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　 ６ 月 ２３ 日現在

機関番号：１４３０１
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２０６９７

研究課題名（和文）古代インド文献成立過程解明に向けた文体計量分析のためのデータベース構築

研究課題名（英文）Construction of Database for Quantitative Analysis of Language with a View to Clarify the Process of Composition of the Ancient Indian Literature

研究代表者
天野　恭子（Amano, Kyoko）

京都大学・文学研究科・人文学連携研究者

研究者番号：８０３４３２５０

交付決定額（研究期間全体）：（直接経費）　　4,900,000円

研究成果の概要（和文）：本研究は、古代インドの祭式文献ヴェーダの計量分析を可能にする、文法解析付きデータ作成を目的とする。デュッセルドルフ大学のO. Hellwigが2009年に発表したサンスクリット語の文法解析プログラムは、自動で文法解析データを作成する可能性を拓いたが、解析結果のチェックと訂正が必須である。そこでHellwigと共同で、web上でinteractiveに作業ができる、自動解析結果のチェック・訂正のためのシステムを構築した。これを用いヴェーダ文献の文法解析付きデータを作成し、Digital Corpus of Sanskritに蓄積している。作成したデータをもとに文献の言語的特徴の分析を行っている。

研究成果の学術的意義や社会的意義
本研究は、成立に謎が多い古代インドのヴェーダ文献の言語について、詳しく分析する可能性を拓く、文法解析付きデータベースの作成を目指したものである。コンピュータによる言語の分析は、多くの資料を扱うことができ、細かい言語的特徴を捉えることが可能であることから、ヴェーダ文献の成立とその背景となる社会の解明に繋がると考えられる。本研究は、ヴェーダ文献の文法解析付きデータを正確かつ効率的に行うシステムの構築を行い、いくつかの重要文献のデータ化を達成した。このデータを用い、文献の言語的特徴から文献成立過程を解き明かす分析の試行を行い、今後の研究発展の可能性を示した。

研究成果の概要（英文）：The purpose of this research is to create data with grammatical analysis that will enable quantitative analysis of the Vedas, the ritual literature of ancient India. A program for grammatical analysis of Sanskrit that was published by O. Hellwig of the Dusseldorf University in 2009 opened up the possibility of automatically generating grammatical analysis data, but checking and correcting the analysis results are essential. Therefore, in collaboration with Hellwig, we constructed a system for checking and correcting the results of automatic analysis that can work interactively on the web. Using this system, we have created data with grammatical analysis of Vedic literature and stored it in the website "Digital Corpus of Sanskrit". Based on the created data, we are analyzing the linguistic features of the literature, to clarify the internal construction (historic layers) in a text and also the relationships among several texts.

研究分野：インド哲学

キーワード：古代インド　ヴェーダ　サンスクリット　文法解析　TEI　タグ付けプログラム　データベース

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ヴェーダ文献は、古代インドの言語、宗教、歴史の研究に豊富な資料を提供するものであるが、約三千年から二千年前に口承のみで成立したものであり、成立の時期や地域、特定の作者を指示する直接の証拠は一切存在しない。近年の計量言語学の発展により、計量分析がヴェーダ文献の成立過程の解明に役立つと考えたが、計量分析には、テキストを単語ごとに分離し、原形と文法形を記したデータ、いわゆる文法解析付きデータ（XML 等の形式）が必要となる。しかし古代インド語（サンスクリット語）は、単語同士の接触部分が融合が音変化を起こし、その際、解釈に複数の可能性を生じることから、自動での解析は極めて困難であると考えられており、すでに文法解析された資料が存在するごく一部の文献だけがデータ化され、他の膨大な文献は手つかずで残されていた。デュッセルドルフ大学の Oliver Hellwig は 2009 年に、サンスクリット語の自動解析プログラムを公開し、その後データ化されたテキストを蓄積するための Digital Corpus of Sanskrit を運営していた。そのため、研究代表者天野は Hellwig と、ヴェーダ文献、特に本研究がターゲットとするマイトラーヤニー・サンヒターのデータ化について協議を行った。その協議では、完全な自動化は依然として困難で、自動解析の結果を専門研究者がチェックし、正しい文法解析を行う手順が必須であること、またそのような作業のためにまず、Hellwig のシステムを web 上で利用できるようシステムを構築することが必要であることを確認し、本研究を計画することとなった。

２．研究の目的

研究の目的は、古代インド文献の成立過程の解明のために、これまでサンスクリット文献学では用いられていない、文体計量分析の手法を用いるための文法解析付きデータを作成することにある。そのためには、自動解析プログラムの文法解析、その結果のチェック・訂正、それをフィードバックしてデータを完成しそのデータを蓄積する、という作業フローを可能にするシステムを web 上に構築することが最初の課題となる。システムが構築できればデータを順次作成し、文献の言語の特徴をあぶり出す計量分析を試行し、より目的に叶ったデータのあり方、分析の方法を探る。
文献の言語の分析により、文献の言語の特徴や、文献同士の類似度、関係性を考察できる。このことは、十分に解明されていない古代インドヴェーダ期の、宗教や社会の変遷を解き明かすことに繋がり、古代インドの歴史、思想史の研究に大きな影響を及ぼす。本研究で対象とされるマイトラーヤニー・サンヒターは、BC900 年前後の成立と推定されているが、その成立年代や作者を示す直接の証拠はない。同文献は一度に作られたのではなく、新しく作成した部分が次々と加わって構成されていったと考えられ、作成された時代によって、同時代の他の文献から様々な影響を受けたと考えられる。時代を縦軸とするならば、外からの影響という横軸があり、その両方の軸の中で文献の成立を考えねばならない。この、縦軸（時代）および横軸（影響関係、地理的要因）に位置付けて文献の変遷を見ることが、ヴェーダ期の社会や思想の変遷を解き明かすことに繋がるのである。
そして、このように複雑な文献の成立過程の解明に、文体の分析が役立つと考えている。そのためにまず、マイトラーヤニー・サンヒター及び、同時代の成立と考えられるカータカ・サンヒター、タイッティリーヤ・サンヒターの文法解析付きデータが必要であり、そのデータを作成する。

３．研究の方法

１）上で述べた、**サンスクリット語自動解析プログラムによる文法解析、その結果のチェック・訂正、それをフィードバックしてデータを完成しそのデータを蓄積する、という作業フローを可能にするシステムを web 上に構築する**ことが最初の課題となる。このシステム構築は、Hellwig が担当する。
２）システム構築後は、上の手順で**データを作成**する。まず、プログラムによる文法解析、その結果をシステムにアプロードするまでを Hellwig が行う。その次の段階では、伏見誠、天野恭子が、文法情報のチェックをシステム上で行う。システム上には、単語ごとの文法情報が表示されるが、複数の可能性を示す場合も少なくなく、その場合は最も適切と考えられるものを選んで、単語ごとに確定させてゆく。**確定した情報を Hellwig がファイナライズし、TEI 形式に準じた conllu データにしてデータベースに加えてゆく**。なお、解析データが増えるごとにプログラムの学習が進み、解析の精度が上がるようになる。本研究の主たる部分は、このデータ作成である。
３）**作成したデータを用いて分析**を行うことを、今後の展望としている。マイトラーヤニー・サンヒターの**章の間の類似度**を分析することにより、マイトラーヤニー・サンヒターの内部構造を

考察すること、マイトラーヤニー・サンヒターと同時代のカータカ・サンヒターとタイッティリーヤ・サンヒターとを比較することにより、**文献間の影響関係**を考察することが目的である。データ作成には時間がかかるため、文献全体の完成データを用いて分析をすることは先のこととなるが、まずは作成できたデータを用いて、分析を試行する。研究開始当時は、N-gram 分析を考えていたが、その後、Doc2Vec，Word2Vec といった、語彙の頻度によってトピックを見つけ、類似度を数値化するツール、そこに重要語彙を析出して類似度の判定に加味する「重みづけ」の手法を、ヴェーダ文献の分析に利用すべく検討した。

４．研究成果

１）サンスクリット自動文法解析プログラムによる解析、チェックと訂正、その結果をデータに反映させてデータベースに組み入れる、という一連の作業を可能にする web システムは、研究期間一年目に Hellwig により完成された。プログラムによる解析結果は、Digital Corpus of Sanskrit の作業用ページにアプロードされ、伏見誠と天野恭子がチェックを行った。web ページ上では、各単語ごとに文法情報を表示させることができる。複数の可能性のある場合は、すべての可能性が表示され、その中から選択できるようになっている。単語の意味も表示され、同じ単語で複数の意味がある場合は、その箇所における適切な意味を選択できるようになっている。単語ごとの文法情報のチェックという非常に手間と時間のかかる作業を、効率よく行うことを可能にする非常に優れたシステムである。このシステムの完成により、２）のデータ作成と、３）の分析が可能となった。サンスクリット語のような、音の融合や変化が多く起こり、随所で複数の解釈の可能性が生じる言語は、プログラムによる文法タグ付けを完全に自動化することは難しい。従って、専門研究者によるチェックと訂正の作業の重要性が非常に高くなるが、その作業の効率を飛躍的に高める Hellwig のシステムは、サンスクリット文献のデータ化を考えている研究者や新しいプロジェクトから興味を持たれており、今後このシステムが広く使われるようになれば、非常に大きな波及効果があると考えられる。

２）データを実際に作成する作業においては、作業開始時にはいくつかの問題が発生した。例えば、一文字で一つの単語となる動詞前綴りの ā が、一つの単語として認識されなかったり、動詞前綴りと動詞本体が、繋がっているか離れているかで別の単語として認識されてしまうなど、サンスクリット語独特の問題があったが、その都度解決策を相談し、より正しく効率的なデータ作りの方法を発展させた。本研究で対象とするマイトラーヤニー・サンヒターは未解読の文献であり、既存の辞書に載っていない語、マイトラーヤニー・サンヒターだけに出てくる語、それも一か所にしか出てこない語がある。そのような語をプログラムが自動解析することはできないが、伏見・天野が正しい文法情報を別に Hellwig にフィードバックし、Hellwig がその情報をデータに反映させることで、正しいデータを作成することに注力した。マイトラーヤニー・サンヒターの大部分、およびカータカ・サンヒターは、未解読の文献であり、タイッティリーヤ・サンヒターは英訳が存在するものの、どれも膨大な文献であり、データ化の完成は簡単ではない。マイトラーヤニー・サンヒターの１巻から順にデータ化すべく作業を開始したが、時間がかかり、目的である３つの文献の文献間関係の考察になかなか至らないため、まずはそれぞれの文献の儀礼解釈部分についてデータ化することにした。これらの文献は、儀礼解釈部分と、マントラ（祝詞）部分に分かれる。それぞれ成立事情が異なるので、別々に考察することが望まれる。儀礼解釈部分は散文で書かれ、文体の分析も比較的やりやすいと考えられる。３年間の研究期間で、英訳のあるタイッティリーヤ・サンヒターの儀礼解釈部分のデータ化が完成し、未解読のマイトラーヤニー・サンヒターとカータカ・サンヒターの儀礼解釈部分についても、およそ半分のデータ化が完成した。これらのデータによって３）の分析が可能になる。このデータは、Digital Corpus of Sanskrit を通じて公開されているため、天野と Hellwig のチーム以外の研究者もデータを利用している。2022-2023 年に印欧語比較言語学の研究者が、統語論研究にマイトラーヤニー・サンヒターのデータを用いた研究を発表した例があり、今後さらに多くの研究者に利用されると考えられる。

３）データを用いた計量分析の目的は、マイトラーヤニー・サンヒターおよび同時代のカータカ・サンヒターとタイッティリーヤ・サンヒターについて、その内部構造と、文献間の関係を知ることである。一つの文献の中の章に着目し、それぞれに固有の特徴を浮き彫りにし、章同士の関係を明らかにすることで内部構造を探る。その内部構造を踏まえた上で、文献同士の影響関係を探る。これまで、これらの文献の比較は、あくまでも文献全体同士を単純に比較するものであった。本研究の狙いは、一つの文献を全体として扱うのではなく、内部に層があることを浮き彫りにし、文献間の比較も、全体としての比較ではなく、ある文献の層と、他の文献の層、という細かな単位で比較していくことである。これは、文献の成立過程についての新しい考え方に根差している。本研究および、本研究から発展した他のプロジェクト（後述）から、この狙いを発信してきたが、これはヴェーダ文献成立の歴史、文献成立を取り巻く社会の歴史を考察する上での、新しい視点を提言するものであったと考えている。
分析ツールを用いた、語彙・文体分析は、研究協力者である京極祐希（ライプツィヒ大学）が行

った。2020 年度に、Doc2Vec, Word2Vec といった、語彙の頻度によってトピックを見つけ、類似度を数値化するツールを用いた章間類似度の分析、2021 年には、重要語彙を析出して類似度の判定に加味する「重みづけ」の手法の検討を行った。いくつかのやり方を比較検討し、最もヴェーダ文献の内容の特徴を捉えているものはどの手法かという検討を続けている。これまでのところ元のデータが文献の一部分であるため、目的としている考察にはたどり着いていないが、これまでの分析で、ある程度ヴェーダ文献の内容の特徴を捉えられることがわかり、今後の発展に期待できる。

本研究によって作成したデータをもとに、本研究で設定した課題にさらに取り組むべく、次の２つのプロジェクトが発足したことは、本研究の大きな成果である。
１つは、京都大学研究支援 SPIRITS（「知の越境」融合チーム研究プログラム）「データ駆動型科学が解き明かす古代インド文献の時空間的特徴」 2020-2021（https://ancientindia-datascience.jp/）である。本研究では、前述のマイトラーヤニー・サンヒターを含む多くのヴェーダ文献について、その地理的・時代的特徴を浮き彫りにし、成立・発展過程を明らかにするために、データサイエンスの手法をどのように利用するかを、広く深く論じるためのプロジェクトである。メンバーは、代表者天野の他、データ可視化の専門研究者である夏川浩明（京都大学学術情報メディアセンター（当時)、現在は大阪成蹊大学)、Oliver Hellwig、京極祐希である。本プロジェクトにおいて、Hellwig と京極が、作成したデータを用いた分析を発展させ、成果を発表した。この成果発表の場となった国際ワークショップは、SPIRITS プロジェクト主催、本研究課題（挑戦的研究）を共催として開催したものである：

2021 年 2 月 12 日：Dynamism of Social Context Deciphered by a Linguistic Analysis of Ancient Literature「古代文献の言語分析から読み解く社会背景のダイナミズム」（オンライン）

天野恭子：Problems in the Formation of the Vedas, Ancient Indian Religious Texts.「古代インド宗教文献ヴェーダの成立を巡る諸問題」

Relationship among Vedic Schools Deciphered by the Visualization of Mantra.「マントラ共起関係の可視化から読み解くヴェーダ学派間の関係性」

京極祐希：Collocation Measuring the Semantic Similarity between the Chapters of Taittirīya Samhita Using a Vector Space Model.「ベクトル空間モデルによる『タイッティリーヤ・サンヒター』の章間類似度比較」

Oliver Hellwig:Dating Vedic Texts with Computational Models: Algorithmic Considerations and Data Selection.

夏川浩明：The Possibility of Information Visualization and Data Analysis for Ancient Indian Literature.「古代インド文献を対象とした情報可視化やデータ分析の可能性」

濵地瞬（京都大学）：Citation Prediction Using Academic Paper Data and Application for Surveys.「学術論文データを用いた引用数予測とサーベイへの活用」

師茂樹（花園大学）：morogram: Background, History, and Purpose of a Tool for East Asian Text Analysis.「morogram: 東アジア文献分析ツールの開発の経緯と目的」

2022 年 2 月 11 日: Ancient India meets Data-Science「古代インドとデータサイエンス」（オンライン）

天野恭子：The Result of the Two-Year SPIRITS Project and Our Vision for the Next Research.「２年間の SPIRITS プロジェクトの成果と今後の研究への展望」

Oliver Hellwig, Sebastian Nehrdich, Sven Sellmer: Dependency parsing of Vedic Sanskrit – Algorithms and linguistic conclusions.

京極祐希：One Step Further: Assessing Semantic Similarity in Sanskrit Using Word Embeddings with a Weighting Factor.「検証の次なる段階へ：重み付けを伴う単語分散表現によるサンスクリット文献の類似度推定」

宮川創（京都大学（当時)、現在は国立国語研究所）：Computational Stylometric Analysis on

Intertextuality in Historical Written Languages: A Case Study of Coptic.
「文献言語における間テクスト性の計算言語学的・計量文献学的分析：コプト語における事例研究」

夏川浩明：Visualization meets Ancient India: Mapping the Structure of Vedic Texts.
「可視化と古代インド研究：ヴェーダ文献の構造のマッピング」

このSPIRITSプロジェクトを足掛かりに、さらに研究を発展させるプロジェクトとして、科研費国際連携研究「ヴェーダ文献における言語層の考察とそれを利用した文献年代推定プログラムの開発」研究代表者：天野恭子(研究期間2021-2026)が発足した。この研究は、挑戦的研究で可能にしたヴェーダ語データ作成と、上述のヴェーダ文献成立過程の解明に情報学的手法を取り入れる研究を受け継ぎ、さらに文献の成立年代推定を目指すプロジェクトである。これまでも共同研究を行ってきたOliver Hellwigが、ドイツで展開しているChronBMM – Bayesian Mixture Models für die Datierung von extkorpora「テキストコーパスの年代推定のためのベイズ混合モデル」とも協働し、コプト語やアイヌ語、琉球語等で人文情報学的研究を発展させている宮川創（国立国語研究所）をメンバーに加えた。この国際連携研究からもすでに下記の２つの発表を行い、今後もさらに発信する予定である。この研究の発足に大きな足掛かりとなったことが、本挑戦的研究の最も大きな成果であると言える。

Kyoko Amano, Hiroaki Natsukawa. "Visualization of the relationship among Vedic texts and observation of the development of Yajurveda texts". A Three Day International Seminar on Paninian Grammar & its Applications
13-15 February 2023. Central Sanskrit University, Ganganath Jha Campus, Prayagraj, UP. 2023/2/14.

Oliver Hellwig, Sven Sellmer, Kyoko Amano. "The Vedic corpus as a graph. An updated version of Bloomfield's Vedic Concordance". The 18th World Sanskrit Conference 2023. 9-13 January 2023. The Australian National University, Canberra. 2023/1/12.

５．主な発表論文等

〔雑誌論文〕　計4件（うち査読付論文　4件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 天野恭子、夏川浩明	４．巻 -
２．論文標題 古代インド文献の文献間影響関係の可視化	５．発行年 2021年
３．雑誌名 可視化情報シンポジウム 2021 講演論文集	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計12件（うち招待講演　6件／うち国際学会　4件）

１．発表者名 Oliver Hellwig, Sven Sellmer, Kyoko Amano
２．発表標題 The Vedic corpus as a graph. An updated version of Bloomfield's Vedic Concordance
３．学会等名 The 18th World Sanskrit Conference 2023. 9-13 January 2023（国際学会）
４．発表年 2023年

１．発表者名 Kyoko Amano, Hiroaki Natsukawa
２．発表標題 Visualization of the relationship among Vedic texts and observation of the development of Yajurveda texts
３．学会等名 A Three Day International Seminar on Paninian Grammar & its Applications 13-15 February 2023（招待講演）（国際学会）
４．発表年 2023年

１．発表者名 Kyoko Amano
２．発表標題 The Result of the Two-Year SPIRITS Project and Our Vision for the Next Research
３．学会等名 Ancient India meets Data-Science. The 2nd and concluding Workshop for the SPIRITS project "Chronological and Geographical Features of Ancient Indian Literature Explored by Data-Driven Science"（招待講演）
４．発表年 2022年

１．発表者名
天野恭子

２．発表標題
Maitrayani Samhita IV 14（kamya-pasuのためのrc）とRgveda I巻

３．学会等名
2021年度インド思想史学会

４．発表年
2021年

１．発表者名
Kyoko Amano

２．発表標題
Historical Background of the Formation of the Vedas in Ancient India as Deciphered from the Visualization of the Influential Relations among the Vedic Texts

３．学会等名
KUDH International Conference. Digital Transformation in the Humanities（招待講演）

４．発表年
2021年

１．発表者名
天野恭子、夏川浩明

２．発表標題
古代インド文献の文献間影響関係の可視化

３．学会等名
可視化情報シンポジウム 2021（招待講演）

４．発表年
2021年

１．発表者名
Kyoko Amano

２．発表標題
Problems in the Formation of the Vedas, Ancient Indian Religious Texts

３．学会等名
Dynamism of Social Context Deciphered by a Linguistic Analysis of Ancient Literature

４．発表年
2021年

１．発表者名
Kyoko Amano

２．発表標題
Relationship Among Vedic Schools Deciphered by the Visualization of Mantra Collocation

３．学会等名
Dynamism of Social Context Deciphered by a Linguistic Analysis of Ancient Literature

４．発表年
2021年

１．発表者名
Kyoko Amano

２．発表標題
Diversity of Vedic ritual. Its different origins, innovations and the composition of the canons

３．学会等名
Letture Vediche: Il dono: croce e delizia dei brahmani（招待講演）

４．発表年
2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

データ駆動型科学が解き明かす古代インド文献の時空間的特徴
https://ancientindia-datascience.jp/

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	ヘルヴィック　オリヴァー (Hellwig Oliver)	デュッセルドルフ大学	

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	京極　祐希 (Kyogoku Yuki)	ライプツィヒ大学	

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2件

国際研究集会	開催年
Ancient India meets Data-Science. The 2nd and concluding Workshop for the SPIRITS project "Chronological and Geographical Features of Ancient Indian Literature Explored by Data-Driven Science"	2022年～2022年
国際研究集会 Dynamism of Social Context Deciphered by a Linguistic Analysis of Ancient Literature.	開催年 2021年～2021年

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			
ドイツ	Dusseldorf University	Leipzig University		