現代日本語のコーパス

—複合動詞の研究に向けて—

ノヘイル・アダム
Adam Nohejl

第35期　日本語・日本文化研修コース

名古屋大学　2016年

目次

第１章　はじめに

どの言語の学習者も、母語にないその言語の特徴を身につけようと苦労するだろう。ある特徴を覚えなければ、ある学習の段階を乗り越えられないだろう。例えば、声調がない言葉の母語話者は中国語を習う際に、中国語の4つの声調を充分練習しなければ、声調だけで区別される言葉を区別できるようにならない。あるいは、SVO の語順の言語の母語話者は日本語を習う際に、SOV の文の作成を充分練習しなければ文を作れない。それは当然であろう。

さらに、言語の基本的な項目以外でも同じような特徴があるだろう。よく知られているのは、英語の put up、move on などの句動詞である。句動詞を使わずにある段階までは済ませられるが、自然な英語ではないし、句動詞がわからなければ、理解力も限られるのである。したがって、英語の教科書では詳しく取り上げられている。

日本語の上記のような特徴の一つは複合動詞であろう。通常学習者は確かに、初級で「～すぎる」（動詞の連用形＋すぎる）、中級で「～はじめる」、「～おわる」[1]、上級で「～える」（「～えない」）「～かねる」（「～かねない」）などというパターンを学習するだろう。しかし、普通の教科書は複合動詞のいくつかの基本的なパターン以外は取り上げていない。さらに筆者の経験では、日本語では様々な複合動詞がよく使われていると気づいても、自分で適切に使えるようになるのはなかなか難しい。

さて、どのような複合動詞が学習者には大事だろうか。どのような整理や分類が学習者に役に立つだろうか。これらの質問に答えるのに、言語学の理論的研究はいうまでもなく、数量的な分析の必要もあると思われる。

以上から、本稿では、学習者の視点を考慮しながら、日本語の複合動詞を概説した上で、複合動詞の研究のためのコーパスを作る。

本稿の内容は下記のとおりである。

第２章では、日本語の複合動詞の特徴と分類を紹介し、学習者の視点から検討する。

第３章では、まず、コーパスと形態素解析について説明し、検討する。検討を踏まえて、複合動詞のためのコーパスを作る。さらに、作ったコーパスを評価し、抽出した複合動詞の語数と数量的な指標を示す。

第４章では、本稿をまとめ、さらなる調査研究の方向を示す。

第２章　日本語の複合動詞と日本語の学習

2.1.　複合動詞とは

本節では、複合動詞を定義し、以下で使用する用語を導入する。

姫野 （1999: 1）によると、日本語の複合動詞とは二つ以上の語彙的意味をもつ部分（形態素）に分析できる動詞である。姫野（1999: 2）は次の構成パターンを取り上げている。

[1] 例えば、坂野他（1999: 第12課）では「～すぎる」が導入されており、三浦・マグロイン花岡 （2008: 第７課）では「～はじめる」と「～おわる」が導入されている。この同じ出版社の初級教科書と中級教科書では他の複合動詞は導入されていない。

構成パターン	前要素	＋	後要素	＝	複合語（例）
名詞＋動詞	目	＋	さめる	＝	目ざめる
動詞＋動詞	書く	＋	始める	＝	書き始める
形容詞＋動詞	近い	＋	寄る	＝	近寄る
副詞＋動詞	ぶらぶら	＋	下がる	＝	ぶら下がる

本稿では和語の「動詞＋動詞」というパターンの複合動詞を中心に考察する。この場合には前項は必ず動詞の連用形である。

	前項動詞	＋	後項動詞
	（Ｖ１）		（Ｖ２）
形：	連用形		自由

上記の形式の動詞を簡単に**複合動詞**と呼び、その要素を**前項動詞**または**Ｖ１**と**後項動詞**または**Ｖ２**と呼んでおこう。

前項動詞自体が複合動詞である可能性もあるから、三つ以上の形態素から成る複合動詞（例えば「見回し始める」）も含まれている。

2.2. 学習者の視点

第１章で、学習者には自分の母語にない言語の特徴は特に覚えにくいと述べた。では、世界の言語の中には日本語の複合動詞のような動詞があるかどうか見てみよう。

言語	複合動詞の有無	例
ヒンディー語	**有**	nikal jana (出る＋行ってしまう→出てしまう)
		ro padna (泣く＋落ちる→泣き出す)[2]
ドイツ語	**希**	spazieren gehen, drehbohren[3]
	（分離動詞もある。	ab|fahren, aus|gehen）
英語	**希**	stir|fry, kick|start, force|feed
	（句動詞もある。	come up, think over）
フランス語	**無**	
	（接頭辞付き動詞はある。	re|venir, sur|charger）
チェコ語	**無**	
	（接頭辞付き動詞はある。	ob|jevit, pro|myslet）

世界の多くの言語を見たわけではないが、複合動詞は多くの言語にない特徴とはいっていいだろう。ない場合と希にしかない場合には、同じような機能は違う構成の動詞（分離動詞・句動詞・接頭辞付き動詞）によって果たされているようである。

[2] Gauri Pawar さんにご教示いただいた（2016年５月13日）。

[3] Fox（1990）による。

2.3. 複合動詞の種類

本節では、日本語の複合動詞の分類を説明する。姫野（1999: 11）は三つの先行研究における代表的な分類を取り上げている。その三つの分類に結果的には共通点があるが、出発点は違う。三つ目の影山（1993）は生成文法の立場から複合動詞が形成される部門によって次の二種類に分類する（影山 1993: 75）。

A類：語彙的（lexical）複合動詞は語彙部門で形成される。

例：飛び上がる、押し開く、泣き叫ぶ、売り払う、受け継ぐ、解き放す、飛び込む

B類：統語的（syntactic）複合動詞は統語部門で形成される。

例：払い終える、話し終わる、しゃべり続ける、食べすぎる、食べそこなう

影山（1993: 76–79）はまず、日本語の複合動詞は複数の語から成る表現でなく、語であることを確認している[4]。そして、A類の動詞は「典型的な〈語〉の特徴――意味の慣習化と語彙的結合制限――を備えている」のに対し、B類の動詞には語であっても、文や句のような働きがある。

> B類ではV1とV2の意味関係は完全に透明かつ合成的であり、『手紙を書き終える＝手紙を書くことを終える』、『雨が降り始める＝雨が降ることが始まる』（中略）のようにいずれも補文関係として分析できる。（中略）B類のほうは、ちょうど文や句が自由に作られるように、語彙的な制限を受けずに、形成される。
>
> （影山 1993: 78）

それとともに、B類は生産性が高く、意味の制限はあるが、A類の動詞の生産性はV2に大きく依存し、生産性の高い場合でも語彙的な結合制限があると影山（1993: 78）は述べている。

最後に、影山（1993: 88–92）は動詞がB類に属する基準を5つ指摘する。代用形 （V1は「そうする」に変えられる）、主語尊敬語（V1は主語尊敬語に変えられる）、受身形（V1は受身形に変えられる）、サ変動詞（サ変動詞のV1が可能である）、重複構文（例えば、「鍛えに鍛え抜く」は可能である）という基準である。

姫野（1999: 18）によるとB類は結局、寺村（1969）の「付属動詞」や山本（1983, 1984, 1992）の「II類」の一部であるが、影山の分類には明確な区別の基準があるということである。

この基準によって、すべての統語的複合動詞のV2（下記の30語）を挙げることもできる。

始動　　：～かける　～だす　～始める　～かかる
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　～果てる

[4] 複合動詞の内部に統語的な要素が介入できない。例：＊飛び〈も〉上がる、＊食べ〈も〉続ける。さらに、等位構造における削除もできない。例：「この夜、兄は神戸で飲み〈歩き〉、弟は大阪で食べ歩いた。」「ちょうど同じ時に妹は本を読み〈終え〉、姉はレポートを書き終えた。」から〈歩き〉や〈終え〉は削除できない。

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　～損ねる
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 1999: 19）[5]

ただし、複合動詞のＶ２がこのリストに含まれていても、統語的動詞だとは限らない。例えば、「～だす」の場合には始動の意味の統語的動詞（例：動きだす、読みだす、作りだす（＝作り始める））も、方向性などを表す語彙的動詞（例：飛びだす、（警察に）突きだす、作りだす（＝創作する））もある。

統語的複合動詞のＶ２は生産性が高く、語彙的な結合制限はないという特徴を学習者の立場から簡単に言い換えれば、統語的複合動詞のＶ２は頻繁に、割と自由に、様々な結合で使われているということである。したがって、通常教科書で取り上げられている複合動詞のパターンは統語的である。（第１章で挙げた教科書で触れている動詞の例はすべて統語的である。）語彙的複合動詞の中でも学習者に役立つ動詞はないだろうか。

姫野（1999）は「主な後項動詞」として語彙的動詞もいくつか取り上げ、「生産性の高い後項動詞には方向性を表すものが多い」と述べている（姫野 1999: 227）。次に、姫野（1999: 227–236）は、語彙的動詞の後項動詞の意味を、方向性、指向性、接着性・対等接合性、創出・完成、単独性に分類をする。方向性を表す動詞の場合にはＶ２を下記の「４種の方向性を表す後項動詞群」に分け、Ｖ１をその４つのグループとの結合の様相から分類している。

①上方向：～あがる　～あげる
②下方向：～さがる　～さげる　～おりる　～おろす　～おちる　～おとす
③内方向：～こむ　～こめる　～いる　～いれる
④外方向：～でる　～だす　　（姫野 1999: 227）[6]

さらに、方向性を表す動詞全体を「形態変化」と「移動」に分けている。

例：⑴　Ｖ１「燃える」はＶ２「あがる」とＶ２「おちる」とのみ結合でき、形態変化を表すので、上下方向で、内外無関係で、「形態変化」である。
　　⑵　Ｖ１「走る」はＶ２「こむ」とＶ２「でる」とのみ結合でき、移動を表すので、上下方向無関係で、内外で、「移動」である。

[5] 27語は影山（1993: 96）に取り上げられており、「かかる」「損ねる」「果てる」という３語は姫野（1999: 19）によって加えられたものである。

[6] この分類の前にある「『～あがる』、『～あげる』および下降を表す複合動詞類」という第３章（姫野 1999: 35–57）では、さらに「～のぼる」「～くだる」「～くだす」にも言及されているが、造語力がきわめて弱いとされている。

方向性を表す後項動詞は生産性が高く、結合が限定されていても、学習者には割と覚えやすいパターンが多いようだ。本稿では、影山の統語的・語彙的分類と姫野の語彙的動詞を踏まえて論を進める。

2.4. Ｖ２の生産性

２.3節で、先行研究を引用しながら何回も「生産性」という用語を使った。しかし、学習者のために、「生産性が高い」後項動詞を整理する場合、生産性は実際にどうやって比べるのが適切であろうか。

生産性は数学的に定義された量ではないので、唯一の決まった計量や比較の方法はない。以下に三つの先行研究で使用された方法を要約しよう。

姫野（1999: 24–25）は野村・石井（1987）に基づいて、異なり語数の上位30位以内に入っている後項動詞を「主な後項動詞」として記載している。「前項動詞との結合の可能性がすべて網羅されているわけではない」としたうえで、「上位の語ほど生産性が高い（＝造語力が強い）と言える」と述べている。つまり、姫野によると、生産性は数量的に結合できる異なり語数に対応する。

淺尾（2007）は数学的に定義されている指標「Baayen の生産性 $\mathcal{P}$」を日本語の複合動詞のＶ2に適用し、次のとおり説明する。

> Baayen の生産性$\mathcal{P}$（Baayen & Lieber 1991; Baayen 1992）は、ある後項をもつ語の総トークン数を N、そのうちただ一度だけ出現した語（hapax legomena）の数を n_1 として、次の式で求められる：
>
> $$\mathcal{P} = \frac{n_1}{N}$$

Tamaoka et. al.（2004）は直接 productivity（生産性）に言及していないが、情報科学と自然言語処理でよく使われているエントロピーを variety of combinations（結合の多様性）と解釈し、日本語の複合動詞のＶ２に適用している。エントロピー H は次の式で定義されている。

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i$$

あるＶ２に適用すると、n はそのＶ２と結合する異なるＶ１の数で、
p_i はその各Ｖ１との結合の確率である。

このエントロピーの値の実際の解釈は平均情報量である。Ｖ２動詞を０と１のみで最適な符号化をすれば、H は平均符号数を表す。

例えば、分析される資料にあるV2「加える」と結合するＶ１は「付ける」、「書く」、「言う」という３語のみであり、それぞれの延べ語数は「付け加える」４語、「書き加える」２語、「言い加える」２語（合計は８語）であるとすれば、エントロピーの計算は次のとおりになる。

$$p_1 = \frac{4}{8} = \frac{1}{2}, \quad p_2 = \frac{2}{8} = \frac{1}{4}, \quad p_3 = \frac{2}{8} = \frac{1}{4}$$

$$H = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i = -\left(\frac{1}{2} \cdot \log_2 \frac{1}{2} + \frac{1}{4} \cdot \log_2 \frac{1}{4} + \frac{1}{4} \cdot \log_2 \frac{1}{4}\right) =$$

$$= -\left(\frac{1}{2} \cdot (-1) + \frac{1}{4} \cdot (-2) + \frac{1}{4} \cdot (-2)\right) = -(-0.5 - 0.5 - 0.5) =$$

$$= 1.5$$

この場合には、一つの最適な符号化は次のようになる。

付け加える↦0，書き加える↦10，言い加える↦11

符号化すれば、半分の場合（「付け加える」の頻度は½）は1符号（0）が必要で、残りの半分の場合（「書き加える」と「言い加える」の頻度は¼＋¼）は2符号（10または11）が必要である。ゆえに平均符号数は1.5であることになる。

上記のように、エントロピーを複合動詞に適用すれば、次の2つの特徴が見出せる。

(1) 異なるV1の数 n を一定にすれば、エントロピーが最大になるのは、すべてのV1の頻度（延べ語数）が等しいときである。最大値は $\log_2 n$ である。ゆえに、すべてのV1の頻度（延べ語数）が等しければ、異なるV1が多ければ多いほど、エントロピーが高くなる。

(2) 異なるV1の数 n を一定にすれば、エントロピーが最小になるのは、1語のV1の頻度が1に近いときである。言い換えれば、1語のV1を除いて、すべてのV1の頻度が0に近いときである。そのとき、エントロピーは0に近くなる。例えば、3つのV1のそれぞれの延べ語数が1,000語と1語と1語であれば、エントロピーは約0.02になる。

エントロピーもBaayenの生産性も資料かコーパスに基づいている。姫野（1999）も淺尾（2007）もTamaoka et. al.（2004）もそれぞれの指標で主な統語的動詞と語彙的動詞を比較している。いずれの場合でも統語的動詞の一部が指標の値が高いという傾向がある。

姫野（1999: 25）では、合計30語に限られており、上位15位の中では12語が統語的なV2で、30位の中では16語が統語的V2である。

淺尾（2007）はデータとして『CD-毎日新聞 '95 データ集』を利用している。調査は延べ語数が1,000語以上の24語に限られており、上位15位の中では統語的V2が9語あり、下位9位の語の中では統語的V2がない。

Tamaoka et. al.（2004）はデータとして（a）1991年から1994年にかけての毎日新聞と（b）青空文庫[7]を使用している。延べ語数が最小限10語という基準で「語彙的V2の88語と統語的V2の21語」から（a）48語、（b）37語に限定し、エントロピーを計算している。

(a) 統語的37語：0.08〜6.73、平均＝4.38　語彙的11語：0.88〜5.76、平均＝2.97

(b) 統語的29語：2.98〜6.07、平均＝4.86　語彙的8語：1.55〜5.13、平均＝3.72

要約すれば、数量的にはどの基準でも、統語的動詞の生産性が高いことは傾向にすぎない。したがって、学習者のために複合動詞を選択するとしたら、統語的動詞でも、実際の

[7] 青空文庫の全体を使用しているようである。青空文庫については第3章で詳しくとりあげる。

生産性の指標を確かめなければならないと思われる。さらに、Tamaoka et. al.（2004）が指摘しているように、エントロピーによって計ることができた多様性は文体によって異なることも重要な側面だと思われる。（例えば、新聞に対して、小説では「〜得る」という複合動詞は非常に少ないので、エントロピーも低い。）

第３章　複合動詞の研究のためのコーパス

日本語の複合動詞の学習のための指針を作ろうとする場合、どのような研究方法を用いればよいであろうか。まず、このレポートの目的をより正確に説明しよう。学習者及び指導者にとって価値のある指針は単なる単語のリストではないだろう。Garnier & Schmitt（2015）を参考にして考えると単語のリスト以外に、下記の情報が不可欠だと思われる。

(1) 対象語の範囲
　例：複合動詞、統語的複合動詞
(2) 語の出所（実例があれば、実例の出所）
　例：新聞記事、均衡コーパス（BCCWJ）
(3) 語の選択の方法
　例：頻度、生産性
(4) 各語の頻度とリストされた語のすべての語に対する生起の比率
　例：「リストの200語（タイプ）の複合動詞は出所の複合動詞の生起（トークン）の５割占める。」
(5) 多義語の区別（各意味が対象語の範囲に当てはまるか確認する）
　例：「作りだす」という動詞を語彙的複合動詞のリストに含めた場合、「作りだす」には統語的動詞（「今年から作りだした品」）もあると指摘する。さらに、語彙的動詞の二つの意味（「製品を作りだす」と「流行語を作りだす」）を区別する。各意味の頻度、実例を挙げる。

(1)〜(4) はもちろん、(5) も多義語が多い日本語の複合動詞の場合に重要だと思われる。Garnier & Schmitt（2015）は英語の句動詞の学習用リストについて同様の点を指摘している。

不可欠とは言えないが、さらに資料の価値を上げるためにもう一つのポイントを付け加えたい。

(6) 語の分類、あるいは特徴による整理
　例：自動詞と他動詞のペア、動作の動詞・変化の動詞などの意味分類

姫野（1999）は学習者の質問をきっかけとして、複合動詞の分類と用法を考察している。このような幅広い研究は確かに指導者に特に参考になると思われる。したがって、(6) のために姫野の研究も踏まえたい。しかしながら、姫野（1999）で取り上げられている複合動詞は優先順位が示されておらず、範囲が幅広いからこそ、そのままでは学習の指針にならないと思われる。

優先順位として、語の頻度がよく使われている。Garnier & Schmitt（2015）は先行研究を踏まえて、シラバスに複数の語から成る表現を含めるかどうかの決定に最も有意義な基準は頻度だと主張している[8]。なお、語の頻度を調べるにはコーパスが必要である。

次に、コーパスと形態素解析を取り上げる。さらに、本稿で使用するデータとツールについて説明する。本章の最後に、コーパスから抽出した語数と数量的な指標を示す。

3.1. コーパスと形態素解析

以上、語の頻度や形態素の生産性の指標に関して、データ・資料・コーパスを取り上げた。なお、コーパスとはデータや資料より狭い意味であり、以下のように定義される。

> ある言語の研究のために、その言語で実際に用いられた用例を大量に偏りなく収集して電子化し、検索用情報を付加したもの。（前川（編）2013: 2）

前川（編）（2013: 13–19）によれば、コーパスの要件は下記のとおりである。

(1) 代表性　　：ある特定の言語か、その言語の特定の変種を反映する。
(2) 均衡性　　：言語を対象とする場合、主要な変種をカバーする。
(3) 規模　　：大規模であることが望まれるが、偏りがないことを優先する。
(4) 真正性　　：実際に話されたことか書かれたことをそのまま含める。
(5) 電子化　　：コンピュータで検索できる。
(6) 公開　　：有償無償を問わず、公開されている。
(7) アノテーション：最も基本的なものは形態論情報である。

２.4節で、データとして新聞記事を使った先行研究にも言及した。それは「新聞の日本語のコーパス」とは言えるが、学習のための調査の場合、やはり特定の日本語の変種だけでなく、できる限り現代日本語全体を代表しているコーパスを使わなければならない。

本稿の目的に不可欠なのは電子化とアノテーションである。複合動詞の生起の自動的な抽出は形態論情報に基づいている。

形態論情報をプレーンテキストに自動的につける作業は自然言語処理で形態素解析（morphological analysis）と呼ばれる。形態素解析は形態素解析を行うツール（プログラム）と辞書と呼ばれている形態論の統計的モデルによって行なわれる。

日本語の場合には、通常、形態素解析はテキストデータを形態素の列に分割し、次のようにそれぞれの形態素に読み、原形（lemma）、品詞の種類、活用の種類、活用形を付ける。例えば、「日本語を勉強し始めた。」という文は MeCab によって次のとおり解析されている[9]。

---

[8] Garnier & Schmitt（2015）は英語の句動詞を multiword item として捉えるが、日本語の語である複合動詞にも適用できる。さらに生産性が重要だということと矛盾していないと指摘したい。生産性は形態素、例えば、後項動詞の特徴で、頻度は語や multiword item の特徴だからである。

[9] MeCab（バージョン0.996）という形態素解析ツールの出力である。使用した辞書は IPAdic（バージョン2.7.0）で、出力フォーマットは ChaSen である。そのウェブサイトは http://taku910.github.io/mecab/（MeCab）と https://en.osdn.jp/projects/ipadic/（IPAdic）である。

日本語	ニホンゴ	日本語	名詞-一般		
を	ヲ	を	助詞-格助詞-一般		
勉強	ベンキョウ	勉強	名詞-サ変接続		
し	シ	する	動詞-自立	サ変・スル	連用形
始め	ハジメ	始める	動詞-自立	一段	連用形
た	タ	た	助動詞	特殊・タ	基本形
。	。	。	記号-句点		
EOS[10]					

このような形態素解析によって、2.1節で示した複合動詞の形に当てはまる形態素の連続を自動的に検索できる。

3.2. 使用するデータとツール

本稿の条件を満たすものに近い、よく使われているコーパスは『現代日本語書き言葉均衡コーパス（BCCWJ）』（国立国語研究所　2009a）であるが、二つの使用の問題点がある。

第一に、BCCWJ は MeCab という形態素解析ツールと BCCWJ のために開発されたUniDicという辞書で解析が行われる[11]。UniDic の特徴は非常に規模が大きいことであるが、多くの複合動詞は一つの形態素として含まれている。

MeCab（バージョン0.996）と UniDic-MeCab（バージョン2.1.2）を使用したとき、ほとんどの語彙的複合動詞と「Ｖ１+合う」型の統語的複合動詞の生起は、一つの形態素として解析された（例えば、「思い到る」「切りわける」「見合せる」「話し合う」「愛し合う」）。したがって、この形態論情報を使用すると、複合動詞は区別できない。

第二に、BCCWJ は公開されているが、無償ではウェブ・インタフェースで「オンライン版」が使用できるが、ウェブ・インタフェースでは、すべての複合動詞を区別できるように形態素解析をやり直せない。そのためには、有償の「オフライン版」が必要である。

本稿の調査では、経済的な理由で有償の BCCWJ の「オフライン版」を使う代わりに、「青空文庫」からコーパスを作ることにした。青空文庫のウェブサイトによる説明は下記のとおりである。

> 青空文庫は、誰にでもアクセスできる自由な電子本を、図書館のようにインターネット上に集めようとする活動です。著作権の消滅した作品と、「自由に読んでもらってかまわない」とされたものを、テキストとＸＨＴＭＬ（一部はＨＴＭＬ）形式に電子化した上で揃えています。（青空文庫 2014）

上記のように、青空文庫は、著作権の関係で、現代日本語で書かれていない作品が多い。青空文庫の利点として、各作品には詳細な情報（出版年・初出年・日本十進分類法のコー

[10] 「EOS」は文の区切りを示すマークである。

[11] UniDic のウェブサイトは http://pj.ninjal.ac.jp/corpus_center/unidic/ である。

ドなど）があるので、初出年によって作品を選ぶことができる。ただし、出版年は付いているが、初出年は付いていないものもある。以上から、コーパスを作るために、初出年が1945年以降の作品を選択した。

次に、それらの作品を XHTML から形態素解析に適切な形式に変換した。UniDic には上記の問題があるため、形態素解析ツール MeCab（バージョン0.996）を使用し、その辞書として IPAdic（バージョン2.7.0）を使用した。IPAdic は新仮名の現代日本語を対象としているため、さらに選択した作品から旧仮名遣いのものも除かなければならなかった。結局13,258の作品から1,185（約９％、表１参照）を選択した。

	作品数			割合		
	旧仮名	新仮名	合計	旧仮名	新仮名	合計
初出年：不明	2,327	3,379	5,706	17.55%	25.49%	43.04%
初出年：〜1944	2,506	3,540	6,046	18.90%	26.70%	45.60%
初出年：1945〜	321	1,185	1,506	2.42%	8.94%	11.36%
合計	5,154	8,104	13,258	38.87%	61.13%	100%

表１． 青空文庫全体：初出年と仮名遣い

3.3. 本稿のコーパス

3.1節で説明したコーパスの要件を順に確認しよう。

(1) 代表性。現代日本語の書き言葉を対象としており、現代の作品に限定されているのは妥当である。ただし、作品の初出年（図１）を見ると、残念ながら1955年以降の作品が少ない。BCCWJ（国立国語研究所 2009d）と比べると、BCCWJ の出版・書籍サブコーパスと図書館・書籍サブコーパスに対応するものしか含まれていない。つまり、BCCWJ に含まれている新聞、雑誌、白書、ブログなどは含まれていないことになる。

(2) 均衡性。BCCWJ の出版サブコーパスは書籍の日本十進分類法（NDC）に基づくサンプリングによって作られている（国立国語研究所 2009b）。青空文庫にも NDC のコードが記載されているから、NDC のコードによる文字数を BCCWJ 出版・書籍サブコーパス（国立国語研究所 2009d）と比較し、表２のようにまとめた。本稿で使用しているコーパスは文学に偏り（BCCWJ の約19％に対して約82％）、特に産業・技術工学・自然科学の書籍が欠けている。

(3) 規模。国立国語研究所（2009c）を参考にすると、規模は BCCWJ の約一割である（BCCWJ の短単位数には句読点は含まれていないので、形態素解析の出力の単位数と多少異なるが、およその比較はできる）。

BCCWJ（短単位数）	104,911,464
– 出版・書籍サブコーパス（短単位数）	28,348,233
– 図書館・書籍サブコーパス（短単位数）	30,377,866
本稿のコーパス（形態素解析の出力の単位数）	12,566,822

(4) 真正性。入力ミスを除いて、書籍がそのまま含まれている（青空文庫の注釈は削除されている）。
(5) 電子化。コンピュータで検索できる。
(6) 公開。青空文庫は公開されている。
(7) アノテーション。上記のとおり、形態論情報がある。

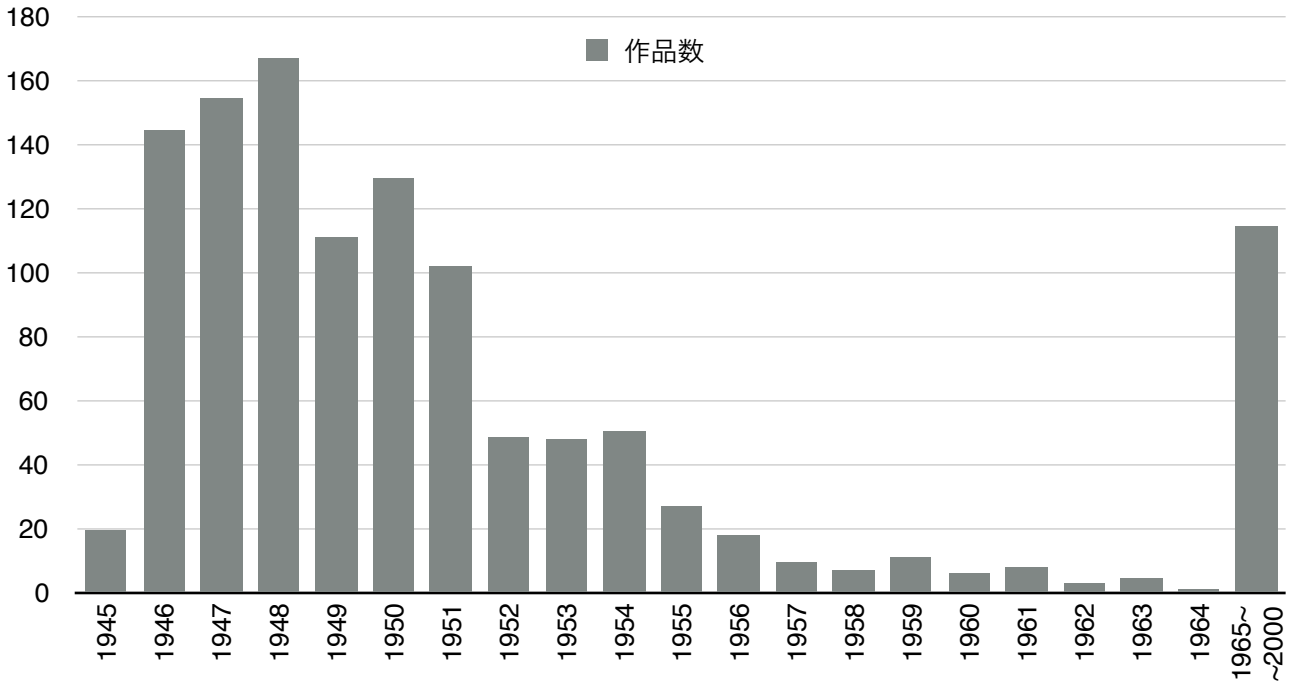

図1．本稿のコーパス：初出年の分布

	BCCWJ出版・書籍 サブコーパスの母集団		本稿のコーパス	
NDC	総文字数	構成比	総文字数	構成比
0. 総記	1,636,414,548	3.37%	9,379,422	5.06%
1. 哲学	2,597,610,813	5.35%	1,390,885	0.75%
2. 歴史	4,301,204,340	8.86%	1,906,327	1.03%
3. 社会科学	12,408,321,943	25.56%	13,588,867	7.32%
4. 自然科学	5,069,594,034	10.44%	1,720,068	0.93%
5. 技術工学	4,615,929,967	9.51%	966,809	0.52%
6. 産業	2,196,387,437	4.53%	127,601	0.07%
7. 芸術	3,258,432,447	6.71%	4,553,471	2.45%
8. 言語	888,800,128	1.83%	423,972	0.23%
9. 文学	9,341,275,486	19.25%	151,477,759	81.64%
n. 記録なし	2,225,954,208	4.59%	0	0%
合計	48,539,925,351	100%	185,535,181	100%

表2．本稿のコーパスとBCCWJ出版・書籍サブコーパスの母集団：均衡性の比較

#	Ｖ２	種類	延べ語数	異なり語数	エントロピー	生産性𝒫
1	出す	S D	3,569	343	6.44	0.037
2	始める	S	2,373	557	7.38	0.122
3	得る	S	2,240	339	5.44	0.087
4	込む	D	1,678	198	6.42	0.038
5	合う	S	1,362	299	6.95	0.106
6	続ける	S	1,016	241	6.60	0.123
7	切る	S	1,011	183	5.91	0.091
8	掛ける	S	951	279	6.70	0.174
9	上げる	D	818	128	6.13	0.049
10	上る	D	774	71	4.57	0.037
11	過ぎる	S	764	199	6.17	0.153
12	付ける	S	620	94	5.22	0.060
13	切れる	S*	604	153	5.91	0.141
13	回る		532	76	4.72	0.058
15	兼ねる	S	480	136	5.77	0.177
16	掛かる	S	442	96	5.31	0.102
17	回す		416	51	3.81	0.046
	以上の会計		19,650	3,443		
	すべてのＶ２（1,094語）の合計		33,054	7,588		

表３．本稿のコーパス：主なＶ２（延べ語数≥400）の延べ語数・異なり語数・エントロピー・生産性𝒫

#	Ｖ１	上方向のＶ２	下方向のＶ２	内方向のＶ２	外方向のＶ２	合計
1	歩く	–	–	5	308	313
2	言う	–	–	1	291	292
3	思う	18	–	4	239	261
4	泣く	–	–	6	187	193
5	笑う	1	–	–	178	179
6	駆ける	48	3	73	54	178
7	考える	–	–	151	17	168
8	引っ張る	24	2	22	98	146
9	起きる	127	–	–	11	138
10	見る	78	20	31	8	137
11	出来る	136	–	–	–	136
12	飛ぶ	89	11	16	10	126
13	流れる	–	18	52	38	108
14	持つ	43	–	–	62	105
15	引く	11	24	23	41	99
	以上の合計	575	78	384	1,542	2,579
	すべてのＶ１（605語）の合計	1,872	577	2,017	3,912	8,378

表４．本稿のコーパス：主な方向性を表すＶ２と結合するＶ１（上位15位）の延べ語数

以上から、青空文庫に基づく本稿のコーパスの最も深刻な問題は、均衡性と代表性だと思われる。例えば、技術関係以外であまり使われていない複合動詞は代表されていない可能性がある。一方で、均衡コーパスである BCCWJ でも技術の書籍は10%以下であるので、主な動詞に関してはあまり影響はないと言えるだろう。ただし、頻度が低い動詞の場合には、注意が必要である。

表３では、本稿のコーパスから抽出した主な複合動詞のＶ２（延べ語数が400語以上）が示されている。統語的複合動詞に使われているＶ２はＳと表示されており、方向性を表すＶ２はＤと表示されている。形態素解析は可能形を区別しないので、多くの場合に「切れる」は実際には「切る」の可能形である。いうまでもなく、それらの動詞のすべての生起が統語的動詞か方向性を表す動詞だというわけではない。それぞれのＶ２の複合動詞の延べ語数・異なり語数・エントロピー・生産性𝒫（2.4節参照）も表示されている。

表４では、主な方向性を表すＶ２と結合するＶ１（上位15位）が示されている。それぞれのＶ１の複合動詞の延べ語数は姫野の方向性を表すＶ２の４つのグループ（2.3節参照）に分けて示してある。

第４章　おわりに

以上、このレポートでは、まず、学習者のための複合動詞の整理や学習の指針を確認した上で、なぜコーパスが必要かを説明した。次に、コーパスの特徴と要件を検討した上で、コーパスを作った。最後に、作ったコーパスを評価し、コーパスから抽出した複合動詞の語数と数量的な指標を示した。

しかしながら、このような情報に基づいて、学習者・指導者にさらに役立つ資料を作るためには、多義語の意味の区別に関する分析（第３章参照）が必要である。この分析を今後の課題としたい。

参考文献

青空文庫（2014）「青空文庫編 青空文庫早わかり」 <http://www.aozora.gr.jp/guide/aozora_bunko_hayawakari.html>（2016年６月26日アクセス）.

淺尾仁彦（2007）「複合語の生産性と文法的性質」『日本言語学会第134回大会予稿集』pp. 416–421. 6月16–17日（麗澤大学）.

影山太郎（1993）『文法と語形成』 ひつじ書房.

国立国語研究所（2009a）『現代日本語書き言葉均衡コーパス（BCCWJ）』 <http://pj.ninjal.ac.jp/corpus_center/bccwj/>（2016年６月26日アクセス）.

国立国語研究所（2009b）『現代日本語書き言葉均衡コーパス（BCCWJ）設計の基本方針』 <http://pj.ninjal.ac.jp/corpus_center/bccwj/basic-design.html>（2016年６月26日アクセス）.

国立国語研究所（2009c）『現代日本語書き言葉均衡コーパス（BCCWJ）DVD版公開データ』 <http://pj.ninjal.ac.jp/corpus_center/bccwj/dvd-index.html>（2016年６月26日アクセス）.

国立国語研究所（2009d）『現代日本語書き言葉均衡コーパス（BCCWJ）サンプリング』 <http://pj.ninjal.ac.jp/corpus_center/bccwj/sampling.html>（2016年６月26日アクセス）.

寺村秀夫（1969）「活用語尾・助動詞・補助動詞とアスペクト―その一―」『日本語・日本文化』 1，pp. 32–48. 大阪外国語大学研究留学生別科.

野村雅昭・石井正彦（1987）『複合動詞資料集』国立国語研究所報告.

坂野永理・大野裕・坂根庸子・品川恭子（1999）『初級日本語げんき（An Integrated Course in Elementary Japanese）』 Japan Times.

姫野昌子（1999）『複合動詞の構造と意味用法』ひつじ書房.

前川喜久雄（編）（2013）『講座日本語コーパス第一巻　コーパス入門』朝倉書店.

三浦昭・マグロイン花岡直美（2008）『中級の日本語【改訂版】(An Integrated Approach to Intermediate Japanese【Revised Edition】)』Japan Times.

山本清隆（1983）「複合語の構造とシンタクス　ソフトウュア文書のための日本語処理の研究―5」情報処理振興事業協会.

山本清隆（1984）「複合動詞の格支配」『都大論究』21，pp. 32–49. 東京都立大学国語国文学会.

山本清隆（1992）「第三部複合動詞辞書　複合動詞結合情報付き動詞辞書作成の試み」『ソフトウェア文書のための日本語処理の研究―11』情報処理振興事業協会技術センター.

Baayen, R. H. (1992) Quantitative aspects of morphological productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of Morphology* 1991, pp. 109–149. Dordrecht: Kluwer.

Baayen, R. H. & Lieber, R. (1991) Productivity and English derivation: a corpus based study. *Linguistics* **29**, pp. 801–843.

Fox, Anthony (1990) *The Structure of German*. Oxford: Clarendon Press.

Garnier, M. & Schmitt, N. (2015) The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research* **19**(6), pp. 645–666.

Tamaoka, K., Lim, H. & Sakai, H. (2004) Entropy and redundancy of Japanese lexical and syntactic compound verbs. *Journal of Quantitative Linguistics* **11**(3), pp. 233–250.