

# Endangered languages documentation: from standardization to mobilization

David Nathan

All correspondence to:

David Nathan

Foreign Language Center, University of Tsukuba

Tennodai 1-1-1, Tsukuba-shi, Japan

## Abstract

Currently, the main arena for computer-based linguistic contribution towards endangered languages is in data encoding and standardization. This phase urgently needs to be complemented by a period of working out how to *deliver* computer-based language support to endangered language communities.

Established linguistic practice has neither sufficiently documented nor strengthened endangered languages; Himmelmann (1998) identified this problem and proposed a new discipline he called *documentary linguistics*. Although this emerging discipline has stimulated digital projects for data standardisation, encoding, and archiving, it lacks two vital aspects: a methodology that provides roles for community members, and new genres of dissemination. Without identifying new ways to mobilize the products of documentation, documentary linguistics will remain indistinguishable from its predecessors in its ability to support language communities.<sup>1</sup>

## 1. Introduction

In recent years there has been a convergence of interest between linguists working with endangered languages and those who use computers as a research tool. A ten-year period has

---

<sup>1</sup> Research for ShoeHorn software has been supported by the the University of Tsukuba Research Project Fund. Research and development of the Spoken Karaim CD has been supported by the ILCAA, Tokyo University of Foreign Studies, and Uppsala University, Sweden. Production of the Paakantyi CD was supported by the LAIP program, Aboriginal and Torres Strait Islander Commission, Australia.

seen language endangerment placed on the linguistic agenda, first as a problem of language diversity (e.g. Krauss, 1992; Nettle and Romaine, 2000), subsequently addressed in terms of human rights and resource distribution (e.g. Skuttnab-Kangas and Phillipson, 1994), software localisation and, recently, increasingly framed in terms of computer-based archiving, portability, access, and, most notably, data encoding and standardisation (e.g. Bird and Simons, 2003). Issues in the cognitive and educational aspects of electronic language resource delivery, such as interface design and software development beyond basic tools for entering or viewing data, have received scant interest. The voices of endangered language communities have not become amplified, and languages are disappearing as fast as ever.<sup>2</sup>

Himmelman's paper 'Documentary and descriptive linguistics' (1998) promoted a new discipline of documentary linguistics, distinguished from traditional linguistic description, in response to 'the recent surge of interest in endangered languages' (Himmelman, 1998:161). Documentary linguistics is aimed at creating records of the 'linguistic practices ... of a speech community', as opposed to description, which attempts to record a language as a 'system of abstract elements, constructions, and rules' (Himmelman, 1998:166). In addition, Himmelman pointed out weaknesses of current linguistic documentary practice, as well as the potential advantages of creating rich records of language behaviour untied to particular analytical presuppositions or even disciplines.

Linguists had already become alarmed about the state of digital records for endangered languages (where records exist at all): their disparate structures and storage formats, lack of documentation within files, and their often-fragile storage conditions—all leading to difficulties in locating, identifying, and preserving data. The increasingly discussed plight of endangered languages, enabling technologies such as increased networking, XML document technologies, multimedia-capable computing, together with the linguistic foundations provided by long-established resources such as TEI (Sperberg-McQueen and Burnard, 1999), ELRA, SIL, and

---

<sup>2</sup> I am grateful for Jeanie Bell for teaching me that support for endangered languages is only realized when community members are choosing again to speak their languages.

DC metadata, were catalysed by the new idea of a new documentary linguistics, leading to the establishment of several new archiving and data encoding projects.

These new projects have a range of emphases, from resource discovery (e.g. IMDI, Broeder *et al* 2001), to language or region-specific coverage (Ega Web Archive, Gibbon (nd)), to creating standard mark-up and documentation formats for various linguistic artefacts such as lexica, interlinear data and media annotation (Linguistic Data Consortium, Bird and Simons, 2003). The aim of this paper is not to describe the history and significant accomplishments of these projects, but rather to identify what needs to be done to mobilize them in the service of language communities, and the reader is referred to the relevant projects for information about their holdings, formats etc.

Five years have passed since the appearance of Himmelmann's paper, and it is difficult to say what concrete contributions—either current or projected—have been made to the state of endangered languages as a result of work done so far. Two crucial aspects of an effective documentary linguistics should be urgently recognized: a methodology that builds in the participation of the language community, and the creation of suitable genres of documentation products. These two ends of the documentation process—the community, and the tangible computer-based resources that they increasingly look to for motivation and support of their language activities—have been neglected at the expense of one part of the chain that links them; data encoding and standardisation.

## **2. Including Language Communities**

It is assumed in this paper that research in endangered languages bears responsibility to the relevant language communities, and that members of such communities can make valuable contributions to research.

How do the projects mentioned in Section 1—the “PEAS” projects (projects for encoding, archiving and standardisation)—factor in roles for community members? Computer-based projects are built upon explicit representations, so their designs can tell us whether the project intends data to be acquired from, mediated, monitored, or used by community members.

methodology for designing and implementing computer systems. By mapping this onto project descriptions, we can see that the community members are *not* clients or end-users of the PEAS projects, because they are not built into the project design.

Some projects explicitly define their ‘target communities’ to be ‘language professionals’ such as ‘the field linguist, the syntactician, the language teacher’ (Ide and Romary, 2001:141-2, Lewis *et al*, 2001: 152). While some project descriptions do mention community members as stakeholders of one kind or another, none seem to provide a formal framework for including the target language communities in consultation, data acquisition, or product design or delivery. A brief survey of 33 abstracts for the ‘Workshop on Resources and Tools in Field Linguistics’ at ELRA’s LREC (Language Resources and Evaluation) 2002 conference suggests that only 10% of papers mention community members as potential users of materials; and only one paper refers to them as active contributors.

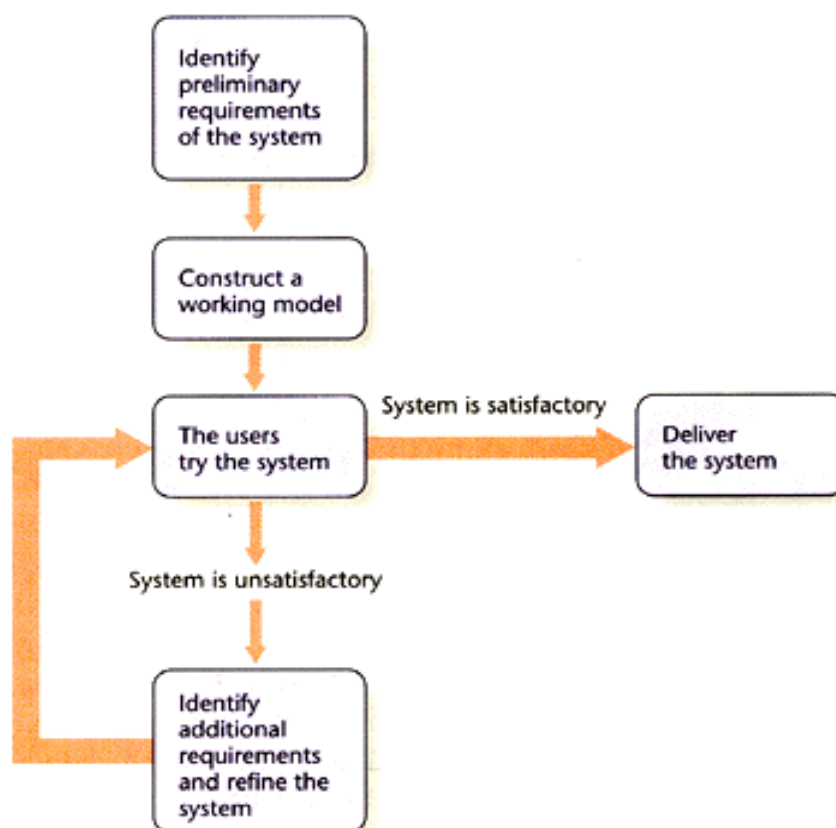


Figure 1 (adapted from Oz, 2002: 600) shows steps in the System Development Life Cycle, a standard IT

The E-MELD project has implemented *ontologies*, a kind of metadata that describes a system’s representations:

An ontology makes explicit what kinds of concepts exist (in this case linguistic concepts) in a domain; it defines what relations can exist between concepts; and it represents knowledge about the target domain.

(Lewis *et al*, 2001: 150). Such explicitness about fundamentals of representation is welcome.<sup>3</sup>

However, while the E-MELD ontology is claimed to be as broad as possible, its categories are limited to morphosyntactic and other linguistic terminology (Lewis *et al*, 2001: 154). This limitation is inappropriate to the description and preservation of endangered languages, which are so characterized by diversity and a shortage of existing linguistic knowledge. Speakers of endangered languages may possess interesting ontological categories— categories yet unrecorded, or categories not normally thought linguistically relevant but relevant in the context of language endangerment.

Standardized formats do provide (at least potentially) a way for any user, including language community members, to locate, browse, and use data. But such users should not be confined to ‘consumers’:

- members of endangered language communities should not just be consumers but should be potentially active participants in the production and evolution of records of their languages;
- the underlying technology, hypertext—first credited to Vannevar Bush (1945), but long existing in scholarly documents in the form of footnotes, references etc—was intended to provide symmetry between creators and users, writers and readers. Bush’s proposal allowed any user to create links between different kinds of information. Pursued through to the early 1990’s (Bolter, 1991; Barret, 1994), this symmetry all but vanished as expectations of hypertext were reshaped by the World Wide Web’s limited implementation of hypertext.

---

<sup>3</sup> Recently fashionable in networked IT, associated with “semantic web” research, ontologies have long been the province of an entire profession (IT systems analysts), and are commonly found in the form of data dictionaries upon which databases are built. Relational databases that store dictionaries (Nathan and Austin 1992), as well as DTDs for SGML files, can also be said to provide ontologies.

### 3. Standardising Language Data

Some standardisation technologies, such as alphabets and writing systems, follow from the digital (or quantized) nature of language itself. These technologies should be clearly distinguished from systems that are used to encode knowledge structures. The latter involve making creative, selective decisions about entities of interest and how the coding system expresses relationships among them, and results in codifying the ways in which we communicate messages with one another.

Reducing a language to a body of data may be as ‘absurd’ for language as it is for an organisation or other social system (Brown and Duguid, 2000:16). It recasts language behaviour as merely individual customisations of data, rather than social action. A cynical view is that focusing on metadata is a strategy for manufacturing new domains of intellectual property to enable research and networked publication while avoiding the potential intellectual property issues of working with authentic, media-based materials. This view can be rejected once language communities are built into project designs.

Preoccupation with data encoding (and especially when derived, selected, and codified—i.e. written—material is not distinguished from the linguistic events that preceded it) represents a reversion to older practice of describing languages for theories’ sake and traditional scholarly exclusivity where, for example, it is unquestioned that earlier records of endangered languages are written by ‘doctors, surveyors, clergymen, and others acting as amateur linguists’ (Romaine and Nettle, 2000: 26).

But while many linguists working with endangered languages have broadened their practice, emphasis in language computing remains with data, despite the potential of new multimedia literacies and expanding networks to expand their scope to include stakeholder communities. This is probably due to three reasons. Firstly, because of the robust history of language description in linguistics. Secondly, it is reinforced by a computing tenet known as *data independence*: data should be independent of any application that uses it. This tenet continues to be observed not because it is universal valid but because typical data structures do conform to a small number of applications—applications that are invisible to literate societies that do not see

writing as a technology and the written page as an interface.<sup>4</sup> And thirdly, most linguistic software does not exploit the potential of data independence anyway; in most cases the interface is a relatively transparent projection of the underlying data.

#### **4. Encoding Needs Physical Implementation**

Systems for encoding data cannot be effective alone: they must be complemented by software that makes the data tangible. The case of HTML and the World Wide Web provides a clear example.

The WWW evolved rapidly and unpredicted into new form of mass media and transcended its enabling technologies. HTML, one of these technologies—the encoding mark-up for hypertext that describes document properties and links—was technically weak, and indeed subsequently judged inadequate, and ‘redeemed’ by newer technologies such as XML and CSS.

HTML was designed before the advent of web browsing software, and would probably have remained obscure without the advent of that software. It was the implementation of a software interface for hypertext that could bridge a user, a page’s links, and the network infrastructure, that made the WWW wildly successful.

In fact, the browser’s underlying model was simply the page metaphor. Other than a simple method for expressing links between texts, browsers do not significantly extend the page metaphor. However, some intrinsic hypertext properties, from the users’ point of view, such as ‘forward’ and ‘back’ navigation, result from software functions that have no correlate in the encoding system. Forward and back navigation, which reflects a user-inscribed relationship between documents, has become an indispensable idiom of hypertext. Thus, navigational devices can be emergent results of software development. Below I will show how the inclusion of community members in the design, development and testing of one CD led to the construction of culturally-based data and navigational structures that could only be articulated post hoc.

---

<sup>4</sup> The written page is, interestingly, an interface that some Indigenous people find overt, and alienating.

The WWW's success consisted in combining HTML, network infrastructure, and browser software in order to transform networks of computing appliances into networks of documents. This was a profound transformation for users (see Nathan, 2000).

For a second example, consider the case of corpus linguistics, which has made great advances since the advent of computing, networks, and the creation of large volumes of on-line text. Arguably, its needs are less urgent than those of documentary linguistics. However, corpus linguistics is far more established in linguistic practice than documentary linguistics, and is associated with important encoding and archive initiatives. Yet Meyer (2002, 78) notes that 'one of the great challenges that corpus linguists must face is the development of software with user interfaces that permit users to browse spoken corpora' (see also Meyer, 2002: 86, 98).

Another example is MP3 sound encoding. Eriksen (2001: 107) describes MP3 as a 'file format for electronic transmission of music' and as 'a concrete example of the logic of the web'—a fragmented, personalized 'neo-liberal' world where 'each user puts together his or her own, personal totality out of fragments.' Rather than the MP3 encoding, it was primarily software—the now infamous Napster, and its descendants—that was responsible for the current revolution in music distribution. While Napster did utilize the MPEG standard (that long predated the MP3 file-sharing era), the real agent for change was the formation of a large, decentralized, file-sharing community based on the capabilities of the software.

To make encoding efforts meaningful, we also need to create suitable "players", just as Mosaic and Netscape were players of hypertext/HTML. Such players may need to be as complex as the data and domain require. The Spoken Karaim application (Csatò and Nathan, 1998; see Figure 2), for example, can be thought of as a player or an 'explorer' for rich, multifaceted, linguistic and cultural data, with some limited (at this stage) capability for user input—a pilot for a new era of linguistic multimedia exploration and acquisition systems ("MEAT").

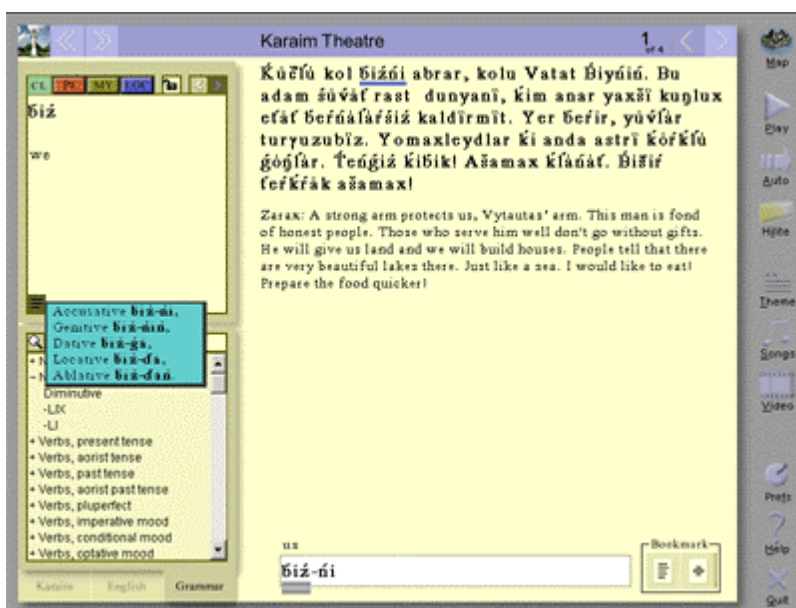


Figure 2 The Spoken Karaim CD-ROM, main screen snapshot. The application contains spoken voice, morphologically-analysed transcriptions linked to a dictionary, finderlist, concordance and grammar, and a morphological generator, all interlinked with relevant cultural texts, images, songs and video, together with user facilities such as bookmarking, forward/back, autoplay, and sound control. It also imports and exports interlinear data in a standard XML schema.

## 5. Sources of Encoding

Standardized encoding is a useful tool that most developers have learned to live without. It is normal for projects to find that data needs to be re-encoded; the ability to convert and restructure disparate forms of data is inextricably part of the skill set required to collaborate with others and to integrate text with other resources. Having standard encodings is not what makes projects doable or not doable, or successful or not. Standardized data formats alone, without appropriate software, will not enable those without suitable skills to create resources (Bird and Simons, 2003: 2).

The major hurdle for real projects is the acquisition and encoding of knowledge, rather than the particular way in which it is encoded. Even classifying and marking up relatively simple data is a major effort; for example, Bird, Jeffcoat and Hammond (2001), describe their dictionary data as 'lexical entries consisting only of a headword ... and a definition, which includes several components all lumped together: the English translation, example sentences, related forms etc.' and note the difficulty of separating out those components (Bird, Jeffcoat and

Hammond, 2001: 33). On the other hand, software developers typically work effectively with structured data from various sources with various implementations of structure (e.g. from relational databases to word processor tables to Shoebox mark-up to other typographic mark-up). For example, while the development of the *Spoken Karaim* CD application involved setting up many specialized data structures and exchange standards, and took more than 2 person-years to complete, the subsequent implementation of a standardized XML export and import module for interlinear data took only about 3 days.

The problem for endangered languages support is not that we do not have enough encoding; standard interlinear format, or even text “marked up” by punctuation, has far more encoded structure than current technologies provide for images or sounds. Rather, the problem is lack of bandwidth for knowledge flow (however expressed) between knowledge holders and end-users of that knowledge.

## 6. Interfaces

Traditional markup, such as SGML typically encodes these aspects of documents:

- i. visual, typographic properties, that are fundamentally procedural in the sense of being directions to typographers, printers etc, although not seen this way by modern, literate readers (especially those who read SGML documents!); or
- ii. logical properties—often signaled by perceptual text properties as in (i)
- iii. specifications for access or usage of texts, to be implemented by a reader, or a computer program

The potential for interactivity expressed in marked-up data is typically that which we associate with traditional paper texts. The rest—the truly interactive part, has to come from somewhere else.

An interface is the framing and handling of the data flow between the user and an information-bearing artefact, usually computer software. Interfaces are much more than conduits for data; they provide the whole theatre for managing the interactions between the user

and the software creators. Using an orchestra analogy, the interface for the performance comprises the stage, set, and arrangement of the performers. The audience know how to react because they understand this interface; if the soprano arrives at centre-stage they expect that she will start singing; if the conductor faces the audience or the curtain falls they begin clapping.

The interface is prior to encoding; it tells us what is to be encoded. Without a conception of the interface, we have a task that is too undefined. Currently, this is controversial, perhaps counterintuitive: we constantly work with linguistic and other data without thinking about the delivery interface, observing the concept of data independence (see Section 3). However, data independence is challenged by interactive multimedia where information, users, and tasks have become interdependent: '[a] user ... is ... performing as task, the very act of which implies information transfer' (McKnight, 1996:215). Data independence is a concept built upon assumptions that come from our deep affinity with text. Most of the data-oriented tasks we do are so embedded in our textual traditions that the interfaces (words, lines, pages etc.) are seen as background rather than foreground, or, rather, are not seen at all, rendered invisible by the tide of literacy that has so affected the cognition of many cultures (Ong, 1982). We do not see the page of paper, the horizontal lines or grids of symbols as being interfaces that preconstruct the possibilities for our "application independent" data.

A well-rounded interactive multimedia application, combining hypertext with time-based media, consists of a lot more than data and its metadata. Anderson (1994) compared multimedia to libraries: 'multimedia are the technical analog of the social construct of libraries'. Following this analogy, metadata is as important for effective multimedia as catalogues are for a working library. However, a library is much more than its catalogue and its books: users scan bookshelves, evaluate, group, compare, and summarize; in addition, there are professional staff who create a pleasant environment and perform functions such as acquisitions, reservations, helping users locate and copy items, borrowing, setting up reading spaces, and navigating the building—many of these activities and functions are implemented in some way in a multimedia application.

The Spoken Karaim CD (see above) has about 6000 lines of custom code (on top of the Macromedia “projector” runtime player) that implements how the data, kept scrupulously separate, is presented to the user.

What should good screen interfaces be like? Cooper (1995) argues that interfaces should not be determined by the underlying data, but by the functionalities needed by users and in terms of their understanding of the domain. In addition, good interfaces support the users’ effective performance of what the user feels is valid, not correctness according to some hidden schema—in other words, the interface should validate users and not insult them or make them feel inadequate (Cooper, 1995:13). The way to achieve this, according to Cooper, is to move the design strategy from models (which tend to recapitulate the underlying data), to metaphors (which are better, but limited by source of the metaphor), to idioms which use ‘gizmos’ whose behaviour must be learnt by users but, once learnt, best support working with the interface (a classic idiom is the car’s steering wheel, whose circular motion in a vertical plane is not a metaphor for a left-right vector but a “gizmo” that can very conveniently be manipulated. In the Karaim CD we use little blocks as gizmos that users can drag onto lexical entries as an interface to the morpho-phonological generator).

A good interface should fade seamlessly into a task for the user, ‘to help users feel like they are reaching right through the computer and directly manipulating the objects they are working with’ to the extent that ‘the interface isn’t even there’ (Mandel, 1997:60-1), or ‘invisible’ (Cooper, 1995:135). In other words, an interface should be in synch with the user’s mental model. Users should be free to focus on the work they are trying to perform, rather than translating tasks into the functions that the software provides (Mandel, 1997:61). Good interfaces will support learning; constructivist approaches to learning propose ‘that learning occurs best as a result of doing, creating, and building ... [especially through] the manipulation of real or virtual objects’ (Goldman, 258).

Interfaces should be customisable to suit users, and, if possible, by the users themselves. Annotations, for example, should be able to be added simply as further sound recordings, or be entered within categories nominated by users. The author’s Shoehorn software under

development will allow users to assign their own “channels” of annotation, which could include paralinguistic, social or other observations (cf. MPI’s Media Tagger, whose annotations allow only simple text-only values; Brugman and Wittenburg, 2001: 65). The aesthetic and emotional responses that are inextricably bound up with language materials, particularly those for endangered languages mean that responses to particular "data" vary from person to person, occasion to occasion, and it may be difficult or inappropriate to render them as conventional data or metadata. These are the cases where effective knowledge acquisition from, and delivery of the material to, community members is most crucial. We need, therefore, to provide linguistic and computer support for such channels.

Too often, however, software interfaces implement models that transparently reproduce the way their data and programs are structured. Software developed by the PEAS projects make no pretension of providing interfaces or access to endangered languages communities: most are bland, transparent projections of underlying categories and assumptions (see, for example, the IMDI BCBrowser, which, however, also offers some access via a clickable map; Broeder *et al*, 2001). They are typically oriented to the technical aspects of language so that interaction with them feels like it is above all about formal correctness, rather than, say, authenticity, relationship, or reminiscence. These are interfaces that are likely to intimidate naïve computer users, and shame the average Aboriginal language speaker or teaching assistant.

It is not surprising that the PEAS projects develop resources for dealing with the categories that primarily interest linguists, and indeed some of them, such as IMDI, are quite explicitly aimed at describing, rather than delivering, resources. Such observations are further evidence that language communities are not considered working partners; an unfortunate conclusion since for this author, consulting software designs with Aboriginal people has resulted in considerable improvement on several occasions. Indeed, Goodall (1996), found ‘the only way we could develop a program which would be comfortably used by Aboriginal people in the north west was to include them in the design process’.

Interactive multimedia applications are difficult to produce firstly because they lack established conventions (Brett, 1997) and the ‘mentors’ that typographers and designers

provided for desktop publishers, while nevertheless having to compete with high expectations created by computer games and television (Schlüsselberg and Haward, 1994: 95, 97). Secondly, a complex variety of inputs, participants, and skills are required. Working with graphics, databases, and programming/authoring languages are high-level and specialized skills that take time and devotion to master. It is an arduous task learning how to use an authoring application, choose and manipulate media in multiple formats, convert and link various kinds of text and structured data, as well as co-ordinating and negotiating between designers, linguists, communities and funding bodies, and dealing with practical matters such as runtime performance, user acceptance testing, and platform, version and hardware variables.

Many linguists can make some initial progress in some of these areas but typically “hit the wall” of a steep learning curve after a short period. On the other hand, “marking up” materials in HTML or XML, or with metadata, is a steadily incremental process that should not severely daunt most linguists. And since a fundamental aspect of standardisation is the decentralisation of the work involved in unifying formats (typically through adding and/or modifying mark-up), it therefore makes sense to delegate this work to linguists.

The contribution required from larger, highly skilled, specialized, or resourced projects and institutions is the development of infrastructure software such as language-tailored software and the provision of technical services to local language initiatives. The real reason that we find interactive multimedia production so rarely used in language documentation is not its complexity—after all, we are surrounded by various genres of media products that result from the collaboration among people with disparate skills, such as newspapers, movies, and recorded music—but a wider lack of appreciation of the importance of the language interface.

## **7. Documentation and Community Involvement**

Himmelmann (1998:166-7) urged the collection of ‘a comprehensive record’ that could potentially support a variety of research disciplines and theoretical approaches. He was not the first to do so. Goldman-Segall, for example, describes her use of video as a tool in providing a ‘thick description’ for ethnographic research (Goldman-Segall, 1994:258), noting Margaret

Mead's use of film 'as data' over 70 years ago, and Mead's later predictions about the merit of emerging technologies:

The emerging technologies of film, tape, video, and, we hope, the 360 degree camera, will make it possible to preserve materials ... long after the last isolated valley in the world is receiving images by satellite.<sup>5</sup> (Mead, 1975:9, quoted in Goldman-Segall, 1994)

Goldman-Segall describes her methodology that involves the camera 'changing hands', putting the 'researched' subjects themselves at the centre of a range of authentic communicative activities. The knowledge holders play a role in documentation by designing, creating, and interpreting their own video narratives in order to create a better, 'thicker' record (1994: 257, 269). This also results in a democratization of the research process, where knowledge holders can become real members of research communities and play a part in creating institutional memories for projects (Landow, 1994: 209).

Himmelmann, on behalf of linguistics, could have gone as far; instead, while recognising that spontaneous communicative events provide the most authentic record (1998: 176ff), he attenuates their importance by his 'pessimistic assessment' that participants will usually not consent to recording (1998: 187). This assessment is not supported by this author's experience in recording materials in Aboriginal communities. Furthermore, the crucial question to be asked is whether a truly worked-out discipline of documentary linguistics *would* grapple with the methodological and political questions of how its most authentic type of data *can* be collected, rather than shouldering community members with responsibility for its scarcity.

For a successful documentary linguistics, community involvement need not be limited to the two polarized possibilities of holding the camera or telling the researcher to turn it off. Community members can be willing and eager *users* of the *products* of documentation. The problem of how

communities can be actively involved in the design of a concrete documentation project ... in such a way that the community not only accept it but also shape it in essential aspects (Himmelmann, 1998:188)

---

<sup>5</sup> Today's Virtual Reality software (such as Quicktime VR) functions like a 360 degree camera.

can be addressed not solely within the project planning phase but also within project development—by ensuring that there *is* a visible, concrete development process—and by making sure that the community is involved so that they can understand, appreciate, and look forward to the concrete *outcomes* of that development (cf. Goodall, 1996).

A project must create its own life and history within the community. The process of building a multimedia project in collaboration with a community can raise a host of relevant linguistic and sociolinguistic issues. For example, when working on the Paakantyi CD (Hercus and Nathan, 2002), issues such as the relationship between spelling systems and contested land claims were raised; when consulting about the Kamilaroi/Gamilaraay Web dictionary (Austin and Nathan, 1996) there were interesting negotiations about the inclusion (or otherwise) of “rude words”, and about the boundaries of Kamilaroi country on the web map.

In addition, documentary linguistics—together with members of endangered language communities, teachers and others—needs more than data and better ways to encode, transmit and process it. It needs an evolution of interfaces and software to deliver the richness and diversity of collected materials and support a diversity of users. Documentary linguistics needs to clarify what (possibly new) genres of publication reflect its particular emphases.

Figure 3 below sums up the range of contributions, from elicitation to “reclamation”, that interactive multimedia (eg on CD-ROM) can make to an endangered language and its community. It draws on the parallel polysemies of voice and retrieval:

*voice:*           utterance   /   influence  
*retrieval:*       fetching     /   recovery of ownership

<b>Retrieval</b> \ <b>Voice</b>	<i>Utterance</i>	<i>Influence</i>
<i>Search process</i> create CD use CD	Elicitation Learning	Participation Validation
<i>Return to owner</i>	Motivation	Reclamation

Figure 3 The language CD-ROM and “voice retrieval”

## 8. Emergent features in community-based projects

Harnessing community input is not only an instrumental strategy for raising acceptance: it can provide a means to improve and innovate. In this section I outline cases in the development of the Paakantyi CD (Hercus and Nathan, 2002) where a fluid process involving the linguist, the author (as multimedia developer), and several Paakantyi community members resulted in positive outcomes that emerged as a result of negotiations about data and methodology.

It was an important part of this dynamic that we prepared and delivered concrete multimedia samples at every stage of the project, each version representing the accruing state of the CD. This was valuable because (i) it is much easier for people to demonstrate their reactions to concrete products than to give opinions about abstractions; (ii) it demonstrated our commitment to the project; and (iii) it helped create a kind of biography for the CD that would give it more motivation for use once delivered. We also workshopped our main participants in some of the techniques by which we recorded, digitized, edited, and linked the sounds they contributed.

The first emergent properties of the CD were the representational and navigational structures that resulted from working with members of the Paakantyi community on the language and graphic content, and the design of the CD. This is summarized in Figure 4. The top layer lists the “old time” speakers whose texts and songs feature on the CD. The lower layer lists the contemporary contributors of art (the two leftmost boxes) and linguistic (the two rightmost boxes, with Badger Bates in both categories) material. The arrows represent linguistic input; the other links represent artwork supplied. Vertical alignment shows ancestry: guided by Badger Bates, it turned out that the key linguistic assets were illustrated and accessed by use of the artwork of their living descendants (both Dutton and the Bates’ are related to Jack ‘Gunsmoke’ Dutton). This not only provided us with explicit data about Paakantyi genealogy, but also, we believe, contributed to the keen acceptance of the CD-ROM by the community.

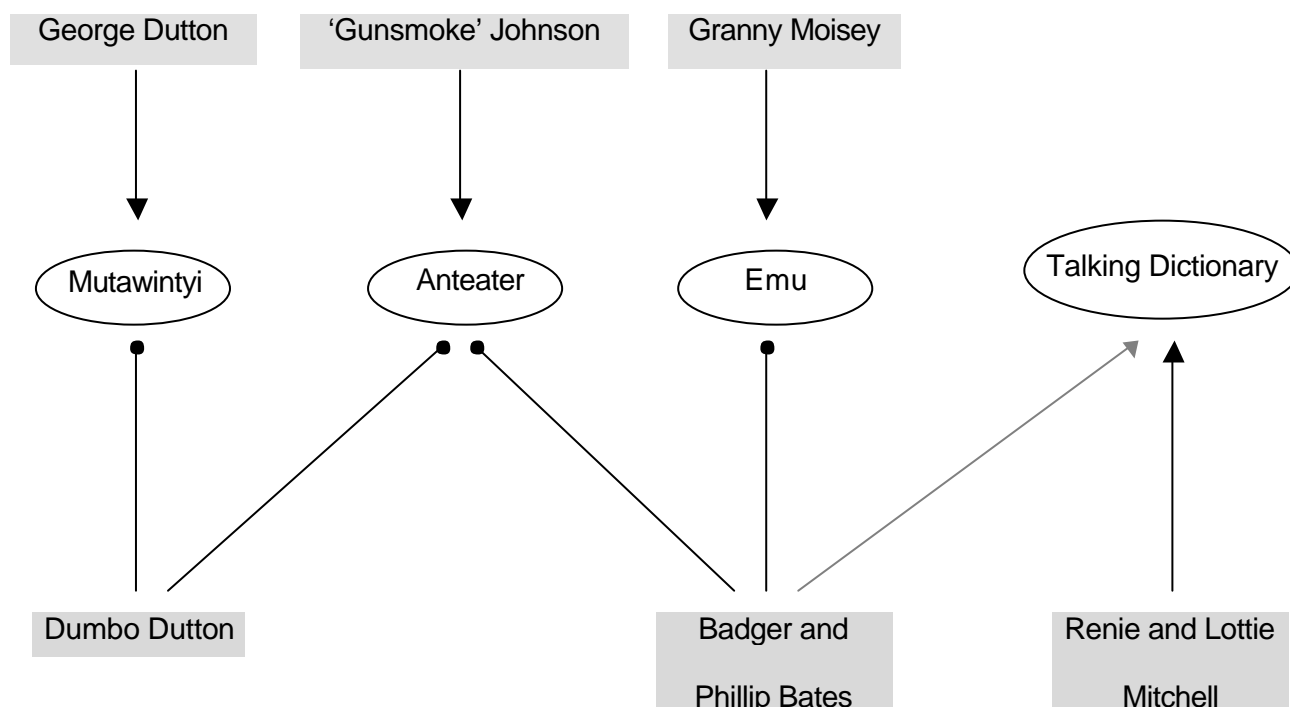


Figure 4 Language, art and genealogy in the Paakantyi CD-ROM

The second emergent aspect of the Paakantyi design was the form of the lexicon. People had asked at the outset if we could produce a “talking dictionary”, and this became a centrepiece of the project. In eliciting and recording the spoken data, I preferred initially not to have people add English translations to words and phrases (or at least I intended to edit them out so that the lexical entries were only uttered in Paakantyi). However, speaker after speaker preferred to add English glosses (for example, they would say: ‘wiimpatya: blackfella’). In constructing the dictionaries, and in informal user testing with Paakantyi students, I found that this method provided access for audiences that it would otherwise have missed: preliterate children (who could not read the written glosses), vision-impaired, and those inevitably gathered around the computers and standing too far away to read.

A third emergent aspect was that when significant sound assets are available, community members tend to treat the textual/linguistic content as a kind of indexing or metadata. In two projects, initial presuppositions about the roles of data have been reversed. In the Paakantyi CD and the Warrungu interactive concordance (Tsunoda and Nathan, 2002), the sound assets for word and phrase pronunciations—initially intended as an resource additional to the dictionary or text—became the primary content for Aboriginal users, who manipulated the text or

dictionary representations as indexes or paths for accessing the sounds. In that way, entire areas of the application had to be reinterpreted as simply providing pathways to access the sound resources. Thus, from the point of view of the Aboriginal users, what was originally data, became a system of metadata that provided access to resources that otherwise are extremely difficult to access. In addition, this helps explain the enthusiastic embracement and effective usage of the Paakantyi interface, despite its significant use of written text (cf. Goodall, 1996).

## 9. What Communities Want

What software developments have made an impact on endangered language communities? In several Australian Aboriginal communities, language software, even quite simple in some cases, has been well received (Auld, 2001; Nathan, 2000, 1999). The publication of the Kamilaroi/Gamilaraay Web Dictionary (Austin and Nathan, 1996), in close consultation with a range of community members, assured that by 2001, the website was regarded in the Kamilaroi community as a stable part of their language infrastructure, where members could readily turn to learn the language (Robert Amery, pc.).

What do endangered language communities want or expect from IT? While there is not enough evidence to answer this question definitively, it seems that people want at least some of the following:

- processes and products that respect their “ownership” of a language
- to not have to pay to “buy back” what they see as parts of their own language
- products that can be used publicly (not likely to cause shame)
- products that do not divulge inappropriate information
- the facility to input their own data or commentary (Michael Christie pc, Goodall 1996)
- useful, everyday expressions, especially with sound
- expressions showing how words are used to formulate real messages
- almost anything with sound

- products that are easy to use. Goodall (1996) interprets this as a matter of using non-text based navigation due to alienation from literacy and often failing eyesight; however, although this seems altogether reasonable, in practice this particular constraint has *not* been suggested by testing in community contexts. For example, the Paakantyi CD uses a contemporary, text-driven navigation system which has been extremely well accepted and found easy to use
- products for particular purposes, for example spell checking (for Arnernte, Manning and Parton, 2001: 167; also for Yolngu, Michael Christie pc).

## 10. Conclusion

Developments in data encoding standards have already demonstrated enormous benefits. But they will only pay dividends for endangered language communities when the energy of those involved in endangered languages turns to providing similar innovation in areas that are more intractable than data handling: good software for connecting real-time media to other resources, with interfaces that are pedagogically effective, not overdependent on literacy, and tailored for flexible use by and within communities. Communities want to use our tools and technologies to help counter language endangerment.

We cannot afford to repeat a “productivity paradox” for endangered languages. This term describes the widely experienced phenomenon of massive business investment in information technology, with little return in output or profit for up to 30 years later, if at all. In Australia, we know that despite over 30 years of enlightened research and documentation of Indigenous languages, they are disappearing faster than ever.

We also need to ensure that our interest in endangered languages is much more than a gritty message for consumption by the mainstream media, that languages are not celebrated globally for their formal genius while being ignored or denigrated on their home ground (Thieberger, 2002: 311; Hornberger and King, 2001: 183).

Is it necessary for linguistics to be accountable to the fate of languages and to be in the service of their speakers? In the case of endangered languages, yes. Firstly, because we have

appropriated the term ‘endangered languages’ into the description of so many projects. Secondly, because (the loss of) biological diversity has provided a pervasive metaphor for endangered languages under which it has been possible to merely describe the phenomena. However, this is probably the wrong metaphor, not least because it is not generally held by members of the respective communities (Tsunoda pc), but also because a medical metaphor is more appropriate: the patient is dying. It is unimaginable that medical science could describe illnesses without simultaneously trying to counter their effects (cf. Krauss, 1992); moreover, we can surely better understand the phenomena through practical efforts and observations of progress.

## References

- Anderson, G.** ‘Dimensions, Context, and Freedom: The Library in the Social Creation of Knowledge’, in Barrett (ed), pp 107-124.
- Aristar, A. and Aristar-Dry, H.** (2001). ‘The E-MELD Project’, in Bird *et al* (eds) (2001), pp 11-16.
- Auld, G. (2001).** ‘What Can We Say about 112,000 Taps on a Ndjebbana Touch Screen?’, *The Australian Journal of Indigenous Education*, Vol 30 Number 1 (2002).
- Austin, P. and Nathan, D.** (1996). *Kamilaroi/Gamilaraay Web Dictionary*  
<http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/GAMDICT/GAMDICT.HTM>
- Barrett, E.** (ed.). (1992). *Sociomedia: Multimedia, Hypermedia, and the Social Construction of Knowledge*. Cambridge, MA: MIT Press.
- Bird, S. and Simons, G.** (2003) (to appear). ‘Seven Dimensions of Portability for Language Documentation and Description’, *Language* 79.
- Bird, S., Buneman, P. and Liberman, M.** (eds) (2001). *Proceedings of the IRCS Workshop on Linguistic Databases*, IRCS University of Pennsylvania, December (2001).
- Bird, Sonya, Jeffcoat, M. and Hammond, M.** (2001). ‘Electronic Dictionaries for Languages of the Southwest’. In Bird, S., Buneman, P. and Liberman, M. (eds) (2001), pp 32-37.

- Bolter, J.** (1991). *Writing Space: The Computer, Hypertext, and the History of Writing*. Erlbaum: Hillsdale NJ.
- Brett, P.** (1997). 'Multimedia applications for language learning - what are they and how effective are they'. In Dangerfield, M. *et al East to West*. pp 171-180.
- Broeder, D., Ofenga, F., Willems, D. and Wittenberg, P.** 'The IMDI Metadata set, its Tools and accessible Linguistic databases'. In Bird, S., Buneman, P. and Liberman, M. (eds) (2001), pp 48-55.
- Brown, J.S. and Duguid, P.** (2000). *The Social Life of Information*. Boston MA: Harvard Business School Press.
- Brugman, H. and Wittenburg, P.** (2001). 'The application of annotation models for the construction of databases and tools: overview and analysis of MPI work since 1994'. In Bird, S. Buneman, P and Liberman, M. (eds) (2001), pp 65-73.
- Bush, V.** (1945). 'As we may think'. In *Atlantic Monthly*, 176 (July), 101-108.
- Cooper, A.** (1995). *About Face: the Essentials of User Interface Design*. Foster City CA: IDG.
- Csató, É.** (2001). 'Karaim'. In Thomas Stolz (ed.) *Minor languages of Europe* [Bochum-Essener Beiträge zur Sprachwandelforschung 30] Bochum: Brockmeyer, 1-24.
- Csató, É. and Nathan, D.** (1998). *Spoken Karaim* (L version). Multimedia CD-ROM. Tokyo University of Foreign Studies.
- Eriksen, T.** (2001). *Tyranny of the Moment: Fast and Slow Time in the Information Age*. London: Pluto Press.
- Garay, K. and Walker, D.** (2000). 'Bringing computing into the Middle Ages: the making of Sybils!, a Multimedia CD-ROM'. In *Literary and Linguistic Computing* vol 15, no 2 (2000), pp 199-218.
- Gibbon, D.** (nd). EGA WEB ARCHIVE <http://coral.lili.uni-bielefeld.de/LangDoc/EGA/>

- Goldman-Segall, R.** (1994). 'Collaborative Virtual Communities: Using Learning Constellations, A Multimedia Ethnographic Research Tool'. In E. Barrett (ed.) (1992), pp 257-296.
- Good, J. and Sprouse R.** (2001). 'Creating a database and query-tools for the TELL multi-speaker linguistic corpus'. In Bird, S. Buneman, P and Liberman, M. (eds) (2001), pp 82-91.
- Goodall, H. and Flick, K.** (1996). 'Angledool stories'. Paper delivered at AUC Academic Conference 'From Virtual to Reality', The University of Queensland.  
<http://auc.uow.edu.au/conf/Conf96/Papers/Goodall.html>
- Hercus, L. and Nathan, D.** (2002). *Paakantyi*. Multimedia CD-ROM. Canberra: ATSIIC.
- Himmelmann, N.** (1998). 'Documentary and Descriptive Linguistics'. In *Linguistics* 36 (1998), pp 161-95.
- Hornberger, N. and King, K.** (2002). 'Reversing Quechua Language Shift in South America'. In D. Bradley and M. Bradley (eds) *Language Endangerment and Language Maintenance*. pp 166-194. London: Curzon.
- Ide, N. and Romary, L.** (2001). 'Standards for Language Resources'. In Bird, S. Buneman, P and Liberman, M. (eds) (2001), pp 141-149
- Krauss, M.** (1992). 'The World's Languages in Crisis' *Language*, (1992).
- Landow, G.** (1994). 'Bootstrapping Hypertext: Student-created Documents, Intermedia, and the Social Construction of Knowledge'. In E. Barrett (ed.) (1992), pp 195-217.
- Lewis, W., Farrar, S. and Langendoen, T.** (2001). 'Building a Knowledge Base of Morphosyntactic Terminology'. In Bird, S. Buneman, P and Liberman, M. (eds) (2001), pp 150-156.
- Mandel, T.** (1997). *The Elements of User Interface Design*. New York: Wiley.
- McKnight, C.** (1996). 'What makes a Good Hypertext?'. In Herre van Oostendorp and Sjaak de Mul (Eds) *Cognitive Aspects of Electronic Text Processing*. (Advances in Discourse Processes Vol LV111). Norwood, NJ: Ablex.

- Mead, M.** (1975). 'Visual Anthropology in a Discipline of Words'. In *Principles of Visual Anthropology*, Paul Hockings (ed.). The Hague: Mouton.
- Meyer, C. F.** (2002). *English Corpus Linguistics: An introduction*. Cambridge: Cambridge University Press.
- Nathan, D.** (2000). 'The Spoken Karaim CD: Sound, Text, Lexicon, and 'Active Morphology' for Language learning Multimedia'. In Göksel, Asli and Celia Kerslake (eds) *Studies on Turkish and Turkic Languages*. (2000). Wiesbaden: Harrassowitz. pp 405-413.
- Nathan, D.** (2000). 'Plugging in Indigenous knowledge: Connections and innovations'. In *Australian Aboriginal Studies* (2000) (1&2): 39-47.
- Nathan, D.** (1999). 'Language support with IT: not a high-wire act'. Paper presented at *Learning IT Together*, Brisbane, April 1999.
- Nathan, D. and Austin, P.** (1992). 'Finderlists, computer-generated, for bilingual dictionaries'. In *International Journal of Lexicography* 5:1, 32-65.
- Ong, W.** (1982). *Orality and literacy: the technologizing of the word*. London: Methuen.
- Oz, Effy.** (2002). *Management Information Systems*. (3rd Ed.). Boston MA: Thomson.
- Romaine, S. and Nettle, D.** (2000). *Vanishing Voices: The Extinction of the World's Languages*. Oxford: OUP.
- Schlusselberg, E. and Haward, V. J.** (1994). 'Multimedia: Informational Alchemy or Conceptual Typography?'. In E. Barrett (ed.) (2001), pp 95-106.
- Skutnabb-Kangas, T. and Robert Phillipson (eds.), with Mart Rannut.** (1994). *Linguistic Human Rights. Overcoming Linguistic Discrimination*. Contributions to the Sociology of Language 67. Berlin & New York: Mouton de Gruyter.
- Sperberg-McQueen, C. and Burnard, L.** (eds) (1999). *Guidelines for Electronic Text Encoding and Interchange*. TEI P3 Revised reprint: Oxford (1999).
- Thieberger, N.** (2002). 'Extinction in Whose Terms?', in D. Bradley and M. Bradley (eds) *Language Endangerment and Language Maintenance*. (Ch 18) (2002). London: Curzon.

**Tsunoda, T. and Nathan, D.** (2002). *Warrungu Stories & Interactive Concordance: Stories from Alf Palmer*. <http://www.dnathan.com/language/warrungu/>

**Warschauer, M.** (1998). Technology and indigenous language revitalization: Analyzing the experience of Hawai'i. *Canadian Modern Language Review*, 55(1), 140-161.