LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 9 (2015), pp. 10–11
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24631

Documentary Linguistics and Computational Linguistics: A response to Brooks

Steven Bird
University of Melbourne

David Chiang
University of Notre Dame

Friedel Frowein
University of Goroka

Florian Hanke
University of Melbourne

Ashish Vaswani
University of Southern California

In mid-2012 we organized a two-week workshop in Papua New Guinea (PNG) to provide training in basic techniques and technologies for language documentation, and to gain understanding of how these technologies might be improved in the future. It was a diverse program, combining the expertise of scholars from ten institutions. It was also a diverse audience, including academics, teachers, students, archivists, translators, pastors, and farmers from across the country. Approximately twenty local languages were represented.

The central idea of Brooks' assessment of the workshop is that its computational goal was incompatible with its documentary goal. However, we would say that there was a single goal, namely, to document languages of PNG. We would particularly guard against the possible misperception that data is collected from PNG languages to fuel machine translation in general. While that would admittedly be interesting, machine translation, in this context, is not an end in itself but a means to an end, which is documentation.

Much of Brooks' commentary addresses our reliance on textual sources. We agree with his reasons, and most were already raised in our article. We had planned to include spoken language recordings among the workshop activities, using 34 voice recorders donated by Olympus, but the recorders turned out to be tied up in student projects. This was one of several logistical challenges of organizing a workshop in PNG, challenges which made the execution of the workshop turn out differently from its conception. As always, there are things we would do differently the second time, including a stronger emphasis on oral language recording. In fact, it was for this purpose that the Aikuma mobile phone app was developed (Hanke & Bird 2013, Bird et al 2014a, b). Nevertheless, texts are a form of

E-ISSN 1934-5275

documentation that some local villagers can readily produce, and several of the workshop participants came with exercise books full of hand-written texts.

Participants were unanimous that the workshop was a success. They voted with their feet, returning each day for intensive language work without expecting compensation. Both participants and presenters came away with experiences and training commensurate with their prior knowledge and skills. Those with limited western-style education learned how to gloss and translate a text using an exercise book, and how to elicit the words of a semantic domain using the Rapid Words Method. Those with computer skills gained new skills glossing with FLEx. Those with linguistics training learned about scalable language documentation workflows and new computational methods for supporting them. Those with computer science training learned about the different sources of noise that are introduced into textual data when languages lack an established orthography, and when the boundaries between adjacent languages and dialects are poorly understood.

While we have argued that computational linguistics and traditional documentary linguistics have the same goal with respect to endangered languages, the two fields offer different methodologies. Our inclusion of machine translation methods in language documentation work is not to supplant linguists, but to increase their productivity and to avoid the need for resources like treebanks and wordnets (Bird & Chiang 2012, Abney & Bird 2010). Unfortunately, there are many opportunities for misunderstandings to arise. But we maintain that both fields offer proven approaches to the analysis of language data. The urgency of the documentary challenge compels us to find effective ways to multiply our efforts.

REFERENCES

Abney, Steven & 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*, Uppsala, Sweden. 88–97.

Bird, Steven & David Chiang. 2012. Machine translation for language preservation. *Proceedings of the 24th International Conference on Computational Linguistics*, Mumbai, India. 125–133.

Bird, Steven, Lauren Gawne, Katie Gelbart & Isaac McAlister. 2014a. Collecting bilingual audio in remote indigenous villages. *Proceedings of the 25th International Conference on Computational Linguistics*, Dublin, Ireland. 1015–1024.

Bird, Steven, Florian R. Hanke, Oliver Adams & Haejoong Lee. 2014b. Aikuma: A Mobile App for Collaborative Language Documentation. *Workshop on the Use of Computational Methods in the Study of Endangered Languages*, Baltimore, USA. 1–5.

Hanke, Florian & Steven Bird. 2013. Large-scale text collection for unwritten languages. *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan. 1134–1138.

Steven Bird
sbird@unimelb.edu.au