Perspectives from cross-linguistic databases and the quantification of phonetic difference

Heggarty, Paul (Max Planck Institute for the Science of Human History in Jena)

This paper ranges over many of the topics for this workshop, offering two particular perspectives.

The first perspective emerges from a cross-linguistic phonetic database of (currently) c. 50,000 individual word recordings, and corresponding phonetic transcriptions, of cognate lists in various language families worldwide. It so far covers some 400 different language, dialectal and accent varieties, many of them poorly documented and highly endangered. The data can be accessed, searched, filtered and customised through a dedicated database 'explorer' website, with search functionality that aspires to be as phonologically and phonetically aware and powerful as possible.

Necessarily, this database and its website have come up against many of the issues that surround the selection and development of standards for phonological and phonetic databases. Transcriptions have necessarily been contributed by different expert transcribers for the various families, and views differ on the pros and cons of narrower or closer transcriptions, especially in the absence of any published (or agreed) phonological analyses of most of the language varieties covered. There is a need for a workable compromise across different priorities in collecting, exploring, analysing and quantifying across large cross-linguistic databases in phonetics. It is also a challenge to reconcile the priorities of a diverse range of end-users: from computational linguists, hungry for off-the-peg data to try out their methods, to traditional qualitative historical and comparative linguists, to outreach audiences among the general public and speakers of the languages themselves, including the objectives of documentation, dissemination and revitalisation. And standards are needed not just in transcription but also in search functionality, including how best to integrate into regular expression syntax a range of phonologically desiderata: phonological symbols, wildcards and features, reconstructed proto-forms, (Unicode) IPA, (competing) orthographies, and so on.

The second perspective on the topics of this meeting comes from one particular research objective that these phonetic cognate databases are intended to serve: the quantification of phonetic difference (or more precisely, of net divergence in phonetics since a common ancestor). This is approached certainly not by some off-the-shelf computational method, grossly applied to pseudo-phonetic data as abstract 'strings'. Rather, I outline some key relevant features of a dedicated and highly complex algorithm, explicitly informed by the architecture of phonetic classification, and by expert comparative and historical linguistic knowledge. This means, not unexpectedly, that although the algorithm does make crucial practical use of the concept of the segment, it also insists on the need to go beyond it, if we are to aspire to a reasonably precise and meaningful expression in numbers of the actual linguistic significance of given phonetic differences (synchronically and diachronically). The segment idealisation fails, and a more realistic, sophisticated representation is needed, not just in cases that involve suprasegmentals, but any of a host of the commonest sound changes that involve quantity, timing, secondary articulations, compensatory lengthening, syllabicity, and changes across the vowel/consonant transition. Finally, the need for cross-linguistic consistency also entails that the search for standards cannot escape the fundamental questions that surround the phonetics/phonology distinction.