Unsupervised acquisition of allophonic rules: the lexicon strikes back

Luc Boruta
luc.boruta@inria.fr

ALPAGE (INRIA/P7) & LSCP (EHESS/ENS/CNRS)
'FdV' interdisciplinary grad. school (P5/P7/FBS)

Keywords: phonology, language acquisition, computational linguistics, data streams.

Phonological rules relate surface phonetic forms to underlying phonological/lexical forms. Infants must learn such rules to bootstrap the acquisition of an accurate lexicon. Consider the following devoicing rule that applies in French:

$$/\mathrm{r}/ \;\rightarrow\; \begin{cases} [\chi] & \text{before a voiceless consonant} \\ [\mathrm{ʁ}] & \text{otherwise} \end{cases}$$

Its application creates two contextual variants for /kanar/ (*canard*, 'duck'): [kanaʁ‿ʒon] (*canard jaune*, 'yellow duck') and [kanaχ‿flotɑ̃] (*canard flottant*, 'floating duck'). Before knowing the rule, children have to stock both [kanaʁ] and [kanaχ] in their emerging lexicon. After the acquisition, they are able to undo allophonic variation and construct a single lexical entry: /kanar/. We are interested in procedures by which infants could extract phonemes and allophonic rules from speech, and in the interaction between lexical and phonological acquisition.

Following the structuralist literature on the discovery of phonological units, we combine distributional, lexical and acoustic cues extracted from speech signal and phonetic transcriptions. We will present two major extensions to Peperkamp *et al.*'s model:

- though ubiquitous in linguistics, minimal pairs are not statistically robust and can be efficiently replaced by an information-theoretic reformulation of functional load;
- using clustering algorithms, rather than statistical filters, allows us to explore the interactions between different cues.

As infants do not wait, in Brent's words, "until the corpus of all utterances they will ever hear becomes available", we will argue that online learning is a sound desideratum for any model of human language processing. As such, we will discuss the relevance of data stream algorithms as models of early language acquisition.