**The Cape York lexical records of Bruce Sommer**

We report on a project which has created a digital version of lexical material on approx. 70 lan­guage varieties of Cape York, from the archival records of Bruce Sommer *[1]*. Our focus here is on methodology.

**Background & aims** Great strides have been made in preparing the lexicons of Australian lan­guages in digitally readable and accessible form, however a notable gap so far is Cape York *[2]*. Bruce Sommer deposited lexical, grammatical and textual materials on some 70 language varieties of central and southern Cape York, comprising 4,950 pages of fieldnotes and summaries, and 203 audio tapes. Ouraimwas to key in Sommer’s handwritten and printed lexical materials, as a first step in the digital representation and eventual audio time-alignment of his invaluable archive.

**Materials** Fryer Library digitised Sommer’s print materials in 2014 and tapes in 2015. We identi­fied 1,520 pages of lexical material. These wordlists range in length from 2 entries to 2635 (mean 485, median 255). Many are numbered, following the Hale–O’Grady 100-item list.

**Methods** Our work plan centred on simultaneous and collaborative data entry. Two researchers entered the same wordlist simultaneously into a Google spreadsheet, where the other’s activity is also visible. Each worker focussed on either the vernacular or English, but also provided constant checking of the other’s work, and assistance when necessary. The spreadsheet contained columns for: speaker, language, tape number, subheadings, page number, language form, notes on language form, English gloss, notes on English gloss, other text and notes on other text. Additional columns were added if wordlists become more complex: language form corrections, number, addi­tional language form columns for lists with two vernacular languages.

**Challenges** 1. Legibility of handwriting was a challenge. To improve accuracy, researchers exam­ined illegible entries together to reach agreement; if needed, other wordlists were consulted, to see if a word appeared elsewhere with a similar form. In rare cases where neither of these solutions worked, a note was entered. 2. Sommer used many abbreviations. These were gradually deciphered as our familiarity increased. 3. Some pages contained extensive corrections, annotations and/or margin notes; some had multiple languages or speakers. Extra columns were added for those docu­ments. 4. Most of the materials were in IPA. This was entered using a convenient set of *as hoc* con­ventions to enable fast data entry, and then transposed into IPA afterwards. Having two researchers dealing collabora­tively with challenges led to rapid and effective problem solving.

**Analysis** Cape York is a notoriously complex region *[3]*. Cross-linguistic datasets such as Sommer’s lexi­cons will make possible automated analyses which can detect diffuse patterns which challenge the obser­vational and memory limitations of human linguists. We present some initial examples, including auto­mated phylogenetic analysis *[4]*; network analysis *[5]*; and admixture analy­sis *[6]*. These do not replace expert manual analysis, but can increase productivity by rapidly highlighting areas deserving partic­ular attention.

**Methodological recommendations** We cannot recommend strongly enough the method of collabo­rative data entry for this kind of data, which enables quick and effective detection and correction of data entry errors. It makes the task more collaborative, and hence enjoyable.

1. Sommer, B. 2003. *Papers, 1964–2003* (item number UQFL476), Fryer Library, St Lucia.
2. Bowern, C. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Lan­guages of Australia*. Language Documentation and Conservation*. Vol 10.
3. P.Sutton (ed.) 1975. *Languages of Cape York*, Canberra:AIAS.
4. Blomberg, S.P., T. Garland & A.R. Ives. 2003. Testing for phylogenetic signal in compara­tive data: Behavioral traits are more labile. *Evolution* 57:717-45.
5. Bryant, D., & Moulton, V. 2004. Neighbor-net: an agglomerative method for the construc­tion of phylogenetic networks. *Molecular biology and evolution*,*21*(2), 255-265.
6. Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.