Programme

Monday 22 May 2006

14:30 – 14:45 JC Roux. Opening; Scene setting; Aims

Area surveys

14:45 – 15:00 D Gibbon. Language documentation in West Africa.

15:00 – 15:15 A Hurskainen Language technology and resource development in East Africa.

15:15 – 15:30 J Emejulu, J Kwenzi-Mikala, F Idiata, S Ndinga Koumba-Binza. Language resources and tools in Central Africa

15:30 – 15:45 JC Roux & S Bosch. Language resources and tools in Southern Africa

Projects per area

West Africa

15:45 – 16:00 A Fleisch & F Seidel. Cologne initiative on language processing in African languages.

East Africa

16:00 – 16:15 N Abdillahi, N Pascal, J-F Bonastre & F Bechet. Speech mining to make African oral patrimony accessible.

16:15 – 16:30 P Wagacha, G De Pauw & K Getao. Developing an annotated corpus for Gikuyu using machine learning techniques.

16:30 – 17:00 Tea / Coffee break

Southern Africa

17:00 – 17:15 A Kasonde. Saving an African language: Computer assisted lexicon and grammar for translation purposes.

17:15 – 17:30 R Gauton. Share and share alike - developing language resources for African language translators.

17:30 – 17:45 S Bosch, J Jones, L Pretorius & W Anderson. Resource development for South African Bantu languages: Computational morphological analysers and machine-readable lexicons.

Larger projects and infrastructures

17:45 – 18:00 K Safir, A Anghelescu, S Murray & J Rett. The African Anaphora project.

18:00 – 18:15 D Yie-Chien Liu, S Chun-Feng Su, L Yu-Hsuan Lai, E Hsiao-Yun Sung, J Yi-Lung Hsu, S Yin-Chi Hsieh, O Streiter. From corpora to spell checkers: First steps in building an infrastructure for collaborative development of African language resources.

Networking language resources

18:30 – 19:00 General discussion on text and speech resources; standardization; network formation, future activities [Facilitator: JC Roux]

Workshop Organiser(s)

Justus C Roux
Centre for Language and Speech Technology (SU-CLaST)
Stellenbosch University
South Africa

Sonja E Bosch
Department of African Languages
University of South Africa (UNISA)
Pretoria
South Africa

African Language Association of Southern Africa, Special Interest Group for Language and Speech Technology (ALASA-SIG)

Workshop Programme Committee

Table of Contents

Author Index

Language Documentation in West Africa

Dafydd Gibbon

Universität Bielefeld
Postfach 100131, D-33501 Bielefeld
gibbon@uni-bielefeld.de

Abstract

The present contribution reports on selected developments in language documentation in West Africa, in particular on general and region-specific documentation criteria, and on specific initiatives in the area.

1. Language Documentation

The present brief overview is based principally on the following selective criteria:

1. The widely accepted distinction between *language documentation* and *language description* introduced by Nikolaus Himmelmann (Himmelmann 1998).
2. The portability criteria put forward by Bird and Simons (Bird & Simons 2002).
3. The *WELD* (*Wordable Efficient Language Documentation*) criteria for language documentation (Gibbon 2002).

This means, essentially, that linguistic work which is either mainly theoretical or descriptive, or idiosyncratic, or based on proprietary or non-open-source software are considered peripheral. A great deal of linguistic work on West African languages is, of course, available, and indeed has more than once proved to be pioneering and paradigm-determining, as with the developments in Prosodic Phonology, from Firth to the Autosegmental Phonology initiated by Leben, Goldsmith and others. However, these developments are outside the scope of language documentation itself, which is a scientific discipline in its own right, closely related to Text Technology, and also the provider of high quality empirical resources for linguistic description, in the form of text and speech data, lexica, grammars and software tools for acquisition, annotation, archiving, dissemination, retrieval and data mining.

The WELD criteria, which were developed in close consultation with colleagues at universities in Ivory Coast and Nigeria, outline a charter for language documentation with specific reference to endangered languages, and specifies that language documentation should be:

1. Comprehensive: In principle this means that language documentation must apply to all languages. But economy is a component of efficiency, and priorities must be set which may be hard to justify in social or political terms: if a language is more similar to a well-documented language than another language is, then the priority must be with the second.
2. Efficient: Simple, workable, efficient and inexpensive enabling technologies must be developed, and new applications for existing technologies created, which will empower local academic communities to multiply the human resources available for the task. A model of this kind of development is provided by the Simputer ("Simple Computer") handheld Community Digital Assistant (CDA) enterprise of the "Bangalore Seven" in India (see http://www.simputer.org/), which could be incorporated into conventional European and US project funding schemes.
3. State-of-the-art: In addition to using modern data exchange formats and compatibility enhancing archiving technologies such as XML and schema languages, efficient language documentation requires the deployment of state of the art techniques of from computational linguistics, human language technologies and artificial intelligence, for instance by the use of automatic classification techniques for part of speech tagging and other kinds of annotation, and of machine learning techniques for lexicon construction and grammar induction. The SIL organisation, for example, has a long history of application of advanced computational linguistic methodologies (see www.sil.org), but more advanced techniques are available, and more research is needed.
4. Affordable: In order to achieve a multiplier effect, and at the same time benefit education, research and development world-wide, local conditions must be taken into account. Traditional colonial policies of presenting “white elephants” to local communities, which must be expensively cared for and then rapidly become dysfunctional, must be replaced by less expensive methods – for instance it is expensive or impossible to download a large, modern software package because of slow networks, and electricity outages and landline interruptions). Net-based software registration and updating is very costly, as is wireless data transfer. However, in some areas modern techniques such as ADSL are becoming available.
5. Fair: If a language community shares its most valuable human commodity, its language, with the rest of the world, then the human language engineering and computational linguistic communities must do likewise, and provide open source software (also to reap the other well-known potential benefits of open source software such as transparency and reliability). The Simputer Public Licence for hardware and the Gnu Public Licence for software are useful models. The development and deployment of proprietary software (and hardware for that matter) and closed websites in this topic domain

is a form of exploitation which is ethically comparable to other forms of one-way exploitation in biology and geology, for example in medical ethnobotany and oil prospecting.

The WELD criteria have been deployed particularly by Eno-Abasi Urua at the University of Uyo, Akwa Ibom State, Nigeria in developing a programme for training in documentation of the languages of South-Eastern Nigeria.

2. Motivations for Language Documentation in West Africa

The language situation in general in West Africa is highly complex, and consequently the motivations for langauge documentation are also highly complex when it comes to the details of which languages to document with highest priority. The now general terminology of *global langauges* vs. *local languages* needs considerable refinement in order to distinguish sensible criteria for language deployment in dependence on the regional social functionality of the languages concerned. The following characterisation is still very general, but is based on practical experience and discussion with colleagues at West African universities:

1. Languages of administration:
    1. Official (e.g. national) standard languages.
    2. Local standard languages.
2. Regional trading languages.
3. Local languages (with dialectal, social and functional variation).
4. Former colonial languages:
    1. European varieties of former colonial languages.
    2. Local varieties of former colonial languages
    3. Creoles based on the pidgins which are related to former colonial languages.
    4. Pidgins based on trade interactions between speakers of colonial and local languages, where the dominant society is the colonial society.

The motivations for documenting each of these language categories are very different. For the languages of administration, i.e. both national languages and local standard languages, a prime motivation for language documentation is the development of educational materials and speech technology applications (including porting existing applications in the case of the former colonial languages English and French). For the local languages without official administrative status, of which there are claimed to be almost 500 in Nigeria alone, for example, the motivations are generally very different and from the linguistic point of view more traditional: the creation of a language and culture heritage archive for future use by linguists and descendants of the current language community.

For the non-administrative local language category, the situation is often extremely difficult, and dependent on highly time and resource intensive linguistic fieldwork, especially if a writing system for the language has not been developed. Also, in some cases, the development of a writing system has been more of a hindrance than a help: well-meaning applied linguists have developed orthographies in a very idiosyncratic fashion, with *ad hoc* character glyphs and almost as often several different fonts which map these characters on to the glyphs in different ways, resulting in non-portable documents (Gibbon, Bow, Hughes & Bird 2004). In other cases, competing orthographies have developed for reasons of political expediency. In the case of Anyi (Agni) which overlaps the frontiers of Ivory Coast and Ghana, for example, the Ivory Coast orthgraphy is based on French orthography, whereas the Ghanaian orthography is based on English orthography. The differences between the two varieties – which are not simply based on dialectal differences but on differences in English and French orthography – may be illustrated by the spellings *Kouadio* (Ivory Coast) and *Kojo* (Ghana) for the same male personal name. Different Christian denominations have also developed competing orthographies for some languages.

It is evident to anyone with experience in language engineering projects that the size of the efficient documentation task is well beyond the abilities of individuals, projects, single consortia or research institutes. The present vision, which needs to be developed further, is for involving wealthy language engineering and computational linguistic communities and for spreading the WELD idea beyond these communities to less well equipped local scientific communities around the world with old computers, unstable electricity supplies, expensive internet links (if any), and little if any contact with recent developments in the language and speech communities. Communities like these need tools which are workable in the local environment (not the latest heavy GUI software with proprietary applications and massive hardware requirements). But it is clear that the benefits of the WELD paradigm would not be one-sided: research and development on portability for such tasks would benefit many local language communities around the world and have spin-off effects for portable speech and text technologies in other applications.

3. Language Documentation in West Africa

Language documentation been under way in West Africa for some time, and builds on a number of more traditional precursors, which include the following:

1. The *Atlas* projects during the 1970s and 1980s in Côte d'Ivoire (Ivory Coast): *Atlas des Langues Gur*, *Atlas des Langues Kwa*, *Atlas des Langues Kru*, and *Atlas des Langues Mande*.
2. The Ford Foundation project roughly during the same period.

In these projects, systematic questionnaires were developed for documenting basic wordlists and small dictionaries, sketch grammars, and language samples. A questionnaire of this kind which has been used rather frequently is the West African Language Data Sheet (WALDS), compiled by Mary Esther Kropp-Dakubu. The projects provided an impressive array of short language descirptions, which have been the starting point for many linguistic treatments of West African Languages since 1980.

These traditional projects had not yet been subjected to the massive influence of the methodologies which the language documentation paradigm has meanwhile adopted from text technology, speech technology and database

technology. Technological influences on the language documentation paradigm arose mainly in the multilingual European projects *Speech Assessment Methodology* (*SAM*), in which the well-known IPA encoding SAMPA (SAM Phonetic Alphabet) was originally developed, in the two phases of the EAGLES project, and in related projects such as MATE and NITE (further information on these is easily available on the internet). Technologically comparable projects elsewhere, e.g. in the USA, tended to be monolingually oriented until the speech-to-speech translation paradigm emerged in the course of the 1990s.

4. Infrastructure

4.1. Universities and archives

The universities of West Africa have well established universities with Departments of Linguistics and of Applied Linguistics (e.g. applications of linguistics to the development of alphabetisation and literarisation materials, and of terminology for communication in local languages in the context of modern political, economic and cultural environments). In addition to university activities, there are also local and national archives. However, a general coordination policy has not, in general, been developed for efficiently handling, reliably storing and disseminating language resources. There is no continuity which would relate earlier Atlas-type activities to modern language documentation programmes, but attempts are being made, for example at the Département de Linguistique / Institut de Linguistique Appliquée in Abidjan, Ivory Coast, to remedy this, and the West African Linguistics Society is also increasingly functioning as a catalyst in this respect.

4.2. West African Linguistics Society

The West African Linguistics Society (WALS) / Société Linguistique de l'Afrique Occidentale (SLAO) meets in different West African countries every two years, and since its meeting in Ibadan, Nigeria, in 2004, where a well-attended panel discussion on the topic took place, has been actively supporting the Language Documentation paradigm. The topic will play a major role at the 2006 meeting in Benin, and it is hoped that established journals such as JWAL (Journal of West African Linguistics) will join forces with the WALS/SLAO in the language documentation community.

4.3. Projects

Since approximately 2000, an increasing number of initiatives of various kinds have been funded world-wide, including, among many others, archive repository metadata development (*EMELD*, Electronic Metastructures for Endangered Languages Data), a metadata repository (*OLAC*, *Open Language Archive Community*), the Language Documentation curriculum at SOAS, the School of African and Oriental Studies, and specific language documentation projects (e.g. In the *DoBeS*, *Dokumentation bedrohter Sprachen* / *Documentation of Endangered Languages* group).

West African languages have featured in these initiatives in a number of ways, including:

1. The project "Ega: A documentation model for an endangered Ivorian language" in the *DoBeS* group pilot phase (Connell & al. 2002), funded by the VW-Stiftung. The project focussed on the development of fieldwork oriented computational phonetic and lexicographic techniques (e.g. a multimodal concordancer for audio data) and the collation of interactive, experimental, and questionnaire data.
2. The curriculum development project *ABUILD* (*Abidjan-Bielefeld-Uyo Introduction to Language Documentation*), funded by the DAAD, Deutscher Akademischer Austausch-Dienst. The project has developed and implemented M.A. curricula which include language documentation, in a triangular configuration of francophone (Université de Cocody, Abidjan, Côte d'Ivoire), anglophone (University of Uyo, Uyo, Akwa Ibom State, Nigeria) and germanophone (Universität Bielefeld, Germany) universities.
3. A doctoral project at SOAS by Sophie Salffner.
4. A doctoral project at the University of Birmingham by Tunji Odejobi.
5. The establishment of ALT-I, the African Language Technology Institute in Ibadan, Nigeria, by Tunde Adegbola.
6. The *LLSTI*, *Local Language Speech Technology Initiative*, initiated and coordinated by *Outside Echo*, managed by Roger Tucker and Ksenia Shalonova, in which resources and a prototype speech synthesiser for Ibibio (Lower Cross, Nigeria) were created.

The documentation spin-off from other linguistic, applied linguistic (e.g. orthography, translation, terminology, educational materials) projects on African languages is currently rather low, at least in the sense of language documentation represented in this overview.

5. Conclusion and outlook

The Language Documentation paradigm as formulated by Himmelmann, together with the Language Resources and Evaluation paradigm established mainly by Antonio Zampolli and collaborators within the European IT framework programmes, have already had considerably impact on the development of language documentation infrastructures, educational programmes and actual data and metadata resources in West Africa.

As one of the barriers to development in this area, the "digital divide" is frequently mentioned. This term must not only be taken to refer to actual hardware and software issues, however, but as denoting a much broader frame of reference, including local infrastructural factors (social and physical) as well as linguistic factors (resources and language typology issues which are less relevant for the dominant technological paradigms which are generally oriented towards the Indo-European languages):

1. Local infrastructure: the differential between local university and archive infrastructures in the

West African context and in those in the affluent areas of the world is enormous:

1. Social factors include:
    1. Fewer educational opportunities.
    2. No collaborative infrastructure.
    3. Low local research and development funding.
2. Physical factors include:
    1. Low-speed or no internet, in general.
    2. Unreliable electricity.

2. Linguistic issues:
    1. Local resources:
        1. Sparse data, most on paper, little electronic material.
        2. Inconsistent formatting and font use.
        3. Orthography: sometimes no tone marking, which is fine for use in native speaker reading contexts, but inadequate for many other uses; historically determined orthographic variants.
    2. Language typology:
        1. Local West African languages typically have functions of pitch (fundamental frequency) which differ considerably from those found in other languages, including phonemic tone (comparable with phonemic tone in Eastern Asian languages), but also morphosyntactic tone which realises grammatical functions and marks structure. The properties are shared with languages of Central and Southern Africa, though lexical tone in general appears to be more highly functional in the West African languages.
        2. Specific phonetic configurations of pitch patterning represented by terraced tone systems (downstep, upstep etc.) and discrete level tone systems.

Many of these issues also apply to language documentation in Central and Southern Africa (and, of course, in many other less affluent parts of the world). The question therefore arises of how to ameliorate this situation. The initiatives already mentioned have the drawback of being anchored in the competitive arena of research and development funding in affluent societies, and are located mainly in Europe and the USA, with funding turnover mainly remaining in Europe and the USA. The life and death of projects in the field is therefore not determined by local needs but by the goals and ambitions of funding agencies and project leaders in the more affluent societies.

This characterisation of language documentation in West Africa outlines a situation in which, gradually, solidarity, cooperation and collaborative infrastructures are developing in the West African universities. Essentially, the way forward is for local infrastructures to be initiated and supported locally, in organisations, lobbies, and funding networks. A first step would be to establish a firm and efficient cooperation network between the regions of Africa, boot-strapped by colleagues in these areas. A starting point could be the establishment of cooperation between the existing linguistics societies in Western and Southern Africa in order to develop strategic policies, in cooperation with other disciplines from speech engineering and comptuter science to anthropology and musicology, for solving the problematic issues in language documentation in these areas. And, not least, the solution of specific technological problems stemming from the language typology of West African languages, for instance in the area of prosody, promises to stimulate a technology push which can benefit speech and text technology for other languages.

6. Relevant publications and references

The following list of publications includes references for sources for this paper, and other materials which are relevant to language documentation issues in West Africa.

Bird, Steven & Simons, Gary (2002) Seven dimensions of portability for language documentation and description. *Language*, 79:557-582.

Connell, Bruce, Firmin Ahoua & Dafydd Gibbon (2002). Illustrations of the IPA: Ega. *Journal of the International Phonetic Association* 32/1, 99-104. With Bruce Connell & Firmin Ahoua.

Gibbon, Dafydd (2002). The WELD paradigm -Workable Efficient Language Documentation: a Report and a Vision. *ELSNews* 11.3 Autumn 2002, 3-5.

Gibbon, Dafydd (2003). Computational linguistics in the Workable Efficient Language Documentation Paradigm. In: Gerd Willée, Bernhard Schröder & Hans-Christian Schmitz, *Computerlinguistik: Was geht, was kommt?* St. Augustin: Gardez! Verlag, 75-80.

Gibbon, Dafydd (2003). A computational model of low tones in Ibibio. In: *Proceedings of the International Congress of Phonetic Sciences*, I: 623-626.

Gibbon, Dafydd, Cathy Bow, Baden Hughes, Steven Bird (2004). Securing Interpretability: The Case of Ega Language Documentation. *Proceedings of Language Resources and Evaluation Conference (LREC) 2004*, Lisbon.

Gibbon, Dafydd, Firmin Ahoua, Eddy Gbery, Eno-Abasi Urua, Moses Ekpenyong (2004). WALA: a multilingual resource repository for West African Languages. *Proceedings of the Language Resources and Evaluation Conference (LREC) 2004*, Lisbon.

Gibbon, Dafydd (2005). First steps in corpus building for linguistics and technology. *Proceedings of the "First Steps..." Workshop, Language Resources and Evaluation Conference (LREC) 2004*, Lisbon.

Gibbon, Dafydd (2006). Problems and solutions in Text-to-Speech for African Tone Languages. Multiling2006, Stellenbosch, South Africa.

Gibbon, Dafydd, Eno-Abasi Urua & Moses Ekpenyong (2006). Morphotonology for TTS in Niger-Congo languages. *Speech Prosody 2006*, Dresden. With Eno-Abasi Urua.

Gibbon, Dafydd (2006). Tone and timing: two problems and two methods for prosodic typology. *Proceedings of the Tonal Aspects of Language Conference 2004*, Beijing.

Himmelmann, Nikolaus P. (1998) Documentary and descriptive linguistics. *Linguistics*, 36:161-195.

Language Technology and Resource Development in East Africa

Arvi Hurskainen

University of Helsinki
Box 59, FIN-University of Helsinki
Arvi.Hurskainen@helsinki.fi

Abstract

The paper discusses the status of language technology (LT) in Eastern Africa. The resource development, including human resources, various kinds of language materials, such as text corpora, manuscripts, dictionaries, thesauri and grammars are discussed and evaluated. Also language manipulation tools already available and those under construction are described. Discussion is carried out on the need and problems of standardization. Also localization programs, such as Kilinux for adapting OpenOffice to Swahili, and the translation of MS Office 2003 and Windows XP to Swahili, are introduced. The shortage of high quality open source platforms as well as the need of African governments to support the development of such platforms are also discussed.

1. Introduction

The development of language technology (LT) is often motivated by a chosen language policy, and technology is developed to languages, which have important roles in communication. But LT may have a bearing on the language policy itself, regardless the official government policy. LT should not be seen only as a set of tools for manipulating language. In a context of rival views on the importance of various languages, especially the foreign vs. indigenous languages, LT may have an important role. The availability of competitive tools for language manipulation strengthens the chances of the language to develop into a viable means of communication. This aspect is particularly relevant in Africa, where colonial languages compete with local languages.

In countries of Eastern Africa there is currently no official plan for developing LT specific to African languages. Yet interest in information technology is great and there is openness also towards developing language-specific technology (e.g. Mapendekzo ya Kongamano la Kiswahili na Utandawazi, Dar-es-Salaam, July 2005)[1]. The LT developed so far is an initiative of individual researchers, although it has been part of established academic research in cooperation with various institutes.

LT in Eastern Africa has concentrated almost exclusively on written text, but initial steps also for developing speech technology have been taken, especially as part of individual PhD studies.

2. Resource development in East Africa

Various kinds of resources are a precondition for developing genuine LT. Human as well as material resources are necessary. In Eastern Africa there is shortage of all kinds of resources.

2.1. Human resources

The first requirement for developing computational language resources is the existence of a sufficient number of people with skills of developing LT. The skills needed vary greatly from fairly simple text editing to system development, which often requires high level skills. The number of computer literate people increases rapidly, and the availability of people for compiling language resources such as text corpora or other kinds of text collections should not be a problem. The departments of computer science in Dar-es-Salaam and Nairobi, for instance, give training in many kinds of skills, including programming, and this activity is a potential resource, which could be made use of in LT.

Although there is no specific training in LT in East African universities, the existing resources could be used for developing LT as well. Occasional courses on how to manipulate language with computers may be given in departments of computer science in Dar-es-Salaam and Nairobi, but there are no degree courses specific to LT. A few students of PhD level have studied abroad, and some are still in training. Yet the big problem is the absence of local undergraduate training in LT.

However, the absence of degree courses in LT does not mean that LT could not be developed without specific training in. Establishing specific degree courses might be fairly costly compared with the solution that computer scientists and linguists work together for the common goal. Examples of this do exist, e.g. in South Africa (Pretorius and Bosch 2003a, b). Similar cooperation can be found also in Dar-es-Salaam and Nairobi.

There is also a need for training people with lower level skills in LT, keeping in mind that most of the actual work in developing LT does not require a PhD degree. There are efforts for arranging this kind of training.

2.2. Material resources

Material resources needed in LT include corpora of written and spoken language, thesauri, dictionaries, wordlists, grammars, etc. Language tools can also be counted among material resources. However, because of their special nature and big importance I shall discuss language tools in a separate chapter.

2.2.1. Text corpora

Before the time of electronic publishing, the creation of text corpora was very labour-intensive. Currently the situation is very different, because practically all publishing makes use of electronic text form. There are potentially vast amounts of text in the Web and in the archives of various publishing houses.

However, there has been very little effort in compiling text corpora in Eastern Africa, at least such corpora that

[1] Recommendations of the International Symposium on Kiswahili and Globalization, held in Dar-es-Salaam on 4-7 July, 2005.

are available for the research community. The only globally available text corpus is the Helsinki Corpus of Swahili (HCS)[2] with more than 12 million words, which was compiled by scanning publications and later by extracting texts from the Web. There is also a Swahili speech archives[3] with 150 hours speech from various Swahili-speaking areas in Tanzania, compiled in 1988 - 1992 by the universities of Dar-es-Salaam and Helsinki.

In addition, there are text collections in private archives of individual researchers, but they are not publicly available for research.

Currently there is a good potential for compiling text corpora in East Africa, especially for Swahili. Cooperation with various publishing companies would yield good quality text for corpora. The web is also a useful source for extracting texts. In recent years, the majority of texts for the Helsinki Corpus of Swahili was retrieved from the Web with a robot. In addition, all texts were read and edited manually, because they contained too many typing errors.

The Open Swahili Localization Team, also known as Kilinux, compiled a corpus of 10 million Swahili words from the Web for the Jambo Spellchecker, using an automatic search engine. However, this corpus has not been made available to other researchers.

2.2.2. Manuscripts

The library of the School of Oriental and African Studies in London holds a large public collection of Swahili manuscripts. The collection includes more than 250 manuscripts dating from the 1790s to the 1970s, contained in the papers of William Taylor, Alice Werner, William Hichens, Wilfred Whiteley, Jan Knappert and Yahya Ali Omar. The library also holds microfilms of the manuscripts that were deposited by JWT Allen at the University of Dar es Salaam.

The SOAS manuscripts are an invaluable resource for the study of cultural and literary history. Many were scribed in Arabic script, and many contain poetry composed in northern Swahili dialects.

SOAS has launched the Swahili Manuscripts Project, which aims to create a comprehensive catalogue of the SOAS manuscripts, thus enabling researchers and more general readers to make focused and, it is hoped, greater use of an illuminating collection. The project, a collaboration between the Department of the Languages and Cultures of Africa and the SOAS Library, is funded by a research project grant from the Leverhulme Trust.

The web page is well organized and it is easy to browse through the material. The fact is, however, that the material itself, i.e. the manuscripts, are not available though the Internet. We have a well organized web catalogue of the material, some description of its contents, but not the material.

2.2.3. Dictionaries, thesauri and grammars

Some Swahili dictionaries have been published in recent years using electronic publishing methods. These are potential sources in compiling electronic dictionaries and language management tools. However, they are not public resources, and their use requires a special agreement with the copyright holder.

An exception is the Internet Living Swahili Dictionary Project[4], hosted by Yale University, which has made the Swahili-English and English-Swahili dictionaries available in the Web with free access. Initiated in 1995, it has developed into a very widely used web dictionary. Its current search facilities include also some morphological analysis – a very useful feature, which can be implemented only in computer dictionaries.

The dictionary project of Yale University makes use of voluntary contributions of interested people, thus using the idea which Wikipedia[5] has later on successfully applied. Within the course of ten years the Swahili dictionary has grown into a large collection of Swahili words used in various Swahili-speaking areas. It is a storehouse of Swahili language, and this is its biggest advantage.

From the viewpoint of the Swahili student, however, the dictionary contains problems. Because it is a collection of dictionary entries, with a considerable amount of non-standard Swahili, the student can be misled to use a word or expression, which is either very rarely used or non-standard. Some kind of frequency marking of synonyms would vastly improve the usability of the dictionary, without the need of restricting the availability of rare and dialectal material.

The idea of compiling and maintaining a free Web dictionary is splendid, but for making it really useful for the main part of users it requires rigorous editing policy.

Perhaps we should add that in the compilation of the Swahili Web dictionary also the competing views of what is 'correct' Swahili play a part. The policy of allowing anybody to become a voluntary editor of the dictionary opens up possibilities for various kinds of zealots to put their fingerprints on the dictionary. I do not think that the introduction of foreign words as such without adapting them to Swahili phonotax develops the language, especially if those words already have a number of synonyms in the language. The need of expanding the vocabulary lies elsewhere, especially in various fields of science. This is, however, the responsibility of language committees of East Africa.

3. Tools

Among the very few language manipulating tools available on East African languages are at least two spell checkers of Swahili. The Open Swahili Localization Project (OSLP), to be discussed more below, prepared a spell checker based on a list of Swahili word-forms found in some texts. The spell checker operates in Linux operating system and it can be downloaded freely.[6] Also the simple spell checkers for 150 African languages, reported by Liu et al in the proceedings of this workshop, make use of word lists.

[2] http://www.csc.fi, visited 26.3. 2006.

[3] The speech texts are in transcribed form and are available in the Helsinki University Language Corpus Server. The application form for a user account is available in: http://www.ling.helsinki.fi/uhlcs/, visited 26.3. 2006.

[4] http://www.yale.edu/swahili/, visited 26.3. 2006.

[5] http://en.wikipedia.org/wiki/Main_Page, visited 26.3. 2006.

[6] Jambo Spellchecker 9D can be downloaded in http://www.yale.edu/swahili/, visited 26.3. 2006.

Lingsoft released in 1999 Orthografix 2 for Swahili[7], a spell checker that includes also a Swahili hyphenator. This spell checker uses finite state methods and two-level description in describing the language (Koskenniemi 1983, Hurskainen 1992).

The fundamental difference between these two approaches is that while the former uses word-lists of inflected words for checking the correctness of the word, the latter performs a full morphological analysis of each word. The spell checker of Lingsoft uses a comprehensive computational lexicon in describing the lexicon and morphology of the language, and a set of two-level rules for taking care of morphophonological alternations in morpheme boundaries. Orthograxic 2 for Swahili is fully compatible with recent versions of Ms Word and other programs of the MS Office family.

Orthografix 2 for Swahili is in fact only a minor offshoot in a comprehensive language management system, the aim of which is to provide an environment for developing many kinds of tools and utilities, including machine translation systems.

The language management system is currently called SALAMA (Swahili Language Manager)[8], which is in fact an environment for developing a variety of tools for manipulating Swahili. It was used for annotating the Helsinki Corpus of Swahili[9], which has a full morphological encoding, including the lemma of each word-form and the glosses in English. The Institute of Kiswahili Research is using SALAMA for compiling domain-specific dictionaries (Sewangi 2001).

There are also other projects that are not strictly LT but nevertheless support it. Those include a tool for storing and managing Swahili poetry texts for research by Tom Hinnebusch[10], and Synchrotext[11], a web-based tool for simultaneous listening and reading of songs and texts with translation.

4. Standardization

4.1. Standardization of language

The developer of the computational tools for language manipulation will necessary encounter the problem of standardization. This question is especially important in choosing acceptable words and word-forms when constructing a spell checker. To be really useful, the spell checker should be covering and accurate. It should include and accept all grammatically correct word-forms of the language, and only those.

This is the basic requirement of the simple spell checker, which works on the word level only, without considering the well-formedness of phrases and sentences. The developer encounters questions related to lexicon and grammar. Should one stick strictly to the lexicon defined by a standardization committee, or should one allow also other words? For a compiler of a language manager or a computational dictionary this is not a problem, because differed types of non-standard words or word-forms can be labelled with appropriate tags. But if the alternatives are inclusion and exclusion, as is the case in traditional spell checkers, the developer is in a really difficult position, especially if language standardization is not up-to-date.

So far such problems have been usually solved by compiling separate dictionaries for different language varieties, English being a good example of this. Another solution would be to add different marking for such words or word-forms, which are not commonly accepted as standard, but which in some areas of the language community are considered correct language. This would make the work of the designer easier and at the same time give a more informative result. The same principle has already been used in grammar checkers, which mark word level mistakes and syntactic mistakes with different colours.

4.2. Standardization of tagging

Another area where the question of standardization arises is the linguistic tagging. Is it possible to construct such a tag set, which all developers of language description would use, so as to make the result immediately familiar to other developers? This is certainly desirable, but it is not clear whether it is possible.

In the course of twenty years of designing a comprehensive description for Swahili, a variety of ideas have emerged and sunk. Some ideas continue to pop up every now and then, and it is perhaps fruitful to dwell on those for a while.

4.2.1. Core tagging

There are linguistic features that are common to all languages. On the level of morphology, part-of-speech categories and singular/plural dichotomy are examples of common features. Similar features can be found also in syntax, although the mutual order of syntactic elements may vary radically.

There is no reason why such commonly occurring features could not be marked with an agreed tag set in different languages. In addition to these global similarities, there are common features specific to a language family. In Bantu languages we find the noun class system, the description of which requires a complex set of tagging. Here again there is no reason why a joint tagging scheme could not be designed for these features. The established identification system based on numbers allows for different kinds of surface realizations of affixes, and if numbers are used in marking classes, identical tags can be used for all Bantu languages.

4.2.2. Secondary tagging

The experience shows that in addition to the core tags, there is also a need for secondary tagging. The need derives from two sources.

First, in spite of many common features of Bantu languages, there are also language-specific and group-specific differences, which require specific tagging. There are languages with augmented noun prefixes, while others are without augment. Tense/aspect categories between languages also differ, as does the affix structure of verbs. In these categories there are many group-specific features, which can be marked with a common tag. However, it is

[7] http://www2.lingsoft.fi/orthografix/, visited 26.3. 2006.
[8] http://www.njas.helsinki.fi/salama, visited 26.3. 2006.
[9] http://www.csc.fi, visited 26.3. 2006.
[10] http://www.linguistics.ucla.edu/people/hinnebus/hinnebus.htm, visited 26.3. 2006.
[11] http://www.smithsonianglobalsound.org/synchrotext/default.htm, visited 26.3. 2006.

hardly possible, or desirable, to force all tagging under a set of commonly agreed tag set.

Second, the description platform may require such extra tagging, which is then used in further processing, such as morphological, syntactic and semantic disambiguation or Machine Translation. Semantic clustering of words has been done in many ways, and very different tag sets have been used. It is hardly conceivable, at least in near future, that agreement on such tagging could be achieved. This is difficult, if not impossible, especially because of the different methods in constructing the systems.

4.2.3. Conversion of tags

The fear of being enforced to use predefined tags is alleviated by the possibility of creating conversion programs, which rewrite the tags of an individual researcher into the commonly agreed tags. This would give the researcher a possibility to use one's own style. Some of us are not at all too keen in adopting a tag set, which is felt to be unsystematic, too coarse, requires too much space or is too cryptic. We have also to remember, that some development platforms require certain types of tags. This also counteracts the idea of a common tag set.

Tag conversion is used in some user interfaces of language analyzers, which give a user the possibility to choose between tag sets. Some users need only a very general description of the phenomenon and they will be given the information in well-formed words of the language. Others need more technical and detailed information and can choose the tag set accordingly[12].

5. Projects in progress

There are various projects going on in Eastern Africa in the field of information technology. Most of them concern Swahili, just because of the relative importance of the language in the area. There is activity in information technology in general as well as in LT.

5.1. Localization projects

Recently there have bee at least two localization project on Swahili. In December 2005 Microsoft released its Windows Office 2003 version of Swahili, and in February 2006 was released the Windows XP operating system in Swahili. Although the terminology used in localization is going to cause discussion and criticism, the project itself is welcome. Although most users of Windows and MS Office would probably be conversant with English, the availability of the Swahili version will make users aware of the fact that computer environments and platforms are in fact language independent, and that they can be used in any language. The Swahili version also promotes discussion on computer terminology in Swahili and the vocabulary will be established in the field.

Microsoft announces that the Ms Office and Windows NT can be downloaded freely from the net. They are not free, however, because only those can download a Swahili version, who already have a purchased legal copy of the same product in some other language. It even sounds as if Microsoft through the localization project aims at controlling whether the user has a legal copy of the product. Being unable to load the Swahili version indicates that the user does not have a legal version installed.

Another localization project in East Africa is the Open Swahili Localization Project (OSLP), also called Kilinux[13], the aim of which is to translate the Linux operating system and OpenOffice to Swahili. In February 2005 the OpenOffice was officially released with the name Jambo OpenOffice. The project also produced its own computer vocabulary, which is different from that produced by Microsoft.

The localization project is funded by the Swedish International Development Agency (SIDA) and The University of Dar-es-Salaam (UDSM). It is coordinated by the Department of Computer Science (UDSM), Institute of Kiswahili Research (IKR) and the Swedish consultancy company IT+46. The project does not only aim to localize free and open source software into Swahili, but also to create awareness of the benefits of using and extending free software.

Noteworthy in the Kilinux project is that it promotes free availability of commonly used computer programs. It is also important that the OpenOffice package can be installed in Linux, Windows and Macintosh operating systems.

Along with OpenOffice, there is also the localization of the Tux Paint software for children. The localization is a result of collaboration between the Kamusi Project at Yale University and the Open Swahili Localization Project (OSLP) at the University of Dar es Salaam. The software is a complete Swahili adaptation of Tux Paint, a free and open source drawing program for children.

One should also mention Ubuntu[14], the project for constructing a Linux operating system that is easy to download and which contains only the most used functionalities of the Linux family. Translations of Ubuntu continue to several languages. At the time of writing, only Tux Paint was translated to Swahili.

5.2. Language technology projects

Language technology, in true sense, is taking just the initial steps in Eastern Africa. SALAMA, discussed above, is probably the only project, which fulfils the criteria of LT proper. By saying this I take the risk of being criticized for defining LT too narrowly. Therefore we need to clarify the concept.

It is true that the term 'language technology' has been used also in a wide sense, meaning any computer program or utility designed for processing language in some way. Roughly speaking, researchers in LT form a continuum, where in one end are the mathematically oriented researchers, who use mathematical methods in language manipulation, and in the other end are the researchers, who take the linguistic theory as a starting point and use it in all relevant phases of processing. Broadly speaking, all those count themselves as language technologists.

If we accept this broad definition, the situation in East Africa is not very bad. For example, there are efforts for designing spell checkers using the simple lookup method,

[12] The Machinese Syntax analyzers and Machinese Phrase Taggers of Connexor are examples of environments, where one can switch from one tag set to another. http://www.connexor.com, visited 26.3. 2006.

[13] http://www.kilinux.org visited 26.3. 2006.

[14] http://www.ubuntu.com, visited 26.3. 2006.

where a word-form is checked against a word list. The form is accepted if found in the list and otherwise rejected. Anybody knowing inflecting Bantu languages knows that using this method the recall and precision will not be sufficiently high for a good quality spell checker. And what is still worse, the system does not form a basis for further development in language analysis. The result is just a stand-alone poor application.

The Kilinux project, discussed above, produced a Swahili spell checker[15] using a word-list method. Also for some official languages of South Africa spell checkers have been designed with a similar method. Although it would seem that in languages with a disjoint writing system the method gives good results, this is not true. The system does not check the correctness of the disjointly written morphemes of the word.[16]

The Internet Swahili Dictionary project of Yale University has also announced that it is planning a language analyzer for enhancing the search with inflected word-forms. This is a very welcome facility, although not very easy to implement for the full grammar.

If we take the term 'language technology'[17] to mean such a computational manipulation of language, where the linguistic theory has a central role in designing algorithms, then we have very few ongoing projects.

As already mentioned above, SALAMA (Hurskainen 2004b) is an ongoing project for facilitating many kinds of high-level linguistic applications. Characteristic to SALAMA is that it is based on detailed and comprehensive linguistic description in all phases of processing. The analysis of text includes morphology, syntax and semantics so as to produce as full analysis as possible.

Because the morphological analyzer is non-deterministic (Koskenniemi 1983), i.e. it produces as many parses as linguistically possible on word level, disambiguation is a major problem. In the current system, also disambiguation makes full use of linguistic information (Tapanainen 1996, 1999; Hurskainen 2004a), i.e. disambiguation rules are in fact linguistic rules.[18]

One major application of SALAMA is Machine Translation. Therefore, semantic disambiguation becomes a major task for ensuring the optimal translation of the text. In defining semantic clusters, the automatic clustering facilitated by the Self Organizing Map (Kohonen 1995; Ng'ang'a 2005) is used. Large masses of corpus texts are needed for achieving reliable clusters. Also the WordNet of English helps in selecting the most preferable translations into English.

Machine translation contains much more than this. When we have an analyzed and disambiguated text, we still have to perform two major operations.

The first problem concerns word order, as well as non-matching of 'words', in source and target language. The word order of the source language is converted to match the word order in the target language. There may also occur double marking of subject and object in Swahili, but not in English. Also the article in English is problematic, because the source language has nothing equivalent to it.

The second major task is to transfer the lexical word into the correct surface form in English. Particularly problematic this is in verbs, which, in addition to having a number of prefixes and suffixes, each with semantic meaning, also have a number of tense and aspect forms. Also serial verbs[19] cause problems, because in English all verbs in the co-ordinated verb chain have the same inflected form.

Multi-word expressions, such as idioms, proverbs, and adjectival expressions, are a large group of constructs and require a special treatment.

6. Open source platforms

The best LT systems available today have been developed in environments, which are not publicly available. At least their use is not free. Therefore, it is no wonder that LT has made advances in commercially attractive languages. Commercial companies have developed applications on their own platforms or on platforms licensed from other companies. Open source platforms are currently not competitive with the commercial ones.

Commercially less attractive languages[20] are not in the program of commercial companies, because these languages do not have sufficient commercial potential. Therefore, there is a real danger that indigenous African languages will be excluded from LT proper.

Xerox has opened a possibility to develop morphological parsers with finite-state methods on its development platform (Beesley and Karttunen 2003), and the platform has been used also in describing some African languages. But we have to remember that the use does not allow the release of any product developed in the environment. Furthermore, the package does not contain any kind of disambiguator, or an environment for constructing sophisticated tokenizers needed in the description of languages with a disjoining writing system.

In SALAMA, discussed above, various components were developed using platforms licensed from Xerox, Lingsoft and Connexor. This was necessary, because there are no appropriate open source platforms.

There is an urgent need for open source platforms in LT. These should cover at least tokenizers, morphological analyzers and disambiguators. Without these, I am afraid, there will be no significant advances in the LT of

[15] http://www.yale.edu/swahili/, visited 26.3. 2006

[16] Prinsloo and de Schryver (2004) tested the performance of some spell checkers for various languages, including isiZulu and Sesotho sa Leboa. The latter shows a recall of over 95%, and the former less than 80%, with a list of 50,000 most common word-forms. This difference derives from the different writing systems, and not from differences in morphological complexity. IsiZulu uses a conjoining and Sesotho sa Leboa a disjoining writing system.

[17] Perhaps a better term would be 'computational linguistics' in stead of 'language technology' because it emphasizes the importance of linguistics in designing applications.

[18] To be precise, also rules that are based on no explicit linguistic theory must occasionally be written.

[19] In Swahili, a serial verb construction is a chain of co-ordinated verbs, where only the first verb inflects and carries all the morphological information, and the following verbs are in infinitive.

[20] I use this term to refer to languages, which by commercial companies are not considered attractive enough for investing in language technology. For example, Finnish with five million speakers is commercially attractive, while Swahili with 100 million speakers is not.

commercially less attractive languages. The need of such platforms was discussed in the Workshop of Finite State Methods and Natural Language Processing in Helsinki, September 1-2, 2005 (Yli-jyrä 2005).

7. Government support in East Africa

There is a positive atmosphere towards information technology in Eastern Africa. This applies to institutes of higher education, such as universities, as well as various government bodies. Despite many doubts, information technology is not particularly vulnerable in Africa. Computers and mobile phones are spreading rapidly and seem to endure local conditions rather well. The poor quality and unreliability of electric supply and high charges of Internet connections are major bottlenecks in developing information technology.

Although the governments have a positive attitude to information technology, financial resources available for developing such technology are very limited. The universities of Dar-es-Salaam and Nairobi at least have departments of computer science, and there is interest in developing information technology and LT. As discussed above, the Department of Computer Science in the University of Dar-es-Salaam was a partner in the Kilinux localization project. The Department of Computer Science in the University of Nairobi is working with Machine Translation, and according to current plans this interest will be integrated into the development of SALAMA.

In information technology, and in language technology in particular, there is a good chance for international cooperation, where the best practice is combined with best and tested developing platforms.

The African governments would make a big service to the future development of LT of African languages by giving strong support to efforts for producing open source platforms for LT. Otherwise there is a danger that African languages will be either totally neglected or pushed to the second class category with 'toy' tools in LT. The development of LT should be made independent of the will of commercial companies.

8. Conclusion

Language technology, with a few exceptions, is taking initial steps in East Africa. There is no coordinated planning or a strategy to guide the development. Also commercial companies have shown very little interest in the area, and so far the companies have considered the area commercially less attractive. The few achievements in LT must be credited to activities of individual researchers.

The development has also been hampered by the repeated doubts of the suitability of Swahili as an official language and as a language of training on all levels of education.

However, Eastern African countries are in a unique position in having Swahili as a lingua franca. The language has ample possibilities in extending even more widely. Successful LT could play a part in convincing the leaders of the suitability of African languages as main means of communication. This would also decrease the dependence of the education sector from foreign countries and would boost local book industry.

All those who deal with language in a multilingual environment would benefit from tools developed in LT. So far most activities have been dealing with the language in text form. Work on spoken forms of language is also necessary, if the aim is to construct speech-to-speech Machine Translation systems. Today this is a technically realistic aim, but its realization requires also political will.

9. References

Beesley, K. and Karttunen, L. (2003). *Finite State Morphology*. Series: CSLI Studies in Computational Linguistics. Stanford: Center for the Study of Language and Information.

De Schryver, G.-M. (2004). Spellcheckers for the South African languages, Part I: The status quo and options for improvement. *South African Journal of African Languages* 24(1l): 57-82.

Hurskainen, A. (1992). A Two-Level Computer Formalism for the Analysis of Bantu Morphology: An Application to Swahili. *Nordic Journal of African Studies* 1(1): 87-122.

Hurskainen, A. (2004a). Optimizing Disambiguation in Swahili. In *Proceedings of COLING-04, The 20th International Conference on Computational Linguistics*, Geneva 23-27.8. 2004. Pp. 254-260.

Hurskainen, A. (2004b). Swahili Language Manager: A Storehouse for Developing Multiple Computatiional Applications. *Nordic Journal of African Studies* 13(3): 363-397. Also in http://www.njas.helsinki.fi

Kohonen, T. (1995). *Selg-Organizing Maps*. Springer-Verlag, Heidelberg, Berlin.

Koskenniemi, K. (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Publications No.11. Department of General Linguistics, University of Helsinki.

Ng'ang'a, W. (2005). *Word Sense Disambiguation of Swahili: Extending Swahili Language Technology with Machine Learning.* Department of General Linguistics, Publications No. 39. University of Helsinki. PhD dissertation.

Pretorius, L. and Bosch S.E. (2003a). Computational aid for Zulu natural language processing. *Southern African Linguistics and Applied Languge Studies*. 21(4): 267-282.

Pretorius, L. and Bosch S.E. (2003b). Finite-State Computational Morphology: An Analyzer Prototype for Zulu. *Machine Translation* 18: 195-216.

Sewangi, S. (2001). *Computer-Assisted Extraction of Terms in Specific Domains: The Case of Swahili*. Ph.D. thesis. Publications of the Institute for Asian and African Studies, 1. University of Helsinki.

Tapanainen, P. (1996). *The Constraint Grammar Parser CG-2*. Department of General Linguistics, Publications No. 39. University of Helsinki.

Tapanainen, P. (1999). Parsing in two frameworks: finite state and functional dependency grammar. University of Helsinki. PhD dissertation.

Yli-Jyrä, A. (2005). Toward a Widely Usable Finite-State Morphology Workbench for Less Studied Languages - Part I: Desiderata. *Nordic Journal of African Studies* 14(4): 479 – 491.

Language resources and tools in Southern Africa

Justus Roux[1] and Sonja Bosch[2]

[1] Centre for Language and Speech Technology (SU-CLaST)
Stellenbosch University, South Africa
jcr@sun.ac.za
[2] Department of African Languages
University of South Africa, Pretoria, South Africa
boschse@unisa.ac.za

Abstract

The aim of this paper is to present an overview of current initiatives in the field of human language technologies (HLT) in South Africa with special attention being paid to the development of tools and resources for the languages spoken in Southern Africa. An overview of the language situation is sketched, after which a web-based survey is done of resources and tools created and used by role players in the field. The importance of standardisation of resources and resource management is discussed with some final conclusions and suggestions for possible wider co-operation in Africa. This paper is largely based on two publications, Bosch & Roux (in press) and Roux and Du Plessis (2005).

1. Introduction

The aim of this paper is to present an overview of current initiatives in the field of human language technologies (HLT) in South Africa, whilst focussing on the development of language resources and tools. It needs to be pointed out that this paper does not claim to be a comprehensive account of all resources or tools available as some websites (as sources) are dated, and some resources may not be accessible. Resources referred to in this paper also do not include any resources that are being developed by private companies.

South Africa, with its population of 44.8 million people, is a multilingual country in which eleven languages are recognised as official languages (cf. *Statistics South Africa,* 2005). These languages are English, Afrikaans, three related Sotho languages, namely Southern Sotho (Sesotho), Northern Sotho (Sepedi) and Tswana (Setswana), four related Nguni languages, namely Zulu (isiZulu), Xhosa (isiXhosa), Swati (siSwati) and Ndebele (isiNdebele), as well as Tsonga (Xitsonga) and Venda (Tshivenda). Figure 1 below gives a breakdown of the official languages as mother tongues. Although English only ranks fifth (9%) as a mother tongue, it is the language of the business environment, especially in urban areas. Despite the fact that many South Africans are multilingual, there are also many who can only express themselves in their mother tongue (especially in the vast rural areas), and many of whom are functionally illiterate. In a national survey on *Language Use and Language Interaction* conducted by the Pan South African Language Board (PanSALB) in 2000, it was found among others that English is used for neighbourhood communication in only three of the nine provinces, and that in fact only 22% of the respondents indicated that they fully understood addresses made in English (PanSALB Annual Report, 2001:6). The implementation of eleven official languages proves that national government is committed not only to honour the democratic language rights of its citizens, but also to create a situation in which information can be managed and disseminated to its citizens in their respective languages of choice and/or understanding. The realities that government is therefore confronted with are the following:

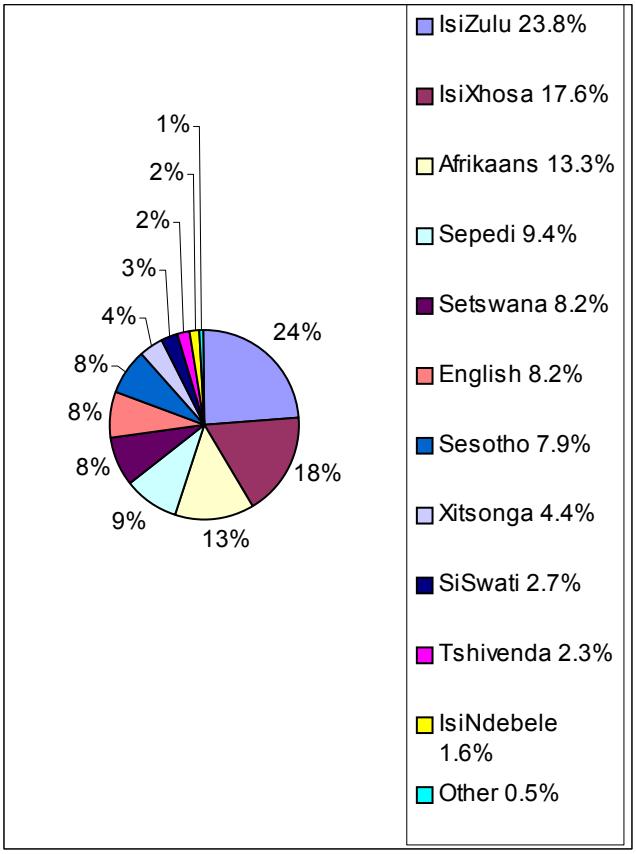


Figure 1: Mother-tongue division as per official language (n = 44,8 million speakers)

- the recognition of eleven official languages that need to be functionally developed and implemented within a society;
- language specific needs for communication within government (from parliament to individual departments), and between government and its citizens (from the highest level to local offices);
- language specific needs in a multilingual education and training environment with enormous historically induced backlogs which can hardly be addressed by means of traditional approaches;
- language specific needs related to service delivery in the public and private sectors – which could be converted into challenges for obtaining an economic edge over competitors;
- communication needs extending across a large geographic region, linking a number of adjacent countries in the sub-continent such as Botswana, Lesotho, Swaziland, Mozambique which share languages;
- a vastly expanding IT industry and specific strategies in place for growth in the telecommunications industry.

It would seem that these realities provide the ideal opportunity for the development and implementation of human language technologies (HLT), especially in view of applications such as voice access to information systems and machine translation systems which are possible given current technology.

2. Language resources and tools

In 1999 the possibility was mooted by a number of academics that an infrastructure should be created for the development reusable language resources. This process was then partly driven by the state, which through a number of policy decisions and various committees (cf. Roux & Du Plessis, 2005) led to a decision in 2003 to establish a **National Language and Speech Resource Centre.** Unfortunately, little progress has since then been made by the Department of Arts and Culture (DAC) in establishing this centre.

The **National Language Service (NLS)** residing within the DAC is involved in the development and promotion of multilingual terminologies. The website has unfortunately not been updated since 2002 and it is therefore not possible to determine the extent of the terminology data available. Terminologies have been created for primary and secondary educational purposes within the following domains: mathematics, natural science and technology, Economic & Management Sciences and Human & Social Sciences. Multilingual terminologies (in all eleven languages have been developed for the following domains: Aids, Geography, Indigenous mammals, while trilingual terminology is available for parliamentary proceedings (Hansard). Word lists that are part of spellcheckers for all the official languages are also available on the particular website. (Cf. Department of Arts and Culture 2002)

The **Pan South African Language Board (PanSALB)** is a DAC supported institution tasked, amongst others, to develop *National Lexicographic Units* (NLUs) for the official languages of South Africa. These units were established by government in March 2001 as autonomous, though governmentally sponsored Section 21 (non-profit) companies. The primary task of these Units is to take responsibility for the development of appropriate lexicons for the different languages. "NLUs must compile user-friendly dictionaries, always keeping the target users in mind. Different types of dictionaries will be compiled by NLUs depending on the needs of each language group, e.g. :

- Monolingual, i.e. comprehensive explanatory dictionaries of the general vocabulary of the language
- Multilingual, i.e. translating between different languages
- School dictionaries
- Dictionaries for learning a new language."

Unfortunately this website (Pan South African Language Board [sa]) is also dated. Although the task is quite clearly defined there are no details on specific aims achieved and/or dictionaries available.

The **Afrikaans Lexicographic Unit**, or Woordeboek van die Afrikaanse Taal, or WAT, as it is commonly referred to, has an online Afrikaans dictionary in place (Elektroniese WAT [sa]).

The **Dictionary Unit for South African English** (DSAE) (The Dictionary Unit…[sa]) has "...compiled extensive archives which reflect the diverse influences which shape South African English. Material from these archives forms the basis for both specialised and general dictionaries which the DSAE produces in conjunction with commercial publishers. Staff also monitor developments in English throughout the world, and maintain contact with international centres for the study of English."

Unfortunately none of the **Lexicographic Units for African languages** have websites of their own, which makes it extremely difficult to assess their respective targets and outputs.

It is indeed unfortunate that, while national government policy acknowledges the official status of eleven official languages, and while it has institutional structures to implement this policy, it is virtually impossible for the average citizen or serious researcher to obtain information on the nature and extent of resources available. This is a deficiency that could be addressed as part of the foreseen network of resources that are to be debated at this workshop.

Whilst knowledge on government language resources is rather limited, researchers at academic and semi-academic institutions have already gone a long way in developing language resources as components of particular research projects. The following institutions are actively involved (directly and indirectly) in the development of text and speech resources for different languages spoken in Southern Africa:

The **Department of African Languages at Pretoria University** has been actively compiling general corpora for all eleven official South African languages. These corpora are solely utilized for student training and academic research, specifically in the fields of lexicography, terminology, linguistics, translation practice and corpus-based translation studies (CTS). The sizes of the various corpora range from 2 million tokens for Venda to 6 million tokens for Northern Sotho and Zulu. Spellcheckers, based on the corpora, have been developed for most of the languages, while the following dictionaries were also created: New Sepedi Dictionary, Popular Northern Sotho Dictionary, Nuwe Sepedi Woordeboek.

(Cf. University of Pretoria…2003). Tools for automatic corpus annotation for Northern Sotho are being developed in cooperation with the Institut für Maschinelle Sprachverarbeitung (Universität Stuttgart…2006).

A multi-disciplinary project on the development of computational morphological analysers for South African Bantu languages is based in the **Department of African Languages at the University of South Africa,** and includes computer scientists from the School of Computing. The project mainly focuses on the development of finite-state morphological analysers for five Bantu languages, namely Zulu, Xhosa, Swati Northern Sotho and Tswana, using the natural language independent Xerox Finite-State Tools (Beesley & Karttunen, 2003). Analyser prototypes for these languages are at various stages of development, with the Zulu analyser prototype (ZulMorph) being most advanced. The development of morphological analysers is based on underlying machine-readable lexicons that conform to common lexical specifications and international standards (cf. Bosch et al., 2006). A data model towards a standardised machine-readable lexicon for all languages in the project has been developed and formulated as an XML DTD. Electronic lemma lists extracted from paper dictionaries have been compiled for Xhosa and Zulu, while a machine-readable Xhosa lexicon in the form of an XML document is in the process of being developed. (Cf. University of South Africa…2005)

The **Department of Linguistics at the University of South Africa** is involved in a collaborative research project with the Linguistics Department of the University of Göteborg. The main objective of the project entitled Spoken Language Corpora for the 9 official African Languages of South Africa, is to develop a platform of computer supported basic linguistic resources for the previously disadvantaged languages of South Africa. The aim is to collect audio-visual recordings of +/-300 hours of spoken language used in a variety of social activities in natural settings for each one of the languages (cf. Spoken Language Corpora…2005). In Allwood and Hendrikse (2003) it is reported that in the pilot study on Xhosa 171 hours of spoken interaction (1.34 million tokens) have been recorded. Furthermore an automatic computer tagger is being developed.

The **Centre for Text Technology at the North-West University**, Potchefstroom has four main activities, namely research (including basic research, strategic research, applied research, and market research); development (including development of sources and end-user applications & products); commercialisation of products and services; and maintenance of products and support to end users/clients. Products include spelling checkers for Afrikaans as well as for five South African languages developed in collaboration with the University of South Africa and the University of Pretoria. Available corpora and technologies are not listed. (Cf. Centre for Text Technology 2006).

The **Unit for Language Facilitation and Empowerment at the University of the Free State** is involved in a joint project with the Province of Flanders and the University of Antwerp entitled *Multilingualism, Informatics and Development (MIDP*. The broad aim of the MIDP is to establish a central system where all provincial (and eventually all local) governmental spatial and non-spatial information (data) can be warehoused. As part of the endeavour all information will be made available in English, Afrikaans and Southern Sotho, the three major languages of the Free State province. (Cf. Unit for Language Management 2005).

The **Department of Computer Science at the University of Limpopo** runs a Centre for Speech Technology, with a specific focus on the construction of synthesisers and automatic speech recognition systems for languages such as Northern Sotho, Tswana, Venda and Tsonga. Unfortunately no information is available on the nature of language resources (or tools) developed for these activities. (Cf. University of Limpopo 2004)

The **Speech Technology and Research group at the University of Cape Town** is involved in a number of projects, some involving African languages. The website is dated and unfortunately it is not possible to obtain a detailed overview of resource developments or research outputs. (Cf. Speech Technology…2003)

The **Human Language Technology Research Group** within the **Meraka Institute** is actively involved in developing open source software for language and speech technology applications, *inter alia*, in three African languages. A particularly interesting tool is the so-called *DictionaryMaker* which can be used to develop an electronic pronunciation dictionary. No further details are available on specific resources developed. (Cf. Meraka Institute 2005).

The **Centre for Language and Speech Technology at Stellenbosch University (SU-CLaST)** has been involved through its predecessor, the Research Unit for Experimental Phonology (RUEPUS), in developing telephone speech databases in five of the local languages, i.e. in Afrikaans, English, Zulu, Xhosa and Southern Sotho. These databases were collected in terms of the well known SpeechDAT formats and protocols, and focused on obtaining data on different accent types in, respectively, Afrikaans (two accent types) and South African English (five accent types). Due to a number of constraints it was not possible to gather data on accent / dialect in the African Languages involved. These eleven databases formed part of the African Speech Technology (AST) project. These databases are all orthographically transcribed and phonetically annotated and are available for research purposes under specific conditions. The Centre will embark on a major speech data gathering project in 2006, focussing on natural spoken language in respectively, Xhosa and Zulu as part of a joint project on high quality speech synthesis with the Institute for Advanced Telecommunication Research (ATR) in Kyoto, Japan. (Cf. Stellenbosch University…2005).

Various types of tools have been developed by SU-CLaST; this includes the following:

- Patana: A deterministic grapheme-to-phoneme converter for nine African languages.
- Astudio: A software toolkit comprising eleven sub-programmes with graphic interface to develop interactive speech applications
- A set of Java-scripts to check the validity of speech databases, i.e. whether annotations are consistently made according to fixed specifications.

For more detail please see African Speech Technology (2005).

Finally, it needs to be mentioned that as for Khoisan language and speech resources, the following website is of

great value: Web Resources for African Languages (2006).
Given the above mentioned resources for African languages, there seems to be a general lack of information on the availability of data and tools for research and/or commercial purposes. Obviously IP issues play a great role in this process. The establishment of the long awaited Resource Centre for Human Language Technologies will to a large extent address some of these pressing issues, however, this does not preclude researchers in the field to set up own networks, such as those that are anticipated to arise during the LREC 2006 Workshop on networking the development of language resources for African languages.

3. Standardisation

When developing reusable resources in text and/or in speech format, it is extremely important that there is consistency as far as the processing of the data is concerned. It is a well known fact, that "linguistic annotation" lacks widely accepted standards (cf. Bird & Liberman, 2001:23), and that there are a variety of speech and text annotation schemes available.
Within the South African context a mirror image committee of ISO TC37 (Terminology, other language and content resources), that is StanSA TC37, was instituted by the Standards South Africa in 2003. The scope of this technical committee (TC) is "Standardisation of principles, methods and applications relating to terminology and other language resources." (Standards South Africa 2006). This committee, which can comprise any individual or institution active in the field, meets regularly once a year. At the recent meeting in February 2006, a workshop on setting standards for the annotation of text in African languages was held. A follow up of this meeting is to take place in September 2006.
The domain of standardisation will also be an important point to discuss at the Workshop in Genoa. Specific attention should be paid to ways and means to co-operate within the African continent.

4. Conclusion

It is clear that although a great amount of work has been done on the development of language resources in South Africa over many years, there is an urgent need for some type of infrastructure to present the scholar, researcher, or developer with up to date information on the status of text, speech or multimodal data available in different languages. Currently, very little information is available on resources created by government or semi-government organizations. Academics are likewise also apprehensive in making resources available that have been developed as part of larger research projects due to the effort and funding that has gone into this development. A factor that is certainly impacting negatively on the availability of multilingual text and speech data is the fact that this type of data is extremely valuable in the commercial sector, i.e. in the development of HLT products. As such it is not easy to control or check on data that have been earmarked for research purposes, reaching the commercial market in one form or the other. A central depository, with mechanisms in place that will coordinate resource development, ensure adherence to international norms and standards, and attend to legal matters and contracts regarding Intellectual Property rights, will to a large extent address these issues. Such an infrastructure has been proposed by the Ministerial Advisory Panel on HLT, however, the Steering Committee that has to oversee its implementation has unfortunately not yet been able to make real progress in this regard.
The diagram in the Appendix presents personal views of the authors (based on original recommendations in the report to the Minister in September 2002) on the possible structure and functions of a Resource Centre for HLT. Such an infrastructure could eventually also form part of a Network for Language Resources to be debated at this workshop.

5. References

African Speech Technology. (2005). [O]. Available: http://www.ast.sun.ac.za/ Accessed on 13 April 2006.

Allwood, J. & Hendrikse, A.P. (2003). Spoken language corpora for the nine official African languages of South Africa. Southern African Linguistics and Applied Language Studies 21(4) pp. 189-201.

Beesley, K.R. & Karttunen, L. (2003). Finite-state morphology. Stanford, CA: CSLI Publications.

Bird, S & Liberman, M. (2001). A formal framework for linguistic annotation. In *Speech Communication*, 33. pp. 23-60.

Bosch, S., Jones, J. Pretorius, L. & Anderson, W. (2006). Resource Development for South African Bantu Languages: Computational Morphological Analysers and Machine-Readable Lexicons. In Workshop Proceedings (Networking the development of language resources for African languages). 5th International Language Resources and Evaluation Conference, Genoa, Italy.

Bosch, S.E. & Roux, J.C. (in press). HLT Initiatives in South Africa – implications for human development, empowerment and democratisation, in T. Bearth et al. (eds): African languages in global society / Les langues africaines à l'heure de la mondialisation, Cologne: Rüdiger Köppe Verlag.

Centre for Text Technology. (2006). [O]. Available: http://www.puk.ac.za/fakulteite/lettere/ctext/index_e.html Accessed on 13 April 2006.

Department of Arts and Culture. (2002). [O]. Available: http://www.dac.gov.za/about_us/cd_nat_language/home.htm. Accessed on 13 April 2006.

Elektroniese WAT. [Sa]. [O]. Available: http://www.woordeboek.co.za/content/tuisblad.aspx Accessed on 13 April 2006.

Meraka Institute. (2005). [O]. Available: http://www.meraka.org.za/hlt_projects.htm Accessed on 13 April 2006.

Pan South African Language Board. [Sa]. [O]. Available: http://www.pansalb.org.za. Accessed on 13 April 2006.

PanSALB Annual Report. (2001) Pretoria, Afriscot Printers.

Roux, .JC. & Du Plessis, T. (2005). The development of Human Language Technology policy in South Africa. In *Multilingualism and Electronic Information Management*. Pretoria. Van Schaik, pp 24 -40.

Speech Technology And Research (STAR, S*) Research Group. (2003). [O]. Available: http://www.star.za.net/White.html Accessed on 13 April 2006.

Spoken Language Corpora for the 9 official African Languages of South Africa Project. (2005). [O]. Available: http://www.unisa.ac.za/Default.asp?Cmd=ViewContent&ContentID=11830. Accessed on 13 April 2006.

Standards South Africa. (2006). [O]. Available: http://www.stansa.co.za/SPS_HTML/037.html#SCOPE Accessed on 13 April 2006

Statistics South Africa. (2005). Report-03-02-01 - Census 2001: Key results, 2001. [O]. Available: http://www.statssa.gov.za/publications/publicationsearch.asp?PN=h4fehj-qt-qs-qr&PM=&PY=&PS=1. Accessed on 13 April 2006.

Stellenbosch University Centre for Language and Speech Technology (2005). [O]. Available: http://www.sun.ac.za/su_clast Accessed on 13 April 2006.

The Dictionary Unit for South African English. [Sa]. [O]. Available: http://www.ru.ac.za/affiliates/dsae/ Accessed on 13 April 2006.

Unit for Language Management. (2005). [O]. Available: http://www.uovs.ac.za/faculties/index.php?FCode=01&DCode=153 Accessed on 13 April 2006.

Universität Stuttgart Institut für Maschinelle Sprachverarbeitung. (2006). [O]. Available: http://www.ims.uni-stuttgart.de Accessed on 13 April 2006.

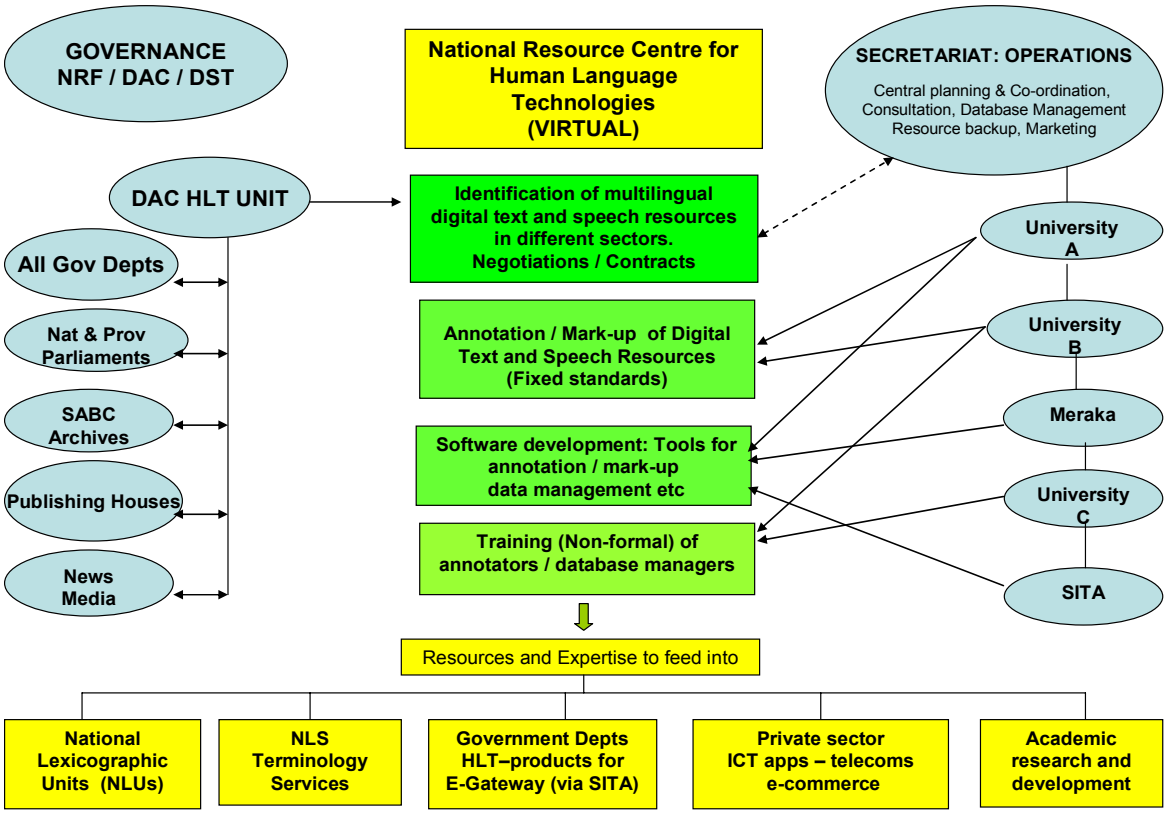
University of Limpopo Computer Science. (2004). [O]. Available: http://www.unorth.ac.za/FacultySchools/sch-comp/disciplines/comp_sci.html Accessed on 13 April 2006.

University of Pretoria Department of African Languages. (2003). [O]. Available: http://www.up.ac.za/academic/humanities/eng/eng/afrlan/eng/initiative.htm Accessed on 13 April 2006.

University of South Africa Department of African Languages. (2005). [O]. Available: http://www.unisa.ac.za/Default.asp?Cmd=ViewContent&ContentID=143 Accessed on 13 April 2006.

Web Resources for African Languages. (2006). [O]. Available: http://goto.glocalnet.net/maho/webresources/khoesan.html Accessed on 13 April 2006.

APPENDIX


GOVERNANCE
NRF / DAC / DST
National Resource Centre for Human Language Technologies (VIRTUAL)
SECRETARIAT: OPERATIONS
Central planning & Co-ordination, Consultation, Database Management Resource backup, Marketing
DAC HLT UNIT
Identification of multilingual digital text and speech resources in different sectors. Negotiations / Contracts
University A
All Gov Depts
Nat & Prov Parliaments
SABC Archives
Publishing Houses
News Media
Annotation / Mark-up of Digital Text and Speech Resources (Fixed standards)
University B
Meraka
Software development: Tools for annotation / mark-up data management etc
University C
Training (Non-formal) of annotators / database managers
SITA
Resources and Expertise to feed into
National Lexicographic Units (NLUs)
NLS Terminology Services
Government Depts HLT–products for E-Gateway (via SITA)
Private sector ICT apps – telecoms e-commerce
Academic research and development

Cologne Initiative on Natural Language Processing in African Languages

Axel Fleisch, Frank Seidel

Institute for African Studies, University of Cologne
50923 Cologne, Germany
axel.fleisch@uni-koeln.de, frank.seidel@uni-koeln.de

Abstract

Four African languages are at the core of an initiative concerned with computatioanl linguistics based at the Institute for African Studies. Our objective is to develop tools for the analysis of South African Ndebele, Bambara (Manding), (Adamawa-)Ful and Lingala. All of these are languages of wider distribution and/or receive backing by national governments to be advanced towards the use of computational methods in order to promote their use and strengthen their status. From a linguistic viewpoint, the selection of languages (with differing degrees of agglutinating versus isolating character) is interesting because their typological make-up presents a variety of challenges for computational modelling.

In the course of the article, we will present Ndebele and Bambara in more detail and address the pertinent issues with regard to the computational processing of these languages. While morphological analysis is a crucial component for the highly agglutinating Ndebele, Bambara—being largely isolating—poses different challenges that lie more in the field of morphosyntax and semantic disambiguation. Major features of Ful and Lingala will be addressed briefly, before discussing further steps, possible co-operations and the creation of language resources in the framework of our project that might be of a wider interest.

1. Introduction

Since October 2005, a group of researchers of the Institute for African Studies at the University of Cologne has engaged in initial steps towards devising working environments for African languages that include spelling checkers, morphological transducers, automatic parsers and taggers, etc. Our focus is, for the time being, on four languages, South African Ndebele, Bambara, Ful and Lingala. These languages have been chosen for several reasons some of which are purely technical (available resources and expertise). Another important reason is that each of these languages poses different typological challenges which makes the selection interesting from a more theoretical angle. This will be outlined for each language briefly in the respective sections. Accordingly, we will also explain how we intend to approach the different challenges and illustrate this with examples.

On a more programmatic note, a practical problem will be addressed in section 3. In the field of African language studies a certain reluctance to move into the domain of applied linguistics can be noticed. To many colleagues, a potential co-operation with computational linguists appears difficult. Past experiences have shown that the latter tended to be interested primarily in computationally challenging problem solutions rather than the creation of language resources. We hope that this gap can be bridged.

2. Project design and targets

Our work falls into three components determined by the respective languages. The following paragraphs will give a brief outline on what has been achieved so far and what is envisaged for the near future.

2.1 South African Ndebele

Work on the development of a finite state morphological transducer for Ndebele (Bantu, Niger-Congo) has been carried out by Axel Fleisch in co-operation with South African scholars working on closely related and typologically similar languages (Sonja Bosch, Laurette Pretorius).

2.1.1 Typological properties of Ndebele

Ndebele is a highly agglutinating language. Nouns, verbs and other parts of speech are built around a nucleus that corresponds to a lexical root, but can be extended by numerous affixes. These affixes serve a broad variety of functions. With regard to the verb, these include semantic derivation (causative, reciprocal, stative, passive, etc.; usually by suffixes) and inflection for tense, aspect, mood, polarity (by pre- or suffixes, or a combination thereof). Subjects and objects are cross-referenced on the verb by means of prefixes. This list is not exhaustive.

(1) *bazongivumela?*

ba-	zo-	ngi-	vum	-el	-a
SUBJ:3pl-	FUT-	OBJ:1sg-	'agree'	-APPL	-FV

'they will admit (<agree to) me?' [e.g. job application]

With regard to the noun the issue is slightly less complex, but there are also a number of features indicated by means of affixes. Noun class marking is obligatory and therefore no root ever occurs just by itself. In addition to this, further derivational affixes and copulative morphemes can be attached to the noun.

(2) *kunelitjana* (+ locative adjunct)

ku-	na-	ili-	tje	–ana
LOC-	ASSOC-	cl.5-	'stone'	–DIM

'there is a small stone'

These affixes can be used very productively, and the amount of possible combinations is high, although not unlimited. While, for example, the passive, applied (=benefactive) and causative extensions appear in a relatively unconstrained way on many verbs, the combination of noun class morphology is lexically specified. Each noun root occurs only in some of the overall 15 noun classes. Typically, a root can combine with at least one singular and the corresponding plural noun class prefix, although some nouns occur in one lexical noun class only (such as collective or abstract nouns), while others may occur in several classes with related, but distinguishable meanings. In addition to this, some noun classes have derivative functions (creating e.g. abstract nouns, derogatory terms, locatives from "basic" nouns). The morphology of these

classes may attach to a wider variety of roots and is sometimes pre-prefixed to the lexically-specified noun classes, leading to noun class prefix-stacking. In summary then, the system of morphologically marked noun classes follows a complex, but limited pattern.

Other relevant characteristics of Ndebele concern the phonology and morphophonological changes. The basic syllable structure consists of alternating consonant-vowel sequences, although a few consonant clusters do occur. Vowel-initial stems are not very common, but where they occur, they lead to assimilatory processes and vowel coalescence (for illustration compare the segmented to the actual surface form in example 2). Some phonological rules are rather intricate, such as those concerning the palatalisation which is triggered by various suffixes and affects labial and alveolar consonants (which are often, but not always followed by a rounded root-final vowel).

(3)
a. diminutive *–ana*
*um-lo**m**o* 'mouth' > *um-lo**ny**-ana* 'small mouth'
*isi-kha**th**i* 'time' > *isi-kha**tjh**-ana* 'moment'
b. locative *-ini*
*um-lo**m**o* 'mouth' > *em-lo**ny**-eni* 'at the mouth'
*isi-khwa**m**a* 'pocket' > *esi-khwa**ny**-eni* 'in the pocket'
(but also: *esi-khwam-eni* 'in the pocket')

What is interesting about these processes is that they represent cases of morphologically driven phonology: they are not purely phonological in that the condition that brings about the palatalisation is not the phonological environment (not any following suffix with initial /e/ or /a/ would lead to palatalisation), but specifically the diminutive and locative endings (besides a few others).

This is linked to a second interesting observation. While the occurrence of the diminutive *-ana* always triggers palatalisation of consonants that are subject to this phenomenon, in the case of the locative ending *-ini* (with the allomorph *-eni* after mid- and low-vowel stems) there is a high amount of variation with regard to palatalisation. Possible conditions include the character of the last stem vowel (palatalisation appears to be more frequent with rounded back vowels), but contrary to the state of affairs in Zulu (cf. Poulos & Msimang 1998: 520) it is not excluded in the case of stems with a different last vowel (cf. *esikhwanyeni* in example 3b above). Thus, whether a given noun palatalises is not predictable on purely phonological grounds. It appears that it is a largely idiosyncratic feature of each noun to either undergo this change if followed by the locative suffix, or to maintain the original consonant. However, free variation concerning one and the same noun root can also be encountered (cf., again, example 3b).

A third characteristic to be observed is that the palatalisation is not confined to consonants which are immediately followed by the conditioning suffix. The occurrence of the palatalised variant can be a long-distance alternation conditioned by, e.g., the passive extension *-w*. In example (4), the palatalising effect of the passive extensions is not barred despite the occurrence of the intervening causative extension *-is* closer to the stem.

(4) *ku-za**m**a* 'to try'
*ku-za**m**-is-a* 'to cause to try'
*ku-za**ny**-is-**w**-a* 'to be caused to try'

2.1.2 Modelling the linguistic facts with Xerox finite state tools

What does the typological make-up of Ndebele imply for the task(s) at hand? With regard to the development of a spelling checker, we obviously need to take care of the morphological complexity of the words to be modelled. Following the practice of South African colleagues we have been using finite state tools (in our case from Xerox; Beesley & Karttunen 2003) in order to handle the morphological complexity of Ndebele. So far, the fairly complex noun is reasonably well covered by our morphotactic architecture. The *lexc* (<lexical compilation) component of the Xerox tools provides a device to model the manifold possible constellations of the available slots in the complex word classes of Ndebele.

The following nouns contain a noun class prefix and a root: *amanzi* (cl. 6) 'water'; *iqanda, amaqanda* (cl. 5/6) 'egg(s)'; *ilitje, amatje* (cl. 5/6) 'stone(s)', *indoda, amadoda* (cl. 9/6) 'man, men'; *imvu, izimvu* (cl. 9/10) 'sheep (sg./pl.)'. Only classes 5, 6, 9 and 10 are considered here. The respective prefixes are *ili-*, *ama-*, *iN-* and *iziN-*. Morphophonological adjustments concern the shortening of *ili-* (cl.5) to *i-* before polysyllabic roots and the assimilation of the underspecified prenasal /N/ in the prefixes of classes 9 and 10 to the following root-initial consonant. Nouns except for mass nouns such as *amanzi* 'water' can receive the diminutive ending *-ana*.

```
Multichar_Symbols
5-  9-  6-  10-
@P.NCL.6@     @P.NCL.5-6@   @P.NCL.9-6@
@P.NCL.9-10@  @R.NCL.6@     @R.NCL.5-6@
@R.NCL.9-6@   @R.NCL.9-10@  @D.NCL.6@

LEXICON Root
0:0               NounPrefix;

LEXICON NounPrefix
5-@P.NCL.5-6@:ili@P.NCL.5-6@       NounStem;
6-@P.NCL.6@:ama@P.NCL.6@           NounStem;
6-@P.NCL.5-6@:ama@P.NCL.5-6@       NounStem;
6-@P.NCL.9-6@:ama@P.NCL.9-6@       NounStem;
9-@P.NCL.9-6@:iN@P.NCL.9-6@        NounStem;
9-@P.NCL.9-10@:iN@P.NCL.9-10@      NounStem;
10-@P.NCL.9-10@:iziN@P.NCL.9-10@   NounStem;

LEXICON NounStem
water@R.NCL.6@:nzi@R.NCL.6@        DIM;
egg@R.NCL.5-6@:qanda@R.NCL.5-6@    DIM;
stone@R.NCL.5-6@:tje@R.NCL.5-6@    DIM;
man@R.NCL.9-6@:doda@R.NCL.9-6@     DIM;
dog@R.NCL.9-10@:nja@R.NCL.9-10@    DIM;
ear@R.NCL.9-10@:ndlebe@R.NCL.9-10@ DIM;

LEXICON DIM
-DIM@D.NCL.6@:ana@D.NCL.6@   #;
0:0                          #;
```

Figure 1: Code written for *lexc* covering selected nouns

In addition to the above-mentioned morphophonological rules, it is important to bear in mind that the diminutive suffix triggers substantial phonological changes (coalescence of adjacent vowels and palatalisation of certain final consonants of the noun roots (e.g. *amatje* 'stones' > *amatjana* 'small stones, pebbles'; *indlebe* 'ear' > *indletjana* 'small ear'). All of these morphophonological

adjustments at morpheme boundaries need to be taken care of as well. This can be handled by regular expressions using the Xerox tools (*xfst*).

```
define Nasal [ n | m ];
define Vowel [ a | e | i | o | u ];
define Labial [ b | p | f | v ];

read regex
b -> t j || _ ( Vowel ) a n a
.o.
Vowel -> 0 || _ Vowel
.o.
N -> 0 || _ Nasal
.o.
N -> m || _ Labial
.o.
N -> n
.o.
i l i -> i || _ $[Vowel ?+ Vowel];
```

Figure 2: Morphophonological alternations in *xfst*

A few tasks are still to be undertaken. Most importantly the highly complex verb needs to be integrated. At the same time, word classes that are less susceptible to morphological affixation are gradually built into the transducer.

As a rule of thumb, morphotactics are handled in *lexc* and phonology by regular expressions in *xfst*. Semantically-based and non-adjacent co-occurrence restrictions are taken care of by flag diacritics. While this has worked properly for the class assignment of noun roots, the matter is different for verbs and their derived variants. Even though, in principle, many of the derivative suffixes are fully productive, numerous combinations of verbs with specific extensions are ruled out. Similarly, in many instances the meanings of the (often multiply) derived forms are not predictable, i.e. their semantics is not simply compositional; cf. *-sebenza* 'work' > *-sebenzisa* 'use (< cause to work)'. With regard to formal peculiarities, it has already been noted in the preceding subsection that certain variations appear to be idiosyncratic and will therefore have to be included in the lexical component rather than by productive rules. For these reasons, it seems to us that derived verb stems should not be handled in the morphotactic *lexc* component, but rather be entered as independent lemmata. Any other approach would either require a highly sophisticated system of a huge number of flag diacritics (many of which would serve to take care of only very few verb stems–sometimes perhaps even only a single one!), or the resulting transducer would hopelessly overgenerate (still not accounting for the many truly idiosyncratic formal variants that have to be entered manually anyhow).

2.1.3 Enlargened machine-readable lexicographic knowledge bases

Once the transducer works at a reasonable rate of reliability, in addition to spelling checking, it could be used for lexicographic purposes. Corpus mining is one of the immediate goals that come to mind. Lexicographers in South Africa have been working on the compilation of digitalised text corpora for a long time. These corpora are a valuable resource for the enlargement of the dictionaries available for that language. In order to provide lexicographers with a corpus-mining tool, root guessers can be constructed on the basis of the morphological transducers that are being built right now.

2.2 Bambara

The team working on Bambara (Manding) consists of Frank Seidel and Mohamed Touré. In contrast to Ndebele, the morphological analysis of words in Bambara does not appear to be too much of a challenge, since this language is largely isolating. However, there are a number of exceptional constructions which will have to be taken care of by a morphological transducer. Compounds and reduplicated forms present challenges for a finite state analysis.

2.2.1 Analysing compounds and reduplications

Among the easier-to-model constructions one finds fully productive inflectional and derivative mechanisms, such as the formation of dynamic verbs from stative verbs by suffixation of *–ya* (5a), the morphological marking of participial forms with the suffix *–len* (5b), the perfective TA marker *-la/-ra* (5c) and the plural marker *-w* attached to nouns (5d), among several others.

(5)
a. *bòn* 'be big' > *bònya* 'grow, enlargen, make big'
b. *sígi* 'seat sth.' > *sígilen* 'be seated'
c. *ń bɛ́ tága* 'I go' > *ń tágara* 'I have gone, I am gone'
d. *jíri* 'tree' > *jíriw* 'trees'

More interesting are compounds and word forms resulting from a combination of compounding, derivation and even conglomerative combinations that contain entire clauses as illustrated in (6).

(6)
a. *wùludenmuso* 'female puppy'
   < *wùlu* 'dog', *dén* 'child' *mùso* 'woman; FEM.'
b. *báarakɛlaw* 'workers'
   < *báara* 'work', *-kɛ* 'do', *-la* 'AGENT NOUN', *-w* 'PLUR.'
c. *báarakɛlamusow* 'workers' wives'
d. *díyanyebaga* 'mistress'
   < *díya* 'please a person'; *ń* 'I, me'; *yé* 'POSTPOSITION: to, from'; *-baga* 'AGENT NOUN'
e. *líwuruwaraba* 'lion out of a book (e.g. a picture of a lion in a book)'
   < *líwuru* 'book'; *wára* 'wild animal'; *-ba* 'AUG.'

The various compounding and word formation mechanisms pose certain problems for formal modeling. One of these is similar to that outlined above for the derivative verb extensions in Ndebele. Although in principle, some of the compounds can be regarded as the outcome of a productive morphological mechanism, there are clear occurrence restrictions. Certain elements can be used as full nouns (cf. ex. (5): *mùso* 'woman; FEMALE'), but—when suffixed to another noun—exert certain functions that attest to incipient grammaticalisation, such as (non-obligatory) gender marking for a certain set of animate nouns. One should take note of the difference in semantics between examples (6a) and (6c). Words of this type probably form a semantically determined closed part-of-speech sub-category.

Be that as it may, the productivity of N-N compounding is still high as illustrated through (6e) which might in

some regards be somewhat unlikely or silly but completely acceptable. Thus as Culy (1985:347) mentions creative word formation in Bambara seems to be limited more by interpretability rather than grammatical rules. This holds for the next issue as well.

Even more problematic for morphotactic modelling are the different possibilities of composition and conglomeration based on a variety of elements when forming nouns. For example in (6d) the first part *díya ń yé* 'please me', an otherwise grammatically acceptable full sentence, is prosodologically altered and treated as a noun that can take a nominal derivative suffix. Such procedures are quite frequent in Bambara and can be productively employed by a skilful speaker. Complicating the matter further is a high number of lexicalised words created through derivation, composition and/or conglomerations. Oftentimes the resulting semantics can not be logically deduced from the process of formation and the meaning of the single parts, such as the second nominal part in (6e) consisting of a noun and an augmentative derivation that is lexicalised as 'lion'.

At the end of the day, it may very well turn out that what looks like obvious instances of compounding should rather be treated as fully lexicalised. Yet, since innovations can be created *ad hoc* if needed, the purely lexical solution is not fully satisfactory either. A possible solution we are currently contemplating is to treat most nouns as lexemes in *lexc* and to create an additional analytic tool that is able to provide guesses as to the morphological segments of the compound or conglomerative nouns. This tool will most likely overgenerate hopelessly, but by externalizing it the user can at least be provided with some control over the application by being able to switch the analytic tool on or off as needed.

Concerning the issue of reduplication in Bambara there exist already some ideas on the solution of the problematic issues. The problem was introduced through Culy's (1985) article to a wider audience in terms of formal linguistic theory and eventually reached finite state morphology computational linguists. Due to space considerations we will limit ourselves to referring to a selection of the literature; consider Walther (2000) and Bachmann (2005)

2.2.2 Beyond the boundaries of the formal word

If most of the morphologically complex forms have to be treated as lemmata in Bambara the question arises, why one should not try a different route to its processing. A spelling checker, for example, could theoretically be devised on a pure wordlist basis, with a few additions (allowing certain parts-of-speech with a plural variant, transitive verbs with the additional perfective ending, etc.). Or, in order to take it to the extreme, it could almost be developed as a tool checking simply for correct phonotactics. As mentioned above words are, with very few exceptions, CV(CV) and certain restrictions apply as to what consonants and vowels may occur in which slot (and co-occur with one another). Generating random strings on the basis of these restrictions might already lead to a fairly comprehensive account of possible or even probable Bambara words.

But what good would it be to have an instrument that checks exclusively for morphological correctness within word boundaries in a primarily isolating language? If the tool that is to be developed to process Bambara data was restricted to morphological analysis, its application would be trivial, since words tend not to be very complex on the morphological level. The whole issue becomes more challenging as we leave the smallest word-level. Certain morphosyntactic phenomena present a much richer field. Word order is fairly strict, and to have a tool which takes morphosyntax into account will prove to be very valuable. One obvious application is grammar checking.

There are a number of phrase-level phenomena. Plural marking *-w*, e.g., occurs only once in the NP, on the last element (cf. the adjective *fítinin-w* 'small-PLUR' in 7a), and only if number is not expressed by an explicit numeral (such as *dúuru* 'five' in 7b).

(7)
a. *nìn lèmuruba fítinin**w*** 'those little oranges'
b. *lèmuruba fítinin (*fítininw)* ***dúuru*** 'five little oranges'

What is perhaps more important in the long run is that it would be insufficient to simply decide whether a given string is a valid word in Bambara, or not. Such an approach would miss out on the lexicographic side. If our tool is intended to contribute to the growth of lexicographic databases, knowing not only whether a string could be a valid word, but also whether it occurs in an appropriate syntactic position will obviously be an enormous help in various ways. These concern automatic part-of-speech tagging, consequently semantic disambiguation and the identification of homonymous lemmata with (certain) meanings that are not accounted for in our lexicon.

2.2.3 Semantic disambiguation and lexical information

Automatic POS-tagging is rendered difficult by the fact that the N-V distinction is not very clear. Many lexical items can be used interchangeably. Besides that, many syntactic constructions are structurally ambiguous—despite a fairly rigid constituent order (cf. the common, but notoriously ambiguous, sentence structures like S-Aux-Complement-Postposition ~ S-Aux-Verb-TA marker(=postposition). There is little doubt that these contructions are diachronically related. They are still a challenge for synchronically reliable automatic POS-recognition. A few additional properties come into play. Bambara has a fairly strict CV(CV) syllable structure and almost no consonant clusters. The number of homonyms is relatively big. The fact that, in addition to real homonyms, tone differences are usually not marked, leads to an even higher number of homographs. We anticipate that, eventually, semantic disambiguation will become the most crucial part in this project.

All of these features indicate that the creation of a machine-readable lexicon is a task of utmost importance. There is a variety of printed sources, but to our knowledge no easily available lexical databases in electronic form. Since Bambara/Manding is a major language of Western Africa, such a resource would be highly welcome by scholars, as well as members of the community outside academia.

In order to ensure that the lexicon that is to be created will be efficient for our purposes, but also eventually serve a broader public that possibly requires a diverse

range of applications, we are presently working on a general database structure that will be the backbone of our machine-readable lexicon. For this task insights from language typology need to be considered as well as the specific make-up of Bambara.

2.3 Further languages

Apart from Bambara and Ndebele, the work on other languages is in its initial phase. One of these languages is Ful that is being worked on by Anne Storch and Doris Richter gen. Kemmermann. In addition to this, Helma Pasch has undertaken initial steps towards the processing of Lingala.

2.3.1 Ful

Ful presents many challenges for computational modelling. Ful varieties used in different regions show a considerable degree of disparity. One of the issues that we intend to pursue concerns the question to what extent it is feasible and practical to accommodate linguistic variation in the tools that we are developing. Another property that makes Ful particularly interesting concerns its rather intricate morphophonological features. Both issues will be dealt with very briefly in the following.

Ful (Fulfulde and Peul are other regional designations for this language, the latter being more commonly used in francophone countries) is spoken in a vast region stretching from the Atlantic coast in Guinea and Senegal far to the East into Nigeria, Cameroon and Tchad. This extensive area in the West African Sahel region is limited to the North by arid areas in which insufficient rainfall hardly allows for pastoralism, let alone permanent settlement of a more sedentary type. In the South, it is limited by areas which are more suitable for horticulture/agriculture and had always been more densely populated by other groups. The mainly pastoralist speakers of Ful thus spread along the semi-arid Sahel belt.

Their history of migration and the presently rather dispersed settlement patterns led to a situation in which Ful is hardly ever the only or dominant means of communication. However, in a few exceptional regions it has attained the status of a *lingua franca* (parts of eastern Nigeria/northern Cameroon; parts of Guinea/Senegal). In certain countries, it is recognised as a national language, usually as one among many others. The overall number of speakers is difficult to estimate, but is considerably high. What is interesting for our purposes is the fact that the speakers use a number of regional variants and there is no agreed-upon standard valid for all of them. Therefore, the issue of language variation is very important. Although we depart from Adamawa-Ful, it would be shortsighted to pick out just one variety because the applicability of the results would be too limited. The intention behind this is to ensure that in order to cover other varieties, one would not have to start all over and develop a fully independent tool from scratch when dealing with another local variety. Therefore, the set-up of the project has to be a modular design, with shared core-components and more specific filters for regional variation. Our approach will be to deal with shared linguistic features first, include local speech forms being as comprehensive as possible, and build in language-specific filters towards the end of the cascading networks, so that they can rather easily be adjusted to specific user requirements and local variation.

In terms of linguistic typology, Ful shows several phenomena that are not yet fully understood: initial consonant mutation and suffix grade variation, to name the most obvious ones. Initial consonant mutation affects the quality of the stem-initial consonant, which changes from plosive to fricative, to prenasalised or vice versa. Since Ful is a suffixing language, the morphophonological influence of the class marker on the initial stem consonant must be explained diachronically as a reflex of a former prefix. The variation of suffix grades in turn may have a semantic dimension, as no morphological or phonological rules can be given for the choice of one of the usually four suffix varieties.

While in the cases of Ndebele and Bambara, we face a situation in which it is at times not easy to decide whether to deal with certain phenomena within the morphotactic *lexc* component of the Xerox tools, or to handle them in the lexicon (cf. the derivative verb extensions in Ndebele, compounds in Bambara), here, we have a case in which the morphophonological system can not be easily described in terms of rule-based phonological processes. Usually, one would opt for handling phonological processes by regular expressions (cf. the Ndebele example illustrated in fig. 2). If, however, the (morpho)phonology shows considerable idiosyncrasies, this may turn out not be feasible. One possibility to deal with this would be to take care of such properties in the lexicon. However, since the phenomenon is very pervasive and affects the entire lexical inventory, we do not favour this option. How to find another solution to this problem (possibly by doing a better descriptive job, which might unearth more complex morphophonological processes, ultimately taking care of the problem in a rule-based manner) is a question the team working on Ful is presently trying to address.

2.3.2 Lingala

Like Ndebele, Lingala belongs to the large group of Bantu languages (Niger-Congo). Accordingly, it shares a number of typological features with Ndebele. It is a highly agglutinative language with a noun class system similar to that of Ndebele, and a considerable amount of inflectional and derivative morphological markers affixed to the verb. The language is used by millions of speakers mainly in Central Africa and Europe.

One aspect that makes the language particularly interesting for us is its typological similarity to Ndebele. We have argued that different varieties of Ful should ideally be covered by one system that adapts to regional differences. There is no doubt about the fact that Lingala and Ndebele are completely different languages. But it is interesting to gain more experience on how similar or different both processing systems will have to be in their underlying architecture. This might lead to interesting insights with regard to further Bantu languages which one might wish to work on at some point in the future.

In a similar vein, teams of scholars working on the Nguni languages in South Africa (Zulu, Xhosa, Ndebele, Swazi) have started to formulate standards for their source codes when working with comparable software. This is done in order to ensure that experts from one team can jump in with relative ease if a sister-team requires help. In

this manner, the projects dealing with languages that have only been worked on for a shorter period of time can take advantage of the experiences made with the more established components of the overall project (in this case the work on Zulu, cf. Bosch & Pretorius 2003a; 2003b). In the case of the Nguni languages, the four varieties that are presently covered in the project are very closely related. This holds for their morphosyntactic constructions, their phonology and their lexicons. In this regard the case of Lingala and Ndebele is clearly different. Despite a high amount of cognates, the lexicons are completely different, as are the specific grammatical structures that have to be modelled. The question is to what extent a high degree of typological closeness will ease the process of constructing the tools that we envisage also for Lingala.

3. Future perspectives, co-operations and language resources

For the time being, we have started to work specifically on Ndebele and Bambara, with two slightly different focus areas. While morphological analysis is the major task in Ndebele, for Bambara we decided to lay more emphasis on the development of the machine-readable lexicon. The reasons for this are the difference in morphological complexity on the one hand, and the different access to language resources on the other hand.

A government-funded lexicography unit for Ndebele is in existence in South Africa and we hope to be able to intensify our co-operation. While access and the permission to use lexical data in electronic format would be an obvious advantage for us, the morphological analyser that we are presently building can then be easily developed further into a root guesser. Such a tool would facilitate the lexicographers' task of extending their lexicographic knowledge base by applying it to their growing corpus of running text. As mentioned earlier, no similar body seems to exist for Bambara. We have not yet discarded the idea to co-operate with possible partners who have developed electronic lexical databases. In the meantime, however, we assume that the creation of such a database will be part of our task.

The type of challenges that we are facing should make it quite clear why specialists in African languages (rather than a computer scientist with a probably rather limited experience with regard to the respective languages) will have to carry out a substantial part of the overall task. The descriptive challenges for the languages that we are working on are still huge. Although reference material is available for all of these languages, it is often hard to filter out the type of information that can serve as input to the formal modelling required in the creation of the language resources that we are planning to provide.

As mentioned in the introduction, it is not easy to bridge the gap between the philological tradition of our own discipline—the study of African languages and cultures with its descriptive and historical bias—and computational linguistics. At the same time, it has been oberved that computer linguists may have dealt with African languages out of specific computational interests, but often appear to lack the perseverance necessary for the development of broader applications and language resources, especially for languages that still have quite a way to go in this regard.

4. References

Bailleul, C. 1996. *Dictionnaire Bambara-Français*. Bamako: Editions Donniya.

Beesley, K.R & Karttunen L. (2003). *Finite State Morphology*. Stanford, CA: Center for the Study of Language and Information.

Bosch, S. & Pretorius, L. (2003a). Finite State Computational Morphology. Finite State Analyser of Zulu. *Machine Translation* 18: 195-216.

Bosch, S. & Pretorius, L. (2003b). Computational Aids for Zulu Natural Language Processing. *Southern African Linguistics and Applied Language Studies,* 21: 267-282.

Breedveld, A. (1995) Form and Meaning in Fulfulde. Leiden.

Culy, C. (1985). The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* 8: 345–351.

Dumestre, G. (1987). *Le Bambara du Mali. Essais de description linguistique.* Paris: Université de la Sorbonne Nouvelle.

Dumestre, G. (2003). *Grammaire fondamentale du Bambara.* Paris: Karthala.

Dzokanga, A. (1979). *Dictionnaire sémantique illustré français-lingala* Leipzig: VEB.

Gottschligg, P. (1997). Nominale Morphophonologie des Ful von Klingenheben bis Paradis: Kleine Rezeptionsgeschichte. *Afrika und Übersee* 80.

Joffe, D. & de Schryver, G.-M. (2004). TshwaneLex. A State-of-the-Art dictionary compilation program. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004*: 99–104. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Joffe, D., de Schryver, G.-M. & Prinsloo, D.J. (2003). Computational features of the dictionary application "TshwaneLex". *Southern African Linguistics and Applied Language Studies* 21 (4): 239–250.

Kawata A.T. (2003). *Bago - Dictionnaire Lingala/Falanse Français/Lingala*, Paris: Editions Le laboratoire de langues congolaises.

Poulos, G. & Msimang C.T. (1998). *A Linguistic Analysis of Zulu*. Cape Town: Via Afrika.

Storch, A. (1995). Die Anlautpermutation in den westatlantischen Sprachen. *Frankfurter Afrikanistische Blätter*, Sonderheft 2.

van Everbroeck, R. (1969). *Le Lingala. Parlé et écrit.* s.l.

van Everbroeck, R. (1985). *Dictionnaire lingala-français français-lingala*. Kinshasa: Editions L'épiphanie.

pdf-files available on the internet:

Bachmann, A. (2005). Nicht Kontextfreiheit des Vokabulars des Bambara. pdf-file available at http://evilazrael.net/~StyxGuardian/acr/documents/2004ws/sprawi/hausarbeit-nicht-cfg-r1.pdf (last accessed 17.03.06).

Karttunen, L. (2002). Lingala script. pdf-file available at http://www.stanford.edu/~laurik/fsmbook/LSA-207/Scripts/lingala.script (last accessed 17.03.06).

Walther, M. (2000). Finite-State Reduplication in One-Level Prosodic Morphology. pdf-file available at http://arxiv.org/PS_cache/cs/pdf/0005/0005025.pdf (Last accessed 17.03.06).

Speech mining to make African oral patrimony accessible

Nimaan Abdillahi *†, Nocera Pascal †, Bonastre Jean-François †, Bechet Fréderic †

† Laboratoire Informatique d'Avignon - CNRS / Université d'Avignon et des pays du Vaucluse
BP 1228 84911 Avignon, Cedex 9, France
* Institut des Sciences et des Nouvelles Technologies - Centre d'Etudes et des Recherches de Djibouti
BP 486 Djibouti, Djibouti
{nimaan.abdillahi, pascal.nocera, jean-françois.bonastre, frederic.bechet}@univ-avignon.fr

Abstract

Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. This paper presents the first steps for automatic transcription and indexing of African oral tradition heritage, particularly the Djibouti cultural heritage.

1. Context of this study

In most African countries, the cultural and historic patrimonies are inherited orally through generations. This ancestral knowledge gathered during centuries is today threatened of disappearing due to the globalization process, the economic situation and the lack of interest of the young generations for the traditional way of life. Several national, regional and international organizations (Unesco, 2003) are elaborated policies to save this human richness. Today, African countries have databases of cultural audio archives coming mostly from radio broadcast sources, and recorded during the last forty years. They are now concerned by two main issues: saving this patrimony by digitalizing the recordings and exploiting the data. Concerning the first problem, the techniques are well known and digitalization is mostly a logistic problem. The second problem is less straightforward as facing a huge amount of data requires automatic tools. Particularly, automatic transcription and indexing tools are necessary for accessing the richness of the databases. These tools are language-dependent and need to be adapted to each of the African languages targeted. This work is focused on the processing of the Djibouti oral patrimony [1]. The republic of Djibouti launched a wide digitalization program of radio broadcast archives [2].

2. Methodology

Accessing the richness of the African oral patrimony requires automatic search engine. Most of the known search engines concern text. So, a transcription phase is necessary for the automatic indexing of audio archives. Nowadays, statistical Automatic Speech Recognition (ASR) systems can reach a good level of performance for a wide range of languages if training data (both for the acoustic and linguistic models) are available. Unfortunately, it is difficult to get enough textual corpora for African languages. This is mainly due to the oral tradition system of these countries. With the development of the information technologies, Internet documents appear like a probable solution. Several works have been undertaken by different teams (Ghani et al., 2000), (Vaufreydaz et al., 1999) for the automatic construction of textual corpus for the resource-scarce languages [3] by using web documents. However, this method is limited by the number and quality of the websites available on Internet for each language. For some of them, there are not enough data. In this case, others solutions must be considered. Anyway, this kind of collected corpus is not adapted to the audio archives we project to transcribe. There are temporal and thematic mismatch between them. This mismatch will produce high out of vocabulary (OOV) rate and high word error rate (WER). However, the NIST Topic Detection and Tracking (Fiscus and Doddington, 2002) and TREC document retrieval evaluation programs has studied the impact of recognition errors in the overall performance of Information Extraction systems for tasks like story segmentation or topic detection and retrieval. The results obtained by (Fiscus and Doddington, 2002) have shown that this impact was very limited, compared to those obtained on *clean* text corpora. Similar results were obtained during the TREC program for a document retrieval task (Barnett et al., 1997). Then, to automatically index African audio patrimony we present the speech mining system shown in figure 1. It is composed of

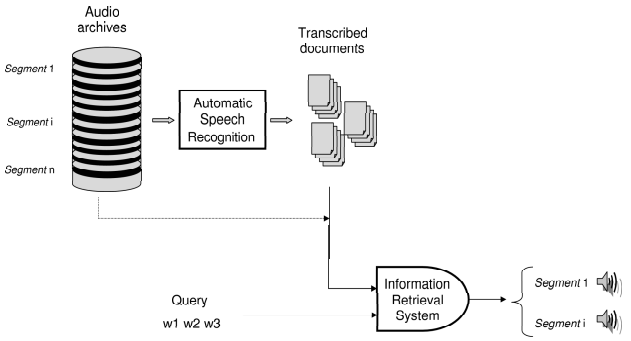


Figure 1: Speech mining system

[1] A part of the Djiboutian oral heritage is in Somali language. Other part is in Afar language

[2] http://www.rtd.dj

[3] In the literature, the terms "minority language", "less-equipped languages" and "resource-scarce languages" appears to design "less computerized" languages

	Labial	Labiodental	Dental	Alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Voiced plosives	b		d		dh		g	q		’
Voiceless plosives		t				k				
Nasal	m			n						
Voiceless fricatives		f		s		sh		kh	x	h
Voiced fricatives						j			c	
Trill				r						
Lateral				l						
Approximants	w					y				

Table 1: Somali-consonants phonetic structure.

two modules: automatic speech recognition (ASR) one and an information retrieval (IR) one. The information retrieval system uses the hypothesis files provided by the ASR module. In this paper, we present, the Djibouti languages and more precisely Somali one. Secondly, we describe our first Somali speech recognizer and the results obtained. Finally, we discuss about using Somali “roots” to deal with the inevitable mismatch between the audio archives and the training corpus available.

3. Djibouti languages

Four languages are spoken in Djibouti. French and Arabic are official languages, Somali and Afar are native and widely spoken. This work is dedicated to process Somali language, which represents half of the targeted audio archives. This language is spoken in several countries of the East of Africa (Djibouti, Ethiopia, Somalia and Kenya) by a population estimated between 12 to 15 millions of inhabitants [4]. It is a Cushitic language within the Afro-asiatic family. The different variants are Somali-somali, Somali-maay, Somali-dabarre, Somali-garre, Somali-jiiddu and Somali-tunni. Somali-somali and Somali-maay are the most widely spread variants (80% and 17%). We only process the Somali-somali variant, frequently known as Somali language and spoken in Djibouti.

The phonetic structure of this language (Saeed, 1999) has 22 consonants and 20 vowels (5 basic distinctions which all occur in front and back versions. These 10 all occur in long and short pairs). Table 1. resumes the consonants structure. Somali is also a tone accent language with 2 to 3 lexical tones (Hyman, 1981), (Saeed, 1993), (Le-Gac, 2001). The written system was adopted in 1972 (SIL, 2004), and there are no textual archives before this date. It uses Roman letters and doesn’t consider the tonal accent. Somali words are composed by the concatenation of a small number of sub words, named “roots” in this paper. Their forms are mostly (Bendjaballah, 1998) CVC, CV, VC and V [5], etc.

4. Automatic Speech Recognition

4.1. Corpora constitution

In order to build an ASR system for the Somali language, we collected an audio and a textual corpus. With the development of the information technologies, many works have been undertaken by using Internet documents for the resource-scarce languages (Ghani et al., 2000), (Vaufreydaz et al., 1999). We applied this kind of strategy and downloaded from Internet various Somali documents. The textual corpus gathered contains 2 820k words and 121K different words. Table 2 shows the distributional properties of this textual corpus.

Unit	Total
Sentences	84.7k
Words	2 820k
Distinct words	121k
Roots	6 042k
Distinct roots	4.4k
Phones	14 104k
Distinct phones	36

Table 2: Distributional properties of the Somali textual corpus.

We also downloaded a subset of text from Internet for the audio recordings. This text was read by 10 speakers. The recordings were done in a quiet environment. We obtain a Somali audio corpus named “Asaas” [6] composed of 10 hours of speech and the corresponding transcriptions in Transcriber format (Barras et al., 2001). It contains 59k words (10k different words) and it is digitalized with a sampling rate of 16 KHz and a precision of 16 bits. The audio corpus is phonetically well balanced. This corpus was divided into two subsets: 9,5 hours for the training subset and 0,5 hours for the evaluation subset. The figure 2 shows the phoneme duration in Asaas corpus.

4.2. Normalisation tools

Several tools (Nimaan et al., 2006) have been developed to process Somali texts for audio and language processing. Somali language is a recent written language and the spelling is not normalised. The same word can be written with a wide range of different forms (*jibuuti, jabuuti, jibbuuti, jabbuuti, jabuudti*). Another difficulty is due to the morphology of Somali words (concatenation of roots). Some words appear sometimes splited in two components

[4] http://www.ethnologue.com

[5] C=Consonant, V=Vowel

[6] Asaas means beginnings in Somali language

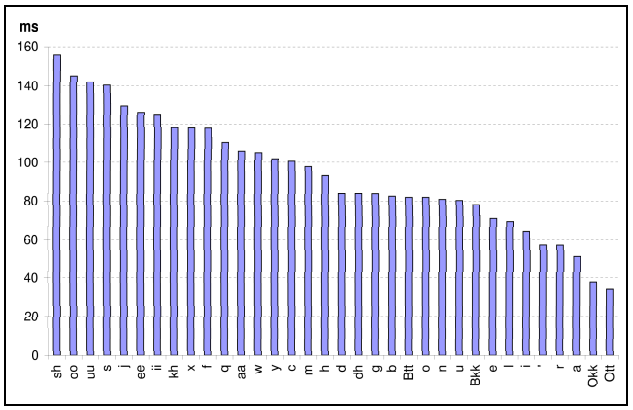


Figure 2: Phoneme duration in Asaas audio corpus.

(*ka dib* and *kadib*). These multi-spelling forms must be taken into account for the development of human language technologies for languages with recent written form. To solve this problem, we have developed a set of tools of Somali text normalization. To each word in a text, is associated its most frequent written form. If the word *dhaw* appears 11 times in the corpus and *dhow* 7 times, *dhaw* will be considered as the exact transcription. A serie of transducers have been developed to transform into textual-form the different abbreviations and numbers which appear in the corpus, like dates, telephone numbers, money, etc. A morphological analyzer has also been developed for extracting roots from Somali words. We choose 4 types of roots : CVC, CV, VC and V. We first extract the CVC roots from words, after the CV roots, and finally the VC and V. This algorithm produces 4400 different roots for the whole corpus. We also developed a Somali phonetizer named SOMPHON to transform text into phonemes, inspired by the French one LIA_PHON (Bechet, 2001), for the audio modelling.

4.3. Experiments

In this section, we describe our first Somali large vocabulary recognition system.

4.3.1. Acoustic models

The first Somali acoustic model was obtained from a French one, and was used, as a baseline, to produce the first audio segmentation of the Asaas corpus. To build this model, we established a concordance table between Somali and French phonemes. The first audio segmentation was used to produce a new Somali acoustic model with the LIA acoustic modelling toolkit. We iterated the segmentation and learning processes many times. We also tried a different initialisation by using the confusion matrix between French and Somali phonemes, to obtain an automatic baseline model. Figure 3 shows the results obtained by the two initialisations (knowledge-based and automatic). After 3 iterations, the results are similar. This confirm the previous studies done for a fast language independent acoustic modelling methods (Beyerlein P., 1999). We adopted 36 models[7]. Acoustic models are composed of 3 states by phoneme, except for the glottal plosive phoneme coded on one state (taking into account its duration). For the moment, we used non contextual models with 128 Gaussian components by state. The speech signal is parameterized using 39 coefficients: 12-mfcc coefficients plus energy and their first- and second-order derivative parameters. The cepstral mean removal and the normalization of the variance have been performed sentence by sentence.

[7]We considered only 10 vowels (5 longs and 5 shorts)

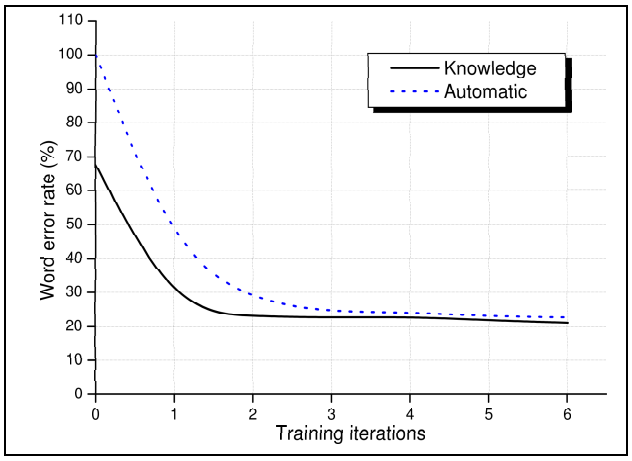


Figure 3: Learning process for the Somali acoustic model with knowledge-based and automatic methods. The decoding was done with a trigram language model.

4.3.2. Language model

A trigram language model trained on the Somali textual corpus with the LIA toolkit and CMU toolkit (Rosenfeld, 1995) has been obtained. We extract a 20K word lexicon from the most frequent words and a canonical phonetic form was produced for each entry using Somali phonetizer. The language model is composed of 726K bigram and 1.75M trigram. The perplexity of the language model on the test corpus is 63.97 with 6.77% of Out-Of-Vocabulary words. Likewise, we trained a trigram language model based on roots. The entire textual corpus was transformed in roots form. We obtained 4.4k, 189k and 996k trigram of roots. The perplexity of this model is 19.05. With the test corpus, we obtained 0.03% of Out-Of-Vocabulary roots.

4.4. Results

This paragraph presents the first results of the ASR system for the Somali language. Speech decoding is made with the LIA large vocabulary speech recognition system Speeral (Nocera et al., 2002). The same speakers are in the test and the training sets. We obtain a word error rate of 20.9% on the 30 minutes test corpus as shown in table 3. This is an encouraging result according to the size of the training corpora (9,5 hours for the audio and 3M words for LM). Without the spelling normalization presented above, the error rate is 32%. This shows that the normalization process is necessary for recent written languages. When the evaluation is done at the root instead of the word level, we obtain a word-root error rate of 14.2% as shown in table 4. It is an encouraging result for indexing the audio archives with roots.

	Correct	Sub	Del	Ins	WER
Not normalized	75.2	19.2	5.6	7.1	32.0
Normalized	84.2	13.9	1.9	5.2	20.9

Table 3: Results of the Somali automatic speech recognition in %, with a normalized and a raw text.

	Correct	Sub	Del	Ins	Error rate
Root	87.8	8.0	4.2	1.9	14.2

Table 4: The Word-root error rate (WRER) of 14.2% is obtained with the word hypothesis files.

5. Information retrieval

Our aim is to make Djibouti audio archives more accessible. Obtaining exact transcriptions of this audio data is an extremely difficult task, the main difficulty being collecting text corpora matching all the different kinds of speech recorded. In most of the cases, no written corpus at all is available. For this reason our goal in this work is not to produce transcriptions of audio archives that would replace the original recording but rather word and sub-word transcriptions that can be use for performing Information Retrieval processes for accessing the audio data. Indeed, even in perfect transcriptions all the non textual information included in the audio data (prosody, emotions, ...) is lost.

6. Conclusions and perspectives

Results of this first Somali large vocabulary recognizer are encouraging. We demonstrate that a normalizing process is necessary for Somali language and probably for all recent written languages. We reduce the WER of about 34%, after the normalization process. This work is the first step for the automatic transcription for indexing Djibouti cultural audio heritage. One perspective is to work in a root-based decoder in order to be more robust to thematic and temporal mismatch between training and testing corpora. We also project to transpose our results to the Afar language spoken in Djibouti. We believe that the work done within this project will be useful not only to the Somali language but to several oral tradition countries.

7. acknowledgment

This research is supported by the Centre d'Études et des Recherches de Djibouti [8] (CERD), the Service de Coopération et d'Action Culturelle [9] (SCAC) and the Laboratoire Informatique d'Avignon [10] (LIA).

8. References

J. Barnett, S. Anderson, J. Broglio, M. Singh, R. Hudson, and S. Kuo. 1997. Experiments in spoken queries for document retrieval. In *In Eurospeech 97*, pages 1323–1326.

C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. 2001. Transcriber : development and use of a tool for assisting speech corpora production. *Speech Communication*, 1-2(33):5–22.

F. Bechet. 2001. Lia_phon : Un système complet de phonétisation de textes. *Traitement Automatique des Langues*, 2(1):47–67.

Sabrina Bendjaballah. 1998. La palatisation en somali. *Linguistique Africaine*, (21 - 98).

Huerta J.M. Khudanpur S. Marthi B. Morgan J. Peterek N. Picone J. Wang W. Beyerlein P., Byrne W. 1999. Towards language independant acoustic modeling. *IEEE workshop on automatic speech recognition and understanding*.

Jonathan G. Fiscus and George R. Doddington. 2002. Topic detection and tracking evaluation overview. *Topic detection and tracking: event-based information organization*, pages 17–31.

Rayid Ghani, Rosie Jones, and Dunja Mladenic. 2000. In *Mining the web to Create Minority Language Corpora*, Berlin.

Larry Hyman. 1981. Tonal accent in somali. *Studies in African linguistics*, (12):169–203.

David Le-Gac. 2001. Structure prosodique de la focalisation: cas du somali et du français.

A. Nimaan, P. Nocera, and J.M Torres-Moreno. 2006. Boîte à outils tal pour des langues peu informatisées : le cas du somali. In *JADT 2006 Journées d'Analyses des Données Textuelles*, Besançon, FRANCE.

P. Nocera, G. Linares, D. Massonie, and L. Lefort. 2002. Brno. In *Phoneme lattice based A* search algorithm for speech recognition*, TSD2002.

R. Rosenfeld. 1995. The cmu statistical language modeling toolkit, and its use. In *ARPA Spoken Language Technology Workshop*, Austin, TEXAS, USA.

John Saeed. 1993. *Somali reference grammar*. Dunwoody Press, MD.

John Saeed. 1999. *Somali (London Oriental and African Language 10)*. Johns Benjamins Publishing Company, Amsterdam/Philadelphia.

International SIL. 2004. *Ethnologue : Language of the World. 14th edition*. USA.

Unesco. 2003. Convention pour la sauvegarde du patrimoine culturel immateriel. *http://www.unesco.org/*.

D. Vaufreydaz, M. Akbar, and J. Roullard. 1999. Asru'99. In *Internet documents: a rich source for spoken language modelling*, pages pp. 177 – 280, Keystone Colorado (USA). Workshop.

[8] http://www.cerd.dj
[9] http://www.ambafrance-dj.org/
[10] http://www.lia.univ-avignon.fr

Developing an annotated corpus for Gĩkũyũ using language-independent machine learning techniques

Peter W. Wagacha*, Guy De Pauw† and Katherine W. Getao*

*School of Computing & Informatics
University of Nairobi - Box 30197-00100, Nairobi, Kenya
{waiganjo,kgetao}@uonbi.ac.ke

†CNTS - Language Technology Group
University of Antwerp - Universiteitsplein 1, 2610 Antwerpen, Belgium
guy.depauw@ua.ac.be

Abstract

Networking the development of computational resources for African languages can be greatly advanced if researchers aim to develop tools that are to a large extent language-independent and therefore reusable for other languages. In this paper we describe a particular case study, namely the development of an annotated corpus of Gĩkũyũ, using language-independent machine learning techniques. The general aim of our work on Gĩkũyũ is two-fold: on the one hand we wish to digitally preserve this resource-scarce language, while on the other hand it serves as a feasibility study of using language-independent machine learning techniques for linguistic annotation of corpora. To this end we investigate established annotation induction techniques like unsupervised learning and knowledge transfer. These methods can provide interesting perspectives for the linguistic description of many other resource-scarce languages.

1. Introduction

Many languages in Africa are resource-scarce. This means that their computerization in this digital world is nearly impossible for the moment. The languages on this continent are as diverse as its people. Many researchers are addressing this though at various levels and stages. Networking the disparate nodes of research on the continent working on African languages is therefore crucial. Through a network, the distributed knowledge and expertise, tools, corpus building strategies can be shared. This unique situation requires that researchers should aim to develop tools that are to a large extent language-independent and therefore reusable. Such a network of researchers should be instrumental in enhancing inter-disciplinary collaboration and synergy between linguists and computer scientists, who traditionally do not always collaborate.

The general aim of our work on Gĩkũyũ is two-fold: on the one hand we simply wish to preserve this resource-scarce language, which is decreasingly being used in verbal and written communication. An annotated corpus for this language can consequently serve to stimulate interest in the language from researchers from linguistics as well as computer science. On the other hand the development of the Gĩkũyũ corpus acts a case study that researches the feasibility of using the aforementioned language-independent machine learning techniques for linguistic processing of text corpora.

To this end we investigate established annotation induction techniques like unsupervised learning and the relatively novel approach of knowledge transfer. These methods aim to minimize human annotator cost and language expert knowledge during corpus construction and can provide interesting perspectives for the linguistic description of many other RSLs. Through the use of these techniques the process of developing an annotated corpus can be significantly expedited.

2. The Gĩkũyũ language

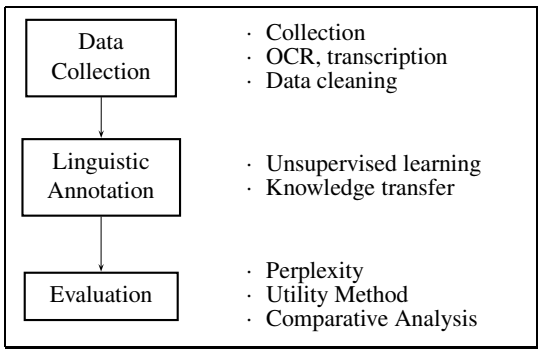
Gĩkũyũ is a language spoken by the Gĩkũyũ people (often written as Kĩkũyũ) who are approximated at 6 million people. Most of these live in the central part of Kenya in East Africa. It is a Bantu language and has been classified as E51 (Guthrie, 1967). The Bantu linguistic group covers the area from South Cameroon to almost the whole of Southern Africa, including Eastern and Central Africa. Gĩkũyũ belongs to the Kamba-Kikuyu subgroup of Bantu. There are six dialects, namely Southern Gĩkũyũ (Kiambu, Southern Murang'a), Ndia (Southern Kirinyaga), Gichugu (Northern Kirinyaga), Mathira (Karatina), and Northern Gĩkũyũ (Northern Murang'a, Nyeri). The Gĩkũyũ language is lexically similar to some closely related languages such as Chuka, Embu, Kamba and Meru.

2.1. Socio-linguistic challenges of Gĩkũyũ

Gĩkũyũ nowadays is more spoken than read. The primary reason for this is the fact that Kenya has two official languages, English and Kiswahili, which have become the default languages of communication both written and spoken. In school, English is the language of instruction. This is indeed a good thing in a country that has about 40 local languages. These two languages unify all and serve as a standard communication medium.

A few decades back, basic primary education was taught in the mother tongue in the rural areas. This is not the case today. Typically for the younger generation, one speaks his/her mother tongue at home (or Kiswahili), Kiswahili to the general public and English in school. This has resulted in a generation that is neither fluent in any language. Moreover, there is an emergent language, Sheng - initially a distortion of Kiswahili and English, but now a murky concoction of languages. It is evident that good native speakers of the mother tongue are lacking, and there is a marked decrease in young language users. There is nevertheless a

Figure 1: Corpus-Building Strategy



great interest in local languages within the country as has been clearly noted from the high number of vernacular radio broadcasts and some vernacular periodic publications. This work is motivated by the need to develop resources and tools that can be used in the language's computerization, as well as its preservation.

It is a fact that language, culture and knowledge are entwined. African languages contain a significant part of the world's historical, cultural, social, botanical, ... knowledge and wisdom, which needs to be preserved. Unlike in South Africa where there is in place a language policy that enshrines local languages in the constitution, this is not the case in Kenya, for its 40 odd languages, indicating a possible lack of political will to preserve local languages.

2.2. Gĩkũyũ as a resource-scarce language

We have adopted the following definition for a resource-scarce language (RSL): *a language for which few digital resources exist; a language with limited financial, political, and legal resources; and a language with very few linguistics experts.* A language that has very few digital resources, poses unique challenges with regard to computerization. With these resources, it would be easy to develop technologies and tools to support a language such as machine translation, speech recognition and text-to-speech systems.

Limited political and financial resources is another aspect that surrounds many local languages. This largely contributes to the lack of significant quantities of language resources. To our knowledge, there are very few trained linguists for many of our local languages. It is interesting to note that, apart from Mugane (1997), there seems to be no other recent and easily available work describing this language. Most of the literature about Gĩkũyũ was developed between 1900 and 1960. This may indicate a heightened lack of interest in Gĩkũyũ within linguistics circles. This is puzzling, since the number of speakers of this language is still relatively high. Academic work on the languages is also not readily available if it exists. Moreover, over time a language changes, necessitating further research. This is indeed the case for Gĩkũyũ.

3. Corpus Construction

We define three phases during the construction of our annotated Gĩkũyũ corpus, illustrated in Figure 1. First of all we have a data collection and clean-up phase, during which we gather as much raw data as possible. Next, we annotate the collected data, where we initially concentrate on providing part-of-speech tag information for the words in the corpus. Finally, we evaluate the corpus using quantitative and qualitative measures.

3.1. Data collection and clean-up

For Gĩkũyũ today, there are few resources available. As in many African languages, the Holy Bible is the default language reference for Gĩkũyũ. Most available material is in print media, such as religious literature, hymn books, poems, basic language learning aides, short stories, novels by renown authors e.g. Ngũgũ wa Thiong'o, etc. There is hardly any ready material in electronic form.

In our data collection phase, we located available source documents, which were run through a scanner and later an optical character recognition (OCR) system. A small collection of the collected data was manually transcribed. The source documents included: religious texts, simple readers, songs, poetry, transcribed radio programs, books, newspapers and magazines. We were also able to find a few web pages.

Gĩkũyũ incorporates diacritics (or accents), to provide two extra vowels (ĩ and ũ). In a typical Gĩkũyũ text more than 50% of words require at least one of the two diacritics. The fact that these diacritics are not readily available on the computer keyboard means that words are typically transcribed and OCRed without the correct diacritic, resulting in 'i' representing ĩ and 'u' representing ũ. Even in the print medium, there is some data that that does not have the diacritics. We therefore developed a language-independent diacritic restoration tool based on machine learning techniques (Wagacha et al., 2006). This tool was able to provide the correct diacritics for more than 90% of the words in a text.

Another technique that can be used in the collection of data is the automatic collation of web pages from the internet. An example of one such application has been used for Kiswahili (Geato and Miriti, 2005). The unavailability of diacritics on the computer keyboard is responsible for the lack of web content in Gĩkũyũ language. The above two examples illustrate that even for the data collection phase, suitable tools need to be developed.

3.2. Corpus Annotation using language-independent induction techniques

Linguistic annotation of large text corpora is nowadays most often done semi-automatically. Computational annotation systems generate a first (automatic) classification, which is manually corrected by human annotators. Not only does this approach increase the speed with which these corpora can be annotated, it also improves the consistency of the linguistic description. A typical first annotation task in corpus construction is part-of-speech (POS) tagging, as this is a primordial component on the one hand for linguistic description, on the other hand as a first processing step for almost all language technology applications. We therefore chose to concentrate our research on this particular annotation task.

The current state-of-the art methods for POS tagging are trained on annotated corpora (De Pauw et al., 2006 submitted). If there is no previously annotated data available however, which is by definition the case for RSLs, these traditional algorithms for POS tagging are largely useless. Rather than go through a costly human annotator phase, we opt to investigate two alternative techniques that can provide an initial annotation of the Gĩkũyũ corpus: (1) "unsupervised learning" and (2) "knowledge transfer".

In **unsupervised learning** classification is performed on the basis of free text. The standard technique is that of conceptual clustering, which describes the linguistic objects as data points in a feature space, which can be subdivided in a number of clusters. These clusters can then a posteriori be interpreted as placeholders for linguistic classes (e.g. parts of speech). Experiments for other languages have shown that this technique is very suitable as a first initialization of POS tag annotation (Schütze, 1993; Yarowsky, 1995). Furthermore, it can also provide interesting novel insights from a (psycho)linguistic point of view as it approaches the data without preconceived linguistic notions (Clark, 2001).

Knowledge transfer techniques try to apply the annotation properties of an adequately described language (= source language, e.g. Kiswahili) to a language for which there is no annotated data available (= target language, e.g. Gĩkũyũ). Knowledge transfer makes use of parallel corpora (the same text in two different languages) which have been aligned on the sentence and word level. The direct correspondence assumption (Hwa et al., 2002) consequently allows for the annotation of the words in the source language to be projected unto the text in the target language. Even though lexical and structural differences between languages prevent a simple one-to-one mapping, knowledge transfer is often able to generate a well directed initial annotation of the target language with a minimal amount of resources (Cucerzan and Yarowsky, 2002). The knowledge transfer technique could be particularly fruitful for transfer of linguistic annotations among Bantu languages since they are known to have grammatical similarities (Nurse and Philippson, 2003).

Unsupervised learning techniques can be directly applied to the existing Gĩkũyũ corpus. For the knowledge transfer methods, we need a parallel corpus. For this purpose we can use the bible and quran, available in both source and target language. The POS tag annotation of the source language can consequently be induced from the annotated Helsinki Corpus of Swahili (Hurskainen, 2004; De Pauw et al., 2006 submitted) and transferred to the target language.

Although unsupervised learning and knowledge transfer methods typically serve the same purpose, the literature hardly describes comparative efforts. Furthermore, they are not often applied as tools in the linguistic description of RSLs, although they are often touted as providing interesting perspectives for them. During the development of the annotated Gĩkũyũ corpus, we aim to evaluate and possibly combine these two techniques. This will give us an idea of their feasibility as language-independent annotation induction techniques. In the next section, we discuss how this feasibility can further be gauged in an extensive evaluation phase.

3.3. Evaluation of the annotated corpus

As noted before, the methods described in the previous section only provide a first initialization of annotation, not unlike established supervised methods would during semi-automatic corpus construction. This means that there is still a certain amount of postprocessing needed by human annotators. Evaluation of the annotation induction methods can be described in terms of how well they are able to minimize this human postprocessing effort.

Again, the evaluation is met with specific challenges when dealing with RSLs. They may have no prior formal linguistic abstractions for parts of speech or faulty or outdated linguistic abstractions for parts of speech. These issues may prevent standard evaluation against an existing gold-standard. Evaluation and postprocessing must therefore be performed manually, preferably by experts of the language. It may however also be challenging to identify and recruit human annotators for RSLs because:

- Typically, the general literacy level is low (few people may be competent in reading and writing the language).
- The language may not be standardized and thus may be represented by several dialects.
- There is a lack of formal linguistic expertise in the language.
- Computer literacy may be low so that capacity building would be needed before human annotators could make full use of automated tools.
- Linguistic expertise in the language is more likely to reside in places where the cost of labor is high.

One way to circumvent this problem is by performing a purely quantitative evaluation on the basis of perplexity measures. These are commonly used in language modeling to evaluate the generalization properties of an annotation system for natural language (Charniak, 1993). While this can provide a solid quantitative estimation of the adequacy of the induced annotation, it still does not guarantee a well annotated corpus, nor does it solve remaining annotation errors. We therefore propose two additional evaluation/postprocessing methods that are intended to minimize human intervention as much as possible.

The first method, the **utility method**, involves the annotation of a small test set using the discovered (or transferred) tags. This test set should also have a manual annotation. The two annotations can then be compared using normal tag evaluation methods. This method requires much less human intervention since the discovered tags do not have to be named. Instead quantitative measures of tag correspondence on the corpus annotation task can be used to match tags (for example how often is a manually-annotated tag, VERB, matched with an automatically induced tag, C1, when using the same test set). Different manually-produced tag sets can be compared with the automatically-produced tag sets to reduce the problem of bias (where an automatically-produced tag set is poorly scored because of a particular linguistic choice of tags.) This evaluation method may be a rich area for research study since learning algorithms, RSLs and tag choice methodologies can all be

research parameters. It is hoped that this research direction will produce some results that can be generalized.

The second method, the **comparative analysis**, compares a manually produced tag set for an RSL with a tag set produced through unsupervised learning or knowledge transfer. This could be done by human intervention, requiring the diagnosis and naming of the discovered tag set and the manual production of an optimized tag set so that the two can be compared. Both quantitative (such as the percentage of optimized tags discovered) and qualitative (such as the 'meaningfulness' of the tags) metrics would apply in this case. One method of comparative analysis uses tag instance lists. An example tag instance list for is: instance-PRONOUN = (he, she, it, I, ?). The tag instance lists produced from manually-annotated corpora can be compared with tag instance lists generated during the course of unsupervised learning. Further drill down analysis can be performed by comparing the automatically induced instance-context lists of all instances of a particular tag for correspondence with the context (bigram or trigram) present in corpus data.

An advantage of the proposed evaluation approach is that it can also be implemented using an exemplar database instead of a hand-annotated corpus. An exemplar database consists of a list of POS tags, each associated with a list of examples. It may require less expertise to manually generate such an exemplar database than it would to manually annotate a sizable corpus. The interaction between the manual and automatic annotation process can be iterative, with each iteration leading to a refinement of the exemplar database.

4. Conclusion

We defined a resource-scarce language as a language for which few digital resources exist; a language with limited financial, political, and legal resources; and a language with very few linguistics experts. We have used Gĩkũyũ as an example of such a language. Since this language has some very closely related languages, we believe that the knowledge, tools and expertise can be easily 'transfered' to these languages and indeed other Bantu languages.

We have proposed the use of state-of-the art unsupervised and knowledge transfer techniques. The proposed evaluation techniques, the utility and comparative analysis methods, combined with an iterative approach of developing an exemplar database, are also designed to automate linguistic processing and minimize expert human intervention.

In summary, the corpus-building strategy that we propose combines data collection, with automated linguistic annotation and semi-automated evaluation of annotated corpora to achieve the following:

- Minimize the use of scarce human resources;
- Maximize the potential of limited linguistic data;
- Develop techniques that speed RSL linguistic annotation by: (a) using learning techniques that do not require a large amount of prior information; and, (b) facilitating transfer of knowledge between already-annotated languages and similar languages that have not yet been well annotated.

If these aims are achieved within this corpus building strategy, it will go a long way towards facilitating the entry of RSLs into the digital world, thus assisting in the preservation of culture, experience and knowledge that is embodied in these languages.

Acknowledgments

The research presented in this paper was made possible through the support of the Flemish Inter-university Council (VLIR), under the VLIR-IUC-UON program.

5. References

E. Charniak. 1993. *Statistical Language Learning*. The MIT Press, Cambridge, MA.

A. Clark. 2001. *Unsupervised Language Acquisition: Theory and Practice*. Ph.D. thesis, COGS, University of Sussex.

S. Cucerzan and D. Yarowsky. 2002. Bootstrapping a multilingual part-of-speech tagger in one person-day. In *Proceedings of CoNLL-2002*, pages 132–138, Taipei, Taiwan.

G. De Pauw, G. M. de Schryver, and P. W. Wagacha. 2006 (submitted). Data-driven part-of-speech tagging of Kiswahili. In *Proceedings of the Ninth International Conference on TEXT, SPEECH and DIALOGUE*.

K. Geato and E. Miriti. 2005. Process for building a Kiswahili corpus from the world wide web. In *Proceedings of the 1st Annual International Conference and Workshop on Sustainable ICT capacity in developing countries 2005.*, pages 148–152, Makerere University, Kampala.

M. Guthrie. 1967. *The Classification of Bantu Languages*. Dawsons of Pall Mall, London.

A. Hurskainen. 2004. *HCS 2004 - Helsinki Corpus of Swahili*. Compilers: Institute for Asian and African Studies (University of Helsinki) and CSC - Scientific Computing.

R. Hwa, Ph. Resnik, A. Weinberg, and O. Kolak. 2002. Evaluating translational correspondence using annotation projection. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 392–399, Philadelphia, PA, USA.

J. M. Mugane. 1997. *A Paradigmatic Grammar of Gĩkũyũ*. CLSI Publications, Stanford California.

D. Nurse and G. Philippson. 2003. *The Bantu Languages*. Routledge Language Family Series. Routledge, London, UK.

H. Schütze. 1993. Part-of-speech induction from scratch. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 251–258, Columbus, OH, USA.

P. W. Wagacha, G. De Pauw, and P. W. Githinji. 2006. A grapheme-based approach for accent restoration in Gĩkũyũ. In *Proceedings of fifth international conference on Language Resources and Evaluation, LREC 2006*.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, MA, USA.

Saving an African Language: Computer Assisted Lexicon and Grammar for Translation Purposes

Alex Kasonde

University of Zambia
PO Box 32379, Lusaka, Zambia
E-mail: mkasond@yahoo.com

Abstract

The paper presents on-going computer research in saving an African language from environment-based degradation and the threat of extinction. The focus of discussion is Zambia, a landlocked country situated in Southern Africa bordering Angola, Botswana Congo DR, Malawi, Mozambique Namibia, Tanzania and Zimbabwe. In terms of language policy, Zambia is like the quasi-totality of former European colonies in Sub-Saharan Africa reluctantly improving its linguistic human rights record. From the era where the former colonial language (e.g., English, French, Portuguese, Spanish) was coercively regarded as superior and the African language (e.g., Amharic, Hausa, Swahili, Twi, Wolof) as inferior the African continent has entered a new epoch of linguistic equality and justice. Despite the newly acquired freedom linguistic projects remain largely neglected. The project discussed involves all the 7 national languages of Zambia, namely Bemba, Nyanja, Tonga, Lozi, Kaonde, Luvale and Lunda (BNTLKLL). The aim of the project is to create Computer Lexicon and Grammar for general reference and translation purposes. With the prospect of internet access the complete BNTLKLL online software resources developed will be made available to end users from around the world as either an independent website or hosting. Considering that BNTLKLL belongs to the Bantu Language Group (Guthrie), the degree of deviation from correct literal rendition during translation of particular words or sentences should be low and manageable.

1 Introduction

The research project discussed in the present paper is a culmination of earlier attempts to develop a versatile lexicon of 7 national languages of Zambia. These in order of statistical superiority are Bemba, Nyanja, Tonga, Lozi, Kaonde, Luvale and Lunda, hereafter simply referred to as BNTLKLL. The aim of the project is the creation of an original research tool available to the public in the form of BNTLKLL Computer Assisted Translation. The project requires a data base comprising the BNTLKLL Lexicon, BNTLKLL Grammar and BNTLKLL Translation System of the written text patented and marketed like any electronic calculator. The primary sources of data for the project are existing written texts and transcribed oral texts derived from oral literature. The written texts include non fiction (dictionaries, histories, ethnographies, newspapers, magazines, etc) and fiction (folktales, novels, plays). Oral sources are Audio-Video Recordings of stories, tales, myths, legends and related literature. The different stages of the project will be managed through consultations during Research Project Development Workshops. Results and research findings of the project will be disseminated in the form of hard and soft copies with an On Line Version bearing interactive capabilities. Current estimates show that the entire project will extend over one calendar year of thousands of pages.

Regarding attempts to build a Linguistic Data Base of official Zambian languages with lexicon, grammar and translation capability, the first notable achievement was a paper presentation in the early 80s under the University of Zambia Department of Literature and Languages Seminar Series chaired by Professor Theresa Chisanga. The paper Lexicography in Zambia took a close look at existing literature and identified certain challenges, including funding and sustainability. The second major initiative was hatched in the early 90s at the University of Zambia, Institute of African Studies (IAS) when the author returned from graduate studies at Sorbonne University Institute of Applied and Theoretical Phonetics and Linguistics (ILPGA).

The second phase marked the first direct contact with various universities abroad. The response from Hawaii University with respect to computer software adapted to multi-lingual lexicons, grammar and automatic translation deserves special mention. The third initiative was a paper presentation in the late 90 at the University of Frankfurt based on the work of Professor Malcolm Guthrie of School of Oriental and African Studies called Comparative Bantu (CB). The paper explored the possibility of utilizing IT to reduce, store, analyze and present CB for teaching and research. The paper among other things opened for the author doors to the

European Association of Lexicology (Euralex) jointly managed from University of Stuttgart (FRG) and Oxford University (UK). The last major development was a stay at Emory University Department of Anthropology as Visiting International Scholar working on a Corpus Linguistics project in collaboration with Professor Debra Spitulnik also of Emory University. The collaborative project from Emory University involving Dr. Spitulnik (Emory University), Dr. Fenson Mwape (Osaka University) and the author (University of Hamburg) also received an overwhelming response during the inaugural meeting of the Open Language Consortium (OLAC) held at University of Pennsylvania under the joint-chairmanship of Professor Steven Bird, formerly of University of Pennsylvania.

1.1. Zambian Society

The Zambian society is characterized by a multi-ethnic complex where multi-lingualism permeates all facets of social life. It is estimated that 72 different ethnic groups can lay a rightful claim to indigenous status in the country. Each ethnic group is generally located in a well-defined territory while individuals remain free to travel and live in the country and abroad on the basis of one common nationality. In addition to their ethnic identities that are generally addressed by the existence of different Local Government regulations Zambians also belong to a higher identity that is regulated by Central Government. The same polarity between central and local authority of government finds expression in the duality of legal systems divided into traditional Customary Law and received statutory law based on English Law.

So far as government function is concerned Zambia is a Republican democracy with an executive president, vice president, cabinet ministers and deputy ministers. Various government departments are headed by permanent secretaries and certain other government officials. The judicature is headed by the Chief Justice and the legislature is headed by the Speaker of the National Assembly.

Since the establishment of the British protectorate of Northern Rhodesia around 1890 by the British South Africa Company (BSA) following the amalgamation of North Eastern Rhodesia and North Western Rhodesia the same territory remains largely geographically unchanged. After independence in 1964 the name of Zambia was adopted and certain development projects were initiated. In the social sector more educational and health facilities were established. The same efforts were also applied to the political and economic arena. Despite the determination and motivation of the Founding Fathers contained in the various National development Plans it is now generally agreed that successive Zambian governments have failed to build a modern nation state where abundance and prosperity could be self-evident. Unemployment is high and poverty, squalor and illiteracy remain disproportionate to the potential of the human resource and natural resource base. Not surprisingly the Zambian government is at the for-front of various economic and political projects in the Southern African region, including the Front Line States (FLS), Southern African Development Community (SADC), and Common Market for Eastern and Southern Africa (COMESA). The recently created New Economic Partnership for African Development (NEPAD) and the African Union (AU) also received tremendous support from the Zambian government.

1.1.1. Socio-linguistic Situation

From a numerical point of view Bemba currently spoken by over 55% of the population represents the most important African language in the country. This figure corresponds to approximately 5,000,000 speakers, based on recent Census of Population in the country. Out of the 9 provinces of Zambia (Northern, Southern, Eastern, Western, Central, Lusaka, Copper belt, North Western, Luapula), Bemba is also spoken by the majority in more provinces (Northern P, Luapula P, Central P, and Copper Belt P) than any other African language in the country. In terms of administrative functions English is recognize by the existing Constitution as the official language of communication. For legal purposes an individual who does not understand English is entitled to free services provided by interpreters and translators as a matter of human rights obligations. The only problem with the law is that the capacity of the state to maintain law and order is limited by certain challenges, including ignorance, illiteracy and human greed.

The multi-lingual character of Zambian society is both individual and social determined. At the individual level the national elite is obliged to master the official language English in addition to the respective ethnic language. The bottom line could be English-African language bilingualism but in most cases travel, schooling and work mean that a third language is

oftentimes acquired naturally. The third language can also be French in view of the fact that French is the only modern European language available in the curriculum of government funded secondary schools, colleges and universities. The support that the Zambian Government offers to French in the educational system has created a certain amount of friction between conservatives and liberals. Conservatives argue that the promotion of French can lead to social problems in relation to existing norms and standards based Anglo-Saxon practices. Liberals feel that the promotion of French is good for integration in the global arena. Suggestions to include a French clause in the Constitution have been unsuccessful. Despite the constitutional impasse French remains popular in Zambian private and public schools. Certain African intellectuals would like to raise concerns regarding need for introduction of different world languages in public life. The United Nations languages (Arabic, Chinese, English, French, Russian, and Spanish) and other world languages such as Dutch, German, Japanese, Turkish and Swahili are also important vehicles of culture and international relations.

Endangered languages of Africa are strictly speaking the small unwritten ethnic languages. Viewed from a broad perspective endangered languages in Africa should include the national languages considering that African national languages are neglected by African governments. African governments generally promote former colonial languages to avoid national disintegration and linguistic conflict. At the same time African governments rightly regard the use of European languages as the gateway to universal civilization. The major problem is that language policies and legislation in place in the vast majority of African countries ignore the experience of peoples of Asia. In countries of South Asia (e.g., Pakistan) and South East Asia (e.g., Indonesia, North and South Korea, Singapore) the use of a local idiom is generally understood as a tool of national development. The language policies of certain African countries such as Botswana and Tanzania indicate that the promotion of African languages through appropriate legislation and media practices could play critical roles in Good Governance and concomitant Macro Economic Stability.

2 References

Givon, T. (1969) Studies in Chibemba and Bantu Grammar. PhD. Thesis, University of California Los Angeles, 263p.

Guthrie, M. (1967) Comparative Bantu. 4 vols.

Kashoki, M. & Sirarpi, O. (ends) (1978) Language in Zambia. London: International Africa Institute, 461p.

Kasonde, A. (2005) Teaching Languages of the World: Did the Summer Institute of Linguistics Make a Big Difference? Paper presented at University of Zambia Linguistics Association Seminar, unpubl.

Kasonde, A. (2000a) How Culture, Experience and Input Shape and Influence Language Use: The Case of Time in Selected African Languages. Paper presented at Emory University, Department of Psychology, and Atlanta GA, unpubl.

Kasonde, A. (2000b) A Classified Vocabulary of icibemba Language. Munich: LINCOM Europa Publishers.

Kasonde, A. (2000c) Language Law and Development in the Third World countries (South Korea and Zambia). Hamburg, London, etc.: LIT Verlag Publishers.

Kasonde, A. (1997) The Use of Computers in Teaching Historical and Comparative Linguistics: The Case of Comparative Bantu. Paper presented at International Conference on the Use of Computers in Historical and Comparative linguistics, Johannes Wolfgang von Goethe University-Frankfurt, Frankfurt A/M, unpubl.

Kasonde, A. (1994) Van Sambeek Revisited: A Historical Account of the Origins of Bemba Grammatical Descriptions. Paper presented at First World Conference of African Linguistics, University of Witwatersrand (Johannesburg, South Africa) and University of Swaziland (Kwaluseni), unpeel.

Kasonde, A. (1981) Lexicography in Zambia. Paper presented at University of Zambia Department of Literature and Languages Seminar, unpubl.

Sharman, J.C. (1963) The Tabulation of Tenses in a Bantu Language (Bemba: Northern Rhodesia), PhD. Thesis, University of South Africa

Sharman, J.C. & Misuse, A.E. (1956) The Representation of Structural Tones, with Special Reference to the Tonal Behaviour of the Verb in Bemba, Northern Rhodesia. Oxford: Oxford University Press.

Van Sambeek, J (1955) A Bemba Grammar, as illustrated by W.A.R. Gorman and amended by Publications Bureau staff. London: Longman, 117p.

White Fathers (1947) Bemba-English dictionary. Chilubula Mission, Northern Rhodesia, 1505p

Share and Share Alike – Developing Language Resources for African Language Translators

Rachélle Gauton

University of Pretoria
Department of African Languages, University of Pretoria, Pretoria, 0002, South Africa
rachelle.gauton@up.ac.za

Abstract

The biggest challenge facing translators working into the previously marginalised and disadvantaged South African Bantu languages, is the lack of (standardised) terminology in most specialist subject fields. The ever increasing demand for translation and localisation into these languages far outstrips the ability of the various national language bodies that are tasked with *inter alia* standardisation and terminology development, to provide translators with the requisite (standardised) translation equivalents for English source language terms. In reality, South African translators working from an international language such as English into languages of so called 'limited diffusion' such as the SA Bantu languages, routinely have to create their own terminology without recourse to standard sources on the language such as (technical) dictionaries, and in the absence of clear guidelines as to which term formation strategies are available to them and/or which strategies are the preferred translation strategies in these languages. Such language resources being created daily mostly go to waste (so to speak) as SA translators are not generally in the habit of documenting their terminology for future reuse by either themselves or others. In this paper, I will indicate:

- which types of language resources the SA Bantu language translator needs to be able to overcome, and/or compensate for, the lack of (standardised) terminology in these languages;
- why there is such a pressing need for SA translators to pool and share language resources for translation purposes, and
- how a start can be made at an institution such as the University of Pretoria to develop such language resources.

1. Introduction

The biggest challenge facing translators working into the previously marginalised and disadvantaged South African Bantu languages, is the lack of (standardised) terminology in most specialist subject fields. The ever increasing demand for translation and localisation into these languages far outstrips the ability of the various national language bodies that are tasked with *inter alia* standardisation and terminology development, to provide translators with the requisite (standardised) translation equivalents for English source language (SL) terms.

In reality, South African translators working from an international language such as English into languages of so called 'limited diffusion' such as the SA Bantu languages, routinely have to create their own terminology without recourse to standard sources on the language such as (technical) dictionaries, and in the absence of clear guidelines as to which term formation strategies are available to them and/or which strategies are the preferred translation strategies in these languages.

It is furthermore no secret that terminology development and elaboration of the SA Bantu languages as technical languages are taking place at grassroots level in the studies and workplaces of numerous translators spread all over the country, not forgetting the trainee translators at various higher education institutions completing translation tasks as part of their training. As Kruger (2004:2) points out:

> In South Africa, translation and interpreting are the main areas in which the technical registers of African languages and Afrikaans are being developed and standardised […]

These language resources being created daily mostly go to waste (so to speak) as SA translators are not generally in the habit of documenting their terminology for future reuse by either themselves or others.

This is most probably because there is no established tradition in South Africa yet of translators making use of electronic translation resources and translators' tools such as, for example, translation memory tools and terminology management tools.

Furthermore, with a few notable exceptions such as the translator training courses at the University of Pretoria, South African (higher) education institutions do not as a matter of course train these translators in the application of Human Language Technology (HLT) in translation practice, and specifically with reference to translating into the SA Bantu languages.

In this paper, I will indicate:

- which types of language resources the SA Bantu language translator needs to be able to overcome, and/or compensate for, the lack of (standardised) terminology in these languages;
- why there is such a pressing need for SA translators to pool and share language resources for translation purposes, and
- how a start can be made at an institution such as the University of Pretoria to develop such language resources.

2. Language Resources Needed by the African Language Translator in the South African Context

When translators find themselves in a position where they are unable to access the requisite translation equivalents in any standard sources on the language concerned, they can turn to parallel texts. These are texts that have already been translated into (or have originally been written in) the target language and that deal with the same subject field as the source text that needs to be translated. Such parallel texts can then be mined for possible translation equivalents. With the increase in such texts that are currently available on the Internet, in all of the official SA languages, the translator (as well as translator trainers), have a ready resource that can be mined with the assistance of corpus query software such as *WordSmith Tools* (Scott, 1999) and *ParaConc* (Barlow, 2003). (See in this regard Gauton & De Schryver, 2004).

There are also an increasing number of sites available on the web that contain parallel corpora in the official SA languages, i.e. sets of texts in a source language (usually English) together with their translations in one or more other languages[1]. Pearson (1998:47) defines parallel corpora as:

> [...] sets of translationally equivalent texts, in which generally one text is the source text and the other(s) are translations.

Such parallel corpora can be used by the Bantu language translator in the following ways:

- By studying parallel corpora, the translator (or the trainee translator in a translator training programme) can gain an insight into the translation strategies utilised by a large number of (professional) translators in creating translation equivalents for source language terms that are not lexicalized (in a standard form) in the target language. By studying translators' behaviour in this manner, the translator / trainee translator will also be able to deduce what the preferred translation strategies are in the SA Bantu languages. (See in this regard Gauton et al., 2003; Gauton et al., forthcoming; and Mabasa, 2005)
- Parallel corpora (either comprising a translator's own translation work or culled from the Internet) can be used as input for translation memory (TM) tools that can be utilised by the translator to work faster, more accurately (particularly concerning consistency in terminology use) and more efficiently. A TM tool serves as a 'store' of previously translated sentences which 'reminds' the translator of what she has previously translated. This type of tool works best when the text to be translated contains many repetitive sentences/repetitions, as would normally be the case with most types of technical translation tasks. Translation memory (TM) tools also usually contain a terminology management (TM) tool. (Cf. Esselink, 2003). Some of these tools furthermore enable the translator to extract a glossary / term list of the source and target language terminology used in a specific translation project. Such glossaries are ideally suited for re-use either by the translator herself and/or for sharing with other translators.

3. The Need for Language Resources for the African Language Translator in the South African Context

One only has to study existing translations (particularly in technical domains) to realise what the consequences are of SA Bantu language translators not having access to the language resources as outlined in this paper.

Having to work in an environment where one cannot rely on the availability of standardised terminology, creates various problems for these translators. A case in point is the Zulu translations of:

(a) the user interface (UI) of the Microsoft operating system Windows XP for which I was project leader; language manager responsible for creating and maintaining the language style guide and managing the terminology to ensure consistency; as well as quality controller of specifically the grammatical and linguistic correctness of the translations (cf. Gauton, forthcoming); and

(b) the Sony Ericsson user guides for the T310, T610 and Z520i mobile phones.

A very high level of inconsistency in terminology used for the same English SL concepts can be found not only between these translations, but also within the same translation, and in the case of the Windows XP translation, within the work of the same translator.

This type of situation invariably results when there is more than one translator working on a specific project, with each translator working in isolation (so to speak); i.e. working without the use of computer assisted translation (CAT) tools such as translation memory (TM) tools, terminology tools and software localisation tools, coupled with a pooling and sharing of resources within the project.

However, as indicated, even when only one translator is involved and doing large amounts of translation (which is usually the case in localisation projects such as these) without the use of a translation memory tool; this usually results in terminological inconsistencies within such a translator's work.

See the following representative examples (culled from the Zulu translations mentioned earlier) illustrating this point:

[1] See De Schryver (2002) for a discussion on African language parallel texts available on the web. Note, however, that since the publication of this 2002 article, many more parallel texts in the official SA languages have become available on the web.

Source Language (English) Term	Translation equivalents culled from the Zulu translations of the Windows XP UI (before consistency checking) and the published Sony Ericsson user guides
network (*n*)	umphambo; inethiwekhi; ukuxhumana; uxhumano
e-mail (*n*)	i-imeyli; i-e-mail; i-emeyili; iposi le-elekthroniki
settings	izinhlelo; okokuhlela; ukuhlelwa; uhlelo; izimiso
set (*v*)	hlela; misa; setha
shortcut	indlela enqamulayo; unqamulelo; ukunqamulela; indlela emfushane; ishothikhathi; ushothikhathi
phonebook	ibhuku locingo; ibhuku lezingcingo; incwadi yezingcingo; ibhuku lefoni; ifonibhuku
memory	inkumbulo; imemori; isiqophi
password	igama lokungena; igama lokuvunyelwa ukudlula; igama lokudlula; iphasiwedi
wizard	umeluleki; umbuzimholi; iseluleki; iwizadi; umthakathi

Table 1: Inconsistent use of terminology in the Zulu translations of the Windows XP UI (before consistency checking) and the published Sony Ericsson user guides

As can be seen from Table 1 above, the level of inconsistency regarding the translation equivalents used for the SL terminology is unacceptably high.

Furthermore, by not being aware of terminology that has perhaps already been coined by other translators, and/or by not getting the benefit of other translators' experience and insight, the unwary translator may end up coining culturally unacceptable target language equivalents such as the term **umthakathi** 'witch' for the source language term 'wizard'. As pointed out elsewhere (Gauton, forthcoming) the term **umthakathi** has extremely negative connotations in the Zulu culture, and is not a suitable translation equivalent for the English concept 'wizard'; particularly as used in the domain of computer studies. Whereas in the Western context a wizard is a wise and magical imaginary fairytale character that is often benign (unless of course he is characterised as an 'evil wizard'), quite the opposite applies in Zulu culture. Within Zulu culture, witches and wizards are evil creatures who practice witchcraft, are intent on doing only harm to their fellow man, and who are shunned and avoided by all wherever possible. In recent years in South Africa, a significant number of people have been persecuted, ostracised and even killed on suspicion of being witches or wizards practising witchcraft. Thus it is clearly inappropriate and culturally unacceptable to use the term **umthakathi** 'witch' to signify an interactive computer programme (i.e. a 'wizard') that fulfils the function of helping and guiding the user through complex procedures.

Another serious drawback that results from translators not having access to shared language resources, is that gross mistranslations can result, e.g. the translation of 'default' as **–nephutha** / **-yiphutha** '(something) that is faulty / a fault / wrong'.

It must be borne in mind that the translators involved in these localisation projects are usually highly experienced translators, which underscores the necessity for translators to be able to access the kinds of language resources mentioned in this paper.

4. Towards Developing Language Resources for African Language Translators in the South African Context

As indicated in this paper, there are various language resources already available on the web that the translator can use to build corpora that can be interrogated with various corpus query tools and that can be recycled / reused by making use of translation memory tools.

What is needed, in this regard, is translator training of the type that is provided at the University of Pretoria (UP), which will equip translators to take advantage of such electronic translation resources and translators' tools.

In addition, every year large amounts of translations in the SA Bantu languages are produced by student translators as part of their training, as is for example the case in specifically the postgraduate UP translator trainee courses. In this regard, and in cooperation with my students, I plan to establish an (interactive) online database containing student outputs in the form of glossaries/term lists. In this way it would be possible to receive input from interested parties regarding the suitability/acceptability of the various terms and also to provide a service to other translators and language workers. In time, such a multilingual student site could become a very large, comprehensive and valuable language resource that will contribute not only to the development and elaboration of the African languages as technical languages, but also towards the standardisation of these languages. The results of this exercise could then be passed on to the relevant language bodies for their consideration.

Such an initiative would go some way towards providing the sort of cooperative resource sharing envisaged by the organising committee of this workshop, and it should be coupled with similar initiatives that may already be taking place at other South African institutions.

5. Conclusion

In conclusion, I wish to refer again to the title of this paper and reiterate my call to African language translators: let's share and share alike in order to advance, develop and elaborate the previously marginalised official South African languages as high function technical languages.

6. References

Barlow, M. (2003). *ParaConc: A concordancer for parallel texts*. Houston, TX: Athelstan. See for this software also http://www.athel.com

De Schryver, G.-M. (2002). Web for/as Corpus. A Perspective for the African Languages. *Nordic Journal of African Studies*, 11(3), pp. 266-282.

Esselink, B. (2003). Localisation and Translation. In H. Somers (Ed.), *Computers and translation: A translator's guide*. Amsterdam: John Benjamins, pp. 64-82.

Gauton, R. (forthcoming). The anatomy of a localisation project in an African language – translating Windows XP into Zulu. *Special HLT Issue of the South African Journal of African Languages*.

Gauton, R. & De Schryver, G-M. (2004). Translating technical texts into Zulu with the aid of multilingual and/or parallel corpora. *Language Matters, Studies in the Languages of Southern Africa*, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), pp. 148-161.

Gauton, R., Taljard, E. & De Schryver, G-M. (2003). Towards Strategies for Translating Terminology into all South African Languages: A Corpus-based Approach. In G-M. de Schryver (Ed.), *TAMA 2003, South Africa. Terminology in Advanced Management Applications. 6th International TAMA Conference: Conference Proceedings. "Multilingual Knowledge and Technology Transfer"*. Pretoria: $(SF)^2$ Press, pp. 81-88.

Gauton, R., Taljard, E., Mabasa, T.A. & Netshitomboni, L.F. (forthcoming). Translating technical (LSP) texts into the official South African languages: a corpus-based investigation of translators' strategies. (To be submitted to *Language Matters*).

Kruger, A. (2004). Editorial: Corpus-based translation research comes to Africa. *Language Matters, Studies in the Languages of Southern Africa*, 35(1) (Special issue: Corpus-based Translation Studies: Research and applications), pp. 1-5.

Mabasa, T.A. (2005). *Translation equivalents for health/medical terminology in Xitsonga*. Unpublished MA dissertation. Pretoria, University of Pretoria.

Pearson, J. (1998). *Terms in context.* Amsterdam: John Benjamins Publishing Company.

Scott, M. (1999). *WordSmith Tools version 3*. Oxford: Oxford University Press. See for this software also http://www.lexically.net/wordsmith/index.html

Sony Ericsson Mobile Phone T310. 2003. ©Sony Ericsson Mobile Communications AB, 2003.

Sony Ericsson Mobile Phone T610. 2003. ©Sony Ericsson Mobile Communications AB, 2002.

Sony Ericsson Z520i User Guide. ©Sony Ericsson Mobile Communications AB, 2005.

Resource Development for South African Bantu Languages: Computational Morphological Analysers and Machine-Readable Lexicons

Sonja Bosch, Jackie Jones, Laurette Pretorius, Winston Anderson

University of South Africa

PO Box 392, UNISA, 0003, South Africa

boschse@unisa.ac.za, jonescjj@unisa.ac.za, pretol@unisa.ac.za, winston.anderson@btgroup.co.za

Abstract

The development of computational morphological analysers for South African Bantu languages is linked to a project funded by the National Research Foundation in South Africa. The main research question in the project concerns the development of finite-state morphological analysers for five Bantu languages, namely Zulu, Xhosa and Swati (belonging to the Nguni group of languages), and Northern Sotho and Tswana (belonging to the Sotho group of languages). This development is based on underlying machine-readable lexicons that conform to common lexical specifications and international standards. Due to the rich agglutinating morphological structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature. The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of standardised XML machine-readable lexicons for the South African Bantu languages, based on an appropriate general data model.

1. Introduction

The development of computational morphological analysers for South African Bantu languages is linked to a project funded by the National Research Foundation in South Africa. The main research question in the project concerns the development of finite-state morphological analysers for five Bantu languages, namely Zulu, Xhosa and Swati (belonging to the Nguni group of languages), and Northern Sotho and Tswana (belonging to the Sotho group of languages). This development is based on underlying machine-readable lexicons that conform to common lexical specifications and international standards.

2. Challenges posed by Morphological Analysis of Bantu Languages

Automated morphological analysers exist for many European languages, but the development of morphological analysers has only been reported for a few Bantu languages, such as Swahili (Hurskainen, 1992) and a few others in southern Africa (for example, Bosch & Pretorius, 2003). Due to the rich agglutinating structures of these languages, the morphological processing poses particular challenges. These challenges are of an orthographical, a morphological as well as of a lexical nature.

In the case of the Nguni languages, a conjunctive system of writing is adhered to with a one-to-one correlation between orthographic words and linguistic words. For example, the Zulu orthographic word *siyakuthanda* (si-ya-ku-thand-a) 'we like it' is also a linguistic word. The Sotho languages on the other hand, are disjunctively written, and the above mentioned single Zulu orthographic word is written as four orthographic words in Northern Sotho, namely *re a go rata* 'we like it'. These four orthographic entities constitute one linguistic word. The **orthographical challenge,** which lies in the writing conventions of the Bantu languages, may according to Hurskainen and Halme (2001:399), be ascribed to the fact that disjunctive writing systems "require a special treatment, before they can be analyzed successfully". Pre-processing of the text in order to identify linguistic words, before morphological analysis takes place, is one of the options of addressing this challenge.

The **morphological challenges** in computational morphological analysis are twofold and comprise the modelling of two general linguistic components, namely morphotactics (word formation rules) as well as morphophonological alternations:

- The morphotactics component includes word formation rules, which determine the construction of words or word forms from an inventory of morphemes. This inventory of morphemes consists of word roots and affixes. Morphemes that make up words cannot combine at random, but are restricted to certain combinations and orders. A morphological analyser is required to recognize valid combinations of morphemes of the language in question.
- The morphophonological alternations component deals with the morphophonological changes between lexical and surface levels. A morphological analyser should identify the correct form of each morpheme since one and the same morpheme may feature in different ways depending on the environment in which it occurs.

The main **lexical challenge** in the building of morphological analysers for the Bantu languages is the fact that machine-readable lexicons, which are fundamental resources, are not readily available in any form. Although online dictionaries for Bantu languages are reported on by de Schryver (2003), such dictionaries available for Zulu and Xhosa for instance, contain a maximum of 2000 to 3000 lemmas and do not include explicit linguistic information, which is essential for a word root dictionary of the analyser. In the case of Northern Sotho, a bilingual online dictionary SeDiPro 1.0 (de Schryver, 2003:10) containing over 20 000 entries is available with linguistic information. However, such online dictionaries are only accessible for look-up of individual words or word stems, and are not accessible as a whole.

3. Meeting the Challenges

3.1 Orthographical Challenges

The Sotho languages pose a pre-processing challenge in that the disjunctive orthographical tradition isolates as separate "words" what are essentially affix morphemes of a lexical unit. Thus in order to correctly analyse multi word lexical units morphologically without causing excessive ambiguity, a multi word tokeniser is required. For Northern Sotho, this was addressed by first constructing regular expressions to deal with all verb constructions and they were then extended to address all predicate constructions. These cater for the most complex multi word tokens in the Northern Sotho language.

The grammars historically cover the verbs and copulatives reasonably adequately but other research theses and more modern study grammars (for example, Louwrens, 1989) had to be consulted to get consolidated views of these rules. None of the sources adequately covered what Ziervogel and Mokgokong (1985) term "deficient verbs", but in other texts are referred to as auxiliary verbs. A new linguistic research project is now under way to examine these in more detail.

There are various computational alternatives to producing a tokeniser (Hurskainen & Halme, 2001). The Northern Sotho morphological analyser team chose the approach of using finite state software to construct the tokeniser (Beesley, 2004a). The tokeniser for Northern Sotho now adequately deals with all predicate clauses (verbs, auxiliary verbs and copulatives). For more information see Anderson and Kotze (2006).

These regular expressions and the tokeniser can now be used as the basis for tokenisers in the other orthographically disjunctive Sotho languages, namely Tswana and Southern Sotho.

3.2 Morphological Challenges

Since human language technology is a novel field of research in South Africa, especially in the field of Bantu languages, a team approach was decided on for the morphological analysis project. Each language team consists of a computer scientist and one or two linguists. In order to facilitate multi-disciplinary team co-operation, all team members participated in at least one of the training courses presented by Dr Ken Beesley, Principal Scientist and Computational Linguist, Xerox Research Centre Europe, Grenoble, France. Three such courses have taken place in South Africa so far, and were organised by the Special Interest Group for Language and Speech Technology Development (SIG) of the African Language Association of Southern Africa (http://www.alasa.org.za/sig)[1].

Morphological analysis in this project is based on a finite-state computational approach, using the natural language independent Xerox Finite-State Tools (Beesley & Karttunen, 2003). This integrated set of tools is used to model and implement the complexities of word-formation rules as well as morphophonological alternations by means of finite-state networks. The latter are subsequently combined algorithmically into larger networks that perform morphological analysis.

The Xerox tools provide a declarative programming language, **lexc** (Lexicon Compiler) for specifying the required natural language lexicon and for modelling the morphotactic structure of the words in the language concerned.

Alternation rules are subsequently needed to map the abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool xfst.

In practical terms this means that all morphemes in the natural language need to be arranged in a cascade of LEXICONs (in a **lexc** description), while each entry in a LEXICON consists of morphological information and either a continuation class (the name of the next LEXICON in the cascade) or the end symbol #, which indicates the end of a valid morpheme sequence, as shown below in the example of a **lexc** description:

```
...
LEXICON NounPrefixes
...
i[NPrePre7]si[BPre7]:^I^SI          NStem;
i[NprePre8]zi[Bpre8]:^I^ZI          NStem;
...
LEXICON  NStem
...
gubhu                            NClass7-8;
...
LEXICON NClass7-8
@U.CL.7-8@                         NomSuf;
...
LEXICON NomSuf
ana[DimSuf]:ana                      #;
...
```

The **lexc** source file is then compiled into a finite-state network. This network recognises morphotactically well-formed, but still abstract morphophonemic or lexical strings such as

i[NPrePre7]si[BPre7]gubhu[NRoot]ana[Dim].

Alternation rules are subsequently needed to map these abstract lexical strings into properly spelled surface strings, as they occur in the natural language. The alternation rules are formulated as regular expressions, and are then compiled into a finite-state network by means of the Xerox tool **xfst.**

The orthographic changes that are manifested between the lexical and surface words when morphemes are combined to form new words or word forms are described as illustrated in the following example:

```
b h [o|u]  -> j  || _ a n a
```

This alternation rule models the change of a bilabial sound *-th-* appearing in the final syllable of a noun stem such as *–gubhu* to a palatal sound *-j-* when the diminutive suffix *-ana* is added to the noun stem.

The final step in the development of the morphological analyser is the combination of the **lexc** and **xfst** finite-state networks by means of composition (see Beesley & Karttunen, 2003) into a single network, a so-called lexical transducer. This transducer constitutes the morphological analyser and represents all the morphological information about the language being analysed.

Figure 1 gives a schematic representation of the application of a morphological analyser.

[1] Work on the development of a finite state morphological analyser for the Nguni language Ndebele, by Axel Fleisch (Fleisch & Seidel, 2006) is a direct result of one of these courses.

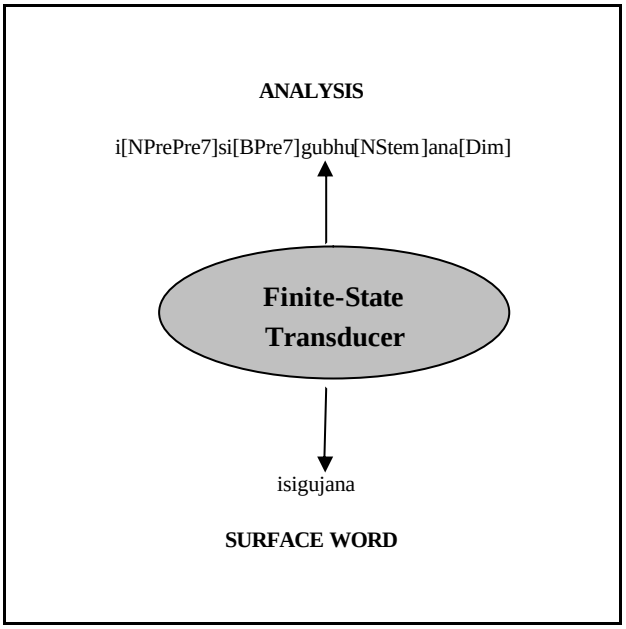


Figure 1: Schematic representation of a morphological analyser

In Figure 1 the morphological analyser maps the Zulu morphemes *i-*, *-si-*, *gubhu* and *-ana* to *isigujana* 'little calabash'. In other words, if the surface word *isigujana* constitutes the input string to the finite-state transducer, the output string is the morphological analysis which consists of the following morphemes in combination with their morphological feature tags: i[NPrePre7]si[BPre7]gubhu[NStem]ana[Dim].

The arrow in Figure 1 indicates the bidirectionality of the transducer and shows that analysis takes place in the upward direction while generation takes place in the downward direction. For more details regarding the Zulu morphological analyser prototype (ZulMorph) see Bosch and Pretorius (2003) as well as Pretorius and Bosch (2003).

3.3 Lexical Challenges

In addressing this problem of unavailability particularly for Zulu and Xhosa, a lemma list in electronic format was extracted from a Zulu paper dictionary (Doke & Vilakazi, 1964). For Xhosa however, the resources available in terms of lemmas were even more limited. This therefore demanded a time consuming exercise of extraction of lemmas from existing Xhosa paper dictionaries. Lemmas were retyped from a number of dictionaries and the scanning and proof reading of these resources increased and contributed to the development of the lemma lists substantially. These various sources yielded data in largely varying formats and forms containing many inaccuracies and errors. The non-existence, but urgent need for lemma lists for Xhosa also created the opportunity for researchers to devise a practical compilation procedure in accordance with appropriate standards in order to ensure reusability. The procedure for producing a large and reliable collection of Xhosa nouns and verb stems from this data consisted of a semi-automated data validation phase and, in the case of nouns, an automated generation phase. Data inconsistencies were identified by means of Perl-style pattern recognition, then scrutinized and corrected by the linguists in the team in the data validation phase. The validated data formed the input to the automated generation phase. Nouns were generated in two formats. The first of these was for human readability and the second was in an XML document. The second of these is particularly important in the creation of reusable lexical resources for future applications.

The only available Swati paper dictionary is being scanned and proofread in stages and then included into an electronic lemma list. Similarly for Tswana a paper dictionary has been scanned and is in the process of being proofread also to be developed into a lemma list for use in morphological analysis.

Regarding word lists for Northern Sotho, the major dictionaries were examined. The largest, the Comprehensive Northern Sotho dictionary (Ziervogel, D & Mokgokong 1985) includes support for the extra vowels (beyond the five standard vowels) marked with a circumflex in Northern Sotho, as well as support for the letter š and their capitalized form. Furthermore, the comprehensive dictionary includes tone markings on each main entry. In order to obtain a word list accurately scanned, these characters needed to be recognised by optical character recognition (OCR) software. No standard Northern Sotho OCR packages are available, so standard language settings were used. The scanning errors are consistent with the incorrectly scanned characters. Therefore, Perl scripts were developed to automatically correct the incorrectly optically recognized text. A further process of human editing is now underway to confirm all corrections. Subsequent to the dictionary scan, many other works have been scanned to add to the Northern Sotho test corpora. Eastern European language recognizers, such as Czechoslovakian, have proved most effective in recognizing characters (due to their adequate handling of š and its capitalized form).

For Northern Sotho, the lexicon structure used by the lexicographers of the Comprehensive Northern Sotho dictionary (Ziervogel & Mokgokong, 1985) does cater for complex lexicographic information. XML Schema (XSDs) was chosen to ensure accurate validation of lexical elements. The design also took other aspects into consideration including the most modern linguistic grammar descriptions, previous tag sets defined for Tswana, the work done by the ISO team on tagging and workshops provided by Xerox on using XML lexicons as a basis for finite state lexicons (Beesley, 2004b).

In terms of the lexical challenges our ultimate aim is to develop from these above-mentioned word lists and paper dictionaries machine-readable lexicons according to a standardised XML format that would be applicable to all the languages under investigation.

4. Current Status of the Project

The current status of the project is reported on, firstly in terms of the development of prototypes of morphological analysers for the various languages, and secondly in terms of the development of machine-readable lexicons for the South African Bantu languages, based on the above-mentioned proposed data model.

4.1 Analyser Prototypes for the various Languages

The Zulu analyser prototype (ZulMorph) at present covers most of the morphotactics and morphophonological alternations required for the automated analysis/generation of: nouns of all classes, the positive and negative forms of verbs, pronouns, the demonstrative and copulative demonstrative, underived adverbs, relatives and adjectives, possessives, conjunctions and ideophones. Word categories that still need to be completed are compound tenses of the verb, and derived adverbs. Preliminary testing of the current prototype was done on a test corpus consisting of 30 000 types. The application of the morphological analyser to the test corpus results in the recognition of approximately 77% of the types in the corpus.

The research aims for the other Nguni languages in the project, i.e. Xhosa and Swati closely follow those for the Zulu, since all these languages follow a conjunctive writing system. This enables the fast-tracking of the development of the morphological analysers for the Nguni languages by adapting the Zulu continuation classes and rules. Implementation and testing of the model in terms of the Xerox finite state tools are already in progress.

Regarding Northern Sotho a framework is under development for the nominal and verbal structures of Northern Sotho, with special emphasis on establishing the order of verbal extensions, reduplication patterns in nouns and verbs, as well as formalising rules for the derivation of morphological processes that involve the phonological process palatalisation in the formation of passives and diminutives. Implementation and testing of the Northern Sotho prototype (NsoMorph) is based on a limited, though representative lexicon, while cleaning up a scanned version of a Northern Sotho dictionary is in progress.

The first prototype of a morphological analyser for Tswana (TsnMorph) is being developed with nouns being treated first, while other word categories are added systematically.

Progress with the development of analyser prototypes for the various Bantu languages in the project has been reported in a number of publications, as listed in the bibliographical references at the end of this paper.

4.2 Development of Machine-Readable Lexicons

By definition the analyser can only recognise and analyse words of which the roots/stems have been explicitly included in its embedded lexicon. Ideally, a comprehensive machine-readable lexicon in the form of an XML document should be available for each language as a basic resource from which word roots/stems may be obtained.

As stated previously, electronic lemma lists for Xhosa and Zulu have been developed albeit on a small scale. To date a lemma list in Zulu extracted from the paper dictionary contains a total of over 28 000 entries. For Xhosa, also extracted and generated from a number of paper dictionaries, some 27 240 Xhosa lemmas were collated of which 20 845 nouns and 6 093 verb stems. With regard to the development of the machine-readable Xhosa lexicon, work continues on the manual capturing of data using an XML editor. A total of 4 138 entries have already been captured. Entering of the nouns is now completed and work on the verbs has just commenced. It is envisaged that the entering of this data, which will include all word categories, will be completed by the end of 2006.

In order to eventually arrive at an XML lexicon structure, the underlying standardised data model needs to be formulated and verified first. This is the subject of recent work in this regard (Bosch, Pretorius & Jones, 2006) where a data model towards a standardised machine-readable lexicon for all languages in the project is developed and formulated as an XML DTD. This model aims to ensure maximum inclusiveness of all linguistic information and to provide flexibility and handle the various representations applicable to Bantu languages in particular and is therefore applicable to diverse uses of electronic lexicons ranging from research in numerous areas resulting in publication. Included in this data model are particular requirements for complete and appropriate representation of linguistic information as identified in the study of available paper dictionaries. As starting point the extent to which the Bell and Bird (2000) data model may be applied to and modified for the above-mentioned languages was investigated. It was found that changes to this data model were necessary to make provision for the specific requirements of lexical entries in the relevant languages. Our model differs in significant ways from the Bell and Bird model and ensures maximum inclusiveness of all linguistic information. It also provides flexibility and handles the various representations applicable to Bantu languages in particular and is therefore applicable to diverse uses of machine-readable lexicons.

The data model we developed andproposed is intended to contribute to the further discussion and development of a common scheme for storing lexical data not only for the South African Bantu languages, but for the Bantu language family as a whole.

5. Conclusion and Future Work

Morphological analysis is generally recognised as a technology that enables the development of more advanced tools and practical applications in various areas of natural language processing, such as part-of-speech tagging, syntactic parsing, text-to-speech systems, information extraction, and machine translation. Research in this project concerning the development of computational morphological analysers for South African Bantu languages has confirmed the importance of comprehensive machine-readable lexicons as fundamental resource of the morphological analysers. The current project in computational morphological analysis includes research into the development of automated morphological analysers for Zulu, Xhosa, Swati, Northern Sotho and Tswana, using finite-state methods in computational morphology. It is envisaged that the project will eventually cover all South African Bantu languages.

Further aims of the project are:

- Wider distribution of the intermediate versions of the electronic lexicon for constructive feedback from the broader community of lexicographers and other users/speakers of the

relevant languages.

- Research into lexicon design and development in order to contribute to the international definition of standards as envisaged by the International Standards Organisation ISO/TC37/SC4, whose goal it is to develop a platform for the design and implementation of linguistic resource formats and processes in order to facilitate the exchange of information between language processing modules (Romary & Ide, 2002).
- Research into place names as occurring in the various languages of the project, for inclusion in the relevant machine-readable XML lexicons.

6. Acknowledgements

This material is based upon work supported by the National Research Foundation under grant number 2053403. Any opinion, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Research Foundation.

7. References

Anderson, W.N. & Kotze, P.M. (2006). Finite State tokenisation of an orthographical disjunctive agglutinative language: The verbal segment of Northern Sotho. In *Proceedings of the 5th International Language Resources and Evaluation Conference*, Genoa, Italy.

Beesley, K.R. & Karttunen, L. (2003). Finite-state morphology. Stanford, CA: CSLI Publications.

Beesley, K.R. (2004a). Tokenizing Transducers. Xerox Research Centre. Europe. Unpublished course notes presented in Pretoria, South Africa, September 2004.

Beesley, K.R. (2004b). Downtranslation of XML dictionaries to lexc lexicons. Second draft. Unpublished course notes presented in Pretoria, South Africa, September 2004.

Bell, J. & Bird, S. (2000). A Preliminary Study of the Structure of Lexicon Entries. [O] Available. http://www.ldc.upenn.edu/exploration/expl2000/papers/bell/bell.html Accessed on 19 September 2005.

Bosch, S.E. & Pretorius, L. (2002). The significance of computational morphological analysis for Zulu lexicography. *South African Journal of African Languages*, 22, pp.11-20.

Bosch S.E. & Pretorius, L. (2003). Building a computational morphological analyser/generator for Zulu using the Xerox finite-state tools. In *Proceedings of the Workshop on Finite-State Methods in Natural Language Processing, 10th Conference of the European Chapter of the Association for Computational Linguistics*, April 13-14 2003, Budapest, Hungary. ACL. pp. 27-34.

Bosch, S., Pretorius, L. & Van Huyssteen, L. (2003). Computational Morphological Analysis as an aid for Term Extraction. In *6th International TAMA Conference Proceedings*, Pretoria: (SF)2 Press, pp. 65-71.

Bosch, S.E. & Pretorius, L. (2004). Software tools for morphological tagging of Zulu corpora and lexicon development. In *Proceedings of the 4th International Language Resources and Evaluation Conference*, Lisbon: ARTIPOL , vi, pp. 1251-1254.

Bosch. S.E., Pretorius, L. & Jones, J. (2006). Towards machine-readable lexicons for South African Bantu languages. In *Proceedings of the 5th International Language Resources and Evaluation Conference*, Genoa, Italy.

De Schryver, G-M. (2003). Online Dictionaries on the Internet: An Overview for the African languages. *Lexikos*, 13, pp. 1-20.

Doke, C.M. & Vilakazi, B. (1964). Zulu-English Dictionary. Johannesburg: Witwatersrand University Press.

Fleisch, A. & Seidel, F. (2006). Cologne Initiative on Natural Language Processing in African Languages. In *Proceedings of the Workshop on Networking the Development of Language Resources for African Languages, 5th International Language Resources and Evaluation Conference*, Genoa, Italy.

Hurskainen, A. (1992). A two-level formalism for the analysis of Bantu morphology: an application to Swahili. *Nordic Journal of African Studies*, 1(1), pp. 87-122.

Hurskainen, A. & Halme, R. (2001). Mapping between Disjoining and Conjoining Writing Systems in Bantu Languages: Implementation on Kwanyama. *Nordic Journal of African Studies*, 10(3), pp. 399-414.

Kotzé, A.E. (2005). Towards a morphological analyser for past tense forms in Northern Sotho: verb stems with final m and n. *Southern African Linguistics and Applied Language Studies*, 23(3), pp. 245-258.

Kotzé, P.M. (2005a). Towards a finite-state network for Northern Sotho deverbative nouns: the morphotactic rules. *Southern African Linguistics and Applied Language Studies*, 23(3), pp. 259-268.

Kotzé, P.M. (2005b). A finite-state transducer for Northern Sotho deverbative nouns: the morphophonemic rules. *Southern African Linguistics and Applied Language Studies*, 23(4), pp. 392-403.

Louwrens, L.J. (1989). Northern Sotho. Study guide for Grammar). University of South Africa: Pretoria.

Pretorius, L. & Bosch, S. (2002). Finite-State Computational Morphology - Treatment of the Zulu Noun. *South African Computer Journal*, 28, pp. 30-38.

Pretorius, L. & Bosch, S. (2003). Finite-State Computational Morphology: An Analyzer Prototype

for Zulu. *Machine Translation*, 18, pp. 195-216.

Pretorius, L. & Bosch, S. (2003). Towards technologically enabling the indigenous languages of South Africa: the central role of computational morphology. *Interactions of the Association for Computing Machinery*, 10(2), pp.56-63.

Pretorius, L. & Bosch, S. (2003). Computational aids for Zulu natural language processing. *Southern African Linguistics and Applied Language Studies*, 21(4), pp. 265-282.

Romary, L. & Ide, N. (2002). Standards for Language Resources. In *Proceedings of the 3th International Language Resources and Evaluation Conference*, 1, pp. 59-65.

Weber, D.J. (2002). Reflections on the Huallaga Quechua dictionary: derived forms as subentries. [O] Available. http://emeld.org/workshop/2002/presentations/weber/emeld.pdf Accessed on 3 October 2005.

Ziervogel, D & Mokgokong, P.C. (1985). Comprehensive Northern Sotho dictionary. Second corrected edition. J.L. van Schaik: Pretoria.

The African Anaphora Project

Ken Safir, Andrei V. Anghelescu, Sarah E. Murray, and Jessica Rett

Rutgers University
Department of Linguistics
18 Seminary Place
New Brunswick, NJ 08901
afranaph@rci.rutgers.edu

Abstract

The goal of the African Anaphora Project is threefold: to elicit a research-directed database of African language data, which is collected and analyzed with native speaker linguist consultants; to organize and present this collected data in a manner such that it is as widely accessible as possible; and, to provide a forum where project directors, consultants, linguists, and in general anyone interested in African languages or linguistics can share an interactive community research space. The achievement of these goals necessitates the development of an interactive, dynamic site capable of allowing many users with many different objectives to access, input, edit, search, and browse data. Such an implementation requires sophisticated language resources, such as a site management component, a data storage component, and a query component, as well as a tool to input and display Unicode correctly, and tools to export the data in a variety of formats. It is our hope that our project design and our technical implementation will be generalizable to other projects that seek to collect complex linguistic data online and make it available to online users.

1. Introduction

The African Anaphora Project (NSF grant BCS-0523102, Ken Safir, Principal Investigator) was established to develop in-depth descriptions of a wide range of African languages in order to facilitate linguistic research into the nature and distribution of anaphora. Anaphora, in the sense intended here, is the phenomenon where one linguistic form, such as a pronoun, reflexive or reciprocal, refers back to a previously mentioned form in the sentence or discourse. This phenomenon is common to every human language; the African Anaphora Project (=:AfrAnaph) aims to explore anaphora — its forms, distribution and interpretive effects — for every African language with a native speaker consultant who is willing to help us by filling out an in-depth anaphora questionnaire (=:AQ).

Although our project is grounded in a generative framework, and our focus has been the phenomenon of anaphora specifically, it is our intention to insure that the data we collect will be accessible to anyone interested in learning about these languages or in more general and/or theoretically oriented typological research.

Our theoretical bias is toward nativist accounts of language competence, hence we posit a universally available language forming capacity in human beings that operates no differently for speakers of African Languages than for speakers of any other language. From this point of view, even if the world's languages vary enormously, thanks to relatively small formal differences with large effects, or because lexicons and phonological forms must differ, there are commonalities which may be extrapolated through careful study. In other words, there is a core plan common to all the world's languages that can be most profitably studied by closely examining how they are similar and how they vary. For example, certain constructions are unavailable in every human language; accordingly, we are just as interested in ungrammatical forms as we are in grammatical ones.

Anaphora is an interesting object of study because of its variation cross-linguistically; languages vastly differ in what forms can be used as anaphors and how the meanings of these forms are distributed. Additionally, its study is of crucial importance because it has provoked vast debate in the generative project, inspiring slews of different theories which account for varying ranges of data. We desire that through careful elicitation and study, the AfrAnaph project will help to shed some light on this debate and inform the current and future theories.

However, the goals of the AfrAnaph project extend beyond furthering generative research into this specific phenomenon. We fully hope and intend that researchers with goals and perspectives different from ours may find our data pertinent and accessible. Thus, this project was designed with broader interest in mind, and both the range of data and the technological contributions will reflect this.

This project is feasible at this point in history not only because there are an unprecedented number of trained African linguists who are potential participants in our project, but also because digital technologies and the resources of the Internet make it possible for more efficient remote participation. However, the technological challenges which Africa is presented with are also a challenge for AfrAnaph, and every effort has been, and will continue to be, made to ensure that even those with limited access to technological resources will be able to contribute to and access our database.

Since the beginning of the project three years ago, we have differentiated several sub-tasks to achieve our goals. They include a) improving the efficiency of data collection; b) standardizing the presentation of data; c) making the collected data searchable for the research community; and d) implementing a forum where both project participants and the general research community can discuss our data and propose new research initiatives.

These tasks are currently being realized with the development of a new website. Making use of the transactional object database of the open source web application server Zope (http://www.zope.org), we are designing Unicode-compliant dynamic HTML templates for use both by consultants — to input the data in a standardized fashion — and by researchers — to search and browse the data. Additionally, to accommodate

possible technological difficulties, we will provide a number of offline resources, like exporting of data and snapshots of the database.

Section Two of this paper discusses the background of the African Anaphora Project as well as the specifics of its goals. Section Three discusses how we have begun to implement these goals, the software we are using—Zope—and how it both helps to achieve the specific project goals as well as the goals of the larger language documentation project. Section Four is a brief summary of where the AfrAnaph project is today and where it will be in the near future.

2. Project Background and Goals

Our data collection begins with invitations to native-speaker linguists of African languages to fill out a comprehensive Anaphora Questionnaire. In follow-up interactions with consultants, we analyze AQ responses, expanding and elucidating whenever issues of particular interest are uncovered. Previously, the AQ was available online for download; most consultants completed it offline on a word processor and then emailed it back. The data would then have to be transposed into a form presentable on the website and consistent across surveys. However, we have recently developed a dynamic input interface through which consultants can directly input the data into our database, returning later to finish the questionnaire or to edit the data. This method is much more accurate and efficient. Once the data is approved, it is accessible to anyone who accesses the AfrAnaph webpage. The technical details of this development will be discussed in Section 3.

In addition to the language data, each language has its own "case files" containing the results of the questionnaire (which is supplemented by follow-up data elicitation focusing on unique and rich aspects of that language), a short grammar sketch of the language, a sketch of the anaphora system (highlighting aspects of anaphora in the language that appear to raise interesting theoretical or analytic issues), a bibliography, and links to related linguistic and cultural resources. In the future we hope that each case file will also contain audio files, papers available on line, responses to additional questionnaires on other topics, and language-specific bulletin boards and venues for online discussion between researchers. Case files for Berber, Bukusu, CiNsenga, Urhobo, and Yoruba are already complete, and case files for 15 more are in development. Completed case files can be viewed on our website, http://www.africananaphora.rutgers.edu.

Our previous method of eliciting data was to send an informant a questionnaire, which he or she would fill out, and then send back to us. After the questionnaire had been filled out, a long process with a fair amount of back and forth to clarify the responses and questions would have to be undertaken. Additional follow-up questions were asked, which were designed, for example, to explore questions that can be profitably asked for those domains where one language has a more articulated set of distinctions than others.

This method of elicitation is clearly limited. Though it has allowed us to collect invaluable data from a number of languages, there is clearly room for improvement. The goals for the new website have been prompted by challenges we've faced eliciting, recording and publishing data over these past three years. Most languages we study have a complex alphabet as well as complex tonal systems. Faithful elicitation and reproduction of this linguistic data is a challenge, especially when the informants don't have access to PDF writers and readers. This required that our informants and interested researchers download a Unicode font, something that was often a challenge for many informants without access to optimal or consistently available internet resources.

Additionally, since each informant was completing his or her AQ on his own word processor, not all AQs had a consistent format. The data, often 70-80 pages, was not at all indexed, much less cross-indexed, and was cumbersome to navigate.

The goals then for the new website are to have a standard method of data entry, so that all of the data will be formatted correctly, will display correctly (with all tones and diacritics), and will be easier to navigate, for both the consultants and project directors during follow-up sessions, and for researchers browsing and searching the data.

It is clear that improving data elicitation and recording will save time, make the data easier to search and study, and improve the accuracy of data recording. At the same time, the new website will provide a venue for research discussion and debate, between the project researchers and informants and the general research public.

We are revising our offline input template for the AQ response and aligning it with the online data entry option using our dynamically generated interface pages. The former option insures that we do not exclude potential consultants whose access to internet and software resources is limited. For researchers, the data will be exportable in a number of formats, including HTML, XML, and PDF. Website discussion boards will be provided to meet objective (d) above. As our methods and our researcher network evolve, we hope they can be applied to a wide variety of linguistic phenomena, while creating a community space for research into African languages.

3. Implementation

As stated above, the goals of the AfrAnaph project necessitated the development of a language resource specifically designed to accommodate it. In creating such a resource, we believe that we have also developed a tool which is generalizable—one that can be easily modified for use by many similar projects.

Essentially, what we are developing is a dynamic version of the Anaphora questionnaire which we have been using (http://www.africananaphora.rutgers.edu/NSF-ques.pdf). This online version of the questionnaire is broken up into subsections organized by topic, and the consultants may fill out any question in any order, save these responses, and return at a later time to finish other questions or edit their responses. This data will be kept private until approved for public viewing, a task done by project director under the consultant's approval. These responses will then be available to anyone who searches or browses our site.

The answers to the questions can be viewed individually, or the whole questionnaire for a language may be viewed. Additionally, researchers may choose to

view the answer to one or several questions from several languages, in order to compare the forms.

We have designed HTML templates for use both by consultants—to input the data in a standardized fashion—and by researchers—to search and browse the data, by any combination of language, language family, grammatical phenomenon, morphemic gloss, or substring. Consultants will be able to input their data in the IPA, complete with diacritics, through use of a customized keyboard or keyboard shortcuts, and both methods can be personalized by individual informants for the needs of their language. Additionally, our morphemic glossing convention, based mostly on the Leipzig standards (Bickel et al, 2006), will make searching easier and more efficient, while minimizing, if not eliminating, the need for tagging.

This remainder of this section presents the technical details about the implementation of the Anaphora online database, as follows: a general introduction to online databases, the technical solution adopted, the hardware and software requirements, the data input tools, and offline access. We conclude with a description of the offline facilities.

3.1. Online databases

The Anaphora server implements a web-based content management architecture that allows end users to collaborate towards building a database of answers to a pre-defined questionnaire.

The database is implemented using the Zope Object Database (ZODB) framework, which follows the object oriented database paradigm. In contrast to the more traditional relational database systems (RDBMS), the data is not represented as a set of tables with relationships designated between them. In fact, the representation of the data is a natural description of the information stored in the database. For example, a questionnaire is structured into sections, which together compose the questionnaire. According to an object oriented representation, the questionnaire would then be represented as a container filled with objects representing sections. Analogously, sections are objects, which may contain other objects (e.g. subsections, questions, etc).

The object oriented representation of such a structure has several advantages over a representation using a relational database model. Conceptually, it is more intuitive, because it follows closely the structure of the questionnaire to be stored. Each item in the questionnaire (e.g. section, subsection, question, etc.), is an object with properties and methods associated to it. This allows for great flexibility in the design of the functionality presented to the user, since the programmer can conveniently write code for any particular class of objects and this code is stored in the database as part of the object itself. The hierarchical representation of data, closely tailored to the purpose of the database (to store and query the answers to a questionnaire), allows for very fast searches, using tree searching algorithms. Such searches are an order of magnitude faster than searches in tables of data when using tree agnostic relational database systems (most common systems available today).

The main interface to the database is network oriented, using a web browser front end over an encrypted communication channel. The user is presented a dynamically generated website, where each page contains the information associated with the particular object displayed, as well as links to the actions allowed for that context. For offline processing purposes, the data may be exported in a text format, described using XML.

3.2. The Implementation of the Anaphora Database

The server is built on top of the Zope 3 framework, which uses Python as a programming language and XML for configuration files. The Zope 3 framework provides utilities for various common tasks, such as template-driven dynamic HTML generation, user authentication and data storage. The remaining tasks, such as data querying, have been written from scratch, taking advantage of the object oriented model provided by the Zope Object Database (ZODB).

The server, a *site* in Zope parlay, is structured in the following main components:

- a site management folder, responsible for user registration and authentication
- a data storage component, containing trees of objects, as follows:
  - user annotations (e.g. user profiles)
  - questions
  - answers to the questions
  - general information about the languages present in the database
- a query component, which handles queries to the database

3.2.1. The data storage component

The site is a hierarchical structure of objects, with the root of the site containing the questionnaire, language information folder and the site management folder. Each of these components is a container on its own right.

The questionnaire is structured in sections (or question groups), each of which may contain questions and subsections. The only restriction is that a question is not a container, in the sense that it may only store its text (description) and information about the type of answer (e.g. yes/no, string). As such, the questionnaire may be arbitrarily deep, although for practical purposes one should devise it so that it doesn't contain more than four levels.

A question is a structure containing the following:

- description: a free-form optional text (HTML tags are honored), containing some information about the purpose of the question.
- text: a mandatory text (HTML tags are honored), containing the actual question.
- answer type: the type of allowed answer. This may be one of the following:
  - string: a one line, free form string. No HTML allowed
  - boolean: yes/no
  - single-choice: choices are provided by the editor of the questionnaire (HTML tags are honored)
  - multi-choice: choices are provided by the editor of the questionnaire (HTML tags are honored)

For efficiency, the user input is converted to the appropriate data type: yes/no become booleans, single choice answers become numbers, and multi-choice answers are stored as sequences of numbers. The

reasoning behind this choice is that it makes searching these data types efficient.

In general, objects may contain HTML tags in their descriptions and contents. Since some of these are dynamically inserted in the HTML page generated on the server, this leaves some of the server code vulnerable to malicious HTML code attacks. For this reason, only the administrator and the editor of the questionnaire are allowed to input such code. The data input by all other users is interpreted literally, and the HTML code is displayed as plain text, with tags included.

The language information folder is a list of objects containing general ethnological information about the languages represented on this site.

3.2.2. The site management folder

The site management folder contains the information associated with the registered users. This information contains the user profile, the data the user contributed to the database and the workspace of the users (e.g. query objects saved by the user). The workspace and the personal information are accessible only to the user, while the contributed data is public (the user may opt to not be publicly associated with the data).

The access to the site is built around the notion of "connection", where users authenticate using a login name/password combination. The action of authentication creates a *user authentication token*, which is transiently stored in the browser and sent with every request. The absence of such a token is interpreted as lack of authentication, the user being then assigned the status of guest by default. The set of resources and actions available in a context may be restricted using security policies. These policies are defined by a privileged user (root). A policy defines what classes of users may access a resource (e.g. view the content of an object or call a method of an object). Subsequently, users are assigned to one or more classes. The granularity of such a system of permissions allows to specify permissions for each object individually, or to define classes of objects and their interactions.

A user is represented as an object that stores the following information:

- authentication data (login name, password)
- personal information
- data contributed by the user
- a user workspace, which contains queries saved by the user

The login name is chosen by the user, and it must be unique in the database. This is a somewhat artificial constraint, since Zope does allow for users with the same login name, but is implemented only for public identifiability concerns, of particular relevance in the forums. As the number of users is expected to be small (by modern standards), this constraint will not be a concern. However, it is important to note that every user is assigned an unique numerical id, which is used for indexing purposes by the system.

This id is also used when exporting the data to XML. In UNIX tradition, the password is never stored on the server: instead the server generates a hash using the Blowfish algorithm, and stores this hash. At every login, the hash is re-calculated from the password input by the user, and the authentication is granted only when the tested hash is identical to the one stored on the server. The Blowfish algorithm allows for passwords of arbitrary length, and it is sufficiently expensive to offset dictionary based attacks, and should provide our users with a requisite level of security.

The minimum of information requested from a user is a login name and a password. There are no requirements of personal information; this option is left to the user, who may also specify the information as private (to be viewed only by the project directors and administrators) or public (able to be access by either other registered members of the site or to all users).

Each user will have a certain set of permissions, which will allow the possibility of entering data, editing it, or simply browsing and searching. By default, registered users are not allowed to input data. To achieve the status of contributor, the user must request such credentials from the administrator of the site. This can be handled entirely online, and will provide safeguards to ensure the accuracy and the integrity of the presented data.

For the user who does not wish to, or for some reason can not, register, the database may be browsed without registration. For users who do not desire to input data, the major benefit to registration is the workspace where the user may save queries.

3.2.3. The query component

Given the multi-variate nature of the data, with diverse features (e.g. language, type of answer, goal of the question asked, user), the most natural and flexible approach to querying it is to represent the available data as an amorphous block.

The user may then use a drill-down method, imposing successive constraints on the data features to reduce the number of matching entries to a set that meets a goal. Each such simple constraint limits the values of one feature of the data. The entire filter is the conjunction of the simple constraints it contains.

As such, the result of applying the filter on the data is the set of entries that meet all the simple constraints specified in the filter. Formally, this can be defined by the following grammar:

< Filter > ::=< SimpleCondition > [∧ < Filter >]
< SimpleCondition > ::=< FeatureName > < operator > < value range >
< FeatureName > ::=< string >
< operator > ::= [not](< | > | = | ~)
< value range > ::= (< number > | < string > | < regular expression >)

Since this is a multiple conjunction, the simple conditions are commutative, so they can be applied in any order. This is important for query optimization, because the condition can then be applied in the order that minimizes the number of operations (i.e. data transferred, which is the bottleneck).

Also, since simple conditions are objects, they can be stored in the database (the workspace of the user), with the entire query being just a list of simple conditions.

In what concerns the performance of the query, the implementation is non-trivial. In particular, the user inputs the constraint as an HTML query (that is, a formatted string), which doesn't match the structure or the data types of the objects in the database. Questions in the database can have one of the following types of answers: free form

strings, booleans (yes/no), and numerical values (e.g. answer no. 3 in a list of five options). As comparing strings is a slow process (even with the fastest matching methods available, it still is linear time complexity), the constraints are firsts compiled into a more efficient representation.

When a constraint is created, the interpreter extracts the properties of the feature to which the constraint applies, and finds the data type stored in this feature. Then, the value range introduced by the user is converted to the data type of the feature, allowing for faster comparisons (constant time) for the nonstring data types.

Since a query filter is evaluated as the conjunction of all simple constraints it contains, the order of applying the constraints does not change the result. As such, the constraints known to be calculated faster (e.g. numerical comparisons) are applied first, reducing the number of expensive comparisons, because those would be applied to already filtered data sets. This is implemented by a simple search procedure, applied every time the filter is run: select all numeric constraints, and apply them in the order found in the list, then repeat for the string constraints. This step could be made marginally more efficient by maintaining a sorted list of constraints, with those affecting numerical features moved to the top. However, this would be a usability nightmare, since the constraints would end up in a different order than input by the user. As such, this minor performance hit is by far the better choice.

The constraints applying to string features may also be regular expressions. This takes advantage of the facilities provided by Python for regular expression matching, by using its re module. Since a regular expression is compiled into a finite state automaton, the structure used to store such an automaton could become somewhat large. For a small number of concurrent users, this will not be a problem. However, it is unclear how it would scale when the number of concomitant users reaches hundreds.

3.3. Language-Specific Input Tools

One challenge for online data collection and presentation is that of character input and display. A common solution to this problem is the use of the character encoding standard Unicode, which can represent the vast majority of scripts, and allow the simultaneous display of many different scripts. However, the problem of data entry still remains: a goal for any online elicitation should be the use of a tool which creates a reasonably simple method of entering characters, eliminating the need for a complex system of keyboard shortcuts. One such available tool is CharWrite (E-MELD 2005), developed by the LinguistList, which allows the user two options for data entry. The first is the user can type in an ASCII character, and then right click to bring up a table of similar characters. The second option is to double left-click on the input field to bring up an interactive IPA chart, which provides the user a workspace in which they may create strings of IPA characters, and then send them back to the input form.

3.4. Software and hardware requirements

The software requires Zope 3.2.0 or newer, and Python 2.4.2 or newer. It has been developed and tested on a SuSE Linux platform (versions 10.1 x86 64, 9.3 i386), but it should run without changes on any platform that provides the required versions of a Python interpreter and Zope. Additionally, the Python cryptography toolkit (Kuchling 2005) — which provides the Blowfish algorithm for password hashing — is required.

In terms of processing power, a commodity PC should provide sufficient power for handling on the order of 100 concurrent connections.

3.5. Offline Access

We are also revising our offline input template for the AQ response and aligning it with the online data entry option using our dynamically generated interface pages. Basically, we can create an offline snapshot of the questionnaire which can then be downloaded (or burned to CD and mailed) and filled out at the consultant's leisure. This insures that we do not exclude potential consultants whose access to internet and software resources is limited. For researchers, the data will be exportable in a number of formats, including HTML, XML, and PDF, and snapshots of the database will also be available. All data representation will be made consistent with all data-sharing protocols such as those suggested by the Linguistic Data Consortium. For example, we want the data we collect to be converted into a standard method of data representation which will allow the data to be used by tools developed in later phases of the project or by independent researchers. Similarly, bibliographic entries will be made consistent with existing protocols.

All data representation will be made consistent with all data-sharing protocols such as those suggested by the Linguistic Data Consortium. For example, we want the data we collect to be converted into an XML data format, which is a standard method of data representation, one that will allow the data to be used by tools developed in later phases of the project or by independent researchers. Similarly, bibliographic entries will be made consistent with existing protocols.

4. Summary (and Future Extensions)

There are substantial plans for the future of the AfrAnaph project. We plan to develop a French language version of the AQ which will facilitate elicitations for those whose second language is French. Additionally, we will devise new questionnaires for other aspects of anaphora, including a logophoricity questionnaire (developed by Oluseye Adesola and Ken Safir) and a specialized reciprocals questionnaire, but we will also commission questionnaires that explore other well-studied, compact empirical domains that have been known to vary in interesting ways, such as the nature of questions or the nature of specialized focus constructions (in the languages that have them). We expect our research platform to be flexible enough to serve the interests of anyone who has a good idea about what can insightfully be investigated using our resources, including phonological or semantic phenomena. We will add audio files to each case file so that an interested participant can hear different forms being spoken. In addition to important information about intonation that can influence anaphoric interpretation or the acceptability of phrases, well-developed audio files may also be a resource for

phonologists interested in some of the languages in our case files.

In addition to the Listserve that we will develop on our current site, we hope to open a chat room for the members of the community we serve, in order to facilitate interaction and draw together researchers with common interests.

As our usership grows, we will have a bulletin board space and a newsletter space reporting what's new on the site and who is up to what. As our project grows and new work employs some of our data base, occasional papers will be published on the site as a series of technical reports connected with our project.

We hope to develop a software library of open source materials that our users can download for their projects, including tree diagram programs, fonts, other forms of graphic representation, etc., and perhaps more ambitious analytic tools for speech recognition and data analysis, as well as links to other site which have similar resources available. Participants in the project would make themselves available as references so that someone unfamiliar with a particular program could ask the listed reference person for advice on how to download it, install it and use it. However, all informants will have the option of anonymity if they so choose.

We will also develop a detailed case file for each language, which will include a grammar sketch, links to linguistic and cultural resources for the language, and a list of linguist consultants who are willing to be contacted by other researchers with questions about their language. In this way it may eventually be possible to have several consultants available for a novel project on a language specific basis. This will also serve to form links between researchers on a given language, organized around the development of the case file for the language they speak.

Our central business, however, is the careful and directed collection of linguistic data that is likely to be of use to researchers interested in specific questions which require sophisticated cross-linguistic data. A great many more languages need to be explored if we are to develop resources for even a partially representative sample of the languages of Africa on our site, a project that can only be result of many years or work with new and current consultants. On the other hand, it is also part of our ambition for the project that the technical implementation of our database, with its online access, data input, static presentation and data search and manipulation functions, will prove useful to other projects that strive to make linguistic resources available to anyone interested in linguistic research, especially those with limited technological access.

5. References

Bickel, Balthasar, Bernard Comrie, and Martin Haspelmath.2006. http://www.eva.mpg.de/lingua/files/morpheme.html

Kuchling, A.M. 2005. Python cryptography toolkit v2.0. http://www.amk.ca/python/code/crypto.

E-MELD. 2005. *E-MELD School of Best Practice: CharWrite Index Page*. Online: http://emeld.org/school/toolroom/software/charwrite-index.html

6. Acknowledgements

This project was initiated with NSF grant BCS-0303447 and is currently funded by NSF grant BCS-0523102. Many thanks go to all of our past, current, and future informants; without their participation, there would be no project. Additional thanks go to Oluseye Adesola, the Assistant Director of the project, and an anonymous editor for helpful comments. All errors are the authors'.

From Corpora to Spell Checkers: First Steps in Building an Infrastructure for the Collaborative Development of African Language Resources

Demi Yi-Chien Liu, Simon Chun-Feng Su, Laurel Yu-Hsuan Lai, Ellie Hsiao-Yun Sung, Jasmine Yi-Ling Hsu, Sibyl Yin-Chi Hsieh, Oliver Streiter

National University of Kaohsiung,
Department of Western Languages and Literature
No.700,Kaohsiung University Rd.,Nan Tzu Dist.,811.Kaohsiung,Taiwan, R.O.C.
just_aquamarine@hotmail.com, ahome0304@yahoo.com.tw, dofin6_6@hotmail.com, kiwibaby@pie.com.tw, miss5410@hotmail.com, sibyl330@hotmail.com, ostreiter@nuk.edu.tw
http://140.127.211.214/xnlrdf

Abstract

The acquisition of language data is a primordial step in the creation of NLP applications. In order to overcome the difficulties in this acquisition of language data from which taggers or electronic dictionaries can be derived for a wide range of languages, we develop a cooperative environment called XNRLDF. Aiming at a Wikpedia-like cooperation of linguists, we thus create an infrastructure necessary to collect, create an distribute data for the processing of under-resourced languages. The data currently available are derived via bootstrapping from Internet documents. Besides testing the infrastructure, we created corpora, word frequency lists, ngrams, list of number words and function words and links to external resources. The compiled data are available under the GNU public license. A wider cooperation with experts on African languages is needed to achieve an in-depth coverage of African languages.

1. Motivations

As the usage of the Internet extends from its main traditional applications, such as emailing, teaching, commerce, reading and distributing news, more and more languages are used in a wider range of language technologies, e.g. Information Retrieval, Web-publishing, e-learning etc. However, languages are, technically speaking, not supported to the same degree around the world.

In Africa, from the 1800 spoken languages, only relatively few have a writing standard. In addition, as non-African languages (Arabic, Afrikaans, English, French, Portuguese) frequently dominate in electronic communication, it is difficult to find electronic documents from which corpora or electronic dictionaries can be derived semi-automatically. If the writing of a language started with a Bible translation using Latin characters, many tasks necessary to port a language into an electronic environment, like creating fonts, input methods and editing tools, are more tractable than for non-Latin scripts, even if covered by Unicode (cf. Prinsloo & Heid 2005, Uchechukwu 2005). Non-Latin Unicode fonts like Abugida, still remain problematic as many potentially useful tools like SPELL or ASPELL have been developed for European languages. Even search engines like Google cannot handle queries formulated in Abugida.

To overcome the limitations in the processing of so many languages, the initiative of a single person or a single research body are of limited impact (cf. Streiter 2005). Therefore, a general infrastructure for the cooperation of linguistics, language activists, computational linguists, language teachers and socio-linguists is currently under development (Streiter & Stuflesser 2005). Basic resources for processing hundreds of languages have started to be collected by researchers and volunteers. Section 2 introduces our main concepts and the feature of the XNLRDF database. Section 3 outlines our approach to data collection, the tools for the elaboration of data and the usage of feedback loops to enrich the database. In Section 4, we will discuss again the notion of developing NLP-data and anticipate how interested students and researchers might assist in the expansion. Finally, a summary of the main achievements and an outlook for the future developments of the database will conclude the contribution.

2. XNLRDF: Resources and Tools in a Collaborative Environment

Our purpose is to provide an infrastructure for the collaborative development of NLP resources for African languages and other languages which lack computational resources. By using a Web-interface to a database (Streiter & Stuflesser 2005), we collect textual examples for major languages (important in terms of native speakers, second language speakers, and web-pages) and convert them into free resources, such as corpora, spell checkers and a language recognition tool. The high number of parallel texts in our textual examples will be used in the alignment of sentences and the automatic extraction of translation dictionaries. Additional tools are planned continuously, having the purpose a) to be useful as such b) to linguistically enrich the database through feedback functions and c) to motivate more researchers to participate in the project in the elaboration of linguistic data in electronic format.

The pivot notion in the data-structure of XNLRF is the writing system, defined as n-tuple of *language, locality, script, writing standard, orthography, range of validity* and *transliteration*. This writing system is the main meta-datum assigned to linguistic resources identified or collected in XNLRDF. After resources are entered into the database, they are available for download under the GNU-Public License after a short time.

For a project of this scope, an individual effort is definitely doomed to fail (cf. Streiter 2005). In addition,

fine-grained linguistic information is needed, e.g. for the development of tag-sets. We just hope that the resources

check the language of a downloaded document (cf. Baroni & M. Ueyama. 2004).

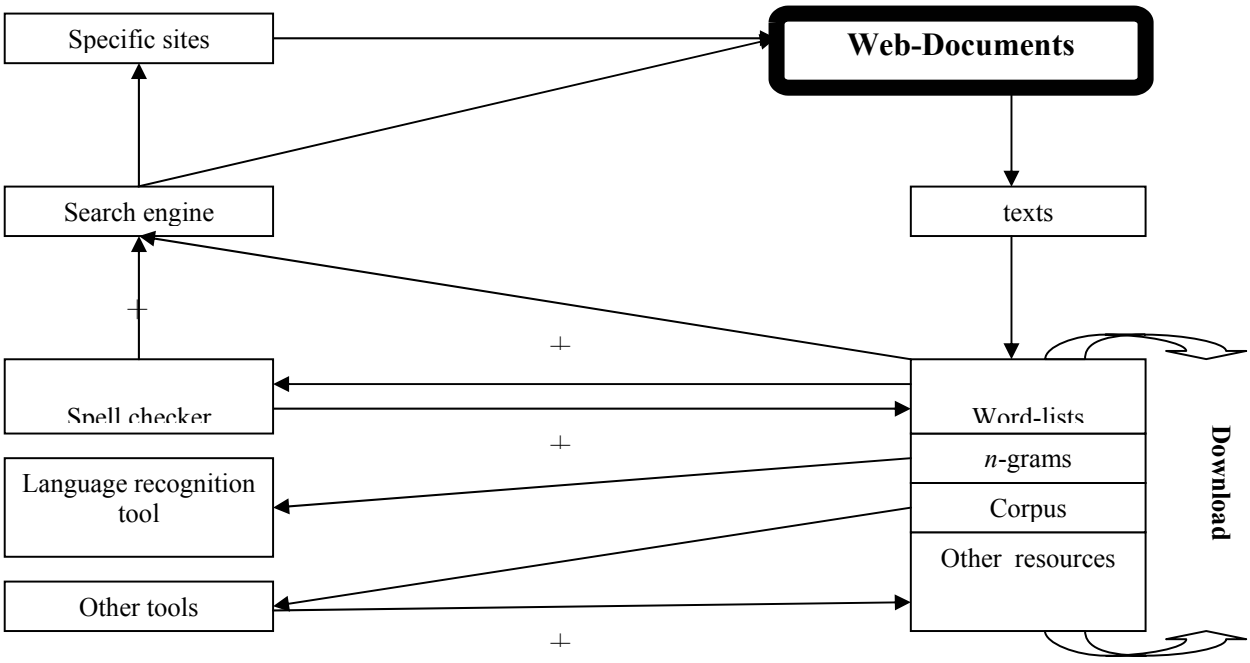


Figure 1: Manual bootstrapping from Web-documents with the help of simple NLP-applications. Tools, in addition, provide motivation to insert more and better data.

would raise the interest of linguists to engage in cooperation and to develop resources or tools. Using simple instructions, researchers can easily learn to manipulate this system. While currently the system requires a password, we will create a Wikipedia-like free system for the creation of language resources. Like In Wikipedia, numerous editors will be required to check the consistency and coherence of the data.

Furthermore, we also anticipate cooperation with projects that aim at providing resources in distinct African languages. Finally, we look forward to collaborating with those who also work in the preservation of written data to make the system more complete.

3. XNLRDF: Data Collection and Elaboration

Manually bootstrapping from word-lists and Web-documents to other Web-documents, we collect texts and derive resources and tools. Simple texts are converted into word-lists, n-grams; and if allowed by the copyright, preserved as corpus. Key words obtained from word-lists and spell-checker can be submitted into a search engine. The search engine then shows either specific multilingual sites that can be exploited or Web-documents which contains the key words. Figure 1 illustrates this.

Moreover, we will develop additional tools such as morphological analyzers and stemmers fo the writing systems with a minimum of resources. Furthermore, parallel corpora will be aligned to extract translation dictionaries and support attempts to apply Machine Translation to unresourced languages. The language recognition tool is directly used through our work, to

4. Results and Applications

Using this approach we collected textual examples for over 600 writing systems and 500 languages, including 150 African languages. Many of the collected texts are parallel texts, some covering about 200 languages. From the textual examples we manually derive rules for the tokenization (identification of words with a text), collect numerical expressions and function words.

Texts are compiled into word frequency lists. These are used for a primitive spell checkers. In addition we compile N-grams for a language recognition system.

4.1 Spell Checker

A simple spell-checker for 150 African languages is thus now accessible on-line at http://140.127.211.214/cgi-bin/spell/spell.pl (cf. Fig. 2). Using nothing but a corpus-derived word lists, the spell checker matches unknown words to a set of similar words using a general n-gram-based matching algorithm. We thus created a spell-checker working for all languages, after our attempts to compile our data into ISPELL or ASPELL failed.

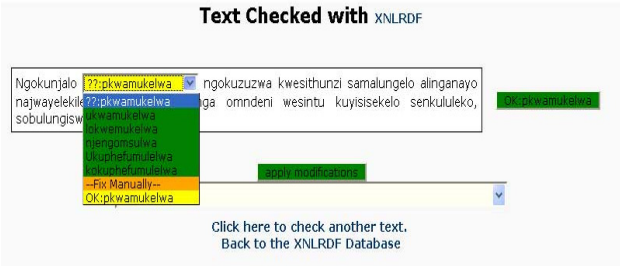


Figure 2: XNLRDF Spell Checker at work.

Although the functionality of the spell checker is limited, given the small size of the wordlist and the lack of a morphological analysis, we hope that the spell checkers might provide a motivation for interested users to cooperate in the collection and elaboration of data. In addition, the spell checker might help to enlarge the wordlist through its on-line usage. Every word not found in the wordlist but confirmed by the user will be stored and evaluated for inclusion in the database. The language recognizer and the word frequencies will work as primary filters before human experts might confirm the new data.

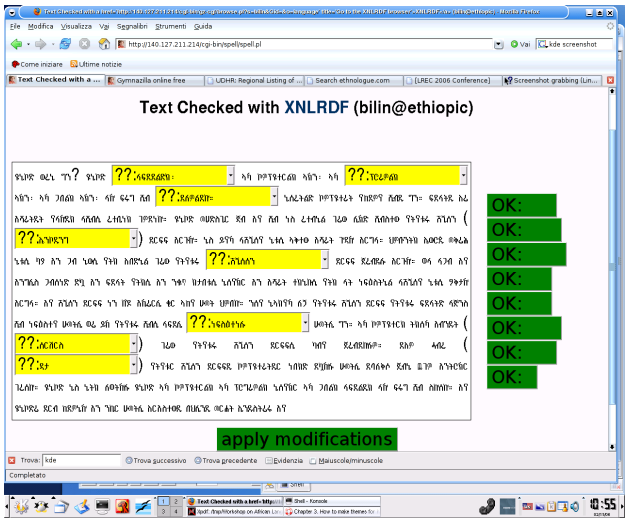


Figure 3: One of 150 simple spell checkers for African languages, here checking Bilin.

4.2 Language Recognition System

A language recognition tool is a system that recognizes the language of a text by comparing the text to a set of text for which the language is known (cf. Cavnar and Trenkle 1994). Such a tool is useful in all multilingual applications, such as information retrieval where queries and returned pages should have the same language. To improve its practical implication for the developers of XNLRDF however, we intend to extend the classification from known to unknown languages. By localizing all languages, with an without textual examples on the three dimensions *time, locality* and *language family*, unknown languages might be identified via the interpolation over this grid. The grid is currently under construction.

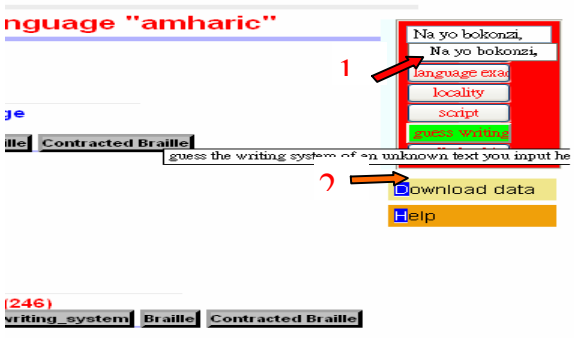


Figure 4: Input of texts in the XNLRDF interface to guess the language.

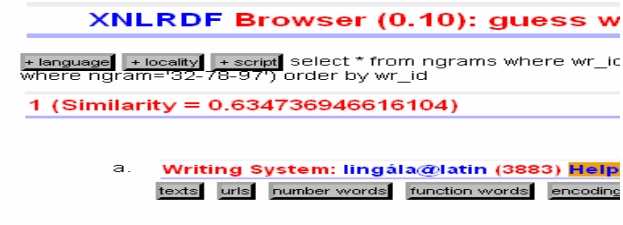


Figure 5: The output of the language recognition system.

5. Summary and Discussion

As we mentioned in the beginning, it is not an easy task to find electronic documents from which corpora or electronic dictionaries can be derived semi-automatically. Consequently, spell checker, language recognition system and other devices among XNLRDF were necessary for an infrastructure for the collaborative development of NLP resources for African languages if everyone's mother-tongue is to be fully functional in electronic environments.

Building corpora via the Internet, we encountered the following difficulties. First, some search engines might not support non-Latin scripts like Abugida. Second, documents exist only for a short time in the Internet. Third, when finding a document, we are unable to judge the language, especially if the language is not already included in our database and thus are unrecognizable to the language recognition system. Fourth, many documents are stored as images and texts cannot be extracted. Finally, only short sentences are found in some languages.

Websites with long duration that provide texts in different languages are efficient tools for us. Such websites may more or less solve the problems mentioned above, for example, we can save time on identifying languages. However, not all languages are covered. Specific websites are still necessary. Websites such as www.ethnologue.com helps us to distinguish languages of same origin, which are viewed as separate writing systems in terms of locality. In fact, people working on such job should always be aware of the differences between language varieties.

For those websites which store texts that as images, we may, in an arduous way, consider typing them with the aid of multilingual word, especially examples of this language are not used widely on the Internet.

This project requires both intrinsic and extrinsic collaboration. In order to fulfill it, we need a system that helps us to keep data coherent when different people collaborate on the database. Using a relational database we can use the internal checks for data-types, uniqueness, coherence and consistency at a level below the interface so that these checks are effective in all interactions with the database. The checks can can be defined to any level of complexity using triggers and functions. For example, changing the time period of a language in XNLRDF will change the time period for its antecedent and it following language as well (thus assuring the coherence). Organizing data into a network makes singular incorrect data modifications difficult or impossible. Declaring an ever growing number of data in this network as

unchangeable, will make the space for incorrect modifications smaller and smaller.

Creating ambiguous metadata becomes impossible through uniqueness constraint. References make it impossible to delete central data, e.g. a language referred to by a writing system. The inclusion of false positives, e.g. pejorative language names, marked as deleted make it impossible to insert or inherit the same value again through the effect of uniqueness constraints.

The hierarchy of collaborators in XNLRDF follows from this: While all users of XNLRDF will be in state to enter new data, a group of language expert or expert in linguistic subfields, language groups etc has in addition the power to 'delete' incorrect entries and thus move them to the false positives, or to assign the status of 'unchangeable' to cornerstone data. These experts thus complete and guide the set of control mechanisms provided by the system. A third group of language and database experts defines the constraints and inheritance mechanisms to account for the completeness and coherence of the data.

We thus hope that the project may raise the attention of students and researchers who are interested in the matter, and can further anticipate cooperation, with linguists of all flavours. Especially through the integration of simple applications, which among others test and show the potential of XNLRDF, we want to motivate researchers to enter the required data e.g. to insert open-licensed texts for a language to download shortly later a simple spell-checker, or to enter morphemes to download a better morphological analyzer. XNLRDF will also supply information of other aspects, such as ISO/DIS encoding of languages, multilingual naming of languages, language classification, numbers and function words.

6. References

Baroni, M. & Ueyama, M. (2004). Retrieving Japanese specialized terms and corpora from the World Wide Web. In *Proceedings of KONVENS* 2004.

Cavnar, W.B. & Trenkle, J.M. (1994). N-Gram-Based Text Categorization. In *Proceedings of Third Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, UNLV Publications/ Reprographics, pp. 161-175, 11-13, April 1994.

Prinsloo, D. & Heid, U. (2005). Creating word class tagged corpora for Northern Sotho by linguistically informed bootstrapping. In I. Ties (ed.) *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*, Bolzano/Bozen, Italy, October 2005.

Streiter, O. (2005). Implementing NLP-Projects for Small Languages: Instructions for Funding Bodies, Strategies for Developers. In I. Ties (ed.) *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*, Bolzano/Bozen, Italy, October 2005.

Streiter, O. & Stuflesser, M. (2005). XNLRDF, the Open Source Framework for Multilingual Computing In I. Ties (ed.) *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*, Bolzano/Bozen, Italy, October 2005.

Uchechukwu, Ch. (2005). The Igbo Language and Computer Linguistics: Problems and Prospects. In I. Ties (ed.) *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*, Bolzano/Bozen, Italy, October 2005.