

A Practical Multi-language Database Program for Lexicography

John M. Durdin

*The Summer Institute of Linguistics
Thailand*

0. Abstract

The preparation of multilingual dictionaries involving more than two dissimilar scripts presents particular problems for the lexicographer. While commercially available database programs have often been extended to cope with the various European alphabets, very few have been adapted for the Indic languages of Southeast Asia. Even those programs which have been adapted for the Thai language cannot easily be extended to allow a further non-roman script.

An alternative approach, which has been taken by many SIL researchers in the past, has been to compile all dictionary data using a multilingual text editor. Although great flexibility is possible with this method, it is impossible to guarantee the integrity of a large database and its record structures when the whole or parts are edited with a text editor. This is particularly so when informants with only limited technical background are being trained to add data to a dictionary database.

This paper describes the development of a multi-language, multi-script, (noncommercial) database program designed particularly (but not exclusively) with lexicography in mind. Data files use plain text format but are coded to represent up to fifteen user-specified languages in each record. Keyboard layouts and character sets are switched automatically according to the language of each data field being edited. Any or all data fields of each record may be displayed as configured by the user, and records to be edited can be selected according to a variety of filtering conditions. Editing maintains the ordering of fields required according to previously defined record structures. On-line, context-sensitive help can be either in English or in a selected second language. Use of the program will be illustrated by its application to a Kmhmu' minority-language dictionary having Lao script, romanized script and phonetic main entries, and definitions in Lao, French and English languages.

1. Introduction

The project, which this paper is reporting, arose following the author's experience in processing Thai, Lao and minority language texts, using computer techniques for the purposes of linguistic analysis. A number of commercial and non-commercial word-processing programs were already available two years ago for use with Thai and Lao scripts on MS-DOS¹ computers. However, while such programs provided an adequate means to produce letters and reports in those languages, they generally lacked more powerful text-processing capabilities such as syllable and word recognition and sorting. In addition, only one of the editor programs available² allowed the use of more than one non-roman script within a single document.

¹ MS-DOS is a registered trademark of Microsoft Corporation.

² SIL's Direct Translator Support (DTS) Editor (ED) provides for up to four languages with distinct scripts simultaneously, but is a text editor rather than a true word processor.

*Apple Macintosh*³ computers provide more flexible script-handling capabilities than those using MS-DOS but are not as widely available or well supported in Thailand or Laos as are MS-DOS machines. Consequently, few field linguists have access to them.

At the start of 1992, the author was assigned to assist in the preparation of a multilingual dictionary for Kmhmu', one of the Mon-Khmer minority languages of Laos. For this dictionary, each entry included romanized, phonetic, and Lao-script Kmhmu', and meanings in English, French and Lao. Use of a text editor to enter and correct a dictionary database is increasingly difficult as the size of the database increases, and accidental, unwanted changes become increasingly likely: it is the kind of application that database programs were created for. However, there were (and, as far as I am aware, still are) no available database programs that could handle the multiple-language, indefinite length dictionary data in any useful manner.⁴ Even those (commercial) database programs that are available for English language materials have difficulty in handling the somewhat loosely-structured data of a real dictionary, which may have nested subentries and semantic variants within each main entry record.

For these reasons, the author was encouraged to produce, specifically for linguistic applications, a new text-database program, which has become known as *Polyglot*.

2. Program Objectives

The *Polyglot* program has the following main design objectives:

- (a) to provide multiple-language, multiple-script capability, both on the screen and in printed output;
- (b) to have adequate control of record structures with sufficient flexibility to enable the safe input and editing of a large text-database;
- (c) to efficiently allow for variable record sizes in a plain ASCII text-database;
- (d) to be user-friendly enough for it to be usable by our national colleagues and language helpers, many of whom are neither highly trained nor particularly fluent in English; and
- (e) to be user-configurable.

Each of these will be discussed in more detail in the following section.

2.1 Multiple Language Capability

When entering text data in multiple languages, it is important to keep track of the codes being used for each language and script. *Polyglot* does this by associating each field type in a record with a particular language, so that whenever that field is viewed, edited, or printed, the correct font (and keyboard layout) will automatically be used. Up to fifteen distinct languages and/or scripts can be accommodated by the program.

Simple screen and printer fonts have been developed for a number of languages, and others may be created relatively easily. *Polyglot* uses graphics mode display according to the capability

³ Apple and Macintosh are registered trademarks of Apple Computers Inc.

⁴ SIL's Shoebox text database program is a highly practical tool for initial dictionary development, but it cannot at present handle multiple scripts simultaneously, and it also suffers from a lack of integrity checks.

of the particular computer (CGA, Hercules, EGA, VGA, etc.) but does not depend on *Windows*⁵ or other specialized operating environments. Similarly, in the initial version, printed reports may be generated on 24-pin dot-matrix printers, using graphics mode rather than text or downloaded characters. Support for other printer types will be developed shortly.

2.2 File Coding: Plain ASCII Text, Indefinite Record Sizes

Data files are coded in a variety of different ways, often specific to particular programs, and it is frequently difficult to convert data for use with different programs. However, if certain rules are followed, the exchange of data between different programs is greatly simplified. In particular, if the use of certain control characters is avoided, text data files can generally be manipulated by different programs without much difficulty, regardless of what the characters used actually represent.

The convention most often followed for a particular language is to represent characters unique to that language by the upper ASCII codes (values 128-255), leaving the lower ASCII codes (values 32-127) to represent English (and punctuation) characters as usual. *Polyglot* wherever possible maintains the established conventions for each language, with the significance of each upper ASCII character being determined by the field in which it is found. In this way, the transfer of text files between *Polyglot* and other programs creates no difficulties, although the processing of *Polyglot* database files by other programs can compromise the structural integrity that *Polyglot* is designed to maintain (see below).

2.3 Record Structure Control

The necessity for record structure control is best illustrated by means of an example. In the following typical entry from the Kmhmu' dictionary, the data has been simplified for the purposes of this paper by the removal of Lao script and phonetic entries.

Kmhmu:	du'
French:	partir
English:	to leave
French:	fuir
English:	to flee
Example:	du' da' hô'
Meaning in French:	vas là-bas
Meaning in English:	go over there
Sub-entry:	du'!
Synonym:	pdu'
Variant:	1
French:	vas-t'en!
Note in French:	pour enfants
English:	go away!
Note in English:	for children
Variant:	2
French:	appel pour chasser des animaux dans un chemin étroit.
English:	call for chasing animals on a narrow path.

⁵ Windows is a registered trademark of Microsoft Corporation.

In most databases, the position of each field within a record is not important, and usually only one field of each type is allowed. However, in data such as this, it is evident that multiple fields of a given type may occur, and the order of fields carries additional information. A semantic variant, or a synonym, may be associated with the main entry, or, as here, with a subentry, as indicated by the order of the fields. If the order of associated fields is allowed to change, information will be lost. Similarly, some fields can only be included if they follow another specific field (e.g., the explanatory note fields above). If such conditions are not observed, printed output generated from such data may be badly garbled. Obviously, the same is also true if incorrect field markers are used.

In a large database, it is very difficult for a user to keep track of the many field markers to be used and their ordering restrictions. The *Polyglot* program has been designed to assist in this task by requiring each record to have a defined, although highly flexible, structure to associate related fields and maintain ordering information.

2.4 Ease of Use

While there are no standards that define the "best" way to interact with a computer program, *Polyglot* has adopted the following common design features to simplify program use:

- (a) Pull-down menus: Program operations may be selected from menus by moving a highlighted cursor bar, pressing a menu letter, or pointing with a mouse. As the cursor is moved around the menus, a brief description of each menu function is provided, either in English or in a specific second language, as selected. More extensive context-sensitive help is also available, again in two languages.
- (b) Dialog-style interaction: Dialog boxes, familiar to users of Windows and other graphic environments, are a means of providing the user with a guided, limited choice of responses for any information the computer requires.

2.5 User Configurability

Since each user, and each application, will have different operating requirements, *Polyglot* is configurable at various levels, as follows:

- (a) During installation the user must specify which languages are to be used, and the type of equipment he is using.
- (b) For each new data file being created, record and field characteristics must be specified.
- (c) Default screen and report formats are provided, but program use is greatly enhanced by creating user-designed screen layouts and reports according to the particular application. The program provides the means for creating such screen layouts and reports, and associates them with the particular database file for which they are designed.
- (d) A variety of filter conditions may also be specified for each database, to enable the selection of particular subsets of records.

- (e) At a lower level, character fonts and keyboard layouts may be modified or completely new fonts prepared, using, for example, SIL's DTS Font System software, and conversion programs.

3. Examples

A user-defined screen layout showing (part of) the record used as an earlier illustration is shown in Figure 1, below. Each field's window may be edited independently and scrolls as required to show longer data.

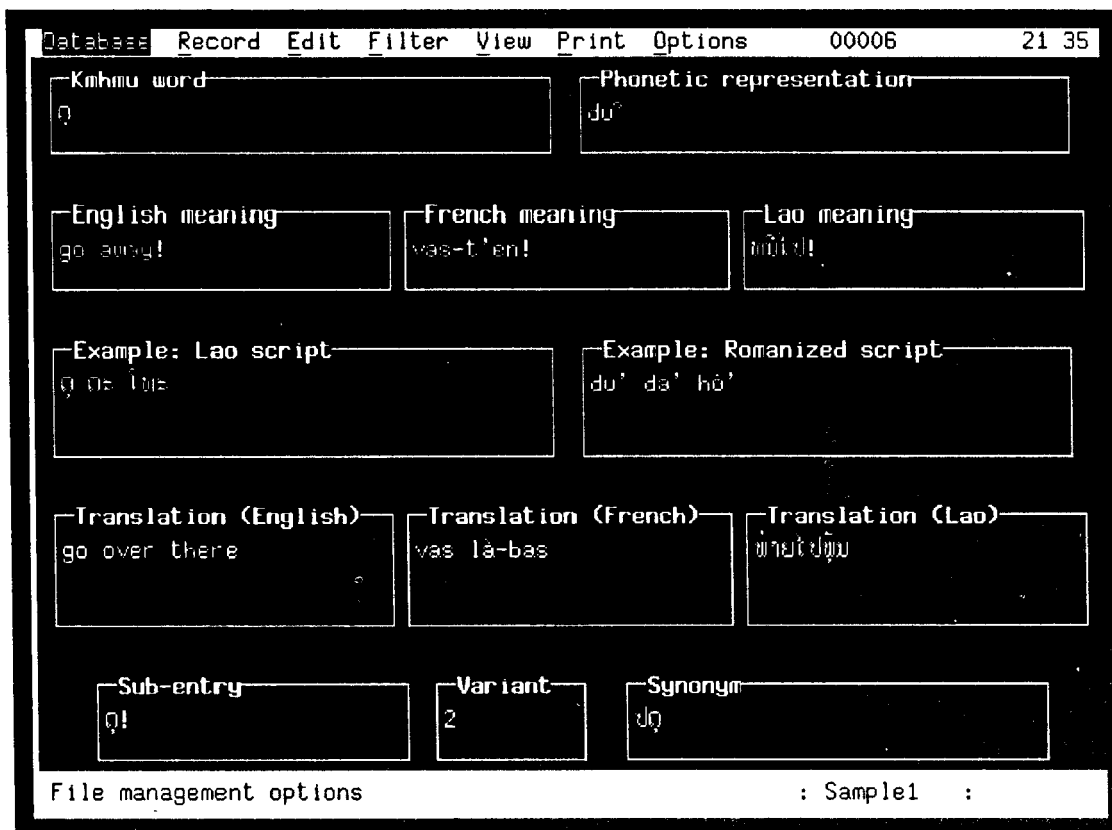


Figure 1: A Single Record View

An alternative format, chosen to display selected fields of a number of records, is shown in Figure 2, below. Such multiple record views are not themselves editable but will switch to an editable view if editing is attempted. The default view of any record simply shows the fields in sequence, each field to a line, as in the tabulated entry illustrated above.

Database	Record	Edit	Filter	View	Print	Options	00001	22 06
	Kuhru	Phonetic	English	French				
	ເືອນ	et	eleven	once				
	ເືອນ	eh	to get infected	s'infecter				
	ເືອງ	eeŋ	self	soi-même				
	ເືອນ	eh	to construct	construire				
	ເື	a	particle	particule				
	ູ	du?	to leave	partir				
	ູນ	dun	crutch of trousers	entre-jambes du pa>				
	ູນ	dul						
	ູນ	duh	all	tous				
	ູນ	dot	short	court				
	ູນ	dooc	to be bent on han>	être plié ou penç>				
	ູນ	dol	cicada	la cigale				
	ູນ	daj	to embark	endiguer				
	ູນ	dsew	line	une ligne				
	ູນ	de?	to take	prendre				
	ູນ	daet	a little	un peu				
	ູນ	deen	terrain	terrain				
	ູນ	daaj		lézard				
	ູນ	daal	dull	émoussé				
	ູນ	andrayh	lightning	les éclaires				
File management options						:	sample2	:

Figure 2: Multiple record 'browse' view

Printed reports may be generated similarly for one or several records, with a variety of formatting options to control field layout and spacing, and record spacing.

4. Conclusion

It is hoped that the program described will make a useful contribution to the work of lexicography in Asia. A prototype version of the program is currently available to researchers, although considerable further development is planned, including, particularly, the use of multiple indexes with language-dependent sorting.

5. Acknowledgements

I wish to acknowledge the sponsorship of the Lao Committee for Social Sciences while developing this program as a contribution to the Kmhmu'-Lao dictionary project.