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　３０　年　　６　月　２１　日現在

機関番号：　６２６１５
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ０１１１４

研究課題名（和文）日本語作文支援システムにおける誤用の検出及び添削に有用な情報の提示法の研究

研究課題名（英文）Study on information presentation method for misuse detection and correction in Japanese writing support system

研究代表者

阿辺川　武（Abekawa, Takeshi）

国立情報学研究所・コンテンツ科学研究系・特任研究員

研究者番号：００４３１７７６

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：本研究課題では、学習者が誤用を犯しやすい様々な言語要素の中から、「接続表現」に焦点を当て、「接続表現辞典」の構築をおこなった。具体的な手法として大規模均衡コーパスや科学技術論文集において文頭から読点までが5形態素以内の表現をすべて収集し、高頻出表現に対し人手で接続表現か否かのラベルを付与した。その結果、頻度上位1,000表現のうち568表現が接続表現として獲得できた。そして、得られた接続表現に対し様々な分析をおこない、中頻度の接続表現に対しては機械的に獲得可能であることを示した。今後、得られた接続表現を作文推敲支援システムに組み込み、誤用の検出、言い換え候補の提示に利用していきたい。

研究成果の概要（英文）：In this research project, we focused on "connection expression" from various grammatical elements which Japanese learner is liable to misuse, and constructed a "connection expression dictionary". As a concrete method, all expressions within 5 morphemes are collected from the beginning of sentence to the reading point in large-scale balanced corpus and science and technology paper collection. And we manually assigned labels of whether or not to express connection to frequently used expressions. As a result, we acquired 568 expressions as connection expressions. We conducted various analysis on the obtained connection expressions. and we showed that it can be obtained mechanically for medium-frequency connection expressions. In the future, we would like to incorporate the obtained connection expressions in the composition support system, to detect misuse and present candidates for paraphrasing.

研究分野：　自然言語処理

キーワード：　日本語学習支援　作文支援　接続表現　誤用検出　アカデミックライティング

１．研究開始当初の背景

日本語を学ぶ学習者は近年増加しており、国際交流基金による調査では 2012 年にはほぼ 400 万人に達し、中でも中・上級者の増加が報告されている。また教材の不足が指摘されている中で、大学や日本語学などの教室で日本語教師から直接教えを受ける学習者だけでなく、海外にいながらオンラインの学習教材を用いて自発的に学ぶ学習者が増えている。第二言語習得の四技能(読む、聞く、話す、書く)のうち、作文は学習期間後半で習得する傾向にあり、Web 上で提供されるサービスを見ても他の技能に比べて作文支援を目的としたサイトは少ない[1]。

このような状況で我々は 2000 年の日本語読解システム「あすなろ」の開発に続いて、2007 年に作文支援システム「なつめ」開発に着手し、作文で使用したい語に関する共起表現の検索と例文参照を可能にした。さらに学習者に利用しやすくするための工夫の一環として、2010 年から作文推敲支援システム「ナツメグ」の開発を行い、計量的手法を取り入れることで、理系レポートなどの作文に含まれるレジスター誤り(分野に不適切な表現の誤り)を指摘できるようにした[2] 。

開発に当たっては複数の科学研究費補助金、「大規模知識資源の体系」21 世紀 COE プログラム(東工大:代表者　古井貞熙)による研究費助成を獲得し、大規模な言語コーパスを利用した従来にない学習システムとしての成果を示した。以降では、関連する諸分野における位置づけを述べる。

<教育工学分野>坂元昴他(日本教育工学雑誌 27(3), 2003)において人間活動のグローバル化に伴う ICT の発達の視点から第二言語学習とその支援に関する教育工学研究の重要性を論じている。この方向性の示唆に基づき、我々は言語処理と日本語教育の知見を加え、新たな技術的、教育的成果にむけて研究を続けている。ICT の発達に伴い、作文を母語話者が相互に添削する Lang-8[3]などのサービスが普及している。この種の添削サイトでは、リアルタイム性に欠け、さらに添削された文のみが返されるため、学習者の意図通りに添削されているかの確認ができず、必要な情報が欠如している。我々のシステムは即時性と学習者の意図を汲むという点で上記の欠点を補填するものとなる。

<言語処理分野>誤用タグ付き学習者作文コーパスの入手が容易になったことから、機械学習や統計的手法を用いた誤用検出の研究が増加している。これらの研究は助詞の誤りや活用の誤りといった文法的な誤用の検出・訂正が多く、主に初級者を対象としている。一方で、本研究課題が対象とするレジスター誤りは、中・上級者でも犯しやすいものであり、日本語学習者全般の支援を可能にするオリジナルな提案である。我々の一連の研究は(水本他, 2012)でも紹介され、他の研究グループにも影響を与えている。

<日本語教育分野>日本語教育の分野では、筑波大学、国立国語研究所、東京外国語大学がコーパスを利用した検索システムを開発しているが、作文をリアルタイムで添削するシステムを行っているのは我々のプロジェクトが唯一であり、日本国内外での多数の利用者が存在する。また、連携研究者仁科は、2012 年にこれらのシステム開発における功績により日本語教育学会賞を受賞している。

２．研究の目的

本研究課題では、日本語学習者向けの作文支援システムにおいて、作文の誤用検出手法を確立し、検出した誤りをどのように指摘すれば効果的な学習につながるかを明らかにする。具体的には、アカデミックライティングを作文課題に設定し、中・上級者でも誤りを犯しやすいレジスター誤りの検出精度を向上させる。本報告書では接続表現辞典の作成を中心に成果内容を報告する。

日本語を第二言語として使用する日本語学習者がアカデミックな場面で論文およびレポートを作成する場合、目的にふさわしい表現かどうかの判断に悩むことが明らかになっている(Hodoscek 他, 2011)。この問題を解決するためにオーセンティクな日本語コーパスから学習支援システムに必要な項目を抽出することで機能表現辞典の構築を目指す。(八木他, 2016)では、科学技術論文という特定分野における表現を「論文のレジスター」と考え、他のレジスターとの異なりが顕著に見られる副詞表現に注目して分析した。これと同様の方法を用いて「接続表現」に注目する。従来、辞典、文法書、学校教育で用いられてきた「接続詞」および「接続語」の概念を整理し、連語を含めた新たな「接続表現」という概念を提示する。次に、科学技術分野のレジスター特有の「接続表現」を抽出し、論文・レポートを意図した学習者作文において、「あまり用いられない接続表現」があれば、それを指摘し、「接続表現辞典」を参照することで、代替表現を示す。これは、学習者が目標とする文章において「よく用いられる」適切な表現例を示し、学習者自身がそのデータをもとに、自ら考えて適切な表現に書き換えるというデータ駆動型学習（data driven learning）の考え方に基づいている。

３．研究の方法

「接続表現辞典」の構築には次の３つのステップが必要である。１）日本語の代表的な書き言葉コーパスと科学技術論文コーパスから「接続表現」を定義に基づいて抽出し、さらに国語辞典、日本語教育における語彙表などと比較する。２）言語におけるレジスターの概念を導入し、抽出した項目について、コ

[1] http://nihongo-e-na.com/jpn
[2] http://hinoki-project.org/
[3] http://lang-8.com/

ーパスにおける頻度を調査した結果から「論文」「論文以外の書き言葉」「硬い話し言葉」「砕けた話し言葉」の４つのレジスターに仕分けする。３）各接続表現について、その意義素と、レジスター別書き換え接続表現候補、および例文を作成する。本報告書ではこの内１）までを説明し、得られた接続表現リストの仔細な分析をした。２），３）については研究成果論文①を参照されたい。

3.1　「接続表現」候補の抽出方法

上記の BCCWJ および科学技術論文に対して形態素解析辞書 UniDic と形態素解析ツール MeCab を用いて形態素解析したものをデータベースとし、次の 1）から 3）の項目を満たす頻度上位 1,000 項目を抽出する。

1）BCCWJ および「科学技術論文」コーパスにおいて文頭から読点までを UniDic の定める語の単位の５単位までの表現を抽出する。例えば、「と｜いう｜の｜は｜、」が文頭にあった場合、この文字列は UniDic４単位の後に、「、」があることから５単位以内にあり、抽出の対象となる。

2）「接続表現」の対象に該当しない記号、数字、アルファベットなどを排除する。

3）2）で示した不要なものを排除した後の 1,000 番目までのリストを対象にして、国語辞典、日本語能力試験「旧出題基準」、日本語教科書、「分類語彙表」などを参照して「接続表現」に該当するか否かを１項目ずつ検討し、次の a）～f）のルールで選り分ける。a）～e）に該当するものは候補から排除し、f）は候補として残す。

a)「人称代名詞」およびそれを含む複合表現-「彼女、僕｜も、私｜と｜し｜て｜は」など

b) 時を表す UniDic 単位の表現および複合表現-「今日、３時｜に」など

c）感情を示す UniDic 単位およびそれを含む複合表現、会話での受け答えなど一般に「感動詞」と言われるもの-「まあ、はい、あの｜ね」

d) 挿入フィラーと判断される UniDic 単位および複数表現-「あの」「ええ｜と」

e)「副詞」のみの機能と判断される UniDic 単位およびそれを含む複合表現-「きっと、ひょっと｜し｜たら、もし｜か｜し｜たら」

f）文頭から抽出された５単位までの表現が複数の品詞からなる場合、接続機能を担っていると判断されれば、「連語」として採用する。「そう|する|と、と|は|いえ」などは「助詞、動詞、名詞」などを含む連語である。

以上の方法で 1,000 項目のリストを得た後、我々で定めた「接続表現」の定義に照らして、各項目が「接続表現」として認められるか否かを研究代表者を含めた４名で検討し、568 項目を採択した。その結果、「こ」「そ」を含む指示代名詞の「連語」、「といえば」のように「助詞・助詞」および「動詞」の活用形の組み合わせなど「連語」が多く選定された。

４．研究成果

得られた「接続表現」の分析結果を研究成果として示す。

4.1　UniDic との比較

表１は選定された 568 項目中で代表的でかつ判定時に問題となった「接続表現」の項目を示している。文字表記は UniDic の語彙素をもとに「公用文作成用例」などで示される標準的な表記に置き換えたものである。568 項目中、UniDic の品詞付与で短単位の「接続詞」は 16 項目（2.82%）である（また、そして、ただし、なお、ただ、さらに、で、が、さて、けど、さてさて、もっとも、すなわち、あるいは、もしくは、しかも）。一方、「接続詞」を含む複数短単位からなる語は 14 項目である（けれど|も、で|もっ|て、さらに｜また、で|ね、しかし|ながら、など）。これらの複合形の要素には、「また、そして、あるいは、さらに、しかし」などの高頻度短単位が多く含まれている。「名詞」単独のものが 23 項目ある（「実際、事実、以上、結果、結局、当初、反面、半面」など）。これに関連して、(水谷・星野，1994)は「名詞」と「副詞」の間に存在する語の性質を整理し、従来からの品詞の境界線の再考を促している。他に UniDic の「名詞・副詞・代名詞・連体詞・動詞」の組み合わせによる「連語」(「そう｜する｜と、具体｜的|に|は、従い|まし|て」など）は、新たな複合語を構成し、「接続表現」の機能をもつ項目として含まれている。なかでも「指示語」が「連語」構成要素となっているものが多い.詳細は 4.5 節で述べる。

4.2　国語辞典との比較

辞典の記述においても複数の品詞の結合形は辞典編纂者によって判断が異なる。表１によると、「副詞」と「接続詞」の判断も異なっており、４種の国語辞典でも品詞の異同がみられる。「接続表現」568 項目中に「これから」「でも」「だが」「といって」（岩国・接続詞）「とすれば」（明鏡・接続詞）「なのに」「にもかかわらず」など「動詞、名詞、助詞、助動詞」など他の品詞との組み合せによる「連語」となる「接続表現」がみられる。

4.3　「旧出題基準」との比較

「接続表現」568 項目と日本語能力検定「旧出題基準」を対照すると、177 項目が合致する。そのうち「旧出題基準」において「接続詞」と明記されているのは１級２項目（だ、だと）、2 級 9 項目（あと、けれど、けれども、従って、すると、だから、で、でも、だが）、3 級１項目（だから）、４級０、計 12 項目と極めて少ない。合致する項目の中で「接続詞」以外の品詞としては「指示語」11 項目、「副詞」１項目、「感動詞」１項目、計 13 語が含まれているが、他は品詞の付与がない。レベル別では２級が「接続表現」リストの７割以

上を占め、他のレベルはいずれも１割以下であり、全体の配置はアンバランスである。

4.4　意味用法における分析との比較

不適切な表現の指摘から修正案を提示するという過程において、使用意図を理解した上で、接続表現を解釈しないと修正案は示せない。このような場合に「接続表現」がもつ「意味用法情報」が必要になる。(石黒, 2015)は、「接続詞」の用法を大きく５種に分けて、「論理、整理、理解、展開、文末」の「接続詞」とし、さらに階層構造で細分化して、例文によって用法を示している。このうち、「文末の接続詞」は、本報告書の「接続表現」の対象ではないが、他の４分類は対応している。「理解の接続詞」を例にあげると、その直下の階層に「換言」、「例示」、「補足」の３種類があり、さらにそれぞれの下に「つまり」系、「たとえば」系、「とくに」系などという「系」の階層が付いている。また一般には「接続詞」とは呼ばれない名詞「事実、結果、本来」、「そうすると、というのは、にも関わらず」など複数単位からなる「連語」も含めて提示している。「と言うと」は、石黒の分類によると、「理解の接続詞」の階層下の「換言の接続詞」のさらに下の階層「つまり」系に相当する。「つまり」系の中には、「すなわち、つまり、言い換えると、言わば」など書き言葉と話し言葉を合わせた類義語が集められている。我々が目指す辞典では、このような類義語をレジスターごとに整理した言い換えのためのシソーラスの装備が必要となる。なお、「文末の接続詞」には「のではない」「と思われる」「必要がある」などの分析整理をしている。本稿の対象にはならないが、「連語」からなる「機能語」として今後取り組むべき課題である。

4.5　他の資料に出現しない「接続表現」

「接続表現」568 項目のうち調査した資料にも出現しない項目が次の 23 項目である。

「あとは、ここで、ここでは、このことは、このことから、このとき、このほか、この結果、この場合、これにより、これに対し、これに対して、そうだ、その一方で、その意味で、その結果、その際、その中で、その点、その場合、その理由は、それと、中には」

これら 23 項目は UniDic では、短単位を組み合わせた「連語」である。電子化されたテキストを UniDic 短単位で処理すると、これらの項目は抽出されないことになる。

次に、注目すべきこととして 23 項目の内、21 項目は「そこ、その、それ、ここ、この、これ」などの「指示語」を含んでいる。

「その」は61種(全コーパス中の頻度21,052)の連語を形成している。「それ」54 種（頻度 25,356)、「この」54 種（頻度 16,903），「これ」36 種（頻度 17,673），「そう」28 種（頻度 5,773)、「そこ」7 種（頻度 11,501）と続き、連語形成要素として多くの「接続表現」を構成し、全コーパス中接続表現 568 項目における割合は 23%である。「指示語」は文章中の前方照応をする機能があり、前後の語句あるいは文の連結を担う「機能語」として働いていると考えられる。

4.6　他の資料にあって「接続表現」568 項目のリストにない表現

本節では他の資料にあって、「接続表現リスト」にないものについて述べる。(石黒, 2008）に掲載されている「接続詞整理表」４種 10 類の接続詞の中の接続詞 117 例および「旧出題基準」において接続詞と明記されている 8 項目と「接続表現」568 項目を対照した。「接続詞整理表」中 26 項目は、本稿の「接続表現」のリストにはない。この中で、文中の名詞間接続の機能をもつ「および、かつ」、やや古風な「いな」など、我々があらかじめ排除したものが数例ある。「そうはいうものの、どっちみち」など砕けた話しことば７例、他は「指示語」を含む表現で、活用形が違うがほぼ同意のものであった。石黒は広い分野から接続詞を抽出していることがわかる。

次に、「旧出題基準」の「接続詞」で「接続詞」リストにない語は、１級「だと」、２級「けれど」の２項目のみであり、他はリストに含まれている。

また UniDic における「接続詞」はすでに述べたように短単位であるため、項目数は少ない。語彙素数が 31 項目、その出現形は 104 項目である。その中で、本報告書の「接続表現」にはない項目としては文中の語と語の接続が３項目（および、かつ、と)、古語に近い表現４項目（さりとて、しかれども、然して、されば)、方言２項目（けんど、ばってん）である。他の 18 項目は本報告書の「接続表現リスト」に含まれている。

他の資料にあって、「接続表現リスト」にない項目は、我々が意図的に制限したものも含めて、限られていることが明らかになった。また石黒の「接続詞整理表」では、我々が「接続表現」と定義した条件に合うものが多数見られるが、多くは「指示語」を含む連語である。以上のことから、本「接続表現リスト」は、現在用いられている標準的な接続表現をほぼ網羅しており、他の辞典に比べて遜色のない「接続表現辞典」といえる。

これに関連して、「接続表現」568 項目の頻度順で100位ごとにUniDic１単位の語の数を見ると、上位 100 位までは 28 項目、200 位までと 300位までは８項目、400位までは４項目、500 位までは３項目、501 位～568 位までは５項目となっている。この傾向をからすると 568 位より低頻度の「接続表現」では UniDic １単位の語はあまり出現せず、連語のパターン分析から得られる生成装置で多くの表現を補完すると考えられる。

4.7 より大規模な接続表現の収集

当初の研究計画では日本語学習者に対し作

文を執筆させ接続表現に関する誤用を検出するという評価実験を行う予定であったが、今回得られた接続表現リストだけでは、学習者の記述する接続表現の多様性に対応させることは難しいことがわかった。そこで学習者に対する評価実験の代わりに、さらなる接続表現の獲得に向けてアルゴリズムによる機械的な収集方法を考案し、機械的に収集した接続表現の獲得精度を測る評価実験を代わりに行うこととした。機械的に接続表現を収集する方法については、各種言語リソースを組み合わせたルールベースの手法を提案した。文頭から最初の読点まで 5 形態素以内で出現するすべての表現を大規模コーパスから抽出すると延べ 976,367 表現、異なり 275,279 表現が収集できた。そこから 5 つの言語リソース(UniDic、分類語彙表、つつじ:日本語機能表現辞書、JUMAN 辞書、IPADIC)を用いて、接続表現とみなす表現をフィルタリングすると異なりで 5,910 表現が抽出された。このすべての接続表現を評価することは現実的に難しいので、2 つの接続表現リストと比較することで抽出精度の評価をおこなった。1 つめは本報告書で獲得した 537 表現(その後の見直しにより 568 表現から減少している)で、このリストと比較すると再現率 0.633、適合率 0.909 となった。2 つめは(石黒, 2008)で紹介されている 121 表現と比較をし、その結果、108 表現(89%)が獲得できることがわかった。適合率が 9 割前後であることを考えると、5,910 表現のうち約 5,300 表現は正しい接続表現であるとみなせ、この接続表現リストを用いれば日本語学習者の作文における多様性に対応できるものと考えられる。

＜引用文献＞

①Bor Hodoscek, 阿辺川 武, Andrej Bekeš, 仁科喜久子 (2011)「レポート作成のための共起表現産出支援－作文支援ツール「なつめ」の使用効果－」『専門日本語教育研究』.
②石黒圭 (2008)『文章は接続詞で決まる』光文社文庫 光文社.
③石黒圭 (2015)「書き言葉・話し言葉と「硬さ/軟らかさ」: 文脈依存性をめぐって」特集ことばの「硬さ」「やわらかさ」『日本語学』34(1), 14-24, 明治書院
Vol.13, 34-40.
④水谷静夫, 星野和子 (1994)「名詞から副詞まで－語類の新しい枠づけ」,『計量国語学』19 (7), 331-340.
⑤水本智也, 小町守(2012) 「なんで日本語はこんなに難しいなの？--リアルな日本語学習者コーパスの分析と言語処理の課題--」『情報処理』, Vol.53, No.3, pp.217-223.
⑥八木豊, ボル ホドシチェク, 阿辺川 武, 仁科喜久子 (2014)「作文推敲支援システムによる誤り指摘への学習者の対処に関する調査」『日本教育工学会研究報告集』日本教育工学会 14(5), 151-156.

５．主な発表論文等

〔雑誌論文〕（計 1 件）

① 仁科喜久子、八木豊、ホドシチェク・ボル、阿辺川武、作文学習支援システムのための接続表現辞典構築、計量国語学、査読有、31 巻、pp.160-176、2017 年

〔学会発表〕（計 2 件）

① 阿辺川武、八木豊、Bor Hodoscek、仁科喜久子、アカデミック・ライティング向け接続表現リストの獲得、日本語教育国際研究大会、2016 年
② 阿辺川武、Bor Hodoscek、八木豊、仁科喜久子、作文推敲支援システム「ナツメグ」の誤用指摘手法の改善、The 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J)、2015 年

〔図書〕（計 1 件）

① 仁科喜久子、八木豊、阿辺川武、ホドシチェク・ボル、くろしお出版、誤用分析からみた作文指導への示唆 in 習ったはずなのに使えない文法、江田すみれ、堀恵子(編集)、pp.211-232、2017 年

〔その他〕
ホームページ等
https://hinoki-project.org/

６．研究組織

(1)研究代表者
阿辺川　武（ABEKAWA, Takeshi）
国立情報学研究所コンテンツ科学研究系・特任研究員
研究者番号：00431776

(2)研究分担者
ホドシチェク　ボル（Hodoscek, Bor）
大阪大学言語文化研究科・講師
研究者番号：10748768

(3)連携研究者
仁科　喜久子（NISHINA, Kikuko）
東京工業大学・名誉教授
研究者番号：40198479

(4)研究協力者
歌代　崇史（UTASHIRO, Takafumi）
八木　豊（YAGI, Yutaka）
曹 紅荃（SOU, Kosen）
Irena Srdanovic

表１：「接続表現」候補高頻度語彙

順位	頻度	代表表記	UniDic	旧出題基準	新明解	三国	岩国	明鏡
1	48,199	また	接	4	副,接	副,接	名,副,接	副,接
2	38,016	しかし	接	4	接	接	接	接
3	18,462	そして	接	4	接	接	接	接
4	12,741	例えば	副	3	副	副	副	副
5	12,675	でも	助｜助	2 接	副,接	連，副助，接	連, 副助, 接助, 接	接,副
6	11,912	と	助	1 接	－	格助，接助，副助，終助，接	格助，接助，接続詞的に文頭	接
8	10,105	そこで	代名｜助	2	接	接	接	接
11	8,844	これは	代名｜助	これ 2	－	感	－	感
14	7,670	だが	助｜助	2 接	接	接，連	接	接
18	6,104	で	接	2 接	格助，接(話)	格助，接助，接(話)，助動，形動	格助，助動，接助(文頭で接続詞的に)	接(話)
21	5,395	まず	副	3	副	副	副	副
61	1,155	そうすると	副｜動|助	－	－	－	接	－
97	686	それにしても	代名|動|助|助	－	接	接	「それ」追込(―にしても)	－
100	684	とはいえ	助|助動	－	連	格助「と」追込「とはいえ」	連,接	－

凡例　名:名詞　動:動詞　接:接続詞　副:副詞　形動:形容動詞　助動:助動詞
助:助詞　格助:格助詞　接助:接続助詞　連:連語　話:話し言葉