LANGUAGE DOCUMENTATION & CONSERVATION
Vol. 9 (2015), pp. 1–9
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24629

On Training in Language Documentation and Capacity Building in Papua New Guinea: A Response to Bird et al.

Joseph D. Brooks
University of California at Santa Barbara

In a recent article, Bird et al. (2013) discuss a workshop held at the University of Goroka in Papua New Guinea (PNG) in 2012. The workshop was intended to offer a new methodological framework for language documentation and capacity building that streamlines the documentation process and accelerates the global effort to document endangered languages through machine translation and automated glossing technology developed by computer scientists. As a volunteer staff member at the workshop, in this response to Bird et al. I suggest that it did not in the end provide us with a model that should be replicated in the future. I explain how its failure to uphold fundamental commitments from a documentary linguistic and humanistic perspective can help inform future workshops and large-scale documentary efforts in PNG. Instead of experimenting with technological shortcuts that aim to reduce the role of linguists in language documentation and that construct participants as sources of data, we should implement training workshops geared toward the interests and skills of local participants who are interested in documenting their languages, and focus on building meaningful partnerships with academic institutions in PNG.

1. INTRODUCTION. *Language Documentation & Conservation* published an article called "The International Workshop on Language Preservation: An Experiment in Text Collection and Language Technology" (Bird et al. 2013) in which the authors discuss a workshop that took place from 21 May to 1 June 2012 at the University of Goroka in Eastern Highlands Province in Papua New Guinea (PNG). This workshop, the first of its kind in PNG, attempted to address the challenge of documenting languages in the world's most linguistically diverse country. The Division of Language and Literature at the University of Goroka was an ideal institutional site for such a workshop due to the experience of its faculty members in certain aspects of language documentation, the interest of its students in receiving training to be able to document their languages, and the personal and professional connections it has with villagers throughout the Eastern Highlands and students and faculty at other universities in PNG. The workshop was unique in that it pursued a new methodological framework for language documentation and capacity building in PNG. This new framework sought to accelerate the documentation of endangered languages by relying on the technological promises of machine translation.

Given the formidable if not insurmountable challenge faced by linguists to document and describe endangered languages in PNG and worldwide, there is no doubt that we need to think creatively and interdisciplinarily so as not to overlook any methods which might

E-ISSN 1934-5275

meet that challenge. The Goroka workshop represents one such undertaking, but in the end, its drawbacks proved to outweigh its anticipated value. As a volunteer member of the workshop staff,[1] I would like to offer an alternative perspective on the workshop, in order to explain why I found this model not one to be replicated in the future. I begin by explaining how the computational research goals and related methodological approach of the workshop, as I understood them, were incompatible with the goals of documentary linguistics. I then draw two important lessons for future language documentation training workshops. First, I suggest that a workshop that provides training in the basic technological and other related methodological skills in language documentation and at investing in the interests and skills of participants would be more likely to succeed in laying the groundwork for the large-scale documentation of endangered languages in PNG. Second, I discuss what we need to keep in mind in order to uphold our commitment to the human element in language documentation training.

2. WHOSE GOALS? SOME PROBLEMS WITH SHORTCUTS IN LANGUAGE DOCUMENTATION AND TRAINING. The Goroka workshop was inspired by the idea that computer science technology in the form of machine translation can advance and even revolutionize the practice of documentary linguistics. Documentary linguistics as practiced today is concerned with the creation of an annotated, multi-purpose, and lasting record of a language based primarily upon digital recordings of communicative events (Himmelmann 1998, 2006; Woodbury 2011, inter alia). But, as computer scientists at the workshop pointed out to linguists like me who participated, one problem with current language documentation methodology is that it is very slow-paced. This is the basic problem the workshop aimed to solve:

> PNG is renowned for the great number and diversity of its languages. Many of these languages are moribund, and there is a critical need to document them before they fall out of use. The scale of this task far exceeds the resources available to documentary linguists. (Bird et al. 2013:155).

The workshop organizers were therefore seeking an interdisciplinary approach which would contribute to the documentation of the languages of PNG through "methodologies that could lead to the partial automation of the documentation process" (157) while contributing to machine translation research by using data from the small, endangered languages of the Papua New Guinean participants. As I understand it, the overriding goal, and the one that primarily guided the workshop methodology, was to develop and train computational models for machine translation. It was hoped that accurate translations between languages might result and so speed up the documentation process.

Machine translation research, it is claimed, can "scale up the pace of the documentary work" (166) by assisting but also reducing the role of the linguist in linguistic analysis. It was hoped that computers might be able to "largely automate word glossing" (159). Where morphological analysis is concerned, it was hoped that machine translation technology might even allow for the "unsupervised morphological analysis" of textual data (159). The

[1] A brief explanation of the roles of the members of the workshop staff and their respective areas of expertise is in order. Among the authors of the Bird et al. (2013) paper, the computer scientists and/or computational linguists include Bird, Chiang, Hanke, Vaswani, and Wan. Bird and Chiang planned the workshop and spearheaded it. The non-resident staff includes those four as well as two documentary linguists—Berez and myself. Eby, Frowein, and Shelby represent the resident members (i.e., expatriate faculty members at the University of Goroka) of the workshop staff who assisted with various aspects of the workshop, gave presentations, and also went out of their way to share with the rest of us their knowledge about cultural life in PNG.

workshop focused exclusively on collecting written texts instead of texts transcribed from audio or audiovisual recordings, "in order to streamline progress towards our goal" (158). The slow pace of transcribing and annotating audio-recorded texts was not compatible with the computer science goals of the workshop and would have produced too small a corpus to be of any use for machine translation. The focus on written texts thus represents the primary and clearest way in which the computational goals of the workshop all but completely supplanted the professed documentation goals (see Appendix). Participants composed and translated as many written texts as they could produce within the approximately 40 hours they spent at the workshop. To maximize textual intertranslatability, they typed up their texts in FLEx, doing their best to gloss at the level of the word. Feeding the typed versions of all translated texts into machine translation software was expected to yield useful results for language documentation by allowing for the automatic glossing of words and morphemes in participants' written texts. Bird et al. do not report on the results of the machine translation efforts, but I would be very surprised if the outcome was successful given the glossing problems experienced by the participants (these are discussed below). Little of what was done in the process of working toward the machine translation goal contributed to documenting or preserving the participants' languages or to laying the groundwork for building capacity for large-scale documentation endeavors in the future. We therefore need to take a closer look at what aspects of the workshop's approach were problematic, in order to see what lessons we can learn for the future.

One problem that Bird et al. identify in their article is the way in which the focus on written texts deformed the documentation of the participants' languages:

> Sentences averaged 9.2 words in length...which is short compared with the spoken languages. The syntax and discourse of Papuan languages is characterized by a phenomenon known as *clause-chaining*, in which many dependent clauses are chained together and anchored by a final clause (Foley 1986). Although we have not studied this systematically, our choice of the written medium appears to have biased several writers to adopt the discourse structure of simple English texts (Bird et al. 2013:164-5).

But based on what I was able to infer from interactions with workshop participants as they wrote their texts, the data collected at the workshop were even more problematic. This is because the deformation of the syntactic and discourse structures was compounded by the lack of transcription uniformity with respect to phonology, morphology, and the lexicon as well—an inevitable result when speakers with variable literacy skills and no standard literary practice write texts in their languages.

In his article on language documentation and capacity building in the New Guinea context, Foley (2004) makes the point that when native speakers with at least basic literacy skills (i.e., in Tok Pisin, the main lingua franca in PNG) write their language, the way in which they represent segmental distinctions and where they choose to place word boundaries can give insight into linguistic structure and thereby enrich the documentary corpus. But there is a flip side to this. When multiple speakers of a language with variable literacy skills write their language, a diverse range of conventions will result. How speakers choose to represent allophonic variation, whether they will represent clitics as separate words or lump them together with their domain constituents, and the potential for speakers to transcribe bound pronominal and other affixes as separate words are just a few of the many ways in which the transcription practices of native speakers might differ from the transcription practices of linguists. Consider the following example from Chini, a language spoken in Madang Province on which I am currently conducting research:

(1) *Nu mɨyi nɨñjiyẽntpmicha?*
nu mɨyi nɨ=ñji-yi-tpmi-cha
2SG what INS=MID-chew.betel nut-IPFV.IRR-IPFV.IRR:Q
'What are you chewing betel nut with?'

When I asked three speakers with literacy skills (i.e., in Tok Pisin) to write the above Chini sentence, each transcribed the morphology of the verb in a different way and represented the phonetic form in quite different ways as well. One speaker wrote <giycmsa>, omitting the instrumental clitic *nɨ=*. Another wrote <n jinpisa>, clearly representing the proclitic as separate from the rest of the verb form. The third speaker wrote <njiypi ca>, leaving it unclear whether the initial <n> represents the proclitic or is instead part of the middle voice prefix which also begins with a nasal. Unlike the other two speakers, he placed a word boundary between *-cha*, a content question suffix used with imperfective irrealis verb forms, and the rest of the verb. The variable orthographic conventions these three native speakers used provide some insight into how native speaker transcriptions might differ from a linguist's transcription. We would not expect texts written by native speakers with variable literacy skills and no linguistic training to result in a uniformly transcribed corpus. Problems like these in the approach to collecting data negatively affected the Goroka workshop, in great part because the workshop was designed and implemented by computer scientists and not by linguists.

Another problematic issue identified by Bird et al. was the way participants glossed words in their written texts:

> Contrary to expectation, phrasal translation turned out not to be the most difficult of the three basic tasks. For each assigned writing task, the stories tended to be short (an average of 9.5 sentences, or 88 words including punctuation). Glossing also turned out to be challenging for a different reason: some of the participants found the task unnatural, and a few even simply wrote their English translations, in English word order, into the gloss line. About 30% of words were left unglossed (Bird et al. 2013:164).

Of course, this is not surprising, since morphologically-encoded distinctions do not always lend themselves to clear word-level glosses in the target language. If I were given the simple English sentence *I'd see* and were to try to gloss the single phonological word *I'd* into French, I would immediately run into trouble since the conditional meaning encoded by the English auxiliary clitic *='d* is encoded by a suffix on the French verb *verrais* 'would see'. And when we consider how different many of the morphologically-encoded categories in languages of PNG are from those in English, it becomes even more apparent why native speakers would run into glossing difficulties.

That is not to say that there is no place for computational methods in the transcription and annotation process. Some recent scholarship, for instance, suggests that computational models might be able to assist linguists in interlinear glossing. In their study, Palmer et al. (2010) describe how computational models might help an experienced field linguist to gloss the unglossed portions of a corpus (i.e., one that has been transcribed and partially glossed by a linguist) and to improve the consistency of the glossing. They point out that the success of the model for such purposes would depend on its close interaction with the linguist, and that an important factor would be whether or not the linguist has expert knowledge of the language. However, the approach of the Goroka workshop was very different from what Palmer et al. propose, because it did not include any real role for the documentary linguists during the text collection process or during the computational process. The latter computationally-driven approach to language documentation is in my view unlikely to produce results of much use to anyone, since what the linguist brings to the process is much too valuable to be significantly reduced by machine translation technology.

Moreover, even the analysis produced by a linguist can be insufficient if it does not incorporate native speaker intuitions. This gets at a fundamental problem of the Goroka workshop, since what linguists and speakers bring to the documentation process were both undervalued. As Hale (1968), suggested years ago based on his fieldwork experience in North America and Australia, contrary to the view of those who consider native speakers to be naïve with respect to their own language, there is a "dependency relationship between effective research and a native speaker's control of the data" (1968: 387):

> Linguistic information about the structures of American Indian and Australian Aboriginal languages...is, with a few exceptions, limited to spheres which are more or less readily accessible to the perceptive non-native speaker—that is, to areas which can be studied with minimal appeal to the native speaker's intuitions. Thus, we have a great deal of excellent information on phonology and morphology, but relatively little on syntax. And the extent to which success has been enjoyed has depended a great deal on the efficiency of the partnership between a linguist who was not a native speaker and an informant who was (Hale 1968:386).

Hale's call for native speakers to have a more formative role in linguistic analysis is readily applicable to language documentation training and capacity building. If the participants at the Goroka workshop had been given a more influential role in the production of their data and had been given access to audio recording devices, a record of greater depth and usefulness both to the participants and to linguists would have almost certainly resulted.

This points to another human complexity of documentary linguistic research that the workshop failed to take into account, and that concerns people's motivations to engage in it. One issue which merits greater reflection is how we should go about assessing what participants' interests and skills actually are, and what it is that they hope to gain from the documentation process and from workshops like the one held in Goroka. Although one can only speculate as to why each individual participant attended the workshop, a reasonable assumption is that many came in response to recruitment materials that were distributed at the University of Goroka, which were similar to the one distributed by the workshop organizers to the rest of the staff (see Appendix):

> This workshop...will provide hands-on training in digital technologies for language documentation, and [the workshop] will create archival documentation...Topics to be covered include: audio and video recording techniques, documenting informed consent, metadata and record keeping, the protocol of Basic Oral Language Documentation, and key software tools for transcription, interlinear text preparation, and dictionary making.

Most of the workshop participants were villagers from communities in the Eastern Highlands region. A minority had more than basic Western-style education, and several were illiterate. The approach of the workshop, with its focus on the potential of machine translation to streamline the documentation process, did little to serve those participants who were illiterate, did not speak English, or were local university students in Goroka who had hoped to gain skills in language documentation but spoke only English and Tok Pisin. Future language documentation workshops organized by outside academics would benefit from considering how the local academic community is to be served during the design and implementation stages of the workshop, since the interests of that particular population are most likely to overlap with those of the researchers.

Thus far, the documentation and description of languages in Melanesia is due to the hard work of dedicated outside field linguists and missionaries. Many of these people have worked in places where community-driven language work is arguably neither feasible nor desired by community members (cf. Dobrin 2008). For most speaker communities in PNG, this state of affairs is the norm, and without the work of outside linguists the languages of these communities will likely never be documented or described. However, there are students representing communities throughout PNG and the South Pacific at institutions such as the University of Goroka, the University of Papua New Guinea in Port Moresby, and Divine Word University in Madang. Among them are students eager to document their languages but who lack basic skills and training in documentary linguistic methods. We know from models like InField/CoLang in the United States and 3L in Europe that though training workshops may not make immediate experts out of novice participants, the training they receive there opens a door for them to become more deeply involved in documentary and descriptive linguistics. This is where future training efforts and capacity building in PNG ought to be focused. Language documentation training workshops—organized for these students by trained and experienced documentary linguists who speak Tok Pisin and take into account Melanesian social practices and values—have real, but up until now untapped, potential to result in a large-scale documentation effort in PNG.

3. CONCLUSION. It can be hoped that the organizers of future language documentation workshops will proceed cautiously if they consider incorporating the methodological approach of another discipline. In the case of the Goroka workshop, the practices and goals of computer science and computational linguistics were clearly quite different from those of descriptive and documentary linguistics. That is not to say that interdisciplinary approaches to language documentation and capacity building will necessarily prove to be as problematic or unsuccessful. But we should be sensitive to the possibility that the methodological imperatives of other disciplines may conflict with our commitments to the quality of the corpora we collect and to the quality of our relationships with communities.

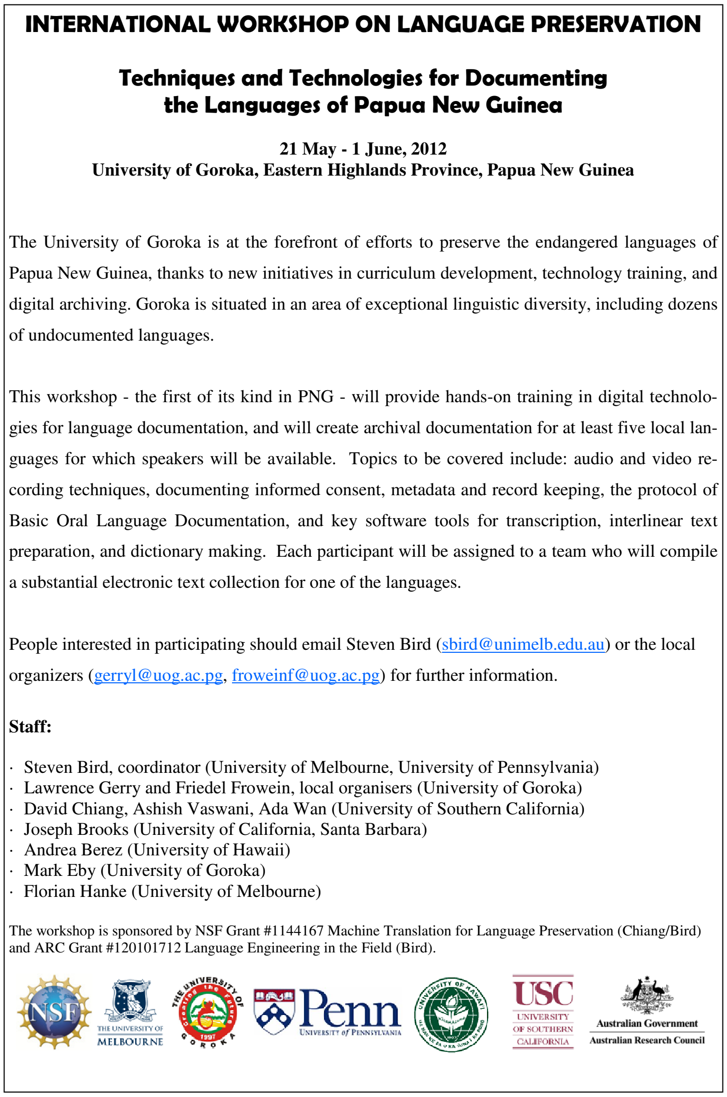
In a field that is now very concerned not just to document endangered languages but also to empower marginalized communities and their speakers, we need to be careful not to misidentify or misrepresent participants' interests. This is particularly critical since our assumptions as outsiders can lead us to falsely conclude that participants' goals are convergent with our own. Unfortunately, the focus of this workshop on implementing a new experimental methodology resulted in a top-down approach that was assumed to simultaneously serve the participants' interests. Bird et al. assert that the workshop participants were: "keen to have written language resources in the form of texts and lexicons, and are more interested in preserving content than documenting endangered speech styles" (Bird et al. 2013:157). However, the comments I heard at the workshop suggest that this statement does not represent what brought the Papua New Guinean participants there. One person was interested in translating Genesis into his language. Another hoped to learn how he might document place names in his community. One small group with prior SIL training expressed to me their discontent, because they had come in hopes of learning more about the structure of their language and creating useful language materials of some kind. At the beginning of the workshop when participants were briefly shown digital recorders (many of them had never seen such devices before), one participant who spoke no English and was illiterate became enthusiastic, but the focus of the workshop on written texts ended up excluding him. We need to think more critically about what workshop approaches would best advance the global documentation effort, while serving participants' goals, however they may be related to that effort.

As we think about how best to build capacity in PNG and beyond, a look into the text collection practices of our intellectual forebears can be revelatory. The collaborative relationship between the linguist Franz Boas and the Kwak'wala speaker George Hunt, for instance, lasted almost 40 years, and the linguistic training Boas provided Hunt and others resulted in one of the most extensive and robust corpora in the history of documentary linguistic fieldwork. The training Boas provided his consultants is something we might consider replicating in PNG, where an (albeit quite small) subset of the population would like to document their own languages. With that long-term goal in mind, the type of workshop that is needed in PNG is one that lays the groundwork for capacity building by providing training in basic documentary linguistic methods.

And in addition to learning from Boas' successes in training native-speaker linguists, we can also learn from some of his mistakes. One cross-cultural and humanistic complexity which the Goroka workshop could have been more sensitive to was the inherent power asymmetry between the outside workshop staff and the local participants. This relates directly to Boas' own textual practices, since he was unable to see those practices as embedded within a context characterized by differences in power. As Bauman & Briggs (2003) point out, Boas' relationship with Hunt may not have been as egalitarian and collaborative as many have assumed. This is because Boas constructed Native American texts in such a way as to "turn those elusive fieldwork encounters into the sorts of public, accessible, and objective observations required of scientific research" (ibid. 2003:281). This resulted in an erasure of just those details which linked the texts to their human and cultural context. Boas' efforts in his time have clear parallels in linguistics today, when we hope that methodological and technological contributions from subfields such as corpus and computational linguistics will help to advance the field as a whole. But we must not forget the human and cultural element in language that cannot ultimately be reduced—it can only be hidden from view.

As we seek to build capacity in marginalized communities, it is crucial that we also work to address the kinds of cross-cultural complexities which Boas and others may not have sufficiently considered. Whether with respect to field research or endeavors like the Goroka workshop, field linguists usually work in places where there is an imbalance of power, in many cases a continuing legacy from colonial times. Recognizing that disparity is a key step to addressing it and to ensuring that we do not replicate the practices that characterized so many relationships between powerful outsiders and indigenous persons in the past, and that of course is at the root of the massive language endangerment problem that we are working to address today. A further step in the right direction will be to provide training which speaks to participants' interests and which gives them the tools they need to become part of the conversation in language documentation, rather than constructing them as sources of data. After all, participants bring with them cultural and linguistic expertise which no outsider can replicate. If we are committed to increasing the capacity for language documentation in places where there is interest on the part of local communities, our efforts and resources intended for training purposes should be spent not on experimenting with shortcuts and quick solutions, but on building meaningful partnerships with those academic institutions, local students and faculty, and community members who are eager to engage with us.

4. APPENDIX .



INTERNATIONAL WORKSHOP ON LANGUAGE PRESERVATION

Techniques and Technologies for Documenting the Languages of Papua New Guinea

21 May - 1 June, 2012
University of Goroka, Eastern Highlands Province, Papua New Guinea

The University of Goroka is at the forefront of efforts to preserve the endangered languages of Papua New Guinea, thanks to new initiatives in curriculum development, technology training, and digital archiving. Goroka is situated in an area of exceptional linguistic diversity, including dozens of undocumented languages.

This workshop - the first of its kind in PNG - will provide hands-on training in digital technologies for language documentation, and will create archival documentation for at least five local languages for which speakers will be available. Topics to be covered include: audio and video recording techniques, documenting informed consent, metadata and record keeping, the protocol of Basic Oral Language Documentation, and key software tools for transcription, interlinear text preparation, and dictionary making. Each participant will be assigned to a team who will compile a substantial electronic text collection for one of the languages.

People interested in participating should email Steven Bird (sbird@unimelb.edu.au) or the local organizers (gerryl@uog.ac.pg, froweinf@uog.ac.pg) for further information.

Staff:

- Steven Bird, coordinator (University of Melbourne, University of Pennsylvania)
- Lawrence Gerry and Friedel Frowein, local organisers (University of Goroka)
- David Chiang, Ashish Vaswani, Ada Wan (University of Southern California)
- Joseph Brooks (University of California, Santa Barbara)
- Andrea Berez (University of Hawaii)
- Mark Eby (University of Goroka)
- Florian Hanke (University of Melbourne)

The workshop is sponsored by NSF Grant #1144167 Machine Translation for Language Preservation (Chiang/Bird) and ARC Grant #120101712 Language Engineering in the Field (Bird).

References

Bauman, Richard & Charles L. Briggs. 2003. *Voices of modernity: Language ideologies and the politics of inequality*. Cambridge: Cambridge University Press.

Bird, Steven, David Chiang, Friedel Frowein, Andrea L. Berez, Mark Eby, Florian Hanke, Ryan Shelby, Ashish Vaswani, & Ada Wan. 2013. The International Workshop on Language Preservation: An Experiment in Text Collection and Language Technology. *Language Documentation & Conservation* 7. 155–167. [http://hdl.handle.net/10125/4593]

Dobrin, Lise. 2008. From linguistic elicitation to eliciting the linguist: Lessons in community empowerment from Melanesia. *Language* 84(2). 300–324.

Foley, William. 1986. *The Papuan languages of New Guinea*. Cambridge: Cambridge University Press.

Foley, William. 2004. Language endangerment, language documentation and capacity building: challenges from New Guinea. In Austin, Peter K. (ed.), *Language documentation and description* 2. London: Hans Rausing Endangered Languages Project. 28–38.

Hale, Kenneth. 1968. Some questions about Anthropological Linguistics: The role of native knowledge. In Hymes, Dell (ed.), *Reinventing Anthropology*. Ann Arbor: The University of Michigan Press. 382–397.

Himmelmann, Nikolaus P. 1998. Documentary and descriptive linguistics. *Linguistics* 36. 161–195.

Himmelmann, Nikolaus P. 2006. Language documentation: What is it and what is it good for? In Gippert, Josh, Nikolaus P. Himmelmann & Ulrike Mosel (eds.), *Essentials of language documentation*. Berlin: Mouton de Gruyter. 1–30.

Palmer, Alexis, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell & Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for Uspanteko. *Linguistic Issues in Language Technology* 3:4. 1–42.

Woodbury, Tony. 2011. Language documentation. In Austin, Peter K. & Julia Sallabank (eds.), *The handbook of endangered languages*. Cambridge: Cambridge University Press. 159–186.

Joseph D. Brooks
josephdbrooks@umail.ucsb.edu