This work examines the use of Tunisian Arabic encoded in Arabizi, a non-standard orthographic system using the Latin alphabet and numbers, which emerged in digital contexts and transformed written communication. Traditionally considered primarily oral, Tunisian Arabic has seen increased use on social media over the past two decades, challenging its conventional definition.

This study examines whether Arabizi is influenced by the specific writing context (blogs, forums, and social networks) and how it facilitates the use of French vocabulary during code-mixing. Indeed, one of the key aspects of the study concerns the quasi-oral nature of Arabizi, considered a system that allows independence from writing traditions, such as that of the Arabic alphabet.

These analyses are corpus-based, and in particular, the Tunisian Arabish Corpus (TArC) was created to observe these linguistic dynamics. TArC collects texts produced over ten years, comprising 43,327 words with various levels of linguistic annotation, allowing for a detailed study of the language. The hybrid methodology adopted to build TArC combines approaches from Arabic dialectology, Corpus Linguistics, and Natural Language Processing. TArC has been annotated using semi-automatic procedures, making it useful for NLP research as well. Finally, the work examines the challenges and limitations of interdisciplinary research, proposing preliminary hypotheses on the linguistic and sociolinguistic trends related to the use of Arabizi in Tunisia. These observations aim to mark a starting point for future research in the field of dialectology and linguistic technologies applied to Tunisian Arabic.

Scarica il PDF in Open Access