Comparative study of Sanskrit and Perso-Arabic loanwords in Modern Hindi with Word Embeddings

Texts contain a vast amount of information about words and their usages and, therefore, part of our knowledge of the world. Since a linguistic corpus is a large collection of texts, it can be assumed to represent — to some extent — the language and its use patterns; but more than that, our collective knowledge, stereotypes, and memory, as embedded in the language. Therefore, linguistic corpora can be seen as exceptionally valuable data from which we can derive cultural understanding.

Furthermore, operating on sufficiently large collections of text using statistical analysis of lexical co-occurence patterns allows us to extract an interpretation of the reality which manifests itself in the language and is closely related to a given cultural context (i.e. certain entrenched cultural beliefs and stereotypes, or at least a trace of them). These realities are sometimes unknown or unnoticed even by native speakers. Moreover, such results can be statistically authoritative, precise and objective, as they are based on extensive amounts of data from diverse sources.

According to the distributional hypothesis, the semantic neighborhood of a word in the form of a word embedding can represent word meanings and how those concepts are perceived and linked to other ideas by a given community. In other words, individual words do not function in isolation, their meaning is associated with the meaning of co-occurring words and the use of them invokes a conceptual framework that is influenced by previous use.

Modern Hindi abounds in parallel use of Perso-Arabic loanwords, which often function similarly to words of Sanskrit origin. Despite their semantic proximity, each of them exhibits different connotational meaning, stylistic distribution and cultural associations and so occupies a slightly different semantic field, with a different range of application.

The purpose of this study is to determine if word embeddings can identify whether lexemes inherited from different languages can embody a different linguistic worldview which can influence their meaning. As a result, even though two words may be synonyms, their embedded cultural meaning can be different, as well as the mental image they invoke.

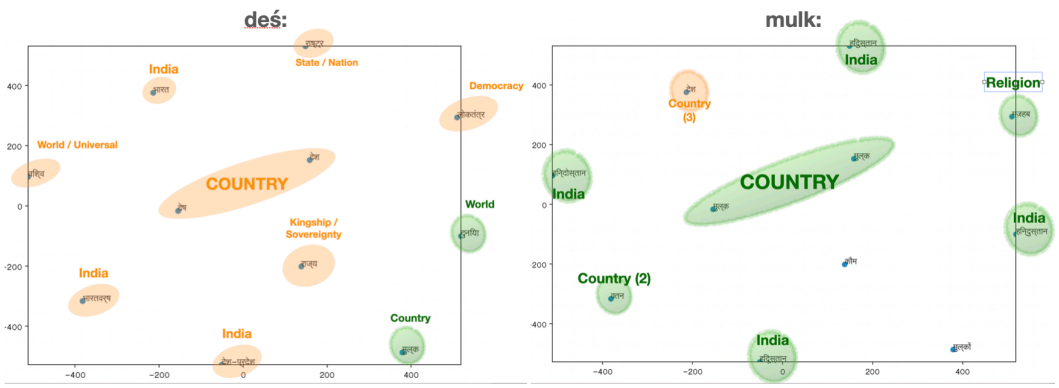

Fig 1.: Comparison of synonyms for the word country: deś (Sanskrit origin) and mulk (Perso-Arabic).