Periodizing Written Arabic Using Automated Methods and Digital Philology

Arabic is traditionally divided into only two major linguistic periods - 'Classical Arabic,' the language of formal texts from the pre-modern corpus and 'Modern Standard Arabic,' a variety of Arabic that was transformed by interactions with European written norms and by the project of the Arabic *nahḍa* (Newman 2013). However, while this periodization is frequently taken for granted, it has not necessarily been tested in a principled way, nor has there been any attempt to divide Arabic further into other periods.

The availability of large collections of Arabic texts allow for new avenues of automated and semi-automated inquiry into the history of the Arabic language. In this talk, we explore the periodization of Arabic using an 800 million word text collection based on the texts from Al-Maktaba Al-Shamela, employing a variety of tools such as word vector comparison, lemmatization and concordancing. We show that in general, the Arabic lexicon changes much more slowly than in English, with words in Arabic lasting on average for 52% of the time from their first appearance to the present, but in English lasting only 41% of that window on average. However, we show that there is significant evidence for a rapid change in the Arabic language in the 19th century based on the increased incidence of novel words and change in the usage of existing words coinciding with this time period, suggesting that there is in fact a disjunction between Classical Arabic and Modern Standard Arabic, even when genre is controlled.

We also explore the periodization of Arabic in the pre-modern era. While there are not such clear changes in the pre-modern record as in the modern record, there does seem to be a period of change which occurs from the 8th to the 11th hijri centuries which may coincide with a change from literary production centered in Iraq and the Levant to Egypt following the Mongol invasions and the rise of the Mamlukes. For example, the proximal locative demonstrative *hā hunā* "here, at this place" is the dominant form in texts through the 8th Islamic century, when it is supplanted by *hunā*. The former is more congruent to the form used the Levant and Iraq, while the latter is more similar to the form used in Egypt. We also explore the changes over the lifespans of individual words and show how there are significant changes in word usage and meaning even in the pre-modern period, though they do not necessarily cluster as closely as the transition from Classical to Modern Standard Arabic, belying some perceptions of Classical Arabic as an immutable entity.

Al-Maktaba Al-Šāmila. http://shamela.ws/. Texts accessed via version curated by Maxim Romanov, Universität Leipzig. Exactly date of download unclear.

Newman, Daniel. 2013. "The Arabic Literary Language: The Nahḍa (and Beyond)." In *The Oxford Handbook of Arabic Linguistics*, edited by Jonathan Owens. Oxford: Oxford University Press.

We are indebted to Maxim Romanov for making available to us a user-friendly version of the Shamela text collection.

Subfields: Computational Linguistics, Historical Linguistics