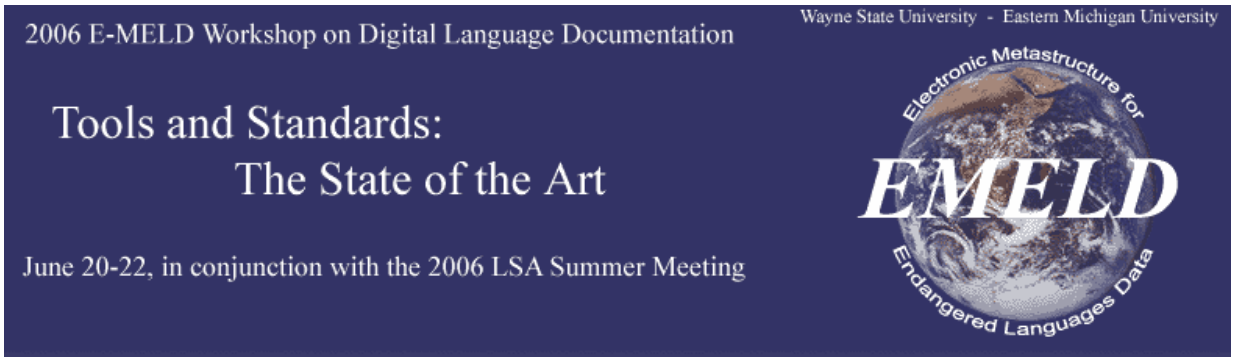


FIELDWORK COMPUTING: PDA APPLICATIONS

By

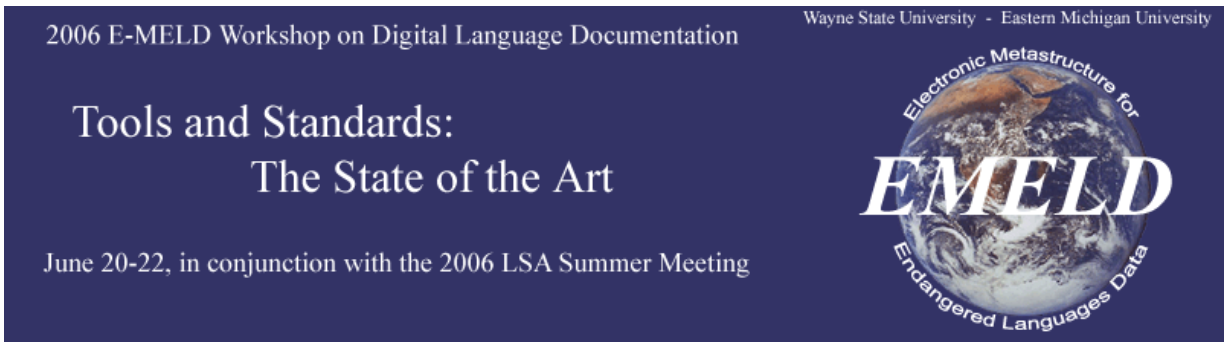
Dafydd Gibbon

Paper presented at

2006 E-MELD Workshop on Digital Language Documentation
Lansing, MI.
June 20-22, 2006

Please cite this paper as:

Gibbon, D. (2006), Fieldwork computing: PDA applications, *in* 'Proceedings of the EMELD'06 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art'. Lansing, MI. June 20-22, 2006.



Fieldwork computing: PDA applications

Dafydd Gibbon, Universität Bielefeld, Germany

1 Objectives and motivation

The objective of the present contribution[1] is to address the issue of portability in the context of linguistic fieldwork, both in the sense of platform interoperability and in the specific sense of ultra-mobility as required in typical fieldwork situations and in relation to the potential provided by handheld PDAs. Based on an initial statement of specifications, a three-layer networked architecture, the UbiCorpus model, for information gathering in the field is described, with the following layers:

1. Resource Archive distributed server layer,
2. Data Processing application client and local server layer,
3. Fieldwork Acquisition layer ("Corpus Pilot" layer) designed to provide PDA support for certain types of fieldwork session under adverse conditions, mainly for on-site questionnaire presentation and metadata acquisition, with database applications and custom applications designed and implemented by the author.

The perspective taken is partly that of a *post hoc* use case evaluation report (towards the end of this contribution), and partly that of the development of detailed specifications for further development based on use case experience. The evaluation leads to a set of requirements specifications for linguistic fieldwork, a subset of which can be fulfilled by a PDA rather than paper and pen tools or a laptop. The use case concerned is a metadata acquisition database application for Palm PDAs which has been used for some 8 years in fieldwork in West Africa; other use cases (e.g. interview prompting) are not discussed in detail.

Like other empirical and applied scientific disciplines, linguistic investigations involve fieldwork, on the one hand to discover realistic, natural, i.e. ecologically sound data, and on the other to conduct field trials in order to evaluate results of both theoretical claims and developments in speech and text technology. Linguistic fieldwork is situated along a scale from experimental laboratory data production through various kinds of data elicitation to 'donnees trouvees' (i.e. data in the naïve etymological sense of 'given'). Fieldwork requires tools, starting with the linguist's mouth, hands and memory, extended by instruments such as pen, paper, and recording and computing tools. The suitability of these instruments is conditioned by many different factors.

A convenient systematisation of the factors involved in computational fieldwork tool design and deployment derives from the fact that fieldwork, as a variety of metalinguistic scientific activity, can be coarsely parametrised, like other varieties of such activity, along three presumably rather uncontroversial dimensions:

1 This contribution is an extension of contributions presented at LREC 2002 and LREC 2004. Detailed discussions with Sandrine Adouakou, Firmin Ahoua, Doris Bleiching, Nadine Borchardt, Bruce Connell, the late Eddy Aimé Gbery, Ulrike Gut, Ben Hell, Sophie Salffner, Thorsten Trippel and Eno-Abasi Urua are gratefully acknowledged.

1. *topic domain* (i.e. the linguistic domain of interest, ranging from phonetics through to discourse activities);
2. *formal methods* (i.e. the explicit or implicit functional and formal descriptive and explanatory paradigms according to which elicited data are generalised);
3. *empirical methods* (i.e. qualitative judgments, quantitative measurements, and procedures for both qualitative and quantitative generalisations, as well as situational participant, observer and tool constellations).

Example 1: A well-known speech signal processing tool such as *Praat* can be characterised by topic domain (acoustic phonetics), formal methods (e.g. formal signal and symbol processing) and empirical methods (e.g. support of quantitative signal analysis and of data categorisation by transcription and annotation).

Example 2: A well-known lexicographic workbench tool such as *Toolbox* can be characterised by topic domain (tabulation of the properties of words and other lexical expressions), formal methods (relational database construction; alignment of parallel information) and empirical methods (corpus analysis).

Other tools such as morphological analysers, part-of-speech taggers, machine learning classification and grammar learning systems, can be similarly classified.

The three dimensions of topic domain, formal methods and empirical methods have changed dramatically in the past decade:

1. Along the topic domain dimension, many changes have taken place, mainly due to an increased interest in issues of language in context, i.e. pragmatics, sociolinguistics, discourse analysis, linguistic anthropology. Some of these changes have been due to discipline-immanent developments, but many have been driven by applications needs in language planning, language education, and language engineering. Computational tools have also opened up new topics (for example in prosodic analysis) which were not previously empirically accessible.
2. Along the formal methods dimension, progress in theoretical linguistics and in fields such as linguistic anthropology and discourse theory has led to a sharpening of categories in linguistic typology and in the fieldwork on which linguistic typology is based. Computational tools, embodying results of formal linguistic modelling are widely used, especially in phonetics and lexicography, but also in more specific areas such taggers and parsers for morphological and syntactic analysis.
3. Along the empirical methods dimension, new corpus-oriented and quantitative methods have arisen, partly due to new portable methods of audiovisual data recording, partly due to conveniently portable computing devices, partly due to the availability of audiovisual experiment and elicitation control software, and to phonetic and linguistic analysis software for the annotation of signal and text corpora. The empirical dimension also benefits from formal archiving and search techniques, and the development of taxonomic instruments in the form of ontologies such as GOLD (General Ontology for Linguistic Description), which have contributed to the utility of formalised markup languages such as XML in providing a strategy for going beyond the dozens of existing mutually incompatible XML codings to a more integrated classificatory framework.

The concepts of topic, formal methods and empirical methods will be used to define an evaluation space for characterising the deployment of tools, including the use of PDAs.

A major contribution to the development of quality standards for linguistic fieldwork was made by scenario design and data collection efforts for the development of language and speech technologies, such as speech-to-speech translation support systems and or dictation software. For these systems, field data are required for statistical processing in system development, and field trials of system performance are needed for evaluation (Gibbon, Mertins & Moore 2000). Two significant formal quality criteria have emerged in this area which will be applied here to linguistic fieldwork:

1. *completeness* (comprehensiveness, broad coverage);

2. *soundness* (correctness, precision, consistency).

These two criteria are in general used to evaluate the interpretation of a formal system with respect to a formal model, but by analogy the criteria can be extended with respect to empirical models to criteria of *empirical completeness* (coverage) and *empirical soundness* (truth), as in the experimental criteria of *recall* (completeness – degree of coverage of predictions) and *precision* (soundness – avoidance of prediction of junk). In a sense, these criteria have always been applied to empirical linguistic work, but completeness and soundness have often been difficult to assess in linguistics because of the prevalence of informal textual formulations.

The formal and empirical criteria are not easy to fulfil, of course, particularly where complex theories or large quantities of data are involved. Consequently, operational computational models are increasingly being used to check the consistency and precision of linguistic descriptions. More extensive application of this technique is becoming feasible with the introduction of high quality formalised archives as well-defined databases whose construction is supported by dedicated tools, for example as hierarchically pre-structured XML databases (enhanced with more complex cross-hierarchy constraints for tables, multi-tier lattices and inter-document links).

In addition to the three linguistic criteria of topic domain, formal methods and empirical methods, and to the two general formal criteria of completeness and soundness, there are further operational, pragmatic criteria for the development of fieldwork tools to be considered.

Formal methods are essentially concerned with what may be called the *syntax* of science, and empirical methods coupled with the topic domain are essentially concerned with the *semantics* of science. However, all three dimensions are also connected with the *pragmatics* of science, in the following senses:

1. *Conventionality* at various levels, including:
    1. *Kuhnian paradigms*, in which communities of scholars determine what is correct (both in the sense of "best practice" and in the sense of "politically correct"!), and therefore fundable, and what is not correct.
    2. *Clashes of civilisations* between the conventions of the economic systems of affluent countries and the subsistence level systems of areas where linguistic fieldwork is typically practised.
2. *Portability* issues, bearing in mind that the term is systematically ambiguous:
    1. *Interoperability* of applications on different OS and hardware platforms.
    2. *Compatibility* of data formats through import and export filters for functionally equivalent or interfaced applications.
    3. *Ubiquity*, i.e. time and place independent mobile deployment, primary focus here, in respect of:
        1. *logistical problems* of travel and equipment transport;
        2. *ergonomics* of equipment use, e.g. touchscreen, inconspicuous and non-distracting small size;
        3. *re-usability*, in the senses of
            1. interoperability of hardware, software and data formats;
            2. interpretability of data (Gibbon, Bow, Hughes & Bird 2004) in terms of both semiotic content (media files are not enough and are mainly of journalistic rather than scientific interest if the media data are not mapped to a functional level such as glossing).
    4. *WELD (Workable Efficient Language Documentation) criteria* (Gibbon 2002, 2003, 2006), (which 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), which postulate that language documentation must 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 pragmatic criteria will also be important in assessing the potential role of PDAs in providing computational fieldwork support.

2 Requirements specification

In this and the following sections, classic software engineering concepts will be employed, distinguishing between requirements specification, software design, and software implementation. There follows a discussion and evaluation of a use case involving metadata collation.

The requirement of portability (in many senses of the term) has become very important in recent years (Bird & Simons 2002). In linguistic fieldwork, conceptually the initial stage in any language documentation procedure, the issue of portability is important in two senses: first, in the common interpretation in terms of *platform interoperability* and second, in the meaning of *ultra-mobility*. Both issues are addressed here.

Likewise, requirements pertaining to corpus standards and resources based on criteria such as those outlined previously have become increasingly important as benchmarks for speech technology

systems (Gibbon, Moore & Winski 1997; Gibbon, Mertins & Moore, 2000). Similar criteria were proposed by Bird & Simons (2002) for areas directly relevant to typical linguistic fieldwork, and new issues such as the role of standardised metadata in resource archiving and reusability have come to the fore. All this has added to the demands on the fieldworker to produce high quality and interoperable documentation.

The linguistic requirements for fieldwork documentation (Himmelmann 1998) traditionally specify the following components:

1. field notes,
2. outline of the situation of the language,
3. transcriptions,
4. sketch grammar of basic phonology, morphology, and syntax,
5. lexicon with glosses and examples, perhaps also with other information,
6. word-field based thesaurus (perhaps).

Computational support for certain aspects of linguistic fieldwork has been available for many years, of course, both for laptop-based data entry and initial analysis on the move or in isolated areas, and for desktop-based detailed descriptive work and document production (with increasing overlap between laptop and desktop functionalities). Software applications have been characteristically in the following areas:

1. Lexical databases, either using general office DBMS such as FileMakerPro and MS-Access, or custom lexicon project software such as SIL's Shoebox; the latter also includes lexical support for textual glossing.
2. Publication support such as DB export functions, fonts.
3. Phonetic software, for signal analysis (e.g. general signal editors such as CoolEdit, or SIL's signal analysis packages), and for the symbol-signal time alignment (labelling) of digital recordings (e.g. Praat, Transcriber).
4. Computational linguistic software for basic phonological, morphological and syntactic processing.

The prompt materials for eliciting much of the documentation generated with these types of software are mainly systematic linguistic and ethnographic questionnaires, and the media for production of the documentation are often office-oriented software such as word processors (MS-Word etc.), DBMS (Access, FilemakerPro etc.), and spreadsheets (Excel, etc., also used for database entry). The guiding objectives of this concept of documentation are applications in what might be called an *educational paradigm* of production of translations, terminologies, and alphabetisation materials for use with the source communities. Presentation of prompt materials, as well as metadata collation, is a prime candidate for PDA applications.

The newer *multimedia paradigm* focusses more on linguistic anthropology and related disciplines, and are not necessarily focussed on these educational issues; their motivation oscillates between general scientific interest, provision of heritage documentation for descendants of the source communities, and journalistic collection of picturesque audiovisual material. Screenplay overviews and metadata collation are again prime candidates for PDA applications.

In order to address the issues specified here, a generic three-layer networked architecture, the UbiCorpus model, for information gathering in the field is specified, with the following layers:

1. *Resource Archive Server* layer, typically non-mobile, and distributed;
2. *Data Processing Application Client/Server* layer, typically a local laptop or desktop;
3. *Fieldwork Acquisition Client* ('Corpus Pilot') layer, designed to support specific fieldwork sessions under adverse conditions with questionnaire presentation and metadata editing, and typically implemented on a handheld PDA.

The Fieldwork Acquisition layer is focussed on in this contribution. One of the main advantages of the Fieldwork Acquisition layer, when implemented on a modern handheld device, is that it provides a convenient, inexpensive, efficient and (important in many fieldwork situations) inconspicuous method for the frequently neglected task of systematic on-site metadata logging. However, the scope of the model is more general. The model supports both the documentation of

spoken language corpora in general, whether in linguistics or speech technology, and also permits systematic incorporation of further comptational corpus processing in the form of computational linguistic methods in lexicography (van Eynde & Gibbon 2000), and in computationally supported grammar testing.

Some of this functionality (lexical databases, document production, computational linguistic processing) overlaps with the Fieldwork Acquisition layer, but the Fieldwork Acquisition layer has the following characteristic additional fieldwork corpus acquisition functionality in common with best practice in speech technology (Gibbon, Moore & Winski 1997; Gibbon, Mertins & Moore 2000), in supporting all phases of data acquisition in various ways:

Pre-recording phase: planning of the overall corpus structure and contents, in particular design of corpus recording sessions, including the preparation of *a priori* scenario metadata, interview strategies, questionnaires, data prompts (for instance with prompt randomisation),

Recording phase: conduct of corpus recording sessions, including session management with the logging of non-*a priori* metadata in a metadata editor and database, questionnaire consultation and data prompt presentation;

Post-recording phase: provision of recorded and logged data for archiving and processing in interoperable formats, including metadata export, transcription, lexicon development, systematic sketch grammar support and document production.

3 Software design

The language documentation model within which the UbiCorpus model is deployed is visualised in Figure 1. The two components of the model with which the UbiCorpus tools are concerned are the *Creation* and *Archiving* component, and the *Fieldwork* information source, which is directly associated with the Corpus Pilot layer described below. The three layers of the UbiCorpus model are characterised in the following subsections.

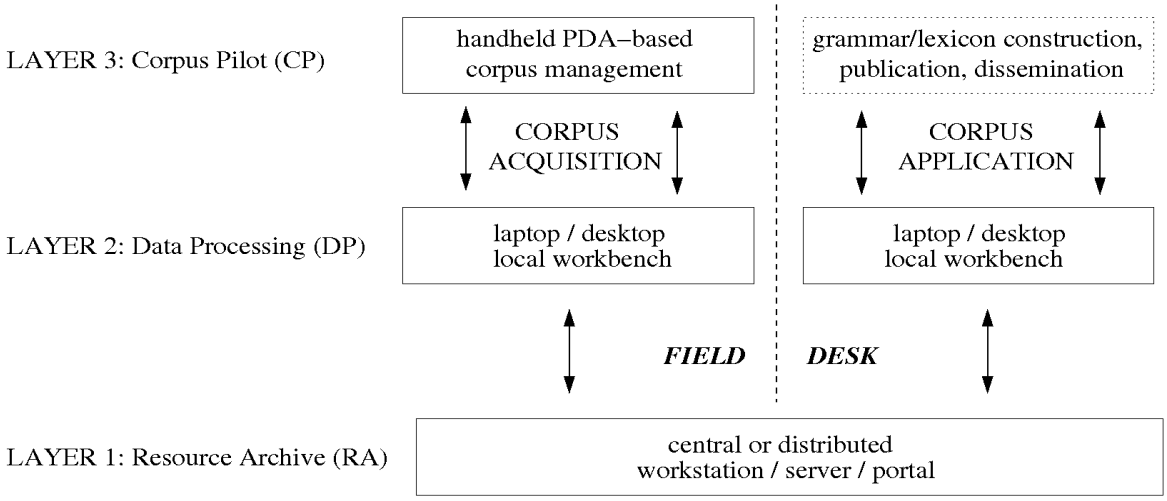


Figure 1: UbiCorpus three layer model of PDA deployment in the field

3.1 Resource Archive Server layer

The Resource Archive UbiCorpus layer represents the archive database and the access and media dissemination functions associated with it. On the declarative side, a number of current language resource and documentation proposals may be assigned to the Resource Archive layer: a single resource database such as a corpus or a lexicon, a multiple resource database such as a browsable corpus or concordance system, a web portal constituting a large and systematic resource world, or an entire dissemination agency. On the procedural side, the Resource Archive layer provides search

functions of various kinds, from standard browsing strategies to intelligent search and concordance construction, with token renderings of resources in any suitable media, whether entire corpora or lexica.

3.2 Data Processing Client/Server layer

The Data Processing layer is the layer which is most familiar to the ``ordinary working linguist''. The data include paper fieldwork logbooks, transcriptions, sketch grammars and card index lexica; word processor and database versions of these; analog and digital audio and video recordings; time aligned digital annotations of recordings, and concordance or browsing software based on annotations; metadata catalogues for all of these Data Processing layer data types. Procedurally, the platforms and applications used at the Data Processing layer are very varied, though there is a tendency to go for platform independence and standardised data interchange formats. By using modern laptops, both the Resource Archive and Data Processing layers can be integrated into a single mobile environment.

3.3 Fieldwork Acquisition Client layer

The top layer of the model represents the functionality which needs to be available in an actual fieldwork situation. This functionality can be very varied, and much - especially free format interviews and film recording - lies outside the range of systematic computational support. However, the following on-site support features can easily be covered:

1. Structured or free format questionnaire presentation (either from a database or in free text format, the latter using a plain text editor or with special formatting and rendering, for example by means of an IPA font, in a PDA word-processor).
2. Transcription (either in plain ASCII IPA encoding such as X-SAMPA - preferred for computational linguistic processing - or using an IPA font, preferably Unicode, though the latter is generally deprecated for general speech engineering use).
3. Metadata editor and database.
4. Participant database for interviewers, interviewees, technicians, etc.
5. Storage of media recordings (on recent microdisk PDA devices).

3.4 Interfaces

Except for media data, the interfaces between the three layers, and modules within the layers, are defined mainly on the basis of generic ASCII formats, including XML annotated text, CSV database tables, and richly formatted documents (including IPA font information). For the interface between a handheld implementation of the Fieldwork Acquisition layer and the Data Processing layer, conversion scripts are provided as required, in order to export handheld database and text formats into generic ASCII formats. Data transfer at the implementation level is via the usual synchronisation functions provided with handheld devices, or via scp, http, and ftp procotols for laptops, desktops and server.

4 Software implementation

4.1 Resource Archive Server layer

The server archive provides web portal access for the local and global linguistic communities, CD-ROM access for the local linguistic community, and analogue selections (in general, tape cassette, print media) for practical applications in the local user community. Currently, the leading models for the Resource Archive level are provided by the LDC and ELRA dissemination agencies; the E-MELD project is developing a general model for best practice in resource collation, and a meta--portal for flexible access to language resources. The local server currently used for initial database collation contains a number of specific search functionalities for corpus analysis, in particular an audio concordance \cite{gibbontrippel2002}.

4.2 Data Processing Client/Server layer

The classical environment for fieldwork data processing is a laptop, often a Mac, but also very frequently an Intel based device configured alternatively with Linux or MS based portable standard software. The kinds of application typically used are for basic corpus processing: Transcriber and Praat for transcription and annotation; Shoebox for lexical database development; MS Office or StarOffice for word processor, database and spreadsheet applications. These may be augmented with custom applications in Java (cf. the TASX engine (Milde & Gut 2001) and Perl (PAX audio concordance, Gibbon & Trippel 2002).

4.3 Fieldwork Acquisition Client layer

The Fieldwork Acquisition layer is implemented as custom-developed Palm compatible PDA applications. The rationale behind the use of the PalmOS based handhelds, as opposed to the use of a laptop, is based on the following considerations:

1. extremely inexpensive (in relation to other computational equipment),
2. ultra-lightweight (lighter than other standard portable fieldwork equipment such as field laryngograph, DAT recorder),
3. long battery charge operating cycle, depending on model,
4. fast and highly ergonomic in use,
5. small and unobtrusive in the interview situation,
6. an integrated environment with other PDA functionalities such as calendar, diary, address and other databases, other custom applications in C and Scheme.

4.4 Networking

The three levels are networked by standard techniques: server-client communication in general via TCP/IP-based protocols (via DSL, Ethernet, WLAN) and mobile or landline telephone. The applications-to-acquisition via dedicated sychronisation software of the kind typically used to link handheld PDAs to desktop installations.

4.5 Evaluation considerations

The general criterion of portability (specifically: ultra-mobility), with its detailed ramifications as previously discussed, is applied to the implementation in respect of several applications. The satisfaction of these criteria points towards a high level of suitability for use in extreme fieldwork situations without power supplies, for instance in isolated outdoor locations (forest, village, etc.). The functionality which has been included in the Corpus Pilot layer so far covers the following:

1. *Metadata editor and catalogue database*. Audio/video recordings, photos, paper notes, artefact require complete and efficient cataloguing. This application is based on a widely used PalmOS DBMS application, HanDBase, which provides a wide range of input support facilities (popups, date picker, free format notes, etc.), as well as cross-table linking.
2. *Questionnaire administration*. In general, free text format has been used for questionnaire administration, and responses have been recorded for later out-of-field processing. For some questionnaire types (e.g. demographic information), the HanDBase DBMS is used.
3. *Lexicon development tools*. Three applications are used for lexical database input (excluding freely formatted notes):
    1. an Excel-compatible spreadsheet, QuickSheet, which permits export in either CSV or Excel format (Excel is widely used in field linguistics as a convenient input tool for lexical databases, because of the ease with which databases may be constructed and restructed, and because it has many database-like functions, as well as built-in arithmetic functions if required for corpus work);
    2. the HanDBase DBMS which is also used for the metadata editor database;
    3. an implementation of the DATR lexicon knowledge representation language in LispMe, a Scheme implementation for the PalmOS platform (this application is a more Data Processing layer oriented tool, but is included in the Corpus Pilot layer implementation suite for convenience).

4. *Transcription support*. In general, transcription in X-SAMPA (Gibbon, Mertins, Moore 2000) is used, but if required, IPA fonts may be used with handheld word processors for PalmOS devices such as WordSmith and QuickOffice.
5. *Statistics package for initial evaluations*. This is also a more Data Processing layer application, but integrated into the Fieldwork Acquisition layer environment. Functions include all the measures used in basic experimental and corpus work (including random sorting, mean, median, standard deviation, standard error, as well as standard pairwise comparison measures).
6. *Context-free parser package for basic grammar development*. This is another Data Processing layer application, which is integrated into the Fieldwork Acquisition layer environment because of the convenience of the LispMe Scheme application for PalmOS in which the parser suite is implemented.

The last two applications were implemented by the present author, and are freely available from the author's download website. The fulfilment of the criterion is discussed with reference to the use case of metadata editing and storage in the following section.

5 Use case: Metadata editor and database application

5.1 Field application requirements specification

The metadata application has been selected for detailed description, because it is most immediately relevant to the issue of language resources.

A metadata editor for audio/video recordings, photos, paper notes and artefact cataloguing was required, based on a standard PalmOS relational database shell (HanDBase). The metadata editor provides a fast and inconspicuous input method for structured metadata for recordings and other field documentation, based on current work on metadata in the ISLE, E-MELD projects, and in the pilot phase of the DOBES project.

5.2 Design

The metadata specifications required for the types of fieldwork which were carried out are detailed in the following table.

Fieldwork metadata specifications for PDA metadata database.	
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 widget
SessionTime:	pick widget
SessionLocale:	string
Domain:	popup: Phonetics; Phonology; Morphology; Lexicon; Syntax; Text; Discourse; Gesture; Music; Situation
Genre:	popup: Artefacts; Ceremony; Dialogue; ExperimentPerception; ExperimentProduction;

Fieldwork metadata specifications for PDA metadata database.	
	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

For the work in hand, standardised metadata specifications, such as the Dublin Core and IMDI sets, were taken into account. However, new resource types such as those which are characteristic of linguistic fieldwork demonstrate that the standards are still very much under development, since some of the standard metadata types are not relevant for the fieldwork data, and the fieldwork data types contain information not usually specified in metadata sets, but which are common in the characterisation of spoken language resource databases (Gibbon, Moore & Winski 1997). In respect of the fieldwork resource type, it appears that it cannot be expected that a truly universal - or at least consensual - set of corpus metadata specifications will be developed in the near future, or perhaps at all, at a significant level of granularity. It may be possible to constrain the attribute list, though the existence of many different fieldwork questionnaire types belies this. However, the values of the attributes are in general unpredictable, covering not only free string types but also new rendering types (e.g. different alphabets; scanned signatures of approval).

Indeed, it may be noted in passing that the expectation of fully standardising the entire metadata specification tends to reveal singularly little awareness of the potential of machine learning and text mining procedures for handling generalisation tasks of this kind. It may be predicted that such procedures will be applied in future not only to extensive resource data sets but also to increasingly extensive sets of metadata. In consequence, the metadata specifications used in the UbiCorpus applications are deliberately opportunistic, in the sense that they are task-specific and freely extensible.

A selection of attributes and values for the current fieldwork application are shown below. Metadata attributes concerned with the Resource Archive layer of archiving and property rights are omitted.

CorpusMetaData	**02-3-12:HanDBase Table Export**
RecordID:	Agni2002a
LANGname(s):	Agni, Anyi
SILcode:	ANY
Affiliation:	Kwa/Tano
Lect:	Indénié, Ndenye
Country:	Côte d'Ivoire
ISO:	CI
Continent:	Africa
LangNote:	
SESSION:	FieldIndoor
SessionDate:	02-3-11
SessionTime:	8:57
SessionLocale:	Adaou

CorpusMetaData	**02-3-12:HanDBase Table Export**
Domain:	Syntax
Genre:	Questionnaire
Part/Sex/Age:	Kouamé Ama Bié f 35
Interviewers:	Adouakou
Recordist:	Salffner, Gibbon
Media:	Laryngograph
Equipment:	1) Audio: 2 channel, l laryngograph, r Sennheiser studio mike 2) Stills: Sony digital 3) Video: Panasonic digital (illustration of techniques)
SessionNote	Adouakou phrases repeat

For current purposes, databases are exported in the attribute-value format shown below and converted into the TASX reference XML format (Milde & Gut 2001). The PDA RDMS exports this DB table rather straightforwardly into XML, with the convention that the attribute is encoded as an element with appropriate tags, and the value as the content of the element. The XML code for the table is as follows:

```
<?xml version="1.0"?>
<CorpusMetaData>
    <Record>
          <RecordID>Agni2002a</RecordID>
          <LANGnames>Agni, Anyi</LANGnames>
          <SILcode>ANY</SILcode>
          <Affiliation>Kwa/Tano</Affiliation>
          <Lect>Indénié, Ndenye</Lect>
          <Country>Côte d'Ivoire</Country>
          <ISO>CI</ISO>
          <Continent>Africa</Continent>
          <LangNote></LangNote>
          <SESSION>FieldIndoor</SESSION>
          <SessionDate>03/11/2002</SessionDate>
          <SessionTime>08:57 am</SessionTime>
          <SessionLocale>Adaou</SessionLocale>
          <Domain>Syntax</Domain>
          <Genre>Questionnaire</Genre>
          <PartSexAge>Kouamé Ama Bié f 35</PartSexAge>
          <Interviewers>Adouakou</Interviewers>
          <Recordist>Salffner, Gibbon</Recordist>
          <Media>Laryngograph</Media>
          <Equipment>
                1) Audio:
                      2 channel, l laryngograph, r Sennheiser studio mike
                2) Stills: Sony digital
                3) Video: Panasonic digital (illustration of techniques)
          </Equipment>
          <SessionNote>f Adouakou phrases repeat</SessionNote>
...
    </Record>
...
</CorpusMetaData>
```

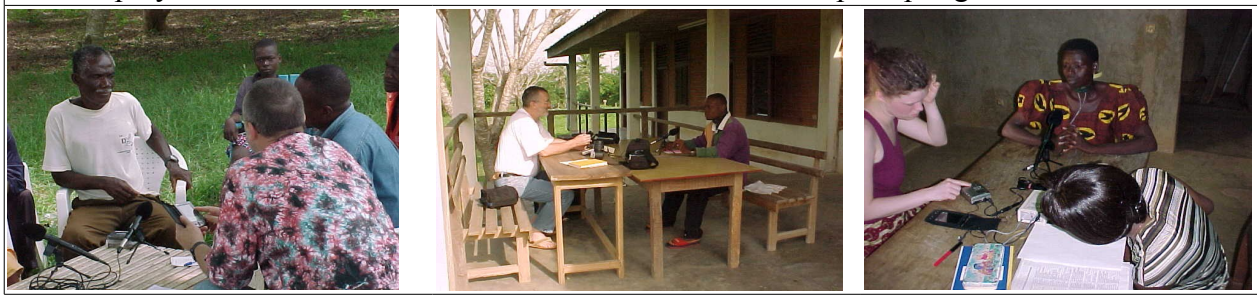
Specific examples of the application of the metadata editor in fieldwork sessions are depicted in the Figures.

5.3 Database implementation

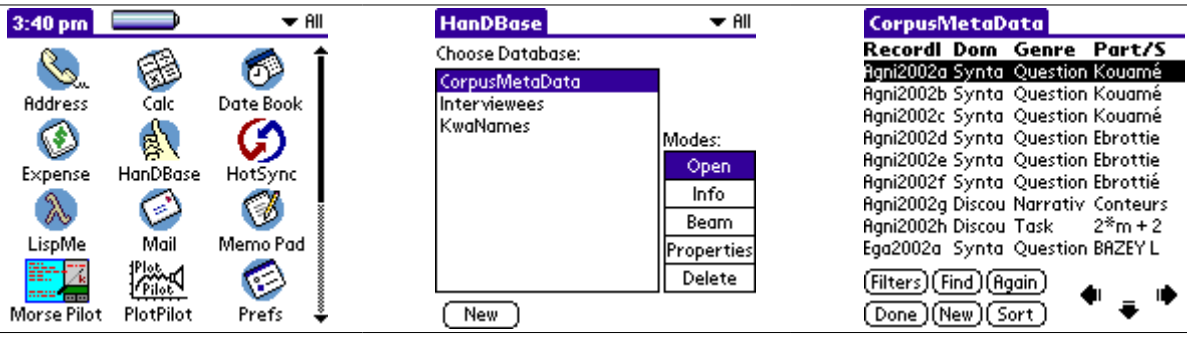
The metadata database was implemented for Palm PDAs with touchscreen input using the commercial HanDBase Relational Database Management System (RDBMS). Since the database on the PDA is mainly used for data acquisition, it would be straightforward to implement this on an open source basis using a developer friendly environment such as PocketC. The following screenshot series shows the Graphical User Interface with different views of the database.

1. Implementation on different Palm PDA models; zoom into data category view:

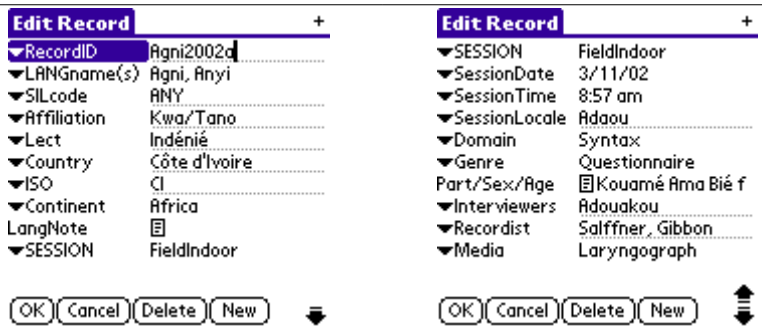
2. Deployment of PDA for metadata collection and interview prompting in the field:



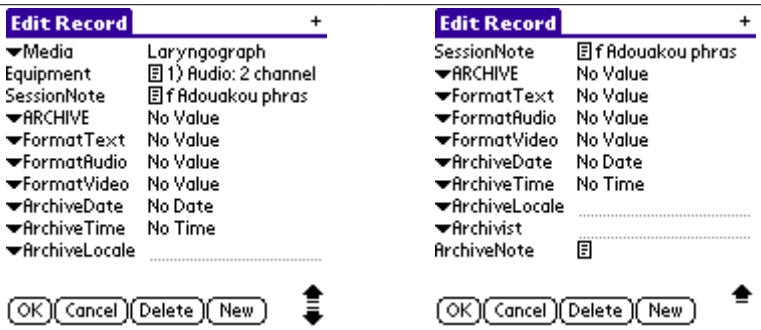
3. PDA screenshots of touchscreen GUI, database list and record list view:



4. PDA screenshots of data categories in metadata records:

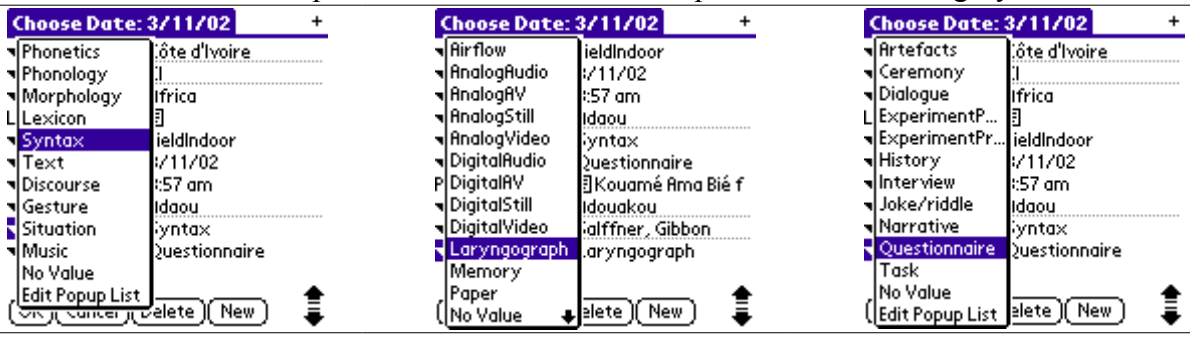


5. PDA screenshots of data categories in metadata records:



6. PDA Screenshots of drop-down menus and date picker for efficient metadata input:

7. PDA screenshots of drop-down menus for efficient input of metadata category values:



The images show the following features:

1. Implementation on the older Palm PDAs IIIc and Vx (in the second case with data transfer via infra-red connection to mobile phone).
2. PDA deployment in the field for metadata documentation and interview prompting:
   1. Interview of village narrator Grogba Marc of Gniguedougou (Ega, putative Kwa, South Central Ivory Coast), 2003.
   2. Interview of consultant Baze Lucien of Gniguedougou (Ega, putative Kwa, South Central Ivory Coast), 2003.
   3. Questionnaire interview of consultant Kouamé Ama Bié (Anyi/Agni, Kwa, South West Ivory Coast), 2003.
3. Graphical user interface environment on Palm applications touchscreen; RDBMS database selection and record selection widgets.
4. First and second parts of metadata attribute list, with specific values for one session.
5. Third and fourth parts of metadata attribute list, with specific values for one session.
6. Drop-down menus and date picker widget for efficient metadata category value entry.
7. Further drop-down menus for efficient metadata category value entry.

5.4 Evaluation

The metadata editor and database application has been tested extensively in fieldwork on West African languages, and has proved to be an indispensable productivity tool, especially in difficult situations where very limited time is available. Typical Data Processing layer outcomes with respect to phonetic data resources are shown in Connell, Ahoua & Gibbon (2002), Gibbon (2003) and Gibbon (2004).

6 Conclusion

This study has motivated, specified and described an implementation of the Fieldwork Acquisition layer of the UbiCorpus architecture for resource acquisition, processing and archiving. Architectures using the Resource Archive and Data Processing layers, e.g. a server configuration and a laptop for use in the field, are very common. The UbiCorpus model is based on and has been tested during extensive fieldwork experience in West Africa. Following the WELD criteria, owing to severe financial and platform resource limitations in practical linguistic fieldwork situations, the general development strategy is to use available freeware or open source components as far as possible, and to augment these with custom applications, which in the Palm PDA world are generally distributed as free demoware for initial testing.

In many situations the laptop is unsuitable because of relatively heavy power requirements which are not immediately available in many fieldwork locations without additional arrangements (solar panel, car battery, etc.). For these applications, handhelds, and in particular the PalmOS based family of handhelds, constitutes the platform of choice because of minimal expense, size and power requirements, permitting several days (or with some models weeks) of use on one charge or with a small external battery (internal on some older models). Although the PalmOS platform is not so suitable for signal processing applications such as time-aligned annotation (though freeware signal processing applications exist) it is well-suited for logging, transcription, reference and general editing purposes.

The power of PDA miniature computing platforms as useful components of laboratory and office environments is often underestimated. It was demonstrated that a number of applications for which even a laptop is clumsy or unsuited for the developing field of computational ethnolinguistic fieldwork may be elegantly provided on the Palm PDA platform. The addition of a foldable keyboard further enhances the text handling capacity of the devices.

In the medium term, it will be possible to integrate the hybrid applications at the Corpus Pilot, Data Processing and Resource Archive levels into a corpus management environment which not only permits seamless dataflow and workflow, a goal already achieved, but also into a non-technical user-friendly prototype which may serve as the basis of a fieldwork management product.

The UbiCorpus architecture has been used as the basic specification for different kinds of language documentation work in a variety of different projects. The Resource Archive layer was originally designed and implemented for web--based lexical database development in the VerbMobil project (Wahlster 2000), funded by the German Federal Ministry of Education and Research (BMBF). The concept has been further developed theoretically and practically in connection with the projects *Theorie und Design multimodaler Lexika* funded by the German Research Council (DFG), *Enzyklopädie der Sprachen der Elfenbeinküste* funded by the German Academic Exchange Service (DAAD) and *Ega: a documentation model for an endangered Ivorian language* in the pilot phase of the DOBES funding programme of the Volkswagen Foundation.

In its local implementation, the current Resource Archive layer version also includes support for telecooperation and web-teaching. The Data Processing layer includes numerous applications which cannot be specified here. The Corpus Pilot layer as described in the present contribution has been informally but extensively field tested at a number of fieldwork locations, most recently in the framework of DAAD (German Academic Exchange Service) funded doctoral thesis work. It is planned to apply the field testing criteria defined in Gibbon, Mertins & Moore (2000) to an extended implementation of the components of UbiCorpus model.

7 Bibliography

Bird, Steven & Gary Simons (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 (1997). *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd (2000). *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation*. Boston, Dordrecht & London: Kluwer Academic Publishers.

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

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 (2006). Language Documentation in West Africa. In: *Proceedings of the Workshop on "Networking the development of language resources for African languages"*. LREC 2006, Genoa.

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

Milde, Jan Torsten & Ulrike Gut (2001). The TASX-engine: an XML-based corpus database for time aligned language data. In: *Proceedings of the IRCS Workshop on Linguistic Databases*. Philadelphia: University of Pennsylvania.

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

Wahlster, Wolfgang, ed. (2000). *Verbmobil: Foundations of Speech-to-Speech Translation*. Berlin: Springer Verlag.