Building Capacities in Human Language Technology for African Languages

Tunde Adegbola
African Languages Technlogy Initiative (Alt-i), Ibadan, Nigeria
taintransit@hotmail.com

Abstract

The development of Human Language Technology (HLT) is one of the important by-products of the information revolution. However, the level of knowledge and skills in HLT for African languages remain unfortunately low as most scholars continue to work within the frameworks of knowledge production for an industrial society while the information age dawns. This paper reports the work of African Languages Technology Initiative (Alt-i) over a five-year period, and thereby presents a proposal for the acceleration of the development of knowledge and skills in HLT for African languages.

1 Introduction

The world is undergoing a transformation from industrial economies to an information economy, in which the indices of value are shifting from material to non-material resources. This transformation has been rightly described as a revolution because of the height of its pace and intensity. Of necessity therefore, the response to the changes brought about by the information revolution has to be commensurate, both in pace and intensity.

At the root of the information revolution is the development of digital technology which has brought about a major shift in the way we conceptualize, describe and anticipate our world.

One of the salient social imperatives of the information revolution is the need for humans to communicate through and with machines. This has brought about the need to make machines capable of handling natural language used by humans as against formal language used by machines. The field of Human Language Technology (HLT) was developed to provide the necessary knowledge and skills that will enhance the effectiveness and efficiency with which machines mediate communication between humans as well as facilitate communication between humans and machines.

So far, developments in HLT have not sufficiently addressed African languages. This is attributable to a low level of awareness of the importance of and lack of interest in HLT and among scholars of African languages and scholars of technology on the African continent. Furthermore, there is little or no immediate economic incentive in working in HLT. Consequently, there is a dearth of scholars with the requisite impetus, knowledge and skills to support the development of HLT for African languages. Hence, even though there are pockets of activities in Africa, the level of development of HLT for African languages remains low.

The circumstances that necessitated the development of HLT have been described, and rightly so as a revolution. The response therefore has to be commensurate both in pace and intensity. Hence, the need for accelerated development of HLT for African languages is urgent.

This paper presents a proposal aimed at accelerating the development of HLT for African languages based on the experiences gathered at African Languages Technology Initiative (Alt-i) over a five-year period of activities mainly in Nigeria.

2 The State of Language technology in Africa

The application of Language technology to African languages is relatively new and most efforts seem to be incidental. The most consistent efforts motivated and guided by national policy come from South Africa while projects in other countries are based primarily on private initiatives. In a report on HLT development in Sub-Saharan Africa, Justus Roux (2008) reported nine organizations involved in HLT activities in South Africa, one organization in West Africa and two in East Africa. Seven out of the nine organizations in South Africa are based in universities, one is a Semi-Government institution and one is an agency of the Government.

The seven universities with HLT projects are:

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 53–58,
Athens, Greece, 31 March 2009.

- University of Cape Town
- University of Limpopo
- University of the North West (Potchefstroom)
- University of Pretoria
- University of South Africa
- University of Stellenbosch
- University of the Witwatersrand (Johannesburg)

The Meraka institute is semi-governmental while there is a Human Language Technology Unit under the Department of Arts and Culture of the Government of South Africa.

In West Africa the only organization reported is Africa Languages Technology Initiative (Alt-i), while the two reported in East Africa are The Djibouti Center for Speech Research and Technobyte Speech Technologies in Kenya.

Apart from the organisations reported above, there are individual efforts in some universities. These include Dr. Odetunji Odejobi working on Text to Speech Synthesis at Obafemi Awolowo University, Ile Ife, Nigeria, Dr. Wanjiku Ng'ang'a working on Machine Translation and Dr. Peter Wagacha working on Machine Learning, both at the at the University of Nairobi, Kenya.

Apart from these organizations and individuals that are strictly located in Africa, there are a number other efforts in various parts of the world that address HLT for Africa languages, usually in cooperation with some organizations in Africa. Examples include:

- Local Language Speech Technology Initiative (LLSTI), a project of Outside Echo in the UK
- West African Language Documentation, a project of the University of Bielefeld, Germany in collaboration with the University of Uyo, Nigeria and the University of Cocody, Cote D'Ivoire.

Also, there are significant short-term activities on language technology for African languages both within and outside Africa which have not been sufficiently publicized. For example, in 2002, there was an undergraduate project in Yoruba-English machine translation at the St Mary's College of Maryland, USA.[1]

[1] It is also necessary to mention the efforts of Bisharat and SIL. Even though both organizations are not strictly founded for developing language technology for African languages, they have both done important work in making various resources and tools for developing language technology for African languages available.

3 On-going Alt-i Activities

Since its inception, Alt-i has done more work in Yoruba than in other African languages. This is due primarily to the ready availability of intellectual and other resources for Yoruba at Alt-i's base in Ibadan. However, work in a few other languages with available local resources have also been undertaken. The main projects undertaken so far are:

3.1 Automatic Speech Recognition (ASR) of Yoruba

The ASR project started in 2001, and is still ongoing. A PhD thesis with the title of Application of Tonemic Information for Search-Space Reduction in the Automatic Speech Recognition of Yoruba is one of the results of the ASR project. The project approaches ASR of Yoruba from the point of view that Yoruba tones carry so much information that "talking drums" can "speak" the Yoruba language, hence, ASR of Yoruba (and probably other African tone languages) should be based primarily on tones or at least should address considerable computational resources towards correct identification of tones. Experiments have shown that a tone-guided search of the recognition space as proposed in the above PhD thesis leads to improvement in recognition speed and accuracy. ASR efforts are continuing within the project "Redefining Literacy"; Alt-i's main on-going project funded by the Open Society Initiative for West Africa (OSIWA).

3.2 Text to Speech (TTS) synthesis of Yoruba

This project was conceived and initiated in 2002. TTS is a major component of the "Redefining Literacy" project but we have not yet succeeded in attracting funding for this component of the project. Hence, it is in abeyance. However, one of our Associates, Dr. Odetunji Odejobi of the Department of Computer Science and Engineering of the Obafemi Awolowo University, Ile Ife is actively working on TTS and we are collaboration with him on this project. Dr. Odejobi applies fuzzy logic to formalise Yoruba prosody.

3.3 Machine Translation

Work is on-going on Igbo-English and Yoruba-English Machine Translation. The machine translation projects are not funded at present, but

they are taking advantage of the efforts of student volunteers from the Department of Linguistics and African Languages as well as the Africa Regional Center for information Science, both at the University of Ibadan. The main thrust of the present stage of the project is identifying and developing formal specifications for various challenges of a rule-based machine translation as it relates to translation between Igbo/Yoruba and English.

3.4 Yoruba Spelling checker

As a member of the African Network of Localizers (AnLoc), Alt-i is developing a spelling checker for Yoruba in Open Office. This has provided the opportunity to undertake a computational study of Yoruba morphology. Staff and students of the Department of Linguistics and African Languages at the University of Ibadan are playing an active role in this project. The work is producing new insights for interpreting the existing literature of Yoruba morphology and is already leading to interests in similar projects for other languages. As at the time of writing, a dictionary file of about 5000 Yoruba root words and over 100 highly productive affix rules have been developed. Even though some important Yoruba morphological rules cannot be efficiently coded in Hunspell (the software on which the spelling checker is based) the modest dictionary and affix files have produced a useful spelling checker. The project is funded by the International Development Research Center (IDRC) of Canada.

3.5 Automatic diacritic application for Yoruba

As a by-product of the Yoruba spelling checker project, Alt-i is developing an automatic diacritic application program, using the Bayesian learning approach. This project is not funded, but it is taking due advantage of some of the resources, particularly the corpus produced in the IDRC funded spelling checker project. Work is ongoing to expand the corpus used for the automatic diacritic application program.

3.6 Localization of Microsoft Vista and Office Suite

Alt-i was appointed by Microsoft as moderators for the localization of Microsoft Vista and Office Suite into Hausa, Igbo and Yoruba. This project is making steady progress.

3.7 Academic assistance to the University of Ibadan

Apart from the above projects, Alt-i offers academic support to the University of Ibadan. The Executive Director of Alt-i teaches post-graduate courses in Artificial Intelligence and Information Networking as well as supervises post-graduate projects at the Africa Regional Center for Information Science (ARCIS) in the University of Ibadan. He also gives various levels of support in the supervision of post-graduate projects, particularly in the area of acoustic analysis of speech at the Department of Linguistics and African Languages.

Some of the HLT issues addressed in the Master in Information Science projects between 2002 and the present include:
Statistical Language Model (SLM) of Yoruba, Man-Machine Communication in Yoruba, Machine Translation between spoken and signed Yoruba for the deaf and impaired in hearing, Yoruba phonology multimodal learning courseware and phonetically motivated automatic language identification.

Many PhD students of the Department of Linguistics and African Languages have also enjoyed intellectual support and use of Alt-i's speech laboratory in their studies

3.8 Support to other universities and scholarly associations

Many staff members and students from far and wide travel to Ibadan to use Alt-i's speech laboratory. PhD students as well as faculty members from the University of Lagos, University of Ilorin, University of Benin and University of Abuja come regularly to use the facilities. At present a teaching staff of the Department of Systems Engineering of the University of Lagos is undertaking a PhD programme on ASR of Yoruba. This PhD candidate visits Ibadan frequently and regularly to use Alt-i's library and speech laboratory, as well as consult Alt-i staff.

Alt-i is involved in the activities of various scholarly associations such as the West African Linguistic Society (WALS), Linguistics Association of Nigeria (LAN) and the Yoruba Studies Association of Nigeria (YSAN). In 2004, Alt-i collaborated with the West African Linguistics Society to organize the West African Languages Congress with the theme: Globalisation and the Future of African Languages. Alt-i's collaboration in the organization of this congress influenced the proceedings towards language technol-

ogy which brought about great awareness of language technology issues among West African linguists.

3.9 Bridge building seminar series

Alt-i runs a Bridge Building Seminar series as a way of encouraging cross-disciplinary studies in universities and research centers. These seminars have so far been run in eight Nigerian universities and the National Institute for Nigerian Languages (NINLAN). The one-day seminars bring together scholars from Linguistics, Literature, Psychology, Mathematics, Physics, Computer Science and other relevant departments to build awareness of language technology problems and the need for knowledge and skills from a wide range of departments for their solutions

4 Observations

While undertaking the above activities in the last five years between 2003 and 2008, the following observation were made:

- The intellectual resources needed for developing knowledge, skills and academic programmes in Language Technology are largely available in Nigerian universities.
- The lack of awareness of the need to address language technology problems has made it difficult to harness and direct these resources towards the development of language technology for African languages.
- Strong sentimental attachments to departmental traditions makes it extremely difficult for scholars to venture far outside their departmental cocoons.
- The importance of linguistics as a field of study and the role of linguists in society are not properly understood. Hence students of linguistic may not be sufficiently motivated to aspire to their important roles in society.
- Inappropriate admission criteria, limited curricular, and low level of formal interaction between different faculties in the universities make it extremely difficult for students of the pure sciences, technology and linguistics to share courses and thereby have opportunities for academic interaction beneficial to the development of HLT.

5 Recommendations

- Intensive and sustained awareness building programmes on the importance of linguistics and language technology should be undertaken in institutions of higher learning. This will make it possible to harness some of the available intellectual resources that remain yet untapped for the development of language technology for African languages in these institutions.
- Admission criteria and curricular should be reviewed in order to encourage and capacitate students to widen their intellectual horizon beyond the artificial traditional departmental boundaries.
- Modern techniques for the management of learning resources should be employed in order to address the logistic challenges that discourage students from taking courses across the faculties of science, technology and arts.

6 Alt-i: a historical perspective

Initial interests in language technology that ultimately led to the founding of Alt-i date back to 1978, but it was in 1985 that a small group of one Electrical Engineer and two Physicist started to investigate Text to Speech synthesis of Yoruba in Ibadan, Nigeria. All they were armed with was a book (Electronic Speech Synthesis by Geof Bristow), a microphone and storage oscilloscope. It was extremely difficult to get the required materials in the Nigeria of those days.

Unfortunately, neither the Physicists nor the Engineer realized the relevance of linguistics in their work because the academic environment within which they grew did not provide the necessary impetus for interdisciplinary or multidisciplinary studies between certain fields of study, certainly not between linguistics, physics and engineering. This brought about a lot of misdirected efforts and frustration. By 2001 however, with better access to the scientific literature of computational linguistics and HLT, it had become clear to what was left of the group that HLT is as much an issue in language as it is an issue in technology.

A careful review of the relevant aspects of the scientific literature of linguistics was then undertaken. Contacts were made with some of the known scholars of Yoruba phonology and their insights brought new impetus to the work, leading to the founding of Alt-i in 2001.

The salient point in this historical perspective is that the academic environment that produced the members of the original group did not provide the necessary impetus for the level of cross-disciplinary cooperation demanded by the solutions to the problems the group was addressing. Even though there were many papers that illustrated fruitful connections between linguistics and computer science in the mid 1980's, the Nigerian economy was in such a bad state that the universities could not afford to keep their libraries updated with current publications in any field. Nigerian universities now have noticeably better access to the global academic literature but a systemic weaknesses that does not encourage inter-disciplinary scholarship still subsists and needs to be addressed.

7 Proposal

A project with advocacy and service components aimed at accelerating the development of language technology for African languages is hereby proposed. The aim of the proposed project is to produce lecturers, researchers and other experts in language technology for African languages.

The advocacy component will identify and develop policy thrusts that will encourage the development of language technology and raise awareness at various levels of the importance of linguistics and language technology. These include raising awareness among secondary school students, university undergraduates, cultural activists and relevant policy makers.

Within the service component, in affiliation with a university, a post-graduate course of study aimed at producing a number of PhDs in language technology/computational linguistics within the space of about five to six years is to be developed. The candidates for this post-graduate course shall be university graduates of various relevant fields. The programme shall start with a one year diploma programme of intensive course work in linguistics, computational and cognitive sciences. These will serve to widen the knowledge-base of the participants thereby creating the necessary connections between their backgrounds and various aspects of language technology. Those that attain a high level of achievement in the diploma course may stay on for another six-months to undertake a practical project in language technology. Success in this project will earn such candidates a master degree in language technology or computational linguistics. Graduates of the master programme that attain a high level of performance in the project will be encouraged to stay on for the PhD programme.

The main faculty for the programme shall be drawn from relevant departments in the university. They shall undergo induction courses (locally and overseas) to re-orientate their knowledge towards applications in language technology.

To kick-start the programme, the support of scholars in the international language technology community shall be sought for curricular development as well as teaching. Occasional or short-term visiting lectureships will be accommodated within sabbatical, fellowship and exchange programmes.

As an on-going experiment in this regard, two students of the university of Ibadan are at present working together on Yoruba-English machine translation. One student is a graduate of Computer Science, working towards a master degree in Information Science, while the other is a graduate of Linguistics working towards a master degree in Linguistics. The two students are jointly supervised by a lecturer in Information Science and a lecturer in Linguistics. Even though the computer science graduate has never had any formal training in Linguistics and the Linguistics graduate has never had any formal training in computing, their collaboration has served to widen their knowledge-bases. The student of linguistics is approaching the project from the point of view of comparative syntax and is now able to express syntax rules in the form of context-free-grammar in Prolog, while the Information Science student is approaching the project from the point of view of predicate logic as a knowledge representation formalism and now has a fair understanding of the principles of Yoruba and English grammars.

Even though the experiment is still on-going, the emerging results suggest that the one year of formal study in the proposed diploma programme will provide adequate knowledge and skills for graduates of the physical sciences, computer science, technology, linguistics and psychology to undertake productive research in language technology.

8 Conclusion

The development of language technology for African languages is at a rather embryonic stage. Apart from the efforts in South Africa, there are

little or no coherent programmes on language technology in African universities. National language policies where they exist do not accommodate language technology issues and there is a generally low level of awareness of the benefits derivable from language technology.

With one-third of the world's languages spoken in Africa, there is an urgent need for the development of new techniques that address the peculiar features of these languages and thereby make it possible for the cultures that use them to benefit from the information revolution without having to adopt foreign languages

The implementation of the proposed academic programme in HLT would require active support of the Nigerian government in cooperation and collaboration with other friendly governments as well as various multilateral agencies for funding and other resources. However, in the absence of a language policy and a coherent language technology programme, the advocacy component becomes be the necessary starting point.

Developments in HLT and computational linguistics present avenues for Nigerian universities to re-invigorate the study of linguistics, provide new impetus for students of linguistics and prepare graduates of linguistics for more roles in society than the traditional teaching of local languages at secondary level. Nigerian universities must therefore play an important role in the necessary advocacy.

As the world moves further into the information age, concerted efforts are need to ensure that developments in HLT takes due account of African languages so that African languages and cultures can benefit from the information revolution.

Acknowledgments

Alt-i acknowledges with thanks the funding support of Tiwa Systems Ltd., Bait-al-Hikma, Open Society Initiative for West Africa (OSWIA) and International Development Research Center (IDRC) in its activities.

References

Tunde Adegbola. 2005. Application of tonemic information for search-space reduction in the automatic speech recognition of Yoruba. Unpublished PhD. Thesis, December, 2005, University of Ibadan, Nigeria.

Tunde Adegbola. 2007. Hitting the right tone. ICT Update http://ictupdate.cta.int/en/Feature-Articles/Hitting-the-right-tone

Tunde Adegbola. 2009. The Future of African Languages in a Globalising World. Presented at the Bamako Summit on Multilingualism, 19 - 21 January 2009. Bamako, Mali.

Tunde Adegbola. 2009 Indigenising Human Language Technology for National Development. To be presented at the Africa Regional Center for Information Science (ARCIS), University of Ibadan, Guest Lecture, March 18, 2009.

eLearning Africa. 2007. Interview with Dr. Tunde Adegbola. http://www.elearning-africa.com/newsportal/english/news56_print.php, International Conference on ICT for Development Education

Justus Roux. 2008. HLT Development in Sub-Saharna Africa. Report to COCOSDA/WRITE Workshop, LREC2008, Marrakesh. http://www.ilc.cnr.it/flarenet/documents/lrec2008_cocosda-write_workshop_roux.pdf

Adam Samassekou. 2007. Linguistic Diversity. ICT Update. http://ictupdate.cta.int/en/Regulars/Carte-blanche/Linguistic-diversity

Roger Tucker. 2003. Local language speech technology initiative. http:/www.llsti.org