How to format an etymology: Some general principles, and some specifics for Toolbox users

by
David Mead

2010

**Sulang Language Data and Working Papers:
Topics in Lexicography, no. 2**

Sulawesi Language Alliance
http://sulang.org/

LANGUAGES

Language of materials : English

ABSTRACT

Based on practical experience working on two indigenous language dictionaries in Indonesia, these are my tips about what kind of information to include in the etymological portion of a dictionary entry, and how to format it in the Toolbox program.

TABLE OF CONTENTS

VERSION HISTORY

Version 1 [28 December 2010] This paper was largely completed in September 2005; slightly modified for publication December 2010.

How to format an etymology:
Some general principles, and some specifics for Toolbox users[1]

David Mead

> **et•y•mol•o•gy** the origin and historical development of a linguistic form as shown by determining its basic elements, earliest known use, and changes in form and meaning, tracing its transmission from one language to another, identifying its cognates in other languages, and reconstructing its ancestral form where possible.

I am not an etymologist. I do not study the history of words. However, in the course of compiling a vernacular language dictionary, I have found it impossible *not* to stumble across etymologies, or at the very least to note intriguing similarities between forms in different languages. Once these have come to our awareness, then the question becomes: will we record these insights for the next scholar down the road? Or will we let that information lapse, leaving it up to someone else to (re)make the connection?

This article is written from the perspective of someone who has started work now on his second vernacular language dictionary and—realizing the shortcomings of some of my first efforts—I wanted from the start to get things right this second time around. Consequently I have been giving some thought about how to format the etymological portion of a dictionary entry. Here’s what I’ve come up with.

Keep things simple

According to Nathan Bailey, author of *The Universal Etymological Dictionary* (1721), an etymology of a (English) word should include the following (cited in Landau 2001:128):

1. Source language or language family
2. First English form and/or immediate source
3. Date or period of entry into English
4. Changes in form and meaning in English
5. Intermediate stages
6. Ultimate known source
7. Semantic development
8. Ultimate underlying or hypothetical form, e.g. an Indo-European root
9. Cognates in related languages also derived from the underlying form
10. Other English words derived from the same base

Apart from perhaps the *Oxford English Dictionary*, however, there is no English language dictionary which even comes close to achieving such lofty goals, and how much less so our own work in vernacular languages. The byword that I use is: keep things short

[1] I have directed my comments to Toolbox users, because this is the program with which I am familiar.

and simple. Since my etymologies will never be ‘complete,’ I might as well aim for ‘short,’[2] since ‘intermediate’ is likely to satisfy no one!

Borrowed versus inherited words

Since its inception as a dictionary tool, Shoebox and its successor Toolbox has suggested that when a word is borrowed, the etymological information about where it came from, etc. needs to be kept distinct (in the \bw field) from the etymological information when a word is inherited (which information is to be placed in the \et field). There are some side advantages to this approach. For example, when you go to write the section in your phonology paper on the ‘Phonemicization of Loan Words,’ then you can very handily search in the \bw (borrowed word) field and draw all your examples from there, and not bother with the contents of the \et (etymology) field.

When it comes times to publish your dictionary, however, there are few advantages to this approach, since the source language will always make it clear whether a word was borrowed or directly inherited. For example, since Malay is not an ancestor language of Kulisusu (southeastern Sulawesi), any Kulisusu word which has Malay as its source must perforce be a borrowing. Conversely, if the source is given as Proto Celebic, a known ancestor of Kulisusu, then the word is inherited.

> **pusi** [Mal *pusing* ‘confused’]
>
> **poniana** [PCel **panianan* ‘parent-in-law’]

Source language, source form, and gloss

Both entries above illustrate the three most important parts of an etymological citation: the source *language*, the *form* in the source language, and a brief *gloss* for the source form. These three parts are also distinguished by formatting: the source language (or abbreviation thereof) appears with a capital letter; the source form appears in a particular typeface, usually italics; and the gloss appears in single quotes. Source forms which are proto forms (that is, reconstructed forms of a proto language for which we have no direct, written evidence) are also by convention preceded by an asterisk.

Do all three parts of an etymological entry need to be present? No, they don’t. The one part which *does* need to be there is the source language. In fact, in a very simple Bobongko lexicon which I prepared (of only about a thousand entries), I chose to indicate only the source language (Malay, Pamona or a Gorontalic language). In this case, a number of words had probably entered Bobongko from earlier stages of Pamona or

[2] According to one school of thought, bilingual and trilingual dictionaries do not need to include etymologies, because they are usually not of interest to their main audience (language learners). However, this assumes that there also exist monolingual dictionaries which *do* include etymologies. Since this is unlikely to be the case in the situations where we work, and furthermore our dictionaries are likely to become one of the foremost authoritative works on the language, it behooves us to include etymologies. However, even for ordinary persons etymologies can give a sense of history as well as demonstrate that languages ‘change’ and are not static.

Gorontalo, and reconstructing these earlier forms was simply beyond the scope of this limited project.

It is perhaps more common to include the source language and the source form, but to drop the gloss. However, you should be consistent about when you do this. For example you may decide you will leave all Malay forms unglossed, which may force some (but perhaps only a small minority) of your readers to consult a Malay or Indonesian dictionary. Or you may decide to drop the gloss when the vernacular and donor language forms agree very closely in meaning (as might be the case with certain plant or animal species). If a word you cite is from a lesser-known language, or in any case your reader is not likely to have access to a dictionary for that language, then this would indicate that a gloss should be included.

You will need some way of formatting the source form differently from the source language and the gloss. In the *Guide for Pacific Linguistics Dictionary-makers* it is suggested (page 10) that you use three separate fields, that is, one field for the source language, another field for the source form, and yet a third field for the gloss of the source form.[3] However, I have found this format too restrictive for presenting other kinds of information which I want to include in an etymology (see below). Therefore, within Toolbox I use |fi{ }, which is one of the supported ways of indicating character formatting (see further the Toolbox *User's Guide*, pages 240-241).[4] Thus corresponding to the two examples given above, in my Toolbox lexical database I have:

```
\lx pusi
\bw Mal |fi{pusing} ‘confused’

\lx poniana
\et PCel *|fi{panianan} ‘parent-in-law’
```

Using the word ‘from’

The word ‘from’ (or its abbreviation ‘fr’ or ‘<’) need not be included at the beginning of the etymology, since this can be assumed. But it is necessary when giving a more distant source, for example:[5]

> **dambu** [Mal *jambu* fr Skt *jambu* ‘a tree (SYZYGIUM)’]

[3] The Multi-Dictionary Formatter export process, on the other hand, allows for only one field for borrowed words (\bw) and two fields (\et etymology and \eg etymology gloss) for inherited words.

[4] Note that character formatting codes given in the Multi-dictionary Formatter user's guide (Coward and Grimes 2000:206) are *incorrect* for current versions of MDF.

[5] In Toolbox, entered as follows. Note the use of |fs{ } for scientific names.

```
\lx dambu
\bw Mal |fi{jambu} fr Skt |fi{jambu} ‘a tree (|fs{Syzygium})’
```

In some cases you may want to further indicate how exactly something is ‘from,’ e.g. ‘derived fr,’ ‘loan transl fr,’ ‘acronym fr,’ ‘clipping fr,’ ‘blended fr,’ etc. (if this would not be clear from ‘fr’ alone). For example:

uwi ngkeu [loan transl fr Mal *ubi kayu* ‘cassava’]

behaa [Mal *beha, B.H.* alphabetism fr Du *bustehouder* ‘brassiere’]

I also use the abbreviation ‘(<met.)’ to indicate that metathesis was involved in the process just presented, for example:

ti’olu [PMP **qiteluR* ‘egg’ (<met.)]

Hedging an etymology

Because the majority of us do not have the time, resources or expertise to research and verify the history of a word, in many cases we will either have to drop an etymology altogether, or we will have to hedge it. The simplest method would be to employ some special symbol (such as a cross, a question mark, or an ‘x’) at the beginning, indicating that the following etymology is doubtful. In my own work, however, I have decided to employ the following fuller set of conventions for giving hedges:

or	*or*	disjunctive possibilities
cf	*compare*	means that some etymological connection may be present, but requires further investigation
sim to	*similar to*	an etymological connection is even more tenuous than with ‘cf’
ult	*ultimately*	the intermediate pathway is not known

For example:

mbo’u [cf Pam *wo’u* ‘also, still, again’ ult fr PMP **baqeRu* ‘new’]

sarai [cf Pam *sarai* ‘a moment, an eyeblink’ sim to Sa’dan Toraja *sarra’i*
‘do quickly’ Rampi *meharai* ‘run hard’]

I also make use of the \ec (etymological comment) and \es (etymological source) fields, but mainly as non-printing fields (that is, to keep track of information which is useful to me in the process of compiling the dictionary, but which will not appear in the final, published version). You may decide for yourself to make the \es field a printing field. Be aware, however, that there are two kinds of ‘sources.’ One kind of source is whence you got a form in another language or proto language. The other kind of source is the source of an etymology itself, e.g. the person who first deduced or speculated about an etymological connection between two different forms. When someone else has led you to a particular etymology, attributing the etymology to them is both proper and a way to hedge your own work; let them take the heat if it’s wrong!

Special characters

When the occasional Arabic, Sanskrit or Chinese form needs to be cited, it is acceptable (and more helpful to most readers) to use a romanized transcription rather than Arabic, Devanagari, or Chinese characters. Two words of caution are in order, however. First, as much as possible, use a single source for your romanized forms, so that you do not inadvertently introduce competing transcription systems into your dictionary.[6]

Second, take an extra few seconds to check and make sure the form you cite has been correctly keyboarded. The transcription systems which have been developed for both Arabic and Devanagari make use of dots below consonants as well as macrons above vowels (and, in Devanagari, also above some consonants). Dots, macrons, tildes, acute accents, etc. should never be omitted. Note also that the first letter of the Arabic alphabet, *alef*, is usually transcribed with a right single quote,[7] while the eighteenth letter, *ain*, is usually transcribed with a left single quote.[8] Don't confuse these, nor let the 'smart quotes' of word processing programs mislead you!

Placement of the etymology

There are two schools of thought on where to place etymological information. In one school—and as in most English language dictionaries—the etymology follows closely after the head word. When placed here (so say detractors), the reader then has to 'jump over' the etymology to get to the information which he or she is really seeking. For this reason, the competing school of thought says to place the etymological information out of the way, at the end of the entry. In this position, however, it can sometimes appear to be part of the final subentry, when in fact the etymology applies to the head word and not the subentry.

Whatever one's opinion on this issue may be, Toolbox follows this second school of thought, and places the etymology at the end of the entry, which will therefore perhaps end up being the default for many of us.

If etymological information is always placed in such a position and in such a format so that the reader will know it is etymology, then in the final copy you can drop (via edit, replace…) the strings *Etym:* and *From:* which the Multi-Dictionary Formatter export process automatically inserts for you (to mark the contents of the \et and \bw fields respectively). Dropping these is in fact my own personal preference (note that I have

[6] This also applies of reconstructed (proto) forms. Particularly the field of Proto Austronesian studies is to be noted for different researchers promulgating different notational schemes.

[7] This transcription convention is ultimately the source for why glottal stop is represented by an apostrophe in the orthographies of not a few indigenous languages of Indonesia. Some Arabic purists prefer instead using a right half ring (Unicode character U+02BE). Whether you choose single right quote or right half ring, you should be consistent throughout all your citations.

[8] Some purists prefer instead using a left half ring (Unicode character U+02BF). Whether you choose single left quote or left half ring, you should be consistent throughout all your citations. See the preceding footnote.

done this in all the examples given above). Alternatively, you could replace *Etym:* and *From:* with a certain symbol (e.g. '<') which marks the following information as etymological. This may be more necessary when etymological information has been shunted off to the end of the dictionary entry. See Van den Berg's published (1996) *Muna-English Dictionary* for an example of this kind of formatting.

Etymologies for a national audience

As long as one is consistent and keeps careful track of what has been placed in the \et and \bw fields, then all the formulaic parts of the etymology can be translated into Indonesian using a consistent changes table, or even by using the Find, Replace… function of Toolbox. The following is a very simple consistent changes table,[9] which changes the abbreviation 'fr' (from) to the abbreviation 'dr' (dari), provided that 'fr' is encountered in a \bw or an \et field.

```
group(main)

 "\bw" > dup use(b)           c when matched goes to group (b)
 "\et" > dup use(b)           c where change will be made

group(b)

 " fr " > " dr " back(1)      c when matched makes change;
                              c must be preceded and followed
                              c by a space
 "\" > dup back(1) use(main)  c when another backslash is
                              c found goes back to main
                              c group so that changes are
                              c only made in the \bw and
                              c \et fields
```

The one part which could not easily be changed this way, however, are the glosses, and for this I don't have a simple solution. If glosses are not omitted altogether, then they would have to be changed manually by inspecting each etymology. However, in a vernacular language dictionary of some thousands of entries, only a few hundred will contain etymological information. Therefore making manual changes may not be an insurmountable task, especially if it is a one-time task just before publishing the dictionary.

Front matter

As you develop the conventions which you will use for formatting etymologies, and while all the choices you made are still fresh in your mind, this is the best time to write up a paragraph or two for the front matter of your dictionary. This is where you explain to your reader how you have formatted the etymological portion of a dictionary entry. This section will also end up being helpful to you, because—as you work on a dictionary off

[9] I would like to thank Barbara Altork for her assistance in developing this consistent changes table.

and on over a span of perhaps several years—you can refer to it from time to time to make sure you are being consistent with your own conventions.

Whatever abbreviations you use in your etymologies, make sure that these are entered in your list of abbreviations which also appears in your front matter.

The front matter is also an appropriate place to make some disclaimer, e.g. how you wish you had had more time to properly research etymologies. Nonetheless, never be ashamed, and never doubt the helpfulness of simply citing similar forms in closely related languages. You will thereby give the real etymologist or historical linguist valuable threads to follow.

References

Berg, René van den. 1996. *Muna-English dictionary*. Leiden: KITLV Press.

Coward, David F. and Charles E. Grimes. 2000. *Making dictionaries: a guide to lexicography and the Multi-Dictionary Formatter*, PDF version. Waxhaw: Summer Institute of Linguistics.

JAARS, Inc. 2000. *The linguist's shoebox: tutorial and user's guide*. Waxhaw: JAARS.

Landau, Sydney. 2001. *Dictionaries: the art and craft of lexicography*, 2nd ed. Cambridge: Cambridge University Press.

Pacific Linguistics. 2001. A guide for Pacific Linguistics dictionary-makers. Unpublished typescript, 16 pp. http://pacling.anu.edu.au/for_authors/for_authors.html (accessed December 14, 2005).