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　元 年　 ６ 月 １１ 日現在

機関番号：３２６２１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ０２５２７

研究課題名（和文）ソシアルメディアのスペイン語の変異の研究

研究課題名（英文）Variation on the Spanish in Social Media

研究代表者
Ｒ・ＴＩＮＯＣＯ　Ａｎｔｏｎｉｏ（R. TINOCO, ANTONIO）

上智大学・外国語学部・教授

研究者番号：８０２９６８８９

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）：いくつかの大陸にまたがる5億人以上の話者がいるスペイン語の言語的変異的な研究は大変困難で、一般的な研究方法は現地調査であるが、共時的な研究としては効率が悪い。
Twitterなどのソーシャルメディアの普及により、スペイン語のテキストを自動的に収集することが可能になったので、本研究では地理情報が含まれる5300万のツイートを収集し、どこから発信されたか、またその日時を正確に把握し、それらをデータベースに蓄積することが出来た。
地理情報システム（GIS）の手法を利用し、収集されたデータの地理的分布を観察し、多くの方言地図を作成した、収集された大量のデータによって量的・質的な研究が可能になった。

研究成果の学術的意義や社会的意義
語彙バリエーションの研究は伝統的に現地調査を通して比較的少数のインフォーマントを選び、狭い地理的な範囲で研究されてきたので、方言のコーパスを作成するには数ヶ月または数年間の作業に専念する多くの専門家の共同作業が不可欠だった。
現在、インターネット、特にツイッターのようなソーシャルネットワーク上でのデジタルテキストの普及で、比較的に短期間で多量のデータを収集することが可能になってきた。 このように作成されたデータベースは言語的変異という現象を、各地域でその特徴を理解するのに大変便利で、大いに応用できる研究方法になってきた。

研究成果の概要（英文）：The study of the linguistic variation of a language such as Spanish is difficult, because it is spoken by more than 500 million people spread over several continents. Current studies of variation are generally based on on-site surveys. This method is too slow for the investigation of synchronic variation.
Thanks to the widespread use of social media such as Facebook, Twitter, etc., which offer the technical possibility of collecting texts in several languages including Spanish, more than 530 million Tweets with geographical information have been collected in this project. We know their exact origin, as well as the date and time when it was written.
Data collected are then processed using geographical information system (GIS) techniques to observe the geographical distribution of the collected cases. In addition, the large amount of data collected makes both quantitative and qualitative study possible.

研究分野：スペイン語学

キーワード：スペイン語学　スペイン語圏　コーパス言語学　変異言語学　方言学　データベース　地理言語学

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

5 億人以上の話者がいるスペイン語の言語的変異的な研究は大変困難で、一般的な研究方法は現地調査である。日本では上田博人、高垣敏博、福嶌教隆、宮本正美などが多くの論文やデータベースの実績を残している。しかし、フィールドワークで大変貴重なデータは得られるが、全スペイン語圏のデータを集めるには長い時間がかかるという弱点がある。Twitter などのソーシアルメディアの普及により、テキストを自動的に収集することが可能になってから特に英語のソーシアルメディアの言語調査は普及したが、スペイン語の調査はまだ少ない。Twitter のようにデータに地理的な情報（主に発信地の経度と緯度）が含まれるようになってから地理言語学的な観点と語彙バリエーションの研究では大変有効である。このように得られたデータを処理するために一般的な自然言語処理の手法の他にインターネット、ネットワーク、地理情報システム（GIS）の技術が必要となったが、完全に共時的なコーパスを作成することも可能になった。

２．研究の目的

全スペイン語圏のインターネット上のソーシアルメディア（主に Twitter）のデータをプログラミングで自動的に収集し、専用のデータベースを構築することである。このような方法で集めたデータはスペイン語の変異言語学的な研究で利用する。具体的には語彙と文法の変異が主な研究対象となる。

収集したデータにはすべて発信地の経度と緯度があるので、GIS（地理情報システム）の技術を使いデジタル言語地図を作成する。データはおよそ 4 年間にわたり収集するので、通時的な研究となるが共時的な部分も否定できない。データベースに蓄積するツイートを分析し、内外の研究者と方法論について検討した上で、単独研究や共同研究の成果を国際学会で発表する。また、構築したデータベースをスペイン語学の研究者に提供することを考えている。目標としては地理情報付きの５０００万以上のツイートで、およそ５億以上の token である。

３．研究の方法

本プロジェクトのアプローチとしてインターネット上のソーシアルメディアのデータを収集することなので、主にツイッターの Streaming API（ver1.1）を利用し、自動的に収集し、大規模なコーパスを作成することにより、広大なスペイン語圏のスペイン語の語彙と文法の地理的な分析をすることである。利用するデータベースに関しては SQL 系（MySQL）と NoSQL 系（Elastic Stack）の両方を利用した。地理的な情報（経度、緯度）が含まれるデータは MySQL のデータベースに蓄積し、他の言語も含まれるデータは NoSQL 系のデータベースに蓄積し、目的により使い分けることにした。例えば、米国のスペイン語と英語の接触、あるいはスペインのスペイン語とカタルーニャ語の言語接触の現象を研究するために、可能な範囲で他の言語も NoSQL 系のデータベースで蓄積した。Elastic Stack のような NoSQL 系のデータベースは、ツイッターの JSON フォーマットをそのまま処理できるので、Kibana などで基本的な可視化もできる。

蓄積したデータをそれぞれの種類の特徴を利用し、token を中心に、正規表現を使って、地理情報などで検索することが出来るようになっている。MySQL 系のデータベースでは Fig 1 のインターフェースで検索した結果を主に CSV フォーマットでダウンロードし、次の処理に進む。

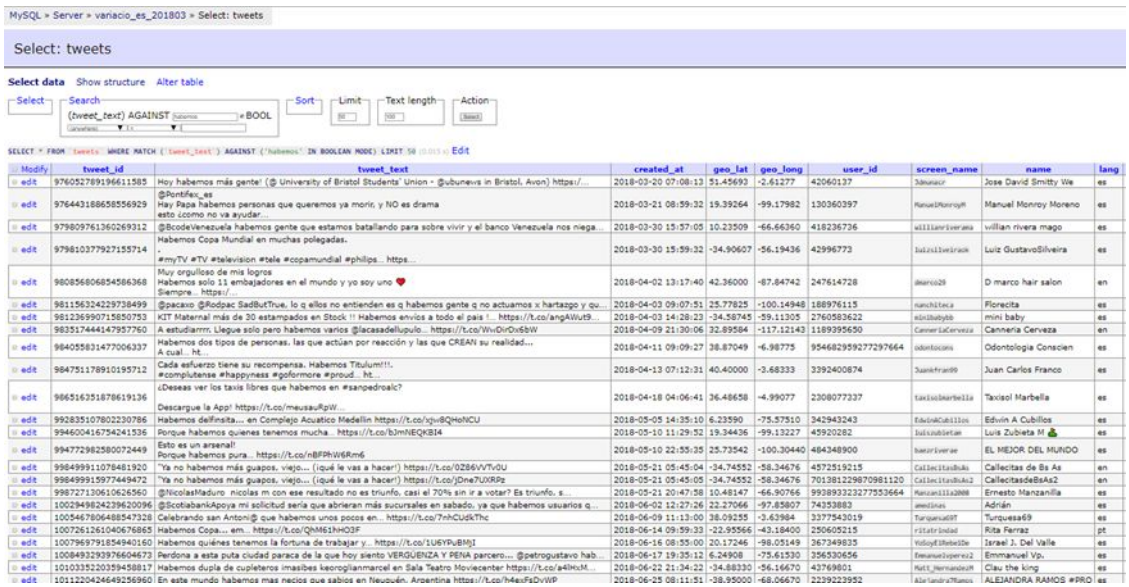

Fig. 1　データベースのインターフェース

ELK のデータベースは複数の検索方法があるが、Kibana を通して検索することがもっとも一般的である。Kibana のインターフェースを使って、ペルーに限って検索した結果は Fig.2 で表示

でき、またその結果を別のフォーマットでダウンロードすることも可能である。テキストの他にツイートのすべてのフィールドを表示することが出来る。この例では text、place.full_name(市、国)と発信の geo.coordinates(経度、緯度)である。

geo.coordinates:[-180 TO 180] AND place.country_code:PE

text	place.full_name	geo.coordinates
Ammmm Almuerzo 😎🍝🍝🍜 @ Metro - Av La Marina https://t.co/4ZhQhJxtNM	San Miguel, Peru	-12.077, -77.089
Time of cousin 🎮😉 #play #pes #cousin @ Ciudad de Dios - SJM https://t.co/Yvbag8VAgG	San Juan de Miraflores, Peru	-12.17, -76.972
Ok (@ Mercado Municipal "Santa Cruz" in Miraflores, Lima) https://t.co/cuvVkS1XZL	San isidro, Peru	-12.111, -77.05
Soy esa mujer la que sueña todo el tiempo despierta con algo mejor; y que es un misterio sin… https://t.co/FXgIObvn2j	Cajamarca, Peru	-7.164, -78.511
Cevichón. (@ La Caleta del Norte) https://t.co/NTHEYNebf1	Lima, Peru	-12.068, -77.074
Y así empezamos Marzo 👯🎀 https://t.co/He1I8Y8zIT	Ate, Peru	-12.029, -76.912
'imprescriptibilidad' es ahora una tendencia en #Lima https://t.co/MW1TnonVxL https://t.co/c4CNaadMDA	Lince, Peru	-12.085, -77.036
Hueon, 2300 pesos un menú completo entrada, fondo, postre, canchita, jugó natural y pisco sour… https://t.co/mV3rFOHE0U	Rimac, Peru	-12.043, -77.028
Hoy salimos así. Causa con ceviche de lapas. Verano peligroso. #catalina555 #comerycompartir… https://t.co/fPYBOQ9LCz	La victoria, Peru	-12.086, -77.019

Fig.2　Kibana（ELK）のデータベースのインターフェースの例

言語地図の作成は GIS 専用の QGIS と可視化（visualization）の Kibana（ELK）で簡単に行うことができる。研究者用のマニュアルを作成したので、既に複数の研究機関の学者や学生が利用している。例えば、語彙のバリエーションとしては南米での chifa（中華料理のレストラン）の仕様分布を Kibana のヒートマップで表示すると Fig.3 のようになる。主にペルーとエクアドルで使われているということが分かりやすい。

Fig. 3　chifa の使用分布

QGIS で hiciste に対して hicistes の過去形の二人称で見られるバリエーション現象も量的に分析し、その分布を示す言語地図（統計色分けマップ、choropleth）は Fig.4 のようになる。

Fig. 4　hicistes vs hiciste の使用分布(色分け)

４．研究成果

本プロジェクトでツイッターのデータを自動的に収集し、大規模なコーパスに蓄積にすることにより、広大なスペイン語圏のスペイン語の語彙と文法の地理的な分析をし、語彙と文法のバリエーションの研究の方法論が確立できた。
例えば、語彙のバリエーションとしてはメキシコの ahorita, luego luego など、ベネズエラの chamo, piche など、またはアルゼンチンでは boludo、mina、または a mi lado es un poroto のような独特な表現を、地理的な分布及び用法についても調べること可能になった。また、hiciste と hicistes のような動詞の過去形の二人称で見られるバリエーション現象も量的な分析と、その分布を示す言語地図を作成することが可能である。
本プロジェクトの４年間でツイッターの Streaming Api から収集したツイートをデータベースに蓄積し、その件数は 5300 万件以上になった。データの検索は正規表現などを利用し、CSV、SQL、JSON などの複数のフォーマットで出力することができる。EXCEL、GIS 系などのソフトで処理することにより、さらに自然言語処理が可能になった。地理情報は QGIS で処理することもできるようになった。
また、大学の学部生および院生に紹介して、授業でスペイン語のバリエーションの語彙・文法だけではなく、文化的な意味などで複数のアプローチを試みた。また、国際学会でスペイン語のバリエーションを語学教育でどういうふうに利用すればいいかについて論じた。
なお、バルセロナ自治大学の研究者とベルリン科学アカデミーのコーパス言語学の研究者との共同研究を進め成果をあげた。
今後は定年退職になり、所属している大学で名誉教授でも研究費を申請することが不可能なので、できるだけ他の研究者や学生にデータベースが使い続けられようにしたいと考えている。

５．主な発表論文等

〔雑誌論文〕（計 5 件）

① Ruiz Tinoco, Antonio.; Barbaresi, Adrien
“Using Elasticsearch for Linguistic Analysis of Tweets in Time and Space”,11th edition of the Proceedings of LREC2018, ELRA, pags. 14-19.2018
http://lrec-conf.org/workshops/lrec2018/W17/pdf/16_W17.pdf
② Ruiz Tinoco, Antonio.
“Variation of the Second Person Singular of the Simple Past Tense in Twitter: Hiciste vs. Hicistes. Special Issue VII, pp. 145-167, Dialectology, 2017.
③ Ruiz Tinoco, Antonio.; Perea, M.P.
“Análisis del uso y distribución de formas léxicas dialectales del catalán en Twitter”, Revista Internacional de Lingüística Iberoamericana, Vol. XIV-28, pp. 49-63. 2016.
④ Ruiz Tinoco, Antonio.
“Geocorpus del español de las redes sociales y cartografía automática”, IX Congreso

Internacional de la Asociación Asiática de Hispanistas, pp. 1-14. 2016
⑤ Ruiz Tinoco, Antonio
“Análisis de más con adverbios negativos en un corpus de Twitter”. Lingüística Española (LEA), Vol. 37-2, pp.201-214. 2015

〔学会発表〕（計 10 件）
① Ruiz Tinoco, A.; Barbaresi, Adrien
“Using Elasticsearch for Linguistic Analysis of Tweets in Time and Space”,11th edition of the LREC、ELRA、Miyazaki, Japan, 2018.
② Ruiz Tinoco, Antonio
“Twitter como recurso para la enseñanza del léxico variable del español 2018”, V Jornadas ELE en Bangkok. 2018.
③ Ruiz Tinoco, Antonio
“Variación léxica del español en las redes sociales”, XXI Congreso de la Asociación Alemana de Hispanistas、Ludwig-Maximilians-Universitat Munchen. Munich. 2017
④ Ruiz Tinoco, Antonio
“Distribución de la variación de la segunda persona singular del pretérito de indicativo en Twitter: hiciste vs hicistes”, 2017 年度日本イスパニヤ学会、第 63 回大会,2017
⑤ Ruiz Tinoco, Antonio
“Análisis de tuits en español con Elastic Stack - variación geográfica y temporal”, 日本イスパニヤ学会第 62 回大会,発表場所神戸市外国語大学（兵庫県神戸市西区）.2016.
⑥ Ruiz Tinoco, Antonio
“Using Geotagged Tweets in Spanish Lingusitic Variation”, 8 Congreso Internacional de Lingüística de Corpus, Universidad de Málaga. 2016
⑦ Ruiz Tinoco, Antonio
“Geocorpus del español de las redes sociales y cartografía automática”, IX Congreso internacional Asociación Asiática de Hispanistas, Universidad de Chulalongkorn, Bangkok. 2016.
⑧ Ruiz Tinoco, Antonio
“El estudio de la variación lingüística del español -Las redes sociales”, Instituto Cervantes, Tokyo. 2015
⑨ Ruiz Tinoco, Antonio
“Cartografía digital en el proyecto VARILEX”, ALFALito, Instituto Cervantes, Tokyo, 2015.
⑩ Ruiz Tinoco, Antonio
“Tratamiento de datos geocodificados - Uso y distribución del español en Twitter”, II Congreso Internacional sobre el español y la cultura hispánica en Japón, Instituto Cervantes, Tokyo. 2015.

〔図書〕（計２件）
① Ruiz Tinoco, A.; Ueda, H; Álvarez, M; González, E; Julia, C; García Mouton, P; Fajardo, A; Zimmermann, K; Perea, M.P.; Aliaga, J.L.; González, J.A.; Carriscondo, F.M.; Corbella, D; Huisa, J.C.; Viejo, X; Le Men, J; Almeida, I; Negri, A; Colón, G; Garriga, C; Werner, R; Sánchez, M.D.; Saramago, J; de Almeida, C.“Léxico dialectal y lexicografía en la Iberromania”, Iberoamerica Vervuert, 2017.

② Nadine Rentel, Tilman Schroder, Ramona Schropf (eds.) Ruiz Tinoco, Antonio
“Kommunikative Handlungsmuster im Wandel? - ¿Convenciones comunicativas en proceso de transformación? 総ページ数 288, Peter Lang Edition. 2015.

〔産業財産権〕
○出願状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
出願年：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年：
国内外の別：

〔その他〕
ホームページ等: http://variaciones.org/

６．研究組織

(1)研究分担者
研究分担者氏名：

ローマ字氏名：

所属研究機関名：

部局名：

職名：
研究者番号（8 桁）：

(2)研究協力者
研究協力者氏名：マリア・ピラル・ペレア
ローマ字氏名：Maria Pilar Perea

(3)研究協力者
研究協力者氏名：アドリアン・バルバレシ
ローマ字氏名：Adrien Barbaresi

(４)研究協力者
研究協力者氏名：宮本正美
ローマ字氏名：Miyamoto Masami