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

平成　２９　年　　６　月　１５　日現在

機関番号：　２２６０４
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３７０６１１

研究課題名（和文）中上級者向け日本語コロケーション教材作成に向けたコーパスからの情報抽出とその応用

研究課題名（英文）Information Extraction from Corpus and its Application for Creating Japanese Collocation Teaching Materials for Intermediate and Advanced Level

研究代表者

長谷川　守寿（Hasegawa, Morihisa）

首都大学東京・人文科学研究科・准教授

研究者番号：５０２７２１２５

交付決定額（研究期間全体）：（直接経費）　　1,100,000円

研究成果の概要（和文）：　中級後半・上級前半の日本語学習者に向けた教材を作成するため、上級者向け教材に模範解答を追加し、不要な指示文を削除したデータを作成した。そのデータを、「リーディングチュウ太」で、語彙のレベル判定を行い、結果の確認を行った。その結果、形態素解析にはよる語の区切りの誤りや、それに起因する語彙レベルの判定の誤りが多数あることが分かった。機械的な修正では不可能で、データを全て個別に修正し、教材の特徴を明らかにした。
　また、特定の中上級者向けの日本語コロケーション教材を作成するため、中上級者が多く、実際に教材が求められている分野として幼稚園の配布文書を選び、文書の特徴を語彙の面から明らかにした。

研究成果の概要（英文）：Make teaching materials for intermediate and advanced Japanese learners. First, I added model answers to advanced materials and created data deleting unnecessary directives. Next, we entered data into "Reading Chuta" and judged the level of the vocabulary. After that, I checked the result. As a result, it was found that there were many mistakes in word segmentation due to morphological analysis and errors in judgment of vocabulary level caused by it. Because it is impossible with mechanical modification, I modified all of the data separately and characterized the teaching material.
Furthermore, in order to prepare Japanese collocation teaching materials for specific intermediate advanced level, there are many middle-advanced level learners, the distribution document of kindergarten is selected as a field to which teaching materials are actually requested. I clarified the characteristics of the document from the vocabulary.

研究分野：　コーパス言語学

キーワード：　コロケーション　語彙　レベル判定

１．研究開始当初の背景

国際交流基金では、世界各国での日本語教育の最新状況を把握するために、3 年に 1 度「海外日本語教育機関調査」を実施している。この 3 年間でも、学習者は 33 万人（9.1%）増加し、速報値で 398 万人を超えるという状況になっている。興味深いのは、教師数が 63,771 人と、2009 年の 49,803 人に比べ、28.0%も増加していることである。(「2012 年度海外日本語教育機関調査」結果 http://www.jpf.go.jp/j/japanese/survey/result/survey12.html)。

全体的な傾向として、高校等の中等教育段階の学習者の拡大が続いている中で、これら学習者の関心、期待に応える教材の開発・提供や教師の能力を維持・向上させる必要などの課題への対応の必要性も強く意識される結果が見られた。逆に言えば、学習者の期待に応える教材の開発が十分されてはいないということである。

研究代表者は、日本語教育に従事していく中で、初級・中級の単語を組み合わせて、複雑な意味を表す表現の習得に関心を持っていた。日本語上級者学習者は、現状では一度学習した語彙に対する関心は低く、「この魚は足がはやい」のように、全て 4 級の語彙からなる表現であるが、このような表現が表す意味は理解できないのが現状である。

そのため研究代表者は、日本語上級・超級者向けのコロケーションに関する日本語語彙教材を共同で作成した。教材開発には、膨大な経費とエネルギー、さらに実際の日本語教育の蓄積が必要で、大変な時間がかかった。幸いにして、その教科書は、国内で２刷りと版を重ね、台湾では、大新書局から廉価版が発行されるまでにいたった。

現在もその教材を使用し、授業を行っているが、上級・超級者向け教材であるため、文脈の難易度も高く、もう少し難易度を下げ中級後半から上級前半に向けた、コロケーション教材の必要性を痛感している。しかし、そのような中級後半から上級前半に向けた教材が存在していないのが現状である。

２．研究の目的

上級者・超級者向けの日本語コロケーション教材の語彙を分析し、そこから、語彙のレベルが高いものを考慮し、中級・上級学習者向けの教材が作成する手順を明らかにすることが本研究の目的である。

その際に、語彙レベルを判定する語彙チェッカーを用いて、語彙教材を分析し、教材の改訂に役立て、中級学習者も学習可能な教材に変えることを試みる。実際には、語彙チェッカーを用いることの利点、問題点を明らかにすることによって、今後の教材の語彙分析に生かしていきたいと考える。

教材作成について、国際交流基金(2008:iv)では以下の 8 つのステップがあるとしている。

１：教材を作る前に
２：教材を設計する
３：企画書を作成する
４：教材開発を進める
５：試作し、評価する
６：完成する
７：使用し、改善点を明らかにする
８：改訂する

関・平高（2015）では、教材開発のプロセスは、教科書開発のプロセスにも当てはまるとし、上記の国際交流基金(2008)と Jolly&Bolitho(2011:113)をまとめ、教科書開発の評価について、以下のように述べている。

出版後も教師、学習者の両者による評価が必要です。授業を通して使いやすさ、学習効果などのチェックを行い、必要であれば例文の差し替えなどを行って、増刷や改訂の機会があればそれに向けて準備をしておきます。
（関・平高（2015,p.18)）

しかし、ステップ７の「改善点を明らかにする」には、関・平高(2015)が述べるように教師や学習者による評価も必要であろうが、出版後の教師による使いやすさや学習効果の評価は、使用している教師を特定するのが難しいため、ごく一部の教師に限られてしまうという問題がある。また学習者による評価は、授業で使用した感想を集めることで可能となるが、そもそもその教材で学習するため、全ての内容が理解出来るとは言いがたい学習者に評価が出来るのかという問題が残される。また、教師や学習者による使いやすさの評価は漠然とした主観的な評価になってしまう恐れがある。

筆者は教材を作成する中で、執筆者の一人として他の執筆者（や時には出版社の担当者）とああでもないこうでもないと議論をする中で、例文を決定するプロセスを経て、教科書を作ってきた。改訂には、網羅的でかつ一貫した作業が要求されるが、改訂の際に執筆者側として同様の人的資源を使用できるとは限らず、逆に割くことが出来ない場合が多いことが予想される。さらに、執筆者が教師として教科書を使用していく中で、改訂する必要がある問題点に全て気づけるわけではない。また、教師や学習者による、やや主観的になりがちな評価では、どこを修正したら良いのかは結局執筆者の判断に任されることになってしまう。そこで、筆者は客観的なツールによる評価を用いて、教材の改善点の候補を明らかにすることが出来るのではないかと考える。本調査では、客観的なツールとして、ある語が学習されるレベルが表示される語彙チェッカーを挙げる。これを用いれば、学習者にとって未習の語彙が判明し、例文の差し替えなどを行うべき箇所の候補を、教科書の中で漏れなく挙げていくことが出来るのではないかと考え、語彙チェッカー

を用いて教材内の語彙の評価を行う。そして、学習効果や使いやすさなどの主観的な観点からでなく、語彙の難易度という観点からの評価を行うことの利点と問題点を明らかにし、実際に教材を評価した結果を基に、差し替えなどを行うか検討すべき箇所を特定し、改訂の機会に向けた準備とする。

３．研究の方法

小野正樹他（2009）『コロケーションで増やす表現―ほんきの日本語―vol.1』（くろしお出版）を対象に、語彙のレベル判定を行う。この教材は、ある語（例えば「手」や「頭」）を中心とした語彙教材で、中心となる語は 30 語である。教材内では「課」という呼び方はしていないが、説明の便宜上 30 課とする。それぞれについて、語彙のレベル判定を行う。

筆者は本教材を授業で使用していく中で、実際の受講者に対して難しすぎるのではないかと感じることが多くなった。昨今の学習者のレベルが、作成当時の学習者と比べて、相対的に低下しているということも考えられるが、現状の学習者のレベルに相応しくない語を含んでいることは確かである。ちなみに、実際に使用している首都大学東京の未習言語科目（日本語Ⅰ（上級））において 2015 年秋のクラスの単位取得を目標とした受講者は N1 取得者が 5 名、N2 取得者が 1 名である。授業のレベルを紹介したホームページには「日本語能力試験 N1 レベル」とされ、「アカデミックな日本語の基礎を総合的に学習します。」とあるが、教科書が学生のレベルにあっていないという印象は否めない。出版から 6 年が経過し、授業で使用してくる中で、教材の中心となる語でもなく、難解で必要性のない語句の存在なども分かってきた。必要な部分に集中できるよう、分かりやすい語句に替え、支障が出ないようにする必要がある。

次に調査の手順について述べる。まず調査対象であるが、出版社への入稿原稿である pdf ファイルからテキストを抜き出し、調査対象のデータとする。入稿原稿では答えを入れるために括弧になっている部分には、「別冊　解答」から答えを入力した。なお、「手＜　＞込んだ模様」のように複数の回答（「に／の」）がある場合は、初めの答えを入力した。また入稿原稿と製本は入稿後の校正により多少表記等が異なる部分があるが、その場合は製本を基準として修正する。

この本は以下の 6 つの部分からなる。これらを指示文と問題番号を除きテキスト化した（さらに学習を深めるための課題「STEP 調べてみましょう」もあるが、実例ではなく指示が多く含まれているため除外した）。

- 意味を記述した「こんな意味があります」（以後「意味」と呼ぶ）
- 問題文の語がどの基本義に該当するか考える「基本語のどれに当たるか考えてみましょう。」（「どれ」）
- 空欄に適切なことばを入れて意味を確認する練習（「問題 1」）
- 選択問題で該当する語と共起しやすい表現を学ぶ練習（「問題 2」）
- 該当する語の複合語などを学ぶ練習（「問題 3」）
- 今までに学んだ表現が新聞や小説の文体などで現れ、内容を確認する問題（「練習 4」）

次にレベルの判定方法について説明する。

本調査では、日本語の表現の難易度を判定するツールとして、「日本語読解学習支援システム　リーディング・チュウ太」(http://language.tiu.ac.jp)（以後「チュウ太」と呼ぶ）を選択した。表現の難易度を判定するサイトやツールには、先行研究で挙げたように多々存在するが、広く長く使用されていること、インストールや辞書の準備などの必要がないこと、問題点も一部明らかになっていることなどその実績を考慮し、チュウ太を選んだ。

チュウ太のテキストボックスに各課のテキストを部分毎に入力し、「語彙レベル」のボタンをクリックする。結果画面は 3 つに分かれるが、本調査では画面右の級別の単語リストと、画面下の単語レベルの判定結果をコピーして使用する。その下にある「総数・語彙総数・級外」等をカウントした表は使用しない。

チュウ太には「茶筌」と「出題基準」で単語の区切り方が異なることから発生する問題が存在する。また、形態素解析システムも 100%正しく解析できるわけではない。そこで、本研究では、少しでもその問題点を解決するため、以下のような手順で後処理を行うこととする。

後処理でチュウ太の結果に加えた修正点を示す。これはチュウ太の問題点を示すことでもあり、今後チュウ太を使用して語彙レベル判定を行う際には必要となる情報である。多く見られた誤りとその原因、その修正の方法について説明する。全ての誤りと修正を載せられないため特徴ごとにまとめて示す。

・読みが正しく同定されてない問題
・文が正しく語に区切れていない問題
・連接規則･辞書･語彙リストに起因する問題
・その他の問題

多く見られた誤りとして、漢字の読みが正しく同定されないことによる誤解析がある。「特定が難しく、合致する語彙が複数候補考えられる場合には、下の級を優先」(川村 1999)したことが原因であるが、単純に下の級に判定されていないものもあるため、文脈から判断して修正を行う。

たとえば「象が十頭いる。」の波線部に対して、チュウ太は「頭」を１級としている。１級で「頭」と記載されているのは「かしら」の読みの場合であるが、この文での読みは「とう」なので正しくない。この場合、頭（と

う）は２級の語なので２級に修正する。同様に「勝ち頭」の頭（がしら）を４級（あたま）と解析しているが、この場合は１級（かしら）に修正する。以下に同様の例を示すが、括弧内左の読みと級を、括弧右内の読みと級に変更した（なお、今回は漢字の読みを考慮した異なり語数の集計は行っていない）。読みが正しく同定されていない例としては、案内所（ところ・４級＝＞じょ・２級）、政治家（いえ・４級＝＞か・２級）、京都市（いち・１級＝＞し・２級）、留学先（せん・１級＝＞さき・２級）などがある。チュウ太を使用する際には、読みが多数ある漢字は以上のような誤った判断が見られるので、注意が必要である。

語の区切り方が正しくないのは形態素解析システムの問題であるが、たとえば、「あんな石頭」の「あんな」は、「ある（4 級）／だ（4 級）」となる。調べてみると「あんな人」でも同じ結果が生じるため、「石頭」の問題ではない。「出題基準」の語彙リストを参照し、この場合は「あんな」一語に統合し、さらに 3 級に修正する。

茶筌の連接規則に原因があると考えられる例がある。たとえば、「コンセプトはいい線行ってるけど」内の「いい」は、「いう（4 級）」と解析されてしまう。ただし、「いい線行ってるけど」だけならば「いい（4 級）」と判断されるため、「は」を含む連接規則による影響と思われるが、このような場合は正しい語形である「いい（4 級）」に修正する。

まず「出題基準」の問題に触れておく。2010 年から実施されている新しい日本語能力試験からは出題基準が非公開になったが、「出題基準」は今でも多く使われているため、問題点を示しておくことは一定の意味があると思われる。処理に困ったのが、基準が複数ある場合であり、たとえば 2 級の語彙リストに「＊～か（家）」（p.60）があり、3 級の語彙リストに「～か（家）」（p.23）がある。これは前出の「店（みせ）」（4 級）・「店（てん）」（2 級）のように、読みによる判定が出来ないという問題が発生する。本調査ではこのような場合、川村(1999)に倣い低い方の判定に合わせた。

最後に本調査に現れたチュウ太の問題点を記しておく。これは上記の問題とは違い、全てのテキストに当てはまるとはいえないのであるが、語と語が改行で区切られている場合、つまり句点「。」で区切られていない場合）、一語としてみてしまうことである。たとえば、ある文が「大」で終わり（句点「。」はなく）、次の行が「家に話を聞く」のような場合、「大家（1 級）」と判断してしまう。教材には箇条書きなども含まれると思うが、そのような項目をチュウ太にかける際には、あらかじめ句点を入れておく等、対策が必要となる。

４．研究成果

チュウ太を用いることの利点が明らかになった。以下に詳細に述べる。

まず、教材の著者の一人としては、表記の揺れを確認できたことに驚かされた。例えば「充分」と「十分」、「気づく」と「気付く」、「１０分」と「十分」等、筆者や編集者の目を通しても漏れてしまった表記の揺れが見つかった。表記の揺れが離れた場所に存在する場合、校正者であっても気づきにくい問題であろう。このようなツールを用いることにより、表記の揺れが明確になった。これは特に複数の筆者によって執筆されている場合などに有効なのではないかと思われる。また、チュウ太では句読点の有無によって、形態素解析に結果が変わるという特徴がある。たとえば「コンサートで、」という部分では、波線部を断定の助動詞「だ」の連用形として解析しているが、読点「、」がなければ格助詞として解析する。このように不要な読点の存在に気付かされ、ツールを通した結果を考察することは、表記の統一にも一役買うのではないかと思う。また、「パンツのチャックが開いていたことに気がついて」という例文中の「チャック」は「ファスナー」の商標名であるが、そのまま教材に掲載されていた。このように級外と判断されることによって、使用されている語の問題に初めて気がつくという面も存在する。

また、集計し数値化することにより、教材の中では特に目立った印象がない話し言葉的な表現の存在が明らかになった。ここは改訂の際に、修正すべきかの検討候補になり得ると思われるが、そもそもの問題として、習得させたかったのは書き言葉でのコロケーションなのか、話し言葉でのものなのか明確ではなかったと思われる。この教材には話し言葉と書き言葉が混在しており、教材の設計の段階でどのジャンルにおける用法か、統一した方針を明確にしておく必要があったと思われる。

なおチュウ太の本来の使用方法が、学習者が読解教材を自習する際の支援や、教師が教材を準備する際の支援であるとすれば、本来の使用方法とは異なるが、少なくとも上記のような項目は改訂の機会があれば修正を検討すべき点であり、このような点に目を向けさせてくれるという意味で有効な手段たり得る。

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２件）

①長谷川守寿・西尾広美「語彙・文型調査を目的とした『幼稚園の配布文書コーパス』の作成について―特定の目的コーパスの作成例として―」、『人文学報』513-7、首都大学東京人文科学研究科、2017 年 3 月、pp.55-71、査読なし

②長谷川守寿「語彙チェッカーを用いた語彙教材の分析とその問題点」、『人文学報』512-7、首都大学東京人文科学研究科、2016 年 3 月、単著、pp.1-20、査読なし

〔学会発表〕（計２件）
①長谷川守寿・西尾広美「語彙・文型調査を目的とした『幼稚園の配布コーパス』の作成」、言語資源活用ワークショップ 2016、国立国語研究所（東京都・立川市）2017 年 3 月 8 日
②長谷川守寿・西尾広美「『幼稚園の配布文書コーパス』の作成と試行調査」、言語処理学会第 22 回年次大会、東北大学（宮城県・仙台市）、2016 年 3 月 8 日

〔図書〕（計１件）
①荻野綱男・伊藤雅光・丸山直子・長谷川守寿・荻野紫穂編『データで学ぶ日本語学入門』編著、「第９章日本語教育」執筆、2017 年 3 月 10 日、朝倉書店、155(96-106)

〔産業財産権〕

○出願状況（計０件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　　件）

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
長谷川　守寿（Hasegawa, Morihisa）
首都大学東京・人文科学研究科・准教授
研究者番号：５０２７２１２５

(2)研究分担者
（　　）

研究者番号：

(3)連携研究者
（　　）

研究者番号：

(4)研究協力者
（　　）