様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 ２３ 日現在

機関番号：３２６２１
研究種目：基盤研究(C)
研究期間：2011～2014
課題番号：２３５２０５２３

研究課題名（和文）スペイン語圏のオンラインコーパスの設計と構築

研究課題名（英文）Designing and building an online corpus of Spanish-speaking areas

研究代表者
Ｒ・ＴＩＮＯＣＯ　Ａｎｔｏｎｉｏ（Ruiz 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）　　3,900,000円

研究成果の概要（和文）：本研究の目的は全スペイン語圏のインターネット上のテキスト・データをデータベース化し、スペイン語の変異を研究するためにコーパスを構築することである。研究期間の前半は新聞のRSSを試したが、後半はTwitterのデータが大変有効な手段だと判明し、TwitterのAPIを利用し、データを収集し、いくつかの語彙・文法のバリエーションの項目を研究することができた。それを内外の学会で発表した。主な特徴としては全スペイン語圏のデータが含まれているという点と、データが発信した場所の経度・緯度も取れたので、GIS（地理情報システム）の技術を利用し、正確な地図を作成する基本的な方法論が確保できた。

研究成果の概要（英文）：The purpose of this research is to make a database of Spanish text data taken from all the Spanish-speaking areas and build a corpus to study lexical and syntax variation. During the first half of the study period,data from RSS newspaper was evaluated. During the second half, we found that Twitter’s data is a very effective means for our purposes.
Data was collected using Twitter’s API. Some basic lexical and syntax variation issues were studied and the results were presented at national and international congresses. The main point of this study is that it includes data from all Spanish-speaking areas, as well as the longitude and latitude of the place of origin. Using GIS (geographic information system) techniques, we also developed a basic methodology to create dialect maps.

研究分野：スペイン語学

キーワード：スペイン語学　変異言語学　コーパス言語学　データベース　方言学

１．研究開始当初の背景

(1)研究開始当初はスペイン語コーパスがインターネット上で検索可能な大きなコーパスは少なかった。規模の大きいコーパスは主にスペイン王立アカデミー（RAE）の CREA（現在スペイン語参考コーパス、書き言葉と話し言葉）と CORDE(スペイン語史)、そして米国の Brigham Young 大学の Mark Davies 氏が作成した Corpus del Español であった。CREA は 90%が書き言葉で、10%程度は話し言葉で構成されており、ラテンアメリカのデータは残念ながら半分以下であった。本研究ではスペイン語の実態をより正確に把握するためにラテンアメリカの割合は 80%を超える予定であった。Brigham Young 大学（BYU）の Mark Davies 氏が 12 世紀から現在までの約１億語のコーパス（Corpus del Español）を 2002 年に発表した。当時も現在も地域別に細かく検索できるコーパスがない。スペイン語の変異の共時的な研究をするには多くのデータが必要である。このために、本研究では、全てのスペイン語圏の国のデータを集め、スペイン本土のデータに対してラテンアメリカのデータが全体の 80%をしめる予定とした。

２．研究の目的

(1)WEB データの抽出の技術（ウェブデータマイニング）を使い、全スペイン語圏の代表的な新聞などのデータをダウンロードし、自動的に整理し、データベースに納める。対象国はスペイン語が公用言語になっている２１ヵ国、および米国のヒスパニック系の新聞から最低 2009 年分の記事を使用する。人口の多い国（メキシコ、スペイン、コロンビア、アルゼンチン、ペルー、ベネスエラ、米国）は２～５紙分をダウンロードする。新聞により記事の数が異なるが、およそ 30 万件の記事で、数千万語になると予想する。

(2)サーバーの必要な環境（OS、データベース、スクリプト言語など）を整え、データベースを完成し、コーパス検索用のインターフェースを作成する。研究期間中に国内と海外の一部のスペイン語学者と大学院生に実験的に使用可能にした後、研究期間終了直後は上智大学の専用サーバーで一般公開し、国内および海外の学会にも公開する。長期的にはデータの種類と量を増やし、インターフェースを改善し、また異なった目的の新しいインターフェースを作成し、スペイン語コーパスを利用しながらスペイン語学の研究を続ける。

(3) コーパス言語学の一つの特徴は、基本的にはあらゆる言語現象を観察するための技術である。ゆえに、研究課題により多くのデータから必要な言語学的な証拠(linguistic evidence)をできるだけ効率よく探したい。言語研究の目的は形態論、統語論、意味論、語用論、教授法・教育など多くあるので、コーパスには強力な検索機能が必要である。また、スペイン語のように多くの国で話されている言語の場合はそれぞれの地域の十分なデータも必要である。本研究のコーパスは標準的なデータベースのタイプ（SQL）が広く使われている環境（LAMPP）で使うので、構造の拡大、改善、新しいインターフェースの追加、データの更新などは特殊なシステムと比べてメンテナンスは容易に管理でき、使い方も短期間で学習できる。

３．研究の方法

(1)全スペイン語圏のデータを収集するために、オンライン上の代表的な新聞をおよそ 30 紙を選択する。各国から最低１紙で、人口・影響力を考慮し、メキシコは 5 紙、スペインは 4 紙、コロンビアは 2 紙、アルゼンチンは 2 紙、ペルーは 2 紙、ベネスエラは 2 紙、米国のヒスパニック系新聞 2 紙。その記事（データ）を 2 回に分けて自動的に収集する。第 1 回はおよそ全体の 20%のデータでシステムの評価と確認をし、第 2 回は残りの 80%を収集する。コーパスを利用する研究者は研究発表をする際、インターネットで得たデータの新聞名、ニュースの見出し、発刊日、ダウンロードした日、そのデータの URL も必要なので、これらの詳細もダウンロードし前処理した後、データベースに投入する。データの前処理は、ニュースを国内、社会、政治、経済、スポーツ、文化などで整理し、分類する。

(2)データを収集するにつれ、語彙・文法の地理的な変異を分析するにはソシアルメディアの Twitter のように API があり、データの発生時も分かり、なおかつ発信地の経度・緯度も分かるので、データの情報源を Twitter に変えてみた。データベース（MySQL）の構造を設計しなおす必要があったが、地理的な情報が分かるので、言語地図を作成することが可能になった。

４．研究成果

(1) 本研究では、ツイッターから抽出されたコーパスデータを分析した。ジオコーディング機能を利用した。任意の地理的位置の周辺を検索制限することにより、マイクロブログの自動抽出が可能になる。スペインとラテンアメリカの都市の 25 キロ圏内を収集し、さらなる分析のために、データベースに格納した。まずは、例として es una lástima que + IND/ SUB などの構造を持つスペインの直説法/接続法のバリエーションを分析した。さらに、delante de mí vs. delante mío の問題を分析した。

(2) 都市は以下のとおりである。スペインのアルカラ･デ･エナレス、バルセロナ、ウエルバ、ラス･パルマス、マドリード、オビエド、パンプローナ、セビリアとサンタ･クルス･デ･テネリフェ。また、ヒスパニックの多いニ

ューヨークのほかにラテンアメリカのアスンシオン、ボゴタ、ブエノスアイレス、カラカス、ラパス、リマ、メキシコシティ、モンテレー、モンテビデオ、キト、サン･ホセ･デ･コスタリカ、サンティアゴ・デ・チリ。２ヶ月の間に 800 万以上のメッセージ、約 120 万語を収集し、MySQL データベースに格納した。前処理として、意味不明の tweets、RT（リツイート）、引用、スパムなどを削除し、各都市の 100 例をランダムに選択し、分析した。

図１．データベースのインターフェース

分析した構造：

(a)Me alegro (de) que… + ind/sub
(b)Me gusta que… + ind/sub
(c)Es necesario que… + ind/sub
(d)Es una lástima que… + ind/sub
(e)Es una pena que… + ind/sub

このような構造の直説法と接続法の使用の割合は地域によりだいぶ異なると判明したが、Es necesario que…+ind/sub のように、ほとんど直説法が使われることもあれば、Es una lástima que…+ind/sub、または Es una pena que…+ind/sub は図３で示すように大きく標準語から離れていた。このような現象を観察すると、地理的にだけではなく、具体的な構造により、直説法の使用率が異なり、モンテビデオのように 27%の例だけ標準語に従い、直説法を使う。

また、スペインよりもラテンアメリカ全体的に接続法の使用率が減りつつあることが判明した。

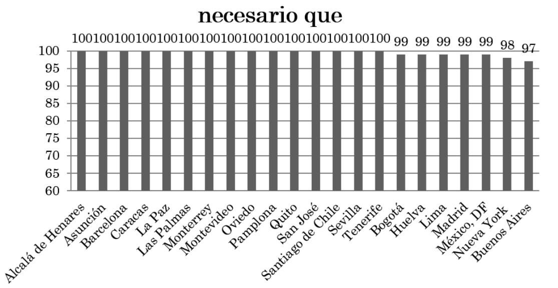

図２．Es necesario que…+ind/sub

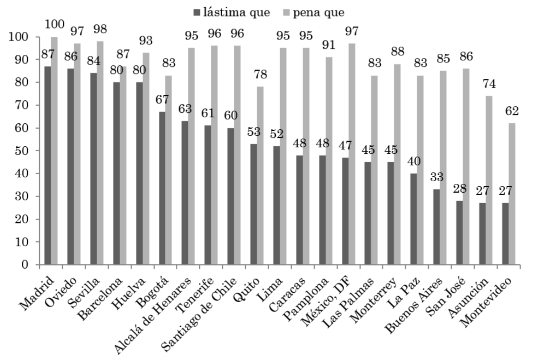

図３． lástima que vs. pena que

(3) delante de mí vs. delante mío の問題について、接続法の例と同様にデータをアメリカ合衆国とメキシコの国境の両側の例を集め、delante、debajo、detrás、cerca などの例を分析した結果、図４のようになった。

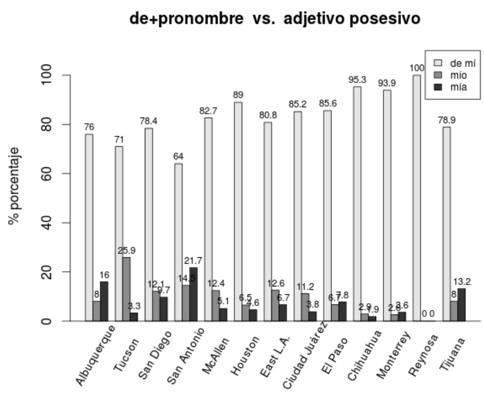

図４．Delante de mí vs. delante mío(a)

国境の両側の傾向は異なるだけでなく、形容詞の男性系と女性系の傾向の使用の差も見られた。

(4)このような方法論を使い、そのほかに次のような問題を分析した：

(a) más nada vs. nada más のように nada、nadie、nunca、ninguno を分析した。
(b)auto，coche，automóvil の使用分布。

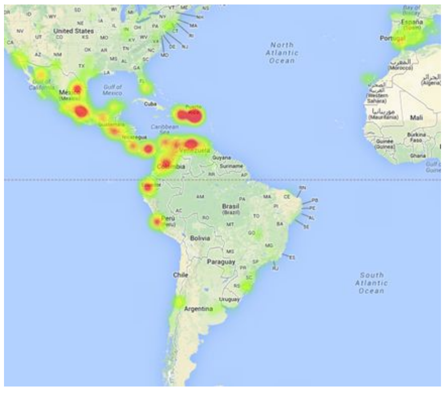

図５．carro の使用分布

(c)Queísmo vs. dequeísmo
(d)その他の語彙と構造。

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 RUIZ TINOCO, Antonio
“Variación léxica y gramática del español peninsular e hispanoamericano”, The Korean Journal of Hispanic Studies, 査読有、Vol. 3 pags. 29-53. 2013. Institute of Hispanic Studies, Korea University, Seoul, 韓国.

② RUIZ TINOCO, Antonio
“Twitter como corpus para estudios de geolingüística del español”, en Sophia Lingüística LX, 査読有、pp 147-163, The Graduate School of Languages and Linguistics, Linguistic Institute for International Communication, Sophia University,2013, Tokyo.日本.

③ RUIZ TINOCO, Antonio
“La variación del español en las redes sociales”, ALFALito en Japón, Actas del Congreso Internacional sobre el español y la cultura Hispánica. 査読有、Instituto Cervantes de Tokio (2013).日本.
http://cvc.cervantes.es/ensenanza/biblioteca_ele/publicaciones_centros/tokio_2013.htm

〔学会発表〕（計 10 件）

① RUIZ TINOCO, Antonio
“Queísmo y dequeísmo en Twitter ―Uso y distribución geográfica ―”, 7th International Conference on Corpus Linguistic, CILC2015, Universidad de Valladolid, 2015 年 3 月 6 日,スペイン。

② RUIZ TINOCO, Antonio
“ Queísmo y dequeísmo en locuciones preposicionales variables ―Uso y distribución en las redes sociales―”, LX Congreso de la Asociación Japonesa de Hispanistas, Universidad de Osaka.2014 年 10 月 11 日.日本。

③ RUIZ TINOCO, Antonio, Maria-Pilar Perea
“How two Catalan dialects coexist in the same areas: - an analysis of the use and distribution of some lexical forms in Twitter”, Methods in Dialectology XV, University of Groningen, The Netherlands、2014 年 8 月 15 日,オランダ

④ RUIZ TINOCO, Antonio
“ Subjective communication on the Internet”, Fourth International Symposium on European Languages in East Asia The Role of Art, Music and Literature in European Studies – A Critical Discourse in Cross Cultural Communication, National University of Taiwan. 2013年11月16日. 台湾。

⑤ RUIZ TINOCO, Antonio
“La variación del español en las redes sociales”, ALFALito en Japón, Instituto Cervantes, Tokio. 2013 年 10 月 3 日.日本。

⑥ RUIZ TINOCO, Antonio
“Variación de la anteposición de más con adverbios de negación”, V International Conference on Corpus Linguistics, Universidad de Alicante.2013 年 3 月 15 日. スペイン。

⑦ RUIZ TINOCO, Antonio
“Variación sintáctica en Twitter en el español fronterizo de Estados Unidos”, Language Contact, Conflict, and Confluence at the Edge of the Nation, 24th Conference on Spanish in the United States and 9th Conference on Spanish in Contact with Other Languages, The University of Texas Pan American.2013 年 3 月 8 日.米国。

⑧ RUIZ TINOCO, Antonio
“ Alternancia indicativo/subjuntivo en Twitter en el español de Estados Unidos”, II ALFALito, Cuestiones lingüísticas en relación con la diáspora latinoamericana”, The Graduate Center, City University of New York.2012 年 9 月 28 日.米国。

⑨ RUIZ TINOCO, Antonio
“ Twitter como corpus de variación geográfica -alternancia modal del español-“, IV Congreso Internacional de Lingüística de Corpus, Asociación Española de Lingüística de Corpus, Universidad de Jaén. 2012 年 3 月 22 日.スペイン。

⑩ RUIZ TINOCO, Antonio
“Variación geográfica del uso del modo subjuntivo”, LVII Congreso de la Asociación Japonesa de Hispanistas, Universidad Komazawa.2011 年 10 月 8 日.日本。

〔その他〕
ホームページ等
http://variaciones.org

６．研究組織

(1)研究代表者

ルイズティノコ　アントニオ
(RUIZ TINOCO, Antonio)
上智大学・外国語学部・教授
研究者番号：

(2)連携研究者

上田博人（UEDA, Hiroto）
東京大学大学院・総合文化研究科・教授
研究者番号：　20114796

高垣敏博　（TAKAGAKI, Toshihiro）
東京外国語大学・外国語学部・教授
研究者番号：　00140070

宮本正美　(MIYAMOTO, Masami)
神戸市外国語大学・外国語学部・教授
研究者番号：　20131477