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

Itamar Kastner and Frans Adriaans  (itamar@nyu.edu, frans.adriaans@nyu.edu)

New York University

INTRODUCTION

- **Word segmentation** mechanism: differs cross-linguistically?
  - Do English and Arabic learners track the same distributions?
- **HYPOTHESIS**: acquiring Arabic is facilitated by dividing the input into consonant and vowel tiers. Identical learning model.
- **Learners track consonant** co-occurrence probabilities. (Newport and Aslin 2004; Bonatti et al. 2005)
- Does this help make progress in natural language acquisition?
- Our computational models quantify if C-representations are:
  - Useful for segmenting different languages.
  - Useful for subsequent phonological learning.

SEMITIC MORPHOLOGY

- ✂ qabla: ‘greeted’ qabila: ‘accepted’
- ✂ taqablab: ‘tribe’ qabila: ‘prayer direction’
- ✂ Nine verbal templates in Modern Standard Arabic.

DATA AND MODEL

- Arabic: parsed newswire. (Graff 2003; Pasha et al. 2014)
- Subset of Emirati CDSS. (Ntelitheos and Idrissi 2015)
- English: CHILDES subset. (Bernstein-Ratner 1987; Goldwater et al. 2009)

Adults and children distinguish consonants from vowels in artificial grammars (Newport and Aslin 2004; Bonatti et al. 2005; Keidel et al. 2007; Hochmann et al. 2011) and acquisition (Werker and Teen 1984; Polka and Werker 1994).

Divide input into consonants and vowels.

Unigram Segmentation Model (Goldwater et al. 2009)

- Language model: performs Bayesian inference on the input.
- Maximize hypothesized segmentation given the data.
  - yawanntoseethеbook
  - yawa.nt.o see.th.ebook
  - ya.want.to see.the.book

ywantarmybook ➞ ya.want.my.book

State of the art model, model-independent results.

**RESULTS**

C-only representation aids Arabic performance, hurts English.

Simulation #1: Segmentation

Two inputs for each language. Evaluate word boundaries.
- Full representation: yawanntoseethеbook
- C-only representation: ywnntsthbk

Simulation #2: Phonology

Does the segmented proto-lexicon support the learning of Arabic phonological patterns? ⇒ OCP

Proto-lexicon aids further acquisition. (Phillips and Pearl 2015)

- Phonotactics learned within the first year (Jusczyk et al. 1994).
- Semitic restriction on homorganic consonant pairs (OCP-Place).
  - *badam: strongly under-represented
  - madad: possible, under-represented
  - tasabah: possible, under-represented

Four segmentations of the Arabic data:
- The result of C-only.
- The result of Full.
- An unsegmented baseline.
- Correct segmentation (gold standard).

Calculate O/E for non-identical labials, coronals and dorsals.

OCP can be learned from O/E in the lexicon (Frisch et al. 2004).

**RESULTS**

C-only is closest to gold standard (horizontal line).

**SELECTED REFERENCES**