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.