Linguistic constraints on statistical word segmentation: the role of consonants in Arabic and English

Itamar Kastner (Humboldt-Universität zu Berlin)
itamar@itamarkast.net
Frans Adriaans (Universiteit Utrecht)
f.w.adriaans@uu.nl
November 2016

Abstract

Statistical learning is often taken to lie at the heart of many cognitive tasks, including the acquisition of language. One particular task in which probabilistic models have achieved considerable success is the segmentation of speech into words. However, these models have mostly been tested against English data, and as a result little is known about how a statistical learning mechanism copes with input regularities that arise from the structural properties of different languages. The current study focuses on statistical word segmentation in Arabic, a Semitic language in which words are built around consonantal roots. We hypothesize that segmentation in such languages is facilitated by tracking consonant distributions independently from intervening vowels. Previous studies have shown that human learners can track consonant probabilities across intervening vowels in artificial languages, but it is unknown to what extent this ability would be beneficial in the segmentation of natural language. We assessed the performance of a Bayesian segmentation model on English and Arabic, comparing consonant-only representations with full representations. In addition, we examined to what extent structurally different proto-lexicons reflect adult language. The results suggest that for a child learning a Semitic language, separating consonants from vowels is beneficial for segmentation. These findings indicate that probabilistic models require appropriate linguistic representations in order to effectively meet the challenges of language acquisition.

Keywords: statistical learning; word segmentation; Arabic; language acquisition

1 Introduction

One of the most pressing cognitive tasks that infants face in the first year of life is that of word segmentation: dividing the continuous speech stream into individual words. How can infants accomplish this task? Experimental studies have shown that infants use a variety of statistical and linguistic cues to detect word boundaries in continuous speech, including transitional probabilities (e.g., Saffran, Aslin, and Newport, 1996), metrical cues (Jusczyk, Houston, and Newsome, 1999), coarticulation cues (Johnson and Jusczyk, 2001) and phonotactic cues (Mattys and Jusczyk, 2001). Computational models have subsequently been used to assess segmentation strategies against natural language corpora. These models typically induce a segmentation of the input via some form of statistical learning, either through the learning of a lexicon based on distributional regularities (e.g., Brent and Cartwright, 1996; Goldwater, Griffiths, and Johnson, 2009; Lignos and Yang, 2010), through the computation of segment or syllable bigram probabilities (Cairns, Shillcock, Chater, and Levy, 1997; Daland and Pierrehumbert, 2011; Swingley, 2005), or through a combination of statistical learning and phonological generalization (Adriaans and Kager, 2010).
While the importance of probabilistic models for our understanding of language acquisition has been established, most models have only been tested against English data, and little is known about how probabilistic models cope with the structural properties of different languages. That is, different languages have different phonologies, different grammars and different surface structures, resulting in different statistical regularities in the input. Are probabilistic models able to find relevant units (in this case, words) in each language, using the same learning strategy? Or do segmentation strategies differ across different languages? Our understanding of cognitive processes pertaining to language stands to benefit from testing models on languages that appear to be very different from each other. Cross-linguistic comparisons allow us to test a cognitive strategy more broadly, avoiding the risk of modeling the idiosyncrasies of one particular language, such as English.

Recent studies have started to examine whether probabilistic models generalize across datasets in different languages. Phillips and Pearl (2015a) evaluated the performance of two segmentation strategies on a number of different languages: English, German, Hungarian, Italian, Japanese, Persian and Spanish. The segmentation strategies considered were the Bayesian model of Goldwater et al. (2009) and the subtractive model of Lignos and Yang (2010). The study showed that the Bayesian strategy performs fairly well across languages, whereas the subtractive model is particularly successful at segmenting English and German. However, the study doesn’t directly test the effect that a particular phonological or morphological characteristic of a language has on word segmentation in that language.

The current study provides an investigation of statistical word segmentation in Arabic, a language whose morphology and phonology is fundamentally different from English. Through a comparison of word segmentation in both Arabic and English, we assess how statistical learning might interact with linguistic structure in order to obtain the most effective segmentation performance in a language. If two languages differ in some fundamental aspect of linguistic structure, then there are two immediate possibilities with regards to the learning mechanism. One possibility is that different computations are performed by learners of the two languages, which would mean that learners of different languages have different learning mechanisms. The alternative hypothesis, explored here, is that the same computations are performed by learners of different languages, but that the learning mechanism operates on different input representations.

Arabic provides an interesting test case for this hypothesis, since it is a Semitic language in which words are formed via a non-concatenative morphology of consonantal roots and vocalic patterns. Consonants and vowels thus play different roles in Arabic word formation; the resulting non-linear structure poses a challenge for models of word segmentation, which assume that the input can be linearly decomposed into individual elements, specifically morphemes. (See the Appendix for details on non-concatenative morphology in Arabic.) We predict that if the learner divides the input into separate phonological representations (“tiers”) of consonants and vowels, acquisition of Arabic, but not English, would be facilitated. The learner of Arabic may concentrate on the consonants at early stages of acquisition, while the learner of English will attune to both consonants and vowels. The segmentation algorithm itself remains identical across languages; only the representation on which it operates changes.

Interestingly, evidence from artificial language learning experiments indicates that human learners are indeed able to separate consonants from vowels, rendering our hypothesis cognitively plausible. One study by Newport and Aslin (2004) found that adult learners are able to track consonant-to-consonant transitional probabilities, as well as vowel-to-vowel transitional probabilities, in a continuous stream of CV syllables. Since low transitional probabilities between segments often indicate word boundaries (Johnson and Jusczyk, 2001; Saffran et al., 1996; Thiessen and Saffran, 2003), the ability to track such probabilities in the speech stream could allow learners to make considerable progress in word segmentation. Importantly, Newport and Aslin’s study provides evidence that co-occurrence probabilities can be learned when segments are not immediately adjacent, but are separated by intervening vowels or consonants in the speech stream.

In a similar study, Bonatti, Peña, Nespor, and Mehler (2005) found that learners used consonant
probabilities but not vowel probabilities for segmentation. This is thought to be due to a functional
distinction between consonants and vowels, namely that consonants are used for word identification
while vowels carry information about syntax (Nespor, Peña, and Mehler, 2003). These findings
indicate that learners might in fact pick up consonantal roots from the speech stream, rather than
complete words consisting of consonants and vowels. In addition, consonants and vowels are learned
by infants at different ages (Polka and Werker, 1994; Werker and Tees, 1984), and possibly have
different roles in early language acquisition (Hochmann, Benavides-Varela, Nespor, and Mehler, 2011).
The emerging picture is one in which consonants contain lexical information and vowels contain
grammatical information (but see Keidel, Jenison, Kluender, and Seidenberg, 2007).

While these studies provide evidence that learners can track consonant co-occurrence probabilities
across intervening vowels in artificial languages, it is unknown to what extent this ability would
help them to solve the word segmentation problem in natural languages. Using the Bayesian model of
Goldwater et al. (2009) as the current state of the art, we present a series of simulations that evaluates
the role of consonantal tiers on statistical word segmentation in Arabic and English. Experiment
1 tests the learner’s performance on a standard segmentation task. The segmentation algorithm
postulates word boundaries in the input and these are evaluated using a number of metrics. The
models are compared to a baseline model trained on English data, where there are no a priori reasons
to expect a beneficial effect of a consonantal tier. Each segmentation of the input also results in
a “proto-lexicon” of hypothesized word forms in the language. In Experiment 2, we asked to what
extent structurally different proto-lexicons can assist in further stages of language acquisition. Here
we examined to what extent each proto-lexicon reflects basic phonological patterns in Arabic. Each
experiment begins by testing the model on a corpus of Modern Standard Arabic, representing an
approximation of the adult grammar, followed by a corpus of child-directed speech of the Emirati
Arabic dialect.

2 Experiment 1: Word segmentation in Arabic and English

Experiment 1 asked to what extent a language-specific representation facilitates the task of segmenting
the input stream into individual words. Datasets were contrasted in two languages, English and
Arabic. In each language the same segmentation model was trained and tested on two different
linguistic representations of the same corpus: a full representation (containing both consonants and
vowels) and a consonant-only representation. Given the different word formation systems of the
two languages, we predicted that the same probabilistic model will show improved segmentation
performance when trained on consonants in Arabic, but not in English. We contrast English child-
directed speech (CDS) with two different corpora of Arabic.

2.1 Methods

2.1.1 Data

The dataset for English was the subset of CHILDES used in previous segmentation work (Bernstein-
Ratner, 1987; Goldwater et al., 2009). Only data uttered by the caregiver were used. Statistics for
this dataset are given in Table 1 under “English”.

Since large phonemically transcribed corpora of spoken Arabic are, to our knowledge, unavailable,
we aimed to obtain a large representative sample of Arabic by using subsets of Gigaword (Graff,
2003), a newswire corpus. The Arabic Newswire corpus derived from Gigaword is made of seven
smaller corpora, each corresponding to a set of news articles from a given online press agency in a
given month. The first 1,200 utterances of each dataset were sampled and concatenated to create
one combined dataset. Basic cleanup and preprocessing ensured all subsets were in the same format,
resulting in a combined corpus of 7,914 utterances as can be seen in Table 1 under “Arabic Newswire”.

3
Combining datasets from different news agencies into one corpus is meant to provide a test on different inputs. The experiments described below were also run on the individual datasets separately, with no qualitative difference in results.

Gigaword is a text-only corpus. Importantly, Arabic has a close grapheme-phoneme correspondence, which warrants the use of text corpora. Documents are available in orthography, which was automatically parsed using MADAMIRA, a state-of-the-art morphological parser for Arabic (Pasha, Al-Badrashiny, Diab, El Kholy, Eskander, Habash, Pooleery, Rambow, and Roth, 2014). This tool provides a morphological parse and a phonemic transcription which have proven useful in other work on Arabic (Gwilliams and Marantz, 2015). All results reported in this paper are based on phonemic representations, not orthography.

Since newswire utterances are longer than CDS utterances, only the first 20 words in each utterance were used. This self-imposed limitation does not follow from any assumptions about the segmentation process itself and does not bias the experiment in either direction; we have also experimented with other ways of making the two corpora more comparable, including matching for utterance length. The results were qualitatively similar and will not be reported here.

The third corpus used in our experiments was the EMALAC corpus of Emirati Arabic child-directed speech (Ntelitheos and Idrissi, 2015). The Emirati dialect of Arabic is the vernacular used in everyday interactions in the Gulf; like other dialects of Arabic, it is different than Modern Standard Arabic in a number of respects, including phonology, lexicon and syntax. To the best of our knowledge, this corpus is the most recent carefully curated corpus of child-directed speech in Arabic. Statistics for this corpus are given in Table 1 under “Emirati CDS”. It can be seen from the table that the English corpus is slightly larger when going by total utterances. The number of words per utterance is remarkably similar across the two corpora, though the Arabic words are longer.

Testing our model on newswire sets up a comparison between the input received by infants and the language used by adults, while using CDS data allows us to mimic the input that the actual child receives in English and Arabic. The English CDS corpus has been used in previous studies of segmentation (Brent and Cartwright, 1996; Goldwater et al., 2009; Phillips and Pearl, 2015a); the similar utility of the Emirati CDS corpus for work on segmentation is highlighted here for the first time.

For each of these datasets two “representations” were constructed: one “full” consisting of consonants and vowels and one “C-only” consisting only of consonants. Table 2 gives a number of examples.

Each unsegmented representation was fed to the segmentation model in turn (two representations per corpus). Performance was compared between representations. The following three metrics were used to evaluate how well the model postulated word boundaries:

- **Precision**: percentage of correct word boundaries out of all boundaries found by the algorithm.
- **Recall**: percentage of correct word boundaries found out of all true boundaries in the corpus.
Table 2: The two representations constructed for each dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Representation</th>
<th>Input</th>
<th>Ideal lexicon output</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>Full</td>
<td>you want to see the book</td>
<td>you.want.to.see.the.book</td>
</tr>
<tr>
<td></td>
<td>C-only</td>
<td>ywntstthbk</td>
<td>y.wnt.t.s.th.bk</td>
</tr>
<tr>
<td>Arabic</td>
<td>Full</td>
<td>jallah il buul il buu fat riin</td>
<td>jallah. il buu. il buu. fat riin</td>
</tr>
<tr>
<td></td>
<td>C-only</td>
<td>jllhl il bbl il bbl fi riin</td>
<td>jllhl. il bbl. il bbl. fi riin</td>
</tr>
</tbody>
</table>

- **F-measure**: harmonic mean of Precision and Recall, \( \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \).

### 2.1.2 Model

Our simulations employed the Bayesian segmentation model developed by Goldwater et al. (2009). This is a generative model of unigram segmentation which infers a lexicon out of which the observed data (the corpus) are assumed to have been drawn. This specific model was chosen as it implements a well-tested Bayesian framework for segmentation (Phillips and Pearl, 2015a, 1). The specific assumptions made by the model are not crucial since it is used here mostly as proof of concept; the general effects of language-specific representations should hold regardless of the exact algorithm employed. Nevertheless, two biases inherent to the model are highlighted below: a preference for a smaller lexicon and a preference for shorter word forms.

For each utterance the model generates the word forms \( w_1, w_2, \ldots, w_n \) sequentially using a Dirichlet Process (Ferguson, 1973). The probability of the word \( w_i \) being generated depends on two parameters and on the number of times this word has appeared previously; this is a “rich-get-richer” algorithm in which words that have been generated previously in the run enjoy a higher probability of being generated again than novel word forms:

\[
P(w_i|w_1, \ldots, w_{i-1}) = \frac{n_{i-1}(w_i) + \alpha P_0(w_i)}{i-1 + \alpha} \tag{1}
\]

In Eq. 1:

- \( n_{i-1} \) is the number of times our word \( w_i \) has already appeared within the previous \( i - 1 \) words.
- \( \alpha \) is a parameter of the model specifying how likely it is that \( w_i \) is a new word. In the simulations of Goldwater et al. (2009) its value was set at 20, a value we retained. It can be seen that as \( \alpha \) approaches zero, the model is less likely to generate a new word, favoring a smaller lexicon instead.
- \( P_0 \) is a parameter of the model describing the “base distribution” of the word \( w_i \), that is, its internal phonemic makeup. \( P_0 \) is the probability that the novel word will consist of the phonemes \( x_1, \ldots, x_m \):

\[
P_0 = P(w_i = x_1, \ldots, x_m) = \prod_j P(x_j) \tag{2}
\]

It can be seen that as the word is shorter (smaller \( m \)), its probability will be higher; we return to this bias in the Discussion, explaining why it does not influence our conclusions.

Next, the word boundaries must be identified through an inference procedure. Goldwater et al. (2009) used Gibbs sampling (Geman and Geman, 1984) with 20,000 iterations, a value we retained. The Gibbs sampler uses Markov Chain Monte Carlo methods to decide on the value of each potential word boundary, i.e. whether a word boundary should be inserted between each two phonemes. The learner iterates through the input, guessing the value of each possible boundary based on the value of all other potential boundaries. The model eventually converges on a set of word boundaries, leading
to a segmented dataset. The model is non-deterministic as it uses a random seed for the initial distribution of word boundaries in the inference procedure.

This kind of model is effectively trained on the sequences of phonemes in the data and then tested on the segmentation. Goldwater et al. (2009) propose a modified version of the model which takes context into account, assuming that each word $w_i$ is also related to the previous word $w_{i-1}$. We set this more complicated model aside for the present study, proceeding with the frequency-based unigram model while noting that the results should generalize to the more elaborate model.

The implementation provided by Goldwater et al. (2009) generates Precision, Recall and F-measure scores for the segmentation produced by the model. Better segmentation performance is usually reflected by higher F-measure scores, which we predicted for the Full representation of English and the C-only representation of Arabic. These results are surveyed next.

2.2 Results

The following plots present the results of running the segmentation algorithm on each of the representations for each of the two languages. The output consists of a set of unsegmented utterances (the original input) into which word boundaries have been inserted, as in Table 2 above.

Starting with English in Fig. 1, the segmentation based on the full representation (“CV”, blue, on the left of each pair) outperforms the segmentation based on the consonant-only representation (“C”, red, on the right of each pair) for each of the three metrics: Precision, Recall and F-measure. Precision drops from 90.8 ($\pm0.23$) to 83.7 ($\pm0.38$) when using the C-only representation, Recall drops from 64.3 ($\pm0.44$) to 51.6 ($\pm0.36$) and F-measure drops from 75.3 ($\pm0.35$) to 63.9 ($\pm0.34$). Error bars show 95% confidence intervals over five runs of the Goldwater et al. (2009) model; variance is low, as could also be seen in the original experiments of Goldwater et al. (2009). We followed the methodology used by these authors in their own evaluation, calculating a Wilcoxon sum-rank $p$-value for comparisons between the representations. All three differences were significant at the $p < 0.01$ level.

In Arabic Newswire a different pattern can be seen, as in Fig. 2. Precision rises from 37.3 ($\pm0.29$) to 57.1 ($\pm0.29$) when using the C-only representation, Recall rises from 77.7 ($\pm0.49$) to 84.6 ($\pm0.25$), and F-measure rises from 50.5 ($\pm0.36$) to 68.2 ($\pm0.24$). All differences were significant at the $p < 0.01$ level. Performance in Arabic is overall worse than in English.

Emirati CDS patterns together with Arabic Newswire, as can be seen in Fig. 3: the consonant-only representation aids segmentation when compared to the full representation, as hypothesized. Using the C-only representation boosts Precision from 44.6 ($\pm0.49$) to 52.9 ($\pm0.23$). Recall takes a hit, dropping from 85.7 ($\pm0.67$) to 76.7 ($\pm0.23$). Crucially, F-measure rises from 58.6 ($\pm0.57$) to 62.6 ($\pm0.26$). All differences were significant at the $p < 0.01$ level. The one fly in the ointment is the higher Recall score for the full representation in our Emirati CDS corpus; this result is discussed immediately below.

2.3 Discussion

The segmentation algorithm performs better on English data than on Arabic data. This result is in line with the crosslinguistic findings of Phillips and Pearl (2015a), who show that performance drops, in absolute terms, when segmenting languages other than English. For Arabic this much is to be expected: Arabic words are longer than English words (see Table 1) and the model assigns lower scores to longer words. Furthermore, Arabic words are consonant-heavy in including many affixes, thereby making the algorithm’s task more difficult.

The results of Experiment 1 support our hypothesis: language-specific representations help the learner in a basic segmentation task. In particular, attuning to the consonants in the input aids the learner of Arabic but hinders the learner of English. The preference for consonants was highly evident
when segmenting Arabic Newswire and persisted in Emirati CDS as well, providing a stronger test of this hypothesis. While the Emirati CDS model saw a drop in Recall, the net result as provided by the F-measure was still positive. Moreover, it is often argued that it is more important for segmentation models to obtain high Precision than high Recall, due to the developmental finding that children’s initial segmentations are often undersegmentations (Peters, 1983). Adriaans and Kager (2010) note that undersegmentation (high Precision but low Recall) could lead to accurate proto-words to which meaning can be attributed, whereas oversegmentation (high Recall but low Precision) would lead to inaccurate lexical entries. The results on the Emirati CDS corpus are thus positive both from a quantitative perspective (due to the higher overall F-measure) and from a qualitative perspective (due to developmentally plausible undersegmentation).

The next set of experiments examined how well the segmented dataset supports further acquisition of the language: is the bias towards part of the input useful for additional cognitive tasks beyond pure word segmentation?

### 3 Experiment 2: The emergent proto-lexicon

The results of Experiment 1 support a view on which the statistical learner of Arabic performs better when considering only the consonants in the input, to the exclusion of vowels. As noted earlier, the output of the statistical segmentation procedure can be used to construct a proto-lexicon of hypothesized word forms in the language: whatever appears between word boundaries in the output should be a word. In Experiment 2 we ask to what extent structurally different proto-lexicons (based
on either full representations or consonant-only representations) could provide a starting point for further stages of language acquisition in Arabic. First, we ask whether the proto-lexicon helps to learn the actual lexicon of the language, focusing on consonantal roots. Then, we ask whether the proto-lexicon supports the acquisition of probabilistic phonological patterns in Arabic. In this way, we examined two distinct cognitive tasks involved in the acquisition of language which might benefit from the statistical learner’s bias towards part of the input.

3.1 Learning the lexicon

Let us start with a representative utterance from the results of Experiment 1. For the example mentioned earlier, *jallah lībūu lībūu fatīrīin*, a morpheme-by-morpheme gloss and translation are provided in Table 3. In the gloss, “PL” stands for plural, “PTCP” stands for the participial form and “2” stands for second person inflection.

<table>
<thead>
<tr>
<th>Arabic</th>
<th>Morpheme-by-morpheme segmentation</th>
<th>Morpheme-by-morpheme gloss</th>
<th>Free translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>jallah</td>
<td>jallah</td>
<td>come.on</td>
<td>‘Go ahead and play, play, you clever ones’</td>
</tr>
<tr>
<td>lībūu</td>
<td>lībūu</td>
<td>play-2PL</td>
<td></td>
</tr>
<tr>
<td>lībūu</td>
<td>lībūu</td>
<td>play-2PL</td>
<td></td>
</tr>
<tr>
<td>fatīrīin</td>
<td>fatīrīin</td>
<td>clever.PTCP-PL</td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Breakdown of the Arabic example.

The segmentation given in Table 3 is the correct segmentation for this utterance. We will go through the word forms one by one. *jallah* is a common interjection. *lībūu* is a second person plural
imperative. The root of the verb is $l^\text{-}\tilde{v}^\text{-}b$, three phonemes which also constitute the stem. The suffix 
$-uu$ marks second person plural. Finally, in $fat^\text{5}riin$, $fat^\text{5}r$ is a noun meaning ‘clever one’, derived 
from the root $f^\text{-}t^\text{5}^\text{-}r$. Inserted in the prosodic pattern XaYZ, this root instantiates the participial form 
$fat^\text{5}r$. The suffix $-in$ marks plurality on nouns.

Now let us see how a randomly selected run of the segmentation algorithm fared on this trial. The 
results of using the full representation and the consonant-only representation are given in Table 
4. The full representation leads to a segmentation generating individual word forms such as $buu$, $fa$ and 
$riin$, which have no independent status in the language. In contrast, the consonant-only representation 
correctly singled out the two roots $l^\text{-}\tilde{v}^\text{-}b$ and $f^\text{-}t^\text{5}^\text{-}r$. This is a welcome result, assuming that these 
patterns need to be acquired by the learner if she is to master the morphology of her language. Recall 
that the model favors shorter words; this might be the reason why the full-representation segmentation 
includes a number of very short postulated word forms such as $fa$ and $t^\text{5}$. Yet this bias does not trip 
up the C-only representation in our example, allowing it to postulate the two roots.

<table>
<thead>
<tr>
<th>Full representation</th>
<th>jallah</th>
<th>$R$</th>
<th>buu</th>
<th>$R$</th>
<th>buu</th>
<th>$fa$</th>
<th>$t^\text{5}$</th>
<th>riin</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-only representation</td>
<td>jllh</td>
<td>$l^\tilde{v}b$</td>
<td>$l^\tilde{v}b$</td>
<td>$f^\text{5}r$</td>
<td>$n$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Segmentations produced by the model for different representations.

This contrast exemplifies how the overall quantitative results of Experiment 1 reflect a qualitative 
aspect of the grammar of Arabic. Attuning to the consonants in the input allows the learner to make 
headway on both acquisition tasks at once, since the morphological and phonological patterns are
highly correlated.

In order to evaluate performance on this task quantitatively, a list of roots in Arabic was obtained and used as reference (Altantawy, Habash, Rambow, and Saleh, 2010; Habash and Rambow, 2006). For each proto-lexicon generated by a run of the segmentation algorithm, a list of postulated tri-consonantal and quadri-consonantal roots was extracted from the segmented word forms. This list was then compared to the reference list of true roots in Arabic. Results are given in Table 5 in the form of Precision, Recall and F-measure, averaged over the five runs of segmentation on the Arabic Newswire corpus in Experiment 1.

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV</td>
<td>43.5</td>
<td>11.2</td>
<td>17.8</td>
</tr>
<tr>
<td>C-only</td>
<td>39.2</td>
<td>12.8</td>
<td>19.3</td>
</tr>
<tr>
<td>Lexicon</td>
<td>21.0</td>
<td>22.5</td>
<td>21.7</td>
</tr>
</tbody>
</table>

Table 5: Root matches for the proto-lexicons produced by different representations of Arabic Newswire.

When considering the tri- and quadri-consonantal roots that can be learned from the different proto-lexicons, even the word list extracted from the actual text (“Lexicon”) shows a low absolute F-measure. This low score is not surprising, since the reference list was extracted from a dictionary, which contains all possible consonantal roots attested throughout the language, including many roots that are not used in day-to-day speech. What is of interest is the relative differences between the representations, not the absolute values. Here, we find that the proto-lexicon based on the C-only representation is significantly closer to the gold standard than that obtained from the full representation ($p < 0.01$ on a Wilcoxon rank-sum test for Precision, $p < 0.02$ for Recall and F-measure).

Evaluation of the CDS corpus showed identical patterns, significant as well, albeit with even lower absolute scores: F-measure scores of 4.0 (CV), 4.6 (C-only) and 8.7 (Lexicon). These lower absolute scores are to be expected for three reasons. First, the corpus is smaller so the sample of roots is smaller. Second, CDS contains a smaller vocabulary than adult speech. And third, the list of roots was based on Standard Arabic, not on the Emirati dialect used in the CDS corpus (Ferguson, 1959).

3.2 Phonotactics: The Obligatory Contour Principle

The next test of our hypothesis focused on the Obligatory Contour Principle (OCP), a phonological principle which restricts the co-occurrence of elements sharing certain features (e.g., Goldsmith, 1976; Greenberg, 1950; Leben, 1973; McCarthy, 1986,8). For example, OCP-Place (McCarthy, 1988) states that sequences of consonants that share place of articulation (homorganic consonants) should be avoided. This is especially the case for consonants that share manner of articulation as well (e.g. stops with stops and fricatives with fricatives). In many languages this constraint has the gradient effect that pairs of homorganic consonants (across vowels) tend to be underattested in the lexicon (Arabic: Frisch, Pierrehumbert, and Broe, 2004; Dutch: Kager and Shatzman, 2007; English: Berkley, 2000; Japanese: Kawahara, Ono, and Sudo, 2006; Muna: Coetzee and Pater, 2008). In addition to under-attestation of OCP-violating words in the lexicons of various languages, the OCP has been shown to affect listeners’ behavior in a variety of tasks. For example, Berent and Shimron (1997) found that root-initial geminates in Hebrew (i.e., repeated consonants at the beginning of a root) affected the rating of nonwords by native Hebrew speakers. Root-initial gemination (X-X-Y) was judged to be less acceptable than a non-geminate control (X-Y-Z). Similar results were obtained in a study of wellformedness judgments by native speakers of Arabic: novel roots containing a violation of OCP-Place were judged to be less word-like than novel roots which did not violate OCP-Place (Frisch and Zawaydeh, 2001). OCP-Place has also been found to affect speech perception by English listeners, particularly in phoneme identification tasks (Coetzee, 2005).
While OCP-Place was originally proposed to be an innate phonological primitive (McCarthy, 1988), more recent work has suggested that it is a probabilistic constraint that emerges from abstraction over word forms (or roots) in the lexicon. In an analysis of OCP-Place in Arabic, Frisch et al. (2004) define a gradient constraint whose degree of violation is a function of the similarity between consonant pairs (in terms of shared natural classes). Highly similar homorganic consonants are strongly underrepresented in the lexicon of Arabic roots, whereas relatively dissimilar pairs are underrepresented to a lesser degree. Frisch et al. (2004) propose that during language acquisition, individual speakers learn an abstract phonotactic constraint as a result of generalization over statistical patterns in the lexicon. The degree of support for such a generalization is quantified using the observed/expected ratio (Pierrehumbert, 1993), a statistical measure of association between observations that is closely related to pointwise mutual information:

$$OE(xy) = \frac{P(xy)}{P(x) \cdot P(y)}$$  \hspace{1cm} (3)

In Eq. 3, $P(x\bullet)$ is the probability of $x$ being the first phone in a biphone and $P(\bullet y)$ is the probability of $y$ being the second phone in a biphone. The log of O/E is simply the pointwise mutual information of the two variables, $x$ and $y$.

OCP-Place is thus taken to be an important concept in Arabic phonology, providing us with a test case for our view of consonant-only word segmentation. The question addressed here is to what extent the proto-lexicons that are formed by various segmentation algorithms provide the appropriate statistical patterns that give rise to gradient OCP effects in Arabic. If the consonant-only proto-lexicon shows OCP effects more closely than the consonant-and-vowel proto-lexicon does, then this provides further evidence for the benefits of learning based on consonants. Not only is segmentation more efficient, but the resulting output also better supports the learning of language-specific phonological generalizations, something that the child learner will have to do. We therefore analyze the statistical phonological patterns in each of the segmented proto-lexicons from Experiment 1.

3.2.1 Methods

Four different datasets were compared, all the result of different segmentation strategies applied to the Arabic corpora. Two proto-lexicons were the result of the segmentation procedures in Experiment 1 (C-only and full CV). In addition, in order to get an idea of lower and upper boundaries on performance, we included a proto-lexicon obtained from an unsegmented baseline (in which each utterance was treated as an entire word) as well as a proto-lexicon obtained from the correct segmentation (the true lexicon). The unsegmented baseline allows us to quantify to what extent word boundaries are relevant at all in obtaining the appropriate statistical patterns, while the true lexicon allows us to determine which representation (C-only or CV) provides the closest approximation to the adult pattern that needs to be learned. The three different proto-lexicons and the true lexicon are listed below, with statistics for the segmented proto-lexicons given in Table 6.

1. Unsegmented baseline (`Unsegmented'; e.g., jallahl\textsl{b}uul\textsl{b}uu\textsl{f}at\textsl{f}\textsl{riin})
2. Segmentation of the full representation (`CV'; e.g., jallah\textsl{.b}uu\textsl{.b}uu\textsl{.f}at\textsl{f}\textsl{riin})
3. Segmentation of the consonant-only representation (`C-only'; e.g., jil\textsl{h}h\textsl{b}.\textsl{b}uul\textsl{.f}at\textsl{f}\textsl{.riin})
4. The correct segmentation based on the true lexicon (`Lexicon', e.g., jallah\textsl{.b}uu\textsl{.b}uu\textsl{.f}at\textsl{f}\textsl{riin})

For each of the four datasets, O/E ratios were calculated for pairs of non-identical labials, coronals and dorsals in each proto-lexicon. Coronals were separated into stops and fricatives, as in Frisch et al. (2004) following the findings of Greenberg (1950) and McCarthy (1988). The inventory for Arabic was as follows (/\textsl{g}/ is a phoneme of Emirati Arabic but is not part of the phonemic inventory of Modern Standard Arabic):
Table 6: Statistics for Arabic proto-lexicons derived from the five runs of the model in Experiment 1.

<table>
<thead>
<tr>
<th></th>
<th>Arabic Newswire</th>
<th>Emirati CDS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CV</td>
<td>C-only</td>
</tr>
<tr>
<td>Utterances</td>
<td>7,914</td>
<td></td>
</tr>
<tr>
<td>Words</td>
<td>267,004.4</td>
<td>192,204.6</td>
</tr>
<tr>
<td>Types</td>
<td>4,057.0</td>
<td>3,443.2</td>
</tr>
<tr>
<td>Words/utterance</td>
<td>33.74</td>
<td>24.29</td>
</tr>
<tr>
<td>Word length</td>
<td>3.71</td>
<td>2.79</td>
</tr>
</tbody>
</table>

- Labials: b f m
- Coronal stops: t d t̠ d̠
- Coronal fricatives: r θ δ s s̠ z z̠ f d̠
- Dorsals: k (g) q χ υ

3.2.2 Results

The O/E ratios for each proto-lexicon are given in Fig. 4 for Arabic Newswire. Following earlier studies on OCP-Place, the results are grouped by place and manner of articulation. For each place of articulation, the four segmentations are plotted as bars.

For labials, the true O/E based on the real lexicon is 0.162, based on 1,513 observations (1,513 pairs of such consonants). The O/E calculated from the proto-lexicon of the C-only segmentation comes closest with 0.185, followed by the full representation (0.151) and the unsegmented baseline (0.318). For coronal stops, the lexicon gives an O/E of 0.347 based on 4,350 observations. The C-only representation is again closest with 0.314, with the full representation noting a low O/E of 0.218 and the unsegmented baseline reaching 0.235. For coronal fricatives, the O/E from the lexicon is 0.51 based on 9,698 observations. The C-only representation fares best with 0.521, ahead of the full representation (0.560) and the unsegmented baseline (0.525). For the dorsals, the lexicon’s O/E is 0.053 based on only 12 observations. The C-only representation has 0.003, the full representation 0.003 and the unsegmented baseline 0.06.

The finding that C-only representation is closer to the gold standard than the full representation is significant for coronal stops ($p < 0.008$) and coronal fricatives ($p < 0.02$). Neither C-only nor full representation are significantly closer than the other to the labial result, but they are different from each other ($p < 0.008$).

Results for the Emirati CDS corpus are presented in Fig. 5. For the three labial consonants of Emirati Arabic, the O/E as would be induced from the true lexicon for the corpus is 0.128, based on 162 observations. The full representation comes closest with 0.128 ($±0.02$), slightly better than the C-only representation with 0.145 ($±0.003$) and ahead of the unsegmented baseline with 0.257. Variance is larger by an order of magnitude for the full representation than for the C-only representation. The C-only and full representations are not significantly different at the $p < 0.05$ level.

For the coronal stops, the O/E from the lexicon is 0.09 based on 128 observations. Here, the unsegmented baseline is closest to the ideal with an O/E of 0.139. The C-only representation and the full representation do not differ significantly, though the latter once again shows greater variance: 0.037 ($±0.011$) for C-only vs 0.041 ($±0.055$) for the full representation.

For coronal fricatives, the lexicon’s O/E is 0.496, based on 1,181 observations. The C-only representation shows an O/E of 0.612 ($±0.006$), statistically indistinguishable from the full representation with 0.601 ($±0.015$). The unsegmented baseline is most similar to the lexicon (0.5).
Figure 4: Experiment 2 results for Arabic Newswire. Averages were calculated over the five runs from Experiment 1 for the CV (full) and C-only representations of the input.

Turning to the dorsals, a true segmentation leads to an O/E of 0.062, based on a low count of seven observations. The C-only representation results in an O/E of 0.042 (±0.002), similar to the full representation (0.043±0.029) with less variance and ahead of the unsegmented baseline (0.102).

### 3.3 Discussion

The results of Experiment 2 provide additional support for the hypothesis that a consonant-only representation benefits the learner of a Semitic language. If basic phonological generalizations such as OCP-Place are learned from the lexicon, then the proto-lexicon that emerges from a segmentation based on consonants (“C-only”) facilitates this aspect of learning the language better than a proto-lexicon resulting from the segmentation of both consonants and vowels (“CV”).

To be sure, the Arabic Newswire results are different than those of Emirati CDS: whereas Experiment 1 showed a clear advantage for the C-only representation in both corpora, the results of Experiment 2 were clear for Arabic Newswire but less so for Emirati CDS. It is difficult to establish the causes of the difference since the two Arabic corpora are different along at least two dimensions. First, they each represent different dialects. The Arabic Newswire corpus is written in Modern Standard Arabic, while the Emirati CDS corpus is in the Emirati dialect. Second, the Arabic Newswire corpus represents adult language, while the Emirati CDS corpus consists of child-directed speech.

We return to these differences in the General Discussion. What is important to note is that both datasets showed clear positive effects for the C-only representation on segmentation in Experiment 1 and that, while the difference between the full and C-only representation is minimal in terms of O/E
for the Emirati CDS corpus in Experiment 2, the removal of vowels turned out to be beneficial in the Newswire corpus, supported learning of the lexicon in the Emirati CDS corpus as well, and did not result in a negative effect on later phonotactic learning in CDS.

4 General discussion

The current study tested the hypothesis that a learner of Arabic would be aided by focusing on consonants in the input. We have put forward a view in which the statistical learning mechanism used for word segmentation is invariant, but operates on representations which may vary crosslinguistically. This possibility was explored by testing a Bayesian model on two languages with different morphological systems (English and Arabic) and two different input representations (consonants-plus-vowels and consonants-only). Our findings indicate that a consonant-only representation of the Arabic input stream is superior to a full representation containing both consonants and vowels. The results of Experiment 1 followed our prediction that a consonant-only representation is beneficial for segmentation in Arabic, but not in English. The beneficial effect on segmentation was found for two different Arabic corpora, one containing adult-directed speech (Arabic Newswire) and one containing child-directed speech (Emirati CDS). Experiment 2 assessed whether the consonantal representation could aid further language acquisition by examining the extent to which the segmented Arabic proto-lexicon contained actual Arabic consonantal roots and phonological patterns. We found that a consonantal proto-lexicon provides a better match with a list of known Arabic roots than a consonant-plus-vowel proto-lexicon. In addition, for our corpus of adult-directed speech we found that the phonological constraint OCP-Place is more apparent in the consonant-only proto-lexicon than in the consonant-plus-vowel proto-lexicon. However, in our corpus of child-directed speech we did not find clear-cut differences in terms of phonological patterns. Taken together, the results support the view that consonantal representations are beneficial for segmentation and root learning in Arabic, and the results partially support the view that consonantal representations might support further phonological acquisition in Arabic.

Our findings support the possibility that linguistic constraints might guide and improve the ef-
fectiveness of statistical learning (e.g., Adriaans and Kager, 2010; Bonatti et al., 2005; Peperkamp, Le Calvez, Nadal, and Dupoux, 2006). While previous studies have shown that human learners can track consonant probabilities across intervening vowels in artificial languages (Bonatti et al., 2005; Newport and Aslin, 2004), it was not known to what extent this ability would be beneficial in the segmentation of natural language. Our simulations indicate that focusing on consonants leads to an improvement in statistical segmentation in some languages, but not in others. For the case of Arabic, focusing on consonants would enable the learner to identify consonantal roots, an integral part of the morphological system of a Semitic language.

Consonants have an important role in carrying lexical information across languages. In a study of Dutch and Spanish, Cutler, Sebastian-Galles, Soler-Vilageliu, and Van Ooijen (2000) found that speakers were more likely to change a vowel in the stimulus than a consonant. In a recent study by Aldholmi (2016), Arabic and English participants were asked to judge whether two nonce words were identical or different. English speakers were equally likely to identify a difference between the stimuli when the words differed by one vowel or by one consonant. In contrast, Arabic speakers were much more likely to detect a consonantal difference than a vocalic one. These findings, as well as our current ones, support a view in which speakers of Semitic attune to consonants in the input over vowels to a larger degree than in European languages.

4.1 Developmental data

Our view of consonant-based acquisition is indirectly supported by developmental data from Semitic languages. Since there is not a large body of knowledge on the acquisition of Arabic, we focus here on studies of a closely related language, Modern Hebrew. At age 2;2 Hebrew-speaking children produce erroneous verbs by using the correct root in the wrong template, often innovating new verbal forms (Berman, 1982,9; Borer, 2004; Levy, 1988). This finding indicates that children first acquire consonantal roots and then learn the grammatical details of the templatic system. Our root-centric model is fully compatible with this description. In fact, the model can explain why this route is taken: it is because roots, as consonantal strings, are easy to pick out from the input stream. It remains to be evaluated how much information, and of what kind, the learner needs in order to progress from one developmental stage to the next.

The developmental data also indicate that infants encounter a substantial amount of morphophonological regularity in the first few years of life (Ravid, Ashkenazi, Levie, Ben Zadok, Grunwald, Bratslavsky, and Gillis, To appear). It appears that infants start off using consonantal roots in the most frequent template, gradually expanding their morphosyntactic system based on the phonological alternations and the additional templates they are exposed to. By age 4 children seem to have good control not only of the template as a morphophonological object but also of its syntactic and semantic properties. Our root-centric model is compatible with these findings: exposure to the complexities of the morphological system is gradual and can be argued to begin with the consonantal root, progressing from it to an expanded system of morphosyntactic primitives. Future work will need to develop a model integrating vowels back into the system. The goal would then be to model how the rest of the morphological system is learned, namely the different verbal templates.

With these developmental data in mind, one might wonder why our results were generally stronger in our corpus of adult-directed speech than in our corpus of child-directed speech. Two differences should be emphasized between these Arabic corpora. First, the adult-directed speech corpus (Arabic Newswire) uses Modern Standard Arabic, which has nine verbal templates. The child-directed speech corpus uses the Emirati dialect, which mostly employs only the three or four most frequent templates. The Emirati dialect is thus relatively less complex in terms of its verbal morphology and as a result the differences between the consonant-only and consonant-plus-vowel representations in the Emirati corpus are relatively small. Second, child-directed speech is simplified on both grammatical and lexical levels when compared to newswire. As can be calculated from the statistics in Table 1, the
average number of distinct word types per utterance is larger in Arabic Newswire than in Emirati CDS (3.3 in Arabic Newswire compared to 1.0 in Emirati CDS). The lexicon used in the Emirati CDS corpus is thus relatively small, which makes the segmentation task less complex. This contributes to relatively small differences found between consonant-only and consonant-and-vowel representations in the Emirati CDS corpus.

4.2 Choosing the right representation

According to the learning mechanisms proposed here, the Arabic learner may focus on the consonants in the input while the English learner considers both consonants and vowels. It has not been specified, however, how the segmentation algorithm should know that it should target consonants in Arabic, but not in English.

In this context, it is important to note that the developmental literature is unclear on the exact nature of infants' representational units, and how these develop over time. It seems that newborn infants primarily retain information about the vowels they hear, while ignoring information about consonants (Benavides-Varela, Hochmann, Macagno, Nespor, and Mehler, 2012). However, older infants (12 months) seem to mainly represent consonantal information about the words they hear (Hochmann et al., 2011). A related unresolved debate in both infant studies and computational studies has been whether initial representations are segment-based or syllable-based (e.g., Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler, 1988; Eimas, 1999; Hochmann and Papeo, 2014; Jusczyk and Derrah, 1987; Phillips and Pearl, 2015b; Saffran et al., 1996).

While it is unknown how infants would develop a language-specific focus on a particular representation, we propose that the learner might initially entertain multiple hypotheses (consonant-only and full representation, but perhaps also considering a syllabic representation), but through experience with the language the learner will recognize that the consonant-only representation is more effective in Arabic. This transition would require some form of internal evaluation metric. For example, the learner might notice that the consonant-only segmentation results in a Zipfian proto-lexicon whereas the full representation segmentation results in an unexpected distribution. In such a case, the learner might prefer the consonant-only hypothesis. If the learner is assumed to entertain both hypotheses with different weights, it might then increase the prior for the consonant-only hypothesis and reduce the prior for the full representation hypothesis. This is a general question as to what guides the learner in the transition from CDS-like patterns to adult-like patterns, one that confronts any study of language acquisition.

4.3 Vowel inventory and prosodic structure

An alternative explanation of our findings is that the difference between Arabic and English is not due to different morphologies between the two languages, but to other phonological properties that could give rise to the consonantal advantage that we saw for Arabic. We discuss here two specific alternative views, and mention what additional languages would need to be investigated in order to determine the most likely source of the consonantal advantage in our results.

It might be that the phonemic inventory of the language is what gave rise to the different results for Arabic and English. Modern Standard Arabic has three vowels and Emirati Arabic has five, compared to the 15 phonemic vowels of English distinguished by Goldwater et al. (2009). Perhaps it is the smaller vowel inventory that leads to a higher reliance on consonants? Consider Spanish, then, which like Emirati Arabic has five phonemic vowels. If the results are driven by number of vowels, Spanish should pattern like Arabic and unlike English in that a consonant-only representation would be more beneficial to the learner than the full representation. Nevertheless, we suspect that Spanish would pattern with English and unlike Arabic due to the morphological nature of the language. While this prediction remains to be tested, it should be noted that even if Spanish ended up patternning similarly
to Arabic with regards to the two tasks in this paper, the learner of Arabic would still need to acquire the consonantal roots. The consonant-only learning strategy would thus be beneficial for the Arabic learner regardless of its feasibility for non-Semitic languages.

Another alternative interpretation of the facts holds that the prosodic structure of Arabic biases the results. Arabic syllables follow a CV(V) pattern, meaning that perhaps more information could be inferred from consonants in Arabic than in a language with a more complex syllable structure such as English. In this case, a comparison could be made with a language such as Japanese, which is also strongly CV-shaped. Like English, Japanese does not base its morphology on roots and patterns. If our findings are the result of the prosodic structure of the language, then Japanese should pattern with Arabic and unlike English. But if we are correct in attributing the difference between English and Arabic to the nature of the morphological system, Japanese would pattern with English.

Cases such as these highlight the importance of investigating multiple languages that differ in terms of linguistic structure when studying cognitive tasks related to language. As Phillips and Pearl (2015a) put it, for a given language acquisition strategy, each language can be seen as an individual data point. Testing multiple languages is necessary in order to properly evaluate different strategies. Our proposal models the word segmentation process in a way that separates the particulars of a specific language from a uniform learning mechanism. In this case, we have argued that separating consonants from vowels is beneficial for the learner of Semitic. The result is a “less-is-more” situation (Newport, 1990; Phillips and Pearl, 2012), in which withholding certain information (the vowels, in this case) helps focus the learner on the signal. Like Nespor et al. (2003), we speculate that the learner might first assign consonantal “chunks” to objects before augmenting the rest of the grammar with vowels. This paper presented a computational test of this hypothesis on natural language data, providing computational support for the view that statistical learning might interact with linguistic representations in order to effectively meet the challenges of language acquisition.
Appendix: Non-concatenative morphology

Contemporary approaches to segmentation treat the input as a string upon which the segmentation algorithm operates. The algorithm inserts word boundaries, generating a segmented representation of the data. While this approach is straightforward in keeping with our assumptions about individual phonological words, it contains a less innocent assumption: that the input can ultimately be linearly decomposed into individual elements, specifically morphemes. The input *thedogs* (to use orthography) can be segmented into the distinct phonological words *the.dogs* and then into the three morphemes *the.dog.s*: ultimately, the child will need to recognize the individual affixes in her language and distinguish between them and lexical roots. This view is innocent enough when applied to languages such as English and Dutch. Yet while the input stream is linear, Semitic morphology is not. The learner’s strategy will need to be modified.

In Central Semitic languages such as Arabic and Hebrew, the word is generally taken to be made up of a consonantal “root” interleaved with a prosodic “pattern”. The following verbs in Modern Standard Arabic, for instance, all share the three consonants *q-b-l*. When instantiated in one of a limited number of verbal “templates” or “patterns”, a verb is derived: *istaqḇal* ‘greeted’, *qabîl* ‘received’, *tagabbal* ‘accepted’. It is traditionally postulated that this root identifies a general semantic field of “receiving”, though not all roots are as regular in their meaning across patterns. The same root can be used in other ”patterns” to derive nouns, adjectives and even prepositions: *qablâ* ‘before’, *qabîlâ* ‘tribe’, *qîblâ* ‘the direction of prayer’.

The exact semantic contribution of each root and each pattern has been a matter of debate for as long as these forms have been studied. Contemporary approaches have gone back and forth on the question of whether the root and the pattern are themselves morphological primitives or epiphenomena of other syntactic and phonological processes in the language (Aronoff, 1994; Bat-El, 1994; Doron, 2003; Kastner, 2016; McCarthy, 1981; Ussishkin, 2005). What is important is that the root generally provides the semantic content of the word while the pattern provides the grammatical content. Both contribute to the formation of the word, to its meaning, to its syntactic frames and to its phonological form. Within a given root or a given verbal template the paradigm is predictable: if the speaker knows the form *istaqḇal* ‘greeted’, she will know how to conjugate the verb in other tenses and other patterns as well.

In his seminal analysis of these patterns, McCarthy (1981,8) proposed to divide the Semitic verb into two “planes” or “tiers”: one consonantal, for the root and affixes, and one containing vowels, for the pattern. Under this influential analysis, the representation of *tagabbal* ‘received’ is as in Fig. 6.

```
CVCVCCVC
```

**Figure 6:** Tier-based representations for *taqabbal*, following McCarthy (1981).

This analysis correctly predicts a range of phonological alternations in the language. Crucially, this account allows us to speak independently of the root and the pattern in a principled way: the consonants are architecturally different from the rest of the verb, leading to an elegant description of the system. Material on the consonantal tier is both phonologically and morphologically different from material on the skeletal tier, and the identity of the vowels is filled in on another tier. In the main text, we suggest that the task of learning roots and patterns can be approximated by learning
combinations of consonants and combinations of vowels. According to our hypothesis, the learner of Arabic is aided by focusing on the consonants in the input, leading to improved segmentation.

References


Aldholmi, Yahya. 2016. The role of non-concatenativeness vs. concatenativeness experience in perception. In *The 30th annual symposium on Arabic linguistics*. Stony Brook, NY.


infants. *Science* 274:1926–1928. 1, 2, 16