Celem niniejszego projektu jest porównanie leksemów (quasi-synonimów) współczesnego języka hindi pochodzących z dwóch różnych języków, mianowicie z sanskrytu i z urdu, i próba rekonstrukcji językowego obrazu świata użytkowników języka widzianego poprzez ich wybory leksykalne, podjęta przy użyciu metodologii językoznawstwa komputerowego, przede wszystkim metody zanurzeń słów.

Hindi jest językiem o długiej i bogatej historii; jego słownictwo jest zapożyczone z wielu różnych języków, głównie z sanskrytu oraz – za pośrednictwem urdu – z perskiego, arabskiego i tureckiego. Zapożyczenia z sanskrytu oraz z urdu często bywają używane równolegle, jako pary synonimiczne, przy czym wybór jednego bądź drugiego synonimu zależy od kontekstu, rejestru oraz bieżącej sytuacji komunikacyjnej. Czasem wybierane jest słowo pochodzące z urdu, podczas gdy jego odpowiednik sanskrycki brzmiałby sztucznie, a czasem to właśnie lexem o pochodzeniu arabsko-perskim jest nienaturalny w danym kontekście. Wiele w tym względzie zależy od pochodzenia społecznego użytkownika języka, wykształcenia, a czasem nawet od indywidualnych wyborów leksykalnych.

W związku z powyższym pary leksemów pochodzące z urdu albo z sanskrytu trudno nazwać synonimami w tradycyjnym sensie. Mimo bliskości semantycznej oba słowa określa nieco inne pole semantyczne, nieco inne konotacje, a także inny odcień stylistyczny i nierzadko inna częstość użycia. Zależne w ogromnym stopniu od czynników pragmatycznych, owe pary leksemów będą zapewne również nośnikami językowego obrazu świata ich użytkowników. Powyższe założenie badawcze jest punktem wyjścia wszystkich analiz przedsięwziętych w ramach niniejszego projektu – celem mianowicie, jaki zostanie tutaj podjęty, jest próba empirycznej weryfikacji hipotezy, że językowy obraz świata daje się zrekonstruować poprzez wielkoskalową analizę wyborów leksykalnych w obrębie par quasi-synonimów z urdu i sanskrytu. W tym celu zostanie zebrany korpus tekstów współczesnego języka hindi i poddany analizie za pomocą najnowszych metod semantyki dystrybucyjnej.

Osadzona w tzn. hipotezie dystrybucyjnej, zgodnie z którą podobieństwo semantyczne między dwoma słowami (lub innymi jednostkami języka) można modelować jako stopień podobieństwa między kontekstami tych słów, współczesna semantyka dystrybucyjna opiera się na wyrafinowanych algorytmach wykorzystujących sieci neuronowe, np. na algorytmie word2vec i pozwala dostrzec przesunięcia semantyczne pomiędzy dowolnymi zbiorami słów. Dzięki temu można w sposób empiryczny wykazać, czy (i w jakim stopniu) quasi-synonimy w hindi rzeczywiście są zależne od użycia kontekstowego.

Żeby móc odkryć systematyczne przesunięcia semantyczne i wyłuskać z tekstów stereotypy językowe, będziemy musieli zebrać względnie obszerny korpus w hindi. Planujemy pozyskać zarówno czasopisma (ze zbiorów dostępnych online oraz z archiwów cyfrowych), jak i źródła literackie (dostępne przez internet), tak żeby pokryć końcówkę XX i początek XXI wieku. Następnym krokiem będzie analiza korpusu. Zaczniemy od pewnej liczby quasi-synonimów które już zostały wstępnie zidentyfikowane. Te słowa zalążkowe, odwołujące się do podstawowych wartości, będą stanowiły podstawę materiałową i punkt wyjścia do dalszych, półautomatycznych poszukiwań kolejnych par leksykalnych.

Znaczenie niniejszego projektu trzeba rozpatrywać w trzech aspektach. Po pierwsze, niniejsze studium w zamierzeniu ma przynieść systematyczny opis quasi-synonimów we współczesnym hindi, przy szczególnym uwzględnieniu czynników pragmatycznych. Po drugie, projekt wypracuje nową metodę śledzenia przesunięć semantycznych między parami słów zapożyczonych z różnych języków – sanskrytu i hindi. Po trzecie, projekt uzupełni teorię językowego obrazu świata o element empiryczny (eksperymentalny i ilościowy), co z kolei sprawi, że będzie można tę teorię albo zweryfikować, albo sfalsyfikować.