Equivalent Malay-Arabic data corpus collection

ABSTRACT

This paper aims to introduce a search strategy and collecting comparable sentences of Arab-Malay corpus data. This method was introduced for the use of students, researchers and amateur translators to search and compare the structure of sentences in Arabic and Malay. The first stage is to collect data corpus with high impact titles from the press and must be able to enlarge the scope of study as stated by Maia (2003). The second stage is to search using the specified key words based on selected high-impact titles such as the Football World Cup year 2010 and 2014. Data search is by using Webcorp engine http://www.webcorp.org.uk/live/ corpus and also open database Google https://www.google.com. The third stage is to filter the data by using Aker et.al (2012) and Braschler's (1998) method based on similar story, related story and similar aspects. At the fourth stage every category is measured by Guidere's (2002) equivalence strength which is strong comparability (SC), medium (MC) and weak (WC). At the last stage comparable sentences between the two languages are compiled in parallel according to Mona Baker's (1992) level of grouping which are sentence level, combination of words, grammatical, pragmatic and textual level. The result from data analysis based on Mona Baker and Vinay - Darbelnet's (1995) comparable theory proved the existence of some sentences in large quantities are on the same level of comparability from the point of information delivery. This can be used as the basis of additional evidence concerning the validity of 'universal theory.' in the science of translation.

Keyword: Software; Comparable; Parallel