 138 |
| 163. | | |
| 164. | | |
| 165. | | |
| 166. | si$^{2}$ | DB 590 |
| 167. | ʔyip$^{L}$ | JAM-TSR 180(b) |
| 168. | {C}-tsit$^{H}$ | JAM-TSR 32 |
| 169. | C-mwat$^{L}$ | DB 637; JAM-TSR 132 |
| 170. | C-sip$^{L}$ | JAM-TSR 129 |
| 171. | C-šik$^{L}$ | JAM-TSR 126 |
| 172. | Ntsit$^{L}$ | JAM-TSR 88(b) |
| 173. | ʔwik$^{L}$ | JAM-TSR 176 |
| 174. | s-nik $^{L}$ | JAM-TSR 146(a) |
| 175. | s-mit | STEDT 139 |
| 176. | V-ci:t$^{L}$ | JAM-TSR 27 |
| 177. | ʔ(c/j)up$^{L}$ | JAM-TSR 73 |
| 178. | s-ga | JAM-TIL 60 |
| 179. | C-ma$^{3}$ | DB 174 |
| 180. | ʔpa$^{1}$≍ʔpa$^{3}$ | JAM-TIL 1 |

## Appendix 3.4 Sources of *Loloish reconstructions cited in previous table

| No. | Etymon | Source | No. | Etymon | Source |
|---|---|---|---|---|---|
| 181. | | | | | |
| 182. | loŋ$^2$/$^1$ | DB 46-2 | | | |
| 183. | | | | | |
| 184. | | | | | |
| 185. | {C}-tsat$^L$ | JAM-TSR 24 | | | |
| 186. | {C}-pwap$^L$ | JAM-TSR 92(b) | | | |
| 187. | | | | | |
| 188. | | | | | |
| 189. | | | | | |
| 190. | | | | | |
| 191. | | | | | |
| 192. | C-ɲap$^L$ | DB 738B | | | |
| 193. | | | | | |
| 194. | ʔtap | | | | |
| 195. | kyap | JAM-TSR 21(a) | | | |
| 196. | | | | | |
| 197. | ra$^3$ | DB 788 (barely) | | | |
| 198. | | | | | |
| 199. | | | | | |
| 200. | s-mut$^H$ | JAM-TSR 143 | | | |
| 201. | | | | | |
| 202. | | | | | |

## Appendix 4.1: Tamang languages of Nepal - Description

The Tamang group is part of the Bodic division of the Tibeto-Burman branch of the Sino-Tibetan family in Shafer's classification (Shafer 1955), spoken in Nepal (Mazaudon 1978).

The reconstructed ancestor, Proto Tamang-Gurung-Thakali-Manang, is abbreviated *TGTM.

Languages of the Tamang group are spoken by about one million people, mostly in the center of Nepal, with a recent extension toward the east. The speakers are divided into ethnic groups which by tradition do not intermarry.

### Phonological Typology

Four modern tones (numbered $^1$ to $^4$) are recognized in the modern languages and two proto-tone categories (labelled $^A$ and $^B$) are reconstructed. The tones of both reconstructed and daughter forms are transcribed before the syllable, e.g. $^A$bap.

The eight dialects used are discussed in detail in Mazaudon 1978. The dialects and their abbreviations are:

| | | | |
|---|---|---|---|
| Risiangku | (ris) | Sahu | (sahu) |
| Taglung | (tag) | Tukche | (tuk) |
| Marpha | (mar) | Syang | (syang) |
| Ghachok | (gha) | Prakaa | (pra) |

Some languages of the Tamang group have preserved, like written Tibetan, a phonological structure which permits final consonants and initial clusters in a syllable.

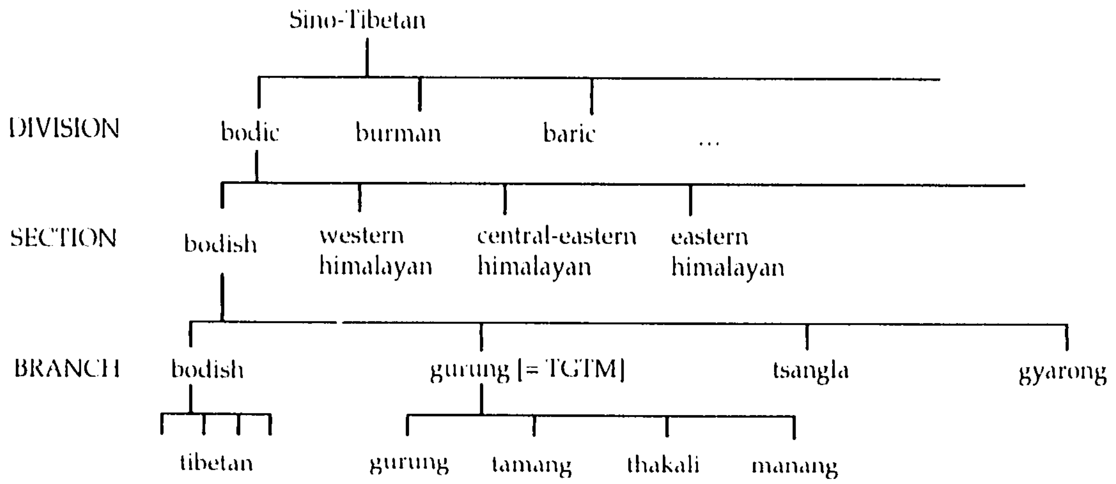
## Appendix 4.1: Tamang languages of Nepal - Description

Like the majority of Tibeto-Burman languages, the TGTM group are largely monosyllabic: monosyllables represent 98% of the verbal roots, and about 60% of the noun forms. Of the polysyllabic forms, many can be etymologized as compounds; others, because they vary greatly at the level of the common language reconstructed for several dialects, indicate that there may have been a significant number of doublets in this common language. This resembles the situation proposed for Proto-Tibeto-Burman by James Matisoff in the formation of compounds from a very reduced phonology.

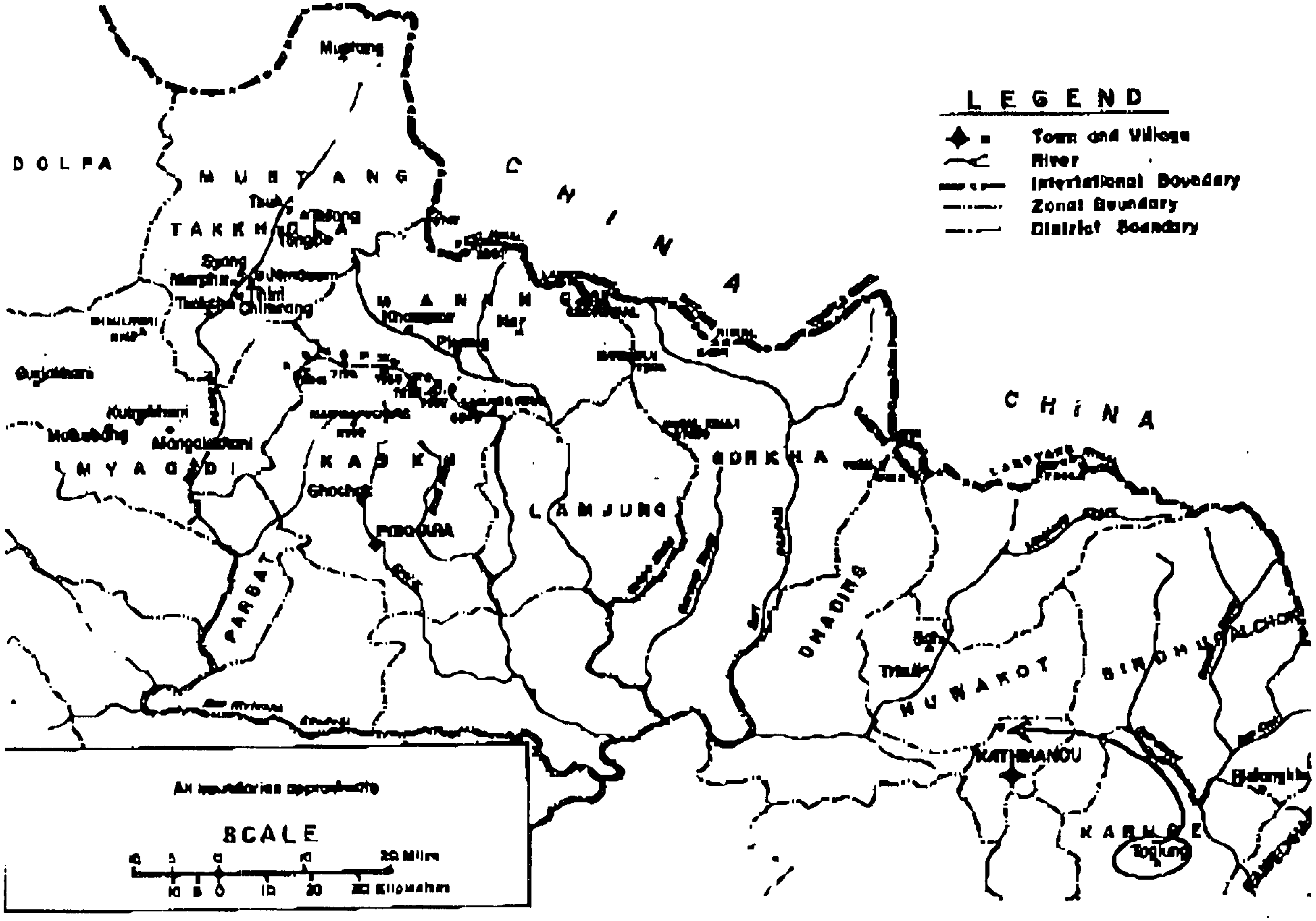
The data presented here are restricted primarily to monosyllabic morphemes.

## Appeindix 4.1 Tamang languages of Nepal - Classification

Shafer's internal classification of the dialects, resting on the ethnic denominations of the speakers and certain geographic considerations:



Appendix 4.2: The *TGTM area (Nepal) (from Mazaudon 1994)



# Appendix 5: NP-completeness, as it applies to cognate set conflation (§6)

A number of computational problems are known to be intractable in the sense that no ways have been found to reduce the number of steps required to solve them; also, these problems have the property that the number of steps required grows at a rapid rate as the number of elements in the problem increases.

Many problems have solutions in which the number of steps required is proportional to the number of items, or to the number of items to some specified power. The time required to solve these problems can therefore be expressed as a polynomial equation. Problems whose solution requires an amount of time that cannot be expressed as a polynomial are called non-polynomial complete, or NP-complete.

An interesting property of NP-complete problems is that solving any version of an NP-complete problem solves all NP-complete problems, and there are many. Below are a two problems which have this property.

**The Knapsack problem:**

Given a list of numbers and a "knapsack size," determine if some subset of the listed numbers adds up to the knapsack size.

If the list is 4, 9, 18, 25, 27, 42 and the knapsack size were 89, the answer is 'YES' because 4 + 18 + 25 +42 = 89. If the number were 90, the answer would be 'NO.'

(This problem requires that all combinations of numbers in the set be added together to see if they yield the knapsack number. There is no short way of doing this.)

## Set covering problem

This is the problem which is similar to the cognate set conflation problem described in §6.

For a given set, a collection of subsets is said to cover the given set if each member of the given set belongs to at least one set in the collection (this is certainly true of the distribution of reflexes in cognate sets).

The problem:

Given a set to be covered, a list of subsets, and a "cover size k," determine if k of the available subsets covers the given set. This is not exactly the problem of finding all the sets which are subsets of other sets, but it is related: in the cognate set conflation problem, the goal is to find the *smallest* value of k which covers the set.

from

## Appendix 6: *TB *s-na 'nose', a large cognate set from the STEDT database

*2 Tibeto-Burman*

| | | | |
|---|---|---|---|
| *Tibeto-Burman | s-na | | STC 101<br>ACST 521c |

*2.1 Kamarupan*

| | | | |
|---|---|---|---|
| Abor Miri | nyé-bung | | JAM-Ety |
| Ao Naga | $^{3}$tu$^{3}$niʔ | | AW-TBT 168 |
| | $^{3}$tu$^{3}$niʔ | | AW-TBT 186 |
| Apatani | $^{1}$ya$^{2}$piŋ($^{1}$) | | AW-TBT 610 |
| Atong | na-kuŋ | | JAM-Ety |
| Chinbok | hŋa-kɔŋ | | JAM-Ety |
| Darang | ha:nya:gom | | JAM-Ety |
| | həna-gam×hnya-gom | | STC 101 |
| | xa$^{31}$ nia$^{33}$ pum$^{55}$ | | JZ-DGdr |
| Gallong | ɟɪɤ•pum | | AW-TBT 610 |
| | ɲepum | | KDG-IGL |
| Geman | min$^{55}$ nioŋ$^{55}$ | | JZ-DGgm |
| Idu | enambo | | JP-Idu |
| | enambõ | | NEFA-PBI |
| | e$^{33}$ɳaŋ$^{55}$bo$^{33}$ | | SHK-Idu 3.5 |
| Khiamngan | $^{21}$ŋan | | AW-TBT 610 |
| Khumi | na-taran | | JAM-Ety |
| Konyak (Tamlu) | nagoŋ | | AW-TBT 186 |
| Lakher | hna | | JAM-Ety |
| Lhoba | ɳapum | | JZ-LBny |
| Liangmei | mai-nu-kuaŋ | | AW-TBT 186 |
| Lotha | Khenoe na okhe nandeng | *nose to fingertip* | WN-LothQ |
| Lotha Naga | kenno | | GEM-Loth 50 |
| | $^{3}$ke$^{3}$no | | AW-TBT 186 |
| | $^{3}$ke$^{3}$no(?) | | AW-TBT 168 |
| Lushai | hna-pa-su | | JAM-Ety |
| | hna-r | | STC 101 |
| Meithei | naton gi koy | *hair* | CYS-Meithei |
| | naton | | CYS-Meithei |
| | nâtôn | | JAM-Ety |

## Appendix 6: *TB *s-na 'nose', a large cognate set from the STEDT database

| | | | | |
|---|---|---|---|---|
| | | nəkhaŋ | *nose bridge* | CYS-Meithei |
| | Mikir | -nò•kàn | | KHG-Mikir 1 |
| | | nokan | | JAM-Ety |
| | Miri | ɲipum | | IMS-HMLG |
| | Monpa Cuona | nʌr$^{55}$ n̥ʌk$^{53}$ | | JZ-MBcn |
| | Monpa Cuona Wenlang | nɑ$^{55}$ | | JZ-CNwl |
| | Monpa Motuo | nawuŋ | | JZ-MBmt |
| | | nawuŋlaŋ | | JZ-MBmt |
| | Monpa Tilang | nauŋ | | JZ-CLtl |
| | Mru | na- | | JAM-Ety |
| | Rongmei | nû-kûaŋ | | AW-TBT 186 |
| | Sampang | nabu | | AW-TBT 168 |
| | Sangtam | $^{2}$na$^{1}$buŋ | | AW-TBT 168 |
| | | $^{2}$na$^{1}$buŋ | | AW-TBT 186 |
| | | $^{2}$na$^{1}$buŋ | | AW-TBT 610 |
| | Tagin | naŋ | | DG-Tag |
| | Tangkhul | nātāŋ | *nose, snout* | JAM-Ety |
| | Tiddim | nak | | JAM-Ety |
| | | nâ | | AW-TBT 168 |
| | | nâ | | AW-TBT 186 |
| | Tsangla=C. Monpa | nawung | | SER-HSL/T 3 |
| | Wancho | ne-kuŋ | | JAM-Ety |
| | Yimchunger | $^{1}$nɯ$^{2}$buŋ | | AW-TBT 186 |
| | | $^{1}$nɯ$^{2}$buŋ | | AW-TBT 610 |
| 2.2 | *Himalayish* | | | |
| | *Tamang | n$^{h}$a$^{1}$ | | MM-K78 12 |
| | Amdo | hnæ | | JS-Amdo 703 |
| | Bahing | noegatsi | | BM-Bah |
| | Bantawa | nabu | | WW-Bant 53 |
| | | nabuk | | AW-TBT 168 |
| | Batang | ŋa$^{55}$ | | DQ-Batang 3 |
| | Chamling | nadipung | | BM-PK7 134 |
| | | nadipü | | AW-TBT 168 |
| | | nañipɔ̃ | | AW-TBT 168 |
| | Chantyal | nak-nha-ri-wa rə̃w | *nose hair* | ChanQ 3 |

## Appendix 6: *TB *s-na 'nose', a large cognate set from the STEDT database

| | | | |
|---|---|---|---|
| | nak-nlia-ri-wa rõw | *nose hair* | ChanQ 3 |
| | nak-ye dāri | *bridge (of nose)* | NPB-ChanQ 3 |
| | nak/nakh | | NPB-ChanQ 3 |
| 2.3 *Lolo-Burmese* | | | |
| *Lolo-Burmese | ʔna$^{1/2}$ + koŋ$^{1/2}$ | | JAM-MLBM 61 |
| *Loloish | s-na$^{1}$ | | DB 93 |
| Achang Lianghe | na$^{31}$ khaŋ$^{55}$ | | JZ-AClh |
| Achang Longchuan | ni$^{31}$ xɔŋ$^{55}$ | | JZ-AClc |
| | ŋ̊ɔŋ$^{55}$ | | JZ-AClc |
| Achang Luxi | na$^{55}$ kaŋ35 | | RJL-DPTB 55 |
| | na$^{55}$ kaŋ$^{35}$ | | JZ-AClx |
| Ahi | no$^{33}$ bo̤$^{21}$ | | LMZ-AY 3.5 |
| Akha | naˇmehˇ(mehˇ) | | JAM-Ety |
| | naˇmehˇnaˇtsmˇ | *ridge of nose* | JAM-Ety |
| | námɛ́ | | AW-TBT 186 |
| Akha (Yunnan) | ná bàŋ | | IH-PL3 222 |
| Bisu | nakháŋ | | PB-Bisu 15 |
| | nákʰaŋ | | ZMYYC |
| Bola | na̤$^{55}$ | | DQ-Bola 103 |
| Burmese (Written) | hna | | JAM-Ety |
| | hna-khòŋ | | JAM-Ety |
| | hna-tam | *ridge of nose* | JAM-Ety |
| Dafang | nɔ$^{33}$ mo$^{55}$ | | DQ-Dafang |
| Dafang (Guizhou) | nɔ$^{33}$ mo$^{5}$ | | JZ-YIdf |
| Gazhuo | na$^{35}$ khy$^{33}$ | | DQ-Gazhuo 3 |
| Hani (Caiyuan) | na$^{55}$ me$^{55}$ | | JZ-HNcy |
| Hani (Dazai) | na$^{55}$ me$^{55}$ | | JZ-HNdz |
| Hani (Gelanghe) | na$^{55}$ mɛ$^{55}$ | | JZ-HNgl |
| Hani (Lüchun) | nDá mɛ́ | | IH-PL3 222 |
| Hani (Shuikui) | nɔ$^{55}$ mɛ$^{55}$ | | JZ-HNsk |
| Hani (Wordlist) | nalmeil | | IH-PL3 222 |
| Jinuo (A) | nɤ$^{31}$ to$^{44}$ | | DQ-JinA 106 |
| Jinuo (B) | ŋ̊o$^{31}$ tu$^{44}$ | | DQ-JinB 106 |
| Jinuo (Youle) | ŋ̊ɔ$^{42}$ to$^{44}$ | | JZ-JNyl |
| Khàtú | nDá tDù | | IH-PL3 222 |
| Lahu | nā qhɔ̂ | | JAM-MLBM 61 |

## Appendix 6: *TB *s-na 'nose', a large cognate set from the STEDT database

| | | | |
|---|---|---|---|
| | nā-qhɔ̂ | | JAM-Ety |
| Lahu (Common) | na̱ | | DB 93 |
| Lahu Na | na$^{31}$ q$^{h}$ɔ$^{33}$ | | JZ-LhNa |
| Lahu Xi | na$^{11}$ q$^{h}$ɔ$^{35}$ | | JZ-LhXi |
| Lalo | ʔnɑ$^{55}$ k$^{h}$ɯ$^{33}$ | | CK-Lalo 3.5 |
| Langsu | nɔ$^{31}$ | | DQ-Langsu 3 |
| Lashi | no$^{33}$ | | DQ-Lashi 3. |
| | no$^{33}$ mou$^{55}$ | *nose hair* | DQ-Lashi 3. |
| Lisu | na̱-bi | | JAM-Ety |
| | na̱-bwe | | JAM-Ety |
| | nɑ$^{44}$ k$^{h}$ɯ$^{44}$ | | JZ-Lisu |
| | na$^{3}$-bē$^{4}$ | | JAM-Ety |
| | na$^{3}$bē$^{4}$ | | DB 93 |
| | nâbī | | AW-TBT 186 |
| Lolopho | ny$^{44}$ gy$^{33}$ | | DQ-Lolopho |
| Maru | nʔò | | AW-TBT 186 |
| Mpi | ŋ$^{4}$khoŋ$^{6}$ | | JAM-MLBM 61 |
| Nasu | no$^{21}$ bi̥$^{21}$ | | CK-Nasu 3.5 |
| Naxi (Eastern) | ɲi$^{33}$gə$^{33}$ | | JZ-NaxiE |
| Naxi (Lijiang)) | ɲi$^{55}$məɹ$^{33}$ | | ZMYYC |
| Naxi (Western) | ɲi$^{55}$mər$^{31}$ | | JZ-NaxiW |
| Nesu | no$^{55}$ ko$^{21}$ | | CK-Nesu 3.5 |
| Nusu (A) | ŋa$^{55}$ kõ$^{35}$ | | DQ-NusuA 10 |
| Nusu (B) | ŋa$^{55}$ kɔ$^{33}$ | | DQ-NusuB 10 |
| Nusu (Central) | ŋɑ$^{55}$ kɑ̃$^{35}$ | | JZ-NUzl |
| Nusu (Northern) | nɑ$^{55}$ kə̃$^{53}$ | | JZ-NUwk |
| Nusu (Southern)) | ŋi$^{55}$ kɔ̃$^{31}$ | | JZ-NUgp |
| Nyiq | nɑ$^{44}$ bɪ$^{33}$ | | DQ-Nyiq 3.5 |
| Pijo | nÐá mí | | IH-PL3 222 |
| Sani (Maa) | nɑ$^{44}$ bɪ$^{33}$ | | DQ-SaniMa 3 |
| Sani (Wu) | nɑ$^{44}$ | | DQ-SaniWu 1 |
| Yi (Xide) | nɑ$^{55}$ k$^{h}$ɯ$^{33}$ | | JZ-YInj |
| | ŋa$^{21}$ bi$^{55}$ | | JZ-YIxd |
| | ŋa$^{21}$-bi$^{55}$ | | CSL-YIzd |
| Zaiwa | no$^{51}$ | | JZ-ZW |
| | nʔó | | AW-TBT 186 |

## Appendix 6: *TB *s-na 'nose', a large cognate set from the STEDT database

*2.4 Jinghpo-Nungish*

| | | |
|---|---|---|
| Dulong Dulonghe | sɯ$^{11}$ na$^{55}$ | RJL-DPTB 55 |
| | sɯ$^{11}$ nɑ$^{55}$ | JZ-DLdlh |

## Appendix 7 *Forms meaning* rainbow *from the STEDT database*

| | | | | | |
|---|---|---|---|---|---|
| *Loloish | ʃi$^{3}$ | DB 322 | Damu OY | dza | JS-TANI |
| Achang | xɔŋ$^{51}$tɕin$^{31}$nam$^{31}$ | ZMYYC 16 | Darang (=Taraon) | tɑ$^{31}$xui$^{55}$gɑŋ$^{55}$ | ZMYYC 16 |
| Angami (Khonoma) | kwesi | GEM-CNL | Dimasa | jengolong-mander | GEM-CNL |
| Angami (Kohima) | pfesei | GEM-CNL | | | |
| Anong=Nung | mu$^{55}$bɛ$^{35}$bɯŋ$^{55}$ | ZMYYC 16 | Dulong | mɯ$^{31}$ɕiŋ$^{55}$ɟɔ̃ʔ$^{55}$ | ZMYYC 16 |
| Ao (Chungli) | tungnusen | GEM-CNL | Dumi (=Dumi Rai) | na:ghɨ | SVD-DUM |
| Ao (Mongsen) | tsungyangsung | GEM-CNL | Ergong | mdza | ZMYYC 16 |
| Apatani "A" | ñi-mé ja-ri | JS-TANI | Ersu | mɛ$^{55}$khua$^{'55}$ | ZMYYC 16 |
| Apatani "S" | ñi-me ja-ru | JS-TANI | Gallong | agre-go-ge | KDG-IGL |
| Bahing | tsit- | BM-BAH | Geman (=Kaman) | i$^{55}$phit$^{55}$ | ZMYYC 16 |
| Bai (Bijiang) | ʂeɹ$^{55}$ko$^{33}$lo$^{33}$ | ZMYYC 16 | Guiqiong | sɿ$^{33}$mpɛ$^{53}$ | ZMYYC 16 |
| Bai (Dali ) | ko$^{42}$ | ZMYYC 16 | Gurung (Ghachok) | ya:hgõ | SIL-GUR 5.A.75 |
| Bai (Jianchuan) | kɑ̃$^{55}$ko$^{44}$tɯ$^{21}$ | ZMYYC 16 | Hani (Caiyuan =Biyue) | tshɤ$^{55}$thɤ$^{55}$lɤ$^{55}$khy̠$^{33}$ | ZMYYC 16 |
| Bengni "S" | uk-ri: (go:-go:) | JS-TANI | | | |
| Bengni "S" | uk-ri: (ta-go:) | JS-TANI | Hpun (Northern) | săyoŋ maʔ | EJAH-HPUN |
| Bokar Adi | u rɛ: | ZMYYC 16 | Idu | ahu | NEFA-PBI |
| Bokar "S" | u-re: | JS-TANI | Idu | ɑ$^{31}$hɯ$^{55}$ | ZMYYC 16 |
| Burmese (Rangoon) | tθɛʔ$^{44}$tã$^{53}$ | ZMYYC 16 | Jinghua Pumi | mɐ$^{13}$sqhuɑ$^{13}$ | ZMYYC 16 |
| Burmese (Written) | cui | PKB-WBRD | Jingpho | nggoilatum | GEM-CNL |
| Burmese (Written) | sak-taṁ | PKB-WBRD | Jingpo | n$^{55}$koi$^{51}$la$^{55}$tum$^{33}$ | ZMYYC 16 |
| Burmese (Written) | saktan | GEM-CNL | | | |
| Burmese (Written) | thɑk$^{4}$tɑm$^{1}$ | ZMYYC 16 | Jinuo | ja$^{33}$mɔ$^{33}$ko$^{44}$ta$^{55}$ko$^{44}$tɕhø$^{33}$ | ZMYYC 16 |
| Chang | milishen | GEM-CNL | | | |
| Chepang | yo | SIL-CHEP 5.A.75 | Konyak | nimlong | GEM-CNL |
| | | | Kulung | riktokom_ | RPHH-KUL |
| Chinese (Mandarin, Simp.) | tsaaihurng | JS-CH 797 | Lahu (Na) | ʌ$^{33}$mu$^{53}$lʌ$^{31}$si$^{35}$dzɔ$^{33}$ | ZMYYC 16 |
| Chokri | meriizho | GEM-CNL | Langsu (=Maru) | ɣɔʔ$^{31}$kɔŋ$^{35}$saŋ$^{31}$ŋjɔ̃$^{31}$ | ZMYYC 16 |
| Cuona Monpa (=Takpa) | ŋʌ$^{53}$ | ZMYYC 16 | | | |

# Appendix 7 *Forms meaning* rainbow *from the STEDT database*

| | | |
|---|---|---|
| Liangmei | tingkhambam | GEM-CNL |
| Lijiang Naxi | mɯ$^{33}$lɯ$^{55}$xɯ$^{55}$dzi$^{31}$ | ZMYYC 16 |
| Lisu | ɑ$^{55}$mu$^{31}$ʃɿ$^{44}$kho$^{55}$ | ZMYYC 16 |
| Lisu (Northern) | a$^{55}$mɔ$^{21}$ ʃʐ$^{44}$khɔ$^{55}$ | DB-LISU |
| Lisu (Northern) | a$^{55}$mɔ$^{21}$ʃʐ$^{44}$ | DB-LISU |
| Lotha | sündraka | GEM-CNL |
| Lushai | chhimbâl | GEM-CNL |
| Manang (Manang) | sya$^{3}$ | YN-MAN 256 |
| Manang (Prakaa) | $^{4}$ɕyə | HM-PRAK 0158 |
| Manipur | chumthang | GEM-CNL |
| Mao=Sopvoma | mare | GEM-CNL |
| Maram * | tingmarangabang | GEM-CNL |
| Maring | langkhutmatin | GEM-CNL |
| Mawo Qiang | mu ʁu tsə thi | ZMYYC 16 |
| Meluri * | arepi | GEM-CNL |
| Mikir | mukak | GEM-CNL |
| Milang | be-ke-be-le | AT-MPB |
| Mile Yi (=Axi) | si$^{33}$mu$^{21}$si$^{33}$lɯ$^{33}$ | ZMYYC 16 |
| Mojiang Yi | ɣo̠$^{21}$kʌ̃$^{55}$ | ZMYYC 16 |
| Motuo Monpa (=Tsangla) | dza | ZMYYC 16 |
| Muya | ndzɛ$^{53}$ | ZMYYC 16 |
| Namuyi | nɛ$^{55}$ŋkhɛ$^{33}$zɿ$^{33}$ | ZMYYC 16 |
| Nanhua Yi | mɯ$^{21}$ɕi$^{33}$dʌ$^{33}$ | ZMYYC 16 |
| Nanjian Yi | ɑ$^{55}$m̩(ŋ̩)$^{21}$tshɑ$^{55}$du$^{55}$ro$^{21}$ | ZMYYC 16 |
| Ntenyi | chamakhokesho | GEM-CNL |
| Nusu (Bijiang) | tshɑ̃$^{31}$gɹi$^{53}$ɑ$^{31}$ | ZMYYC 16 |
| Pattani | ɖrug | DS-PATT |
| Puiron | sangpok | GEM-CNL |
| Queyu | ndza$^{35}$ | ZMYYC 16 |
| rGyarong | ndʒa | ZMYYC 16 |
| Rongmei | pongsing | GEM-CNL |
| Sangtam | mürükingking | GEM-CNL |
| Tibetan (Sde-dge = Khams) | ndza$^{53}$ | ZMYYC 16 |
| Sema | milesü | GEM-CNL |
| Shuikui Hani (Haoni) | xɯ$^{55}$tɯ$^{55}$lɯ$^{55}$mɔ$^{33}$ | ZMYYC 16 |
| Sulong (=Sulung) | kə$^{33}$leɳ$^{33}$ | ZMYYC 16 |
| Tagin | hugritago | KDG-TAG |
| Tangkhul | changji-khamawut | GEM-CNL |
| Taoba Pumi | mɐ$^{35}$kha$^{35}$ | ZMYYC 16 |
| Taoping Qiang | χmə$^{55}$qu$^{31}$tsu$^{33}$ | ZMYYC 16 |
| Thulung | naŋguŋ | NJA-TR |
| Thulung | pla(l) | NJA-TR |
| Tibetan (Amdo /Bla-Brang) | ndza | ZMYYC 16 |
| Tibetan (Lhasa) | tɕa$^{13}$ | ZMYYC 16 |
| Tibetan (Written) | 'jathsoŋ | GEM-CNL |
| Tibetan (Written) | ja.tshon | JS-TIB 797 |
| Tibetan (Written) | ɦidzaɦ | ZMYYC 16 |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

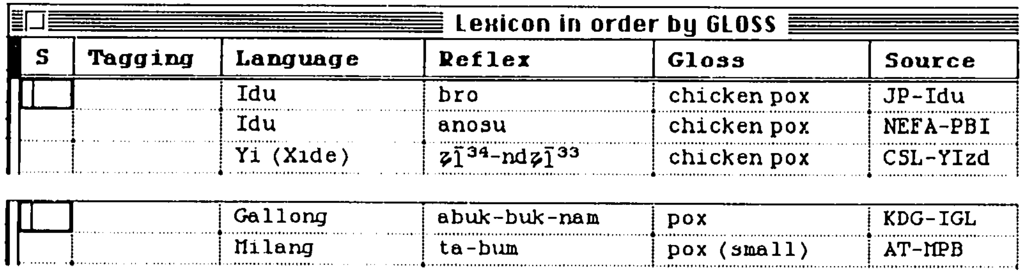
At STEDT, the interlinking of lexical items is made explicit though a process of semantic classification (cf. §xxx above): each word is assigned to a specific semantic category based on a "culturally appropriate" thesaurus hierarchy created to structure the database. Since the database is quite large, providing such a classification represents a substantial data processing problem, which we will address in this section.

## A Terminology

The following terms will be used in the discussion which follows. A *lexical entry* is the computer equivalent of a dictionary article: it contains a *headword* or *lexeme*, a short definition called a *gloss*, and other information such as *grammatical function* (including part of speech). A gloss is composed of a *word* or *phrase* (several words) which may or may not provide in the *glossing metalanguage* a form that is substitutable for the given lexeme. That is, some glosses in the STEDT database are terse, merely markers of the most common sense of the lexeme, while others are long and verbose. A *semantic category* (or *semcat*) is an element (such as "Diseases" or "Body Part Noun") in a hierarchical description of the semantic universe, and is designated by a mnemonic label (such as Ndis or Nbp). These notions are illustrated in the following example:

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

*(xx)*  *5 lexical entries from STEDT DB*



Lexicon in order by GLOSS

| S | Tagging | Language | Reflex | Gloss | Source |
|---|---|---|---|---|---|
| □ |  | Idu | bro | chicken pox | JP-Idu |
|  |  | Idu | anosu | chicken pox | NEFA-PBI |
|  |  | Yi (Xide) | ƶɿ̄$^{34}$-ndƶɿ̄$^{33}$ | chicken pox | CSL-YIzd |
| □ |  | Gallong | abuk-buk-nam | pox | KDG-IGL |
|  |  | Milang | ta-bum | pox (small) | AT-MPB |

Note that in the five lexical entries above (for five different lexemes in four languages), there are three glosses:

*(xx)*

CHICKEN POX
POX
POX (SMALL)

and three words used in these glosses:

*(xx)*

CHICKEN
POX
SMALL

The gloss *chicken pox* (as a two-word phrase) occurs three times; elsewhere the word *chicken* alone occurs 66 times, and *chickenpox* (as one word) occurs once. These distinctions will be important in creating the semantic clasification of the database (below).

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

## 3.2.2.3.2. Statistical analysis of the STEDT lexicon

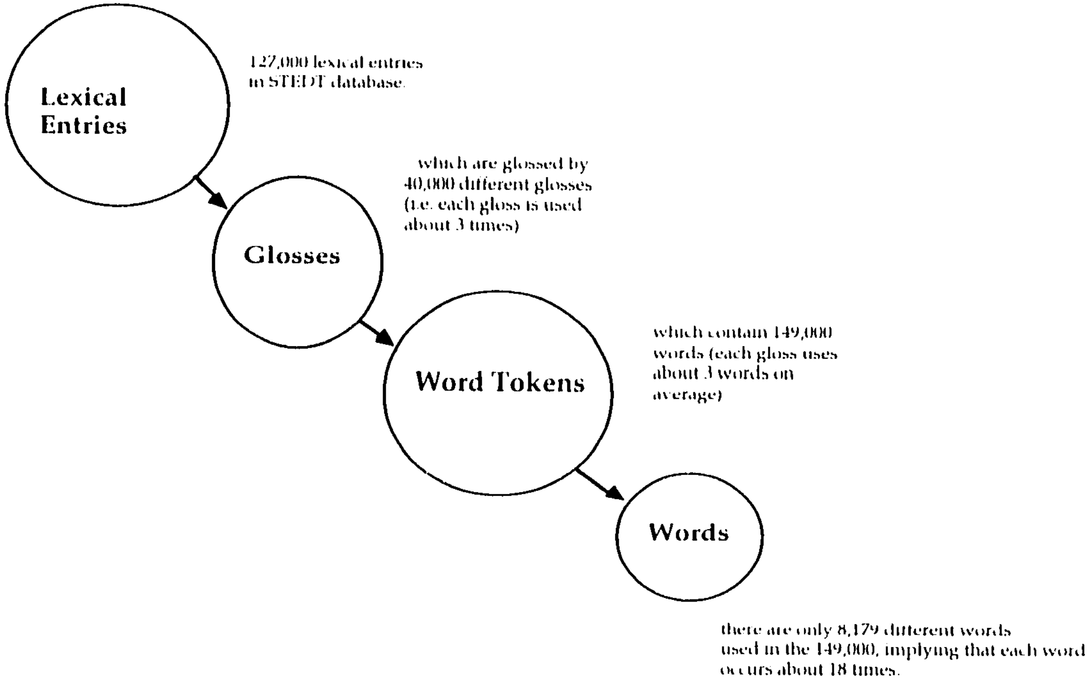
To get an idea of the scope of the data set requiring semantic classification in the STEDT database, an analysis of the words found in the gloss fields of the Lexicon file in the STEDT database was performed.

There are about 125,000 lexical entries in the database, each with its own gloss. However, there were only about 40,000 *different* (i.e. unique) glosses. For example, the gloss *chicken pox* (noted above) occurs three times in the database.

In these 40,000 distinct glosses there are 149,419 words (tokens) . These consolidate to 8,179 different words (types), implying that each word occurs an average of 18-odd times. Of these 8000+ words, 831 are used in glosses which were 'recognized' as body parts and as such were already classified in the first phase of research. Of the 7,200 remaining words, most belong to other realms of the semantic universe.

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

**!! fix and move stats on sem cat...**


Lexical Entries
127,000 lexical entries in STEDT database.
Glosses
which are glossed by 40,000 different glosses (i.e. each gloss is used about 3 times)
Word Tokens
which contain 149,000 words (each gloss uses about 3 words on average)
Words
there are only 8,179 different words used in the 149,000, implying that each word occurs about 18 times.

### 3.2.2.3.3. Outline of approach to semantic classification

Carrying out a semantic classification is a time-consuming business. Assuming for example that a grad student can classify 20 glosses a minute (1 every 3 seconds), it would take 167 hours (four workweeks of 8-hour days!)

to classify 200,000 lexemes. And this presumes that there is no "overhead" in the classification process, i.e. that all necessary categories pre-exist for classification or can be immediately created in the appropriate places when needed. How can we minimize the work involved in carrying out this classification?

Certainly from a database point of view it would be inefficient to classify each gloss-token anew every time it appears: it is better to make such a decision just once, based on a list of *consolidated* glosses—the 8,179 gloss-words mentioned in the preceeding section. Classifying just these 8,179 gloss-words into semantic fields permits in turn a classification of the entire list of 40,000 glosses, which in turn permits a classification of all 119,000. In fact, since Zipf's Law ({Zipf, 1939 #813}:xx) applies, a classification of the 1400 most frequently occurring words (those which occur more than 18 times) in turn classifies 122,545 of the occurrences of words or about 83% of the database.

Thus several steps of data reduction and analysis are proposed:

(xx) *STEPS FOR SEMANTIC CLASSIFICATION OF THE STEDT DATABASE*

1) Review existing sources, and identify and (to the extent possible) utilize existing classificatory schemes and elicitation aids, especially those which are geared toward the ST area, such as CALMSEA (!!cite) and the existing semantic classification of roots in the Sino-Tibetan Conspectus (!!cite Matisoff ms). Create an overarching, coarse semantic categorization.
2) Create a *consolidated* list of all occurring glosses (n = approx. 40,000).
3) Analyze glosses into words and/or phrases (n = approx. 8,179).
4) Classify these gloss-words into the coarse semantic categories; then subdivide each category into salient subcategories. This will provide a finely detailed semantic grid.

5) The list of glosses is in turn classified in terms of these same categories and subcategories, based on the gloss-words the gloss contains.
6) It is inevitable that some multi-word glosses will be assigned to conflicting categories. Some means of reconciliation (either automated or manual or both) must be provided to either identify the dominant category, or to assign a gloss to multiple categories.
7) Within each semcat, order the glosses, probably in alphabetical order by the "dominant" gloss word.
8) Once the glosses are classified, the lexical entries in the database can be themselves classified.
9) Final verification and evaluation of processing; use of semcats in tagging.
10) Maintain and refine categories.

The following sections will discuss each of these steps in turn.

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

## B Creating a semantic structure for (semcats) classification

Usable semantic classifications of specific areas do exist (cf. GET REFERENCES!). However, since these represent only partial classifications of the semantic universe which the STEDT project addresses, they were not used as the foundation for a classification. Thus, existing thesauri and etymological dictionaries based on semantic fields were found to be based on schemas which were too abstract to be useful for classifying words in the languages of SE Asia.

Matisoff's previous work at semantic classification provided a basic skeleton for the present classification. For example, Matisoff classified all reconstructions in the STC into semantic fields. His noun classification is reproduced below.

(xx) *The top level of semantic classification of ST Nouns (Matisoff xxx)*

I. NOUNS

| | Description | "SemCat" |
|---|---|---|
| A. | Pronouns and nouns referring to humans | (Npro) |
| B. | Kinship terms (Nkin) | |
| C. | Body-parts, bodily secretions, bodily excrescences (human or animal) | (Nbp) |
| D. | Foodstuffs (vegetable) | |
| E. | Animal-names or animal products | (Nanim) |
| F. | Natural objects or natural phenomena | (Nnat) |
| G. | Artifacts; objects or institutions made by human beings | (Nart) |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

| | | |
|---|---|---|
| H. | Spatial or directional nouns | (Nspat) |
| I. | Measure nouns | (Nmeas) |
| J. | Time nouns | (Ntime) |
| K. | Numerals and quantifiers | (Num) |
| L. | Abstract nouns | (Nabst) |

STC cognate sets were assigned to these categories (*semcats*). The reconstructions from STC assigned to 1H and 1I, for example, are given below:

(xx) *A portion of Matisoff's classification of TB roots from the STC using the above semantic categories*

| | | | | |
|---|---|---|---|---|
| H. | Spatial or directional nouns | | Set No. | (Nspat) |
| | bay | left (side) | (47) | |
| | r-gu:ŋ | edge | (395) | [also Nbp 'shin'] |
| | ki:l | angle | (373) | [also Vtr 'bend, twist, roll] |
| | lam | road; direction | (87) | [also Nnat 'road'] |
| | la:y | center | (287) | [also Nbp 'navel'] |
| | ok | below | (110) | |
| | ren | line, row | (346) | [also V 'be equal'] |
| | riy | boundary | (429) | [also Vtr 'draw, mark'] |
| | syar | east | (28) | [also Vmot 'rise'] |
| | l-tak | above | (52, 110, 123) | [also Vmot 'ascend'] |
| | tsyu:ŋ | inside | (390) | |
| | tu:ŋ | inside | (390) | |
| | g-ya~g-ra | right (side) | (98) | |
| I. | Measure nouns | | | (Nmeas) |
| | hap | mouthful | (89) | [also Vtr 'bite, snap at] |
| | mu:k | arm's-length | (394) | [also Nbp 'arm'] |
| | twa | span | (165) | [also Vtr] |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

| | | | |
|---|---|---|---|
| um | mouthful | (108) | [also Vtr 'hold in the mouth'] |
| duŋ | length | (20) | [also Vstat 'be long'] |

Extending this simple structure to the STEDT framework, the major grammatical distinctions (N, V, A), taken as a first cut, were re-analyzed into finer categories and laid out as 5 "fascicular groups". The result was the provisional grouping and ordering of semantic fields that has been used for the volumes of STEDT (Matisoff 1992x):

*(xx)*

1. **Body-parts, functions, and products**

2a. **Animals**
2b. **Kinship/pronouns/social roles** (e.g. *younger sister, we, spirit-doctor...*)
2c. **Natural objects** (e.g. *sun, star, grass...*)
2d. **Plants and foods** (e.g. *cotton, indigo, taro, rice...*)

3a. **Inanimate verbs** (e.g. *blow, melt, freeze, boil...*)
3b. **Verbs of motion** (e.g. *swim...*)/ **transportation** (e.g. *carry...*)/ **orientation in space** (e.g. *stand...*)
3c. **Verbs of manipulation and production** (e.g. *build, make, hold, take, twist, roll...*)

4a. **Artifacts** (e.g. *trap, crossbow...*)/ **clothing** (e.g. *shirt, sash, button...*)/ **dwelling** (e.g. *house, granary, chickencoop, village, road...*)
4b. **Psychological/emotional/sensory verbs/verbs of utterance** (e.g. *hope, think, feel, shout...*)
4c. **Interpersonal/social verbs** (e.g. *promise, help...*)

5a. **Adjectival/stative verbs (including colors)** (e.g. *fat, smooth, red, pointed...*)
5b. **Numerals and quantifiers**
5c. **Abstractions: Space/time/existence/possession/grammatical functors**

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

These fascicles and fascicle subsections were further refined into 65 smaller categories as follows:

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

*(xx)*

| Chap | Heading | SemCat |
|---|---|---|
| 1a | BODY PART NOUNS | Nbp |
| 1b | BODY PART VERBS | Vbp |
| 2a | ANIMALS | Nanim |
| 2a | ANIMATE VERBS | Vanim |
| 2b | KINSHIP TERMS | Nkin |
| 2b | RELATIONSHIPS | Nrel |
| 2b | WORDS FOR HUMAN BEINGS | Nhum |
| 2c | NATURAL OBJECTS | Nnat |
| 2d | VEGETABLES AND PLANTS | Nveg |
| 2d | VERBS REFERRING TO PLANTS | Vveg |
| 3a | ACTIONS OF NATURAL OBJECTS | Vnat |
| 3a | VERBS OF FRICTION | Vrub |
| 3a | VERBS OF REDUCTION | Vburn |
| 3b | STRIKING, HITTING VERBS | Vhit |
| 3b | VERBS OF MOTION | Vmot |
| 3c | ACTION VERBS | Vact |
| 3c | CULINARY VERBS | Vfood |
| 3c | VERBS OF CONSTRUCTION | Vmake |
| 3c | VERBS OF CUTTING | Vcut |
| 3c | VERBS OF DIGGING | Vdig |
| 3c | VERBS OF HOLDING | Vhold |
| 3c | VERBS OF JOINING | Vjoin |
| 3c | VERBS OF PUTTING | Vput |
| 3c | VERBS OF REPAIR | Vmend |
| 3c | VERBS OF SEPARATION | Vsep |
| 3c | VERBS OF THROWING | Vthrw |
| 3c | VERBS OF UNDOING | Vundo |
| 3c | VERBS WITH DIRECT OBJECTS | Vdo |
| 3d | VERBS OF CARRYING/LIFTING | Vbear |
| 3d | VERBS OF CLEANING | Vwash |
| 3d | COPULATIVE VERBS | Vcop |
| 4a | ARTIFACTS | Nart |
| 4a | ARTIFACTUAL VERBS | Vart |
| 4b | PSYCHOLOGICAL VERBS | Vpsi |
| 4b | VERBS OF FEELING | Vfeel |
| 4b | VERBS OF UTTERANCE | Vutt |
| 4c | HUMAN ACTIONS | Vhum |
| 4c | VERBS OF ACQUISTIONS | Vacq |
| 4c | VERBS OF EXCHANGE | Vexch |
| 4c | VERBS OF FIGHTING | Vfigh |
| 4c | VERBS OF GIVING | Vgive |
| 4c | VERBS OF HELPING | Vhelp |
| 4c | VERBS OF HARM | Vharm |
| 4c | VERBS OF SOCIAL ACTION | Vsoc |
| 5a | ADJECTIVES, ATTRIBUTIVES | Vadj |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

| | | |
|---|---|---|
| 5a | CHANGE OF STATE | Vcos |
| 5a | CHANGE OF X | Vcox |
| 5a | SOUNDS | SOUND |
| 5a | COLORS | COL |
| 5a | VERBS REFERRING TO SIZE/SHAPE | Vshiz |
| 5a | VERBS REFERRING TO SURFACES | Vsurf |
| 5a | VERBS REQUIRING LIQUIDS | Vliq |
| 5a | VERBS REFERRING TO QUALITIES | Vqual |
| 5b | MEASURE WORDS | MEAS |
| 5b | NUMERIC EXPRESSIONS | NUM |
| 5c | ABSTRACT NOUNS | Nabst |
| 5c | ABSTRACT VERBS | Vabst |
| 5c | SPATIAL TERMS | SPAT |
| 5c | TIME | Ntime |
| 5c | VERBS OF POSSESION | Vhave |
| ? | ERRORS | ERROR |
| ? | Ambiguous | ambi |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

### 3.2.2.3.x. Creating a consolidated list

A list of the 40,000 unique glosses was extracted from the database. An excerpt of this "consolidated list" is given below.

(xx) *Glosses from STEDT DB beginning with the characters CORP-*

| | |
|---|---|
| corpse | 69 |
| corpse (archaic language) | 1 |
| corpse (human) | 1 |
| corpse (shi) | 2 |
| corpse, bier | 1 |
| corpse / dead body | 35 |
| corpse, carcass; residue, sediment | 1 |
| corpse, dead body | 17 |
| corpse, lie as a corpse | 2 |
| corpse/body | 1 |
| corpse/dead body | 2 |
| corpulent, plump | 1 |

### 3.2.2.3.5. Extracting individual words (decomposing complex lexical items)

A concordance of words found in the consolidated list was created. From the list of glosses above, for example, the following list of words would be extracted:

*(xx)*

archaic
human
shi
a
as
bier
body
carcass
corpse
corpulent
dead
language
lie
plump
residue
sediment

We can take advantage of the monosyllabic, analytic nature of TB languages to considerably simplify the problem of identifying the morphological constituents of lexical items. Consider first the general, abstract case, where "complex" (i.e. multimorphemic) words are glossed

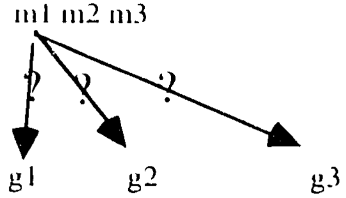
*(xx) Three lexical entries, containing four different morphemes and five gloss words*

| Entry | Form | Gloss |
|---|---|---|
| $e_1$ | $m_1$ $m_2$ $m_3$ | $g_1$ $g_2$ $g_3$ |
| $e_2$ | $m_1$ $m_4$ | $g_4$ |
| $e_3$ | $m_3$ | $g_2$ $g_4$ $g_5$ |

Can we match the morphemes to the words of the glosses in some meaningful way that will help sort out the meaning of the constituents? That is, do any of the morphemes of the source languages have a meaning similar to one of the words of the glossing phrase, as illustrated below?

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

*(xx)*



Suppose we were to create an "exploded" list in which the entries consisted of ordered pairs (morpheme, gloss), with each morpheme matched with each possible gloss word. If we created this list using the scheme of data in (xx) above, the set's "outer product" (usually written M X G, where M is the list of morphemes and G is the list of glosses) would have fourteen entries:

*(xx)*

| | | |
|---|---|---|
| $e_1$ | $m_1$ | $g_1$ |
| $e_1$ | $m_1$ | $g_2$ |
| $e_1$ | $m_1$ | $g_3$ |
| $e_1$ | $m_2$ | $g_1$ |
| $e_1$ | $m_2$ | $g_2$ |
| $e_1$ | $m_2$ | $g_3$ |
| $e_1$ | $m_3$ | $g_1$ |
| $e_1$ | $m_3$ | $g_2$ |
| $e_1$ | $m_3$ | $g_3$ |
| $e_2$ | $m_1$ | $g_4$ |
| $e_2$ | $m_4$ | $g_4$ |
| $e_3$ | $m_3$ | $g_2$ |

| | | |
|---|---|---|
| $e_3$ | $m_3$ | $g_4$ |
| $e_3$ | $m_3$ | $g_5$ |

*(xx)*

Total: $(3 \times 3) + (2 \times 1) + (1 \times 3) = 14$
(N.B. stopwords omitted)
**!!describe "stopwords" or give reference...**
In general, the number of entries in this cross product is:

*(xx)*

$$\sum_{i=1}^{m}(M_i \times G_i)$$

where: m is the number of items in the consolidated list
$M_i$ is the number of morphemes in the ith entry
$G_i$ is the number of words used in the gloss of the ith entry

For the STEDT database as a whole, m is 225,000.[1] The total number of morphemes is approximately yy, or about qq morphemes per lexical entry. 45,000 different phrases (i.e. sequences of one or more words) are used to gloss these entries; and the total number of words used in the glosses is zz, giving an average of yy words per gloss. The actual

[1] See the statistical summary of the STEDT database in chapter xx section yy.

number of entries in the cross product is 559,000, which agrees well with what would be expected given the mean numbers of morphemes per entry.

**!! finish this section...**

*(xx)*

| Morpheme List<br>All meanings for a<br>particular morpheme | | | Gloss Word List<br>All morphemes with a<br>particular meaning (or part thereof) | | |
|---|---|---|---|---|---|
| $m_1$ | $g_1$ | $e_1$ | $g_1$ | $m_1$ | $e_1$ |
| | $g_2$ | $e_1$ | | $m_2$ | $e_1$ |
| | $g_3$ | $e_1$ | | $m_3$ | $e_1$ |
| | $g_4$ | $e_2$ | | | |
| | | | $g_2$ | $m_1$ | $e_1$ |
| $m_2$ | $g_1$ | $e_1$ | | $m_2$ | $e_1$ |
| | $g_2$ | $e_1$ | | $m_3$ | $e_1, e_3$ |
| | $g_3$ | $e_1$ | | | |
| | | | $g_3$ | $m_1$ | $e_1$ |
| $m_3$ | $g_1$ | $e_1$ | | $m_2$ | $e_1$ |
| | $g_2$ | $e_1$ | | $m_3$ | $e_1$ |
| | $g_3$ | $e_1$ | | | |
| | $g_2$ | $e_3$ | $g_4$ | $m_1$ | $e_2$ |
| | $g_4$ | $e_3$ | | $m_4$ | $e_2$ |
| | $g_5$ | $e_3$ | | $m_3$ | $e_3$ |
| $m_4$ | $g_4$ | $e_2$ | $g_5$ | $m_3$ | $e_3$ |

Note that morpheme $m_4$ occurs only once in the morpheme list, since it occurs in only one lexical entry ($e_2$), and the gloss used with it ($g_4$) comprises only one word. Morpheme $m_3$, on the other hand, occurs six times: it occurs in two lexical entries, whose glosses comprise a total of six words altogether.

Consider now a more realistic application of this technique. Below are listed some words in TB languages containing a morpheme composed of a labiodental fricative initial plus open rhyme and having a meaning related to *pig* (the data in this contrived example is related to PTB *p-wak STC xxx). Note that these words (unlike the previous, abstract example) are monosyllabic and monomorphemic.

*(xx)*

| Morph | Language | Reflex | Gloss | Source | Set |
|---|---|---|---|---|---|
| vɑ̰$^{21}$ | Bai | vɑ̰$^{21}$ tui$^{33}$ | *fat (person)* | JZ-BAIbj | |
| vɑ̃$^{21}$ | Bijiang Bai | vɑ̃$^{21}$tui$^{33}$ | *fat (of people)* | ZMYYC | 849 |
| va$^{53}$ | Ergong (Northern) | va$^{53}$<br>snɔ$^{53}$~va$^{53}$<br>ŋ̊tɕʰe$^{53}$ | *snout (pig)* | SHK-ErgNQ | 3.5.5 |
| va$^{44}$ | Jinuo | va$^{44}$ni$^{44}$ | *pig* | ZMYYC | |
| va$^{55}$ | Jinuo (A) | va$^{55}$ ni$^{55}$ a$^{44}$ ʃɔ$^{44}$ | *pork* | DQ-JinA | 587 |
| $^{21}$vaʔ | Lahu | $^{21}$vaʔ | *pig* | JAM-TSR | 168 |
| vá6 | Lisu | á2-vá6 | *pig* | JAM-TSR | 168 |
| va 55 | Nasu | va 55 | *pig* | JAM-TSR | 168 |
| va̰$^{11}$ | Nusu (A) | va̰$^{11}$ kʰi$^{55}$ | *excrement (pig)* | DQ-NusuA | 337. |

# APPENDIX 8: SEMANTIC CLASSIFICATION OF THE STEDT LEXICON

| | | | | | |
|---|---|---|---|---|---|
| va̠$^{53}$ | Nusu (B) | va̠$^{53}$ kʰ'i$^{55}$ | *excrement (pig)* | DQ-NusuB | 337. |
| vɑ̠$^{53}$ | Nusu (Bijiang) | vɑ̠$^{53}$ | *pig* | ZMYYC | |
| vɑ̠$^{55}$ | Nusu (Bijiang) | de$^{55}$vɑ̠$^{55}$ | *wild boar* | ZMYYC | |
| va$^{13}$ | Yi (Dafang) | va$^{13}$ | *pig* | ZMYYC | |

Taking a few of the forms (as shown in (xx)) and expanding and sorting them according to the algorithm given above, we would get morpheme and gloss word lists like those in (xx):

*(xx)*

| | | | | | |
|---|---|---|---|---|---|
| va̠$^{53}$ | Nusu (B) | va̠$^{53}$ kʰ'i$^{55}$ | excrement (pig) | DQ-NusuB | 337. |
| vɑ̠$^{53}$ | Nusu (Bijiang) | vɑ̠$^{53}$ | pig | ZMYYC | |
| vɑ̠$^{55}$ | Nusu (Bijiang) | de$^{55}$vɑ̠$^{55}$ | wild boar | ZMYYC | |
| va 55! | Nasu | va 55! | pig | JAM-TSR | 168 |

*(xx)*

| Morpheme List | | Gloss List | |
|---|---|---|---|
| Morpheme | Gloss Word | Gloss Word | Morpheme |
| de$^{55}$ | boar | boar | de$^{55}$ |
| de$^{55}$ | wild | | vɑ̠$^{55}$ |
| kʰ'i$^{55}$ | excrement | excrement | kʰ'i$^{55}$ |
| kʰ'i$^{55}$ | pig | | va̠$^{53}$ |

| | | | |
|---|---|---|---|
| va 55 | pig | pig | k$^{h}$'i$^{55}$ |
| va̱$^{53}$ | excrement | | va 55! |
| va̱$^{53}$ | pig | | va̱$^{53}$ |
| vɑ̱$^{53}$ | pig | | vɑ̱$^{53}$ |
| vɑ̱$^{55}$ | boar | | |
| vɑ̱$^{55}$ | wild | wild | de$^{55}$ |
| | | | vɑ̱$^{55}$ |

*!!also note: WT sde 'field, expanse, zone'* *GSTC 95* *da:y

*(xx)*

PLB *wak PIG [TSR 168]

| | |
|---|---|
| Na | va$^{55}$ |
| LC | a$^{55c}$ |
| Wo | ma$^{33}$ |
| Bi | wà |
| Li | á$^{2}$-vá$^{6}$ |
| Ha [HT] | ɣa$^{21c}$ |
| Ha [K] | βa$^{21}$ |
| Sa | ve$^{22S}$ |
| Ahi | vie$^{44S}$ |
| Ak | a-za$^{1}$-$^{1S}$ |
| Lh | vàʔ |
| | |
| Ch | pje |
| WT | phak |

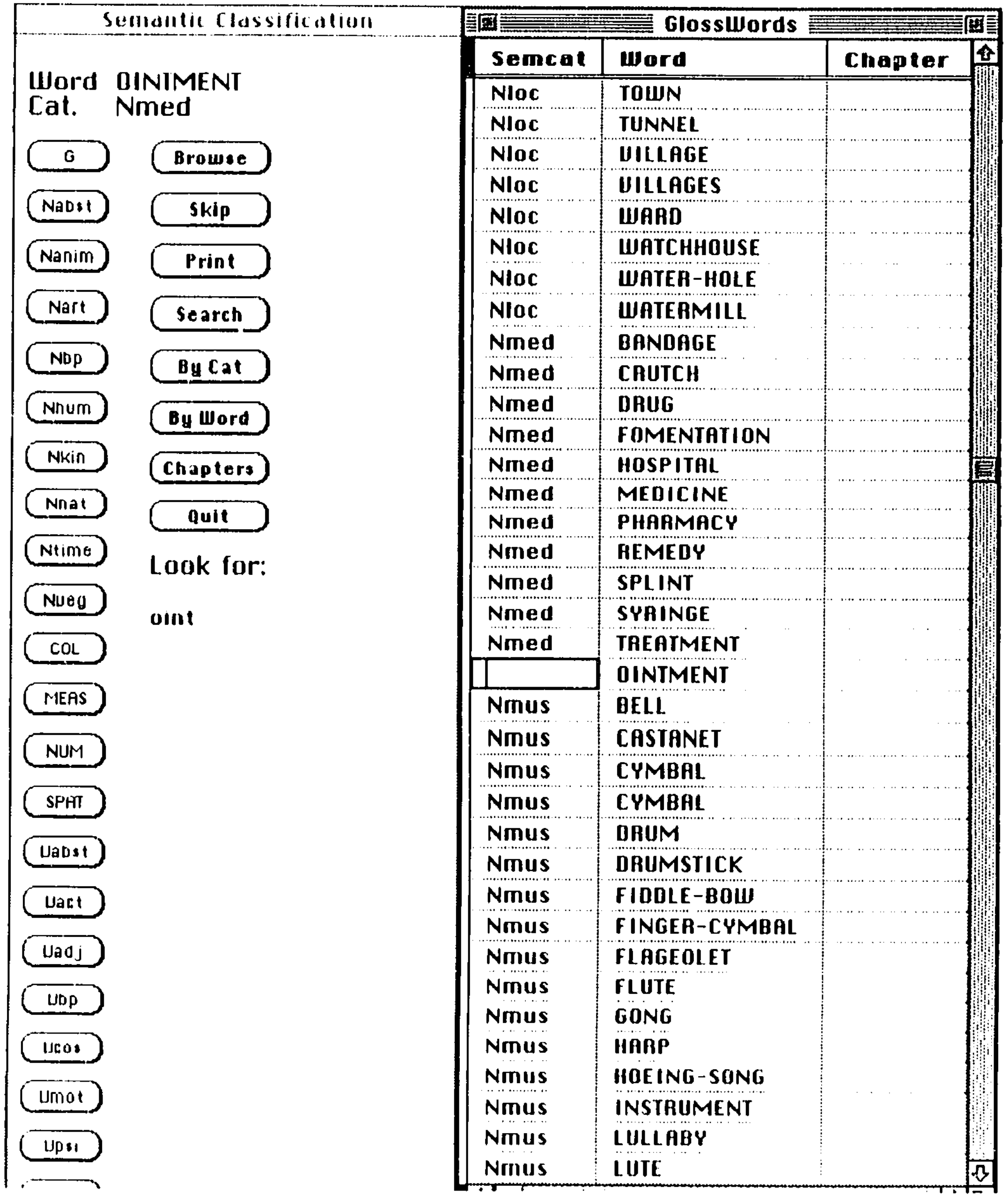
| | |
|---|---|
| Tr | Ua$^{43}$ |
| WB | wak |

### 3.2.2.3.6. Assigning words to major semcats

Each word in the consolidated list of gloss-words (n = 8,179) was assigned to one of the 65 semantic categories presented in section 3.2.2.3.4 (Table xx), with some exceptions. Software was developed to facilitate the assignment process: the user could select a category and browse the list of words looking for words to assign to the given category. This ensured consistency between the words assigned and the semcats used, and sped up the process substantially.

*(xx) Semantic Classification software:*

*The word OINTMENT will be tagged as Nmed (medical terminology) when the TAG button is clicked.*



Word OINTMENT
Cat. Nmed

Look for:

oint

| Semcat | Word | Chapter |
|---|---|---|
| Nloc | TOWN | |
| Nloc | TUNNEL | |
| Nloc | VILLAGE | |
| Nloc | VILLAGES | |
| Nloc | WARD | |
| Nloc | WATCHHOUSE | |
| Nloc | WATER-HOLE | |
| Nloc | WATERMILL | |
| Nmed | BANDAGE | |
| Nmed | CRUTCH | |
| Nmed | DRUG | |
| Nmed | FOMENTATION | |
| Nmed | HOSPITAL | |
| Nmed | MEDICINE | |
| Nmed | PHARMACY | |
| Nmed | REMEDY | |
| Nmed | SPLINT | |
| Nmed | SYRINGE | |
| Nmed | TREATMENT | |
| | OINTMENT | |
| Nmus | BELL | |
| Nmus | CASTANET | |
| Nmus | CYMBAL | |
| Nmus | CYMBAL | |
| Nmus | DRUM | |
| Nmus | DRUMSTICK | |
| Nmus | FIDDLE-BOW | |
| Nmus | FINGER-CYMBAL | |
| Nmus | FLAGEOLET | |
| Nmus | FLUTE | |
| Nmus | GONG | |
| Nmus | HARP | |
| Nmus | HOEING-SONG | |
| Nmus | INSTRUMENT | |
| Nmus | LULLABY | |
| Nmus | LUTE | |

At this point, a number of issues arise, in some cases requiring ad hoc solutions.

1) Words which were high in frequency and low in information (and which would thus normally occur on a stop list) were assigned to a special "dead end" category; such words would not be used in further semantic processing.

2) Many words were ambiguous with regard to their category assignment; these were assigned to a separate, additional category for special treatment.

3) The categorization of gloss-words into the major grammatical categories of the glossing metalanguage (i.e. English) can be somewhat problematic for Asian languages. For example, the distinction between adjective and verb is often irrelevant (*cold* and *be cold* are often the same lexical item).

4) If a word is clearly homonymous or homographic in English (e.g., 'tear'), two entries will eventually be needed. As a temporary expedient, the symbol "?" is used.

"?" is used also for entries which were problematic in any other way. Such "?" entries must be discussed individually.

5) For sets like BEAT/BEATEN/BEATING, where a single lexeme occurs in several inflected forms straddling several parts of speech, one semantic classification is chosen (e.g., V act). (Other examples: BOIL/BOILED/BOILING: V act; BRIGHT/BRIGHTNESS: V adj.)

6) Abstract nouns like COMPLETION were temporarily assigned to a cover category VERB.

7) Some sem. classes, e.g., meteorological phenomena, will be entirely, e.g., verbal (there will be no noun for 'thunder'). These will be handled at a later stage of refined classification.

8)

lightning, thunder
make a sound, thunder
roar (of animals), thunder
roar (of tiger), to thunder
roar of thunder and crackle of
roar, thunder
thunder
thunder
thunder (verb)
thunder and lightning
thunder and lightning
thunder and lightning
thunder strike
thunder v.
thunder, cloudy
thunder, lightning
thunder, lightning strike
thunder, thunderstorm.
thunderbolt
thunderhead, nimbus cloud
thunderingly
thunderstorm

*Glosses from the STEDT database containing the string THUNDER*

8) Some words were assigned to certain categories by fiat; thus gods/ghosts/demons/spirits were classified as "N hum".

classifying some of the 8,719 different words found in glosses in the STEDT database. I removed the 2,761 words which occur only once and have set them aside for treatment later (note that some of these have already been tagged as a result of being identified as body parts). 5,418 words remained

### 3.2.2.3.7. Assigning glosses to categories based on the words they contain

Matching the classified gloss-words to the glosses in the lexicon which contain them, 106,550 glosses (= 90%) can be assigned a semantic classification (12,468 remain unassigned since the words they are composed of do not match anything). The quantitative breakdown of how many words are assigned to each semcat is as follows:

*(xx)*

| In descending frequency order | | In alphabetical order | |
|---|---|---|---|
| Classification | Count | Classification | Count |
| Nbp | 54878 | unassigned | 12468 |
| unassigned | 12468 | ? | 1264 |
| Vact | 10541 | COL | 40 |
| Vadj | 9744 | Error | 14 |
| Nart | 4516 | G | 2487 |
| Nanim | 4277 | NUM | 1353 |
| Nabst | 3831 | Nabst | 3831 |
| Nnat | 3720 | Nanim | 4277 |
| Nveg | 2633 | Nart | 4516 |
| G | 2487 | Nbp | 54878 |
| NUM | 1353 | Nhum | 1169 |
| ? | 1264 | Nkin | 1201 |

| | | | |
|---|---|---|---|
| Nkin | 1201 | Nnat | 3720 |
| Vabst | 1186 | Ntime | 442 |
| Nhum | 1169 | Nveg | 2633 |
| Vmot | 1141 | SPAT | 13 |
| Vbp | 1080 | Vabst | 1186 |
| Vpsi | 1020 | Vact | 10541 |
| Ntime | 442 | Vadj | 9744 |
| COL | 40 | Vbp | 1080 |
| Error | 14 | Vmot | 1141 |
| SPAT | 13 | Vpsi | 1020 |
| Total | 119018 | Total | 119018 |

Some of the smaller categories reflect a semantic subdivision which is already fine enough (such as Ntime). Others, such as active verbs (Vact), require considerable further refinement before they can be useful for further research.

It is at this point that the efficacy of this approach is evident (Figure xxx). For example, by classifying 136 words into the category for numbers (NUM), 2,783 glosses are consequently identified and brought together for further potential analysis.

| Semcat | N glosses in SemCat | N words in Semcat | Semcat | N glosses in SemCat | N words in Semcat |
|---|---|---|---|---|---|
| ? | 1191 | 350 | G? | 29 | 8 |
| COL | 48 | 10 | MEAS | 92 | 14 |
| Err? | 4 | 3 | NUM | 2783 | 136 |
| Erro? | 2 | 2 | Nabst | 704 | 53 |
| Error | 1023 | 40 | Nani? | 1 | 1 |
| G | 9941 | 197 | Nanim | 5206 | 313 |

|  |  |  |
|---|---|---|
| Nart | 7510 | 809 |
| Nbp | 51828 | 874 |
| Nbp? | 1 | 1 |
| Nhum | 2755 | 337 |
| Nkin | 1883 | 70 |
| Nnat | 5577 | 328 |
| Nnat? | 13 | 2 |
| Nrel | 372 | 43 |
| Ntime | 1870 | 95 |
| Nveg | 4025 | 308 |
| Nveg? | 11 | 4 |
| OK | 36 | 10 |
| SPAT | 3647 | 142 |
| V | 3 | 1 |
| Vabst | 1459 | 31 |
| Vacq | 29 | 7 |
| Vact | 14477 | 1194 |
| Vadj | 14164 | 1636 |
| Vadj? | 1 | 1 |
| Vanim | 3 | 2 |
| Vart | 3 | 1 |
| Vbp | 8780 | 192 |
| Vbp? | 3 | 1 |

|  |  |  |
|---|---|---|
| Vcatc | 4 | 1 |
| Vcos | 90 | 16 |
| Vdo | 18 | 3 |
| Vfood | 12 | 1 |
| Vgive | 9 | 2 |
| Vharm | 8 | 2 |
| Vhave | 2 | 1 |
| Vhelp | 7 | 2 |
| Vhum | 20 | 1 |
| Vjoin | 31 | 3 |
| Vliq | 6 | 3 |
| Vmake | 4 | 1 |
| Vmot | 1824 | 132 |
| Vnat | 1 | 1 |
| Vpsi | 3158 | 295 |
| Vqual | 274 | 46 |
| Vshiz | 640 | 35 |
| Vsoc | 2 | 1 |
| Vutt | 1360 | 145 |
| Vveg | 59 | 4 |
| Vvp | 4 | 1 |
| ambi | ! | 1 |
| error | 2 | 1 |

### 3.2.2.3.8. Resolving conflicting assignments

Of the classified glosses (n = 127,000) , 24,214 are "ambiguously classified"—that is, the words which make up the glosses belong to more than one semantic class (e.g. *Kidney (of animal)* = Nbp + Nanim ; *Lead (by the hand))* = Vact + Nbp).

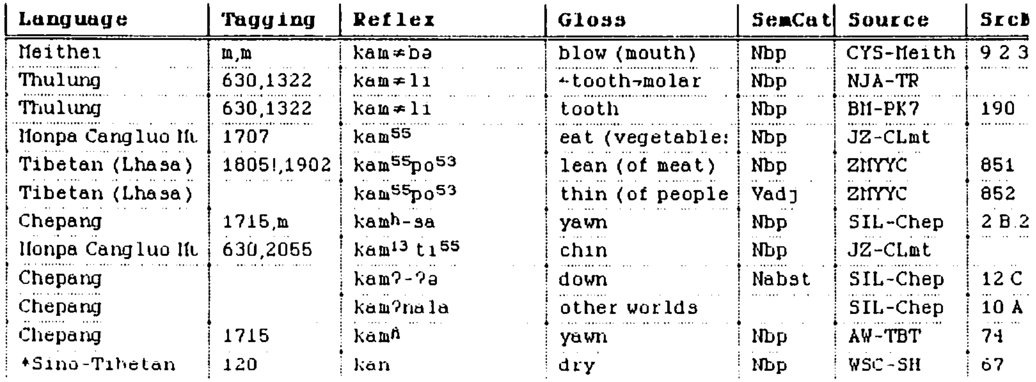
### 3.2.2.3.9. Ordering glosses with semcat and subcat

Since there may be several glosses assigned to even the most specific category, these glo ses must be ordered with respect to each other.

### 3.2.2.3.10. Propagating categorization of glosses into database

When the categorization of the gloss-words is completed, the gloss-words are matched against the glosses in the lexical entries and the relevant semcat(s) is/are attached to each entry, as illustrated below.

(xx) *Some lexical entries in the STEDT database showing their tentative semantic categorization*



| Language | Tagging | Reflex | Gloss | SemCat | Source | SrcD |
|---|---|---|---|---|---|---|
| Meithei | m,m | kam≠bə | blow (mouth) | Nbp | CYS-Meith | 9 2 3 |
| Thulung | 630,1322 | kam≠li | ⌐tooth¬molar | Nbp | NJA-TR | |
| Thulung | 630,1322 | kam≠li | tooth | Nbp | BM-PK7 | 190 |
| Monpa Cangluo Mt | 1707 | kam$^{55}$ | eat (vegetable: | Nbp | JZ-CLmt | |
| Tibetan (Lhasa) | 1805!,1902 | kam$^{55}$po$^{53}$ | lean (of meat) | Nbp | ZMYYC | 851 |
| Tibetan (Lhasa) | | kam$^{55}$po$^{53}$ | thin (of people | Vadj | ZMYYC | 852 |
| Chepang | 1715,m | kam$^{h}$-sa | yawn | Nbp | SIL-Chep | 2 B 2 |
| Monpa Cangluo Mt | 630,2055 | kam$^{13}$ ti$^{55}$ | chin | Nbp | JZ-CLmt | |
| Chepang | | kam?-?ə | down | Nabst | SIL-Chep | 12 C |
| Chepang | | kam?nala | other worlds | | SIL-Chep | 10 A |
| Chepang | 1715 | kam$^{ɦ}$ | yawn | Nbp | AW-TBT | 74 |
| *Sino-Tibetan | 120 | kan | dry | Nbp | WSC-SH | 67 |

### 3.2.2.3.11. Evaluating the categorization

**Accuracy**

- How many errors were made either by humans or by limitations of the approach?
- If a word was tagged incorrectly, all glosses with that word might be tagged incorrectly. Give some examples.

1) e.g. YOUNGER only in Nkin, but tagged Vadj
2) CORN tagged as Nveg, CALLUS (polysemy)
3) SCORCH, WHEEDLE, PERSUADE, CHOOSE (BRIDE) semantic breadth -> need multiple semcats

**Speed**

- How much time was actually saved (if any) by this method?

**Are there any other benefits of this scheme?**

Yes! There is now

1) a consistent and updatable description of the semantics of the entire database.
2) a means for organizing the etymologized forms into thesaurus format: the semantic classification of the etyma themselves (described in section xxx) will parallel the classification of their glosses (just accomplished).

### 3.2.2.3.12. Using the "final product"

It will soon be possible to use the semantic categorization to subdivide the database into semantic fields, making it easier to study groups of words having related meanings. This is very useful for identifying morphemes that are shared by seve ral lexical items in a single language.

(xx) *Portion of STEDT database (sorted by language) categorized Nliquid ("Nouns referring to Liquids")*

| Language | Reflex | Gloss | Source | SrcNo | RecNo |
|---|---|---|---|---|---|
| Milang | asi-bi-da | water line | AT-MPB | | 955 |
| Milang | (L)asi kol | water pipe | AT-MPB | | 954 |
| Milang | (L)asi-kol | water point | AT-MPB | | 940 |
| Milang | diŋ-go | waterfall | AT-MPB | | 936 |
| Miri | kai-ne shite | river | IMS-HMLG | | 21749 |
| Miri | belɣ kybe | valley , below | IMS-HMLG | | 22139 |
| Miri | ish'-lek | water | IMS-HMLG | | 22567 |
| Miri | ish[i] | water | IMS-HMLG | | 22560 |
| Miri | ish'-byc' | water snake | IMS-HMLG | | 22566 |
| Moyon | nʌp | liquid mucus | DK-Moyon | 3 5 6 | 150234 |
| Moyon | rurəm | amniotic fluid | DK-Moyon | 10 4 1 | 150471 |
| Mpi | tɯʔ$^{2}$ | drop | JAM-MLBM | 39 | 30640 |
| Mpi | tɯʔ$^{2}$ | drops | JAM-MLBM | 5 | 30541 |
| Mpi | tɕhoʔ$^{5}$ | water | IH-PL1 | 497 | 134065 |
| Muya | sɯ$^{55}$mbɘ$^{35}$ | liquid mucus | SHK-MuyaQ | 3 5 6 | 129712 |
| Namuyi | ɲi$^{33}$ŋgɛ$^{55}$ɦɛ$^{55}$ | liquid mucus | SHK-NamuQ | 3 5 6 | 129887 |
| Nasu | no$^{33}$ tsɔ$^{33}$ ʑi$^{33}$ | liquid mucus | CK-Nasu | 3 5 6 | 9505 |
| Naxi (Lijiang | tʰə$^{35}$ | drop | ZMYYC | | 140636 |
| Naxi (Lijiang | dʑi$^{33}$xo$^{31}$ | river | ZMYYC | | 140775 |
| Naxi (Lijiang | dʑi$^{31}$ | water | ZMYYC | | 140855 |
| Nesu | no$^{55}$ ʑʅ$^{55}$ | liquid mucus | CK-Nesu | 3 5 6 | 9718 |
| Newari | sa: | amniotic fluid | SH-KNw | 10 4 1 | 62379 |
| Newari | nhi | liquid mucus | SH-KNw | 3 5 6 | 62139 |
| Newari (Dolak | khwo | river/valley | CG-Dolak | | 8898 |
| Newari (Dolak | lokhu | water | CG-Dolak | | 8911 |
| Newari (Dolak | naapẽ | water leech | CG-Dolak | | 8885 |
| Newari (Kathm | khusi | river/valley | CG-Kath | | 9132 |
| Newari (Kathm | lɔ: | water | CG-Kath | | 9145 |
| Nung | thi | water | STC | 55 | 69540 |
| Nusu (Bijiang | dzɑ$^{53}$ | drop | ZMYYC | | 141674 |
| Nusu (Bijiang | kʰɹɔ̃$^{55}$ɑ̃$^{31}$ | river | ZMYYC | | 141809 |
| Nusu (Bijiang | ɹi$^{31}$gɹɑ$^{53}$ | water | ZMYYC | | 141890 |
| Nyiq | nɑ̱$^{44}$ bɪ$^{33}$ ʐʅ$^{33}$ | liquid mucus | DQ-Nyiq | 3 5 6 | 16479 |