The SEAlang Projects, Spring 2008

http://sealang.net
Doug Cooper[1], *CRCL Inc.*[2]

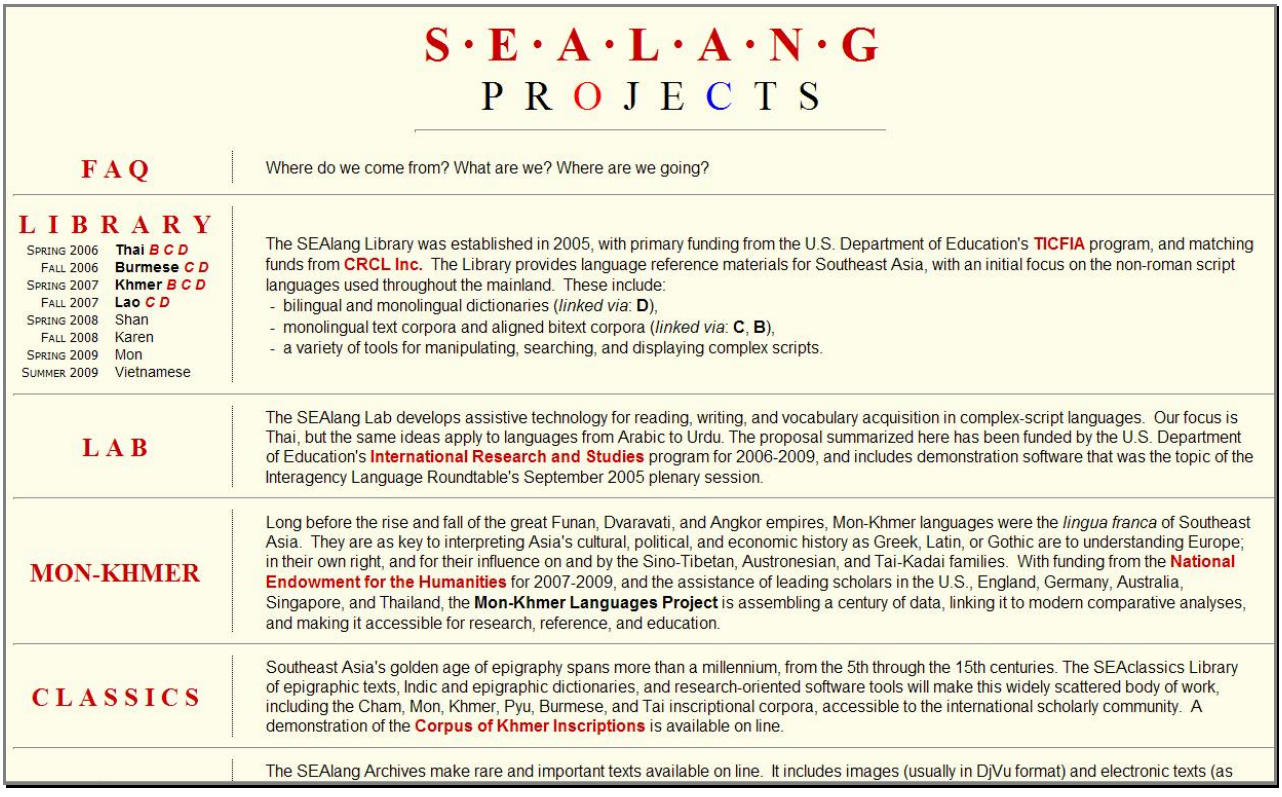
The *SEAlang Projects* are dedicated to gathering and sharing data relevant to Southeast Asian languages: modern or historic, and written, spoken, or digitally stored. They include:

- the **SEAlang Library**, funded by the U.S. Department of Education's *TICFIA* program 2005-2009 (UW-Madison Prof. Robert Bickner, PI).
- the **SEAlang Lab**, funded by the US/ED *International Research & Studies* program 2006-2009 (Doug Cooper, PI),
- the **Mon-Khmer Languages Project**, supported by the National Endowment for the Humanities *Division of Preservation and Access* 2007-2009 (Paul Sidwell, PI), and
- the **Southeast Asian Linguistics Archives**, now in preliminary development.

SEAlang also includes internally funded, collaborative efforts focused on more specialized resources (particularly related to epigraphy and historical manuscripts), and an upcoming series of conferences: in 2009 CRCL will co-host (with the University of Wisconsin-Madison) the annual *TICFIA Projects Conference* in Madison, and (with Mahidol University) the *4th International Conference on Austro-Asiatic Linguistics* in Bangkok.

The SEAlang Library has been the essential catalyst for all of these efforts; providing enabling data and software infrastructure that make new initiatives practical. In return, the projects have helped expand the resource base available to SEA researchers and students, and opened many doors for long-term collaboration. All in all, the success of SEAlang comes down to one simple fact: the projects fulfill real needs.

None of the work described here would have been possible without the collaboration, cooperation, or support of the University of Wisconsin-Madison Center for Southeast Asian Studies, faculty and staff of the Southeast Asian Studies Summer Institute, the Foreign Service Institute, Defense Language Institute and others; the Council of Teachers of Southeast Asian Languages, the Committee on Research Materials on Southeast Asia, and other professional groups, the Mon-Khmer Studies Journal, Pacific Linguistics Publishing, Dunwoody Press, and other publishers, the David Thomas Library, and other organizations and scholars around the world.

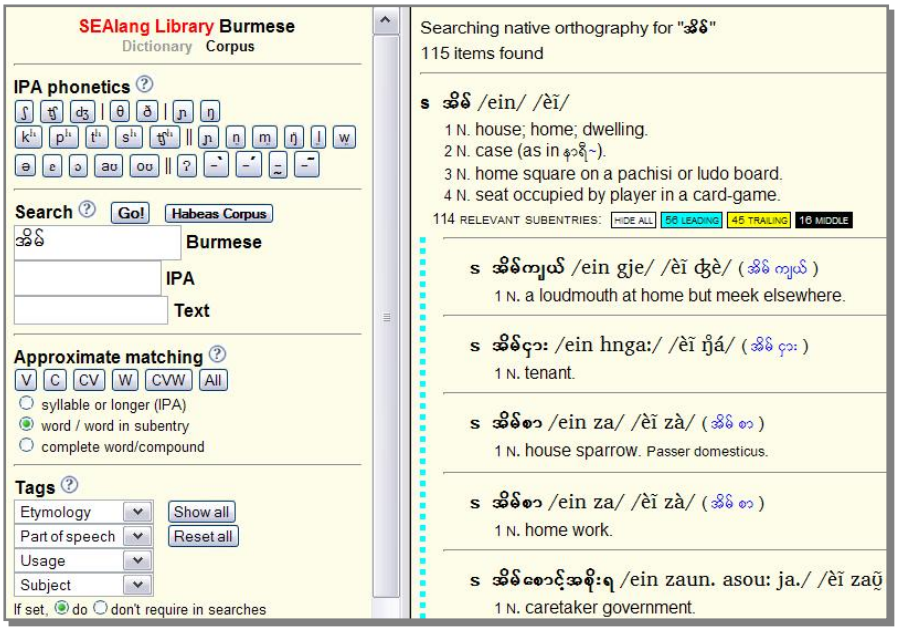


[1] The author is grateful for support from the U.S. Department of Education and the National Endowment for the Humanities. Any opinions expressed here do not necessarily reflect those of these agencies.
[2] CRCL Inc. is the *Center for Research in Computational Linguistics*, a US 501(c)3 nonprofit organization.

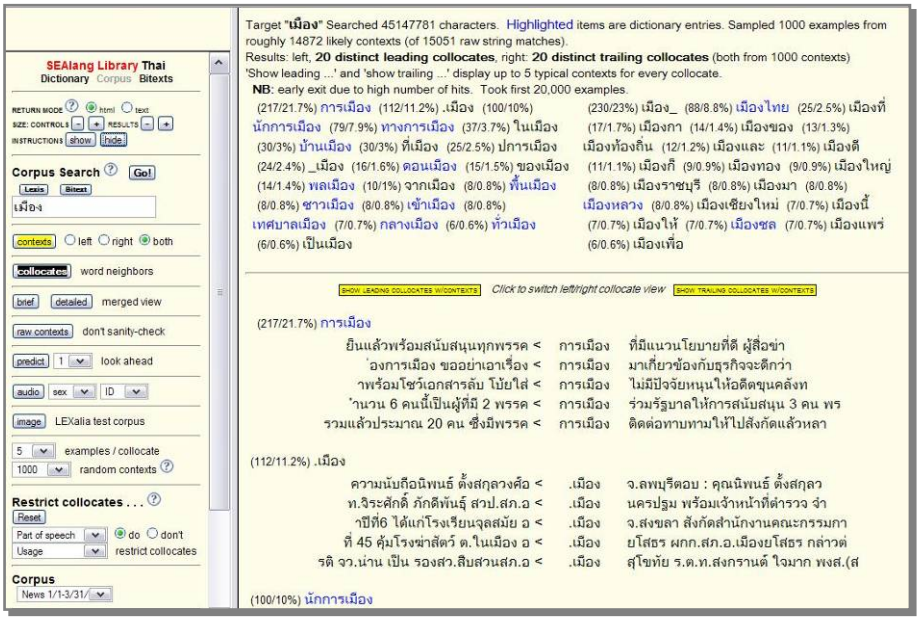
SEAlang Library

About the project The Southeast Asian Languages Library was conceived as the practical solution to a longstanding problem: the lack of a national *Language Resource Center* for SEA. Its basic charter is to provide digital bilingual dictionaries, searchable corpora, aligned bitext corpora, and software for accessing Southeast Asian languages, beginning with the most difficult: the complex-script languages of the mainland.

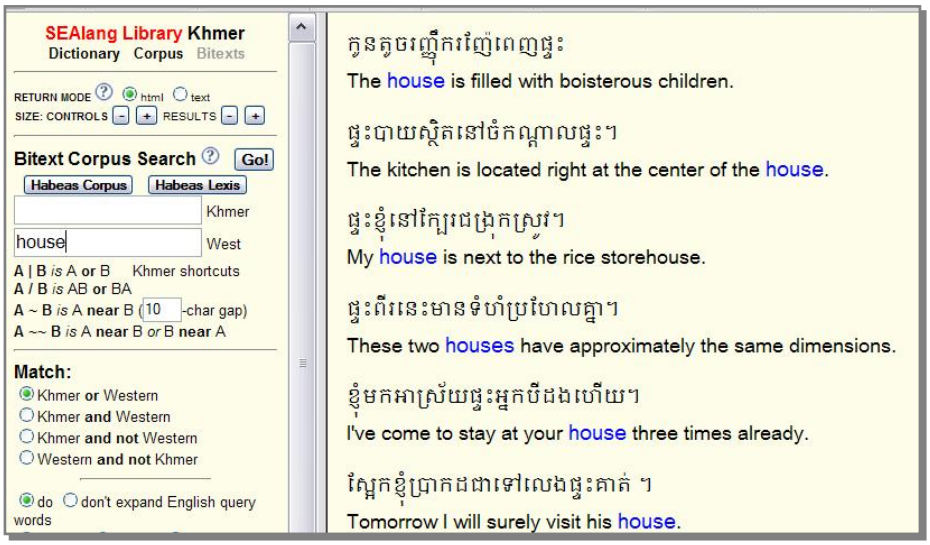
DIGITAL DICTIONARIES provide the most familiar view of SEAlang: a dictionary-query interface, and dictionaries for Thai, Khmer, Lao, Burmese, and Shan are available on-line. All share features that make the dictionaries useful for research applications as well as ordinary lexicographic reference. For example, queries may be entered in local orthography, IPA phonetic, or any combination. Wildcards are allowed, so that particular orthographic / phonemic traits can be precisely specified.



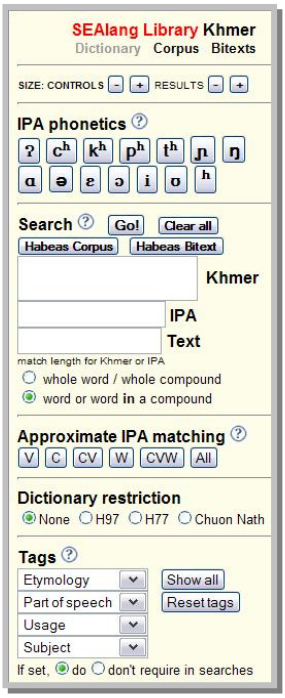
MONOLINGUAL TEXT CORPORA are the SEAlang Library's second major component. Here, we've queried the Thai word nominally defined as 'country' in a 50-megabyte sample of newspaper text. We see a quick overview, with samples both of immediate left and right *collocates*, and of the target word in longer *contexts*. Items in blue are known compounds. Ranking reflects frequency: **201/20.1%** means that a pair was found 201 times in a 1000-item subsample (randomly selected on each query from nearly 16,000 appearances in the corpus).



BITEXT CORPORA are SEAlang's third main resource. We find close translations, align them sentence-by-sentence, and allow searches in either language. Bitexts originate in a variety of sources, and range from sentence examples mined from dictionaries (for some minority languages, our only source), to parallel texts of thesis abstracts. These have been enormously useful in teaching applications (see the *SEAlang Lab*, below), as well as prompting research applications: our colleague Lwin Moe is completing his thesis on automated alignment of Thai-English bitexts.



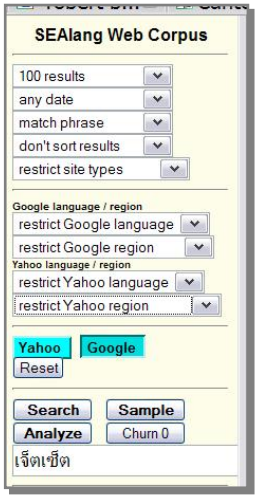
SEALANG LIBRARY RESOURCES are deep and authoritative. For example, our Khmer tools include both of Headley's major references (the 1977 edition is notable for its grammatical and etymological analysis, including nearly 10,000 Pali Sanskrit citations), while the 1997 edition is enriched with far more usage tagging and more than 2,000 example sentences). Below left, note that hundreds of relevant subentries (compounds that include the search term) are also found. This 'just-in-time' data-mining dramatically increases the utility of even the best print dictionaries, which usually only list compounds that begin with the current head.



We have also extended the Library's original plan by seeking out monolingual resources. For example, CRCL funded the Buddhist Institute of Phnom Penh in digitizing and tagging the 1966 Chuon Nath dictionary; the high-water mark of Khmer lexicography. We continue to work together on a more ambitious series: digitizing the major Khmer-French historical works (Aymonier 1872, Bernard 1902, Guesdon 1930, and Tandart 1935), with the cooperation of our colleagues at the EFEO. Similar initiatives for Lao (thanks to the Mahasila Viravong Library) and Thai (thanks to the Royal Institute) are also underway.

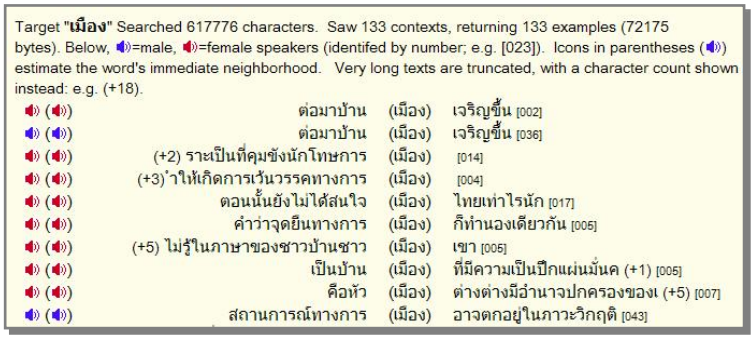
Another set of extensions derive from our roots in computational linguistics. For example, each term in the Thai dictionary has a **WebRank** following the head. This figure is calculated by looking at the number of Web pages located by Google and/or Yahoo (how we retrieved these for more than 100,000 terms is a story in itself!), and serves as a first approximation to the word's difficulty. This helps make sense of compound-word cross-references (we list them by WebRank), and has an extraordinary application, discussed below, in extracting lexicons from arbitrary texts in the SEAlang Lab.

We also use the Web as a corpus – after all, no fixed collection can match its size, lexical coverage, or immediacy. SEAlang's **Web Corpus** is an experimental tool specifically designed for less-common language research. It launches parallel search threads that collect up to 1,000 Google or Yahoo pages, then segments, analyzes, and reports on both collocates and contexts, just like the fixed corpus, all in about 10 seconds. We estimate that we're able to access a *terabyte* Thai corpus in this manner (as researcher Rikker Dockum is doing for our **Thai in Transition** project, which is investigating official vs. popular perceptions of, and responses to, language change).



Reference materials aren't just texts, of course. Below left we see the results of an *audio corpus* search. It draws on resources prepared for research in machine-speech understanding, but re-purposed (as part of the SEAlang Lab) to provide content for listening practice and quizzes. We've also begun to collect *lexalia* – 'realia you can read' – with the help of researcher Frank Smith and as seen below right (read more at http://www.lexalia.org)

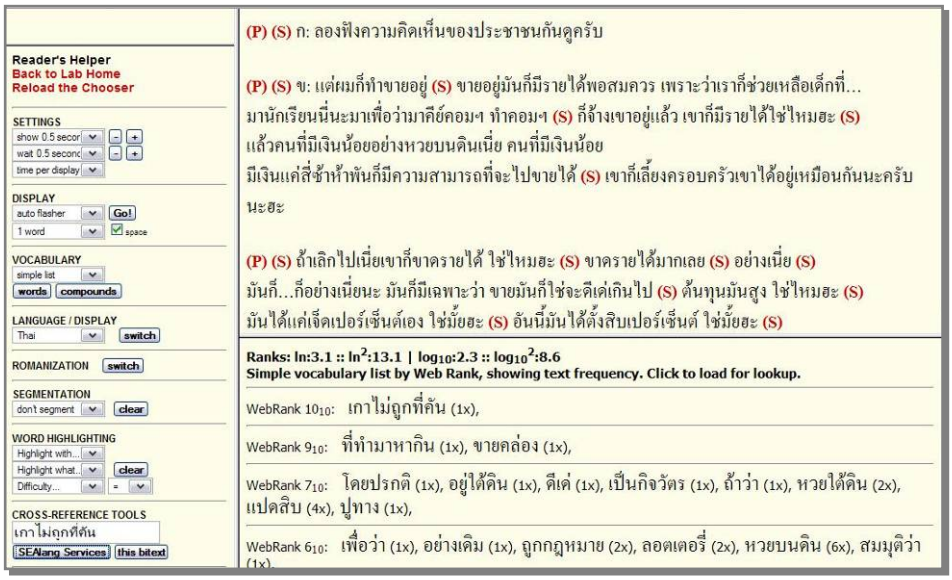
LOOKING AHEAD From the start, SEAlang Library development was conceived as a two-stage process, focusing first on the complex-script languages of the mainland, then moving outward to the major insular Southeast Asian languages, and 'inward' to minority mainland languages (including Hmong, Mien, Wa, and the Chinese dialects Wu and Fouzhou). We welcome suggestions for further coverage and services from the SEAlang Library.



SEAlang Lab

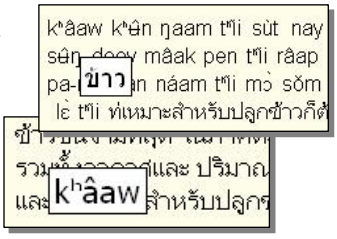
ABOUT THE PROJECT The Lab focuses on the three most difficult areas for students of the many less commonly taught languages that require complex scripts: *reading*, *writing*, and *vocabulary acquisition*. Very little work has been done on the special hurdles such languages present. Thus, we build as wide a variety of tools as we can: not only to help students now, but to encourage research into the effectiveness of different approaches to *assistive technology* for language learning and use.

The Lab probes the limits of autonomous, data-driven learning: closely coupling extensive dictionary, audio, corpus, and bitext reference resources with automated generation of review and test material. It could not have been conceived without the Library's resources, and has proven to be an excellent testing ground for our on-line data-access mechanism. By the same token, Lab development has made significant contributions to the Library, especially in the extent and quality of its aligned Thai-English bitext corpus.

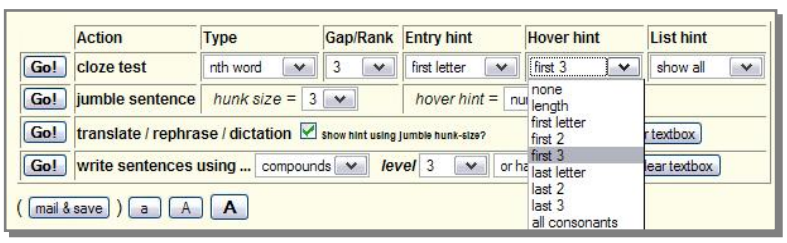


READER'S HELPER The variety of available data resources can be seen above right; each dataset works, if at all possible, with every Lab tool. For example, aligned bitexts are obviously suited for reading practice, but they are also ideal for work on translation or rephrasing in either language direction. By extracting a wordlist from the text, we have grist for our vocabulary mill.

Above, we can see the Reader's Helper as invoked on an ILR-rated text. Texts may be presented in a variety of commonly encountered fonts, with or without segmentation, and using romanization and/or popup tool-tips to help less-sure readers. Indeed, the 'auto-flasher' (a **Display** option) can even flash the text in arbitrary chunks, at nearly any display / pause rate. Thus, the Reader's Helper is designed not only to assist and encourage broad reading at all levels, but to provide necessary resources for research in how students master complex-script languages.

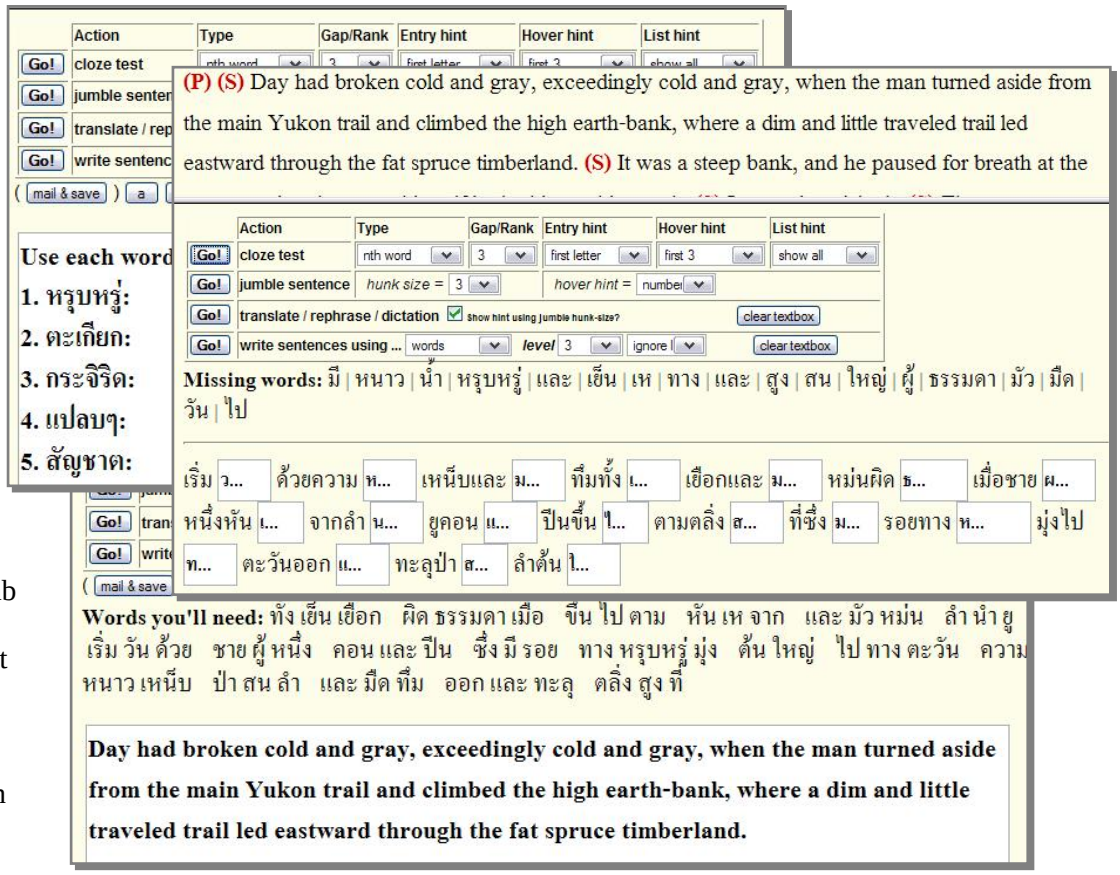


We've also had the Reader extract a vocabulary list (of compounds), ranking it using the **WebRank** statistic. We're investigating using the same approach in generating a *reading difficulty* metric for the text, which will some day be very helpful for locating class or study material on the web.



WRITER'S HELPER Also designed for students at all levels, the Writer's Helper treats every kind of composition – even cloze tests, or reordering of jumbled sentences – as a necessary part of developing complex-script skills. Like all Lab tools, these exercises are generated on the fly, using aligned bitexts and other word boundary and frequency information obtained from the SEAlang Library.

Here, we see several of the exercise types implemented in the Lab project's first year. Those that involve original composition, translation, rephrasing, or writing example sentences can be mailed to the instructor.



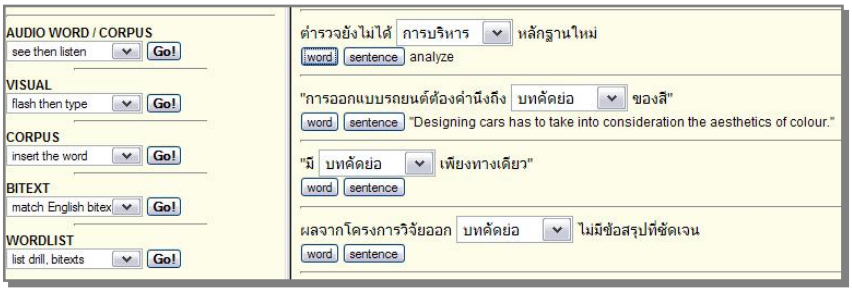
It's important to remember that the Lab is concerned with *all* less commonly taught complex script languages. Appearances to the contrary, very little in the wide variety of exercises seen here depends on Thai-specific development.

VOCABULARY HELPER Tools that focus on vocabulary and drill round out the Lab. Our goal is to go well beyond traditional flashcard drill methodologies by *a*) drawing on every available SEAlang Library resource, including text and bitext corpora and audio samples, as a source of drill content, and *b*) using these resources to create the vocabulary lists themselves.

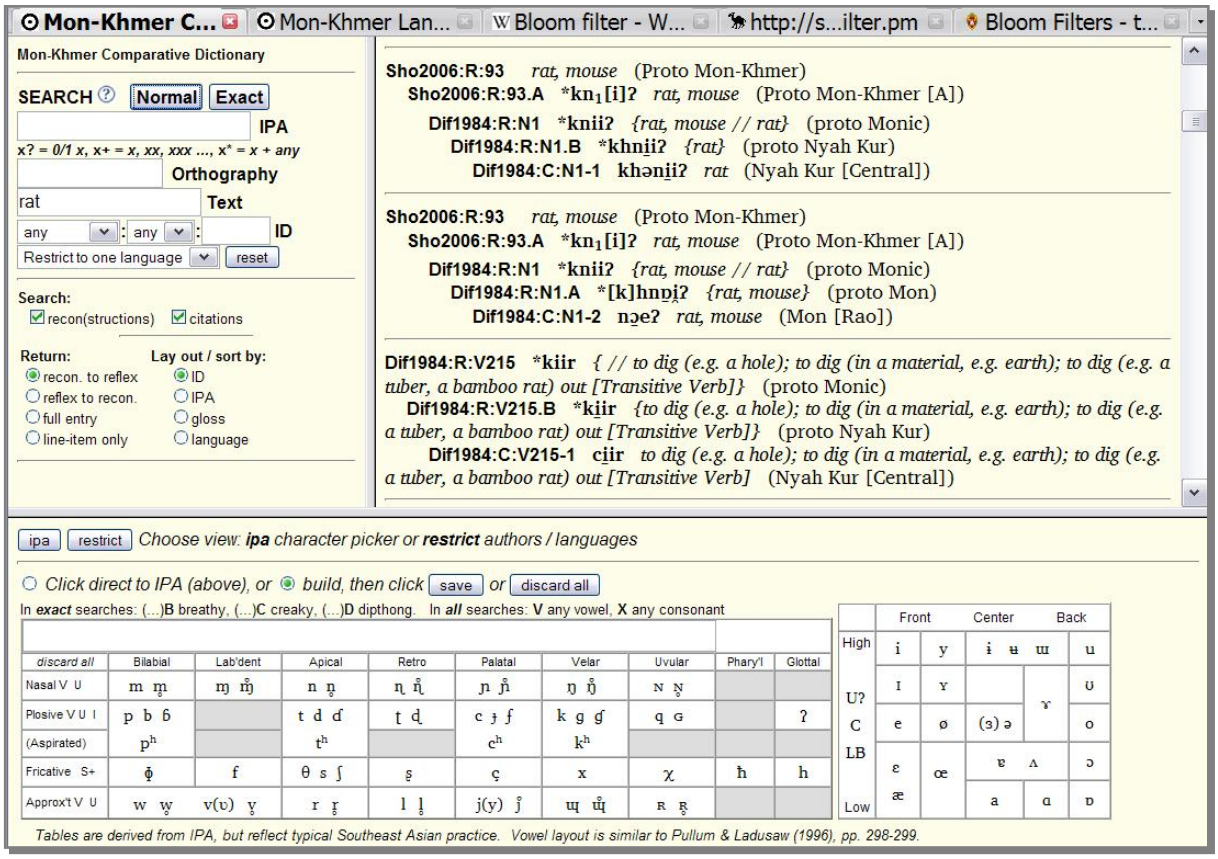
For example, below left we use sentences pulled at random from the Thai corpus for one exercise type; below right, randomly drawn bitext examples serve the same purpose.

Our Thai Academic Wordlist (**AWL**) project provides a typical example of the use of Library resources in Lab development. Historically, most effort language resource development goes to introductory materials, while advanced students work their way through texts, acquiring lexicon as encountered. Such incidental word acquisition is less than effective; coverage becomes increasingly difficult, and retention declines as words are encountered less frequently. The alternative approach is to identify words that may be relatively infrequent, but nevertheless characterize academic or other advanced writing.

Once we've defined a list of words and glossed, we can use the bitext corpus to a content-rich exercise, complete with word and sentence-translation clues, as is seen below. We can do the same with *any* wordlist – even a list that has been automatically extracted from an arbitrary text file or Web page.



Mon-Khmer Languages Project

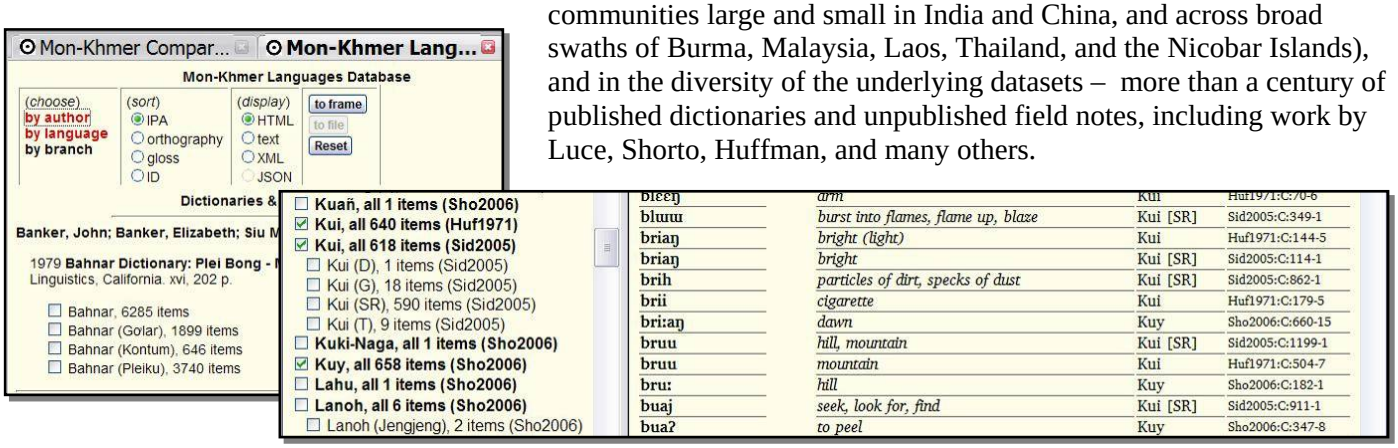


ABOUT THE PROJECT This major initiative is funded by the National Endowment for the Humanities (initial phase 2007-2009) to build on-line etymological and comparative databases of the roughly 150 Mon-Khmer languages, mostly endangered, spoken throughout Southeast Asia. Data and expertise are contributed by collaborating scholars worldwide, including France, India, England, Vietnam, Germany, Australia, Singapore, Thailand, and the U.S.

Like the Lab, the MK project is made practical by data, software tools and infrastructure developed by the SEAlang Library. Khmer, Vietnamese, and Mon are all essential MK branch exemplars. But the critical ingredient in designing a useful etymological and comparative database has been the approximate phonetic search technology originally created to help students deal with unfamiliar modern orthography in the Library

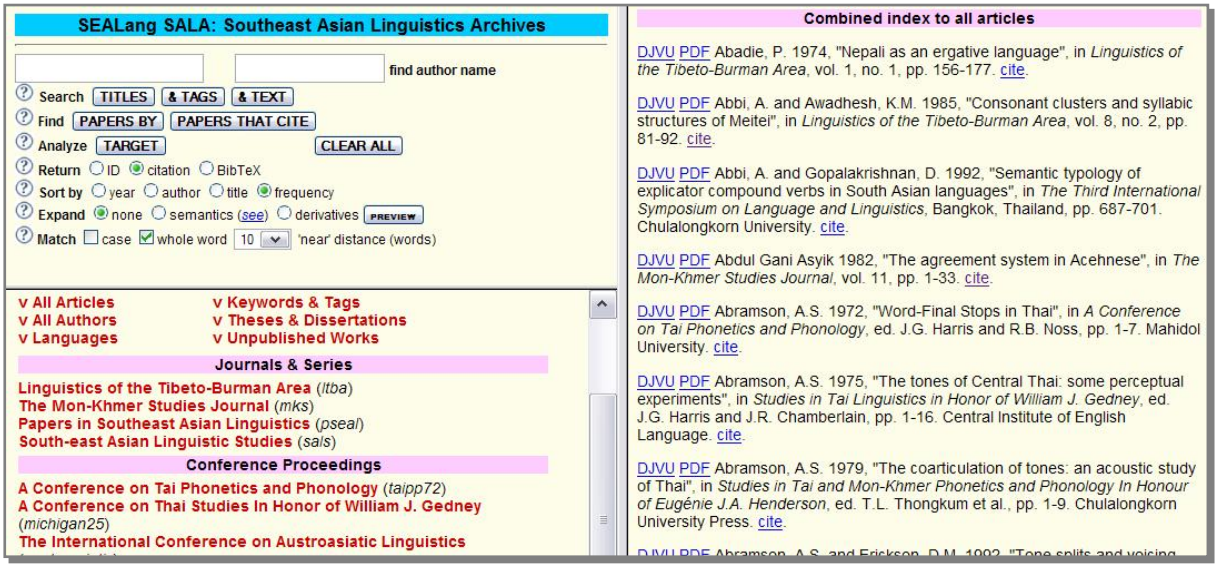
MK COMPARATIVE DICTIONARY Above, we see the results of searching for *rat* as part of a gloss. Results are grouped (as possible) into etymological sets from past to present; this can be reversed (present to past) for use in extending modern dictionaries. For phonetic search, the 'allophone set' builder on the bottom is used. A wide variety of natural classes, which take predictable historical and areal variation into account, are predefined and accessible simply by clicking on rows, columns, or abbreviations.

MK LANGUAGES DATABASE Below, we see examples from the MK Languages Database. Each data item is tagged with its source, and full sets can be aggregated, sorted, and returned either as plain text, or in a reusable tagged format; below right we merge various Kui/Kuy sets. The database stores an extraordinary range of data, both in geographical distribution (Mon-Khmer languages are spoken universally in Vietnam and Cambodia, and in communities large and small in India and China, and across broad swaths of Burma, Malaysia, Laos, Thailand, and the Nicobar Islands), and in the diversity of the underlying datasets – more than a century of published dictionaries and unpublished field notes, including work by Luce, Shorto, Huffman, and many others.



Southeast Asian Linguistics Archives

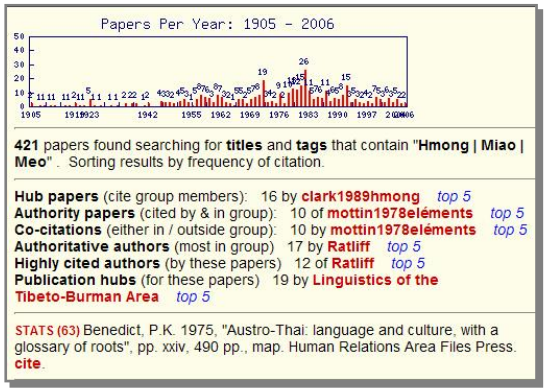
ABOUT THE PROJECT The *SEAlang SALA* collects, scans, indexes, and disseminates scholarly publication on Southeast Asian language and linguistics, and devises and tests innovative approaches to aggregating and assisting discovery of the field's scattered literature. The SALA includes:



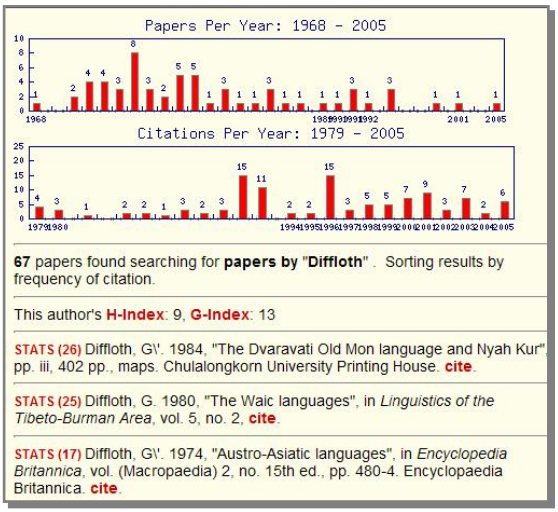
- A *searchable on-line archives* of major journals, conference proceedings, series, and collections that focus on (mainland) Southeast Asian linguistics, and unpublished material (field notes, theses) unavailable in the US.
- A *bibliography* (in *BibTex* and reusable XML-tagged formats) of related language and linguistics publication. This provides much deeper coverage than existing alternatives, and extends well beyond mainland SEA.
- A *corpus of citations* and *analytical citation index*, with Web-based tools that assist in discovery and evaluation of texts (*who does this article cite? who cites this article? what articles have similar citations?*), and calculates impact factor, G-index, and other measures of influence.
- An *API* for computer-to-computer interoperability between the SALA and other tools (e.g. *Multi-Tree*, *LL-MAP*, other SEAlang projects), as well as data sharing and citation tracking with other digital archives.

Open, on-line access to publication is essential to the health of the field for many reasons: the unusual extent to which basic data appear in journals and conference proceedings, the increasing difficulty and cost of obtaining materials published outside of the US, and the equal difficulty faced by our colleagues in Southeast Asia who do not have access to expensive aggregator services like JSTOR. Most important is the critical role the literature plays in educating young scholars from the region.

The SALA's facilities for *semantic query expansion* (above) are of particular practical and research interest. Individual Southeast Asian language (and even families) are identified in the literature by a dizzying variety of names; sometimes due to simple spelling variation, but often the result of differences between speaker self-identification and (often derogatory) outsider names.

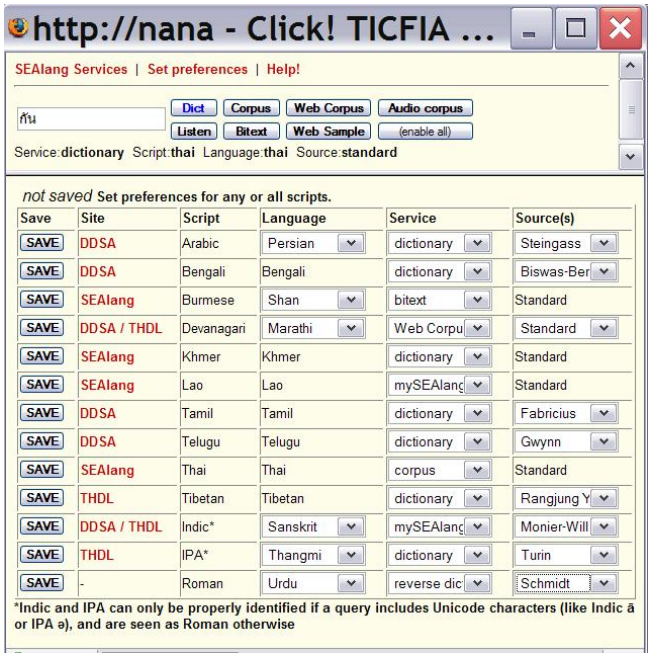


But it is *citation graph analysis* (left and right) that will seem most magical. We're just at the beginning of a lengthy process of gathering and correcting data, but the SEAlang SALA can already demonstrate the extraordinary power of this technique. Such tools help us all – authors, readers, and publishers – find a way to join forces in supporting ready access to printed knowledge.



SEAlang Services: *Click!SEAlang, myGloss*, and Beyond

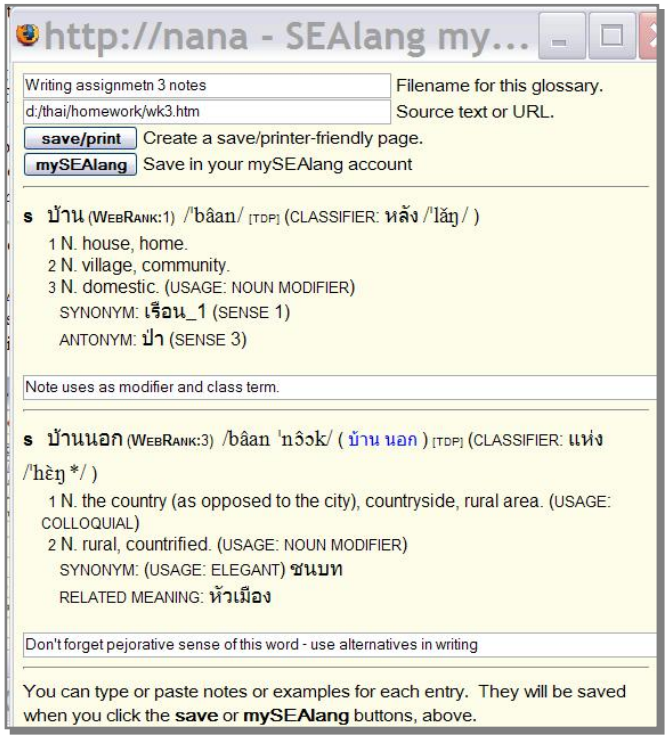
ABOUT THE PROJECT The SEAlang Library takes a leadership role in making data resources *interoperable* – providing mechanisms to let Library content be accessed and reused by direct, computer-to-computer communication. The critical component of this effort is a Web-based *application program interface*, or Web API, that lets simple queries search the Library directly, without requiring the user to work though an on-line Web page.



The SEAlang Web API is being applied to all of the services we offer, including dictionary lookup, corpus queries and the like. A request to **http://api.sealang.net/?service=dictionary&query=บ้าน …** looks up the word in the appropriate dictionary, then returns a result that can be displayed, or parsed and re-purposed.

We are also creating 'middleware' functionality that lets us satisfy requests from other projects that do not publish their own APIs. The screen capture on the left shows how a user sets preferences for one-click access to dictionaries supported by our sister TICFIA projects *Digital South Asia Library* and *Tibetan-Himalayan Digital Library*. The lookup is initiated from any Web page through a right-click, button, or bookmarklet. A highlighted term *in any language* is sent to our server, where its script – from Arabic to Thai – is detected automatically, and a preset service (dictionary, corpus, sound, or any other available) is invoked.

myGloss One of our most frequent feature requests has been for a way to collect and save dictionary query results. A glossary-oriented **save** feature is now built into the SEAlang dictionaries: words can be saved in named sets (so that a teacher can create study lists), individually annotated, and then printed or stored on the SEAlang server.



mySEAlang mySEAlang also has an experimental provision for creating, saving, and sharing richly annotated vocabulary lists. When a word is saved, the user also adds a gloss, hint, monolingual and bitext corpus example, image reference – any and all information that might be useful for generating review and drill later on. Such lists can be prepared collaboratively, with students sharing responsibility for each week's vocabulary portion, and shared within a class, or released to the general public.

What's next? The SEAlang Projects continue to explore ways of adapting data or skills developed in one context to meet the needs of another, and to support both independent and affiliated projects. Among other efforts, we're working with Philip Jenner on lexicography and epigraphy of Angkor and before, with Seang Sokha on Khmer-French lexicography, with Saowapha Virawong on library search and access issues, with Noosai Inthimas on Thai video lexicon recording, and with Stephen Morey and Zeenat Tabassum on early Tai languages. We're also hosting the websites of the *Mon-Khmer Studies Journal* and the *Journal of the Southeast Asian Linguistics Society*, and archiving an increasing store of unpublished and out-of-print field notes and texts.

Contact Doug Cooper at **doug.cooper.thailand@gmail.com** or visit the SEAlang website at **http://sealang.net**