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　24 年　5 月　31 日現在

機関番号：１２６０８
研究種目：基盤研究（B）
研究期間：　2009　～　2011
課題番号：21320092
研究課題名（和文）　大規模コーパスを利用した日本語学習支援システム『ひのき』構築と評価
研究課題名（英文）　Construction and Evaluation of Japanese Learning Support System HINOKI Using Large Scale Japanese Corpus
研究代表者　仁科　喜久子　　（NISHINA KIKUKO）
東京工業大学・留学生センター・教授
研究者番号：40198479

研究成果の概要（和文）：
日本語教育における e-Learning 分野の研究として、言語学・自然言語処理・教育工学など学際的な背景をもつ研究者が大規模コーパスを利用した日本語統合学習支援システム「ひのき」を構築することを目指した。既に公開している読解支援システム「あすなろ」、作文支援システム「なつめ」（http://hinoki.ryu.titech.ac.jp/）という二つのシステムを統合し、学習者コーパス開発、評価実験によって、システムの複眼的な評価を行い、新たな分析方法を確立し、教授方法への示唆を行った。また、学習者コーパスの構築の必要性を認識し、学習者コーパスおよび自動添削システム構築の完成を今後の課題とした。

研究成果の概要（英文）：
The project aimed to combine a large-scale Japanese language corpus with research in e-Learning systems to construct the comprehensive Japanese language learning system HINOKI through joint collaboration between linguists, Natural Language Processing researchers, and educational engineers. We developed a comprehensive and effective learning system for reading and writing Japanese by synthesizes the Asunaro reading support system and Natsume writing support system, which we had developed previously. We evaluated the system on Japanese learners and found it to be efficient .In conclusion, we proposed not only new learning methods applicable to e-Learning, but also made clear the necessity of further development of learner corpora and writing correction system.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	4,600,000	1,380,000	5,980,000
2010 年度	3,700,000	1,110,000	4,810,000
2011 年度	4,200,000	1,260,000	5,460,000
年度			
年度			
総　計	12,500,000	3,750,000	16,250,000

研究分野：人文学
科研費の分科・細目：言語学・日本語教育
細目表キーワード：教育工学、教材、教育メディア
細目表以外のキーワード：e-Learning

１．研究開始当初の背景

言語教育において Data Driven Language Learning の関心が高まってきていた。また言語処理研究分野では、辞書・検索・形態素解析・構文解析技術の進展に伴い、言語教育への応用が検討される機運が高まっていた。教育工学においては CALL としての日本語教育における教示・評価の方法の開発に関心がも

たれていた。
　また、海外の第 2 言語習得研究会議においても、「日本語教育」誌、自然言語処理特集号「コーパス言語学・言語教育と言語処理」(2004)、「教育・学習を支援する言語処理」(2008)において編集委員を務めたことから、言語学・自然言語技術専門家の日本語教育におけるコーパス研究とへのアプローチが急速に高まっていることを確認した。
　その一方では、学習支援システムを構築する上で必要な日本語学習者コーパスの収集が遅れ、誤用タグに関する研究も希薄であることがわかった。

２．研究の目的
(1)国内外の留学予定者・留学生を主たる対象として、「大規模日本語コーパス」などのコーパスを用い、日本語総合能力を養成するシステムとして開発を進める。すでに公開している作文支援システム「なつめ」、読解支援システム「あすなろ」に大規模コーパスを利用し、さらに学習者コーパスを収集、誤用検索システムを構築する。
(2) 学際的な研究組織によるシステムの複眼的な評価を行い、新たな学習者への支援および教授方法を見出すことを目指す。
(3)海外調査の成果も踏まえて、これらの海外機関においてシステムの評価実験を行い、遠隔学習の試行を最終的な目標とする。

３．研究の方法
(1) 既存コーパスに加え、理系論文コーパスを独自に収集する。
(2)読解と作文のための文章の分析：コーパスを文章（ディスコース)、文、語の各層で分析する。
(3)誤用分析
曹・仁科（2006）で検討した母語転移の問題に関連して、学習者作文を収集し、文章構成、文構造、語彙選択、表記の各層の分析を行い、学習者の母語、習得レベルなどとの相関も分析する。
(4)共起表現の分析
・現在までに様々なコーパス中に現れる副詞と呼応する述部に含まれるモダリティの関係の分析
・名詞と格助詞＋述語の共起
(5) 文の難易度の分析
日本語能力 1-4 級の語彙・文型リストを文の難易度とし，難易度別の例文表示機能を実現したが，さらに Readability 研究の成果も参照しながら，文長と接続詞の関係や係り受けの深さといった文のわかりやすさを示す基準を再検討し，新尺度を考究し，結果を「なつめ」に実装する.
(6)ジャンルによる表現の差異の分析
・BCCWJ,青空文庫、学術論文、新聞、日本語教科書のコーパスから、統計的手法などを用いて、各ジャンルの表現の差違を実際に検証する。(Srdanovic、Bekes,仁科)
(7)評価
・コーパスの評価　利用したコーパスが適切かどうかを評価する。
・誤用検索システム評価
　閉じたシステムとしての誤用データベースのインターフェースを構築し、日本語教育・言語研究の専門家からの評価を得る。
・学習者評価　学習者が作文をしているときに瞬時に誤入力の判定をして、正解候補を表示するインターフェースを実現させるため、学習者自身の使い勝手を学習者に評価させる。
・システム全体の評価
　応答のスピード、正解を返す精度、使いやすさなどシステムとしての質の評価をする。
(8) 学習者コーパス構築
　学習作文データを収集し、学習者情報とタグ付き作文データを作成することで、学習者コーパスを構築する。タグ付けについては、まずタグセットの設定を行い、それにしたがってアノテータ(日本語教育専門家 3 名)が作業を行うこととした。当初作文データから Excel 上に目視、手作業で行ったが、2010 年度からは汎用アノテーションツール {SLATE}（東工大徳永研究室開発）を利用して、タグをつけることとした。
　このデータ結果をもとに、誤用検索サイトを構築し、さらに自動作文添削システム構築へと開発を進める。

４．研究成果
　日本語教育、自然言語処理、教育工学の専門家がそれぞれ学習者の言語分析、現存の日本語コーパスの分析、習得のための方略の設計、学習者コーパス構築、独自に用意したコーパスを含む大規模コーパスを用いての言語処理技術を利用した学習システム構築、システムに利用したコーパス、国内外の大学組織で学習システムの評価など様々な成果を得た。
(1) 理系論文コーパスの収集とシステム搭載
J-Stage に掲載されている電気学会誌、土木学会誌、自然言語処理などの論文について転載許可を得て、共起表現データ、例文表示用データとして利用し、Web 上での表示を実現した。
(2)コーパス分析、「なつめ」システム構築、視覚化
　使用するすべてのコーパスについて形態素解析と構文解析を行い、共起表現データベ

ース、例文表示用文を整備した。それらがジャンルごとに検索できるようにしたことで、「なつめ」において、ある分野で用いられる共起表現の傾向を、頻度情報などで検査することを実現した。

① コーパス分析

国研で開発した現代日本語均衡コーパス（以後 BCCWJ）および我々が独自に収集した J-Stage 上の学術論文を下記のように形態素解析し、共起表現の抽出、難易度による分類、ジャンルによるテキストデータの分類を行った。文の難易度については、readability に関する先行研究を参照し、例文表示をするときにレベル、母語などの諸要因を考慮に入れつつ、学習者に理解できる可能性を検討した。

② システム構築

ある語についてどのよう語が良く共起し、どの助詞を伴うかを知りたい場合、キーワードを入力すると共起する語が格助詞とともに頻度あるいは様々な確率指標値の順に表示するようにした。2011 年までは名詞か動詞の一方をキーワードとして、対になる語を検索する仕組みであったが、2012 年には共起する 2 単語を両方入力する機能を実現した。名詞と動詞がすでに想起できており、その共起対が日本語として正しいものか、または該当するジャンルで使用頻度が高いものかを判断したいことも多々あり、そのような状況に対応できるようにした。

③ 視覚化

(1)図 1 は、「なつめ」のインターフェースの一画面であり、学習者にとってわかりやすいように工夫して視覚化したものである。図 1 は画面左上に名詞あるいは動詞、形容詞の「キーワード」を入力することで、その下に、格助詞「が、を、に、で」などの項目ごとに、共起する動詞、形容詞あるいは名詞の共起語が提示される。

ポップアップによる機能の選択

2011 年までは例文表示では、画面下部のジャンル別表示の頻度表をクリックすることで表示されていた。しかし上級学習者や共起語を検索する日本語教師から、共起語リストから直接例文を表示させたいという声があったため、2012 年にはポップアップにより例文表示ボタンを表示し、簡易に表示できるようにした。

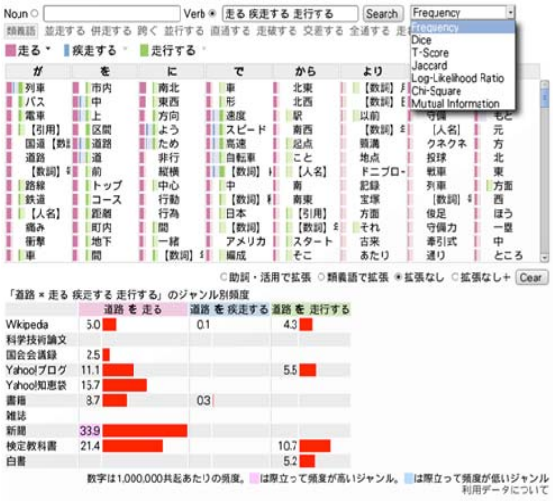

図 1 「なつめ」スクリーンショット

(3)評価実験

「なつめ」を使用した学習効果を実験により明らかにした。

「なつめ」使用群と非使用群に分けて、2 種類の実験を行った。一つは、話し言葉からなる単文を論文らしい文体に書き変える課題を留学生（大学生）に実施した。その結果、「なつめ」を使用することで、アカデミックな文生成に効果があることが明らかになった。一つは、事柄の是非を議論する文章作成課題を「なつめ」を使用群と非使用群の間で評価比較する実験である。この実験においても、「なつめ」使用群が適切な表現をし、豊富な語彙によって文章を生成したことを明らかにした。

（4）学習者コーパス構築

学習者コーパスを「なたね」と命名し、次の作業を行った。誤用分析のために学習者コーパス収集をする過程で、国内外において日本語学習者コーパスと誤用分析手法、データベースが皆無に近いことが明らかになり、その構築を行った。(a)学習者作文データの収集は理系大学生のレポート作文を中心に 5500 文を収集した。(b)学習者コーパスの必須要件としての「学習者情報」を作成した。個々の作文と作文作成者の母語、日本語レベルなどが作文とリンクしているものである。(c)誤用タグをつけるに当たり、誤用タグセットを体系から検討し、「誤用の対象」「誤用の内容」「誤用の要因」という 3 つの視点からタグセットを分けた。言語要素としては、「音素」「文字」「品詞」までの語と文のレベル、さらに待遇表現、レジスター、接続、論理展開、文体などディスコースの対象となる項目を加えた。さらに、この「なたね」コーパスを利用して、自動添削システム「ナツメグ」構築へと開発を進めた。このシステムでは作文目的、学習者レベル、母語などの情報画面上から入力し、学習者が作文すると、そ

の文章と正用および誤用データを照応し、確率計算をすることで、誤用らしいものには警告を発するという機能を実現するものである。正式公開は、2012 年 8 月の予定である。

(5)言語学的知見から第 2 言語習得への適用
本研究での
フィルモアの格文法理論、ハリデイの機能文法を学習支援システム構築に取り入れて、共起表現の効果的な表示法、コミュニケーションを目的とする視点でのレジスター概念の導入をした結果、システムのコンセプトが明瞭になった。

(6)学習者コーパスの必要性の認識
作文支援システムを開発する中で、日本語のオーセンティックな大規模コーパスだけではなく、学習者の誤りやすい箇所のデータを収集する必要性を認識した。そのため、我々のプロジェクトでは独自に学習者作文コーパスを収集し、誤用タグを付与した。そのためには、綿密で一貫性のあるタグセットの設計が重要であることがわかり、試行錯誤の結果、誤用の箇所、誤用の内容、誤用の理由という視点の異なる層と、音素からディスコースに至る多層からなる言語要素によるタグセットを提案した。現時点で、5000 文の学習者作文のタグ付データからなるコーパス「なたね」を構築し、さらにこれを用いた自動作文添削ツール「ナツメグ」を構築した。
(7)成果の発表
開発の成果は、下記の通り雑誌論文、学会発表を多数行い、国内外の日本語学習者の利用を促した。その結果、海外でも多数の国や地域で利用する学習者が増えてきた。利用者の要望を聞き、本科研費課題研究終了後も、改善を進めて行く予定である。
(8)今後の課題
・「あすなろ」のもつ辞書機能、多言語辞書、読解教材と「なつめ」がもつ共起検索機能、例文表示機能をさらに融合して、統合的に構築することは今回の研究機関では十分に果たせなかったため、今後の課題とする。
・共起対の検索では、入力された共起語とともに共起類似度の高い語を順に表示しているが、共起類似度の計算が我々の直感を反映できていない。この共起類似度の算出の改善が急務であると考えている。
・当初予測しなかった学習者コーパスの国内外の現状から、利用可能なコーパスを構築することが急務であると考えている。

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 20 件）
① ホドシチェク・ ボル、阿辺川 武、アンドレ・ベケシュ、仁科 喜久子、レポート作成のための共起表現産出支援、専門日本語教育学会、Vol．13、2011、pp.33-40、
②スルダノヴィッチ・イレーナ、ホドシチェク・ボル,ベケシュ アンドレイ,仁科喜久子、ウェブコーパスと検索システムを利用した推量副詞とモダリティ形式の遠隔共起抽出と日本語教育への応用、自然言語処理、言語処理学会 Vol．16 No．4、2009、pp.29-46
③ スルダノヴィッチ・イレーナ、ベケシュ アンドレイ、仁科喜久子、コーパスに基づいた語彙シラバス作成に向けて―推量的副詞と文末モダリティの共起を中心にして―、日本語教育、No．142、2009、pp．69-79
④ 村岡貴子、因京子、仁科喜久子、専門文章作成支援方法の開発に向けて：スキーマ形成を中心に、専門日本語教育研究 第 11 号、No．11、2009、pp．23-30

〔学会発表〕（計 14 件）
① 八木豊、Hodoscek Bor、仁科喜久子 BCCWJ と学習者作文コーパスを利用した日本語作文支援 第一回コーパス日本語学ワークショップ 2012 年 3 月 6 日、国立国語研究所
② 曹紅荃、仁科喜久子、文産出と質問紙調査から見る「通」で表記する和語動詞の習得について 、日本語教育方法研究会誌、Vol．19No．1、2012 年 3 月 10 日、pp．8-9、国際基督教大学
③仁科喜久子 日本語コーパスに基づいた日本語学習支援システムにおける語の提示、語彙・辞書研究会 第３８回研究発表会、語彙・辞書研究会、2010 年 11 月 20 日、新宿 NS ビル 308 会議室、pp．9-16

〔図書〕（計 3 件）
① 仁科喜久子(監修),鎌田美千子、曹紅荃、歌代崇史,村岡貴子（編集)、凡人社、日本語学習支援の構築- 言語教育・コーパス・システム開発-2012、285
② KIKUKO NISHINA, Publishing House Japan, Construction of Speech Database for Second Language Learning of Japanese, Computer Processing of Asian Spoken Languages, 2010,pp.147-150
③ 仁科喜久子 言語処理学事典 解説 日本語教育支援、言語処理学事典、共立出版 2009、pp．434
〔産業財産権〕
○出願状況（計 0 件）
○取得状況（計 0 件）

〔その他〕
ホームページ等
http://hinoki.ryu.titech.ac.jp/

６．研究組織

(1)研究代表者

仁科　喜久子（NISHINA KIKUKO）
東京工業大学・留学生センター・教授
研究者番号：40198479

(2)研究分担者

阿辺川武（ABEKAWA TAKESHI）
大学共同利用機関法人情報・システム研究機構・特任准教授
研究者番号：00431776

村岡貴子　（MURAOKA TAKAKO）
大阪大学・国際教育交流センター・教授
研究者番号：30243744

(2) 連携研究者

西方敦博（NISHIKATA ATSUHIRO）
東京工業大学・社会理工学研究科・
准教授
研究者番号：60260535

因京子（CHINAMI KYOKO）
九州赤十字国際看護大学・看護学部・
教授
研究者番号：60217239

テリー・ジョイス（Terry Joyce）
多摩大学・言語文化研究科・教授
研究者番号　20418677