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

Identification of Text Domains and Register Variation through the Analysis of Lexical Distribution in a Bangla Mass Media Text Corpus

Authors : Mahul Bhattacharyya, Niladri Sekhar Dash

Abstract : The present research paper is an experimental attempt to investigate the nature of variation in the register in three major text domains, namely, social, cultural, and political texts collected from the corpus of Bangla printed mass media texts. This present study uses a corpus of a moderate amount of Bangla mass media text that contains nearly one million words collected from different media sources like newspapers, magazines, advertisements, periodicals, etc. The analysis of corpus data reveals that each text has certain lexical properties that not only control their identity but also mark their uniqueness across the domains. At first, the subject domains of the texts are classified into two parameters namely, 'Genre' and 'Text Type'. Next, some empirical investigations are made to understand how the domains vary from each other in terms of lexical properties like both function and content words. Here the method of comparative-cum-contrastive matching of lexical load across domains is invoked through word frequency count to track how domain-specific words and terms may be marked as decisive indicators in the act of specifying the textual contexts and subject domains. The study shows that the common lexical stock that percolates across all text domains are quite dicey in nature as their lexicological identity does not have any bearing in the act of specifying subject domains. Therefore, it becomes necessary for language users to anchor upon certain domain-specific lexical items to recognize a text that belongs to a specific text domain. The eventual findings of this study confirm that texts belonging to different subject domains in Bangla news text corpus clearly differ on the parameters of lexical load, lexical choice, lexical clustering, lexical collocation. In fact, based on these parameters, along with some statistical calculations, it is possible to classify mass media texts into different types to mark their relation with regard to the domains they should actually belong. The advantage of this analysis lies in the proper identification of the linguistic factors which will give language users a better insight into the method they employ in text comprehension, as well as construct a systemic frame for designing text identification strategy for language learners. The availability of huge amount of Bangla media text data is useful for achieving accurate conclusions with a certain amount of reliability and authenticity. This kind of corpus-based analysis is quite relevant for a resource-poor language like Bangla, as no attempt has ever been made to understand how the structure and texture of Bangla mass media texts vary due to certain linguistic and extra-linguistic constraints that are actively operational to specific text domains. Since mass media language is assumed to be the most 'recent representation' of the actual use of the language, this study is expected to show how the Bangla news texts reflect the thoughts of the society and how they leave a strong impact on the thought process of the speech community.

Keywords : Bangla, corpus, discourse, domains, lexical choice, mass media, register, variation
Conference Title : ICCLM 2018 : International Conference on Corpus Linguistics and Methodology
Conference Location : Venice, Italy
Conference Dates : April 12-13, 2018