World Academy of Science, Engineering and Technology
International Journal of Cognitive and Language Sciences
Vol:12, No:04, 2018

Compilation and Statistical Analysis of an Arabic-English Legal Corpus in Sketch Engine

Authors : C. Brierley, H. El-Farahaty, A. Farhan

Abstract : The Leeds Parallel Corpus of Arabic-English Constitutions is a parallel corpus for the Arabic legal domain. Analysis of legal language via Corpus Linguistics techniques is an important development. In legal proceedings, a corpus-based approach to disambiguating meaning is set to replace the dictionary as an interpretative tool, and legal scholarship in the States is now attuned to the potential for Text Analytics over vast quantities of text-based legal material, following the business and medical industries. This trend is reflected in Europe: the interdisciplinary research group in Computer Assisted Legal Linguistics mines big data collections of legal and non-legal texts to analyse: legal interpretations; legal discourse; the comprehensibility of legal texts; conflict resolution; and linguistic human rights. This paper focuses on ‘dignity’ as an important aspect of the overarching concept of human rights in current constitutions across the Arab world. We have compiled a parallel, Arabic-English raw text corpus (169,861 Arabic words and 205,893 English words) from reputable websites such as the World Intellectual Property Organisation and CONSTITUTE, and uploaded and queried our corpus in Sketch Engine. Our most challenging task was sentence-level alignment of Arabic-English data. This entailed manual intervention to ensure correspondence on a one-to-many basis since Arabic sentences differ from English in length and punctuation. We have searched for morphological variants of ‘dignity’ (□ □□□□, karāma) in the Arabic data and inspected their English translation equivalents. The term occurs most frequently in the Sudanese constitution (10 instances), and not at all in the constitution of Palestine. Its most frequent collocate, determined via the logDice statistic in Sketch Engine, is ‘human’ as in ‘human dignity’.

Keywords : Arabic constitution, corpus-based legal linguistics, human rights, parallel Arabic-English legal corpora
Conference Title : ICCLM 2018 : International Conference on Corpus Linguistics and Methodology
Conference Location : Venice, Italy
Conference Dates : April 12-13, 2018