Language Documentation and Description

ISSN 1740-6234

Current issues in language documentation

PETER K. AUSTIN

Current Issues in Language Documentation

Peter K. Austin

1. Introduction[1]

This chapter is an introduction to the field of linguistics that has come to be known as 'language documentation' or 'documentary linguistics', covering its main features and giving examples of what it involves. The difference between language documentation and descriptive linguistics is discussed, and an argument presented that the two are complementary activities that can cross-fertilise one another. We then look at some current challenges in the field of language documentation, including issues that are the subject of on-going research.

2. Language Documentation

Language documentation (also known by the term 'documentary linguistics') is the subfield of linguistics that is 'concerned with the methods, tools, and theoretical underpinnings for compiling a representative and lasting multipurpose record of a natural language or one of its varieties' (Himmelmann 2006:v). A similar definition is given by Woodbury (2010) as 'the creation, annotation, preservation, and dissemination of transparent records of a language'. Language documentation is by its nature multidisciplinary, and as Woodbury (2010) notes, it draws on 'concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts, and more' (see Harrison 2005, Coelho 2005, Eisenbeiss 2005 for examples).

Documentary linguistics has developed over the past 15 years in response to the need to make a lasting record of the world's endangered languages (estimated to be as many as 90% of the 7,000 languages spoken on earth today), and to support speakers of these languages in their desires to maintain

[1] I am grateful to Lise Dobrin, Lenore Grenoble, David Nathan and Tony Woodbury for detailed comments on an earlier draft of materials incorporated into this chapter; I alone am responsible for any errors.

Peter K. Austin 2010. Current issues in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description*, *Vol 7,* 12-33. London: SOAS.

them (Austin 2007, Whalen 2004). Its establishment is not only driven by the pressing need to record languages while speakers continue to use them, it is also fuelled by advances in information, media, communication and archiving technologies (see Nathan 2010a, 2010b) which make possible the collection, analysis, preservation and dissemination of documentary records in ways which were not feasible previously. Language documentation also fundamentally concerns itself with the rights and needs of language speakers and their direct involvement in the documentation and support of their own languages (see Austin 2010).

Himmelmann (2006:15) identifies five significant features of language documentation:

- **focus on primary data** – language documentation concerns the collection and analysis of an array of primary language data to be made available for a wide range of users;
- **explicit concern for accountability** – access to primary data and representations of it makes evaluation of linguistic analyses possible and expected;
- **concern for long-term storage and preservation of primary data** – language documentation includes a focus on archiving in order to ensure that documentary materials are made available to potential users into the distant future;
- **work in interdisciplinary teams** – documentation requires input and expertise from a range of disciplines and is not restricted to linguistics alone;
- **close cooperation with and direct involvement of the speech community –** language documentation requires active and collaborative work with community members both as producers of language materials and as co-researchers.

As language documentation projects have been initiated, reports on their progress become available and their results start to be deposited in archives, it is becoming clear that a further aspect is equally or more important (Dobrin et al. 2009, Nathan 2010b):

- **diversity** – as researchers respond to the unique and particular social, cultural and linguistic contexts within which individual languages are spoken, documentation projects are showing a diversity of approaches, techniques, methodologies, skills and responses.

Before continuing, it is perhaps also useful to identify what language documentation is not:

- **it is not about collecting material to preserve it without analysing it** – Himmelmann (1998:166) has argued that language documentation should strive 'to provide a comprehensive record of the linguistic practices characteristic of a given speech community' but to do so requires the application of analytical theories and techniques to transcribe, translate, annotate and disseminate these records;
- **it is not language description plus technology** – some commentators have suggested that language documentation is just descriptive linguistics, as practised by early 20th century scholars, for example, with the addition of technologies such as digital audio and video recording. This is a misrepresentation which fails to appreciate those aspects of language documentation which differentiate it from description, such the close attention to methodology (Lüpke 2010) and to the nature and role of data and metadata within the analytical processes (see 4 below for further discussion);
- **it is not necessarily about endangered languages per se** – the principles and practices of language documentation can be applied to all languages, large or small, endangered on non-endangered. While the field has developed as a response to language endangerment, it is not fundamentally restricted to endangered languages;
- **it is not a fad or passing phase** – language documentation is a development that is not a temporary aberration within linguistics but represents a paradigm shift within the discipline.

There is a lot of evidence to support this last point, including the following (see also the Appendix in Dobrin et al. 2009: 47-50):

1. **student interest** – the MA in Language Documentation and Description which was established in 2003 at SOAS has so far graduated 62 students, and in 2009-10 it has attracted its largest enrolment of 17. The PhD in Field Linguistics at SOAS has graduated 7 students to date, with a further 12 currently enrolled. Other post-graduate programmes that support language documentation, such as the University of Texas at Austin and the University of Hawaii at Manoa, report similar growth in student numbers. Further evidence comes from summer schools such as the 3L Summer School held in London in June-July 2009 which attracted 100 participants, or the InField Workshop held at the University of California, Santa Barbara in summer 2008 that attracted 75 attendees;

2. **researcher interest** – growing numbers of researchers have been attracted to the field, including post-doctoral fellows and more senior researchers. This is reflected in the growing number of publications, conferences, and workshops on language documentation topics;
3. **publications** – there is an increasing number of publications on language documentation theory, practice and outcomes, such as Gippert et al. 2006, Harrison et al. 2008. The journal *LanguageDocumentation and Description* was established at SOAS in 2003 and has so far published seven volumes and sold over 1,500 copies. The on-line journal *Language Documentation & Conservation* is published by University of Hawaii and has produced six issues to date. Several major reference works are in press, including Austin & Sallabank 2010, and Austin & McGill 2010;
4. **research funding** – since 1996 both government and non-government funding for language documentation research has grown at astronomical rates so that currently millions of pounds/dollars are available annual to support documentation work, specifically:

- the Endangered Languages Documentation Programme (ELDP) at SOAS has to date funded 195 documentation projects on endangered languages worth £7.25 million (see Figure 1);
- the Volkswagen Foundation DoBeS project has funded 50 research projects to a value of over 30 million euros (see Figure 2);
- the Documenting Endangered Languages (DEL) inter-agency programme of the National Science Foundation and the National Endowment for the Humanities has funded 60 projects worth approximately $10 million;
- the European Science Foundation EuroBABEL initiative (Better Analyses Based on Endangered Languages) is funded by 8 million euros;
- smaller, more modest funders, such as the Endangered Language Fund (ELF), Foundation for Endangered Languages (FEL), Gesellschaft für bedrohte Sprachen (GfBS) and Unesco, have provided hundreds of thousands of pounds in support for dozens of research and community-based documentation projects.

Figure 1: Map of ELDP-funded projects for 2003-2007

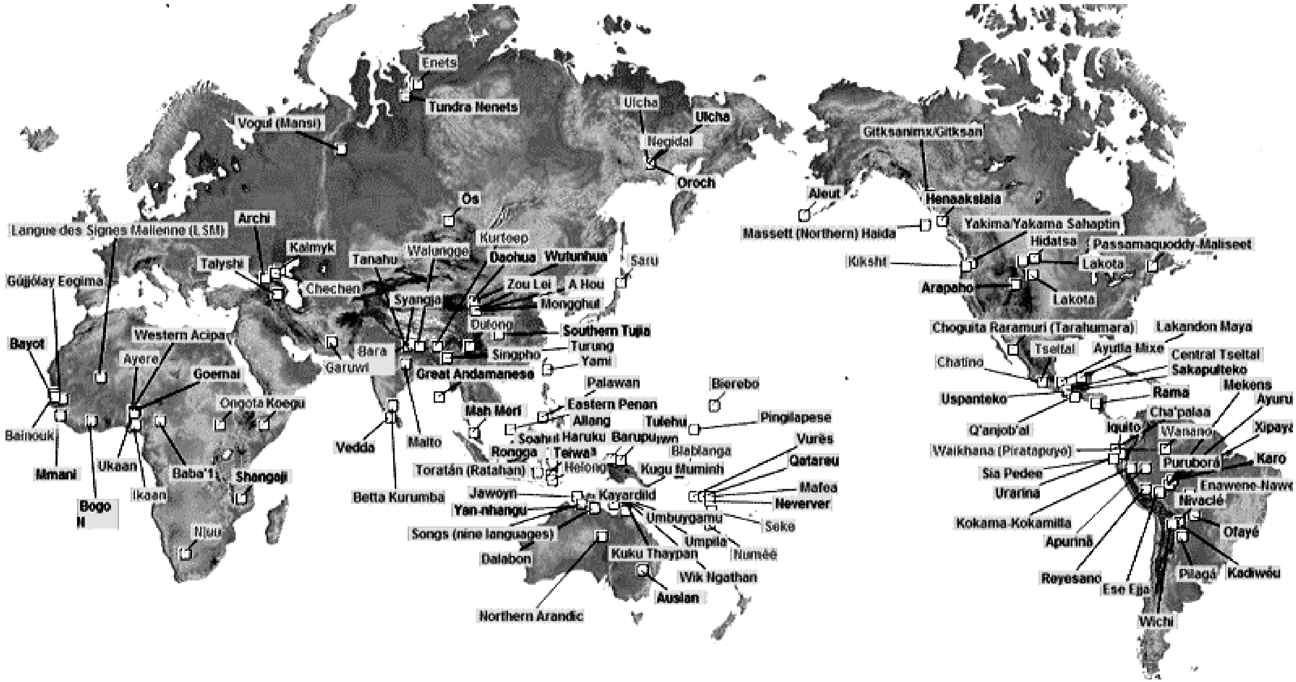
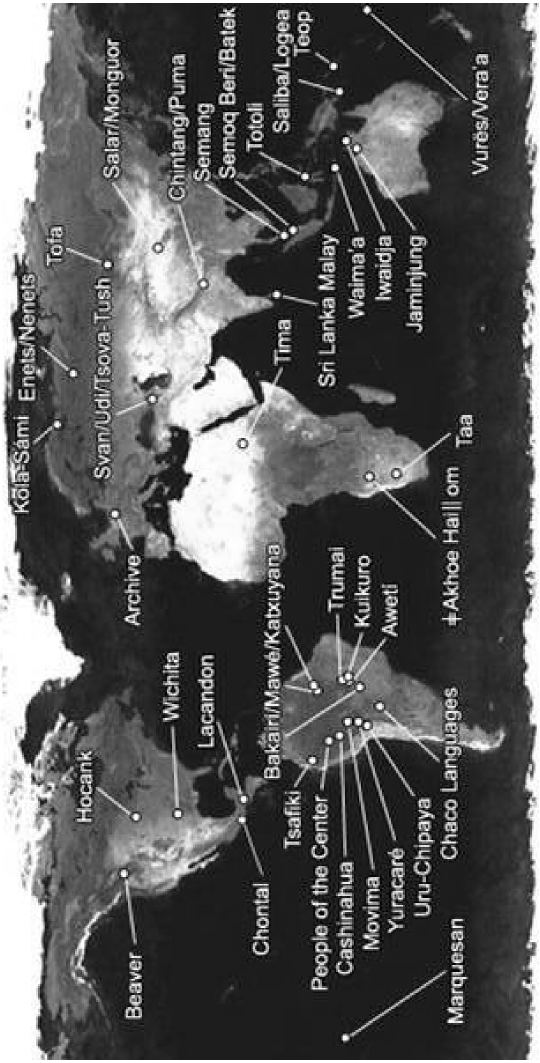


Figure 2: Map of Volkswagen-funded DoBeS projects



There are several reasons why language documentation as a field has emerged over the past 15 years (Dobrin et al. 2009, Austin & Grenoble 2007, Woodbury 2003). The first of these is the growing concern among linguists and others about the crisis facing the world's languages and threats to linguistic diversity, including the possibility that 50-90% of the 7,000 languages spoken on earth today will cease to be passed to children, or used by anyone at all, by the end of this century (Austin 2007). Secondly, linguists have shown increasing interest, both in theoretical research (see Sells 2010) and in linguistic typology (Bond 2010) in incorporating data from as wide a range of languages as possible to ensure that claims about human language and linguistic abilities are not areally or genetically biased, but represent the true diversity of the human language capacity. Thirdly, there has been a move towards a more humanistic view of language that pays attention to language as use rather than language as system. Fourthly, as mentioned above, the emergence of extensive funding resources and the requirements of funders to adopt a documentary perspective and to archive recorded data and analyses has had an influence on the topics that linguists (and others) have chosen to research, and the research methods they are employing. Technological developments in such areas as digital recording equipment, portable computing equipment (including netbooks and mobile phones), software tools, and storage devices have created the possibility of recording and analysing massively greater amounts of data, and being able to access and link it in ways not previously possible. Finally, increasing ethical and social concerns for the rights and needs of stakeholders across the language documentation domain have also played a role in its emergence and particular focus.

3. The documentary record

According to Himmelmann's 1998 original specification, language documentation aims to record the linguistic practices and traditions of a speech community, along with speakers' metalinguistic knowledge of those practices and traditions. This includes systematic recording, transcription, translation and analysis of a variety of spoken (and written) language samples collected within their appropriate social and cultural context (Austin 2006, HRELP 2006). Analysis in documentation is aimed at making the records accessible to a broad range of potential users which includes not only linguists but also researchers in other disciplines, community members and others, who may not have first-hand knowledge of the documented language. The record is thus intended for posterity (and hence should be preservable and portable, in the sense of Bird & Simons 2003), and so some level of analysis is required, in particular glossing and translation into one or more languages of wider communication, and systematic recording of metadata to make the archived documents understandable, findable and usable.

The core of a language documentation is generally understood to be a corpus of audio and/or video materials with time-aligned transcription, annotation, translation into a language of wider communication, and relevant metadata on context and use of the materials. Woodbury (2003) argues that the corpus will ideally cover a *diverse* range of genres and contexts, and be *large*, *expandable*, *opportunistic*, *portable*, *transparent*, *ethical* and *preservable*. As a result, documentation is increasingly done by teams, including community members in various roles (see Austin 2010). Lexico-grammatical analysis (description; see 4 below) and theory construction is contingent on and emergent from the documentation corpus (Woodbury 2003, 2010).

The components of documentation research consist of the following five activities (see also Austin 2006, 2008):

- **recording** – of media and text (including metadata) in context;
- **transfer** – to a data management environment. Increasingly recordings are 'born digital' (Nathan 2010a) and transfer involves moving files from storage media such as flash cards to computer hard disks with proper file and folder naming within a data management system (Nathan 2010b section 3.7);
- **adding value** – the transcription, translation, annotation and notation and linking of metadata to the recordings. Increasingly this is done with computer software such as ELAN and Toolbox;
- **archiving** – creating archival objects (or 'bundles', see Nathan 2010b) and assigning them access and usage rights;
- **mobilisation** – creation, publication and distribution of outputs, in a range of formats for a range of different users and uses.

Due to the diversity of documentation projects noted above, it is difficult to identify a 'typical' language documentation project, however it may be useful for those entering the field to have some idea of the kinds of projects undertaken within documentary linguistics (see www.hrelp.org/projects for some short descriptions). Here I present an overview of one project, the four year PhD research carried out at SOAS by Stuart McGill. The project involved both documentation and descriptive/theoretical research on Cicipu, a Niger-Congo language spoken in north-west Nigeria, carried out in collaboration with several native speaker researchers (see www.cicipu.org for further information). The project involved two fieldwork trips to Nigeria, a longer one of 8 months in the second year and a shorter follow-up trip in the third year of 4 months. During the first fieldwork only audio and written records were created – before Stuart's research no work had been done on Cicipu and its phonology (including the complex tonal system) and morphology required much initial research. On the second field trip Stuart recorded video, including conversational and culturally significant materials. There are sample videos available on the website www.cicipu.org.

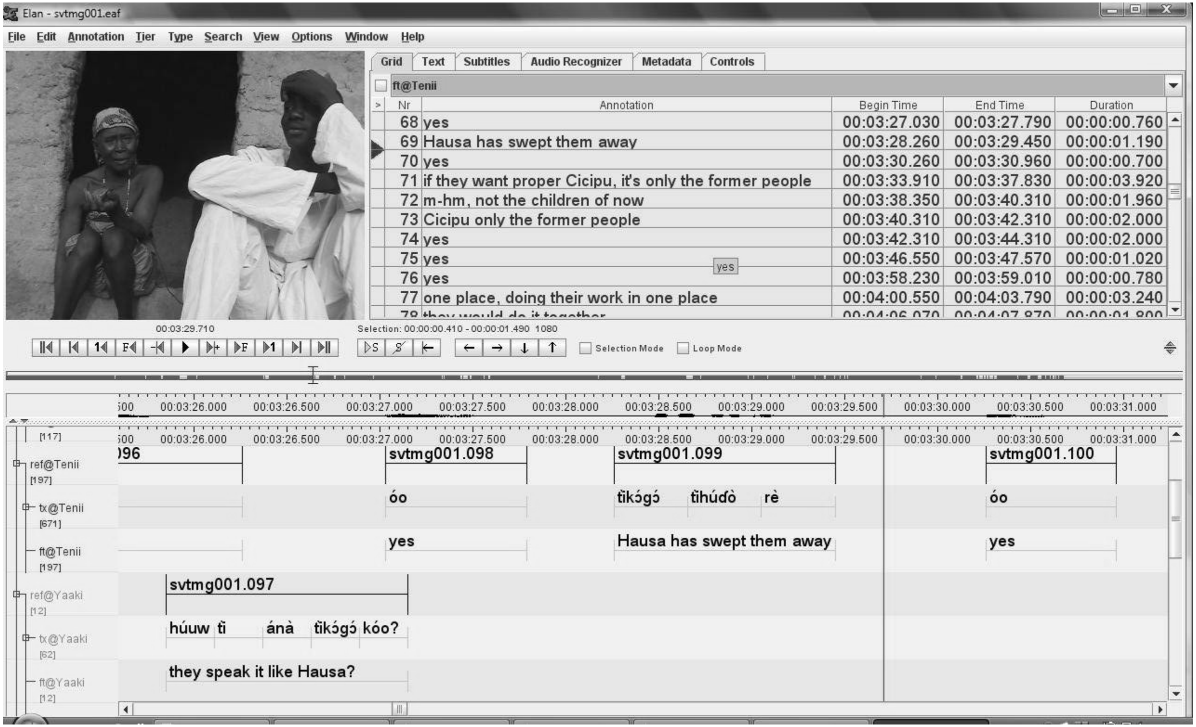
The outcomes of the project were the following:

- a corpus of recorded and written materials, including six hours of narrative and conversational texts recorded on audio or video, all of which has been fully transcribed using ELAN software (see Figure 3) and annotated using Toolbox software (see Figure 4);
- a 2,000 item lexicon with Cicipu headwords and Hausa and English glosses, stored in Toolbox field-oriented standard format (FOSF);
- a digital archive of 956 files comprising 50 Gbytes, deposited in the ELAR archive at SOAS;
- an overview grammar of the language (part 1 of the PhD, comprising 134 pages);
- an analysis of the verbal agreement system of Cicipu (part 2 of the PhD, 158 pages) describing the complex system of choice between person and gender (noun class) agreement on the verb in the third person, and showing how typological and theoretical syntactic accounts (in particular LFG) do not deal with these alternations. A theoretically-informed analysis in terms of information structure is presented;
- a website (www.cicipu.org) that includes recordings, glossed texts linked to the audio (with a choice of interlinear glossing or not); cassette tapes of text materials and songs for community members; books of folk tales; an orthography proposal and a workshop on orthography and spelling.

4. Documentation and description

Language documentation and description differ in terms of their goals, areas of interest, research methods, workflows, and outcomes. Language description typically aims at the production of grammars, dictionaries, and collections of texts, the intended audience of which is usually linguists, and the materials produced are sometimes written in frameworks accessible only to trained linguists. In contrast, documentation is discourse-centered: its primary goal is the direct representation of a wide range of discourse types (Austin 2008; Woodbury 2003, 2010; Himmelmann 1998). Although description may draw on a corpus, it involves analysis of a different order: description provides an

Figure 3: screenshot of ELAN transcription and annotation



Nr	Annotation	Begin Time	End Time	Duration
68	yes	00:03:27.030	00:03:27.790	00:00:00.760
69	Hausa has swept them away	00:03:28.260	00:03:29.450	00:00:01.190
70	yes	00:03:30.260	00:03:30.960	00:00:00.700
71	if they want proper Cicipu, it's only the former people	00:03:33.910	00:03:37.830	00:00:03.920
72	m-hm, not the children of now	00:03:38.350	00:03:40.310	00:00:01.960
73	Cicipu only the former people	00:03:40.310	00:03:42.310	00:00:02.000
74	yes	00:03:42.310	00:03:44.310	00:00:02.000
75	yes	00:03:46.550	00:03:47.570	00:00:01.020
76	yes	00:03:58.230	00:03:59.010	00:00:00.780
77	one place, doing their work in one place	00:04:00.550	00:04:03.790	00:00:03.240
78	they would do it together	00:04:06.070	00:04:07.870	00:00:01.800

ref@Tenii:)96 | svtmg001.098 | svtmg001.099 | svtmg001.100

tx@Tenii: óo | tìkɔ́gɔ́ tìhúɗò rè | óo

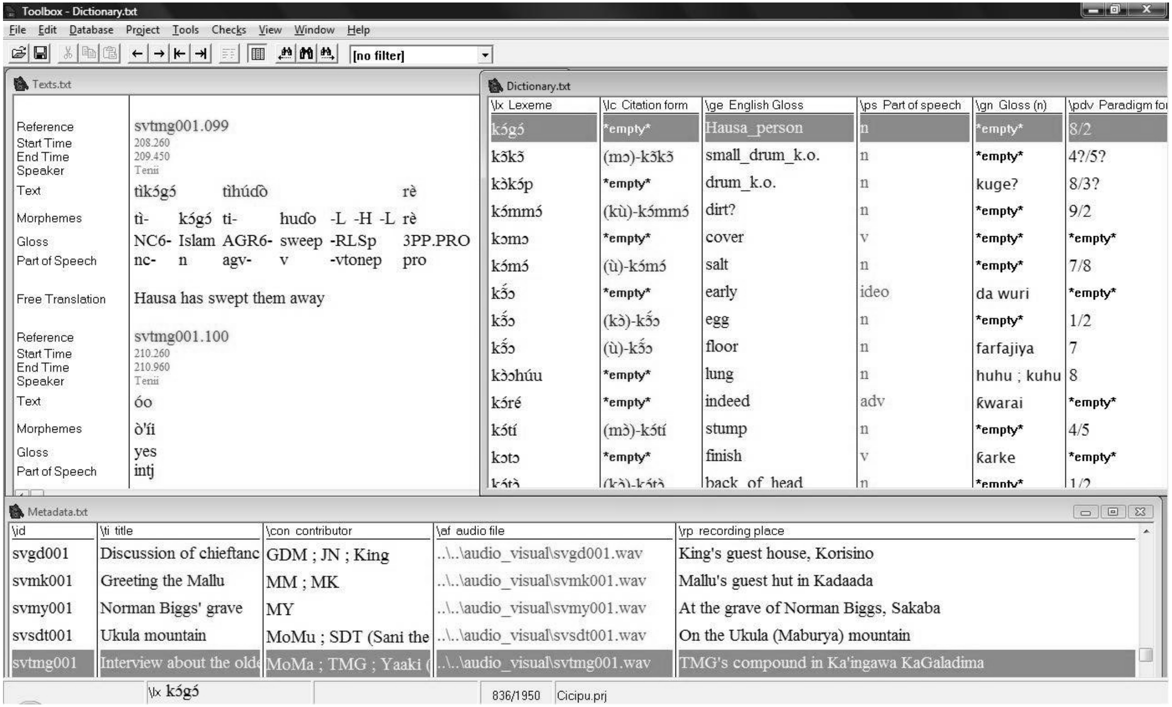
ft@Tenii: yes | Hausa has swept them away | yes

ref@Yaaki: svtmg001.097

tx@Yaaki: húuw tì ánà tìkɔ́gɔ́ kóo?

ft@Yaaki: they speak it like Hausa?

Figure 4: screenshot of Toolbox annotation and metadata files



Texts.txt

```
Reference          svtmg001.099
Start Time         208.260
End Time           209.450
Speaker            Tenii
Text               tìkɔ́gɔ́        tìhúɗò                     rè
Morphemes          tì-    kɔ́gɔ́  ti-     huɗo  -L -H -L   rè
Gloss              NC6-   Islam  AGR6-  sweep -RLSp       3PP.PRO
Part of Speech     nc-    n      agv-   v     -vtonep     pro

Free Translation   Hausa has swept them away

Reference          svtmg001.100
Start Time         210.260
End Time           210.960
Speaker            Tenii
Text               óo
Morphemes          ò'íi
Gloss              yes
Part of Speech     intj
```

Dictionary.txt

\lx Lexeme	\lc Citation form	\ge English Gloss	\ps Part of speech	\gn Gloss (n)	\pdv Paradigm fo
kɔ́gɔ́	*empty*	Hausa_person	n	*empty*	8/2
kɔ̃kɔ̃	(mɔ)-kɔ̃kɔ̃	small_drum_k.o.	n	*empty*	4?/5?
kɔ̀kɔ́p	*empty*	drum_k.o.	n	kuge?	8/3?
kɔ́mmɔ́	(kù)-kɔ́mmɔ́	dirt?	n	*empty*	9/2
kɔmɔ	*empty*	cover	v	*empty*	*empty*
kɔ́mɔ́	(ù)-kɔ́mɔ́	salt	n	*empty*	7/8
kɔ̃́ɔ	*empty*	early	ideo	da wuri	*empty*
kɔ̃́ɔ	(kɔ̀)-kɔ̃́ɔ	egg	n	*empty*	1/2
kɔ̃́ɔ	(ù)-kɔ̃́ɔ	floor	n	farfajiya	7
kɔ̀ɔhúu	*empty*	lung	n	huhu ; kuhu	8
kɔ́ré	*empty*	indeed	adv	ƙwarai	*empty*
kɔ́tí	(mɔ̀)-kɔ́tí	stump	n	*empty*	4/5
kɔtɔ	*empty*	finish	v	ƙarke	*empty*
kɔ́tɔ̀	(kɔ̀)-kɔ́tɔ̀	back_of_head	n	*empty*	1/2

Metadata.txt

\id	\ti title	\con contributor	\af audio file	\rp recording place
svgd001	Discussion of chieftanc	GDM ; JN ; King	..\..\audio_visual\svgd001.wav	King's guest house, Korisino
svmk001	Greeting the Mallu	MM ; MK	..\..\audio_visual\svmk001.wav	Mallu's guest hut in Kadaada
svmy001	Norman Biggs' grave	MY	..\..\audio_visual\svmy001.wav	At the grave of Norman Biggs, Sakaba
svsdt001	Ukula mountain	MoMu ; SDT (Sani the	..\..\audio_visual\svsdt001.wav	On the Ukula (Mabuṛya) mountain
svtmg001	Interview about the old	MoMa ; TMG ; Yaaki (	..\..\audio_visual\svtmg001.wav	TMG's compound in Ka'ingawa KaGaladima

understanding of language at a more abstract level, as a system of elements, rules, and constructions (see again Himmelmann 1998). Description and analysis can be seen as contingent by-products of documentation and will change and develop over time as research progresses (Woodbury 2003, Austin 2005). Documentary support for description can reduce the risk that it is sterile, opaque and untestable (as well as making it preservable for future generations and valuable for language support activities including revitalisation).

Austin & Grenoble (2007:22) point out that, in addition:

> [d]ocumentation projects must rely on the application of theoretical and descriptive linguistic techniques in order to ensure that they are usable (i.e. have accessible entry points via transcription, translation and annotation), as well as to ensure that they are comprehensive. It is only through linguistic analysis that we can discover that some crucial speech genre, lexical form, grammatical paradigm or sentence construction is missing or under-represented in the documentary record. Without good analysis, recorded audio and video materials do not serve as data for any community of potential users.

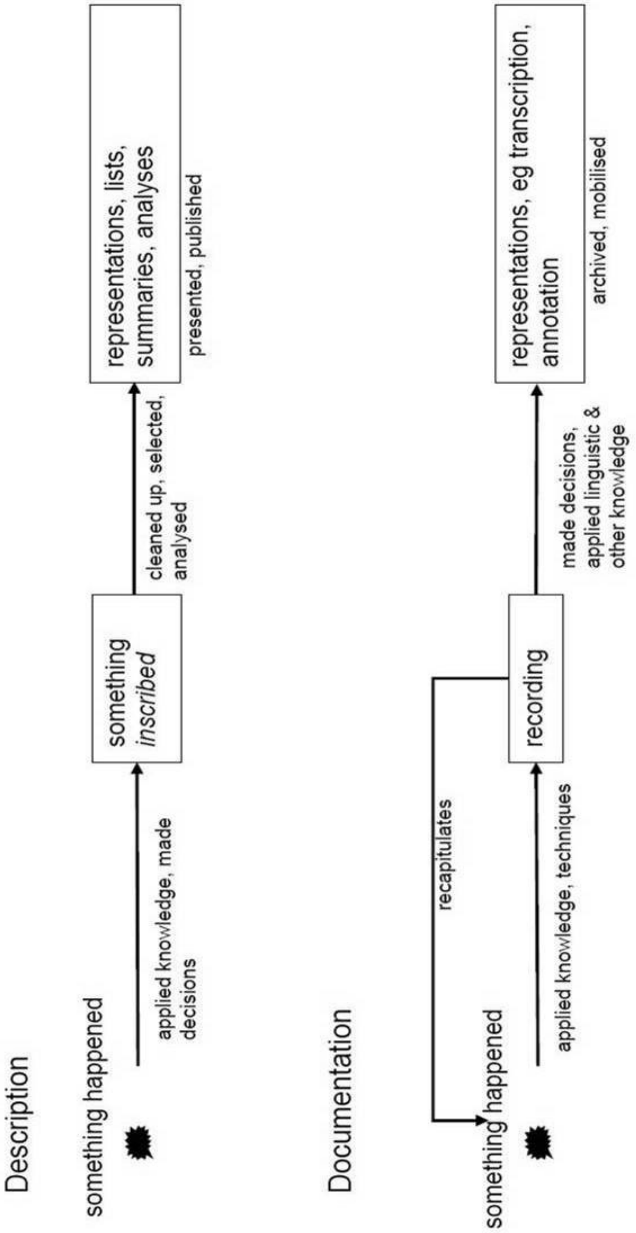
In other words, documentation and description are complementary activities with complementary goals and outcomes. In terms of workflow, they also differ. Figure 5 sets out the differences[2]:

- in description, linguistic knowledge and decision-making is applied to some event in the real world to make an inscription (e.g. an audio recording) that is not itself of interest[3] but serves as a source which can then be selected, analysed and systematised in order to create analytical representations, typically in the form of lists, summaries and analyses (e.g. statements about phonology, morphology or syntax). It is these representations which are the main focus of interest and which are then presented and distributed to users, typically other linguists;

[2] This figure owes much to input from David Nathan and Robert Munro.

[3] Nathan 2010a:267 argues that for many descriptive linguists audio is 'presently seen as an **inconvenience** on the way to transcription, annotation, selection or analysis'.

Figure 5: Workflow in description and documentation


Description
something happened
applied knowledge, made decisions
something *inscribed*
cleaned up, selected, analysed
representations, lists, summaries, analyses
presented, published
Documentation
something happened
recapitulates
applied knowledge, techniques
recording
made decisions, applied linguistic & other knowledge
representations, eg transcription, annotation
archived, mobilised

- in documentation, linguistic knowledge and documentary techniques are applied to some event in the real world to make a recording (audio or video) that recapitulates aspects of the original event (such as spatial relationships – see Nathan 2010a) and is itself a focus of interest (e.g. for archiving and preservation). In relation to the recording, the researcher makes decisions and applies linguistic and other knowledge to create representations, typically in the form of transcriptions, translations and annotations. These representations are the second major focus of interest and may be archived or mobilised, or otherwise used to meet language documentation and support goals. The representations could, of course, also be the input to the selection and analytical procedures of description, thereby linking the descriptive outcomes to the documentary corpus.

5. Some current challenges

There are a number of unresolved theoretical and practical issues relating to language documentation; in this section, I highlight just four of the challenges which face documentary linguistics today (see also Austin & Grenoble 2007): (1) the quality and quantity of the documentary record; (2) interdisciplinarity and cross-discipline collaboration; (3) meta-documentation; and (4) recruitment, training and sustainability. There are other issues which remain to be resolved and will undoubtedly emerge as practices and experiences of language documentation develop (see Harrison et al. 2008 for discussion of some results from the DoBeS documentation projects).

5.1 Quality and quantity of the documentary record

Language documentation is defined by Himmelmann 1998 as aimed at providing a ‘comprehensive record’ of a language or one of its varieties, but it is unclear how the quality or quantity of such a record could be determined[4].

As Nathan (2010b) notes, there is a tendency among some researchers to equate documentation outcomes with properties of archival objects (part of what he has termed ‘archivism’; see also Dobrin et al. 2007), e.g. the number and volume of recorded digital audio and/or video files and their related transcriptions and annotations. Clearly, mere quantity of archival files is not a good proxy for quality of research. Some commentators would argue that outcomes which contribute to language maintenance and revitalisation are

[4] See also Austin & Grenoble (2007:21).

better measures of the quality of a documentation project (i.e. we could ask what better success of an endangered language project could there be than that the language continues to be used). While there is growing interest in the creation of documentary corpora, and indeed professional bodies such as the Linguistic Society of America have recently passed resolutions urging Departments to take such corpora into account in determination of faculty appointments, and tenure and promotion decisions, it is not clear what parameters might be employed to determine the quality of a documentary corpus. One could imagine the following as possible metrics:

- compliance with some widely agreed standards in data and metadata representation – currently Unicode for character encoding and XML for text encoding are widely recognised as de facto standards in language documentation (and elsewhere), however there seems to be little other agreement about any possible standards and compliance. Certainly, such things as the GOLD ontology for interlinear glossing (see http://linguistics-ontology.org/) have been put forward as standards, but the community of documenters has been slow to adopt them;
- architecture of the data and modelling of the knowledge domain so that representations comply with some expressed data model and show internal and rigorous structural consistency;
- range and comprehensiveness of the data and analysis, in terms of such things as the genres present in a speech community as determined by a well-grounded ethnography of speaking;
- the ethical context of the project, including training and involvement of native speakers in the project outcomes, including the corpus.

In 2007 the Committeee on Endangered Languages and their Preservation (CELP) of the Linguistic Society of America was presented with a proposal for assessing ‘adequacy of documentation’ which proposed to answer the question for language documenters of ‘how do we know when we’re done?’ It suggested an adequate documentation should cover:

1. all the basic phonology, both low-level and morphophonemic
2. all the basic morphology
3. all the basic syntactic constructions (in context)
4. a lexicon which (a) covers all the basic vocabulary and important areas of special expertise in the culture, and (b) provides at least glosses for all words/morphemes in the corpus
5. a full range of textual genres and registers’

It offered a set of ‘accounting standards’ to determine adequacy, including quantitative measures such as a figure of 10,000 items for a lexicon, and a text corpus of 1 million words (around 1200 hours of recorded speech). Other qualitative measures were suggested such as: ‘[o]ne is done when nothing new is coming up in non-elicited material and when any apparent lacunae in the phonological system can be shown to be real and not an accident of data collection’. It is doubtful that for non-minority languages linguists would ever suggest it is possible to qualitatively and quantitatively determine when a research project is ‘adequate’ (has English been ‘done’ after so many years of work by some many linguists?), yet this is precisely what was suggested for language documentation, especially that involving endangered languages in particular.

5.2 Interdisciplinarity

Himmelmann and others have argued that language documentation fundamentally requires a multidisciplinary perspective potentially involving researchers, theories and methods from a wide range of disciplines, including linguistics, anthropology, (oral) history, musicology, psychology, ecology, applied linguistics, computer science and so on (see Harrison 2005, Coelho 2005, Eisenbeiss 2005 for examples). However, as Austin & Grenoble (2007:22) point out:

> in our experience, true interdisciplinary research, especially in teams carrying out fieldwork in remote locations, is difficult to achieve, both because of theoretically different orientations, and practical differences in approach (ranging from the trivial where linguists’ and anthropologists’ practices concerning payments for consultants traditionally have differed, to more significant differences in academic paradigm that make communication and understanding fraught).

Whether these differences of theory and practice can be resolved in meaningful ways remains an open question, and one that documentary linguistics needs to grapple with. Unfortunately, over the past 60 years mainstream linguistics has tended to turn away from these other disciplines and to emphasise its ‘independence’ by concentrating on theoretical concerns that are of discipline-internal interest primarily to linguists alone (Liberman 2007). Language documentation opens new doors to multi-disciplinary collaboration but we need to work out how to achieve it.

5.3 Meta-documentation

Documentary linguistics researchers have been clear that alongside the collection of data it is necessary to record metadata, data about the data, to ensure that its context, meaning and use can be properly determined. As Nathan (2010b:196) states:

> [m]etadata is the additional information about data that enables the management, identification, retrieval and understanding of that data. The metadata should explain not only the provenance of the data (e.g. names and details of people recorded), but also the methods used in collecting and representing it.

Notice that metadata is required not only for archiving but also for the very management, identification, retrieval and understanding of the data within the documentation project once the transfer process (see above) is undertaken and value-adding is to be done. The way files are named and structured in folders is itself a type of metadata (see Nathan 2010b, section 3.7), and as Nathan and Austin 2004 argue, any data added to the recordings (including transcription, translation, annotation etc.) should be seen as 'thick metadata' (contrasted with the 'thin' cataloguing metadata often promoted in discussions of language documentation, e.g. by the E-MELD School of Best Practice).

Nathan (2010b:196) also proposes that:

> [a]nother way to think of metadata is as meta-documentation, the documentation of your data itself, and the conditions (linguistic, social, physical, technical, historical, biographical) under which it was produced. Such meta-documentation should be as rich and appropriate as the documentary materials themselves.

If we extend the concept of meta-documentation to include as full as possible documentation of the documentation project itself (cf. Good 2010), then it is clear that the following aspects at least should be covered:

- the **identity** of the stakeholders that were involved and their roles in the project
- the **attitudes** of language consultants, both towards their languages and towards the documentation project;
- the **methodology** of the researcher, including research methods and tools (see Lüpke 2010), any theoretical assumptions encoded through things such as abbreviations or glosses, as well as relationships with

the consultants and the community (Good 2010 mentions 'the 4 Cs': 'contact, consent, compensation, culture');

- the **biography** of the project, including background knowledge and experience of the researcher and main consultants (eg. how much fieldwork the researcher had done at the beginning of the project and under what conditions, what training the researcher and consultants had received). For a funded project, the project biography would include the original grant application and any amendments, reports to the funder, email communications with the funder and/or any discussions with an archive, such as the reviews of sample data described by Nathan (2010b, section 3.3);
- any **agreements** entered into, whether formal or informal (such as a Memorandum of Understanding, payment arrangements, and any promises and expectations issued to stakeholders).

This kind of information is invaluable, not only for the researcher and others involved in a project, but also for any other future parties wishing to make sense of the project and its history and context. Unfortunately, linguists have typically been poor at recording and encoding this kind of information, meaning that work is often difficult with so-called 'legacy data', especially materials that only become available once the researcher has died (see Bowern 2003, Innes 2009, O'Meara & Good 2009). This is an area for further development within language documentation theory and practice.

5.4 Sustainability

One of the forces leading to the development of language documentation as a research field was the concern shown my linguists for the current threats to the world's linguistic diversity. While some work has been done on language revitalisation, that is, the theories and practices that can be developed and employed to strengthen the position and use of already endangered or moribund languages, little research has been carried out on how language documentation can contribute to sustaining endangered languages and the communities who want to maintain and develop them. International development practitioners such as Boven & Morohashi 2002 have argued for 'participatory development' practices that will sustain communities, however most of the research in this area remains unknown to linguists, as does 'resilence theory' (Van Breda 2001) employed in a range of fields, including development studies and social work, to help strengthen fragile communities and groups so they can weather threats to them, both internal and external.

A further dimension of sustainability of language documentation concerns issues relating to the available human resources:

- how can we recruit new contributors to the discipline since there are many more languages in need of documentation than there are researchers to document them? Post-graduate programmes such as SOAS, UH Manoa, and UT Austin, mentioned above, have attracted increasing numbers of students to join, however this is a narrow group of recruits and more and different contributors are needed;
- how can we train language documenters so that they become proficient (Nathan 2006:57-61) in the theory and practices of the field? Some experiments in this area have been carried out (Austin 2008) but more and better training is needed;
- how can we sustain these recruits through fulfilling career paths beyond their initial training, be it via post-graduate degrees or native-speaker training within a documentation project? Although increasing numbers of students are entering the field, there are not enough post-doctoral fellowships or academic jobs to employ them all at universities or research centres, and we risk the loss of committed and enthusiastic participants if they cannot find sustaining careers. Similarly, local participants, including native speakers, are increasingly receiving training and employment as part of documentation projects, however these projects typically have a 3-4 year life span and we need to work out how their interest and involvement can be sustained beyond the end of the project.

The issue of sustainability, in its various forms, will be one that challenges documentary linguistics for some time to come.

6. Conclusions

The past 15 years has seen the emergence and gradual development of a new field of research called documentary linguistics or language documentation. For many researchers and communities, especially those speaking endangered languages, the focus of language research has shifted to a new attention to recording, analysing and preserving records of language in use in ways that can serve a wide range of constituencies, particularly the speaker communities themselves. The field has come about due to a change in the vision of what the goals, methods and outcomes of linguistic research can be, changes in the relations between researchers and those whose languages they study, and has benefited from various developments in technology. A number of challenging issues will need to be addressed as documentary linguistics as a field matures further in the future.

References

Austin, Peter K. 2006. Data and language documentation. In Jost Gippert, Nikolaus Himmelmann and Ulrike Mosel (eds.) *Essentials of Language Documentation* (*Trends in Linguistics. Studies and Monographs*, 178), 87-112. Berlin: Mouton de Gruyter.

Austin, Peter K. 2007. Survival of Languages. In Emily F. Shuckburgh (ed.) *Survival: Darwin College Lectures*. Cambridge: Cambridge University Press.

Austin, Peter K. 2008. Training for language documentation: Experiences at the School of Oriental and African Studies. In Margaret Florey & Victoria Rau (eds.) *Documenting and Revitalising Austronesian Languages*, 25-41. Language Documentation & Conservation Special Publication No. 1. Hawaii: University of Hawaii Press

Austin, Peter K. 2010. Communities, ethics and rights in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 34-54. London: SOAS.

Austin, Peter K. & Lenore Grenoble. 2007. Current trends in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 4*, 12-25. London: SOAS.

Austin, Peter K. & Stuart McGill. 2010. *Essential readings in endangered languages*. London: Routledge.

Austin, Peter K. & Julia Sallabank. 2010. *Cambridge Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Bird, Stephen & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3):557-82.

Bond, Oliver. 2010. Language documentation and language typology. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7,* 238-261. London: SOAS.

Boven, K. & J. Morohashi. 2002. *Best practices using indigenous knowledge*. Paris: Unesco.

Bowern, Claire 2003 ‘Laves’ Bardi Texts’ Foundation for Endangered Languages. In Joe Blythe & M. Brown (eds.) *Maintaining the links: Language, identity and the land. Proceedings of FEL VII*, Broome, Western Australia: FEL

Coelho, Gail 2005. Language documentation and ecology: areas of interaction. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 3*, 63-74**.** London: SOAS.

Committee on Endangered Languages and their Preservation. 2007. Adequacy of documentation. Ms.

Dobrin, Lise, Peter K. Austin, & David Nathan. 2007. Dying to be counted: commodification of endangered languages in documentary linguistics. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 6*, 37-52. London: SOAS.

Eisebeiss, Sonja. 2005. Psycholinguistic contributions to language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 3*, 106-140. London: SOAS.

Gippert, Jost, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) 2006. *Essentials of language documentation* (*Trends in Linguistics. Studies and Monographs*, 178). Berlin: Mouton de Gruyter.

Good, Jeff. 2010. Documenting consent, access and rights. Presentation at LSA Annual Meeting OLAC workshop on archiving, Baltimore.

(available at http://www.ailla.utexas.org/site/lsa_archiving10.html)

Hans Rausing Endangered Languages Project 2006. What is language documentation?
Available at http://www.hrelp.org/documentation/whatisit/, accessed 2006-05-15.

Harrison, K. David 2005. Ethnographically informed language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 3*, 22-41. London: SOAS.

Harrison, K. David, David S. Rood & Arienne Dwyer (eds.) 2008. *Lessons from documented endangered languages*. Amsterdam: John Benjamins.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36:161-95.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Jost Gippert, Nikolaus P. Himmelmann and Ulrike Mosel (eds.) *Essentials of Language Documentation* (*Trends in Linguistics. Studies and Monographs*, 178), 1-30. Berlin: Mouton de Gruyter.

Innes, Pamela 2009 Ethical problems in archival research: Beyond accessibility. *Journal of Language and Communication*.

Lieberman, Mark. 2007. The future of linguistics. Plenary talk given at Linguistic Society of America Annual Meeting, Annaheim, California.

Lüpke, Friederike. 2010. Data collection methods in language documentation. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 55-104. London: SOAS.

Nathan, David. 2006. Proficient, permanent, or pertinent: aiming for sustainability. In Linda Barwick & Tom Honeyman (eds.) *Sustainable data from digital sources: from creation to archive and back*. 57-68. Sydney: Sydney University Press.

Nathan, David. 2010a. Sound and unsound practices in documentary linguistics: towards an epistemology for audio. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 262-284 London: SOAS.

Nathan, David. 2010b. Archiving and language documentation: from disk space to MySpace. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 172-208. London: SOAS.

Nathan, David and Peter K. Austin. 2004. Reconceiving metadata: language documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 2*, 179-187. London: SOAS.

O'Meara,Carolyn & Jeff Good. 2009. Ethical issues in legacy language resources. *Journal of Language and Communication*.

Sells, Peter. 2010. Language documentation and linguistic theory. In Peter K. Austin (ed.) *Language Documentation and Description, Volume 7*, 209-237. London: SOAS.

Van Breda, Adrian D. 2001. *Resilience theory: A literature review*. Pretoria, South Africa: South African Military Health Service. (available at http://www.vanbreda.org/adrian/resilience.htm)

Whalen, Doug. 2004. How the study of endangered languages will revolutionize linguistics. In Piet van Sterkenburg (ed.) *Linguistics Today – Facing a greater challenge*. Amsterdam: John Benjamins.

Woodbury, Anthony C. 2010. Language documentation. In Peter K. Austin & Julia Sallabank (eds.) *The Handbook of Endangered Languages*. Cambridge: Cambridge University Press.

Woodbury, Anthony C. 2003. Defining documentary linguistics. In Peter Austin (ed.), *Language Documentation and Description*, *Volume 1,* 35-51. London: SOAS.

Discussion questions

1. Why has language documentation developed as a subfield of linguistics in the past 15 years and how is it defined?

2. Go to www.hrelp.org/projects and look at some of the project descriptions – what kinds of topics are researchers working on in their ELDP-funded projects. Do these differ from projects funded by Volkswagen as part of the DoBeS programme. Do the research methods differ between the two programmes?

3. Sustainability has become a major concern of researchers in ecology, agriculture, development studies etc. Are any of the concepts and methods developed in those fields applicable to sustainability of language documentation?