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

平成　３０　年　　６　月　２２　日現在

機関番号：　１２６０８
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2017
課題番号：　２６３７０５３０

研究課題名（和文）和歌用語シソーラスの開発と用語空間記述に関する基礎研究

研究課題名（英文）A development of thesaurus and fundamental study of description of longitudinal spacial chages of classical Japanese poetic vocabulary

研究代表者

山元　啓史（Yamamoto, Hilofumi）

東京工業大学・リベラルアーツ研究教育院・教授

研究者番号：３０２４１７５６

交付決定額（研究期間全体）：（直接経費）　　2,100,000円

研究成果の概要（和文）：本研究の目的は、従来の和歌用語のシソーラス八代集対応版を、二十一代集(905年頃~1439年)対応版に拡張することである。同時に、その開発を通して、和歌を基盤とする534年間におよぶ古代語の用語空間分析（語彙体系変遷）の理論化を試みた。シソーラスデータ入力を二十一代集のすべてについて完了し、シソーラスコードシステムとその体系ができあがった。シソーラス体系における単語の記述が統一的かどうかを確かめた結果、シソーラスの体系は現代語の体系としては良好ではあるが、古代語の単語間の距離計算に矛盾があることがわかった。

研究成果の概要（英文）：The purpose of this research is to extend the conventional version of the Hachidaishu thesaurus (ca.905-1205) to include the Nijuichidaishu (ca.905-1429). At the same time, through its development, we attempt to develop a theory of the term space analysis (vocabulary system change) encompassing 534 years of ancient words based on classical Japanese poetry. Thesaurus data input was completed for all of the Nijuichidaishu, and the thesaurus code system and its vocabulary index system were completed. As a result of checking whether the description of the word in the thesaurus system is unified or not, we found that the system of the thesaurus itself is consistent with modern language. However, there exist contradictions when utilizing the distance calculation between ancient words.

研究分野：　言語学，通時言語学

キーワード：　和歌　言語記述　シソーラス　系列比較　漸近的語彙対応　グラフ理論　通時言語学　ネットワーク

１．研究開始当初の背景

従来、古代語の研究は、研究者が自分の目で読み、研究者が生まれながらにして持つ現代語の知識から推論して、古代語を研究してきた。言語の意味、語と語の相互関係は、その言語が使われた当時の文化に依存するにもかかわらず、現代語の知識を使って分析してきた。人間が言語を見て分析する方法では、ありのままに分析したいところではあるが、意識的に現代語の知識を使わずに分析しようとしても分析できるものではない。当時の文化に依存した語と語の相互関係は文脈に現れる。文脈とは、言語情報だけでなく、人間社会での常識、すなわち言語にならない情報も含まれる。現代語であれば、文中の情報だけでなく、現代人の知識とともに文脈が意味あるものとして理解される。常識であるから記述されない。古代においても普通なら常識は記述されることはない。当時の人々のみが知り得るものである。もちろん、現代において古代の常識を得る方法はむずかしいと考えられる。たとえば、ある語が美しいことばなのか、醜いことばなのか、驚きを秘めたことばなのか、日常的なことばなのか、判別することはむずかしい。意味だけでなく、語の持つ価値観を見つけることは難しそうだ。昔のことばがどんな感じでやりとりされていたのかは、（わかったつもりになったとしても）本当はわかったわけではない。

本プロジェクトの研究者の興味は素朴である。「当時の日本語の姿は実際どうだったのか」である。現代人の知識に依存することなく、データサイエンスとして、古代日本語の姿を追求するための環境作りが本研究の目的であり、以上の説明がその背景である。

表 1: シソーラスなしでは同語として計算できない例

かな表記	実際に和歌に出現する実例
たつた	立田, 竜田, 龍田, ...
たつらむ	立つらん, 立らん, 立覧, ...
ちぎりけむ	契りけん, 契けむ, 契けん, 契剣, ...
おもふてふ	思ふてふ, 思てふ, 思ふ蝶, 思蝶, ...
えてしがな	得てしかな, 得てし哉, ...

２．研究の目的

本研究の目的は、従来の和歌用語のシソーラス八代集対応版を、二十一代集(905 年頃~1439 年)対応版に拡張することである。代表者はこれまでに 10 年以上の歳月をかけて、和歌用の形態素解析辞書とシソーラス（語彙体系用語集）を開発してきた。形態素解析用の辞書は、25 年度（基盤研究Ｃ）までに二十一代集対応版が完成している。一方、シソーラス（さまざまな単語の表記を同一視するか、あるいは異なる意味を持つものか、他などを判定するための語彙一覧）については、二十一代集に対応できていない。

本研究において、シソーラスの二十一代集対応版を目指すとともに、その開発を通して、和歌を基盤とする 534 年間におよぶ古代語の用語空間分析（語彙体系変遷）の理論化を試みた。

３．研究の方法

平成２６年度では、シソーラスデータ入力と計算処理の検討を行った。その計算処理は、２段階からなり、形態素解析と分類コードづけを行った。和歌の単語切りだし、未知語を追加登録しつつ、分割単位の修正作業を行った。シソーラスデータ入力については t2c（token to code）というプログラムを用いて、形態素解析されたそれぞれの単語に分類番号（国立国語研究所開発分類語彙表準拠）をつけ、目視による確認と修正を行う作業を行った。意味の検討も行い、任意の２語は互いに意味が近いか、同じと扱ってよいか、表記のみの違いであるか、表記は異なるが、意味は近いか、など、手作業にて行った。八代集以後に初出する語のデータは当初含まれていなかったのでそれらの語のデータ追加を重点的に行った。

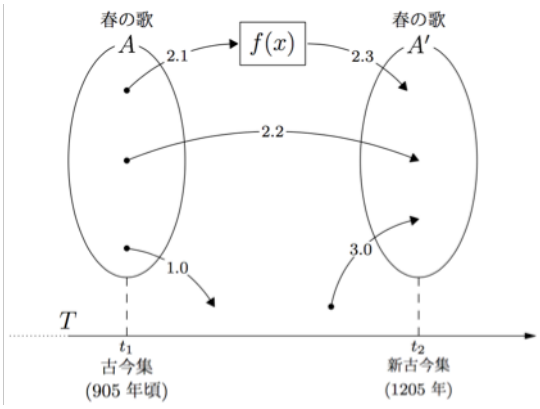

図 1: 通時軸における系列比較モデル: 横軸 T は時間を表す。$f(x)$ は A の任意の要素 x を A' の要素とするための関数。A は t_1 の時に発生した、あるまとまりを持った内容 (例: 古今集の春の部)、A' は t_2 時に発生した、A に対応するまとまりを持った内容（例: 新古今集の春の部)。

４．研究成果

シソーラスデータ入力を二十一代集のすべてについて完了し、シソーラスコードシステムとその体系ができあがった。

つぎにシソーラス体系に均一的な基準や矛盾がないかどうかを確かめるため、さらに和歌語彙空間分析を行うために、単語相互間の距離の計算方法の検討をおこなった。計算方法は、ネットワーク分析で行われる

linkcomm（Ahn et al. 2010）に加え、Google の Word2vec（Mikolov, et al.)の方法論を用い、和歌データにおいても、単語間距離が取り扱えることがわかった。

系列比較モデルを開発し、それを用いて、時代の異なる語の相違・差分の分析を行った。また、linkcomm モデルを利用し、任意の２時代の比較（差分）の計算がシソーラスによって可能であるかどうかを確かめた。さらに、Word2vec を利用し、機械学習による語の空間の計算が可能かどうかを実験した。その結果、シソーラスの体系は現代語の体系としては良好ではあるが、古代語の単語間の距離計算において、矛盾をきたす結果となった。

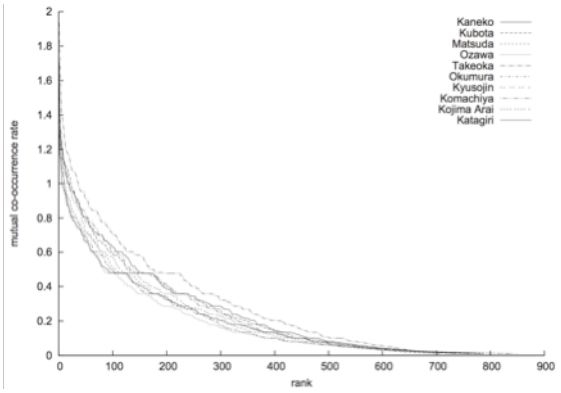

図 2: 古今和歌集とその現代語訳 10 種を対応処理した結果、得られた対応推定値の分布。縦軸は推定値、横軸はランク。

そこで、シソーラス開発の評価は、漸近的語彙推定システム（現代語と古代語のパラレルコーパスを用いて、文字面によらぬ用語対応関係を計算し、シソーラスコードの割り振りを決定するシステム）による方法、系列比較モデルによる方法、２段階によって行った。ネットワーク分析に用いられる手法のひとつである、コミュニティ発見の手法を用いて、従来の語の相互関係を（単語は人々、関連付けを社会・街・集団といった）コミュニティと見做して、分析する方法を用い、古代語の語の空間分析を行った結果、対比する語（桜、梅、橘）の間の距離や、従来述べられている語相互の関係だけでなく、必須と考えられている関係には実はさらに必要な条件があったこと（たとえば、山吹には、蛙・井手と共に使われるという常識があるが、実はそれには単に使えばよいのではなく、八重といういくつにも重なるという条件が必要であること、山吹は実は一重の山吹ではなく、八重山吹であること、歌ことば辞典ではそれらが述べられていないこと等）がわかった。

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　3　件）

① ホドシチェク，ボル．　山元 啓史．歌ことば「橘」「梅」「桜」における関連対の抽出，人文科学とコンピュータシンポジウム論文集，情報処理学会，Vol. 2017，No. 2，pp. 207-212，Dec. 2017.（査読有）

② Hilofumi Yamamoto. Bor Hodoscek. Development of the dictionary of poetic Japanese description, Digital Scholarship in History and the Humanities, the 6th conference of the Japanese Association for Digital Humanities, pp. 44-46, Sep. 2016.（査読有）

③ 山元 啓史，村井源，ボル ホドシチェク．二十一代集シソーラスのための漸近的語彙対応システムの開発，人文科学とコンピュータシンポジウム論文集，Vol. 2014，No. 3，pp. 157-162，Dec. 2014.（査読有）

〔学会発表〕（計　5　件）

① Hilofumi Yamamoto, Bor Hodoscek. Relationships between Flowers in a Word Embedding Space of Classic Japanese Poetry, JADH2017 Proceedings of the 7th Conference of Japanese Association for Digital Humanities “Creating Data through Collaboration”, Faculty of Culture and Information Science, Doshisha University, Vol. 2017, pp. 70-72, Sep. 2017.（査読有）

② Hilofumi Yamamoto, Bor Hodoscek. Thesaurus of classical Japanese poetic vocabulary for the Nijuichidaishu (ca. 905-1439), 14th International Conference of European Association for Japanese Studies BOOK OF ABSTRACTS, p. 86, Aug. 2014.（査読有）

③ Bor Hodoscek, Makiro Tanaka, Hilofumi Yamamoto. A Visualization and Analysis System for Japanese Language Change: Quantifying Lexical Change and Variation using the Serial Comparison Model, JADH Conference 2014 ABSTRACTS, p. 3, Sep. 2014.（査読有）

④ Hilofumi Yamamoto, Bor Hodoscek, Hajime Murai. Development of an Asymptotic Word Correspondence System between Classical Japanese Poems and their Modern Translations, JADH Conference 2014 ABSTRACT, p. 40, Sep. 2014. （査読有）

⑤ Hilofumi Yamamoto, Bor Hodoscek. Thesaurus of classical Japanese poetic vocabulary for the Nijuichidaishu (ca. 905-1439), 14th International Conference of European Association for Japanese Studies, 14th International Conference of European Association for Japanese Studies BOOK OF ABSTRACTS, p. 86, Aug. 2014. （査読有）

〔図書〕（計 1 件）

① 山元 啓史. 通時コーパスによる言語の研究, コーパスと日本語史研究, ひつじ書房, Vol. 127, pp. 17-35, Oct. 2015.

〔その他〕（6 件）

① ホドシチェク・ボル, 山元 啓史「歌ことば『橘』『梅』『桜』における関連対の抽出」人文科学とコンピュータシンポジウム, じんもんこん 2017 ベストポスター賞受賞. 2017.12.9.

② 山元 啓史「視野が広がる！？大学研究室探検隊 Vol.6, 東京工業大学山元啓史研究室: コンピュータを利用して言語を可視化する研究」サクセス 15, グローバル教育出版. pp.16-19. 2018.2 月号

③ 山元 啓史「目で見てわかる昔の日本語と今の日本語：タイムマシンに乗らずに行ける昔の世界」ひらめき☆ときめきサイエンス実施, 2017.8.2.

④ 山元 啓史「目で見てわかる昔の日本語と今の日本語：タイムマシンに乗らずに行ける昔の世界」ひらめき☆ときめきサイエンス実施, 2016.8.3.

⑤ 山元 啓史「二十一代集シソーラスのための漸近的語彙対応システムの開発」2015 年度情報処理学会山下記念研究賞受賞

⑥ 山元 啓史「目で見てわかる昔の日本語と今の日本語：タイムマシンに乗らずに行ける昔の世界」ひらめき☆ときめきサイエンス実施, 2015.8.5.

６．研究組織

(1)研究代表者

山元 啓史（Yamamoto Hilofumi）
東京工業大学・リベラルアーツ研究教育院・教授
研究者番号：30241756