LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 7 (2013), pp. 155-167
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/4593

The *International Workshop on Language Preservation*: An Experiment in Text Collection and Language Technology

Steven Bird[(1)], David Chiang[(2)], Friedel Frowein[(3)], Andrea L. Berez[(4)], Mark Eby[(3)], Florian Hanke[(1)], Ryan Shelby[(3)], Ashish Vaswani[(2)], and Ada Wan[(5)]

(1) University of Melbourne, (2) University of Southern California,
(3) University of Goroka, (4) University of Hawaiʻi at Mānoa,
(5) City University of New York

With hundreds of endangered and under-documented languages, Papua New Guinea presents an enormous challenge to the documentary linguistics community. This article reports on a workshop held at the University of Goroka in May and June of 2012. The workshop aimed to collect written texts and their translations for several languages, while building local capacity through hands-on training, and improving our understanding of the appropriate use of technology. The majority of participants were mother tongue speakers who seek to preserve their languages through the preparation of written language resources.

1. INTRODUCTION.[1] Papua New Guinea (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. This problem is compounded by the fact that there are few opportunities for PNG-based scholars to receive training in language documentation.

The *International Workshop on Language Preservation* was a two-week training course held in May 2012. The workshop took place at the University of Goroka in the Eastern Highlands of PNG, and it operated under the auspices of the Language and Literature Department. The workshop had three goals: (1) to collect some documentation for at least five local languages; (2) to provide hands-on training in digital technologies for language preservation; and (3) to investigate methodologies that could lead to partial automation of the documentation process.

The concept of the workshop grew out of a three-day workshop held in February 2010.

[1] The workshop was sponsored by NSF grant IIS-1144167 *Machine Translation for Language Preservation*, and ARC grant DP120101712 *Language Engineering in the Field*. We are grateful to the University of Goroka for hosting the workshop, and to the following staff for their support: the Pro Vice Chancellor (Academic), Dr Michael Mel; the Acting Dean of Arts, Ms Helen Vetunawa; the Head of the Language and Literature Department, Ms Anne-Marie Wanamp; and the secretary of the department, Ms Velda Feka. We thank Steven Inapelo and Brian Kaeho for helping us to recruit workshop participants. We gratefully acknowledge the others who gave presentations in the workshop: Joseph Brooks (UC Santa Barbara), Bill and Sandra Callister (SIL PNG), Don Daniels (UC Santa Barbara), Kevin Poke (U Goroka), and René van den Berg (SIL PNG).

E-ISSN 1934-5275

University staff and local community members were trained in the use of digital voice recorders for recording stories and oral translations, following the methodology of 'Basic Oral Language Documentation' (Reiman 2010). After this workshop, the staff embraced the goal of language documentation; in the words of the department head:

> The department is eager to pursue an in-depth and far reaching program of language documentation. Our unique ability to access all provinces in Papua New Guinea, as well as the other Pacific Islands through our students, makes us the most logical 'hub' of language documentation in the region. We can serve as the central point in facilitating research, development, symposiums, translations, and publication of such data that pertains to the status of language, the linguistic properties of our languages, and the evidence of the intellectual or epistemological foundations of our languages and cultures. The department can also serve as the center of cross fertilization of languages and be involved in developing policies that concern with language, literacy, and culture. Our department members are eager to participate in any training on best practices for language documentation. It would help us immensely to document our efforts the 'correct' way the first time. (Anne-Marie Wanamp, pers comm, 2011)

Each staff member identified a local language that he or she would document. Collectively, they identified courses in each year level 1–4 where language documentation activities would be integrated into the curriculum. This was the ideal scenario: locally employed staff training students representing over 100 languages, while conducting language documentation themselves, in a place that is surrounded by dozens of under-documented languages.

The workshop had several outcomes. First, we prepared a small collection of written texts with translations and published a booklet. Second, we provided basic training in language documentation to interested scholars and language workers in the Goroka area and further afield, and we raised awareness concerning the needs and goals of language preservation amongst leaders in the university and local community. Third, we gained a deeper understanding of the technological requirements for mounting a locally-sustainable activity in language documentation.

This paper is organized as follows. We begin by describing the goals and methodology for our approach (Sections 2 and 3), then give the details of the program and participants (Sections 4 and 5). Next, we explain how the collected data was processed and provide summary statistics (Section 6). The paper concludes with a reflection on what was achieved, and future prospects.

2. GOALS. In view of the linguistic riches of Papua New Guinea, one would ideally like to replicate the range of documentary activities that exist elsewhere, including graduate programs in language documentation, annual training institutes, and sponsorship for the documentation of specific languages. Until this happens, we need lightweight methodologies that permit minimally trained speakers of endangered languages to collect materials that will have future documentary value. Our goals grow out of this desire to start the documentation work now, using the available resources as effectively as possible.

The first goal of the workshop was to collect documentation for at least five local languages. We understand *documentation* to consist of “comprehensive and transparent records supporting wide ranging scientific investigations of the language” (Woodbury 2010). However, the focus of the documentary effort needed to accommodate the interests of the university, the capacities of the participants, and the technological exploration being conducted by the researchers. The University of Goroka’s ongoing agenda is to foster language research in its region and to build strong links with local language communities. The language workers are 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. The outside researchers wanted to explore the use of technology in collecting written texts to complement the earlier workshop’s focus on spoken texts.

The second goal was to provide *education* for participants in a methodology that they can continue to use in the absence of external support. Without funding, and without local expertise in language documentation, volunteers can still perform tasks having documentary value. After all, the textual resources in literate societies past and present have not been produced by linguists, yet they serve as an important body of linguistic evidence. What minimal intervention could trigger the ongoing production of materials that would ultimately support linguistic description and analysis once the language has fallen out of use? This amounts to a limited version of what Bernard (1996) has advocated, with the establishment of independent local publishing industries. Another point of comparison is the way in which ephemeral Australian Aboriginal sand art was preserved thanks to limited external encouragement and resources (Bardon and Bardon 2006).

The third goal of the workshop was to investigate methodologies that could lead to partial *automation* of the documentation process (Abney and Bird 2010; Bird and Chiang 2012). The main idea was to see to what extent the texts produced by language workers could be used to train computational models similar to those used in machine translation systems, so that those same models could be used to aid language workers in producing new texts.

These goals led us to make a number of choices that shaped the execution of the workshop, as described in the next section.

3. WORKFLOW. We presented a simplified documentary process consisting of five components: spoken language, written language, translation, lexicon, and archiving. All workshop activities were referenced back to these so that participants could see what purpose was being served by each task. Translation was presented as a bidirectional task, for bidirectional access to knowledge, and for generating new content in the local language in order to challenge the widespread assumption that the local language is for traditional genres only. The lexicon was identified as a separate task because participants were interested in their language’s stock of words and the associated material culture, and because this would provide a reference for orthographic representation. The five activities were mapped onto the workshop schedule as explained in Section 4. The remainder of this section describes some key features of the workflow.

3.1 HANDS-ON EXPERIENCE. The most distinctive aspect of the program was the 5–6 hours of documentation on most days. We felt that participants would learn from doing, and that some issues would only become apparent once a body of documentation had already been collected. We wanted to avoid the situation where people had only practiced initial tasks and were unable to proceed further (once the trainers had departed). We also wanted to avoid the situation where people listened to days of presentations about language documentation without applying themselves to the task.

A large part of the time (for many participants) was spent directly digitizing texts using Fieldworks Language Explorer, more commonly known as FLEx (Butler and van Volkinburg 2007). This served our first goal of documentation by removing the bottleneck that would have been created by limiting computer use to the workshop organizers; it served our second goal of education by giving participants training in standard linguistic software. It also served our third goal of investigating future automation, as we needed to assess what kinds of user interfaces would be suitable for the widest range of users.

3.2 SPOKEN VERSUS WRITTEN LANGUAGE. In a previous workshop, participants were trained in oral language documentation using digital voice recorders (Bird, 2010). However, the methodology was too onerous: collecting metadata and collating digital files required regular external support that was not available, and careful transcription using voice recorders without fine-grained control of rewinding was too difficult. For the present workshop, in order to streamline progress towards our goal of documentation, we elected to focus on the collection of written text and to avoid spending the bulk of the time on painstaking annotation for a single text, which would have been a discouragement to participation. As it turned out, the focus on writing was almost exclusive, as most of the audio recorders previously provided to the university were unavailable at the time of the workshop owing to their use in ongoing coursework.

The choice to focus on writing was also compatible with our goal of education, in that we would be building on participants' own skills and interests. Many already had extensive experience with writing their languages, including literacy teaching and Bible translation. They spoke of the need for written resources, especially the need for educational materials including lexicons and text collections. An important driver for their participation was that the workshop would have a written product. Finally, in view of our goal of investigating future automation, we wanted to test the feasibility of a text-only workflow, to avoid the added complexity of processing speech data.

3.3 TRANSLATION. In order to develop a documentation workflow that would be both productive during the workshop and sustainable after the workshop, using only volunteer workers and not trained professional linguists, we decided that the primary finished product of the workshop would be interlinear glossed text, without morphological analysis or morpheme glossing. The omission of morphology is sure to be controversial; however, given the lack of an available morphological analysis for most of the languages, and the impossibility of developing one in the available time, morphological analysis is beyond our reach. It was not even clear *a priori* whether word glossing was a realistic goal (see Section 6 below for our findings).

This choice is also consistent with our goal of developing technologies for automation. Current research in unsupervised morphological analysis is beginning to produce promising results (e.g., Snyder and Barzilay 2008), and unsupervised word alignment – which is basically equivalent to word glossing – is a highly successful area of machine translation research (e.g., Brown, 1993). Both kinds of models are able to operate using only parallel text (text with phrasal translations). Our hope is that, with a sufficient amount of parallel text, computers will be able to assist linguists with morphological analysis and will largely automate word glossing.

4. PROGRAM. Most days had the same organization, consisting of an opening presentation, a two-hour documentation activity, a lunch break, then a second presentation (first week only), followed by 2–3 hours of further documentation activities. Most days closed with reporting back and review. The documentation activities were typically done in groups of 2–4 speakers who spoke the same or related languages.

The presentations were at an introductory level, and did not assume prior knowledge of linguistics. They lasted 45–60 minutes and were delivered in English or Tok Pisin. Those delivered in English were interpreted into Tok Pisin whenever possible. The following presentations were included in the program, and were mostly delivered during the first week.

1. Opening address on language preservation (Michael Mel and Helen Vetunawa, U Goroka)
2. The languages of Papua New Guinea (Joseph Brooks, UC Santa Barbara)
3. Introduction to language documentation (Andrea Berez, U Hawaii)
4. The effects of language loss in Kamano Kafe (Kevin Poke, U Goroka)
5. The language documentation workflow (Steven Bird)
6. SIL in Papua New Guinea (Bill and Sandra Callister, SIL)
7. Rapid word collection (René van den Berg, SIL)
8. Lexicography (René van den Berg)
9. Fieldworks Language Explorer – FLEx (Steven Bird)
10. Audio recording techniques (Andrea Berez)
11. Video recording techniques (Mark Eby)
12. Comparative kinship (Don Daniels, UC Santa Barbara)
13. Machine translation (David Chiang)
14. Language archiving (Andrea Berez)
15. Documenting informed consent (Steven Bird)
16. Language endangerment (Friedel Frowein)

Video recordings of these presentations are lodged in the Internet Archive (http://archive.org/details/IWLP2012). The presentations on lexicography, FLEx, audio recording, and kinship were immediately followed by corresponding documentation activities.

The documentary activities were mapped onto the schedule as follows. An opening activity on day one was to form teams of three, all speaking a different language, and to jointly complete a Swadesh wordlist. (This activity was particularly apt, falling on the UN World Day for Cultural Diversity.) Apart from breaking the ice, this time-limited activity was designed to give people practice at working quickly, not being overly concerned about the correct way to write a word, and skipping over cases where the answer was unclear. On day two, participants were instructed in how to format interlinear glossed texts using their A5 sized exercise books, using the left side of each opening for the source text, triple-spaced to allow room for word-level glosses, and the right side for phrasal translation. The method was introduced in stages, with each stage explained first in English then in Tok Pisin. On day three we returned to lexicography, this time trying out the Rapid Words Method (http://rapidwords.net/). On day four the focus reverted to written texts, but this time with the aid of digital technologies, for audio and video recording and for digitization of texts using FLEx. On day five, to round out the first week, all the activities were brought together on a single day with lexicography (this time on kinship), audio recording, transcription, and keyboarding.

The second week involved few presentations, and instead, the bulk of the time was spent on text collection: generating story ideas, composing texts in the exercise books, glossing and translating the texts, keyboarding, and displaying them on the wall for others to read. These intensive activities were broken up with a variety of group activities including storytelling and reading aloud, designed to get people thinking about new topics to write about, and to publicly recognize those who had taken their story through the complete workflow. To maintain people's interest, we varied the topic of the texts, including: traditional stories, daily life (gardening, markets), and current affairs (the election, a ferry disaster). We varied the composition of the groups: some activities called for same-language groups (e.g. editorial work), while other activities called for different language groups (e.g. ideas for more stories). We also varied the location: we were fortunate to have access to a variety of spaces including a large auditorium, sheltered outdoor seating around tables, a well-equipped computer laboratory, and meeting rooms.

The final day opened with a tongue-in-cheek announcement that we were gathered to mark the centenary of a workshop on language preservation, held 100 years ago in 2012; many of the ancestral languages of the region had now fallen out of use but we were fortunate to have some records of these languages that were collected at that workshop and preserved in the university archive. After some further elaboration we came back to the present, but now with a long-term view of our two weeks of work and its potential significance. Discussion moved naturally into a brainstorming session about further language preservation activities the participants want to pursue. These included documentation activities: eliciting texts from elders; recording narratives, dialogues, and songs; composing original narratives, poems, and songs; preparing a dictionary; and depositing these materials in a planned archive at the University of Goroka. They also included many revitalization activities: preparing literacy books for use in schools and churches; creating CDs of narratives, dialogues, and songs for use in schools and churches; establishing language schools alongside elementary schools; translating songs into local languages; running 'language camps', establishing orthographies. They also discussed the required approach to these activities: moving forward in small steps, helping each other; raising awareness of language shift so that community leaders support language revitalization activities; attending the SIL PNG dictionary making course. This list of activities was written up on a single page and distributed to participants. Following this discussion of future work, we presented the printed booklets containing all the texts and translations, along with a certificate and the financial gift, and this concluded the workshop.

During the workshop, documentary activities proceeded in tandem with a small amount of elicitation work by individual presenters in the areas of phonology, morphology, and syntax. After the workshop, seven of the language consultants stayed on for a further week of text collection work.

Goroka worked well as the workshop location, given the local university support, the agreeable climate, and the airport with twice daily flights to the capital. Unfortunately the timing of the workshop was not suitable for several of the university staff given their examination responsibilities. Earlier dates would have overlapped with the teaching semester when staff and venues were fully occupied, while later dates would have run up against the national election when travel was considered unsafe.

5. PARTICIPATION. Participants were identified in several ways. The event was promoted across the university and attracted participants from the Department of Social Science and Commerce, and the Department of Mathematics and Computing, in addition to the host department. Some of the organizers already had contact with people who speak the languages they are investigating, and we contacted these speakers directly. University staff invited older speakers of their mother tongue. We also used the church network: pastors from the Alekano community of Goroka got word out to their colleagues across the Eastern Highlands Province, and they sent people to represent the local languages. Some of these people had been involved in translation and literacy work years ago, and were eager to return to language work. All of them were literate. A further source of participants was other professional organizations including the Institute of PNG Studies and the University of PNG (both in Port Moresby), and SIL. Several additional staff and students from the University of Goroka attended workshop sessions. A list of the languages represented at the workshop is given in Table 1.

The requirements of the workshop called for two main kinds of participant: trainee language documenters (university staff), and speakers of endangered languages. In many cases, the staff were affiliated with a local language, but few were old enough to have a good command of the oral literature. The inducement for the staff to participate was to receive training in language documentation. The inducement for the speakers to attend was to support language development, and to have their writing published in the story book that we published at the end of the workshop. The speakers were reimbursed for their travel expenses, and paid a small daily allowance; those who stayed to the end of the workshop received a financial gift, partially compensating them for lost time cultivating their crops. The level of financial compensation proved to be about right: it was low enough that we only attracted participants who were enthusiastic to work on their language, and it was high enough that most participants attended for the whole two weeks of the workshop.

A final comment should be made about how participants were compensated for their time. University staff were not compensated, because the workshop was aligned with the goals of the Language and Literature Department and scheduled during working hours. Workshop funds were limited, and we could not afford to pay the non-university participants very much, yet we did not want to turn away anyone who wanted to participate. Equally, we did not want to attract people who were only seeking employment and were not committed to language preservation work. Furthermore, we wanted people to commit to attending the whole workshop. After wide consultation, we decided to compensate participants as follows. Those within commuting distance from Goroka were reimbursed for their bus fares; those from further afield had to stay in Goroka and were paid an allowance to cover local accommodation with relatives. We paid a K10 ($5) allowance each day (about the level of a day's laboring work in the village). Additionally, we wanted people to stay on site for the day instead of leaving for lunch, so we provided lunch (costing K20 per head). We gave a further K200 in the form of a gift on the last day. This approach avoided the sense that people were being paid a specific rate for their time, thus preventing an unhelpful precedent for future projects that may have even fewer resources. It also fitted the culture of gift giving. The amount of the gift was not disclosed in advance, and it was clear that participants were motivated by their enthusiasm for the work rather than financial gain.

Language	**ISO**	**Province**	**Population**	**Participants**
Adzera	adz	Morobe Province	28,900	1
Alekano	gah	Eastern Highlands Province	25,000	8
Benabena	bef	Eastern Highlands Province	30,000	1
Boiken	bzf	East Sepik Province	31,000	1
Dano	aso	Eastern Highlands Province	30,000	2
Huli	hui	Hela Province	71,000	1
Kafe	kbq	Eastern Highlands Province	63,000	2
Kuman	kue	Simbu Province	115,000	2
Motu	meu	Central Province	31,000	2
Nii	nii	Western Highlands Province	12,000	2
Siane	snp	Eastern Highlands Province & Simbu Province	29,000	2
Siwai	siw	Autonomous Region of Bougainville	6,600	1
Toaripi	tqo	Gulf Province	23,000	1
Tokano	zuh	Eastern Highlands Province	6,000	2
Usarufa	usa	Eastern Highlands Province	1,300	3
Yaweyuha	yby	Eastern Highlands Province	2,000	1
Yate	ino	Eastern Highlands Province	10,000	2

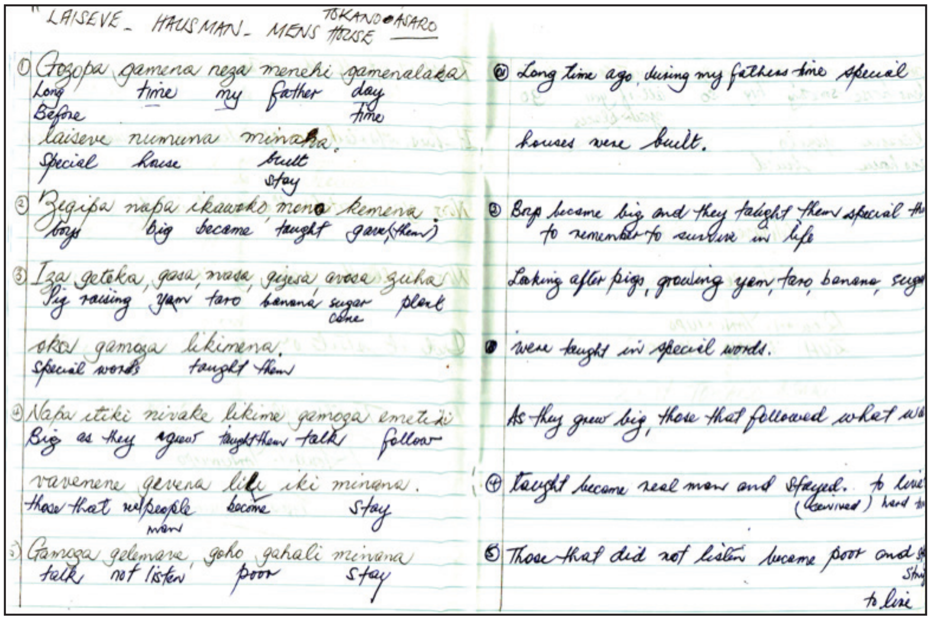
TABLE 1: Languages represents at the workshop

6. DATA PROCESSING. For most tasks, participants composed a text in their own language then added glosses in English (or Tok Pisin), and finally added a phrasal (or "free") translation in English (or Tok Pisin).

Initial documentation work was done using pen and paper, as shown in Figure 1. After the first few days, participants began entering data into a computer using SIL FieldWorks

Language Explorer (FLEx), as shown in Figure 2. When FLEx encounters a word it has seen before, it suggests possible glosses for that word. We do not have measurements of how much this speeds up the glossing process, if at all, but we intend to increase the level of automated feedback in the future using custom software.

The language workers had varying degrees of familiarity with computers. A few composed directly into FLEx. Some copied their stories from their notebooks into the computer. Others were paired up with more computer-literate partners who served as typists. This arrangement worked fairly well, since older people tend to have the most authoritative knowledge of their language, whereas younger people tend to have more experience with computers. However, some of our most skilled typists reported that data entry work was not intellectually challenging enough for them. Once the texts were digitized, they were printed, read aloud, and displayed for others to read.



"LAISEVE – HAUSMAN – MENS HOUSE TOKANO ASARO

① Gozopa gamena neza menehi gamenalaka
Long time my father day
Before time
laiseve numuna minaka.
Special house built
stay
② Zegipa napa ikawoko mena kemena
boys big became taught gave(them)
③ Iza geteka, gasa, masa, gijesa, avosa, guha
Pig raising yam taro banana sugar cane plant
oka gamoga likimena.
Special words taught them
④ Napa itiki nivake likime gamoga emetiki
Big as they grow taught them talk follow
vavenene gevena like iki minana.
those that ~~ne~~ people man become stay
⑤ Gamoga gelemave, goho gahali minana
talk not listen poor stay

① Long time ago, during my fathers time special
houses were built.
② Boys became big and they taught them special th
to remember to survive in life
Looking after pigs, growing yam, taro, banana, sug
were taught in special words.
As they grew big, those that followed what w
④ taught became real man and stayed. to live
(survived) hard t
⑤ Those that did not listen became poor and st
to live

FIGURE 1: Interlinear text, with source text and word-level glosses (left) and phrasal translation (right)

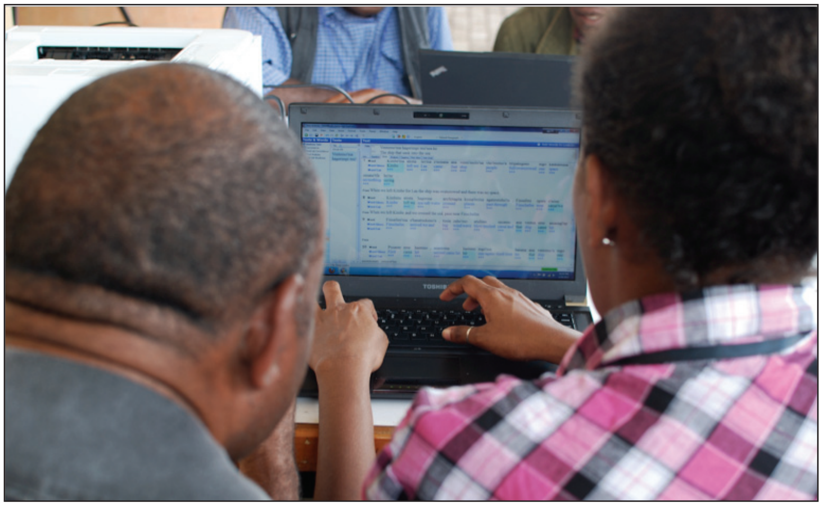


FIGURE 2: Entering interlinear text into Fieldworks Language Explorer (FLEx)

Contrary to expectation, phrasal translation turned out not to be the most difficult of the three basic tasks. Composition was much less productive than expected: 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 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.

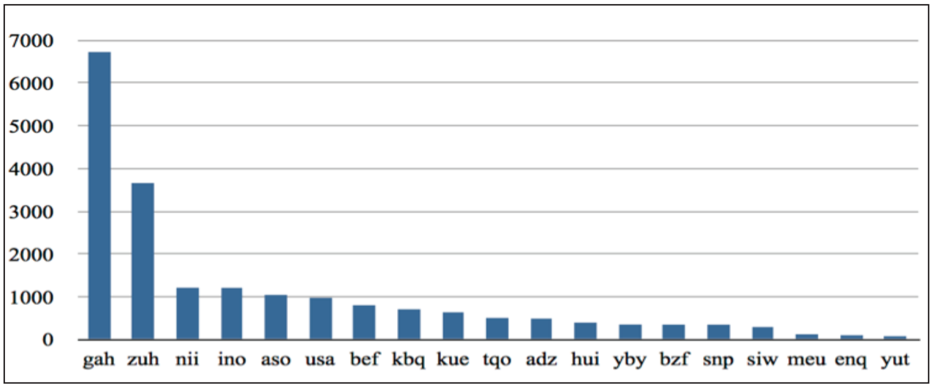


FIGURE 3: Words collected, by language

More seriously, the style of composition was limited. Sentences averaged 9.2 words in length (including punctuation), 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.

It seems likely that the medium of composition biased writers toward the style of elementary English texts that they were familiar with. Some even composed their stories in English before translating them into their mother tongue.

	stories	sentences	words (source)
composed	226	2,156	19,805
glossed			13,730
translated		1,778	15,956

Table 2: Amounts of linguistic data collected at each stage

In all, we collected about 20,000 words of source text (see Table 2). Of this, about 16,000 words were translated into another language (mostly English, with some into Tok Pisin and some into Alekano), and about 14,000 words were glossed (at the word level). The distribution of data collected across languages was highly skewed, which is not surprising (see Figure 3). Alekano is the primary language in the Goroka area, and Tokano is also spoken nearby and is closely related to Alekano.

The collected data was left at the university in hardcopy and digital format, and there are plans at the university for establishing a digital archive to support preservation and access in future. The materials are also being lodged in the PARADISEC archive (http://catalog.paradisec.org.au/collections/IWLP12). All presentations and source materials are also available in the Internet Archive (http://archive.org/details/IWLP2012).

Figure 4: Workshop Participants

7. CONCLUSION. With hundreds of endangered and under-documented languages, Papua New Guinea presents an enormous challenge to the documentary linguistics community. Local scholars are an untapped resource: they are already affiliated with these linguistic groups, and they are often involved in teaching and research which is consistent with documentary activities. Literate elders are another untapped resource: they have the best knowledge of oral literature, they are highly motivated to preserve this knowledge, and they have available time to transcribe and translate stories. A small amount of training promises to yield big dividends in terms of the quantity and quality of language documentation.

The *International Workshop on Language Preservation* was designed to take advantage of these opportunities. Approximately 40 language workers enthusiastically volunteered their time (Figure 4), for little more than travel and subsistence expenses. Several participants travelled for a day each way in order to attend, while several others commuted for over two hours each day. We delivered 12 hours of presentations and provided 40 hours of hands-on training. This proved sufficient to cover a range of documentary activities in some detail, and was not so long that people were overcome by fatigue. Ninety percent of participants stayed to the end.

The work reported here represents the beginnings of a new approach to the problem of language documentation, and it has close ties to the fields of corpus linguistics, computational linguistics, and data-intensive experimental linguistics (cf. Abney 2011, Bird 2011). As in documentary linguistics, progress in these fields depends on having broad samples of language use. Practitioners routinely collect and analyze corpora of a million words or more, and they bring their own computational methods for working with languages. Combining these methods with those used by documentary linguists may yet provide the most effective means for scaling up the pace of the documentary work while there is still time.

It is too early to evaluate the long-term impact of the workshop. We have shown how local scholars and literate elders in Papua New Guinea can be trained to prepare parallel text collections, and several of them are continuing in our absence. We were pleased with the quantity of data that we were able to collect, though it may be more effective to return to working with the spoken language directly, and building up a larger pool of transcribers. The university community has experienced a new model for engagement with the linguistic communities in its region, and this could be developed further in the context of the university's ongoing teaching and research activities.

References

Abney, Steven. 2011. Data-intensive experimental linguistics. *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/lilt/article/view/2578.

Abney, Steven and Steven Bird. 2010. The Human Language Project: Building a universal corpus of the world's languages. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 88-97. Association for Computational Linguistics. http://www.aclweb.org/anthology/P10-1010.pdf.

Bardon, Geoffrey and James Bardon. 2006. *Papunya: A place made after the story: the beginnings of the Western Desert painting movement*. Aldershot: Lund Humphries.

Bernard, H. Russell. 1996. Language preservation and publishing. In Hornberger, Nancy (ed), *Indigenous literacies in the Americas: Language planning from the bottom up,* pp. 139–156. Mouton de Gruyter.

Bird, Steven. 2010. A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, Gold Coast, Australia, pp. 5–14. Springer.

Bird, Steven. 2011. Bootstrapping the language archive: new prospects for natural language processing in preserving linguistic heritage. *Linguistic Issues in Language Technology* 6. http://elanguage.net/journals/lilt/article/view/2580.

Bird, Steven and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of the 24th International Conference on Computational Linguistics*, 125–133, Mumbai, India. http://www.aclweb.org/anthology/C12-2013.pdf.

Brown, Peter F., Vincent J. Della Pietra, Steven A. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19, 263–311. http://www.aclweb.org/anthology/J93-2003.pdf.

Butler, Lynnika and Heather van Volkinburg. 2007. Review of Fieldworks Language Explorer (FLEx). *Language Documentation and Conservation* 1, 100–106. http://hdl.handle.net/10125/1730.

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

Nettle, Daniel. 1999. *Linguistic Diversity*. Oxford: Oxford University Press.

Reiman, D. Will. 2010. Basic oral language documentation. *Language Documentation & Conservation* 4, 254–268. http://hdl.handle.net/10125/4479.

Snyder, Benjamin and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies,* pp. 737–745. http://www.aclweb.org/anthology/P08-1084.pdf.

Woodbury, Anthony. 2010. Language documentation. In Peter Austin and Julia Sallabank (eds.), *The Cambridge Handbook of Endangered Languages*, Cambridge University Press.

Steven Bird
sbird@unimelb.edu.au

David Chiang
chiang@isi.edu

Friedel Frowein
froweinf@uog.ac.pg

Andrea Berez
andrea.berez@hawaii.edu

Mark Eby
ebym@uog.ac.pg

Florian Hanke
florian.hanke@gmail.com

Ryan Shelby
shelbyr@uog.ac.pg

Ashish Vaswani
avaswani@isi.edu

Ada Wan
adawan919@gmail.com