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

The input stream: **linear**. Semitic morphology: **non-linear**.

istaqbal	'greeted'	*qabil*	'received'	*taqabbal*	'accepted'
qabla	'before'	*qabiːla*	'tribe'	*qibla*	'prayer direction'

- ➤ **Nine** verbal templates in Modern Standard Arabic.
- ➤ Distinct morpho-syntax (McCarthy 1981; Ussishkin 2005; Kastner in prep).

Data and Model

- ➤ Arabic: parsed newswire. (Graff 2003; Pasha et al. 2014)
- ➤ Subset of Emirati CDS. (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 Tees 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.
  - yawanttoseethebook
  - yawa.ntt.oseeth.eb.ook
  - ya.want.to.see.the.book

book see the
to want ya

- ➤ yawantmybook ⇒ ya.want.my.book
- ➤ State of the art model, model-independent results.

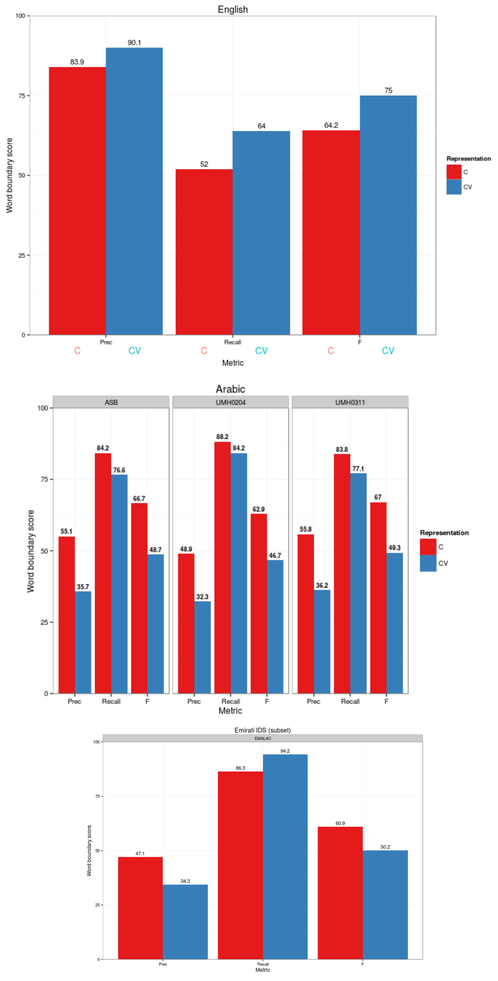
Simulation #1: Segmentation

Two inputs for each language. Evaluate word boundaries.

- **Full** representation: yawanttoseethebook
- **C-only** representation: ywnttsthbk

Results

✓ C-only representation **aids Arabic** performance, hurts English.



Simulation #2: Phonology

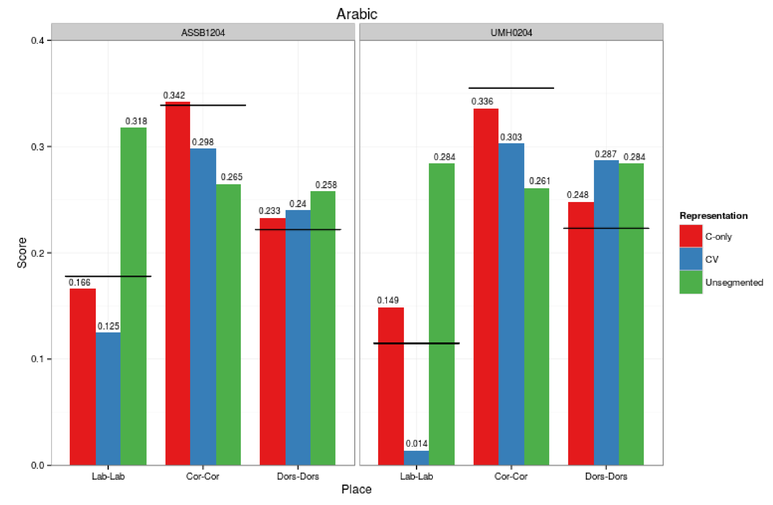
Does the segmented proto-lexicon support the learning of Arabic **phonological patterns**? ⇒ OCP

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**). (Greenberg 1950; McCarthy 1989; Berent and Shimron 1997; Frisch et al. 2004)
  - **dadam*: strongly under-represented
  - *madad*: possible, under-represented
  - *tasaba*: 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

Goldwater, Griffiths & Johnson (2009). A Bayesian framework for word segmentation. *Cognition* 112:21–45. **Hochmann, Benavides-Varela, Nespor & Mehler (2011).** Consonants and vowels. *Developmental Science* 14:1445–1458. **Newport (1990).** Maturational constraints on language learning. *Cognitive Science* 14:11–28. **Ntelitheos & Idrissi (2015).** Language Growth in Child Emirati Arabic. *29th Annual Symposium on Arabic Linguistics*. **Phillips and Pearl (2015).** Evaluating language acquisition strategies: A cross-linguistic look at early segmentation. *Ms., UC Irvine*.

Conclusions

➤ **In learning Semitic, separating consonants from vowels is beneficial.** ➤ Provided **explicit model** of how the learner attends to Cs over Vs. ➤ **Evaluation methods** for such a model. ➤ **Consistent segmentation algorithm crosslinguistically** with different representations for different languages. ➤ **"Less-is-more"** result: withholding information helps the learner (Newport 1990; Phillips and Pearl 2012). ➤ **Learner might** first assign consonantal chunks to objects, then fill in the grammar with vowels (Nespor et al. 2003). ➤ **First computational test** of this hypothesis on natural language data. ➤ **How does the learner know which hypothesis to prefer?**

BUCLD 40, November 2015