

Arabic Morphology Parsing Revisited

Suhel Jaber¹ and Rodolfo Delmonte¹

¹ University Ca' Foscari, Dept. Language Sciences, Laboratory Computational Linguistics,
Ca' Bembo, Dorsoduro 1705, 30123 Venezia, Italy
{jaber, delmont}@unive.it

Abstract. In this paper we propose a new approach to the description of Arabic morphology using 2-tape finite state transducers, based on a particular and systematic use of the operation of composition in a way that allows for incremental substitutions of concatenated lexical morpheme specifications with their surface realization for non-concatenative processes (the case of Arabic templatic interdigitation and non-templatic circumfixation).

Keywords: Arabic, morphology, non-concatenative, finite state, composition.

1 Introduction

In this paper we propose a new approach to the description of Arabic morphology using 2-tape finite state transducers, based on a particular and systematic use of the operation of composition in a way that allows for incremental substitutions of concatenated lexical morpheme specifications with their surface realization for non-concatenative processes (the case of Arabic templatic interdigitation and non-templatic circumfixation). Then we compare it with what in our opinion represents the state-of-the-art among the 2-tape finite-state implementations, that of Xerox [1], which is mainly based on the operation of intersection. We intentionally limit ourselves to the evaluation of 2-tape strictly finite-state implementations for this paper, leaving out n-tape implementations such as [2] and [3], and those based on extended finite-state automata, such as [4]. In any case we believe that our approach could be trivially adapted to n-tape implementations as well.

In this paper we argue that:

1. the use of composition allows to overcome certain technical problems inherent to the use of intersection;
2. the method of incremental substitutions through compositions allows for an elegant description of all main morphological processes present in natural languages including non-concatenative ones in strict finite-state terms, without the need to resort to extensions of any sort;
3. our approach allows for the most logical encoding of every kind of dependency, including traditional long-distance ones (mutual exclusiveness), circumfixations and idiosyncratic root and pattern combinations;