Comparative Evaluation of Tools for Arabic Corpora Search and Analysis

Abdullah Alfaifi
University of Leeds
scayga@leeds.ac.uk

Eric Atwell
University of Leeds
e.s.atwell@leeds.ac.uk

Abstract

Learner corpora are considered one of the resources used in language learning and teaching (Nesselhauf, 2004) and second language acquisition (Granger, 2003). The Arabic Learner Corpus is a recently developed project that includes 282,732 words of written and spoken materials, collected from 942 learners of Arabic in Saudi Arabia. Different tools have been investigated in order to evaluate their functionalities and suitability for searching and analysing Arabic corpora, such as the KACST Arabic Corpora Processing Tool [Al-thubaity, 2014], aConCorde [Roberts, 2014], AntConc [Anthony, 2005], WordSmith Tools [Scott, 2012], The Sketch Engine [Kilgarriff et al., 2004], IntelliText Corpus Queries [Sharoff, 2011], CQPweb [Hardie, 2008] and others. This paper aims to cover three aspects: (1) an overview on learner corpora uses in language research, (2) a brief summary of the second version of the Arabic Learner Corpus, (3) and a particular focus on evaluating search and analysis tools that can be used on the Arabic corpora. This evaluation is based on specific criteria in order to conclude whether the tools provide functions suitable for Arabic, a right-to-left reading language, in order to be effectively searched and analysed. The criteria include, for example, the ability to read Arabic texts, showing right-to-left texts in the correct order and the availability of uploading a personal corpus.

References


http://www.comp.leeds.ac.uk/scayga
http://www.comp.leeds.ac.uk/eric

