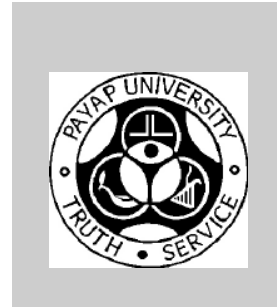


LEXICOGRAPHY CONFERENCE ABSTRACTS



Payap University Linguistics Department
Chiangmai, Thailand (24-26 May 2004)

Final Update: 3 May 2004

Abstracts Listing (29 total)

Listed alphabetically by presenters name.

An English-Thai Dictionary Arranged by Semantic Areas

AMON Thavisak and Robert S. BAUER

Institute of Language and Culture for Rural Development, Mahidol University, Bangkok,
Thailand and the Department of Chinese and Bilingual Studies, Hong Kong
Polytechnic University

This English-Thai dictionary takes an unusual and rarely used approach in the making of bilingual dictionaries in that its lexical entries are categorized by the semantic areas to which they belong. The rationale behind this method of semantic categorization is that human minds to some extent order their mental lexicons in the same way, so that lexical items that are linked through their meanings can also be brought together as a set of words; therefore, we believe that foreign students of Thai should find this dictionary to be a language-learning tool as effective as (if not more so than) traditional dictionaries in which lexical items are mechanically but arbitrarily arranged in alphabetical order. The dictionary includes nine major semantic areas of *Food, Language, People, Life, Materials, Nature, Numbers, Time, and World*; within these are listed 110 subcategories, for example, classifiers for nouns, English loanwords in Thai, pronouns; people by family relationships and occupations; human bodyparts and bodily excretions; diseases and disabilities; tools and instruments; modes of transport; time words and expressions; geography; places; weather phenomena; animals and animal bodyparts; fruits, plants, vegetables, condiments, methods of preparing foods, and prepared foods associated with Thailand. The Table of Contents provides a complete index to all the semantic areas and subcategories.

Within each semantic area the arrangement of lexical entries is in alphabetical order of the English word since this is an efficient way by which to look up a particular lexical item. Because semantically related items are joined together as a set, the dictionary user is able to discover lexically-encoded semantic distinctions that s/he may not realize exist in the Thai lexicon. The lexical entry comprises the English word followed by its Thai semantic equivalent, the pronunciation of which is first romanized in phonemic transcription with IPA symbols (our romanization system has

modified Mary Haas' system to conform to IPA) and then transcribed in standard Thai orthography. In addition, example sentences are provided for grammatical items, such as verbs associated with people, stative verbs, adverbs, time words, prepositions, conjunctions, and question words and phrases.

The dictionary comprises the following sections: (1) *Preface* which explains why we decided to compile this kind of English-Thai dictionary and how we selected semantic areas and lexical items for inclusion; (2) *Introduction to Thai Pronunciation* which compares Thai and English phonologies and instructs the foreign student in the pronunciation of Thai consonants, vowels, and tones; (3) the main body of *The Dictionary*; (4) *References* which lists some Thai dictionaries and textbooks which can be regarded as helpful for foreign students studying Thai; (5) *Alphabetical Index to English Vocabulary* which lists all English items from the dictionary in alphabetical order with page numbers on which they appear in the dictionary; and (6) *Alphabetical Index to Romanized Thai Vocabulary* which lists all Thai entries in the alphabetical order of their romanized transcriptions with page numbers in the dictionary.

The compilation of the dictionary is now almost complete, and the authors are now editing the manuscript (over 500 pages in length, not including the two alphabetical indices) in preparation for sending it to the publisher.

Moving from Mienh-English through Mienh-Thai toward a Trilingual Dictionary

Daniel T. ARISAWA

Payap University, Chiangmai, Thailand

This paper reports on the embryonic project of an Iu-Mienh-Thai dictionary. Iu-Mienh (henceforth shortened as Mienh), belongs to the Hmong-Mien (Miao-Yao) language family, and is spoken in five southern provinces of China, the northern parts of Vietnam, Laos, Thailand, and the immigrant communities on the west coast of the U.S.A. The Thailand speech variety is used for this dictionary. Although there are three Mienh-English dictionaries (Lombard-Purnell 1968, Panh 1995, and the forthcoming one by Burgess, Pienh, and Purnell), the Mienh in Thailand have expressed their frustration about the lack of a Mienh-Thai dictionary. Stemming from such a background this paper discusses the cross-national orthography issue, the influence of existing dictionaries, monolingual definition through monolingual elicitation, and the expected impact on language development.

Lombard, Sylvia J. (compiler) and Herbert C. Purnell, Jr. (editor). 1968. Yao-English dictionary. New York: Southeast Asia Program, Department of Asian Studies, Cornell University.

Panh, Smith (Aka Koueifo Saephan). 1995. Mienh In-Wuonh Dimv Nangc Sou--- Mien-English Everyday Language Dictionary. 1st edition. Visalia, CA: privately published.

Report on the Iu Mien - Chinese - English Dictionary Project

Greg AUMANN and PAN Chengqian

SIL International and Central University for Nationalities

This dictionary is a dictionary of Iu Mien as spoken in Laibin County, Guangxi Zhuang Autonomous Region, China. The intended audience includes Chinese and English speaking linguists. But the main audience is Iu Mien speakers. The dictionary has some unusual features designed to make it simpler and more useful for Iu Mien speakers. The first is that it has definitions in Iu Mien. Thus is like a combined monolingual and bilingual dictionary. The second is that the Chinese part of entry includes a Chinese gloss, pinyin of the gloss and a Chinese definition. It is hoped that the dictionary will be useful in helping Iu Mien to learn to read their own language and also to learn Chinese pinyin and characters.

This dictionary will have reverse Chinese - Iu Mien and English - Iu Mien indexes. Generating the reverse Chinese - Iu Mien index requires being able to sort Chinese in a standard order. The order most frequently used in modern dictionaries in China sorts characters first by pronunciation, then by stroke count and thirdly by stroke categories. This sort order is easy to use for Chinese speakers but difficult to implement. It requires disambiguation for characters with multiple pronunciations and the stroke counts and types for each character in the reverse index. The available databases of character information were generally not accurate and didn't contain all the necessary data so character data suitable for sorting modern simplified Chinese was developed.

The ABC Colloquial Cantonese-English Dictionary

Robert BAUER, Cheung Kwan-hin, Tang Sze-wing, Roxana Fung, Cathy Wong
Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University

The Hong Kong Polytechnic University research team is now in the process of compiling a computerized Cantonese-English lexical database which forms the basis of *The ABC Colloquial Cantonese-English Dictionary*. In this dictionary all lexical items will be listed in the alphabetical order of their romanizations, that is, according to the ABC principle. This principle, which is downright revolutionary within Chinese lexicography but pretty much taken for granted elsewhere, strictly arranges Chinese words in the alphabetical order of their romanized spelling and ignores their corresponding Chinese characters. The extraordinary beauty of the alphabetical arrangement of the romanized forms of Chinese words is that it allows the user who has heard a word spoken to look it up without needing to know how it is written in Chinese characters. Prof. John DeFrancis of the University of Hawaii pioneered the application of the alphabetical principle to Putonghua in his *ABC Chinese-English Dictionary* (University of Hawaii Press and Chinese University of Hong Kong Press, 1996; the revised and expanded edition was published in 2003). Today no comprehensive bilingual Hong Kong Cantonese-English dictionary of this kind is available to students and scholars of Hong Kong Cantonese. Thus, this bilingual dictionary will put the Cantonese language into the hands of international visitors, business people, students, and scholars who would like to learn to speak Cantonese.

The ABC arrangement of Cantonese words follows the alphabetical ordering of romanized forms of words and disregards the Chinese characters by which words are written as indicated below:

- baa1** 巴 n. shortened form of *baa1 si6/2* □□ bus
- baa2** 把 clf. for small bundle
- baa3** 霸 vb. occupy; take by force
- baa1 bai3** 巴閉 sv. showy; acting high and mighty; grand and magnificent
- baa2 gwai2** 把鬼 sv. useless; of no value
- baa6/2 laa3** 罷啦 utterance-final particle for polite suggestion
- baa2 mak6** 把脈 vb.o. (in Chinese medicine) feel pulse
- baa1 neoi5/2** 吧女 n. bargirl
- baa2 paau3** 把炮 n. confidence; skill; ability
- baa2 pei3** 把屁 des.ph. utterly useless (stronger than *baa2 gwai2*)
- baa1 si6/2** 巴士 n. bus (English loanword)
- baa3 wai6/2** 霸位 vb.o. occupy a seat; hold a seat for another
- baa3 wong4** 霸王 vb. do something without paying for it
- baa3 wong4 faa1** 霸王花 n. kind of lotus (its bud used in making a Cantonese soup)
- baa1 zaa1** 巴撻 sv. talkative, likes to comment on what is right and wrong; meddlesome

The numbers at the end of each syllable indicate the tone categories. According to the ABC principle of alphabetization, letters take precedence over symbols for tone categories.

In the computerized database each lexical entry includes the following items: romanized Cantonese word; Chinese characters with which the word is ordinarily written in Hong Kong; part of speech; English semantic equivalent; source of the Cantonese word; phrase or short sentence exemplifying the use of the word; source of the lexical example; and serial number. The dictionary will comprise about 10,000 lexical entries. Lexical material is now being collected from published dictionaries, glossaries, lexical databases, studies on the Cantonese lexicon, as well as the research team members' own intuitive knowledge of Cantonese. The lexical items come from the colloquial language of daily conversation used by educated speakers. The Linguistic Society of Hong Kong's romanization of Cantonese pronunciation reflects the conservative standard as recognised and accepted by educated speakers. The colloquial lexical material covers a wide range of semantic areas, such as: prepared foods, drinks, condiments, vegetables, fruits, meats, flavours, meals of the day, methods of preparing foods; domestic and wild animals, insects, fish, bodyparts of animals; festivals, holidays; house furnishings and parts of the house; illnesses, physical conditions, medications, human bodyparts; weather, topography, plants, other natural phenomena; people by age, sex, family and social relationship, occupation, ethnicity, activity engaged in; grammatical categories of noun classifiers, deictics, directions, exclamations, onomatopoeia, relative pronouns, personal pronouns, quantifiers, question words, adverbs, action verbs, prepositional verbs, stative verbs, modifying phrases, other kinds of verbs, time words, utterance-final particles, utterance-initial particles, verb-aspect markers, slang expressions, other grammatical forms; time words; tools, instruments, modes of transport, etc.

From Language to Ethnolect: Maltese to Maltraljan – A case study in Cross Continental Lexicography

Roderick BOVINGDON
Merrylands NSW, Australia

This paper looks into the first ever formal compilation of a select glossary, of one of a number of newly emergent ethnolects, within an Australian sociolinguistic environment: a direct result of Australia's multi-ethnic and multi-linguistic population composition.

All the lexical items extracted from the written form presented here, were selected solely from locally produced émigré literature. The material from the spoken idiom is my personal record, noted in face-to-face meetings with individuals during normal day-to-day conversations, as well as from Maltese language radio programmes from different regions within Australia.

The dictionary form is my preferred approach as the most scientific methodology for the glossary discussed in this paper. It features many of the more salient language traits of the lexical material collated. Such approach facilitates for further delving into this particular language deviation while laying the groundwork for a more comprehensive record for future similar lexical compilations.

Bilingual Dictionary of Computer Terms: A Need for the Intellectualization of the Filipino Language

Imelda P. de CASTRO
De La Salle University- Manila, Philippines

It cannot be denied that Science and Technology at present times is rapidly changing, continuously developing and becoming more meaningful as more and more people are able to avail of the new entities it can offer. To prove my point, it is actually the current century, which brought us from horse drawn carriages to automobiles, which utilize gasoline, from stoves, which needed wood and have to be fanned all the time to microwave ovens and from Victor alphas to DVD players which allow us to listen to good quality music. All these things mentioned are just a few of the many changes that actually occurred and continue to occur in the present society. And along with side-by-side these changes are the emergence of new terminologies, meanings and definitions and alternative terminologies, thus, the inevitable development of Language. This is actually the main reason why this researcher opted to make a project on Bilingual Dictionary of Computer Terms. This researcher observed the need for a supplementary material for people who are both technologically inclined and challenged. With the emergence of new technological terminologies, the intellectualization of language is seen by this researcher to be essential as any other basic needs.

The first lexical elaborations were presented: 1) direct borrowing of technical terminologies from a foreign language, 2) Spelling borrowed technical terminologies based on the native alphabet, and 3) Formulation of new technical terminologies. It is actually these three possible lexical elaborations, which this researcher used in the development and creation of a bilingual dictionary.

The first lexical elaborations, the direct borrowing, require that the spelling and pronunciation of the particular terminology that has been borrowed will not be modified. For example, The English term "*computer*", if borrowed by another language, will remain as it is.

The second lexical elaboration, the spelling of the borrowed technical term based on the native alphabet, can be described by the phrase "spell as pronounced." Using the English term *computer* to be spelled using the Filipino alphabet, it will be spelled as *kompyuter*. The pronunciation remains the same.

Last is the formation of new technical terms. This particular lexical elaboration is non-limiting for it gives freedom in terms of the creation of new words because pronunciation and spelling can be altered and modified. For example, the English term "computer" can this researcher create "Makinang Datos Taguan".

The following is an example of a bilingual entry that has been included in this research Using *English* as the source language and *Filipino* as the target language:

Download- download (pd) (kol) 1. Ito ay ang pagkuha ng mga impormasyon sa kompyuter. 2. Proseso ng pagpapadala ng impormasyon sa kompyuter. 3. Pagkuha ng impormasyon sa internet. 4. Ang halimbawa nito ay pag-apgreyd ng sistem ng inyong kompyuter upang makapunta sa bagong programa lalo na kung windows ang gamit, ang kabaligtaran ng prosesong ito ay apload.

The preceding entry, translated into English, is as follows:

Download- download (v) download (v) (col.) 1. Getting information from the computer. 2. Process of sending information to the computer. The opposite of this is uploading.

The current policy in the bilingual education states that in order to attain competence in two languages, the native and foreign, the two should be used as mediums of instruction in the academe at all levels. Given this, it is important that the language teacher must be aided and supported by materials that will greatly help in the efficient edification of the language and effective learning of the language learners. A Bilingual Dictionary can do this.

Digital Resources for SEA Lexicography: the SEALANG Projects

Doug COOPER

Center for Research in Computational Linguistics, Bangkok, Thailand

The SEALANG projects build digital lexical resources for the historical and modern languages of Thailand, Burma, Laos, and Cambodia, and for loan languages like Sanskrit and Pali. Resources include dictionaries in both e-text and digital-image format, modern and historical text corpora, and an extensive set of software tools, all tied together with server-based "scholar's workbenches." Applications include modern lexicography, etymological research, and development of innovative teaching/study tools. This talk will demonstrate many of our resources, and discuss underlying design issues and future plans.

Typological Classification of Dictionaries

Shalom DEVAPALAN
TAFTEE Bangalore, India

The present paper strives to study and analyse the different important typological classification of dictionaries proposed by Malkiel (1959, 68); Thomas Sebeok (1962); Zgusta (1971); Ali-M-Al-Kasimi (1971); Bo Svensen (1993) and other scholars, and an attempt is made to arrive at a new typological classification which tries to include all the most important types of dictionaries which are classified on the basis of **internal** and **external** features. The purpose of the new typology is to locate the proposed dictionary with all its different aspects (in the correct slot) in the new typology.

The Munda Lexical Archive

Patricia Donegan and David STAMPE
University of Hawai'i at Manoa

The occasion of the SEALS and the Asian Lexicography Conferences in Thailand in May 2004 will also be the occasion of the opening of public access on the web to the Munda Lexical Archive at the University of Hawai'i at Manoa. The Archive will provide online access to the available lexicographic data on the Munda family of Austroasiatic languages of India. It consists of the merged lexical materials of eight of the ten major Munda languages. For the other two, Santali and Mundari, the encyclopaedic dictionaries of Bodding and Hoffmann have been announced as available or forthcoming on other sites, and it is hoped that mutual arrangements for mirroring each other's sites can be arranged so that all the materials can be accessed at once by scholars and students anywhere in the world who share an interest in the South and South-East Asian linguistic areas or the Austroasiatic languages.

Although the Archive includes data from many previously published wordlists and dictionaries (some very rare), it mainly makes available previously unpublished materials, particularly those collected by members of the joint Indian and American field project organized in the early 1960s by Norman Zide. In addition to materials on the individual languages there will also be etymological materials, published or unpublished, assembled from comparative studies.

The materials in the Archive will be published in paper form when it seems appropriate, but they are far more usefully accessed in electronic form. Furthermore, as its name suggests, the Archive is intended as a public resource not only for students and scholars, but also for the speakers of the languages, in the hope that they will vastly improve the scope and accuracy of what has been done so far. To that end, the only restriction on the use of the Archive will be that its users may not restrict its availability to others.

This brief description of the Archive will present an overview of its format, its content, its many contributors, and how it can be used and improved. Some examples will be presented to illustrate the great diversity of structure within Munda (in some respects as great as that of Austroasiatic) and to show how much remains to be learned while there are still speakers of these languages.

Monolingual dictionaries in south Pacific languages

Robert EARLY

Pacific Languages Unit, University of the South Pacific, Vila, Vanuatu

The South Pacific region is an area of widely dispersed human communities living in an environment that is sometimes idyllic, and sometimes threatening. The physical, social and cultural integrity and viability of these communities cannot be taken for granted. Languages in particular may be endangered on a wholesale basis.

Of the 1200 or so languages identified in the region (non-Austronesian; Oceanic Austronesian [including Polynesian, Micronesian, and most of the languages of Melanesia], and Australian) the large majority have had very little attention from linguists. Most are still unwritten. Of those which have had some work done on them, the materials consist of basic descriptive materials such as bilingual wordlists, and missionary translations.

Since the time of European contact, the best described languages have been those of the Central Pacific, namely Fijian and the Polynesian languages. In more recent decades, a new phase of development has led to interest in the production of monolingual dictionaries for some of these languages.

At this point, there is no full monolingual dictionary in print for any of these languages. However, monolingual dictionary projects are underway for at least Fijian, Niuean, Samoan, Tongan, and Tuvaluan. These projects are of special interest because unlike most other linguistic work on these languages, they are characterised by a high level of ownership and participation by motivated native speakers.

This paper describes these projects, and analyses the particular approaches to lexicographic work that have been adopted in each case. Answers are given to other questions:

- What is seen as the purpose of a monolingual dictionary for the language community?
- What kinds of training and skills do the compilers have?
- What approaches have been taken to the task of compiling entries, writing definitions, selecting information to include in entries?
- How have issues like computer use and data management been handled?
- What about other logistical matters like project organisation and funding?

The Dictionary of Dunhuang Characters: Towards a sample base of characters from medieval Chinese manuscripts

Imre GALAMBOS

International Dunhuang Project, The British Library, United Kingdom

The creation of the Dunhuang Character Dictionary is a pilot project of the International Dunhuang Project at the British Library. It offers a sample base of characters found on dated medieval manuscripts from Dunhuang and Chinese Central Asia in general. Although only a fraction of the documents found in this region carry a date, the large size of the corpus still yields a sufficient amount of dated manuscripts that can be utilized. The dictionary is essentially a collection of

characters from these dated documents arranged into a database format. Its main application is dating other manuscripts by means of comparing their orthographic and calligraphic attributes to those listed in the dictionary.

The dictionary breaks with the Chinese lexicographic tradition in that it lists *all* characters from the utilized documents, not only those different from each other. This way, researchers have the ability to both detect distinctions and match similarities. Moreover, they can now make decisions on the basis of a statistically significant amount of data, which significantly increases the accuracy of their results. The inclusion of all character forms also shifts the task of determining the identity of certain character forms from the lexicographer to the researcher; thus the dictionary serves as a research tool, instead of a set of predigested material.

An important aspect of the work is that all character forms appear as photographs. On the one hand, this eliminates many of the common mistakes associated with tracing and redrawing the original characters by the lexicographers. On the other hand, the dictionary can be used by itself as a first-hand reference work, without the need of locating the original characters in other sources.

Formerly, a dictionary of this size was simply not feasible because of the time and cost associated with extracting and collating the individual character forms. In recent years, however, as a result of the work of the International Dunhuang Project at the British Library and the National Library of China, an increasing number of manuscripts has been digitized and put online. Beside the availability of high quality digital photographs, the other major milestone in the compilation of the dictionary was the development of a visual software tool used for extracting the coordinates of each character in a manuscript. Instead of having to cut out, name and save each character individually, now we can detect the location of a character within the manuscript page semi-automatically. Then, using the extracted coordinates, we use a fully automated command-line tool to cut out the characters and name the new files. The semi-automatic process not only considerably speeds up work but also eliminates mistakes in naming files.

In the course of compiling the dictionary, we have developed a set of tools and established an efficient procedure which can be applied to the creation of additional dictionaries, including one in languages other than Chinese. The prime candidates for such new projects would be the languages used in the same corpus of Central Asian manuscripts (e.g. Tibetan, Sogdian) but other scripts could be considered too.

Ouch! Don't print that! Political correctness in Gurung lexicography

Warren GLOVER
SIL South Asia

Controversy over certain words included in a recent (May 2003) dictionary of the Gurung language of Nepal, leading to a court case against the editors and all the advisory committee, alerts one to some hazards facing the compiler of dictionaries. This article traces the development of Gurung lexicography and decisions regarding dialect and orthography. It shows that the controversy and litigation arises from a combination of language and culture change over a few decades, together with competition and envy within the community.

Cognitive Grammar and Lexicography

Douglas INGLIS
Payap University and SIL International

In general lexicographical practice, theory is not often used to support lexicographic characteristics (Geeraerts 1987:1). This paper shows how theory, Cognitive Grammar specifically, can be used to support lexicographic decisions. The paper first considers Langacker's analysis of English episodic nominalization and verbalization (Langacker 1991:24-5). It then shows how the semantic intuitions of these two processes established from the theory can be characterized in a lexicographic entry. An episodic nominalization takes what Langacker calls a perfective verb and uses it as a noun. An example is found in the pair of sentences, *He will **walk** around. He will go for a **walk**.* The verbalization process is seen in the pair of sentences, *He tasted the **salt**. He **salted** the food.* Each pair of examples, though quite similar, represents a different semantic process developed conceptually in the analysis. These two processes can then be accounted for in lexicographical practice by standard conventions of range and sense as practiced by Newell (1995). Cognitive Grammar, in as much as possible, uses theoretical notions founded in cognitive psychology. The goal of these notions is to capture linguistically marked semantic nuances and intuitions of a language which makes this a good theory for applications such as lexicography. In this way, lexicography finds a suitable counterpart in Cognitive Grammar for the motivation and explanation of its intuitions.

- Geeraerts, Dirk. 1987. Types of semantic information in dictionaries. In R. Ison (ed.) *A Spectrum of Lexicography*. Amsterdam: Benjamins.
- Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar*, Volume 2, *Descriptive Application*. Stanford: Stanford University Press.
- Newell, Leonard E. 1995. *Handbook on Lexicography for Philippine and Other Languages*. Linguistic Society of the Philippines and SIL, Manila.

The treatment of cultural elements in English-Tamil dictionaries

Gregory JAMES

Language Centre, Hong Kong University of Science and Technology

Nowadays, lexicographers are acutely aware of the necessity for a 'user perspective' in dictionary compilation, and a good deal of research has been carried out by academics and publishers to determine what users need from a dictionary, what frustrates them, and what makes for success in the dictionary consultation process. This paper examines a number of English-Tamil dictionaries from the perspectives of a Tamil-speaker wanting to know the meaning of an English word, of an English-speaker wanting to know how to express a particular word or concept in Tamil. Although the linguistic needs of these two groups of users are different, lexicographers often focus their dictionary entries towards one or other group. For example, most modern English-Tamil dictionaries are explicitly written for Tamil-speakers learning English. The lexicographer's focus enables certain assumptions to be made about a user's knowledge, which, in turn, informs the orientation of the dictionary, and the representation of definitions in the entries. However, what if, now, we superimpose a dimension of unpredictability in users' knowledge, even within one speech community? The realm of socio-cultural vocabulary, specifically that of religion, where many terms are 'untranslatable' from one language to another, provides such a dimension. Since lexicographers are unable to predict the particular religious allegiances of the users of their dictionaries, entries are often couched in vague, neutral terms, because no assumptions can be made about background knowledge on the part of the users. Some terms used in three major religions - Hinduism, Islam and Christianity - are compared and contrasted, to show how English-Tamil lexicographers have resolved the often lack of one-to-one correspondences in concept and terminology between the two languages, and how their solutions in terms of definitions and translations may help or hinder a dictionary user. From the evidence, conclusions are drawn relating to general principles of cross-cultural and bilingual lexicography.

The Dewan Malay-English Dictionary: a fresh departure

Fadilah JASMANI

Institute of Language and Literature of Malaysia, Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia.

This paper reports on the first Malay-English Dictionary project to be undertaken by Dewan Bahasa dan Pustaka, Kuala Lumpur, Malaysia. The target audience of this dictionary are English-speaking second language learners of the Malay language (bahasa Melayu or bahasa Malaysia as it is known in Malaysia).

One of the essential features of a L2 learner dictionary is the presentation of meaning. In this new dictionary, we explain the meaning of words in full Malay sentences, following the practice attributed to Collins COBUILD. This is a fresh departure from the way meanings have been presented in previous dictionary projects undertaken by DBP. Also, the corpora used in the compilation of the dictionary is the Bank of Malay, a collection of 95 million words of written and spoken Malay held on

computer in Dewan Bahasa dan Pustaka. This dictionary will contain 40,000 headwords, a considerable coverage of the basic and semi-technical words in the Malay language.

UniSpell and Lexical Databases

Tom LASKOWSKE

SIL International

Where does one look to know how to spell a word correctly? In the dictionary! Because spell checking is one of the popular uses of a dictionary, anyone compiling, or helping others to compile, a dictionary should know how to ensure that everything in the dictionary is spelled consistently. But for most of the world's languages there is currently no commercial spell check program available. UniSpell has been created to address the need for spell checking in such languages.

UniSpell has features that set it apart from other spell check programs and these will be looked at briefly. It can check specific fields in a database that uses standard format markers. It can check multiple languages in the same document, which is particularly useful for spell checking dictionaries. And, it is Unicode compatible.

But a major challenge to those working with vernacular languages is the process of building a spell check database. This process is what the presentation will concentrate on. How can we avoid introducing errors? How can we keep the spell check database from "running ahead" of the lexical database? Using a Shoebox* parser, we will demonstrate a method of how to keep the entries in your lexical database up to date while at the same time building a spell check database in UniSpell that you can be confident in.

Challenges of the Sherpa-Nepali-English-Tibetan Dictionary Project

Sang Yong LEE

Korean Research Institute in Language and Culture

Sherpa, which is a Tibeto-Burman language of Nepal, does not have standardization yet in script, dialect, and orthography. Involvement of the language group is important in the very first processes of dictionary production. Without consensus from the whole language group the completed dictionary is difficult to get approval from the main society.

Dialect: In sociolinguistic survey to find out the more prestigious dialect among three different dialects, the southern one was chosen.

Script: Even though a field test showed that 88% of the people wanted Devanagiri script, and that script was developed and used for several years, the Sherpa society recently decided to use Tibetan script for their language in their primary school textbook. So, painful change was accepted from Devanagiri to Tibetan script. This change will have effect not only in the script of the main entry, but also in sorting order in dictionary. This incident will be a good teaching to the field lexicographers in similar situations.

Orthography: Stimulating the language group to see the need for standardization of their language, and to make some appropriate actions, is very crucial in dictionary making process. In this sense dictionary production is a basic groundwork for literacy and translation of written documents in underdeveloped language groups.

Multi-Dictionary Formatting in an eXtensible, Open Standards world

Greg LYONS

Payap University and SIL International

The Multi-Dictionary Formatter (MDF) has been in wide use for many years. It represents both a small set of schemas for structuring multilingual lexicographical databases and a tool for quickly producing from such a database many different types dictionaries targeting different purposes, audiences, and media. The development of "eXtensible MDF" brings this proven utility into the world of Open Standards and thus provides a platform for addressing many outstanding technical challenges to be faced in the process of "Making Dictionaries".

A key challenge is complex writing systems which can not supported via MDF 4.0's export to Microsoft Word. One MDF XSLT option will address this thru Graphite enabled Open Office.

How Many Words Does Bisu Have? – An Asian Hilltribe Encounters the “Dictionary Development Program”

Kirk PERSON

Payap University and SIL International

During the late 1990's, Ron Moe of SIL International began experimenting with what he termed the “mass production” of dictionaries. Underlying his approach was the assumption that “the mind has the incredible ability to assimilate, organize, and recall massive amounts of information” (Moe 2003: 56). But how could one tap into these vast mental resources?

Moe's solution involves the use of an impressive list of some 2000 semantic domains, all accompanied by questions intended to spark spontaneous mother tongue speaker elicitation. For example, a mother tongue speaker (or, better yet, a group of mother tongue speakers) would be presented with a domain such as ‘sky,’ and then be asked to write down all the items found in or related to ‘sky.’ In the case of the Lunyole people of Africa, a two-week long workshop yielded about 17,000 lexical items—a vast improvement in the ‘normal’ rate of lexicon compilation.

This paper chronicles the application of Moe's techniques in a new setting, among the Bisu people of Southeast Asia. During March of 2004, some 35 Bisu people from three countries, representing four distinct but mutually intelligible dialects, came together in Thailand for a lexicography workshop. Moe's materials, translated into Thai, were utilized for the workshop. The results of the workshop were encouraging, both in terms of linguistic data collection and positive impact upon the language attitudes of this small (total population 6,000), endangered language group.

Pan-dialectal databases: Mlabri, an oral Mon-Khmer language

Jørgen RISCHÉL

University of Copenhagen

The success of comparative research on Austroasiatic is dependent on the accessibility of reliable lexical and grammatical information on the numerous oral languages and dialects that are spoken by minority groups across Mainland Southeast Asia. The documentation of many of these is limited to basic word lists of varying quality. Even when one has satisfactory access to word lists or to speech data from one or more varieties of such a language, other varieties may differ so significantly that one cannot claim to have access to a representative sample or a vocabulary reflecting the language in its totality.

The Khmuic (Northern Mon-Khmer) language Tin, for example, has split into two lexically divergent main dialects. Still, there may be a tendency to think of Tin as one entity when searching for cognates across Austroasiatic. Since the most dramatic sound-shifts that have happened in Tin are shared by its two branches, this lumping together certainly makes sense phonologically but it is controversial from a lexical point of view. As dictionaries become available one may expect each such volume or file to cover one or the other dialect but hardly both. Thus, even in the future comparativists may have to search through different sources before it is warranted to make statements about the survival of lexical items in Tin.

Mlabri (the so-called “Phi Tong Luang”-language spoken on both sides of the northern border between Thailand and Laos) looks as if it may be a sister language of Tin which branched off some centuries before Tin itself split up. Unlike Tin, Mlabri is phonologically extremely conservative and retains features that can be reconstructed for Ancient Tin, so it is imperative to consider the two jointly in comparative work. Recent research has identified three varieties of Mlabri, which differ very little in phonology but considerably – and very interestingly – in everyday lexicon. There is reason to assume that this reflects a deliberate polarization in terms of jargon among clan-like subgroups (the picture is complicated by additional dichotomies of male vs. female, and obsolescent versus current usage). As with Tin, word lists for Mlabri are typically not representative of the language as a whole but at best of one variety.

The present paper will outline an ongoing project with the goal of establishing a fairly extensive lexical database covering all three varieties of Mlabri, with consistent indication of the provenance of each form, and with very extensive use of phrasal and sentential examples to document the sub-meanings (polysemy) and associated syntactic characteristics of each lexical item. A derived product is a compact cross-dialect dictionary, the prototype of which has already proved useful in comparative work and as a tool in the analysis of narrative texts (since these often contain words which have not been recorded from daily use in the same variety of Mlabri but only in another variety). In the paper various linguistic challenges posed by such work across ethnic varieties of a “small” language will be presented, ranging from the difficulty of retrieving genuine speech data from timid speakers with very limited bilingualism to the handling of idiosyncratic variation over forms or meanings.

A project to establish an on-line Mon-Khmer Comparative Etymological Database and Language Documentation repository as a basis for international research cooperation

Paul SIDWELL

Department of Linguistics, Research School of Pacific and Asian Studies, Australian National University, and Centre for Research on Language Change

The paper discusses a new project to establish an online Mon-Khmer Comparative Etymological Database Language Documentation repository. The thrust of the project in the first place is historical, but it has many aspects and consequences that relate directly to lexicography. The project will create a web-site which will present not only a comparative etymological database, but also is planned to include access to many of the lexicons which underlie the etymological comparisons. This is possible because in many cases the available documentation for comparative historical analysis consists of unpublished wordlists and draft dictionaries, and as these are being converted into electronic format and edited for use in historical research it also becomes practical to make them browser friendly, effecting direct electronic publication, greatly improving their usefulness. Practical questions then arise concerning the appropriate scope of content and format of such electronic lists, access and intellectual property rights, and the relationship to printed versions derived from or otherwise related to the web-based lexicons. The data repository may also store simple scans of unprocessed data and other media such as compressed video and/or audio files.

The etymological database will be interactive through common browsers, so that potentially a cooperating researcher anywhere in the world will be able to upload and manipulate data, subject to appropriate protocols. This should facilitate ease of research cooperation, giving a boost to an area that has otherwise languished due to reliance on individual projects and outmoded and cumbersome models of presentation and publication. Various technical challenges remain to be overcome to achieve complete functionality and this may lead to a phased implementation. A preliminary gateway to the project has been posted at:

http://www.anu.edu.au/~u9907217/db/db_project.html.

Sedang Dictionary

Kenneth SMITH

SIL International

Ken Smith will discuss various features he sought to develop in his recently published dictionary of the Sedang language (Vietnam, Mon-Khmer): *Sedang Dictionary with English, Vietnamese, and French glossaries: A Thesaurus-Alphabetical Listing of Sedang Words and Word-Groups* (*Mon-Khmer Studies Journal*, Special Issue No. 1; 2000; xlv, 567 pp.), a companion volume of his *Sedang Grammar*.

The development of a multilingual database for a Chinese-Tamil dictionary

Bronson SO Ming-Cheung

Language Centre, Hong Kong University of Science and Technology

The advance of computer technology is now playing an important role in the development of dictionary writing systems, and providing us with a variety of ways to manipulate them. Although the value of dictionary writing systems has long been recognised, such systems are still being developed in which lexicographic elements can be optimised and text can be characterised for special languages. In this paper, the presenter will first describe his role in designing and developing a relational database and writing system for a Chinese-Tamil dictionary (compiled by Sridharan Madhusudhanan of the Indian Foreign Service), which supports Chinese, pinyin, Tamil and English, and in which contents are specialised for each component in the dictionary, such as text, transcription, definition etc. He will then examine this system with respect to model, structure and presentation. Lastly suggestions will be made as to how to optimise this system to suit other dictionary projects.

An ideal monolingual learner's dictionary of English for Thai speakers

SUMITTRA Suraratdecha

University of Hawaii at Manoa

This paper discusses the design of an ideal monolingual learner's dictionary of English for Thai speakers. In the last decade, there have been significant improvements in the monolingual learner's dictionary in two senses: 'the description of a language that a dictionary provides corresponds more closely to reliable empirical evidence regarding the way in which the language is actually used; [and] the presentation of this description corresponds more closely to what we know about the reference needs and reference skills of the target user' (Rundell 1999:316). In order for this ideal dictionary to reflect the improvements mentioned above, its design will be based on the needs of Thai speakers and discussions of improved features in contemporary monolingual learner's dictionaries. This paper aims to give an outline of linguistic and cultural information needed by Thai speakers learning English. The scope of the discussion covers the following topics: target users and their needs, learners' reference skills, pronunciation, morphology, syntax, usage, and culture-specific items.

Thesaurus and Dictionary Series of Khmu Dialects in Southeast Asia

SUWILAI Preamsirat

**Institute of Language and Culture for Rural Development, Mahidol University,
Bangkok, Thailand**

Khmu is an Austroasiatic language spoken by 600,000 people living across countries in northern Southeast Asia. This paper presents the objectives, data collection and data presentation of Thesaurus and Dictionary Series of Khmu dialects.

The field research on Khmu was carried out on seven Khmu dialects in Thailand, Laos, Vietnam and China (Yunnan). The main objectives were to document the Khmu dialects as thoroughly and completely as possible in order to form a complete picture of Khmu language and to present its lexicon using the Thesaurus semantic format of grouping words by related meaning. Apart from that the lexicon of the major Khmu dialects spoken in each of the four countries where the Khmu people live are also alphabetically arranged using the Dictionary format so that it would be convenient for further language development including orthography development and literature productions. The description of Khmu prosodic, lexical and syntactical characteristics is also provided.

More than plain words: A report on the emerging lexicography of Austronesian languages in Taiwan

Josef SZAKOS

**Department of English Language, Literature, and Linguistics, Providence University,
Taichung and Department of Aboriginal Language and Communication, National Dong
Hua University, Hualian, Taiwan**

Taiwan is regarded as the homeland of Austronesian languages. There are still at least fifteen languages spoken, fragmented into almost 50 dialects. Due to the tumultuous history of the Island and the population, dictionary making for the aboriginal languages has remained a luxury for a few devoted scholars up to the recent decade.

In my presentation, I intend to give a short history of the major stages of this development, when half of these languages stand on the threshold of disappearance. I introduce the major practical dictionaries created by missionaries in the second half of the 20th century for Bunun, Paiwan, Amis, Puyuma, Taruku languages.

Then I plan to introduce the projects started by the Aboriginal Affairs Commission for the lexical documentation of some of the languages. Scholars were relatively slow in this field, also for the lack of recognition connected with this kind of work, but recently some word-lists done by Japanese colonial researchers are being published e.g. for the Pazeh language. A very important milestone is the work on Thao, created by Robert Blust, published last year, which integrates the vast historical linguistic knowledge of the author with the huge amount of data he was able to gather in a few years' time.

My own projects on the Tsou, Kanakanavu and Saaroa lexicon will soon reach the users - and I also intend to use this occasion to seek solutions to some unresolved problems in the analysis and presentation of these underdocumented languages.

The morphology of Austronesian languages is centered around the verbal roots, somewhat reminding us of the Semitic lexicography. But this fact leads to the problem of practical arrangement of entries in the dictionary. It is still an unresolved problem in Taiwan, how to preserve the balance of an alphabetic order presentation and a root-based arrangement.

Presenting grammar and cultural information is another difficulty, where we are searching for solutions for the diversity of theories. We try to satisfy the need for practically adequate bilingual dictionaries for the Aboriginal/Chinese speakers, while trying to present the pairs of typologically so divergent languages on a similar scale. I also intend to give an account of corpus use (spoken and written) in the creation of the newest dictionaries for these languages.

There are many points of convergence, where, in my opinion, we have to join the mainstream of Asian Lexicography and I feel this introduction could be a good chance for this. The field may be very broad, but my decades of experience tell me that we are moving in the right direction, and hope we succeed before there is nothing more to record or to document.

Serving as a Consultant to Dictionary Compilers

Doug TRICK
SIL International

This session deals with various practical aspects of how to offer consultant help to someone who is compiling a dictionary. It is recommended that a consultant check be conducted in at least two stages. Procedural issues are presented dealing with: purpose and audience, language-related matters, front/back matter, principles of lexicography, and dictionary entries. The paper concludes with a 26-item checklist, which may be of some value to compilers and consultants alike.

The presenter, a member of SIL Philippines, has consultant-checked two bilingual dictionaries documenting Philippine languages. Since 1990, he has also been compiling a dictionary of Southern Sama (Tawi-Tawi, Philippines and Sabah, Malaysia).

Lexical Database of the Torwali Dictionary

Inam ULLAH
Pakistan

Torwali is one of at least twenty four lesser known languages of the northern Pakistan, that have been remained unwritten and previously less exposed to the academic community of the world. According to L.R Turner, 'these languages may not be significant politically, but historically they are of great importance'. There is hardly an institution in Pakistan that supports such activities as preservation and documentation of these lesser languages. Smaller groups or individuals strive to preserve and document their mother tongues, but due to the lack of a national policy

on languages and low priority given to them, only foreigners take pain to explore and analyze these languages. *The Torwali-English Urdu-Dictionary Project* is an example of an indigenous initiative, that of a non-linguist lexicographer from within the speech community without any institutional support.

In the paper that I propose, I will deal with several aspects of the lexical database of the Torwali dictionary. My objective is to highlight the beginning of the dictionary project by providing some background on the geographic setting and genetic affiliation of the Torwali language as well as the methodology used for collecting and compiling lexical material and the problem of deciding orthography for a previously unwritten language. I shall also talk about the linguistic software 'shoebox' which is a wonderful lexicographic tool for entering, parsing and interlinearizing the database that, I adapted for the compilation of the dictionary at a later stage. I will outline the various lexical fields used in the database and the items, which have been included in different files, such as, the main database, clans, plants, trees, proverbs, affixes and doubtful entries etc. I will then mention my involvement with the University of Chicago for recording sound files of the example sentences in the database and its inclusion among the Less Commonly Taught Languages of Pakistan on the World Wide Web. I will conclude my discussion by describing some the problems presently faced as well as some comments regarding the future of the this project.

The Wa Dictionary Project Report: A Dictionary of the Wa Language with English, Chinese, and Burmese (Myanmar) Glosses and Internet Database for Minority Languages of Burma (Myanmar)

Justin WATKINS

School of Oriental and African Studies, London, UK

Wa is a member of the Northern Mon-Khmer language family. More specifically, it is the major language of the Palaungic branch of Mon-Khmer. (See map and chart of Austroasiatic languages, which includes, primarily, the Mon-Khmer family.) Wa is spoken by about one million people in an area on the border between China's Yunnan Province and the Shan State of the Union of Myanmar (Burma). The Wa language has been the subject of several linguistic studies, most of them in Chinese (see our table of publications in and about the Wa language). As noted in Justin Watkins' recently-published monograph *The Phonetics of Wa* (Pacific Linguistics 531, Canberra: Australian National University, 2002), the Wa language has featured in the phonetic literature chiefly on account of its contrastive use of vowel phonation, or register, variously described as breathy vs. clear or lax vs. tense. The only Wa dictionaries which exist are a few Wa-Chinese dictionaries which are either limited or out-dated.

The SOAS Wa Dictionary Project is a three-year effort (2003-2006), funded by the Arts and Humanities Research Board to produce a high-quality dictionary, translating Wa into Chinese, Burmese/Myanmar and English. The project will use advanced techniques in corpus-based lexicography, centred on a database and Internet resource, which will also be suitable for other languages spoken in Burma/Myanmar besides Wa after the life of the project.

The academic objectives of this project are:

- to compile an electronic Wa language corpus from printed and recorded materials suitable for linguistic and textual analysis
- to produce two dictionaries of the Wa language from the corpus: one scholarly and one for Wa-speaking end-users
- to record, document, preserve and disseminate the corpus using the Internet
- to establish, by the end of the project, an expandable template for corpus-driven lexicography and linguistic research on other minority languages of Burma.

The project benefits Wa speakers by:

- providing a dictionary targeting Wa-speaker end-users
- supporting literacy in Wa, Burmese, Chinese and English
- bridging the divide between the two main Wa orthographies (PRC and Revised Bible orthographies)