The Saudi Novel Corpus
Progress of Compilation and Preliminary Results

Dr. Tareq Alfraidi

Islamic University of Madinah, Saudi Arabia

Abstract

There are several specialized Arabic corpora available that are built for various purposes, and they represent different genres. Nonetheless, online newspapers have been the dominant genre among the existing Arabic corpora (Al-Thubaity 2015, El-Khair, 2016). The Arabic novel genre, especially the Saudi novel, however, has been largely overlooked. This has led to the absence of Arabic corpus stylistics research. Such a lacuna has motivated us to create the first version of the Saudi Novel Corpus (SNCorpus) to contribute to the analysis of Saudi literary works and facilitate stylistic and linguistic studies in this area (Alfraidi et al., 2022).

In this paper, we aim to (i) present the procedures for expanding the current version of the corpus and (ii) report some results that emerged from the analysis of the content of the corpus. This is in order to demonstrate the value of using the corpus approach to exploring Arabic literature, especially Saudi novels.

To peruse the first aim, we will present the recent advancements achieved to improve the corpus quality. We will give particular attention to presenting the data collection procedure, text pre-processing, annotation, and interface programming.

The main recent additions to the corpus so far include:

1. Increasing the size from 3M words to 5M words.
2. Annotating the words of the corpus with POS tags.
3. Building a web interface that facilitates the search in the content of the corpus.

To fulfill the second aim, we will present the main results that emerged from the empirical investigation of the corpus. Using the functions available in the web interface, we generated a word list of the top 100 words and keywords. This has been followed by examining the concordance lines of the top ones. As a result, we managed to spot several textual patterns.

Our hope is that this work will prove valuable to the Arabic stylistics and linguistics research communities and bridge the gap between the fields of corpus linguistics and Arabic literature.

Keywords: corpus linguistics, saudi novels, stylistics, wordlist, keywords