The Kamusi Project Edit Engine: A New Tool for Collaborative Lexicography

By Martin Benjamin and Ann Biersteker, Yale University

Published in: *Journal of African Language Learning and Teaching*, v1 n1 p75-88 Spr 2001

The Kamusi Project's *Internet Living Swahili Dictionary* was conceived in 1995 as a work in which Swahili speakers and scholars could participate to create new lexicographic resources for the Swahili language. The project's original model involved having remote participants submit glossaries and suggestions that were then compiled on a central computer using the *Excel* spreadsheet program. Catholic University Press then granted the project copyright permission to incorporate Charles Rechenbach's *Swahili-English Dictionary* (1968) into the lexicon, necessitating much additional data entry and manipulation in *Excel*. Updates of the work-in-progress were posted to the project's web site, www.yale.edu/swahili, about once a year. The Kamusi lexicon quickly grew beyond the capacities of *Excel*. After an evaluation of other existing software, and in consideration of emerging Internet technologies, the project employed programmer Joe Rodrigue who worked with a *Unix* platform and the *Mysql* database to assist in the creation of a web-based software system uniquely suited to the needs of Swahili collaborative lexicography.

This essay discusses the design and implementation of this proprietary software, the Kamusi Project Edit Engine. The paper first describes the Edit Engine, including organization of the lexicon and the mechanics by which participants use the system. Next, the paper discusses a few of the philosophical issues confronted in the design of the system, particularly considerations of maintaining control over the integrity of the data

while yet striving for democratic inclusiveness. Finally, the essay points to additional applications for which the Edit Engine model can be expanded with further programming and funding, including more in depth work on Swahili and the creation of on-line lexicons for other African languages.

Description of the Edit Engine

Kamusi Project participants access the Edit Engine through the dictionaries at the project's web site. When users browse the dictionaries alphabetically or search for particular terms, they are served web pages that display the appropriate data, with keywords and glosses highlighted as hypertext links. Clicking on a gloss results in a reverse dictionary search, whereby the gloss becomes the keyword. (For example, if the original look-up is an English-Swahili search for "house," clicking on the gloss "nyumba" produces a Swahili-English look-up for "nyumba.") By clicking on the keyword ("house," in the first example, or "nyumba" after the reverse look-up), users enter the Edit Engine.

The fields in the Edit Engine include Swahili Word and English Word, Swahili Sortby and English Sortby, Swahili Plural and English Plural, Part of Speech, Class, Swahili Definition, Swahili Example and English Example, Related Words, Dialect, Used In, and Comment. Additional fields that will be included as project resources allow, especially a system to rank and group entries, are discussed in the final section. The following list discusses the function of each field that is included in the Edit Engine as of this writing:

- Word. Each entry must contain both a Swahili word and an English word. These are the major facets of the entries that appear to users of the online (and forthcoming

print) dictionaries. Items in the Word fields are often quite long, for example glossing Swahili word *kanga* as English word "cotton cloth (with designs in several colors) worn by women." Because these items can differ from the headword under which they should be listed or for which a user may search, the Word fields are not used as the primary fields for searching or ordering the online dictionaries. The Word fields are used for secondary ordering of the dictionaries as well as for advanced searches (searches for all entries beginning with or containing the search term). The Edit Engine maintains a one-to-one relationship between Swahili and English words, meaning that each English gloss for a Swahili word must have its own entry, and vice versa. A user who wishes to add an additional gloss for a Swahili or English word must enter the information in a new line.

- Sortby. The Sortby fields tell the database where to display each entry in the dictionaries, but are never actually seen by ordinary dictionary users. For most entries the choice of Sortby is intuitive, but sometimes the decisions to date have been arbitrary and will need to be reconsidered. What, for example, should be the English Sortby for "United States of America"? In such a case, a duplicate line may be necessary, so that the entry will appear under both "United States" and "America". Furthermore, to speed access time for web users almost all Sortbys have been restricted to a maximum of sixteen characters, an efficiency-oriented limit based on log records that show virtually all user searches begin with strings of shorter length.
- Plural. Swahili plural forms often differ from their singulars by the addition of prefixes. The Edit Engine includes fields for plurals so that searches of the online Swahili-English dictionary can usually, by searching the plural field when looking for

terms beginning with the plural prefixes, give users the results they are seeking. In addition, learners of both Swahili and English can learn the appropriate plural forms when the Plural fields are returned as lexical items when they search or browse the dictionaries.

- Part of Speech. The Edit Engine provides a restricted list of parts of speech from which the user selects one for each entry. If a word can be more than one part of speech, for example being both an adjective and an adverb, then each should have a separate entry. The choice of part of speech is less straightforward than it might appear, especially when a single-word term in one language requires a torrent of words in the other. For example, the adjective "brown" in English has among its colorful Swahili translations, "rangi ya udongo" (lit.: the color of dirt) and "rangi ya damu ya mzee" (lit.: the color of an old person's blood). Do the Swahili glosses map as adjectives, or as phrases, or as adjectival phrases? The current version of the lexicon subsumes adjectival phrases as adjectives, noun phrases as nouns, verb phrases as verbs, etc.. A main reason for such a simplifying algorithm has been to reduce the confusion that multiple choices will generate for multiple editors. The editorial rule is that if a term is clearly a unique part of speech in either language, it will be entered as such in the single Part of Speech field that applies to both languages.
- Class. A distinguishing characteristic of Bantu languages, including Swahili, is the grammatical noun class system. Each noun is a member of one of fourteen classes that determine the pattern of plurals, adjectives, pronouns, and verb conjugations within each sentence. Learners often find that an indication of class helps them figure

out how to employ nouns correctly. The Edit Engine allows participants to select a noun class for each noun from among a numerical system that is used by linguists. Future computer programming may enable users to view entries using other classification systems they find more helpful. The Edit Engine also allows users to select a "class" for verbs, to indicate when a verb entry is shown in a derived form such as passive or causative. If the participant ignores the class field, the entry will appear in the database as "unspecified."

- Swahili Definition. An eventual goal of the project is to produce a comprehensive Swahili dictionary of Swahili words. The Swahili Definition field is a newly introduced feature that stands as an invitation to participants to write definitions in clear and succinct Swahili.
- Examples. The current lexicon includes several thousand Swahili examples derived from various sources, but very few English translations. In addition, for reasons peculiar to the use of *Excel*, many example sentences were inadvertently replicated in entries for which they are erroneous. The Example fields provide an opportunity to translate and increase the quality of the existing Swahili examples, as well as increasing the quantity of examples for both languages.
- Related Words. Originally shown as "Derived," the Related Words field is in the process of being converted to a feature that will link the user to words with similar roots. When it is possible to discern that a word is borrowed from another language, such as Arabic, Persian, or English, the Edit Engine will link the user to available information or online dictionaries of the appropriate language. This feature is still under development at the time of writing.

- Dialect. Swahili has several regional dialects that are notable for some differences in vocabulary and pronunciation. For want of a better term, the Edit Engine calls these variations "dialects," and provides a restricted list of possible options from which the participant may select. The default category is "standard," which indicates a term that is in common use by Swahili speakers throughout Tanzania and Kenya.
- Used In. One objective of the Kamusi Project is to create specialized vocabularies for terms that are used in specific fields of endeavor, such as medicine or law. The Edit Engine allows users to select a vocabulary for which an entry should be included, or if it should be listed as rare or archaic. Unfortunately, including a term in more than one vocabulary will necessitate giving it more than one entry in the database. The default vocabulary is "general."
- Comment. Many entries require special explanatory notes about usage or cultural context. The Comment field allows an open-ended space for participants to add such usage notes.

The first screen that users see in the Edit Engine allows them to select among one or more lines to edit. The lines that appear on this screen are all those that match the same "sortby," or headword, as the original keyword that the user clicked. On this screen the user can see the entire contents of the database for each line, including all the lexical items discussed above. The user checks a small box next to each line to include that entry in the selection to edit, or can choose to edit all the lines on the screen. If the participant wants to suggest a new or additional entry, s/he can also generate blank lines from this screen.

The second screen is an editable table of all of the lines the user selected from the first screen. Each line displays the original content of the database in an editable cell, as well as displaying the original content in fixed form immediately below the editable cell. If the original entry is a noun that does not specify a plural form or a noun class, or is a verb that does not specify a class, the Edit Engine highlights the appropriate cell in yellow to indicate that the item might need to be fixed. If the Edit Engine determines that the original entry does not contain a value for one of the required fields (Swahili word, Swahili sortby, English word, English sortby, or Part of Speech), or if a value is given in the Plural field but the entry is not given as a noun or pronoun, the appropriate cell will be highlighted in red to indicate that the item must be fixed. The participant may choose to make changes to any or all cells in the table, regardless of whether the cell is blank or already contains data.

After the user has completed editing the chosen entries, the Edit Engine compiles his/her work on a new screen that shows all the edited fields in peach-colored cells, as well as showing cells that still must be completed in red and showing fields that may need to be completed in yellow. The user has the option to be brought to a new edit screen to make further changes, or continue to the final stage of proofreading. On the final page, the participant sees all of his/her changes as they will be submitted to the editor. Users must enter their "screen names" (or register with the project to select a screen name at this point), after which they can submit their changes.

At this point in the process, the Edit Engine writes a unique web page for the pending submission. The program sends an email to the editor that lists the edited headword, the participant's screen name, and the web address for the submitted entry.

Clicking on that address brings up a page that shows the changes that the user has suggested, as well as the original entry. The editor can choose to accept the changes as they were submitted, reject the changes entirely, or enter into the editing process to make further changes. If the editor chooses to edit the submission, the Edit Engine will cycle through the same series of steps that the participant encountered. Once the editor is satisfied with the entry, it is immediately incorporated into the database, and the Edit Engine runs a set of processes to make the updated information available online.

Philosophy behind the Edit Engine

The complexity of the Edit Engine programming model is largely due to the need to reconcile two competing objectives. First, the Kamusi Project must maintain control over the integrity of the lexicographical data. Second, the project strives to include the expertise of as many Swahili speakers and scholars as possible, while also opening to others a window to the Swahili-speaking world. The program must therefore be easy to understand and use, comprehensive in enabling participants at all levels to make useful contributions, but secure from inappropriate manipulation. All this must be accomplished within a restrictive programming environment that will provide functionality at any time on any number of computer platforms running a variety of software interfaces. Finally, the program must attend to the complicated lexicographic model described above.

The Edit Engine accomplishes the goal of protecting project data through a system that buffers the actual database from remote users. When a user accesses the Edit Engine, s/he is able to see the contents of the database exactly as it exists. At no point, however, can the user make changes directly to the central data files. Instead, any

changes that participants wish to suggest are bundled into unique web pages, with web addresses that include long strings of random numbers for security purposes, where they sit pending approval by the editor. In this way, the editor is able to check the submissions of trusted participants for such problem as input errors, and is also able to discard entirely any submissions from people who are intent on mischief. The Edit Engine will only incorporate changes to the database when they come from a computer that it recognizes as belonging to the editor. The review process thus checks for errors by following this routine: 1) User decides to edit the entry for item X; 2) User previews all entries with the same Sortby as item X and chooses which to edit; 3) User makes changes; 4) User reviews changes, and has the option to return to step 3 or; 5) User proofreads changes and submits; 6) Editor receives email; 7) Editor's email opens web page that shows submitted changes; 8) Editor proofreads changes in the context of all entries that share Sortbys with submitted entries, and updates database or; 9) Editor makes further changes; 10) Editor reviews changes and has the option to return to step 9 or; 11) Editor proofreads changes and updates the database; 12) User receives optional email notification that the database has been updated and can review the revised entry.

By designing a program that can be used from any computer linked to the Internet, the Kamusi Project fosters an approach to scholarship that seeks a high level of democratic inclusiveness. The primary audience for the Edit Engine consists of scholars and speakers of Swahili who wish to share their expertise and their love of the language. Dictionaries are usually compiled by university-trained linguistic authorities, a group from which the project eagerly solicits contributions. Scholars with a rich understanding of language provide a depth of knowledge that will translate into richly informative

dictionary entries. The Edit Engine is also designed, however, to elicit contributions from non-"experts" who nevertheless are closely attuned to the language as it is used and as it evolves. Although it will be many years before most Swahili-speakers have access to the Internet, more East Africans are able to participate every day. The Edit Engine is a tool to achieve the goal of a "living" dictionary that chronicles the Swahili language not only as it was or as purists wish it to be, but also as it is spoken. A participant in Dar es Salaam can submit a term as heard today, and it will be available online globally tomorrow. This democratic approach is intended to produce a lexicography that is both linguistically dense and actively current.

The Edit Engine also encourages participation by people who have no particular expertise in Swahili. Such individuals can contribute by filling in such items as English plural forms or suggesting English example sentences. Teachers can work with their students of Swahili on projects such as suggesting noun or verb classes, or helping find related words. The editorial buffer system, of course, prevents erroneous student submissions from harming the actual database. Through the process of including learners as participants, the Edit Engine is designed to arouse interest in the language and to help students in the process of learning it.

The final decisions about the entries that appear in the Kamusi Project lexicon remain in the hands of a single editor. Consistent oversight is necessary so that errors are minimized and so that the project remains focused, especially given the complexity of administering the Kamusi system. The project is currently programming a method to place difficult entries in a special web zone where participants can discuss specific lexicographic challenges. (Meanwhile, the editor manually places difficult submissions

in a file for pending work.) The Kamusi Project has also applied for funding to transfer direct editorial control of the lexicon from Yale to the Institute for Kiswahili Research (TUKI) at the University of Dar es Salaam, the institution with the largest body of active researchers of the Swahili language. Even after the transfer, however, final editorial authority will need to remain vested in one person or cohesive group. The editor must remain committed to following the guidance of project participants, and participants must accept that some editorial decisions must be made at an executive level. The Edit Engine is a system that democratically accepts input from all quarters, but maintains its quality through rigid scholarly oversight.

Applications for Expanding the Edit Engine

The Edit Engine model can be expanded in a variety of directions. Immediate possibilities include more in depth work on Swahili, and the creation of on-line lexicons for other African languages. These various applications would depend on further computer programming, the availability of funding, and the interest level of qualified scholars.

Within the basic Edit Engine model, the Kamusi Project lexicon can be enhanced in several ways. The most important missing feature is a method to group and rank entries. Currently, entries are arranged alphabetically first by the Sortby field, then by Part of Speech, and finally by the Word field. This current system is unsatisfactory because it does not separate out homophones, distinguish shades of meaning, or indicate which are the preferred entries. Unfortunately, the programming necessary to rectify this deficiency will be quite intricate, though preliminary efforts are underway. Additional programming also will be necessary to make the Related Words field have the capacity to

link to multiple dictionary entries. The project currently has the capability to link recorded sound files to dictionary entries, but must devise a convenient method for remote users to contribute spoken Swahili examples before instituting an audio feature, or devote considerable resources to recording and linking tens of thousands of individual files. Similar considerations must be given to how best to incorporate contributions of visual files, including photographs, illustrations, and video. The structure of the Edit Engine interface with the *Mysql* database enables a wide range of potential applications to which remote participants can contribute. Decisions about which features to undertake first will be subject to ongoing evaluation and prioritization.

An immediate application that requires competent scholars to administer is the potential to incorporate translating dictionaries between Swahili and additional European languages. The Kamusi Project currently links to a static Swahili-Russian dictionary compiled by Dmitri Polanyov, and plans soon to host a similar Swahili-Spanish dictionary compiled by Chege Githiora. Work can begin to make each of these dictionaries fully web searchable and editable with minor modifications to the Edit Engine. An editor is needed to oversee the work on each language, and each language may require some specific changes to the lexicographic model. The core Swahili database and the simplicity of the Edit Engine system can then make the inclusion of these languages a relatively straightforward process. Other languages, including French and German, can also be added to the Kamusi Project in the near future.

The Edit Engine can also be modified as a tool to build online dictionaries for other African languages. Such work will need to accommodate the challenges of designing appropriate lexicographic models, working with non-standard characters and

alphabets, and accounting for grammatical and linguistic features that may differ from Swahili. Further, not all concepts in the Swahili dictionary will have direct equivalents in other African languages, so it will be impossible to develop lexicons that provide smooth glosses from both Swahili and English to languages that are to be incorporated. However, the Edit Engine and the Kamusi database provide a pre-existing tool and compendium of data that can, with appropriate modifications, be expanded for a wide variety of African languages. Funding proposals are currently being considered to begin with a Yoruba dictionary and an Oromo dictionary in the second half of 2000.

The Edit Engine and the Kamusi database also have potential as resources for research in linguistics. Using wildcard searches it is possible to search the database for lexical items and examples that incorporate specific affixes. For example, in research on reciprocal verbs in Swahili it is possible to search the database for verb forms and examples by doing a search for the reciprocal verb suffixes using the search items "*an*" and "*ani*." Similar searches could be done for other verbal suffixes as well as for nominal and subject prefixes. As additional examples are incorporated this capability will be expanded.

The major constraints to expanding the Kamusi Project to other African language applications are funding and personnel. While the current Kamusi administrators are eager to work with other scholars to establish such projects, there are limits to the time and expertise they can offer. It will therefore be necessary for interested others to step forward in order for such work to proceed. Unfortunately, African languages have not historically generated much interest among funding agencies. We hope that by offering the Edit Engine as a model and a tool for future projects, future work will be stimulated

that scholars will want to undertake and funders will be excited to sponsor, in the furtherance of cooperative scholarship for African languages.