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　２８　日現在

機関番号：　８２６３６
研究種目：　研究活動スタート支援
研究期間：　2015～2016
課題番号：　１５Ｈ０６２５８

研究課題名（和文）構文形態論の形式モデルの構築に関する研究

研究課題名（英文）Studies on a formal model of construction morphology

研究代表者

淺尾　仁彦（Asao, Yoshihiko）

国立研究開発法人情報通信研究機構・ユニバーサルコミュニケーション研究所データ駆動知能システム研究センター・研究員

研究者番号：１０７５５１１９

交付決定額（研究期間全体）：（直接経費）　　1,450,000円

研究成果の概要（和文）：本研究では、形態論（単語の内部構造）に関して、言語経験がどのように言語知識を形作るかを探究するための基礎となる、日本語の語構成データベースの構築を行った。語構成データベースは、例えば「招き猫」という語が動詞「招く」と名詞「猫」から成るといった情報を網羅的にカバーするもので、幅広い研究用途を持つものであるが、複合動詞など一部のカテゴリーに限定されたものを除外すれば、これまで日本語には存在しなかった。また、このデータベースのための検索ツールを公開した。これを利用することで、例えば「歯磨き」のように「名詞＋動詞」という語構成をもつ複合名詞を全て検索するといったことが容易にできるようになった。

研究成果の概要（英文）：In this study we built a database of word structures in Japanese, which will be a basis for investigating how speakers' linguistic experience relates to their knowledge on morphology (i.e. word-internal structures). With a database of word formation contains information that, for example, the word "manekineko" consists of the verb "maneku" (to invite) and the noun "neko" (cat). Despite of the potential usefulness of such a database in a wide range of research topics, Japanese has not have one, except those for specific categories such as compound verbs. A web-based search interface was also built for this database. With this interface, one can easily search for, for example, all instances of compound nouns that are made of a noun followed by a verb, such as "hamigaki" ("ha" (tooth) + "migaku" (to brush), "tooth brushing").

研究分野：　言語学

キーワード：　形態論　日本語　コーパス　言語資源　用法基盤モデル　構文文法

１．研究開始当初の背景

言語知識が個々の言語経験からボトムアップ的に形作られるという視点（用法基盤モデル）は、広く認知言語学分野では共有されている。この視点は、より具体的には以下のように要約することができる。

- 言語知識は構文(慣習化された形式と意味との組み合わせ)のネットワークから成る。
- 構文は具体的な個々の言語経験から共通性を見いだすことによって得られる。多くの言語経験によって支えられた構文は定着度が増していく。
- ある表現の自然さは、その表現を具体例とするような定着度の高い構文があるかどうかによって決まる。

しかしながら、用法基盤モデルについて、実際に形式的・定量的なモデルが与えられることは稀であるため、どのような予測が得られるのか、反証可能性のある形で提示することが難しい。またそれに関連して、心理学研究や工学応用など他分野との接続も十分に行われていないという現状がある。

また、用法基盤モデルの枠組みで、現実に話されている日本語などの言語について具体的な予測を行うためには、話者がどのような言語経験に晒されているかを知るため、個々の言語的要素の頻度などに関する情報を得ることが不可欠であるが、とくに形態論分野（単語の内部構造を扱う分野）においては、現状利用可能な言語資源からは、必要な情報を得ることができないという現状がある。例えば形容詞から名詞を派生する「-み」という接尾辞の頻度について調べようとしても、既存の形態素解析辞書には「強み」のような単語が一語として登録されており、この単語が「-み」という接尾辞を含むという情報は得られないためである。

２．研究の目的

本研究課題では、用法基盤モデルに基づく自然言語の形態論に対して、反証可能な予測を行うような基盤となる枠組み及びデータを作成することを目標とした。現実に話されている日本語などの言語について、そのような基盤を得るためには、以下の組み合わせが必要である。

- 話者の言語経験の近似としての、コーパスにおける形態論レベルでの頻度等の情報
- 話者の言語知識を表す、容認度判断などのデータ
- 両者を結びつける形式的な理論

当初は上記の 3 点の準備を順次進め、容認度判断などのデータとコーパスベースの頻度とを比較することにより、頻度が容認度判断に与える影響について定式化することを計画していたが、研究代表者の所属変更によって当初の計画通りのエフォートを割くことが難しくなったため、予算を縮小し、1 点目に重点を置き、日本語の形態論情報データベースを網羅的に整備する方略を採った。語構成情報データベースは英語等には以前から存在しているが日本語にはなく、本研究のみならず形態論・音韻論・語彙論の幅広い研究において意義が大きいと思われるためである。

３．研究の方法

本研究では、コーパスや形態素解析用辞書に基づいて、個々の形態素の生起頻度やその文脈についての情報を容易に得られる環境を構築することが主要な課題となった。先述のように、既存の形態素解析辞書では、言語学的な意味での形態素（意味の最小単位）の情報は得られない。そこで、形態素解析用辞書（具体的には UniDic）の見出し語に対し、その内部構造の情報を付与した語構成情報データベースを構築する。これとコーパスの検索とを組み合わせることで、従来、形態素解析済みコーパスであっても検索できなかった、語より小さい形態論的な単位での検索を可能にする。

４．研究成果

(1) UniDic への語構成情報の付与

まず、既存の形態素解析辞書である UniDic の見出し語を利用して、語構成情報データベースの構築を行った。UniDic を利用したのは、加工再配布の自由なライセンスで公開されていること、日本語書き言葉均衡コーパス(BCCWJ)等の主要な大規模コーパスで利用されていること、単語の採録基準が言語学的な考慮に基づいており斉一的であること、語種やアクセントなど言語研究に有用な情報が付与されていることなどが挙げられる。語構成情報は、UniDic のある見出し語の構成要素がまた UniDic の見出し語になっている場合、再帰的にリンクを張る形で表現した。例えば、「飛び箱」という見出し語の内部構造は、見出し語「飛ぶ」および見出し語「箱」へのリンクとして実現した。そのほか、以下のような情報も合わせて付与した。

- 形態素境界情報。例えば「雨傘（あまがさ）」であれば、「雨（あま）」と「傘（がさ）」のあいだに形態素境界があるという情報。

- 語に生じている形態音韻論的現象についての付属情報。「雨傘（あまがさ）」であれば、「雨（あめ）」と「雨（あま）」が交替するいわゆる被覆形・露出形の交替と、「傘（かさ）」と「傘（がさ）」が交替するいわゆる連濁。付加情報には他に、半濁音化、音便（促音便、撥音便）、音挿入（促音挿入、撥音挿入、ノ挿入）等を用意している。

「飛び箱」という単語を例に、本研究で付与する情報の概念図を図 1 に示した。

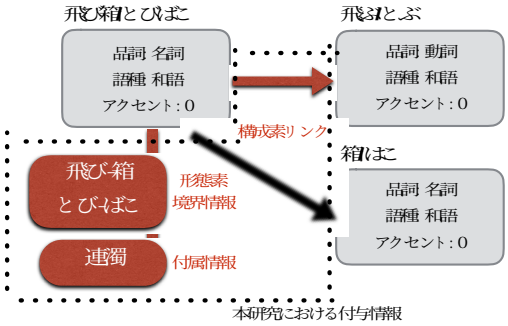

図 1　語構成データベースの概念図

本研究課題の期間終了時点で、構成要素も UniDic の見出し語になっているような複合語については基本的に整備済みである（人手でチェックしていない部分は一部残っている。また、音韻変化や表記が例外的で予測できないものは除く）。漢語熟語（「電車」など）や派生語（「疑わしい」など）については、現段階ではその語構成情報をデータベースに含めていない。ひきつづき整備を行う予定である。

現段階での整備状況は以下の通りである。UniDic の見出し語を、表記揺れなどを吸収して独自にまとめ、196,486 語とした。このうち 96,275 語は記号、固有名詞、外来語であり、本データベースでは語構成情報の付与は行わない。残りの 100,211 語から、24,826 語の複合名詞、8,099 語の複合動詞、204 語の複合形容詞を認定している。

以下の表は、複合名詞・複合動詞・複合形容詞について、語構成ごとの頻度と例を示したものである（主要な品詞から成る複合語のみ掲載）。

名詞	名+名	11,246	横顔
名詞	動+名	3,839	出前
名詞	形+名	1,241	早口
名詞	名+動	5,517	目隠し
名詞	動+動	2,249	置き引き
名詞	形+動	304	早起き
名詞	名+形	283	品薄
名詞	動+形	27	切れ長
名詞	形+形	21	遠浅
動詞	名+動	232	目立つ
動詞	動+動	7,842	話し合う
動詞	形+動	24	近寄る
形容詞	名+形	153	根強い
形容詞	動+形	23	蒸し暑い
形容詞	形+形	28	重苦しい

これら本研究で新しく付与された情報と、UniDic に元々付与されている品詞、活用タイプ、語種（和語・漢語・外来語等の区別）、アクセントなどの情報を組み合わせることにより、従来の言語資源では容易には実現できなかったさまざまな検索が可能になった。

例えば、「手書き」や「物書き」のような「名詞＋動詞という語構成をもつ名詞」については、これまで、異なる語形成のタイプや、連濁の有無・アクセントを決める条件などについて研究の蓄積がある。しかし、従来はこのカテゴリ全体を辞書ないしコーパスから検索するということは容易ではなかった。このような分野では、この言語資源は極めて有用である。

合わせて、この語構成情報データベースのための検索ツールをウェブ公開し、学会発表を通じて情報共有を行った。検索ツールはテキスト処理やデータベースに関する専門的な知識がなくてもデータベースが直感的に利用できるようにすることを意図している。

ある程度の網羅性が確保できた時点で、データ本体も含め公開する予定である。

(2) コーパス検索との統合

上記の語構成情報データベースの整備と並行して、日本語書き言葉均衡コーパス(BCCWJ)をローカルで検索する簡易ツールを作成した。BCCWJ は「中納言」など、オンラインの検索ツールも提供されているが、独自の検索ツールを用意したのは、API 等が現段階では提供されておらず、スクリプトを用いた自動的なアクセスには適さないためである。このツールを先述の語構成情報データベースと組み合わせることにより、形態素解析によって得られる単位よりも細かいレベルの形態素を指定したり、語構成パターンについての抽象的な情報を指定したりして、コーパスにおける頻度や出現文脈の情報を機械的に得ることのできる環境を整えた。

(3) 生産性と文法的性質の関係に関する研究

データベースの整備と並行して、構築中のデータベースや検索ツールを利用しつつ、理論的な示唆をもつ具体的な研究として、「彼頼み」「ここ止まり」など、代名詞を含む合成語についての研究を行った。

複合動詞においては、「そうする」という代用形によって前部要素を置き換えることができる「-始める」「-続ける」などの動詞と、置き換えることができない「-取る」、「-込む」のような動詞があることが知られてお

り、前者を統語的複合動詞、後者を語彙的複合動詞と二分する分析が知られている。研究代表者はこれまでに、用法基盤モデル的な視点から、このような文法的性質の違いは、頻度の違いを原因として生じているという分析を示してきた。「彼頼み」のような語は従来、「そうし始める」のような統語的複合動詞の例とは並行的に捉えられていないが、実際には類似の現象であると考えた。そのうえで、統語的複合動詞と代名詞を含む複合語に共通する用法基盤モデル的な説明が成立するかどうかについて検討した。

一部の代名詞に関して見れば、「彼名義」「彼女任せ」などの合成語を作る「-名義」「-任せ」などの語ないし接辞には、生産性の高さや句の包摂を許すなどの性質との相関が見られ、これらは統語的複合動詞の性質と並行的である。ただし、「そっち」などさまざまな代名詞について検討すると、複合動詞の場合と異なり、代名詞の使用可否は生産性とは単純には相関しないことがわかる。動詞の代用形である「そうする」と異なり、代名詞は数が多く、その意味によって振る舞いが変わってくる。語が構成的に解釈されるかどうかは、生産性によっては決まらない部分があることを示している。

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 0 件）

〔学会発表〕（計 3 件）

① 淺尾仁彦、「フリーな形態論情報データベースと検索ツールの構築」、形態論・レキシコンフォーラム 2016、2016 年 9 月 11 日、慶応大学（神奈川県横浜市）

② Yoshihiko Asao、Word-internal pronouns in Japanese、The 24th Japanese/Korean Linguistics Conference、2016 年 10 月 14 日、国立国語研究所（東京都立川市）

③ 淺尾仁彦、「日本語語構成情報データベースの構築」、言語資源活用ワークショップ 2016、2017 年 3 月 7 日、国立国語研究所（東京都立川市）

〔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
http://asaokitan.net/jmorph/

６．研究組織

(1)研究代表者

淺尾 仁彦 （Asao, Yoshihiko）
国立研究開発法人情報通信研究機構・ユニバーサルコミュニケーション研究所 データ駆動知能システム研究センター・研究員

研究者番号：10755119

(2)研究分担者

(3)連携研究者

(4)研究協力者