

(1) Paper title:

'Sound[s/ed] like ...?' Approximate Phonetic Search in the Mon-Khmer Languages Project

(2) Sub-field

computational linguistics, phonology

(3) Name(s) of author(s)

Doug Cooper

(4) Affiliation(s) of author(s)

Center for Research in Computational Linguistics

(5) Email address for each author

doug.cooper.thailand@gmail.com

‘Sound[s/ed] like ...?’ Approximate Phonetic Search in the Mon-Khmer Languages Project

The wealth of data made available by modern language documentation and preservation projects opens the door to comparative linguistic analysis on a grand scale. Yet even with (some would say ‘despite’) the aid of computers, coming to terms with this largesse has been an unexpected struggle. We must on occasion ruefully admit that, however tedious, the pencil-and-paper methods employed by our predecessors helped them develop the intimate knowledge of form and content necessary to navigate and classify tens or hundreds of thousands of citations.

The *Mon-Khmer Languages Project* confronts this problem of overabundance in aiding discovery of the content of our language database. It includes data from a growing number of roughly 150 languages in the dozen or so branches of Mon-Khmer, the major component of the Austroasiatic family, as well as proto-language reconstructions whenever available.

Our difficulties are partially due to the raw material: to the plethora of traditions, skills, and standards of the five full generations of scholars whose work we record, their mix of phonemic and phonetic notation, and the bracketed ambiguous segments so often required for provisional reconstruction.

But most problematic are the uses to which linguists wish to put these data. Beginning with renditions whose diversity brings Shaw’s *fish = ghoti* example to mind, we must respect the endless capacity of human speakers to introduce variation in articulation as we seek the cognate daughters of an elusive Mon-Khmer parent, or compile and analyze the regular patterns of correspondence that are the grail of the comparative method.

Our solution focuses on these three points – *notation*, *realization*, and *historical variation* – in implementing approximate phonetic search. Notational inconsistency is managed by building in specific equivalence sets (e.g. **g/g** or **ŋ /hm**), and by expanding bracketed reconstructions as part of the initial storage process (e.g. a head originally listed as ***b[h]raap** is implicitly indexed as **braap** and **bhraap**). Variations in phonation, vowel length and the like are managed with approximate match letters and special-purpose search controls.

The most interesting innovation is a graphical interface, modeled after the standard IPA manner-of-articulation chart, that lets the user incorporate extensible pre-built *allophone sets* into search queries. These are carefully designed to reflect predictable historical and areal variation, and include many natural classes of consonants and vowels that characterize Mon-Khmer language variation. We will describe the design of the system and demonstrate its operation.