The Cambridge Handbook of Endangered Languages

Edited by
Peter K. Austin
and
Julia Sallabank

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521882156

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data
The Cambridge handbook of endangered languages / [edited by] Peter K. Austin,
Julia Sallabank.
p. cm. – (Cambridge handbooks in language and linguistics)
ISBN 978-0-521-88215-6 (hardback)
1. Language obsolescence–Handbooks, manuals, etc. 2. Language and
languages–Handbooks, manuals, etc. I. Austin, Peter. II. Sallabank, Julia.
III. Title. IV. Series.
P40.5.L33C36 2011
409–dc22
2010051874

ISBN 978-0-521-88215-6 Hardback

9

Language documentation

Anthony C. Woodbury

9.1 What is language documentation?

LANGUAGE DOCUMENTATION is the creation, annotation, preservation and dissemination of transparent records of a language. While simple in concept, it is complex and multifaceted in practice because:

- its object, LANGUAGE, encompasses conscious and unconscious knowledge, ideation and cognitive ability, as well as overt social behaviour;
- RECORDS of these things must draw on concepts and techniques from linguistics, ethnography, psychology, computer science, recording arts and more;
- the CREATION, ANNOTATION, PRESERVATION and DISSEMINATION of such records pose new challenges in all the above fields, as well as information and archival sciences and;
- above all, humans experience their own and other people's languages viscerally and have differing stakes, purposes, goals and aspirations for language records and language documentation.

Language documentation, by this definition, is at least as old as writing. Attestations of the Homeric poems, for example, are records of what were for a long time verbal performances that reflect a once-held competence for a language and its use. Their inscription was a feat of linguistic analysis, and their passage to us through nearly three millennia a triumph of preservation and dissemination. For philologists, they have served as the basis for editions, retellings, translations, concordances, dictionaries and grammars. And these in turn have been valued by many as poetry, rhetoric, narrative and logic; and as history, politics, psychology, religion and ethnic ideology.

I am grateful to Christine Beier, Stuart McGill, Keren Rice and Julia Sallabank for their extensive comments on earlier drafts of this paper; and I acknowledge a debt to colleagues too numerous to name for discussion of the general issues of this chapter over many years.

Likewise, the spread of writing to vernacular languages, along with ideologies of language standardization and practices of manuscript curation, constituted language documentation on an enormous scale over a millennial time frame.

In this chapter my concern is language documentation as it applies to endangered languages. I take LANGUAGE ENDANGERMENT to be the en masse, often radical shift away from unique, local languages and language practices, even as they may still be perceived as key emblems of community identity. As applied to endangered languages, language documentation is accelerated, enlarged, popularized and transformed. More and more communities have sought documentation of their languages just as they slip away; the languages for which documentation is sought show ever more genetic and typological diversity from one to the next; and the communities themselves are usually small in terms of population. Thus, what once might have been accomplished as a national project by many people with specialized training over many years, is instead forced to happen in only a few years, village by village, at the hands of a few people, often with little or no technical training. For their part, many linguists and scholars in related fields have been inspired by the human and scientific dimensions of the issue, have called for renewed attention to language documentation, and along with it a substantial reordering of their own disciplinary priorities and practices (Dobrin and Berson, Chapter 10).

Before evaluating critically the scholarly and popular contexts of endangered language documentation, let us first draw some basic corollaries and concomitants of the idea, in order to be clear what is at issue. We begin with RECORDS of a language, the products of language documentation. Minimally, such records are any kind of preservable representation of lexico-grammatical form, typically WRITING; or any kind of preservable, real-time replica of speech, typically (nowadays) a VIDEO-RECORDING or AUDIO-RECORDING. To be TRANSPARENT, interpretable by its future perceivers, a record requires what philologists call an APPARATUS: systematic information about the creation and provenance of the record and of the event it represents (technically termed METADATA; see Conathan, Chapter 12, and Good, Chapter 11); and TRANSLATION(S) into other languages, including at least one language of wider communication. To this, nearly all linguist practitioners would add TRANSCRIPTION (or retranscription, if the original is written) in the terms of a scientific analysis of the lexico-grammar, so that the identities of sounds and lexical elements are systematically elicited from contemporary speakers and thence transmitted to future perceivers. Nevertheless, users interested in the records for reasons other than the analysis of the lexico-grammatical code may consider transcription expendable (as is standard, for example, in films with audio in one language and subtitles in another). Beyond the bare minimum (such as the metadata set specified by the Open Language

Archives Consortium (OLAC), see Good, Chapter 11), metadata can be enriched with primary and secondary commentaries and links to other records (Nathan and Austin 2004), and translations can be multiplied at different compositional levels, passage-by-passage, clause-by-clause, word-by-word, or bit-by-bit of meaning, and supplemented with commentaries and with commentaries on commentaries from different perspectives (Evans and Sasse 2007, Woodbury 2007).

Language documentation is still language documentation whether or not the records that are produced 'add up' in some way; nevertheless, we do well to explore the many different ways that sets of records could cohere. For example, a set of records resulting from an endangered-language documentation project could:

- be tailored to certain interests of community members, or of scholars of different kinds, or of publics variously conceived;
- be assembled so as to tell a specific story, like the images in a photographic essay;
- comprise samples of talk in a specific community regardless of the language, or follow just one lexico-grammatical code across several communities;
- comprise samples of different speakers, or speakers of different social categories, or sample different genres;
- comprise samples of purely naturalistic, fly-on-the-wall records, or records of talk that is staged in different ways, or both; and
- comprise samples of speech from one moment in time, or (with the right resources) a sample across time.

The sets of records, coherent or not, are often called LANGUAGE DOCUMENTATIONS; but since that is what we are calling the activity as a whole, I will call such sets LANGUAGE DOCUMENTARY CORPORA (or just CORPORA); and I will call the ideas according to which a corpus is said to cohere or 'add up' its (CORPUS) THEORIZATION. Corpus theorizations, and even principles for corpus theorization, can both offer a space for invention and become a matter of contention and debate; we will return to these in Sections 9.4.3 and 9.4.4.

Endangered-language documentation, the activity, can be an isolated occurrence, as when a person creates and keeps a few scraps or tapes, or when word lists are scribbled down during brief encounters, or when records emerge as by-products of other activities. But of special interest is the range of concerted, programmed documentary activities motivated by impending language loss and aimed at creating a final record. These activities raise issues of corpus theorization; but in addition, they raise questions about the participants, their purposes and the various stakeholders in the activity or programme of activity or project: we may refer to this set of questions as the PROJECT DESIGN (see Bowern, Chapter 23) of a language-documentation activity.

Regarding PARTICIPANTS (see Grinevald and Bert, Chapter 3, and Dobrin and Berson, Chapter 10), a language-documentation activity can be carried out by one or many people and raises questions of competence, capacity and entitlement: must a documenter be a native speaker (Ameka 2006), or at least a second-language speaker, of the language being documented? A community member or traditional political ally? A person making a common purpose or cause with some, many or all community members? Must a documenter be a linguist? An ethnographer? An oral historian, or specialist in verbal art, or ethnomusicologist, or educator? An audio- and video-recording artist and technician? An archivist? All of these, or at least one or several of them? (See further discussion in Section 9.4.2.)

Regarding PURPOSES, documentation can mean different things to different people. A project may be aimed at preservation, or revitalization, or the scientific study of language use or acquisition or grammar or lexical knowledge, or the reconstruction of linguistic or social history. It can be ideologically keyed to the establishment and maintenance of identity, or as a symbol of progress or global participation, or as art, reality, nostalgia or a general quest for knowledge; and this just scratches the surface.

Finally, regarding STAKEHOLDERS (see Austin 2003: 8–9, and Dobrin and Berson, Chapter 10), that is, who a project is for, and who takes part in shaping its design: can a project be conceived narrowly as just for the community being documented, or some sector of it, or just for science, or just for a generic wider public? Is there a compact among stakeholders that mediates among their different purposes, and how might those purposes intersect, or fail to intersect? And does being a 'stakeholder' of one sort or another give people equal say over how documentation is to proceed (or not proceed)?

9.2 How is documentation related to traditional academic projects and orientations?

As noted above, language documentation as defined here is as old as writing. But it has evolved considerably in the context of the massive, world-scale language contact of the past 500 years, leading to a scholarly discipline or framework now increasingly termed DOCUMENTARY LINGUISTICS, for which carrying out endangered language documentation has been the defining project or DISCIPLINARY CHARTER.[1] Much insight could be gained from a detailed study of the early origins and antecedents of documentary linguistics; but we will begin with Franz Boas, whose charter for ethnography encompassed a prototype of the modern notion of language documentation, and whose influence has been especially significant.[2]

Boas (1911: 59–73) saw the study of languages, including especially the collection of texts, as both a practical and a theoretical component in the study of aboriginal ethnography in the Americas: practically, as a way to obtain information about complex topics in contexts where neither investigators nor the most knowledgeable tribal members knew each others' languages well or at all; and theoretically because, in his view, much of the content of culture, e.g. rituals, oratory, narrative, verbal art and onomastics, was linguistic in nature. Furthermore, he considered linguistics itself a domain of ethnology, which he defined as 'the science dealing with the mental phenomena of the life of the peoples of the world' (Boas 1911: 63).

From a modern point of view, Boas's conception of language was both broad and interestingly free of dichotomization: there is no strong theoretical division between language use versus linguistic knowledge. There is an acknowledgement of a universal core of grammatical concepts, structures and categories, alongside an openness to areas where these may vary, and in the areas where they vary, an openness to both genetic inheritance and contact-based diffusion. In turn, his focus on particulars within this broadly conceived whole allowed for inferences about the histories of individual traits in preference to long-range, essentialist, all-or-nothing reckonings of the 'origins' of whole peoples, 'races', nations or cultures.

Despite this aversion to line-drawing in a theoretical sense, he advocated the creation of texts, grammars, and dictionaries (the so-called Boasian trilogy or triumvirate) as his theorization of language documentary corpora, as in Boas (1917: 1):

> We have vocabularies; but, excepting the old missionary grammars, there is very little systematic work. Even where we have grammars, we have no bodies of aboriginal texts … it has become more and more evident that large masses of texts are needed in order to elucidate the structure of the languages.

All three were interrelated parts of a documentary whole, treating, in different ways, overlapping empirical domains; and it would be a mistake to project from any one of these a specific theoretical domain or level of analysis.

The *International Journal of American Linguistics* (IJAL), together with university and museum monograph series, were to be the archiving mechanism for such corpora. For example, in IJAL's second year Speck (1918) published a 58-page collection of Penobscot texts with interlinear and free translations, the first of many text publications in IJAL's early years. Lexicons and sketch grammars were likewise published. Moreover, field notes were frequently archived for posterity (for example, in the extensive collection of original field notes on Native American languages at the Library of the American Philosophical Society, Philadelphia, which includes many of Boas' own notes).

Boas' corpus theorization included a broad view of so-called texts. He chafed at the limitations imposed by dictation (Boas 1917: 1):

> The slowness of dictation that is necessary for recording texts makes it difficult for the narrator to employ that freedom of diction that belongs to the well-told tale, and consequently an unnatural simplicity of syntax prevails in most of the dictated texts.

He was somewhat happier with texts written directly by native speakers. But still he complained:

> On the whole, however, the available material gives a one-sided presentation of linguistic data, because we have hardly any records of daily occurrences, every-day conversation, descriptions of industries, customs, and the like. For these reasons the vocabularies yielded by texts are one-sided and incomplete. (Boas 1917: 2)

He later elaborates:

> The problems treated in a linguistic journal must include also the literary forms of native production. Indian oratory has long been famous, but the number of recorded speeches from which we can judge their oratorical devices is exceedingly small. There is no doubt whatever that definite stylistic forms exist that are utilized to impress the hearer; but we do not know what they are. As yet, nobody has attempted a careful analysis of the style of narrative art as practiced by the various tribes. The crudeness of most records presents a serious obstacle for this study, which, however, should be taken up seriously. We can study the general structure of the narrative, the style of composition, of motives, their character and sequence; but the formal stylistic devices for obtaining effects are not so easily determined. (Boas 1917: 7)

He also advocated the study of other kinds of speech, including song words, speech distortion and play, and ritual language. Clearly the lack of practical recording techniques impeded this programme, but not for any want of basic conception; indeed his conception prefigures the current mainstream as will be described below.

For Boas, linguistics was one of four anthropological fields (alongside archaeology and physical and cultural anthropology) for which students were to receive training; and at least as carried out by Boas, this represented a robust, if ultimately temporary, disciplinary establishment of linguistic documentation. A further important feature, again prefiguring contemporary practice, was Boas's personal commitment to the training of native speakers as documenters: George Hunt produced volumes of written text material in Kwakwa̱ka̱'wakw (discussed critically in Briggs and Bauman 1999); Ella Deloria co-authored with him a grammar of Dakota (Boas and Deloria 1941); and Zora Neale Hurston (1935) collected

texts in African-American communities in Florida and elsewhere which were published as folklore and literature.

As structuralism (including eventually generative structuralism) took hold, linguists increasingly distinguished lexico-grammatical systems from language use and a subtle but important retheorization (or indeed detheorization) of documentary corpora and of the traditional text–dictionary–grammar trilogy took hold. The relationship, originally mutually reinforcing, becomes hierarchical: texts, elicited data, judgements and other exemplifications of use are the 'raw data' which allow for the extraction of lexical information for the dictionary and grammar. The dictionary generalizes over the lexical knowledge presupposed in the texts, and the grammar generalizes over the categories and relations instantiated in the texts and presupposed in the presentation of the dictionary: the sound system, general morphophonemics, word structure, parts of speech, the system of regular inflection, phrase and sentence formation and the like. This left texts and other 'raw data' corpora untheorized, except as they might inform the dictionary and grammar; for example, Samarin (1967: 46) pithily calls only for the publication with a grammar of 'enough texts to permit a verification of the analysis'. With text and other 'raw data' documentation theorized so narrowly, the grammar and dictionary themselves remain as documentation of internalized speaker knowledge or of a shared system, which in turn serves ends even higher on the hierarchy, including genetic classification, typology, or the testing of cross-linguistic theories. Alternatively, grammars and dictionaries might be recognized themselves as a level of analysis, in which case there is nearly nothing at all that is theorized as a documentary corpus, rendering texts or other data as epiphenomenal.

This made it possible to pursue grammar in a more or less non-documentary framework (see Himmelmann 2002: 3–4 for an analysis of how this approach came to be known as DESCRIPTIVE). Grammars were often published without texts, and the data in grammars were not always sourced or even drawn from texts at all. But even more significantly, grammars themselves became less highly valued within the disciplinary economy, so that by the 1980s there were debates as to whether the 'compilation' of a grammar could even serve as a doctoral dissertation in many major linguistics departments, while the work of grammar writing was, in significant measure, abandoned by secular academic linguists to their missionary colleagues, especially members of SIL International. More commonly, grammatical analysis was pursued in the context of typology and theory, presented in article-length works, and was often not even founded on systematic lexicographic analysis, let alone documentary records curated for long-term preservation or easy access.

Nevertheless, even as theoretical perspectives changed, there was a remarkable degree of persistence of the Boasian theorization, or at least of its main procedures, among Americanists. Both Edward Sapir, a

student of Boas, and Leonard Bloomfield produced voluminous Boasian-style documentation despite their vanguard roles in the theoretical shifts. At mid century, Murray Emeneau and Mary Haas, both students of Sapir, presided at the University of California at Berkeley over a veritable factory of graduate students who produced Boasian grammar–dictionary–text trilogies published by the *University of California Publications in Linguistics*. These texts were linked to audio-recordings which, along with field notes and slip-files, were archived with the Survey of California Indian Languages. Michael Krauss established in the 1970s, as part of the Alaska Native Language Center at the University of Alaska, Fairbanks, 'an archival library whose holdings include almost every printed document, and much of the unpublished material, that has been written in or on an Alaska Native language' (Krauss 1980: 31–2), and which also included a significant sound archive.

Many scholars' life work touched every corner of Boasian practice. Consider, for example, Knut Bergsland's documentation of Aleut, a language spoken mainly by small pockets of elders on the Alaska Peninsula, the Aleutian chain, and some nearby islands. His documentary *oeuvre* begins with a dense philological presentation of Atkan and Attuan Aleut materials, including a catalogue of documentary sources, an exegesis of proper names, and interlinear texts from century-old sources written by Aleut church men, and from people whom he and others audio-recorded directly (Bergsland 1959). He continues, after a series of analytic, theory-oriented articles in the 1960s and 1970s, with:

- a pedagogical dictionary and grammar co-authored with Moses Dirks, a native speaker of Atkan Aleut (Bergsland and Dirks, 1978, 1981);
- a 715-page edition (with free translation only) of the texts written or wire-recorded by Waldemar Jochelson's expedition in 1909–10 (Bergsland and Dirks 1990);
- a 739-page dictionary covering all Aleut varieties, with extensive text-keyed exemplification and coverage of stems, productive derivation and special lexical areas including place names keyed to maps of the entire Aleut territory, technical terminologies and loans (Bergsland 1994);
- a 360-page grammar also covering all varieties and with copious text-keyed exemplification, often long and complicated (Bergsland 1997); and
- a monograph analysing and interpreting personal name data gathered by the Billings Expedition in 1790–2 (Bergsland 1998).

Although informed by post-Boasian structural linguistics, the philology, the breadth of focus, the interleaving of text, dictionary and grammar, and the concern for speaker training, represent a magnificent (and exceedingly brilliant) rendering of Boasian documentary theorization, decades after it stopped being forcefully articulated within

the international linguistic mainstream. Moreover, to the extent that endangered-language research was pursued, Bergsland's Boasian orientation was hardly atypical.

Despite the generally counterdocumentary trend in the mainstream of linguistics from the 1950s onwards, there nevertheless were contexts in which something like the Boasian theorization of texts was taken up and elaborated. The ETHNOGRAPHY OF SPEAKING had as its charter the creation of comprehensive, grammar-like descriptions of language use in speech communities, notably defined by Gumperz (1962) as 'a social group which may be either monolingual or multilingual, held together by frequency of social interaction patterns and set off from the surrounding areas by weaknesses in the lines of communication' (see also Michael, Chapter 7, Dobrin and Berson, Chapter 10, and Spolsky, Chapter 8). Within this framework, descriptions were to be made in terms of parametric categories such as WAYS OF SPEAKING, FLUENT SPEAKERS, SPEECH COMMUNITY, SPEECH SITUATION, EVENT and ACT, and such components of speech itself as MESSAGE FORM (including language or code) and CONTENT, SETTING, SCENE, GOAL, CHANNEL, and PARTICIPANTS (Hymes 1974a: 45–58). This theorization at least implied balanced, selective documentary corpora. But it also would be fair to question whether documentation itself was its goal: writings on the ethnography of speaking were not explicit about methods for record creation, annotation, archiving or dissemination, nor was systematic grammatical investigation a part of the programme, especially given the focus on community rather than code per se. Thus work done within the framework approached these issues in a range of ways, and in this respect, many practitioners notwithstanding, the ethnography of speaking mirrored the structural linguistics of the same period.[3]

By the 1970s, with language documentation receded from the limelight and its parts parochialized, redefined or repurposed, several trends began emerging, most noticeably among students of endangered indigenous languages of the Americas and Australia. First and foremost, their work was conducted in a context of heightened concern, awareness and activism by both communities themselves, and outside linguists (Alvarez and Hale 1970, Hale *et al.* 1992, Krauss 1980, Wilkins 1992). Second, attention was increasingly drawn to the role of FIELD LINGUISTICS in exploring new structures relevant to theoretical concerns (e.g. Dixon 1972, 1976, Hale 1975, 1983), and from there to the methods of field linguistics, including the role of texts in language documentation (Heath 1985). Third, a reconceptualization of language documentation as a unified field of endeavour in its own right was under way (notably Sherzer 1987), challenging its subservience to more specialized kinds of inquiry in both ethnography and linguistics:

> Both linguists and anthropologists have traditionally treated discourse as an invisible glass through which the researcher perceives the reality

of grammar, social relations, ecological practices, and belief systems. But the glass itself, discourse and its structure, the actual medium through which knowledge (linguistic and cultural) is produced, conceived, transmitted, and acquired, by members of societies and by researchers, is given little attention. My stance here is quite different from the traditional one, and reflects a growing interest in discourse in many disciplines. I view language, culture, society, and the individual as all providing resources in a creative process which is actualized in discourse. In my discourse-centered approach, discourse is the broadest and most comprehensive level of linguistic form, content, and use. This is what I mean by saying that discourse and especially the process of discourse structuring is the locus of the language–culture relationship. Furthermore, it is in certain kinds of discourse, in which speech play and verbal art are heightened, as central moments in poetry, magic, politics, religion, respect, insult, and bargaining, that the language–culture–discourse relationship comes into sharpest focus and the organizing role of discourse in this relationship is highlighted. (Sherzer 1987: 305–6)

In this view documentation itself becomes the goal, and yet is theorized well beyond sampling or surveying by proposing a special status for verbal art and speech play. It is also essentially open as to whether documentation is community-based or lexico-grammatical code-based, leaving that to depend on participants' goals. Sherzer furthermore makes proposals for text transcription, annotation, and analysis that build on ideas of Jakobson (1968) and others for seeing linguistic constructs as primes in discourse artistry. This raises representation from an ad hoc matter to a theoretical issue.

All these trends among scholars of endangered languages were ingredients in a wholesale revival of interest in documentation; and they did not arise in a vacuum. There was a continuing increase of interest in language preservation by communities undergoing shift, a growing intellectual focus on diversity and diversity issues in linguistics more widely (Nichols 1992) and on neo-Whorfian approaches in linguistic anthropology (Lucy 1992), and rapid advances in computational archiving and analysis of linguistic mega-corpora, exemplified by the work of the Linguistic Data Consortium at the University of Pennsylvania, among others.

By the mid 1990s an explicit disciplinary ideology and set of practices for endangered-language documentation had emerged, from which point it underwent rapid, global institutionalization as an academic discipline or framework, termed DOCUMENTARY LINGUISTICS (or DOCUMENTARY AND DESCRIPTIVE LINGUISTICS). An important workshop organized by David Wilkins at the Max Planck Institute for Psycholinguistics in Nijmegen, The Netherlands, in October 1995 asked what the 'best record' of a language would look like. This is a question that can only be raised

when documentation itself, rather than specific lines of linguistic or social inquiry, is the goal of study. The responses received included a number of disciplinary manifestos (Himmelmann 1998, Lehmann 2001, A. Woodbury 2003), all of which in one form or another pointed to the relative neglect of documentation and proposed documentation-centred approaches. Himmelmann in particular argued for a stronger division between documentation and description. He considered the creation and preservation of multipurpose records to be an endeavour distinct from dictionary-making and grammar-writing, although interrelated, and he argued that the contents of a documentation must be theorized beyond the immediate needs of description. In particular, dictionaries, grammars and perhaps ethnographies of speaking were to evolve as part of the documentary apparatus, much as in Boasian times or in the philological practice of even earlier, and even the conduct of paradigm elicitation or of a psycholinguistic experiment was to become amenable to treatment as a raw event recordable in the same ways as more traditional narrative or conversational texts.

Himmelmann (1998, elaborated somewhat differently in Himmelmann 2006a) also offers a quite specific format for a 'documentation', a comprehensive documentary corpus that focuses on a speech community (and thus not necessarily a single code) and includes a corpus of recorded events, lexical or other databases, and notes. It also includes metadata, commentary and annotation for these records, as well as general metadata about the community and, optionally, a descriptive grammar, ethnography and dictionary. This basic idea served as a charter for what was to be the first of a major series of funding efforts for endangered-language documentation, Dokumentation Bedrohter Sprachen, funded by the Volkswagen Foundation, the results of which are archived at the DoBeS Archive in Nijmegen.[4]

This was followed by other efforts in Canada (the Community-University Research Alliance and the Aboriginal Research Programme, both sponsored by the Social Sciences and Humanities Research Council of Canada), Japan (Vanishing Languages of the Pacific Rim, funded by the Japanese Ministry of Education), the UK (the Hans Rausing Endangered Languages Project (HRELP), funded by Arcadia Trust, which includes new granting, archiving, and academic programmes), and the US (the Documenting Endangered Languages (DEL) programme of the National Science Foundation and the National Endowment for the Humanities); and two smaller private charitable endeavours, the Foundation for Endangered Languages (FEL) in the UK, and the Endangered Language Fund (ELF) in the US. Each of these has led to projects animated by different conceptions of documentation: Pacific Rim and DEL with more allowance for grammars and dictionaries as documentation, HRELP with a strong recorded-text emphasis but allowing a more heterodox range of theorizations and project designs, and FEL and ELF with an emphasis

on documentation in the context of community-driven language-preservation efforts.

Alongside these initiatives, archiving projects were developed. Some, as already noted, arose in conjunction with documentation projects, and some were continuations of older archives, including the archives of the Smithsonian, the Survey of California Indian Languages, Alaska Native Language Center, and SIL International. Yet others were new, regional initiatives such as the Archive of the Indigenous Languages of Latin America (AILLA) and the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). All arose as part of an effort to organize and coordinate the archiving and cataloguing of endangered language documentation, including (see Nathan, Chapter 13 for the abbreviations): OLAC, IMDI, E-MELD and DELAMAN. Not surprisingly, the professional practices of archivists were unfamiliar to field linguists, especially given the wider, anti-documentarian trends mentioned; and even more so because of uncertainties about the nature and vulnerabilities of electronic documentation. Bird and Simons (2003) is an important general statement of the problems of making documentation last (see also Nathan, Chapter 13, and Good, Chapter 11).

Finally, documentary linguistics developed a general literature dealing with a wide range of issues, including a comprehensive handbook with thoughtful treatments of a set of issues and practices (Gippert *et al.* 2006), and important collections on field work (Newman and Ratliff 2001a) and grammar writing (Ameka *et al.* 2006, Payne and Weber 2007). Beginning in 2003, the yearly series *Language Documentation and Description*, edited by Peter K. Austin and published by SOAS, covered conceptual and technical questions, with special focus on community capacity building, multidisciplinary cooperation, archiving, documenting language variation and contact, meaning and translation, and literacy issues. The *Language Archives Newsletter* edited by David Nathan (SOAS), Peter Wittenburg and Paul Trilsbeek (MPI Nijmegen) and Marcus Uneson (Lund) published ten online issues between 2004 and 2007, covering technological and related issues. The year 2007 saw the founding of the online journal *Language Documentation and Conservation*[5] which publishes papers and reviews on a wide range of topics.

In summary, an influential academic charter for language documentation was articulated by Boas a century ago, leading in several important intellectual directions even as academic and popular attention to language loss waned. Documentary linguistics in its modern academic sense is an ambitious rewelding of the splintered pieces of the Boasian framework for language study. It arose as speaker communities and linguists drew more attention to language endangerment, as scientific interest in linguistic diversity renewed, as information management technologies burgeoned, and as linguists began to engage with the social and ethical dimensions of their work (see Dobrin and Berson, Chapter 10).

Documentary linguistics is new enough, however, that its scope, its scientific and humanistic goals, its stakeholders, participants, and practices are still being explored and debated both inside and outside academic contexts. After examining endangered language documentation in community contexts in Section 9.3, we will consider how a broad, inclusive idea of endangered-language documentation might be framed in order best to realize its potential, avoid pitfalls, and meet its challenges.

9.3 How is documentation related to community and other non-academic endangered language projects and perspectives?

We have considered the context and development of endangered-language documentation in academic research. But the stakeholders in documentation include the communities in which endangered languages are spoken. They may also include a wider array of publics with interests of different kinds, including friendship, literature, music, science, tourism, entertainment, nationalism, education policy, literacy, economic relations, development, subsistence, land acquisition, law enforcement, military conscription and religious conversion; and that only scratches the surface. In what follows I will focus mainly on documentation and community stakeholders and will have little to say of wider publics except as they may form part of the community context of endangered-language documentation.

In principle, the documentation of any language might be interesting or useful to anyone. If it is one's own or ancestral language, there may be special dimensions of identity, territory, spirituality, aesthetics, utility or nostalgia. And if one's own language is endangered, these may all be amplified and bring to the fore further personal, political, scientific and humanistic questions. Documentation, so far as it exists, is almost always a matter of interest in endangered-language communities, and sometimes a matter of controversy.

Certainly the form of language documentation longest at issue in endangered language communities has been writing, and through it the creation of indigenous language literatures, while audio- and video-documentation is of a more recent vintage. It is often the case that the development, support and teaching of writing is central in community-based language preservation programmes and a powerful emblem for language loyalty (see Lüpke, Chapter 16), whereas electronic documentation is often less debated or even accessible (see Bennett (2003) and Hinton (2001a) for some perspective, and Holton, Chapter 19).

Nevertheless, it is important to recognize that, within the wider realm of language activism, language planning and language revitalization and maintenance, documentation (written or otherwise) need not play

a central role or even any role at all (Grounds 2007). And, when it does play a role, it will not necessarily be created, preserved or disseminated in ways that are professionalized. Rather, we find individuals in communities, insiders or outsiders, needing to make a case for documentation in relation to wider ideologies and aspirations for language and community, and doing so on the basis of their own conceptions of what documentation is, and how different forms of it are created, preserved and valued. For example, Hinton (1994a) (see also Conathan, Chapter 12) tells vividly the story of John Peabody Harrington, a dogged, skilled, but imperious language documenter of the early and mid twentieth-century American west who left almost a million pages and many recordings of material on over ninety languages; and how his work was later appreciated as it became central in community projects in California to reawaken languages falling out of use (see Hinton, Chapter 15, on the *Breath of Life Workshops* that trained community members on how to access and use such archival materials).

An excellent handbook by linguists for language maintenance and preservation work, with surveys of such efforts around the world, is Hinton and Hale (2001) (see also Grenoble and Whaley (2006) with partly similar goals). Its major section headings include language policy, language planning, immersion in schools and in private settings, and training, none of which centrally involve documentation; documentation comes up directly in a section on media and technology and, implicitly, literacy. The intellectual diversity of community-based (and often school-based) language revival and preservation work is evident in the edited proceedings of a continuing series of language revitalization conferences at Northern Arizona University, organized by Jon Reyhner (Burnaby and Reyhner 2002, Cantoni 1996, Reyhner 1997, Reyhner and Lockard 2009, Reyhner *et al.* 1999, 2000, 2003). Here too, documentation is but one among many tools, with a different range of uses in different conceptions and approaches to language revitalization.

Wilkins (1992), Hill (2006), Rice (2009) and Grenoble (2009b) describe and discuss differences in worldviews and agendas among speakers and non-speaker linguists. The differences can be described at times as irreducible; and seem the more so when individuals talk past each other. They argue for ethnographic awareness and openness to different goals and purposes on the parts of linguists, and for flexibility in designing projects that meet participants' goals (Grenoble, Chapter 2, Hinton, Chapter 15). This is a stance which I will pursue in Section 9.4.1 below.

J. Hill (2002), Dobrin (2008), and Dobrin *et al.* (2009) further turn the spotlight on presuppositions, assumptions, attitudes and discourses held by outside researchers, often in the deep background, that can affect their interactions in communities even with the best of intentions. J. Hill (2002) assesses metaphors of languages as riches that are enumerated, pointing out uncomfortable connections of this rhetoric with colonial

extirpation. Dobrin *et al.* (2009) pursue the question of enumeration and its reductionistic extension, in academic discourses, to the valuation of documentary corpora. Dobrin (2008) discusses the high valuation of autonomy and self-determination in outsiders' assessments of the common good in their interactions, at times failing to parse their relationships in more locally familiar value systems, including especially systems where exchange is highly valued.

Finally, a key factor in any documentation, including endangered-language documentation, is the danger it may present, particularly in communities that feel marginalized and vulnerable at many levels (Conathan, Section 12.5). If data is collected and preserved for wide use and is genuinely multipurposed, then who is to stop it from getting into the wrong hands? Whereas from a utilitarian viewpoint this can be framed simply as leading to 'limitations' on documentation (Himmelmann 2006a: 16–17) and can be responded to within an ethical framework of INFORMED CONSENT, it must be seen in connection with the growing ubiquity of electronic record-making and surveillance in the contemporary world, and the kinds of trade-offs they present between having your voice heard or forging wider connections on the one hand, and a loss of control or fear of 'digging your own grave' on the other. Responses to this dilemma may differ from person to person, community to community, and time to time; but as technology continues to change no one ever fully imagines, let alone becomes 'informed,' of all the possible long-term effects of electronic record-making.

9.4 Toward a broad and inclusive view of endangered language documentation

The formulation of endangered-language documentation that I have presented is intentionally broad. At its core is the impulse people may have to make and keep records of languages that are falling into disuse through rapid language shift. That motivation serves many different goals and constituencies, and gives rise to many different disciplinary charters and programmatic approaches, as noted in the previous two sections. I defend a broad formulation because, even amidst such evident heterogeneity, there is a danger and even a tendency for individuals to establish and stipulate more specific practices aimed at just the situations they are most accustomed to, losing track of the greater whole.

In view of this, I think one of the principal functions of DOCUMENTARY LINGUISTICS, or better, of the whole patchwork of enterprises chartered in some way to see to or use endangered-language documentation, is to try to know, understand, acknowledge, analyse, and coordinate the enterprise in all its various, multifaceted forms. In some respects, this draws on particular bodies of expertise; for example, linguists have expertise

in areas like transcription and lexical documentation, and archivists in record organization and preservation that are likely to be broadly applicable, whatever brand of documentation is to be undertaken. But it also requires an imagination for difference of purpose and aspiration, a flexibility in understanding data structuring and management and using expert tools (Good, Chapter 11), an awareness of quite different ideologies of language and speaking among academics and non-academics alike, and an openness to the social complexities of what often are radically multicultural projects (Dobrin and Berson, Chapter 10). My sense of the current state of affairs is reflected well by Dobrin *et al.* (2009: 45) when they write, in a somewhat more specific context:

> Even despite systematizing conceptual efforts within linguistics, such as Himmelmann's [1998] careful distinguishing of description from documentation, a set of agreed upon principles of language documentation with associated methods does not exist. The resulting questions that this leaves open are fundamental: are our goals activist or scientific? Is documentation a research activity, or is it more closely aligned with art and practice of creative media? Does our data consist of symbols or of audio and video? How should archives prioritise dissemination across the potential constituencies they serve (academics of various persuasions, speaker communities)? On what basis could we decide?

And I am very much in sympathy with their conclusion:

> Resolving the tensions we have been describing will require an approach to documentation that is more closely tied to the guiding vision that continues to attract linguists to the language endangerment problem. However, this goal is not well served by a totalising theory that distinguishes documentary work from the rest of linguistics as a distinct and separate entity (Himmelmann 1998, cf. Austin and Grenoble 2007). Linguistics already has theoretically-informed ways of comparing languages for a host of reasons that are orthogonal to their moral value, and it is by distancing themselves from these that documentary linguists have been led to ask confused and unproductive questions such as 'how do we know when to stop documenting?' or 'how many recording hours should I put in the archive?' (page 45)

Dobrin *et al.* (2009: 46–7) continue:

> What is needed instead is an explicit recognition that the singularity of languages is irreducible, and that the methods used to study them must be singular as well. Each research situation is unique, and documentary work derives its quality from its appropriateness to the particularities of that situation. Rather than approaching endangered languages with preformulated standards deriving from their own

culture, documentary linguists must strive to be singularly responsive – both to what is distinctive about each language as an object of research, and to the particular culture, needs, and dispositions of the speaker communities with whom their work brings them into contact.

In this spirit, I wish to consider some ways documentary linguistics, in the broadest, most inclusive sense, might best realize its potential and address its challenges and pitfalls. To be more concrete, I will frame my discussion always with an eye to different PROJECT DESIGNS, and their accommodation and coordination within the larger whole.

9.4.1 Coordinating academic, community, and popular agendas for the design of documentation projects

As is clear from Sections 9.2 and 9.3 above, it is usually linguists who initiate systematic language-documentation projects, and as linguists, whether community members or not, we design and propose projects organized around our disciplinary agendas. But, as noted, various linguists have argued for ethnographic awareness and openness to different goals and purposes and for flexibility in designing projects that meet participants' goals. Wilkins (1992: 186) sums up as follows his PhD fieldwork 'under aboriginal control' after setting aside his linguist's agenda:

> It would be misleading, indeed it would be a boldfaced lie, to claim that my approach to learning, documenting, and building up a picture of Mparntwe Arrernte grammar was very systematic (especially from the point of view of an idealized, and generally antiseptic, field methods course) … [T]he development and directions of my research have not been independent of the changing developments and demands of my research for the Yipirinya School. I have just pointed out that much of the research for the school was joint research done by a team of people. The data-gathering techniques used for any individual project were established through conferencing among members of the group and through consultation with the Yipirinya Council … In these cases, then, I had input to, but did not determine, the research methods which would yield the information which I was to work with.

Among the advantages Wilkins cites for his approach were the opportunity to put linguistic skills to practical use, to gain a deeper knowledge of the language and its context, to have better access to community members, and to benefit from collaboration and teamwork (see Bowern, Chapter 23).

I see this approach as both a blueprint for the design of projects in an inclusive documentary linguistics, but also an institutional challenge to linguistics. It operates not only in practical terms but intellectual terms

too (see Woodbury 2010 for more discussion). On a practical level, agendas and activities are guided by goals, enthusiasms and capacities of all those involved. Linguists must be flexible and inventive (and their institutions more tolerant) about how and when to accomplish traditional linguistic agendas, and training takes centre stage as projects involve many people, with different expertise, roles and levels of training (Jukes, Chapter 21). On an intellectual level, documentation and related linguistic discovery require serious attention to language ideology (how people conceive of language and speaking, and relate them to territory, identity, aesthetics, spirituality and other domains), and to the relationship of language and context; and a focus on speaking in given communities problematizes a disciplinary focus on specific lexico-grammatical codes.

To the extent I am able, I try to keep sight of these issues in all of what follows.

9.4.2 Participants and training

Modern endangered-language documentation projects depart notably from earlier efforts, in involving interdisciplinary teams rather than a lone linguist. DoBeS made it a requirement to compose teams not only of linguists of various expertise, but other specialists such as ethnographers, musicologists, videographers and the like, and Himmelmann (2006a: 15) points out the need for such expertise and the unlikelihood that just one linguist can adequately handle it all (see also Austin 2007; Jukes, Chapter 21). Meanwhile, team participation has become a feature of documentary training in many university contexts (e.g. Woodbury and England 2004), and teams increasingly train community members as researchers or as trainers for other community members. For example, Kaufman *et al.* (2001) undertook a survey in 2007–9 of Zapotecan languages under the sponsorship of the Mexican Instituto Nacional de las Lenguas Indígenas that trained about twenty young speakers of Zapotec and Chatino languages to conduct surveys and transcribe and translate the interview. Our Chatino project, based at University of Texas at Austin and funded by ELDP, recruited the Chatino trainees from among people already working with us, making possible a 'ladder' of academic experiences.

This is clearly a way in which documentary linguistics has in its sights its potential for broadness and inclusivity. Yet it still raises significant challenges. Austin and Grenoble (2007: 22–3) suggest that interdisciplinary cooperation can be difficult to achieve. For example, with a language that is difficult for non-speakers to hear and transcribe, must non-linguist team members become fully proficient? Or, as in our Chatino project, where linguists are spread thinly over many significantly different varieties, must the team in fact be a family of 'lone linguists'? Moreover, how does participation in documentation projects fit into people's lives and future career goals? Are community members volunteers? Employees?

Entrepreneurs? Are academics professors, or postgraduate students? Are there coherent rewards for biologists, psychologists, musicologists, videographers, audio-recording specialists, and so on, who might enhance a project? To take one example, are there valued genres of filmmaking that fit with the documentarian's (perhaps futile) goal of creating records in which the depiction of context is stabilized by keeping the camera or cameras in one place for long periods of time and avoiding highly interpretive pans, sweeps and zooms? [6]

Moreover, what does documentary training look like for children and adults without a full secondary educational background? And at university and postgraduate levels, how, in two or three years, is it possible to train a good lexico-grammarian who can also find, handle, and thrive in a field situation, in their own community or somebody else's, and then record and archive properly? It may be too much to ask (but see Jukes, Chapter 21 for some discussion).

Somehow, all of these questions must be approached and addressed if endangered-language documentation is to reach its full potential.

9.4.3 Lexicogrammatical code, language use, nostalgia and contemporary realism

We now turn to a set of conceptual issues. Relatively unheeded in Himmelmann's (1998) programme for linguistic documentation is his idea that 'a language documentation ... aims at the record of the linguistic practices and traditions of a speech community' (Himmelmann 1998: 166), wherein, as he points out in a footnote, a speech community may share more than one language.

In fact though, endangered language projects are almost always focused on one specific (endangered) language. As such they are instances of what I have called DOCUMENTATIONS OF THE ANCESTRAL CODE (Woodbury 2005: 257). We may make several observations about such projects in order better to see their relationship to what Himmelmann proposed, as well as several other possibilities. In turn, this allows us to see certain controversies as misplaced, given the broader conception of documentary linguistics I am exploring here.

Documentation of the ancestral code requires that the language be 'put through its paces', studied in various contexts of use, and supplemented with elicitation in order to fill lexical fields, paradigms and the like. Dictionaries and grammars can follow efficiently from such documentation. And at a community level, it supports orthography creation, the preparation of pedagogical dictionaries, grammars and readers, and efforts to get the language taught in schools or recognized politically: in general, it performs well in contexts where communities wish to assert that their language, despite disparagement, is a language in the same sense that Spanish or English or Russian are languages.

Despite falling short of Himmelmann's call for documentation irrespective of code, I think it is important to defend the status of such work as documentary, as long as the documentation is curated (Conathan, Chapter 12). Likewise, to push the limit still further, a project that archives its data properly may count as documentary even if it is mainly or solely focused on making dictionaries, and uses elicitation rather than text collection as its main method, such as the *Project for the Documentation of the Languages of Meso-America* (Kaufman *et al.* 2001).

Debates also arise as to whether grammar-writing and dictionary-making should be avoided in favour of more text collection or elicitation. In a review of Gippert *et al.* (2006), Evans (2008:348) defends the role of these activities:

> Something about the definitive appearance of these products brings out a higher level of scrutiny and a leap to new levels of accuracy in transcription and translation. Both times that I have been involved in producing dictionaries of Australian Aboriginal languages, there was a sudden upsurge in interest and in the supplying of new or extended lexical entries at the point where speakers of the language held in their hands a properly-produced book in their language.

He concludes:

> For these reasons I think it is a mistake for documentarist linguists to argue that they should consecrate all their time and effort to pure documentary activities at the expense of preparing descriptive grammars or other reference materials. A much more apt strategy is Colette Grinevald's (2001) vision of an eternal spiralling upwards through the elements of the classic Boasian trilogy – grammar, texts (now = documentary corpus), and dictionary – with each step forward producing advances and refinements in how the other steps proceed. (page 348)

Indeed, to dichotomize 'description' versus 'documentation' to such an extent as to exclude or restrict code-focused documentation amounts to a kind of back-door structuralist recidivism, reenergizing the very distinction that detheorized primary linguistic records in the first place.

Finally, documentation of the ancestral code, like the endangerment construct itself, can be termed, without any intention to disparage (see Williams 1973), as NOSTALGIC, in the sense that it selects as important from among all the speech in a community that speech which gives evidence of a feature of the past which may not persist long into the future, namely the ancestral code (see also Dobrin and Berson, Section 10.2). On the academic side, we may see linguistic reconstruction, or a focus on the most traditional variant forms, as nostalgic tendencies, while purism and assertions of the linguistic code as intrinsic to ethnic or spiritual identity or to traditionalism are forms of nostalgia in a popular sense.

Himmelmann's speech-community perspective, mentioned earlier, is in keeping with the ethnography of speaking and its successors, which offer useful precedents. However, we might identify (at least) two possible models of this kind, one nostalgic, the other not (the discussion here draws on a typology presented in Woodbury 2005). One, a DOCUMENTATION OF ANCESTRAL COMMUNICATIVE PRACTICES, would focus on ENDANGERED WAYS OF SPEAKING, in any code, and in that sense would also count as nostalgic. This may include formal genres or speech situations falling out of use, but it may also involve informal conversation among people perceived as traditional, or speech in connection with traditional activities, and it may involve traditional forms of multilingualism, such as remnant uses of Russian in Alaska. Although basic documentation would involve translation and presumably transcription and some annotation, grammar-writing and dictionary-making would then not be an intrinsic part of the project, in keeping with Himmelmann; although it certainly could be.

Such an approach is consonant with any ideology that locates language mainly in its verbal products, and may more directly address a community's feelings of language loss than would a focus on lexico-grammatical code. As a scholarly approach, it fits well with Boas's broader 'ethnological' framework, but at the same times is susceptible to critiques of a more contemporary kind, voiced by Garrett (2004):

> A paradox lies near the heart of documentary linguistics. As an enterprise it relates to language ecology; it is founded on the same commitment to the sociocultural embeddedness of language. But a discipline of 'documentary sociocultural analysis' or 'documentary anthropology' could hardly exist without buying into the myth that cultures are static and there is some endangered moment that merits documentation: the discredited Boasian ethnographic present. What justifies the documentary enterprise in the case of language? The only clear rationale is the distinctness of language, its systematic character and structural integrity: Documentary linguistics presupposes the same assumption of linguistic autonomy that it purports to eschew.

While agreeing with this line of commentary, I think nostalgia plays a key role, for while it fits well with the essentializing tendency of structural analysis, in sociocultural realms nostalgic selectivity leads directly to the problems mentioned. Moreover, the issue can be placed in a different light by considering a second construction of Himmelmann's position, alluded to above, namely a rational attempt to produce a DOCUMENTATION OF CONTEMPORARY COMMUNICATIVE ECOLOGY. Such an approach would aim in some sense at the 'real' or immanent as opposed to the nostalgic, even if plagued with the problem of having to select just what, from among everything, to document. It might serve in a scholarly sense as a form of sociolinguistic survey of a community, irrespective

of endangerment issues, or as a way of putting them in context, and it might fit with community ideologies where contemporary modes of speaking are a source of interest (for example, the popular interest in the United States in Spanish–English code switching). And although it would restrict itself to 'now', nothing in principle confines such a project to operating only in a single moment in time.

Finally, to round out the typology, there is the case where lexico-grammatical code is again the focus, but with a contemporary rather than nostalgic orientation: DOCUMENTATION OF AN EMERGENT CODE, that is, a focus on lexico-grammatical systems with an emphasis on their contemporary state, including the emergence of new forms, neologisms, coinage, syntactic innovation, contact convergence, borrowing and even indigenized versions of the language of wider communication (e.g., so-called 'Aboriginal English' or 'Indian English', see Woodbury 1993, 1998 for extended discussion; see also O'Shannessy, Chapter 5, on 'light Warlpiri'). While seemingly most appropriate to studies of creolization or sign language formation (Meir *et al.* in press) rather than language endangerment, it can be relevant and even essential to the study of so-called semi-speakers, and of the variation in communities undergoing rapid language shift. And it can support community efforts to grapple with linguistic purism and to accomplish such aspects of language planning as coinage and translation of foreign texts into the endangered vernacular. On the other hand, it may go against the grain of anyone, academic or not, with a strong sense of nostalgia. It also raises profound theoretical issues (Le Page and Tabouret-Keller 1985) about the focus, or degree of conventionalization, that a linguistic code may have; and the line between emergent code documentation as opposed to documentation of a contemporary communicative ecology may be blurry indeed.

In summary, a broad, inclusive documentary linguistics can stand a few paces back from the ideological fray in order to coordinate, in a rational way, different kinds of endeavours that contribute to endangered-language documentation.

9.4.4 Corpus theorization: adequacy, comprehensiveness, complementarity, quality and quantity in documentary corpora

There is considerable focus in the literature (Himmelmann 1998, 2006a, Lehmann 2001, Rhodes *et al.* 2006, A. Woodbury 2003) and in the design of specific projects on assembling corpora that are adequate and comprehensive overviews, much as grammars, dictionaries or ethnographies of speaking may be said to be. Assuming we make allowance for differences on such basic parameters as code versus community focus, nostalgic versus contemporary, and probably others too, there is nothing wrong with this in principle, nor is it wrong for there to be some amount of contention or variety in how corpora are to be theorized. Perhaps

most comprehensively, Hymes (1974a) proposes an 'etic grid' of essentially orthogonal parameters, implying a scheme for sampling all values with all values. Or sampling can be boiled down to different kinds of speech (Himmelmann 1998), or certain kinds of speech can be privileged on principle such as verbal art (Sherzer 1987), or conversation (Levinson 1983: 284–5); or collection can be monitored in part by the dictates of dictionary and grammar making (Rhodes *et al.* 2006).

If there is one and only one chance ever to document a language, it makes especially good sense to strive for a result that is as complete as may be possible. All things being equal, a corpus should be DIVERSE, and any of the proposed theorizations would offer that. But for many (perhaps most) languages, some documentation already exists, and more may be done in the future. It is therefore worthwhile for project designs to take complementarity into account and recognize that corpus building should be ONGOING, DISTRIBUTED and OPPORTUNISTIC. For example, the documentation project of Taff (2004) for Aleut takes note of the fact that Bergsland's corpus, albeit exemplary, is long on narrative and lexicon but very short on conversation. Accordingly her project focuses almost exclusively on conversation, nearly the only Aleut genre now available, and thereby adds enormously to the overall documentation of Aleut. Likewise, within Bergsland's own corpus is his philological monograph analysing and interpreting personal name data gathered by the Billings Expedition in 1790–2 (Bergsland 1998). And in my own documentation of Central Alaskan Yupik, I spent an intensive three-year period assembling a large corpus of experimental productions designed to elucidate the intonational system, yielding, I think, a reasonable adjunct to the narratives, conversations, and music I had documented earlier. There is no reason not to engage in specific, narrow-cast documentation projects of this kind, especially as complements to a larger corpus.

Moreover, corpus theorization, and indeed the very design and conduct of documentation projects, is also driven by social, aesthetic and humanistic values in the speech community itself, as well as those that develop within the emergent communities of practice conducting the documentation (see Hill 2006). My own documentation of Central Alaskan Yupik began with a focus on recording elders' retellings of traditional myth and folk tales, a strong community interest at the time, and part of a tacitly agreed-upon function for documentation. It was only once a traditional men's house had been built that I was invited to record the conversation that took place there, licensed, so it seemed, by the setting the men's house offered.

By contrast, community interests and values have led to quite a different corpus theorization in the documentation project I am currently engaged in (Woodbury 2010). This project is focused on Chatino, a shallow family of languages spoken in Oaxaca, Mexico, and besides me it involves a group of six University of Texas graduate students, two of whom are

native speakers. Writing Chatino emerges as a core goal throughout the region and a rejoinder to colonial claims that Chatino is just a dialect incapable of being written. Accordingly, our documentation has aimed at least to sample each substantially distinct variety through text recording and lexical elicitation in order to adapt an orthography to its segments, phonotactics, tones and tone sandhi well enough that we can teach it in communities.

At the same time, Chatino speakers hold in extremely high esteem the rapid-fire, classically Mesoamerican parallelistic oratory of traditional community political authorities, and our awareness of those linguistic practices and their 'leakage' into ritual, prayer and everyday speech has led Hilaria Cruz to seek out such instances and compile them into a coherently focused body of documentation (Cruz 2009).

Likewise, Emiliana Cruz, working from a sense that Chatino linguistic knowledge is significantly organized and ordered by its connection to territory, created an extensive corpus of audio- and video-recorded Chatino-language narratives, interviews, and atlas and dictionary research gained during long mountain forays with traditional land users to document flora, fauna, land forms, population distributions, land claims, trade routes, and ethnohistory in one sprawling municipality. She hopes to present the text and lexical materials linked to maps, along with exegeses, translations and interpretations as a LINGUISTIC ECOLOGY that draws on ethnographic discourses developed in Basso (1996), complements them with an emerging documentary genre, and addresses her community's understandings of language as it links to territory.

At the same time, alongside these fairly structured goals, we often find ourselves looking, in more or less abstract or general terms, for ways to diversify our corpus.

In the end, our corpus theorization involves a combination of considerations of various kinds represented here. Moreover, the fact that corpora are strongly shaped by the goals of their creators does not prevent them from becoming general, multiuse resources.

Finally, we consider the question of QUANTITY. All things being equal, more is better than less, although we hardly have good generic measures of what counts as adequate quantity. Liberman (2006) shows that documentary text and speech corpora are typically at least an order of magnitude smaller than corpora for major world languages or even than text corpora for major dead languages like Latin. He proposes technological approaches for rapid corpus acquisition in large quantities, some of which involve such shortcuts as the use of small, cheap (but widely distributed) hand-held digital recorders; oral interpretation on-the-fly into a language of wider communication; and recording slow, careful, (hopefully) philologically accessible 'respeakings' of recorded texts in place of written transcriptions made by, or in the presence of, native speakers (e.g. a project called *Basic Oral Language Documentation* that is

being carried out by Stephen Bird in Papua New Guinea on various local languages has adopted this methodology)[7]. There is a tendency to dismiss such 'substandard' record-making and annotation; Liberman's point is that such records, in quantity, would appreciably augment the total documentary corpus and supplement the records made to standard specifications. His point, from a philological perspective, hardly seems controversial, and it contributes to the broadening of our notion of language documentation.

9.4.5 Annotation

Currently the typical annotated record is an audio- or video-recorded text that is described by metadata and annotated with transcription that is time-aligned to the transcription by chunks delimited by pause-, clause- or sentence-breaks. On the model of Bickel *et al.* (2004), there is a free-translation into a language of wider communication, as well as a morphological segmentation and a morphological gloss or parse line somehow aligned to it; and these analytic elements may be connected, via a look-up system such as Toolbox, to an independent lexicon. Schultze-Berndt (2006) gives an excellent discussion of these and other attributes of annotation (see also Good, Chapter 11).

On the one hand, representation and annotation are rich, complex topics and presuppose general theorizations of almost every area of grammar. A fully broad and inclusive framework for language documentation could undertake to build into its annotation practices more of what is commonly studied in typological and theoretical linguistics (e.g. Lieb and Drude 2000). For example, most syntacticians agree that phrasal constituency or dependency is a major feature of syntax and basis for semantic interpretation; and the stock examples of constituent ambiguity (like Groucho Marx's line, *I shot an elephant in my pajamas; what it was doing in my pajamas I'll never know!*) show that constituency is sometimes covert and verifiable only by test when context or translation fails to disambiguate it. Yet in their practice, documentary linguists rarely include constituency mark-up in their annotations (as is common, e.g. in corpus-based and computational linguistics) or perform the tests needed to complete it. Nor is the lack of this information perceived as a loss in the same way as a failure in transcription to note tone or test for a quiescent vowel would. There is a perception around that theoretical linguists are engaged in some debates that are irrelevant to the serious business of documentation. Theoreticians themselves may even agree, but they may also agree that they could play a key role in the framing of more general approaches to annotation, as well as the training of documenters interested in annotation.

On the other hand, it is worth assessing soberly to what extent annotation is required to meet minimal standards of record transparency.

Liberman (2006) raises this issue in connection with documentation in quantity. Austin and Grenoble (2007) raise questions about the choice of the language of wider communication. Within a broad view of endangered language documentation might fit projects whose leading actors are not professional linguists and whose primary aims are not code-focused, or even text-focused in the usual sense. A minimal documentation (as noted in section 9.1) might consist of just a recorded text and a free translation into a language of wider communication. It might be very hard to use for linguistic purposes, or for literacy-based language learning, but with respect to some aims, as discussed in Section 9.4.3 above, it may not be entirely useless. Particularly in an enlarged framework that is academically multidisciplinary as well as popular, one may ask whether 'useful for linguistic analysis' is a non-negotiable component of a multipurpose corpus, however non-negotiable it may be within linguistics and among linguists.

9.4.6 In what forms should documentation be stored and disseminated?

A related question is that of the forms in which documentation is to be stored and disseminated. An important distinction may be drawn between an archival form intended for preservation, and a presentational form, intended for engagement with a range of possible publics (Bird and Simon 2003; Good, Chapter 11). Books, monographs, and journal articles are the long-time default, some examples of which have already been cited. They can range from general collections to more specific projects. Among digital resources, the DoBeS presentation of its material via its IMDI tree structure offers a browsable collection of holistic documentations (viewable through the IMDI browser software). Other digital archives, such as AILLA, offer corpora that are not uniformly thematic, but are browsable and searchable (see also Nathan, Chapter 13). Meanwhile many projects make their materials accessible in ways that are locally appropriate, via the web or in print. Accessibility is one very important area for innovation and may best be accomplished with an eye toward specific themes and tailored to specific audiences.

9.5 Conclusion: toward a broad and inclusive view of endangered-language documentation

In this chapter we have explored several different dimensions along which our conceptions and practices of endangered-language documentation could acknowledge and extend its own potential for breadth and inclusivity. This must start with a recognition by linguists, whether native speakers or not of the language of study, of the interests of the

people with whom they work and of the ways in which they might share projects and agendas (see Dobrin and Berson, Chapter 10; Bowern, Chapter 23).

Community context also bears on deeper intellectual and ideological questions underlying the design of projects focused variously on the speech of a specific community and the use of a specific lexico-grammatical code, and animated by varying humanistic stances. I try to emphasize that these intellectual and ideological questions face anyone considering language documentation, whether scholar, community member, both, or neither; and that different framings of these issues lead in many different, and I believe productive, directions.

Inasmuch as endangered language documentation is propelled by a sense of urgency, one may wish to construct the ultimate linguistic 'Noah's Ark', a final corpus theorization according to which any language can be encapsulated. But while it is possible to list desirable attributes for a corpus, all other things being equal, such a theorization is unlikely to fairy span the diversity of agendas and the intellectual and ideological stances we know to exist, and, correspondingly, unlikely uniformly to engage all those who might contribute.

Likewise, overly categorical stipulations of so-called 'best practice' in corpus creation may overlook tradeoffs between quality and quantity in documentation; and when applied to annotation and interpretation, may channel efforts unequally toward some uses of corpora over others, or even understate how ultimately difficult it is ever to fully translate or interpret a record of human behaviour in context.

Notes

1 The term DOCUMENTARY LINGUISTICS could reasonably be applied to any linguistic work that creates or uses documentation, including most work on major world languages in lexicography, corpus linguistics, language acquisition studies or sociolinguistics; nevertheless the term is ordinarily used in the context of research on endangered languages or other languages where linguistic field methods are involved. I will follow the ordinary usage.

2 Special thanks to Victor Golla for an illuminating exchange on shifting perspectives toward documentation in early and mid twentieth-century Americanist linguistics.

3 In striking contrast, Malinowski (1935) decades earlier offered a text-based approach to ethnography that was linguistics-free on principle, taking the view that meaning only exists in context (thus making dictionaries and grammars futile). Following that view (or perhaps in spite of it!) he presents a copious, explicitly theorized, varied documentary corpus of narratives, descriptions, technological discussions,

prayers, and magical incantations as part of an integrated treatment of agricultural practice and ritual in the Trobriand Islands that can still stand as a model for the ethnographic use of a documentary corpus.

4 See www.mpi.nl/DOBES (24 January 2009).

5 nflrc.hawaii.edu/ldc.

6 An interesting example of mutually rewarding collaboration is that between sound artist John Wynne, linguist Tyler Peterson and artist/photographer Denise Hawrysio to make documentary recordings of speakers of Gitxsanimaax, an endangered indigenous language in northern British Columbia, Canada (in a project funded by ELDP). The materials from this research are incorporated into Wynne and Hawrysio's installation *Anspayaxw* (www.sensitivebrigade.com/Anspayaxw.htm) which is part of the *Border Zones: New Art Across Cultures* exhibition showing at the Museum of Anthropology, Vancouver, from 23 January to 12 September 2010. When the exhibition travels to the 'Ksan Gallery in Gitxsan territory in 2011, Wynne and Peterson will also deposit at 'Ksan an archive of their materials for community use.

7 See www.boldpng.info (17 November 2010).