ỌBÁFẸ́MI AWÓLỌ́WỌ̀ UNIVERSITY

Ilé-Ifẹ̀, Nigeria

Joint Collaboration Between
Computer Science & Engineering and
Linguistics & African Languages

Contents

Abstract

This proposal documents a research collaboration between the Computer Science and Engineering Department of the Faculty of Technology and the Linguistics and African Languages Department of the Faculty of Arts, both of the *Ọbáfẹ́mi Awólówọ̀* University, *Ilé-Ifẹ̀*. This research collaboration is motivated by two interrelated interests. First, the need to explore the possible contributions of African languages and culture to modern computing. Second the potentials for the development of technologies for improving African indigenous languages users' experiences of modern Information and Communication Technologies. The collaboration will provide a platform for manpower development and computer based applications generation.

1 Introduction

The ubiquity of computing and the pervasive use of Information Communication Technologies (ICT) in modern times present a number of opportunities and benefits to human users. However, a number of challenges accompany these benefits and opportunities. Some of these challenges relate to the ease of use and the consequent effects of the technologies on indigenous languages, culture, knowledge resource and social interactions. The need to put in place a systematic and proactive mechanism for properly responding to these challenges emerges in the context of current trends in globalisation.

In respect of the social implication of of ICT, experience in recent times has shown that the Internet, particularly social network and communication tools, mobile phones, etc. have great potential to create profound changes to the character and evolution of African indigenous culture and language. The use of ICT in this context provides opportunities and challenges for African Indigenous Languages AIL. This is in the background that human languages remain the most potent means of expressing and solving complex tasks. The ability to receive and carry out instructions given in the form of human language sets the computer technology apart from those before it. For example, the explosion of web-based information services, commerce and business as well as mobile phone-based dialogue, are vital to functioning of modern society. Human languages remain the medium for developing, deploying and exploiting these technologies.

We think that two critical issues are crucial to appropriately responding to these challenges: (i) Manpower development and (ii) Language resource and application development. To this end, a collaboration between relevant research and academic departments becomes important. This document set out the modality for that collaboration.

1.1 Aim of the Collaboration

The aim of this collaboration is to : "*address intellectual, academic and developmental challenges relating to African Indigenous Languages (AIL) and culture in the context of modern Information Communication Technologies*"

Appropriate research and development activities will be deployed achieving this aim. Such activities will comprise:

1. Those that explore the use of theories from various aspects of African Indigenous Languages (AIL) studies, including: linguistics, phonetics and phonology, orthography, etc. in the development of new computing techniques and devices.

2. Those that develop specific technologies for facilitating the use of African indigenous languages, including, in the aspect of user interfacing, information sharing, processing and language translation.

1.2 Objectives

The specific objectives of this research collaboration are captured within three keywords: Conceptualisation, Development and Share (CDS). These are expanded as follows:

1. Conceptualisation: Identify, analyse and document various aspects of African indigenous languages and culture, particularly those with potential for the development of computing artefacts.

2: Development: Specify, design, implement, evaluate and execute relevant research and development activities to realise 1.

3. Share Deploy appropriate pedagogical and communications instruments and techniques for demonstrating and sharing the products developed in 2.

1.3 Justification

The study of African indigenous knowledge and the development of language technologies for African indigenous languages (AIL) becomes necessary in the context of modern trends in ICT. Except for a few languages, such as *Kwashili* and *Hausa*, majority of over two thousand five-hundred $(2,500)$ languages in Africa are categorised as *low resourced languages* [9, 6] (LRL). LRL is a term used to describe languages with none or limited digital resource of the quality and quantity required for the development of language technologies (cf. Table 3). In speech and language research, especially in the development of language technologies products, progress is largely dependent on commonly available large language resource [9]. Aside from these, the specific justification for the research activities in this collaboration include the following.

1. Wider access to global information resource: The development of language technologies, such as machine translation (MT) systems, for AIL has the potential of improving access to database and information provided on the internet as most of these information and databases are written in languages that are not indigenous to Africa. The effect of this will be appreciated if one notes that more that 60% of the population, and in some cases larger, are not literate. However these indigenous people do engage in intellectual discourse and can express themselves very effectively using their indigenous languages. A combination of Machine Translation (MT), speech synthesis and speech recognition techniques can, for example, provide a new information and communication paradigm. Within such paradigm, one does not need

to be literate to access modern information infrastructure. Such paradigm will redefine the process of education and enlightenment.

2. **Contribution to global information infrastructure:** Intellectual resources that are embedded in African culture and knowledge systems, which can further enrich global information, will remain inaccessible to the global communities due to the fact that the people with these knowledge cannot communicate them effectively without appropriate technologies. Such barrier can be reduced or eliminated using language technologies thereby facilitating the indigenous peoples' contribution to the global information infrastructure.

3. **Computing tasks simplification:** Ability to exploit the power of modern technology provides the potential of reducing the difficulties encountered during, for example, commerce and banking transactions by indigenous people. This tasks will become seamless if people can perform them using their native languages.

4. **Educational applications:** The educational benefits that is derivable from language technologies are unquantifiable. Aside from producing language education materials for students and teachers alike in their indigenous languages, it will also be possible to improve access to education through the use of mobile devices that have become ubiquitous in most African countries. The benefit of this will be greatly felt among the nomadic populations in Africa.

5. **Language preservation:** The problem confronting African Indigenous languages in the modern era of globalisation have been extensively discussed [3]. The development of language technologies for AIL will improve their functional loads by improving their usage in day-to-day tasks and inter-personal interactions. The development of language technologies will no doubt reduce the risk of the languages going into extinctions. Ability for people to issue commands to the computer in their indigenous languages demystified the computer and makes it more readily accessible as a tool for communication.

1.4 Scope

The task of conducting research and developing technologies for all AIL in one sweep is a project that requires skills and resource beyond a small project group such as ours. To this end, our work will be focused on a class of AIL in the *Edekiri* [7] language family. However, the Yorùbá language will be the center of most of our studies. Members of the *Edekiri* family of languages are listed in Table 1. This family of languages are indigenous to West African countries.

A common feature of the *Edekiri* languages is that they are *tone languages*. They are so called because lexical tones play significant role in the semantics of the written and spoken forms of the languages [13, 11]. It is important to note that all human languages use tone at the sentence level to convey some para-linguistic information such as the mode (e.g. question, exclamation) and mood (happy, sad) of an expression. In addition to these, tone languages also use tones to function at lower levels as its speakers are able to perceive more granular tonal signatures. For instance, it has been shown that native language speakers of a tone language

Table 1: The Edekiri Languages (Source [7])

Name :	Edekiri
Code :	edki
Code Standard :	LINGUIST List
Documentation :	Private Use
Families :	Niger-Congo (Niger-Kordofanian)
Parent Subgroup:	Yoruboid; c (yrbd)
Child Languages:	Yorùbá; Yoòbá; Yariba; Ede-Yorùbá (yor) Ifé; Ana-Ifé; Ana; Baate; Ana-Ife; Anago; Ede Ife (ife) Ede Ica; Ica (ica) Isekiri; Itsekiri; Ishekiri; Shekiri; Jekri; Chekiri; Iwere; Irhobo; Warri; Iselema-Otu; Selemo (its) Lucumi (luq) Ede Nago; Nago; Nagots; Nagot (nqg) Ede Nago, Kura; Nago (nqk) Ede Ije; Holi; Ije (ijj) Ede Cabe; Caabe; Cabe (cbj) Ede Idaca; Idaca; Idaaca; Idaasa; Idáítsá (idd) Ulukwumi (ulb) Ede Nago, Manigri-Kambolé; Manigri; Ana; Southwest Ede; Kambolé (xkb) Mokole; Mokollé; Mokwale; Monkole; Féri (mkl)

are better able to imitate (through singing) and perceptually discriminate musical pitch [8].

Tone, whose main acoustic correlate is the fundamental frequency, and its associated timing and intensity, is anchored with a syllable. In tone languages, tonal signatures are applied at the syllable level to convey and generate semantic variations that serve role of determining the meaning of an utterance. For example the *Yorùbá* words *rí* (to see), *ri* (to weep uncontrollably), *rì* (to sink) are phonetically identical but semantically different because of the tone associated with each syllable (see Table 2). In a carefully articulated multi-syllable word, each syllable will carry a distinct tone. A change in the tone associated with a syllable changes the meaning of the word. For example, the words: *igbá* (calabash), *igba* (two hundred), *ìgbà* (season, time, period), *igbà* (climbing rope) have different meaning because the tone associated with its constituent syllables.

Table 2: Sample of tone on mono-syllabic expressions

Syllable	L	M	H
Ri	*rì* (sink)	*ri* (weep)	*rí* (see)
Ki	*kì* (praise)	*ki* (tick)	*kí* (greet)
Mi	*mì* (shake)	*mi* (mine)	*mí* (bread)
Si	*Sì* (fade)	*si* (doll)	*sí* (open)

These features are common to all the tone languages but they differ in types and scale as well as along the following dimensions:

1. The number of distinct and discrete tones in the language. For example, in *Yorùbá*, there are three(3) tones and in *Igbo* there are two(2).

2. The co-articulation behaviour of the tones in fluent speech.

3. Tono-tactics, the rule guiding or constraining the pattern of occurrence of tone in the language. For example some tones cannot follow or be preceded by other specific tone in some lexical construct.

4. Phono-tactic the rule guiding the pattern of occurrence of phonetic elements in an utterance

5. Syllabi-tactic the rule guiding the formation of syllables in an utterance. It is a combination of the tono-tactic and the phono-tactic rules.

Our study will be in the remit of characterising, analysing, synthesising and in other ways generating knowledge about the languages by borrowing from more well research and technologically developed languages such as English. The generated knowledge will be used to develop software and other computing artefacts.

1.5 Methodology for Engagement

To achieve our objectives, we will pursue the following methods.

1.5.1 Methods for Objective 1

The first objectives relates to conceptualisation. Our method will be to use the forum of intellectual engagement between members of the collaboration team to generate ideas and concepts. These engagement will be in the form of intellectual discussions and brainstorming sessions on possible research and development ideas. The ideas agreed upon will be compiled and a research proposal and proof concepts will be written. The collaborating team will invite and co-opt resource persons from sister Departments and Faculties such Departments of Philosophy, Psychology, Mathematics, Education and Curriculum Development, whenever necessary.

1.5.2 Methods for Objective 2

The second objective will be achieved through application, specification, design, implementation, evaluation and deployment. Whenever a software will be developed, we will adopt the open source software development platform based on the *Ubuntu* operating system. This is to enable our computing artefacts and software to be freely available without strict copyright restrictions.

1.5.3 Methods for Objective 3

Two methods will be deployed in this regard. The first will be targeted at manpower development in the aspect of students' and staff training as well as interdisciplinary

student project supervision. A staff and student information exchange and education scheme focusing on knowledge acquisition and update will be implemented. Project topics that addresses issues of concern to the collaboration team will be generated for joint supervision. The second aspect of our method for development will comprise the sharing of the research finding and product through interactions between members of the collaborating team and other interested parties. We will host: (i) Meetings, (ii) Seminars, (iii) Public and course based lectures, (iv) Workshops, (v) Product show and (vi) Conferences. The collaborating team will create and maintain a publication that will be named: "*African Languages & Technologies: Work in Progress*". This will be an open publication that will be available over the internet. This publication will document our ongoing research and development activities as well as other relevant information relating to the team. Matured work will also be written up and submitted for publication in reputable national and international journals and conferences.

2 African Indigenous Language Technologies

We define *human language technologies* as those that exploit the knowledge of human languages in its development or those, whose, through their applications, enhance the use of human languages. Technologies that exploit human language in their development are built based on some theory of human language or human communication system. Such knowledge arises from the scientific studies of languages. Applications of the knowledge of human languages abound in computing. This includes computer programming languages, computer system design and development techniques, formal language and methods as well as computer communications and resource sharing protocols. The knowledge from the study of human languages is used to create and improve computer technology while the products of computer technology helps to enhance the use of human languages. Therefore, *the computer* and *human languages* are in a symbiotic relation whereby the knowledge from one feeds into the other.

It is easy to take for granted human computer interaction through medium such as an operating system prompts. Without a good knowledge of the items in the prompts, commands and icons, it will be difficult to put the system to a fruitful use. These items are expressed using different forms of human languages and graphical illustrations. A project on localisation of the Microsoft Vista [1] operating systems suggested that language use in computer interface is fundamental to improving user experience.

A list of some human language technologies is provided in Table 3. Most of the technologies listed in Table 3 can be deployed through various platforms including: (i) Stand-alone systems and networks, (ii) the Internet and (iii) Mobile applications including, mobile telephones.

2.1 Some Technologies Relating to Specific Language Medium

The development and deployment of *human language technologies*, as is the case in any other technologies, require tangible and intangible resources. These should normally include the materials for their conceptualisation, specification, design, implementation, evaluation and deployment. It is interesting to note that, what

Table 3: Language Technologies

Ser. No	Technology
1.	Digital mono-lingual dictionary
2.	Digital multi-lingual dictionary
3.	Speech recognition systems
4.	Speech synthesis system
5.	Machine translation systems
6.	Digital mono-lingual dictionary
7.	Text processor- spell checker, grammar checker, morphological analysers Lexical analysers, etc.
8.	Automatic Dialogue systems

counts as *language technology* product in a context can become a resource for other language technologies. For example, a dictionary can becomes a resource for a machine translation system.

The collaboration team will consider the following language medium as the object of indigenous language studies: Speech (oral), Text (orthography) and Sign (graphical). We shall also consider a combination of these medium in a hybrid context. The language technologies related to the mediums are listed below.

Speech:
1. Speech recognition
2. Speaker recognition
3. Spoken language recognition
4. Speech control systems
5. spoken language translation system

Text:
1. Text to speech synthesis
2. Text summarisation
3. Machine translation (text to text)
4. Text dialogue systems
5. Automatic story (narrative) generation system
6. Text processor

Sign
1. Gesture recognition system
2. System state to speech synthesis
3. Sign language to speech or text generation
4. Sing based control systems

Hybrid: Combines any two or more of the above

1. Multi-media (text, audio, image) dictionary

2.2 Issues in African Indigenous Languages Technologies

The following are peculiar issues relating to research challenges for African Indigenous languages (AIL):

Large number: Africa has more than two thousand five hundred $(2,500)$ languages. Some of these languages have diverse dialects that are sometimes not mutually intelligible. For example, there are about two hundred and fifty (250) indigenous languages in Nigerian alone. One of these languages, i.e. *Yorùbá* [2], has more than ten (10) dialects. It is interesting to also note that, in Africa, *Yorùbá* is natively spoken in two other countries namely: Benin (Dahomey) and Togo. This issue is further complicated by the fact that the exact number of the indigenous languages in Africa and their corresponding dialects is not known definitively as this is still a subject of research.

Documentation: This is poor due to a number of reasons, including inadequate orthographic system, low literacy, non-utilisation in formal education, etc.

Functional load: Low functional load as a result of reducing capacity of modern technology, unofficial status [10, 3].

Expert and manpower: The experts and manpower on AIL are either low or not available in most cases. The works of the few that are available is yet to crystallise into tangible resource or standards due to lack of coordination on the one hand and inability to reach consensus on important issue, such as theory and research strategies, on the other hand.

Many different genre and expression forms: Some AIL have many different forms of expressions. For example, aside from the normal usage form of expression, the *Yorùbá* language has many other forms such as incantations (*ọfọ̀*, *èpè*, *ògèdè*, *Ìyẹ̀rẹ̀*, etc.), praise (*orírìkì*), proverb (*òwe*), deep expression (*ìjìnlè ọ̀rọ̀*), analogies (*àlọ́*), etc.. Each of these require different levels of competences in the language.

Language resource availability: All the above issues militate against the development and deployment of a viable high quality publicly available language resources for the majority of AIL. There are private collections in research laboratories or academic institutions for specific projects. It is not known whether efforts to make these public are under-way. Even when or whether that effort will be achievable will depend on agreeing on a standard for collection, annotation and storage.

There has been a number of works reported on the development of language resource for language technologies applications [5, 4, 12]. In [5] a large-scale Japanese speech database that serves as language resource for speech recognition and synthesis was presented. The work provides materials for studies into the acoustic, phonetic and linguistic evidence that served as basic data for speech technologies.

3 Courses to be floated by the collaboration

To achieve the goals of our third objectives, the collaboration team will introduce the following courses to be taught at undergraduate and postgraduate levels.

3.1 Undergraduate level courses

The following undergraduate level courses will be floated by the Collaboration team:

U1. Fundamentals of Computing for Language Studies This is an introductory course that present computing as a tool for language data capturing, storage and processing. The course will explore the fundamental principles of modern computer systems including software and hardware in the context of African languages studies.

U2. Introduction to Computational Linguistic This course will introduce the concept of computing as a tool in the linguistics analysis of African languages. Various software tools that are commonly used and are freely available for computational linguistics will be demonstrated. Course U1 is a prerequisite to this course.

U3. Corpus Development for language engineering The development of practical speech and language technologies depends on a robust, well designed and properly deployed language resource. This course will explore how such resources are developed. The instruction will be demonstrated using African Indigenous languages.

3.2 Postgraduate level courses

The following Postgraduate level courses will be floated by the Collaboration team:

P1. Philosophy and Theories of Computing in African Culture: This course will explore and discuss various aspects of the philosophy and theories underlying computing in the context of African cosmology. The concepts of data, their creation, storage and organisation, manipulation, numbering and numeration, computational models of language, games, timing and calendar, message and information communication, the structure of discourse and intellectual engagements, etc in modern computing science will be explained from an African perspective drawing examples and analogues from indigenous languages.

P2. African Culture and Artificial Intelligence: This course will look into the fundamental concepts of Artificial intelligence, from its basic definitions to the various sub-fields and the techniques employed in the development of modern intelligent systems. Issues such as the concept and process of decision making, NP problems and games will be discussed in details using case example from African culture.

P3. Computation Principles of Patterns and their Design in African Culture: This course will discuss the general principles of patterns in terms of their perception, creation, expression and use and in African settings. The computational principles underlying these topics will be discussed in the context of patterns on African fabrics, building and works of arts as well as markings and signs. Fractals and related patterns will also be discussed.

P4. African Languages and Language Processing Technologies: This course will present and discuss fundamental and advanced issues in human languages technologies using African indigenous languages as case studies. Topics to be covered will include, but not limited to, the fundamental concept of language and grammar, lexical and morphological analysis, word sense disambiguation, speech synthesis, speech recognition, abstract communication devices (e.g. the talking drum), machine translation, etc. Automata theory based techniques such as finite state transducers will be used to demonstrate the design and implementation of selected applications.

4 Funding

Our funding profile will have a low foot print as we will endeavour to use the most cost effective mean to implement our tasks while at the same time not compromising academic quality and intellectual rigour. The team will put in place a mechanism to respond to call for research funding proposals both locally and internationally. Team members will also be encouraged to bear personal, but affordable, costs that may arise as a consequence of their research activities, though this will be exception rather than the rule.

5 Team Members

Table 4: Team Membership

Ser. No.	Name.	Function
1	Prof. Adéwọlé, L. O.	Team leader and convener
2.	Dr. Ọdẹ́jọbí O. A.	Coordinator Software development
3.		
4.		
5.		
6.		
:	:	:

References

[1] T. Adégbọlá, K. Owólabí, and O. A. Ọdẹ́jọbí. Localising for Yorùbá: Experience, challenges and future direction. In *Conference on Human Language Technology for Development*, pages 7–10, Alexandria, Egypt, May 2011.

[2] O. L. Adéwọlé. *The Yorùbá language Published works and doctoral dissertations 1843-1986 - African linguistic bibliographies*. Helmut Buske Verlag, 1987.

[3] A. Bámgbóṣé. African languages today: The challenge of and prospects for empowerment under globalisation. In E. G. Bokamba, editor, *Cascadilta Selected Proceeedings of the* 40^{th} *Annual Conference on African Linguistics*, pages 1–14, 2011.

[4] S-Q. Chen and L. Xu. A full and efficient machine tractable dictionary for natural language processing: A revised version of the cuvoald. *Computers and the Humanities*, 28(3):141–152, 1995.

[5] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano. ATR japanese speech database as a tool of speech recognition and synthesis. *Speech Communication*, 9:357–363, 1990.

[6] T. McEnery and A. Wilson. *Corpus Linguistics: An Introduction*. Edinburgh University Press, University of Edinburgh, 2^{nd} edition, 2001.

[7] MultiTree. Multitree: A digital library of language relationships. the URL:http://multitree.org/codes/edki, 2009. retrieved: 14-Oct.2012.

[8] P. Q. Pfordresher and S. Brown. Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics*, 2009.

[9] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes. The design and collection of COSINE, a nulti-microphone in situ speech corpus recorded in noisy environment. *Computer speech and Language*, 26:52–66, 2012.

[10] C. Taylor. Typesetting African languages. the URL:www.conradiator.com/downloads/pdf/Afrolingua_full.pdf, 2000.

[11] Y. Xu, J. T. Gandour, , and A. L. Francis. Effects of language experience and stimulus complexity on the categorical perception of pitch direction. *Journal of the Acoustical Society of America*, 120:1063–1074, 2006.

[12] D-H. Yang, I-H. Lee, and P. Cantos. On the corpus size needed for compiling lexicon by computational comprehensive lexical acquisition. *Computers and the Humanities*, 36:171–190, 2002.

[13] M. Yip. *Tone*. Cambridge University Press, Cambridge, 2002.