Building a Predictive Model of the Arabic Plural System

Lisa Hesterberg & Janet Pierrehumbert, Northwestern University
lisah@u.northwestern.edu

The noun plural system in Modern Standard Arabic lies at a nexus of critical issues in morphological processing. In this work, we examine two questions. First, can we account for the Arabic plural system within a default vs. irregular framework where there is a categorical distinction between the regular and the irregular? Our data show that this framework does not fit well with the Arabic system, and a competition-based model is implemented. Second, what types of linguistic information are relevant in the pluralization process? Using this competition-based model, we find that the CV template and the frequency distribution of noun forms are the key components.

In Arabic, there are two plural types: the suffixed (sound) plural and the broken plural, called such because it changes the internal structure of the singular. Although many researchers claim that Arabic is a minority default system, wherein the default regular process (the sound plural) occurs less frequently than the irregular process (McCarthy & Prince 1990), analysis of the Corpus of Contemporary Arabic (Al-Sulaiti 2009) shows that the sound plural is statistically in the majority by type (74%, N=4957) and by token (61%). Given the high rate of broken plurals (26% by type, N=1640), a categorical distinction between the regular and irregular does not fit this system well.

Competition-based models have shown high accuracy for systems like this (Ernestus & Baayen 2003). We implemented four models based on the Analogical Model of Language (Skousen 1993), which uses segmental string-edit distance to determine the similarity of a test form and a group (island) of forms that pattern together, e.g. “spring” -> “sprung”, “fling” ->”flung”. The test form is predicted to pattern like the island with which it has the highest similarity rating.

There are two main factors that we test: frequency distributional information and templatic/segmental information. We achieved the highest accuracy when the model is restricted to islands with the same template, and when it considers all forms in an island rather than the single best match (91%). We then tested an additional model where the test form is predicted to pattern like the largest island with the same CV template. This model’s accuracy (91%) is not significantly different than when segmental information is used, which suggests that the segmental information is not adding to the accuracy.

These overall results show that the raw CV template and the distribution of the morphological islands are the most important factors in pluralization in Modern Standard Arabic. This supports a theory of morphological processing based on competition and not default vs. irregular, and also shows that distributional information plays a large role in this process. This intersection of analogy and template is the critical factor that drives processing, which has broad implications for morphological processing.

References

Al-Sulaiti, L. (2009). Corpus of Contemporary Arabic [data file]. Retrieved from http://www.comp.leeds.ac.uk/eric/latifa/CCA_raw_utf8.txt

Ernestus. M. & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language, 79*, 1, 5-38.

McCarthy, J., & Prince, A. (1990). Foot and word in prosodic morphology: The Arabic broken plural. *Natural Language and Linguistic Theory*, *8*, 209-283.

Skousen, R. (1993). Analogy and structure. Dordrecht: Kluwer.