LSA 2008 Annual Meeting Tutorial, January 2008:
Mobilizing linguistic resources within speaker communities

Efficient Language Documentation: creation of local multipliers

Dafydd Gibbon, Universität Bielefeld

1 Introduction (p. 1)

1.1 Background (p. 1)

Language documentation can be enhanced by creating “local multipliers”: a network of participants from the language community and from local universities who are team-trained in modern language documentation methods. Participants can be drawn from three constituencies:

- Colleagues from local universities who are informed about modern documentary linguistic methods and, in the best case, are documentary linguists themselves;
- Graduate students, both from one’s home institution and from local universities;
- Helpers from local communities who in the best case provide structured dictionary material and text transcriptions.

The colleagues from local universities in the regions concerned are not only linguists, but colleagues from other departments with whom transdisciplinary work on modern computational language documentation methods is possible.

The development of such local multipliers is especially effective for established researchers who are committed to investigations and collaborations in a particular area over the long term, but even more junior researchers with a long-term commitment to a particular area or community will be able to make use of some of the methods on which the approach is based, both to achieve short-term goals and to set the stage for their future work. This presentation will relate key experiences and the results of this approach in West Africa, in order to give advice to those interested in applying this collaborative model elsewhere.

To date, this approach has proven especially successful in the development of multply re-usable dictionaries for use in linguistics and in speech technology, for example in speech synthesis applications for Nigerian languages.

1.2 Overview (p. 2)

After an overview of the general and specific goals of my talk, I will outline my own background in order to clarify my motivation for combining a number of different transdisciplinary elements in the approach I am advocating. I will proceed – as if in a software development context, because this is what resource development is, to a large extent, with texts as the core component of software – by outlining resource specifications and drawing some preliminary conclusions about resource creation in a transdisciplinary context. In the final section, I will discuss a number of examples of cooperation and make some suggestions for future developments.

1.3 General goals (p. 3)

The overriding goal of this presentation is to sketch an integrative, transdisciplinary and transnational programme for Documentary Linguistics in the 21st century, on the basis of state of the art techniques and strategies in the field.

Why is this an opportune time to suggest such a programme? One obvious reason is that Documentary

Linguistics, as outlined by Himmelmann (1998) and represented in many language documentation projects in the 1990s and 2000s, is in a sense coming of age as an identifiable linguistic sub-discipline. A second reason, however, is that Documentary Linguistics has taken up many heterogeneous components of and influences from a wide variety of other disciplines, as well as the traditional sub-fields of linguistics. These include:

1. Field Linguistics: methods of note-taking, interviewing, recording, with the goal of creating resources (in the context of descriptive linguistics, typological linguistics and various areas of applied linguistics) in the form of
    1. field notes for further exploitation;
    2. linguistic sketch descriptions of phonology, morphology, syntax, language in use;
    3. dictionaries;
    4. collections of transcribed and written texts.
2. Speech Technology: methods of recording spoken data, annotating with time-aligned transcriptions, induction of statistical models of language structure from such data, for the purpose of creating Automatic Speech Recognition and Text-to-Speech (TTS) Synthesis systems. This is the origin of the well-known Praat, Transcriber and WaveSurfer phonetic workbenches.
3. Natural Language Processing and Computational Linguistics: methods of analysing, generating and translating texts, and for operationally testing the coherence of linguistic theories, including Machine Learning, i.e. automatic induction of grammars and lexicons – not a fantasy, but a well-established field of research – in this area, finite state grammars are standard models and (contrary to assumptions made by many linguistics) are computationally adequate for a wide range of phenomena.
4. Document Modelling and Text Technology: for automatic document classification, text mining and information retrieval in text archives, the internet, to which (computational) linguistic enterprises such as the *Text Encoding Initiative* (TEI) have made significant contributions.
5. Word processor development: the development of document format style models (text linguistics) as well as spell checkers, grammar checkers, lexicons and thesauri, by linguists working for large organisations (cf. MS-Office; OpenOffice.org).

1.4 Specific goals (p. 4)

The specific goal of the presentation is to outline the workable, efficient language documentation, involving transdisciplinary and transnational partnerships, which my team uses in Computational Documentary Linguistics.

How is this done? The essential elements are covered by four dimensions:

1. Social: a collaborative multi-level partnership (for creating local multipliers and “human payback”).
2. Empirical: the use of well-tried, valid empirical procedures of interviewing, corpus collation and analysis, etc., for descriptive accuracy, soundness and completeness.
3. Formal: the use of explicit linguistic models – category sets, metadata sets, ontologies such as GOLD, the General Linguistic Ontology for Linguistic Descriptio (for consistency, archiving, search, lexicon and grammar induction, ...).
4. Operational: use of technologically up-to-date archiving and processing procedures and data structures, including
    1. speech annotation conventions and accepted tagsets for different levels of linguistic description;

2. productive uses of speech and language technology, for validating formal and empirical resources and for applications, involving necessary criteria of *completeness* (relative to a specified domain) and *soundness* (i.e. without spurious over-generalisation).

In the area of operational criteria, Steven Krauwer's BLARK concept has been very influential, and is currently being further developed for African languages: *Basic LAnguage Resource Kit*; further information on this is easily accessible on the internet.

1.5 By way of explanation: my background (p. 5)

The strategy I am advocating is of course highly dependent on my own background and experience. Briefly:

My core research areas are in the areas of fieldwork methods for West African Languages, and in more theoretical areas of computational phonology/prosody/morphology/lexicography (also including African languages).

But I am also interested in practical applications, as a form of “payback”: language resources must be machine processable, not only for heritage documentation, but also for speech and language technology.

The languages I have been concerned with creating resources for include German & English (Verbmobil project) and for West African languages (Côte D'Ivoire, Nigeria).

I have been and am involved in various infrastructure-oriented projects, including:

- COCOSDA, the international Coordinating Committee for Speech Databases and Assessment, of which I am currently coordinator;
- the European Union SAM project (cf. the SAMPA IPA encoding for speech engineering) and the EAGLES projects (two handbooks of resources, mainly for speech engineering, but also relevant for Documentary Linguistics.

2 Specifications for language resources (p. 6)

2.1 General specifications for language resources (p. 7)

These specifications for local language resource creation, developed in cooperation with Nigerian and Ivory Coast colleagues, were first published in the Newsletter of the European Language and Speech Network of Excellence (ELSNET) newsletter a few years ago.

By local language I just simply mean any language, endangered or not, which does not have a global trading function like English or French or Spanish or, increasingly, Chinese and so on. Language resources (i.e. annotated speech and text corpora, grammars, lexicons) should be:

- *comprehensive,* with respect to whichever application domain they are intended for;
- *effective,* in terms of human and computing resources;
- *state of the art,* not necessarily the latest internet-dependent software and hardware, but intellectually, of course (not everyone can afford and work with the latest internet registration procedures);
- *affordable,* for example, older computing facilities may be available, as in the countries where I have worked in West Africa, on which newer software and internet registration and updating techniques do not work;
- *fair,* i.e. provide payback, either as remuneration or as applications of the resources in printed or electronic (documents, software) media.

One consequence of this work has been that I have supported the Open Archive Initiative, specifically the Open Language Archive Community, which was founded by and is being continued by Steven Bird and Gary Simon. Of course I have encouraged the use and development of Open Software. My own software is of this category.

3 Cooperative resource creation (p. 8)

3.1 Models of resource creation (p. 8)

The “HEROIC LONE FIELDWORKER MODEL”, which is necessarily an essential ingredient of much linguistic fieldwork activity, needs to be supplemented by more sophisticated cooperative activities. There are of course many partially realised initiatives of this kind; my aim here is to indicate an ideal complex activity texture to aim for (Table 1).

Table 1: Models for resource creation in Documentary Linguistics.

MODEL TYPE	***AGENT***	***SOURCE***
LONE FIELDWORKER MODEL	fieldworker	community
MULTI-LEVEL COOPERATIVE MODELS		
TRANSDISCIPLINARY MODELS	linguistics (descriptive; corpus linguistics)	local colleagues
	phonetics speech technology	
	text technology	
	computer science / computational linguistics	
TRANSNATIONAL INFRASTRUCTURAL MODELS	linguistics student	local linguistics student
	linguistics department	local/regional linguistics department
	funding organisations	regional/international funding organisations
	political institutions	regional political institutions

3.2 Development of corpus resource methods (p. 9)

The bar chart (Figure 1) simply shows – very roughly – the development of corpus resource creation methods over the past two centuries, and the convergence of linguistic and technological requirements during this time.

The real start of corpus linguistics was in the mid-19th century. Here is a bit of linguistic “trivial pursuit”: it is not widely known (except among linguists) that the Grimm brothers, were indeed linguists (of “Grimm's Law” fame), and that collected their fairy tales in different German dialects primarily as a source of data for their reconstruction of the history of the German language, and only secondarily as a source of income for the Disney Corporation. Indeed, like many modern linguists they were also politically revolutionary, and expelled from their university – Universität Göttingen – for their activities.

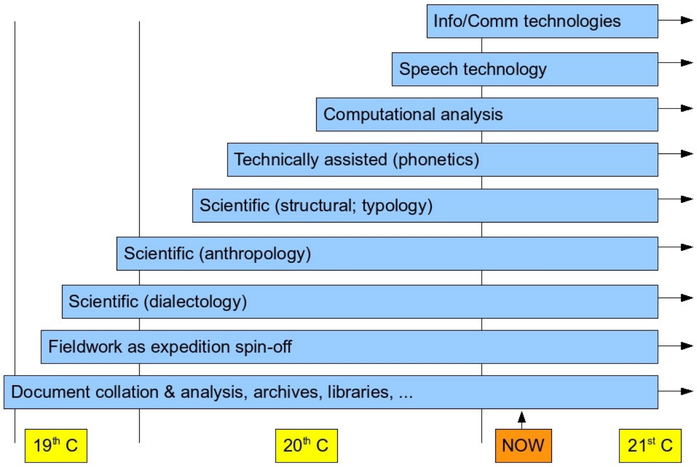

Info/Comm technologies
Speech technology
Computational analysis
Technically assisted (phonetics)
Scientific (structural; typology)
Scientific (anthropology)
Scientific (dialectology)
Fieldwork as expedition spin-off
Document collation & analysis, archives, libraries, ...
19th C
20th C
NOW
21st C

Figure 1: Development of resource creation strategies.

3.3 Transdisciplinary shared goals for resources (p. 10)

The chronological convergence of linguistic and technological requirements is manifested in a wide range of shared goals in speech and language processing research and development, on the one hand, and in (computational) Documentary Linguistics on the other. Both areas necessarily use very similar speech and text resources (though in speech and language technologies the corpora are usually vastly larger, though not necessarily more complex). Human Language Technology systems are used for research-oriented problem-solving in empirically based linguistic theory development: a speech synthesiser will not work if the underlying models are wrong, or incomplete, or unsound. Both are employed in practical infrastructure developments of various kinds, particularly educational media. Practical uses of resources for speech and language technologies in the contexts within which field linguists also work include health and agriculture information services.

3.4 Transdisciplinary R&D context (p. 11)

To summarise: it should be fairly obvious from the preceding discussion that the main disciplines involved in this cooperation are linguistics and phonetics, with fieldworkers, text technologists (for archiving, search), speech technology engineers, computer scientists and computational linguists.

3.5 Speech resources: technology AND linguistics (p. 13)

As a quick illustration of one of the most well-known interfaces between linguistic and phonetic resource creation on the one hand, and speech technology resource creation on the other: the Praat phonetic workbench is used in both fields, mainly for the time-aligned annotation of speech data, as shown in the illustration (Figure 2).

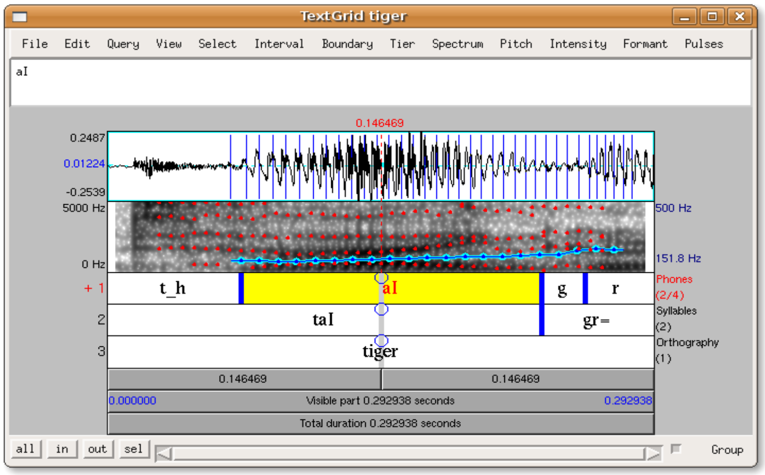


Figure 2: Time-aligned annotation with Praat.

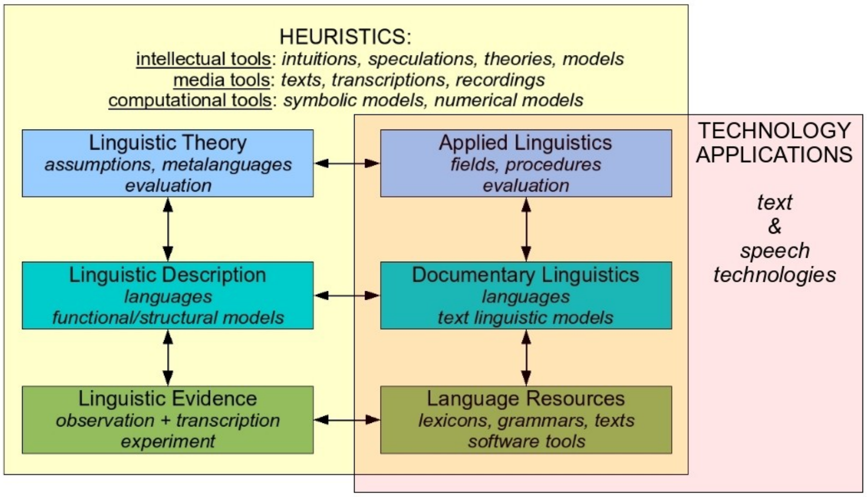


Figure 3: Cooperations between Linguistics and Computational Linguistics.

3.6 Cooperation with Computational Linguistics (p. 14)

The overlaps and cooperation potential between theoretical and descriptive linguistics, documentary linguistics and the text and speech technologies are considerable. A model for relating these areas is showin in Figure 3.
I will leave Figure 3 for further discussion, without further comment at the moment.

3.7 Current resource creation in Documentary Linguistics (p. 15)

Thieberger and Nash have developed useful workflow models for language documentation, one of which is shown in Figure 4; this is included for reference, and will also not be further commented here, for lack of time.

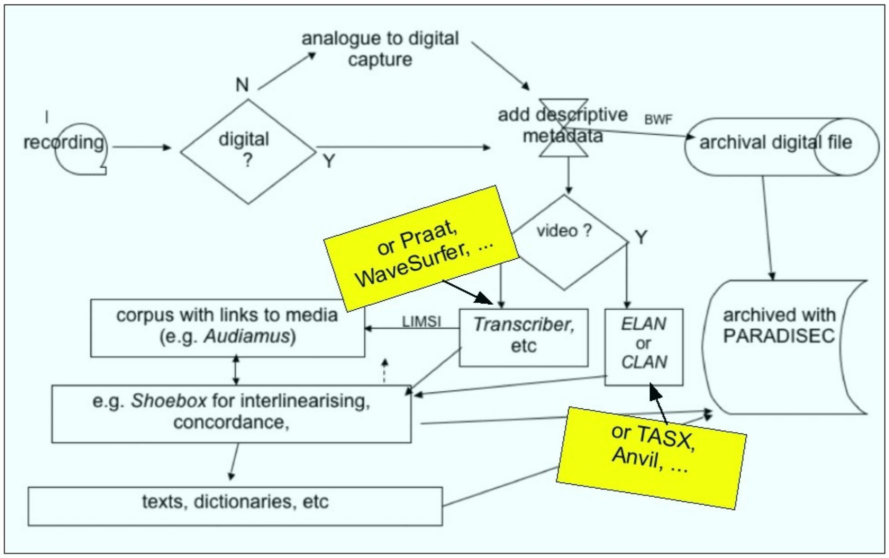


Figure 4: Thieberger and Nash workflow proposal for PARADISEC.

3.8 Integrated computational resource development (p. 16)

The present proposal goes further than the PARADISEC proposal in terms of the linguistic and computational domains covered. The integrated model (Figure 5) outlines how contributions from existing areas of computational linguistics are related to resource development, from speech corpus creation (time-aligned annotation) through lexicon creation to text annotation (tagging). Clearly, computational techniques can take the resource creation process much further, but the basic resources which are currently created, both with individual tools such as SIL tools and in mainstream computational linguistics, are shown in the Figure.

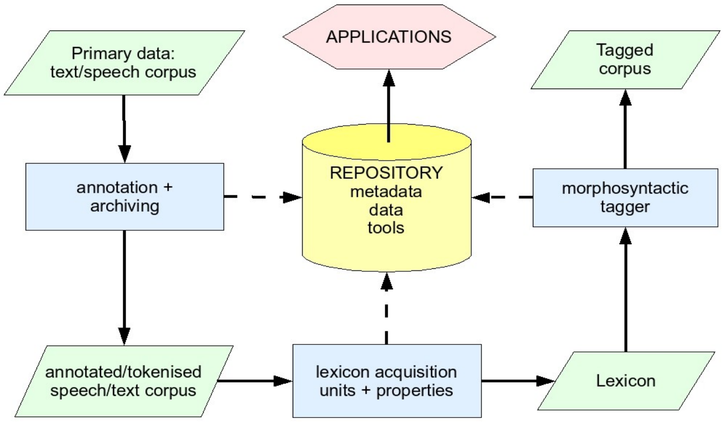

Primary data: text/speech corpus
APPLICATIONS
Tagged corpus
annotation + archiving
REPOSITORY metadata data tools
morphosyntactic tagger
annotated/tokenised speech/text corpus
lexicon acquisition units + properties
Lexicon

Figure 5: Computational linguistics based workflow for resource creation.

4 For example... (p. 17)

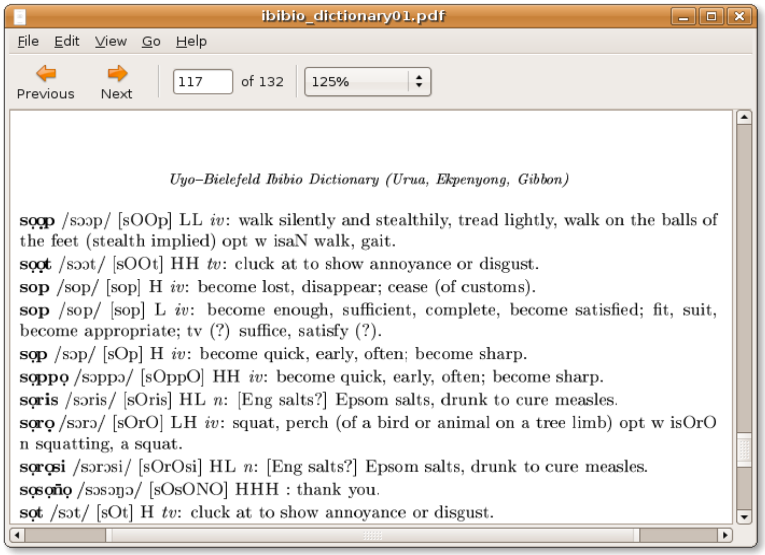
4.1 Some transnational resource projects (p. 18)

The projects listed here are simply for reference, and will not be commented on in detail here; nor need they be read out... The important point is that these are or were resource-creation partnerships with universities in other continents, in which universities, colleagues and students were explicitly involved as partners in the projects, rather than being purely national projects with “missions” to other regions.

- DAAD funded international projects:
  - 1990s: Design for a new atlas of Ivory Coast languages
    - with Christian Lehmann
    - Université de Cocody, Côte d'Ivoire
  - 2001-2005: Development of an MA curriculum for Computational Language Documentation
    - Université de Cocody, Abidjan, Côte d'Ivoire
    - University of Uyo, Akwa Ibom State, Nigeria
  - 2002: DoBeS pilot project: EGA: A Documentation Model for an Endangered Ivorian Language
  - 2002-2003: Data Mining on Large Spoken Language Corpora
    - University of Campinas, Brazil
- Outside Echo funded international project:
  - 2002-2003: Speech Synthesis for Ibibio
    - University of Uyo, Akwa Ibiom State, Nigeria
    - also: Nairobi, Johannesburg, Hyderabad partners

4.2 Cooperation with Computational Linguistics: lexical databases (p. 19)

The 4000 word Ibibio dictionary, from which an excerpt is shown in , was created automatically from a simple database, typed by a secretary into a spreadsheet, and processed straightforwardly with UNIX scripting tools such as Perl to produce a custom made LaTeX document, which was then printed in the usual way.



Uyo–Bielefeld Ibibio Dictionary (Urua, Ekpenyong, Gibbon)

sọọp /sɔɔp/ [sOOp] LL *iv*: walk silently and stealthily, tread lightly, walk on the balls of the feet (stealth implied) opt w isaN walk, gait.
sọọt /sɔɔt/ [sOOt] HH *tv*: cluck at to show annoyance or disgust.
sop /sop/ [sop] H *iv*: become lost, disappear; cease (of customs).
sop /sop/ [sop] L *iv*: become enough, sufficient, complete, become satisfied; fit, suit, become appropriate; tv (?) suffice, satisfy (?).
sọp /sɔp/ [sOp] H *iv*: become quick, early, often; become sharp.
sọppọ /sɔppɔ/ [sOppO] HH *iv*: become quick, early, often; become sharp.
sọris /sɔris/ [sOris] HL *n*: [Eng salts?] Epsom salts, drunk to cure measles.
sọrọ /sɔrɔ/ [sOrO] LH *iv*: squat, perch (of a bird or animal on a tree limb) opt w isOrO n squatting, a squat.
sọrọsi /sɔrɔsi/ [sOrOsi] HL *n*: [Eng salts?] Epsom salts, drunk to cure measles.
sọsọñọ /sɔsɔŋɔ/ [sOsONO] HHH : thank you.
sọt /sɔt/ [sOt] H *tv*: cluck at to show annoyance or disgust.

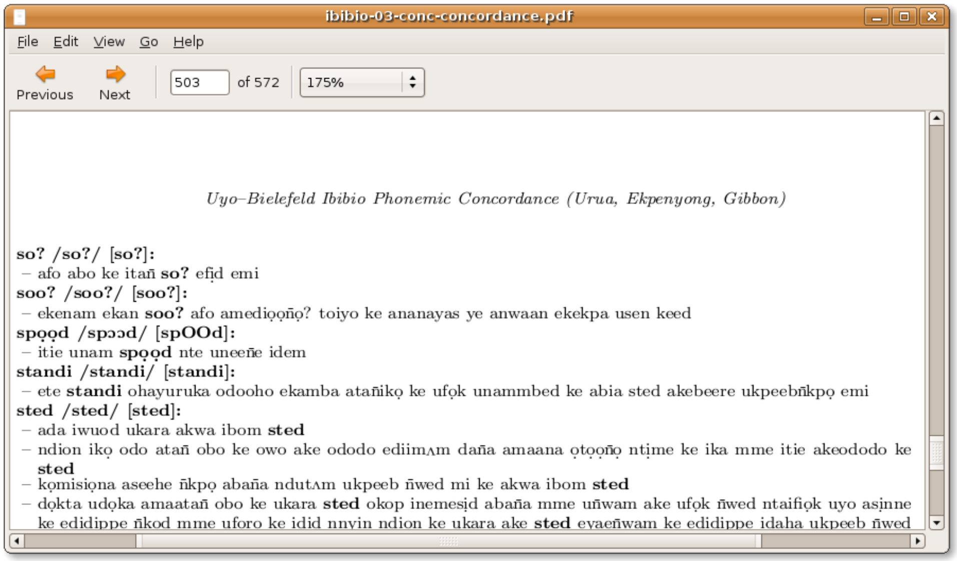
Figure 6: Ibibio dictionary excerpt.

The advantage of this procedure is that is is very straightforward to convert the database into a hypertext internet dictionary from the same database; this was in fact done by an Ibibio software engineer now working in the USA; the result can easily be found on the internet by searching for the keyword “MyIbibio”.

The required computational skills are rather widespread; in the project concerned, we cooperated with linguists and computer scientists at the University of Uyo, Akwa Ibom State, Nigeria.

4.3 Cooperation with computational linguistics: concordancing (p. 20)

Another very basic application of computational techniques to resource creation is the building of a concordance from a corpus. There are many tools for creating concordances, and a basic concordance tool is not hard to create. Why should concordances be part of a linguistic resource collection? The main reason, from the linguistic point of view, is that a concordance is an essential tool for the lexicographer and for the grammarian when it comes to examining detailed, empirically gained corpora for information about the language. Concordances figure far too rarely in specifications of language resources – they are perhaps the most elementary kind of dictionary, and of course standard dictionaries contain concordance-like elements in the examples given in dictionary articles. An extract from a concordance made from the Ibibio corpus is shown in Figure 7.



Uyo–Bielefeld Ibibio Phonemic Concordance (Urua, Ekpenyong, Gibbon)

so? /so?/ [so?]:
– afo abo ke itañ **so?** efịd emi
soo? /soo?/ [soo?]:
– ekenam ekan **soo?** afo amediọọñọ? toiyo ke ananayas ye anwaan ekekpa usen keed
spọọd /spɔɔd/ [spOOd]:
– itie unam **spọọd** nte uneeñe idem
standi /standi/ [standi]:
– ete **standi** ohayuruka odooho ekamba atañikọ ke ufọk unammbed ke abia sted akebeere ukpeebñkpọ emi
sted /sted/ [sted]:
– ada iwuod ukara akwa ibom **sted**
– ndion ikọ odo atañ obo ke owo ake ododo ediimʌm daña amaana ọtọọñọ ntịme ke ika mme itie akeododo ke **sted**
– kọmisiọna aseehe ñkpọ abaña ndutʌm ukpeeb ñwed mi ke akwa ibom **sted**
– dọkta udọka amaatañ obo ke ukara **sted** okop inemesịd abaña mme uñwam ake ufọk ñwed ntaifiọk uyo asịnne ke edidippe ñkod mme uforo ke idid nnyin ndion ke ukara ake **sted** eyaeñwam ke edidippe idaha ukpeeb ñwed

Figure 7: Ibibio concordance extract.

4.4 Some results of cooperation (p. 21)

Briefly, the results of this kind of transnational, transdisciplinary cooperation include

- a lexical database for use in linguistic and speech technology applications,
- a prototype speech synthesiser for Ibibio (check http://www.llsti.org/),
- an MA course “Computational documentation of Local Languages” at Université de Cocody, Abidjan (Côte d'Ivoire) and University of Uyo, Akwa Ibom State (Nigeria).

The MA course has aroused the interest of UNESCO, and has been presented by my Nigerian colleague Prof. Eno-Abasi Urua at two UNESCO conferences (Mali, Ethiopia).

The speech technology work and the educational work have influenced the establishment of a section on Documentary Linguistics in Johannesburg, and the establishment of a PhD course and the “EthioBLARK” project in Addis Ababa, Ethiopia.

5 Where can we go from here?

I have advocated the development of an integrated, transdisciplinary, transnational model of language resource creation involving Documentary Linguistics and the Human Language Technologies. The simple justification is that the task is too great for individual researchers and traditional methodologies. The approach is viable, though it takes effort. To conclude, I will simply list a few references to relevant work in this area:

Just a few references to conclude:

The Local Languages Speech Technology Initiative (LLSTI): http://www.llsti.org/

India: the Simputer project: http://www.simputer.org/

South Africa: several initiatives for computational lexicography, automatic speech recogntion and text-to-speech synthesis for the 11 official languages.

Free software (basic resources, often overlooked in this context): operating systems such as Linux (localisation to many local languages), generic applications such as OpenOffice (likewise many linguists employed in localisation), specific applications such as Praat (very widely used in spoken language documentation in linguistics, phonetics and speech technology).

The Open Archive Movement: the Open Language Archives Community (OLAC) as a spin-off of the library and archive serving Open Archive Initiative (OAI): http://www.language-archives.org/

SPICE (Tanja Schulz): a web-based, any-language, automatic text to speech synthesis & automatic speech secognition prototype creator: http://csl.ira.uka.de/index.php?id=29&L=1

And now – many thanks, and I am looking forward to further discussion!