A predictive account of masdar formation for Arabic form I verbs

Lisa Dawdy-Hesterberg
Northwestern University
lisah@u.northwestern.edu

The Arabic masdar (verbal noun) system for form I (underived) verbs is relatively understudied, yet offers important insight into general principles of morphophonological pattern-learning in Arabic. First, there are a large number of patterns available in the system, with as many as 44 patterns cited in classic grammars (Wright, 1988). Second, the range of potential cues to masdar pattern that have been noted in the literature is large, including phonological, syntactic, and semantics features of the verb (ibid), yet the system has been frequently written off as unpredictable (e.g. Grenat 1996). However, to my knowledge, there has been no study of masdar predictability using modern statistical and modeling techniques.

Using a set of 1408 verb-masdar pairs from the *Hans Wehr dictionary of modern written Arabic* (Wehr, 1976), I examined the distribution of masdars across the lexicon, and modeled masdar predictability using morphological and phonological features of the verb. The dictionary set was first filtered using Aralex, a lexical database of Arabic word frequencies (Boudelaa & Marslen-Wilson, 2010). The dataset includes only masdars with token frequency greater than zero. Overall, there are 26 masdar patterns in this set, with 62% of verb types taking the masdar pattern [CaCC]. One major cue noted in the literature is the verb pattern. Form I verbs all have the CV template [CvCvCv], and have three patterns: [CaCaCa], [CaCiCa], and [CaCuCa]. We find that each vowel pattern shows a distinct distribution of masdar forms (Figure 1), with [CaCaCa] verbs most frequently taking the masdar [CaCaCa], [CaCiCa] verbs the masdar [CaCaC], and [CaCuCa] verbs the masdar [CaCaCaT]. Overall, statistics on the verb pattern predict over 80% of masdar forms in the dataset.

In order to examine whether other phonological features of the verb (i.e., the consonants in the verbal root), contribute additional predictive power, I compared three predictive analogical models based on the Generalized Context Model (GCM) (Nosofsky, 1990), which has proven successful in modeling the Arabic noun plural (Dawdy-Hesterberg & Pierrehumbert, 2014). The GCM, generally, predicts the output pattern for an unseen form (test item) based on similarity to known forms (comparison items), weighted by the number of existing forms taking a given pattern. The first model is the GCM, which considers all verbs in the dataset, and uses featural string-edit distance to assess similarity to known forms. The second model is the Pattern-restricted GCM, which uses the same similarity metric as the GCM but considers only verbs with the same pattern as the test item. The third model is the Simple Pattern Match, which selects the most frequent masdar pattern among verbs with the same pattern. Overall, both models using the pattern perform significantly better than the GCM, t(9.64)=27.27, p<0.001. In addition, model accuracy is not significantly different between the two pattern-based models, t(9)=0.94, p=0.37, with both models predicting the correct masdar for roughly 83% of verbs.

In conclusion, the masdar of form I Arabic verbs is not unpredictable. Using a representative dataset of verb-masdar pairs and analogical modeling, I have shown that the masdar for a given verb is best predicted using type statistics on the verb pattern. In addition, I have shown that the segmental features of the verb beyond those defined by the verb pattern do not contribute additional predictiveness for the masdar of a given verb. This provides additional support for the predictive power of type statistics on abstract representations in Arabic morphology, which has also been demonstrated for the noun plural (Dawdy-Hesterberg & Pierrehumbert, 2014).