Machine translation for language preservation

Steven BIRD[1,2] *David CHIANG*[3]
(1) Department of Computing and Information Systems, University of Melbourne
(2) Linguistic Data Consortium, University of Pennsylvania
(3) Information Sciences Institute, University of Southern California
sbird@unimelb.edu.au, chiang@isi.edu

ABSTRACT
Statistical machine translation has been remarkably successful for the world's well-resourced languages, and much effort is focussed on creating and exploiting rich resources such as treebanks and wordnets. Machine translation can also support the urgent task of documenting the world's endangered languages. The primary object of statistical translation models, bilingual aligned text, closely coincides with interlinear text, the primary artefact collected in documentary linguistics. It ought to be possible to exploit this similarity in order to improve the quantity and quality of documentation for a language. Yet there are many technical and logistical problems to be addressed, starting with the problem that – for most of the languages in question – no texts or lexicons exist. In this position paper, we examine these challenges, and report on a data collection effort involving 15 endangered languages spoken in the highlands of Papua New Guinea.

KEYWORDS: endangered languages, documentary linguistics, language resources, bilingual texts, comparative lexicons.

Proceedings of COLING 2012: Posters, pages 125–134,
COLING 2012, Mumbai, December 2012.

1 Introduction

Most of the world's 6800 languages are relatively unstudied, even though they are no less important for scientific investigation than major world languages. For example, before Hixkaryana (Carib, Brazil) was discovered to have object-verb-subject word order, it was assumed that this word order was not possible in a human language, and that some principle of universal grammar must exist to account for this systematic gap (Derbyshire, 1977). In spite of the scientific importance of the world's languages, computational linguistics research has only touched about 1%. In 100 years, 90% will be extinct or on the way out (Krauss, 2007). Linguists are addressing this problem by *documenting* the world's endangered languages (Woodbury, 2010). What can computational linguistics offer to support this urgent task?

Machine translation (MT) is directly relevant to the process of language documentation (Abney and Bird, 2010). First, when source texts are translated into a major world language, we guarantee that the language documentation will be interpretable even after the language has fallen out of use. Second, when a surviving speaker can identify errors in the output of an MT system, we have timely evidence of those areas of grammar and lexicon that need better coverage while there is still time to collect more. These tasks of producing and correcting translations can be performed by speakers of the language without depending on the intervention of outside linguists. Furthermore, we sidestep the need for linguistic resources like treebanks and wordnets, which are expensive to create and which depend on the existence of morphological, syntactic, and semantic analyses of the language.

For over a century, an early task in describing a new language has been to collect and translate texts, where a "text" could be a written document or a transcribed recording. Despite the documentary value of such data and its usefulness for linguistic research, for most languages there is no collection of texts and translations. Now, transcribing and translating audio recordings takes upwards of ten times real time. It is evidently not practical for an expatriate linguist to do such work, based on the track record of past language documentation projects in which the text collection only amounts to a few thousand words. We would need a thousand times as much primary data in order to support wide-ranging investigations of a language once it is no longer spoken, equivalent to 10 million words, or 1,000 hours of speech (Liberman, 2006) Yet a small team of bilingual speakers should be able to transcribe and translate a substantial collection of texts in a few months. The questions then shift to the following: (a) how can we harness the efforts of minimally trained bilingual speakers to create and share bilingual texts? (b) how can we maximise the consistency of the data in the absence of an orthography or a dictionary? (c) how can we tell when enough of the right kind of data has been collected?

These are difficult questions to answer. In this paper we point a way forward. After a background discussion, we discuss a simplified workflow for language documentation and the role that MT can play in that workflow, then we report on our experience of collecting bilingual spoken and written texts in Papua New Guinea.

This work represents a new approach to language preservation. It begins with the observation that linguists will probably not be able to collect an adequate sized corpus. It leverages local capacity to get started on the work rather than waiting until outside linguists to arrive. It puts the work in the hands of locals, who can make their own decisions about what should be preserved. And it offers a plausible way to limit the "observer effect" which occurs when an outsider comes into a language situation and starts eliciting data (Himmelmann, 1998, 184ff).

2 Background

A statistical translation model is simply a model of parallel text, that is, a model that knows what sentence pairs are more likely than others to occur as translations of each other. Accordingly, a prerequisite for building a statistical MT system for any language pair is to collect texts and their translations into a reference language. However, this coincides with a key activity in documentary linguistics, and harks back to the early days of 19th century descriptive linguistics in which text collection is a major component.

A *language documentation* consists of "a comprehensive and representative sample of communicative events as natural as possible" (Himmelmann, 1998, 168), or "comprehensive and transparent records supporting wide ranging scientific investigations of the language" (Woodbury, 2010). The ideal form of the primary data is video, though audio is a good second-best, and requires less expertise and less expensive equipment, and produces smaller data files. To facilitate access, the raw data is usually transcribed and translated. It should be clear that language documentation is not the same as linguistic description, which calls for linguistic expertise and which produces systematic presentations of the phonology, morphology, syntax, and semantics of the language. Nevertheless, the descriptive work cannot proceed without the language documentation. This documentation – the bilingual text collection – is the same as what is needed for statistical MT and we can expect to apply MT algorithms to the data from linguistic fieldwork (Xia and Lewis, 2007; Palmer et al., 2010).

The workflow for language documentation and description has never been standardised, but there is general agreement that it involves at least the following activities: (a) recording communicative events; (b) transcribing and translating the recordings; (c) performing basic morphosyntactic analysis leading to a lexicon and to a collection of morphologically-glossed text; (d) eliciting paradigms, i.e. systematic tabulations of linguistic forms designed to reveal underlying patterns; (e) preparing descriptive reports to show how the language is structured. These activities are well understood and widely practiced, and provide the empirical foundation for linguistic theory and for the preparation of language resources such as treebanks and wordnets. However this workflow does not scale up. Languages are falling out of use before linguists can get to them.

This leaves the question of what quantity and quality of documentation is required. Here the only consensus amongst linguists is that more is better. Yet linguist-driven documentation projects only produce a tiny fraction of the quantity required for corpus-based studies. Linguists stress the importance of quality, which includes the accuracy and consistency of transcriptions and glosses, but do not report explicit measures of transcription quality (e.g. the Kappa coefficient, widely used for inter-annotator agreement). Since the documentary linguistics community does not provide objective methods and measures of quantity and quality, we need to develop these ourselves.

Note that the agenda is not to remove linguists from the language documentation process. Without specialised training, speakers of endangered languages will never produce the lexicons, morphologically glossed text, treebanks, and wordnets that we would like to have. Instead, we want to capture enough bilingual text to enable documentation and description even after the language has fallen out of use and only the archived documentation is available.

3 A simplified workflow for language documentation

How could minimally-trained speakers of a language create a useful corpus for their language? From the earliest days of corpus construction for English, the first step was to have a digital text collection, from which a balanced corpus could be selected and further annotations applied. However, most endangered languages lack any kind of text collection. Thus, we would like to find a way to produce a substantial text collection for a language without external staffing and resourcing. We envisage that members of the speech community could create documentary artefacts – recordings and transcriptions – using locally available technology, even if it is only a pen and exercise book, or an inexpensive recording device.

The first step is to create a text, either by recording then transcribing, or by composing directly onto paper. Chances are that the speaker will have no experience at IPA transcription and that no standardised orthography for the language exists. Thus, transcription needs to use whatever orthography people know. This practice has some documentary value, for it shows meaningful sound contrasts and word boundaries, and serves as a rough finding aid. In cases where more than one speaker transcribes content in a language, we can try to clean up the transcriptions automatically (Foda and Bird, 2011).

The second step is to translate the text, providing word by word glosses plus a phrasal translation. The correspondence between this literal and "free" translation amounts to training data for an alignment model, and does not require a separate translation model. The final step is to prepare a lexicon, in order to help fix the inconsistencies in spelling and glossing between. SIL's *Fieldworks Language Explorer* (FLEx) software is ideal for this purpose, though it currently lacks support for synchronisation and conflict resolution between databases.

An important refinement is to conduct the above workflow within a cluster of closely related languages. Speakers often produce a wealth of information about lexical correspondences with neighboring languages, as illustrated in Figure 1. Armed with these correspondences, we can pool knowledge about all the languages in the cluster (Nakov and Tiedemann, 2012). We can also try to guess word translations by leveraging regular sound correspondences.

eng	aso	bef	gah	ino	kbq	snp	yby	zuh
sun	ho	yege	ho	yake	zge	fo	homa	ho
water	noso	nagami	nagami	tina	tina	no	noma	nosa
fire	olo	logo	lo	ata	teve	soo	iizo	olo
earth	misumbo	mei	mikasi	mopa	mo'pa	mika	mika	mikesupa
tree	ya	yafa	za	yosa	zafa	yaa	yah	yah
mountain	golo	kosa	agoka	akoya	agona	obura	bora	gola
house	numuno	nohi	numuni	nona	nona	numuna	numuda	numuna
food	nosonite	nosena	nosa'neta	neya	ne'zane	aáwa'a	nodenesa	nosaneta
pig	ije	yaga	iza	afu	afu	savu	izah	iza
man	we	bo	ve	ve	ve'nene	wee	we	vemoha
woman	vene	amo	vena	a'ne	a're	wena	mena	vena
father	meneho'we	afonifu	ahono	afo nimo'e	nenfa	wemeteuo	ahone	meneho
mother	ijeneho	itonifu	izo'no	ita	anta'nimo	wena otevo	idone	izeneho

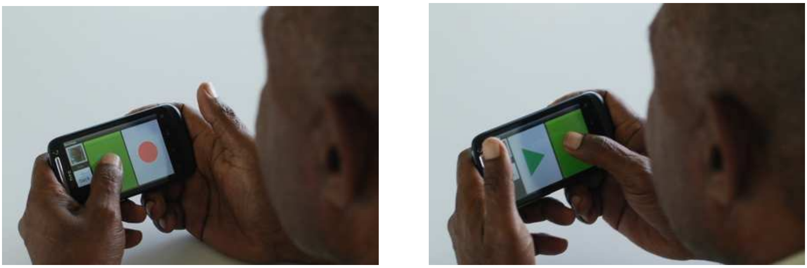
Figure 1: Comparative wordlist for the languages spoken near Goroka. Languages are identified by ISO 639-3 code. It is likely that, for some language pairs (e.g. aso-zuh, ino-kbq), many wordforms are related to one another by regular sound correspondences.

4 Collecting parallel and comparable texts in Papua New Guinea

Papua New Guinea (PNG) is home to the greatest number of languages and the greatest diversity of language families in the world (Nettle, 1999), including many languages with only a few hundred speakers. Although there is a long history of linguistic description in PNG (Foley, 1986), few of these languages have been comprehensively documented. There is no up-to-date picture of language vitality across PNG, and no systematic efforts to preserve them on the kind of scale that would be required. Some small languages are clearly vital: for example, the Nen language, spoken in the Morehead District, has a population of just 300, and the language is reportedly being transmitted completely to the younger generation (Nicholas Evans, pers. comm.). Nevertheless, many languages – perhaps even the majority – are already moribund and are quickly being overtaken by Tok Pisin, an English-based creole. In the face of this language shift, there is almost no local capacity for language documentation.

Bird trained university staff and students, adult literacy workers, and retired professionals, to collect oral literature using 100 digital voice recorders (Bird, 2010). Participants learned the technique of "respeaking", which involves listening to an original recording and repeating what was heard carefully and slowly (Woodbury, 2003), resulting in a secondary recording which is much easier to transcribe later on. The respoken version plus a phrase-by-phrase spoken interpretation are captured on a second voice recorder. Each voice recorder comes with an A5 exercise book which is used for logging recordings, and keeping track of the different linguistic genres that have been collected. Genres included dialogue, narrative, procedural discourse, oratory, and singing (Johnson and Aristar Dry, 2002).

The result of that work has been a set of phrase-aligned audio files for approximately 50 languages. One significant shortcoming of this approach is that it is virtually impossible to manage files that are collected on 100 voice recorders. Instead, we have developed a mobile phone interface, as shown in Figure 2. It can be used for audio collection and sharing, and for respeaking and interpreting (Hanke and Bird, 2012).

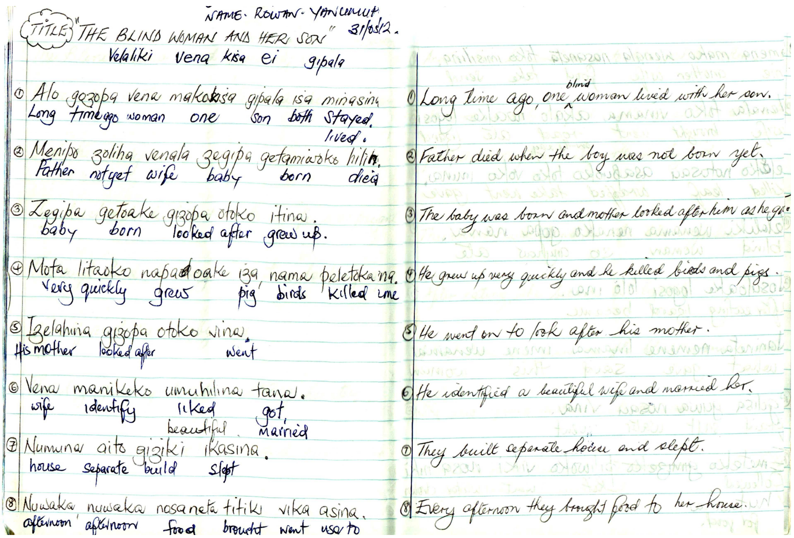


(a) Audio playback

(b) Respeaking and Interpretating

Figure 2: Mobile phone interface: (a) press and hold the play button to hear the original recording (b) press and hold the record button to record the respeaking or interpreting

However, voice recorders and mobile phones can only collect bilingual audio, while machine translation technologies require bilingual text. We organised a two week workshop at the University of Goroka involving approximately 40 speakers of 15 undocumented languages (Bird et al., 2012). We elicited comparable texts across the languages with a variety of tasks, for example: (a) write about the national election or about a traditional legend; (b) listen to someone's story and put it in your own words, e.g. the *Rabaul Queen* disaster; (c) listen to dictation in English and Tok Pisin, but translate each sentence into your language, e.g. a story about a visit to the chicken market. Each text was set out using the format shown in Figure 3.



NAME. ROWAN. YANUMUT
TITLE "THE BLIND WOMAN AND HER SON" 31/05/12.
Velaliki vena kisa ei gipala

① Alo gezopa vena makokisa gipala isa minasing
Long time ago woman one son both stayed. lived.

② Menipo zoliha venala zegipa getamiaoko hilin.
Father notyet wife baby born died

③ Zegipa getoake gizopa otoko itina.
baby born looked after grew up.

④ Mota litaoko napadoake iza, nama peletokaina.
Very quickly grew pig birds killed ime

⑤ Izelahina gizopa otoko vina.
His mother looked after went

⑥ Vena manikeko umuhilina tana.
wife identify liked beautiful got, married

⑦ Numuna aito giziki ikasina.
house separate build slept

⑧ Nuwaka nuwaka nosaneta titiki vika asina.
afternoon afternoon food brought went use to

① Long time ago, one blind woman lived with her son.

② Father died when the boy was not born yet.

③ The baby was born and mother looked after him as he gr

④ He grew up very quickly and he killed birds and pigs.

⑤ He went on to look after his mother.

⑥ He identified a beautiful wife and married her.

⑦ They built separate house and slept.

⑧ Every afternoon they brought food to her house.

Figure 3: Interlinear Text Layout: (a) the title, translated title, author, and date are written at the top; (b) the source text is written on the left page, with three-line spacing, numbering each sentence; (c) the gloss is written beneath each word (omitted if no simple gloss is possible); (d) the phrasal translation is written on the right page, and coindexed with the source.

We were able to categorise the speakers into four types based on linguistic and technical capabilities. The first category, *monolinguals*, consisted of elders with no functional knowledge of Tok Pisin, who probably have good knowledge of their oral literature but who are so culturally different that it is difficult to tap their knowledge; they are not particularly comfortable in the university setting. The second category, *village-based bilinguals*, consisted of elders with basic literacy in Tok Pisin or English, and no formal education beyond primary school. The third category, *retired professionals*, consisted of bilingual speakers with post-secondary education who have moved around the country and held various professional roles, with solid literacy in English. Finally, the fourth category, *young professionals*, consisted of bilingual speakers who are studying or are employed in town, with English literacy, computer literacy, but limited fluency in their ancestral language and almost no knowledge of oral literature.

Texts and translations were keyboarded by people working in pairs, as illustrated in Figure 4. Once it was finalised, each text and translation was printed and displayed on a wall. This served three purposes: (a) participants were publicly recognised for completing a text; (b) corrections could be marked for later editing; and (c) ideas for writing topics were shared. On the last day, we published a booklet containing all the texts.

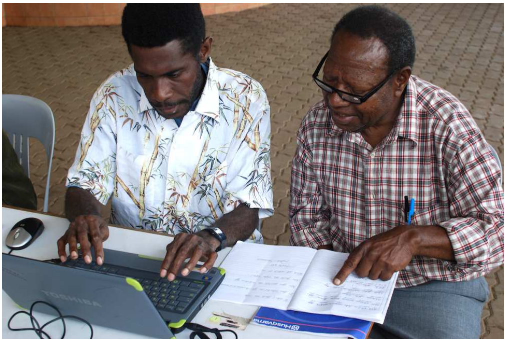


Figure 4: Interlinear Text Entry: an Adzera speaker who is a competent typist (left) enters interlinear text for the Asaro speaker (right) who dictates the words and glosses and checks that they are correctly entered. (The handwritten source text is shown in Figure 3.)

A sample of the interlinear text is shown in (1).

(1) Velaliki veena kisa ei gipala (The blind woman and her son)
Alo gozopa vena makokisa gipala isa minasina.
long time ago woman one son both lived.
A long time ago, a blind woman lived with her son.

Menipo zoliha venala zegipa getamiwoko hilina.
father not yet wife baby born died.
The father died when the boy was not yet born.

Zegipa getoake gizopa otoko itina.
baby born looked after grew up.
The baby was born and the mother looked after him as he grew.

Mota litaoko napaoake iza, nama peletoka ana.
very quickly grew pig, birds killed came.
He grew up very quickly and he killed birds and pigs.

Izelahina gizopa otoko vina.
his mother looked after went.

He went on to look after his mother.

In the two weeks of the workshop, we only managed to collect a total of 20k words of source text (16k translated) for the 15 languages. Many participants found it relatively difficult to compose directly into the written form, and so they did not produce much writing. For the languages where we had more than one speaker, there was some dialect variation and this was reflected in spelling. There was also some variation in the marking of word boundaries, and with the writing of glottal stop (apostrophe, *q*, or omitted). We lacked the time and the language-specific information required to perform morphological glossing, and this would have been quite challenging given the systems of switch reference, serial verbs, and clause chaining in many of these languages (Foley, 1986; Payne, 1997). Perhaps because of these morphological issues, word-level glossing was slower than phrase-level translation. In any case, for these reasons it proved impossible to construct useful translation models for the languages.

In order to scale up the work to generate a quantity of data that would be more useful for machine translation experiments, the following steps would be required. First, the primary textual sources should be audio recordings, and transcribed using a tool that preserves the audio alignment (for later verification) and which links wordforms to lexemes (for consistency in spelling, word breaks, and glosses). Second, the transcription and glossing software should operate in tandem with curating a shared *n*-language lexicon to speed up the process and encourage consistency across speakers, possibly using the structures described in (Baldwin et al., 2010; Abney and Bird, 2011).

5 Conclusion

Most of the world's languages will fall out of use before the world's linguists and computational linguists are able to collect sufficient data. However, we have been investigating simple methodologies and supporting software that are helping speakers of endangered languages in Papua New Guinea to produce usable documentation on their own. The primary data type is bilingual text – or interlinear glossed text – which serves the dual purpose of documenting a language and developing translation models.

Once the translation models reach an adequate level, they could be usable as the basis for post-editing work, and may speed the translation process. More importantly, system errors will draw attention to those areas of the grammar and lexicon that are not yet well represented in the data. They may prompt speakers to provide more data of the required kind, without requiring the intervention of an outside linguist. While is still difficult to imagine being able to do this work on the required scale, it represents a promising approach for shaping the effort of non-specialist language speakers in creating a documentary record of their languages while there is still time.

Acknowledgements

We are grateful to the many colleagues and students who helped run the workshop discussed in section 4. This research is sponsored by NSF 1144167 *Machine Translation for Language Preservation* and ARC 120101712 *Language Engineering in the Field*.

References

Abney, S. and Bird, S. (2010). The Human Language Project: building a universal corpus of the world's languages. In *Proceedings of the 48th Meeting of the Association for Computational Linguistics*, pages 88–97. Association for Computational Linguistics.

Abney, S. and Bird, S. (2011). Towards a data model for the Universal Corpus. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 120–127. Association for Computational Linguistics.

Baldwin, T., Pool, J., and Colowick, S. (2010). PanLex and LEXTRACT: Translating all words of all languages of the world. In *Proceedings of the 23rd International Conference on Computational Linguistics, Demonstrations Volume*, pages 37–40. Tsinghua University Press.

Bird, S. (2010). A scalable method for preserving oral literature from small languages. In *Proceedings of the 12th International Conference on Asia-Pacific Digital Libraries*, pages 5–14.

Bird, S., Chiang, D., Frowein, F., Eby, M., Hanke, F., Shelby, R., Vaswani, A., and Wan, A. (2012). Language preservation in Papua New Guinea: Report from a workshop at the University of Goroka. Unpublished.

Derbyshire, D. C. (1977). Word order universals and the existence of OVS languages. *Linguistic Inquiry*, 8:590–99.

Foda, A. and Bird, S. (2011). Normalising audio transcriptions for unwritten languages. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 527–535, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Foley, W. A. (1986). *The Papuan Languages of New Guinea*. Cambridge University Press.

Hanke, F. and Bird, S. (2012). Preserving endangered oral culture: speech annotation on mobile devices. Unpublished.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Johnson, H. and Aristar Dry, H. (2002). OLAC discourse type vocabulary. `http://www.language-archives.org/REC/discourse.html`.

Krauss, M. E. (2007). Mass language extinction and documentation: the race against time. In Miyaoka, O., Sakiyama, O., and Krauss, M. E., editors, *The Vanishing Languages of the Pacific Rim*. Oxford University Press.

Liberman, M. (2006). The problems of scale in language documentation. Plenary talk at TLSX Texas Linguistics Society 10: Computational Linguistics for Less-Studied Languages, `http://uts.cc.utexas.edu/~tls/2006tls/`.

Nakov, P. and Tiedemann, J. (2012). Combining word-level and character-level models for machine translation between closely-related languages. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 301–305. Association for Computational Linguistics.

Nettle, D. (1999). *Linguistic Diversity*. Oxford University Press.

Palmer, A., Moon, T., Baldridge, J., Erk, K., Campbell, E., and Can, T. (2010). Computational strategies for reducing annotation effort in language documentation. *Linguistic Issues in Language Technology*, 3(4):1–42.

Payne, T. E. (1997). *Describing Morphosyntax*. Cambridge University Press.

Woodbury, A. C. (2003). Defining documentary linguistics. In Austin, P., editor, *Language Documentation and Description*, volume 1, pages 35–51. London: SOAS.

Woodbury, A. C. (2010). Language documentation. In Austin, P. K. and Sallabank, J., editors, *The Cambridge Handbook of Endangered Languages*. Cambridge University Press.

Xia, F. and Lewis, W. D. (2007). Multilingual structural projection across interlinearized text. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 452–459. Association for Computational Linguistics.