様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １０ 日現在

機関番号：１２６０３
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３２００７４

研究課題名（和文）ことばを測る　―ヒンディー語とウルドゥー語の語彙属性に関する研究―

研究課題名（英文）To what extent are Hindi and Urdu the same language on the base of their vocabularies?

研究代表者
町田　和彦（MACHIDA, Kazuhiko）

東京外国語大学・アジア・アフリカ言語文化研究所・教授

研究者番号：７０１３４７４９

交付決定額（研究期間全体）：（直接経費）　　12,800,000円

研究成果の概要（和文）：本研究は、言語的には同一言語とされるヒンディー語（インドの公用語、デーヴァナーガリー文字）とウルドゥー語（パキスタンの国語、アラビア系文字）それぞれに含まれる語彙の語源別含有比率をコンピュータを使って計測した。
対象とした19世紀から21世紀における散文に関しては、使用頻度上位８０％を占める語彙に関しては両言語において事前の想定よりも差がなく、近代インド・アーリア語系と若干のペルシア系語彙であることが確認できた。

研究成果の概要（英文）：This study aims to make clear to what extent Hindi and Urdu languages differ from each other from the etymological viewpoint of the vocabulary the respective languages contain.
The output of the investigation shows that both languages do not differ in most frequent words, consisting of modern Indo-Aryan words and a few Persian loanwords, occupying more than 80 percent of total running words, contrary to anticipation.

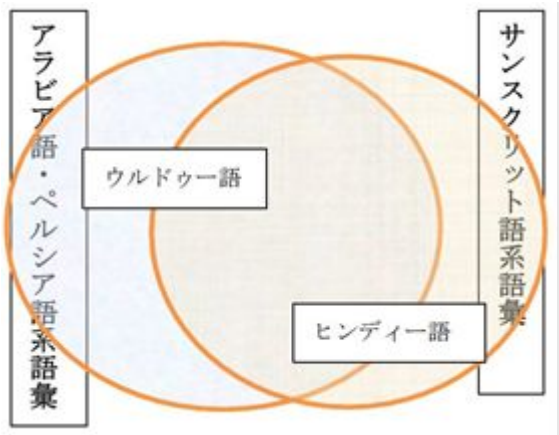
研究分野：言語学

キーワード：ヒンディー語　ウルドゥー語　語彙

１．研究開始当初の背景

(1)今日まで約 200 年間にわたるヒンディー語とウルドゥー語の関係について、コミュナリズムを背景にした政治的な言説は別にして、南アジアや欧米の言語学者の間では「別個の言語」から、「方言の差」、「文体の違い」あるいは「同じ言語の変異形」までさまざまな説が提出されていた。しかしどの説も共通しているのは、両言語の形態論・統語論における本質的な違いの主張ではなく、主に語彙の特徴、具体的には特定の語源に帰属する語彙の含有傾向に根拠を置いていることである。

(2)従来、ヒンディー語ではサンスクリット語系からの語彙が、ウルドゥー語ではペルシア語・アラビア語系の語彙が多いという漠然とした事実認識があった。しかし問題はそれほど単純ではなく、ヒンディー語の基本語彙にもペルシア語・アラビア語からの借用語がかなりあり、ウルドゥー語にも音韻変化を受けたサンスクリット語語源の語彙がある。これまで語彙特徴の科学的な抽出やそれらの含有率を客観的に計測した先行研究はなく、また前述した 200 年間における両言語のダイナミックな変容を追跡した先行研究もなかった。



２．研究の目的

(1)本研究の目的は、ヒンディー語とウルドゥー語の言語テキストに含まれる語彙属性の定量分析をとおして、両言語の特性を客観的に測定しその差異の内容を明らかにすることである。

(2)言語の根底では同じでありながら特に使用語彙の傾向をめぐってこれまで約 200 年間続いてきた両言語の異同に関する論争に客観的な測定結果をもとに終止符を打つと同時に、時代ごとに変容した可能性のある差異や傾向の実態をも明らかにすることを目指す。

(3)我が国において東京外国語大学と大阪外国語大学（現大阪大学外国語学部）ではヒンディー語とウルドゥー語は別個の言語として教授されてきた歴史がある。このため、教員や研究者は南アジア研究にとって両言語の習得・理解が必要不可欠であるという認識は共有していても、両言語の客観的な差異に基づいた効率的なカリキュラム、教材、辞書の作成が十分ではなかった。そこで本研究では、ヒンディー語・ウルドゥー語の使用語彙を中心に、「何がどれくらい同じで、何がどれくらい違うのか」ということを客観的に計測し、これからの両言語の教育と研究に役立たせることをも目的とした。

３．研究の方法

本研究の計画・方法の柱は、(1)語彙属性の設定と機械辞書作成、(2)時代別のヒンディー語・ウルドゥー語散文テキストの電子化、(3)語彙属性の定量分析である。

(1)ヒンディー語・ウルドゥー語の定量分析の対象である語彙属性に関し語源情報（供給源言語、音韻変化の程度など）をもとにきめ細かく設定(約 30 種類)し、最終的に約 2 万 5,000 語の機械辞書（見出し語形、語彙属性、品詞、活用変化情報、語義など）を作成する。見出し語形は両言語の相互参照の便を考慮しヒンディー語（デーヴァナーガリー文字）とウルドゥー語（アラビア文字）を併記する。

(2)設定した時代ごと（19、20, 21 世紀）に選定された両言語の散文テキスト（合計約 300 万語）を電子化し蓄積する。

(3)電子化された散文テキストに対し、機械辞書を利用した自動タグ付け(tagging)の出力結果をもとに、以下の要領で、語彙属性の定量分析を実施する。

①機械辞書の活用変化情報を使いテキストの構文解析を実行し、見出し語形とそれに後続する活用変化にタグをつける。

②タグの付けられた語形から、延べ語(tokens)、異なり語形(types)、見出し語形に直した語形(lemmas)とそれぞれの頻度を求める。

③見出し語語形に直した語形(lemmas)に対し機械辞書に記述してある語源情報を自動参照させ語源情報のタグ付けを行う。これにより、散文テキストに含まれる各語彙の語源情報（供給源言語、音韻変化の程度など）の客観的な数値化と、各テキストごとの語彙属性の分布の特徴を把握することができる。

４．研究成果

研究実施の中で成果として得られた主な知見は以下の通り。

(1)研究対象として電子化された 19 世紀から 21 世紀にいたる期間におけるヒンディー語・ウルドゥー語の各散文テキストは、それぞれの量的なばらつきを補正しても、語彙の多様性をあらわす指標といわれる TTR (Type-Token Ratio) の値にはかなりの幅があった。

(2)TTR の値とは別に、各テキストにおける出現頻度上位 80%を占める語彙（type レベル、lemma レベルそれぞれ）を抽出し、それらの語彙の語源別含有比率を計測した。語彙の多

様性には関係なく、上位 80%を占める語彙の顔ぶれはほとんどが一定であった。
(3)上位 80%の語彙に関する語彙属性は、時代別、言語別、テキスト別の TTR に無関係に、安定した内容であることが確認できた。
(4)上位 80%の語彙の語彙属性は、ほとんどが、近代インド語派に分類されるもので、サンスクリット語やペルシア語からの借用は非常に少なかった。特にサンスクリット語からの借用語は、この出現頻度の範囲において、ヒンディー語においてもほとんどゼロであった。
(5)使用頻度上位 80%を除いた、下位 20%を構成する語彙の内容が、いわゆるヒンディー語・ウルドゥーの差異を特徴づけるものであることを確認した。つまりこれらの語彙は、出現頻度は非常に少ないが個別数は極端に多く、語彙属性は借用が多い。
(6)研究の結果、ヒンディー語とウルドゥー語の両言語においての語彙の語源的片寄りについては、事前の想定よりも差がなかったということがわかった。特に出現頻度上位 80%までの範囲では、両言語とも圧倒的に近代インド語派が多数を占め、若干の古くから借用されているペルシア語系語彙が上位に来ることが確認できた。
(7)新たに得られた知見として、合成語の語彙属性の設定に関わる問題がある。ヒンディー語・ウルドゥー語の主に問題となる借用の供給源言語はサンスクリット語とペルシア語（アラビア語含む）である。これらの中に外形からは典型的な借用語彙とみなされるものが、実は南アジアにおいてそれぞれ特定の時期に造語されたいわばネオ・サンスクリット語やネオ・ペルシア語とでも言うべき語彙がかなり含まれていることがわかった。また使用頻度では、これらの語彙は比較的高い数値で分布していることもわかった。本研究では、これらの合成語については分かる範囲で語彙属性の独自のカテゴリーを立てて対応した。ただし、これらの語彙の扱いについては、造語の時期を含めてさらに検討する必要がある。
(8)本研究遂行の副産物として、ヒンディー語・ウルドゥー語の機械辞書、散文テキストの電子化、開発した解析用のプログラムなどがある。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計７件）
① 萬宮　健策、ウルドゥー語の所有・存在表現：接尾辞 wala を用いた表現が表すもの、東京外国語大学語学研究所論集第 18 号、査読有、2013、121-139

② 萬宮　健策、ウルドゥー語、ことばの読み方事典、査読無、丸善出版、2014、326-331

③ 町田　和彦、ヒンディー語、ことばの読み方事典、丸善出版、査読無、2014、258-263

④ 町田　和彦、インド系文字、ことばの読み方事典、丸善出版、査読無、2014、406-409

⑤ 町田　和彦、文字からことばへ－17 世紀のフィールド言語調査を読み解く－、人文学のフィールドサイエンス、査読無、2014、1－20

⑥ 町田　和彦、文字からことばへ、人はみなフィールドワーカーである、査読有、2014、134-149

⑦ 萬宮　健策、ウルドゥー語における他動性、東京外国語大学語学研究所論集第 19 号、査読有、2014、265-275

〔学会発表〕（計３件）
① 町田　和彦、萩田　博、萬宮　健策、ヒンディー語とウルドゥー語における語形成の問題について、日本南アジア学会第 26 回全国大会、広島大学、2013 年 10 月 15 日

② 萩田　博、町田　和彦、萬宮　健策、ウルドゥー語散文の歴史的展開についての一考察、日本南アジア学会第 27 回全国大会、大東文化大学、2014 年９月 28 日

③ 萬宮　健策、パキスタンの大学における英語、ウルドゥー語教育の現状、「アジア諸語を主たる対象にした言語教育法と通言語的学習達成度評価法の総合的研究」第９回研究会、東京外国語大学、2014 年６月６日

〔図書〕（計２件）
① 萬宮　健策、言語問題とアイデンティティ－シンディー語の事例から－、現代インド５　周縁からの声、査読有、2015、277－296

② 萩田　博、石田　英明、マイノリティ文学からの発信、現代インド５　周縁からの声、査読有、2015、251－271

〔産業財産権〕
○出願状況（計０件）

名称：
発明者：

権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕
ホームページ等
http://www.aa.tufs.ac.jp/~kmach/hirdu/html/MorphologicalAnalysis.htm

６．研究組織
(1)研究代表者
町田　和彦（MACHIDA, Kazuhiko）
東京外国語大学アジア・アフリカ言語文化研究所・教授
研究者番号：７０１３４７４９

(2)研究分担者
三上　喜貴（MIKAMI, Yoshiki）
長岡技術科学大学工学系研究科・教授
研究者番号：７０２９３２６４

萩田　博（HAGITA, Hiroshi）
東京外国語大学大学院総合国際学研究院・准教授
研究者番号：８０１４３６１８

萬宮　健策（MAMIYA, Kensaku）
東京外国語大学大学院総合国際学研究院・准教授
研究者番号：００４０３２０４

(3)連携研究者
（　　）

研究者番号：