科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　２０　日現在

機関番号：２２６０４
研究種目：基盤研究(C)
研究期間：2011～2013
課題番号：２３５２０６４０

研究課題名（和文）文構造を考慮した日本語コロケーション情報の抽出とその応用

研究課題名（英文）Japanese collocation extraction of information that takes into account the sentence structure

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（研究院）・准教授

研究者番号：５０２７２１２５

交付決定額（研究期間全体）：（直接経費）　　1,200,000円、（間接経費）　　360,000円

研究成果の概要（和文）：文構造を考慮に入れた日本語のコロケーション情報を抽出する方法を考察した。データの検証として、新聞データおよび実際の新聞紙面はどんな違いがあるか確認した。次に、明らかになった新聞データの問題を修正し、2種類のデータ、1文単位に分割されたデータ、及び変更されていないデータを用意した。各々に対し形態素解析を行った後、データベースに変換し、高頻語語彙についてのコロケーション情報を抽出し、結果を考察した。
　さらに、対象による結果への影響を考察するために、上記と同様の処理をBCCWJを対象に行う。まず１文単位に分割するために、BCCWJの文構造タグを検証し、サブコーパス単位で問題点を挙げた。

研究成果の概要（英文）：How to extract Japanese collocation information taking sentence structure into consideration was considered by this research. First, newspaper data and an actual newspaper checked what kind of differences there were. Next, newspaper data was corrected and two kinds of data, the data divided per one sentence and the data which has not been changed, were made. After the morphological analysis of each data, they were changed into the database and the collocation information on a high frequency vocabulary were extracted and the results were considered. Then, in order to divide BCCWJ per sentence, the sentence structure tags contained in BCCWJ were checked.

研究分野：人文学

科研費の分科・細目：言語学・日本語教育

キーワード：教育工学　教材　教育メディア　コロケーション　教材作成

１．研究開始当初の背景

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っても、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、大変な時間がかかる。作成の中で、コーパスからの有効な情報の抽出を何度か試みたが、従来の方法では有効な情報は得られず、日本語教育者としての直感を元に、辞書の情報や WEB の情報などを選り分けていくしか方法がなかった。また調査対象とすべき均衡のとれたコーパスがなく、研究代表者の場合、新聞記事データである『CD-毎日新聞』や、『CD-ROM 版　新潮文庫の 100 冊』（新潮社）から発行が昭和元年以降の作品に限定したものや、さらに総合研究大学院大学の『小松左京コーパス』（http://aci.soken.ac.jp/databaselist/BC001_01.html）など、入手が可能なデータを検索対象とした。しかし、ブラウンコーパスなどに代表されるような、バランスのとれたコーパスではないため、データの網羅性に関する保証がないまま行わざるを得なかった。また、作成された教材は、日常生活に使われるような表現を対象としたものであり、学習者の専門日本語を学習したいというニーズに応えることができているとはいえない。

しかし、大規模でバランスのとれたコーパスの需要が高まり、2006 年に開始された国立国語研究所による『現代日本語書き言葉均衡コーパス』（http://www.tokuteicorpus.jp/）のプロジェクトは、2011 年に完成が予定されていた。既にサンプルデータや領域内公開データを元にした、現代語の研究は盛んに行われるようになった。しかし、まだその使用法については、ツールの開発などが中心であって、教材開発は十分検討されているとはいえなかった。本研究は、『現代話し言葉均衡コーパス』（Contemporary Written Balanced Corpus of Japanese、以後 BCCWJ と略す）の日本語教育への応用方法の一つと位置づけ、BCCWJ から日本語教育教材作成に有効となるコロケーション情報を抽出すること、さらにジャンルを限定して、コロケーション情報を抽出することにより、専門日本語教育への応用も考察したいと考えた。このようにバランスのとれたコーパスを使用することにより、語彙教育においては、実際に使われている日本語としての真正性（Authenticity）が保証された教材の作成が可能となると考えた。

２．研究の目的

本研究の当初の目的は、真正性を備えたコロケーション情報の抽出を行う際に必要となるデータの整備法について提案し、その有効性と問題点を明らかにすることであった。さらに、主に身体語彙のコロケーション情報の抽出を行い、日本語教師が考える学習項目と、テキストから得られたコロケーション情報によって導かれる学習項目の照合を行い、本研究の有効性について明らかにする。最後に、コーパスから自動的に学習項目を抽出する方法について提案することであった。なお、以下に述べるように目的の一部を変更した。

３．研究の方法

本研究で提案するコロケーション情報抽出法の妥当性を明らかにする。

従来のコロケーション情報は、形態素 N-gram を用いた方法が従来採用されてきた。このような隣接ペアからの情報抽出では、頻度が高い語彙の場合、有効である。なぜなら、たくさん用例が出現するため、文構造上、その語に関係しないような情報、いわゆるゴミ情報は目立たなくなるからである。しかし、「日本語能力試験出題基準」（国際交流基金・日本国際教育協会著　凡人社　1991 年）に含まれるような日本語教育で対象となる語彙は、全てが頻度で決められているわけではなく、中頻度の語彙も含まれている。そのような語の場合、コロケーション情報を求めるならば、正確さが必要となる。例えば、深田 淳 (2007) で説明されている茶漉（http://tell.fll.purdue.edu/chakoshi/public.html）のように、茶筌をベースに設計された、用例およびコロケーション情報を抽出するシステムでは、「頭」のコロケーション情報で、後 3 語を「青空文庫」を対象とした時には、「ながら」がコーパス頻度で見た場合 8 位に位置してしまう。これは(1)のような情報を抽出しているためと思われる。

(1)彼は静かに頭を振りながら、怪訝そうな目でじっと相手の顔を見詰めているのであった。（『街頭の偽映鏡』佐左木俊郎）

また、被修飾語となる場合を探すために、前 3 語のコロケーション情報を得ようとすると、同じ資料では、「ちょっと」がコーパス頻度、期待頻度でともに 1 位になってしまう。また、形態素に区切ってから調査するため、「頭」を含む「石頭」や「頭金」などは、抽出されない。そこで、中・低頻度の語彙の正確なコロケーション情報を抽出するには、文字列を対象としていたのでは不十分であり、文構造を考慮したコロケーション情報の抽出方法が必要になると考え、その方法の有効性を明らかにする。

研究を開始した段階では、BCCWJ の DVD 版を入手する事ができなかった（実際に入手できたのは、2012 年３月になってからであ

る)。そのため、方法を確立するために、新聞データを対象として、研究を開始した。まずは、対象とするデータのクリーニングを行った。これは、電子化された新聞データは、実際の紙面をどの程度正確に反映しているものか確認するためであり、さらには、「CD-毎日新聞データ集」に収録されている毎日新聞記事データの特徴を調べ、言語研究に使用する際に注意すべき点を明らかにすることにある。なお、新聞社の発行している記事データ集は、ある年全ての新聞記事を収録したものであるという設計上、コーパスに求められる要件である、言語資料としての代表性や均衡性は有していない。しかし、記事は新聞社各社の校閲の過程を経ており、文の正しさという点では保証されていて（それでも含まれる誤植には目をつぶらざるを得ない)、さらに非常にまれに古い作品を引用の形で含むこともあるが、新聞の特性上、記事が執筆された時期と、それが発表された時期が近いという特徴を持っている。

その結果、４．１に示すような問題点が明らかになったので、解決可能な問題点を修正の上、構造を考慮に入れたコロケーションの抽出を行った。節の情報までを考慮に入れたかったが、節単位の抜き出しには、困難点があり、また節自体への区切りにも問題が見られた。そこでまず初めに、文単位でコロケーション情報を抽出することを行った。その方法としては、文単位に整形し、それを元にコロケーション情報を抽出するという方法である。データは、１年分の新聞記事で、比較のために２種類のデータを用意した。一つは、句点を元に文単位のデータに整形したものである。もう一つは文単位に区切るような加工は施していないデータである。この２種類のデータに対して、MeCab と UniDic を用いて、形態素解析を行った。次にその形態素解析の結果を、ChaKi.NET を用いてデータベースに変換し、高頻度語彙（上位２０語）のコロケーションを抽出した。その結果、以下のような点が明らかになった。「言う」などの引用句を伴う語の場合は、文単位のデータを用いた方が、修正を加えていないデータよりも、実態に即した、ゴミの少ない結果が得られた。ただし、他の動詞に関しては、抽出された助詞のデータに大きな違いは見られず、また、抽出された名詞のデータは、対象としたデータの性質による影響がどのくらいあるのか、見定めることができないという問題が発生した（なお、この部分については現在原稿を準備中である)。

そこで、その時点では入手可能になっていた BCCWJ-DVD 版を対象として、新聞データと同様の調査を行なうことを計画した。まず、紙ベースの情報と電子化データの確認であるが、これは比較する紙ベースの情報がないため、省略した。次に文単位での抜き出しである。この段階で問題が発生した。BCCWJ-DVD 版を入手し文タグ(<sentence>と</sentence>で一つのペア)を用いて、検索用プログラムを作成していたところ、文タグに問題があり、完全には文単位に分けられたデータが得られないことが判明した。最初に発見したのは、(2)である。入力の誤りがあるため（“「いや、よしてし”の部分は“「いや、よして」”が正しいと思われる）理解しづらいが、明らかに文タグが示す一文中（<sentence>と</sentence>に挟まれた間）に二つ以上の文が入っている。

(2)<sentence>これを「いや、よしてし(ママ)と読めば、いやがっているという意味である。はたしてそれだけだろうか。句読点をすこし動かして「いやよ、して」と読めば、(中略)。</sentence>　(PB10_00030)

このように、元々のデータに文境界を示すタグが欠けているという問題があることが分かった。なお、今回は、句点（。）のような、明確な文境界を表す記号を使用したが、OY などには、(3)のように文境界を示す記号が含まれていない用例が多数ある。この文は本来ならば、「思いもかけないことを発見しています。多くの方はご存知のことかもしれません」のように、句点「。」によって文境界を示されるべきであるが、(3)には、それらを示すタグが欠如している。これらについては、言及できなかったが、文構造を考える上では、避けては通れない問題である。

(3)思いもかけないことを発見しています多くの方はご存知のことかもしれません
(OY14_05003)

そこで、文タグが欠如している箇所はどのくらいあるのか、明らかにするために、出版サブコーパス、図書館サブコーパス、特定目的サブコーパス全てを扱う。その際データ形式は、ファイル数が多く形態素情報のタグがついていない C-XML のデータを使用する。さらに、文単位を対象とするため可変長（Variable）を用い、ファイルに含まれる全ての文を対象とする。そして、文タグが入力されていない箇所を探す正規表現を使用し、どのような場所で文タグが挿入されていないか検討する。その際、修正が必要な箇所はどのサブコ－パスに多いのか、それはどのような出現傾向が見られるのか、明らかにする。

４．研究成果

４．１．新聞について

新聞紙面と新聞データの比較によって、以下の点が明らかになった。まず、新聞データには、データが重複して入力されている問題が見られた。この場合、ほぼ内容が同一のデータが続けて現れるという傾向があった。これはＩＤ（数字 8 桁でユニークな情報）が異なり、記事見出しの文言が多少異なり、さらに記事見出しの末尾に【大阪】という文字列

を持つかどうかの違いがあるだけであり、重複するデータをどのように扱うか、方針を決める必要がある。

また、記事の見出しの一部が記事本文になっている問題が見られた。本来ならば、記事の見出しは全て、見出しの中に入っているべきであるが、(4)のように、見出しの一部が記事に含まれている問題が明らかになった。

(4)　＼Ｔ２＼告］生きる<ENTER>

また、新聞紙面では表の中に記述されていた文を、新聞データにテキスト化する際に、改行が正しくなく、(5)のような部分が文として含まれてしまっている。修正を要する箇所である。

(5)　農家民宿で「どぶろ<ENTER>

また、読点後に改行記号が入っている問題点として、文が読点「、」で終わる文には(6)のような記事が含まれ、文が正しく検出されなくなっている。

(6)ついで官邸の施設に触れ、<ENTER>
「いたれり尽くせりで……」<ENTER>
と言うと、それも、<ENTER>

さらに、新聞記事データベースの中に、「小説」という文のスタイルとしてはかなり異なるジャンルのテキストが含まれている問題が発見された。これは著作権の問題で含まれないであろうと（使用する側が勝手に）判断していたデータに、何らかの手違いと判断していいかもしれないが、新聞社が著作権を持っていないデータが載っているのである。これらは二度と修正されることがないため、これらが検索結果に影響を与える可能性についても考慮しなければならないことが分かった。

以上、データのクリーニング作業を通して、「CD-毎日新聞データ集」に含まれる毎日新聞記事データの特徴と言語研究に使用する際の注意点が明らかになった。

４．２．BCCWJ（DVD 版）の文構造タグに関して

文タグを修正してから、BCCWJ のデータを研究対象として使用する場合、修正が必要な箇所はどれくらいあるのか、修正箇所の多寡はサブコーパスや媒体により違いはあるのか、そして目的にもよるが、これらのことから文タグを修正して使用するには、どのサブコーパスや媒体が適当かを明らかにした。

その結果、修正が必要な箇所の多寡は、サブコーパスに依存し、修正して使うのであれば、出版サブコーパス（PB・PM・PN）や図書館サブコーパス（LB）が適していると思われる。PM（出版-雑誌）・PN（出版-新聞）は修正箇所数が少なく、PB（出版-書籍）・LB（図書館-書籍）は、修正箇所数は多いが真に修正が必要な箇所である可能性が高く、文タグの追加などの修正後、データとして用いることが可能である。逆に、特定目的サブコーパスは、OM（国会議事録）・OL（法律）・OV（韻文）のように修正箇所数が少ないものも多いが、OY（Yahoo!ブログ）や OC（Yahoo!知恵袋）は修正箇所数が多い割には、その場所が文の終端ではない可能性も高く、さらに修正して文として一様に扱うためには、タグの追加だけではなく、タグの削除やデータ自体の修正が必要になり、困難が予想される。また、OY（Yahoo!ブログ）では、文の終端が様々で、文末を探すのが難しいという問題がある。

(7) □３兄妹めちゃいいわぁ～（＾＾ゞあのＴシャツ、マジでガチで欲しいですw笑　にしても可愛い♥（略）　(OY14_21962)
(8) 傑作ポチを戴けるととても嬉しく思います♡　(OY14_28649)
(9)早く元気になってねぇ～♥　（略）「カシャカシャ」って、ケーキなんかも作りたい♫　デジカメ持って散歩もしたい♫　(OY14_35446)

例えば、上記の(7)(8)(9)では“♡”や“♥”、“♫”“♪”　“ｗ笑”などが文の終端に位置し、これらのサブコーパスを対象とする際の、文境界の決定の難しさが明らかになった。文単位のデータの正確な抽出にはこれらの解決が必要となる。

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 「BCCWJ の文構造タグに関する一考察」、長谷川守寿、『人文学報』488、首都大学東京、査読なし、pp.23-48、2014 年 3 月
② 「CD-毎日新聞データ集」に含まれるデータの特徴と使用上の注意点について」、長谷川守寿、『人文学報』473、首都大学東京、査読なし、pp.31～pp.49、2013 年 3 月
③ 「新聞紙面と新聞記事データ集の相異について」、長谷川守寿、『人文学報』443、首都大学東京、査読なし、pp.20-45、2011 年３月

〔学会発表〕（計　２件）

① 「BCCWJ のタグ情報の修正について」、第５回コーパス日本語学ワークショップ、2014 年 3 月
② 「CD-毎日新聞データ集」に含まれるデータの特徴について、言語処理学会第 19 回年次大会、2013 年 3 月

〔図書〕（計　０件）

〔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　なし

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