Toward a formal markup standard for etymological data

Sean Crist, Ph.D., Swarthmore College
scrist1@swarthmore.edu
LSA Annual Meeting, January 2005

Many lexical markup schemes ignore etymological information. This is often a perfectly reasonable design choice. For many applications (text-to-speech, part of speech tagging, machine translation, etc), this kind of information is of no obvious use.

The specialist in historical linguistics, however, might have considerable use for online etymological data.

> **Sample application 1:**
> Joe wants to know the relative chronology of X and Y, which are sound changes in the prehistory of Old English. He writes a program to project the entire reconstructed vocabulary of Proto-Germanic downstream through the Old English sound changes, excluding those forms for which there is no Old English reflex. The program does this twice, once for each ordering of X and Y. The program outputs words for which the two orderings make different predictions.
>
> **Sample application 2:**
> Joe wants to find the mistakes in an etymological dictionary. One kind of mistake is that pairs of words are wrongly claimed to be cognate. Joe writes a program to project an attested word upstream through the known sound changes, producing a set of possible forms in the proto-language which would have surfaced as the observed form. If two words are claimed to be cognate, and if the intersection of the respective two sets of possible proto-forms is empty, then the program outputs the pair as a likely error.

Applications of this kind are not a new idea. Smith (1969) implements a set of string substitutions (in SNOBOL4, on a CDC 6400) to project Proto-Indo-European forms into modern Russian for the purpose of testing hypotheses regarding the relative chronology of sound changes, *etc*. Other applications

are discussed *e.g.* in Campanile and Zampolli (1973), Stubbs (1985), Lowe and Mazaudon (1994), Covington (1996).

This paper is not about any particular application. Rather, it discusses how etymological data can be structured in machine-readable form which is general enough for use in a broad range of applications, including the two representative cases described above.

Most general-purpose markup schemes for dictionary data treat etymological data as a field of unstructured prose, perhaps delimited as an <etym> field (see appendix for an extensive review). This is a perfectly adequate strategy in a dictionary intended for human consumption (the electronic equivalent of an ordinary desk dictionary), where no special processing of the etymological data is expected beyond lookup and presentation. Such cases, however, are not the topic of this paper.

The model described here is etymology-centric. The structure of the entry directly embodies specific etymological relationships between words (cognation, inheritance, and borrowing). A skeletal entry obligatorily includes one or more these etymological relationships. Other information about words (gloss, morphological class, *etc.*) is of secondary importance and can be omitted or included at the user's option without affecting the overall structure of the entry.

The basic mathematical relationships

(*See, e.g.,* Trask, p. 205, Hock, p. 380)

A language L_i is an **ancestor** of a language L_j iff L_j developed from L_i over time through an unbroken chain of first language acquisition.

> *Examples:* Old English is an ancestor of Modern English. Latin is an ancestor of French. Proto-Celtic is an ancestor of Old Irish.
>
> The ancestor relationship is transitive. If Old English is an ancestor of Middle English, and Middle English is an ancestor of Modern English, then Old English is an ancestor of Modern English.

Two languages L_j and L_k ($L_j \neq L_k$) are **related** iff they share an ancestor.

Example: Old English is related to Gothic because the two languages share an ancestor, Proto-Germanic.

The relatedness relationship is also transitive, because every language has a single line of descent.

A word W_j (in L_j) is a **reflex** of a word W_i (in L_i) if it has been transmitted over time from L_i to L_j through an unbroken sequence of first language acquisition.

Example: The word *shall* is the Modern English reflex of the Old English word *sceal.*

An alternative wording is that L_j **inherits** W_j from L_i.

The relationship "is an **etymon** of" is the converse relationship. W_i is an **etymon** of W_j iff W_j is a reflex of W_i.

The earlier form W_i can be either attested or reconstructed. (The same is true for W_j, for that matter. **wiraz* "man" is the Proto-Germanic reflex of Proto-Indo-European **wiros,* even if this is a somewhat atypical use of the terminology)

The reflex relationship *does not include borrowing.* English *herb* is not a reflex of Latin *herba,* because Latin is not an ancestor of English. English has not inherited a single word from Latin!

The words W_j (in L_j) and W_k (in L_k) are **cognate** iff:

- L_j and L_k are related, and
- W_j is the reflex in L_j of a word W_i (in L_i), and
- W_k is the reflex in L_k of the word W_i

Example: English *day* and Gothic *dags* are cognate because they are the reflexes within the respective languages of Proto-Germanic **dagaz*.

Abuse of the term *cognate* is common. For example, there are papers describing the use of "cognates" such as *generation/génération* or *error/erreur* to automatically align bilingual texts. But these aren't

cognates at all; they are loans. *Cognate* does *not* mean "similar in sound and meaning" or "somehow connected in etymology". It is a technical term whose use entails a very specific claim about the type of etymological connection.

Word W_j (in L_j) is a **borrowing** (or **loan**) from L_i iff W_i (in L_i) has been adopted into the vocabulary of L_j, and L_i is not an ancestor of L_j.

Example: English *succotash* is a loan from Narragansett *msiquatash.* English *skirt* is a loan from Old Norse *skyrta.*

Discussion

Beyond these basic relationships, what are some of the other special considerations for etymological data?

Etymologies are a kind of analysis, and are not observations. Knowledge changes over time. This has at least two consequences.

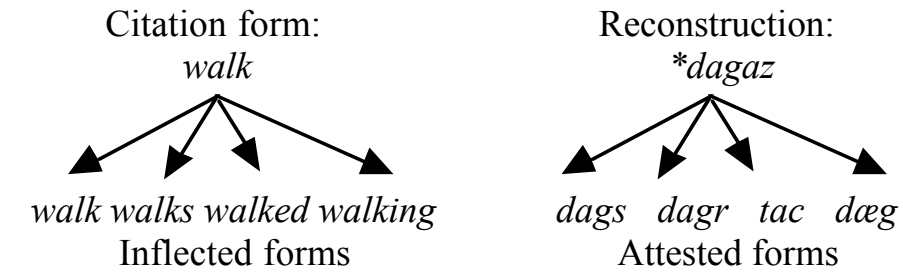
First, some etymological claims are more certain than others. In a synchronic lexicon (particularly one of a modern language), there is usually no reason to indicate a level of confidence; the claim that German *singt* is a form of the verb *singen* is utterly uncontroversial and has been accepted as fact by all for many years. Etymological claims, by contrast, are subject to controversy; probably no two experts agree on every claim. From early work in computational historical linguistics (Campanile and Zampolli, 1973), there have been attempts to encode confidence levels.

Certainty or uncertainty can be a property of an entire etymology. It can also be a property of a portion of an etymology: often, some sub-parts of an etymology are more certain than others. Authorities might differ as to the parts of an etymology which they accept or reject, and to what degree. As far as I am aware, the proposal below is the first to adopt a specific model addressing this issue. (Amsler and Tompa 1998 do include a <cert> element to bracket words such as *Prob.* which indicate a level of confidence within a prose etymology, but there is not a clearly defined model of the scope for such qualifiers. In the model of Campanile and Zampolli 1973, a confidence

level has a clearly defined scope: it applies to exactly one claim of cognation between exactly two words, where one such claim is encoded per punch card. Neither system defines a system of inheritance or scope for confidence levels across multiple claims.)

Second, it might initially seem like an attractive and obvious move to structure cognate sets around reconstructions. This would be similar to the strategy in synchronic dictionaries of using the citation form of a word as a rigid designator for the entire lexical item (perhaps appending an integer to disambiguate homonyms); the inflected forms are organized as dependents of the citation form. We might naively adopt a parallel structure between etyma and reflexes:

A false parallelism:



However, reconstructions change over time. The PIE word which was once reconstructed as *patḗr or *pətḗr "father" is now standardly reconstructed as *ph_2tḗr (but many older references containing the earlier reconstruction are still in use). Further, at any point in time, two competent historical linguists might have differing views which lead to differing reconstructions. This variation is the norm, not the exception. For this reason, it would be unwise to use a reconstruction as the rigid designator for a cognate set. Further, it would be desirable for a markup scheme to allow multiple versions of a reconstruction to be simultaneously included (both *pətḗr and *ph_2tḗr, for example), with information on the authority for each. Also, in some cases, an author might want to indicate that words are cognate but omit the implied etymon.

Cognation among substrings

The cognate relationship can apply between entire words, but it can also apply between substrings of words. For example, the *were-* in Modern English *werewolf* is cognate with Latin *vir* "man". It's not entirely accurate

to say that *werewolf* is cognate with *vir; vir* is cognate with only a substring of *werewolf*.

Compounding and derivational morphology make this a common situation. A markup scheme should provide a way to morphologically decompose words and discuss the etymology of each separately.

A complication is that morphemes are not always linearly contiguous. For example, concatenation followed by metathesis can result in a situation where the segments comprising the reflex of one morpheme are no longer a contiguous substring:

PIE	*kom + *-yós (cf. Latin cum, co(n)-)
Proto-Greek	*κομjός (*mj > *nj; *nj > *jn)
Classical Greek	κοινός "common, shared in common"

A similar problem can happen when two segments are fused into one. These cases can be handled by doing the decomposition on an etymon where the morphemes are still contiguous, which is the solution I adopt below. However, it does not address the more general problem of non-contiguous morphemes such as circumfixation or templatic morphology. I do not see an immediate solution to the problem and am open for suggestions.

Qualifications

Some claims of cognation require qualification. For example, a pair of words might derive from different morphological forms of the same root in the parent language. Or, a set of words might be obviously cognate, but one of the words might contain an unexplained anomaly. A markup scheme might define a standard battery of qualifications of this kind (and I am open to suggestions on what a standard list should contain, beyond the two cases just given).

There might also be provision for qualifications which are specific to a particular language family. For example, in Proto-Indo-European, there is a common ablauting pattern where a stem can occur in e-grade, o-grade, or zero-grade (*e.g.* *seng^{wh}-, *song^{wh}-, *sn□g^{wh}- "to chant", which are the etyma of English *sing sang sung*). It is common for one ablaut grade to survive in one branch of IE, and another ablaut grade of the same word elsewhere. Qualifiers indicating these specific ablaut grades would be use of

to the Indo-Europeanist, but would obviously be of no use to specialists in other language families. Since it would be impossible to anticipate all of the qualifiers which would be needed for every language family, the set should be extensible.

It is probably also desirable to allow the nature of the etymological connection to be partially or fully unspecified. There might be a relationship between two words whose semantics are "these words are in some kind of etymological relationship whose nature is not specified."

Designing a markup

Like Ide et. al. (2000), I discuss an abstract mathematical model for entries, hierarchical in form, which is not committed to any particular concrete markup scheme. Any context-free markup will do, but my own choice of would be XML (Bray et. al. 2004), because of its general flexibility and the availability of tools. For legibility, I will use an indented form to illustrate the mathematical model rather than XML.

A basic cognate set

```
cognate-set
     word
          form: hound
          language: Modern English
     word
          form: Hund
          language: Modern German
     word
          form: hunds
          language: Gothic
     word
          form: hundr
          language: Old Icelandic
     etymon
          form: hundaz
          attested: no
          language: Proto-Germanic
```

Note that the `etymon` node (and its descendants `form, attested,` and `language`) is optional. If it were omitted, then the claim is merely that the

four attested words are cognate. This claim of cognation implies that there is an earlier form in some parent language from which the listed forms are inherited, but the form of that word and/or the identity of the parent language can be left unstated.

No privileged frames of reference

A body of etymological claims can be seen as a network of connections of various types between lexical items, often between lexical items in multiple languages. Under the present proposal, one can enter this network at any point. There are no privileged frames of reference.

The following three examples encode precisely the same claims, but they do so from different vantage points. A program's internal representation of the etymological relationships would be the same regardless of which example it was given as input.

1. (From the vantage point of Modern English) Modern English *stone* is a reflex of Old English *stān,* which is a reflex of Proto-Germanic **stainaz:*

```
word
     form: stone
     language: Modern English
     etymon
          word
               form:     stān
               language: Old English
               etymon
                    word
                         form: stainaz
                         language: Proto-Germanic
                         attested: no
```

2. (From the vantage point of Old English) Old English *stān* is an etymon of Modern English *stone*, and is also a reflex of Proto-Germanic **stainaz:*

```
word
     form:     stān
     language: Old English
     reflex
          word
               form: stone
               language: Modern English
     etymon
          word
               form: stainaz
               language: Proto-Germanic
               attested: no
```

3. (From the vantage point of Proto-Germanic) Proto-Germanic **stainaz* is an etymon of Old English *stān,* which is an etymon of Modern English *stone:*

```
word
     form: stainaz
     language: Proto-Germanic
     attested: no
     reflex
          word
               form:     stān
               language: Old English
               reflex
                    word
                         form: stone
                         language: Modern English
```

Inheritance of attributes

Ide et. al. (2000) define a formal model for dictionary entries. Their model is hierarchical, and permits attributes to be associated with nodes in the tree. A child node implicitly inherits all of the attributes of its parent. However, a child node can explicitly *overwrite* the parent's value for an attribute. In such a case, the new value (not the old value) propogates to all descendants of that node. I adopt this property into the model I propose here.

Example: the Germanic verb meaning "to grip" is a strong verb (class I) in all of the early Germanic languages, as it was in Proto-Germanic. However, it changed to a weak verb in Modern English (*grip/gripped,* not *grip/*grope/*grippen*). In the following example, the "`morphological-class`" property of the "`cognate-set`" node is inherited by all of the descendants of this node. Thus, the Gothic, Old Icelandic, Old High German, and Old English words are all implicitly specified by inheritance as strong verbs of class I; the Proto-Germanic etymon inherits this attribute as well. However, the "`word`" node for the Modern English word overwrites the "`morphological-class`" property of its parent.

```
cognate-set
     morphological-class: strong verb, class I
     etymon
          form: grīpanã
          language: Proto-Germanic
          attested: no
     word
          form: greipan
          language: Gothic
     word
          form: grīpa
          language: Old Icelandic
     word
          form: grīfan
          form: crīfan
          language: Old High German
     word
          form: grīpan
          language: Old English
          reflex
               word
                    form: grip
                    language: Modern English
                    morphological-class: weak verb
```

If the `word` node for the Modern English word had further descendants, then those nodes would inherit the `weak verb` attribute.

Inheritance of claims of confidence

Example: Everybody accepts that Old English *cēn* "torch" is cognate with Old High German *kēn* "torch". Everybody accepts that Russian *sosná*

“pine” is cognate with Polish *sosna* “pine”. Hirt (1931, cited in Ringe 1984) claims that the Germanic grouping is cognate with the Slavic grouping, but Ringe (1984) rejects this claim (and rightly so; PIE *ḱ can come out as /s/ in Slavic, but its reflex in Germanic is *h by Grimm’s Law). Ringe (1984) actually doesn’t mention the Polish cognate for the Russian word, but for the sake of the example, let’s pretend that he mentions it and accepts it as cognate with the Russian:

```
cognate-set
     accepted-by: Ringe 1984
     word
          form: cēn
          language: Old English
          gloss: torch
     word
          form: kēn
          language: Old High German
          gloss: torch
     is-cognate-with
          accepted-by: Hirt 1931
          rejected-by: Ringe 1984
          cognate-set
               accepted-by: Ringe 1984
               word
                    form: sosná
                    language: Russian
                    gloss: pine
               word
                    form: sosna
                    language: Polish
                    gloss: pine
```

The “is-cognate-with” node overwrites the confidence attribute of its parent. The inner “cognate-set”, in turn, overwrites the confidence attribute of *its* parent. Thus, we capture that the Germanic cognate set and the Slavic cognate set are individually secure, but that the claim of cognation between the two sets is not secure.

Handling morphological decomposition

The following example encodes the claim that the *were-* of English *werewolf* is cognate with Latin *vir.*

```
word
     form: werewolf
     language: Modern English
     morphological-decomposition:
          morpheme-1
               form: were
                    comment: cranberry morpheme
               is-cognate-with
                    word
                         form: vir
                         language: Latin
                         gloss: man
          morpheme-2
               form: wolf
```

A case where metathesis has applied

```
word:
     form: κοινός
     language: Classical Greek, Attic dialect
     gloss: common, shared in common
     etymon
          form: κομjός
          language: Proto-Greek
          attested: no
          morphological-decomposition:
               morpheme-1
                    form: κομ
                    is-cognate-with
                         word
                              form: co(n)-
                              language: Latin
                              attested: yes
               morpheme-2
                    form: jός
```

This paper has sketched the broad outlines of a formal markup of etymological data. Obviously, many details have not been filled in, such as an exhaustive enumeration of the standard qualifiers and attributes, or the specification of default values for node attributes. Comments and criticisms are requested.

Appendix: Survey of the treatment of etymological data in existing markup systems

With regard to etymology, lexical markup schemes can be divided for convenience into three types:

> Type I. Markup schemes which make no provision for etymological data
> Type II. Markup schemes where etymological data is delimited as such, but is treated as unstructured prose
> Type III. Markup schemes where the mathematical relationships recognized in historical/comparative linguistics are somehow embodied in the markup system in (semi-)machine-readable form

Type I markups

Type I schemes are common and are probably the majority type. Studying these schemes does not reveal anything about the kinds of things that someone might want to mark up within an etymology, so I will not cite specific examples or consider this type further.

Type II markups

The TEI Guidelines

There are several markup schemes which allow etymological data to be treated as unstructured prose. Perhaps the best known is the TEI scheme (Sperberg-McQueen and Burnard, 2002), which is intended to accommodate a broad range of texts, of which print dictionaries are only one type. The TEI scheme was originally defined in SGML but has been adjusted to conform with XML.

The TEI standard includes an <etym> tag. The TEI documentation defines this tag as follows:

> The element <etym> marks a block of etymological information. Etymologies may contain highly structured lists of words in an order indicating their descent from each other, but also include related

> words and forms outside the direct line of descent, for comparison. Not infrequently, etymologies include commentary of various sorts, and can grow into short (or long!) essays with prose-like structure. This variation in structure makes it impracticable to define tags which capture the entire intellectual structure of the etymology or record the precise interrelation of all the words mentioned. It is, however, feasible to mark some of the more obvious phrase-level elements frequently found in etymologies, using tags defined in the core tag set or elsewhere in this chapter. (p. 299)

Obviously, there is a difference in philosophy between the *TEI Guidelines* and the present paper with regard to etymologies, motivated by differing goals. It may well be true, as the above paragraph states, that there will never be a markup scheme capable of encoding every scrap of etymological information in machine-readable form. But this is surely just as true of synchronic lexical data; every synchronic markup scheme leaves a residue of information which must be swept into comment fields as prose, or omitted. Lexical data is multifarious, but there is enough regularity in lexical data for there to have been considerable success in deploying machine-readable lexicons.

The TEI definition for <etym> goes on to list sample TEI tags which may be of use within a prose etymology:

> <lang> The name of any language mentioned in the prose
> <date> A date in any form, with attributes to indicate the calendar system, a standardized form of the date, and the degree of certainty of the date (which can have "any appropriate value" such as *ca., approx, after, before*)
> <mentioned> "marks words or phrases mentioned, not used"
> <gloss> Defines some other word or phrase
> <pron> Pronunciation
> <usg> Usage information
> <lbl> A label such as "abbreviation for", "contraction of", "literally", *etc.*

Of these tags, only <etym> and <lang> are not declared in other sections of the TEI specification.

The text gives the following example of the use of the <etym> tag:

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...

```
<entry>
        <!-- ... -->
        <etym>
                <lang>F</lang> fr. <lang>ML</lang>
                <mentioned>pneuma</mentioned>
                <mentioned>neuma</mentioned> fr. <lang>Gk.</lang>
                <mentioned>pneuma</mentioned>
                <gloss>breath</gloss>
                <xr type=“etym”>more at <ptr target=“pneumatic”/></xr>
        </etym>
        <!-- ... -->
</entry>
```

Zhang (1995)

Zhang (1995) discusses dictionary entries as tree structures which lend themselves to markup in SGML. Zhang includes an <etym> tag, but the contents of this element are not explicitly defined. Within a sample entry, the following use occurs:

```
<etym> ... <lg> lat. </lg> ... </etym>
```

The <lg> element can also occur within other elements such as <form> and <sense>. Although etymologies are not otherwise specifically discussed, it appears that Zhang is assuming a model like that of the *TEI Guidelines* where etymologies are treated as a kind of prose.

Bell and Bird (2000)

Bell and Bird discuss some of the variation in dictionary entry structure, and seek to define a “general purpose data model for lexical entries”. Under the model they adopt, an entry is primarily divided between **head** and **body**. The **body** can contain one or more of five types of element, of which **Aux** is one type:

> **Aux** contains the various types of miscellaneous information which may be included in an entry. This includes such things as etymology, obsolescence, cross-references, register, informant identity, and so forth. Some, such as **obsolete**, are marked by a binary attribute; others, such as **Etymology**, will need to allow prose within, and hence are sub-elements of **Aux**.

Thus, a prose model for etymological data is assumed.

Ide, Kilgarriff, and Romary (2000)

Ide et. al. (2000) criticize the general approach to formalizing dictionary structure of which the *TEI Guidelines* are an instance. Schemes such as *TEI* are informed both by study of variation in the structure in existing printed dictionaries, and by the requirements of a particular markup format such as SGML. Ide et. al. argue that a markup scheme should be an instance of a clearly defined mathematical model of dictionary entry structure.

Ide et. al. therefore define a model in which the elements of a dictionary entry are hierarchically organized, with an explicit model of attribute inheritance. A child node automatically inherits the attributes of its parent (for example, two numbered definitions might each inherit the same headword and pronunciation from their parent node). However, a child node can overwrite the parent's value for an attribute. They provide the following example:

> gendarme (...) n.m. (XV°; *gendarmes;* de *gens,* et *arme*) ... II. (1790) *Mod.* Militaire appartenent à...

Within the tree structure for this entry, the first etymology field (XV°; *gendarmes;* de *gens,* et *arme*) is an feature of a node near the top of the tree. The numbered definitions (I, II, *etc.*) are daughters of this node, and thus inherit the etymology feature from the parent. However, definition II contains its own etymology feature, namely (1790). This feature overwrites the etymology value of the parent for this particular node. Any lower nodes beneath this node inherit the new value, not the old value.

While Ide et. al. assume what is essentially a prose model for etymological information, there is an interesting system of attribute inheritance which allows etymological data (among other attributes) to be explicitly structured

in interesting ways for which most markup schemes make no provision. I have adopted this property of their model into my proposal.

Type III markups

The markup schemes which I label as "Type III" are those which make some provision for the formal encoding of etymological relationships between words.

Campanile and Zampolli (1973)

Campanile and Zampolli develop an etymological encoding scheme for their study of the lexicon of Old Cornish. The authors perform a number of statistical analyses over the data from an etymological dictionary of Old Cornish which has been encoded on punch cards. The columns on the cards are as follows:

A) a non-Cornish word, and the language to which the word belongs
B) a Cornish word in some kind of relationship with the word in **a**
C) the type of relationship between a and b; and whether the relationship is affirmed, denied, or uncertain
D) a binary field which indicates whether the Old Cornish word is a nominal compound (which Campanile and Zampolli claim to be the only kind of compound in Old Cornish)
E) a breakdown of the elements in a nominal compound, if D is true (otherwise field E is empty, presumably)
F) the page number of the dictionary from which the information was taken

The values for column C are:

1 = The relationship between the two words is etymologically certain.
2 = The relationship between the two words is etymologically very probable
3 = The relationship between the two words is etymologically probable
4 = The relationship between the two words is etymologically doubtful
5 = The relationship between the two words is etymologically not very probable
6 = The relationship between the two words is etymologically improbable

7 = The relationship between the two words is etymologically non-existent
8 = The Cornish word was borrowed from item A
80 = The Cornish word is a calque[1] on item A
82 = The nature of the relationship between A and B is undetermined and could be either of cognation or borrowing
9 = A and B are cognate ("co-radical"), but A is in a Celtic language
0 = B is not Cornish, but is a Cymric word which has "crept" into the Old Cornish glosses.

An obvious modern criticism of this encoding is that the data are shoehorned into an overly rigid columnar form which would not scale to more general applications. However, given the extreme limitations of the technology, it was certainly defensible to sacrifice generality in favor of compactness, encoding one claim per punch card. The resulting schema is quirky, but unlike all of the "Type II" markup schemes, it is obviously designed around the particular considerations of historical/comparative linguistic study. Many of the kinds of information are ones which we might want to include in a modern markup scheme in a more general form:

- Distinction between cognate, loan, and calque
- Degree of certainty
- Morphological decomposition
- Bibliographic information

The Oxford English Dictionary

It appears that the OED uses an <ET> field to delimit etymologies (Blake 1992; Amsler and Tompa 1998).

Stubbs (1985) states that the following items are to be delimited within the OED description of a loan:

- source of loan (presumably meaning both an indication of the donor language, and the form of the word in the lexicon of the donor language)

[1] A calque, or loan translation, is a special kind of borrowing where the morphemes of a foreign word are individually translated into native morphemes. An example is Latin *com+passio*, a calque on earlier Greek *sym+pathia* (both meaning "with" + "suffering").

•date of first citation
•author of first citation
•part of speech
•"subject label" (Stubbs does not explain this term, but it probably refers to the domain in which the word is used, such as Geology or Music)

Stubbs describes various queries which might be run over these data. For example, one might extract the dates and donor languages for all of the loan words into English from Native American languages. From the names of the donor languages, inferences could be made about the geographical area where the loans occurred. The historical patterns of impact of Native American languages in time and geography could thus be deduced.

Coward and Grimes (1995)

Coward and Grimes document the Multi-Dictionary Formatter, a software package intended to aid in the creation of printed dictionaries. Their data format develops on earlier work; it provides a number of tags related to etymologies:

\et Delimits a reconstructed etymon of the headword. Coward and Grimes stipulate that this field should be used only for previously published reconstructions. They recommend that that the \nt (general notes) or \ec (etymology comment) fields should be used to "posit your own guess at a reconstruction," and caution the user: "There is a whole science to the principles and procedures of comparative and historical linguistics, and simply trying to work from what looks obvious can quickly get one mired in muck."

\eg Gloss of an etymon found in \et

\es The bibliographic source for the etymon in \et

\ec Comment on the etymon in \et.

\bw "Borrowed words". This contains the name of the donor language, such as Arabic, and can optionally contain the form of the word in the donor language. In one example, this field contains the

string “Arabic via Malay fi:bahasa” (where “fi:” indicates italic font, thus mixing semantic and presentation markup).

The format of the dictionary entries is a single column where each row begins with a field label such as \bw, and where a value associated with that label follows. Some field labels can occur more than once in an entry, but in none of the sample entries do etymology-related fields occur more than once each.

Amsler and Tompa 1998

Amsler and Tompa survey the state of disarray among markups in existing monolingual English dictionaries, and propose an SGML-based markup standard for this specific purpose. The authors describe in some detail the problems in determining the structure of etymology fields:

> “Thirdly, there is the problem of our own shortcomings in understanding the structure of a dictionary’s entry, when the documentation of that structure is so sparse. In our work to date, this problem was apparent when designing the encoding for etymologies. In spite of well-written prefatory material in the several dictionaries ... and extensive reference books about lexicography and computational linguistics ..., we were unable to uncover a definitive description of the structure within a typical etymology (*e.g.*, the meaning of punctuation symbols and the scope of language names). The solution we adopted is to include tags for *etymons* (the word forms, with language as an attribute), *etymological units* (**eu**, the equivalent of a lexical entry, including form, pronunciation, meaning, and so forth) and *etymological segments* (**es**, branches of a hypothetical universal etymology tree, including information about the relationships **rel** among the components). We must wait for other experts to help us determine whether or not this organization is adequate for the standard.”

Amsler and Tompa propose the following elements:

> <E> encloses the entire etymological section of a dictionary entry
>
> <epart> etymology of one variant form of the entry

<es> etymological segment

<eu> etymon unit

<etymon> a word, morpheme, or phrase cited in an etymology
attribute: **lang** the name of a language

<rel> relation name (*e.g.* fr[om], *etc.*)

<cert> degree of certainty (*e.g.* prob., ?)

<basis> basis for the etymologist's belief (*e.g.* "by folk etymology", "assumed", "according to")

The following sample etymology for the word *apple* is provided:

```
<E>
    <es>
        <etymon lang=ME>appel</etymon>
    </es>
    <es>
        <rel>fr.</rel>
        <etymon lang=OE>&aelig;ppel</etymon>
    </es>
    <es>
        <rel>akin to</rel>
        <eu>
            <etymon lang=OHG>apful</etymon>
            <deftext>apple</deftext>
        </eu>
        <eu>
            <etymon lang=OSlav>abl&breve;ko</etymon>
        </eu>
    </es>
</E>
```

This markup can be seen as a sort of middle ground between the fully prose, human-readable etymologies of Type II markups, and a markup intended to put etymological information into machine readable form. The units here are specific to etymologies, unlike the elements which can occur within the

TEI <etym> element. On the other hand, it would be fairly difficult for a program to recover the fact that the Middle English word is a reflex of the Old English word, or that the English words at all three stages are cognate with the Old High German and Old Slavic words. A program would probably have to use various error-prone heuristics to extract this information.

Good and Sprouse (2000)

The Comparative Bantu Online Dictionary (CBOLD) is a complex database on multiple Bantu languages. As a starting point, the team digitized existing print dictionaries and word lists, and marked up these texts according to the *TEI Guidelines*. The team added a few tags to the standard set to accomodate the particular needs of their project.

The team prepared a standardized list of standardized Bantu reconstructions. The <etym> field of an entry from a digitized dictionary can contain an <xr> element (a standard TEI element referring the reader to some other location in the same text or another text). The team use <xr> to create a pointer from a dictionary entry to an item in the standardized list of reconstructions.

Thus, Good and Sprouse's model is essentially a directed graph. An etymological claim is an arc of a particular type between two lexical items. From a practical standpoint, this model is convenient to implement, especially in the case where one is using digitized synchronic dictionaries as a starting point.

Since the model is non-hierarchical, I do not see an obvious way to implement a model of implicit attribute inheritance. It is certainly possible to associate e.g. a level of confidence attribute with an arc, but this would presumably need to be redundantly specified on each arc.

Jacobson and Michailovsky (2002)

Leaving aside details, the model of Jacobson and Michailovsky with regard to etymology is fundamentally the same as that of Good and Sprouse: an etymological claim is encoded as an arc of a particular type between entries in a lexicon.

Jacobson and Michailovsky's lexical markup scheme is XML-based, and includes some specific provisions for use by field linguists. Jacobson and Michailovsky borrow some of their tags from the large set in the TEI Guidelines, but note that TEI's provisions for dictionary markup are made with print dictionaries in mind. Some of the conceptual structure also borrows from the heritage going back to the 1980s which Coward and Grimes (1995) develop on.

Like Good and Sprouse, Jacobson and Michailovsky allow lexical entries to contain links to other objects, either inside the same document (<ptr>) or outside it (<xptr>). These elements can contain an attribute field indicating the type of pointer, and one of the permitted types is `cfetym`, an "etymological reference," perhaps to another entry in the same dictionary. Following is an example:

```
<ptr type="cfetym" target="tumma_2"/>
```

This is a reference to the verb meaning "to be mature," found within the same dictionary.

Since the headwords in most dictionaries typically represent a single stage in the history of a language, it could be argued that the connection in this particular case is one of synchronic derivation rather than historical etymology. This is not a criticism; the general model of Jacobson and Michailovsky could readily be used for claims which are unambiguously etymological.

Sources Cited

Amsler, Robert A. and Frank Wm. Tompa. 1998. A SGML-based Standard for English Monolingual Dictionaries. *Information in Text,* Proc. 4th Conf. of Univ. of Waterloo Centre for the *New OED* (October 26-28, 1988), pp. 61-80.

Bell, John and Steven Bird. 2000. *A Preliminary Study of the Structure of Lexicon Entries*. Paper presented at the workshop on Web-Based Language Documentation and Description, 12-15 December 2000, Philadelphia, USA.

Blake, G. Elizabeth, Tim Bray, and Frank Wm. Tompa. 1992. Shortening the OED: Experience with a Grammar-Defined Database. *ACM Transactions on Information Systems,* 10.3:213-232.

Bray, Tim, Jean Paoli, C. M. Sperberg-McQueen, Eve Maler, François Yergeau. 2004 *Extensible Markup Language (XML) 1.0,* Third Edition. http://www.w3.org/TR/REC-xml/

Campanile, Enrico and Antonio Zampolli. 1973. Problems in Computerized Historical Linguistics: The Old Cornish Lexicon.

Covington, Michael A. 1996. Alignment of Multiple Languages for Historical Comparison. *Proceedings of the 36th conference on Association for Computational Linguistics,* 1:275-280.

Coward, David F. and Charles E. Grimes. 1995. *Making dictionaries: A guide to lexicography and the Multi-Dictionary Formatter (Version 1.0).* Waxhaw: Summer Institute of Linguistics. ix, 234 p.

Good, Jeff and Ronald Sprouse. 2000. *SGML markup of dictionaries with special reference to comparative and etymological data.* Paper presented at the workshop on Web-Based Language Documentation and Description 12-15 December 2000, Philadelphia, USA.

Hock, Hans H. 1991. *Principles of Historical Linguistics,* second edition. Mouton de Gruyter.

Ide, Nancy, Adam Kilgarriff, and Laurent Romary. 2000. *A Formal Model of Dictionary Structure and Content.*

Jacobson, Michel and Boyd Michailovsky. 2002. *Linking Linguistic Resources: time aligned corpus and dictionary.* International Workshop on Resources and Tools in Field Linguistics, Las Palmas, Canary Islands, Spain, 26-27 May 2002.

Lowe, John. B and Martine Mazaudon. 1994. The Reconstruction Engine: A Computer Implementation of the Comparative Method. *Computational Linguistics,* 20.3:381-417.

Ringe, Donald. 1984. Germanic $*\bar{e}_2$ and **r*. *Die Sprache* 30:138-155.

Smith, Raoul N. 1969. *Automatic Simulation of Historical Change.* International Conference on Computational Linguistics.

Stubbs, John. 1985. The New Oxford English Dictionary and its Potential Users: Some Preliminary Comments. *Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval*, 78-81.

Sperberg-McQueen, C.M. and Lou Burnard, eds. 2002. *Guidelines for Electronic Text Encoding and Interchange,* Vol. 1. The TEI Consortium.

Trask, R. L. 1996. *Historical Linguistics.* London: Arnold.

Zhang, Jian. 1995. Application of OODB and SGML Techniques in Text Database: An Electronic Dictionary System. *Sigmod Record,* 24.1:3-8.