様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ５ 月 １２ 日現在

機関番号：１４４０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ００５４２

研究課題名（和文）ツリーバンクを利用したヒンディー語と日本語のとりたて詞の機能の対照研究

研究課題名（英文）A Contrastive Study of Functions of Hindi and Japanese Toritateshi (particles) Using a Treebank

研究代表者

西岡　美樹（Nishioka, Miki）

大阪大学・大学院人文学研究科（外国学専攻、日本学専攻）・准教授

研究者番号：３０４５２４７８

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：本研究では、既存のウェブコーパスCOSHのデータの一部に、IIIT Hyderabadが開発したHDTB（Hindi Dependency Treebank）から作成したモデルを利用して依存構造解析を行い、UDツリーバンク（Universal Dependency Treebank）を構築した。この統語情報が付与されたデータを統語構造ベースで検索できるツールも併せて開発し、統語レベルでの量的研究を行える環境を整えた。これらを利用し、ヒンディー語の『とりたて詞（小詞）』の代表であるto、hii、bhiiと、それに相当する日本語とりたて詞「は」「も」「こそ」等の意味機能の対照分析を行った。

研究成果の学術的意義や社会的意義
本研究で開発したヒンディー語のUD（Universal Dependency）ツリーバンクと専用の検索ツールにより、当該言語の非母語話者である研究者でも統語レベルでの量的研究が行えるようになった。言語研究面では、とりたて詞を含む統語構造（名詞句、動詞句等）について、研究代表者の言語経験をもとに統語モデルを作成し、このツリーバンクを利用して短時間でとりたて詞を含むこれらの統語モデルの例の検索が行えた。これにより量的研究による客観性を担保しつつ、ヒンディー語のとりたて詞の意味機能について、日本語のそれらと対照させながら分析を行えた。この知見は言語学のみならず、高度な外国語教育にも活かせる。

研究成果の概要（英文）：In this research, we constructed a UD treebank (Universal Dependency Treebank) by performing dependency structure analysis using a model created from the HDTB (Hindi Dependency Treebank) that was developed by IIIT Hyderabad on some data from the existing web corpus COSH. We also developed a syntactic structure-based search tool for the data with syntactic information to provide an environment for quantitative research at the syntactic level.
Using the treebank and its tools, we analyzed the semantic functions of 'to', 'hii', and bhii', representative Hindi "toritate" particles ("nipaat" in Hindi grammar), by comparing them with Japanese 'ha', 'mo', 'koso', etc. and by discussing them with research collaborators who are native Hindi speakers and have learned Japanese.

研究分野：言語学

キーワード：ツリーバンク　ヒンディー語　日本語　とりたて詞　機能　対照研究　コーパス

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

（1）日本語には、「は」「も」「こそ」「しか」「さえ」等、かつては副助詞、係助詞等と呼ばれたとりたて詞ばれる助詞群がある。インドの主要公用語であるヒンディー語にもそれに相当する機能を持つ小詞（ヒンディー語伝統文法では'avyay'または'nipaat'と呼ばれるが、ここでは便宜上「とりたて詞」と呼ぶ）がある。代表的なものは to、bhii、hii の 3 つである。これらについては、例えば古賀勝郎・高橋明編（2005）『ヒンディー語＝日本語辞典』（大修館書店）では、to に 2 つ品詞があり、1 つが接続詞の「と」「たら」、もう 1 つが副助詞の（1）「は」「なら」等の強調的にとりたてた表現をしたり対比や対照的な提示、限定的な意味を付加する、(2)「せめて」「ぐらい」(程度を示す)、のように説明されている。ヒンディー語のとりたて詞は、語、句、節、文に対して付加され、何らかの「強調」の意を加える機能を持つが、これは、語や句、一部の節や文に接続した場合の具体的な訳語及び品詞を掲載しているに過ぎず、このように日本語に引き寄せた意味を記述するのは「読む・聞く」の受動的な言語能力には有効だが、「話す・書く」の能動的な言語運用能力を解明することはできない。特に、ヒンディー語のとりたて詞は、名詞句内あるいは動詞句内にも挿入され、さらに文までとりたてる。これが複文をつなぐ働きをしているように見えるため、上掲辞書のように「接続詞」として品詞分類されることになる。ヒンディー語では、Koul（2009）がこれまでの伝統文法家の記述と違い、統語パターンに基づき網羅的に例を挙げているが、英語の対訳列挙に終始しており、非母語話者（特に非英語母語話者）向けの意味機能を解明する体系的な分析や説明はなされていなかった。

（2）これまで、非母語話者の研究者は対象言語の母語話者的直観が欠如しているため、個人の内省のみではその研究分野に踏み込みにくかった。というのは、非母語話者の研究者個人で実施可能なインフォーマントとのインタビュー調査では、人数が限られるため量的証拠及び収集した例文の客観性が担保しにくいからである。日本のみならず、19 世紀から 20 世紀に音韻、形態、統語の面でのヒンディー語を含む南アジア諸語の記述に大きく貢献してきた欧州や米国でも、これらの研究分野において意味や機能面からの体系的な記述はほぼ手つかずできた。

２．研究の目的

本研究では、ヒンディー語のウェブコーパス COSH（Corpus of Spoken Hindi）に統語情報の注釈を付与したツリーバンクを構築し、専用の検索ツールを開発することで、名詞句や動詞句等の統語レベルでの検索を可能にし、短時間でとりたて詞を含む句の検索を行い、量的研究により客観性を担保しながら、ヒンディー語のとりたて詞の機能について、日本語のそれと対照させながら解明する。具体的には、主語、目的語、後置詞句（副詞句）を成す名詞句や、主に述語を形成する動詞句の中に割り込むものに対し、日本語のとりたて詞をどれだけ対訳として使用できるか、できない場合は、どのような方策でその強調を表すことになるのかを、フォーカス解釈、否定のスコープ等を手掛かりに、対照的視点から両言語のとりたて詞の機能を分析し、双方の非母語話者の言語教育にも有用な意味記述を行う。

３．研究の方法

まず、概ね日本語のとりたて詞と平行して逐語訳できる語（名詞や形容詞）に付加するとりたて詞を整理し、次に、Koul（2009）を参考に、既存の COSH を利用し、予想される統語パターンを名詞句（NP）と動詞句（VP）別にモデル化する。特に日本語のとりたて詞とヒンディー語のそれが平行させにくくなる①NP（修飾語句＋被修飾語句）と②VP（二つ以上の動詞連続）にとりたて詞が割り込むものに重点を置く。①は日本語の「どんな時でも」が斜格名詞句の kisii bhii samay［any + particle + time］になる例が典型だが、とりたて詞 bhii が句末に置かれる例もあるため、このような場合の意味解釈の違いを解明する。②では様々な時制、相等が絡むが、中でも使用頻度が高い現在形／完了形／進行形と、語彙的アスペクトを表す補助動詞（V2：jaanaa「行く」、denaa「与える」、lenaa「取る」等）を伴ったいわゆる複合動詞の動詞句に焦点を当てる。Koul の指摘も参考にし、VP では主動詞（V1）がとりたてられやすいが、未完了・完了のような相や語彙的アスペクトを含むとりたてがどれだけ可能か（例えば「作ってはある」に相当するもの）も調査・分析する。さらに、動詞句が否定語を伴う場合、それが句外のとりたて詞と呼応するため、否定のスコープも視野に入れて分析を行う。なおツリーバンクの検索ツールで NP と VP をチャンク構造ごと検索できるようになった段階で、それを利用して研究を進める。ツリーバンクの開発は最初の 2 年ほど行う。COSH のデータからヒンディー語の文を、自動翻訳のヒンディー語文を極力排除しながら選定する。次にこのコーパスの一部（開発予算とサーバー運営維持費に応じて規模を調整）に Hindi Dependency Treebank（HDTB）を用いて、①構文解析パーサ MaltParser の feature model を作成、②のデータの一部に統語情報を付与、③統語情報付きコーパス用のインターフェース（現在の Corpus Query Language（CQL）でなく、XPath 言語によるもの）を、専門業者に委託して開発する。

４．研究成果

（1）初年度、まず開発担当の研究協力者とツリーバンクのアノテーションに使用するツールの

選定と現在稼働している Web コーパス COSH（Corpus of Spoken Hindi）のデータの選別とサーバーの管理体制について打ち合わせをし、次に、コーパスを使い南アジア諸語の言語研究を進めている研究協力者らとも打ち合わせをした。その際、当初ツリーバンクのアノテーションに使用する予定だった MaltParser 以外にも SpaCy や Stanza 等があることが判明したため、研究協力者と検討し、MaltParser 版と Stanza 版の 2 種類のツリーバンクを作ることにした。データについては、研究代表者が中心となり既存の COSH から、ヒンディー語のとりたて詞に相当する小詞（ヒンディー語文法の nipaat、主に hii、bhii、to の 3 つ）を含むデータを吟味、選別し、業務に委託してツリーバンクのアノテーションを行った。

（2）言語研究面では、Koul（2009）に挙げられているとりたて詞の例文を参照しながら、これらとりたて詞が文のどの要素をとりたてているかに着目し、統語パターンをモデル化した。そのモデルに沿ってヒンディー語と日本語の逐語訳を作成した。主に扱ったとりたて詞は、日本語の「は」（対比の to に相当）、「も」（bhii に相当）、「こそ」、「だけ」、「しか」（おおよそ hii に相当）である。これらの例文について、ヒンディー語のそれが文法的で意味を成しているか、日本語のとりたて詞を伴った逐語訳が同じ意味を成しているか等、研究協力者とオンラインで議論した。

（3）次年度は、初年度に研究協力者の協力で用意した例文に、さらに COSH と既存のインターフェース COSH Conc で検索可能な範囲で例文を収集しながら、ヒンディー語母語話者の研究協力者たちとの自然会話やオンライン動画等からも例文を収集し、それらに対し日本語の「は」、「も」、「こそ／だけ／しか」のようなのとりたて詞が逐語訳的にいかに使用できるかを、同研究協力者らと議論した。その成果をヒンディー語の強調表現の一部として試験的に論文にまとめ、国際会議で口頭発表（オンライン）をした。その際、対照となる日本語のとりたて詞の意味分類の指標として、Noda（2017）の restriction vs. anti-restriction、extremes vs. anti-extremes、similarity vs. anti-similarity を参照した。

（4）2 年目はツリーバンク用の検索インターフェースを年度中ごろから開発する予定だったが、開発業者の都合（CentOS 8 のサポート期限の前倒しに伴うサーバの OS 移行作業による）で、検索用のインターフェース開発に遅延が生じた。しかし、同年 12 月に開発時期と実装内容の変更についてオンライン打ち合わせを行い、インターフェースの開発を最終年度の前半までに終えることにした。

（5）最終年度は、ツリーバンクの MaltParser 版と Stanza 版のアノテーションの精査を研究協力者らと行いつつ、前年度からずれ込んでいた検索インターフェースの開発を行った。これは 6 月下旬には完成したため、研究協力者らにもモニター実験の協力を仰ぎ、速やかに試験運転を開始した。なお、MaltParser 版には以下の①を利用して feature model を作成し、Stanza 版は②で公開されている feature models を使用している。試験運用の中、MaltParser 版と Stanza 版を比較すると Stanza 版の方が有用であることが明らかになった。また、計画段階で予定していた XPath は、検索式が非常に難解であるため、検索式の作成が比較的わかりやすい dep_tregex という検索ライブラリを一部修正して利用したが、それでもなおこれを効果的に使用するには検索式に慣れる必要があることが判明した。

（6）本課題の成果の一つ COSH UD Treebank は、品詞、形態素、構文依存性を通言語的にアノテーションするためのフレームワーク Universal Dependencies（UD）が利用されており、今後他のインド諸語への適用拡大とともに比較研究の促進が期待される。しかし、COSH での品詞、形態素解析レベルの検索と違い、同時に文法関係等が指定できるため検索式が複雑になる。そこで開発班の研究協力者と再度打ち合わせをし、とりたて詞を含む構造を抽出するのに適した検索式モデルを考案した。試行錯誤の結果、主語、目的語、斜格名詞句（副詞句）に付加または挿入されるとりたて詞を丸ごと抽出できる検索式を編み出し、試験運転中の UD ツリーバンクで実際に例文を抽出することができた。ヒンディー語のとりたて詞を含むパターン抽出を可能にするこの検索式については、関連の国際学会に投稿を予定している。

（7）言語研究については、現在論文を投稿中だが、引き続き研究協力者 Dr. Narsimhan（University of Delhi）や他のヒンディー語母語話者らとの自然会話から to、hii、bhii を含む例文をさらに収集している。特に、動詞句に付くヒンディー語のとりたて詞の例に重点を置き、日本語のとりたて詞や、副詞を使った対訳との対応関係について議論を重ねている。

＜参照サイト・文献＞

① GitHub - UniversalDependencies/UD_Hindi-HDTB
URL: https://github.com/UniversalDependencies/UD_Hindi-HDTB

② Stanza, Available Models & Languages - Stanza (stanfordnlp.github.io):
URL: https://stanfordnlp.github.io/stanza/available_models.html

③ UD Hindi HDTB
URL: https://universaldependencies.org/treebanks/hi_hdtb/index.html

④ Koul, Omkar N.（2009）*Modern Hindi Grammar*. Delhi: Indian Institute of Language Studies (Indian reprint).

⑤ Noda, Hisashi. (2017) Toritate: Focusing and defocusing of words, phrases, and clauses. In Masayoshi Shibatani, Shigeru Miyagawa& Hisashi Noda (eds.), *Handbook of Japanese syntax*, pp.123-156. Berlin/Boston: De Gruyter Mouton.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　1件／うち国際学会　1件）

１．発表者名 Miki Nishioka, Ranjana Narsimhan
２．発表標題 Emphatic Particles and Emphatic Expressions: A Comparative Study of Hindi and Japanese
３．学会等名 International Conference on Hindi Grammar and Lexicon（招待講演）（国際学会）
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

COSH UD Treebank https://treebank.cosh.site/ Corpus Of Spoken Hindi (COSH) and COSH Conc http://www.cosh.site/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	ナラシマン　ランジャナ (Narsimhan Ranjana)		University of Delhi
研究協力者	赤瀬川　史朗 (Akasegawa Shiro)		Lago NLP

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	バット　ラジェーシュ (Bhatt Rajesh)		University of Massachusetts at Amherst
研究協力者	カスカート　チャンドラ (Cathcart Chundra)		University of Zurich

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関