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　６ 年　 ５ 月 ３１ 日現在

機関番号：１２６０１
研究種目：若手研究
研究期間：2018～2023
課題番号：１８Ｋ１２３６１

研究課題名（和文）深層学習を用いたスペイン語の通時的研究

研究課題名（英文）Diachronic Studies of Spanish using Deep Learning Methods

研究代表者
川崎　義史（KAWASAKI, Yoshifumi）

東京大学・大学院総合文化研究科・准教授

研究者番号：４０７９４７５６

交付決定額（研究期間全体）：（直接経費）　　1,700,000円

研究成果の概要（和文）：（1）シミュレーションなどの計算的手法により言語変化を研究する計算歴史言語学という新たな分野を開拓した。（2）計量文献学的手法により，スペイン文学史において非常に重要な位置を占める文学作品三点（『贋作ドン・キホーテ』，『アマディス・デ・ガウラ』，『ティラン・ロ・ブラン』）の成立過程に新たな光を当てた。（3）単語の時空間分散表現を活用して，中近世スペイン語古文書の作成年代・地点を正確に推定するアルゴリズムを開発した。

研究成果の学術的意義や社会的意義
（1）計算歴史言語学（シミュレーションなどの計算的手法による言語変化の研究）という新たな分野を開拓した点が，言語の歴史的研究への貢献である。（2）スペイン文学史において非常に重要な位置を占める文学作品三点の成立過程に新たな光を当てた点が，スペイン文学研究への貢献である。（3）中近世スペイン語古文書の作成年代・地点を正確に推定するアルゴリズムを開発した点が，スペイン語の歴史的研究への貢献である。

研究成果の概要（英文）：(i) We developed a new field of "computational historical linguistics" to study language change through computational methods including simulation techniques; (ii) we shed new light on the genesis of three literary masterpieces in Spanish and Catalan ("Quijote" de Avellaneda, "Amadis de Gaula", and "Tirant lo Blanc") using stylometric methods; and (iii) we developed an algorithm to accurately estimate the date and location of Spanish notarial documents by utilizing spatiotemporal embeddings.

研究分野：スペイン語史

キーワード：スペイン語　ロマンス語　深層学習　計算言語学　言語変化　計量文献学　計算歴史言語学　計算歴史方言学

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

研究開始当初，自然言語処理分野などで深層学習技術の目覚しい発展が遂げられつつあった。深層学習技術を用いることで，機械翻訳をはじめ様々なタスクにおいて大幅な性能向上が達成されるようになっていた。深層学習では，大規模コーパスから単語や文などの密なベクトル表現を学習する。これにより，従来の機械学習手法では困難だった意味や用法の扱いや定量化が容易になった。また，深層学習技術を用いることで，言語現象の柔軟なモデル化も可能になった。深層学習技術を活用することで，申請者の専門とするスペイン語史の課題にも新たな視点から取り組むことができるのではないかと考え，この研究を開始するに至った。

２．研究の目的

本研究の目的は，深層学習技術を用いることで，スペイン語の通時的研究に新たな知見を提供することである。具体的には，下記の 3 つの課題に取り組むことを予定していた：

（1）類推による不規則動詞の規則化のモデル化

本課題の目的は，類推による不規則動詞の規則化のメカニズムの解明である。この過程を深層学習モデルにより計算機上で再現し，その振る舞いを観察・分析する。言語変化だけでなく子供の言語獲得などにおいても観察される類推のメカニズムの解明には，大きな言語学的意義がある。

（2）中近世スペイン語品詞解析器の開発

本課題の目的は，中近世スペイン語用の品詞解析器の開発である。各単語に品詞を自動で付与する品詞解析器は，文献の計量的分析には不可欠なツールである。中近世スペイン語用の品詞解析器は存在しないため，その開発には大きな意義がある。品詞解析器を利用して中近世スペイン語文学作品の計量文献学的研究を行うことが，本課題の最終目的である。

（3）中近世スペイン語古文書の年代推定・地点推定アルゴリズムの性能向上

本課題の目的は，中近世スペイン語古文書の年代推定・地点推定アルゴリズムの性能向上である。具体的には，年代誤差を 10 年以下，地点誤差を 50㎞ 以下にすることを目標とする。現存する文献史料に立脚する歴史学的研究において，文献の真贋や作成年代や作成地点の信頼性は第一義的なものである。文書の作成年代と作成地点を正確に推定するアルゴリズムの開発は，スペイン語文献学への大きな貢献となる。

３．研究の方法

（1）類推による不規則動詞の規則化のモデル化

当初は，不規則動詞の規則化に注目していたが，最終的には動詞活用全体まで分析射程を広げた。実験の流れは以下の通りである。まず，過去のある時点における対象言語の動詞に関して，活用情報（法・態・時制・人称・数）と全変化形を列挙したデータセットを作成する。次に，入力を不定詞と活用情報，出力を活用形（文字列）として，深層学習モデルの一つ系列変換モデルに動詞活用パターンを学習させる。その後，テストデータで評価実験を行い，モデルの出力の正誤分布や誤出力形を実際の言語変化のパターンと比較対照する。そして，頻度や強勢位置や活用情報など潜在的に類推に影響する要因を検証する。

（2）中近世スペイン語品詞解析器の開発

当初は，深層学習モデルの一つ再帰型ニューラルネットワークを用いて，中近世スペイン語品詞解析器をゼロから開発することを目標としていた。しかし，研究開始後に，高性能な現代スペイン語用の解析器が利用可能になり，データの前処理や事後修正を行うことで，中近世語にも十分に適用できることが判明した。そのため，自前の中近世語用の解析器を開発するという当初の予定を変更し，最終的な目的であった計量文献学的研究に注力することにした。

（3）中近世スペイン語古文書の年代推定・地点推定アルゴリズムの性能向上

深層学習技術を用いた文書分類と回帰の枠組みで，この課題に取り組む。文書分類の枠組みでは，年代（10 年単位）と地点（行政単位の州や県）を離散変数として扱う。回帰の枠組みでは，年代（年）と地点（緯度・経度）を連続変数として扱う。コーパスとして，CODEA（Corpus de Documentos Españoles Anteriores a 1800）を使用する。

４．研究成果

（1）類推による不規則動詞の規則化のモデル化

ラテン語とロマンス語を対象として，前者から後者への動詞活用の形態変化のモデル化に取り組んだ。両言語を対象としたのは，利用可能なデータが多く，先行研究の蓄積が充実しているためである。上述の手法での実験の結果，全体として，各活用の正答率と実際の言語変化の大きさは強い負の相関を示した。つまり，モデルの正答率が低い（高い）活用ほど，実際の変化度合いが大きく（小さく）なる傾向が見られた。また，散発的ではあるものの，不規則動詞の規則化も観察された。規則化した形が，ロマンス語学で想定される形と一致していたことは興味深い。これらの結果は，モデルが実際の言語変化を部分的に再現できていることを示唆する。本研究は，計算歴史言語学（計算的手法による言語変化の研究）を開拓した点で意義がある。当初は英語や日本語やスペイン語を対象とした分析も予定していたが実施できなかったため，今後の課題としたい。（研究業績［2］）

（2）中近世スペイン語文学作品の計量文献学的研究

上述の通り，当初の予定を変更し，最終的な目的であった中近世スペイン語文学作品の計量文献学的研究を実施した。具体的には，①『贋作ドン・キホーテ』，②『アマディス・デ・ガウラ』，③『ティラン・ロ・ブラン』の三作品を分析対象とした。ただし，『ティラン・ロ・ブラン』はカタルーニャ語の作品である。いずれも，スペイン文学史において非常に重要な位置を占める作品であり，その成り立ちを解明することには大きな意義がある

①『贋作ドン・キホーテ』（1614 年出版）は，スペイン黄金世紀を代表するセルバンテスの小説『ドン・キホーテ』（前編 1604 年出版，後編 1615 年出版）のパロディ作品である。作者として表紙に記されているアベジャネーダはペンネームであると考えられている。そのため，ペンネームの裏に隠れている真の作者は誰なのかについて様々な仮説が提示されてきた。候補者は，セルバンテスを含め，同時代の作家達である。計量文献学の見地からの先行研究も存在するが，（i）文体特徴として内容語を使用している，（ii）文体特徴が数千種類にも上る，（iii）ハイパーパラメータの影響が検証されていないなど多くの方法論的問題があった。そこで，本研究では，（i）内容の影響を受けにくい品詞 n-gram を文体特徴として使用し，（ii）その種類数を数百に抑え，（iii）ハイパーパラメータの影響も考慮した上で検証を行なった。サポートベクトルマシンとロジスティック回帰を分類器として教師あり学習を行なった。分析の結果，驚くべきことに，『贋作ドン・キホーテ』の真の作者がセルバンテス自身である可能性が否定できないことが示唆された。一方で，セルバンテス作の可能性が指摘されてきた「にせの伯母さん」（『模範小説集』所収）は，セルバンテスの作品である可能性が極めて高いことが示唆された。併せて，品詞 n-gram により，スペイン語散文作品の作者をほぼ完璧に判別できることを実験的に示した。（研究業績［3］）

②中世スペイン語騎士道物語『アマディス・デ・ガウラ』（1508 年出版）とその続編『エスプランディアンの武勲』（1510 年出版）は，ともにモンタルボによる作品とされているが，その成立過程には不明瞭な点が多く存在する。文献学の定説では，『アマディス・デ・ガウラ』の第 1 部から第 3 部はモンタルボが既存の版を改訂したもので，『アマディス・デ・ガウラ』第 4 部と『エスプランディアンの武勲』はモンタルボ自身が書き加えたものだと考えられている。しかし，この定説には計量文献学の見地からの裏付けはない。そこで，本研究では，両作品を世界で初めて計量文献学的手法で分析した。具体的には，品詞 n-gram を文体特徴として，クラスタリング分析と主成分分析による教師なし学習を行なった。分析の結果，（i）モンタルボによる『アマディス・デ・ガウラ』第 1 部の改訂は最小限である，（ii）第 2 部と第 3 部の改訂は大規模である，（iii）第 4 部と『エスプランディアンの武勲』はモンタルボの自筆のものである可能性が高いことが示唆された。（研究業績［4］）

③中世カタルーニャ語文学の金字塔『ティラン・ロ・ブラン』（1490 年出版）に関して，文献学の見地から，単一作者説（マルトゥレイの単著）と二重作者説（マルトゥレイとガルバの共作）の二つの仮説が提示されてきた。計量文献学的手法を用いた先行研究では，複数作者説を支持する結果が報告されていた。しかし，（i）著者推定において有効性が確認されていない単語長分布を文体特徴として使用している，（ii）地の文と会話を区別していない，（iii）モデル選択法が不適切である等の問題点が存在した。そこで，本研究では，（i）有効性が確認されている品詞 n-gram を文体特徴として使用し，（ii）地の文と会話を区別し，（iii）適切なモデル選択法を採用して検証を行なった。その結果，先行研究とは反対に，二重作者説よりも単一作者説の方が蓋然性が高いことが示唆された。（研究業績［7］）

（3）中近世スペイン語古文書の年代推定・地点推定アルゴリズムの性能向上

①年代と地点を離散変数として扱う文書分類に基づく推定方法は以下の通りである。まず，各単語を年代と地点に埋め込み，時空間分散表現を獲得する。これにより，各単語ベクトルは，出現した文書の年代と地点の情報を保持するようになる。分散表現の学習には，文字列の情報を考

慮したアルゴリズム fastText を採用した。これにより，未知語の分散表現の獲得も可能になった。次に，各文書を，そこに含まれる単語の分散表現の平均ベクトルとして表す。最後に，文書ベクトルを年代と地点に関する離散確率分布に変換することで，作成年代と作成地点を予測する。実験の結果，年代誤差は 24 年，地点誤差は 104km となった。ある程度正確に年代推定・地点推定が可能になったことには，文献学的に大きな意義がある。しかし，推定誤差の目標の 10 年と 50km は達成できていないため，更なる性能改善が今後の課題である。本手法の特長は二つある。一点目は，単語ベクトルのノルムに基づき推定に大きく寄与する単語を特定できるため，推定結果が文献学的に解釈しやすい点である。二点目は，推定結果を確率分布として提示することで，推定の信頼度を提示できる点である。（研究業績［6］）

加えて，本研究では，計算歴史方言学（計算的手法による歴史方言学の研究）を開拓した。具体的には，推定の副産物として得られる単語の時空間分散表現を活用して，各年代・地点に特徴的な単語群を網羅的に抽出し，スペイン語文献学の知見と比較対照した。その結果，多くの点で一致を示しており，本手法の有用性を裏付けている。（研究業績［5］）

②年代と地点を連続変数として扱う回帰では，何らかの形で表現された文書ベクトルを入力として，回帰モデルにより年代と地点を推定する。様々な文書ベクトルと回帰モデルの組み合わせで網羅的に実験を行い，最適な組み合わせを検証した。文書ベクトルには，文字 n-gram，単語 n-gram，Doc2Vec，BERT による分散表現を選択した。回帰モデルには，加重平均 k-NN，ガウス過程回帰，リッジ回帰，サポートベクター回帰を選択した。網羅的な検証の結果，文書ベクトルとしては，Doc2Vec や BERT による分散表現ベースのものよりも，n-gram の方が優れていることが判明した。これは，文書の全体的な情報よりも，n-gram が捉える具体的な単語や文字の連続が推定に効果的であることを示唆している。回帰モデルとしては，最も単純なモデルである加重平均 k-NN の性能が最も高くなった。性能はやや劣るが，予測の信頼度として推定分散が求まる点で，ガウス過程回帰の有効性も確認された。最良の場合，年代誤差は 24 年，地点誤差は 118km となった。文書分類と比べ，年代推定の性能は同等だったが，地点推定の性能はやや悪化した。推定誤差の目標は達成できていないため，更なる性能改善が今後の課題である。（研究業績［1］）

研究成果の多くは学会発表や論文として発表済みである。しかし，一部の研究成果の英語論文化が遅れているため，今後の課題としたい。

主な研究業績

［1］川崎義史，永田亮．(2022)．「スペイン語古文書の年代・地点推定のための最適な手法の探求」．『言語処理学会 第 28 回年次大会 発表論文集』，2022 年 3 月，pp. 1606-1611.

［2］川崎義史．(2023)．「RNN はラテン語からロマンス語への活用変化を再現するか？」．『言語処理学会 第 29 回年次大会 発表論文集』，2023 年 3 月，pp. 1913-1917.

［3］Kawasaki Yoshifumi. (2021). Stylometric Analysis of Avellaneda's *Don Quijote*. *12th International Conference on Corpus Linguistics*. University of Murcia, Spain (online). April 2021.

［4］Kawasaki Yoshifumi. (2022). A Stylometric Analysis of *Amadís de Gaula* and *Sergas de Esplandián*. *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities (NLP4DH 2022)*, pp. 1-7.

［5］Kawasaki Yoshifumi. (2023). Computational Historical Dialectology Using Spatiotemporal Embeddings. *The 56th Annual Meeting of the Societas Linguistica Europaea (SLE 2023).* National and Kapodistrian University of Athens, Greece. August-September 2023.

［6］Kawasaki Yoshifumi. (2023). Dating and Geolocation of Medieval and Modern Spanish Notarial Documents Using Distributed Representation. *Quantitative Approaches to Universality and Individuality in Language*, Makoto Yamazaki and Haruko Sanada (eds.), Berlin, Boston: De Gruyter, pp. 63-72.

［7］Kawasaki Yoshifumi. (2023). Revisiting Authorship Attribution of *Tirant lo Blanc* Using Parts of Speech *n*-grams. *Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities & 8th International Workshop on Computational Linguistics for Uralic Languages (NLP4DH & IWCLUL 2023)*, pp. 16-26.

５．主な発表論文等

〔雑誌論文〕　計6件（うち査読付論文　4件／うち国際共著　1件／うちオープンアクセス　4件）

１．著者名 Yoshifumi Kawasaki & Andres Enrique-Arias	４．巻 －
２．論文標題 El calco de los pluralia tantum del hebreo en las traducciones biblicas castellanas medievales y renacentistas	５．発行年 2022年
３．雑誌名 Traduccion biblica e historia de las lenguas iberorromanicas	６．最初と最後の頁 55-84
掲載論文のDOI（デジタルオブジェクト識別子） 10.1515/9783110770766-003	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 該当する

１．著者名 Yoshifumi Kawasaki	４．巻 －
２．論文標題 Dating and geolocation of medieval and modern Spanish notarial documents using distributed representation	５．発行年 2022年
３．雑誌名 Quantitative Approaches to Universality and Individuality in Language	６．最初と最後の頁 63-72
掲載論文のDOI（デジタルオブジェクト識別子） 10.1515/9783110763560-006	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 －

１．著者名 Yoshifumi Kawasaki	４．巻 －
２．論文標題 A Stylometric Analysis of Amadis de Gaula and Sergas de Esplandian	５．発行年 2022年
３．雑誌名 Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities	６．最初と最後の頁 1-7
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 オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 川崎義史	４．巻 －
２．論文標題 RNNはラテン語からロマンス語への活用変化を再現するか？	５．発行年 2023年
３．雑誌名 言語処理学会 第29回年次大会 発表論文集	６．最初と最後の頁 1913-1917
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 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 川崎義史，永田亮	4．巻 －
2．論文標題 スペイン語古文書の年代・地点推定のための最適な手法の探求	5．発行年 2022年
3．雑誌名 言語処理学会 第28回年次大会 発表論文集	6．最初と最後の頁 1606-1611
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 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 Yoshifumi Kawasaki	4．巻 －
2．論文標題 Revisiting Authorship Attribution of Tirant lo Blanc Using Parts of Speech n-grams	5．発行年 2023年
3．雑誌名 Proceedings of the Joint 3rd International Conference on Natural Language Processing for Digital Humanities & 8th International Workshop on Computational Linguistics for Uralic Languages	6．最初と最後の頁 16-26
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 オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計13件（うち招待講演　4件／うち国際学会　5件）

1．発表者名 Yoshifumi Kawasaki
2．発表標題 Datacion crono-geografica de los documentos hispanicos antiguos
3．学会等名 Faculte des lettres - Section d’espagnol, Universite de Lausanne（招待講演）
4．発表年 2022年

1．発表者名 川崎義史
2．発表標題 『アマディス・デ・ガウラ』の計量文献学的分析
3．学会等名 日本イスパニヤ学会第67回大会
4．発表年 2021年

１．発表者名
Yoshifumi Kawasaki

２．発表標題
Dating and geolocation of medieval and modern Spanish notarial documents using distributed representation

３．学会等名
QUALICO 2021（国際学会）

４．発表年
2021年

１．発表者名
Yoshifumi Kawasaki

２．発表標題
Stylometric analysis of Avellaneda’s Don Quijote

３．学会等名
12th International Conference on Corpus Linguistics（国際学会）

４．発表年
2021年

１．発表者名
川崎義史

２．発表標題
『ティラン・ロ・ブラン』の計量文献学的分析

３．学会等名
日本イスパニヤ学会第66回大会

４．発表年
2020年

１．発表者名
川崎義史

２．発表標題
文献学への自然言語処理技術の応用 ―スペイン語文献を例にして―

３．学会等名
言語処理学会第27回年次大会（招待講演）

４．発表年
2021年

１．発表者名
川崎義史

２．発表標題
黄金世紀散文作品の計量文献学的分析

３．学会等名
日本イスパニヤ学会第65回大会

４．発表年
2019年

１．発表者名
Yoshifumi Kawasaki

２．発表標題
Modelos probabilisticos para la datacion crono-geografica de documentos antiguos hispanicos

３．学会等名
VI Congreso Internacional de la Red CHARTA: Documentos y monumentos para la historia del espanol（国際学会）

４．発表年
2019年

１．発表者名
Yoshifumi Kawasaki, Andres Enrique-Arias

２．発表標題
Pluralia tantum en los romanceamientos medievales

３．学会等名
Congreso internacional “Biblias hispanicas: traduccion vernacula en la Edad Media y Renacimiento”（国際学会）

４．発表年
2018年

１．発表者名
川崎義史

２．発表標題
『贋作ドン・キホーテ』の計量文献学的分析

３．学会等名
日本イスパニヤ学会第64回大会

４．発表年
2018年

１．発表者名
Kawasaki Yoshifumi

２．発表標題
Computational historical dialectology using spatiotemporal embeddings

３．学会等名
The 56th Annual Meeting of the Societas Linguistica Europaea（国際学会）

４．発表年
2023年

１．発表者名
高村大也，永田亮，川崎義史

２．発表標題
深層学習時代の言語研究：基礎と導入

３．学会等名
日本英語学会第41回大会シンポジウム（招待講演）

４．発表年
2023年

１．発表者名
永田亮，川崎義史

２．発表標題
深層学習時代のコーパス分析：コーパス検索では見えないこと、深層学習で見えること

３．学会等名
英語コーパス学会第49回大会（招待講演）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
researchmap
https://researchmap.jp/16211665/

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

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			
スペイン	Universidad de Alcala	Universitat de les Illes Balears		