Valuing technology:

Finding the linguist's place in a new technological universe

Jeff Good

1 Introduction

The focus of this paper is on the relationship between technology and language documentation and description.[1] It is largely inspired by Bird and Simons (2003), or more specifically, section 4 of that paper (Bird and Simons 2003:570–2), where they suggest a general model for developing recommendations for the use of technology in the creation of linguistic resources—which I term the values-desiderata-recommendations model. Unlike Bird and Simons (2003), however, my primary goal is not to use that model to generate a set of so-called best-practice recommendations for linguistic resources. Rather, I seek to expand upon the conceptual foundations and assumptions of the model so that its general

[1] This paper would not have been possible were it not for many profitable hours I have spent attending meetings of the Open Language Archives Community (OLAC), the Electronic Metastructure for Endangered Languages Data initiative (E-MELD), and the Linguistic Society of America's Conversation on archiving. Many of my ideas with respect to the relationship between technology and language documentation and description are rightfully attributed to the collective wisdom of all the other participants of these meetings. Of course, however, I am solely responsible for the content of this paper.

applicability to issues in language documentation technology can be made clearer to a general linguistics audience.

Though primarily trained as a linguist, the perspective I will try to take on in this paper is that of an idealized "technician". The reason for this is that in my experience the greatest barrier to the proper use of technology in language documentation and description by linguists is that they tend to focus on technical "details" rather than taking a broader view of the general structure of what one might call the "linguistics-technology" interface. In taking on this perspective, I hope to give so-called "ordinary working linguists" a better sense of how the technician understands and reacts to the needs of linguists, thereby putting them in a better position to make informed choices about how they use technology in their work. I also hope to give such linguists some useful conceptual tools for participating in debates about the role of technology in language documentation and description. Up until now, the bulk of this discussion has been largely conducted by a handful linguistic "digerati", which is unfortunate because its outcomes are likely to have consequences for all linguists using language documentation technology in their work.

Much of the paper will be summarizing in nature, rather than attempting to present new arguments and ideas. I will specifically contrast two recent lines of work which I think are likely to have a particularly strong impact in future work on technology and language documentation and description. (In fact, they already have had major impact.) The first of these is work done by Steven Bird, Gary Simons, and their associates within the Open Language Archives Community (OLAC). The second of these is work which has argued for the necessity of a new academic field focusing on issues relating to language documentation,

for example, Himmelmann (1998) and Woodbury (2003). The two lines of work differ from each other in ways that are potentially significant for future developments in language documentation technology. Exploring this "clash" will offer a useful case study illustrating the kinds of tensions that will arise as our field becomes more and more dependent on technological developments that it has little control over.

Throughout the paper, I will often make recourse to the heuristic devices of an idealized "technician" and an idealized "linguist" for expositional purposes, but these should not be taken too literally. In section 2, I summarize important features of the first of the two lines work just discussed, with a focus on Bird and Simons (2003). In section 3, I summarize important aspects the second line of work on issues in language documentation. In section 4, I conclude by highlighting potential conflicts between the two lines of research.

A potentially surprising feature of this paper, given its stated focus on technology and language documentation and description, will be the relative lack of discussion about specific technologies. "Technobabble" terms like XML[2] or Unicode[3] may come up in some spots. But, the ultimate conclusion will be that in the end, for the linguist, technical problems are secondary. The primary concerns instead revolve around *values*. In other words, the question linguists should be asking is, "What am I trying to do here?" rather than, "How am I going to do it?"

2 Technology and linguistics

[2] <http://www.w3.org/XML/>

[3] <http://www.unicode.org/>

2.1 *Introduction*

In this section, I will discuss a line of work on the relationship between language documentation technology and linguistics well exemplified by Bird and Simons (2003). While this work is not devoid of reference to speaker-community concerns regarding language documentation and description (see, e.g., Bird and Simons 2003:576), its primary interests to revolve around research uses of language resources, with a specific focus on issues regarding the preservation of language data (as opposed to languages themselves).

Other work that I classify as belonging to this category are the collected documents of the Open Language Archives Community (OLAC)[4], work done within the Electronic Metastructure for Endangered Languages Data initiative (E-MELD), in particular the E-MELD School of Best Practices in Digital Language Documentation[5] (see Boynton et al., this volume), and more recent work on linguistic ontologies (see, e.g., Farrar and Langendoen (2003) and Farrar and Lewis (2005)). Thieberger and Jacobson (this volume) would also fit fairly comfortably into this category.

2.2 *Overview of Bird and Simons*

Bird and Simons (2003) represented a breakthrough for the linguist's understanding of the relationship between technology and linguistic resources. There were three important achievements in the paper. The first, which has already received a good deal of recognition

---

[4] <http://www.language-archives.org/documents.html>

[5] <http://e-meld.org/school/>

(see, e.g., the E-MELD School of Best Practices in Language Documentation and Description as discussed by Boynton et al. (this volume)), is the presentation of an important set recommendations for best practices for the use of language documentation technology. These recommendations range from the relatively specific and easy to follow (e.g., use Unicode character encodings in electronic text resources) to the more general and complex (e.g., markup data using XML accompanied by a DTD or Schema describing the XML markup) (Bird and Simons 2003:575).[6]

The second achievement of Bird and Simons (2003) was its codification of a number of important practical issues raised by the increasing use of digital resources under the rubric of *portability*, a notion encompassing portability across different computational environments, communities, domains of usage, and time. They break down the notion of portability, as it applies to linguistic data, into seven dimensions: *content*, *format*, *discovery*, *access*, *citation*, *preservation*, and *rights*. It would seem to be too soon to say if these seven dimensions will be generally accepted as an adequate breakdown of the concept of portability. Nevertheless, they have already proven to be useful both as a pedagogical tool in language documentation technology instruction and as a useful way of organizing best practice recommendations, as seen in the E-MELD School of Best Practices in Digital Language Documentation (see, for example, E-MELD (2005c)). The third achievement of

---

[6] Since it is liable to misinterpretation, it is worth pointing out here that the term best practices should not be understood to mean something like "required practices". Rather. it refers to a set of practices that are considered to establish an ideal way of working, given our present understanding of a technical problem. As such, they are often liable to change and, in some cases, deviation from best practices will be well justified.

Bird and Simons (2003)—and the one which will be of greatest interest here—lies not in any of its specific recommendations but, rather, in its attempt to develop a general model through which linguists can devise best practice recommendations for digital linguistic resources. I label this the *values-desiderata-recommendations model* and discuss it in detail in section 2.3.

2.3 *The values-desiderata-recommendations model*

In the values-desiderata-recommendations model (VDR), best-practice recommendations for the use of language documentation technology are not understood to exist in isolation. Rather, they are conceptualized as deriving from general desiderata for best practices, divorced from any particular technological context, which are, in turn, derived from statements about the values of the linguistics community. In (2) I give an example of the application of the VDR model, adapted from Bird and Simons (2003), covering the issue of accountability for the content of a grammatical description.

(2) VALUES, DESIDERATA, AND RECOMMENDATIONS FOR *ACCOUNTABILITY*

a. **Values:** Linguists value the ability of researchers to verify language descriptions. (Bird and Simons 2003:571)

b. **Desiderata:** Best practices deriving from this are those that result in access to the documentation that lies behind the description. (Bird and Simons 2003:571)

c. **Recommendations:** (i) Provide the full set of documentary resources on which language descriptions are based; (ii) When texts are transcribed, provide the primary recording; (iii) Transcriptions should be time-aligned to the underlying recording in order to facilitate verification; (iv) When recordings have been significantly edited, provide the original recordings to guarantee authenticity of the materials. (Bird and Simons 2003:574)

The VDR model is schematized in Figure 1 which illustrates a relationship between values, desiderata, and recommendations where values determine desiderata which, in turn, determine best-practice recommendations.

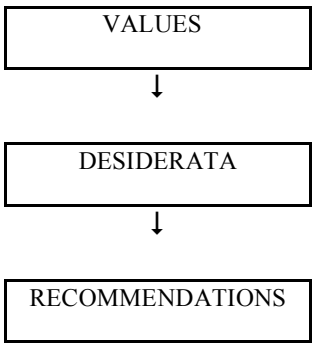


Figure 1: The values-desiderata-recommendation model

From the perspective of the ordinary working linguist, in fact, there should probably be a fourth component to the model in Figure 1 in which abstract recommendation statements are translated into concrete instructions. This component would be labeled *implementation*.

In addition to providing a methodological blueprint for the development of sound technological recommendations, the VDR model is useful in two other ways: (1) helping to pinpoint the source(s) of disagreements; and (2) delimiting the responsibilities of the linguist and the technician.

In the first case, a claim embodied by the VDR model is that, when the technological solutions employed by two groups of linguists are in conflict, the source of the conflict should be traceable either to decisions in how best practice recommendations realize particular desiderata, how desiderata are derived from values statements, or, in the most extreme case, differences in the underlying values themselves. This last possibility will be discussed in section 4.

In the second case, the VDR model has very clear implications as to what the responsibilities of the linguistics community are with respect to the development of recommendations for the use of language documentation technology. Specifically, it indicates that it is the responsibility of linguists, not technicians, to devise statements of their values for linguistic resources, how those values translate into desiderata for technological solutions, and how those desiderata translate into broad best-practice recommendations. The technician only comes in at the "end" of the process—and turns recommendations into an actual implementation. This is not to say that the community of linguists needs to agree on one set of values or desiderata, or one set of recommendations realizing those desiderata.

This scarcely seems possible, let alone advisable. Rather, it simply places the burden on linguists to devise clear statement of their own needs before beginning to implement a technological solution to a problem.

Of course, practice may often deviate from the abstracted process implied by the VDR model, and it will often make sense to bring in a technician into the discussion before implementation. For example, while the burden is clearly on linguists to formulate value statements, translating them into desiderata and recommendations which make sense both to linguists and to technicians will typically require at least some technical input, either from a linguistically-informed technician or a technically-inclined linguist (or, ideally, both).

2.3.1 *Implementing the model*

Typically, when we talk about "implementation" we mean the creation of a purely technological solution to help solve a particular problem. However, if abstract models like the VDR model are to result in concrete recommendations, they, too, will typically need to be implemented in one way or another. One option is to make use of an informal process: Over time, as the result of published work and discussions at scholarly meetings, the relevant community reaches consensus with respect to recommendations in much the same way that practitioners of a given linguistic theory reach consensus about the best way to analyze certain phenomena within a given framework.

However, in the technological realm, this sort of informal implementation has an important drawback: It typically does not result in recommendations which are precise enough for consistent technical implementation. In particular, recommendations devised

informally often suffer from the problem of not being *authoritative*—that is, there is no one place a technician can turn to in order to obtain all of the information they need to produce an appropriate implementation. This will force them to fill in some of the details on their own, inevitably resulting in two kinds of problems. First, some of these details will be implemented incorrectly from the perspective of the linguist—not due to any incompetence on the part of the technician, crucially, but, rather, due to a lack of proper understanding of the linguistic problem. Second, different technicians will each create their own reasonable, but ultimately incompatible, implementations—often along dimensions of minimal relevance to linguistic research—which will hinder collaboration among linguists who happen to have adopted different technological implementations to address the same basic problems.

Within linguistics, one of the most well-known examples an authoritative source of recommendations is the phonetic transcription system of the International Phonetics Association (IPA). It is clearly authoritative in the literal sense giving it a critical advantage over other transcription standards. Everyone knows where to look to discover the latest version of the standard (in the case of the IPA, the International Phonetic Association's (1999) handbook, for example), and the standard is sufficiently detailed that there is relatively little ambiguity in the appropriate interpretation of the symbols used.

At the same time, the realm of the IPA is also instructive in understanding the problems that arise when there is no authoritative standard. Before the advent of Unicode character encoding standard, there was no authoritative recommendation regarding how to encode non-standard characters—including many phonetic symbols—in computer fonts. The result was that it was quite difficult to shift between different phonetic fonts since there was

no guarantee that each would encode non-standard characters in the same way. The encoding for the character ɛ in one font might correspond to the character ə in another, for example. Transferring data between fonts was, therefore, not a simple matter of "changing the font" in a word processor. It also included manually re-encoding characters using the encoding scheme of the new font. The lack of an authoritative standard for special character encoding impeded, in particular, the sharing of documents between two linguists who did not make use of precisely the same set of fonts.

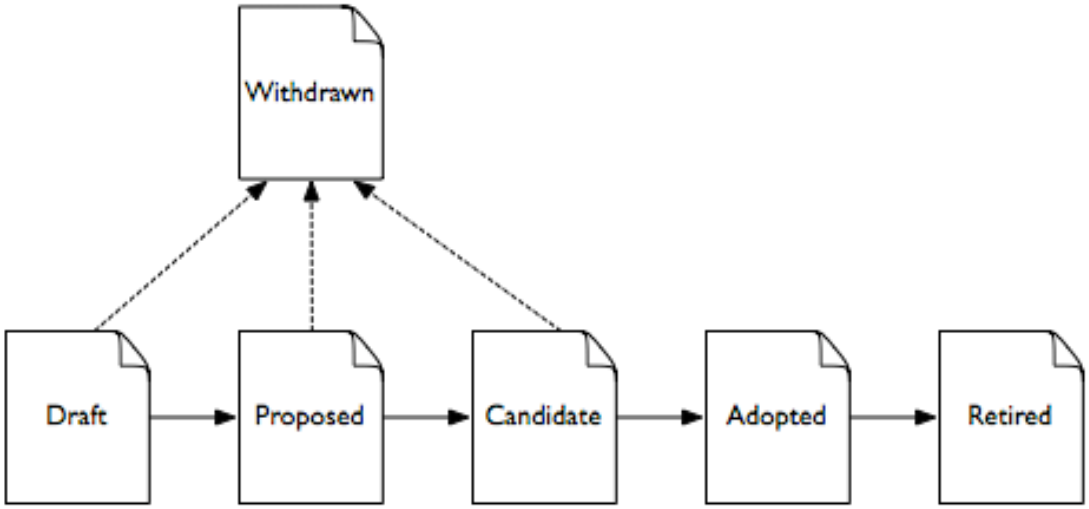
The acceptance of the Unicode Consortium's character encoding standard, has, in great part, solved this problem. Any two fonts containing IPA characters using the Unicode standard should, in principle, be interchangeable, without requiring re-encoding of any characters. This is not to say that the problem of character encoding is completely solved—far from it, as indicated by the need for a group like the Script Encoding Initiative (Anderson 2003), which works to get minority and historical scripts and characters into the Unicode standard. However, despite such issues, it is clear that the authoritative Unicode standard represents a vast improvement over the earlier situation.

Returning to the VDR model, in at least one domain of linguistics and technology—the domain of metadata standards—it has been implemented in a formal sense. The relevant authority is the Open Language Archives Community (OLAC)[7]. The implementation of the model is described in Simons and Bird (2006) which lays out the process through which different types of OLAC documents can be created, and these documents contain explicit

[7] <http://language-archives.org>

recommendations for linguistic metadata.[8] Metadata is "data about data", the sort of information that is used for indexing, searching, and sorting, including information like titles, authors, dates of creation, etc., associated with a resource of well-defined set of resources. (For a non-technical introduction to metadata in the OLAC context see Good (2002).) Critically, these OLAC documents are intended to describe a set of recommendations for linguistic metadata with sufficient detail that they can be implemented consistently by different technicians. The most important of these documents is Simons and Bird (2008), which defines the OLAC metadata standard.

Simons and Bird (2006) is of particular interest here because of the way it implements the VDR model via a series of community-developed documents. The "life cycle" of an OLAC document is schematized in Figure 2.



[8] Many of the details of this implementation are not at all unique to OLAC. Similar models can be found in the document creation processes of other technical bodies like the World Wide Web Consortium or the International Organization for Standardization. The OLAC example is used here because it was developed by an organization whose aims are to specifically create standards for the linguistics community.

Figure 2. The OLAC Document Process

While, on the surface, the schematization in Figure 2 looks to simply represent a procedure for creating a succession of documents. In fact, it is something more powerful: It represents a method to achieve working balance between the needs of the linguists to be able to debate and discuss potentially contentious issues relating to language documentation technology and the needs of technicians to be able to look "in one place" to discover the consensus of the linguistics community with respect to particular problems. (In the case of OLAC, the method depicted in Figure 2 has been applied primarily to the development of metadata standards—but the procedure could be used more generally.) When a document is *Adopted*, it represents a set of recommendations for linguistic metadata that are officially recognized by the Open Language Archives Community.[9] Debate takes place within the stages of document creation labeled *Draft*, *Proposed*, and *Candidate*. In the OLAC context, the debate is open to anyone who takes an active role in commenting on and editing the documents. This reflects a decision on OLAC's part to encourage broad community involvement in the development of its standard. However, the determination of what community will be allowed to participate in document creation is, in principle, independent of the determination of the appropriate succession of documents en route to standards creation.

---

[9] Informally, the distinction between a "standard" and a "recommendation" is that a standard is a recommendation that is required for participation in a given community—in this case OLAC.

There is no automatic expectation that a draft document will eventually be adopted. If the members of OLAC can not reach consensus on a given issue, then no standard or recommendation can be adopted in that area. There is nothing inherently problematic about this, though obviously problems could arise if consensus is not reached in a domain where multiple implementations are likely to be developed even in the absence of a standard.

The documents created as a result of the OLAC process may or may not contain explicit statements of the values and desiderata that resulted in particular recommendations (though one could imagine adding requirements to this effect within the OLAC guidelines if the community deemed it desirable). However, the process itself through which these documents are created is intended to make sure the community has clearly thought through its values and developed desiderata based on those values. Since much of the document creation process is conducted via online mailing lists, in many cases, discussions archived on those lists will record the relevant debates.

Though I am not aware of any other cases within linguistics, one can imagine other ways to formalize the VDR model. Any successful formalization will share a critical property with the OLAC document process, however: It will have to result in authoritative recommendations that any technician can refer to when designing an implementation.

2.3.2 *Summary*

In this section, focusing on Bird and Simons (2003), I discussed a line of recent research on the relationship between technology and language resources which is primarily

concerned with developing digital standards for preserving linguistic data. Of particular interest was the idea that technological best-practice recommendations should be viewed as, ultimately, emanating from values shared by the linguistics community, which inform the creation of desiderata, which, in turn, result in concrete, implementable recommendations. In the next section, I will turn to another line of research where understanding the relationship between technology language documentation and description also plays an important role, which I refer to as *language documentation studies*.

3 Language documentation studies

3.1 *Introduction*

In this section, I discuss a line of research, which I refer to here as *language documentation studies*, which is focused on issues regarding language documentation (as opposed to language description). People working in this area, to this point, have typically been linguists (though language activists have also played an important role). However, unlike the work discussed in section 3, they are less concerned with understanding how technology fits into linguistics as presently understood than with mapping out a new academic discipline whose existence, in large part, is dependent on the emergence of new recording technologies.

Within a relatively short span of time, language documentation studies has shifted from being a proposed research program to a distinctive domain of inquiry with many of the

hallmarks of a true academic discipline: thematic volumes (e.g., Gippert, Himmelmann, and Mosel (2006)), periodicals (e.g., the journal *Language Documentation and Conservation* and the *Language Documentation and Description* series of the Hans Rausing Endangered Languages Project), degree programs (e.g., the Endangered Languages Academic Program at the School of Oriental and African Studies)—and, perhaps most importantly, funding initiatives, in the form of, for example, the Dokumentation Bedrohter Sprachen program (funded by the Volkswagen Stiftung), the Endangered Languages Documentation Program (funded by the Arcadia trust), and the Documenting Endangered Languages program (funded by the United States National Science Foundation and National Endowment for the Humanities). Two important programmatic works in the subfield are Himmelmann (1998) and Woodbury (2003), which lay out various foundational issues.[10]

3.2 *The fuzzy boundary between documentation and description*

While it is straightforward to make a conceptual distinction between language documentation and language description (as done by Himmelmann (1998)), there are a number of activities routinely undertaken by linguists which seem to fall in between the two categories. The most obvious of these is the task of phonetic transcription. On the one hand, transcription clearly

[10] In this context, it is worth nothing that, as pointed out by various authors, including Woodbury (2003:34), documentation, in a broad sense, has been important to the discipline of linguistics since at least the time of Franz Boas. What is new about language documentation studies is a shift in perspective wherein documentary work is treated not simply as a necessary practical step towards descriptive work but, rather, as an independently theorizable domain of inquiry in its own right.

involves a good deal of analysis of raw data, insofar as it involves parsing the speech stream into segments and segments into words and sentences. This would seem to disqualify transcription from being part of language documentation. On the other hand, a transcription—even if it includes word and sentence-level parsing—is not a prototypical instance of description since it describes only one speech event and does not constitute the sort of generalized grammatical statement canonically associated with "language description".

A similarly ambiguous resource is a recorded grammatical elicitation session. To the extent that such a recording would contain primary linguistic data it would seem to qualify as language documentation. However, the fact that the structure of such an interaction would be driven primarily by descriptive linguistic concerns would seem to place such recordings outside of the core set of resources collected when conducting language documentation (see, e.g., Himmelmann 1998:170).

The existence of such categorial misfits is of great practical importance since they help us to understand why the distinction between language documentation and language description has not always been made. Furthermore, the fact that certain important resources straddle the boundaries between these two categories has meant that most field linguists have also straddled that boundary, serving simultaneously as language documenters and language describers. Real-world practice has led to conceptual conflation.

But why does this conflation matter and why was it more or less ignored until recently? The short answer is: Technology—specifically, recording and dissemination technologies (see Woodbury (2003:36) for further discussion on this point). Such technology

breaks the practical link between recording and writing. Before the advent of recording technology, writing was the only method available for recording linguistic events. And, writing meant transcribing. Thus, the primary record of an event was typically one of the "in-between" resources, with some properties of documentation and some of description, allowing documentation and description to be conflated with relatively little practical consequence.

However, recording technologies alone were insufficient to result in the creation of a new subfield—after all, audio recording has been widely available for decades, but language documentation studies has only emerged recently. A second critical set of technological developments also had to take place facilitating the creation and dissemination of recordings alongside linguistic analysis of those recordings. This was made possible by the development of digital audio, video, and text encoding, allowing recordings to be stored on computers directly associated with relevant text-based documentation and description. The rise of the internet, further allowed such resources to be easily disseminated making them much more valuable than they would have been otherwise. The conjunction of these technologies resulted in a shift in conceptions regarding what constituted a "record" of a language from one where traditional print outputs were largely privileged to one where a more complex object comprising primary recording accompanied by low-level and high-level analysis was considered the ideal. This conceptual shift created an opportunity for a new round of theorizing regarding the nature and goals of language documentation versus language description.

3.3 *The consequences for linguistics and language documentation*

The rise of much more powerful language documentation technologies raises at least two related issues: (1) How should this affect the practices of language documentation? And (2) what are the consequences of the existence of these more "faithful" audio and video primary documentary resources for linguistic theory and practice?

As far as I am aware, one finds only one general class of responses to the first question in the published literature. This is, roughly speaking, that the nature of language documentation has been so profoundly affected by these new technologies that it should no longer be considered only a set of practices. Rather, it should consist of practices derived from underlying theories and, as such, it constitutes a new academic field of inquiry. This point of view is espoused by Himmelmann (1998), Woodbury (2003), and Austin (2003a, 2006), for example, thus allowing Himmelmann (1998:184) to write, "In language documentation, as in many other *sciences*. . . [emphasis added]".

Some of this same literature explicitly or implicitly answers the second question posed above by saying that this new, theorizable kind of language documentation is a subfield of linguistics. The clearest indication that, for example, Himmelmann (1998) or Woodbury (2003) take this to be the case is their name for this field of inquiry: *documentary linguistics*. This label can be opposed to the more agnostic one found in Austin (2006) of

language documentation. To avoid making a judgment on this issue myself, I follow Austin's lead and use the term *language documentation studies* here.[11]

3.4 *Some features of language documentation studies*

So, what does this field of study look like? In answering this, we should first accept Austin's caveat that "language documentation is a developing field that has emerged only recently and that is undergoing rapid change in terms of both theory and practice (Austin 2006:88)". Nevertheless, there seems to be a set of core issues that can be expected to form some of the field's central concerns in at least the medium term. These are given in (3).

(3) a. For a given language, what constitutes an ideal documentary corpus?
b. What potential uses of language documentation need to be anticipated and supported?
c. What methodological practices constitute "best practice" in language documentation?

[11] Of course, it is not immediately obvious that the "field of language documentation" is truly a proper academic field as opposed to being, perhaps, a collection of good methodological practices. I accept the idea that there is such a field, here, following the arguments in Himmelmann (1998) and Woodbury (2003)—however, I believe it is an open question whether most linguists would accept this.

The question in (3a) has been addressed from both theoretical and practical viewpoints. On a theoretical level, for example, Himmelmann (1998:176–183) makes use of the notion of *spontaneity* as a universal parameter along which different speech events can be categorized, and he argues that this parameter may be useful in gauging the extent to which a given corpus constitutes a representative sample of the linguistic practices of a community. On a more practical level, Woodbury (2003:47) has addressed the question by producing some desiderata for a good corpus including that it should be portable (in the sense of Bird and Simons (2003)) and ethical.

The most conspicuous feature of the way authors in language documentation studies have answered the question in (3b) is their focus on community uses for documentary and descriptive materials. This can be seen in, for example, Himmelmann (1998:188–9) and Woodbury (2003:43–46), and it comes through particularly clearly in work like Nathan (2004) which places importance on the mobilization of documentary and descriptive materials "into usable materials for practical language support (Nathan 2004:154)". This is also an important theme in the recommendations of Golumbia (this volume) to consider the role that language websites may play in shaping a community's internal and external identity. Csató and Nathan (2003:74) even tie the ultimate success of language documentation studies to whether or not it has a positive effect on the vitality of the languages being documented:

> Documentary linguistics is expected to evolve into a specialised pursuit whose success will be measured in part by the vitality of the languages described and by the successful impetus to new research and publication on the language. It should be

differentiated from a linguistics that works with derived data in pursuit of theoretical, technical, or even archival concerns.

Of course, authors like these are also concerned that the products of language documentation and description should be of value to the academic community. But, this is to be expected in work by academic authors. Their additional emphasis on speaker community concerns is noteworthy precisely because it differentiates them from the linguistics community in general which, though certainly not lacking in work discussing the relationship between community needs and linguistic research (see, e.g., Dauenhauer and Dauenhauer (1998), and England (1998)), can not be uniformly characterized as having an interest in speaker community uses of linguistic resources.[12] Though, in principle, an interest in speaker community uses of linguistic materials should be independent of new technological developments, in practice, there is a connection here: Audio and video recordings are likely to be of more interest and use to many speaker communities—in particular to those communities without a tradition of literacy in their native language—than more traditional transcriptions. Thus, technology is allowing the linguist to create outputs of a type that are more readily adapted to community needs than earlier forms of documentation, making it possible to address community concerns in ways that were not feasible before.

With respect to the question in (3c), much current work in language documentation studies focuses on technological aspects of language documentation methodology. This work can be highly specific in nature, for example, consisting of reviews of different recording

[12] TLadefoged's (1992) response to Hale et al. (1992) is indicative of such a division.

devices for use in the field (see, for example, the three reviews in Nathan (et al. 2005:3–7)). It can also be quite general discussion as in discussions of what kind of metadata is needed for language resources (Nathan and Austin 2004) or surveys of current practice in tool use by linguists (Salffner 2005). Thieberger and Jacobson (this volume) would also fall into this latter category.

Outside of technological issues, an important area of debate within language documentation studies with respect to the question in (3c) focuses on what we may broadly label ethical issues. As with technological practices in language documentation, this set of issues can range from the quite specific to the dauntingly general. On the specific side, Johnson (2004:147), for example, discusses practical strategies for ensuring that language resources are associated with appropriate intellectual property rights information. On the general side, work like that of Grinevald (2003) discusses a range of concerns about appropriate relationships between field linguists and the speakers of the languages they are documenting and describing. Thus, while many of the methodological questions raised in work on language documentation studies focus on technological issues, this is far from an exclusive focus.

3.5 *The values of language documentation studies*

One of the main points of this paper is elaborating the values-desiderata-recommendations (VDR) framework, discussed in section 2.3, for use in developing recommendations for the use of technology in language documentation and description. While I am aware of no work

within language documentation studies that explicitly lays out a set of values for the field, it is fairly easy to discern at least some core values shared by its practitioners. One of the most important values of those working in the area is probably the one given in (4).

(4) VALUE STATEMENT FOR DOCUMENTARY CORPUS COMPOSITION

The sum of the documentary and descriptive resources of the language variety of a community should be representative of the speech practices of that community during the period of documentation.

The statement in (4), at first, might seem so obvious that it hardly needs stating. However, given that many practitioners of linguistics, particularly within the generative tradition, see mental representations of grammar as their primary domain of study, the focus of language documentation studies on documentation of speech practices as opposed to, say, abstract grammatical systems clearly needs to be recognized as a value specific to this subcommunity.

The value statement in (4) would seem to be what underlies some of the desiderata that have been developed to in response to question (3a), discussed in section 3.4. I give some proposed desiderata emanating from (4) in (5).[13]

(5) DESIDERATA FOR DOCUMENTARY CORPUS COMPOSITION

[13] Woodbury (2003:46) uses the label *values* for what are classified as desiderata in (5b). Following the discussion in section 2.3, I believe they are better classified as desiderata in the model developed here.

a. SPONTANEITY: The corpus of documentary and descriptive materials of a language should encompass speech events of differing spontaneities. (Himmelmann 1998:176–183)

b. ONGOING, DISTRIBUTED, AND OPPORTUNISTIC: There should be no set limit to collection of materials and documenters, the work of collection should be spread among many individuals, and documenters should be prepared to take advantage of opportunities to record language use as they arise. (Woodbury 2003:47)

Another important value which seems to be held by those working in language documentation studies is given in (6). Some desiderata emanating from (6) statement are given in (7).

(6) VALUE STATEMENT FOR PORTABILITY OF DOCUMENTARY CORPUS CREATION

The materials collected as part of a documentary corpus should be portable across time, medium, and community.

(7) DESIDERATA FOR PORTABILITY OF A DOCUMENTARY CORPUS

a. TRANSPARENCY: The corpus should be annotated in a way which would allow a philologist in the distant future to interpret its content. (Woodbury 2003:47)

b. PRESERVABLE: The resources in the corpus should be archivable and archived to ensure their longevity. (Woodbury 2003:47)

An important aspect of the value statement I have given for language documentation studies in (6) is that it would seem to be shared by a wider community of linguists as well, as evidenced by Bird and Simons (2003). This overlap may lead one to believe that the values of language documentation studies and linguistics are largely compatible. However, as discussed in section 3.4, a good deal of the work language documentation studies appears to be motivated by something like the additional values statement given in (8).

(8) VALUE STATEMENT FOR COMMUNITY AGENDAS IN LANGUAGE DOCUMENTATION

A documentary agenda should be informed by both researchers' agendas and speaker community agendas.

The value statement in (8) seems to be implicit in Woodbury (2003:43–46) as well as Himmelmann (2006:17) and comes through quite clearly in Nathan (2004:155) who views mobilization of resources as even more important than documentation and archiving in cases where "languages are ceasing to serve social and cognitive functions". This value is also apparent in some of Golumbia's (this volume) recommendations for language websites.

Of the three value statements given in (4), (6), and (8), the one in (8) seems to be the most likely not to be shared among all researchers in language documentation studies. Himmelmann (1998:188–189), for example, discusses the role of community agendas in documentation but it is not clear whether he views this as something that needs to be considered for practicality or truly sees a statement like the one in (8) as underpinning this new field. More strikingly, the high value that Johnson (2004) places on archiving documentary materials is seemingly at odds with the importance that Nathan places on mobilization since the multimedia resources which are ideal for mobilization "are not easily archived" (Nathan 2004:156). Whether this apparent conflict stems from different values, different emphasis placed on potentially competing values, or less fundamental causes is simply not clear.[14] For linguists, this lack of clarity may set the stage for an exciting debate. For technicians, it may leave them at a loss when deciding what implementational paths they should choose.

3.6 Summary

In this section, I introduced some of the major features of the new field of inquiry of language documentation studies. While the issues that this field attempts to tackle are not all

[14] In this context, it is worth noting that Johnson (2004:140) views archives as having an important role in language maintenance and revitalization. So, perhaps she shares the value in (8) with Nathan (2004) and the primary locus of disagreement are desiderata or recommendations emanating from that value, with Johnson, but not Nathan, seeing archiving as a critical step in eventual mobilization.

technological in nature, technology has played an important role in its inception. In the next section, I will compare and contrast some of the technical requirements for language resources suggested by work like that of Bird and Simons (2003) with those suggested by work within language documentation studies.

4 A case study in conflicts

4.1 *Similarities in the two lines of research*

Neither of the two lines of research discussed above would exist if it were not for the tremendous technological advances of the last several decades. Work like Bird and Simons (2003) is quite explicitly a reaction to problems of data preservation and access which have been triggered by the use of new technologies for linguistic research. And, work in language documentation studies would be mostly theoretical if it were not for the rise of new technologies which allow linguistic communication to be recorded and disseminated in ways which were previously impossible. A second feature these two lines of research have in common is that each is being undertaken primarily by linguists—even though this would not appear to be a logical, or even a practical, necessity. Finally, broadly speaking, each shares the value of portability for language resources (though they might differ on details on what aspects of portability are most critical).

A result of these similarities is that, if we look at "surface" aspects of work in both these areas, there is a good degree of convergence. For example, both lines of research are concerned with audio and video recording techniques and standards (compare, for example, Austin (2006) within language documentation studies with E-MELD (2005, 2006)), both are concerned with the diverse roles different individuals may have in resource creation (see, for example, Nathan (2004:158) within language documentation studies and Johnson (2003) within OLAC), and both are especially concerned with issues relating to resources documenting and describing endangered languages (see, for example, Woodbury (2003:37–39) and Bird and Simons (2003:570)).

4.2 *Differences in the two lines of research*

If we step back from such particulars, however, we see a critical difference in values of the two lines of work. Specifically, language documentation studies values community involvement while work like Bird and Simons (2003) is agnostic on this count. In fact, it seems appropriate to be more general and say the field of linguistics itself is agnostic on this issue. This is not to say that individual linguists, or even whole linguistic subcommunities, do not value considering community agendas in their research—most authors in language documentation studies, to this point, have been linguists, of course. Similarly, the Australian Linguistics Society (ALS) officially recognizes the linguistic rights of (Australian) Aboriginal and Torres Strait Islander communities, which, among other things, includes the right to "request the linguist to consult with relevant community organizations where

appropriate."[15] ALS's linguistic counterpart in the U.S., however, the Linguistic Society of America (LSA), has adopted no statement of ethics on appropriate relationships between field linguists and Native American communities, leaving the issue to be decided on an individual basis.

From the perspective of the technician, this difference in values is of tremendous consequence. Technical support for academic linguistics is quite distinct from technical support for speaker communities. Documenting the full range of distinctions is outside the scope of the present paper. However, Table 1 lists some different technical requirements across six of Bird and Simons' (2003) seven dimensions of portability. Listed requirements for the academic community are adapted from ones found in Bird and Simons (2003). Requirements for speaker communities are developed based on discussion found in the references on language documentation studies given in section 3. As discussed, there are also many ways in which the requirements overlap—Table 1 is specifically exemplifying cases where they do not.

DIMENSION	ACADEMIC REQUIREMENTS	COMMUNITY REQUIREMENTS
CONTENT	Map terminology used to common ontology of linguistic terms	Avoid specialist terminology
FORMAT	Provide one or more human-readable versions of the material	Provide versions of the material in formats the community can use

[15] The entire statement of the linguistic rights that ALS recognizes for these groups is published in no. 84/4, October 1984 and can be found online at: <http://www.als.asn.au/activities.html#rights>.

DISCOVERY	List all language resources with an OLAC repository	List all language resources in a community-accessible location
ACCESS	Document the process for access as part of the metadata	Devise and implement strategies for resource mobilization
CITATION	Provide a means for citation of all produced resources	Develop a means of citation that clearly indicates the community's role in resource creation
RIGHTS	Ensure that resources may be used for research purposes	Subordinate scientific interests to community interests

Table 1. Comparing possible academic and community requirements for language resources

I have, of course, "stacked the deck" in Table 1 to emphasize differences rather than similarities. Nevertheless, it should be clear that there are substantial distinctions in academic versus community desiderata for technological solutions to problems in documentation and description. In many cases, the desiderata are not fundamentally incompatible. For example, in the dimension of format, one of the great advantages of digital resources is the ease with which a given resource can be expressed in multiple formats. A paper dictionary only gives one method of access to a lexical description of a language. An electronic lexical database can generate many layouts for the same basic data with relative ease.

However, there are also cases where the desiderata appear to be fundamentally incompatible—for example, the desiderata listed under the dimension of rights. This incompatibility, though, turns out not to be of great technological consequence since both desiderata, in the end, would result in the need for the implementation of a system of rights

management. The broad outlines of such a system would be the same for researchers or community members. The differences would simply lie in the area of setting parameters of access and properly associating different users with appropriate access rights.[16]

The most problematic incompatibility between academic and community requirements, however, lies in the "missing" dimension in Table 2, which covers only six of Bird and Simons' (2003) seven dimensions of portability. This is the dimension of *preservation*. In the context of language documentation and description—and especially the documentation and description of endangered languages—the term preservation is dangerously ambiguous. It can refer to the preservation of language resources or the preservation of languages themselves. The former is the sense intended by Bird and Simons (2003:567). The latter, however, may be of most interest to the speaker community.[17] If the community's greatest concern is preventing language loss, they might, for example, place a higher value on mobilization than archiving. This could result in the loss of improperly archived resources over time and in the collection of less material than would have otherwise been collected. Then again, if mobilization allows the language to survive, the ultimate result could be an abundance of materials of the sort that can only be collected from a healthy, thriving speaker community. From the technician's perspective, there is no right or wrong choice here. There is, however, a choice to be made, one with important technological

[16] This is not to say that digital rights management is easy to solve technically. Rather, it is simply the case that academic needs and community needs are broadly similar in this area. So, a solution for one community will likely carry over into the other community.

[17] The Linguistic Society of America's Committee on Endangered Languages and their Preservation makes use of both senses in its mission statement.

consequences.

For example, it is immediately apparent that mobilization has requirements that are technically quite distinct from archiving. This can be clearly seen in the discussion in Nathan (2004:158) who cites the specific need for a multimedia developer and a graphic designer in certain kinds of mobilization projects. A purely academic project would probably favor an electronic archiving specialist over, for example, a graphic designer. This reflects the simple fact that from a pure data preservation perspective, an "ugly" resource is better than no resource, but from a language maintenance and revitalization perspective, an ugly resource may have little value at all. Since money available for documentation projects is generally quite limited, the ideal solution to this problem of hiring both a graphich designer and an archivist will, of course, generally be unavailable.

4.3 *Evaluating values*

We have, then, arrived at the point where it should be clear why a paper with "technology" as its focus would spend so much time talking about non-technological matters like new lines of research in linguistics and linguists' "values". This is because the central questions that linguists need to address with respect to the role of technology in language documentation and description are, in fact, questions about their values. If their values are made explicit, they will be able to develop informed desiderata which will help them choose appropriate technology for their work. If their values are implicit—and, therefore, quite likely somewhat

muddled—their technological projects may very well be doomed to failure when the technicians they hire make reasonable, but ultimately inappropriate, choices.

It is likely to be the case that different groups of linguists will have incompatible values—as we have seen the value of preserving data is not always consonant with the value of preserving languages. There's no reason to believe such disagreements are inherently a problem. However, it certainly would be a problem if a project whose primary intent is to preserve data gets advice from a technician who thinks its intent is to preserve languages. Different technological choices are required in each case, emanating from distinct values, as we have seen.

In sum, while it's probably a good idea for documentary and descriptive linguists to get a basic education in the technology that their work will rely on, their real concerns shouldn't be technological but are more fundamental in nature: What do they value in a documentation project? The technician can help them answer questions about XML, Unicode, or workflows. But only linguists can figure out why they're doing all this work in the first place.

Emended references

Austin, Peter. 2006. Language documentation and your data. In J. Gippert, N. Himmelmann, and U. Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter. 87–112. (=Austin (in press))

E-MELD. 2005. Digitization of Audio Files. E-MELD School of Best Practice. <http://emeld.org/school/classroom/audio/>. (replaces E-MELD (2005a))

E-MELD. 2006. Digitization of Video Files. E-MELD School of Best Practice. <http://emeld.org/school/classroom/video/>. (replaces E-MELD (2005b))

REMOVE: E-MELD. 2005c. E-MELD School of Best Practice: What are Best Practices? <http://emeld.org/school/what.html> (replaced with reference to E-MELD paper in this volume)

Simons, Gary, and Steven Bird. 2006. OLAC Process. <http://www.language-archives.org/OLAC/process-20060405.html> (=Simons and Bird 2003b)

Simons, Gary, and Steven Bird. 2008. OLAC Metadata. <http://www.language-archives.org/OLAC/metadata-20080531.html> (=Simons and Bird 2003b)

Additional references

Gippert, Jost, Nikolaus Himmelmann, and Ulrike Mosel (eds.). 2006. Essentials of language documentation. Berlin: Mouton de Gruyter.