The AusPhon-Lexicon project: Two million normalized segments across 300 Australian languages

Round, Erich R. (University of Queensland)

We describe the logic, methods, tools and products of the AusPhon–Lexicon project, which at writing has normalized 166 Australian language varieties (1.2million segments) and aims at 290–320 varieties (~2.1million segments).

Materials: Input data is printed and electronic vocabularies, including all modern data of Bowern's(2016) CHIRILA database, plus ~100 additional varieties.

Processes: Data is scrubbed; graphemically tokenized (additional output: language grapheme token-izers), phonemicized (additional output: customizable language phonemicizers, which can then phonemicize textual material for example), and post-produced with a variety of mark-up, e.g. for duplicates, names, ideophones.

Tools & workflow: We use existing and customized tools in R, and extensive 'human computing' (Michelucci and Dickinson 2016): alternating between machine and humans, each doing what they excel at. Specific innovations include entropy-based junk filtering during scrubbing, and visualiz-ation-based error-checking at multiple points (additional output: language data visualizations).

Logic of the process: While it may seem that the task is merely to make an 'electronic version' of the lexicons of a continent, we argue that the task is deeply more complicated. Effective normalization across 300 languages has demanded meta-analysis of disparate analytical traditions spanning half a century, in order to understand why previous analyses are how they are, and therefore what implications will accompany decisions to normalize one way or another. We discuss in particular the bleeding of phonotactic, areal–historical, and theory-based concerns into the analysis of inventories, and the responses/solutions we've identified, which where possible enhance rather than degrade information relative to the un-normalized original.

Resulting representations: The most salient outputs are normalized segmental strings. Yet to be meaningful, these are networked to annotations that link decisions back to original sources, and record all steps in the ingest–scrub–tokenize–phonemicize trail. We consider the result to be not a 'database' containing answers to pre-existing questions, but a 'data warehouse'(Cooper 2014) to be engaged with through queries. Thus, whereas a traditional typological project might produce after much labour a single database on consonant clusters (Hamilton 1996), we can create a new database — on clusters, vowel harmony, position neutralization, OCP effects, etc. — with a simple query.

Data querying: To this end, we use tools for customizable, on-the-fly featurization; a 'phonex' extended regular expression language; and n-gram extraction to create character-based datasets for use, e.g. in phylogenetic phono¬tactic research(Macklin-Cordes and Round 2015). A layer of 'archi-phonemicization' or 'Firthianization' scripts handle positional neutralization, which can otherwise lead to misleading disparities even in an otherwise well-normalized segmental representation.

Constraints: Sister projects were run to gauge the prospects of integrating sub-phonemic information. We find the mining of information on allophony from descriptive grammars to be largely infeasible unfortunately, given results from both (i) surveys of what gets reported, and (ii) ground-truthing such reports against instrumental phonetic study. Nevertheless, there are exceptions. Highly salient features such as nasal/lateral pre-stopping and major allophonic place-of-articulation differences may be viable for integration (as an overlay) into this kind of segmental dataset.

Quantitative analysis: Finally, we illustrate the extraordinary potential of such data for linguistic research, showing that the Australian dental/palatal contrast, long thought to be thoroughly areally distributed(Dixon 1970), in fact contains ample phylogenetic signal throughout Pama-Nyungan.

References

Bowern, Claire. 2016. Chirila: Contemporary and Historical Resources for the Indigenous Languages of Australia. *Language Documentation & Conservation* 10: 1–44.

Cooper, Doug. 2014. *Logistics of the Asia-Pacific Linguistic Data Warehouse*. MPI Nijmegen.

Dixon, R. M. W. 1970. Proto-Australian laminals. *Oceanic Linguistics* 9: 79–103.
Hamilton, Philip J. 1996. *Phonetic constraints and markedness in the phonotactics of Australian Aboriginal languages*. University of Toronto [PhD dissertation].
Macklin-Cordes, Jayden L. and Erich R. Round. 2015. High-Definition Phonotactics Reflect Linguistic Pasts. In *Proceedings of the 6th Conference on Quantitative Investigations in Theoretical Linguistics*, Wahle Johannes, Marisa Kollner, Harald Baayen, Gerhard Jäger, and Tinaka Baayen-Oudshoorn (eds.). Tübingen, doi:10.15496/publikation-8609.
Michelucci, Pietro and Janis L.Dickinson. 2016. The power of crowds. *Science* 351: 32–33.