Lexical changes in Arabic newspaper writings: A corpus-based comparison of 2018 and 1950 Arabic newspapers

Sarah Alajlan,
School of Languages, Cultures and Societies,
University of Leeds, UK

Claire Brierley,
School of Languages, Cultures and Societies,
University of Leeds, UK

Abstract

Studying language change is one of the most challenging topics in the field of corpus linguistics. The challenge stems from two facts. First, there is a shortage of ancient texts in a machine-readable format, and second there is the difficulty of identifying linguistic changes electronically across two or more corpora. This paper attempts to shed light on lexical changes that have happened in Arabic newspaper writings by comparing two sub-corpora that have been built specifically for this study (Alajlan, 2019). It is part of an ongoing PhD study that uses automatic analysis followed by manual investigation of the lexicon and the syntax of Arabic newspaper writings.

The corpus texts were collected from three Arabic newspapers published in two Arabic countries. Al-Ahram was chosen to represent Arabic newspaper writing in Egypt in 1950 and in 2018. Alriyadh was chosen to represent Arabic newspaper writing in Saudi Arabia in 2018. Because Alriyadh was not published back in 1950, Umm Al-Qurā was selected to represent Arabic newspaper writing in 1950.

New methodology has been presented for the purpose of identifying lexical changes. At the beginning of the analysis, word frequency lists were generated using the Sketch Engine toolkit in order to obtain a clear picture of the nature of linguistic data in the corpus. Six lists were generated for each sub-corpus: (1) the entire 1950 sub-corpus list; (2) the entire 2018 sub-corpus list; (3) the sub-corpus of Al-Ahram 1950 list; (4) the sub-corpus of Umm Al-Qurā 1950 list; (5) the sub-corpus of Al-Ahram 2018; and (6) the sub-corpus of Alriyadh 2018.

Since the lists were too long (approximately between 1,800 words and 8,000), manual analysis was not a good option. Instead, analysis of word lists was done via a Python script. The set.intersection() method was used to return the set of shared words between any two input lists, and the set.difference() method was used to identify the set of words unique to each of two input lists. Three return lists were generated for each comparison: the entire 1950 versus the entire 2018 corpus; the 1950 versus the 2018 Al-Ahram sub-corpus; and the 1950 Umm Al-Qurā versus the 2018 Alriyadh sub-corpus.

The results show that there have been visible changes in the lexicon of Arabic newspapers. The changes were classified into five categories: (1) increase or decrease in the frequency of occurrence; (2) semantic changes; (3) changes in form; (4) new emergence; and (5) disappearance of lexical items. Automatic extraction of frequent words and identifying shared and unique words in both sub-corpora was not sufficient for identifying the semantic changes especially in the intersection lists; manual investigation was needed to highlight some changes that happened. The manual analysis reveals some linguistic trends that were noticed in semantics of some lexical items. The phenomenon of “monosemisation” was noticed in this

study where semantic shift has happened with respect to some polysemic terms such that they have become monosemic.

References

Alajlan, S. (2019). *Compiling a Diachronic Corpus of Arabic Newspaper Texts: Methodology and Challenges.* Presented at: CL2019: International Corpus Linguistics Conference, Cardiff, Wales, UK, 22-26 July 2019.