The Sino-Tibetan Etymological Dictionary and Thesaurus: STEDT Project Data Resources and Protocols

Richard S. COOK
rscook@socrates.berkeley.edu
John B. LOWE
jblowe@socrates.berkeley.edu
Linguistics Department
University of California
Berkeley, USA

Introductory Remarks

The *Sino-Tibetan Etymological Dictionary and Thesaurus* (STEDT, [stɛt]) Project began in August 1987 at the University of California, Berkeley, under the direction of Professor James A. Matisoff. More than thirteen years later the concrete goal of the project remains the same: the publication of an etymological dictionary of the Proto-Sino-Tibetan (PST) ancestor language. Scientific advances do not always happen over night, and so it may be no surprise that success in this colossal endeavor should require the protracted concentration of hundreds of linguists over such a long period of time. In fact, when the many sources whence STEDT derives its data are considered, the amount of time required for the reconstruction of PST far exceeds these thirteen years and the lives of STEDT's contributing researchers. The lexicographic task to which STEDT is committed is only the newest contribution in this field of linguistics which begins with the earliest inscriptions in the ST daughter languages.

The earliest inscriptions in any ST daughter language are the Oracle Bone Inscriptions (OBI) of Shang China, commonly dated to approximately 3500 years ago, to the middle of the second millennium BCE.[1] In this very early period the peoples speaking the early ancestors of the Sinitic and Tibeto-Burman languages were already separated from each other by perhaps 2500 years of pre-history.[2] Writing would not appear in Tibet for some 2100 years, in the seventh century.[3] Writing did not appear in Burma until the twelfth century, almost 900 years ago.[4] A systematic writing system did not appear for the Lolo-Burmese language Black Lahu until 1988, nearly twelve years ago.[5]

---

[1] Keightley(1978:91).

[2] "The Proto-Sino-Tibetan (PST) homeland seems to have been somewhere on the Himalayan plateau, where the great rivers of East and Southeast Asia (including the Yellow, Yangtze, Mekong, Brahmaputra, Salween, and Irrawaddy) have their source. The time of hypothetical ST unity, when the Proto-Han (= Proto-Chinese) and Proto-Tibeto-Burman (PTB) peoples formed a relatively undifferentiated linguistic community, must have been at least as remote as the Proto-Indo-European period, perhaps around 4000 B.C." (Matisoff, <http://stedt.berkeley.edu/html/STfamily.html>).

[3] Beyer (1992:40).

[4] Cf. Luce (1978).

[5] Matisoff(1988).

And so it is today that STEDT continues in this long tradition, striving to assemble, systematize, preserve and make publicly available lexical information on the daughter languages of Proto-Sino-Tibetan. This huge task would not have been even remotely possible before the age of computers, and so it is ironic in some respects that computers, the great English-language based cultural levellers of our modern day are making possible now what was before impossible. Many ST languages even today have been far less fortunate than Black Lahu-- they are either very poorly or almost completely undocumented. In our modern world minority languages are being lost at an alarming rate. Through the magic of computers STEDT is able to capture and preserve the accumulated knowledge of these dying languages in a form which lives beyond the walls of books in libraries, available at lightning speed, in the comfort of one's own home, electronically, all over the world. And even for modern languages of the family which are not presently endangered, STEDT preserves and disseminates priceless knowledge of their rich cultural heritages.

As an etymological dictionary project, it may not seem immediately clear that STEDT need be involved in such issues as minority language preservation. And yet, in this paper which is primarily concerned with describing the STEDT computer systems and the protocols developed for the treatment of over 13 years of accumulated linguistic data, we shall demonstrate that for the ultimate validity and longevity of this hard-won data, STEDT must also preserve electronic archives of its source materials in as pristine a state as possible. This type of language preservation, i.e. preservation of the captured source materials in electronic format, provides the only assurance that future generations will find the data and the conclusions based upon it as reliable as we do today.

The work of reconstructing Proto-Sino-Tibetan is much larger than any single person, and will occupy more time than is available in any individual lifetime. The STEDT Project Data Resources represent the contributions of a great many people over the course of many years. Some extent of our indebtedness is evident in the 465 items in the STEDT *Source Bibliography*. We are eternally indebted to those who have worked so hard over so many years to collect and prepare this data, and look forward hopefully to the work of future contributors. In addition to those named in the *Source Bibliography*, we must also acknowledge the work of the following STEDT Project contributors, including past and present staff members, support staff, graduate students, and volunteers, without whom, none of this work would have been possible.

Madeleine Adkins, Jocelyn Ahlers, Shelley Axmaker, Karin Beros, Balthasar Bickel, Brian Bielenberg, Leela Bilmes, Robert Bowen, Michael Brodhead, Bodine Brown, Jeff Chan, Melissa Chin, DAI Qingxia, Jeff Dale, Amy Dolcourt, Julia Elliott, Jonathan Evans, Jerilyn Foushée, GONG Hwangcherng, Daniel Granville, Joshua Guenter, William H. Baxter, Kira Hall, Nell Haskell, Timothy Hayes, Annie Jaisser, Zev Joseph Handel, Matthew Juge, Kerttu K. McCray, Nina Keefer, Jean Kim, George Kraft, Randy LaPolla, Jennifer Leehey, Anita Liang, Liberty Lidz, LIN Yin-ching, Liu Guangkun, Martine Mazaudon, Jean McAneny, Anthony Meadow, Boyd Michailovsky, Pamela Morgan, Karin Myrhe, NAGANO Yasuhiko, Ju Namkung, Peggy Nelson, Zhalie Nienu, Toshio Ohori, Weera Ostapirat, Stephen P. Baron, Jeong-Woon Park, Jason Patent, Karma Penjore, Chris Redfearn, S. Ruffin, Keith Sanders, Marina Shawver, Elizabeth Shriberg, Helen Singmaster, Tanya Smith, Gabriella Solomon, Silvia Sotomayor, SUN Hongkai, Laurel Sutton, Jackson Tianshin Sun, Nicolas Tournadre, Prashanta Tripura, Nancy Urban, Kenneth Van Bik, Kenneth Whistler, Blong Xiong, XU Xijian, YABU Shiro, ZHANG Jichuan, ZHANG Liansheng.

STEDT Data Systems: Resource Overview

The STEDT Project Databases are accessed through the "STEDT" database file. Opening this document presents the user with a palette entitled *The Sino-Tibetan Linguistics Digital Reference Library*. Within this palette are a number of buttons which bring the user to various portions of the data. Most broadly, the databases are divided into Primary and Ancillary groups.

- The **Primary Databases** are those which relate to the centralization of lexical entries for etymological purposes.
- The **Ancillary Databases** are those databases constituting STEDT's most pristine, most complete, highly indexed electronic versions of the original source documents, from which lexical entries for the Primary Databases are culled.

In the pages which follow, detailed descriptions of these resources are given, in the hope that this information may make data access an easy and pleasurable experience.

The six relational databases comprising the Primary STEDT Project electronic resources have been designed around a color-coded scheme which is consistent among the various related files. In all these files, *circular* buttons within particular records are used to call up *related* records. At the bottom of each database window are several *rectangular* buttons which are used to *navigate* among the separate databases, without actually altering the records currently displayed in a given window. The color codes for both small circular buttons and larger rectangular buttons are listed in the table below.

The following database files (detailed descriptions of which follow below) are now being beta-tested by STEDT Project members in a preliminary release CD-ROM as stand-alone Apple Macintosh applications. The Primary STEDT Database Files are also being developed as stand-alone applications for Windows (95, 98 and NT). Because of typographic complexities, world-wide-web delivery is now limited to Adobe Acrobat PDF files, generated on demand.

PRIMARY STEDT Database Files				
#	Database Name	Abbrev.	Total Records	Color Code
01	MAIN LEXICON	Lexicon	376,191	YELLOW
02	ETYMA	Ety	2,066	ORANGE
03	Language Names	Lgnames	1,786	YELLOW
04	Source Bibliography	Srcbib	465	GREEN
05	Language Groups	Lgrp	58	BLUE
06	STEDT Font Reference	SFR	231	PURPLE

The six items tabulated above under "PRIMARY STEDT Database Files" are relational databases which constitute the principal STEDT etymological research environment. Each of the six files serves as a window on the others, allowing researchers easy access to analyzed linguistic forms and reference data. All of the above items are now fully searchable in the unified color-coded graphical user interface. Researcher response has been overwhelmingly positive, with users expressing the feeling that the new STEDT database environment breathes new life into the linguistic data. Detailed description of these Primary STEDT Database files is resumed below.

ANCILLARY STEDT Database Files				
#	Database Name	Abbrev.	Total Records	Syllable Canon
07	*The Electronic Dictionary of Lahu*	JAM-EDL	38,907	yes
08	*Grammata Serica Recensa Electronica*	BK-GSRE	8,442	yes
09	*Written Burmese Rhyming Dictionary*	PKB-WBRD	4,096	yes
10	*Classical Tibetan Lexical Data*	JV-WT	58,757	yes

The four items tabulated above under "ANCILLARY STEDT Database Files" are complete independent lexica developed for the purpose of extracting data for the MAIN LEXICON. These four relational database lexicons represent STEDT's commitment to the production of electronic archives of its reference materials. Detailed descriptions of each of these four ancillary lexica follow immediately.

ANCILLARY STEDT Database Files: Descriptions

- **JAM-EDL** is the STEDT Project's new electronic version of *The Dictionary of Lahu* (DL), James A. Matisoff's pioneering work (1988; in its 1999 incarnation called *The Electronic Dictionary of Lahu*). The original print publication computer files (now preserved on CD-ROM) were reworked, and converted to standard STEDT database format by Richard S. Cook. The new JAM-EDL includes the English reversal work of Dr. Zev Handel (University of Washington), as well as the addenda and corrigenda to the original printed work. It includes also references for each entry to the original computer files, as well as precise page references for each entry to the print version of DL. This fully searchable version of DL is now being tagged according to semantic class, and the extensive etymologies of this work have been isolated for incorporation into STEDT's Etyma Database. In addition, a Lahu Syllable Canon has been built in relational database format, allowing the user to fully explore *The Dictionary of Lahu* via its analyzed syllabic structures. <http://stedt.berkeley.edu/pdf/JAM/DLproto.pdf>

- **GSRE** is *Grammata Serica Recensa Electronica*, a complete electronic version of Bernhard Karlgren's 1957 classic *Grammata Serica Recensa* (GSR). This electronic version of the text is searchable by multiple field criteria, including Old Chinese, Middle Chinese, Karlgren's Mandarin, Pinyin, Big5 Chinese characters, English Gloss, and GSR set number. The index is linked to high-resolution color images of the complete GSR text. Several of the indices were produced by researchers in association with Tor Ulving (of the University of Göteborg, Sweden; donated to STEDT and used by permission), while the proofing and conversion of this raw data, as well as the scanning and indexing of the image files was done by Richard S. Cook. In addition, an Old Chinese Syllable Canon has been built in relational database format, allowing the user to fully explore GSR via Karlgren's reconstructed OC syllabic structures.

- **PKB-WBRD** is a complete revision of the STEDT version of the late Paul K. Benedict's *Rhyming Dictionary of Written Burmese*, searchable on multiple fields, with precise page references to both the original typescript and the LTBA (1976) revision. In addition, a Written Burmese Syllable Canon has been built in relational database format, allowing the user to fully explore PKB-WBRD via the analyzed syllabic structures.

- **JV-WT** is STEDT's version of James Valby's electronic dictionary of Written Tibetan (WT). The raw data files (donated to STEDT and used by permission) contain nearly 59,000 records, painstakingly compiled on the bases of several printed dictionaries of Classical Tibetan,

most notably, Das (1902) and Jäschke (1881). Reworked into STEDT database format, this lexical material has been augmented with a Written Tibetan Syllable Canon in relational database format, which allows the user to fully explore WT via the analyzed syllabic structures. WT lexical entries are displayed in both Tibetan script and in Romanization. In addition, another 60,000 record WT lexical database is linked to these via the syllable canon. Copyright issues remain to be resolved with regard to this latter half of the data, and so this data is inaccessible in the public release.

In addition to the four Ancillary Databases just described, a number of other such files are either in preparation, or not available in the current release due to copyright limitations. Information on some of the databases currently in production is given in the section below entitled "On-going STEDT Project Projects". Information on some of the data not available due to copyright limitations is available within the STEDT start-up document.

PRIMARY STEDT Database Files: Structural Overview

Databases, Fields, Records, Layouts, Relations.

1.) MAIN LEXICON Database

Among the 6 PRIMARY STEDT Database Files listed above, the MAIN LEXICON (henceforth simply "LEXICON") is the heart of the data. This is the main database table in which the lexical data extracted from STEDT Source Bibliography items is stored. An understanding of the basic structure of this file will allow users to access and manipulate the data they are seeking, not only in this file itself, but in the related files as well.

The LEXICON Database file is composed of primary and secondary local[6] fields. The secondary fields are involved with certain data-tracking, verification, analytic and publishing functions which need not be immediately introduced to the end-user. The following primary local fields are however of more fundamental importance:

MAIN LEXICON Database: Local Fields

NAME	TYPE	INDEX	FONT	BRIEF DESCRIPTION
Reflex	Text	ASCII	STEDT	lexical entries (i.e. reflexes when in cognate sets)
Gloss	Text	English	STEDT	English definitions for lexical entries
Gfn	Text	English	Roman	grammatical function code
Srcabbr	Text	English	Roman	source abbreviation (in Source Bibliography)
Srcid	Text	English	Roman	source identifier (location in source document)
Lgabbr	Text	English	Roman	language abbreviation (in source)
Analysis	Text	English	Roman	etymological tags, comma separated
Rn	Numeric	Serial #	Roman	unique record number
Status	Text	English	Roman	error and verification codes

[6]Fields are termed "local" (as opposed to "related") because the data in these fields is stored within this particular database file, rather than in one of the related files (what is meant by "related files" should become apparent in the discussion below).

• The Reflex Field

The "Reflex" field contains lexical entries for the various languages of the various sources. When the syllables of a given lexical entry have been analyzed, it is at this point that the lexical entry contains reflexes of the protoforms. The transcription employed in this field is essentially that of the source document. "Stick to copy!" is the relevant STEDT Project dictum. In effect, this means that the STEDT computerization of the source transcription adheres as closely as possible to that of the original source transcription, within the limits of the STEDT Font.[7] Interpretation of the computerized source transcriptions is undertaken with reference to both the original source documents, and also with reference to the STEDT Project *Phonological Inventories*.[8] In some cases STEDT has already preserved electronic images of the source documents[9], which may be accessed via the Ancillary Databases. In other cases, the scanning of the original documents remains to be completed. The STEDT Project *Phonological Inventories* will appear in future data releases, related to the Language Names Database (described in detail below).

The ASCII indexing of the Reflex field means that users performing searches directly on this field should enclose their search criteria in double-quotes. In addition, wild-card searches on this field may execute rather slowly, depending on the complexity of the query. Several of these slowness issues for wild-card searches will be eliminated in the near future through rearrangement of the database structures (physical reordering of fields). As the data in the Phonological Inventories is refined and syllable canons for the various languages are built, more complex searches will be possible, and such searches will execute exceedingly quickly. Proto-syllable-canons will also be possible, based on data in the soundlaw databases, at which point even higher level searches (accessing also the SEMCAT semantic classificational data) will be possible.

• The Srcabbr Field

Also among the above tabulated fields, the "Srcabbr" field contains values conventionalized in the *STEDT Source Bibliography* in the bipartite form ABC-XYZ, where "ABC" is usually the initials of the author's name, and "XYZ" is a string representing a particular work by that author. Thus, "GHL" stands for "Luce, G. H. (Gordon Hannington)", and "PPB" indicates his work *Phases of Pre-Pagán Burma languages and history, Vol. 2*. The Srcabbr values are unique for a given source, and so the string "GHL-PPB" uniquely identifies the following item in the *STEDT Source Bibliography*:

Luce 85 **LUCE, G. H. (Gordon Hannington).**
1985. *Phases of Pre-Pagán Burma languages and history,Vol. 2*. School of Oriental and African Studies. Oxford: Oxford University Press.

• Calculated Key Fields

In addition to these locally stored text and numeric data fields, there are also locally stored calculation fields which are employed as "key" (i.e. 'matching') fields for the connections to the related databases. Two of these are of chief importance:

---

[7]Notable among those contributing ideas over the years to STEDT's methodological treatment of transcriptional issues is Martine Mazaudon (1987, PC).

[8]A publication in the STEDT Monograph Series.

[9]The STEDT Ancillary Database "GSRE" described above is an example of this.

MAIN LEXICON Database: Calculated Key Fields				
NAME	TYPE	RESULT	INDEX	CALCULATION
BINOM	Calculation	Text	Indexed	=Srcabbr &"|"&Lgabbr
TAG MATCH	Calculation	Text	Indexed	=Substitute(Analysis,",","¶")

• The BINOM[10] calculated key field is a concatenation of the Srcabbr and Lgabbr values for a given record, separated by a "|" (pipe) which serves to assure completely unique concatenations. Thus, Srcabbr values are unique for each source, while Lgabbr values may vary from source to source. By concatenating these two values, the name of a given language appearing in the source document is unambiguously identified and linked to the STEDT name and sub-grouping classification for this language. This is done via the relation to the Language Names Database (described below).

• The TAG MATCH calculated key field is a calculated substitution performed on the locally stored Analysis field values. As the result of this calculation, the comma separated values in the records of the Analysis field are converted to hard-return separated values. Thus, if a record in the Analysis field contains the comma separated array "3,5,8" (indicating syllables analyzed as corresponding to etymon 3, etymon 5, and etymon 8, respectively, then the calculated result on this string is

3¶
5¶
8¶

in which the array elements now appear to be stacked vertically, as they are "¶" hard-return separated. This is one way in which many-to-one (and also many-to-many) relations may be set in the current database environment, permitting such relations to be set without the use of intervening "match" databases.[11] A relation to the Etyma database file is then set, matching this calculated field to etymon numbers stored in the Etyma database. Effectively, this means that when the syllables of a given Reflex have been etymologized in the Analysis field, all of the Etyma for a given Reflex may be called up immediately in the related database. Likewise, all supporting forms for a given Etymon may be called up from the relation set within the Etyma database. This is one way in which cognate sets may be summoned.

02.) ETYMA DATABASE

The Etyma Database stores the reconstructed etyma (protoforms, i.e. etymological "roots"), their reconstructed meanings (protoglosses), and the "TAG" numbers uniquely identifying each root. These TAG numbers are placed in the Analysis field in the LEXICON when a given lexical entry is etymologized, and thus a syllabic reflex of the etymon is identified within the data. As described above under the "TAG MATCH" field, this data is linked to the LEXICON via a matching relation on that "TAG MATCH" field. Each etymon has associated notes, relating that etymon to its source, as well as noting connections with other etyma. A click of the yellow circular button next to a record in the Etyma Database calls up a given STEDT Cognate Set in the

[10]STEDT's binome concept derives ultimately from Boyd Michalowsky (1987 PC).

[11]We are grateful to Dr. Sheila Greibach of the Computer Science Department of the University of California, Los Angeles (UCLA) for her kind explanations and helpful suggestions in the preliminary stages of development of this portion of the FileMaker Pro implementation of the STEDT Databases.

LEXICON. Each etymon record also contains various statistics, including a count of supporting forms.

02.1) ANALYSIS DATABASE

In addition to the relation to the LEXICON, within the Etyma Database a relation also exists to the Analysis Database. This relation, accessed via the circular white button of a given Etymon record, brings up a STEDT Cognate Set in the Analysis Database. The Analysis Database is employed to segment the syllables of the Reflex field according to the etymological analyses (tags). The algorithms developed by Richard Cook for this segmentation are described in Appendix 1. Segmentation of the Reflex field syllables permits the researchers to evaluate the consistency of etymological tagging in a graphical environment. This Analysis Database is also used to prepare a particular STEDT Cognate Set for publication. Due to the complexity and specialized utility of this database tool, access to this portion of the Database environment is at present limited to members of the STEDT Project staff.

03.) LANGUAGE NAMES DATABASE

The Language Names Database was mentioned above in connection with the concepts behind the BINOM field of the LEXICON. The Language Names Database holds data relating to the name of a language in a particular source, and STEDT's determination of how this name may relate to the standard STEDT Language field name. As an example, consider the following data, representing the results of a search for "Tangut" in the Language field.

SAMPLE DATA FROM THE LANGUAGE NAMES DATABASE

Count	Srcabbr	Citation Form	**Language**	Lgabbr	Groupabbr
1004	ZMYYC	Sun H 91 ZMYY	Tangut [Xixia]	Tangut13	X
339	NT-SGK	Nishida 64 Tan	Tangut [Xixia]	Tangut1	X
148	MVS-Grin	Sofronov 97	Tangut [Xixia]	Tangut	X
136	DQ-Xixia	Dai 89 Xixi	Tangut [Xixia]	XIXIA	X
70	NT-SGK	Nishida 64 Tan	Tangut [Xixia]	Tangut2	X
5	JAM-MLBM	Matisoff 78 MLBM	Tangut [Xixia]	Hs	X
1	STC	Benedict 72 STC	Tangut [Xixia]	Hsi-hsia	X

The results of a search on the Language field for "Tangut".

It can be seen that there are four different Lgabbr values for the same language in the four different sources, while the standard STEDT Language field name for this language is "Tangut [Xixia]". In this case, if one searches in the Language field for "Xixia", one will also achieve the same results.

As mentioned above in connection with the Reflex field of the PRIMARY LEXICON, the data in the Language Names Database will be supplemented relationally with data from the STEDT Phonological Inventories. In addition, Information contained in *Languages and Dialects of Tibeto-Burman* will also be available relationally within this database.[12] Each Language Name record also contains various statistics, including a count of forms for this BINOM in the LEXICON.

[12] A publication in the STEDT Monograph Series.

The rightmost column of the above table indicates that language sub-grouping information is also stored in the Language Names Database. Clicking the blue button in a given record will bring up the Language Groups Database, highlighting the position of the current record's Group within the classificational scheme.

Clicking the yellow button in a given Language Names database record will bring up all of the forms having this particular BINOM in the LEXICON.

04.) SOURCE BIBLIOGRAPHY

The Source Bibliography (Srcbib) Database file contains records for each of the sources consulted in creation of Reflex records in the LEXICON. As mentioned above with regard to the Srcabbr Field of the LEXICON, this information is accessed via a related database's Srcabbr field data. Additionally, the Srcbib also contains records for sources among the Ancillary Databases, and records for sources in various stages of the data inputting/proofing process. Current statistics are available for the current electronic records deriving from a particular source.

05.) LANGUAGE GROUPS

The Language Groups Database is the smallest of the STEDT Databases, and yet that with the most far-reaching implications. With the classification scheme set forth in this database, present understanding of the genetic relations among the various daughter languages is presented. All of the data for a given subgroup may be isolated from within this database, with a click of a relational button. The classification scheme elaborated in this file largely reflects the state of modern consensus on these difficult questions, and does not necessarily reflect the opinion of STEDT researchers. For example, the position of the Tangut language within a "Tangut-Qiang" group seems highly questionable to some who have studied the problem.[13] Likewise, although the Karenic languages are off by themselves within this scheme, their exclusion from the general Tibeto-Burman family has been questioned by recent fieldwork.[14]

06.) STEDT FONT REFERENCE

The STEDT Font Reference is primarily a resource for those engaged in data transcription using STEDT Font (now in the kerned version 5.1.2), the Apple Macintosh TrueType® phonetic transcription font.[15] These may be researchers working to input either published texts, or questionnaires of elicited field data. Inputting work is done both by STEDT Project staff and also by collaborators at other institutions. The STEDT Font Reference contains numerous fields allowing access to transcription characters by various criteria, including such information as "decimal number", "symbol name", "symbol type", "symbol shape", "place of articulation", and "PSG[16] Page Number". Fields giving statistics for character frequencies in the LEXICON Reflex and Gloss fields are found here: relations from this database allow isolation of character occurrences in records within the LEXICON. Also within the STEDT Font Reference is management of data relating to typographic development. STEDTwin, a Windows compatible version of the current STEDT Font is in the final stages of development, and should be available in November 2000. STEDTX, a cross-platform Sino-Tibetan phonetic transcription typographic solution, the next generation of STEDT font, is also being prepared from within this database. Further font related information is available below in the section entitled "STEDT Typographic Systems".

---

[13]LIN Yingchin (2000, PC).

[14]MATISOFF (2000, FM [JAM-FM 00]).

[15]http://stedt.berkeley.edu/stedtfont/

[16]The Phonetic Symbol Guide, PULLUM and LADUSAW (1996).

STEDT Data Resources: Searching and Sorting

FINDING WHAT YOU WANT LOCALLY, AND IN THE RELATED FILES.

This brief introduction to Searching the STEDT Databases will be confined to searching within the LEXICON Database File. The search techniques discussed here may however be used with little or no modification within any of the related PRIMARY or Ancillary databases. This tutorial applies to both Macintosh and Windows platforms, the only difference being that when the key stroke uses the "command" (⌘) key on the Macintosh, this is the Windows "control" key. In the descriptions following, if only the Mac key combination is given, it should be understood that the windows key combination is the same, but using the "control" key.

• STRUCTURE OF THE DATABASES: NAVIGATION

Before performing a search, the user would do well to become familiar with the various databases, and the kinds of data found in the various fields of these databases.The descriptions given above provide basic information in this regard. The user is encouraged to browse records in the LEXICON and the other related files in order to learn what data constitutes valid Find Criteria for a given field. Moving between databases is accomplished using the rectangular navigational buttons. Browsing is accomplished by using vertical scroll bars. Remember also that the small circular buttons within particular records are "relational", which is to say, these buttons are used to move between related databases, displaying the related data within the related database. All buttons (circular and rectangular) are color-coded, as described above.

Before initiating a search, the user should first make sure that the foremost database window is that of the particular database which the user wants to search. Navigate to a particular database from within the STEDT start-up palette or from within the current database, using the rectangular navigation buttons.

• PERFORMING A SEARCH

When the desired database window is foremost, searches in the STEDT Databases may be initiated by entering Find Mode. Find Mode is entered in one of two ways: by 1.) clicking the "Find" button; by 2.) typing command-f (⌘-f) on Macintosh platforms (control-f on Windows platforms). To exit find mode without performing a search, simply type command-b.

When the user has entered Find Mode, a single empty Find Request will be visible. The user may now type or paste the appropriate data in one or more of the fields of the Find Request.

When all Find Criteria have been entered into the appropriate fields, the user then hits the "Return" or "Enter" key on the keyboard, at which point, the search begins to execute. Depending upon the complexity of the Find Request, the Find Results will appear quickly or slowly.

• COMPLEX SEARCHES.

Searches may involve not only Find Criteria in multiple fields, but also multiple Find Requests. Choosing "New" when in Find Mode presents the user with a new empty Find Request, the criteria for which may be filled in as above. Additional Find Requests may be employed also to omit certain records from the Found Set, by clicking the "Omit" check box.

As described above under the description of the "Reflex Field", the ASCII indexing of this particular field puts additional restrictions on searches using this field. Most importantly, all Find

Criteria for the Reflex field must be entered between straight double-quotes.[17]

Wild-Card searches are performed on any field by using the "*" asterisk symbol (= 'zero or more occurrences of any character'), or the "@" at-sign (= 'one or more occurrences of any character'). Wild-Card searches on the Reflex field may execute slowly, also due to the relative complexity of its (ASCII) indexing.

Since the LEXICON Database is considerably larger than the other PRIMARY Databases, users are encouraged to perform searches within the related files whenever possible. Corresponding records in the LEXICON may then be called up using the small circular "relational" buttons.

Find Requests from the previously executed Find may be restored by typing command-r.

To browse all records after a find has been performed, type command-j "Show All". The currently selected record in the database will then be shown in its context within the LEXICON.

• SORTING

Records in most databases may be sorted on a given field by clicking on the name of the field. It is highly recommended that only portions of the LEXICON data be sorted. Sorting the entire LEXICON on the Reflex field, for example, may take a very long time. Rather, isolate the desired records, and then sort the data. At present, sorts on multiple fields are only available by typing command-s, and specifying the sort criteria.

---

[17] Do not use "smart-quotes", i.e. left and right double curly quotes, because these symbols have special usage in STEDT Font. Also, using these 2 special characters in Find Requests employing the Reflex field may result in invalid Find Requests. In future versions of STEDT Font this limitation will be removed.

STEDT Typographic and Encoding Systems

TRANSCRIPTION AND ORTHOGRAPHY.

In addition to the information given above with regard to transcription in the description of the STEDT Font Reference database, the following typographic information may also be given. STEDT Data encodings at present are moving towards Unicode compliance, although a number of difficult issues remain to be resolved before this can be possible (see below). Work is progressing toward the development of a variable width CID font which will encompass not only the rich phonetic transcription symbols of the STEDT Source documents, but also orthographic scripts such as Chinese, Tibetan, and Burmese. The CID (double-byte) font format, which is currently only supported on Macintosh computers in mono-width font faces, allows character sets of approximately 64,000 characters.

At present, STEDT Chinese data is encoded in both Big5 and GB encoding formats, specifically, in those versions of these encodings currently supported in Mac OS 9. In addition, STEDT employs the Tibetan Language Kit, produced by Otani University in association with Peter Lofting of Apple Computer. The Tibetan data available in and via the JV-WT (Ancillary) Database exhibits field data in this format. STEDT is also involved in the production of TrueType Burmese and Karen fonts for use in Ancillary Lexica currently in development.

PDF DELIVERY OF STEDT DOCUMENTS

- STEDT is now delivering a number of documents relating to STEDT research findings in electronic PDF format on the world-wide-web. Through the sponsorship of Adobe Systems, Inc. <http://www.adobe.com> and specifically, through the kind generosity of Dr. Kenneth Lunde of Adobe's CJKV Type Development Division, STEDT has been provided with technical support and a license for use of Adobe Acrobat 4 to deliver project documents in platform independent format. Additionally, STEDT has been granted a license for use of Adobe's "ATM Deluxe" software, typographic software employed to manage the various typographic systems in development. Dr. Lunde was also kind enough to provide STEDT with an electronic copy of his book CJKV Information Processing, a text proving instrumental in addressing the complex typographical problems surrounding STEDT data sources.

A FORMAL PROPOSAL TO THE UNICODE CONSORTIUM

- A further collaborative effort involves the preparation of a formal proposal to the Unicode Consortium <http://www.unicode.org>. This work is being undertaken in association with Martin Heijdra <mheijdra@princeton.edu> of the Gest Oriental Library at Princeton . This proposal seeks to bring certain linguistic transcription symbols (those for the alveolo-palatal place-series) currently omitted from the Unicode Standard to the attention of the standards body of the Unicode Consortium. Inclusion of these symbols in future versions of the Unicode Standard will facilitate future computerization of STEDT source transcriptions in a universally accessible format. A recent draft of a paper on this subject is available on-line at:

<http://stedt.berkeley.edu/work/curly-tailed-tdnlcz.pdf>

ON-GOING STEDT PROJECT PROJECTS

Lai (Hakha Chin) / English Dictionary Project

• STEDT Project member and current UCB doctoral candidate Kenneth VanBik obtained independent funding (from two sources: The Open Society Institute (OSI), 400 W 59th St. NYC, NY 10019; Endangered Language Fund, Yale University) for the creation of the *Lai (Hakha Chin) / English Dictionary*. Mr. VanBik, who is a native speaker of the Lai (Hakha Chin) language, working with materials developed in association with his father David VanBik (under the supervision of Prof. Matisoff and with the computer assistance of Richard Cook) is creating the first comprehensive dictionary database of materials relating to the Lai language. A relational syllable canon similar to those described for other STEDT Ancillary Databases will be produced for this database.

The Tangut Reconstruction Project

• In the past spring, STEDT was fortunate enough to be visited by Prof. LIN Ying-chin 林英津 of Academia Sinica (Taipei, Taiwan). Prof. LIN, an expert in the 西夏(唐古特)語 (Xi Xia, a.k.a. Tangut) language worked during this time to computerize reconstructions of the Tangut language developed by her in association with Prof. GONG Hwang-cherng 龔煌城 (also of Academia Sinica). These reconstructions serve to fill a gap in the STEDT data for the source ZMYYC (《藏緬語語音和詞匯》傅懋勣主編 SUN Hongkai et al., *Zangmianyu yuyin he cihui* [Tibeto-Burman phonology and lexicon], 1991), which previously had only place-holders for the Tangut forms in this large collection of data. In addition, Prof. LIN was gracious enough to provide STEDT with the Traditional and Simplified Chinese glosses for the ZMYYC lexical items. This means that the STEDT version of ZMYYC, with its English glosses by Jackson SUN, is the most complete version of this text available. This data will be absolutely invaluable for future reconstruction work, and will be a welcome boon to the research community.

The WT-WT Electronic Dictionary Project

• In collaboration with Dr. Robert Taylor <r.taylor@asianclassics.org>, Assistant Director of *The Asian Classic Input Project*, STEDT is in the process of preparing a new electronic edition of the text *Bod-Rgya Tshig Mdzod Chen Mo* (《藏漢大辭典》= *Zang-Han Da Ci Dian*) [Tibetan-Chinese Great Dictionary] (張怡蓀 Zhang Yisun, ed. Beijing: Nationalities Press, 1993. ISBN: 7-105-02036-9/Z.136. All of the Tibetan text of this 3,146 page Tibetan-Tibetan dictionary of the Written Tibetan (WT, i.e. "Classical Tibetan") language (also with Chinese glosses in the original print edition) has been input, omissions and errors in the electronic text having been corrected by STEDT. In coming months this data will be converted to standard STEDT database format, and will provide ready access to a wealth of hitherto unknown lexical information.

Field Methods 2000-2001

• The year-long UC Berkeley Linguistics Department "Field Methods" Seminar, in which graduate students study a particular little-known language, learning it from an informant from the ground up, is now being planned to continue the tradition of previous seminars by focusing upon a Tibeto-Burman language crucial to STEDT Project work. As described in the previous Interim Report, Karenic specialist Dr. David Solnit is currently engaged at STEDT to work upon Karenic languages. As good fortune would have it, the Karen community in the San Francisco Bay Area is thriving, such that it seems they will be able to provide the "Field Methods" Seminar with a Sgaw Karen speaker. This work will greatly advance knowledge of the Karenic branch of the family, and fruits of this work will certainly occupy a significant place in future Interim Reports.

Conclusion

It is hoped that the picture painted here of the STEDT Project Data Resources has served as a gentle introduction to many of the complex issues surrounding this data and its use. Further documentation is availablefrom within the Help system of the databases. Readers with questions and/or comments are encouraged to send email to <stedt@socrates.berkeley.edu>. In closing, on behalf of the STEDT Project staff, we would like to invite readers with access to data which they believe might be of interest to historical Sino-Tibetan reconstruction to contact STEDT. We invite data submissions of all kinds, and would very much like to have you among our contributors. STEDT Questionairres are available on-line at <http://stedt.berkeley.edu/>.

Acknowledgements

This research was supported in part by grants from:

• The National Science Foundation (NSF), Division of Behavioral & Cognitive Sciences, Linguistics, Grant Nos. BNS-86-17726, BNS-90-11918, DBS-92 09481, FD-95-11034, SBR-9808952 and BCS-9904950;

• The National Endowment for the Humanities (NEH), Preservation and Access, Grant Nos. RT-20789-87, RT-21203-90, RT-21420-92, PA-22843 96 and PA-23353-99.

Appendix 1

Computational Algorithms for STEDT Etymological Analyses

Richard S. COOK
rscook@socrates.berkeley.edu
UC Berkeley Linguistics Department
July 2000, revised 2000/08/22/21:03 PDT
http://stedt.berkeley.edu/

The following is a discussion of algorithms developed to separate the Reflex field (in these cases the related field LEXICON::Reflex) into 3 fields (preSYLinQ, SYLinQ, postSYLinQ) based upon the etymological analyses in the related field LEXICON::Analysis. The formulae presented below reflect the FileMakerPro4-5 (FMP) calculated field implementation. The field SYLinQ is the primary field on which calculations of the other 2 fields preSYLinQ and postSYLinQ are based. SYLinQ ('syllable in question') in turn is based on 2 other fields, ReflexSpaced and nthTAGinQ. The field ReflexSpaced is calculated based on the related field LEXICON::Reflex, and employs conditionals and nested substitution formulae to assure that the syllables in the ReflexSpaced field are separated with a single space character (these formulae are given at the end of this article). The field nthTAGinQ is calculated based upon 2 fields comTAGcom and comTAGinQcom.

The formula for the field comTAGcom is simply

comTAGcom=
```
"," & Analysis & ","
```

which is to say that commas are pre- and post-concatenated to the Analysis field values ("**&**" is the string concatenation operator).

The field comTAGinQcom is similarly calculated

comTAGinQcom=
```
"," & gTAGinQ & ","
```

based on the global field gTAGinQ which is equal to the number of the Etymon in question (i.e. the current cognate set).

Thus, the formula used to compute nthTAGinQ is as follows:

nthTAGinQ=
```
Case(
Position(comTAGcom, ",", 1, 1) = Position(comTAGcom, comTAGinQcom, 1, 1), 1,
Position(comTAGcom, ",", 1, 2) = Position(comTAGcom, comTAGinQcom, 1, 1), 2,
Position(comTAGcom, ",", 1, 3) = Position(comTAGcom, comTAGinQcom, 1, 1), 3,
Position(comTAGcom, ",", 1, 4) = Position(comTAGcom, comTAGinQcom, 1, 1), 4,
Position(comTAGcom, ",", 1, 5) = Position(comTAGcom, comTAGinQcom, 1, 1), 5,
Position(comTAGcom, ",", 1, 6) = Position(comTAGcom, comTAGinQcom, 1, 1), 6,
Position(comTAGcom, ",", 1, 7) = Position(comTAGcom, comTAGinQcom, 1, 1), 7,
Position(comTAGcom, ",", 1, 8) = Position(comTAGcom, comTAGinQcom, 1, 1), 8,
Position(comTAGcom, ",", 1, 9) = Position(comTAGcom, comTAGinQcom, 1, 1), 9
)
```

Starting with the 1st character of the string in a record of the field comTAGcom, the above formula calculates the position of the nth instance of the string "," (a comma), and tests whether this position is equal to the position in comTAGcom of the nth instance of comTAGinQcom. If the nth test is true, the nth result (1-9) is returned. This formula at present assumes at most 9 tags per reflex, but can easily be extended to accomodate more tags, should the need arise.

In addition to these fields, the following is also employed:

```
totTAGs=
PatternCount(Analysis,",")+1
```

The totTAGs calculated field value represents the 'total number of tags' in the analysis of a given reflex, and relies upon the rule that "all tags in records of the Analysis field must be uniformly separated from one another by only a single comma".

Once the above field values have been calculated, the values of the three fields preSYLinQ, SYLinQ, and postSYLinQ may be calculated. The basic formulae for these three fields (the 1st and 3rd being dependant on the 2nd) are as follows:

```
preSYLinQ=
Case(
totTAGs=1,"",
Middle(LEXICON::Reflex,1,Position(LEXICON::Reflex,SYLinQ,1,1)-1)
)
SYLinQ=
Case(
totTAGs=1,LEXICON::Reflex,
Middle(
ReflexSpaced,
(Position(ReflexSpaced," ", 2, nthTAGinQ-1)+1),
(Position(ReflexSpaced," ", 2, nthTAGinQ)-
(Position(ReflexSpaced," ", 2, nthTAGinQ-1)+1))
)
)
postSYLinQ=
Case(
totTAGs=1,"",
Middle(LEXICON::Reflex,(Position(LEXICON::Reflex,SYLinQ,1,1)+
Length(SYLinQ)),Length(LEXICON::Reflex))
)
```

In order to account for cases in which the Reflex may contain 2 or more identical syllables, either identically or differently tagged, i.e. the case

```
PatternCount(LEXICON::Reflex,SYLinQ) > 1
```

special care must be taken, since, depending on the order of the tags, incorrect preSYLinQ and postSYLinQ might result. Thus, preSYLinQ and postSYLinQ formulae are rewritten as follows:

```
preSYLinQ=
Case(
totTAGs=1,"",
PatternCount(LEXICON::Reflex,SYLinQ)>1,
Middle(LEXICON::Reflex,1,Position(LEXICON::Reflex,SYLinQ,1,nthTAGinQ)-1),
Middle(LEXICON::Reflex,1,Position(LEXICON::Reflex,SYLinQ,1,1)-1)
)
postSYLinQ=
Case(
totTAGs=1,"",
PatternCount(LEXICON::Reflex,SYLinQ)>1,
Middle(LEXICON::Reflex,(Position(LEXICON::Reflex,SYLinQ,1,nthTAGinQ)+
Length(SYLinQ)),Length(LEXICON::Reflex)),
Middle(LEXICON::Reflex,(Position(LEXICON::Reflex,SYLinQ,1,1)+
Length(SYLinQ)),Length(LEXICON::Reflex))
)
```

Note that in these cases the value of nthTAGinQ is relevant, and corresponds to the nth occurrence of SYLinQ with which the **Position()** function is concerned.

As mentioned above, the field ReflexSpaced is calculated based on the related field LEXICON::Reflex. This calculation employs conditionals and nested substitution formulae to assure that the syllables in that Reflex field are separated with a single space character. The conditionals test for specific source transcriptions requiring treatment by specific substitution formulae. In particular, those sources employing preposed tone diacritics require separate substitution formulae. The **Case()** functions below could also be written in FMP with the "**=**" and "≠" "Comparison operators", which also operate on text strings, though with "English" (rather than "ASCII") indexing rules. Given the Binom field conventions, this is not a problem, though to avoid the possibility of future problems, the functions **Exact()** and **not(Exact())** are used instead. These formulae are given here:

```
Case(
(
(Exact(LEXICON::Srcabbr,"AW-TBT")) and
(not(Exact(BINOM,"AW-TBT|W.Bw"))) and
(not(Exact(BINOM,"AW-TBT|PL"))) and
(not(Exact(BINOM,"AW-TBT|LB"))) and
(not(Exact(BINOM,"AW-TBT|PLB")))
)
or
(
(Exact(Left(LEXICON::Srcabbr,3),"MM-"))
or
(Exact(LEXICON::Srcabbr,"HM-Prak"))
),
(
Substitute({43}
(LEXICON::Reflex & " "),
"¡", " ¡"),
"™", " ™"),
"£", " £"),
"¢", " ¢"),
"∞", " ∞"),
"§", " §"),
gDec166, " " & gDec166),
".", " ."),
"¡ ¡", "¡¡"),
"¡ ™", "¡™"),
"¡ £", "¡£"),
"¡ ∞", "¡∞"),
"™ ¡", "™¡"),
"™ ™", "™™"),
"™ £", "™£"),
"™ ¢", "™¢"),
"™ §", "™§"),
"£ ¡", "£¡"),
"£ ™", "£™"),
"£ £", "££"),
"£ ¢", "£¢"),
"£ ∞", "£∞"),
"¢ ¡", "¢¡"),
"¢ ™", "¢™"),
"¢ ¢", "¢¢"),
"¢ ∞", "¢∞"),
"∞ ¡", "∞¡"),
"∞ ™", "∞™"),
"∞ £", "∞£"),
"∞ ¢", "∞¢"),
"∞ ∞", "∞∞"),
"≠", " "),
"-)", ") "),
"–", " "),
"±", " "),
"=", " "),
"/", " "),
",", " "),
"·", " "),
"( ¡", "(¡"),
"[ ™", "[™"),
"   ", " "),
"  ", " ")
),
(
Substitute({41}
(
LEXICON::Reflex & " "),
"¡", "¡ "),
"™", "™ "),
"£", "£ "),
"¢", "¢ "),
"∞", "∞ "),
"§", "§ "),
gDec166, gDec166 & " "),
".", ". "),
"¡ ¡", "¡¡"),
"¡ ™", "¡™"),
"¡ £", "¡£"),
"¡ ∞", "¡∞"),
"™ ¡", "™¡"),
"™ ™", "™™"),
"™ £", "™£"),
"™ ¢", "™¢"),
"™ §", "™§"),
"£ ¡", "£¡"),
"£ ™", "£™"),
```

```
"£ £", "££"),
"£ ¢", "£¢"),
"£ ∞", "£∞"),
"¢ î", "¢î"),
"¢ ™", "¢™"),
"¢ ¢", "¢¢"),
"¢ ∞", "¢∞"),
"∞ î", "∞î"),
"∞ ™", "∞™"),
"∞ £", "∞£"),
"∞ ¢", "∞¢"),
"∞ ∞", "∞∞"),
"≠", " "),
"–)", ") "),
"–", " "),
"±", " "),
"=", " "),
"/", " "),
",", " "),
"·", " "),
"  ", " "),
"  ", " ")
)
)
--END--
```

References (Selected)

BENEDICT, Paul King
1972 *Sino-Tibetan A Conspectus* (STC), Contributing editor James Alan Matisoff. Princeton Cambridge Studies in Chinese Linguistics. Cambridge: Cambridge University Press. ISBN 0 521 08175 0

BEYER, Stephan V.
1992 *The CLASSICAL TIBETAN Language*. New York: State University of New York Press. ISBN: 0 7914-1099-4.

KEIGHTLEY, David N.
1978 *Sources of Shang History: The Oracle-Bone Inscriptions of Bronze Age China*. Berkeley and Los Angeles: University of California Press, 1978. ISBN 0-520-05455-5.

LUCE, G. H. (Gordon Hannington)
1981 *A Comparative Word-List of Old Burmese, Chinese, and Tibetan*. London: School of Oriental and African Studies, University of London.

MATISOFF, James Alan
2000 *The Electronic Dictionary of Lahu*. Berkeley: Sino-Tibetan Etymological Dictionary and Thesaurus (STEDT) Project.
1988 *The Dictionary of Lahu*. (University of California Publications in Linguistics, v. 111.) Berkeley, Los Angeles, London: University of California Press.
1978 *Variational Semantics in Tibeto-Burman: the 'organic' approach to linguistic comparison*. Occasional Papers of the Wolfenden Society on Tibeto-Burman Linguistics, Volume VI. Philadelphia: Publication of the Institute for the Study of Human Issues (ISHI). xviii + 331 pp.

PULLUM, Geoffrey K. and William A. LADUSAW.
1996. Phonetic Symbol Guide, Second Edition. Chicago, London: The University of Chicago Press.

Sino-Tibetan Etymological Dictionary and Thesaurus Monograph Series
General Editor: James A. Matisoff, University of California, Berkeley

STEDT Monograph 1:
1989 *Bibliography of the International Conferences on Sino-Tibetan Languages and Linguistics I-XXI*. Randy J. LaPolla and John B. Lowe with Amy Dolcourt. lix, 292 pages.

STEDT Monograph 1A:
1994 *Bibliography of the International Conferences on Sino-Tibetan Languages and Linguistics I-XXV*. Randy J. LaPolla and John B. Lowe. lxiv, 308 pages.

STEDT Monograph 2:
1996 *Languages and Dialects of Tibeto-Burman*. James A. Matisoff with Stephen P. Baron and John B. Lowe. xxx, 180 pages.

STEDT Monograph 3:
1996 *Phonological Inventories of Tibeto-Burman Languages*. Ju Namkung, editor. xxviii, 507 pages.