WALA: a multilingual resource repository for West African Languages

Dafydd Gibbon[1], **Firmin Ahoua**[2], **Eddy Gbéry**, **Eno-Abasi Urua**[3], **Moses Ekpenyong**[3]

[1]Fakultät für Linguistik und Literaturwissenschaft, Universität Bielefeld, Germany
gibbon@spectrum.uni-bielefeld.de
[2]Université de Cocody, Abidjan, Côte d'Ivoire
fahoua2003@yahoo.fr
[3]University of Uyo, Akwa Ibom State, Nigeria
ekpenyong_moses@yahoo.com, enourua@skannet.com

Abstract

The West African Language Archive (WALA) initiative has emerged from a number of concurrent projects, and aims to encourage local scholars to create high quality decentralised repositories documenting West African languages, and to make these repositories available to language communities, language planners, educationalists and scientists via an internet metadata portal such as OLAC (Open Language Archive Community). A wide range of criteria has to be met in designing and implementing this kind of archive. We discuss these criteria with reference to experiences in documentation work in three very different ongoing language documentation projects, on designing an encyclopaedia, on documenting an endangered language, and on creating a speech synthesiser. We pay special attention to the provision of metadata, a formal variety of catalogue or housekeeping information, without which resources are doomed to remain inaccessible.

1. Objectives

This contribution describes a number of experiences in documenting West African languages in Ivory Coast and Nigeria, and aims to promote the *WELD* (*Workable Efficient Language Documentation*) paradigm in which the following principles are practised in the documentation of local, in general unwritten, languages (Gibbon, 2002b):

Comprehensive: Language documentation must apply to all languages. But linguistic economy may dictate priorities may be hard to justify socially or politically: if one language is more similar to a well-documented language than another is, then surely the priority must be with the second.

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 magnify the potential of human resources available for the task. A model of this kind of development is provided by the Simputer,[1] and could easily be incorporated into European and US project funding.

State-of-the-art: In addition to using modern 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 from computational linguistics, human language technologies and artificial intelligence, for instance by the use of machine learning techniques for lexicon construction and grammar induction.[2]

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 dissemination methods: at third world Internet prices it can cost hundreds of Euros or indeed be impossible to download a large, modern software package, and net-based server registration and support is very costly, as is wireless data transfer.

Fair: If a language community shares its most valuable commodity, its language, with the rest of the world, then the human language engineering and computational linguistic communities must do likewise, with open source software and open data (simultaneously reaping the other well-known 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 references. 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 mineral and agricultural resources, medical ethnobotany and oil prospecting.

2. Motivation

Modern language and speech resource repositories have generally been motivated by the human language technologies, and are restricted to languages which have large-scale commercial interest, in particular a small set of European and East Asian languages, though there are exceptions (in particular for Bantu and Indian languages). This contribution describes the ongoing creation of a multilingual repository, currently designated the West African Language

[1] "Simple Computer" — handheld Community Digital Assistant (CDA) enterprise of the "Bangalore Seven" in India. See http://www.simputer.org/.

[2] The SIL organisation, for example, has a long history of application of computational linguistic methods (see www.sil.org).

Archive (WALA), for the documentation of West African local languages based on the quality standards of existing repositories in terms of their data and metadata specifications.

The repository creators are located mainly in francophone and anglophone West African countries (Ivory Coast and Nigeria), and required that the repository should include local resources from both areas and be accessible in both areas. The languages which have so far been and are currently being processed with the methodology described here are Iko and Ibibio (Nigeria), Abbey, Anyi, Baule, Ega (Ivory Coast).

The repository model is intended as a basis for adaptation to the creation of repositories for other local languages. Optimal conformance to EAGLES standards, the newer OLAC repository specifications, and portability specifications (Bird and Simons, 2003) was taken as the initial quality criterion for the repository. The motivations for the creation of this resource differ from those associated with languages typically (though not exclusively) stored in existing professional language resource repositories such as ELRA and LDC, and fall into four main categories:

1. Heritage documentation for an ethnic group and contribution to the group's own knowledge of its historical identity.
2. Provision of language materials for applications in local education and, increasingly, information technologies such as speech synthesis based agricultural and health information systems for pre-literate rural communities.
3. Creation of a high quality empirical basis for corpus driven linguistic research and teaching in local universities.
4. Training of local specialists in language documentation with a multiplier effect on the development of the repository.

These motivations relate to each of the projects from which this initiative sprang: design of an encyclopaedia of Ivory Coast languages, documentation of an endangered language (Ega, Ivory Coast), creation of a speech synthesiser for use in pre–literate rural communities (for Ibibio, Nigeria), and a joint Ivorian–Nigerian–German multilingual curriculum development project in language documentation.

3. Challenges

The repository creators had to face many kinds of challenging constraints in a number of the areas which had originally motivated the WELD principles described above in the introduction. These challenges are outlined in the following sections.

3.1. Discovery

Constraint: The collation of data *in loco*, in the study or in the laboratory, and the methods used in the design, creation and processing of resources should be adapted to the local research and elicitation environments.

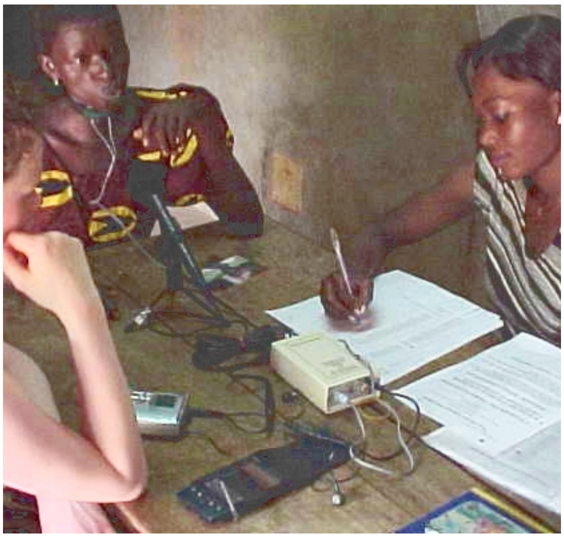


Figure 1: Practical interview–based Anyi phonetic documentation session with DAT recorder, laryngograph, Palm handheld metadata DB.

Situations vary from laboratory type experimental and interview data elicitation to the unobtrusive participant recording of multimodal communicative interaction while respecting ethical standards. Typically, classical laboratory and interview techniques are combined, as illustrated in Figure 1.

3.2. Workability

Constraint: Data acquisition, storage, dissemination and processing procedures should use the WELD principles.

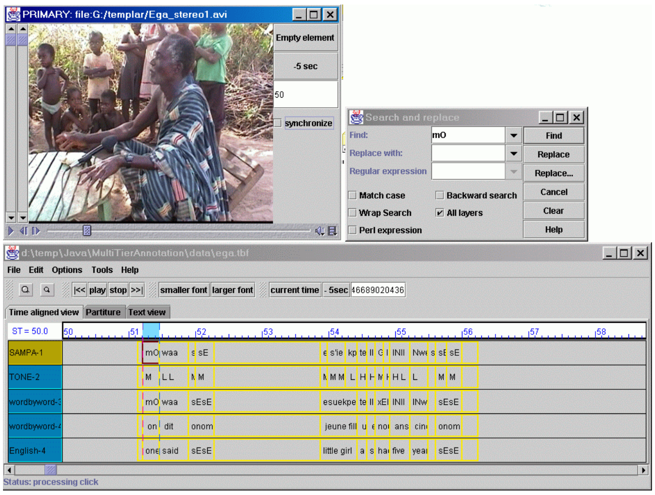
This means that these procedures should be genuinely available in local environments: comprehensive, efficient, and state of the art, but also affordable (not needing constant updating with the latest hardware or proprietary software) and fair (yielding results which flow back into local ethnic and academic communities).

Local conditions vary in terms of power supply reliability, and the cost and availability of internet connections, and a practical PC + modem mode of cooperation was developed, with the repository on a server in Europe and mirrored via CD-ROM on local machines.[3] Rather than using dedicated software, existing familiar office applications were used for the creation of basic database relations (specifically: well-defined table objects in word processors or spreadsheet tables) with export to textual character-separated value (CSV) database relation formats.

In order to generate human and machine usable documentation, the CSV formatted files are converted by means of tools written in suitable scripting languages into standardised formats with the aim of preserving the interpretability of the documentation for future users:

- XML archives with well-defined metadata (Gibbon et al.,);
- HTML for web display of database tables, lexicon, both directly and via XML or LaTeX and

[3] Currently www.spectrum.uni-bielefeld.de/langdoc/WALA/.



PRIMARY: file:G:/templar/Ega_stereo1.avi

50

Search and replace

Find: mO

Replace with:

Regular expression

□ Match case □ Backward search

□ Wrap Search ☑ All layers

□ Perl expression

d:\temp\Java\MultiTierAnnotation\data\ega.tbf

current time -5sec 46689020436

ST = 50.0

SAMPA-1	mO	waa	s	sE	e	s'ie	kp	te	ll	G	l	lNll	Nwe	s	sE	sE		
TONE-2	M	L L	M	M	M	M M	L	H	H	M	H	H L	L		M	M		
wordbyword-3	mO	waa	sEsE		esuekpe			te	ll	xEl	lNll	lNw		sEsE				
wordbyword-4	on	dit	onom		jeune fill			u	e	noi	ans	cin		onom				
English-4	one	said	sEsE		little girl			a	s	ha	five	year		sEsE				

Status: processing click

Figure 2: A video recorded Ega narrative session annotated with open-source software (TASX annotator).

latex2html;

- text format (LaTeX, RTF, ultimately XML + stylesheets) for print media.

3.3. Practicality

Constraint: Compromises in terms of familiarity of tools and degree of training should be made, in preference to the imposition of the standards of high-tech globalisation in language documentation (as in other areas).

Current 'ideal standards' such as XML and Unicode should be appraised relative to their practical value in the local context, and if necessary compromise formats used, with well-defined conversion functions (Gibbon et al.,).

Plain ASCII formats with minimal special markup, such as CSV files, were selected as the basic interchange formats. This necessitated adapting the EAGLES standard ASCII rendering of the IPA, X-SAMPA, to the languages concerned, including some simplifications to make manual data entry more efficient. Much of the available dictionary data used mixtures of proprietary fonts in unsystematic ways and had to be re-coded semi-automatically. Legacy material which had been visually tabulated with white space was re-structured in CSV formats. Some useful written material was re-typed or scanned and edited. For each of these cases, scripts to convert modified X-SAMPA to Unicode and CSV formats to XML were specified and implemented. Separate name spaces are used for the separate languages, and bilingual glossaries were created for francophone-anglophone accessibility.

This is not a plea for sticking with the simplest possible procedures; on the contrary, a development process needs to be bootstrapped. Techniques of this kind are indeed emerging, which permit, for example, sophisticated types of documentation at the level of annotated audio and video signals (see Figure 2).

3.4. Social convention

Constraint: The prime constraint on language resources lies in the language community which is responsible for the language.

Local priorities and ethical standards had to be discovered and taken into consideration. For example, with one of the languages dealt with, historical narratives are not permitted to be publicised outside the circle of elders, perhaps for practical reasons to do with inherited rights to territory or lack of them.

In consultation with local experts catalogues of text types were prepared. For the initial repository core it was sufficient to use standardised questionnaires such as the *West African Data Sheets*, developed for a survey of Ghanaian languages, which were used for interviews and manually transcribed, but also recorded on DAT tape. This traditional "laboratory genre" was supplemented by audio and video recordings of other interaction types (non-taboo narratives, riddles, salutations, artisan work, cookery).

3.5. Linguistic characterisation

Constraint: Minimal conventions for transcription, lexicon construction and grammar creation should be defined.

Data types for corpus definition needed to be adapted to conventional genres common in less technologically oriented descriptive linguistic communities, like interlinear texts, field notes, sketch grammars, core dictionaries.

In designing the repository, the creators started with a two–way distinction (Himmelmann, 1998) and made a heuristic four–way distinction between two documentary and two analytic levels of language characterisation:

Primary data documentation: recording, transcription, text collation, speech annotation, archiving.

Descriptive documentation: Sketch grammars (including tabular phonetic, phonological, orthographic, prosodic, morphological, syntactic description), and core dictionaries of about 1000 words.

Descriptive analysis: extensive linguistic studies of specific languages.

Theoretical analysis: detailed modelling of specific problems of language structure on typological or universalistic principles.

The creators restricted their attention to levels 1 and 2, the documentary levels, but included bibliographical material for levels 3 and 4, the analytic levels.

3.6. Consistency and interpretability

Constraint: The resulting data should have a higher quality than traditional "shoebox and desk drawer" field notes and typescripts to ensure their sustainability.

The data created for the archive should be transformed into a consistent resource (Trippel et al., 2004) which can be accessed reliably both by human searchers and automatic search devices. At the same time, the interpretability of the resource must be sustainable over time, both in terms of *language interpretability* (i.e. the association of recorded forms with meanings) and in terms of *archive interpretability* (i.e. the decodability of formats and accessibility of computational platforms) (Gibbon et al.,).

The genres of documentation are not restricted to purely electronic resources, however. The traditional channels of publication yield sustainable resources of their own, which

Table 1: Handheld fieldwork metadata DB specifications.

Attribute	Type
RecordID:	string
LANGname(s):	popup: Agni,Agni; Ega
SILcode:	popup: ANY; DIE
Affiliation:	string
Lect:	string
Country:	popup: Côte d'Ivoire
ISO:	popup: CI
Continent:	popup: Africa; AmericaCentral; AmericaNorth; AmericaSouth; Asia; Australasia; Europe
LangNote:	longstring
SESSION:	popup: FieldIndoor; FieldOutdoor; Interview; Laboratory
SessionDate:	pick
SessionTime:	pick
SessionLocale:	string
Domain:	popup: Phonetics; Phonology; Morphology; Lexicon; Syntax; Text; Discourse; Gesture; Music; Situation
Genre:	Artefacts; Ceremony; Dialogue; Experiment-Perception; ExperimentProduction; History; Interview; Joke/riddle; Narrative; Questionnaire; Task
Part/Sex/Age:	string
Interviewers:	string
Recordist:	string
Media:	popup: Airflow; AnalogAudio; AnalogAV; AnalogStill; AnalogVideo; DigitalVideo; DigitalAudio; DigitalAV; DigitalStill; DigitalVideo; Laryngograph; Memory; Paper
Equipment:	longstring
SessionNote:	longstring

can—modulo copyright!—be compatible with electronic channels (Connell et al., 2002).

3.7. Metadata

Constraint: Without identification of the available resource itself according to appropriate cataloguing criteria the resources would remain effectively invisible to language community and scientific community alike.

Standard sets of metadata categories for corpus and lexicon resources are slowly emerging, based originally on the Dublin Core library oriented set, and in the meantime augmented by the IMDI linguistic fieldwork oriented set and the OLAC (Open Language Archive Community) linguistic resource metadata portal.[4] There is no ideal set for language documentation, however. The application needs which define the *raison d'être* for metadata are varied, from fieldwork in inaccessible places to laboratory recordings for speech technology, and a pragmatic approach consequently should be taken (Gibbon, 2002a), necessitating opportunistic mapping from one category set to another, as with the set in Table 1, which has been applied in the WALA context in a large number of fieldwork sessions, using a Palm handheld database. Admittedly this runs counter to much contemporary wisdom, but then a straitjacket is not optimal equipment for facing the future.

4. Summary and conclusion

The repository will contain digital recordings and transcriptions for all the languages concerned, annotations for a small number of these recordings, sketch grammars and dictionaries for each of the languages, and bibliographies and mediographies of all available language-specific material. Currently, the conventions for OLAC cataloguing have been prototyped and tested, and metadata records are currently being entered into the OLAC distributed repository network. After further field testing, it is hoped that WALA will be seen as an attractive model for both local and global players in language documentation.

Currently available funding is but a drop in the ocean of the world's languages. But to conclude on a hopeful note: to attain the goals discussed in this paper, computational linguists speech and text engineers in wealthier situations will, ideally, 'adopt' local linguistics and computer science departments in areas of need, and materially support their documentation efforts. We can assure them that both sides will benefit enormously.

5. References

Bird, Steven and Gary Simons, 2003. Seven dimensions of portability for language documentation and description. *Language*, 79:557–582.

Connell, Bruce, Firmin Ahoua, and Dafydd Gibbon, 2002. Illustrations of the IPA: Ega. *Journal of the International Phonetic Association*, 32.1:99–104.

Gibbon, Dafydd, 2002a. Ubiquitous multilingual corpus management in computational fieldwork. In *Proc. LREC2002 Satellite Workshop on Portability*.

Gibbon, Dafydd, 2002b. The WELD Paradigm—Workable Efficient Language Documentation: a report and a vision. *ELSnews*, 11.3 Autumn:3–5.

Gibbon, Dafydd, Catherine Bow, Steven Bird, and Baden Hughes. Securing interpretability: the case of Ega language documentation.

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

Trippel, Thorsten, Dafydd Gibbon, and Felix Sasaki, 2004. Consistent storage of metadata in inference lexica: the metalex approach. In *Proc. LREC2004*. Paris: European Language Resources Association.

Dedication

We dedicate this paper to our colleague, co-author, friend and brother Eddy Aimé Gbéry, who died suddenly in our midst during the preparation of the paper. We remember him for many things, especially for his essential contributions to the WALA idea and to this paper, his leadership as Directeur du Département de Linguistique, Université de Cocody, Abidjan, and—not least—his infectious humour and unshakable loyalty. *Aimé, tu nous manques.*

Acknowledgment

This work is co–funded by the German Academic Exchange Service (DAAD) grant *M.A. in Documentation of Local Languages* and the Universitites of Bielefeld (Germany), Cocody (Abidjan, Côte D'Ivoire) and Uyo (Akwa Ibom State, Nigeria), and a LLSTI scholarship.

[4] Cf. www.language-archives.org/.