Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages

Martin Benjamin
École Polytechnique Fédérale de Lausanne
Lausanne, Switzerland
martin.benjamin@epfl.ch

Paula Radetzky
Kamusi Project International
Geneva, Switzerland
paula@kamusi.org

Abstract

The Kamusi Project, a multilingual online dictionary website, has as one of its goals to document the lexicons of endangered and less-resourced languages (LRLs). Kamusi.org provides a unified platform and repository for this kind of data that is both simple to use and free to researchers and the public. Since Kamusi has a separate entry for each homophone or polyseme, it can be used to produce sophisticated multilingual dictionaries. We have recently been confronting issues inherent in contact language-based lexicography, especially the elicitation of culturally-specific semantic terms, which cannot be obtained through fieldwork purely reliant on a contact language. To address this, we have designed a system of "balloons." Based on a variety of factors, balloons raise the likelihood of revealing terms and fields that have particular relevance within a culture, rather than perpetuating linguistic bias toward the concerns and artifacts of more powerful groups. Kamusi has also developed a smartphone application which can be used for crowdsourcing contributions and validation. It will also be invaluable in gathering oral data from speakers of endangered languages for the production of monolingual talking dictionaries. The first of these projects is planned for the Arrernte language in central Australia.

1 Introduction

The Kamusi Project is a multilingual online dictionary and language-resource website at www.kamusi.org, whose primary purpose is to provide a unified platform designed for documenting the lexicons of the world's languages. The main goal of this effort is a set of monolingual written and audio dictionaries for both large languages and less-resourced ones (LRLs), connected together at the concept level to produce viable bilingual dictionaries between each language in the system, as well as bedrock linguistic data that can be used in advanced machine applications. Linguistic data is contributed by individual researchers and also via crowdsourcing. As a massively multilingual dictionary project, Kamusi has been wrestling with the conceptual challenge of how to elicit terms in a way that minimizes cultural bias but results in lexicons that can be linked between languages. At the same time, we have been developing tools that will enable citizen lexicography without necessarily involving a field researcher. Such tools need to be highly systematic in order to yield usable and trustworthy dictionaries.

In this paper, we first provide an overview of Kamusi (§2); describe "balloons," our system for overcoming the problems of using a contact language to elicit endangered language lexicons (§3); introduce our smartphone application, designed to gather oral data from non-literate speakers and both oral and written data from literate speakers (§4); and, finally, discuss our efforts to produce monolingual talking dictionaries, the first of which involves the Arrernte language (Pama-Nyungan; central Australia) (§5).[1]

2 Kamusi as a Platform for Endangered Language Lexicography

Several technological resources provide good data-gathering solutions for individual lexicographic projects, including Max Planck's LEXUS;[2] TLex;[3] WeSay;[4] and SIL's triad of Lexique Pro, Toolbox, and FLEx.[5] Yet each of these solutions leaves gaps for the individual projects making use of them, and none is suitable for development of sophisticated multilingual dictionaries as envisioned by Kamusi. The learning curve can be steep, particularly the initial effort to set up an effective structure for a language. Each project must reinvent the entire process of bilingual translation, choosing which contact language terms to treat, working out anew how to reference different senses, and coping with or ignoring non-equivalence between languages. The comparison between two languages in different projects is impractical or impossible, even if the two dictionaries share one of their languages. For example, using Lexique Pro to find terms in Bakwé and cross-border Bambara (both Niger-Congo; Côte d'Ivoire) that correspond to English *light* 'illumination' is a Herculean research task. One must visit the multiple entries glossed as 'light' in each dictionary, then compare the Bakwé with Bambara words to try to discern which definition or term matches with which.[6] The dissemination of data becomes an exercise in reinventing multiple wheels: creating a website and finding hosting or using the limited services of Lexique Pro, publicizing the data's availability, finding a publisher who is interested in a language without a market. More extensive ambitions, such as mobile applications or ongoing expansion of the lexicon, are unlikely to be addressed for underfunded LRLs.

The Kamusi Project speaks to each of these gaps. Anyone who is able to purchase an airline ticket online has the technical skills to use the editing system, although some concepts (such as the difference between a definition, a translation, and a definition translation) must be mastered with the aid of tutorials.[7] Setting up a language involves a few hours of back-and-forth with Kamusi staff to configure the parts of speech and the fields for inflections and attributes that vary from language to language. The editing system handles all of the data fields that have been identified for the thirty-odd languages currently configured for the system, with the possibility of adding more data categories if necessary. Lexicon development can proceed via elicitation from an English priority list (§3 below), or directly from deeper lexical research. There is little ambiguity about translation senses; each English sense of *light* (homophones and polysemes) is its own entry with a clear definition, as is each German *Licht*, each Mandarin or Urdu homophone and polyseme, and so on. Equivalence between languages is shown by labeling translations as either parallel, similar, or an explanation in language B of a term in language A (or vice versa). When a concept in language A is linked to a term in language B, the links from language B to other languages are carefully tracked, along with the degree of separation; in this way, were a Bambara term and a Bakwé term both linked to a particular English sense of *light*, they would inherently be shown as second degree links to each other, with the possibility to validate or reject the computer pairing. Each piece of data is published immediately upon validation, so there is no need for the lexicographer to spend time setting up a website, find hosting and pay for it indefinitely, update files, manage a server, attempt search engine optimization, etc. Each language will share access to new tools and resources as they are rolled out on Kamusi.org, such as custom printing, integration with social media, and mobile apps and other improved methods for collecting linguistic data from community members (§3-§5 below).

3 Balloons: Addressing Problems of Contact Language-Based Elicitation

It is a trope in the field-linguistics world that LRLs, especially those that are spoken by small-

[1] Our app, described in §4, will be demonstrated at the present meeting, the ComputEL Workshop of the Association for Computational Linguistics, June 2014.

[2] http://tla.mpi.nl/tools/tla-tools/lexus/release-notes/

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

[4] http://wesay.palaso.org

[5] http://www-01.sil.org/computing/catalog/show_software_catalog.asp?by=cat&name=Data+Management

[6] http://www.bambara.org/lexique/index-english/main.htm; http://bakwe.org/e107_files/LexiquePro/bakwe_lexicon/index-english/main.htm

[7] Rather than suffer through a dry description of the editing process, registered users are invited to click "Edit this entry" on any entry where they see opportunities for improvement at http://kamusi.org, or add new terms or senses through the form at http://kamusi.org/node/add/dictionary-term.

er linguistic minorities, pose special challenges for efforts at documentation. These include scarcity of speakers and researchers, remoteness of field sites, lack of funding, and academic evaluation systems in the humanities and social sciences which reward only certain kinds of investigation—to the exclusion of, notably, lexicographic research, linguistic resource- and website-building, and any sort of research product that is the result of a significant number of participants or community-based input.[8]

Due to the scarcity of speakers and researchers (and especially native-speaker researchers) of endangered languages, the process of lexicographic documentation for such languages almost always begins with elicitation of terms from a major contact language—English, Spanish, Thai, Swahili, etc.—with or without a tool such as a word list.[9] Definitions or, more often, translation equivalents are then recorded in the major contact language as well. It is rare to find dictionaries with own-language definitions for endangered or small minority languages.[10]

There are, however, several problems that are inherent in using a major contact language as the starting point for eliciting LRL lexical items. One problem is that it inhibits the discovery of terms and entire semantic fields which exist in the field language but not in the contact language. In a sense, this is akin to an archaeologist using a metal detector—the technology will reveal iron objects, but ceramic artifacts will remain hidden. Another issue is the cultural imperialism of an approach that privileges the concepts and categories that are important to politically-, religiously-, and economically-dominant sociolinguistic groups. (For a discussion of these and other issues relating to contact language-based elicitation, see Calvet (1974), Raison-Jourde (1977), Fabian (1983), Geeraerts et al. (1994), Errington (2001), Anderson (2003), Enfield (2003), Bowern (2010), and Mosel (2011), among others.) Below, we describe how Kamusi is using a device we call "balloons," so that contributors can avoid these pitfalls and expedite the production of a dictionary with terms derived as much as possible from the local lexicon.

Our springboard into lexicographic elicitation is a prioritized list of English concepts that combines corpus results together with other term sets with particular foci, such as the Comparative African Word List[11] and the basic Special English vocabulary list of the Voice of America.[12] Our master list has some drawbacks, however (Benjamin, 2013). As a starting point for endangered languages, many highly-ranked terms are indisputably useful: *wind, bird, dry.* Other terms, however, do not exist in these languages, nor do their speakers have much need of referencing them: *baseball, subway, century.* The advantages of a cross-cutting, English-biased concept list certainly outweigh a haphazard butterfly-collection approach, but rigid adherence to such a list would foist irrelevant terms on a language documentation team while simultaneously causing them to miss many concepts of local importance.

To rectify the weaknesses of the English-centric approach, we have designed a system of "balloons" to prioritize terms more relevant to a particular language. The simplest balloons attach to the overall number of languages in which a particular concept has been submitted.[13] In addition, balloons provide lift in one language for terms deemed important by contributors in other languages related in some manner—for instance, balloons can attach based on geography, language tree proximity, shared cultural spheres, or other aspects of affinity. When contributors are fed a list of lexical items to elicit, balloons levitate certain terms to higher positions on the list, based on a variety of factors selected by the language moderator or individual contributor. A team working on a river language of Cameroon, for example, could set balloons to raise terms that have been treated by other Cameroonian

[8] Although the hard sciences (including computer science) value collaborative resource-building, the traditional role of the lone-wolf researcher persists as a powerful image among linguists (see Crippen and Robinson (2013) and also the rest of the ink spilled against this ideal in the journal *Language Documentation and Conservation*).

[9] An exception to the wordlist method is the Dictionary Development Process (DDP, http://www-01.sil.org/computing/ddp/) developed by Ron Moe at SIL, which steps away from wordlists to focus on semantic domains.

[10] Some exceptions are monolingual dictionaries of K'ichee' [Quiché] (Mayan; Guatemala) (Ajpacajá Túm, 2001) and Yiddish (Joffe & Marq, 1961-1980). The latter was abandoned after the publication of four volumes, all devoted to the letter *alef.*

[11] https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf/Snider_silewp2006-005.pdf

[12] http://www.manythings.org/voa/words.htm

[13] Features are under development at the time of writing that are expected to be completed for showcasing at ComputEL in June 2014. However, software delivery schedules are notoriously slippery, particularly in a non-profit environment, so features such as balloons for related cultural characteristics may remain temporarily promissory.

languages, by related Bantu languages, or by other groups with a fishing economy.

Central to the mechanism of balloons is that contributors always have the option of skipping on the priority list terms that they do not know or do not deem important. For example, they could provide a term equivalent to English *plant* as a living organism but skip the homophonous *plant* referring to an industrial processing facility. The vegetal sense of *plant* would then float upward as more languages validate its importance, while the industrial sense would linger in the depths.

Languages do not enter Kamusi only when a contributor adds terms by working through a priority list; terms from other languages can become incorporated via the merging of existing lexical data sets. A team from one language could use balloons to find concepts that exist in related languages already in Kamusi, such as terms glossed with explanatory translations. For example, if the Bakwé data set is merged into Kamusi, *-srüpö* 'rolled up dead leaves or cloth, used to cushion the carrying of loads on the head or shoulder' would become available to Bambara and other languages of the region, and the concept would rise in importance as participants around Africa recognized the item and provided their equivalent term. Kamusi's system of balloons, then, ensures that the concept base available to a given language will include many items and semantic fields that would not otherwise come to light.

While development of the balloons system will still be a work in progress at the time of the workshop for which this paper is a contribution, and the task of choosing categories for balloons and the amount of lift they provide will involve ongoing adjustments in response to testing with field lexicographers, we nevertheless want to highlight it as a method for overcoming certain aspects of bias in the selection of vocabulary in a multilingual dictionary. In particular, it is proposed that this method will tend to float concepts that are most universal, while also encouraging the development of vocabularies that have special cultural relevance. We acknowledge, however, that this approach will not elicit concepts that are unique to a culture and are therefore not represented in either the English priority list or the lists that we incorporate from other sources; for a fine-grained investigation of local concepts, there can be no replacement for researcher-directed field study. Kamusi.org has other established tools for adding such indigenous terms, as many as lexicographers can catch in their nets.[14] One of these tools is our smartphone application, discussed in the following section.

4 The Smartphone App: Rapid Elicitation and Validation from the Crowd

The Kamusi Project began as an online bilingual dictionary between English and Swahili. The experience of building a resource for Swahili led to an expansion of the system to other languages, with the technical capacity to document the full lexical scope of any language. One of Kamusi's objectives is to move beyond lists of translations between languages by creating monolingual dictionaries with own-language written and/or spoken definitions for each lexical item. In conjunction with this, we have developed a range of tools designed to support online collection of sophisticated data.[15]

The Big Data ambitions of this project rely on numerous inputs of very small data, most of which must come directly from a language's speakers (including through fieldwork), rather than from digitized data sets.[16] For reasons discussed in Benjamin and Radetzky (2014), relying only on experts using Kamusi's advanced online tools will not be a successful strategy for the expedited production of lexicons for many LRLs. Instead, much data collection will occur through crowdsourcing, using validation procedures to

[14] The only caveat is that a translation link must be provided to English or another contact language in order for the new term to be understandable by people who do not speak the source language, which may necessitate the additional task of creating a new entry on the contact language side.

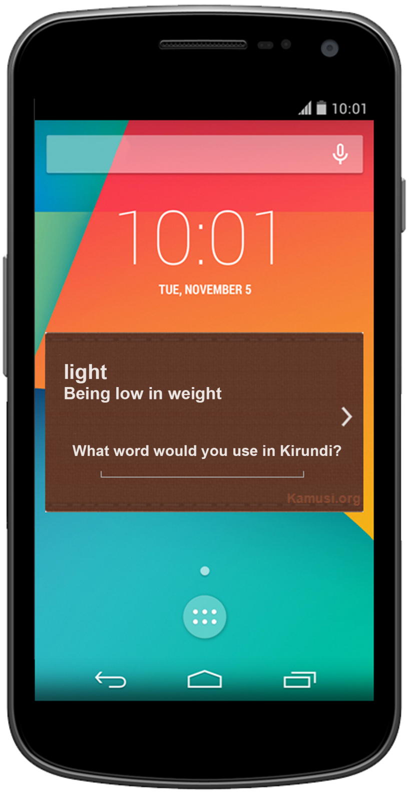
[15] There do exist online projects for baseline documentation, but not actual lexicography, of the vocabularies of endangered languages, such as LEGO (http://lego.linguistlist.org) and PanLex (http://panlex.org), with whom we work collaboratively. To date, these are involved in linking wordlists and are rarely involved in collecting new or rich data. In their disclaimer at http://lego.linguistlist.org/disclaimer, LEGO states, "[W]e are primarily interested in allowing existing lexical data to be included in our datanet and promoting standards to allow others to construct comparable datanets.... [W]e have converted a number of legacy resources..., but we have not engaged in collecting new lexical data...." Regarding PanLex, Kamholz et al. (2014) write: "[P]rojects that are designed to be panlingual tend to have specific and limited objectives... PanLex, with its objective of documenting only the lemmatic forms of lexemes, is no exception."

[16] Each entry is a container for dozens of fine-grained data elements, ranging from inflections to geo-tagged pronunciations to videos, multiplied by tens of thousands of terms in thousands of languages, with complex translational, semantic, and ontological interconnections for every concept.

ensure that the data is reliable prior to its being integrated into the system.

In order to collect millions of pieces of linguistic microdata, we have created a mobile smartphone application, the Kamusi Fidget Widget, that asks users specific, targeted questions about their language.[17] This app gathers data for integration into the project's online multilingual resources, and it is designed for participants who access networks through handheld devices—a major mode of connectivity for many oases of endangered languages.

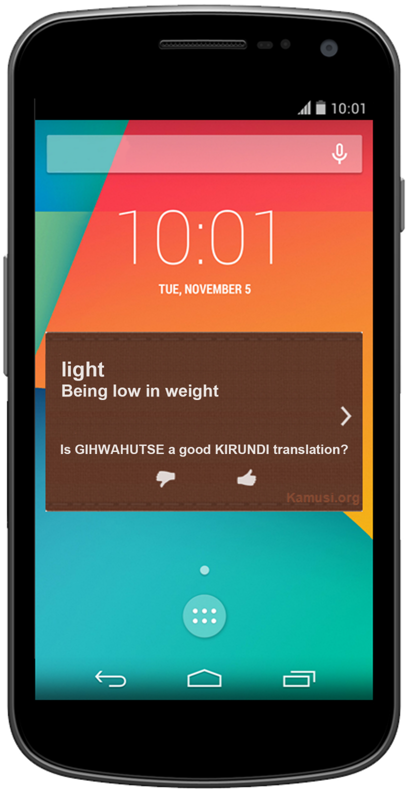
The Fidget Widget pilots a new approach to eliciting terms and definitions that accelerates data collection for LRLs and advances talking dictionaries into monolingually-useful reference resources, while also using Kamusi's ballooning to address issues of cultural bias within lexicographic data collection. Version 1.0 of the app loops through a circumscribed set of question types, beginning with questions geared toward the collection of individual terms. First, we present terms and their definitions from the balloon-modified English priority list (e.g., *light* 'being low in weight') and ask, "What word would you use in [your language]?" (Figure 1).



light
Being low in weight
What word would you use in Kirundi?

Figure 1. Initial request for translation.

If the system is set to field-collection mode, the term will be accepted as is, without passing through crowd validation procedures; in this way, a field researcher can use the tool with one or more consultants to rapidly generate an initial term list. If the system is set to crowd mode, some participants will then be asked to rate the validity of terms submitted by other contributors (Figure 2).

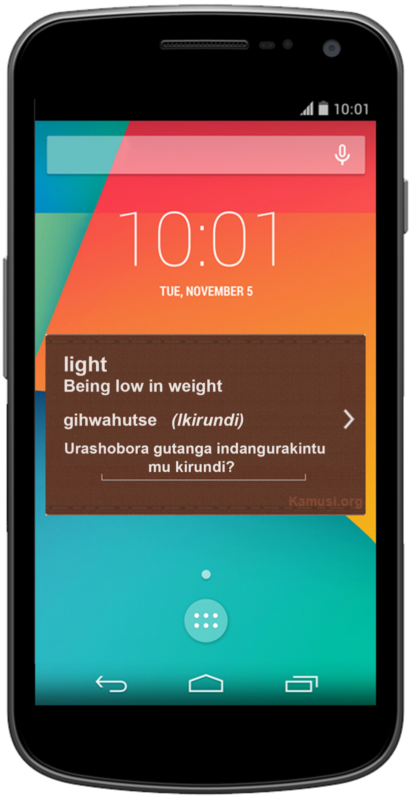


light
Being low in weight
Is GIHWAHUTSE a good KIRUNDI translation?

Figure 2. Rating of submitted term.

Once a translation has passed the validation threshold, further contributors will be asked to provide an own-language definition (Figure 3, localized to Kirundi) or to rate definitions submitted by others.

[17] All programming features discussed in this section are anticipated to be functional by June 2014.



light
Being low in weight

gihwahutse *(Ikirundi)*

Urashobora gutanga indangurakintu mu kirundi?

Figure 3. Localized request for own-language definition.

This system is well-tailored for researcher-driven fieldwork or written languages with numerous speakers who have persistent smartphone network access.

In many cases, the people involved in preserving a language speak it but do not write it. The app's 2.0 version is intended to extend the mobile technology to languages that are not commonly written or do not have a critical mass of participants, or both.[18] Although network access is currently necessary to use the system, an offline version is anticipated when synchronization and funding issues are resolved. The principal new feature of version 2.0 will be the collection of structured audio data, including pronunciations, own-language definitions, and possible retellings of the definitions in a contact language. This is discussed in the following section.

5 Monolingual Talking Dictionaries

Talking (or audio) dictionaries are an important technology for preserving the sounds of particular languages and dialects. Traditionally, for LRLs, the sound files simply appear in association with contact-language descriptions or translations of the terms.[19] Other language preservation projects endeavor to record stories as told by speakers of endangered languages, with italk library (italklibrary.com) providing an excellent example for Australia. The Fidget Widget's version 2 approach to talking dictionaries combines the idea of codifying the sounds of a language and the practice of preserving narratives about what a culture's concepts represent.

The new version will proceed as follows. After finishing version 1's set of questions focused on the gathering of lexical terms, the app will request that each term be pronounced. The smartphone will provide a visual countdown and a beep. This process will yield data on a par with most (if not all) talking dictionaries for endangered languages: as mentioned above, this consists of the written lexical item, the contact-language gloss, and a sound file of the term being pronounced in the indigenous language. The third step will ask the user to explain the concept in their language, with a timer to encourage brevity. It will be necessary to tinker with the timing system to find a sweet spot that allows answers of good quality, minimizes stress, and does not cut speakers off mid-stride but still discourages rambling. The fourth step will ask the user for a similarly pithy explanation of the concept in a contact language that they know. In the Australian case, where almost all participants also speak English, the contributors will also be asked to provide an English resume of their own-language definition, providing a gateway for people who are not already familiar with the language—including many of their own youth. With these simple procedures that integrate the basic capacities of smartphones with the data design of Kamusi, talking dictionaries will become valuable internal reference sources for their own communities, as well as repositories that enable

[18] The existence of some sort of functional Unicode-supported orthography and the involvement of at least one person who can bridge writing and orality are minimum conditions for participating in the system. Where orthographies are still in contention, Kamusi's internal structure is programmed to support multiple writing systems.

[19] See, for example, the Koasati Digital Dictionary (Koasati and English) (http://koasati.wm.edu); the Nganasan Multimedia Dictionary (Nganasan and Russian) (http://www.speech.nw.ru/Nganasan/); and the Talking Dictionary of Ainu (Ainu and Japanese, with further translations of the Japanese glosses into English) (http://lah.soas.ac.uk/projects/ainu/).

interested others to support the language's continued existence and revitalization.

Working with italk library, Kamusi's design of the app's version 2.0 is being developed for the Arrernte community as first users; a field trial with the initial one hundred terms from the Kamusi priority list is scheduled to be completed before late June 2014. The Arrernte are interested in preserving the specific terms of their language as well as the way they are expressed in context, and they also want to revitalize the tongue's use among its younger generations. The stories that italk library has been recording have proven enjoyable, but they do not provide a structure for accessing content in a way focused on reference or language learning. The mobile app will be used to term by term elicit pointillistic accounts and working definitions, keyed to each term's dictionary data container, that together paint a full picture of the language and culture. For example, Kamusi's English definition of the adjective *yellow* is 'being of the color of sunflowers or ripe lemons, between green and orange in the visible light spectrum'.[20] A contributor from the Australian group might say (in Arrernte or in English), "Yellow is, like, it's the color that we see, it's the color of the sun when it's going down, before it turns orange like a cooling fire," and such an unedited vignette would serve as the talking definition.

In addition to own-language definitions, recorded English definitions will make it possible to transcribe the meanings of each term with the aid of a wider crowd (who do not necessarily speak the endangered language); as text, transcriptions can be indexed and searched to provide access to the dictionary data through technological tools. Transcription of the own-language definitions is not planned in the near term, but it remains a desirable possibility with time and resources, especially as a community activity. In the first iteration, terms will be elicited via the mobile app from the English list (since little "ballooning" will be in effect for the first language from the Australian continent to enter the Kamusi system), but it will also be possible to upload native concepts to the system and then use the app for gathering audio.

Monolingual dictionaries have not generally been conceived of as practical for endangered languages, and sophisticated multilingual dictionaries have long been deemed impossible (Zgusta, 1971: 210; Haensch, 1991; Landau, 2001: 11). The tools and methods discussed in this article, however, make it possible to document endangered languages effectively, both by (a) generating term lists rapidly and in association with concepts from related languages, and (b) incorporating spoken definitions that encapsulate the essence of each idea. As mentioned above, the result will be a useful resource for the community, something that can be understood by the segment that does not read the contact language and used by younger generations interested in revitalization. Additionally, the system produces bridges to many other languages, allowing local knowledge to endure beyond the boundaries of shrinking linguistic communities.

6 New Directions

Kamusi.org has a long task list, with a goal of providing a full range of lexical resources for both people and machines. Some of these objectives, such as detection of malicious users and validation procedures for crowdsourced data, are informatics challenges. Many other objectives are technical, and will apply across all languages—enhancements to the data model for bridging concepts that are expressed with different parts of speech in different languages (e.g., where colors act as verbs), and a host of improvements to the editing system based on lessons learned during the multilingual pilot phase. A few are noteworthy in this concluding section because of their specific interest to endangered languages.

Determining the boundary between a language and a dialect is frequently problematic. In the case of two tongues (Kinyarwanda and Kirundi) that are often considered dialects separated by a political border, Kamusi discovered in the process of creating separate dictionaries that there are substantial differences between the two which had not previously been documented. However, it would be impractical to create full dictionaries of every dialect of a language when a large portion of their vocabularies are shared. We will therefore produce a system to geo-tag entries based on where a term is known to be in use. As the map becomes populated with zones of use, it will be possible to visualize where one dialect fades into the next, and where one language territory ends and the next begins. Similarly, programming is planned for geo-tagging the specific location where a participating speaker in a talking dictionary acquired their language (in-

[20] http://kamusi.org/define?headword=yellow&to_language =366.

cluding people contributing pronunciations for well-resourced languages). This will build an audio portrait of dialect, sociolinguistic, gender, and other variation. These mapping features, combined with expanded data collection, will enhance the possibilities for linguists to study language contact, spread, and historical change. Other improvements and new features, such as an app to upload photos of cultural items directly to a dictionary entry, or the expansion of audio features to the existing open-ended cultural notes field, will allow contributors to flesh out dictionary entries with relevant ethnographic information that contextualizes a language within the lives of the people who speak it.

In terms of innovations to the system itself, we see as a priority the development of offline input systems, both for contributors who want to use an interface like the one at www.kamusi.org, as well as those who wish to use the smartphone app when not connected to the internet. In fact, we did release offline software for the bilingual dictionary between English and Swahili, but the multilingual model added so many complexities that the program must be completely rewritten. Synchronization and the management of large data sets on small devices are major technical challenges, which can only be tackled with solid funding. Similarly, money permitting, we aim to code the system architecture to include a privacy system for linguistic groups who wish to document but also restrict access to certain lexical items (e.g., taboo words) or even their entire language. In addition, as we discussed in Benjamin and Radetzky (2014), we are committed to incorporating gamification, or games with a purpose, into both mobile and web platforms (Castellote et al., 2013; Paraschakis, 2013; Hamari et al., 2014). This will propel the accumulation of data and its validation by the crowd, pushing the project along the path toward obtaining as much open data for as many languages as possible.

Market forces will never support the creation of widely-available print dictionaries for most LRLs, and scholarly interest and available funding for online dictionaries will remain hit-or-miss, even as languages fade away. The tools presented in this paper are offered as methods for rapidly and reliably developing lexicographic resources for the world's endangered languages.

References

Pedro Florentino Ajpacajá Túm. 2001. *K'ichee' Choltziij.* Cholsamaj, Ciudad Guatemala, Guatemala.

Earl Anderson. 2003. *Folk-Taxonomies in Early English.* Fairleigh Dickinson University Press, Madison, New Jersey.

Martin Benjamin. 2013. http://kamusi.org/priority-list

Martin Benjamin and Paula Radetzky. 2014. Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. In *Proceedings of the International Conference on Language Resources (LREC '14),* Reykjavik.

Claire Bowern. 2010. Fieldwork in contact situations. In Raymond Hickey, editor, *The Handbook of Language Contact.* Wiley-Blackwell, London, pages 340-357.

Louis-Jean Calvet. 1974. *Linguistique et Colonialisme: Petit Traité de Glottophagie.* Payot, Paris.

Jesús Castellote, Joaquín Huerta, Javier Pescador, and Michael Brown. 2013. Towns conquer: A gamified application to collect geographical names (vernacular names/toponyms). In *Proceedings of the 15th AGILE International Conference on Geographic Information Science,* Leuven. http://www.agile-online.org/index.php/conference/proceedings/proceedings-2013.

James Crippen and Laura Robinson. 2013. In defense of the lone wolf: Collaboration in language documentation. *Language Documentation and Conservation,* 7:123-135.

Nick Enfield. 2003. *Linguistic Epidemiology: Semantics and Grammar of Language Contact in Mainland Southeast Asia.* RoutledgeCurzon, Oxon, UK.

Joseph Errington. 2001. Colonial linguistics. *Annual Review of Anthropology,* 30:19-39.

Dirk Geeraerts et al. 1994. *The Structure of Lexical Variation: Meaning, Naming, and Context.* Mouton de Gruyter, Berlin.

Johanne Fabian. 1983. Missions and the colonization of African languages: Developments in the former Belgian Congo. *Canadian Journal of African Studies/La Revue Canadienne des Études Africaines,* 17:165-187.

Günther Haensch. 1991. Die Mehrsprachigen Wörterbücher und ihre Probleme. In Franz Joseph Hausmann et al., editors, *Wörterbücher: Ein internationales Handbuch zur Lexikographie,* vol. 3. Walter de Gruyter, Berlin, pages 2909-2937.

Juho Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does gamification work? – A literature review of empirical studies on gamification. In *Proceedings of the 47th Annual Hawaii International Conference on System Sciences,* pages 3025-3034, Waikoloa.

Juda Joffe and Judl Marq, eds. 1961, 1966, 1971, 1980. *Groyser Verterbukh fun der Yidisher*

Shprakh, vols. 1-4. Yiddish Dictionary Committee, New York.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the International Conference on Language Resources (LREC '14),* Reykjavik.

Sidney Landau. 2001. *Dictionaries: The Art and Craft of Lexicography,* 2nd ed. Cambridge University Press, Cambridge.

Ulrike Mosel. 2011. Lexicography in endangered language communities. In Peter Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages.* Cambridge University Press, Cambridge, pages 337-353.

Dimitris Paraschakis. 2013. Crowdsourcing cultural heritage metadata through social media gaming. Master's thesis, Malmo University.

Françoise Raison-Jourde. 1977. L'échange inégal de la langue: La pénétration des techniques linguistiques dans une civilisation de l'oral (Imerina, début du XIXe siècle). *Annales: Économies, Sociétés, Civilisations,* 32:639-669.

Ladislav Zgusta. 1971. *Manual of Lexicography.* Mouton, The Hague.