

# Handbook for dictionary consultants and compilers

## SIL Philippines

This handbook was initially prepared in 2001 by Scott Burton in consultation with Len Newell, and edited in 2003 by Doug Trick.

**Introduction:** These are preliminary guidelines and helps to enable compilers of lexical materials to more efficiently meet their goals and to assist and enable consultants in guiding language teams as they prepare their dictionaries and bring them to completion. The purpose and goal of this handbook is to help guide teams to produce quality and useful dictionaries.

Because the compilation of a dictionary is such a large project, it is crucial that a compiler have a fairly clear sense of direction in the very initial stages. Otherwise, there is a very real danger that work will progress but in the wrong direction, and the further it progresses, the greater will be the eventual loss. It is also important that each of the individuals involved in the project have a clear understanding of what is expected of them; it can be very disheartening after having worked hard and long, and thinking that the work is almost complete, to be told that there is more to be done (e.g. a grammar sketch).

### I. Stages of checking

It is recommended that a consultant interact with the compiler(s) at least twice during the course of the compilation. The first check (referred to in this handbook as the Initial Check) should be conducted after 300-1000 entries have been entered; the second check (referred to here as the Intermediate Check) should take place approximately one year before anticipated publication. (One final pre-publication check should be done immediately before publishing, though it may be that this final check will be conducted by a copy editor and not necessarily by a lexicography consultant.)

The consultant should specify well in advance the format in which the compiler should submit the material to be checked. In most cases, the entire current dictionary file in electronic format should be made available to the consultant; this will normally be in database format (using “backslash codes”). In addition, for both the Initial Check and the Intermediate Check, the compiler should submit to the consultant at least 300 entries as a formatted document. This formatted document should include a good cross-section of verbs, nouns, adjectives, particles, affixes, conjunctions, and some phrases (e.g., idioms, sayings, or set phrases). Several of the entries should illustrate various lexical functions such as generic-specific (e.g., various ways of ‘carrying’ or of ‘cutting’), part-whole (e.g., parts of a body, or of a house), synonyms, antonyms.

### II. Procedures

The consultant should discuss with the compiler(s) at least the following topics. Some of these will be discussed in greater detail during the Initial Check, but even during the Intermediate Check they should be briefly reviewed.

#### ***Purpose and audience:***

1. What is the purpose of the dictionary? Who is the target audience? For example, is it a learners’ dictionary (primarily for outsiders engaged in learning the language and translating materials into the language)? Or is it being compiled primarily for the benefit of native

- speakers? Have various potential audiences (e.g., primary audience, secondary audience, tertiary audience) been identified? Will the dictionary be bilingual, trilingual? etc.
2. Has any testing been conducted with respect to the format of the published volume? (It is usually advisable to format a sample of the dictionary in two or three different ways and ask for members of the target audience to react to the different formats.) If this has not been done, are there definite plans to do this?
  3. What is the specific target dialect to be documented in the dictionary? If dialect variants will be included, how is this to be done?

### ***Language-related issues:***

4. What is the orthographic form used for the vernacular language? (Is it essentially phonemic<sup>1</sup>? If not, what are the major ways in which it differs from a phonemic form?) Are any “special characters” to be used (including typographical characters such as em-dash, en-dash, non-breaking spaces and hyphens)?
5. What is the grammatical form of the dictionary entries? For example, are the entries organized according to roots, or roots plus stems, or all forms which occur most frequently in the language?
6. How will the entries be organized? The most common format is alphabetic; if this is used, should the alphabetic arrangement correspond to that used by a regional or national language?

### ***Front matter / Back matter:***

7. What material is being planned for the volume in addition to the dictionary proper? (For example, a brief description of the orthography and a list of abbreviations is essential. Other valuable material might include a grammar sketch, ethnographic sketch, reversal index<sup>2</sup>, bibliography.) The consultant may need to see some of this material (e.g. list of abbreviations) in order to conduct the Initial Check. (The compiler and consultant should also have a clear understanding as to who is responsible to check this additional material; some dictionary consultants prefer to not check a grammar or ethnographic sketch.)

### ***Lexicography issues:***

8. Does the compiler have a grasp of basic lexicography issues? For example, in considering two apparently identical forms, it is essential to be able to distinguish whether they are homonyms, or they illustrate multiple senses of a single lexeme. (And of course the compiler must know how to encode the material accordingly.) If the compiler has had no specialized training in lexicography, what possibilities are there for providing such training?
9. How are procedural decisions recorded? For example, the compiler will have to decide on how to distinguish “loan words” and whether or not to include them in the dictionary. If illustrative sentences are given, are they included for every major entry, or just for some (and in that case, for which entries)? Because the compilation of a dictionary may take several

---

<sup>1</sup> This would include consistent symbolization of glottal stop, even in word-initial position, and explicit marking of stress.

<sup>2</sup> For example, if the body of the dictionary lists vernacular entries and gives English definitions, a reversal index will list English words and their corresponding vernacular entries.

years, it is essential that the compiler records such decisions and is able to refer to them from time to time.<sup>3</sup>

### **Dictionary entries:**

10. What methods were used to determine which words are entered into the dictionary? Is there a significant corpus of natural text material on which the dictionary is based? If not, the consultant should (before or during the Initial Check) interact with the compiler(s) on this issue, and be prepared to give some assistance in how to begin developing and analyzing such a text corpus.
11. Was a standard set of format markers (SFM's, or "backslash codes") used? Was each one used according to its intended purpose, and consistently? Do they occur in the correct order within each record? (These are extremely important issues. Inconsistencies here are often not obvious in the database format, but they will cause major problems when attempting to convert the database to any other format.) Dictionaries to be processed by SIL Philippines should use either Philippine Branch standard backslash codes, or else LinguaLinks standard codes. (MDF is another possible system of codes; however, Philippines Branch Academic Publishing is currently not able to support this system.) An SIL computer program, FINDSFM.COM, can be used to generate a list of all SFM's in a file.
12. What measures were taken to ensure consistency within each field? For example, if the computer program, Shoebox, was used, were range sets employed for fields such as "part of speech"?
13. In most cases, each entry should include a pronunciation field. The pronunciation field should use standard IPA symbols, and indicate syllable breaks, stress, and/or length. The pronunciation field is not necessary if the orthography used for the target language is consistently phonemic, or if it is nearly consistently phonemic and the few exceptions are clearly stated in the front matter. (Rationale for including a pronunciation field is to provide necessary linguistic information to researchers doing comparative and historical linguistics.)
14. Most dictionaries will make use of minor entries<sup>4</sup>; these document variant forms, and irregular inflected or derived forms. These should be included unless the compiler has valid reasons for not including them. Every minor entry must be cross-referenced to a main entry; a significant number should be checked to ensure consistency.
15. A consultant should be aware of certain terms which are widely considered to be "politically incorrect," and ensure that such terms are not used inappropriately<sup>5</sup>. The consultant should advise the compiler to investigate whether there might be legal restrictions on the publication of certain types of indigenous cultural information.
16. The consultant must check definitions with special care, since these are among the most important features of the dictionary. In most cases, a good definition will refer to a genus (a

---

<sup>3</sup> If the compiler is using Shoebox, version 5 or later, some of these decisions can be recorded as part of the "description" information for "marker properties" (under "database type properties"). This facilitates very easy access to such notes.

<sup>4</sup> A "minor entry" is different from a "sub-entry". In database terms, a minor entry is a separate record (like a main entry – though with only a subset of the fields usually included in a main entry). A sub-entry, on the other hand, is included within a main entry. For example, an American English dictionary might have *aluminium* as a minor entry, considering it as a (British) variant form of *aluminum*; *aluminum* would be a main entry. On the other hand, the form *pianist* might be a sub-entry under the main entry, *piano*.

<sup>5</sup> (In some settings, this might include terms such as *primitive*, *folk doctor*, *folk tale*, *myth*, *fiction*, *tribe*, *native*, *costume*, etc.)

term slightly more generic than the particular entry), along with relevant criterial features. It is recommended that the genus and criterial features be highlighted (e.g., by bolding or underlining). Where an illustrative sentence is given (in the vernacular), the lexical form being defined, and its corresponding gloss (in the translation) should be similarly highlighted. (This highlighting will greatly facilitate the dictionary user in finding the most relevant information.)

17. A good definition is accurate and concise without being vague, and avoids ambiguous terms. Normally a single gloss should not combine both active and passive English constructions.
18. The consultant should check that there is adequate correlation between the citation form, its part of speech, and the lexical form which occurs in illustrative sentences.
19. Terms that are used in Scripture to refer to key theological concepts should be included in the dictionary. However, translated portions should not normally be used as illustrative sentences.
20. Definitions (and translations of illustrative sentences) should avoid the use of loan words which tend to become part of the active vocabulary of the expatriate community within the host country. (In the Philippines, this would include terms such as *viand*, *merienda*, *carabao*, *tsinelas*, etc.) Outside the host country, these terms may have no meaning, or a different meaning.
21. A good illustrative sentence should be formed in such a way as to reinforce the definition (including the range of meaning) of the entry. When possible, the source for illustrative sentences should be natural text material. If the illustrative sentences were composed specifically for the dictionary, much care should be taken to ensure that they are well-formed, natural examples.
22. Complex forms, or “derivations” (especially those with prefixes) should be handled in a consistent manner. (Note: this paragraph is still “under construction; see footnote at end of paragraph.) Such a form may occur as a minor entry (in which case it must be cross-referenced to at least one corresponding main entry). It may also occur as a sub-entry (that is, as part of a main entry). If it is probable that a user of the dictionary may have difficulty finding a particular complex form under its main entry, then it should be entered as a minor entry. Many such derivations will be entered both as minor entries and also as sub-entries under a main entry<sup>6</sup>.

---

<sup>6</sup> (From Scott Burton): Using the Philippine Branch standard format markers, a derived form usually occurs after a sense number (**\ms**) and following **\ode**, **\oco**, **\oid** or **\ose** fields. The derived form should be put alphabetically following **\lx** as a minor entry and cross-referenced to the root where it is described. If more than one root is involved, list the form at the end of the entry (**\lx**) of the second (third, etc.) root following **\ld**, **\lc**, **\li**, or **\ls**. There should not be a full description following an **\l?**. If the derivation is particularly uncommon, it should be listed following **\l?** with a simple gloss (and no illustrative sentence).

(Further note from Len Newell): Under "Dictionary entries", #22, the suggested handling of complex lexical forms seems to be different from how I would handle them. As I see it, there are two possibilities. 1) The complex form is fully describe as a "sense" of the main entry (with or without a sense number) of its included root (or one of its included roots in the case of compounds, idioms or set expressions). If necessary for convenience of finding the form, it would occur as a minor entry and referred (including a sense number if there is one) to the main entry. 2) The complex form appears alphabetically and is fully described as a major entry. In the case of derivatives, the entry form should be immediately followed by a root form (**\rf**) and derivational affix(es) (**\ad**). At the end of the entry of the root (if the root occurs as a major entry form) the derivative (as well as compounds, set expressions and idioms) would be cited as a sub-entry (an alternate term: run-on-entry). This would be a very brief description (not a full description). The use of sub-entries is "to display the morphological and grammatical structure of complex forms."

23. The lexemes of a given language are interrelated in numerous networks, and a dictionary should normally reflect some of these relationships, called “lexical functions” (e.g., synonyms, antonyms, generic-specific, part-whole). Each such lexical set must be consistently cross-referenced. That is, if Entry A indicates that Entry L is a synonym, then Entry L should indicate Entry A as a synonym. Failure to consistently cross-reference these relationships is one of the most common problems in compiling a dictionary.
24. For many lexemes, collocation information should be provided. (E.g., a language may have a generic word meaning ‘wash’ which collocates with many objects, but not with clothing – i.e., there is a specific word meaning ‘laundry’ which must be used to describe washing clothing.) The consultant should ensure that this type of detail is not overlooked.
25. In most cases, a bilingual or multilingual dictionary should have at least one index. (For example, a vernacular-to-English dictionary should have an English-to-vernacular index.) There are various ways to keyboard information from which such an index can be extracted; the consultant should ensure that a consistent method has been used throughout the dictionary.
26. The consultant should interact with the compiler about spell-checking. If the dictionary has been compiled in database format, then fields which consist primarily of English (or other major-language) data can be exported and then spell-checked with commercially-available programs. SIL’s Unispell program can be used to spell-check vernacular data.

The following page provides a suggested check-list which a dictionary consultant can (adapt and) use to record consultant-check progress for a particular project.

---

(Handbook on Lexicography, Sec. 8.3.5) Some compilers confuse the use of \l... fields and \o... fields and, in some cases, use only \l... fields where the appropriate field would be \o... .

## Dictionary Consultant Check-List

Language Name \_\_\_\_\_ Compiler(s) \_\_\_\_\_

Anticipated Publication Date: \_\_\_\_\_

Date of Initial Check: \_\_\_\_\_ Format for material to be submitted: \_\_\_\_\_

Material submitted for Initial Check: \_\_\_\_\_

Date of Intermediate Check: \_\_\_\_\_ Format for material to be submitted: \_\_\_\_\_

Material submitted for Intermediate Check: \_\_\_\_\_

|     | Topics discussed with compiler(s)   | Initial<br>Check | Intermediate<br>Check |
|-----|---|------------------|-----------------------|
| 1.  | purpose, target audience(s), languages employed in the dictionary                 |                  |                       |
| 2.  | format testing completed  |                  |                       |
| 3.  | target dialect; how to handle dialect variants                                    |                  |                       |
| 4.  | orthographic form: phonemic? "special characters" (incl. typesetting char's)?     |                  |                       |
| 5.  | grammatical form of entries: roots? roots and stems?                              |                  |                       |
| 6.  | organization of entries: alphabetic? correspond to regional/national language?    |                  |                       |
| 7.  | front/back matter – what's to be included? who will check it?                     |                  |                       |
| 8.  | compiler has good grasp of lexicography issues                                    |                  |                       |
| 9.  | procedural decisions recorded and easily accessible                               |                  |                       |
| 10. | sources: substantial text collection? other sources?                              |                  |                       |
| 11. | consistency in SFM's: each record internally consistent                           |                  |                       |
| 12. | each field internally consistent (e.g. punctuation; part-of-speech abbreviations) |                  |                       |
| 13. | pronunciation field included?   |                  |                       |
| 14. | minor entries included, and consistently cross-referenced to main entries         |                  |                       |
| 15. | politically correct terminology   |                  |                       |
| 16. | well-structured definitions; consistent special formatting conventions            |                  |                       |
| 17. | definitions accurate and concise  |                  |                       |
| 18. | correlation between citation form, part of speech, and form used in examples      |                  |                       |
| 19. | terms used for "key theological concepts" are included in the dictionary          |                  |                       |
| 20. | English definitions, etc. use good, standard English                              |                  |                       |
| 21. | illustrative sentences reinforce definitions; well-formed, natural examples       |                  |                       |
| 22. | complex forms correctly handled (and cross-referenced)                            |                  |                       |
| 23. | lexical networks ("lexical functions") adequately described                       |                  |                       |
| 24. | collocational information provided where necessary                                |                  |                       |
| 25. | adequate indexing provided  |                  |                       |
| 26. | provision made for pre-publication matters (e.g., spell-checks)                   |                  |                       |