Lexicography without Lexicographers: Crowdsourcing and the Compilation of a Multilingual Dictionary

Martin Benjamin
École Polytechnic Fédérale de Lausanne
LSIR
BC 114, Station 14, Lausanne
1015 Switzerland
email: martin.benjamin@epfl.ch

This paper discusses the experiences of the Kamusi Project with engaging the public in the production of the Internet Living Swahili Dictionary (ILSD), and the new crowd features being introduced as the project transitions toward a massively multilingual Global Online Living Dictionary (GOLD). The project has always been conceived as collaborative but controlled – users are encouraged to contribute new or improved entries, but those contributions are ultimately subject to authoritative review. The upcoming incarnation, however, will open the project to any language with an ISO-639-3 code; with a potential for 7000 languages, there is no possibility for scholarly oversight of all incoming data, and thus new systems are necessary to ensure that data from the crowd is acceptable.

Kamusi began in the same week in December 1994 that Netscape 1.0 was released, and has evolved along with the Internet since that time. The original idea was to send out packets of English wordlists to the roughly 30 known Swahili speakers online at the time, and compile the results into a static file that could be shared on a Gopher server. The idea was marginally successful, though more data was completed in-house on the basis of existing glossaries than by contributors. After copyright permission was granted to use the data from Rechenbach's Swahili-English print dictionary (1968), active public elicitation was paused for two years while student research assistants manually typed the OCR-hostile entries into an Excel spreadsheet. Once the Rechenbach data entry was complete, however, the project was re-opened to the public with the introduction of the Edit Engine in 1998. The Edit Engine has been the main tool for both staff and public contributions ever since. In the GOLD era, the Edit Engine will continue to be crucial for contributions by specialists, but will give way to much different systems for eliciting public contributions.

The Edit Engine was designed to channel web users toward contributing information in a consistent format, though it was not originally conceived as a source for data that could be operated on for NLP purposes. Growing out of the project's early phase in Excel, data was compartmentalized into fields for headword, part of speech, Swahili definition, etc. (Benjamin and Biersteker 2001). Users could edit any field for any word. Upon submission, the editor was notified, and opened a page that showed the original and revised versions. Changes could be accepted, rejected, or further modified, and contributors were sent notice when their submissions were processed. In this way, new contributors could be coached about any mistakes, and trusted users were nevertheless subject to a final round of proofreading before their contributions were incorporated to the live database.

Our experience showed the necessity of a controlled review process even within the context of a single language. For Swahili, for instance, several different valid systems can be employed for cataloging verbs and treating noun classes. Verbs might be shown in the infinitive (kufikiri/ to think), or purely as a root (fikiri), or with a hyphen to indicate where inflectional affixes can appear (-fikiri). We decided to show the hyphen as the display form, but have a separate "sortby" field that, among other purposes, did not contain the alphabetization-confounding hyphen. This is a very simple convention, but it is not obvious, and needs to be explained repeatedly to new users. Perhaps fortunately, there were never so many contributors to the monolingual ILSD that we could not keep pace with training new members. On the other hand, when one trusted member set out to add English plurals for every noun, it took months to work through the accumulated submissions.

In this decade we have transitioned the Edit Engine to support the multilingual data model described at http://kamusi.org/molecular_lexicography. The GOLD version of the Edit Engine has been used most extensively to produce pilot data for twenty languages, and by Translation Studies students at the University of Ngozi in Burundi to work on data development for Kirundi. We have not pushed more widespread use because we do not have the manpower to review submissions for languages that the staff does not know. Instead, we have focused on developing crowd systems that can bypass the Edit Engine, while continuing to develop "pro" features to provide a robust dictionary development platform for specialists.

The essential challenge we must address is the difference between knowing a language, knowing something about a language, and knowing how to document a language. During the pilot data phase for GOLD, we worked with a number of highly educated native speakers who had never thought systematically about their language. One of the goals for GOLD is a monolingual dictionary for each language, which means writing an own-language definition for each sense of each term. We soon learned that brilliant oral skills and a PhD in mathematics do not necessarily equate to an inherent ability to compose a well-written definition in one's mother tongue. Submitted definitions could be too long (as an astrophysicist describing the properties of a star), too short (often just a synonym), given in English instead of the subject language, poorly worded (X is…), or combining multiple senses in a single entry, in addition to having typos, or being plain wrong in relation to the English elicitation concept. We have been able to overcome these problems with the students in Burundi through intensive training and review, about 50 entries per student to master the platform, understand all the relevant fields, and write competent definitions, but that level of involvement is not possible for languages that do not have funded components. For languages where someone is available who understands lexicography, trusted users can be given moderation privileges, either to work on their own or to review data as it comes in. For other situations, we have been working on systems to harvest knowledge at a base level from people who speak their language but have never studied it, and at a somewhat more detailed level from people who can tell apart their transitive and intransitive verbs.

Our approach to crowdsourcing has three central aspects. First, we break the effort into well-specified microtasks that are, we hope, difficult to misunderstand. Second, we subject every answer to a consensus process, only considering an answer to be valid after it has passed a threshold of repetition or upvotes (depending on the type of

data sought). Third, we are working to make the process fun and compelling. Crowdsourcing is being deployed on three platforms. Within Facebook, we have an expanding set of games, for which users receive points and recognition when their answers win. On mobile devices, we are focusing on very quick questions that users can answer to unlock their phones or spend a few minutes feeling busy – does X (a word that has been proposed in your language as a match for “pen”) mean “a place for the confinement of people being held in detention”; swipe right for yes, left for no, up to postpone, or down for do not know. On the Kamusi website, we intend (but have not yet begun coding) to pose questions relevant to the entries that a user looks up, with an answer requested in order to return to the search interface.

Taken together, we posit that we will be able to obtain data that is both valid and well formatted. For example, if the answer for a microtask could be either “sunglasses” or “shades”, one game will validate the majority answer “sunglasses”, while a second game will discover that “shades” is accepted with the same meaning, but the misspelled “sungalsses” is rejected by the crowd. Such data can then be advanced in two directions. First, we can use validated answers as the starting point for another round of questions: now that we know the term for pen (jail) in a language, what is the plural form? what is the own-language definition (for which we have a game that guides users toward submitting quality answers, described in Benjamin 2015, section 3.4)? Second, we can use crowd data as a launching point for specialist expansion via the Edit Engine; it is much easier to add details to an entry that has been started than to start from scratch. The goal for an individual member of the crowd is not, however, a perfect and complete dictionary. Rather, the goal is a complete data element that can be published and used when it passes consensus. We suggest that many members of the public will be motivated when they start seeing basic terms come into GOLD for their language, and will work toward increasingly complex entries as they become aware of the potential for more in-depth lexicographic production. In this way, we are nervously preparing to open the gateways to lexicography for numerous languages where lexicographers are not available.

References

Benjamin, M. & Biersteker, A. (2001). The Kamusi Project Edit Engine: A New Tool for Collaborative Lexicography, *Journal of African Language Learning and Teaching*, Volume 1, Number 1

Benjamin, M. (2015). Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. *Proceedings of AsiaLex 2015*, Hong Kong

Rechenbach, C. (1968). *Swahili-English Dictionary*. Catholic University of America Press, Washington, D.C.