Defining Documentary Linguistics

Tony Woodbury

1. Preamble[1]

In the last fifteen years, we have seen the emergence of a branch of linguistics which has come to be called Documentary Linguistics. It is concerned with the making and keeping of records of the world's languages and their patterns of use. This emergence has taken place alongside major changes in the technology of linguistic data representation and maintenance; alongside new attention to linguistic diversity; alongside an increasing focus on the threats to that diversity by the endangerment of languages and language practices around the world, especially in small indigenous communities; and perhaps most importantly of all, alongside the discipline's growing awareness that linguistic documentation has crucial stakeholders well beyond the academic community; in endangered language communities themselves, but also beyond.

The purpose of this paper is to discuss documentary linguistics, how it has been emerging, and where it may be headed.

2. Documentation is old

Of course there has **long** been concern for the perspicuous documentation and description of the world's languages. We see this in the now century-old tradition of monograph series and journals of record in which texts, dictionaries, grammars, vocabularies, and other works have been published.

We can see too that such work has been foundational for the discipline's more theoretical endeavors since at least the time of Franz Boas. Dictionaries, grammars, and texts have informed historical linguistics and the reconstruction of linguistic prehistory, of genetic language families, and of patterns of prehistoric linguistic contact. They have informed inquiry into the methods and tools for linguistic description and discovery. And they have informed the development and testing of theories of linguistic typology and of universal grammar.

---

[1] This is a lightly edited version of a plenary address given at the Annual Meeting of the Linguistic Society of America, Atlanta, Georgia, on January 3, 2003, and also delivered at the Workshop on Endangered Languages, SOAS, London. I wish to thank Chris Beier, Nora England, B'alam Mateo-Toledo, Lev Michael, and others who provided helpful comments and discussion on a version given in Austin, Texas, on December 5, 2002; and Wally Chafe, Bill Poser, Doug Whalen, and others in Atlanta, for comments leading to further revisions.

Documentation and description have been foundational too in having kept linguists in the field, observing language in its social context, and through that it has led directly to work on the use and function of language in specific speech communities.

Finally, practitioners of documentary and descriptive linguistics have always operated in an atmosphere of urgency and impending language loss, making lasting records and in some cases taking part in community efforts at language preservation, teaching, planning, and revival.

3. But a new conception has been emerging

Nevertheless, these antecedent areas of concern have become aligned and focused in a fundamentally new way in a very short time—perhaps as short as a decade—into a field that has come to be known as documentary linguistics.

4. Elements of the shift

Perhaps it's best to start by looking at what has been happening **around** the emergence of a documentary linguistics. What new things have become possible? What ideas have been "in the air"? What is the value of linguistic documentation? To whom? And what do they want from it? In short, what changes in the general scene surrounding linguistic documentation in the last decade and a half have set the stage for its reconceptualization?

4.1 Technology

Let's start with technology because it, more than anything, has changed our thinking about the physical **possibilities** for linguistic documentation. Suddenly, with powerful laptops, digital audio, video, and the worldwide web, it at least **seems** that we should be able to capture and store enormous amounts of information; we should be able to search through this information with unprecedented speed and precision; we should be able to link transcriptions with audio- and videotapes, and entries and dictionaries or statements in grammars with large databases of illustrative examples; we should be able to disseminate around the globe the material now collecting dust in attics or rotting in basements; and we should be able to keep huge amounts of information safe in perpetuity. While reality has turned out to be more complex—it's clear we need to agree on and coordinate our practices before this can happen—this revolution in both the magnitude and the quality of linguistic documentation has brought about permanent changes in what people plan and hope for.

4.2 Diversity

A second change in the general scene surrounding documentation is an increasing emphasis on **diversity** as a central, organizing question in linguistics. To be sure, the study of universal grammar has **also** shed light on the ways languages can differ, but as

something of a side issue. More recently, work on universal grammar has taken increasing responsibility for charting and explaining typological patterns; Bruce Hayes' (1995) book *Metrical stress theory* would be just one nice example; the work of Paul Hopper and others on functional relationships among grammatical categories would be another (Hopper and Thompson 1980). More radically still, work such as Johanna Nichols' (1992) book *Linguistic Diversity in Space and Time* has placed diversity on center stage by asking how typological and genetic diversity can be measured, how it can be that world regions differ markedly in the amount of diversity they show, how areal influence, genetic relatedness, and universal grammar all affect patterns of linguistic difference, and how different geographic, social, and population patterns affect linguistic diversity. Naturally, all such theorizing calls for documentation of the world's languages.

4.2.1 Social diversity

Related to this is a focus on diversity in a slightly different sense — the focus by sociolinguists on social diversity, and on the ways ideology about language and linguistic practice constitute and embody peoples' sense of their social, ethnic, personal, and even spiritual identity. It is perhaps this aspect of 'linguistic diversity' that is most directly relevant to contemporary social and political concerns about diversity within US society, and diversity as a value affected by globalization and other homogenizing tendencies.

4.2.2 Neo-Whorfian concerns

In a related way, it is increasingly asserted — among linguistic anthropologists (Lucy 1992, Gumperz and Levinson 1996) and in society more widely — that linguistic diversity has humanistic value, and that it is critical to intellectual, literary, and aesthetic creativity. These questions might be called neo-Whorfian although their roots go much farther back. To the extent this is the case, the study of linguistic diversity — diversity of linguistic codes as well as of the uses and potentialities of those codes — becomes important.

4.3 Endangerment

Of course, all of these scientific, social, and intellectual concerns are set in a wider contexts where linguistic diversity itself is under threat. Language practices in communities all around the world are shifting so quickly away from traditional heritage languages and toward the 50 or so most important regional, national, and world languages, that we see the number of living languages shrinking from the 6500 or so counted today (SIL International 2002) to as few as half or less in the time of only a century. To linguistics even as it is traditionally conceived, the loss is devastating, whether the scientific focus is on universality or diversity. It is the loss of tens of millennia of natural development over the entire earth under conditions where intercommunication was local at best; exactly the conditions which put to the test our **potential** for diversification.

Endangerment is nothing new; but, after a period of relative inattention within the discipline as a whole, our late colleague Ken Hale and others forcefully and insightfully called attention to the situation at an LSA symposium in 1991, published in *Language* in 1992 (Hale et al. 1992). The effect on the discipline of this and other similar calls around the same time was galvanizing. The LSA has established a Committee on Endangered Languages and their Preservation and has sessions dedicated to "Field Reports." Field methods classes reemerged in graduate departments. And there was an unprecedented level of public outreach as well as public response, including press attention, the establishment of the Endangered Language Fund here in the US, the Foundation for Endangered Languages in the UK, the Volkswagen Foundation's unprecedentedly well-funded *Dokumentation Bedrohter Sprachen (DoBeS—Documentation of Endangered Languages)* project,[2] and, most recently the Lisbet Rausing Charitable Fund's *Endangered Languages Documentation Programme*, administered by the School of Oriental and African Studies at London University.[3]

As Ken Hale and his co-authors made very clear, however, the galvanizing forces behind the renewed activity were several decades of activism in small, endangered-language communities. Many speakers of endangered languages who have spoken and written on the subject, and others belonging to communities where the heritage language has already been lost, have described the loss as a loss of identity, and as a cultural, literary, intellectual, or spiritual severance from ancestors, community, and territory; and as an example or symbol of the domination of the more powerful over the less powerful. I would mention among these activists those in the indigenous language immersion movement, inaugurated in New Zealand in the late 1970s in the *Te Kohanga Reo* ('Language Nests') where Maori-speaking elders spoke only Maori to preschoolers entering with English only; and those struggling to spread the immersion model to Native communities in the US; I would mention the Maya movement in Guatemala, which has worked to fashion standardized literary languages for indigenous language teaching and maintenance across the country; I would mention AILDI, the American Indian Language Development Institute, that a number of our members are involved in, where community language activists inside and outside the schools provide and receive training in language issues (McCarty et al. 1997); the efforts of communities and tribal governments to fund language work; the efforts of individuals to learn their ancestral languages even as adults through master-apprentice relationships; and the emergence in endangered language communities everywhere of grass-roots documentation by tape recorder, video camera, and now laptop computer.

I would like to argue that community language activist agendas have had a profound effect on the new documentary linguistics; and I would like to point out ways they can become even more a part of it. Already, it is becoming less and less viable for

[2] http://www.mpi.nl/DOBES/

[3] http://www.eldp.soas.ac.uk/

linguists to think of the stakeholders in language documentation to be constituted only of a vaguely-conceived scientific posterity. From the point of view of the discipline, I see this as a part of a change of direction away from the Saussurian solipsism that has been our past tendency, toward a much broader involvement in language as it appears on the world stage.

5. So what is this change in conception, anyway?

So what is the change in conception that characterizes the new documentary linguistics? I think the essential principle is that it is, in my colleague Joel Sherzer's term, 'discourse-centered' (Sherzer, 1990) That is, direct representation of naturally occurring discourse is the primary project, while description and analysis are contingent, emergent byproducts which grow alongside primary documentation but are always changeable and parasitic on it.

This orientation contrasts with a traditional view, where linguists have equated documentation with the traditional products of linguistic **description**, namely a grammar, a dictionary, and a set of texts. The relationship to each other is hierarchic. At the top is the grammar, documenting the broadest generalizations; next is the lexicon, serving as an appendix to the grammar : e.g., Bloomfield (1933) called it a list of basic irregularities; and last, 'enough texts to permit a verification of the analysis' (Samarin 1967:46), including dictated narratives and perhaps some proverbs, riddles, or songs. The term 'documentation' remains accurate as long as the proper **object** of documentation is considered to be the internalized or shared lexico-grammatical system; as such, it serves as the input to higher order work on the reconstruction of the linguistic past, or on the range and limits of human linguistic competence. But texts, speakers' commentaries on word meanings, speakers' grammaticality judgments, translations, and proffered examples remain secondary and epiphenomenal in terms of the whole project, even though they may serve as the data.

One weakness of the grammar-dictionary-texts model has been the difficulty of knowing the full of extent of the system to document while you are doing your field work and writing your dictionary and grammar. This problem has been addressed in several ways. One has been the construction of checklists or protocols for field investigation, e.g., Comrie and Smith 1977. It has also been addressed through the construction of linguistic theories, including theories of universal grammar and of typology, since these can establish quite explicit expectations for what to expect in a previously uninvestigated language, often on the basis of what has already been encountered in better-studied languages. Likewise, residual problems arising through dialect difference, social variation, genre, and first language acquisition prompt wider field investigation with reference to social structure, social activity, and social meaning. All these refinements add new types of data to the bottom of the pyramid—from systematic experimental productions, informal elicitation and elicitative experimentation; to more natural speech texts. Even so, much of this work has not been conceived of as the making of linguistic records, nor, I think, has its importance as documentation been fully appreciated by grammar-dictionary-texts

paradigm descriptivists, let alone the second tier grammatical analysts. Moreover, it is the (low level) generalizations in each of these areas—and their relationship to grammar—that are the object of documentation. The data themselves still remain largely epiphenomenal.

Data itself isn't independently theorized, and is ultimately neglected on a number of theoretical and practical levels. In effect, what we called "data" had not itself been independently theorized *as* documentation, apart from its (low level) analytic applications; the record of the speech of an individual or of a language or a dialect or a community had not been thought of as a coherent body. If it had been more fully **theorized** in the sense I intend here, then there would have been a body of agreement (as well as some debate) over just what the whole "data" of a language consists of, and of how one might approach the enormous task of sampling it, given current resources and given a farsighted and integrated view of the uses to which it might be put. Because this was **not** a usual arena for debate, we should conclude that data, as a notion, was under-theorized, or only theorized with respect to relatively specific, parochial uses. For example, I remember as a graduate student in the late 70s talking with graduate student colleagues from other departments about collecting natural speech data on tape pretty much for its own sake, just to have as documentation, and being told I was being 'scientifically naïve,' that there was no such thing as data independent of a theory which uses data.

This is not to say that for any given language, there haven't always been individuals with a practical command of a whole corpus—aware of texts, field notes, tapes, and early written records—who worked effectively with them all. Nor is it to say that practitioners of specific fields failed to theorize corpora relevant to their special interests, e.g., segmental phonetic contrasts, phonological variation, syntactic typology, language ontogeny, or ethnography of speaking in a given community. It is simply to say that there was neglect of the nature of the final record (and record producing efforts), the comparison of such records across communities and languages, and the evaluation of them and their production.

Consider for example the institutional response to "data" *per se*. Higher level analysis, rather than documentation itself, has been the coin of the realm: in nearly all graduate departments of linguistics it has been grammars—or better still, grammars of grammars—that have been suitable as doctoral dissertations, whereas dictionaries, or, heaven forbid, text collections—the low end of the hierarchy—have not. The usual justification for this—that text collection and even dictionary making are only clerical activities—was patently just an artifact of the undertheorization of documentation. For if properly theorized, new instances of documentation—just like new grammars or syntheses of grammars—would have informed additions or revisions of established doctrine.

6. And just how IS data/documentation theorized?

So how is documentation being theorized? In the somewhat reductive terms of the received paradigm, documentation is increasingly coming to be seen as text curation. It is text

curation in approximately the Boasian sense, though much enlarged. Boas emphasized to his students the importance of writing down dictated mythological and ethnographic texts in original languages as a basis for all inquiry in both linguistic and cultural anthropology. This basic idea led to further theorization of speech events, situated in socially coherent speech communities, by Dell Hymes, John Gumperz, and others (Gumperz and Hymes 1964). Speech events were classified in terms of the language or languages (or dialects or varieties) they involved, in terms of genres, participants, participant roles, communicative purpose, and other features; and the theorization itself came to be known as the ethnography of speaking. In the 1980s my Texas colleagues Joel Sherzer, Greg Urban, and others broadened this idea under the heading of 'a discourse-centered approach' to language and culture (Sherzer 1990, Urban 1991), placing increasingly more theoretical emphasis not on any one particular final use for discourse, but on an openness to the range of **possible** uses; as well as an emphasis on how natural discourse data was to be represented, transcribed, preserved, disseminated, and made accessible through interpretive apparatuses, including catalogs, translations, notes commentaries, exegeses, and summaries. This is not to deny that there are specific goals, programs, or agendas which documentation may meet, or which may guide documentation; quite the contrary. It only means that that there can be a wider, coordinated conception of the endeavor irrespective of specific goals.

These issues have all emerged in connection with digital archiving. The LINGUIST List's NSF-funded E-MELD project [4] —*Electronic Metastructure for Endangered Languages Data*—and OLAC (*Open Language Archives Community*)[5] have been conducting discussions on establishing useful, recurrent, searchable metadata categories, that is, the categories to be used in the electronic equivalent of card-catalog information for any given item of text or other linguistic data in a digital archive; and in doing so they have inherited many issues from the ethnography of speaking. Likewise, there has been significant work via the Linguistic Data Consortium[6] by Steven Bird, Mark Liberman, and others, on what they call annotation graphs, the formal properties of speech data transcription and data mark-up (Bird and Liberman 1999). Likewise too, E-MELD is pursuing work specifically on data annotation and mark-up. All of this is done, to be sure, with an eye toward ensuring that data will be universally useable and accessible for a long time to come; but it rests crucially on, and in fact partly constitutes, this emergent theorization of documentation.

6.1 Recasting traditional grammar/text/dictionary research

What are the implications of this conception of documentation for grammar and dictionary research, and for linguistic discovery more generally?

[4] http://saussure.linguistlist.org/cfdocs/emeld/

[5] http://www.language-archives.org/

[6] http://www.ldc.upenn.edu/

From the point of view that is emerging, grammars and dictionaries cease to be the end-product of documentation; rather, they are part of the apparatus—the descriptive and explanatory material—that **annotates** the documentary corpus. It is already the widely held view that no one grammar or dictionary or set of analyses is necessarily final and immune from revision; yet in the pyramidal model this is implied when we say that a language has been documented because it has a grammar and dictionary.

As the term 'apparatus' implies, the grammar and dictionary are contingent on the corpus and evolve with it. Corpus study thus become the primary modality for grammatical analysis, which, as Shobhana Chelliah (2001) has argued in her paper in Paul Newman and Martha Ratliff's recent collection on field work, leads to a better sense of "what is out there" but also requires a range of different methods.

Does this mean then that grammatical elicitation is done away with? Not at all. As Chelliah argues, corpus observation is best done in conjunction with meta-corporeal and metalinguistic discussion; for example, if you are making a thesarus, you don't want to just find the names of different grasses in your corpus, you also want the resultant list to be discussed and gone over by speakers who are authorities on grasses to make sure you have the field properly covered and to generate good definitions. Rather, what a documentation-oriented view says is that the discussions of grass names should themselves be videotaped or tape recorded and should themselves become a part of the whole corpus; as should any and all grammatical elicitation of the traditional kind. Moreover, years from now, it will be the grass name attestations and grasses discussion tapes, and not the dictionary, that you will consider as the final document on grass names. Likewise, we might expect that if you refer to a generalization about a given language, you will not only cite the source of the assertion, with perhaps an illustration or two; but link to a whole pattern of data which **entails** that generalization.

This means, then, that there is a dialectical relationship between corpus and apparatus—the corpus informs the analytic apparatus; but analysis—including everything you bring to the table when doing grammatical and lexical elicitation—in turn also informs the corpus. Likewise, almost any presentation of documentary work requires grammatical analysis—transcription requires a phonological analysis, and lexical presentation in the form of a thesaurus or dictionary requires morphological and lexical analysis.

7. Huge challenges

Summarizing so far, there has been a major change in conception about documentation and linguistic discovery. Furthermore, new technological capacities, and the sheer amount of attention that documentation and language preservation are receiving—including increased funding—are having a profound effect on the magnitude of the enterprise. I imagine it's something like what happened in the 1960s when major attention suddenly turned to ecology and the environmental sciences.

All this presents huge challenges for us. What are our values about our work? What are our working paradigms? What constitutes a valuable and feasible project, or a type of project? What kind of training should we be offering? And, most importantly, how can we come to terms with the wide range of **agendas** surrounding language documentation, above all, in endangered-language communities, so that we as linguists can build coalitions and be of greatest service to others as well as to ourselves?

7.1 Documentation agendas

Let me start by being a bit negative. There's a strong temptation to ask (as in fact was done in an October, 1995 conference at Max Planck Institute for Psycholinguistics, Nijmegen), What constitutes the **best record** of a language? If we had a finite amount of time to document language X and seal it into a time capsule, what would we put there?

My feeling about that question is that it forces thinking about generic documentation situations and that that is premature. I would rather begin by considering how **different** documentary situations can be, and how different documentary agendas can be. I want to acknowledge having been profoundly influenced on this by a paper by David Wilkins (1992) called 'Linguistic research under aboriginal control: a personal account of field work in Central Australia.' Let me make my point by discussing two very different situations that I have been involved in, one directly, and one indirectly.

7.1.1 Chevak

My doctoral dissertation (Woodbury 1981) was a description of the phonology and morphology of Cup'ik, a variety of Central Alaskan Yup'ik spoken only in Chevak, Alaska, a village of under 1000 people on the Bering Seacoast. I first went to Chevak in 1978, with plans to write my dissertation, but also, influenced by work on the ethnography of speaking, to begin audiotape documentation of language **use** in Chevak in as broad a way possible, and to base as much of my description as I could on the textual materials I collected. At that time, kids were entering kindergarten speaking Cup'ik but very little English; and adults over 45 were all monolingual in Cup'ik or nearly so. At precisely the time I came, there was an oral history movement afoot to tape-record elders, and, with the encouragement of Mike Martz—then a teacher there and now a documentary filmmaker—high school kids were going around taping their parents and grandparents telling formal narratives and reminiscences. I became involved with the kids' work, and also found many people enthusiastic about my own efforts to visit and make recordings with elders, to transcribe and translate them with community members, and to publish them in a bilingual book that ended up as the only thing I've ever had go into reprintings. At the same time, as much as I felt valued by my friends in Chevak, it was made very clear to me that most people did not feel good about recording ambient daily conversation or even narratives and talk by younger adults and children. Which was fine with me because what people **were** interested in kept me plenty busy over continued visits to Chevak through the 80s.

In Chevak today—I was last there two months ago—few people under 25 speak Cup'ik, although many parents as young as their 30s make a point of not using English to their kids. The kids generally understand day to day Cup'ik and when cajoled to say something—like their Cup'ik names—their pronunciation seems flawless.

Through the 1990s, as the reality of this radical language shift set in, there was growing unease at how much was being lost, and the inefficacy, as a means of language preservation, of an hour of Cup'ik class every day, or even daily Cup'ik-language lectures in the school by elders on cultural, historical, and environmental topics. Slowly, and inspired by successes in several nearby Yup'ik speaking communities including Bethel and Mekoryuk, the idea developed to start Head Start preschoolers in Cup'ik-only language immersion; and, because virtually all the elementary school teachers are Cup'ik speakers, to continue it, year by year, as the first cohort worked their way up the grades. Although Chevak is nearly unique in having created a school district all its own under local control, the development of a consensus for immersion took a long time and faced many hurdles, including an unfriendly regulatory environment, non-Cup'ik or Yup'ik speaking administrators with little enthusiasm for immersion; and other issues. As recently as last spring it looked like a sure thing, although now, with the emergence of George Bush's *No child left behind* program[7]—that declares schools as "in crisis" if their kids don't perform to a certain level on benchmark tests beginning at the third grade level in—yes, that's right—IN ENGLISH—then schools get defunded, can be closed down, or put in political receivership. So this has led to a new round of cold feet about immersion.[8]

Nevertheless, I have been working with John Pingayaq, Rebecca Nayamin Kelly, Peter Tuluk, and Leo Moses, teachers and community leaders, to design a documentation project whose purpose is be to serve as a resource for immersion education and for the preparation of materials for an eventual immersion program. Our goal too is to try to do as much as we can to record what is often expressed as the glimpse of ancestral life and knowledge which the eldest of the elders possess most clearly, and in so doing to carry out a vision of Cup'ik maintenance in the face of diversity that was set out in the 70s by Joe

---

[7] http://www.nochildleftbehind.gov/

[8] Writing in *The Tundra Drums*, a Bethel, Alaska, weekly newspaper, Chris Meier, co-principal of Bethel's Ayaprun Elitnaurvik immersion school, puts it eloquently:

"At its worst, the No Child Left Behind Act is an assault on the indigenous languages of America and the children and parents who speak them. The practical effect of these laws will be to further destroy the Yup'ik language and demean rural Alaskan schools. As the law is now written, it may potentially close schools, fire principals and teachers, and cut funding to schools whose parents value the use of Yup'ik as much or more than they value the use of English.

"This is not only immoral, it is illegal, and is in direct conflict with the Native American Languages Act. This law states, 'The right of Native Americans to express themselves through the use of Native American languages shall not be restricted in any public proceedings, including publicly supported education programs.'" (*The Tundra Drums*, August 29, 2002, page 6).

Friday, Ulric Nayamin, and several other Cup'ik elders. As a practical matter—following this mandate or agenda—our work will be to curate a huge collection of tapes arising not only from the students and my work in the 1970s, but even more, from vast amounts of recordings of original materials produced by Peter Tuluk in his 15 years of Cup'ik language production work at KCUK, a radio station he founded in Chevak. Furthermore, we want to extend current KCUK production to include videotaping and to include focus groups on lexical topics in order to develop a thesaurus and cultural encyclopedia.

You will notice that the mandate or agenda that we are working with clearly privileges ancestral connection and in fact fully motivates the unease that greeted my interest in younger people's speech decades earlier.

Equally interesting, for my purposes here, is what we **don't** plan to do in this project. We **don't** plan to produce interlinear glosses because we consider it a waste of time, given the specifics of the situation. Cup'ik is pretty well documented and hence, providing interlinear glosses will be something that philologists 500 years from now will be able to handle. Instead we will use the time of the few elder Cup'ik translators with wide English and Cup'ik vocabularies to produce running UN style translations of many more materials, and then have younger speakers flag the obscure words or usages for special attention. We are also considering not transcribing everything—instead, starting with hard-to-hear tapes and asking elders to "respeak" them to a second tape slowly so that anyone with training in hearing the language can make the transcription if they wish. In this way, we plan to document the documentation not by formula, but in keeping with specific needs. Part of my technical input as a linguist is to make guesses about what the "philologist 500 years from now" is going to need; that concept makes sense to me as a linguist; but as you can see, it is also consonant with the transmission agenda that informs the project.

7.1.2 Iquito

I'd now like to turn to a very different type of project. This is the *Iquito Language Documentation Project* in the Peruvian Amazon,[9] a collaborative project of the indigenous community of San Antonio de Pintuyacu, in Loreto, Perú, and a group of Texas graduate students, Chris Beier, Lev Michael, Mark Brown, and Lynda de Jong. This project was initiated in 2001, when Chris and Lev, linguistic anthropology students already involved in work elsewhere in lowlands Peru, responded to a call from San Antonio for a linguist to help them establish an Iquito language teaching program in their tiny community—a daunting task because only about 20 elders, all over 50, still speak Iquito, and because Iquito was only scantily documented in the early 1950s. After a preliminary visit to San Antonio, Chris and Lev negotiated a commitment simultaneously to work with elder speakers; train several younger teacher-field workers to write Iquito and do basic

[9] http://www.iquito.org

dictionary and text collection work; inaugurate daytime classes for school kids and evening classes for adults; and involve teachers in the teaching effort so that they can eventually take it over. Last summer—summer of 2002—Chris and Lev went back, joined by Mark Brown and Lynda de Jong, linguistics students. They raised money not only to fund their involvement in the project, but also to pay field workers to work throughout the year and to build a language study house in the community. They conducted the daytime and evening classes, which were well attended, and made considerable progress on the analysis of the language, which has been the subject of several papers by Lev and of Ph.D. qualifying papers by Mark and Lynda. They have also made some very basic language teaching materials. The most salient quality of the work—as you can easily guess—has been its bootstrapping nature, teaching even before the grammatical and lexical research is done; and teaching, indeed, before even **knowing** the language. Clearly, the design and the **needs** of a project like this, the nature of documentation, the emphasis on grammar and lexicon—contrast markedly with the Chevak case. Yet both are sensitive to both community agendas and linguists' agendas.

These are just two examples and I think it is easy to see that the more we consider different projects—particularly when local agendas are front-and-center—the less inclined we will be to think of "best records" in the abstract. Likewise—if I can go a bit negative one more time—I think we will be less inclined to design one-size-fits-all documentary programs with their own quite specific agendas and practices; this has been the tendency, in my view, of the Volkswagen DoBeS project.

On a more positive note, what I **do** find worthwhile is for linguists to **articulate** their disciplinary agendas—their system of values—when approaching documentation. This too may ultimately reveal surprising diversity, but to take a stab at it, let me mention the following as some common values on what would constitutes a good documentary corpus.

8. Linguists' Agendas: Some widely agreed-on value for documentation

First, all things being equal, a good corpus is **diverse**; diverse in situations; in participants—people carrying various different social roles; in channels such as speech, writing, e-mail; in speech genres, including conversation, narrative, oratory, verbal art, formal and informal interaction, and so on; and perhaps in different dialects or varieties or codes, if the community in question is multilingual, and if the documentary focus is on the community rather than one particular language or code.

Second, a good corpus is **large**. How large? Now, more than ever before, the technology is there for it to be arbitrarily large. It is unlikely, for example, that endangered language corpora will *ever* be as large as the mega-corpora routinely used in the

investigation of widespread world languages, yet corpora that large have proven useful even for grammatical and lexical investigation.

Third, good corpus production is **ongoing, distributed,** and **opportunisitic**. It continually grows. Many people contribute to its development. And documenters take advantage of any opportunity to record, videotape, or otherwise document instances of language use. For this to happen, documentation projects must be designed to put easily available, easy-to-use, well-diffused technologies in the hands of as many people as possible, and to train them to make high quality recordings. This is the opposite of the traditional model, where someone from outside the community controls documentation and the means for documentation.

Fourth, materials should be **transparent**. They should be properly annotated. In short they should be useable by the philologist 500 years from now. As noted, this does not necessarily mean that every text should be glossed interlinearly.[10] But minimally, everything should be competently translated into a language of wider communication; and transcriptional practices should be elucidated with links to phonetic and phonological data. Likewise, mark-up schemes should be elucidated and tested for intersubjective reliability. Moreover, it is never adequate, from a linguist's point of view, to collect lots of text on audio- or videotape and consider the documentation accomplished.

Fifth, material should be **preservable, ethical,** and **portable**. Proper metadata information should be given about each item of data, whether text, audio, video, or any other medium. Data should be archived. It should be handled so that that it migrates easily to the new technologies that emerge like clockwork every few years. In short, all efforts should be made so that data is portable in the sense developed by Bird and Simons (2002).

Sixth—to put it very broadly—a good corpus is **ethical**. It means documentation must be carried out ethically. This includes that data ownership be protected, that is, data should not disseminated to those its owners or producers do not want to have or use it. And it means—I think—that documenters should work with and respect the agendas of those with whom they are involved, especially those producing or owning the data, or having a hereditary or ancestral stake in it.

8.1 Archiving

I have written so far about data collection projects. But documentary linguistics is coming to include other central endeavors, one of which is archiving. Space prohibits me doing this

[10] I do not at all mean to suggest that segmentation and interlinear glossing can always be dispensed with. Segmentation in certain languages is so difficult that it makes practical sense to segment all texts, or at least a very large training set: Iroquoian languages are an excellent case in point, as should be clear by comparing word-level transcriptions with the morphological segmentations in H. Woodbury's (1992) rendition of a long text in Onondaga.

topic justice (but see Wittenburg's contribution to this volume). A related concern is ethical issues of internet dissemination of linguistic digital archives. Both topics require further discussion and elaboration.

8.2 Training

The final issue I wish to raise is training, the complexity — and the possibilities — of which have come very clear to me as our program in documentary linguistics at the University of Texas has grown.

There is, first of all, the issue of courses. Originally, students learned most of their documentary linguistics in a one or two semester field methods course, where a consultant speaking some language would work with the class in order to develop basic descriptive materials, or else topical papers, alongside a basic analysis. Along the way, various analytic, practical, and ethical issues might also be touched on.

As documentary linguistics has grown, I have found it harder and harder to fit everything into field methods. "Should we deal with computational data bases?" I ask myself at the beginning of a course, or should I just show them slip filing? (the programme *Linguist's Shoebox*, vs. **a** shoebox). After some disasters, I dropped it. I also started teaching a second course, *Tools for Linguistic Description*, which had the unexpected but happy additional effect of providing basic analytic training for students taking introductory graduate phonology and syntax. When my colleague Nora England joined us in 2002, we started offering courses with even more breadth, including a course she is teaching this Spring on field grammars and how to write them. We hope to offer a course on computational approaches to field work. And we **should** be offering formal training in pedagogical materials development, since that is central for many projects we and our students become involved in. In all, we feel very lucky at least to have help and to be in a linguistics department where 15 of 16 faculty members have done significant field work and can help us impart perspective.

Likewise, we are very concerned with genuinely supporting the full range of documentary research possibilities for our postgraduate students. It has sometimes been lamented that there are departments that do not allow descriptive grammars as PhD dissertations. That certainly hasn't been our problem, nor the problem of a great many departments like us. But there can be an upping of the ante. A well-articulated open letter from postgraduate students to the Australian Linguistic Society Newsletter last May to which Peter Austin (2002, see also the p14 above) calls attention, argues that not just a grammar, but more holistic documentation, including audio, video, and written text material and community usable lexical presentations, should be allowed as a PhD dissertation in Australian institutions. It is not quite something we have done; but I think it is a reasonable challenge that can be met.

Finally, perhaps the most important thing for us at the University of Texas — and what I hope will be our enduring contribution to documentary linguistic education—has been Nora's founding of CILLA, the Center for the Indigenous Language of Latin America. Rather than just a research umbrella for herself, Joel Sherzer, Megan Crowhurst, and me, the central mission of CILLA is the graduate training in documentary linguistics of speakers of Latin American indigenous languages; and by agreement, the University of Texas has generously offered us support for two new indigenous students a year for the next few years. Right now, although only in our second year, we already have five PhD students from Latin American communities with interests ranging from materials development to language planning to lexicography to discourse documentation to corpus-based syntactic research, alongside interests in the linguistics of their languages and in linguistics in general. I think that our greatest challenge will be making sure that we evolve a program that genuinely trains these students for the things they want to do, and that their communities want them to do, with their linguistic educations.

References

Austin, Peter K. 2002. Grand vision: Some issues. Text of remarks at the LREC Workshop on Computer Tools for Field Linguists, May 24, 2002. Melbourne: University of Melbourne MS.

Bird, Steven and Gary Simons. 2002. *Seven Dimensions of Portability for Language Documentation and Description* Proceedings of the Workshop on Portability Issues in Human Language Technologies, Third International Conference on Language Resources and Evaluation. http://arXiv.org/abs/cs/0204020

Bird, Steven, and Mark Liberman. 1999. *A Formal Framework for Linguistic Annotation* Technical Report MS-CIS-99-01. Philadelphia: Department of Computer and Information Science, University of Pennsylvania. http://www.ldc.upenn.edu/Papers/CIS9901_1999/revised_13Aug99.pdf

Bloomfield, Leonard. 1933. *Language.* New York: Holt, Rinehart and Winston.

Chelliah, Shobhana. 2001. The role of text collection and elicitation in linguistic fieldwork. In Paul Newman and Martha Ratliff, eds., *Linguistic fieldwork*. Cambridge, UK: Cambridge U.P.

Comrie, Bernard, and Norval Smith. 1977. Lingua descriptive series: questionaire. Lingua 42:1-72.

Gumperz, John J. and Dell Hymes. 1964. Eds. *The ethnography of communication. American Anthropologist* 66(6): Part 2.

Gumperz, John J. and Stephen C. Levinson. 1996. *Rethinking linguistic relativity.* Cambridge, UK: Cambridge U.P.

Hale, Kenneth, Colette Craig, Nora England, LaVerne Jeanne, Michael Krauss, Lucille Watahomigie, and Akira Yamamoto. 1992. Endangered languages. Language 68(1):1-42.

Hopper, Paul, and Sandra Thompson. 1980. Transitivity in grammar and discourse. *Language* 56:251-299.

Karetu, Timoti S. 1994. Mauri language rights in New Zealand. In Tove Skutnabb-Kangas and Robert Phillipson, eds., *Linguistic human rights: Overcoming linguistic discrimination.* 208-18. Berlin: Mouton de Gruyter.

Lucy, John A. 1992. *Language diversity and thought: a reformulation of the linguistic relativity hypothesis.* Studies in the social and cultural foundations of language 12. Cambridge, UK: Cambridge U.P.

McCarty, Teresa L, Lucille J. Watahomigie, Akira Y. Yamamoto, and Ofelia Zepeda. 1997. School-community-university collaborations: The American Indian Language Development Institute. In Jon Reyhner, ed., *Teaching indigenous languages*. Flagstaff: Northern Arizona University. http://jan.ucc.nau.edu/~jar/TIL_9.html

Newman, Paul, and Martha Ratliff. 2001. Linguistic fieldwork. Cambridge, UK: Cambridge U.P.

Nichols, Johanna. 1992. Linguistic diversity in space and time. Chicago: University of Chicago Press.

Samarin, William J. 1967. *Field linguistics, a guide to linguistic field work.* New York: Holt, Rinehart and Winston.

Sherzer, Joel. 1990. *Verbal art in San Blas: Kuna culture through its discourse.* Cambridge, UK: Cambridge U.P.

SIL International. 2002. *Ethnologue: Languages of the world.* 14th Edition. http://www.ethnologue.com/

Urban, Greg. 1991. *A discourse-centered approach to culture: Native South American myths and rituals.* Austin: University of Texas Press.

Wilkins, David P., 1992. Linguistic research under Aboriginal control:. A personal account of field work in Central Australia. *Australian Journal of Linguistics* 12:171-200.

Woodbury, Anthony C. 1981. *Study of the Chevak dialect of Central Yup'ik Eskimo.* Berkeley: University of California Doctoral Dissertation.

Woodbury, Hanni, ed. 1992. *Concerning the League: The Iroquois League Tradition as dictated in Onondaga by John Arthur Gibson. Algonkian and Iroquoian Linguistics, Memoir* 9.