

令和 元年 6 月 17 日現在

機関番号：34418

研究種目：研究活動スタート支援

研究期間：2017～2018

課題番号：17H07270

研究課題名(和文) Design and collection of a written corpus of learner Spanish in Japan

研究課題名(英文) Design and collection of a written corpus of learner Spanish in Japan

研究代表者

VALVERDE Pilar (VALVERDE, Pilar)

関西外国語大学・外国語学部・講師

研究者番号：10588205

交付決定額(研究期間全体)：(直接経費) 2,000,000円

研究成果の概要(和文)：本研究では、外国語としてのスペイン語研究のためのCELENという新しいツールを紹介する。これは大学でスペイン語を学ぶ日本語話者によって書かれた文を集めたコーパスであり、今回紹介するバージョン1は、二つの異なる教育現場から得られた文で構成されている。一校目はスペイン語が専攻語として学ばれている関西外国語大学であり、学習者数は459名、収集した文の数は1840、収集した単語数は140,000に上る。もう一校は、スペイン語が第二外国語として学ばれている京都大学であり、学習者数は278名、収集した文の数は2111、収集した単語数は144,000である。

研究成果の学術的意義や社会的意義

本コーパスは、スペイン語を学ぶ日本語話者を対象とした初のコーパスである。このコーパスは次の二点で重要な役割を果たしている。一点目は、本コーパスを用いることで教育現場においてスペイン語教師が調べたい構文、語彙リスト、コロケーションなどに関するデータを、学生らが実際に在学中に書いた文から自動的に採取することが可能となり、さらに教師自身の教育方針を向上させるための結論を導き出したり、対策を探ることが可能になるという点である。二点目は、このコーパスは外国語としてのスペイン語習得研究に役立つという点である。

研究成果の概要(英文)：We present a new tool for research on Spanish as a foreign language: CELEN, a corpus of texts written by native speakers of Japanese who study Spanish at the university. In version 1, the corpus is made up of texts from two institutions in Japan: Kansai Gaidai University (459 learners, 1840 texts, 140,000 words), where Spanish is studied as a major, and Kyoto University (278 learners, 2111 texts, 144,000 words), where it is studied as a second foreign language. Each text is associated with a large amount of information about the learner who wrote it and on the conditions in which the text was written, so that it can be used by researchers with different needs. It is possible to consult the corpus on-line or to download it under a CC BY-NC 4.0 license.

研究分野：コーパス言語学

キーワード：スペイン語 コーパス 学習者コーパス 言語習得

様式 C - 19、F - 19 - 1、Z - 19、CK - 19 (共通)

1. 研究開始当初の背景

言語学分野において、コーパスと呼ばれる多数の文の集合体は、言語学習のみならず、辞書や文法書などの作成にも欠かせないものであると言える。外国語教育分野では、学習者による文を集めたコーパスは、辞書や文法書や教科書などの教材開発や、外国語学習過程についての研究にも必要とされている。

しかし、スペイン語学習者のコーパス開発はまだ初期段階であり、文コーパスに限定すると、現時点で誰でも使用可能な文コーパスは、CEDEL2 と CAES の二つのみ存在する。CEDEL2 は第二言語としてのスペイン語の文コーパスであり、英語が母語である学習者から収集した 800,000 単語が収められている。また CAES は外国語としてのスペイン語を学ぶ学習者によるコーパスで、アラビア語、標準中国語、フランス語、英語、ポルトガル語、ロシア語の六言語を母語とする学習者から収集した 575,000 単語が収められている。

日本における外国語としてのスペイン語の現状に関して、セルバンテス協会の最新の報告によると、日本全国には約 60,000 名のスペイン語を学ぶ学生がいると伝えている。本研究のコーパスには、大学におけるスペイン語教育に関するデータが収集されており、また一般的に大学でのスペイン語は主に第二外国語科目としてのスペイン語と専攻科目としてのスペイン語の二種に分けられる。日本にある 780 校の大学の中で、200 校以上でスペイン語が選択語として学ばれている一方で、たった 15 校でのみスペイン語が専攻語として学ばれている。つまり大多数の学生はスペイン語を第二外国語として学び、約 4500 名だけが専門的にスペイン語を学んでいると考えられる。

2. 研究の目的

本研究では、日本語が母語である大学生が書いたスペイン語のコーパスを作成していくことと、また複数の大学で、第二外国語としてのスペイン語クラスと、専攻語としてのスペイン語クラスの双方から文を収集することを目的としている。特に、異なるレベルのスペイン語という点を重視し、ヨーロッパ言語共通参照枠の A1 レベル(入門)から B2 レベル(中上級)までと言った複数のレベルから文を収集することを目指している。

本コーパスの二つの課題は、まず今回の文収集に協力してもらった大学で教鞭をとる外国語教師たちに検索ツールを提供することであり、次に日本語話者のスペイン語習得に関する研究を向上させることである。

3. 研究の方法

本コーパスに収録されている文は、大学での一学年を通じて集められたものに限られている。またポートフォリオを用いて、学習者たちの学習の歩みを一年間に渡り記録しているため、ある一定の間継続したコーパスであると言えるだろう。

本コーパスのデータはスペイン語の学習量も生徒数も異なる二校の大学から収集されている。

1) 一校目の関西外国語大学は私立の外国語大学で、スペイン語を専攻する学科があり、約 30 組から 508 の回答を得ることができた。要件を満たさない者、また日本語が母語ではない学生を除くと、有効な参加者数は 459 名であった。ここから得られた情報は、年齢、性別、学年、学生の母語、学生の両親の母語、家庭で用いる言語、スペイン語圏の人々との交流の有無、スペイン語圏での滞在経験の有無、スペイン語を習い始めた年齢、スペイン語または英語の資格の有無、その他の言語の知識である。一方、文に関しては、ネイティブ教師が担当する一年生から四年生までの授業から文が集められている。収録されている文のスペイン語レベルは、その学生が文を書いた時点で受けていた授業のレベルと一致している。収集された文のうち 65% は A1 レベル、25% は A2 レベル、10% は B1 レベルに相当する。また文のソースに関して、77% が小テストから、17% が課題から、4% が授業アクティビティから、2% が期末テストからであり、その内の 78% は教科書、辞書、インターネットなどの補助を用いずに書かれている。さらに、文の大多数は学習者によって手書きされているため、OCR ソフトを用いて、あるいは手動でデジタル化し、さらに内容の確認と匿名化をした後、各文に関連するメタデータと共にデータベースに入力した。各文のメタデータとして収録されているのは、日付、該当する授業のレベル、学年、学期、その文を書くまでに受けた授業の期間数、描写、物語や発表といった文のタイプ、ジャンル、題名、宛名、文の現実性、社会的特徴、教科書と課、割り当てられたテーマ、状況、時間制限、辞書や教科書などの使用の有無、指定された単語数、提出手段(手書きか電子媒体か)である。

2) 二校目の京都大学では第二外国語としてスペイン語が学ばれており、一学年を通じて学生が書いた課題を基にデータを収集し、さらに学生の年齢や言語知識についての情報も収集した。集められた文はすべて課題を基にしているため、教科書や辞書などを用いて書かれており、電子媒体を通して提出された。従って、文のデジタル化の必要はなかったものの、

個人情報の保護と、該当するメタデータの割り当てを目的としたデータ処理を行った。

学習者と文に該当するメタデータを各文に割り当てた後、本コーパスをスケッチエンジン (Sketch Engine) というプラットフォームに導入した。そして検索インターフェイスが個に応じた設定され、検索画面でユーザーに表示される検索オプションとその検索結果、またそれらに用いたタグが表示されるようになっている。

4. 研究成果

表1に見られる通り、最終的に約700名の学習者から約4000の文が集められ、合計約284,000語の単語が本コーパスに収められている。

	学習者数	集められた文の数	およその単語数
関西外国語大学 (A1-B2)	459	1840	140,000
京都大学 (A1)	278	2111	144,000
合計	737	3951	284,000

表1. 学習者と、集められた文と単語の数

本コーパスは、多くの言語の専門家が調べ物やコーパス作成で用いているプラットフォームのスケッチエンジン (<https://www.sketchengine.eu>) で調べることが可能である。またこのスケッチエンジンを用いると、コンコーダンス (Concordance)、語彙リスト (Wordlist)、コロケーション、語の組み合わせ (Word Sketch)、同義語や同類語 (Thesaurus)、二語のコロケーション間の比較 (Word Sketch Difference)、N-gram と言った様々な結果を導き出すことも可能である。図1にあるように、スペイン語学習者の間で頻発する誤りである動詞 *ser* と副詞 *bien* の組み合わせを検索した結果、以下のようなコンコーダンスの例が表示される。

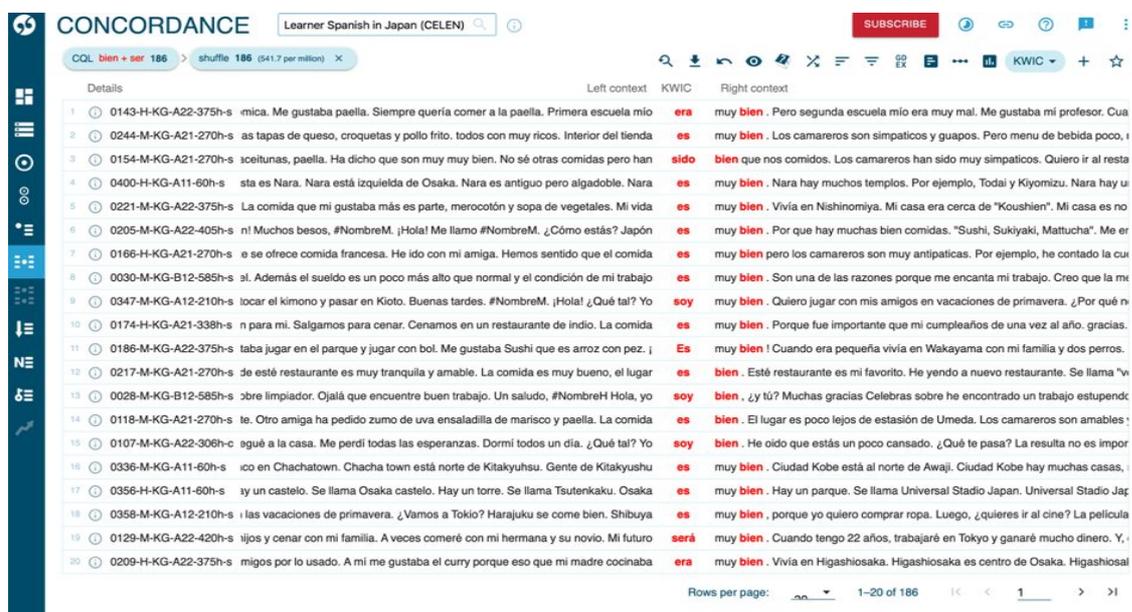


図1. コーパスに収録された “ser + bien” のコロケーションの例

また研究者は全コーパスをダウンロードすることができ、さらに CC BY-NC 4.0 というライセンスがあれば、必要に応じてコーパスの操作が可能である。

表示される文は、調べたい単語や文法カテゴリー (POS tagging) に応じて自動的にタグ付けされるが、通常の用法から逸脱している文の場合、マーキングの正確さは通常の用法が用いられた文よりやや低くなることも考えられる。従って、今後の課題の一つに自動マーキングの修正が挙げられるだろう。またこれにより、将来はコーパスに収める文の範囲を他の教育機関へと広げ、外国語学習のためにブログやソーシャルメディアといった日常的な文脈に由来する文を含めていくことも考慮に入れていこうと考えている。

謝辞

本研究を進めるに当たり、関西外国語大学と京都大学のスペイン語教師の方々の多大な協力を賜りました。厚く感謝を申し上げます。また快くデータを提供してくれた学生の皆様にも感謝の意を表します。

5 . 主な発表論文等

〔雑誌論文〕(計 3 件)

Valverde, P. Diseño y creación de un corpus de aprendices de ELE en Japón (CELEN). E-AESLA Revista digital, 査読有, vol. 6, 2020 (to appear).

Valverde, P. Design Principles and Data Collection for CELEN: A Corpus of Learner Spanish in Japan. In Tono, Y & Isahara, H. (Eds.), Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC2018), 査読有, 2019, pp. 485-491.

Valverde, P. Un corpus de blogs de aprendices japoneses de español. In María Bargalló Escrivá, M. Esther Forgas Berdet Antoni Nomdedeu Rull (Eds.), Léxico y cultura en LE/L2: corpus y diccionarios, 査読有, 2018, pp. 845-858.

〔学会発表〕(計 4 件)

Valverde, P. Diseño y creación de un corpus de aprendices de ELE en Japón (CELEN). 37 International Conference of the Spanish Society for Applied Linguistics (AESLA). March 27-29, 2019. University of Valladolid, Spain.

Valverde, P. Primeros pasos en la construcción de un corpus de aprendices de ELE en Japón. 日本イスタニヤ学会 第 64 回大会, October 13-14, 2018. Nanzan University, Nagoya, Japan.

Valverde, P. Design Principles and Data Collection for CELEN: A Corpus of Learner Spanish in Japan. Fourth Asia Pacific Corpus Linguistics Conference (APCLC2018), September 17-19, 2018. Sunport Hall Takamatsu, Kagawa, Japan.

Valverde, P. Un corpus de blogs de aprendices japoneses de español. XXVIII Congreso Internacional de ASELE. September 6-9, 2017. Universitat Rovira i Virgili, Tarragona, Spain.

〔その他〕

ホームページ等

<https://sites.google.com/view/celen>

科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属されます。