様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年　６月２０日現在

機関番号：２３９０１
研究種目：研究活動スタート支援
研究期間：2010～2012
課題番号：２２８２００４７

研究課題名（和文）　日本人スペイン語学習者の作文に含まれる誤りの自動検出

研究課題名（英文）　Automatic Error Detection in Spanish Texts Written by Japanese Learners

研究代表者
バルベルデ　ピラール（VALVERDE PILAR）
愛知県立大学・外国語学部・講師
研究者番号：10588205

研究成果の概要（和文）：この研究は、日本人スペイン語学習者の作文に見られる文法的エラーを自動検出する分析器の開発を目的とする。そのため 10,000 語のコーパスに対してエラー情報のアノテーションを手動で行い、そして、制約文法の形式化に基づいて一致のエラーを自動的に検出するための規則を記述した。このような分析器およびタグ付きコーパスの評価により、数の一致のエラー検出については、言語の習熟度が上がるにつれて、句レベルより節レベルにおいてそれが頻繁に起こることから、性の一致のエラー検出よりも深い言語分析が必要とされることが明らかになった。

研究成果の概要（英文）：The purpose of the research is to develop an analyzer capable of automatically detect grammatical errors in Spanish texts written by Japanese learners. To do that, we have manually annotated with error information a 10,000 words corpus and have written a set of rules for automatic agreement error detection, using the Constraint Grammar formalism. With the evaluation of the analyzer and the annotated corpus we have confirmed that the detection of number agreement errors requires a deeper linguistic analysis of the texts than gender agreement errors, since number agreement errors occur with more frequency at the clause level than at the phrase level as the language proficiency level improves.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２０１０年度	900,000	270,000	1,170,000
２０１１年度	104,940	31,482	136,422
２０１２年度	795,060	238,518	1,033,578
総　計	1,800,000	540,000	2,340,000

研究分野：コーパス言語学
科研費の分科・細目：言語学・外国語教育
キーワード：スペイン語・言語習得・コーパス・誤り検出・制約文法

１．研究開始当初の背景

本研究はコンピューター支援語学学習の分野に分類されるものであるが、自然言語処理の基本原理やコーパス言語学、言語教育や言語習得とも関連し、より具体的にはそうした方法論に基づいた「学習者の作文に含まれる文法的エラーの自動検出」に焦点を当てるものである。当該分野は主に英語において展

開しつつあるが、外国語としてのスペイン語学習者が増加傾向にあり、情報ツールが重要視されている現状を考慮すれば、スペイン語においても拡張されるべきだと思われる。

研究代表者は、外国人教師として学習者の作文に含まれる誤りの分析を行ってきたが、その過程において、博士学位論文の成果（コーパスに基づくスペイン語統語的機能に関する研究）と Carlsson の制約文法の形式化をもとにした文法記述（意味論的機能を自動判別するための文法体系）とを結びつけることで、文法的エラーが（半）自動的に検出可能となる着想を得たため、本研究を申請するに至った。

２．研究の目的

本研究は日本人スペイン語学習者の作文に含まれる文法的エラーを自動検出できる分析器および学習者エラーコーパスの構築を目的とする。エラー情報を含む大量のテキストを手にしていれば、これを、第二言語習得についての調査の実施や、より良い教育目的の資料（辞書、文法書、文法チェック機能など）を作り出すことにも利用可能である。

エラー分析器の開発に際しては、以下の点を明らかにしなければならないと本研究は考える。

(1) 学習者のテキストはエラー分析の前段階として、形態・統語解析器によって処理される必要がある。そのため、ネイティブのテキスト処理に対応する既存の（主として形態・統語解析器）自動処理ツールを、どのように学習者の習得に応用できるかを調べなければならない。

(2) 日本人スペイン語学習者のエラータイプのうち、どのタイプが最も信頼できる方法で自動処理できるか。

(3) それらのエラーはどのような構造を持っているか、どのような状況において起こるか、そして、(エラー分析器の出発点である) 制約文法の規則としてどのように形式化できるか。

(4) 学習者によって書かれた実際のテキストを解析する際のエラー分析器の処理能力について、また、得られた情報からどのような言語学的あるいは教育学的結論を引き出すことができるか。

３．研究の方法

(1) 規則を記述する前段階として、まずネイティブのテキストに対応する既存のテキスト自動処理ツールを、どのようにして学習者の言語習得に適応できるかを検討した。具体的には、スペイン語に対応する二つの形態・統語解析器 FreeLing と HISPAL（それぞれの語の文法範疇と屈折に関する情報の入ったタグを付与するツール）の性能を調べた。

(2) 次に、学習者の起こすさまざまなエラーのうち、エラー分析器が検出するであろう一番目のエラーを選定した。このエラーは、作業方法を確立するために、また、今後直面する可能性のある別種のエラーに対する問題を見つけるために役立つものと考えられる。

この一番目のエラーの選出においては、次の三つの要因を考慮した。第一に、エラーをエラーと見なすかどうかのネイティブの間での合意の程度、第二に、そのエラー（あるいは学習者に対して想定される問題）の解決に要する言語学的知識の程度、第三に、このエラーが起こる頻度とそこから図れる教育学的視点から見た学習者の関心度、である。

これらの三つの要因を考慮したところ、選ばれたのは性と数の一致に関するエラーであった。これには極めて高い合意が得られており、ある程度の言語学的知識（語の範疇や性と数、語と語の間の統語関係について）を要するもので、日本人学習者の間でかなり頻繁に見られるエラーであることと符合した。

(3) 意味分析器におけるエラー自動検出器には二つの主要なアプローチがある。一つは確率に基づくものであり、もう一つは言語学的規則に基づくものである。本研究のエラー分析器は後者のタイプであり、制約文法（Carlsson, 1995）の形式化をもとにしている点が特徴である。

制約文法の規則の記述にあたっては、教師として経験上得た学習者のエラー情報だけでなく、主にコーパスから引き出した情報も利用している。実際には、一致のエラーを検出する規則を記述、評価する目的で、コーパス CORANE から抽出した 25,000 語の文章を用いて、性と数に関するエラー情報をタグ付けした。

抽出した文章はレベル A2 から B1 までの日本語を第一言語とする 47 人によって書かれた 133 のテキストで構成されている。そのテキストのうち、文法の構築に 15,000 語を、その評価には 10,000 語を利用した。

設計基準において最も重要なのは、本研究の文法が実際のエラーを検出する際、たとえわずかな数であっても、検出誤りを起こさないことである。換言すれば、エラーでないものをエラーと取り違えてはならないのである。

４．研究成果

(1) 形態・統語解析器は非ネイティブ学習者のテキストを処理にかけると、非常に高い検出精度を示す。FreeLing では、ネイティブのテキストにおける精度は 99%、学習者のテキストにおいては 92,6%である。一方 HISPAL では、ネイティブのテキストにおける精度は 99%、学習者のテキストでは 95%であった。

分析器の結果から、ネガティブに影響する学習者のエラーは正書法に関するもの（アクセント、句読点、大文字と小文字の用法を含む）であることがわかったが、一方では、一致エラーを含む語や文法的範疇の選択ミスも 80%以上の精度で正しく分析されていた。

結論として、ネイティブのテキストに対応する分析器は、かなり高い精度において学習者のテキストに適用可能であると言え、また、テキストが事前に正書法のチェックを受けていたならば、さらに高精度の検出が期待できただろう。

(2) そこで、形態・統語レベルでテキストを分析するため HISPAL を用い、さらにそれに対して一致エラーを検出する制約文法を適用した。

句レベルにおいても節レベルにおいても一致のエラーを検出する本研究の文法は、性の一致のエラーを見つける 31 の規則と数の一致のエラーを見つける 50 の規則で構成されている。これらの規則は、どのコンテクストにおいてエラーが起こり、そのエラーがどのような特質を持っているかについて、また、エラー検出時にどのタグをテキストに付与するかについて示している。例えば、表１の規則は、ある一連の条件を満たす場合、形容詞男性形に“%agr-f”（“女性形でなければならない”）というタグを指示するものである。

表１．制約文法の形式主義をもとにして
数の一致エラーを検出する規則

ADD (%agr-f) TARGET ADJ-M (*-1 NP-HEAD-F BARRIER ALL-N-RIGHT LINK NOT p PRP);
形容詞男性形にタグ“%agr-f”を付ける。 形容詞男性形の左に、前置詞を伴わない名詞か代名詞の女性形がある場合。 形容詞と名詞あるいは代名詞との間に置かれるのは、形容詞か分詞、あるいは関係副詞でない副詞のみである。

(4) 我々の文法の性能をコーパス CORANE の 10,000 語からなる文章（B2 レベルの 5,000 語と C1 レベルの 5,000 語)を用いて評価した。結果は次の通りである。性の一致のための規則の精度は 64.52%､検出率は 71.43% であり、数の一致に関する精度は 58.62%、検出率は 31.48%であった。

一致エラーの情報が付与された学習者エラーコーパスから、我々の文法の影響下にあると同時に教育に関りのある学習者が起こすエラータイプについて、一応の結論を引き出すこともできる。それは、スペイン語教師と学生が一致の問題に直面した際、最も習得しがたい問題として、性の一致の方に焦点を置く傾向がある、ということである。なぜなら、学生にとって習得上労力がかかるのは、コンテクストに従って正しい数を選ぶことより、それぞれの名詞の固有の性がどちらであるかを判別する方だからである（形態的なヒントがあったとしても、性は恣意的であるので暗記しなければならない)。

コーパスデータの元テキストを作文した日本人学習者の性の一致に関するエラーは、スペイン語が上達するにつれて減少する。一方、数の一致に関するエラーは上級者においてすらかなり頻繁に起こる。その上、学習者は句レベルでも節レベルでもそうしたエラーを引き起こすことがあるため、本研究が構築したエラーコーパスは、学習者の習得レベルが上がるにつれて、深い言語分析が必要な節レベルでのエラーが多くなることを示しており（一致する語と語がかなり離れているか、もしくは主語が目的語と混同されている)、それゆえ、この種のエラーの自動検出は最も困難なものであるということが言えるであろう。

今後の研究方針として、上記と同じ方法を、特に日本人学習者がよく起こす冠詞あるいは前置詞（本研究と比較すべき英語向けシステムが存在する）のような、極めて高度な別種のエラーを検出する分析器にも適用することを考えている。

本研究が開発した分析器は、エラー情報を含む学習者のコーパスの自動記述やコンピューター支援語学学習の情報化システムへの導入に向けて、さまざまな方法で利用できることが見込まれる。

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 Valverde, M. P. and A. Otani. Automatic Detection of Gender and Number Agreement Errors in Spanish Texts Written by Japanese Learners. Proceedings of the 26th Pacific Asia

Conference on Language, Information and Computation (PACLIC 26). 査読有. 2011. 299-307.

② Ohtani A. and M.P. Valverde. Nominative-marked phrases in Japanese Tough Constructions. Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26). 査読有. 2011. 272-279.

③ Valverde, M. P. An Evaluation of Part of Speech Tagging on Written Second Language Spanish. In Gelbukh, Alexander (ed.), Computational Linguistics and Intelligent Text Processing. 12th International Conference CICLing 2011, Proceedings, Part I. Springer Verlag Lecture Notes in Computer Science. 査読有. 6609. 2011. 214-226.

〔学会発表〕(計３件)

① Valverde, M. P. and A. Otani. Automatic Detection of Gender and Number Agreement Errors in Spanish Texts Written by Japanese Learners. The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26). November 8, 2012. Pullman Bali Legian Nirwana, Indonesia.

② Ohtani A. and M.P. Valverde. Nominative-marked phrases in Japanese Tough Constructions. The 26th Pacific Asia Conference on Language, Information and Computation (PACLIC 26). November 8, 2012. Pullman Bali Legian Nirwana, Indonesia.

③ Valverde, M.P. An Evaluation of Part of Speech Tagging on Written Second Language Spanish. The 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2011), February 21, 2011. Waseda University, Tokyo, Japan.

〔図書〕(計１件)

① Valverde, M. P. and E. Bick. A Web Corpus of Spanish Automatically Annotated with Semantic Roles. In A. Sánchez and M. Almela (eds.). *A Mosaic of Corpus Linguistics. Selected Approaches.* Chapter 6. Berlin/ Frankfurt: Peter Lang. 2010. 249-268.

６．研究組織

(1)研究代表者
バルベルデ ピラール (Valverde Pilar
愛知県立大学・外国語学部・講師
研究者番号：10588205

(2)研究分担者

(3)連携研究者