COBHUNI Lecture

ERC-Project | Universität Hamburg
COBHUNI
Contemporary Bioethics and the History of the Unborn in Islam

in cooperation with

DECO type
linguistic experts & designers of computer-aided typography

Making search work in Arabic-scripted text

Python Library Release: ya-kabikaj

Lecture by Thomas Milo and Dr. Alicia González Martínez

You are doing a web search for a known Arabic phrase, but you can't find it. Did that ever happen to you? If you did find it, do you realise that you could have had many more hits than the ones that you saw? Incomplete search results are just the tip of the digital iceberg. In practice, the potential of academic research is limited by conceptual constraints.

The reason is that the standard for digital encoding of language information, Unicode, evolved from a typographic approach to language. This is problematic because typography is a technique to reproduce images of writing that stems from the 15th century, when nobody could possibly foresee today's information technology. There is no longer a need to deal with language as typography.

Being a collective effort, Unicode is the sum of its contributions. In Arabic studies, scholars have been acting as competent consumers rather than as contributors to fundamental functionality: we are able to work wonders even with dysfunctional software. However, we are facing two serious problems: 1. Only contemporary everyday use is covered, and that with a typographical approach: Unicode encodes multiple Arabic letters (bases + points) 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 with an Arabic web search.

It is Unicode that made the internet a truly global network. Because of its collective nature and its architecture, it is possible to make it even better. This is the opportunity for a contribution from the field of Arabic studies: disambiguated, normalised Arabic Unicode. As a first step, COBHUNI and DecoType have developed a search utility that disambiguates and normalises Arabic text in real time. In passing, we added a novel feature to handle optional diacritics.

Thursday, February 21st 2019, 2:15 pm
Edmund-Siemers-Allee 1, Room 136

20146 Hamburg

www.cobhuni.uni-hamburg.de

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Horizon 2020
European Union funding
for Research & Innovation

European Research Council
Established by the European Commission