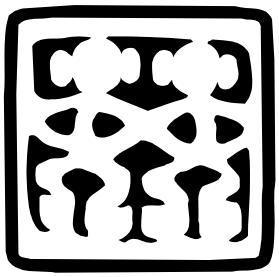
Draft only

A guide for Pacific Linguistics dictionary-makers



Pacific Linguistics

October 2001

Contents

Overview

This guide has been prepared to help you in the preparation of a dictionary for publication by Pacific Linguistics. It should be read in conjunction with *A guide for Pacific Linguistics authors* (http://pacling/for_authors/guide_for_PL_authors.pdf).

The first part of that guide deals with the stages of book production. These are the same for a dictionary, except that we ask you to send us a sample of your dictionary in advance of the other stages of book production (see below).

The second part of with *A guide for Pacific Linguistics authors* describes PL's wordprocessing requirements. These are the same for a dictionary as for any other book, except where otherwise noted below.

The third part of that guide gives details of PL style. Much of the information given there also applies to a dictionary, especially to its introductory part (see below).

Sending us a sample

If you intend to submit a dictionary to Pacific Linguistics for publication, you should send us a sample of at least 500 entries from both the vernacular–English and English–vernacular sections together with a copy of the introduction to the dictionary. The sample should be in accord with the guidelines below. A member of our Editorial Board will check the sample and, more often than not, will make recommendations for improvement. Obviously, it is in your interests to have these recommendations to hand before you do the final editing of your dictionary for submission.

A dictionary's parts

Most dictionaries fall into two parts—an introductory part and the dictionary proper part.

The introductory part

Despite its name, the introductory part is usually like a small book in its own right, with many of the components discussed in the section on PL style in *A guide for Pacific Linguistics authors*. That is, it will have its own

- preliminary material, including titlepage, dedication (if any), contents, foreword (if any), preface, acknowledgements, list of abbreviations, and possibly lists of maps, figures, tables (in this order);

- main body, consisting of an introduction to the dictionary and containing such things as information about how to use the dictionary (with explanation of

symbols and abbreviations), cultural information and maps, a phonology and an explanation of the orthography used in the dictionary, and a grammar sketch;

- end matter, consisting at least of a list of references.

Because these components are the same as those in other books, the information about them in the section on PL style in *A guide for Pacific Linguistics authors* also applies here.

The dictionary part

The dictionary part of a PL dictionary should include at least the first two of the following:

- the dictionary proper: vernacular–English (or vernacular–English–national language)

- a finderlist: English–vernacular (or two finderlists: English–vernacular and national language–vernacular)

- a thesaurus: entries organised in terminologies, e.g. kin terms, buildings, fish, animals and so on

The rest of this guide is devoted to the dictionary part and its production.

Data collection, analysis and compilation

We are often asked for advice about the techniques of compiling a dictionary, and these brief notes are intended to answer some of the questions most often raised.

It is important to keep separate the processes of (i) data collection, analysis and compilation and (ii) formatting the dictionary for publication. However, the second is very much dependent on the first: if you have not planned the first stage adequately, you will run into difficulties at the second stage.

The first set of decisions that needs to be made with regard to data collection, analysis and compilation concerns the categories of information to be collected. Remember that the lexicon (vocabulary) of a language is simultaneously the repository of a great deal of cultural information, as it encapsulates the ways that its speakers see the world, as well as the set of building blocks from which the clauses and sentences of the language are constructed.

Cultural information

Cultural information is encoded in a number of ways. The most important is perhaps that lexical items form parts of terminologies. The simpler terminologies are those which have to do, for example, with the parts of a traditional house: roof, ridgepole, rafter, crossbeam, housepost, balcony, door, ladder, and so on. Somewhat more complicated are layered texonomies, for example of edible plant products. Thus in English—for some native speakers in Australia, at least—edible plant products are divided in the first instance into *fruit* and *vegetables*. *Fruit* is for most speakers probably a more or less indivisible category: it simply has members like *apple, pear, peach, mango, pineapple* and so on. *Vegetables*, on the other hand, are divided for some speakers into *root vegetables, leafy vegetables* and various edible plant products like asparagus that do not fit into the other categories. For some speakers, too, there are edible plant products like the tomato whose membership of either *fruit* or *vegetables* is questionable. The most difficult terminologies often consist of verbs with similar and sometimes overlapping meanings. Oceanic Austronesian languages sometimes have a number of verbs of cutting: ‘chop with an adze’, ‘cut off, sever’, ‘cut off (softer objects like hair, taro tops)’, ‘clip off (protruding growth)’, ‘lop off (branches)’, ‘cut lengthwise’, ‘cut up (meat)’, ‘cut into, incise’, and so on.

Another kind of cultural information is encoded in complex lexical items (see below): spkakers of a language which lexicalises ‘I fell over’ as ‘the ground hit me’ may perceive causation rather differently from speakers of English (although one would want much more evidence than a single lexical item to make a case for this).

Terminological information is typically captured in a dictionary in two ways. One is in entries in the dictionary proper, the other in a thesaurus.

In entries in the dictionary proper, it is often helpful to relate a word to others with which it has a close terminological relationship by

- indicating the terminology to which it begins—if there is a thesaurus, this can be by a crossreference to the relevant thesaurus section;
- indicating the position of the item in the taxonomy to which it belongs: it may seem trivial to define a potato as 'k.o. root vegetable', but, for example, some Oceanic Austronesian languages have a complex taxonomy of marine life, and part of the definition of a marine item is to place it in its indigenous category;
- particularly in cases like the cutting terms mentioned above, indicating synonyms, near-synonyms, and other cutting terms from which the item being defined as a clear difference in meaning.

Grammatical information, complex lexical items

Ironically, many linguists treat lexicon as peripheral to the main business of linguistics, yet a number of modern theories of syntax are 'lexicalist'. That is, they claim that it is the valencies (combinatorial possibilities) of words that determine how clauses are constructed, and that many seemingly syntactic processes, e.g. passivisation in English, are actually the manifestation of partially predictable relationships in the lexicon (between *break* and *broken*, *love* and *loved*, *hit* and *hit* [!]). There is good evidence that linguistic systems are not quite this simple, but part of this evidence also has to do with lexicon. In any language there are probably thousands of 'sentence stems'—parts of sentences that consist of a sequence of lexical items, like *keep an eye on* or *keep tabs on* ('observe'), *have a seat*, *tell the truth, kick the bucket* ('die'). Each of these is a complex lexical item, in the sense that its existence and its use cannot be predicted from its components, and does not necessarily engage in the lexical relationships of its component parts (*kick* can be passivised, but *kick the bucket* can't). We sometimes find fully lexicalised clauses like *be that as it may* and lexicalised sentences like *Long time no see* or *A stitch in time saves nine*. Many of these complex lexical items go under the rubric of 'idioms' in the literature.

Ideally, all this information should be included in the dictionary proper. The introductory part will include a sketch grammar, which will indicate the most important and predictable lexical relationships in the language, and the entry for *break* should tell the reader that it is the active form of a transitive verb, and that the corresponding passive participle (as described in the grammar sketch) is *broken*. The dictionary proper should also include sentence stems, together with a note of their grammatical limitations, as well as lexicalised clauses and sentences. Obviously, measured by this standard, most dictionaries will be significantly incomplete, but they are more likely to approach completeness if these factors are recognised from the beginning.

One question which arises with regard to complex lexical items is where to put them. Given that most dictionaries are organised as an alphabetised sequence of simple lexical items, each complex item is usually included under a simple item, typically the semantically narrowest component of the complex item. That is, *keep an eye on* and *keep tabs on* will be listed under *eye* and *tabs* respectively, rather than under *keep*, which is semantically broader and contextually less constrained (crossreferences may also be given under *keep*).

A word of warning: some dictionary-makers, especially some with anthropological rather than linguistic training, neglect the recording of grammatical information because it seems esoteric. The result is a dictionary (and there are a number that we know of) that is linguistically of little use because essential information about how words are used is missing.

Categorising lexical information

It should be clear from the discussion above that there are a number of potential categories of information to be collected for each lexical item. These include (in their typical dictionary order rather than in the sequence of the discussion above):

the headword;

its pronunciation (this is unnecessary if the practical orthography is such that the pronunciation can be deduced from the description in the introductory part of the dictionary);

part of speech and valency (in terms of the categories outlined in the grammar sketch, e.g. intransitive/transitive/ditransitive verb, relational noun, and so on) and lexical relationship (e.g. **broken**, past participle of **break**);

dialect, speech register or other sociolinguistic information;

gloss (definition), including

- scientific name;
- terminology membership (including crossreference to thesaurus), position in taxonomy, synonyms, near-synonyms and differences in meaning from them, antonyms;
- explanatory notes;

etymology (the source of the item, if known);

example(s) with free translation (a literal translation may also be useful to the reader);

subentry for a complex lexical item which contains the headword (e.g. *keep an eye on* under the headword *eye*); this may also include

- literal (morpheme-by-morpheme) translation;
- valency and other grammatical information (e.g. cannot be passivised);
- dialect, speech register or other sociolinguistic information;
- gloss (definition);
- example(s) with free translation (a literal translation may also be useful to the reader);

This listing is a simplification in two ways. First, because words are often polysemous, many dictionary entries will contain more than one gloss, and each gloss will have its own example(s) and sometimes its own subentry. Secondly, under a given gloss there may be more than one subentry (with all the components a subentry brings with it).

The complexity of this listing demands that data collection, analysis and compilation be very systematic. The best way to guarantee that it is systematic is to use a skeleton entry which includes fields for each of the kinds of information which potentially needs to be collected.

The majority of linguists who produce dictionaries today use a computer program to do so. This is usually a database program in which each dictionary entry has a set of fields which form the skeleton of the entry. It is possible to use either a dedicated linguistic database program, like *Shoebox* from the Summer Institute of Linguistics (http://www.sil.org) , or a relational database program like *Filemaker Pro*. Either way, one has to create one's own skeleton, although there is ample advice on how to so this with *Shoebox*.

One advantage of *Shoebox* is that other linguists have used it in the past and it is possible to draw on their experience. Another is that *Shoebox* can also be used for the interlinearisation of texts, and produces a basic dictionary at the same time. This basic dictionary can be integrated into or used as the basis of a larger dictionary. A third advantage is that *Shoebox* keeps its files in text format, so that they can be read by other programs (e.g. word processors), and this guarantees greater stability of data storage than in some proprietary database programs. Finally, *Shoebox* is accompanied by the *Multi-dictionary formatter*, which allows you to produce formatted (RTF) output from your database (not that this is the guarantee of a perfect dictionary: the output will still need to be checked).

A *Shoebox* file is a text file in which one field is separated from the next by a carriage return, and one entry is separated from the next by two or more carriage returns. Each field is labelled with a 'backslash code'—a code consisiting of the backslash immediately followed by one or more letters indicating the contents of the field. The examples below are based on entries from a *Shoebox* file containing a database of Mauwake (a Papuan language of the Madang Province, Papua New Guinea) dictionary entries compiled by Liisa Järvinen of the Summer Institute of

Linguistics. The codes used in the example below are: `lx` lexical entry/headword, `lc` related lexical item, `ps` parsing/grammatical category, `ge` English gloss, `re` entries for reversal/finderlist, `xv` (vernacular) example, `xe` English translation of example, `cf` compare/see also, `dt` date entry made:

```
\lx aakun
\lc aakuniya
\ps v2.tr
\ge talk
\re talk ; discuss ; speak
\xv Wi iperowa-ke aakunep maemik, ...
\xe The middle-aged men discussed and (then) said, ...
\xv Aakunem-ikaiwkin wia miimam.
\xe As they were talking I heard them.
\xv Yena koora-pa Mauwake opora aakunimik.
\xe In my house/home we speak Mauwake.
\cf maakiya, maiya
\dt 20/Oct/2000
```

Note that two examples (`xv` and `xe`) are given here.

Not all these fields will appear in the published dictionary. The `dt` (date entry made) field is for the dictionary-maker's personal use. The `re` (entries for reversal/finderlist) will be used to generate the finderlist, i.e. it tells the program to list **aakun** under *talk, discuss* and *speak* in the finderlist.

As we mentioned above, some items may have more than one gloss, each followed by its own set of fields. In the example below, the field `sn` contains a numeral indicating the start of a new gloss. It occurs twice here. Under one gloss we also find the field `sd`, which contains a thesaurus category (here `bp` 'boddy parts'). This will also be used in organising the thesaurus.

```
\lx afifa
\ps nn
\sn 1
\sd bp
\ge hair
\re hair
\xv Emeria nain afif maneka, me puukiya.
\xe That woman's hair is long, she doesn't cut it.
\sn 2
\ge feather
\re feather
\dt 26/Oct/2000
```

We also mentioned above that an entry may contain subentries for complex lexical items. The entry below (which we have truncated, as it contains numerous banana varieties each containing the word **akia**) contains three subentries, each introduced

by the field `se`. It also makes use of the field `sc`, containing a scientific name, and three instances of the field `ee`, which contains explanatory notes.

```
\lx akia
\ps nn
\sd pl
\sd fo
\ge banana
\re banana
\sc Musa spp.
\se akia aroguma
\ps np
\ge banana sp.
\re banana_sp
\ee (Dark red skin.)
\se akia biliwa
\ps np
\ge banana sp.
\re banana_sp
\ee (Sweet, eaten without cooking.)
\se akia enuma
\ps np
\sd fo
\sd pl
\ge banana sp.
\re banana_sp
\ee (Cooking_banana.)
```

If these entries are compated with the listing above, it will be clear that other fields could be added: for example, for the source of a borrowing (the name of a neighbouring or colonial language), for the borrowed word (the form in the source language) and for the meaning of the borrowed word in the source language. Note that three separate fields will be needed here because of the probability that different fonts will be used for them in the final output. Thus a number of Papua New Guinea languages have borrowed the Tok Pisin auxiliary form *mas*. This would show up in a dictionary as something like

Source: TP *mas* 'have to'

The point of presenting this collection of Mauwake examples is not to recommend that you use *Shoebox*, nor to provide a 'how to do it' example, but to give an indication of why a database approach is needed in compiling a dictionary. Whether one uses a linguistic or a proprietary database program, the database approach has several advantages:

- it provides the possibility of collecting parallel information for every lexical item;
- if it is set up appropriately, it guarantees consistency, as you can use the program to limit the choices available to you;

- it may allow automated formatting of the dictionary for publication (although the formatted version will still need to be carefully checked).

Occasionally, would-be dictionary-makers tell us that they are compiling their dictionary with a word processor like Microsoft Word. This is fine, as long as a database skeleton is being used. But all too often it is clear that the dictionary-maker is conflating the two stages of data collection/analysis/compilation and formatting the dictionary for publication, and this may well lead to the neglect of detail in the first—and scholarly—stage.

Formatting the dictionary for publication

The dictionary proper

Organisation of entries

The organisation of entries varies from dictionary to dictionary (depending on the nature of the language and the amount of information to be included) but we use the following layout style as a guide to setting out entries in (bilingual) dictionaries (triglot dictionaries are discussed below).

> **headword**, *part of speech,* (dialect, speech register or other linguistic information). 1. gloss. *Example* with free translation (followed by literal translation if one is to be provided). **Idiom** with *example* and free translation. 2. gloss. *Example* with free translation. **Idiom** with *example* and free translation. 3. gloss. *Example* with free translation. [*scientific name;* synonym or antonym]

With respect to this note that:

- The headword is bolded and is in lower case (unless it is a proper name). It is also separated from what follows by a comma.

- The part of speech is abbreviated (with a full stop) and italicised. It is also separated from what follows by a comma unless there is no bracketed information following, in which case there is no comma (because there is a full stop after the part of speech).

- Dialect, speech register and/or other linguistic information is abbreviated (with a full stop) (e.g. E. = eastern dialect; coll. = colloquial, pl. = plural).

- Glosses are numbered from '1' onwards where there is more than one gloss. In that case we use a full stop after the number indicating the particular gloss. We also use a full stop at the end of the preceding gloss, unless there is already a full stop or an equivalent punctuation mark there (e.g. ! or ?). For example:

  > **duringa**, *n.* 1. base (of something). 2. stump (of tree). 3. reason. *Pinu duringa ditimopa?* Do you know why (lit. the reason) he did that? 4. meaning.

- Examples are italicised and are normally punctuated as for English, that is, if the example is a sentence it begins with a capital letter and ends with a full stop.

- Literal translations follow free ones and are given in parentheses.

- Idioms associated with a gloss are bolded and illustrated in the same way as glosses are.

- Scientific names are given in square brackets in italics with genera names capitalised and species names in lower case, e.g. [*Hibiscus tililaceus*]. Where the species name is not known or is regarded as irrelevant the abbreviation for 'species' is not italicised, e.g. [*Hibiscus* sp.].

The following examples illustrate these principles:

bukako, n.,hibiscus [Hibiscus tiliaceus]

mama, n., (pl. **mamuhe**), father. da mame my father

toroka, adj., 1. strong, hard. idi toroka strong (or hard) wood. 2. difficult (to understand). A votere dahina torokamavavaho. Your speech is very difficult for me to understand. n., 3 strength, power no mame ahu torokave God's (lit. our father's) power

va[1], n., 1. sky. Vare ubuiamima. It's getting dark (lit. sky is darkening). 2. rain. Vare dobivima. It's raining. **Vare hivavanu**. It's lightning. **Vare kukuvanu**. It's thundering (or it's threatening to rain).

va[2], n., fighting, warfare. Yabu va otinua. They went fighting (or making war).

va[3], post., 1. to. Yagavage da otima. I'm going home. 2. from. Igauge Haverivage ahu orovonu. One came from the Haveri clan. 3. at. Horehe, da yagevago. It's up there, at my house. 4. with, by means of. Idiva to vama! Hit the dog with a stick!

For triglot dictionaries we suggest the following variations:

headword, *part of speech,* (dialect, speech register and/or other linguistic information). 1. gloss (in English)/*gloss (in Tok Pisin, Bislama, Indonesian or other third language)*. **Example** with free translation in English/*in Tok Pisin, Bislama, Indonesian or other third language*) (followed by literal translation if one is to be provided in English/*Tok Pisin, etc*. **Idiom** with example and free translation (as for examples above)

For example:

mama, *n*. (pl. **mamuhe**), father/*papa*. **da mame** my father/*papa bilong mi*

va[1], *n*. 1. sky/*heven* **Vare ubuiamima**. It's getting dark (lit. sky is darkening)/*Tudak is kamap*. 2. rain/*ren*. **Vare dobivima**. It's raining/*Ren i pundaun*. **Vare hivavanu**. It's lightning/*Klaut i ait*. **Vare kukuvanu**. It's thundering (or It's threatening to rain)/*Klaut i pairap* or *Ren i laik pundaun*.

Homophonous words

Homophonous words are treated as separate entries and are superscripted. For example:

aba[1], *n.,* bush turkey
aba[2], *n.,* hole

Notes on definitions

Different glosses of the same word Where there are gloss discriminations based on grammatical category all glosses are numbered consecutively and all glosses of the same category are grouped together with only the first preceded by the category. For example:

mas, *n.* 1. a charm...2. ... 3. ... *v.* 4. attract by beauty

*Capitalisation of names of plants and animals*In general, species names should be capitalised. This will sometimes result in a word being capitalised when it occurs in the name of a species, e.g. Barking Deer, Jungle Cat, but not when it is used as a common name (deer, cat). This runs against the tendency amongst scientists not to capitalise, but is appropriate in contexts like the following:

> The area inhabited by deer and sandpipers is...The Barking Deer and Marsh Sandpiper can be seen in...Where this latter species and the Red-fronted Lorikeet, the Little Red Lorikeet and the Western Black-capped Lory occur...The Red Deer is not a barking deer and does not...

However, our policy is not rigid: our main concern is that the dictionary is consistent.

*Borrowings*If an item is identified as borrowed, not only the name of the source language but, if possible, both the form and meaning of that item in the source language should be given. This allows readers to see the relationship between the forms and meanings of the borrowed item and the source. If both form and meaning are identical in the two languages, then there is no need to supply both but there should be a covering statement in the introduction drawing the reader's attention to this convention.

Avoiding etc. *in dictionary entries*Our policy is to avoid the use of *etc.* either by using other words like 'for example' or by giving more examples in the list that *etc.* is supposed to cover.

The finderlist

Every dictionary should have an English–vernacular finderlist or index. This finderlist is, as the name suggests, a guide to finding information in the vernacular–English section. It is not an English version of the vernacular–English

part. As a result the entries are much simpler than those in the vernacular–English section. For example, they do not repeat part-of-speech information, examples, synonyms and other details given in the vernacular–English part. They merely list the vernacular items (headwords, subentries, idioms) that the author thinks the reader should check to find the vernacular word or words that best fit(s) the meaning s/he has in mind. In presenting these you should follow the following principles.

English headwords should be in bold and the corresponding vernacular items separated from them by two spaces (without punctuation) and italicised. For example:

door*va*

Where there are subentries these should be listed under a generic headword, indented two spaces and typed in roman (not bolded). For example:

door
door of house*va*
doorway or opening*va mauk*

Where it is necessary to distinguish different parts of speech in English subentries explanations, adjuncts or abbreviations can be used, for example, if one wants to distinguish between different forms of 'cross' this could be done in one of the following ways:

cross
cross (oneself in prayer), make sign of the cross*sibitava*
cross over (bridge)*segoti*
make s.o. cross*snragi*
(a) cross *sibita*

cross
cross (v.)*segoti*
cross (n.)*sibita*

Alternative vernacular items are listed one after the other and separated by (roman) commas. For example:

house*umah, hausi, yaga*

Note that a finderlist is not simply an automated reversal of vernacular–English entries. An automated reversal is only a beginning, and needs to be edited to make sure it enables readers to find what they are looking for in the vernacular–English part of the dictionary. The finderlist is often the weakest point of a dictionary manuscript. Consequently publication is delayed while the manuscript is sent back to the author for revision.

The thesaurus

[This section is still to be constructed]

Format

We normally set dictionaries in two columns. However, there may be cases where other formats are acceptable. In those cases authors should discuss their ideas with us well before submitting the manuscript.