INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

* **Corresponding author**.

kaiser.giri@islamicuniversity.edu.in

Natural Language Processing Resources for the Kashmiri Language

Nawaz Ali Lone[1], Kaiser J Giri[2]*, Rumaan Bashir[2]

1 Research Scholar, Department of Computer Science, Islamic University of Science & Technology, Kashmir
2 Associate Professor, Department of Computer Science, Islamic University of Science & Technology, Kashmir

Abstract

Objectives: The main objective of this paper as a maiden attempt is to identify the basic resources necessary for undertaking Natural Language Processing (NLP) specific research activities pertaining to Kashmiri language. The paper also deliberates on key issues related to Natural Language Processing of Kashmiri language such as complex linguistic phenomena, the lack of standard linguistic tools, documented as well as standardized resources and the influence of some dominant languages mostly Urdu and English on Kashmiri language. **Methods:** As there is no substantial work reported in literature specific to NLP of Kashmiri language, a holistic research strategy was adopted to explore the possible sources as potential means for creation of basic resources to undertake the NLP research for Kashmiri language. **Findings:** After thorough investigation, it was observed that there has been some trivial work reported in the literature related to Machine Translation of Kashmiri language. Further there are few newspapers published in Kashmiri language which can be used as a means for creation of Kashmiri corpus. Moreover crowdsourcing could be used a potential means for development of digital linguistic resources for Kashmiri language. **Novelty:** The present study is a maiden attempt towards identification of NLP resources for Kashmiri language and will be of immense importance for the research community interested to work for development of Kashmiri language in digital domain.

Keywords: Natural Language Processing; Transliteration; Kashmiri Language; Scheduled Languages; crowdsource; Tag set; P-o-S Tagging

1 Introduction

The field of Natural Language Processing (NLP) is an emerging area of research and has thus received significant attention of researchers over past few decades[1]. It consists of diverse language specific research aspects such as Text and Speech Processing, Morphological Analysis, Syntactic Analysis, Lexical Semantics, Relational Semantics, etc. The field of Natural Language Processing has immense research potential owing to its applications across diverse domains such as Education, Healthcare, Military, and