Abstract Title

Participatory Language Technologies as Core Systems for Sustainable Development Activities

Martin Benjamin[1]

[1] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Presenting author's email address: martin.benjamin@epfl.ch

Preferred Presentation Format (Full Paper, Poster, Either): Full Paper

Session Number: TH2-SE02-12

Request for Financial Assistance (Yes / No): No

Key Words: language, ICT, access to knowledge, infrastructure, services

Brief Biography of Presenting Author: Martin Benjamin is the founder and executive director of the Kamusi Project, an international effort to produce learning and lexical resources for languages worldwide. He is a senior scientist in the Distributed Information Systems Laboratory at EPFL. Trained as an anthropologist, his doctoral research (Yale 2000) examined the relations between development programs and the communities they intend to serve. He has conducted extensive field research in Tanzania, and leads language and technology cooperation projects with partners in Africa and beyond.

Abstract Text:

Introduction and purpose: Language is the medium by which people interact with all aspects of their worlds, whether economics, health, the environment, or technology. In both development programs and technology, however, language is usually given secondary consideration, if any at all. As a result, people who do not speak a major language are excluded from full participation in development programs and from technologies such as ICTs that could enhance their economic and social circumstances. In Africa, for example, where only a small minority speaks English or French, few development programs have the resources to devote to the most basic of language considerations, such as translating health information into local languages. Language technology can be a fast and cost-effective way of overcoming knowledge and communication gaps that underlie many other aspects of the development agenda.

Design and methods: The most efficient way to address language development is through public tools and vocabularies that can be reused, revised, and repurposed for multiple domains. We discuss a universal multilingual dictionary that is designed to build a parallel vocabulary of core concepts across languages, with a special focus on languages with few existing resources. The lexicons are built in close cooperation with local partners. Much attention is paid to a data structure that will enable downstream technologies. Further, a system develops domain-specific terminologies through a participatory process, so that complicated concepts can be communicated clearly and consistently. Data is made available to the public for free, with strong efforts to develop systems for access via least-cost technologies with the widest reach along the bottom of the pyramid.

Results: When successful, a focus on core language development can improve the outcomes of many other projects. In health, for example, translation is often too expensive and too difficult, because basic resources such as dictionaries do not exist and technical terms do not have adequate local-language equivalents. For the one-time cost and effort of building the lexicons and terminologies, in conjunction with the free tools being created to access those vocabularies, the infrastructure opens for cheap and rapid translation of health material. Similarly, students are able to use the lexicons to access knowledge that has previously been blocked behind linguistic barriers, reducing future language-based inequalities. Participatory data collection methods mean that vocabularies continue to grow in response to the expressed needs of particular linguistic and development communities. The perpetual accessibility of the public resource means that development programs can make sustainable use of the data in multiple languages, for multiple purposes, with no further investment.

Conclusion: Language is a hidden aspect of the development equation; language technology in itself does not cure a disease or put food on a table. However, whether communicating agricultural techniques, delivering government services, or performing numerous other activities that fall under the rubric of development, attention to developing language technologies for underserved language populations can be the difference between working together and

talking past each other – the difference between failing to communicate and succeeding in expressing the path toward accomplishing common goals.

Text Full Paper

Introduction and purpose: Language is the medium by which people interact with all aspects of their worlds, whether economics, health, the environment, or technology. In both development programs and technology, however, language is usually given secondary consideration, if any at all. As a result, people who do not speak a major language are excluded from full participation in development programs and from technologies such as ICTs that could enhance their economic and social circumstances (Synha & Hyma 2013). In Africa, for example, where only a small minority speaks English or French, few development programs have the resources to devote to the most basic of language considerations, such as translating health information into local languages. Language technology can be a fast and cost-effective way of overcoming knowledge and communication gaps that underlie many other aspects of the development agenda.

The linguistic environment facing development actors can be exemplified by the case of Malangali, a village in Iringa Region in the southern highlands of Tanzania. The residents of the Malangali area have historically spoken one of two interintelligible Bantu languages, Hehe or Bena, as a mother tongue. As a result of economic mobility and national efforts to promote Swahili, all but the most elderly residents now also speak that language fluently. In fact, due to adult literacy campaigns in the 1970s and universal primary education in most years since, most people of Malangali are literate in Swahili, but not in their mother tongues. Primary education occurs exclusively in Swahili. No curricular materials have ever been produced in Hehe or Bena, either for children who speak those tongues as their first languages, or for the rare outsider who wishes to learn. Secondary education, available to only a small proportion of local children, is primarily in Swahili, with materials in English particularly in the upper years. A number of residents of the area come from elsewhere in the country, and speak a variety of the country's 120-odd other languages. These non-local Tanzanians constitute much of the professional cadre of doctors, nurses, teachers, agricultural extension agents, water engineers, foresters, and others with post-secondary training. They communicate with the local population in Swahili, but might communicate with their families in their own mother tongues, especially when using their mobile phones to call or text back home. They also have a working knowledge of English, as do the small number of local residents who have been able to attain secondary education and also remain in or return to the village despite better economic prospects elsewhere. Additionally, Malangali has hosted many visitors from external development agencies, for periods ranging from hours to years, who generally can communicate in English and may also know some Swahili.

This multilingual environment is ripe for miscommunication (Benjamin 2000). The clearest cases of language difficulty occur in the local health facility, where doctors who receive medical training in English engage patients whose deepest level of understanding is in Hehe. In practice, without a common language to discuss intimate health details, patients often communicate minimal information about their condition (I have a fever, my stomach hurts), and doctors communicate minimal information about treatments (take this pill after you eat); language failure can result in a woman with gynecological cramps receiving treatment for gastroenteritis. In agriculture also, experts arrive with limited capacity to communicate their suggestions to the people who might benefit from their expertise, and limited ability to appreciate what farmers can tell about their local experiences. Communication difficulties are rampant in the schools, in forest conservation activities, in police investigations, in tax and permitting offices, and at many other interstices of the development encounter.

Malangali is not a technological center, but, like much of Africa in 2014, many residents have access to much more computing power than went to the moon on the Apollo missions. Twenty years ago, mobile phone service did not reach Malangali, and electricity for most people came in the form of car batteries that were charged at a market town forty kilometers away; a missionary not far away ran into legal trouble for using a satellite phone in contravention of the national telecom's monopoly on international communications (Benjamin 1997). Ten years ago, mobile signal could be picked up from a distant tower when standing near a certain tree at the secondary school, and the few available phones could be charged during the short evening hours when the school ran its generator to provide light in the dorms and staff housing. Today, three different cellular companies have towers on top of a nearby hill, signal is impeccable, and the security guards at the cell towers do a brisk business charging devices from the generators that keep their stations running. A mobile phone is one of the most important investments that a family makes, enabling the maintenance of economic and social networks with kin around the country (Pfaff 2010, Brinkman et al. 2009). In terms of the technological capacity, those phones are increasingly smart. In terms of language, however, the technology remains exceptionally stupid; a few services are available with some Swahili keywords, but device features such as cameras, and almost all apps, are relegated to English and other non-African languages. The full power of the Android devices that rule the market therefore largely

bypasses the Malangali residents who invest in them, whether for using core features or for accessing potentially useful services such as health, agriculture, financial, or weather information. Will this language bias continue for the next ten years? What are the implications either way? And what is the path forward for producing language resources that make technology useful for the people of Malangali, and enable them to interact more effectively with the development programs that seek to serve them?

Design and methods: The most efficient way to address language development is through public tools and vocabularies that can be reused, revised, and repurposed for multiple domains. The Kamusi Project is a universal multilingual dictionary that is designed to build a parallel vocabulary of core concepts across languages, with a special focus on languages with few existing resources. The lexicons and terminology sets are built in close cooperation with local partners. Much attention is paid to a data structure that will enable downstream technologies. Further, a system develops domain-specific terminologies through a participatory process, so that complicated concepts can be communicated clearly and consistently. Data is made available to the public for free, with strong efforts to develop systems for access via least-cost technologies with the widest reach along the bottom of the pyramid.

Is a dictionary an essential technology for development? If necessity is demonstrated by funding priorities, then the answer is a categorical "no" – very few international development agencies devote meaningful resources to language tools, training, or even translation. Partial exceptions are the United States Peace Corps, which provides language training for all its volunteers as necessary and often supports the development of language training materials by volunteers, and the US State Department that has produced public domain language material in conjunction with language training for foreign service officers. Contrary evidence, however, comes from language needs that donors feel domestically, such as the 330 million Euros that the European Union spends annually translating official documents alone (Slavcheva 2013), or the $2.4 billion that Canada spends for bilingual services (Vaillancourt & Veldhuis 2012).

Digging more deeply, we can see clear correlations between the prosperity of a society and the development of its language resources. Japan serves as a prime example:

> After the first contact with Western scholars and engineers, … education in Japan, from elementary school to university, was conducted entirely in the Japanese national language: Japanese intellectuals of the Meiji era translated into Japanese many European words such as philosophy, electricity, railway, locomotive, car, pencil, fountain pen, etc. At the beginning of the Meiji era, more than ten thousand new Japanese words were created as a way to translate Western concepts and terms. (Kawada 1988)

Rather than associating Western knowledge with Western language, Japan brought the technologies it admired into the linguistic milieu that its people understood, while becoming one of the world's major industrial powers. Russia educates its children in Russian, and its people access technology in Russian. Romania uses Romanian. China uses Mandarin and other regional tongues. Mexico uses Spanish. Switzerland uses German, French, or Italian. Though the language policies of these countries often work to the disadvantage of minority languages, they are based on the premise that most activities of most people should be conducted principally in a language that is widely understood by a broad swath of the populace. Only at the very highest levels of education and industry, where international communication is essential, are discussions of a technical nature regularly conducted in the common denominator of English.

Is it any accident that so many of the societies that are less prosperous are also those that do not have educational or technological resources in their own languages? A look around Africa shows clear linguistic/economic divisions (Osborn 2010). Generally, the small minority that speaks well-resourced international languages corresponds to the prosperous elite. Similar tendencies prevail worldwide, with high poverty among many linguistic minorities. The issues are too complex to explore in this short space – even if we could demonstrate a linear relationship between language and economic status, we could not begin to demonstrate causality in the relationship, given the many other complex historical factors at play. However, we can make a few blanket statements. First, people are undeniably more successful navigating any terrain in a language they understand. Second, most of the world's seven billion people will not become proficient in a major international language in this lifetime; those who propose that everyone should just learn English are condemning billions to linguist marginalization for many decades to come. Third, it is much easier, faster, and cheaper for a few people to prepare materials in a particular language than for all the speakers of that language to learn a foreign tongue well enough to benefit from its resources. Fourth, technology can reduce the cost, accelerate the pace, and improve the quality of producing resources for any language. Given the evident if unprovable proposition that linguistic exclusion is an important

element of economic exclusion (Brenzinger 2009, Zhang & Grenier 2012), and the relatively low costs of addressing linguistic equity through language technology (Wolff 2011), we approach the development of language resources as essential to enhancing the success of many other programs geared toward the poorest segments of the population.

The Kamusi Project addresses language needs from three directions:

1) general lexicons, with an emphasis on languages that currently have few resources
2) domain-specific terminologies
3) data for downstream technologies

All of the systems are participatory, calling on the public to share a little of their own linguistic knowledge in exchange for free access to all of the data produced by the collective. In addition, the system is sustainable in the long term; once data is collected and validated, it is safeguarded to be available continuously to the public under an open non-commercial license, with commercial interests required to pay licensing fees in order to keep the servers running.

The basic design of the system is inherently egalitarian (Benjamin 2008). Any language can be sparked in Kamusi as soon as a person emerges who can devote about two hours to fleshing out the peculiarities of the language's structure, such as the number of plural forms a noun can have. After that, the system is designed to accept contributions from anyone, with validation procedures designed to route out bad data before it becomes public. There is no technical reason to prevent a small language from having an extensive resource in its hands long before a major language gets organized.

General lexicons have an enormous range of applications. Within the scope of development, the most immediate use for dictionaries is translation. In Malangali, many encounters between English-speaking development practitioners and Swahili-speaking area residents are mediated by Tanzanian development workers with post-secondary training in fields such as agriculture. With no skilled interpreter on hand, many ideas are skipped or improperly conveyed. Many are the times that a dictionary would clear a communications impasse. Often too are the times that agencies think to prepare brochures or instructional materials, but do not translate them to local languages because they do not have the language resources. Kamusi is designed to fulfill on-the-spot translation needs, not only between primary pairs like English and Swahili, but also, through transitive internal linkages, between any pair of languages that may intersect in the field, such as Bena and Japanese (the Japanese International Cooperation Agency has conducted activity in the Malangali area). General lexicons also fill the needs of students in their classroom studies, enhancing their long term scholarly success and understanding. The secondary school for Malangali area residents has one Swahili-English dictionary and one monolingual Swahili dictionary for its 400 students, but many of those students have access to a telephone that could be used to look up information on a data network. Further, dictionaries can support microbusiness and microfinance, providing the terms to trade with buyers and suppliers beyond the local linguistic circle. In fact, the potential uses for language tools remain unexplored in the field, because the technology has not been available to even make such tools conceivable. As general lexicons become available for more languages within Kamusi, and become available to more people on more devices, consulting the dictionary will transform from an unfulfillable desire to an unremarkable regular occurrence in all aspects of development communications.

This said, the language needs of particular communities in particular situations are not necessarily known in advance. The Kamusi platform, as seen in its name, the "Global Online Living Dictionary", is designed as a form of living laboratory that can expand in response to the needs and experiences of its users. The process is not entirely democratic, because all contributions must eventually pass through the watchful oversight of a language specialist. However, all well-intentioned users who participate according to directions will find their contributions building the resource toward the satisfaction of their expressed needs. This can occur in several ways:

1) Log files record user searches on an anonymous basis. Searches that return null results are ranked by frequency and queued for editorial review.
2) User input is solicited in a variety of ways. The most direct method is for users to go directly into the editing system and add or improve entries for the terms that are of interest to them. Through user interest, for example, Kamusi has built an ornithology vocabulary for every bird known in East Africa, and is currently working on tree names that will be of use in forestry projects. We are currently working on crowdsourcing systems to elicit information from the public about their languages, including gamification; while our crowd methods (as opposed to direct editing) do not yet enable people to select their contributions according to their areas of interest, they do bring people into the production of their own language resources. Enhanced ownership and enhanced data are expected to lead to a virtuous circle of enhanced use and usability.

3) Users or groups can work on terminologies for specific domains of interest. The terminology system relies on participation under the non-democratic supervision of subject and language specialists. Terminology differs from general vocabulary, because it involves proposing new terms in the target language that match to specific concepts in a source language; the new terms must be acceptable to the user community, or they will not be adopted. The KamusiTERMS system enables experts and the community to work through a term set to achieve consensus and make the results publically accessible (Benjamin 2011). When certain programming issues are completed, the system will be opened to groups that wish to develop terminologies for their domains and languages of interest. A human rights NGO in Senegal, for example, could develop a legal terminology set for Wolof; the Wolof vocabulary would be immediately available for their translation and service needs, while the source set would be available for other groups working in other countries on other languages. While the general lexicon grows largely from a list derived from English frequencies, terminology sets can be built in response to the specific felt needs of development actors and agencies.

The Kamusi data structure is intended to support numerous downstream technologies, which will have further use value for development activities. Entries are not simply words, but rather containers that gather a large amount of information about each term. For example, all of the inflected forms of a verb will, in principle, be accessible in conjunction with the primary entry. This rich information can interact with other language technologies to enhance services such as machine translation and voice recognition. Such services will typically be defined by their developers, not their consumers. As such, technological uses stand a step removed from the concept of a living lab; if an organization interested in mobile health develops an app to transcribe and translate medical histories using project lexical data, this is unlikely to involve an introspective relationship between the organization that builds the lexical data and the patients narrating their health situation. However, even such second-order relationships can be designed to return some data or user feedback to the further enhancement of Kamusi services.

An example of addressing a symbiotic relationship between language and development needs is currently in the planning stages, in association with an initiative called PlantVillage. PlantVillage aims to provide comprehensive information about the cultivation of food crops. Their aim is to reach small farmers in tropical regions. They have a large collection of information for hundreds of plants, including information about pests, diseases, and methods of cultivation. All of this information is in English. To reach their intended audience, they need to produce their data in local languages. Local language production requires both general vocabulary and technical terminology, with terms that can be understood by farmers who are not highly literate. Consider the introductory sentences about cassava cultivation:

> Cassava, *Manihot esculenta*, is a perennial shrub in the family Euphorbiaceae grown primarily for its storage roots which are eaten as a vegetable. The cassava plant is a woody plant with erect stems and spirally arranged simple lobed leaves with petioles (leaf stems) up to 30 cm in length. The plant produces petal-less flowers on a raceme. (PlantVillage 2014)

Many words, such as cassava, storage, and vegetable, are part of a general vocabulary that will be useful to any dictionary user. Other terms, such as perennial, family, and raceme, must be treated as terminology specific to agriculture or botany. It is important to have all of these items in the dictionary in a consistent format because of the next step in the process, working with Translators Without Borders (TWB) to produce the PlantVillage documents in a variety of local languages. Without a consistent vocabulary set, different translators will make independent stabs at the relevant terms, with the results that will sow endless confusion among end users. With a unified approach, both translators and users will be able to look up terms in both the source and target language, including the definitions in both languages. This information will be available from both the PlantVillage and Kamusi websites, with pathways that let PlantVillage users learn more about language and Kamusi users to learn more about plants.

The needs of the collaboration between PlantVillage and Kamusi will not initially be defined by the farmers themselves, but rather by the university-based personnel of the two organizations. Further input will come from the volunteers working for TWB. Farmers will have access to the data via the mobile devices they work so hard to afford. Whether farmer feedback enters the project, such as calls for more graphics, simpler text presentation, or more vocabulary, will partially depend on the design of the services that PlantVillage offers. The level of interactivity with the end user will vary from one technical partnership to the next, with the potential ripe to grow from consumer feedback in many instances.

Results: When successful, a focus on core language development can improve the outcomes of many other projects. In health, for example, translation is often too expensive and too difficult, because basic resources such as

dictionaries do not exist and technical terms do not have adequate local-language equivalents. For the one-time cost and effort of building the lexicons and terminologies, in conjunction with the free tools being created to access those vocabularies, the infrastructure opens for cheap and rapid translation of health material. Similarly, students are able to use the lexicons to access knowledge that has previously been blocked behind linguistic barriers, reducing future language-based inequalities. Participatory data collection methods mean that vocabularies continue to grow in response to the expressed needs of particular linguistic and development communities. The perpetual accessibility of the public resource means that development programs can make sustainable use of the data in multiple languages, for multiple purposes, with no further investment.

It is too early, as Kamusi moves from data design to data collection, to discuss tangible results. The major product available to the public currently is the general-purpose Swahili dictionary, and we know that Kamusi has a large usership in East Africa that has been growing steadily as internet access has improved in the region. We take it as a given that speakers of other languages will make extensive use of the resource when significant services are available to them. However, while we will be able to discuss metrics such as hits and word counts, it is not obvious how one could quantify measures of impact. How much will the tomato crop in Malangali improve because farmers had the linguistic tools to participate in a discussion thread about highland cultivation techniques? What is the impact on infant mortality because of an app that improves the ability of doctors and patients to discuss symptoms and cures? Such questions cannot be answered numerically. Case studies from the field may provide some indication of success, but will more likely point to aspects that need improvement; the impact of the availability of linguistic data for a language spoken by millions cannot be encapsulated in a few stories about students using the product to improve their research, while the impact of not having a vocabulary set for emergency service providers in an area hit by an earthquake will be evident in the negative.

Language technology is analogous to electricity. One can quantify the amount of electricity that people use, but there is no way to measure its impact on the lives of its users. Electricity users might benefit from charging mobile devices, from refrigerating their food, or from being able to study into the night – the fact that global impacts cannot be measured by no means indicates that development efforts should shy away from bringing solar panels to rural households. Similarly, the goal of Kamusi is to wire the world for linguistic access. What follows will be interesting, but out of the project's ability to monitor or control.

Conclusion: Language is a hidden aspect of the development equation; language technology in itself does not cure a disease or put food on a table. However, whether communicating agricultural techniques, delivering government services, or performing numerous other activities that fall under the rubric of development, attention to developing language technologies for underserved language populations can be the difference between working together and talking past each other – the difference between failing to communicate and succeeding in expressing the path toward accomplishing common goals. The project under discussion proposes producing core general vocabularies of 20,000 terms for hundreds of languages throughout the developing world, as well as domain-specific terminologies for topics of special interest to development agencies active in the field. Creating language resources for users at the base of the pyramid is currently a priority for almost nobody – it is considered something that would be nice to have, but not anything that most organizations would invest in or take on for themselves. In the absence of a concerted effort to fill the void in language data and services, language inequity will continue indefinitely, and many billions of people will be excluded from knowledge and services purely based on what they speak. We propose that addressing communications deficiencies through a comprehensive focus of technology on the languages spoken by the consumers of development programs will improve both the programs and the overall prospects for people to prosper in an increasingly technologically-oriented global economy.

References:

Benjamin, M. (1997). Malangali and the Cyberians: Reflections on the Internet Living Swahili Dictionary. *Africa Today*, 44 (3), 339-356.

Benjamin, M. (2000). *Development Consumers: An Ethnography of "The Poorest of the Poor" and International Aid in Rural Tanzanzia*. PhD Dissertation, Yale University. New Haven.

Benjamin, M. (2008). PALDO: Moving Toward a Unified Dictionary of African Languages. In D. Prys & B. Williams (Eds), *Global Understanding in Multilingual, Multimodal, and Multimedia Contexts* (pp. 119-126). Conference Proceedings, GUM3C, Bangor University.

Benjamin, M. (2011). Toward a standard for community participation in terminology development. In *Proceedings of The First Conference on Terminology, Language, and Content Resources* (pp. 86-96). Seoul, Korea.

Brenzinger, M. (2009). Language Diversity and Poverty in Africa. In W. Harbert (Ed.), *Language and Poverty* (pp. 37-49). Bristol: Multilingual Matters.

Brinkman, I, Bruijn, M, & Bilal, H (2009). The mobile phone, 'modernity' and change in Khartoum, Sudan. In M. Buijn, F. Nyamnjoh, & I. Brinkman (Eds.), *Mobile phones: The new talking drums of everyday Africa* (pp. 69-91). Leiden: African Studies Centre.

Kawada, J. (1988). Development and Culture – is Japan a Model?. *Discussions of the Inaugural Programme of the Africa Leadership Forum*. Ota, Nigeria, p.25

Osborn, D. (2010). *African Languages in a Digital Age*. Cape Town: Human Sciences Research Council and Ottawa: International Development Research Centre.

Pfaff, J. (2010). A mobile phone: mobility, materiality and everyday Swahili trading practices. *Cultural Geographies*, 17 (3), 341-357.

PlantVillage (2014). Cassava (manioc). Resource document. PlantVillage. https://www.plantvillage.com/topics/cassava-manioc/infos. Accessed 23 January 2014.

Slavcheva, D. (2013). Translation in the European Union - Facts and Figures ; The translation flow is an enormous challenge to the EU. http://one-europe.info/translation-in-the-european-union-facts-and-figures. Accessed 12 January 2014.

Sinha, C. & Hyma, H. (2013). ICTs and Social Inclusion. In L. Elder et al. (Eds.), *Connecting ICTs to Development, The IDRC Experience* (pp. 91-116). Ottawa : Anthem.

Vaillancourt, François and Niels Veldhuis, 2012. Federal and provincial bilingualism requirements cost Canadian taxpayers $2.4 billion annually; provinces spend $900 million to provide dual-language services, January 16, 2012, http://www.fraserinstitute.org/publicationdisplay.aspx?id=2147484098

Wolff, F. (2011). *Effecting Change Through Localisation*. Pretoria : Translate.org.za.

Zhang, W. & Grenier, G. (2012). How can language be linked to economics ? A survey of two strands of research. Working Paper #1206E, Department of Economics, Faculty of Social Sciences, University of Ottawa.