Modelling sound change with the help of multi-tiered sequence representations

It may seem to be a straightforward analogy to compare sound sequences with the sequences of (Wheeler and Whiteley 2015), permitting a transfer of models and methods from the computationally more developed discipline of evolutionary biology to the relatively young discipline of computational historical linguistics. A closer look at sequences in linguistics and biology, however, quickly reveals striking differences between the status of sequences in each disciplines. Biological sequences are drawn from character sets (*alphabets*) which are *limited* in size, such as the 20 amino acids of proteins, and *universal*, recurring in all organisms. This does not hold true for human languages, in which character sets are language-specific and vary greatly in size, with characters being not discrete entities, but abstract representations of multidimensional and continuous information. In order to deal with the difficulty of modelling sound sequences, linguists have elaborated a large variety of feature systems (Jakobson et al. 1951; Chomsky and Halle 1968; Ladefoged and Maddieson 1996), none of which has become standard.

Given the complexity of sound change as a *systemic process* that can be conditioned by various factors, ranging from immediate phonetic context, via suprasegmental conditions, up to the interaction with morphology (Blevins 2004), it is obvious that a proper modelling of sound change phenomena cannot be achieved when relying on pure biological methods that can only handle *replacement*, *loss* or *gain* of discrete segments (Durbin et al. 2002, 13). While immediate phonetic context could be handled with help of *bigram* models (Bouchard-Côté et al. 2013), no satisfying ways to handle accent, tone, or other suprasegmental properties have yet been proposed in computational approaches.

Building on earlier attempts to handle phonetic context in phonetic alignment and linguistic reconstruction (List and Chacon 2015), we propose a formal representation of sound sequences which allows for an extremely flexible handling of phonetic context, including suprasegmental information. The basic idea is to represent a sequence with the help of multiple *tiers*, which can be used to represent contextual and phonetic aspects relevant to a given analysis. A tier is thus a layer of simple or complex information defined for each segment of a given sound sequence. Being stored as sequences aligned to the original sound sequence, they can be related to mathematical embeddings by vectors.

Complex sound environments can be represented by an arbitrary number of tiers, as demonstrated in *Table 1*. The advantage is that these need not be tied to any given theory or system: while tiers are designed for studying interactions, they are themselves independent, and competing theories or feature systems can be modelled simultaneously, if desired.

In this paper, we introduce a preliminary framework for multi-tiered sequence representation, showing how it can be used to pre-process linguistic data, or to identify potentially miscoded cognates, undetected borrowings, erroneous alignments, or peculiar sound changes.

Tier	**Description**	**Alignment**				
SOURCE	source sounds	s	w	e	r	d
CV / _X	previous sound C or V	#	C	C	V	C
CV / X_	following sound C or V	C	V	C	C	$
SOUND CLASS / _X	previous sound class	#	S	W	V	R
SOUND CLASS / X_	following sound class in source	W	V	R	T	$
STRESS	stress in source	1	1	1	1	1
TARGET	target sounds	ʃ	v	eː	r	t

Table 1: Representing differences between Proto-Germanic **swerd-* and German Schwert with help of multi-tiered sequences. By aligning the proto-sequence with its descendant, we can transparently encode additional context information, not only preceding and following context, but also suprasegmental aspects, such as stress.

References

Blevins, J. 2004. *Evolutionary Phonology: The Emergence of Sound Patterns*. Cambridge University Press.

Bouchard-Côté, Alexandre, David Hall, Thomas L. Griffiths, and Dan Klein. 2013. "Automated Reconstruction of Ancient Languages Using Probabilistic Models of Sound Change." *Proceedings of the National Academy of Sciences of the United States of America* 110 (11): 4224–9.

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English*. New York; Evanston; London: Harper; Row.

Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchinson. (1998) 2002. *Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids*. 7th ed. Cambridge: Cambridge University Press.

Firth, John Rupert. 1948. "Sounds and Peosodies." *Transactions of the Philological Society* 47 (1). Wiley Online Library: 127–52.

Harris, Zellig Sabbettai. 1963. "Structural Linguistics." Chicago University Press.

Hartman, Lee. 2003. "Phono (Version 4.0): Software for Modeling Regular Historical Sound Change." Santiago de Cuba.

Jakobson, Roman, C Gunnar Fant, and Morris Halle. 1951. "Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates."

Ladefoged, P., and I. Maddieson. 1996. *The Sounds of the World's Languages*. Phonological Theory. Wiley.

List, Johann-Mattis. 2014. *Sequence Comparison in Historical Linguistics*. Düsseldorf: Düsseldorf University Press.

List, Johann-Mattis, and Thiago Chacon. 2015. "Towards a Cross-Linguistic Database for Historical Phonology? A Proposal for a Machine Readable Modeling of Phonetic Context." Leiden.

Wheeler, W. C., and Peter M. Whiteley. 2015. "Historical Linguistics as a Sequence Optimization Problem: The Evolution and Biogeography of Uto-Aztecan Languages." *Cladistics* 31 (2): 113–25.

Word Count: 490