A Purely Surface Oriented Approach to Handle Arabic Morphology

Yousuf Aboamer and Marcus Kracht
Department of Mathematical and Computational Linguistics
Bielefeld University

We introduce a complete lexicalist approach to deal with Arabic morphology. This purely surface oriented treatment is a part of a comprehensive mathematical approach, referred to as "agreement morphology", to integrate Arabic syntax and semantics using overt morphological features in the string-to-meaning translation. Here we report the first step in our work which will end up with a morphosemantic lexicon for a fragment of Arabic. To maintain the surface orientation, we deal with the language as a sequence of strings, which are allowed only to be concatenated or duplicated, but no rule can delete, add or modify any string.

Arabic morphology is very well known for its complexity and richness. The word formation in Arabic poses real challenges because words are derived from roots, which are supposed to bear the core meaning of their derivatives, by inserting vowels and may be other consonants. Let us consider the following example: the triliteral (3 consonant) root/ك ت ب/ [k t b] is supposed to bear the meaning of writing and from which words like /كتب/ [kataba, he wrote or kutiba, was written], /كاتب/ [kAtib, writer or kAtab, correspond with] and /مكتوب/ [makotUb, is written] are derived.

In this approach, we deal with language as a sequence of glued strings rather than only only strings. A glued string is a string that has left and right context conditions. Optimally morphs are combined in a definite and non-exceptional linear way, as in many cases in different languages (e.g. plural in English). The process of Arabic word formation is rather complex; it is not just a sequential concatenation of morphs by placing them next to each other. But the constituents are discontinuous. Vowels and more consonants are inserted between, before and after the root consonants resulting in what we call "fractured glued string"; i.e. as a sequence of glued strings combined in diverse ways; forward concatenation, backward concatenation, forward wrapping, reduction, forward transfixation and, beyond the MCFGs reduplication.