Third International Conference of Austroasiatic Linguistics
26-28 NOVEMBER 2007, Deccan College Post-Graduate & Research Institute, Poona, India

Data Sharing in the Mon-Khmer Languages Project

Doug Cooper
Center for Research in Computational Linguistics[1]
doug.cooper.thailand@gmail.com

The Mon-Khmer Languages Project is a broad plan to support research in comparative linguistics and lexicography. It was created to provide a practical means for sharing lexicographic data and comparative analysis, including both confirmed and edited results, and the 'dark matter' of working data and partial results that are, in some cases, our only available resources.

The Project provides two linked, Web-accessible resources with the usual array of search and presentation tools:

- The **Mon-Khmer languages database** is an on-line store of lexicographic data. Drawn from both published and unpublished sources, the database will ultimately provide a snapshot of relevant (for comparative purposes) knowledge of each of the Mon-Khmer languages, including glossing and phonetic transcription.
- The **Mon-Khmer etymology database** serves a similar role for analysis. It will initially be based on data extracted from Shorto's *Mon-Khmer Comparative Dictionary* (2006); the most extensive such resource, and a fitting starting point for this effort.

The MKL Project is intended to be both accessible and extensible. 'Source filtering' lets resource sets be defined as narrowly or broadly as desired; for example, searches might include only data from a particular dictionary, or incorporate *all* data available for a given language. However they are defined, resource sets can also be extracted and downloaded for off-line research.

New datasets that follows a simple XML tagging protocol can also be added to the MKL Project databases. Every item is identified by its contributor's name, so the obvious issue of quality control is dealt with in a transparent, elegant manner: source filtering can include, or just as readily exclude, any individual's contributions. Thus, only sources the user trusts, or items that been vetted by scholars the user trusts, will actually figure in any response to user queries.

The Mon-Khmer Languages Project is, above all, a collaborative venture. We have received wide support in the linguistics community in planning and acquiring initial data for the project, and generous funding from the U.S. National Endowment for the Humanities in launching it as of May, 2007. I look forward to describing the project's implementation, and to soliciting advice and comment on how it can best meet its goal of enabling timely sharing of data and analysis by Mon-Khmer language researchers.

Please note that this **discussion paper** describes a **preliminary system** that is still in the initial stages of development and has not yet been publicly released. Both unpublished data and partial subsets of published data are being used for illustrative purposes. The author takes full responsibility for any misrepresentation or error

[1] CRCL and the Mon-Khmer Languages Project gratefully acknowledge the support of the National Endowment for the Humanities. Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of the NEH.

1. Introduction

The Mon-Khmer language family is the larger subgroup of the Austroasiatic stock; the Munda languages, spoken primarily on the Indian subcontinent, form the other. The roughly 150 Mon-Khmer languages are of great antiquity, major linguistic interest, and primary importance for the study of Southeast Asian history and culture. Mon-Khmer languages are the national languages of Vietnam and Cambodia, and are found in communities large and small in India and China, and across broad swaths of Burma, Malaysia, Laos, and Thailand.

As might be expected, the historical depth and geographical diversity that make the Mon-Khmer languages so central for linguists, historians, archeologists, and other academic specialties have also made it extraordinarily difficult to gather the broad set of lexicographic resources required for detailed comparative work. Data has been gathered for more than a century, but not all of it has been formally published, and none of it has been available in electronic form.

The Mon-Khmer Language Project was developed in order to address this issue: to collect and digitize the widest possible range of lexicographic and comparative resources. First announced in 2004, the project received initial two-year funding from the U.S. National Endowment for the Humanities in 2007, following our (Sidwell, Cooper, and Bauer) editing and publication of Shorto's *Mon-Khmer Comparative Dictionary* (Shorto 2006). It has received broad support from the field, with pledges of data and assistance coming from around the world.

The project has three focal points. It will create:

- a **Mon-Khmer languages database** that makes all language reference materials, including phonetic transcription, glosses, and citations, freely available. We anticipate compiling initial datasets representing all Mon-Khmer branches in the first two years.
- a **Mon-Khmer etymological database** that provides an on-line hierarchical reference that puts language data in context. It will be based on – and ultimately extend greatly – Shorto (2006).
- a **collaborative worksite** for Mon-Khmer language research, that provides an architecture for extension, comment, and correction of language and etymological data.

This paper describes the operation of the databases, and begins to document the mechanisms that will be provided for data sharing in the Mon-Khmer Languages Project.

For the first two years of the project we will treat this as a preliminary specification, which may be modified in response to the needs of the linguistics community. Of necessity, the overview provided here will be supplemented by more detailed documentation of technical issues, including in particular project standards for etymological markup, and the design and implementation of phonological search.

We begin section 2 with a brief survey of relevant literature, then introduce some terminology, and discuss certain design considerations that impact on making data available as rapidly as possible. We describe the language and etymology databases from the user perspective in sections 3 and 4. Section 5 discusses how data may be accessed and redistributed, while section 6 introduces the underlying coding of database contents. Section 7 deals with additions to the database. Finally, section 8 gives an overview of current and planned database contents.

2. Preliminaries

We open by briefly citing relevant research and references, then discuss initial design considerations.

2.1 Previous Work

The only general reference to Mon-Khmer etymology is Shorto's ambitious *Mon-Khmer Comparative Dictionary* (Shorto 2006). Prior to this, only branch and sub-branch level

reconstructions had been attempted; these include North Bahnaric (Smith 1972), Mnong (Blood 1966), East-Katuic (Thomas 1967), Viet-Muong (Barker 1963, Barker & Barker 1970), Jeh-Halang (Thomas & Smith 1967), Semai (Diffloth 1977), Waic (Diffloth 1980), Monic (Diffloth 1984), South Bahnaric (Sidwell 1998), Katuic (Sidwell 2005), and Vietic (Ferlus ms.).

Etymological projects of regional interest that have some digital component include three initiated in the mid-1980's: the *Sino-Tibetan Etymological Dictionary and Thesaurus* (STEDT, 1987), the *Austronesian Comparative Dictionary* (ACD, 1990), and the *Munda Lexical Archive* (MLA, initially funded by NSF for Sora in 1979, completed 1985); access points for the latter two are noted below. Earlier works that have since been digitized (although purely as text references under the auspices of the *Digital Dictionaries of South Asia* project) include extensive analyses of Dravidian (Burrow and Emeneau 1964) and Indo-Aryan (Turner 1966-69). Work in the Tai family is not available in any digital form, and includes Li Fang Kuei's proposed reconstruction of proto-Tai (1977) and Ostapirat's reconstruction of proto-Kra (Ostapirat 2000).

Handling of etymological data for computer database applications has received relatively little interest. Crist (2005) provides a good survey of the topic, and of the minimal consideration for etymological markup given by existing systems. His proposed model includes the definition of hierarchically tagged cognate sets with explicit specification of word values and etymon/reflex relations, and uses *accepted-by* and *rejected-by* tags as a way of approaching the problem of encoding confidence levels. Other useful references include Good & Sprouse (2000), Bell & Bird (2000), Ide et al (2000), and Wittenburg et al (2002).

In recent years consideration of preservation and reuse of data have begun to occupy an increasingly important role in project planning. The *Electronic Metastructures for Endangered Languages Data* (E-MELD, emeld.org) has been extremely influential in promoting basic 'best practices;' see also Bird & Simons (2003). The *Open Language Archives Community* (OLAC, www.language-archives.org) and the *Open Archives Initiative* (http://www.openarchives.org) focus on metadata harvesting. Other efforts include the *Documentation of Endangered Languages* project (DoBeS, http://www.mpi.nl/DOBES), and the *Pacific and Regional Archive for Digital Sources in Endangered Cultures* (PARADISEC, http://paradisec.org.au).

On-line resources for comparative linguistics are rare, and include the *Tower of Babel* etymological database project (http://starling.rinet.ru/main.html), and the *Indo-European Etymological Dictionary* (IEED, http://www.indo-european.nl) at Leiden University. Others mainly provide parallel lexicographic resources, and include the *Austronesian Basic Vocabulary Database* (http://language.psy.auckland.ac.nz/austronesian), the *Intercontinental Dictionary Series* (IDS, http://lingweb.eva.mpg.de/ids), and the *Munda Lexical Archive* (http://www.ling.hawaii.edu/faculty/stampe/aa.html).

2.2 Author and Item Identification

As we will see below, the Mon-Khmer Languages database differs from typical approaches to storing and redistributing lexical data. Rather than clustering data into groups that essentially mirror the layout of paper dictionaries (sometimes referred to as the *editorial view*), we instead reduce the entire dataset to three basic elements:

items are citations or reconstructions, and may include orthography, phonemic rendering, glosses, and other lexical information;

links encode the relationship between items; e.g. they point citations to reconstructions;

notes may comment on items, links, or other notes.

The resultant structure is highly suited to representing the genealogical trees associated with etymological data, yet lexical data is readily reused simply by adding or ignoring linking data.

The unique identification of every element is an essential requirement. We define two necessary terms:

authored: all sources are identified by an abbreviation that follows the form **XxxYYYY**:

Xxx is a three-letter abbreviation of the author's name, and is not case sensitive;

YYYY is the year of publication. Preliminary material that is intended for discussion but not citation is dated **YxxY**, e.g. **2xx7**.

itemID: individual items are identified within the project database, *every* data item, including citations, reconstructions, links, and notes, is uniquely identified using this basic format: **authorID:X:number**.

X gives the *itemID type* and is one of the letters **C***itation*, **R***econstruction*, **L***ink*, or **N***ote* (discussed below).

number is usually the entry number in the original source. If the original author has an identifying letter or number (or combination) it is used instead. In both cases, the subcomponents of an entry are numbered sequentially, e.g. **17-2** or **V215-1**.

For example, **Sin1906:C:24** is the twenty-fourth citation in U. Nissor Singh's Khasi-English dictionary (Singh 1906), while **Sho2006:R:24** is the twenty-fourth reconstruction set in Shorto's comparative dictionary of Mon-Khmer.

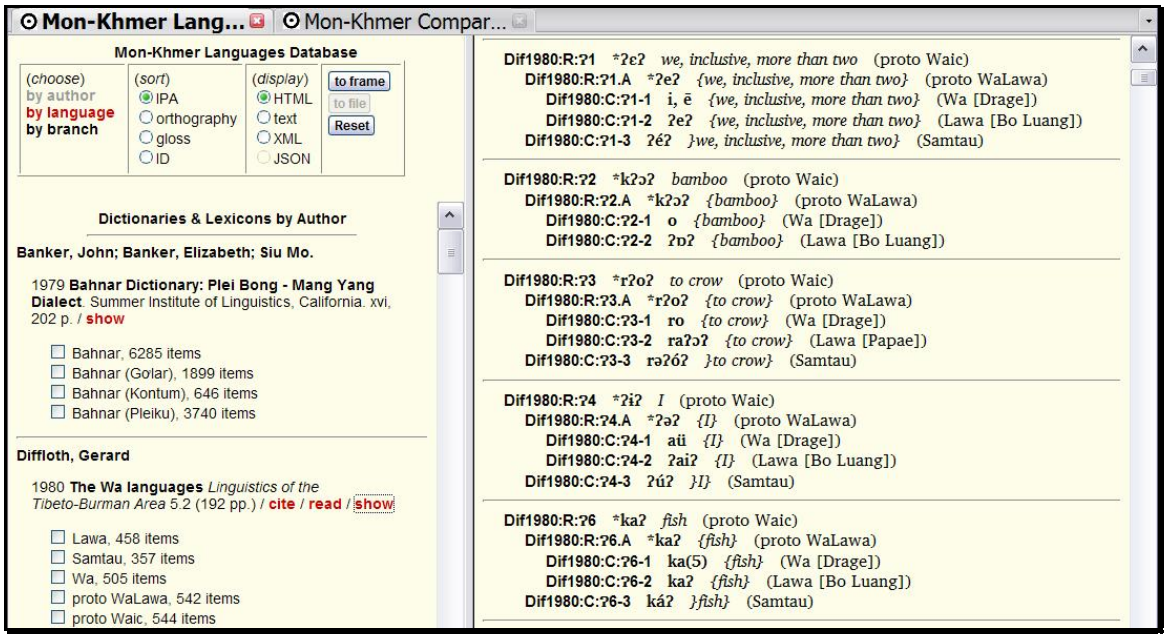
2.3 Other Preliminary Design Issues

A central consideration has been to make the data collected under the auspices of the project available to the broader community as rapidly as possible. A certain degree of tension results from balancing the desire for immediate access with the long-term goals of providing a retrospective survey of existing data, and incorporating it into a fully documented comparative analysis. Points of temporary compromise include:

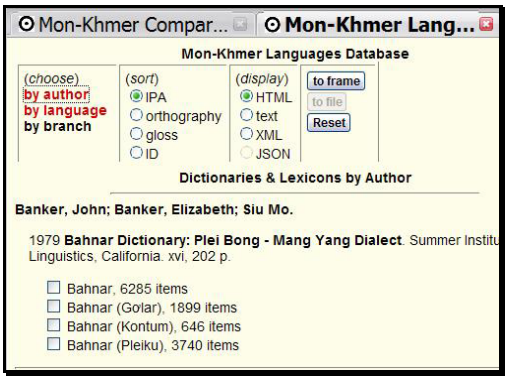
- *normalizing data* Although we use Unicode for character encoding, there is still inconsistency in the conventions used for both orthographic and phonemic data. For the moment, we preserve original formats. Note that data may still be exposed by searching via the original author or publication even if orthographic or phonemic data do not adequately conform to search tools designed for the project as a whole.
- *partial etymological grouping* Lexicographic (as opposed to truly comparative) sources, as well as preliminary field notes, do not necessarily posit higher-order grouping of data into etymological sets. Where possible, we insert partial information to help group individual data items; e.g. adding dummy root entries as described in section 6.3.1. Of necessity such additions are incomplete; in the long run the dummy entries will be replaced by pointers to the Mon-Khmer etymological tree.
- *selective acquisition* In some datasets a very large number of citations from different dialects or sources within a single language are provided. In some cases we have selected representative items in order to populate the database, and will return later to complete data collection.

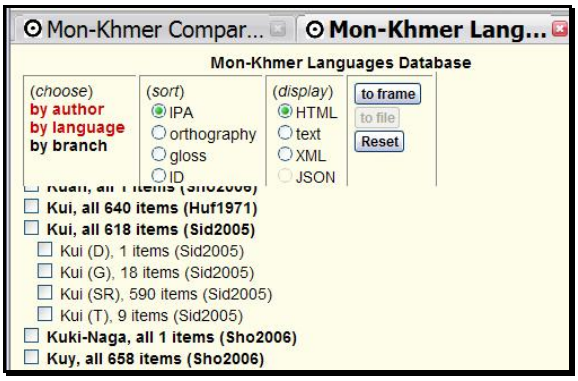
3. The Mon-Khmer Languages Database

The languages database is a collection of published and unpublished texts, comprised of ordinary dictionaries as well as etymological and comparative works. All data in the languages database is also accessible via the etymology database; however, the languages database is more suitable for viewing or extracting datasets, as opposed to searching *across* datasets. The basic user interface is shown below: controls are on the left, and results are on the right.

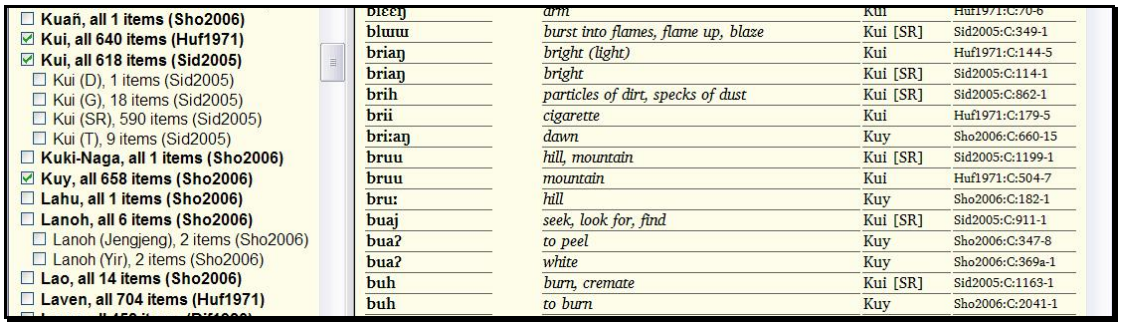


In this case, we have chosen the **show** link from the entry for Diffloth (1980); this reformats and displays all Diffloth 1980 entries currently in the database (which is still incomplete). Other links provide formal citations for each text in a variety of formats, as well as PDF or DjVu files of the original published source.





The contents of the languages database may be listed by author or source (as seen above left), or by language and dialect (above right). For convenience, all dialects cited by a particular author are both aggregated and individually selectable. Above right, note that Kui has three major sets (one listed as "Kuy") consisting of 640, 618, and 658 items respectively. The second set draws on four sources, which can be inspected separately. Checking any set of boxes (such as the Kui and Kuy group) lets us print a merged lexicon, as seen below sorted by IPA:



The languages database can be used both for on-line browsing and for extracting and downloading (possibly merged) datasets. Data may be returned in four formats:

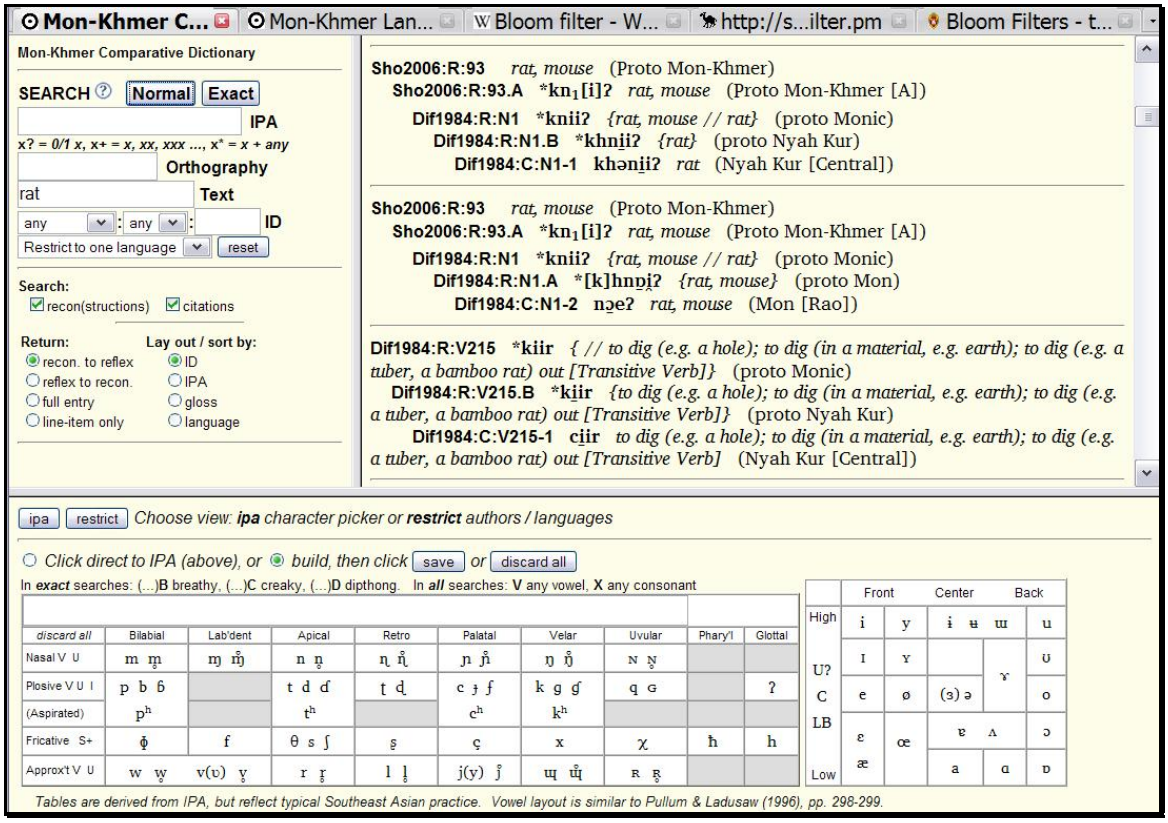
- HTML is best for on-screen viewing, or for re-use in Web pages. The text is tagged using standard HTML tables; a CSS stylesheet is embedded in the page.
- XML returns data marked with the tagset used internally and described in section 5.1.
- Text returns data as tab-separated values, without tagging.
- JSON returns data in *JavaScript Object Notation*, and is intended to be called programmatically in order to be incorporated into Web pages (not yet available).

Data may be sorted by IPA, orthography (if available), gloss (ignoring leading punctuation), or internal ID number.

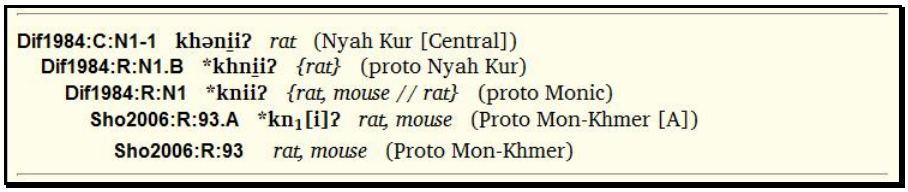
4. Mon-Khmer Etymological Database

The etymological database fulfills a dream long held by Southeast Asian linguists – a resource that can begin to help unravel the tangled web of language influence and development in the region.

The basic user interface is shown below. Controls are on the left and bottom, while results are returned to the upper right of the screen. We see the results of a search for the word **rat** in any gloss. Results are show in sets, ordered from the earliest reconstruction to modern-day reflexes.

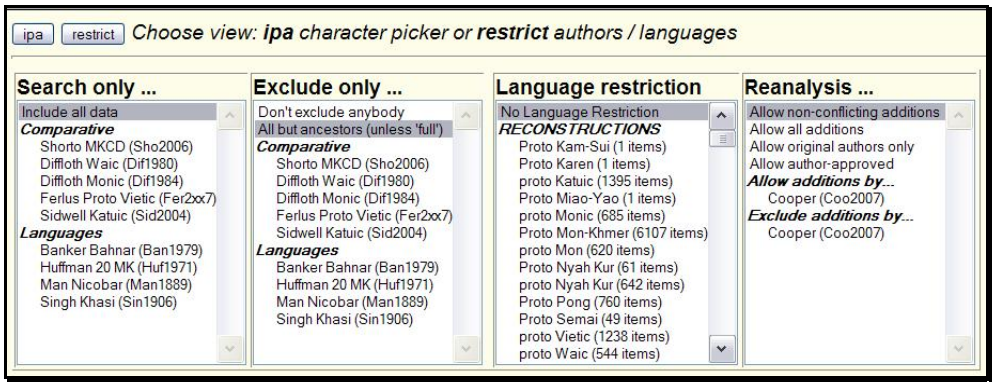


A variety of other options are supplied. For example, set ordering can be reversed and displayed from reflex to reconstruction by setting an alternative **return** option:



The lower part of the screen shows an innovative mechanism for constructing phonemic searches. Both the consonant and IPA tables are similar to the standard IPA charts, but are intended to be more useful for Southeast Asian practice. Items may be selected individually or in sets. For example, clicking on "**S+**" in the **Fricative** box at the lower left returns the set **θsʃʂçχcɟfcʰ**, which may then be used as part of a search target. Similarly, the **U?** set in the vowel panel returns the set **ʔaəiu**, which is designed to specify an optional unstressed vowel.

An alternative lower panel is shown below. It provides the option of either including or excluding particular authors and sources, and may be used to provide language restrictions as well. Note in particular the **Exclude only …** choice ***All but ancestors***, which lets searches be confined to a particular subset of modern languages or intermediate reconstructions, while still allowing linked citations to further up the Mon-Khmer tree.



The *search only, exclude only,* and *reanalysis* controls enable one of the project's fundamental design goals: that it be open to all user contributions, subject to the limitations of good taste and common sense. Such contributions may include both sets of lexical data – citations or reconstructions – and analysis of the historical relations between existing items.

In effect, the database is rebuilt every time it is consulted, and is thus able to take into account any additional links, reconstructions, or commentary. The same mechanism that allows this flexibility also helps ensure that database users can control just whose contributions are reliable enough to be incorporated into the data set. Every relation (and every reconstruction, comment, and citation) is marked with the identity of its contributor. The user can choose to include and/or exclude any subset of analysis (and data) providers.

5 Access to Data

A key design goal of the Mon-Khmer Languages Project is to expose project data both interactively (via the user interfaces just seen), and *programmatically*, using the kind of application programming interface known as a *Web API*. The Web API reveals all of the functionality of the user interface to program-generated queries, and or to URIs that may be embedded in texts (and are discussed below).

5.1 The Web API

A Web API defines a set of *queries* and *responses* that can be managed using the same HTTP protocol that Web browsers rely on. A key feature of this design is that it is stateless, and does not require that the query-response connection be managed or maintained. A type of link called a

URI, which is similar to the more familiar **URL** but can point to abstract instances of data (as opposed to very real Web pages) is all that is required.

Both the languages and etymological databases are accessible in this manner. Queries are handled by SEAlang's default request URL, **http://api.sealang.net** . A typical request URI is:

http://api.sealang.net?resource=monkhmer&show=full&text=rat

The **show** link that helped generate a screen capture seen earlier used this very method:

http://api.sealang.net?resource=monkhmer&show=full&include=Dif1980&exclude=all

The preliminary specification for request attributes is shown in Table 2.

The response to a Web API query may be provided in several forms, including HTML, plain text, and XML. Again, this matches the database's on-screen functionality. How this response is handled is up to the user: HTML output may be redirected to a window or frame, text output may be saved to a file directly or via copy-and-paste, or XML output may be reprocessed and repackaged.

A point that is essential for non-programmers to grasp is that using the Web API need be no more difficult than embedding a link into a Web page, or typing the URI into a Web browser. Indeed, we expect that two particular applications will become commonplace: using URIs to specify references, and using browsers to copy-and-paste dictionary texts.

The XML response to a Web API query matches the tagging scheme to be discussed in section 6, and is outlined in Table 3. This is a preliminary specification, and was developed largely in the context of encoding the data found in Shorto (2006), a task which is not yet complete. The tag, attribute, and value sets may be modified in response to discussion within the linguistics community.

Request URL

http://api.sealang.net
http://api.sealang.net?resource=monkhmer … *&attribute=value pairs*

Table 1. Request URL and example.

Request Attributes

Attribute [=default]	*Values*	*Effect*
resource	*monkhmer*	use the Mon-Khmer databases
show [=*self*]	*full*	derive and return the complete etymological set, ordered from etymon to reflex
	self	return only the matched item
	parents	return the matched item and direct ancestors, ordered from etymon to reflex
	children	return the match item and direct ancestors, ordered from reflex to etymon
phone	*any*	the phonemic value to search for
orth	*any*	the orthographic value to search for
text	*any*	a string value found in the definition
id	*AuthorID:[RCLN]:#*	match a specific item's ID
include [=all]	*all* *AuthorID* *AuthorID*|*AuthorID…*	one or more |-separated AuthorID values
exclude [=none]	*none* *ancestors* *AuthorID …*	
lang	*name* *name*|*name*	restrict the search to one or more languages
dialect	*name*	restrict the search to the dialect, toponym, or source name conventionally used to denote a language subset

Table 2. Request attributes. These are the parameters supplied with any use of the Web API. They reflect the functionality supplied by the interactive Web pages.

Response Tags

tag	subtag	attribute=	value	explanation
<item …> *content* **</item>**				a citation (type **C**) or reconstruction (type **R**).
		id	*itemID, type* **:C:** *or* **:R:**	the item's unique identifier
		lang	*language name*	language or proto-language name
		dialect	*dialect name*	commonly used dialect, toponym, or source name
	<ipa …>			IPA representation
		source=	*string*	original, non-IPA representation
	<orth …>			orthography as supplied
		script=	*name*	an identifier for orthography if necessary
	<gloss …>			definition or gloss
		pos=	*string*	part of speech
	<refer>			reference
<link … />	*none*			a link between two <item> entries
		id	*itemID of type* **:L:**	the item's unique identifier
		f[rom]	*itemID*	the historically more recent item
		t[o]	*itemID*	the historically more removed item
		type	*reflex* *instance* *phon(ological)* *deriv(ationall)* *inflect(ional)* *compound* *loan*	*reflex:* direct descendent of. *instance:* realization of a (possibly unstated) reconstruction. *phon, deriv, inflect:* indirect relations *compound:* unanalyzed compound form. *loan:* borrowing to/from.
		subtype	*assimilation* *dissimilation* *metathesis* *lenition* *fortition* *sandhi* *leveling* *epenthesis* *elision* *affix* *back(-formation)* *secondary(-deriv)*	a more detailed note, typically associated with a phonologically or derivationally related item. **Values of the type and subtype attributes are limited, but not closed.** These are practical sets, based on actual reference literature, and are not intended to define formal ontologies.
		certainty	*uncertain* *unlikely = 0.25* *possible = 0.5* *likely = 0.75* *probable = 0.9*	certainty of the analysis. "*uncertain*" does not have a specific numerical value.
<note …> *content* **</note>**		**id**	*itemID of type* **:N:**	
		r[eference]	*itemID*	

Table 3. XML elements used to tag responses. These reflect the tags used in the internal database, and will eventually be required for new data submissions. Note that the **<link />** tag has only attributes, and no contents. As a general rule, controlled-vocabulary data is stored in attributes, with free text between tags.

6. The Mon-Khmer Database

It is convenient to think about the language and etymological databases as distinct entities. In some ways this is true. Each has its own Web interface, and each has distinctive applications. They are also conceptually distinct: one consists of attested or proposed citations with glosses and phonemic and/or orthographic realizations, while the other is primarily composed of *relations* between citations, and commentary about the nature of the relationships.

In reality, though, there is only one underlying set of data. It contains:

items, which are citations or reconstructions, and may include orthography, phonemic rendering, and glosses;

links, which encode the relationship between items, and

notes, which may comment on items, links, or other notes.

The internal format of items, links, and notes reflects the response-tag specification shown in Table 3. While other details such as timestamps may be employed locally, one can assume that any format returned *by* the database will be acceptable as input *to* the database.

6.1 Lexicographic Data

Within the database, all data is held in plain text files, which can be opened and read using ordinary text editors, and do not require any specialized database software. As a rule, texts are initially received (or are typed by the project) using traditional positional formatting:

orthography /phonetic/ *part-of-speech* 'gloss' commentary

We then tag the data using a simple XML tagset that marks the type and boundary of each part of the entry in a transparent manner. Conceptually, the first step of the tagging process is:

```
<entry>
        <orthography>orthography</orthography>
        <phonetic>phonetic</phonetic>
        <pos>part of speech</pos>
        <gloss>gloss</gloss>
        <note>note</note>
</entry>
```

As a practical matter the tagset is more richly annotated, as we saw in table 3. This lets us preserve additional information relating to authorship, original identification or numbering, language dialect and/or source, citations, references, and the like. In final form, a typical lexicographic entry looks like this:

```
<item id="Ban1979:C:48-1" lang="Bahnar" dialect="Pleiku">
   <ipa>pətaŋ</ipa><orth>potăng</orth><gloss>a boil</gloss>
</item>
```

Additional notes, if any, may point to items. The note's **id** field gives its source, usually within a larger set of notes, and the itemID of the entry it **r**(eferences).

```
<note id="Coo2007:N:1" r="Ban1979:C:48-1">
   This is a note about Banker, Bahnar, or boils.
</note>
```

This separation of data and commentary would also apply to any note that might have appeared in the original text. Our intention is to separate specific language data – which might be reused somewhere else – from commentary about the data. But the commentary is not lost; it can always be recovered by tracking the reference by its *itemID*.

6.2 Etymological Data

Entries in the etymological database are handled in a similar manner. Again, the overarching intent is to separate reusable data from commentary.

reconstructions follow the <item> format used for citations, but the *itemID* has type **R:** rather than **C:**. This distinction is useful in searching the database.

links include *f(rom):* and *t(o):* attributes, and connect citations to reconstructions and derivative items to citations. *type* (and possibly *subtype*) attributes define the nature of the relationship.

notes may add additional commentary. Typically they refer to the certainty of reconstructions, the nature of links, and the sources of citations.

For example:

```
<item id="Dif1984:N1:R" lang="proto Monic"><ipa>*knii?</ipa></item>
<item id="Dif1984:N1:R.B" lang="proto Nyah Kur"><ipa>*khnii?</ipa></item>
   <link f="Dif1984:N1:R.B" t="Dif1984:N1:R" id="Dif1984:N1:L-2" type="reflex/>
<item id="Dif1984:N1:R.A" lang="proto Mon"><ipa>*[k]hnɒi?</ipa></item>
   <link f="Dif1984:N1:R.A" t="Dif1984:N1:R" id="Dif1984:N1:L-3" type="reflex/>
<item id="Dif1984:N1:C-1" lang="Nyah Kur" dialect="Central"><ipa>khənii?</ipa><gloss>rat</gloss></item>
   <link f="Dif1984:N1:C-1" t="Dif1984:N1:R.B" id="Dif1984:N1:L-4" type="reflex/>
<item id="Dif1984:N1:C-2" lang="Mon" dialect="Rao"><ipa>nɔe?</ipa><gloss>rat, mouse</gloss></item>>
   <link f="Dif1984:N1:C-2" t="Dif1984:N1:R.A" id="Dif1984:N1:L-5" type="reflex/>
```

Note that links point backward, from child to parent (or from derivative to source), as opposed to the *headword -> list of reflexes ....* relations traditionally seen in print dictionaries. Above, Mon points to proto Mon, which points to proto Monic; similarly, Nyah Kur points to proto Nyah Kur which also points to proto Monic. From the computational point of view, this subtle alteration greatly eases the task of generating an internal tree of the relations between items. By recursively walking this tree, we can readily identify sisters (they point to the same parent), or show historical relations in either order.

6.3 Extensions to Content and Tagging

Original content will sometimes be extended in specific ways in order to enhance its utility for the etymological database. In particular:

phonemic rendering: when orthography alone is supplied by the original source, we provide a preliminary phonemic rendering.

'dialect' tagging: there is not always a clear distinction between the identification provided by dialect, author, and place names. In practice, all serve to denote language subsets that are treated as being distinct for the purpose of analysis. In preparing datasets, we use the most appropriate information available to preserve this distinction under the rubric 'dialect.'

grouping: comparative dictionaries (and even more so field notes) often group items without either proposing earlier reconstructions or elevating a particular citation to a unique primary status. In such cases it is convenient to generate a 'dummy' root for the sake of grouping, as discussed below.

inferred glossing: when reconstructions and citations are presented in sets individual items are not always glossed. In preparing datasets, we infer glosses so that extracted subsets will be comprehensible. Inferred glossed are *always* given between braces: { … }. Glosses inferred from more than one item are separated: { … // … }.

For example, the items below are from Diffloth 1980 (which glosses reconstructions) and 1984 (which glosses citations). Inferred glosses are shown between curly braces:

***kɔn**	*child*	proto Waic	Dif1980:R:N4
***kɔɔn**	*{child, offspring,… // (one's own) child;…}*	proto Monic	Dif1984:R:N168
kawn	*{child}*	Wa [Drage]	Dif1980:C:N4-1
kɔ̱n	*child, offspring,…*	Mon [Rao]	Dif1984:C:N168-2

In all cases, any additions to or modifications of the preliminary texts are thoroughly documented.

6.3.1 The Dummy "root" Link

Diffloth's Monic and Waic datasets consist of values that have been explicitly linked together. But it is frequently the case that several items are known or believed to be etymologically related, but do not have an explicitly stated etymon. In such cases, a dummy entry whose itemID value is "**root**" can be used to group the citations.

For example, Huffman's unpublished vocabulary list (1971) is a broad collection of citations from some 20 Mon-Khmer languages. A fairly cursory survey is sufficient to mark the occasional Tai or Indic reflex, and to group the remainder into likely etymological sets. The same process can be applied to Banker (1979).

```
<item id="Ban1979:C:58-1" lang="Bahnar" dialect="Kontum"><ipa>hooj</ipa><orth>hôi</orth>
    <gloss>"a few</gloss></item>
<item id="Ban1979:C:58-2" lang="Bahnar" dialect="Pleiku"><ipa>huj</ipa><orth>hŭi</orth>
    <gloss>a few</gloss></item>
<item id="Ban1979:C:58-3" lang="Bahnar" dialect="Gơlar"><ipa>huj</ipa><orth>hŭi</orth>
    <gloss>a few</gloss></item>
<link f="Ban1979:C:58-1" to="root" id="Ban1979:L:58-1"/>
<link f="Ban1979:C:58-2" to="root" id="Ban1979:L:58-2"/>
<link f="Ban1979:C:58-3" to="root" id="Ban1979:L:58-2"/>
```

Internally, the root reference is automatically replaced by a dummy reconstruction (e.g. named **Ban1979:R:58**). The root and target are only used in collecting sisters from the same language (albeit different sources or dialects). At a future date, when this data is more thoroughly analyzed by the project, the 'to' reference is readily replaced by the ID of a formal reconstruction.

7 Collaboration in the Mon-Khmer Languages Project

An important goal of the project is to provide a collaborative workspace for researchers in the field. Just as anonymous peer review is an essential component of quality *in* publication, the open distribution and discussion of preliminary data and results can be an important stage in the evolution of work intended *for* publication.

The functionality provided by the language and etymology databases, including the ability to aggregate data across dialects, languages, or branches, to perform sophisticated phonemic searches, and to propose and view the implications of new analyses, are not intended to be reserved for previously published data or theories. Rather, the databases provide an important forum for seeing ideas in action, in an on-line context that can easily be accessed and teststed by colleagues around the world.

The idea that data of different vintages, so to speak, may be intermingled in a database without becoming irrevocably intermixed is not commonly encountered. Nevertheless, this functionality is easily provided by the *authorID* and *dataID* mechanisms, and is discussed further below.

7.1 Adding Content to the Database(s)

The Mon-Khmer Languages Project welcomes additions of data and analyses to the databases. Elaborate formatting of such additions is not required. Our experience thus far has been that most

datasets can be tagged purely on the basis of their existing internal layout, regardless of whether they are columnar, labeled and indented, or tagged in some other manner.

During the first phase of the project (2007-2009) we will assume responsibility for extracting data and tagging it in MKLP format. In the future, we will continue to do so, but will also specify simple submission formats (e.g. tabbed or labeled values) that can be tagged automatically. Submission of data in electronic form is highly preferable.

Submissions may be copyrighted, of course, but contributors should have the expectation that data may be reused if full attribution is given, in accordance with traditional academic procedure.

7.2 Types of Additions

Additions take three forms:

> **data** consists of citations and reconstructions. Each should include at a minimum a language or proto-language name, a dialect identifier if appropriate, phonemic and/or orthographic rendering, and a gloss. Additional information, such as part-of-speech, may also be supplied.
>
> **relations** have three parts, as seen in the response tag documentation in table 3: *f[rom]* and *t[o]* fields (from is always more recent, to is always older), and a *type* (and possibly a *subtype*) field.
>
> **notes** must include a *r[eference]* attribute that identifies an item, relation, or note.

Note that contributions may consist solely of relations (and/or notes). The link below makes the claim that the Dif1984:N1 group is a reflex of Sho2006:R:93.A. Similar links might provide commentary. In principle, hundreds or thousands of such links might be submitted

```
<link f="Dif1984:R:N1" t="Sho2006:R:93.A" id-"Coo2007:L:1" type="reflex" />
```

The Mon-Khmer Languages Project is committed to accepting all such contributions, subject to the limitations of good taste and common sense. The ID mechanism is used to include or exclude contributions, just as it can include or exclude or exclude ordinary data sets. A subtler degree of control is provided, seen earlier in the **Reanalysis** area of the **restrict** tab.

Allow non-conflicting add-ons	Additional links and notes are allowed, but only when they do not conflict with the original author
Allow all additions	Additions may override original authors
Allow original authors only	Only original authors may override
Allow author-approved	Original data suppliers may certify additions
Allow additions by ...	Specify IDs for inclusion
Disallow additions by ...	Specify IDs for exclusion

8. Database Contents

The sources originally proposed for entry during the first two years of the Mon-Khmer Languages Project are shown below. Asterisked items have already been at least partially entered, as have Huffman (1971), Sidwell (2005), and Shorto (2006).

ASLIAN BRANCH (Geoffrey Benjamin, advisor**)**

Temiar Means, Natalie. 1999. *Temiar-English, English-Temiar Dictionary*.

Semai/Senoi Means, Nathalie & Paul B. Means. 1987. *Senoi-English English-Senoi Dictionary*.

BAHNARIC BRANCH (Paul Sidwell, advisor**)**

Sedang	Smith, Kenneth. 2000. *Sedang Dictionary.*
*** Bahnar**	Banker, Banker & Mo'. 1979. *Bahnar Dictionary, Plei Bong-Mang Yang Dialect*.
Chrau	Thomas, David & Dorothy Thomas 1961. *Chrau-Vietnamese-English.*

KATUIC BRANCH (Paul Sidwell, advisor**)**

Ngeq	Smith, Ron. 1976. *Ngeq dictionary*.
Pacoh	Watson, Watson, Cubuat. 1979. *Pacoh Dictionary: Pacoh-Vietnamese-English*.

KHASIC BRANCH (Anne Daladier, advisor**)**

*** Khasi**	Nissor Singh, U. 1906. *English-Khasi Dictionary.*

KHMERIC BRANCH (Robert K. Headley, Philip Jenner, advisors**)**

Khmer	*Parallel development of Headley and Jenner in SEAlang Library*

KHMUIC BRANCH (Suwilai Premsirat, advisor**)**

Khmu	Suwilai Premsirat. 2002. *Thesaurus of Khmu Dialects in Southeast Asia*

MONIC BRANCH (Christian Bauer, advisor**)**

*** Proto-Monic**	Diffloth, G. 1984. *The Dvaravati Old-Mon language and Nyah Kur*.
Nyah Kur	Luang-Thongkum, Theraphan. 1984. *Nyah Kur – Thai – English Dictionary*

NICOBARIC BRANCH

*** Car**	Whitehead, George. 1925. *Dictionary of the Car-Nicobarese language.*

PAKANIC/MANGIC BRANCH (Jerold Edmondson, advisor**)**

Bolyu	Edmondson, Jerold, 1995. *Lexica: English-Bolyu Glossary*. *Mon-Khmer Studies*
Lai	Liang, M, 1984. *A brief description of the Lai language, Minzu Yuwen*.

PALAUNGIC BRANCH (Justin Watkins, advisor**)**

*** Proto-Waic**	Diffloth, Gérard. 1980. *The Wa Languages.*
Palaung	Milne, Leslie. 1931. *A dictionary of English-Palaung and Palaung-English*.
Wa	*SOAS Wa Dictionary Project* database.

PEARIC (Suwilai Premsirat, advisor**)**

Chong , Chung, Kasong, Son	Suwilai Premsirat, ms. *A Comparative lexicon of 8 Endangered Pearic*
Samre, Su'ung, Pear	*Languages*.

VIETIC BRANCH (Michel Ferlus, Mark Alves, advisors**)**

*** pViet-Muong**	Ferlus, Michel. ms. *Proto-Viet-Muong reconstruction and comparative lexicon*.
Muong	Barker, M. E. & M. A. Barker, 1976. *Muong-Vietnamese-English Dictionary*.
Ruc	Nguyên Phú Phong, Tràn Trí Doi, Ferlus. 1998. *Lexique vietnamien-ruc-français*.
	Solncev, V. M., N. V, Solnceva & I. V. Samarina, 2001. *Jazyk ruk.*

9. References

Banker, John, Elizabeth Banker & Siu Mo'. 1979. *Bahnar Dictionary, Plei Bong-Mang Yang Dialect.* Huntingdon Beach, Summer Institute of Linguistics.

Barker, Milton E. 1963. Proto-Vietnamuong initial labial consonants. *Văn-hoa Nguyêt-san* 12.3:491-500.

Barker, Milton E. & Muriel A. Barker 1970. Proto-Vietnamuong (Annamuong) final consonants and vowels. *Lingua* 24.3:268-285.

Bell, John and Steven Bird, 2000. *A Preliminary Study of the Structure of Lexicon Entries*, Paper presented at the workshop on Web-Based Language Documentation and Description, 12-15 December 2000, Philadelphia, USA.

Bird, Steven and Simons, Gary. 2003. *Seven Dimensions of Portability for Language*

Documentation and Description. Language, Vol. 79, no. 2 (2003), pp 557-582.

Blood, Henry F. 1966. *A Reconstruction of Proto-Mnong (Including Tentative Reconstruction of Proto-South-Bahnaric).* M.A. Thesis, Department of Linguistics Indiana University.

Blust, Robert 1995. *Austronesian Comparative Dictionary* (ACD). Unpublished computer files, Honolulu: University of Hawai'i.

Burrow, T. & Emeneau, M. B. 1964. *A Dravidian Etymological Dictionary*. 2d ed. Oxford University Press.

Crist, Sean 2005 *Toward a formal markup standard for etymological data*. LSA Annual Meeting, January, 2005.

Diffloth, Gérard. 1977. Towards a History of Mon-Khmer: Proto-Semai Vowels. *Tônan Ajia Kenkyû (Southeast Asian Studies)* 14.4:463-95.

Diffloth, Gérard. 1980. The Wa Languages. *Linguistics of the Tibeto-Burman Area*. Vol. 5.2. Berkeley: University of California.

Diffloth, Gérard. 1984. *The Dvaravati-Old Mon Language and Nyah Kur* (Monic Language Studies 1). Bangkok, Chulalongkorn University Printing House.

Edmondson, Jerold. 1995. English-Bolyu Glossary. *Mon-Khmer Studies* 24:133-159.

Ferlus, Michel. 2005. *Proto viet-muong (Proto-Vietic)*. Manuscript provisional reconstruction.

Good, Jeff and Ronald Sprouse. 2000. *SGML markup of dictionaries with special reference to comparative and etymological data.* Paper presented at the workshop on Web-Based Language Documentation and Description 12-15 December 2000, Philadelphia, USA

Huffman, Franklin 1971 *Unpublished vocabulary lists*. 495.9, Mon-Khmer, Austronesian general folder, dated 1976. David Thomas Library, Bangkok, Thailand

Ide, Nancy, Adam Kilgarriff, and Laurent Romary, 2000. *A Formal Model of Dictionary Structure and Content* .

Li Fang kuei. 1977. *A Handbook of Comparative Tai*. Oceanic Linguistics Special Publications. Honolulu, University Press of Hawaii.

Liang, M. 1984. *A brief description of the Lai language*. Minzu Yuwen 1984.4.

Luang-Thongkum, Theraphan. 1984. *Nyah Kur (Chao Bon) - Thai - English Dictionary*. Bangkok, Chulalongkorn University Printing House.

Means, Natalie & Paul Means. 1987. *Senoi-English, English-Senoi Dictionary*. The Joint Centre on Modern East Asia, University of Toronto and York University.

Means, Natalie. 1999. *Temiar-English, English-Temiar Dictionary*. Minnesota, Hamline University Press.

Milne, Leslie. 1931. *A dictionary of English-Palaung and Palaung-English*. Rangoon, Superintendent, Government Printing and Stationary.

Nguyên Phú Phong, Trần Trí Doi, Michel Ferlus. 1998. *Lexique vietnamien-ruc-français*. Paris, Sudestasie.

Nissor Singh, U. 1906. *Khasi-English dictionary*. Shillong: E. Bangal and Assam Secretariat Press.

Ostapirat, Weera. 2000. Proto-Kra, *Linguistics of the Tibeto-Burman Area*, 23.1.

Shorto, Harry L. 2006. *A Mon-Khmer Comparative Dictionary*. Canberra, Pacific Linguistics.

Sidwell, Paul. 1998, *A Reconstruction of Proto-Bahnaric*. University of Melbourne PhD thesis.

Sidwell, Paul. 2005. *The Katuic Languages: classification, reconstruction and comparative lexicon*. Munich, Lincom Europa.

Smith, Kenneth, D. 1972. *A phonological reconstruction of Proto-North-Bahnari*c. Dallas, Language Data Series, Summer Institute of Linguistics.

Smith, Kenneth, D. 2000. *Sedang Dictionary.* Mon-Khmer Studies Special Volume No.1. Mahidol University at Salaya and the Summer Institute of Linguistics Dallas.

Smith, Ron. 1976. *Ngeq dictionary*. Huntington Beach, Summer Institute of Linguistics.

Solncev, V. M., N. V, Solnceva & I. V. Samarina, 2001. *Jazyk ruk.* Moscow, Nauka.

Suwilai Premsrirat. 2002. *Thesaurus of Khmu Dialects in Southeast Asia*. Institute of Language and Culture for Rural Development Mahidol University, Thailand.

Thomas, David, & Marilyn Smith. 1967. Proto-Jeh-Halang. *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 20: 157–175.

Thomas, David & Dorothy Thomas 1961. *Chrau-Vietnamese-English.* Saigon, Summer Institute of Linguistics.

Thomas, Dorothy M. 1967. *A Phonological Reconstruction of Proto-East-Katuic*. MA thesis, University of North Dakota.

Turner, R. L. 1966-9. *A Comparative Dictionary of the Indo-Aryan Languages*. London: Oxford Univ. Press.

Watson, Richard, Saundra K. Watson and Cubuat. 1979. *Pacoh Dictionary: Pacoh-Vietnamese-English*. Trilingual Language Lessons, No.25, part 1. Manila, Summer Institute of Linguistics.

Whitehead, G. 1925. *A Dictionary of the Car Nicobarese Language*. Rangoon

Wittenburg, P., W. Peters, S. Drude. 2002 *Analysis of Lexical Structures from Field Linguistics and Language Engineering*