COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2017) 1–25

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12521

Linguistic Constraints on Statistical Word Segmentation: The Role of Consonants in Arabic and English

Itamar Kastner,[a] Frans Adriaans[b]

[a]*Department of English and American Studies, Humboldt University of Berlin*
[b]*Utrecht Institute of Linguistics OTS, Utrecht University*

Received 1 July 2016; received in revised form 20 March 2017; accepted 8 June 2017

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. This 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; Morphology

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

Correspondence should be sent to Itamar Kastner, Institut für Anglistik und Amerikanistik, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany. E-mail: itamar@itamarkast.net

variety of statistical and linguistic cues to detect word boundaries in continuous speech, including transitional probabilities (e.g., Saffran, Aslin, & Newport, 1996), metrical cues (Jusczyk, Houston, & Newsome, 1999), coarticulation cues (Johnson & Jusczyk, 2001), and phonotactic cues (Mattys & 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 & Cartwright, 1996; Goldwater, Griffiths, & Johnson, 2009; Lignos & Yang, 2010), through the computation of segment or syllable bigram probabilities (Cairns, Shillcock, Chater, & Levy, 1997; Daland & Pierrehumbert, 2011; Swingley, 2005), or through a combination of statistical learning and phonological generalization (Adriaans & 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, Phillips and Pearl (2015a) did not directly test the effect that a particular phonological or morphological characteristic of a language has on word segmentation in that language.

This study provides an investigation of statistical word segmentation in Arabic, a language whose morphology and phonology are 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.

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 & Jusczyk, 2001; Saffran et al., 1996; Thiessen & 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, & 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 & Werker, 1994; Werker & Tees, 1984), and possibly have different roles in early language acquisition (Hochmann, Benavides-Varela, Nespor, & Mehler, 2011). The emerging picture is one in which consonants contain lexical information and vowels contain grammatical information (but see Keidel, Jenison, Kluender, & 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 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

Table 1
Statistics for the different datasets in Experiment 1

	English	Arabic Newswire	Emirati CDS
Number of utterances	9,790	7,914	8,403
Number of word tokens	33,399	132,390	30,931
Number of word types	1,421	26,267	8,395
Number of words/utterance	3.41	16.73	3.68
Word length in phonemes	3.24	7.48	5.34

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." This combined dataset is meant to provide a test that is representative of the entire corpus. 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 et al., 2014). This tool provides a morphological parse and a phonemic transcription which have proven useful in other work on Arabic (Gwilliams & 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 & 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 Emirati words are longer.

The use of two different Arabic corpora allows us to not only make a general comparison between English and Arabic, but also to make a comparison between adult language and child-directed speech. It should be noted that, while the English CDS corpus has been used in various previous studies of segmentation (e.g., Brent & Cartwright, 1996; Goldwater et al., 2009; Phillips & Pearl, 2015a), the Emirati CDS corpus is used in a statistical segmentation study for the first time.

For each corpus two representations were constructed: one "Full" representation consisting of consonants and vowels, and one "C-only" representation consisting only of consonants. Table 2 gives a number of examples. The examples are given in orthography (English) or transliteration (Arabic), for ease of exposition. Each unsegmented phonemic representation (two representations per corpus) was fed to the segmentation model in turn. The output consists of a set of unsegmented utterances (the original input) into which

Table 2
The two representations constructed for each dataset. The Arabic example was taken from the Emirati CDS corpus and reads "Go ahead and play, play, you clever ones"

Dataset	Representation	Input	Ideal Lexicon Output
English	Full	youwanttoseethebook	you.want.to.see.the.book
	C-only	ywnttsthbk	y.wnt.t.s.th.bk
Arabic	Full	jallahlʕbuulʕbuuʃatˤriin	jallah.lʕbuu.lʕbuu.ʃatˤriin
	C-only	jllhlʕblʕbʃtˤrn	jllh.lʕb.lʕb.ʃtˤrn

word boundaries have been inserted. 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.
- F-measure: harmonic mean of Precision and Recall, $\frac{2 \cdot Prec \cdot Rec}{Prec + Rec}$

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 & Pearl, 2015a,b). Two biases are inherent to the model: 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:

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

In Eq. 1:

- n_{i-1} is the number of times our word w_i has already appeared within the previous $i - 1$ words.
- α 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 α 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.

Next, the word boundaries must be identified through an inference procedure. Goldwater et al. (2009) used Gibbs sampling (Geman & 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, that is, whether a word boundary should be inserted between each of two phonemes. The learner iterates through the input, guessing the value of each possible boundary based on the value of all other potential boundaries. This 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 this 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 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

Fig. 1 presents the results for the English corpus. 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 (±0.23) to 83.7 (±0.38) when using the C-only representation, Recall drops from 64.3 (±0.44) to 51.6 (±0.36) and F-measure drops from 75.3 (±0.35) to 63.9 (±0.34). 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 evaluation, calculating a Wilcoxon sum-rank p-value for comparisons between the representations. All three differences were significant at the $p < .01$ level.

A different pattern of results can be seen for the Arabic Newswire corpus in Fig. 2. Precision rises from 37.3 (±0.29) to 57.1 (±0.29) when using the C-only representation,

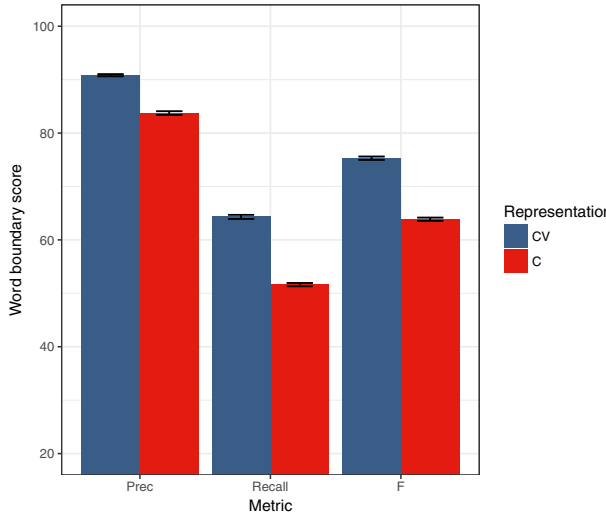

Fig. 1. Experiment 1 results for English (CDS).

Recall rises from 77.7 (±0.49) to 84.6 (±0.25), and F-measure rises from 50.5 (±0.36) to 68.2 (±0.24). All differences were significant at the $p < .01$ level.

The results for the Emirati CDS corpus are shown in Fig. 3. Using the C-only representation boosts Precision from 44.6 (±0.49) to 52.9 (±0.23). Recall takes a hit, dropping from 85.7 (±0.67) to 76.7 (±0.23). Crucially, F-measure rises from 58.6 (±0.57) to 62.6 (±0.26). All differences were significant at the $p < .01$ level. The consonant-only representation thus aids segmentation when compared to the full representation, as

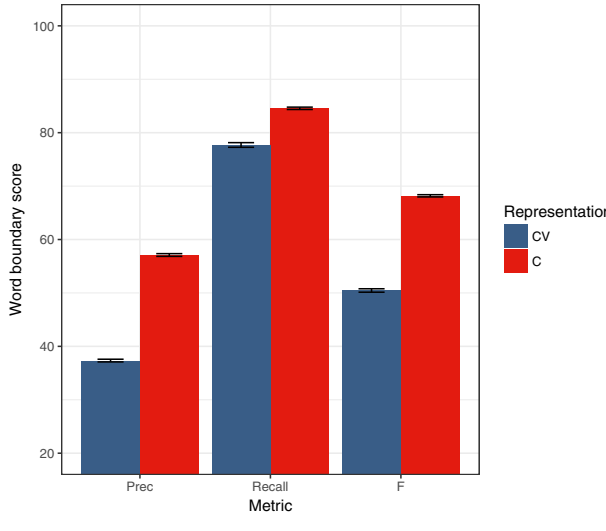

Fig. 2. Experiment 1 results for Arabic Newswire.

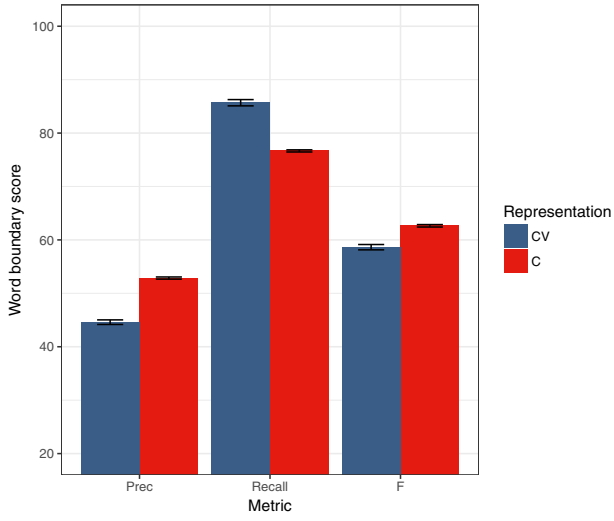

Fig. 3. Experiment 1 results for Emirati CDS.

hypothesized. The difference from the Arabic Newswire corpus is that the positive effect (as reflected by the higher F-measure) is due only to higher Precision.

2.3. Discussion

The segmentation algorithm performs better on English data than on Arabic data. This result is in line with the cross-linguistic 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 complex in that they include 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.[1] 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 (i.e., high Precision but low Recall) could lead to accurate proto-words to which meaning can be attributed, whereas oversegmentation (i.e., 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 segmentation results support the view that the learner of Arabic is aided by focusing on consonants in the input. In Experiment 2 we asked whether this bias towards part of the input could be useful for additional cognitive tasks beyond pure word segmentation. We present a series of simulations that examine whether consonant representations support further lexical and phonological acquisition of Arabic.

3. Experiment 2: The emergent proto-lexicon

In Experiment 2 we tested to what extent different input representations (full representations or consonant-only representations) lead to a proto-lexicon which could support further stages of language acquisition in Arabic. In this case the proto-lexicon is constructed by interpreting whatever appears between word boundaries in the output of the segmentation procedure as a hypothesized word form in the language. First, we asked whether the proto-lexicon helps to learn the actual lexicon of the language, focusing on consonantal roots. Then, we asked whether the proto-lexicon supports the acquisition of probabilistic phonological patterns in Arabic. We thus examined two distinct cognitive tasks involved in the acquisition of language which might benefit from the statistical learner's bias toward part of the input.

3.1. Learning the lexicon

To illustrate some of the lexical information the learner eventually has to learn, we provide a lexical annotation of the utterance *jallah lʕbuu lʕbuu ʃatˤriin* in Table 3.

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. *lʕb-uu* is a second-person plural imperative. The root of the verb is l-ʕ-b, three phonemes which also constitute the stem. The suffix *-uu* marks second-person plural. Finally, in *ʃatˤriin*, *ʃatˤr* is a noun meaning "clever one," derived from the root ʃ-tˤ-r. Inserted in the prosodic pattern XaYZ, this root instantiates the participial form *ʃatˤr*. The suffix *-iin* marks plurality on nouns.

How did the segmentation algorithm perform on this example utterance? The results of using the full representation and the consonant-only representation are given in Table 4. The full representation led to a segmentation generating individual proto-words such as *buu*, *ʃa* and *riin*, which have no independent status in the language. In contrast, the consonant-only representation correctly singled out the two roots l-ʕ-b and ʃ-tˤ-r. Proto-words in the

Table 3
Breakdown of the Arabic example. "PL" stands for plural, "PTCP" stands for the participial form, and "2" stands for second-person inflection

Word segmentation	jallah	lʕbuu	lʕbuu	ʃatˤriin
Morpheme segmentation	jallah	lʕb-uu	lʕb-uu	ʃatˤr-iin
Morpheme gloss	come.on	play-2PL	play-2PL	clever.PTCP-PL
Translation	"Go ahead and play, play, you clever ones"			

Table 4
Segmentations produced by the model for different representations

Full representation	jallah	lʕ	buu	lʕ	buu	ʃa	$t^ʕ$	riin
C-only representation	jllh	lʕb		lʕb		ʃ$t^ʕ$r		n

form of roots are a welcome result, since these patterns need to be acquired by the learner in order to master the morphology of the 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 $\int a$ and t^{ς}. Yet this bias does not trip up the C-only representation in our example, allowing it to postulate the two roots.

This example illustrates 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, & Saleh, 2010; Habash & 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. For the full representation this was done by removing all vowels from the proto-lexicon. Since no vowels occurred in the C-only representation, roots could be extracted without further processing. The list of hypothesized roots 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.

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 (the Lexicon) than that obtained from the full representation ($p < .01$ on a Wilcoxon rank-sum test for Precision, $p < .02$ for Recall and F-measure). The C-only representation thus provides a closer match to Arabic word roots than the full representation.

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

	Precision	Recall	F-measure
CV	43.5	11.2	17.8
C-only	39.2	12.8	19.3
Lexicon	21.0	22.5	21.7

Evaluation of the Emirati CDS corpus showed identical patterns, albeit with 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 phonemes sharing certain features (e.g., Goldsmith, 1976; Greenberg, 1950; Leben, 1973; McCarthy, 1986, 1988). For example, *OCP-Place* (McCarthy, 1988) states that sequences of consonants that share place of articulation (homorganic consonants) should be avoided. In many languages this constraint has the effect that pairs of homorganic consonants across vowels (e.g., *pVm*) are relatively unlikely to occur within words (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]). Importantly, the OCP has been shown to affect listeners' behavior in a variety of cognitive 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 & Zawaydeh, 2001). OCP-Place has also been found to affect speech perception by English listeners, particularly in phoneme identification tasks (Coetzee, 2005).

Recent work has suggested that OCP-Place is a probabilistic constraint that emerges from abstraction over word forms (or roots) in the lexicon. In a study on OCP-Place in Arabic, 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 (see also Albright & Hayes, 2003; Hayes & Wilson, 2008). The degree of support for such a generalization is quantified using the Observed/Expected (O/E) ratio (Pierrehumbert, 1993), a probabilistic measure of association between observations that is closely related to transitional probability and pointwise mutual information (Adriaans & Kager, 2010):

$$OE(xy) = \frac{P(xy)}{P(x\bullet) \times P(\bullet y)} \tag{3}$$

In Eq. 3, $P(x\bullet)$ is the probability of x being the first phoneme, and $P(\bullet y)$ is the probability of y being the second phoneme in a phoneme pair xy. Strong OCP effects are reflected in

low O/E ratios, indicating that *x* and *y* are unlikely to occur as a pair within words of the language.

Since OCP-Place is an important concept in Arabic phonology, it provides us with a test case for our view of consonant-based word segmentation. The specific question addressed here is to what extent the proto-lexicons that are formed by different input representations provide statistical patterns that match probabilistic OCP effects in the Arabic lexicon. To this end, we compare O/E ratios in the proto-lexicons to O/E ratios in the gold standard lexicon (which was derived from the gold standard segmentation of the corpus). If the consonant-only proto-lexicon more closely resembles the gold standard lexicon in terms of OCP effects, 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.

3.2.1. Methods

Statistical phonological patterns were analyzed and compared for a total of four different datasets, all derived from the same corpus. Two of these datasets were the proto-lexicons that resulted from the two segmentation procedures in Experiment 1 (C-only and 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 gold standard lexicon obtained from the correct segmentation of the corpus. The unsegmented baseline allows us to quantify to what extent word boundaries are relevant at all in obtaining the appropriate statistical patterns, while the gold standard 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 gold standard lexicon are listed below, with statistics for the segmented proto-lexicons given in Table 6.

1. Unsegmented baseline ("Unsegmented"; e.g., jallahlʕbuulʕbuuʃatˤriin)
2. Segmentation of the full representation ("CV"; e.g., jallah.lʕ.buu.lʕ.buu.ʃa.tˤ.riin)
3. Segmentation of the consonant-only representation ("C-only"; e.g., jllh.lʕb.lʕb.ʃtˤr.n)
4. The correct segmentation based on the true lexicon ("Lexicon," e.g., jallah. lʕbuu.lʕbuu.ʃatˤriin)

Table 6
Statistics for Arabic proto-lexicons derived from the five runs of the model in Experiment 1

	Arabic Newswire		Emirati CDS	
	CV	C-only	CV	C-only
Utterances	7,914		8,403	
Words	267,004.4	192,204.6	52,938.0	42,016.4
Types	4,057.0	3,443.2	1,434.2	1,031.6
Words/utterance	33.74	24.29	6.30	5.01
Word length	3.71	2.79	3.18	2.49

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 (Frisch et al., 2004; Greenberg, 1950; McCarthy, 1988). The inventory for Arabic was as follows (/g/ is a phoneme of Emirati Arabic but is not part of the phonemic inventory of Modern Standard Arabic):

- Labials: b f m
- Coronal stops: t d $t^{ʕ}$ $d^{ʕ}$
- Coronal fricatives: r θ ð s z $s^{ʕ}$ $z^{ʕ}$ ʃ dʒ
- Dorsals: k (g) q χ ɣ

3.2.2. Results

The O/E ratios for each proto-lexicon are given in Fig. 4 for Arabic Newswire. For each place of articulation (Labial, Coronal-Stop, Coronal-Fricative, Dorsal), the different proto-lexicons are plotted as bars. Lower bars (closer to zero) indicate stronger OCP effects.

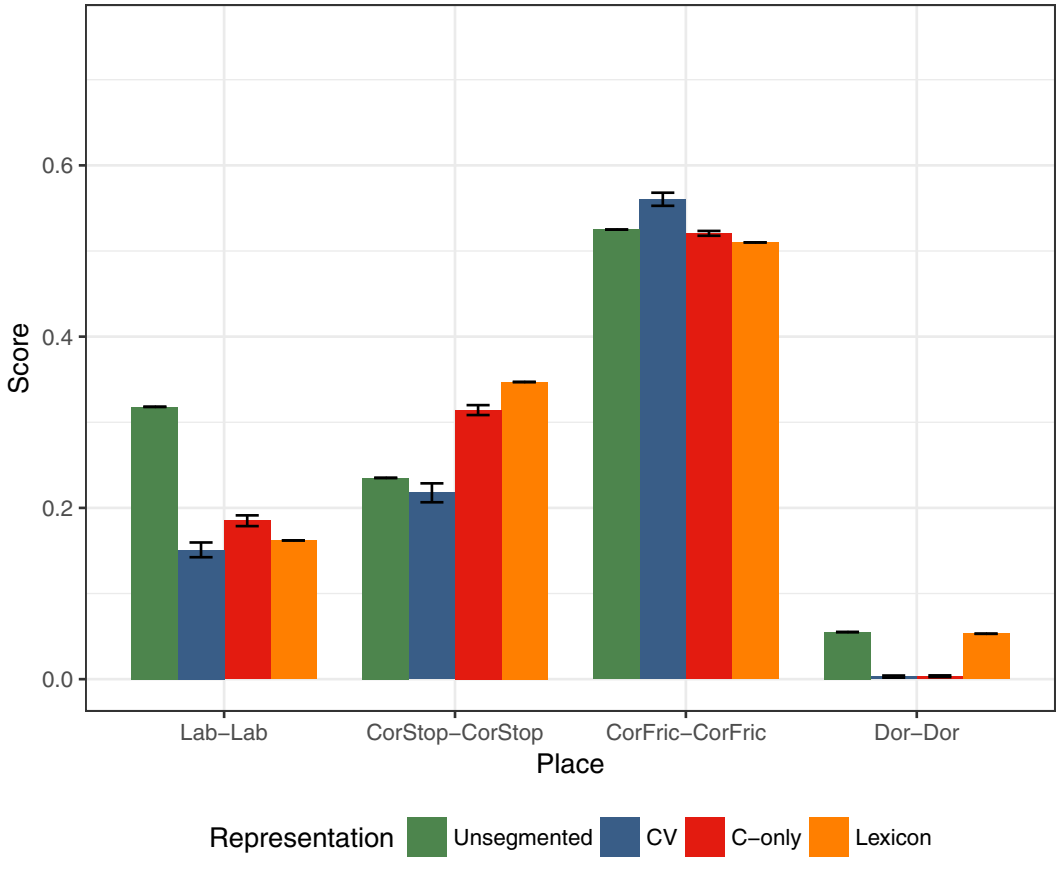

Fig. 4. Experiment 2 results for Arabic Newswire. Averages were calculated over the five runs from Experiment 1 for the CV (full) and C-only representations of the input.

For labials, the true O/E based on the gold standard 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 (±0.007), followed by the full representation (0.151 ± 0.01) 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 (±0.007), with the full representation noting a low O/E of 0.218 (±0.013) 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 (±0.003), ahead of the full representation (0.560 ± 0.009) 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 (±0.001), the full representation 0.003 (±0.001), and the unsegmented baseline 0.06.

The finding that the C-only proto-lexicon is closer to the gold standard than the full CV proto-lexicon is significant for coronal stops ($p < .008$) and coronal fricatives ($p < .02$). Neither the C-only nor the full representation are significantly closer than the other to the labial result, but they are different from one another ($p < .008$). These findings indicate that the statistical patterns in the C-only proto-lexicon provide a closer match to the gold standard lexicon for coronals, but there are no differences between the C-only and full proto-lexicons for labials and dorsals.

Results for the Emirati CDS corpus are presented in Fig. 5. For the three labial consonants of Emirati Arabic, the O/E ratio in the gold standard lexicon is 0.128, based on 162 observations. For the coronal stops, the O/E ratio in the lexicon is 0.09 based on 128 observations, and for coronal fricatives the O/E ratio is 0.496, based on 1,181 observations. Turning to the dorsals, we find an O/E ratio of 0.062, based on a low count of seven observations. In this corpus neither the C-only nor the full proto-lexicon provides a close match to the gold standard lexicon, and the C-only and full representations are not significantly different from each other for any of the consonant groups.

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").

It should be noted that the results of Experiment 2 were stronger for the Arabic Newswire corpus than for the Emirati CDS corpus. This is a difference with the outcome of Experiment 1: 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

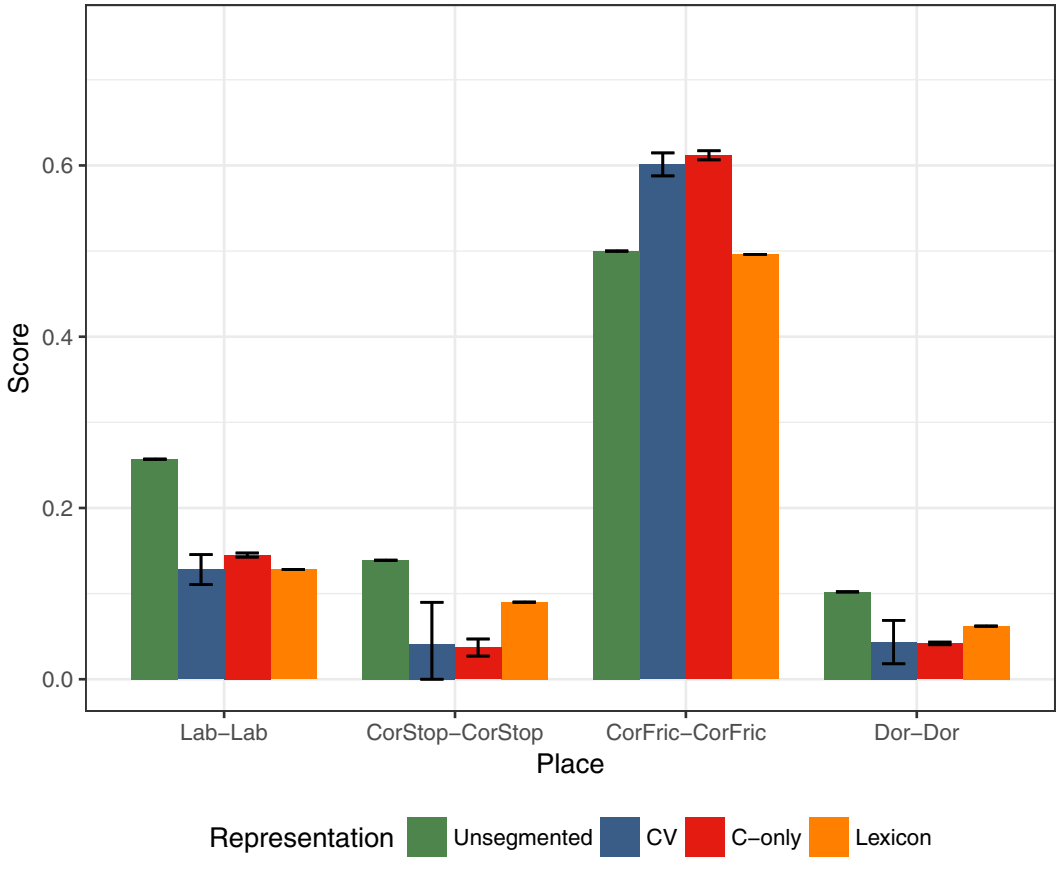

Fig. 5. Experiment 2 results for Emirati CDS.

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, in addition, the removal of vowels supported the learning of roots in both corpora, and it gave rise to accurate OCP patterns in the Arabic Newswire corpus in Experiment 2.

4. General discussion

This 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 cross-linguistically. 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 thus underline the possibility that linguistic constraints might guide and improve the effectiveness of statistical learning (e.g., Adriaans & Kager, 2010; Bonatti et al., 2005; Peperkamp, Le Calvez, Nadal, & 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 & 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 pattern, often innovating new verbal forms (Berman, 1982, 1993; Borer, 2004; Levy, 1988). This finding indicates that children first acquire consonantal roots and then learn the grammatical details of the non-concatenative 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 et al., 2016). It appears that infants start off using consonantal roots in the most frequent pattern, gradually expanding their morphosyntactic system based on the phonological alternations and the additional patterns they are exposed to. By age 4 children seem to have good control not only of the pattern 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 patterns.

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 patterns. The child-directed speech corpus uses the Emirati dialect, which mostly employs only the three or four most frequent patterns. 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, at least when compared to Arabic Newswire.

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 ought to 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, & 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, & Mehler, 1988; Eimas, 1999; Hochmann & Papeo, 2014; Jusczyk & Derrah, 1987; Phillips & 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 faces 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 patterning 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 & 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.

Acknowledgments

Most of the work for this study was carried out while the authors were at New York University. We would like to thank our colleagues there for helpful discussion and comments. This research was supported in part by DFG award AL 554/8-1 to Artemis Alexiadou (I.K.).

Note

1. We also experimented with a model that segments the input based on only the consonants (like the C-only representation), but which then reintroduces the vowels to create a segmentation of the full dataset. The results were highly similar to those of the C-only representation, for both the Arabic Newswire and the Emirati CDS corpus. In both cases overall performance was superior to that of the Full representation.

References

Adriaans, F., & Kager, R. (2010). Adding generalization to statistical learning: The induction of phonotactics from continuous speech. *Journal of Memory and Language*, *62*, 311–331. https://doi.org/10.1016/j.jml.2009.11.007.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*, 119–161. https://doi.org/10.1016/S0010-0277(03)00146-X.

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

Altantawy, M., Habash, N., Rambow, O., & Saleh, I. (2010). Morphological analysis and generation of arabic nouns: A morphemic functional approach. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*, Valletta.

Aronoff, M. (1994). *Morphology by itself: Stems and inflectional classes*. Cambridge, MA: MIT Press.

Bat-El, O. (1994). Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory*, *12*, 571–596. https://doi.org/10.1007/BF00992928.

Benavides-Varela, S., Hochmann, J.-R., Macagno, F., Nespor, M., & Mehler, J. (2012). Newborn's brain activity signals the origin of word memories. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 17908–17913. https://doi.org/10.1073/pnas.1205413109.

Berent, I., & Shimron, J. (1997). The representation of Hebrew words: Evidence from the Obligatory Contour Principle. *Cognition*, *64*, 39–72. https://doi.org/10.1016/S0010-0277(97)00016-4.

Berkley, D. (2000). Gradient obligatory contour principle effects. Unpublished doctoral dissertation, Northwestern University.

Berman, R. (1982). Verb-pattern alternation: The interface of morphology, syntax and semantics in Hebrew. *Journal of Child Language*, *9*, 169–191. https://doi.org/10.1017/S030500090000369X.

Berman, R. (1993). Marking of verb transitivity by Hebrew speaking children. *Journal of Child Language*, *20*, 1–28. https://doi.org/10.1017/S0305000900008527.

Bernstein-Ratner, N. (1987). The phonology of parent-child speech. In K. Nelson & A. van Kleeck (Eds.), *Children's language* (vol. 6, pp. 159–174). Hillsdale, NJ: Erlbaum.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General*, *117*, 21–33. https://doi.org/10.1037/0096-3445.117.1.21.

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. https://doi.org/10.1111/j.0956-7976.2005.01556.x.

Borer, H. (2004). The grammar machine. In A. Alexiadou, E. Anagnostopoulou, & M. Everaert (Eds.), *The unaccusativity puzzle: Explorations of the syntax-lexicon interface* (pp. 288–331). Oxford, UK: Oxford University Press.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*, 93–125. https://doi.org/10.1016/S0010-0277(96)00719-6.

Cairns, P., Shillcock, R., Chater, N., & Levy, J. (1997). Bootstrapping word boundaries: A bottom-up corpus-based approach to speech segmentation. *Cognitive Psychology*, *33*, 111–153. https://doi.org/10.1006/cogp.1997.0649.

Coetzee, A. W. (2005). The obligatory contour principle in the perception of english. In S. Frota, M. Vig′rio, & M. J. Freitas (Eds.), *Prosodies* (pp. 223–245). New York: Mouton de Gruyter.

Coetzee, A. W., & Pater, J. (2008). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory*, *26*, 289–337. https://doi.org/10.1007/s11049-008-9039-z.

Cutler, A., Sebastian-Galles, N., Soler-Vilageliu, O., & Van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory and Cognition*, *28*(5), 746–755. https://doi.org/10.3758/BF03198409.

Daland, R., & Pierrehumbert, J. B. (2011). Learning diphone-based segmentation. *Cognitive Science*, *35*(1), 119–155. https://doi.org/10.1111/j.1551-6709.2010.01160.x.

Doron, E. (2003). Agency and voice: The semantics of the Semitic templates. *Natural Language Semantics*, *11*, 1–67. https://doi.org/10.1023/A:1023021423453.

Eimas, P. D. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustical Society of America*, *105*, 1901–1911. https://doi.org/10.1121/1.426726.

Ferguson, C. F. (1959). Diglossia. *Word*, *15*(2), 325–340. https://doi.org/10.1080/00437956.1959.11659702.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Frisch, S. A., Pierrehumbert, J. B., & Broe, M. B. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory*, *22*, 179–228. https://doi.org/10.1023/B:NALA.0000005557.78535.3c.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-place in Arabic. *Language*, *77*(1), 91–106. https://doi.org/10.1353/lan.2001.0014.

Geman, S., & Geman, R. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741. https://doi.org/10.1109/TPAMI.1984.4767596.

Goldsmith, J. A. (1976). Autosegmental phonology. Doctoral dissertation, MIT, Cambridge, MA.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54. https://doi.org/10.1016/j.cognition.2009.03.008.

Graff, D. (2003). *Arabic gigaword corpus*. Philadelphia, PA: Linguistic Data Consortium.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, *5*, 162–181. https://doi.org/10.1080/00437956.1950.11659378.

Gwilliams, L., & Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. *Brain and Language*, *147*, 1–13. https://doi.org/10.1016/j.bandl.2015.04.006.

Habash, N., & Rambow, O. (2006). Magead: A morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics* (pp. 681–688).

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*(3), 379–440. https://doi.org/10.1162/ling.2008.39.3.379.

Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Consonants and vowels: Different roles in early language acquisition. *Developmental Science*, *14*(6), 1445–1458. https://doi.org/10.1111/j.1467-7687.2011.01089.x.

Hochmann, J.-R., & Papeo, L. (2014). The invariance problem in infancy a pupillometry study. *Psychological Science*, *25*, 2038–2046. https://doi.org/10.1177/0956797614547918.

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*(4), 548–567. https://doi.org/10.1006/jmla.2000.2755.

Jusczyk, P. W., & Derrah, C. (1987). Representation of speech sounds by young infants. *Developmental Psychology*, *23*, 648–654. https://doi.org/10.1037/0012-1649.23.5.648.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207. https://doi.org/10.1006/cogp.1999.0716.

Kager, R., & Shatzman, K. (2007). Phonological constraints in speech processing. In B. Los & M. van Koppen (Eds.), *Linguistics in the Netherlands 2007* (pp. 100–111). Amsterdam: John Benjamins.

Kastner, I. (2016). Form and meaning in the Hebrew verb. Unpublished doctoral dissertation, New York University, New York, NY. (lingbuzz/003028).

Kawahara, S., Ono, H., & Sudo, K. (2006). Consonant cooccurrence restrictions in Yamato Japanese. *Japanese/Korean Linguistics*, *14*, 27–38.

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning? Commentary on Bonatti, Peña, Nespor, and Mehler (2005). *Psychological Science*, *18*(10), 922–923.

Leben, W. R. (1973). Suprasegmental phonology. Doctoral dissertation, MIT, Cambridge, MA.

Levy, Y. (1988). The nature of early language: Evidence from the development of Hebrew morphology. In Y. Levy, I. M. Schlesinger, & M. D. S. Braine (Eds.), *Categories and processes in language acquisition* (pp. 73–98). Hillsdale, NJ: Lawrence Erlbaum.

Lignos, C., & Yang, C. (2010). Recession segmentation: Simpler online word segmentation using limited resources. In *Proceedings of the fourteenth conference on computational natural language learning* (pp. 88–97).

Mattys, S. L., & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, *78*, 91–121. https://doi.org/10.1016/S0010-0277(00)00109-8.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, *12*, 373–418.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, *17*, 207–263.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica*, *43*, 84–108. https://doi.org/10.1159/000261820.

McCarthy, J. J. (1989). Linear order in phonological representation. *Linguistic Inquiry*, *20*, 71–99.

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. https://doi.org/10.1418/10879.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28. https://doi.org/10.1207/s15516709cog1401 2.

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, *48*(2), 127–162. https://doi.org/10.1016/S0010-0285(03)00128-2.

Ntelitheos, D., & Idrissi, A. (2015). Language growth in child Emirati Arabic. In *The 29th annual symposium on Arabic linguistics*. The University of Wisconsin, Milwaukee, WI.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., & Roth, R. M. (2014). MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Language Resources and Evaluation Conference 2014 (LREC)*, Reykjavik.

Peperkamp, S., Le Calvez, R., Nadal, J.-P., & Dupoux, E. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*, B31–B41. https://doi.org/10.1016/j.cognition.2005.10.006.

Peters, A. M. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Phillips, L., & Pearl, L. (2012). "Less is more" in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In N. Miyake, D. Peebles, & R. P. Cooper, (Eds.), *Proceedings of the 34th annual conference of the Cognitive Science Society (COGSCI 34)* (pp. 863–868). Austin, TX: Cognitive Science Society.

Phillips, L., & Pearl, L. (2015a). Evaluating language acquisition strategies: A cross-linguistic look at early segmentation. MS., UC Irvine.

Phillips, L., & Pearl, L. (2015b). The utility of cognitive plausibility in language acquisition modeling: Evidence from word segmentation. *Cognitive Science*, https://doi.org/10.1111/cogs.12217.

Pierrehumbert, J. B. (1993). Dissimilarity in the Arabic verbal roots. In A. Schafer (Ed.), *Proceedings of the North East Linguistics Society* (vol. 23, pp. 367–381). Amherst, MA: GLSA.

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 421–435.

Ravid, D., Ashkenazi, O., Levie, R., Ben Zadok, G., Grunwald, T., Bratslavsky, R., & Gillis, S. (2016). Foundations of the early root category: Analyses of linguistic input to Hebrew-speaking children. In R. Berman (Ed.), *Acquisition and development of Hebrew: From infancy to adolescence* (pp. 95–134). Amsterdam: Benjamins.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month old infants. *Science*, *274*, 1926–1928. https://doi.org/10.1126/science.274.5294.1926.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132. https://doi.org/10.1016/j.cogpsych.2004.06.001.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716. https://doi.org/10.1037/0012-1649.39.4.706.

Ussishkin, A. (2005). A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory*, *23*, 169–218. https://doi.org/10.1007/s11049-003-7790-8.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63. https://doi.org/10.1016/S0163-6383(84)80022-3.

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. 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 patterns, a verb is derived: *ista**qbal*** "greeted," ***qabil*** "received," *ta**qabbal*** "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: ***qabl**a* "before," ***qabiil**a* "tribe," ***qibl**a* "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.

In his seminal analysis of these patterns, McCarthy (1981, 1989) 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 *ta**q**a**bb**a**l*** "received" is as in Fig. A1.

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 vocalic 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 statistical learning and segmentation.

Fig. A1. Tier-based representations for ***taqabbal***, following McCarthy (1981).