Consonant representations aid in learning segmentation and phonology for Arabic but not English

The problem of segmenting speech into words has received much attention in the computational modeling literature (Brent & Cartwright, 1996; Goldwater, Griffiths, & Johnson, 2009). Yet it is not clear to what extent the segmentation mechanism differs across languages, nor is it well understood to what extent the segmented proto-lexicon aids in learning phonological patterns. For example, do English learners track the same distributions as Arabic learners, whose language is built around nonconcatenative morphology of consonantal roots and prosodic templates (McCarthy, 1979)?

The current study explores the hypothesis that the representations on which the segmentation mechanism operates are language-specific. Specifically, we hypothesize that acquisition of Arabic is facilitated by dividing the input stream into separate consonant and vowel tiers. It is known that learners can track consonant co-occurrence probabilities across intervening vowels (Newport & Aslin, 2004; Bonatti et al., 2005), but it is unknown to what extent this ability helps them in making progress in natural language acquisition.

We conducted a number of simulations that aimed to model an infant's behavior and which make predictions for future acquisition studies. By training and testing computational models on different representations for multiple languages, we aim to quantify how useful consonant representations are for segmentation in different languages, and for phonological learning from the resulting segmentations.

In Experiment 1, subsets of parsed Standard Arabic newswire (Graff, 2003) and Emirati Arabic infant-directed speech (Ntelitheos and Idrissi, 2015) were compared to the English subset of CHILDES used by Goldwater et al. (2009). For each dataset we constructed a consonant-only representation and a "full" (consonant + vowel) representation. The segmentation algorithm from Goldwater et al. (2009) was run on each of these inputs. This algorithm performs Bayesian inference on the input, trying to maximize the probability of a hypothesized segmentation of the corpus given the observed data. Table 1 shows that the consonant-only representation aids in detecting word boundaries in Arabic, but hampers segmentation of English, in accordance with our hypothesis.

In Experiment 2, we tested whether the segmented proto-lexicon supports phonological generalizations that the infant would need to learn. Specifically, we asked to what extent the learned Arabic lexicon provides statistical evidence for a restriction against homorganic consonant pairs ("OCP-Place"). We calculated observed/expected ratios for labial, coronal, and dorsal pairs (Frisch, Pierrehumbert, & Broe, 2004) in four different segmentations: unsegmented utterances, full-representation segmentation, consonant-only segmentation, and the correct segmentation. As Table 2 shows, the consonant-only segmentation performs closest to the gold standard.

Our results suggest that for a child learning a Semitic language, separating consonants from vowels is beneficial for segmentation and phonological learning. We speculate that distinguishing consonants from vowels might further aid in bootstrapping semantic information and morphosyntactic rules: the learner will first assign consonantal chunks to objects, and then fill in the grammar with vowels (Nespor, Peña, & Mehler, 2003). These results provide a first computational test of this hypothesis as applied to natural language data, supporting a view in which acquisition relies on learning mechanisms that operate on language-specific representations.

(Word count: 499)

Tables

Representation	Arabic (newswire)			Arabic (IDS)			English (IDS)		
	Prec	Recall	F-score	Prec	Recall	F-score	Prec	Recall	F-score
Full	31.7	74.3	44.4	34.3	**94.2**	50.2	**90.1**	**64.0**	**75.0**
C-only	**55.1**	**84.2**	**66.7**	**47.1**	86.3	**60.9**	83.9	52.0	64.2

Table 1. Precision, recall and F-score for word boundaries in three datasets (best scores in boldface).

Representation	Labial-Labial	Coronal-Coronal	Dorsal-Dorsal
Continuous	0.318 (+0.140)	0.265 (-0.074)	0.258 (+0.036)
Full	0.125 (-0.053)	0.298 (-0.041)	0.240 (+0.018)
C-only	0.166 **(-0.012)**	0.342 **(+0.003)**	0.233 **(+0.011)**
Gold standard	0.178	0.339	0.222

Table 2. O/E ratios (observed/expected, a statistical measure of under- or over-attestation) by category for non-identical biphones in a representative Arabic newswire dataset (difference from gold standard in parentheses, smallest difference in boldface).

References

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: the role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*(6), 451-459.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1), 93-125.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, *22*(1), 179-228.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21-54.

Graff, D. (2003). *Arabic Gigaword Corpus (*Linguistic Data Consortium, Philadelphia, PA.).

McCarthy, J. J. (1979). *Formal problems in semitic phonology and morphology* (Doctoral dissertation, MIT Cambridge, Mass.).

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue e linguaggio*, *2*(2), 203-230.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive psychology*, *48*(2), 127-162.

Ntelitheos, Dimitrios and Ali Idrissi. 2015. Language Growth in Child Emirati Arabic. *29th Annual Symposium on Arabic Linguistics*. The University of Wisconsin–Milwaukee, April 9-11, 2015.