Title: The development of bilingual dictionaries at MRM Inc.

By: Dr. R. David Zorc
Senior Linguist
MRM Inc.
3910 Knowles Avenue
Kensington, MD
USA 20895

Phone: (301) 864-1411
Fax: (301) 864-8956

Abstract:

This paper summarizes the development of bilingual dictionaries at MRM Inc., specifically that of the Somali-English Dictionary (on an IBM) and the Armenian-English Dictionary (on a Mac). In the genesis of these projects, the database had to be defined (and redefined) to include numerous fields. When any new field is introduced, editing of previous entries becomes an arduous and time-consuming task. It is hoped that by reviewing all the fields we have found necessary, we can help other lexicographers in the planning stages of their projects to determine which may be essential to them (and why), and allow them to proceed without extensive back-tracking at a later date. Besides tables and printouts illustrating various screen entries, a computer demonstration will be available.

The development of bilingual dictionaries at MRM Inc.

R. David Zorc
MRM Inc.

1. BACKGROUND INFORMATION

MRM Inc. is a language research firm located in Maryland just outside of Washington, DC. The corporation includes: the **Chinese Research Center** (which has developed 250,000 entry Chinese-English lexical databases of general and technical terms), the **Language Research Center** (among whose publications are a *Somali-English Dictionary* (now in its third edition), an *Uzbek-English Dictionary*, a *Tagalog Slang Dictionary*, with a forthcoming *Armenian-English Dictionary* and a *Kazakh-English Dictionary*), the **Optilex Division** (which has developed a data-management system for retrieval and display of dictionary databases and released a CD ROM version of the Chinese-English Dictionary for DOS on 1 April 1994), and **Dunwoody Press** (which publishes foreign language materials).

The above-mentioned databases have been developed on both Apple Macintosh computer systems (utilizing 4th Dimension Software) and on IBM/DOS computer systems (utilizing Revelation Software). During the development of the Somali-English Dictionary (on an IBM) and the Armenian-English Dictionary (on a Mac), the database had to be redefined to include several fields that proved essential, but were not originally projected. When any new field was introduced, editing of previous entries became an arduous and time-consuming task. It is hoped that by reviewing all the fields we have found necessary, we can help other lexicographers in the planning

stages of their projects to determine which may be essential to them (and why), and allow them to proceed without back-tracking at a later date.

2. LIST OF POTENTIAL FIELDS

The following is a comprehensive list of the fields that we found necessary in all of our lexicographic projects combined. Although not all are in use in any single database, the researcher should consider the necessity of each for any given project.

1. **Headword** (the primary dictionary entry key field)

1a. **True headword** (that part of the headword which may result in a different alphabetical sort, e.g., Somali preposition + verb phrases like **is dhaafso** *swap, exchange with each other* is best alphabetized under **dhaafso**, not **is**).

1b. **Headword tags** (affixed forms that help the user identify the headword more accurately, e.g., gender affixes, case suffixes, plural forms, past tense verb, etc.)

1c. **Sort field** (an indication to the computer what sort order is intended, e.g., the alphabetical order or ASCII codes of non-Roman scripts such as Armenian, Azeri, Georgian, Chinese Pinyin, etc.)

1d. **Phonetic representation** (Romanized transliteration for languages with their own script for users who are beginning in the language and may need this assistance).

1e. **Homograph Number** (for homonyms or homographs).

2. **Part of Speech** (as determined by the needs of the language under study, usually includes noun (+ declension number), verb (+ conjugation number), adjective (pure or derived), positional (preposition or postposition), adverb or adverbial expression,

question word, conjunction, discourse particle, or interjection).

3. **Definition** (English gloss, translational equivalent, or descriptive explanation).

3a. **English key** (key English word for sort purposes or for an English to target language glossary).

3b. **Sense Number** (when polysemy is extensive).

4. **Example** (for appropriate context, special uses, idiomatic phrases, etc.)

5. **Cross Reference** (to synonyms, antonyms, alternates, hyponyms, related entries).

5a. **Computer Cross-Reference** (inflected forms that have undergone morphophonemic changes for database use only, e.g., Somali **biyo** *water* has common forms like **biyaha** *the water*, **biya** + modifier phrase, whereas **bil** *month* has alternates like **bishan** *this month*, which, with an appropriate command, can be used to retrieve the respective headword).

6. **Edit notes** (items or problems to be followed up by the editor(s)).

7. **Sources** (reference to newspaper article, grammar, other dictionary, etc.).

8. **Codes** (may include frequency information from available frequency lists, monolingual dictionary consulted, etc.).

9a. **Created by** (the computer automatically keys in initials of editor or consultant as per log-on information).

9b. **Created on** (the computer automatically keys in today's date).

10a. **Modified by** (the computer automatically keys in initials of editor or consultant as per log-on information when he/she makes any modifications).

10b. **Modified on** (the computer automatically keys in the date when any revisions are

made).

11. **Etymology** (may include dialect, loan status, or etymological information).

12. **Slop** (material such as complex cross-references not to be included in this edition, but which may be useful in a later or more thorough edition).

3. THE SOMALI-ENGLISH DICTIONARY

The first edition was based on a coordinated translation of the 18,000-entry monolingual *Qamuuska* (Keenadiid, 1976) and was edited by Virginia Luling (1987). It had just five fields, called: **headword** (1), **also** (1b), **tag** (2), **English** (3), and **Resources** (7).

That database proved helpful for statistical and grammatical work on the *Somali Textbook* (Zorc & Issa, 1990), and grew by some 6,000 entries as a result of additional research. In the development of the second edition (Zorc, Osman & Luling, 1991) the grammatical field (2) was incremented to include the eight noun declensions and five verb conjugations elucidated by Saeed (1987) in the hopes of allowing the user greater control over the complex inflections and morphophonemics of the language. To familiarize the user with synonyms, antonyms or related words, we added a **cross-reference** field (5). We also included information about the dialect provenance of some words and the donor language for identifiable Arabic, Italian and English loans (11). As a result of the necessity to move preposition + verb phrases around manually in the text-editing process, a **true headword** field (1a) was introduced to allow sorting by computer. We introduced a **date** field (10b) so we could track how

many entries were edited (altogether and on a monthly basis) for completing monthly project reports, as well as **keyboarder** (9a), and **edit notes** (6) fields to keep staff abreast of problems encountered. Dating (10b) also helped if minor shifts in methodology were introduced so we could back-track to entries within a given time frame; those entries without a date also represented the number left to go prior to completion and final printout. Fields within Revelation can be indexed by the computer for instantaneous access by a backslash (\), but extensive indexing costs valuable save time, so only the Somali and English data were computer indexed (including a new field (5a) to find common misspellings and morphophonemic alternates). While this latter does not appear in print, it helps any database user find a greater variety of Somali vocabulary, such as that which may be encountered in the press. Other searches (which take about 20 minutes each) could be accomplished by straightforward commands [like: find tags containing "n2" (second declension noun) or "adj-der" (derived adjective), or keyboarder not containing "DZ"].

Continuous requests for an English to Somali resource led to the third edition (Zorc & Osman, 1993) with an English index. Since the computer cross-referenced every occurrence of an English gloss, such that **gayax** *ocean fish sp*. appeared at OCEAN, FISH and SP., which therefore led to the need for heavy text editing to insure appropriate matches between English and Somali, we have seen the need for an **English key** field (3a) for other projects.

4. THE ARMENIAN-ENGLISH DICTIONARY

The first 4,500 entries of this database were the result of a search through Armenian newspapers. Since the monolingual resources have in excess of 150,000 words, it was felt this procedure would yield current and useful vocabulary. This has since been supplemented by an Armenian consultant's selection of an additional 10,000 headwords from a 20,000 entry frequency dictionary and/or abstraction from the massive monolingual dictionaries available.

From the onset most of the fields listed in section 2 were utilized except:

1a - **true headword** is not necessary, since Armenian phrases generally wind up being "alphabetically correct".

1b - **headword tags** are not being introduced in this preliminary edition, but inclusion of information on the eight possible noun declensions is planned for a later stage.

3b - **sense number** is currently being handled by the use of commas and semicolons to separate polysemous glosses.

5a - **computer cross-reference** could later be used to tie compounds to their composite roots, especially in less transparent cases when vowel loss occurs.

One of the problems encountered involved developing a reliable program to isolate duplicate headwords. Since the consultant was initially working through newspaper articles, some vocabulary items would be encountered again and again, but with most of the Armenian characters being in the upper ASCII range (where computer codes also lurk) numerous bugs had to be worked out so that previously entered material could be correctly called up and rechecked for possible new senses or

examples, as well as addition of source information. Other problems involved developing a sort program to respect Armenian alphabetical order and yet not distinguish between capital and lower-case letters.

The above summarizes some of our problems and refinements in the computerization of bilingual databases. Neither time nor space allow a digression here into the insights and experience we have gained in the linguistic aspects of bilingual lexicography. Suffice it to say that while I came to MRM with a background in Philippine (Aklanon Bisayan) and Australian Aboriginal (Yolngu Matha) dictionary work, and despite the unique demand of each language, there was considerable carry-over in the approach to use and user, handling of exotic language material, translational equivalents and explanatory glosses, etc. What has changed is the amount of data one can usefully handle in a relatively short time. The 11,000-entry *Aklanon-English Dictionary* (Zorc, 1969) took four years, dozens of informants, and ten shoeboxes of data slips; whereas the miracle of computer processing allowed the *Somali-English Dictionary* (op.cit) to achieve 26,000 entries within three years with three consultants, and the *Armenian-English Dictionary* to reach 15,000 entries within two years with two language researchers.

5. REFERENCES.

Baghdasarian, Louisa and R. David Zorc. In progress. *Armenian (Eastern)-English Dictionary*.

Keenadiid, Yaasiin C. 1976. *Qaamuuska Af-Soomaaliga* (A Dictionary of the Somali Language). Mogadisho.

Luling, Virginia. 1987. *Somali-English Dictionary*. Kensington, MD: Dunwoody Press.

Saeed, John I. 1987. *Somali Reference Grammar*. Kensington, MD: Dunwoody Press. (Second Revised edition, 1993)

Zorc, R. David, Vicente Salas Reyes, et al. 1969. *Aklanon-English Dictionary*. Kalibo, Aklan.

-- 1986a. *Yolngu-Matha Dictionary*. Batchelor: School of Australian Linguistics.

-- 1986b. "Linguistic 'Purism' and Subcategorizational Labels in Yolngu-Matha," *Lexicographica* 2:78-84.

--. and Abdullahi Issa. 1990. *Somali Textbook*. Kensington, MD: Dunwoody Press.

-- and Rachel San Miguel. 1991. *Tagalog Slang Dictionary*. Kensington, MD: Dunwoody Press.

-- with Madina Osman and Virginia Luling. 1991. *Somali-English Dictionary* (Second Revised and expanded edition). Kensington, MD: Dunwoody Press.

-- and Madina Osman. 1993. *Somali-English Dictionary with English Index*. (Third edition). Kensington, MD: Dunwoody Press.

Appendix 1: ARMENIAN DICTIONARY DATABASE PROJECT SUMMARY

PURPOSE: to expand the 4,500-entry wordlist (originally gathered from Armenian newspapers) to a 14,000 entry dictionary.

SOURCES: the Armenian Frequency Dictionary (AFD), limited to
(1) those words used more than 50 times or
(2) terms of military, political, economic or social importance;
additional vocabulary in these four areas may be added from the 4-volume Contemporary Dictionary of Armenian (CDA), especially those marked as մամուլ (press).

PROCEDURE: For each entry the Fourth Dimension database has been structured to include:

1. **Headword** selected on the basis of the criteria above.
2. **Romanized transliteration** (phonetic representation automatically supplied by the computer program).
3. **Part of speech** [adj, adj-cmp, adv, conj, deic, dp, expr, intj, n, n-cmp, neg, np, num, pn, postp, prep, pro, qw, v(-intr ~ -tr), vn, vp, v-pass]
4. **English definition** (translation or explanation), doublechecking the meaning of each candidate in the CDA.
5. **Cross-references**, as appropriate, to Armenian **synonyms, antonyms,** extensions of meaning, or **alternates** (words that are alphabetically close but which will not appear separately in the dictionary).
6. **Example** (phrase or sentence examples, if context helps).

========== not to appear in print ==========

7. **Sources** (newspaper selection number, PCV, GAO, SEA, etc.).
8. **Codes 1**: AFD (with frequency statistics), CDA, etc.
9. **Codes 2**: Etymological information (Arabic, Greek, English, French, Italian, Latin, Russian, Turkish, etc.)
10. **Keyboarders** (LB, DZ; formerly TG)
11. **Create Date** (computer inserts when entry is created)
12. **Modify Date** (computer updates when any changes are made).
13. **Sequencer** (next number in sequence supplied by the computer)
14. **Open** (data occuring in the *Armenian Newspaper Reader*)
15. **Edit notes** (points to check up on)
16. **Slop** (Armenian hyponyms, paraphrases and extended cross-references originally keyed by Mrs. Baghdasarian, which will not appear in the published version).

rdz (mrm - 93.08.31)

Appendix 2: STRUCTURE FOR OPTILEX DATABASES