Excluded Linguistic Communities and the Production of an Inclusive Multilingual Digital Language Infrastructure

Martin Benjamin, Senior Scientist
École Polytechnique Fédérale de Lausanne

The consequence of linguistic digital exclusion is the inability of billions of people to access vital knowledge and economic resources that contribute to prosperity in an era of globalization. However, rectifying linguistic inequity is mostly absent from development discourse and the agendas of governments and agencies that undertake development activities. Most efforts to produce content for excluded languages depend on the haphazard occurrence of a commercial, academic, or programmatic purpose for an activity in a given language at a particular moment. The Kamusi Project seeks to address the digital linguistic divide by engaging communities in the systematic collection of codified data for any language – linguistic information that can be used in many kinds of advanced knowledge and technology resources. This paper explores assumptions about participants' motivations and behaviors that underlie the project's methods, including participation in online games and interactive mobile apps intended to elicit speakers' knowledge of their own languages in ways that can be shared by others.

While the Kamusi system aims to welcome all, disparities may continue to exclude those without substantial time, network access, equipment, digital experience, or literacy, leaving international members of a diasporic language group as its most active contributors. Further, smaller and more remote languages have, by definition, fewer potential participants and less access for participation, thus perpetuating their inability to jump the digital divide. Without external support for the time and effort necessary to gather linguistic knowledge, even the most carefully constructed tools will fail for thousands of languages spoken by millions of people, including many languages near extinction. This paper raises, without definitively resolving, the social challenges of a multilingual digital infrastructure platform that has the technical capacity to document every word in every language, but can only approach accomplishing this objective through the involvement of those who have the least access to taking part.[1]

You are not typical. You speak English at an advanced level. You are concerned with matters pertaining to the intersection of language and technology. You have access to high-level knowledge resources. You are also probably immersed in the digital realm, thinking little of events such as checking the weather in a distant city,

---

[1] For a full description of Kamusi Project design details, please refer to the reference section. In accordance with the call of the 11th Language and Development Conference, this paper walks a different path, aiming to stimulate thought based on project experience, rather than to present research findings.

downloading your boarding pass to get there, extracting currency from a cash machine on arrival, and navigating by GPS to a hotel you booked on your phone. Through happenstance of birth and opportunity, you can accomplish almost anything in a language you understand well, aided by technology designed to support you in your quest. Your experiences show how technology can change lives, by connecting you to people, ideas, information, and tools. You are a member of the information elite.

Most people have very different experiences than yours, with both language and technology. The typical communication device is now a phone that transmits voice and text, though these are being steadily supplanted by smarter devices. For most people in most countries, the costs of purchasing their equipment, keeping it charged, and paying for service represent a significant investment. The things they can accomplish with their technology are limited, and the tools to accomplish them are in a language they may not have mastered. They cannot find basic information in their language, such as weather or bus schedules, to say nothing of health, markets, or school subjects. Such resources do not exist. They have no way to learn more detailed information about their languages, as you expect with the dictionaries and other references available to you. Their linguistic information has probably never been gathered. If some has been collected, it has not likely been shared in an affordable format, much less digitized in a way they can access. Their languages almost certainly do not exist as data that can be used in technological applications. The typical person is linguistically excluded, from knowledge about their own language, from resources in other languages, and from the technologies that might open opportunities for them.

Access to technology, a common concern of governments and development agencies, is frequently seen as a question of physical infrastructure: fiber optic cables, microwave towers, computers in schools. Software applications are a second tier of concern: medical records, emergency response monitoring, mapping clean water. Agencies do not build systems themselves. Their leaders, like you, are comfortable in the major languages of technology. Attention to language is an additional expense. Since technology has not leapt the linguistic divide, perhaps it seems unrealistic that any given program could be customized to the language of its constituents.

At the same time, Human Language Technology (HLT) is growing impressively, for English and a few other charmed languages. Your ability to speak into your device, have your sounds recognized and your meaning parsed, is a testament to outstanding efforts in Natural Language Processing (NLP) in recent years. But again, few expect that these advances can or should be available in more than a few languages. Many in the information elite assume that consumers will learn English, or muddle through without it. There are too many languages to support, and their individual markets are too small. Language is a hornet's nest, even for language technology, that is most easily approached by building on the incredible tools and knowledge set developed for the languages that can pay their own way.

Beneath these rationalizations for why language resources and technology are not available for most people, though, is one consistent truth: we do not have the data. We do not have the data even for English, in the form of one open, interoperable set of linguistic referents that can ensure meanings are transmitted among projects and technologies. We do not have the data for other major languages, either for their consistent internal development or to interact with English or other big players. And we most certainly do not have the data for the thousands of languages in which the majority of the world's people conduct their lives every day.

This lack of data can be attributed to:
1. Lack of method. Gathering sufficient useful data for any language is a difficult task, and doing it for thousands is thousands of times harder.
2. Lack of will. With few thinking it possible to gather comprehensive data for most languages at the scale needed for effective technology, few are motivated to try.
3. Lack of money. Gathering data involves costs, and these costs must be borne somewhere. Who is willing to pay to develop a low-market language with few paying consumers, few demanding voters, or few research specialists? Corporations prefer to invest their resources in avenues that have more evident payback. Agencies would sooner spend scarce resources where the most urgent needs can be visualized – providing medicines, building clinics, planting trees.

While the Kamusi Project has been a colossal failure at addressing the third problem, we have produced an answer to the first that could provide a solution to the second. Kamusi has developed a data platform that can, in principle, accommodate a full panoply of shareable, processable information for every word in every language. The project has also designed systems to gather that information directly from the speakers of each language, in a straightforward and systematic manner that is intended for many people to enjoy. Questions remain, however, about the best ways to bring the people who hold the data for their languages in their heads to the games and tools we hope they will want to use.

People typically do not have unlimited bandwidth, storage, power, and processing capacity at their disposal. Those who can afford access do not necessarily have time to devote to non-lucrative activities. Those who have the time might not have the inclination. Those with the inclination might not have the skills to make maximal use of their equipment. Those with the technological skills might not have the knowledge needed to convey particular linguistic information. Those with the knowledge in their heads might not master the methods for transmitting it. For all these reasons, the number of people who will participate in developing resources for their language is only a fraction of the numbers who speak it. The fewer people who speak a language to begin with, the harder it is to find and excite that fraction. Similarly, languages spoken by people with lower incomes have a smaller fraction who can afford to participate, and languages spoken in areas away from communication hubs have a smaller fraction with the physical equipment to be

involved. A language with no tradition of literacy, spoken by a handful of elders far from any cell tower, will not be documented by its speakers playing games online.

For these smaller languages, a separate set of tools is on the Kamusi task list. These tools, including apps for recording "talking dictionaries" for the words and meanings of unwritten languages, will require field researchers or passionate native speakers. It is tempting to assign responsibility for the preservation of an endangered language to the members of its community, but this is a recipe that is guaranteed to fail frequently. What can be offered instead is a research project in a box: all of the tools a linguistics graduate student would need to embark on a documentation project. Hundreds of universities around the world have linguistics students in search of useful research opportunities, but not the time or capacity to set up their own lexicographic system from scratch. Were there organizations committed to funding field researchers for endangered languages, it is readily conceivable that numerous languages could have essential data systematically added to a global linguistic infrastructure. However, supporting field research requires money be guided toward the interests of people who have none of their own to spare and no voice to demand it, from funders who have not yet demonstrated enthusiasm about investing in resources for small languages.

It is midsized languages where Kamusi hopes to make more immediate progress. While the factors enumerated above mitigate against a huge uptake by players for such languages, it only takes a small cadre of devotees to make noticeable progress. One mobile data collection tool, for example, asks device users to type in their language's term for a concept in order to unlock their phone. Even if just a dozen speakers of a language enjoy having their knowledge tapped in this way, data will flow at a rate equal to the users' tendency to check their devices. The predominant beginning point for most languages is games developed for the Facebook environment. These games, ready to be deployed in any language listed in Ethnologue the moment resources are available to service speaker communities and manage the incoming data flow (that is a polite way of spelling "money"), are built around several social premises. First, you probably enjoy talking about the finer points of your language, and so do many people throughout the world. Second, you may enjoy playing word games, and so do many others, as seen with the international success of Scrabble and other popular online games. This premise has not been tested for most languages because no such games have ever been created, so a third hypothesis is that the games will satisfy a previously vacant niche for lovers of languages that have no game market. Fourth, your impulse to give something to your society is matched by even the poorest people, especially if the contribution is not financial, and more-so if the result is something tangible that will visibly benefit the community over the long term; this may prove especially true for expatriates who have migrated from their linguistic homeland, have good digital access, and wish to remain connected to their mother tongues. Fifth, as you like being recognized for your work, many will appreciate that the game announces their successful contributions to their Facebook friends. Sixth, the people in your circle tend to share your languages, so progress for a language will accelerate as players

bring in friends who they invite or who see their posted achievements. Seventh, some people are motivated by competition, so will enjoy timed games, being awarded points, and seeing their position vis-à-vis other players and other languages. Eighth, people like mental challenges but not complex instructions, so the games are built around straightforward tasks that need little technical wherewithal. Ninth, with over a billion members, Facebook is a global common ground where many will feel at ease to play games embedded in the ecosystem. Some languages have user groups on Facebook with thousands of members; a tenth premise is that offering those groups a fun, compelling game that enables them to give something back to their communities will result in some languages growing their datasets at a rapid rate.

Universal systems for gathering data do not equate to universal needs for what to do with it. Endangered languages might not need whiz-bang machine translation in order to communicate with other groups, but might find that speech-to-text technologies provide a pathway to language preservation. People without bank accounts might have little short-term demand for e-commerce, but could immediately benefit from e-health applications in their language. Localization of technology is widely interesting; you are not alone in relying on written help to take photos with your mobile device and figure out the steps with which to share them. The subsequent lives of linguistic data, once gathered in a useable repository, will often not be made by the speakers of a language, but by the academic, administrative, or market forces that decide which languages to pursue for development. However, the platform is in place to support an interlinked lexical database for any language, with a beta version hosting well over a million terms in some two dozen languages containing 10,000,000 joints among 100,000 concepts. The act of systematically seeking the same extent of codified data for excluded languages as for the languages of the wealthy, using methods designed to attract at least a subset of their speakers, and making these words available to posterity as a perpetual digital knowledge base, is one that (when funding supports it) is now ready for implementation.

References elaborating Kamusi data and methods

The Kamusi Big Data Beta, 2015. https://kamusi.org/big_data_beta

Molecular Lexicography: A lexical data model for Human Language Technology, 2014. http://kamusi.org/molecular_lexicography

Problems and Procedures to Make Wordnet Data (Retro)Fit for a Multilingual Dictionary. Global Wordnet Conference, Bucharest, 2016. (Forthcoming, available in advance by request)

Looking forward by looking back: Applying lessons from 20 years of African language technology. 7th Language and Technology Conference, Poznan, Poland, 2015. (with Mohomodou Houssouba). (Forthcoming, available in advance by request)

Kamusi Pre:D – Source-Side Disambiguation and a Sense Aligned Multilingual Lexicon. Translating and the Computer 37, London, 2015. (with Amar Mukunda and Jeff Allen)

The Particles of Language: "The Dictionary" as elemental data for 7000 languages across time and space. CERN invited speaker, Geneva, 2015. Video: https://cds.cern.ch/record/2054123

Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. In: Proceedings of AsiaLex 2015, Hong Kong

Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Languages. 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, Maryland, USA, 2014. (with Paula Radetzky). Available at http://infoscience.epfl.ch/record/200377

Participatory Language Technologies as Core Systems for Sustainable Development Activities. 2014 Tech4Dev International Conference, EPFL, Lausanne, Switzerland, 2014. Available at http://infoscience.epfl.ch/record/200379

Collaboration in the Production of a Massively Multilingual Lexicon. 9th edition of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014. Available at http://infoscience.epfl.ch/record/200376

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification. LREC, 2014 (with Paula Radetzky). Available at http://infoscience.epfl.ch/record/200375

Elephant Beer and Shinto Gates: Managing Similar Concepts in a Multilingual Database. 7th International Global WordNet Conference, Tartu, Estonia, 2014. Available at http://infoscience.epfl.ch/record/200381