Graphemic ambiguous queries on Arabic-scripted historical corpora

Alicia González Martínez
Hamburg University
Edmund-Siemers-Allee 1, 20146 Hamburg
alicia.gonzalez@uni-hamburg.de

Abstract

Arabic script is a multi-layered orthographic system that consists of a base of archigraphemes, roughly equivalent to the traditional so-called *rasm*, with several layers of diacritics. The archigrapheme represents the smallest logical unit of Arabic script; it consists of the shared features between two or more graphemes, i.e., eliminating diacritics. Archigraphemes are to orthography what archiphonemes are to phonology. An archiphoneme is the abstract representation of two or more phonemes without their distinctive phonological features. For example, in Spanish, occlusive consonants loose their distinctive feature of sonority in syllabic coda position; the words *adjetivo* 'adjective' [aDxe'tiβo] and *atleta* 'athlete' [aD'leta] both shared an archiphoneme [D] (in careful speech) in their first syllable, corresponding to the phonemes /d/ and /t/ respectively. In some cases, the neutralisation of two phonemes may cause two words to be homophones. For example, *vid* 'vine' and *bit* 'bit' are both pronounced as [biD]. In paleo-orthographic Arabic script, consonant diacritics were not written down in all positions as it happens in modern Arabic script, where they are mandatory. Consequently, homographic letter blocks were quite common. An additional characteristic of early Arabic script is that graphemic or logical spaces between words did not exist: Arabic orthography preserved the ancient practice of scriptio continua, in which script tries to represent connected speech. Diacritics are signs placed in relation with the archigraphemic skeleton. From a functional point of view, there are two basic types of diacritics: a layer of consonant diacritics for differentiating graphemes and a second layer for vowels. In early script, diacritics are marked in a different colour from the one of the skeleton. Strokes were used for consonant diacritics, whereas dots were used for indicating vowels. In modern Arabic script, dots are instead used for consonant diacritics and they are mandatory. On the other hand, vowels are marked by different types of symbols and are usually optional. Unicode, the standard for digital encoding of language information, evolved from a typographic approach to language and its main concern is modern script. Typography is a technique to reproduce written language based on surface shape. As a consequence, it represents an obstacle for dealing with script from a linguistic point of view, since the same logical grapheme may be rendered using different glyphs. The main problems that arise are the following: 1. Only contemporary everyday use is covered, and that with a typographical approach: Unicode encodes multiple Arabic letters (archigraphemes + consonant diacritics) as single printing units. 2. Some calligraphic variants for the same letter were allowed to have separate Unicode characters. In practice, this means that a search for an Arabic word may yield nothing when typed in a Persian or an Urdu keyboard. This is also why you may find only a fraction of all the results when searching in an Arabic text. 3. There are currently no specialised tools that allow scholars to perform searches on Arabic

Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjuction with RANLP-2019, pages 1–2,
Varna, Bulgaria, Sep 5, 2019.
http://doi.org/10.26615/978-954-452-059-5_001

historical orthography: archigraphemes. Additionally, in order to study early documents written in Arabic script, we need to have search tools that can handle continuous archigraphemic representation, i.e., Arabic script as a scripto continua. In collaboration with Thomas Milo from the Dutch company DecoType, we have developed a search utility that disambiguates and normalises Arabic text in real time and also allows the user to perform archigraphemic search on any Arabic-scripted text. The system is called Yakabikaj (traditional invocation protecting texts against bugs), and show the new perspectives it opens for research in the field of historical digital humanities for Arabic-scripted texts.