The goal of this project is to compare pairs of lexemes (quasi-synonyms) in contemporary Hindi that are borrowed from either Urdu or Sanskrit, in order to reconstruct the worldview of the language users through their lexical choices, using advanced computational linguistic techniques such as word embeddings.

Hindi is a language with a long and complex history, and consequently, it's vocabulary is borrowed from several other languages, mostly from Sanskrit and – via Urdu – from Persian, Arabic and Turkish. Oftentimes, words borrowed from both Urdu and Sanskrit are used concurrently as pairs of synonyms; however, the choice of either word depends on the context, register and current communication situation. While sometimes the word borrowed from Urdu is preferred over its unnaturally-sounding Sanskrit synonym, in some other contexts it is the Urdusim that would sound oddly. Depending on the social class of a language user, on the level of education the she/he has received, and sometimes even on personal preferences, one variant might dominate over the other.

Consequently, the pairs of lexemes in Urdu and Sanskrit can hardly be referred to as pure synonyms. Despite semantic proximity, each word occupies a slightly different semantic area, and exhibits different connotations, stylistic flavor, as well as frequency. Being so highly dependent on pragmatic factors, then, the words in question should be also reflecting the users' worldview. The above assumption is a departure point for all the investigations to be undertaken in this project – namely, the project is aimed at verifying empirically whether or not a worldview can be reconstructed through large-scale analysis of users' choices between Urdu and Sanskrit quasi-synonyms in Hindi. A large corpus of texts in contemporary Hindi will be compiled, and analyzed using state-of-the-art distributional semantics methodology.

Rooted in the "distributional hypothesis", according to which the degree of semantic similarity between two words (or other linguistic units) can be modeled as a function of the degree of overlap among their linguistic contexts, some sophisticated neural-networks algorithms, e.g. word2vec, allow for identifying semantic shifts between any set of words. Therefore, it can be empirically shown to which extent the quasi-synonyms in Hindi are indeed dependent on contextual usage.

In order to discover systematic shifts in meaning, and consequently, to exhibit traces of language stereotypes, we will collect a relatively large corpus in Hindi. To this end, we plan to acquire both newspapers (from online editions and from digitized archives), as well as literary sources (available on the internet), covering the last decades of the 20th century and the beginning of the 21th century. Next step will involve the analysis of the corpus. We will first analyze a number of quasi-synonym pairs that we have already identified manually. These seed words, representing basic values and meanings, will serve as a baseline in our further investigations.

The significance of the project is threefold. Firstly, the study will systematically describe the system of pairs of quasi-synonyms in Hindi, while keeping in mind that their choice depends on pragmatic factors. Secondly, the project will develop a novel method of tracing semantic shifts between pairs of words borrowed from two different languages: Urdu and Sanskrit. Thirdly, the project will supplement the "worldview in the language" theory with an empirical (quantitative and experimental) dimension, with the hope that it will help to either verify or falsify the theory.