Collaboration in the Production of a Massively Multilingual Lexicon

Martin Benjamin
EPFL – Swiss Federal Institute of Technology
EPFL IC LSIR
BC114
Station 14
CH-1015 Lausanne, Switzerland
E-mail: martin.benjamin@epfl.ch

Abstract

This paper discusses the multiple approaches to collaboration that the Kamusi Project is employing in the creation of a massively multilingual lexical resource. The project's data structure enables the inclusion of large amounts of rich data within each sense-specific entry, with transitive concept-based links across languages. Data collection involves mining existing data sets, language experts using an online editing system, crowdsourcing, and games with a purpose. The paper discusses the benefits and drawbacks of each of these elements, and the steps the project is taking to account for those. Special attention is paid to guiding crowd members with targeted questions that produce results in a specific format. Collaboration is seen as an essential method for generating large amounts of linguistic data, as well as for validating the data so it can be considered trustworthy.

Keywords: dictionaries, crowdsourcing, gamification, multilingual

1. Introduction

This paper discusses the multiple approaches to collaboration that the Kamusi Project is employing in the creation of a massively multilingual lexical resource. The project has set its ultimate goal as documenting "every word in every language", which, while admittedly unattainable, nevertheless defines the target. Several strategies will be intermixed in working toward the goal, through a combination of crowdsourcing and expert input. These strategies will address both data generation and quality control, with a special emphasis on building resources for languages that have previously been marginalized.

2. Multilingual Data Structure

With the Kamusi Global Online Living Dictionary (GOLD), we have reengineered the dictionary to enable a concept-by-concept matrix of human linguistic expression. In summary, each language is to have its own monolingual dictionary, with each sense of a term elaborated in its own entry. Those entries can all be richly populated with relevant data, tailored to the structures of each individual language, that will provide detailed information for people and HLT applications. A term in one language can then be linked to a similar concept in another language; if the concept in the second language is further connected, transitive links are established between the new language and the third, fourth, fifth, etc. Relations between terms are marked for their level of equivalence (whether they are parallel, similar, or explanatory), and tracked for whether the relations are human-confirmed or separated by a number of automated linkages. The intended result is a rich standardized lexical resource for each language that further provides detailed, harmonized paths to every other language.

The Kamusi data structure is designed to accommodate the many inconsistencies in lexicons among languages. When concepts align neatly, it is possible to string transitive links among languages indefinitely: English *sun* matches to Swahili *jua* matches to French *soleil*, and adding the parallel concept in other languages produces a consistent chain of meaning. When terms are parallel, Kamusi displays believable connections between languages, with a graphic depiction of the degrees of separation between pairs that have not been human-confirmed. Links between languages are murkier when concepts do not align well, such as Swahili *mkono* encompassing a part of the body that English divides into *arm* and *hand*, a distinction that is basically consistent between Bantu and European languages. Kamusi continues to show the transitive links, but flags the increased level of uncertainty. Concepts that exist in one language but not in another, such as Swahili *kanga*, are shown as "explanatory" when given a working gloss such as *fabric wrap for women* in English, and are shown in searches from the source to the explaining language, but not as part of the lexicon of the second language. A further alignment problem arises because concepts may be framed as different parts of speech in different languages, such as the Swahili verb *–furahi*, which matches to the non-lexicalized English *be happy*. *Happy*, of course, is an adjective, so the data system enables the establishment of a bridge between the different methods of expression, with appropriate indications throughout the multilingual connections.

The data structure enables the inclusion of large amounts of rich data within each sense-specific entry. Each entry should include a monolingual definition, which can be further translated to any other language. Each language can also be configured to input the morphemes and inflections of its particular language structure, such as Bantu noun classes or Arabic singular, dual, and plural forms; these elements can become available to machine translation and other HLTs for improved lexical precision. The system handles many other data elements, including tones, multiple scripts, alternate spellings, intra-language

relationships such as synonymy, pronunciation, etymology, and dialect, with some of these features slated for improvements as programming continues. Several features distinguish the Kamusi system from other large lexicons, such as WordNet (with which we are working to embed cross-links between specific concepts, use attributed definitions to seed our data when appropriate, and push improved data back to that project) or Wiktionary (from which we incorporate definitions, with attribution, when appropriate), including the ability to track differentiated senses within and among languages, the inclusion of word forms and other extended data as part of the machine-readable data structure, and the opportunity for users to add or update data in an easy-to-master format that is nonetheless subject to validation procedures.

The architecture, intended to handle any peculiarity that we have been able to identify, is necessarily complex. However, the interfaces through which users interact with the system are the subject of continued efforts to make simple and intuitive.

3. Sourcing Data: Mining and Minds, Experts and Crowds

With the structure established, the challenge becomes filling the system with data. While humans are remarkably skilled at transmitting lexical data from one cranium to the next, we have done a poor job of downloading that data into forms that can be stored and operated on by machines. Most lexical data does not exist in any digitized form; for those languages that have any documentation at all, a wordlist of a few thousand terms is much more likely than a deep compendium that truly attempts to represent the language. What does exist is rarely in a commensurate form from one resource to another, much less among languages. Data that exists in a digitally useful form is nevertheless often barricaded by copyright. Kamusi GOLD will often be seeded through the harvesting of copyright-available data sets, but the major effort will rely on human knowledge to review that data and add as much new information as possible. As parents pass their linguistic knowledge to their children word by word over many years, our task is to systematically transmit this knowledge into an open database for each language.

We start with three types of data, all of which come with their own problems:

1. Existing data. The first problem with existing data is often the copyright, not just for contemporary digitized data sets, but also for print dictionaries within the seventy year copyright window. Older data also presents difficulties for optical character recognition, because of both poor quality print sources and the lack of spell checkers for languages without good digitized data. If clean data is available, we then must determine the fields to which different parts of an entry belong, and maneuver the components of each record into those fields; this is often made more difficult by dictionaries that were not designed around a database or spreadsheet model, and tend to vary their format at whim. For example, it is often impossible to tell the difference between a definition and a usage example in an automated manner. Assuming we can get a data set neatly into consistent fields, we are then faced with the challenge of aligning one language to the next. It is not enough to know that a term in a data set is glossed by a word such as *light* in English; we must match directly to the right sense. In most cases, imported data will not include even all the minimum data items required for a standard Kamusi entry (lemma, headword, part of speech, and own-language definition), and rich extended data will be totally lacking. These problems can only be fixed by human review, whether experts or the crowd (Hernandez and Stolfo 1998, Lee et al 1999, Dong and Naumann 2009). Over 100 recent data sets have already been made available to Kamusi, mostly for African languages, including Osborn, Dwyer, and Donahue (1993), Ahamer (2001), and lexical training material compiled by the US Peace Corps. Excitingly, a Swahili-Chinese data set has been offered, with the original lexicographer taking charge of incorporating the data; adding Chinese via this route will be an interesting test case for using a language other than English as the index of reference to the system. Many older data sets are available in various Internet archives. Merging can begin when communities and funding to accomplish the work have been identified.

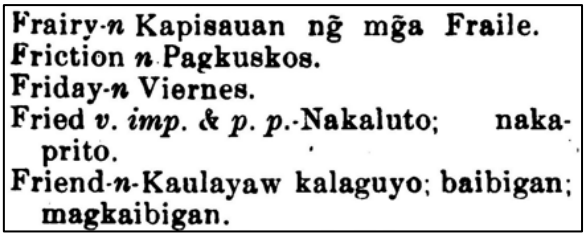
Frairy-*n* Kapisauan nḡ mḡa Fraile.
Friction *n* Pagkuskos.
Friday-*n* Viernes.
Fried *v. imp.* & *p. p.*-Nakaluto; naka-prito.
Friend-*n*-Kaulayaw kalaguyo; baibigan; magkaibigan.

Figure 1: Nigg (1904), an English-Tagalog example of copyright-available data, in scanned PDF format. Problems include interpretation of line breaks, hyphens, and semicolons, e.g. whether nakaluto and nakaprito are synonyms or alternate forms, and capitalization, i.e. determining which are the proper nouns in the data. Nouns are shown in three different ways: -n, n, and –n-.

```
Frairy-n Kapisauan ng mga Fraile.
Friction n Pagkuskos.
Friday-n Viernes.
Fried v. imp. & p. p.-Nakaluto; naka-

prito.
Friend-n-Kaulayaw kalaguyo; baibigan;

magkaibigan.
```

Figure 2: The same data in the Google Books OCR version. Optical character recognition has converted the data into manipulable text, but removed diacritics that convey important information.

2. Language specialists. The ideal data collection method is for specialists for each language to contribute rich data for each entry. Such contributions can be

considered authoritative, and provide the full range of information needed for the term to be understood by humans and manipulated by HLTs. Specialists can work from the English-derived list of parallel concepts, or bring in terms that are unique to their language. They add depth and nuance that cannot come from existing static data and might not be elicited from the crowd. However, we cannot rely solely on experts for several reasons. First, many languages have no dedicated specialists. Second, many linguists who have studied a particular language in detail might be too busy to participate in the project, or uninterested, or might wish to keep their data proprietary. Third, specialists are often professionals who need and deserve compensation for their time, so the work they can contribute will be limited by the funds we can raise. Fourth, specialists do not know every last word of any language, and can take decades to document all that they do know, so we will need input from community members to supplement the items they are able to contribute.

3. Crowdsourcing. Everyone is an expert in their own language, to the extent that they often discuss the words they use, and actively pass them on to the next generation. Getting that expertise into a form that can be used for scholarship and HLT, though, is a difficult undertaking. First, we need to find speakers of a language, or they need to find us – not such a problem for, say, Turkish, but perhaps impossible for some remote or endangered languages. Second, the people need to have cheap and reliable access to electricity, communications networks, and input devices, which is again a large constraint for languages on the long tail. Third, people need to be motivated, and that motivation needs to continue for a long time. Fourth, they need to have tasks that match their knowledge and skill sets. Fifth, they need some form of training or feedback to make sure their contributions are consistent. Sixth, their contributions must be checked for stylistic errors. Seventh, their contributions might be incorrect, so must be checked for factual errors. Eighth, opening data development to the crowd also opens it to malicious users, so systems must prevent or remove vandalism. We cannot address problems of access to ICT, but we are building systems to motivate users and guide them toward producing high-quality vetted data.

4. Collaborative Lexical Data Collection

The collaborative system we are designing has four main components:

1) An in-depth system for editing all aspects of any entry. This system is open to anyone, but is generally expected to be used by specialists. The edit engine is as easy to use as an online hotel booking system. However some of the tasks involved are inherently complicated at the conceptual level, because we are seeking detailed, non-obvious data. For example, users must understand the difference between a definition (an explanation of a term in its own language, written according to certain stylistic and content guidelines consistent with practices discussed in Zgusta (1971), Landau (2001), and Svensén (2009)), a translation (a gloss of a term in another language), and a definition translation (an explanation of a term in another language); e.g., many novice users would contribute English *dog* in the "definition" field for Swahili *mbwa*, or use that space to write an English definition of *mbwa*, rather than using the translation mechanism to link the correct sense of *dog* from the English side or using the "definition translation" field to provide an English explanation of the Swahili term *mbwa*. Other tasks require specialized knowledge, such as IPA or tone spellings. Consequently, users must experience some training, either through online tutorials or directly from a person, and submissions must be reviewed and confirmed by human moderation. Any sense of any term in a language can be submitted via the edit engine, regardless of whether it has a pre-existing translation equivalent; moderators can prevent slang or obscure senses from entering the lexicon until the usage has been documented.[1]

2) A system to pay specialists for their work as funding becomes available. We know how long an average entry takes to produce, so we can calculate a fair per-entry wage for a language professional. Our biggest obstacle is finding the funds to pay for experts' time. Once we have funds for a given number of terms in a given language, we can feed them to specialists from a prioritized queue of concepts as defined in English (Benjamin 2013)[2], and issue payment credits as entries are completed. Submissions can be monitored for quality, and donors can view the entries they have sponsored. We are designing a system where people can sponsor any number of words for the language of their choice, which we hope will then be quickly supplied by hungry linguists.

3) "Play to Pay" and the "Fidget Widget" are variations on the core crowdsourcing element. Kamusi data is available to the public for free, but no-cost will come with the price of sharing knowledge. The Fidget Widget, in testing as of March 2014, is designed to build credits during the idle moments when many people look compulsively for something to do on their mobile devices. Play to Pay will be more aggressive in requiring participation in exchange for access to data; users will either need to answer a question online in order to continue, or answer a question of the day that they will receive via email. Users are asked targeted questions in

[1] Specialized terminology differs from lexicography in substantial ways, and is therefore treated differently in Kamusi (Benjamin 2011). A stand-alone participatory terminology development system was implemented at http://terms.kamusi.org before the current multilingual structure was properly coded using a different content management system. Programming is currently underway to integrate the terminology system within the overall Kamusi architecture.

[2] Using an English wordlist as a starting point is admittedly a methodologically imperialistic means of generating parallel data that yet does not preclude indigenous concepts from being added. A method to levitate concepts for consideration in a particular language based on their occurrence in related languages is discussed in a paper currently under review (Benjamin and Radetsky 2014 pending)

association with the terms they look up, and will earn points that they can exchange for more access. The most difficult part of the model is targeting questions that direct users to provide data in exactly the format required, appropriate to the user's knowledge and skill set. Some of the questions are open-ended, while others are yes/no or multiple choice. An open-ended question in a user's language might be, "What term would you use for [defined English word] in [your language]". After several users have answered that same question differently, other users see, "For the English term [defined word], would you use [A], [B], [C], or do you propose [another]"? Many of our questions will serve to clean and expand upon data that we are trying to import (Hung et al 2013). Advanced users will be asked to produce definitions, while other users are only asked to vote on definitions that others have submitted. User ratings will provide a metric for judging a contributor's competence at a particular task, which can then be used to optimize the types of questions we ask each person, and for ferretting out vandalism. Users who consistently produce good answers will earn trust, and trusted users will earn advanced privileges, including the right to moderate contributions that correspond to their demonstrated skill sets. The crowdsourcing elements are being built and tested over time, with participation strictly voluntary in the early phases, and modifications to the above description anticipated as we learn what approaches elicit the maximum amount of quality data.

4) Gamification. Some aspects of data generation can be turned into sport (Castellote et al 2013, Parashakis 2013). In one game, an English term and definition will be sent to the players. Players will send back their translations of the term, and after ten answers agree, people will be awarded points based on the order in which their entry was received, as well as points based on the order in which their language team completed the task. Games can be used for new data, or to clean up imported data. The important element for games will be that the data be reliable upon its completion – that is, that players end up participating in both data production and error checking as part of the game.

5. Conclusions

The collaborative processes we are designing are intended to overcome a variety of problems that we have encountered in our own project and in others. Because Kamusi has always required user submissions to pass through moderation, we have never experienced the vandalism and reliability problems that are endemic to Wikimedia-style projects. However, even in the absence of the complexities demanded by projects based on the Wiki markup language, we have encountered difficulty conveying to users the exact elements they need to supply for good dictionary entries. While many crowdsourcing projects ask participants to conduct small and simple tasks, many of the questions we ask will be complicated and open-ended. We will therefore work to guide users through tasks they prove competent to manage, rather than simply hoping people migrate toward their lodestars. Finally, while we hope to make collaboration fun through gamification, we will nevertheless make it mandatory. Dictionary users (Bergenholtz and Johnsen 2013), in keeping with users of other web resources (van Mierlo 2014), prefer much more to receive data than to contribute to its development. With a lexical resource intended for as many as 7000 languages, it is not feasible to wait for the occasional person who wants to contribute a few words for their language – such a method would never produce the necessary data, and therefore we would never have any data to offer to the public at large. We are instead creating a system that imposes on users the price of contributing a little of their knowledge, in exchange for access to all that others have already contributed. While mandatory participation may limit the number of people who choose to access the site, that is an intentional trade-off in the effort to generate the massive amount of data that constitutes "every word in every language". Through a combination of approaches to collaboration, including paid experts, knowledgeable volunteers, word game enthusiasts, and ordinary users, we are embarked on building a rich, high-quality lexical resource for numerous languages worldwide.

6. References

Ahamer, F. (2001). K'amus na Hausa/ Hausa Database, http://www.univie.ac.at/Hausa/oracle/KofarHausaE3a.html.

Benjamin, M. (2011). *Toward a Standard for Community Participation in Terminology Development*. Proceedings of the First Conference on Terminology, Language, and Content Resources, Seoul, Korea.

Benjamin, M. (2013). *How We Chose a Priority List for Dictionary Entries*. http://kamusi.org/priority-list

Benjamin, M., and Radetsky, P. (2014 pending). *Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages.* Paper submitted to ComputEL, 52nd Meeting of the Association for Computational Linguistics.

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.

Castellote, J., Huerta, J., Pescador, J. and Brown, M. (2013). *Towns conquer: a gamified application to collect geographical names (vernacular names/toponyms)*. In: AGILE 2013, 14-17 May, 2013, Leuven, Belgium.

Dong, X., Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, pages 1654–1655.

Hamari, J., Koivisto, J., & 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*, Hawaii, USA, January 6-9, 2014.

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

Hung, N., Tam, N., Miklós, Z, Aberer, K. (2013). On leveraging crowdsourcing techniques for schema matching networks. In *Database Systems for Advanced Applications*, pages 139–154.

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

Lee, M., Lu, H., Ling, T., and Ko, Y. (1999). Cleansing Data for Mining and Warehousing. In *Database and Expert Systems Applications*, pages 751–760.

Nigg, C. (1904). A Tagalog English and English Tagalog Dictionary. Imp. de Fajardo y comp.

Osborn, D., Dwyer, D., and Donahue, J. (1993). A Fulfulde (Maasina)-English-French lexicon. East Lansing: Michigan State University Press, 1993.

Paraschakis, D. (2013). *Crowdsourcing cultural heritage metadata through social media gaming.* Malmo University, School of Technology, Department of Computer Science, Master Thesis.

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

van Mierlo, T. (2014). *The 1% Rule in Four Digital Health Social Networks*. Journal of Medical Internet Research, 16 (2).

Zgusta, L. (1971). *Manual of Lexicography.* Berlin: de Gruyter.