

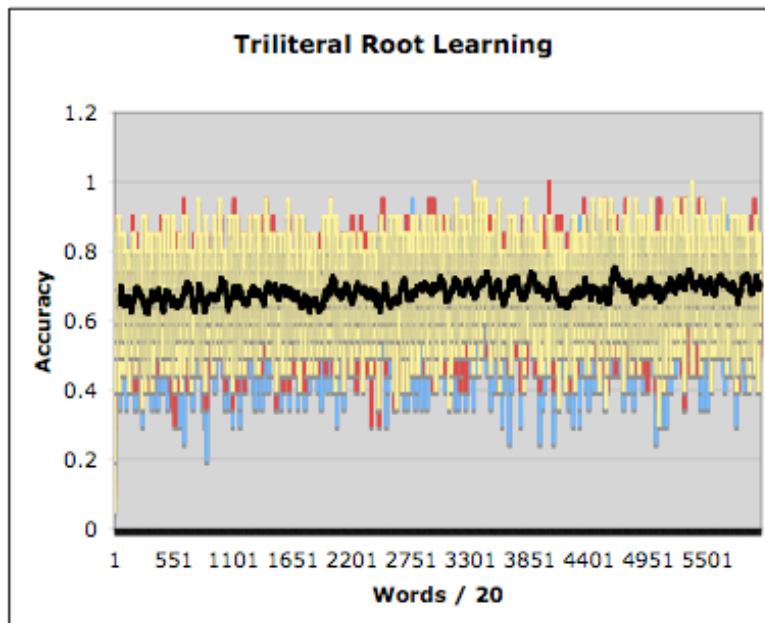
## Statistical Learning of Semitic Morphology Using Autosegmental *Orthography*

Paul Rodrigues  
Indiana University

### Abstract

The root and pattern system, as well as the system of reduplication, are essential to the morphological analysis of Arabic words. (McCarthy 1979, 1981) Few computational morphology systems have been designed to parse concatenative morphology, as well as roots and reduplication simultaneously, without the help of a dictionary. By using simple statistics, we show an algorithm that can learn both the concatenative morphology as well as the roots and template. This paper shows an approach that is analogous to the tier-based autosegmental approach developed by Goldsmith (1976), and applied to Semitic languages in McCarthy (1979).

The root system is learned by comparing frequency statistics. Evidence is weighed for and against a trilateral being declared as the root. Positive evidence includes: the summation of the ratios between a letter being a root and being an affix, the summation of the frequency that a letter has shown up as a possible root, and the summation of the probabilities that the letter belongs to the root. This is then divided by the negative evidence: the summation of the probabilities that the letter is an affix, and the summation of the frequency that a letter has shown up as a possible affix yielding a "score" for the trilateral. (Elghamry, 2004) The trilateral that has the highest score is determined to be the root of the word. The chart below shows that this algorithm learns the trilateral root with nearly 70% precision.



All characters within the first and last radicals is then replaced by a placeholder, *X*. All characters to the left and right of this *X* are now considered to be concatenative morphemes.

To process the concatenative morphology, we used Alignment-Based Learning (ABL) (van Zaanen, 2000) to generate the hypotheses. The information theoretic algorithms of Description Length (DL), Relative Entropy (RE), and Mutual Information (MI) were then used to constrain the hypotheses. DL shows how much our grammar would change if we considered a substring to be a morpheme. RE also measures the cost of divergence, but additionally factors in the frequency of a morpheme relative to the frequency of other morphemes. MI shows the dependency and predictability of one substring to another. By maximizing the MI, minimizing both the DL and the RE, and maximizing the length of a morpheme, the optimal grammar is chosen. (Ćavar 2004, 2005)

The reduplication system is learned by generating hypotheses representing every substring match within a word. A skeleton of the pattern structure in "abc" form is created for each possible parse, and the hypothesis of shortest length wins. For example, galal returns two results, [abcbc, abb], and *abb* wins, because it is the shorter of the two.

We show that with these simple statistics, root lateral identification reaches a high precision, and concatenative morphology identification yields a high precision and a low recall.

We conclude that these simple statistics can be used to identify many semantic and syntactic features of previously unseen Arabic words. We also conclude that the autosegmental approach to Semitic works for orthography.

Training and evaluations were performed on the Arabic Treebank and verified on the Buckwalter Arabic Morphological Analyzer Database.

## Bibliography

Buckwalter, Tim. *Buckwalter Arabic Morphological Analyzer Version 1.0* LDC2002L49 FTP FILE. Philadelphia: Linguistics Data Consortium, 2002.

Elghamry, Khaled. 2004. *A Constraint-based Algorithm for the Identification of Arabic Roots*. Proceedings of the Midwest Computational Linguistics Colloquium. Indiana University. Bloomington, IN. 2005

Damir Čavar, Paul Rodrigues, Giancarlo Schrementi. *Unsupervised Morphology Induction for Part-of-Speech Tagging*. U. Penn Working Papers in Linguistics. Volume 10.1, 2005. Philadelphia, PA, USA.

Damir Čavar, Joshua Herring, Toshikazu Ikuta, Paul Rodrigues, Giancarlo Schrementi. *On Induction of Morphology Grammars and its Role in Bootstrapping*. Proceedings of the 9th Conference on Formal Grammar 2004. Nancy, France. August 2004.

Yu Hu, Irina Matveeva, John Goldsmith, and Colin Sprague. 2005. *Using Morphology and Syntax Together in Unsupervised Learning*. Ms. University of Chicago.

Goldsmith, John. 1976. *Autosegmental Phonology*. Ph.D Dissertation. MIT. 1976

Goldsmith, John. 2001. *The unsupervised learning of natural language morphology*. Computational Linguistics 27(2): 153-198.

McCarthy, John J. 1981. *A Prosodic Theory of Nonconcatenative Morphology*, Linguistic Inquiry 12:373-418.

McCarthy, John J. 1979. *Formal Problems in Semitic Phonology and Morphology*, University of Texas--Austin, Doctoral Diss. Distributed by the Indiana University Linguistics Club. Bloomington, IN. 1982

van Zaanen, Menno Matthias. 2000. "ABL: Alignment-Based Learning", Proceedings of the 18th International Conference on Computational Linguistics, Saarbruecken, Germany.