コーパスからの複合動詞の自動抽出の試み
―近現代作家の文学作品からの用例抽出を例に―

チャクマク　ビルギル・ニハル（アンカラ大学/麗澤大学）†
千葉庄寿（麗澤大学）††

An Extraction of Japanese Compound Verbs from Corpus:
An Example of Modern Literary Works

Nihal Cakmak Bilgili (Ankara University/Reitaku University)
Shoju Chiba (Reitaku University)

要旨

これまで谷崎潤一郎作品に出現する複合動詞をマニュアルにて抽出し、データベース化してきた。今後、さらに多くの谷崎作品から複合動詞を網羅的に取得するにあたり、検索作業の自動化を試みた。本稿では、既存の複合動詞データベース 3 件(野田 2013, 山口 2013, 国立国語研究所 2015)のデータを統合して検索・処理に使用し、谷崎作品に応用した結果に基づいて評価をおこなう。また、複合動詞と単純動詞を区別せず、一般動詞の語彙素として認定している UniDic の解析データの問題点の克服を試みるとともに、さらに、既存の複合動詞データベースに含まれない「新しい」複合動詞の候補を効率よく発見するための工夫についても紹介する。

1. はじめに：谷崎作品からの複合動詞用例抽出

複合動詞とは、動詞の前に何か別の単語を組み合わせたもので、日本語では、大別すると「嘆き悲しむ」、「笑い飛ばす」等の動詞＋動詞型の複合動詞と「名づける」、「旅立つ」などの名詞＋動詞型の複合動詞の 2 つのタイプに分かれる（由本 2013:59）。本研究で取り上げる複合動詞型は「動詞連用形＋動詞」型である(「テ形」＋動詞は扱わない)。代表的な先行研究には、日本語の「動詞＋動詞」型複合動詞の研究として影山（1993:78-96）に代表される、「語彙的」複合動詞と「統語的」複合動詞との 2 分類が挙げられる。「語彙的」複合動詞の特徴は、2 つの動詞が一体化して意味の不透明化や語彙化が進んでいることであり、レキシコンに登録されているものと考えることができる（影山 1993:78）。一方、統語的複合動詞では、そのような意味の習慣化が見られず、2 つの動詞の意味関係は完全に透明であり、2 つは使用される際に統語的に組み合わされるものと考えられる（同上）。

本研究では、既存の複合動詞データベース 3 件(野田 2013, 山口 2013, 国立国語研究所 2015)のデータを統合し、複合動詞のリスト(「複合動詞統合データベース」と呼ぶ)を作成した。そのリストに「茶まめ」を利用し UniDic による形態素解析を施し、検索処理のためのプログラムの参照元となるデータを作成した。

UniDic では、登録されている複合動詞は単純動詞から区別されず、一般動詞と同様に 1 つの語彙素として認定されている。さらに、UniDic においては、登録されていない複合動詞が 2 つの語彙素に分割されたり、ケースによっては誤解析されたりする。本稿では、コーパスから対象となる複合動詞をコーパスから網羅的に検索するための方策として、

† nihalckmk@gmail.com
†† schiba@reitaku-u.ac.jp

UniDic に登録されていない複合動詞についても抽出できるよう検索プログラムを工夫した。

発表者によるこれまでの谷崎潤一郎作品における複合動詞の分析においては、短編小説『刺青（しせい）』、『少年』と『春琴抄』から複合動詞を手作業にて抽出し、用例データを作成してきた。このデータと本研究で作成した検索プログラムを用いて谷崎の作品全般からの用例抽出に実際に応用した結果を比較することで、検索ツールとしての評価をおこなうことができる。

発表者の研究の射程は統語的複合動詞を含めた「動詞+動詞」型複合動詞の全体像を記述することであり、研究課題にはどのような統語的複合動詞が文献にどの程度用いられているかといった問題も含まれる。既存の複合動詞データベースには統語的複合動詞が含まれず、主として語彙的複合動詞のみが収集されている（山口 2013, 国立国語研究所 2015）。そこで、その複合動詞データベースに含まれない統語的な複合動詞、およびこれらのデータベースにない「新しい」複合動詞を発見することも重要であり、本研究のプログラムが求めるところである。

2. 既存の複合動詞データベース 3 種の特徴

本稿では既存の複合動詞データベース 3 件(野田 2013, 山口 2013, 国立国語研究 2015)のデータを統合し、検索のベースとなる複合動詞リストを作成した。使用されたデータベース 3 件はその目的や収集方法等が異なるので、まず本節では、これらのデータベースの特徴と統合の過程と結果について示す。

野田（2013）が「日本語教育のための資料として提示することを目的」に作成した複合動詞一覧は、約 2,400 語の複合動詞を含む。一覧には、国語辞書の複合動詞の語彙項目に加え、「辞書に載らない、意味の分かりやすいものをできるだけ拾」っている(野田 2013:38)。

『Web データに基づく複合動詞用例データベース』（山口 2019）とは、山口昌也氏によって作られたデータベースであり、2013 年より公開されている。用例データベースは日本語の複合動詞と、それを構成する動詞との意味的な関係を分析するための基礎データの構築することを目的とする。山口(2019)によれば、データベースは「動詞連用形＋動詞」型の 3,371 [1] の語彙的複合動詞を含み、それぞれの複合動詞について表記、読み、語構成や用例等の情報が付加されている。用例収集の際には、JUMAN version 6.0 を用いて形態素解析し、構文解析・格解析には KNP version 3.01 を用いているが、収集された複合動詞が語彙的複合動詞であるかどうかは、影山(1993)による代動詞「そうする」による置換テストを用い(山口 2019:19-18)、最終的には手作業で選別している(山口 2019:22)。

『複合動詞レキシコン』 [2] は国立国語研究所が作成・公開しているデータベースである。データベースには「現在の日本語でよく使われる」、2,759 語の「動詞＋動詞」型複合動詞が含まれ、古語・古典語、特殊な専門分野や文学作品に出現が限られ一般性がない語彙は除外されている(ホームページより)。収録されているのは「動詞連用形＋動詞」型のみで、「読んでみる、止めておく」のように前項動詞がテ形のものは含まれない。その一方で、テ形を含む「見て取る」のような意味が慣習化した語彙は収録されている。データベースには語彙的複合動詞だけが収録されており、「～始める、～かける、～忘れる、～合う」

[1] オンライン版のデータベースには、3,757 語が収録されている。
2 https://vvlexicon.ninjal.ac.jp/

のような統語的複合動詞は除かれている。複合動詞が語彙的か統語的かの区別については、データベースのマニュアルに影山(1993)に基づく詳細な記述(3.3 節参照)がみられる。データベースには、意味的・文法的情報も付与されている。

『複合動詞レキシコン』と『Web データに基づく複合動詞用例データベース』は語彙的複合動詞のみを含むが、野田の（2013）データベースには複合動詞の語彙的・統語的区別についての言及はない(3.3 節参照)。

3. 複合動詞検索システムの構築

3.1 用例検索システムの設計と実装

本研究では、統語的複合動詞を含めた複合動詞の用例を網羅的に検索するため、2 つの方策でツールを準備する。

- 複合動詞データベースの用例を統合し、これらに登録されている複合動詞をもれなく検索するためのプログラムを作成する。
- 統語的な複合動詞については、動詞連用形＋動詞というパターンで検索をおこない、上記複合動詞データベースの項目を差し引いた項目を「可能性のある」複合動詞の候補としてコーパスから取得する。そのうえで、V2 動詞(姫野によれば 30、影山は 27 を挙げている)をリストアップし、検索された用例から統語的な複合動詞を取り出す。

上記複合動詞データベースや統語的複合動詞のパターンに含まれない、未知の複合動詞を検索するためのアルゴリズムとして、本研究では以下の方法をもちいて検討する。

- 複合動詞データベースの用例の UniDic による解析結果を参照し、誤解析や 2 語彙素以上の要素として解析されているデータをもちいて、(1)データベースに含まれず、(2)動詞連用形 + 動詞という連鎖としても解析されていないパターンを検索するためのアルゴリズムを検討し実際に検索を試みる。

検索システムの構築にあたり、本研究では UniDic (伝ほか 2007)による形態素解析データをベースとすることにした。UniDic には、登録されている複合動詞は 1 形態素(短単位)で収録され、「動詞-一般」という一般的な品詞が付与され、従って複合動詞としてのマーカーがない。山口(2019)では、このことを踏まえて、UniDic ではなく Juman と KNP を解析ツールとして使用している。本研究では、古典作品・近現代むけの UniDic などが開発されている現状に合わせ、UniDic の解析を積極的に使用しながら、複合動詞の処理に必要な情報を追加付与する形でツールを整備することにした。

3 つの複合動詞のデータベースに含まれる複合動詞は、一つのリストにまとめ、UniDic で解析をおこなって、検索のもとになる語彙データベースを作った。その結果、上述のとおり、1 語彙素としての複合動詞、2 語彙素以上に分けて解析される「動詞複合」、動詞+動詞として解析されないもの、さらには誤解析されるものなど、解析結果はさまざまであった。

特に問題となるのが、統合した複合動詞データベースを UniDic で解析したところ、表記によって誤解析されることである。例えば、複合動詞「有り余す」の前項要素「あり～」

の品詞は UniDic によって「名詞-普通名詞-一般」として解析され、全体は 2 語彙素となる。一方、前項動詞がひらがなで書かれている場合(「あり余る」)、品詞は「動詞-一般」となり、1 語彙素として解析される。

さらに、前項要素が誤解析される例もある。例えば、「きっ立つ」は「切り立つ」の前項要素が音変化したものであるが、前項要素「きっ～」は（副詞）と誤解析される。本研究では、これらの解析違いのパターンを「そのまま」用例解析に活用することで、未知の複合動詞を含めた検索を試みることにした。

3.2 複合動詞データベースの統合について

2 節で紹介した 3 つの複合動詞データベースに登録されている語彙について、以下のように一部修正をおこなった。

1. 受け身形で登録されている複合動詞は基本形で統一した。

   思いやられる、焼け出される、打ちひしがれる(2 語彙素→1 語彙素)、並び称される(3→2 語彙素：並び+称する)

2. 否定形で登録されている複合動詞は肯定形で統一した。

   煮え切らない、数え切れない

また、野田の複合動詞リストには、いくつか語彙の重複があった(「使い分ける」「見出す」)ので修正をおこなっている。同様に、「抱きしめる」「抱き合う」「注ぎ込む」は前項動詞の読みの違いで異なる場所に 2 つリストされているが、本研究では UniDic で同一の語彙素に解析されることをふまえ 1 つの項目にまとめた。

3.3 統語的複合動詞の選定について

影山（1993:96）によると、統語的複合動詞の後項動詞は限られており、統語的複合動詞を形成すると思われる後項動詞は以下の 27 語である。

表 1：統語的複合動詞の後項動詞(27 語)

始動：～かける、～だす、～始める
継続：～まくる、～続ける
完了：～終える、～終わる、～尽くす、～きる、～通す、～抜く
未遂：～そこなう、～損じる、～そびれる、～かねる、～遅れる、～忘れる、～残す、～誤る、～あぐねる
過剰行為：～過ぎる
再試行：～直す
習慣：～つける、～慣れる、～飽きる

相互行為：～合う
可能：～得る

姫野（2018:20-21）は、上記の影山（1993：96）の統語的複合動詞 27 語にさらに 3 語を追加している(始動「～かかる」、未遂「～損ねる」、完了「～果てる」)。

統語的複合動詞の判断基準として、影山(1993)は以下の 5 つを提案している。

a. 代用形「そうする」(影山 1993:80)

「そうする」は意図的な行為を表すから、意味的に排除され、統語的複合動詞の場合は、前項動詞を「そうする」で代用しても全く問題が生じない。例えば「調べ終える」は、「そうし終える」のように適切に置き換えることができる。
語彙的複合動詞の場合、例えば「遊び暮らす」の前項動詞の代わりに「そうする」を置換することはできず、「*そうし暮らす」のような不適切な表現になる。

b．主語尊敬語(影山 1993：83-84)

統語的な複合動詞のもう一つの特徴として、前項動詞では主語尊敬語が可能である。
（例）「歌い始める」は、「お歌いになり始める」のように可能である。
一方、語彙的複合動詞では、「書き込む」の前項動詞を主語尊敬にすることは不可能である(「*お書きになり込む」)。

c．受身形

複合動詞の前項動詞では受身形で活用することには、語彙的複合動詞には（「書き込む」vs.「*書かれ込む」）制限がある。それに対して、統語的複合動詞の場合は、制限がなく、「名前が呼ばれ始めた」のように前項動詞では受身形が可能である。ただし、統語的複合動詞がすべての前項動詞に受身形を許すわけではない（影山 1993:87）。なお、姫野（2018：20）は使役形も受身形と同様に考えることができる（「書かせ始める）と指摘している。

d．サ変動詞(影山 1993:88)

サ変動詞とは、「雑談する、テストする、立ち読みする」のような動詞類であり、語彙的複合動詞の前項動詞を同義的なサ変動詞と置き換えることはできない。統語的複合動詞の場合は、前項動詞としてサ変動詞に自由に置き換えることが可能である。

e. 重複構文(影山 1993:91)

統語的複合動詞は前項動詞に動詞重複を許すことができるが、語彙的複合動詞はできない。ただし、統語的複後動詞であっても、意味的な制限があれば不適格になる。
（例）大臣はそれをひた隠しに隠し続けた。

＊行方不明の子供を探しに探し歩いた。

以上、統語的複合動詞は、影山（1993）が指摘している前項動詞の 5 つの文法的な特徴を使って判定することができる。これらの特徴のいずれか一つでももっている場合には、その複合動詞は統語的であると判断できる。逆に言えば、語彙的複合動詞には、これら統語的複合動詞の文法特徴は見られず、また意味が透明でないなどの意味的な制限がかっていることになる。

本研究では、原則としてコーパスから取得された複合動詞の候補のうち、上記 30 の後項動詞を含む用例について、3 つのデータベースのいずれかに記載があれば「語彙的」、なければ「統語的」複合動詞として分類する。(実際には、語彙的複合動詞として分類される用例には、統語的な動詞としての解釈がふさわしいものがある(詳細は後述する)が、本研究では、この分析はおこなわない。)

一部の複合動詞は解釈により統語的・語彙的複合動詞の両方の特徴をもつと考えることができる。例えば、「～出す、～かける、～合う」のような複合動詞の後項要素は、語彙的複合動詞にも統語的複合動詞にも見える（国立国語研究所 2015, 姫野 2018:83-172）。本研究では、原則として複合動詞データベースに(前項動詞つきで)登録されている動詞はいったん「語彙的」複合動詞としてカウントし、統語的な複合動詞かどうかのチェックをおこなう追加分析を別におこなうことにする。

なお、野田のリストには、以下のような、統語的な複合動詞と考えられるものが含まれている。以下は前項動詞「する」の例であるが、下線の複合動詞は他の 2 つのデータベースには含まれておらず、意味が規則的であることから統語的複合動詞と見るのが妥当である。

> 仕上がる、仕上げる、し終える、し掛かる、し掛ける、しかねる、し損なう、し損じる、し出す、し尽くす、し付ける、し遂げる、し直す、し慣れる、し残す

なお、野田はリストの作成にあたり、語彙的・統語的の区別をおこなっていない。

3.4 語彙データベースに含まれない複合動詞の検出

データベースに含まれない複合動詞については、検索対象となるコーパスを UniDic で解析したデータから、以下のような手順で「未知の複合動詞」候補の用例の抽出をこころみる。

- 統合した複合動詞データベースや統語的複合動詞のリストに含まれない「動詞連用形＋動詞」
- 前項動詞が動詞として解析されないケース：

  3.1 節で述べたように、UniDic で複合動詞データベースに登録されている複合動詞を解析すると、複合動詞として 1 語彙素で検出されるもののほか、2 語彙素以上に分けて解析されたり、動詞+動詞以外の語彙素の組み合わせとして解析されるもの、さらには誤解析されるものがある。

  そこで、2 形態素の前項動詞が名詞として解析されるものは前項要素として動詞も、また動詞として解析されるものは名詞も検索候補として加える。後者により、

「連用形＋動詞」のような一般的なパターンでは検索されない複合動詞を検出することができる。

- 上記と同様に、UniDicの解析間違いがおこなっている前項要素について、後項要素として、データベースに共起している動詞以外が出現するパターンも検索する。以下の例では前項動詞が動詞ではなく記号、副詞や代名詞として解析されている。

表 2：前項動詞が UniDic で動詞以外の品詞で解析される複合動詞候補の例

V1 書字形	V1 品詞	V2 書字形	V2 品詞	V1 語彙素	V2 語彙素
おん	記号-一般	出す	動詞-非自立可能	オン	出す
きっ	副詞	立つ	動詞-一般	きっ	立つ
ずり	副詞	下ろす	動詞-一般	ずり	下ろす
ずり	副詞	出る	動詞-一般	ずり	出る
ずり	副詞	落とす	動詞-一般	ずり	落とす
のっ	副詞	掛かる	動詞-非自立可能	のっ	掛かる
ぶち	副詞	飛ばす	動詞-一般	ぶち	飛ばす
ぶっ	副詞	千切る	動詞-一般	ぶっ	ちぎる
ぶっ	副詞	通す	動詞-非自立可能	ぶっ	通す
われ	代名詞	返る	動詞-一般	我	返る
息せき	副詞	切る	動詞-非自立可能	息急き	切る

3.5 検索システムによる用例検索の手順

作成した複合動詞データベースをもとに、コーパスから以下の手順で用例を抽出する。

- 検索対象とするコーパスを UniDic をもちいて解析しておく(4 節参照)。
- 統合したデータベースの解析結果を使ってパターンマッチをおこない複合動詞の候補を含む用例を抽出する。その際、1 語彙素、2 語彙素、3 語彙素以上の複合動詞データをそれぞれ検索するほか、動詞連用形+動詞のパターンも検索する。3.4 節で示した候補となるパターンも検索する。
- 検出された複合動詞を含む用例は、UniDic の出力結果にコーパス名と出現位置(語彙素番号)、文内の位置(語彙素番号)を記し、複合動詞にあたる語彙素にフラグをつけた形で出力する。UniDic の解析済みデータの形で出力しておけば、その後 KWIC コンコーダンスの形式などに加工することができる。

4. 谷崎潤一郎作品からの複合動詞の抽出実験

4.1 谷崎作品について

上記で述べた複合動詞検索システムを用いて、現在複合動詞の用例を収集している文学

作品を実験データとして、複合動詞の抽出をおこなってみる。今回対象とする谷崎潤一郎(1886-1965)とその作品群について述べる。

谷崎は明治期末から亡くなる昭和 40 年までの非常に長い期間、執筆活動をおこなっている。その間発表された作品における言語表現には、分析において問題になる要素がいくつかある。まず一つは、カタカナ表記される文章を含むことである。例えば、谷崎潤一郎の作品「鍵」は、日記のスタイルで執筆し、妻が書いている部分は平仮名で、夫が書いている部分は片仮名で書き分けられる。それにより、後者の文章に含まれる複合動詞は、「引キ留メタ」「坐リ込ンダ」のように書かれており、一般に用例収集を困難にしている。

現在までに、用例を手作業で収集した谷崎の小説『刺青』『少年』『春琴抄』のほか、以下の作品について、作品テキストをコーパスとして複合動詞の自動検索をおこなう。

表 3：検索する谷崎潤一郎作品の書籍情報

作品名	作品名読み	発表年	仮名
刺青	しせい	明治 43（1910）	新字新仮名
少年	しょうねん	明治 44（1911）	新字新仮名
春琴抄	しゅんきんしょう	昭和 8（1933）	新字新仮名
猫と庄造と二人の女	ねことしょうぞうとふたりのおんな	昭和 11（1936）	新字新仮名
細雪　上巻	ささめゆき	昭和 18（1943）	新字新仮名
細雪　中巻		昭和 22（1947)	新字新仮名
細雪　下巻		昭和 22（1947)～ 昭和 23（1948）	新字新仮名
鍵	かぎ	昭和 31（1956）	新字新仮名
瘋癲老人日記	ふうてんろうじんにっき	昭和 36（1961）	新字新仮名

4.2 用例抽出の準備：テキストの前処理と UniDic による解析

今回作業するのコーパスの元データとなる本文テキストは『青空文庫』より取得した。青空文庫のフォーマットから本文以外の要素を取り除き、電子化にあたって付与されているルビは削除し、コメントとして記述されている特殊文字を以下のように Unicode 文字に変換した(以下は Python のコードの一部である)。

```
line = re.sub(r'｜([^《]+)《[^》]+》',r'¥1',line)
line = re.sub(r'《[^》]+》','',line)
line = re.sub(r'※［＃コト、1-2-24］','ㄱ',line)
line = re.sub(r'※［＃二の字点、1-2-22］','ゝ',line)
line = re.sub(r'※［＃「言＋墟のつくり」、第 4 水準 2-88-74］','譟',line)
line = re.sub(r'※［＃「插」でつくりの縦棒が下に突き抜けている、第 4 水準 2-13-28］','挿',line)
```

```
line = re.sub(r'※［＃「魚＋鑞のつくり」、第4水準2-93-92］','鱶',line)
line = re.sub(r'※［＃「奚＋隹」、第3水準1-93-66］','雞',line)
line = re.sub(r'※［＃「王＋干」、第3水準1-87-83］','玕',line)
line = re.sub(r'※［＃「勹＜夕」、第3水準1-14-76］','匆',line)
line = re.sub(r'※［＃「足へん＋宛」、第3水準1-92-36］','踠',line)
line = re.sub(r'※［＃「倫のつくり／糸」、第3水準1-90-24］','纍',line)
line = re.sub(r'※［＃「口＋它」、第3水準1-14-88］','咤',line)
line = re.sub(r'※［＃「口＋穢のつくり」、第3水準1-15-21］','噦',line)
line = re.sub(r'※［＃「兵」の「丘」に代えて「白」、第3水準1-14-51］','皐',line)
line = re.sub(r'※［＃「りっしんべん＋刀」、第3水準1-84-38］','忉',line)
line = re.sub(r'※［＃トモ、38-8］','ドモ',line)
line = re.sub(r'※［＃トキ、72-13］','トキ',line)
line = re.sub(r'※［＃「口＋云」、第3水準1-14-87］','呍',line)
line = re.sub(r'※［＃「藹」の「言」に代えて「月」、第3水準1-91-26］','腸',line)
line = re.sub(r'※［＃「さんずい＋鬲」、U+6EC6、383-4］','滆',line)
line = re.sub(r'※［＃「ころもへん＋施のつくり」、第3水準1-91-72］','袘',line)
line = re.sub(r'※［＃「足へん＋母」、U+27FF9、63-14］','跇',line)
line = re.sub(r'※［＃トキ、97-2］','トキ',line)
line = re.sub(r'※［＃「土へん＋敦」、第3水準1-15-63］','墩',line)
line = re.sub(r'※［＃「くさかんむり／奥」、第4水準2-86-89］','薁',line)
line = re.sub(r'※［＃「日＋麗」、第4水準2-14-21］','曬',line)
line = re.sub(r'※［＃「齒＋乞」、第4水準2-94-76］','齕',line)
line = re.sub(r'※［＃トキ、33-6］','トキ',line)
line = re.sub(r'※［＃トキ、146-15］','トキ',line)
line = re.sub(r'※［＃トキ、36-12］','トキ',line)
line = re.sub(r'※［＃トキ、37-2］','トキ',line)
line = re.sub(r'※［＃トキ、36-7］','トキ',line)
line = re.sub(r'※［＃「くさかんむり／嬰」、第4水準2-87-16］','蘡',line)
line = re.sub(r'※［＃トキ、43-3］','トキ',line)
line = re.sub(r'※［＃「てへん＋僉」、第3水準1-84-94］','撿',line)
line = re.sub(r'※［＃トキ、75-11］','トキ',line)
line = re.sub(r'※［＃「てへん＋峯」、第4水準2-78-12］','搓',line)
line = re.sub(r'※［＃トモ、38-11］','ドモ',line)
line = re.sub(r'※［＃「てへん＋鼻のへん」、第4水準2-13-55］','擤',line)
line = re.sub(r'※［＃「くさかんむり／生」、U+82FC、19-4］','茋',line)
line = re.sub(r'※［＃「飮のへん＋稻のつくり」、第4水準2-92-68］','餡',line)
line = re.sub(r'※［＃「肄のへん＋欠」、第3水準1-86-31］','欬',line)
```

このようにして準備した文献ファイルを UniDic により解析した。解析辞書には、現代語辞書を用いた。(UniDic で公開されている近現代口語小説、旧仮名口語といった古い文献用の電子辞書による解析結果の比較は稿を改めておこないたい。)

4.3 抽出結果 (1) 統合リストを用いた用例検索

統合した既存の複合動詞データベース 3 件(野田 2013, 山口 2013, 国立国語研究 2015)のデータの比較結果は以下の通りである。

- 3 つのデータベース全てに含まれる複合動詞：1447 語
- 野田（2013）のデータベースにしか登録されていない複合動詞：571 語
- 山口（2013）のデータベースにしか登録されていない複合動詞：1432 語

- 国立国語研究所（2015）にしか登録されていない複合動詞：558 語

その他に、2 つのデータベースにのみ収録されている複合動詞もある。以下にまとめて図で示す。

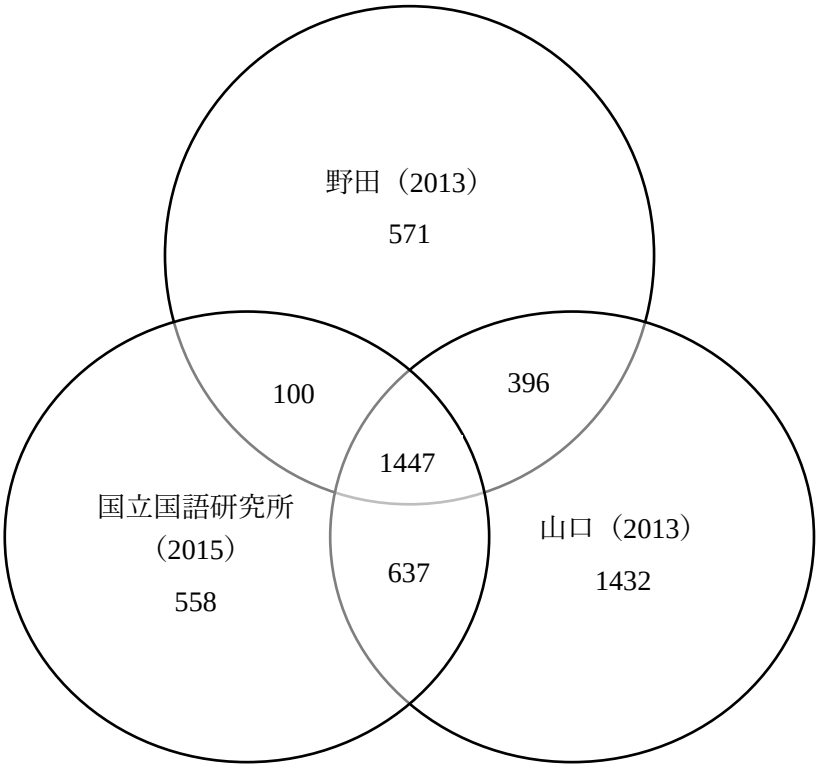

図 1 データベースの収録語彙の比較結果

統合したデータベースを UniDic で解析した結果、統合したデータベースでは UniDic で 1 語彙素として登録されている複合動詞の合計は 7,873 であり、登録されていない複合動詞の合計は 1,312 であった。

表 4：統合データベースに収録された複合動詞の UniDic の登録状況

	国立国語研究所 (2015) [3]	山口(2013)	野田(2013)	総計
UniDic に語彙登録あり	2,456	3,264	2,153	7,873
UniDic では 1 語彙素でない	303	648	361	1,312
総計	2,759	3,912	2,514	9,185

このようにして解析された「複合動詞統合データベース」の複合動詞を用いて谷崎コーパスを検索した結果、1 語彙素の複合動詞として 2,852 例（異なり 847 種類）、2 語彙素から

[3] 表 4 の国立国語研究所(2015)のデータは、「選り抜く」のような読みが異なる複合動詞を分けて数えている。図 1 では読みの違いを相殺しているため、合計は 2742 となる。

なる複合動詞として 278 例(85 種類)の用例が検索された。また、2 語彙素で解析された結果の前項動詞について、データベースの解析結果(名詞+動詞)とは異なる品詞(動詞+動詞)で検索したところ、「落ち着き払う」が 1 例検索された。

表 5　UniDic による「落ち着き払う」の解析例(句点を省略している)

	書字形	語彙素	品詞	活用形	語種
データベースの解析	落ち着き	落ち着き	名詞-普通名詞-一般		和
	払う	払う	動詞-一般	終止形-一般	和
コーパスの解析	落ち着き	落ち着く	動詞-一般	連用形-一般	和
	払っ	払う	動詞-一般	連用形-促音便	和

(例) と、妙子がことさら落ち着き払った口調で云った。(『細雪』下巻: 37649-37661)

このような用例は、UniDic でデータベースの動詞を解析した結果をそのまま用いて検索すると取得できないことになる。同様に、谷崎の作品における「跳び出す」は 2 語彙素で「飛ぶ」と「出す」として、いっぽうデータベースに収録されている「飛び出す」は 1 語彙素で解析される。「着替える」も、谷崎作品では書字形が「着換える」として入っているが、この書字形は UniDic では 2 形態素となる。いっぽう、データベースの「着替える」は 1 語彙素で解析される。

さらに、今までマニュアルで作成した谷崎潤一郎作品の複合動詞データベース（『刺青（しせい）』『少年』『春琴抄』と『細雪上巻（約半分）』）を比べると、「複合動詞統合データベース」に入っていない複合動詞が見つかった。

例えば『刺青』に載せられている「軋み合う、誇り合う、疼き出す、薄らぎ初める、暮れかかる」のような複合動詞は 3 つのデータベースのいずれにも入っていない。それら 5 語の後項要素は統語的複合動詞の候補であるので、データベースに掲載されていないのは自然である。

このような例文も網羅的に検索するため、本研究では、「動詞連用形+動詞」という品詞パターンでの検索もおこない、用例の補完を試みている(3.3 節参照)。しかし、これらのうち、この補完的検索から得られたのは「誇り合う」「疼き出す」「薄らぎ初める」「暮れかかる」であった。「軋み合う」は UniDic による解析が今回の想定と異なってしまった。このような解析のゆれをどのようにコントロールし網羅的検索につなげるかが、課題として残った。

いっぽう、影山、姫野が挙げる統語的複合動詞の後項要素(3.3 節参照)にない「生き代わる、死に代わる、結い繞る（ゆいめぐる）」のような複合動詞も、手作業による用例検索で見つかっているが、これらも「複合動詞統合データベース」を使った今回の検索からは得られなかった。

「結い繞る」は統合データベースにはなかったため取得できなかったのであるが、「生き代わる」「死に代わる」が検索できない問題には、3.1 節で述べた原文の送りがなの表記

の問題があることがわかった。以下の UniDic の解析例にみるように、UniDic には語彙素として「生き変わる」「死に変わる」が登録されている。しかし、谷崎作品『刺青』に収録されているのは「生き代る」「死に代る」という異なる漢字表記であり、送りがなも異なっている。結果として、今回の実験では用例としての検索がうまくいかなかったことになる。UniDic は階層的見出しを導入しており、書字形情報をもつことで表記のゆれに柔軟であるという特徴をもつ(伝ほか 2007:108)が、複合動詞の検索において、漢字の異表記が検索において依然として課題であることがわかった。

表 6：UniDic による漢字表記の異なる複合動詞の解析例(句点を省略している)

文境界	書字形	語彙素	品詞	活用形	語種
B	生き	生きる	動詞-一般	連用形-一般	和
I	代わる	変わる	動詞-一般	終止形-一般	和
B	生き変わる	生き変わる	動詞-一般	終止形-一般	和
B	死	死	名詞-普通名詞-一般		漢
I	に	に	助詞-格助詞		和
I	代わる	変わる	動詞-一般	終止形-一般	和
B	死に変わる	死に変わる	動詞-一般	終止形-一般	和

同様の例には「齧り着く」「噛み着く」「食い着く」「飛び着く」のように、現代であれば後項動詞を「付く」と表記するものなどが挙げられる。これらは統合データベースに収録されているものの、漢字表記の関係で検索できず、全て「動詞連用形+動詞」の用例として取得することができた。

なお、興味深いことに、上記表 6 で前項動詞が動詞として解析されている「生き代わる」であるが、「動詞連用形+動詞」による検索によっては該当する用例を取得できなかった。漢字や送りがなの異表記のため、前部要素が動詞ではなく名詞と解析されたことによるもので、このような、文脈によって解析パターンが異なるケースの存在は、網羅的検索のための方策を考える上での重要な課題である。

4.4 抽出結果 (2) 未知の用例の検索の結果

谷崎潤一郎作品における「動詞連用形＋動詞」パターンを用いた検索では、異なり語数で 164 種類(延べ 1.722 語)の後項動詞が見つかった。その 164 種類の複合動詞の後項動詞の中には、統語的複合動詞候補として影山(1993)、姫野(2018)が挙げている 30 語種類中 22 語が含まれていた(述べ 786 語)。

作品から得られた統合的複合動詞候補：

～出す（234），～かける（106），～始める（78），切る（74），～付ける（55），～合う（41），～過ぎる（35），～得る（34），～続ける（33），～かかる（30），～かねる（17），～尽くす（9），～慣れる（9），～直す（8），～通す（7），～終わる（6），～終える（5），～抜く（5），～損なう（5），～そびれる（5），～果てる（5），～誤る（1），～忘れる（1）。

統語的複合動詞を形成する以下の後項動詞は谷崎作品コーパスには見つからなかった：～まくる，～損じる，～遅れる，～残す，～あぐねる，～飽きる，～損ねる

さらに、語彙的複合動詞と考えられる後項動詞として、後項動詞「置く」の例(「(依頼)致し置く」)のような統合データベースに収録されていない動詞が複数見つかっている。

5. 複合動詞検索システムの評価と今後の課題

本研究では、谷崎潤一郎作品からの複合動詞の用例を網羅的に収集する目的で構想された。既存の 3 件の複合動詞データベースを統合して複合動詞リストを作成し、そこに収録された全ての語彙的複合動詞を抽出するとともに、統語的複合動詞を含む、統合データベースに含まれない新たな複合動詞の用例収集も試みた。

統合データベースに含まれる複合動詞の用例の検索には、UniDic による解析を介しておこなった。複合動詞は UniDic において単純動詞と区別されていないため、その検索には工夫が必要である。UniDic の解析内容に準じた検索により、データベースにある多くの複合動詞はその用例が検索できるものの、網羅的な検索には至らないことが示された。本研究でおこなった副次的候補による検索で収集できた例文もあったが、解析内容を完全に網羅した検索には至らなかった。(小説『刺青』の用例データを用いた検索システムの評価については論文末の付録を参照されたい。)

収集された新たな用例のなかには、「～為(な)さる」「～下(くだ)さる」「～やはる」「～致す」「～申す」のような終助詞的要素と考えられる後項動詞も多く見出される。谷崎コーパスから収集された複合動詞の詳細は稿を改めて分析したい。

今後は本研究では実施に至らなかった、UniDic や Sudachi の辞書をカスタマイズし検索に利用する方法、さらに今回試みた、解析の副次的な候補をさらに充実させる手法について、さらに検討を加え、より網羅的な例文の収集にむけとりくんでゆきたい。また、UniDic の品詞体系に含まれない複合動詞のマーカーをどのように解析情報に加えていくかという問題についても検討していきたい。

謝　辞

本研究は、国際交流基金の日本研究フェローシップにより共同研究として実現したものである。記して感謝する。また、ポスター発表の際、有益なコメントをいただいた参加者の皆さまに深く感謝申し上げます。

参考文献

影山太郎（1993）『文法と語形成』ひつじ書房.

伝 康晴, 小木曽 智信, 小椋 秀樹他 (2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22, pp.101-123.

野田時寛（2013）「日本語動詞用法事典について(4)―複合動詞一覧の試み―」『人文研紀要』75, pp.31-62. 中央大学人文科学研究所.

姫野昌子（2018）『新版複合動詞の構造と意味用法』研究社.

野村雅昭・石井正彦（1987）『複合動詞資料集』国立国語研究所.

山口昌也（2019）「『Web データに基づく複合動詞用例データベース』の構築と評価」『国立国語研究所論集』17, pp.15-34.

由本陽子（2013）「動詞＋動詞型の複合動詞」『レキシコンフォーラム No:6』（編：影山太郎）pp.59-78. ひつじ書房.

関連 URL

国立国語研究所（2015）『複合動詞レキシコン』 https://vvlexicon.ninjal.ac.jp/
『Web 茶まめ』 https://chamame.ninjal.ac.jp/
『Web データに基づく複合動詞用例データベース』 https://csd.ninjal.ac.jp/comp/

付録：**谷崎『刺青』(1910)を用いた複合動詞の用例の検証メモ**

谷崎の小説『刺青』(1910)には 54 種類の複合動詞が現れる(うち 3 種類は今回の検索プログラムで新たに抽出された)が、その中には今回使用している 3 つの複合動詞データベースのいずれにも見つからないものもある。以下に複合動詞の用例として手収集したリストをもとに、検索プログラムを用いて収集した用例との差異際をまとめる。

- UniDic には 1 語彙素の動詞として登録されているものの、複合動詞のデータベースには収録されておらず、用例として抽出できなかった複合動詞は以下の 3 つである：打ち倒れる、明け放れる、軋み合う
- データベースと漢字や送り仮名や漢字と仮名の違いがあるもの：

	『刺青』に現れた複合動詞基本形	DB における見出し	UniDic の解析結果	DB に見つからなかった理由	例文として検索できたか	備考	DB に登録されているか
1	喰いしばる	食いしばる（K,N,Y）	1 語彙素「食い縛る」	漢字の違い	1	1 語彙素「食い縛る」	1
2	繰り展げる	繰り広げる（K,N,Y）	1 語彙素「繰り広げる」	漢字の違い	1	2 語彙素の候補として検出「繰り広げる」	1
3	見馴れる	見慣れる（K,N,Y）	2 語彙素「見慣れる」	漢字の違い	1	「見慣れる」2 語彙素の候補として検出	1
4	惹きつける	引きつける（N,Y） 引き付ける（K）	1 語彙素 引き付ける	漢字の違い	1	1 語彙素「引き付ける」	1
5	脹れ上がる	膨れ上がる（K,N,Y）	1 語彙素「膨れ上がる」	漢字の違い	1	「脹れ上る」2 語彙素の候補として検出	1

6	うち込む	打ち込む（K,N,Y） 撃ち込む（N,Y）	1 語彙素「打ち込む」	漢字と仮名の違い	1	2 語彙素の候補として検出「内|込む」	1
7	探りあてる	探り当てる（K,N,Y）	1 語彙素「探り当てる」	漢字と仮名の違い	1	1 語彙素「探り当てる」	1
8	さし置く	差し置く（K,N,Y）	1 語彙素 差し置く	漢字と仮名の違い	1	1 語彙素「差し置く」	1
9	出来上る	出来上がる（K,Y）	1 語彙素「出来上がる」	送り仮名の違い	1	1 語彙素「出来上がる」	1
10	蹈みつける	踏みつける（N,Y） 踏み付ける（K）	1 語彙素「踏み付ける」	漢字と仮名の違い	0	-	1
11	堪えかねる	耐えかねる（N）	2 語彙素「耐え兼ねる」	漢字の違い	0	-	1
12	刺り込む	刺さり込む（Y）	2 語彙素「刺さり込む」	送り仮名の違い	0	-	1
13	見出す	見出す（N） 見出だす（K） 見いだす（Y）	1 語彙素「見出だす」	 送り仮名の違い 送り仮名の違い	1	1 語彙素「見出だす」	1
14	見つめる	見つめる（N,Y） 見詰める（K）	1 語彙素「見詰める」	 漢字と仮名の違い	1	1 語彙素「見詰める」	1
15	待ち構える	待ち構える（K,Y） 待ちかまえる（N）	1 語彙素「待ち構える」	 漢字と仮名の違い	1	1 語彙素「待ち構える」	1

- データベースと語彙の違いが観察されるもの：

	『刺青』に現れた複合動詞基本形	DB における見出し	UniDic の解析結果	DB に見つからなかった理由	例文として検索できたか	備考

1	生まれ出づ	生まれ出る（K,N,Y）	1 語彙素「生まれ出る」	語彙の違い	1	2 語彙素の候補として検出「生まれ出る」
2	なし終える	し終える（N）	2 語彙素「し終える」「為る」「終える」	語彙の違い	1	2 語彙素の候補として検出「なし終える」
3	打ち捨る（うちやる）	打ち捨てる（K,N,Y）	1 語彙素「打ち捨てる」	語彙の違い	0	-
4	打ち倒れる	ぶっ倒れる（N）	1 語彙素「打っ倒れる」	語彙の違い	0	-
5	縛(ゆ)いつける	縛(しば)り付ける（K,N,Y）cf. 結い付ける	1 語彙素「縛り付ける」	語彙の違い	0	-

- データベースに入っていないもの：

	『刺青』に現れた複合動詞基本形	例文として検索できたか	候補として検出	備考	DB に登録されているか
1	具え始める	1	2 語彙素「具え始める」	-	-
2	誇り合う	1	2 語彙素「誇り合う」	-	-
3	薄らぎ初める	1	2 語彙素「薄らぎ始める」	-	-
4	暮れかかる（暮れかゝった）	1	2 語彙素「暮れ|掛かる」	-	-
5	躍りたつ	1	2 語彙素「躍り経つ」	-	-
6	疼き出す	1	2 語彙素「疼き出す」	-	-
7	死に代わる	1	3 語彙素「死|に|変わる」	-	-
8	生き代わる	0	-	-	-
9	得堪える	0	-	-	-
10	明け放(はな)れる	0	-	UniDic「明け離れる」	-
11	軋み合う	0	-	UniDic にあり	-

- データベースに入っているもの：

	『刺青』に現れた複合動詞基本形	DB における見出し	UniDic の解析結果	例文として検索できたか	備考	DB に登録されているか
1	引き出す	引き出す（K,N,Y）	1 語彙素	1	-	1
2	引き入れる	引き入れる（K,N,Y）	1 語彙素	1	-	1
3	引き立てる	引き立てる（K,N,Y）	1 語彙素	1	-	1
4	引っ込める	引っ込める（K,N,Y）	1 語彙素	1	-	1
5	喰いしばる	食いしばる（K,N,Y）	1 語彙素	1	1 語彙素「食い縛る」	1
6	繰り展げる	繰り広げる（K,N,Y）	1 語彙素	1	2 語彙素の候補として検出「繰り広げる」	1
7	繰り返す	繰り返す（K,N,Y）	1 語彙素	1	-	1
8	見つめる	見つめる（N,Y） 見詰める（K）	1 語彙素「見詰める」	1	1 語彙素「見詰める」	1
9	見守る	見守る（K,N,Y）	1 語彙素	1	-	1
10	見出す	見出す（N） 見いだす（Y） 見出だす（K）	1 語彙素「見出だす」	1	1 語彙素「見出だす」	1
11	見馴れる	見慣れる（K,N,Y）	2 語彙素	1	2 語彙素の候補として検出「見慣れる」	1
12	見入る	見入る（K,N,Y）	1 語彙素	1	-	1
13	見つかる	見つかる（Y）	1 語彙素	1	1 語彙素「見付ける」	1
14	さし置く	差し置く（K,N,Y）	1 語彙素	1	1 語彙素「差し置く」	1

15	似通う	似通う（N,Y）	1語彙素	1	-	1
16	持て囃す	持て囃す（Y）	1語彙素	1	-	1
17	惹きつける	引きつける（N,Y） 引き付ける（K）	1語彙素「引き付ける」	1	1語彙素「引き付ける」	1
18	とり出す	取り出す（K,N,Y） 採り出す（Y）	1語彙素 2語彙素	1	1語彙素「取り出す」	1
19	取り出す	取り出す（K,N,Y）	1語彙素	1	-	1
20	出来上る	出来上がる（K,Y）	1語彙素	1	1語彙素「出来上がる」	1
21	掻き立てる	掻き立てる（K,N,Y）	1語彙素	1	-	1
22	うち込む	打ち込む（K,N,Y） 撃ち込む（N,Y）	1語彙素 1語彙素	1	2語彙素の候補として検出「内込む」	1
23	待ち構える	待ちかまえる（N） 待ち構える（K,Y）	1語彙素「待ち構える」	1	--	1
24	探りあてる	探り当てる（K,N,Y）	1語彙素	1	1語彙素「探り当てる」	1
25	注ぎ込む	注ぎ込む（K,N,Y）	1語彙素	1	-	1
26	脹れ上がる	膨れ上がる（K,N,Y）	1語彙素	1	2語彙素の候補として検出「脹れ上る」	1
27	追いかける	追いかける（N,Y）	1語彙素	1	1語彙素「追い掛ける」	1
28	追い返す	追い返す（K,N,Y）	1語彙素	1	-	1
29	通りかかる	通りかかる（N,Y）	1語彙素	1	1語彙素「通り掛かる」	1
30	突き刺す	突き刺す（K,N,Y）	1語彙素	1		1

31	つきのける	突きのける（N）	1語彙素	1	1語彙素「突き除ける」	1
32	のぞき込む	覗き込む（K,N,Y）	1語彙素	1	1語彙素「覗き込む」	1
33	肥え太る	肥え太る（Y）	1語彙素	1	-	1
34	抱きしめる	抱きしめる（N,Y）	1語彙素	1	1語彙素「抱き締める」	1
35	立てかける	立てかける（N）	1語彙素	1	1語彙素「立て掛ける」	1
36	流れ込む	流れ込む（K,N,Y）	1語彙素	1	-	1
37	老い込む	老い込む（K,N,Y）	1語彙素	1	-	1
38	蹈みつける	踏みつける（N,Y） 踏み付ける（K）	1語彙素「踏み付ける」	0	-	1
39	堪えかねる	耐えかねる（N）	2語彙素	0	-	1
40	刺り込む	刺さり込む（Y）	2語彙素	0	-	1

今回の用例検索プログラムで抽出できた複合動詞の用例は、手作業で抽出した用例中86.27% (51例中44例)であった。検索できなかった用例のいくつかは、語彙の書字形のバリエーションをふまえた検索対象の拡張など、今後のプログラムの改善で収集できる見込みであるが、全ての用例の収集には、未だ課題が残っている。

- まとめ(『刺青』に出現する複合動詞の検索結果)

手作業で収集した用例リストにあり	DBにあり	書字形の違いなし	検索できた	25	51
			検索できなかった	0	
		書字形の違いあり	検索できた	12	
			検索できなかった	3	
	DBになし	検索できた		7	
		検索できなかった		4	
なし	DBになし	検索できた		3	3
合計					54

以上