Workable Efficient Language Documentation: a Report and a Vision

Dafydd Gibbon, Universität Bielefeld, Europe

DRAFT 3 (FINAL), 1 October 2002

Documentation projects: the WELD paradigm

In a number of projects since around 1997, one funded by the *Deutscher Akademischer Austauschdienst*, another by the *Deutsche Forschungsgemeinschaft* and one by the *Volkswagenstiftung*, the Computation and Spoken Language group at Universität Bielefeld, Germany, has been developing efficient techniques for language documentation, and applying these to West African languages in pursuit of the "Workable Efficient Language Documentation" (WELD) paradigm. The techiques range from hypertext formats for the presentation of lexical and grammatical information through signal annotation of audio and video recordings of speech in context, and an annotation−based audio concordance, to palmtop metadata and questionnaire administration software for use in adverse environments, e.g. in tropical rural and forest villages (see Figures 1 and 2).

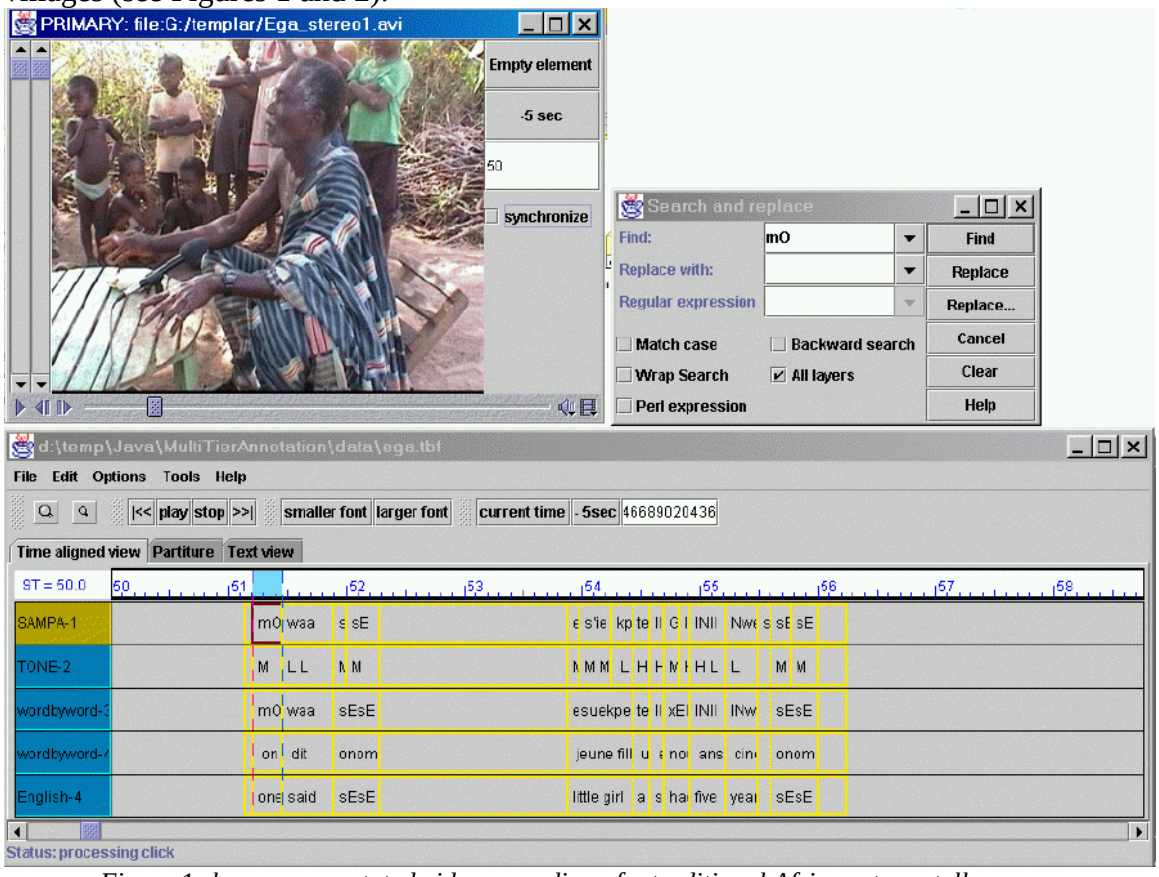

Figure 1 shows an annotated video recording of a traditional African story−teller, made with open source (GPL) signal annotation software TASX developed by Jan−Torsten Milde at Universität Bielefeld (see <`http://tasxforce.lili.uni-bielefeld.de/`>).

Figure 2 shows a typical fieldwork situation in Ivory Coast using portable DAT recorder and laryngograph, and metadata logging with a Palm Pilot database application.

Sample documentation of the endangered Ega language (also Ivory Coast) is given at <http://www.spectrum.uni-bielefeld.de/LangDoc/EGA/>.

In order to disseminate further information about projects and ongoing work in the field, and to provide a showcase of best practice in endangered language documentation, the E-MELD portal has been established by the Linguist List (see <http://saussure.linguistlist.org/cfdocs/emeld/>).

Criteria for language documentation

But why "language documentation" and, in particular, "efficient" and "workable" language documentation? The goal of documenting all the world's languages ist just as important as documenting diversity in the biosphere or in geology: independently of the cultures associated with languages, each of the world's 7000 currently catalogued languages is a community-created abstract work of art in itself; it may be compared to a complex crystal, in that subtle fractal variations in structure distinguish it from all other languages and offer insights into unknown areas of the human cognitive potential. In the linguistic fieldwork community a very simple and practically motivated heuristic distinction is now made between *documentation* of a language and its *description*. Briefly, the former constitutes the empirical foundations (recordings, "sketch grammar" with basic phonemic, morphological and grammatical analyses, and a basic lexicon) for a linguistic description, while the latter follows the patterns of scientific theory formation.

From a more sophisticated perspective, empirical foundations of this kind necessarily interact in complex ways with theoretical assumptions, but there are urgent reasons for making compromises here and sticking to a simple distinction in the interests of efficiency and workability.

The urgency of language documentation

One of the urgent reasons for efficiency is that fieldwork on undescribed languages spoken, for instance, in rural or forest tropical areas or other relatively hostile climates requires complex logistics before the actual work can even start, therefore time is of the essence, and a high quality (thus time−consuming) linguistic description cannot always be made on location. Another consideration is that endangered languages are not likely to be with us for very long, and of the 7000 currently catalogued languages of the world hundreds (several thousand in the medium and long term) fall into this category (see <http://www.ethnologue.org>). High quality and efficient documentation is crucial for endangered languages, because when a language is dead there is evidently no hope of collecting additional data and documentation must serve the purposes of science and the descendants of the extinct community *in the permanent absence of native speakers*.

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. A vision needs to be developed for involving wealthy language engineering and computational linguistic communities and for spreading the idea of Workable Efficient Language Documentation beyond these communities to poorly equipped local scientific communities around the world with old computers, unstable electricity supplies, extremely 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.

Towards a WELD Charter: a vision for language documentation

A Charter for the WELD paradigm would include at least the following five benchmark principles of *comprehensiveness*, *efficiency*, s*tate of the art*, *affordability* and *fairness*:

1. *Language documentation must be 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. *Language documentation must be 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 easily be incorporated into Eurpean and US project funding.

3. *Language documentation must be 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. The SIL organisation, for example, has a long history of application of advanced computational linguistic methodologies (see <www.sil.org>), and more research is needed here.
4. *Language documentation must be 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 inexpensive dissemination methods − at third world Internet prices, it can cost hundreds of Euros to download a large, modern software package (not counting landline interruptions), and net−based registration and support is unthinkably costly, as is wireless data transfer.
5. *Language documentation must be 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, 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 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 biology and geology, for example in medical ethnobotany and oil prospecting.

Outlook

Naturally, things are not so simple in real life. Some of the principles in the WELD Charter outline may well be in conflict in some situations, requiring careful cost−benefit analysis. And there are in fact communities, fortunately not too common, who would be shocked at the thought of sharing their language with outsiders, just as there are of course R&D communities, unfortunately far more common, who would be shocked at the thought of sharing their resources with outsiders even in a context such as WELD. Intellectual property rights must be taken very seriously, of course, and the issues are far from simple. But the good news is that the dominance of these attitudes is slowly being replaced by WELD−friendly Open Archive, Open Resource, Open Source, and Open Data initiatives (just check the web here!), and that these are gradually being taken up by funding agencies as hallmarks of quality.