SOLAB

Register Variation in Persian: A Corpus-Driven Study of Slang, Verbs, and Lexical Items across Informal and Formal Texts

Hossein Fallah Yakhdani[1]*, Elham Mizban[2]

1. MA Student, Department of Linguistics, Allame Tabataba'i University, Tehran, Iran
2. PhD in Linguistics, Ferdowsi University of Mashhad, Iran

Abstract

This study explores the rich tapestry of Persian lexical variation by analyzing the contrast between formal written language and the vibrant, ever-evolving vernacular found on social media. The research centers on slang and dialectal expressions that typically escape traditional corpora. It employs a corpus-based methodology that compares the formal Bijankhan Corpus with the informal Large-Scale Colloquial Persian (LSCP) corpus made of Persian tweets. Two major Persian corpora are compared in this study: the formal Bijankhan Corpus and the informal LSCP. Both datasets were tokenized, cleaned, and normalized through rigorous natural language processing (NLP) preprocessing. Frequency analyses were also conducted to uncover lexical items distinctive to each register. Especially attention was given to slang and colloquial terms prevalent in LSCP. This work sheds light on the vocabulary richness found in informal Persian, contributing to a more nuanced understanding of language variation. It also supports the use of different language forms in the NLP pipelines. Integrating such registers promises to improve the accuracy and cultural relevance of Persian language technologies. This comparison of corpora offers valuable insights into Persian lexical variation, emphasizing the need to augment linguistic analysis and enhance NLP tools with more informal language data.

Keywords: Persian lexical variation, Slang analysis, Bijankhan Corpus, LSCP, Corpus linguistics, Register variation in pers

Email: Hfallah.Civil@gmail.com

doi.org/10.30514/icss.27.0.8