Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Martin Benjamin[†], **Paula Radetzky**[‡]
†EPFL — Swiss Federal Institute of Technology, Lausanne, Switzerland
‡Kamusi Project International, Geneva, Switzerland
martin.benjamin@epfl.ch, paula@kamusi.org

Abstract

This paper looks at the challenges that the Kamusi Project faces for acquiring open lexical data for less-resourced languages (LRLs), of a range, depth, and quality that can be useful within Human Language Technology (HLT). These challenges include accessing and reforming existing lexicons into interoperable data, recruiting language specialists and citizen linguists, and obtaining large volumes of quality input from the crowd. We introduce our crowdsourcing model, specifically (1) motivating participation using a "play to pay" system, games, social rewards, and material prizes; (2) steering the crowd to contribute structured and reliable data via targeted questions; and (3) evaluating participants' input through crowd validation and statistical analysis to ensure that only trustworthy material is incorporated into Kamusi's master database. We discuss the mobile application Kamusi has developed for crowd participation that elicits high-quality structured data directly from each language's speakers through narrow questions that can be answered with a minimum of time and effort. Through the integration of existing lexicons, expert input, and innovative methods of acquiring knowledge from the crowd, an accurate and reliable multilingual dictionary with a focus on LRLs will grow and become available as a free public resource.

Keywords: multilingual lexicography, less-resourced languages, data mining, validation, crowdsourcing, gamification

1. Introduction

Humans do a remarkable job of transmitting linguistic data from one generation to the next. Not just parents, but entire communities, transfer innumerable lexical elements, including pronunciation, grammar, syntax, and usage information. We have done a remarkably poor job, however, of downloading such data into forms that can be stored and operated on outside of our Cranial Processing Units (CPUs).[1] This paper looks at the challenges that the Kamusi Project faces in acquiring open lexical data for less-resourced languages (LRLs) of a range, depth, and quality that can be useful within Human Language Technology (HLT).[2] These challenges include accessing and reforming existing data sets into interoperable data, recruiting language specialists to work with new and existing data, locating and working with non-specialist speakers, and funding the requisite work. We lay out the issues facing data collection for LRLs, then look in particular at a crowdsourcing schema, including our mobile phone application, which we have developed to elicit high-quality structured data directly from each language's speakers.

[1] Using your brain, you understood the wordplay with CPU almost immediately. It is unlikely that today's best artificial intelligence could decode the linguistic subtleties embedded in the pun.

[2] The Kamusi Project began as *The Internet Living Swahili Dictionary* at Yale University in 1994. In 2007, the project spun off as an independent non-governmental organization dedicated to the production of language knowledge resources. Kamusi Project USA is registered in Delaware as a 501(c)(3) non-profit corporation, and Kamusi Project International enjoys the equivalent status in Geneva, Switzerland. As of 2013, the informatics aspects of the project are housed at EPFL, the Swiss Federal Institute of Technology in Lausanne.

2. Acquiring Lexical Data for LRLs

Even for well-resourced languages, much recorded lexical data is neither available nor codified in a deeply interoperable form; for example, from the *Oxford English Dictionary* on, no source of English lexical data has been at once open, reliable, well-structured, and richly elaborated. LRLs are even less likely to have comprehensive lexical data. Most LRL dictionaries are small and basic, with few terms and little information beyond a part of speech and a major-language gloss. Exceptions exist in the form of print tomes researched over many decades (e.g., Young & Morgan, 1987; Matisoff, 1988; Hill et al., 1998; Coupez et al., 2005; Cole & Moncho-Warren, 2012), but most such works are not available in machine-usable format, nor are they economically accessible to most LRL speakers. Furthermore, the lexical data published within the past seventy years that has been digitized for LRLs is generally copyrighted, and if the owners can be located, they are often reluctant to share.

In the effort to create a massively multilingual online dictionary, the Kamusi Project has established a system that can accommodate an unlimited amount of lexicographic data within a single, consistent data structure (Benjamin, 2014). The system is designed around the production of monolingual dictionaries for each language, interlinked to other languages at the level of the concept. With each concept in a particular language treated as an individual entity, we are able to elaborate associated data that can be used for natural language processing, machine translation, and other HLTs. Any feature of a particular language, such as the numbers and types of possible morphemes and inflections for each part of speech, alternate scripts, or tone spellings, can be

handled by the project architecture. Over time, each monolingual entry can come to contain a large amount of rich structured data, including intra-language relations, etymologies, examples, and geo-tagged pronunciations and sightings, as well as unstructured information such as usage and cultural notes. Once a monolingual entry has been created, it can be linked to a concept in another language, with a degree of equivalence specified as parallel, similar, or explanatory. Kamusi then shows the train of transitive links from the second language, marking degrees of separation. In this way, each language develops as a full internal monolingual resource that is simultaneously a multilingual communications gateway to every other language in the system. Once an entry is approved into the system, it becomes part of an open-access data set that is available to the public and to machines through a raft of emerging technological tools for online, mobile, and offline use.

Data for Kamusi comes from three types of sources: (1) existing data sets; (2) direct input from language specialists; and (3) controlled input from the crowd. There is substantial interplay among these categories (Nguyen et al., 2013); imported data may be used as part of the process of validating crowd submissions, experts may approve or revise imported or crowd data, and the crowd helps validate imported data and adds details such as pronunciations, examples, and images to entries produced by specialists. A major method for eliciting entries from specialists and the crowd is via reference to a prioritized list of concepts derived from English, using data from both corpus analysis and topical word lists (Benjamin, 2013). Using English as a starting point can be methodologically problematic and is being addressed by ongoing programming, but it is not possible to use corpus approaches to generate wordlists for many LRLs due to a paucity or absence of digitized written material.[3] For languages with a written record, corpus-based lexicon development can occur when a team is in place that can take on the intensive tasks of assembling the records or gaining copyright permissions to an existing corpus; future plans include tools to harvest lexical data from online sources and, when users grant permission, from translation services that interact with Kamusi. In the near term, however, the English-based list gives us a starting point that enables the rapid growth of lexicons that bring together many languages, with the challenges discussed in the following sections.

2.1 Existing Data Sets, Incommensurate Data, and Intellectual Property

Existing data sets offer substantial benefits, but also considerable challenges, to the multilingual dictionary project. The benefits of bootstrapping the project with data that has already been researched and digitized go beyond the obvious savings of time and effort. Much invaluable work currently languishes in isolation, whether in a field researcher's shoebox, a print volume on a library shelf, or even a web page devoted to an individual LRL. The multilingual dictionary provides a central home where all such data can be readily located, and a platform to link the work produced as open data for one language to a great deal more work on the same and other languages (potentially including non-lexical data, such as items in the ELAR and PARADISEC archives),[4] thereby augmenting the utility of previous accomplishments. Lexicography can be the labor of years, often in remote field settings, producing data that cannot be replicated and should not be lost. In many cases, dictionaries from decades past are historical documents that preserve language data prior to contemporary influences such as migration and assimilationist language policies. Preserving data, making it accessible, multiplying the power of what can be done with it, and accelerating the inclusion of LRLs in the multilingual framework are all advantages conferred by mining previous lexicons.

The challenges of existing data, however, are manifold. The Kamusi Project is refining a system for merging existing data sets into our structure—but perhaps "data sets" is a poor description of what is available. Traditionally, the author of a dictionary determines which elements to include, in what format, and in which order for their publication. As Haspelmath (2014) points out, individual dictionaries for LRLs are not readily comparable even for side-by-side perusal. Many entries are composed as undifferentiated text blocks, often without a consistent structure from one line to the next. For example, this is an entry from a Swahili-Mandarin data set that is currently being prepared for incorporation into Kamusi, with evident difference between the type of data that comes after the 1 and the 2: "**-amba I** *kt* 1. 说某人的坏话，议论某人 *Usikae bure na kuamba watu.* 你别干坐着说别人坏话。 2. <旧> 说." Determining what the fields are, and converting scanned or text-file dictionary blocks into data that can be categorized in a database, can itself be an enormous undertaking. Furthermore, most dictionaries group polysemous items together under a single headword, while Kamusi's multilingual structure requires each sense to be disaggregated concept by concept, polyseme by polyseme. Prior to merging, many data sets demand a tremendous amount of manipulation, much of which cannot be automated (see Hernández & Stolfo, 1998; Lee et al., 1999; Dong & Naumann, 2009). For instance, in the Swahili-Mandarin case, we have been able to isolate and segment the data (more than 10,000 entries) into individual data points, but not distinguish automatically between glosses, example sentences, and special usage explanations. The lexicographer is left with the task of manually shifting the Mandarin elements to their correct fields within a spreadsheet prior to importing to the online system.

[3] To address these issues, Kamusi is developing a system of "balloons" to levitate concepts that are important in languages related by linguistic family, geography, or cultural features (Benjamin & Radetzky, under review).

[4] http://www.elar-archive.org/index.php and http://www.paradisec.org.au/home.html

Once a data set is ready to be merged, each entry must be reviewed individually. Even in the best cases, when data has been curated using software such as TLex[5] or SIL's Toolbox or FieldWorks[6] and therefore does not need cleansing, it remains impossible to align senses without a human eye. It is not enough to know that a particular term has been defined, for example, by English *light,* which has a great number of homophones. Without disambiguation of the specific sense ('not heavy', 'not dark', 'not serious', etc.), the entry cannot be brought into the multilingual system. The merging engine, still under development, will display the definitions of possible matches to another language of the user's choice, not necessarily English, or offer the option to add a new matching sense in a linking language. This process requires humans who know the language well, whether an expert working for love or money, or a large enough number of crowd members to produce a reliable consensus.

After merging, the data may still be inadequate for the requirements of the multilingual dictionary; in particular, most data sources do not include own-language definitions needed to build the monolingual core for each language. Additionally, most bilingual data sets, which constitute the bulk of existing data for LRLs, include terms that do not yet have translations in Kamusi, so a provisional sense indication in a language already in the system is necessary in order to prevent those terms from hiding as orphans outside of the multilingual framework.

Beyond the technical challenges lie issues of intellectual property. In some cases, ownership of the data cannot be determined. For example, Sacleux (1939) was written by a priest who died in 1943 without heirs. Neither his religious order nor the successor to the museum that published his dictionary wished to prevent use of the data, but neither would take responsibility for authorizing its release. A decade after first attempting to obtain permission, the data is finally in the public domain as of this year (2014). Researching the ownership trail of each LRL data source and then writing letters and awaiting responses, or waiting until seventy years after the death of the author, all to secure permission for works that must then be scanned, cleaned, and converted from text to data, is not a winning strategy for data acquisition.

Even when copyright ownership is clear, acquiring usage rights can be difficult. Publishers do not easily relinquish data that they have obtained under contract, even for an out-of-print work in a small-market language. When publishers are willing some LRL lexicographers (or the organizations they are affiliated with) do not want to share their product. After decades compiling the definitive reference work for a particular LRL, many authors wish to keep rights to hypothetical royalties and retain control over how the data will be presented. Conversations can stretch for months and then break down when the author places an untenable condition on the release of the work, such as the ability by the author to remove data after it has already been merged into the system, or a copyright license different from the Creative Commons Attribution Non-Commercial Share Alike license[7] that has been established for data within the larger Kamusi Project.[8]

It is hoped that authors and organizations will become more interested in sharing their data as the Kamusi multilingual dictionary, maintained by a non-governmental organization with a charter to produce language resources to be shared with the public for free in perpetuity, grows and is able to demonstrate the advantages that joining the project can bring to a language community. For example, work to integrate more than one hundred LRL lexicons developed by the US Peace Corps is expected to begin when the merging engine is complete, after optimizing our mobile app (discussed in §3.2) for low-bandwidth African telecommunications. Again, however, securing the blessing to use existing data only brings it to the point where it must face the technical challenges discussed above.

2.2 Language Specialists

The ideal way to collect lexical data is to have language specialists contribute rich data for every entry, using a comprehensive online edit engine constructed with standard web form elements customized for each language. Such contributions can be considered authoritative (Kleinberg, 1999) and can provide the full range of information needed for the term to be understood by humans and manipulated by HLTs. Specialists can work from the above-mentioned list of concepts derived from English, or they can use another reference language as in the Swahili-Mandarin case above, or bring in terms that are unique to their LRL (Bergenholtz & Nielsen, 2013).[9] The specialists add depth and nuance that cannot come from existing static data and might not be elicited from the crowd. However, working with experts is not without its challenges.

The first problem is identifying people to work on a language. The world's leading authority on a given language may not be the person to bring it into a multilingual dictionary. To begin with, the person may have already published a dictionary that is encumbered by copyright or that they do not wish to share. Additionally, such experts are often academics tied up with other research and teaching. Furthermore, in contrast to books and articles, dictionaries do not weigh highly in tenure and promotion considerations, and participation in a col-

[5] http://tshwanedje.com/tshwanelex/

[6] http://www-01.sil.org/computing/toolbox/ and http://fieldworks.sil.org

[7] http://creativecommons.org/licenses/by-nc-sa/3.0/

[8] Source code is not currently open because we do not have enough staff resources to vet incoming contributions, and it is problematic to release code that would lead to other versions of what must function as a unified project. The code base will be opened when the project has the staff capacity to manage externally-developed programming components.

[9] Kamusi's revised approach to the methodological difficulties of starting with a concept list keyed to English is addressed in footnote 3.

laborative project with indeterminate authorship contributes even less to a CV. Sometimes the leading expert is best equipped to offer guidance and perhaps references to people with the time to do the work.

In addition, knowledge of a language does not necessarily imply the ability to document it within the Kamusi framework. Lexicography is a complicated endeavor to begin with, and Kamusi's multilingual model adds new complexities in the pursuit of creating a detailed matrix of the human linguistic experience. While the current project is built on an editing input system that strives to be clear and user-friendly, aspects remain difficult or non-intuitive. Training is necessary so that contributors, even PhDs with experience in lexicography, can understand the purpose of each field and the formats required for the data to be useable and consistent. It is especially difficult, and particularly important, to teach participants how to write good own-language definitions. Before contributors can be given moderator privileges to confirm data as finalized, they must go through a period of training and observation to determine that they understand the technical and philosophical aspects of producing high-quality data.

It is possible to find volunteer participants who are both interested in, and capable of, rigorous lexicographic work; however, expert contributors are more likely gotten with remuneration. Producing a high-quality entry, including an own-language definition, takes five minutes or more. At that speed, ten thousand entries is a year of labor. Few people have a year or more to donate to their language. Although a volunteer might start out with the best of intentions, financial incentives are a more reliable way of ensuring that the work is accomplished (Bederson & Quinn, 2011). A system is under design to pay experts per lexical term, although, ironically, we have not yet been able to fund the coding through to implementation. Quality control will be a challenge because project management cannot possibly know all the languages in which data is supplied, so this is integrated into the crowdsourcing elements discussed below.

The largest hurdle with language specialists, then, is funding. The costs are not especially high per term, and become infinitesimal when extrapolated to clicks over time, but they are a substantial up-front obstacle when the number of words in a language is multiplied by the number of languages worldwide. Funders have many priorities, among which language resources generally rank low. The Kamusi Project has internal task forces to find funds for particular languages or regions and welcomes all suggestions.

Language specialists are being recruited from a variety of institutions, with more than twenty institutions represented in the multilingual pilot phase completed in February 2013. The invitation is open to academics who are actively working on projects for their languages, or who wish to develop a joint proposal to take advantage of the resources that the Kamusi Project offers. We also solicit citizen linguists, that is, people who are both passionate about their language and have the time and skills to invest. (These citizen linguists using the expansive edit engine are not the same as "the crowd" using the constricted app, discussed in §3.2 below.) One of our models is DEX Online, a monolingual Romanian dictionary, which has built a compelling resource with much volunteer labor from Romanian Scrabble players.[10] Retirees with computer skills and spare time are another community that might be tapped for particular languages, providing a stimulating activity in support of a cultural legacy. In terms of remunerated efforts, the Kamusi Project is currently using NEH grant funds to provide student stipends and training at the University of Ngozi in Burundi, in exchange for data development in the Kirundi language. A related method well-suited for LRLs would be grant support for graduate field researchers. More expensive, but benefiting from contracts and enforceable expectations, is the possibility of working with professional translators. In all cases, the challenge is to match people who can do the work with an appropriate reward for getting it done well.

2.3 Crowdsourced Data Collection

For many languages, reliance on language specialists will be too slow to generate useful data, even if a specialist can be located. Furthermore, specialists do not know and do not have the ability to provide every detail of each word in their language. In fact, certain data elements such as regional pronunciation recordings can only come from a wide assortment of contributors. In order to speed progress and provide greater depth and range, techniques are under development to generate linguistic data from the crowd, as discussed below in §3.2. However, crowd-generated data is notoriously unreliable, so the system is being designed with numerous redundancy and reliability checks. Crowd data must always be subject to rigorous validation procedures, labeled for provenance, and be editable by specialists.

Wiktionary provides a case study in the dangers of crowdsourcing a dictionary. The project is to be commended for seeking a fantastic range of linguistic data. Yet, the open architecture invites mischief and mistakes, and inhibits error-checking. For example, as of this writing, a spam English definition of *spring* as 'erection' has persisted in various forms since being added by an anonymous user in 2006. Definitions are sometimes circular, or one-word synonyms. It is simple to add erroneous translations, which then propagate bad automated data throughout the system. The majority of elements are written in wiki markup language, which is a near-impenetrable barrier to most people's participation. While Wiktionary continues to improve, its laissez-faire approach to crowdsourcing leaves it inconsistent and unreliable as a source for lexical information. As a worst-case example, the Malagasy Wiktionary contains an ever-expanding collection, three million pages and counting, of robot-generated gibberish that the organization has been unable to limit or expurgate (Andrianja-

[10] http://dexonline.ro

nahary, 2013).

Crowdsourcing involves these and several other issues, enumerated here. First, most users prefer to receive information rather than contribute their own knowledge. Second, channeling users to contribute specific types of data requires a well-developed process flow. Third, users can introduce inadvertent errors, such as spelling or formatting problems. Fourth, complex tasks such as writing definitions require training and are not suitable for all users. Fifth, malicious users can intentionally introduce bad data. Sixth, even well-intentioned users can introduce data that turns out to be wrong. Seventh, finding a crowd that is large enough to support the redundancy necessary for validation is difficult for many LRLs, especially those with few speakers or poor communications infrastructure. Eighth, the enthusiasm of individual members of the crowd will be difficult to maintain over the years it takes to collect tens of thousands of terms for a language. With the proper methodology and safety checks in place, however, the crowd can become an important source of data for hundreds of languages. In §3, we present our crowdsourcing model to address these issues.

3. A Preliminary Crowdsourcing Model for LRLs

LRLs face a special challenge: With few existing resources, most LRL Internet users do not expect to encounter their own language, nor do they have a history of participating in its resource development. The crowdsourcing model we are developing is designed to change that by making lexicon development fun, easy, and rewarding. Here, crowdsourcing denotes the completion of specific targeted tasks, as distinct from making use of the in-depth editing system that is anticipated to be mostly for citizen linguists and language specialists.

3.1 Motivating Crowd Member Participation

The first incentive of the system will be to channel users to register for the site. Users will have two options, registering for free access or paying a subscription fee. Free access will come with an asterisk—people can earn usage points by answering questions. This "play to pay" system will give users points for proposing translations, writing definitions, or providing other data such as usage examples culled from online sources. Points can be exchanged for dictionary lookups, and high earners may also win material prizes or rewards that appeal to the psyche, such as congratulatory postings on Facebook. Start-up points will be awarded for registering and providing initial survey data, including language experience.

Points will also be awarded for participating in games (Castellote et al., 2013; Paraschakis, 2013; Hamari, Koivisto, & Sarsa, 2014). One game will be a word race, where an English term and definition will be sent to the players, who will be competing both individually and as part of a same-language team against those working on other languages. When players receive the term and definition, they will send back a translation of that term in their language. When ten answers agree, the person who sent in that answer first will get ten points, the next will get nine, etc. Additionally, all the members of the same-language team will get points based on the order in which their language has completed this task (and slower teams will be given an easy form to recruit more members). Another game will then put the term out for own-language definitions, which will be voted on, with points awarded to the winning author and the people who voted on that definition. Similar gamification will be designed to flesh out other data elements. These games will evolve from the logic of the mobile application discussed below.

Motivation will also be stimulated through social rewards. Users who contribute winning definitions will have their accomplishments posted on their favorite social media (Antin & Shaw, 2012). They will also appear on leader boards on the site, with rankings shown within a language and among languages.

Finally, when we can find sponsors to cover the costs, material prizes such as shirts and clocks will be periodically awarded to the winners of specific limited-time competitions. Competitions for these prizes will often focus on quickly augmenting lexicons for new LRLs as they join the multilingual dictionary. This combination of motivations will be experimented with and and successful approaches expanded, in order to stimulate as much participation as possible.

3.2 Steering the Crowd with Targeted Questions

The researchers at the Kamusi Project have developed a mobile phone application with targeted questions that direct users to provide data in exactly the format required. With the working name "Fidget Widget," the app is envisioned to be used by language aficionados in the small moments when they look to their phones to fill time. The app is in testing as of this writing, with the expectation that it will be demonstrated for LRLs at the May 2014 LREC workshop, "CCURL 2014 – Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era." Increased functionalities will be added to the app over the course of time, including additional types of questions and features for field researchers to collect oral data for talking dictionaries (Benjamin & Radetzky, 2014).

Initial testing of the app will provide data that we can use for determining the thresholds at which we accept crowd data as good enough to either queue for expert moderation (lower threshold) or publish as world-viewable, validated data (higher threshold). While the relative costs of majority decisions versus control groups in crowds have been modeled (Hirth, Hoßfeld, & Tran-Gia, 2013), a numerical standard does not yet exist for statistically determining the point at which different types of crowdsourced data can be considered valid. We expect experiments will show crowd validation can accurately indicate that an item is either very good or very bad, but

that ambiguous evaluation results from the crowd will be useful mostly to indicate entries to be queued for specialist review.

For the initial version of the app, we are interested in two types of information: (1) What is the target language equivalent of a defined source language concept? (2) What is the definition of the term in its own language (i.e., its own-language definition)? We are consciously postponing using crowdsourcing to address lexicographic questions that require subtle understanding of complex ideas, such as the degree of equivalence between the source and target term—even such basic questions as the part of speech of terms proposed by the crowd might be better left to specialist review.[11] However, we are interested in seeing whether this method yields independently-generated own-language definitions of target terms—ones that will allow readers to understand, for example, the subtle differences between *connaissance* and *savoir* in French—or whether crowd definitions tend to be close translations of the source definition, in this case the definition we provide for *knowledge,* which would not be fine-grained enough to distinguish *connaissance* from *savoir*. (See also Haviland (2006) and Svensén (2009) for a discussion of such issues.) The data generated by the app in the current stage is intended to provide a starting point for richer dictionary entries and deeper lexicons that will be expanded later.

To find a target equivalent of source language term, we first ask an open-ended question to several users. The ideal crowd member is a native speaker who is also comfortable in the source language, but people who have acquired the target language later in life cannot be excluded, on the premises that (1) language learners have much to offer based on the concerted efforts they often make to master concepts that native speakers might never notice, and (2) errors will be weeded out by the bulk of the crowd and by contribution analysis. We present the source language term and definition (e.g., *light* 'low in weight') and ask, "What word would you use in [your language]?"[12] If we receive a critical mass of identical answers (we have not yet defined the precise number), then the term will be advanced to the next level of moderation or crowd review. However, if we obtain differing responses to the same question, we next show another set of users the source term (here, *light*) and definition ('low in weight'), and ask, "Is [term] a good translation in [your language]?" For this question, counting thumbs up or thumbs down will allow us to evaluate whether a submission is a near-synonym or a mistake.

After a translation term passes the validation threshold, we seek the target language definition by displaying the original term *(light)*, its definition ('low in weight'), and the target language term, and ask, "Can you give a definition in [your language]?" It is important to show the source original in order to ensure that we do not get a definition for a homophonous term in the target language. (Before writing a definition, each user sees a screen that explains the basic lexicographic requirements, with the choice to opt out of the task.) After receiving the first submission, we display the term and proposed definition to other users and pose the question, "Is this a good definition?" If subsequent users approve, then we advance the definition to the moderator or to validated status. However, if other members of the crowd are dissatisfied, then we solicit the definition anew. When we have two definitions in competition, we show both and ask, "Which definition is better?" Through a formula that will be established when test data is available for evaluation, a winning definition will be advanced to moderation or validated status after reaching a certain level of satisfaction among the crowd.

In the future, many questions for the app and games will be generated by information arriving from existing data sets. For example, if an imported bilingual lexicon indicates that a word in a user's language matches a word that has multiple English senses, the user will be asked to choose the most appropriate sense or suggest a new one. Once enough users have agreed on a sense disambiguation for imported data, the system will steer toward adding definitions, morpheme information, and other elements to fill out the entry. Other questions will seek to group and rank search results that yield multiple entries. On the premise that many crowd members will use the app in short bursts (for example, to answer a single question in order to unlock a device from idle), the questions will be designed to elicit either very short text answers, evaluations of mined data, or judgments about other users' contributions through yes/no or X vs. Y questions. As the system grows, it will be possible to expand questions to demonstrated user interests—for example, asking about other terms in a semantic field that a user has accessed in their current session. Tailoring questions will require some experimentation to discern what strategies are effective (Bergenholtz & Johnsen, 2013).

3.3 Contribution Analysis

Central to the crowdsourcing model will be the analysis of user contributions. It is important to know which users provide consistently good answers versus who comes in wide of the mark. Good answers are those that climb toward a consensus opinion. Bad answers are those that are severely rejected by the crowd. Some answers may be ambiguous—for example, if contributors propose essentially synonymous translations for the same term. (In the model, competing answers that both gain upvotes have equal opportunity to move toward incorporation into Kamusi, with the more popular answer winning primacy in the display hierarchy.) Users who consistently produce good answers will earn trust; trust levels will

[11] While Kamusi has a simple method for matching concepts represented by different parts of speech, such as linking the Swahili verb *–furahi* with the English adjective *happy* via the translation bridge 'be happy', this nuance is not obvious to untrained users. Parts of speech are given provisionally based on the source language, but flagged as questionable until confirmed by a moderator.

[12] In principle, all questions will be localized to the target language.

be displayed on site and optionally on a user's chosen social media. These participants will have their votes on other users' contributions weighted more heavily, and they will have their answers integrated more quickly: their submissions will require fewer votes for validation. On the high end, trusted users will earn the right to moderate contributions that correspond to their demonstrated skill sets, gaining the rank of language specialists with the authority to finalize data as valid for incorporation into the master database.

Conversely, users who consistently score poorly will be diverted to questions that more closely match their skill sets. Easier questions might include evaluation of illustrative photos for appropriateness; voting on whether other users' comments are useful or spam; or recording the pronunciation of a word in their native language. The objective will be to find a level for each user at which they provide useful data and feel comfortable. Having multiple users effectively scoring each other's contributions will result in error checking that builds in good data, weeds out the bad, and creates incentives for users to submit their best possible answers.

Some users are intentionally malicious, and refinements to Kamusi's crowd system are on the drawing board to ferret out these out. Intentional subversion of the system is expected to be less than in previously-studied crowd situations, where paid contributors benefited financially by quickly submitting sloppy work (Kittur, Chi, & Suh, 2008; Suri, Goldstein, & Mason, 2011). However, our ongoing battle against spam registrations and discussion posts shows that some maliciousness is inevitable. In addition to normal channels to flag suspect submissions, including wrong data submitted in good faith, analysis of crowd responses will alert moderators to patterns consistent with abuse. Vandalism might sometimes be difficult to detect because malicious users can mix in valid responses with their spam. They might also attempt to slide their handiwork into obscure concepts in low-volume LRLs, as happens in Wikipedia, or distribute their damage across languages. Algorithms for monitoring ill intent will need to evolve. What is certain is that users who are determined to be vandals will be banished, and all of their submissions will be removed or, if their items have been expanded on subsequently, isolated for further review.

Contribution analysis will require us to keep careful track of the interacting histories of users and entries. This is an informatics challenge rather than a linguistic one, the design of which will be tasked to computer science partners.

4. Conclusions

In order to transfer human linguistic knowledge from people to their machines in a massively multilingual data resource, a number of integrated strategies must be implemented. Existing data sets offer a starting point but require extensive manipulation and human review. Language specialists bring much-needed expertise but can be difficult to locate and engage. Crowd sources have a great diversity of knowledge, but that knowledge is extremely difficult to collect in a systematic and structured fashion. A system to elicit and validate the maximum amount of high-quality linguistic data must therefore combine tools for data import and merging, detailed expert contributions, and regulated crowdsourcing. The Kamusi Project has implemented a web platform and mobile app to address these issues for any language, with refinements constantly in progress. The project is now beginning to use these tools for the collection of reliable data for numerous languages. Through this integrated approach, it will be possible to build in-depth, open lexical data sets and related HLTs for any language, and in particular for currently under-resourced languages where data, specialists, and crowd members can come together in a common resource working toward shared goals.

5. References

Andrianjanahary, R. (2013). My history of the Malagasy Wiktionary. http://terakasorotany.wordpress.com/2013/03/27/my-history-of-the-malagasy-wiktionary.

Antin, J., Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems,* pp. 2925--2934.

Bederson, B., Quinn, A. (2011). Web workers unite! Addressing challenges of online laborers. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems,* pp. 97--106.

Benjamin, M. (2013). How we chose a priority list for dictionary entries. http://kamusi.org/priority-list.

Benjamin, M. (2014). Collaboration in the production of a massively multilingual lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14).*

Benjamin, M., Radetzky, P. (under review). Small languages, big data: Multilingual computational tools and techniques for the lexicography of endangered languages.

Bergenholtz, H., Johnsen, M. (2013). User research in the field of electronic dictionaries: Methods, first results, proposals. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 556--568.

Bergenholtz, H., Nielsen, S. (2013). The treatment of culture-bound items in dictionaries. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 469--481.

Castellote, J.; Huerta, J.; Pescador, J. and Brown, M. (2013). Towns conquer: A gamified application to collect geographical names (vernacular names/ toponyms). In *Proceedings of the 15th AGILE Internation-*

al Conference on Geographic Information Science.

Cole, D., Moncho-Warren, L. (2012). *Macmillan Setswana and English Illustrated Dictionary*. Johannesburg: Macmillan South Africa.

Coupez, A.; Kamanzi, T.; Bizimana, S.; Sematama, G.; Rwabukumba, G.; Ntazinda, C. and collaborators. (2005). *Dictionnaire rwanda-rwanda et rwanda-français / Inkoranya y ikinyarwaanda mu kinyarwaanda nó mu gifaraansá.* Butare, Rwanda: Institut de Recherche Scientifique et Technologique and Tervuren, Belgium: Musée Royal de l'Afrique Centrale.

Dong, X., Naumann, F. (2009). Data fusion: Resolving data conflicts for integration. In *Proceedings of the VLDB Endowment* 2(2), pp. 1654--1655.

Hamari, J.; Koivisto, J. and Sarsa, H. (2014). Does gamification work? – A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*.

Haspelmath, M. (2014). Dictionaries as open-access databases: A vision. http://dlc.hypotheses.org/676.

Haviland, J. (2006). Documenting lexical knowledge. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation*. Berlin: de Gruyter, pp. 129--162.

Hernández, M., Stolfo, S. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1), pp. 9--37.

Hill, K.; Sekauquaptewa, E.; Black, M. and Malotki, E. (1998). *Hopi Dictionary/Hopìikwa Lavàytutuveni: A Hopi Dictionary of the Third Mesa Dialect with an English-Hopi Finder List and a Sketch of Hopi Grammar.* Tucson: University of Arizona Press.

Hirth, M.; Hoßfeld, T. and Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms for-crowdsourcing platforms. *Mathematical and Computer Modelling* 57(11), pp. 2918--2932.

Kittur, A.; Chi, E. and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453--456.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5), pp. 604--632.

Lee, M.; Lu, H.; Ling, T. and Ko, Y. (1999). Cleansing data for mining and warehousing. In T. Bench-Capon, G. Soda, & A.M. Tjoa (Eds.), *Database and Expert Systems Applications: 10th International Conference, DEXA'99 Florence, Italy, August 30 – September 3, 1999 Proceedings.* Berlin: Springer-Verlag, pp. 751--760.

Matisoff, J. (1988). *The Dictionary of Lahu*. Berkeley: University of California Press.

Nguyen, Q.; Nguyen, T.; Miklós, Z. and Aberer, K. (2013). On leveraging crowdsourcing techniques for schema matching networks. In W. Meng, L. Feng, S. Bressan, W. Winiwarter, & W. Song (Eds.), *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, Wuhan, China, April 22-25, 2013, Proceedings, Part II.* Berlin: Springer-Verlag, pp. 139--154.

Paraschakis, D. (2013). *Crowdsourcing Cultural Heritage Metadata Through Social Media Gaming.* Master's Thesis in Computer Science. Malmo University.

Sacleux, C. (1939). *Dictionnaire swahili-français*. Paris: Institut d'Ethnologie.

Suri, S.; Goldstein, D. and Mason, W. (2011). Honesty in an online labor market. In *Human Computation: Papers from the 2011 AAAI Workshop,* pp. 61--66.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Young, R., Morgan, W. (1987). *The Navajo Language: A Grammar and Colloquial Dictionary,* revised ed. Albuquerque: University of New Mexico Press.