INDIAN JOURNAL OF APPLIED LINGUISTICS
VOL. 45, NO. 1-2, JAN-DEC 2019 – (ISSN 0379-0037)

Some Quantitative Findings Obtained from Analysis of a Hindi News Text Corpus

VANDANA
NILADRI SEKHAR DASH
Indian Statistical Institute, Kolkata, *India*

JAYSHREE CHAKRABARTY
Indian Institute of Technology, Kharagpur, India

ABSTRACT

In recent times, corpus-based quantitative approach towards the analysis of languages data has been a new widespread methodology to language study. This has been possible due to the availability of a huge amount of language data in machine-readable format. This present paper is directed towards this goal as it tries to present some important quantitative findings of the patterns of use and distribution of characters and words in a corpus made with Hindi newspaper texts. The Hindi News Text Corpus (HNTC) is developed as a part of ongoing research that tries to address various discourse related issues and phenomenon as revealed in the news texts. The information about the usage and distribution of linguistic elements likes characters and words have a direct role in shaping up the observations and analyses of the discoursal properties in the texts. The HNTC shows that although it contains a large set of characters of different types (i.e. vowels, vowel allographs, consonants, consonants clusters, sound modifiers, and punctuation markers, etc.) their frequency of occurrence, patterns of usage, and nature of distribution are different based on text type or text genre. The present paper desires to look into this issue through analysis of characters and words using some basic corpus-based techniques to reflect on frequency count and distributional patterns of these linguistic properties. Moreover, a few empirical observations are also made to reflect on the