科学研究費助成事業　　研究成果報告書

平成　２６　年　　６　月　１４　日現在

機関番号：　１４４０１
研究種目：　挑戦的萌芽研究
研究期間：　2011～2013
課題番号：　２３６５２０９８

研究課題名（和文）大規模コーパスを使用した日本語複合動詞データベース構築法に関する基礎研究

研究課題名（英文）Fundamental issues in construction of Japanese compound verb database from Japanese corpus

研究代表者
今井　忍（Imai, Shinobu）

大阪大学・日本語日本文化教育センター・准教授

研究者番号：２０２９４１７６

交付決定額（研究期間全体）：（直接経費）　　1,500,000円、（間接経費）　　450,000円

研究成果の概要（和文）：任意の日本語コーパスから自動的に複合動詞のデータベースを構築することのできるソフトウェアの開発においてどのような点が問題になるかに関する基礎的な諸問題について検討した。複合動詞の前項動詞、後項動詞、および複合動詞全体のそれぞれについて格パターンの抽出を行う手順を開発し、簡便なソフトウェアの実装を行った。

研究成果の概要（英文）：This study investigates the fundamental issues in the development of the software which automatically constructs the compound verb database from any Japanese corpus. We constructed the procedure for extracting the case frames of the first elemnt, the second element and the whole of each compound verb and implemented a simple software which executes the procedure.

研究分野：　人文学

科研費の分科・細目：　言語学・日本語学

キーワード：　日本語　複合動詞　データベース

１．研究開始当初の背景

近年の言語研究の主要なトピックの一つとして、レキシコンと統語論の境界設定の問題がある。日本語の複合動詞についても、この問題設定を中心に様々な研究がなされてきている（『文法と語形成』（影山太郎、ひつじ書房、1993）など）。一方、認知言語学の枠組みの中でも、複合動詞構文の分析は重要な地位を占めている。特に、複合動詞後項の多義性は、認知言語学の主要な問題の一つである意味拡張という観点から様々な分析がなされている。申請者も「複合動詞後項の多義性に対する認知意味論によるアプローチ」（『言語学研究』第 12 号、京都大学言語学研究会、1993）において、複合動詞後項の「～出す」の起動の意味の分析において、Langacker の「主観化（subjectification）」という概念が有効であることを示した。

さらに、複合動詞研究は、単に日本語学の領域にとどまるものでなく、英語の句動詞構文、朝鮮語の複合動詞構文、中国語の動詞連続構文といった、他言語の構文との対照にも広がっている。

このように、複合動詞は、日本語の記述的・理論的研究、日本語と他言語との対照研究において重要な位置を占めつつあるが、常に問題となるのは複合動詞のデータベースの不足である。特に、どの複合動詞がどのような意味的・統語的特性を持っているかを知る必要があるが、そのための基礎的なデータが不足しているのが現状である。『複合動詞の構造と意味用法』（姫野昌子、ひつじ書房、1999）や『日本語複合動詞ハンドブック』（Tagashira & Hoff、北星堂、1986）などの研究書・参考書についても、複合動詞のリストや意味や統語的特性の記述はあるものの、収録されている語彙数に限界があり、詳細な分析を行う場合には基礎となるデータとはなり難く、新奇な複合動詞用法を知ることもできない。

２．研究の目的

このような現状に鑑み、本研究では以下のような特徴を持つシステムの構築を目的とした：

A. データベースそのものの構築が目的ではなく、データベースを構築する方法の開発
B. コンピューターの操作にあまり詳しくない研究者や学習者にも使うことができるシステムの開発

(1)のような目的を設定するのは、複合動詞には次のような重要な特性があるからである。複合動詞は、レキシコンと統語論の狭間にある現象であり、純粋に語彙的な特異性だけからなるものでもなく、統語的な規則性のみに支配されるものでもない。したがって、ある程度形が固定されていると同時に、ある程度の規則性（生産性）も持つという中間的な性質を示す。例えば、複合動詞後項の「込む」には、「前項が表す動作にかかり切りになる、集中してそれを行う」といった意味がある（「考え込む」「（毎日 10km）走り込む」）。しかし、この意味における「込む」は、いわゆる統語的複合動詞ではなく（「*思索し込む」「*（先生が生徒を）走らせ込む」）、どのような前項動詞とも複合するというわけではない。しかしながら、決められた動詞としか複合しないということでもなく、常に新しい用法が作り出されているという側面もある。例えば、「この人形は丁寧に作り込まれている」のような「作り込む」は新しい用法であり、一般の国語辞典にも、前出の姫野 (1999) や Tagashira & Hoff(1986)にも収録されていない。

このような性質は、データベースを構築する際に、大きな問題点を投げかける。すなわち、ある特定のコーパスに基づいてデータベースを構築したとしても、時間が経つにつれて、実際に使われている形式との「ずれ」が生じてくるのである。例えば、あるコーパスから、上述の「作り込む」という形式を抽出してデータベースに含めることができたとしても、その後「（ケーキを）焼き込む」や「（家を）建て込む」などが仮に生まれたとしたら、それらの新用法を含むコーパスを基にして、その都度データベースを更新する必要が出てくるということになる。

このような問題は、結果としてのデータベースを得ることを目的とする限りは解決することができない。発想を転換し、データベースを「構築する方法を構築する」ことを考えなければならない。すなわち、データベースを自動的に構築することのできるソフトがあれば、時期ごとに区別されたコーパスそれぞれにソフトを適用することによって、時期別のデータベースを自由に作成することができるし、分野別のコーパスを用意すれば、分野別の複合動詞データベースを作ることができる。それにより、複合動詞の語彙的特異性と統語的性質の関係を動態的に捉えることも可能になり、従来から指摘されてきた語彙的複合動詞から統語的複合動詞への意味拡張過程やそれに伴う統語的特性の変化をも、より実証的に捉えることが可能になる。さらには、言語学において現在最も興味深い問題の一つであるレキシコンと統語論のインターフェースの問題に対しても、大きな貢献をなすことができると考えられる。

(2)は、今後のコーパスに基づく日本語研究に必要になると考えられる条件である。従来のコーパス分析は、ある程度情報処理の知識がある工学系の研究者が中心になりがちだったが、これまで日本語学や日本語教育学で蓄積されてきた知見を直接的な形で検証するためには、電子データの扱いに慣れていない研究者たちでも容易に操作できるようにする必要がある。

３．研究の方法

本研究で構築しようとしたデータベースは少なくとも以下のような情報を含むものである。

A. 見出し語
B. 前項の形式と格フレーム
C. 後項の形式と各フレーム
D. 複合動詞全体の格フレーム
E. それぞれの複合動詞の用例

そのために、まず複合動詞の取る格についての先行研究を参照し、どのようなパターンがあるかを整理することにした。特に多くの前項動詞と結びつく後項動詞に関して、その用法と格パターンとの関係について明らかにすることを目指した。

４．研究成果

各年度の成果は以下のとおりである。

初年度は、まず、これまでの複合動詞に関する研究から意味と文法的特徴（特に格支配）との相関性を明らかにした。計画としては、「～出す」「～込む」「～つける」「～かける」を対象とする予定であったが、特に「～出す」と「～込む」について考察した。「～出す」については、『日本語基本動詞用法辞典』から、「外部移動」（「太郎がポケットから財布を出した」）、「発現」（「太郎は不満をすぐに表情に出す」）、「産出」（「子どもが熱を出した」「太郎が本を出した」）の 3 つの基本用法を設定した。それぞれの用法の格支配のパターンは決まっており、「外部移動」は＜ガ格＋カラ格＋ニ格＋ヲ格＞、「発現」は＜ガ格＋ヲ格＋ニ格＞、「産出」は＜ガ格＋ヲ格>である。これらの格パターンは、「出す」が複合動詞後項として使用される場合にも引き継がれており、前項動詞の格パターンに応じて、特定の意味と結びついて具現化することが明らかになった。また、「込む」については、現代語における単独用法が限られた用法しか持たないため、複合動詞後項としての用法に基づいて、「内部移動」（「水が側溝に流れ込んだ」）、「縮小」（「お腹が引っ込んだ」）、「強化」（「父も最近はめっきり老け込んだ」）を認定した。それぞれの用法の格パターンは、「内部移動」が＜ガ格＋ニ格（＋ヲ格）」、「縮小」と「強化」が＜ガ格（＋ヲ格）」である。

次に、データベース作成のための技術的な側面に関して検討を行った。本研究は格支配という構文上の特性を扱うため、構文解析という観点から現在の研究の状況を検討した。

第二年度は、1)データベース作成方法の具体的な手順に関する考察、2)今年度に公表された 2 種類の複合動詞データベース（「Web データに基づく複合動詞データベース」「複合動詞レキシコン」、いずれも国立国語研究所のホームページからアクセス可能）の検討を行った。

1）については、小規模なテキストデータを使って複数の方法でデータ抽出を行った。その結果、形態素解析と係り受け解析によって複合動詞全体の格支配の抽出はある程度自動化できるものの、人手によるデータのチェックを行わなければ十分なデータが得られないことが分かった。また、表記上、一語として扱われるもの（「認める」「率いる」など）をどのように扱うかについても問題が生じることが分かった。2)に関して言えば、「Web データに基づく複合動詞データベース」は格解析の手順について参照すべき点が多いことが分かったが、動詞の組み合わせが限られている点、受身・使役の形式が含まれない点に問題があると考えられる。また、「複合動詞レキシコン」については、格支配だけでなく意味情報を含んでいる点で本研究の目的に合致するものであるが、収録されている形式が限られており様々な動詞の組み合わせを網羅的に抽出するものではないことがわかった。また、これらのデータベースはいずれも特定のコーパスから一回的に抽出されたものであり、本稿が目的とする抽出の手順そのものの構築とは目的が異なることも分かった。

最終年度は、1) 国際シンポジウム「日本語及びアジア諸言語における複合動詞・複雑動詞の謎」（国立国語研究所）への参加、2) テキストデータからの複合動詞の抽出を行うプログラムの実装、を行った。

1)では、日本語とその他のアジア諸言語における複合動詞の基本的性質についての知見を深め、参加した他の研究者との交流からデータベース構築に当たってどのような情報が必要になるかを考察することができた。また、2)については、「Web データに基づく複合動詞データベース」の構築方法（山口昌也 2013）を参考にした。このプログラムによって、複数のテキストデータから複合動詞の形式、それを含む文、および、格要素をデータ化することができた。

これらの成果の一方で、今後解決すべき課題が多く残されていることも判明した。受身形や使役形については、結局適切な検索方法を開発することができなかった。また、意味情報との関連づけについてもほぼ着手できずに終わった。格要素の出現パターンと意味の間には一定の関係があると考えられ、その関係を明らかにすることで有益なデータが得られることが予測されるが、市販の日本語の辞書の記述は意味と格要素の関連がほとんど記述されていないため、その関連づけを明らかにするのが困難である。『日本語基本動詞用法辞典』のような格要素が明確に記述された辞書の利用を含めた検討が必要である。また、コンピューターの扱いにあまり詳しくない研究者や日本語学習者も利用できる簡便なシステムの開発も十分に達成できなかった。

今後の展望としては以下の点が挙げられる。まず、すでに公開されている「Web データに基づく複合動詞データベース」および

「複合動詞レキシコン」との連携である。これらのデータベースの公開が本研究の期間の中途で行われたこともあり、方針の転換が間に合わず、十分な検討ができなかったが、いずれのデータベースも、本研究が目指していた方向性と軌を一にするものであり、本研究が実現しようとしていた機能の多くがこのデータベースに含まれていると言ってよい。したがって、本研究の目的であった、データベース構築のためのソフトウェアを開発するに当たって、これらのデータベースの開発過程を参照することで、有益な示唆が得られると考えられる。しかし、これらのデータベースの開発方法から分かるのは、有用なデータを抽出するためには、ある程度の人手による処理が必要になるということである。当初の目的は、ソフトウェアによる自動的なデータベースの構築であったが、今後は人手による後処理をも考慮に入れる必要があると考えられる。

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）

６．研究組織
(1)研究代表者
今井　忍（IMAI, Shinobu）
大阪大学・日本語日本文化教育センター・准教授
研究者番号：２０２９４１７６