〈全文〉 近代語コーパス設計のための文献言語研究 成果報告書

著者	田中 牧郎，岡島 昭浩，小木曽 智信，小野 正弘，小島 聡子，島田 泰子，朱 京偉，高田 智和，張 元哉，陳 力衛，近藤 明日子，須永 哲矢
ページ	1-260
発行年	2012-10-31
シリーズ	国立国語研究所共同研究報告 ; 12-03
URL	http://doi.org/10.15084/00002759

国立国語研究所
共同研究報告 12-03
ISSN 2185-0127

近代語コーパス設計のための文献言語研究
成果報告書

田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・

朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢

2012 年 10 月

大学共同利用機関法人 人間文化研究機構
国立国語研究所
National Institute for Japanese Language and Linguistics
NINJAL

近代語コーパス設計のための文献言語研究　成果報告書

目　次

本報告書の目的と概要

田中　牧郎　（国立国語研究所言語資源研究系）[1]

１．本報告書の目的

国立国語研究所では、2006 年から『現代日本語書き言葉均衡コーパス』の開発に取り組み 2011 年にこれを完成させたが、2009 年 10 月の大学共同利用機関法人人間文化研究機構への移管を機に、日本語の史的研究に幅広く活用できる通時的なコーパスを構築することにも手を広げることになった。移管に際して始まった、基幹型共同研究プロジェクト「通時コーパスの設計」（プロジェクトリーダー：近藤泰弘客員教授）では、古代から近世までを対象とした「通時コーパス」の設計を行う研究を進めている。一方、ここに研究成果を報告する、独創・発展型共同研究プロジェクト「近代語コーパス設計のための文献言語研究」（プロジェクトリーダー：田中牧郎）においては、明治初年（1868 年）以後を「近代」と扱い、近世までのコーパスと『現代日本語書き言葉均衡コーパス』とをつなげる役割を持った「近代語コーパス」を設計するための論点を整理し、その設計図を描く道筋を付けることを目指して、近代の文献資料とその言語の研究を進めてきた。なお、『現代日本語書き言葉均衡コーパス』の開発以前に、国立国語研究所では最初の近代語のコーパスとして『太陽コーパス』を 2005 年に完成させている。このコーパスの実績を踏まえながらも、最新のコーパス研究の成果を取り込んだ、次世代の近代語コーパスを設計するための研究に主眼を置いてきた。３年のプロジェクト期間が終了するにあたり、その主要な研究成果をまとめて公表するのが、この報告書の目的である。

２．本報告書の概要

２．１　全体の構成

本プロジェクトは、今後構築されるべき近代語コーパスをどのように設計するのかについてと、構築される近代語コーパスをどのように活用するのかについての両側面から、近代の文献資料と言語の研究を行った。本報告書は、その二つの側面に即して「第 1 部 コーパスの設計」と「第 2 部 コーパスの活用」の２部構成とし、共同研究者による論文を集成する形で編集した。第１部では、「資料選定」「文字処理」「形態素解析」「モデルコーパス」の四つに分類し、２編ずつ計８編の論文を掲載した。また、第２部では、「語彙研究」「文法研究」「日中韓対照研究」の三つに分類し、３編ずつ計９編の論文を収録した。各論文の扱っている内容を簡単に紹介しながら、本プロジェクトの成果を概説しよう。

[1] mtanaka@ninjal.ac.jp

２．２　「第１部　コーパスの設計」の概要

第１部の「資料選定」に掲げた２編のうち、田中論文は、近代語コーパスにおける資料選定においては、当時の言語に対する代表性と各資料が有している独自性の両側面が尊重される必要があることを述べ、数千件からなる資料リストに基づいた資料選定の実施例を示し、実際に選定を行う際の論点について研究している。もう１編の岡島論文は、近代語コーパスの対象とすべき資料を探索する演習を、大学院生とともに行った記録を著したもので、院生から提案された資料の特徴とコーパス化する意義を資料ごとに具体的に示している。この２編で扱っているように、今後、近代語コーパスのための資料選定を実際に行う際には、多様な資料の性質を分類整理することと、多くの研究者が参加できる形で幅広い議論を行うことが、重要になるだろう。

「文字処理」には、近代語文献を電子化するために必要な文字セットと、異体字処理について論じる２編を掲載した。文字セットを扱う高田論文は、先行する『太陽コーパス』が、JISX0208（JIS 第１水準・第２水準）で電子化した際に外字となっていた文字が、その後普及した JISX0213 による JIS 第３水準・第４水準でどこまで電子化できるようになったのかを調査し、依然として外字になるものとともに、文字一覧を示している。モデルコーパスとして作成した『明六雑誌コーパス』の文字処理の基準と実際を記す須永論文は、外字をできるだけ減らすために行った、包摂規準の追加と別字による代用の全容を記録している。近代語の資料の多くは活字文献でありながら異体字が非常に多いため、この問題への対処法を明確化しておくことは、コーパス構築の基本問題としてきわめて重要である。この２編で打ち出された方向性は、『太陽』と『明六雑誌』だけにとどまるものではなく、近代語コーパス全体に適用されるべきものであり、さらには、近世以前を扱う「通時コーパス」の設計にも直接役立つものになろう。

「形態素解析」の論文としては、近代語テキストに形態素解析を施す実際の作業とその問題点を述べる小木曽論文と、形態素解析を実現させるのに必要となる形態論情報付与の規程を説明する須永・近藤論文の２編を載せた。それぞれ『太陽コーパス』開発当時は不可能であった近代語テキストへの形態素解析を実用化させるための技術開発を行い、その処理のために必要になる単語や品詞の認定基準を立てる研究である。そこには、自動形態素解析結果に人手修正をかけて形態素解析辞書と機械学習用データを整備していくことで、近代語の形態素解析が十分に実用化可能であることの見通しが明確に示されており、本プロジェクトで最も成果があがった部分である。近代語の言語状況は複雑であるため、今後多くのテキストを対象に人手による作業を重ねることが求められるものの、この技術が確立しつつあることによって、『太陽コーパス』の仕様を大きく進めた、次世代の近代語コーパスを設計できることが確実になったと言ってよい。

「モデルコーパス」においては、明治７（1874）～明治８（1875）年に発行された学術啓蒙雑誌『明六雑誌』の全文を対象とした『明六雑誌コーパス』を作成することを通して、今後構築していく近代語コーパスのモデルを提示する２編を執筆した。コーパスの仕様を扱う、近藤・田中論文は、上述した文字セット、包摂規準、形態素解析をはじめとした、『太

陽コーパス』から大きく発展させた仕様の全体を網羅的に記述したものである。コーパスの概要を記す近藤論文は、形態素解析が実現したことで明らかになった、語種構成・品詞構成をはじめとした、『明六雑誌』の詳細な言語状況が示され、形態論情報付きコーパスが持つ高い価値を印象づけるものになっている。この『明六雑誌コーパス』は、本報告書の公開と同時に、本プロジェクトのホームページを通してダウンロード公開を開始した。コーパス検索ツール『ひまわり』に搭載できるデータも公開することによって、誰でもが容易に利用できるようにした。

２．３　「第２部　コーパスの活用」の概要

よいコーパスを設計するには、コーパスをどのように活用してどのような研究を展開するのかを考えながら研究することが不可欠である。そのような考え方に立ち、本プロジェクトでは、コーパスを活用した新しい研究領域の開拓にも力を入れた。

「語彙研究」に収録した田中論文は、『太陽コーパス』に形態素解析を施したデータを用いて年次別の語彙頻度調査を行い、明治後期から大正期にかけて漢語が減少し和語が増加していく実態を明らかにした上で、個々の語が語彙全体の中に占める位置がどう変わっていくのかという観点から語彙を類型化している。また、小野論文は、『明六雑誌コーパス』から作成した漢語リストをもとに、『日本国語大辞典第２版』の初出時代と比較することで、語史の視点から漢語を階層化する研究である。さらに、近藤論文は『明六雑誌コーパス』から一人称代名詞を抽出し、統計的指標を用いて代名詞一つ一つの性格を詳細に明らかにしている。これら３編はいずれも、従来のテキストコーパスだけでは行えなかった、形態論情報付きコーパスの利点を生かした語彙研究の事例となっている。

「文法研究」にも３編をおさめた。まず田中論文は、やはり『太陽コーパス』の形態素解析データを使って、言文一致が進行するのにともなって、文語助動詞から口語助動詞への推移がどのように進んでいくのかを記述したものである。推移の過程は、個々の助動詞によって異なり、同じ助動詞でも活用形によって異なる事実が種々発見されており、文体変革期における書き言葉の文法変化の記述という、新領域の開拓が期待できる。島田論文は、終止形による準体法が近代語において多様に発展していた事実を、コーパスからの豊富な事例によって明らかにしている。ここでも、品詞や活用形を指定して抽出・検索できるようになった形態素解析データが真価を発揮しており、コーパスが重要テーマでの集中的な議論を可能にする利点が示されている。そして、小島論文は、東北出身の宮沢賢治と濱田廣介の文学作品のコーパスを独自に作成し、標準語のコーパスである『太陽コーパス』と比較して、地方出身者の言語に方言的な特徴がどの程度見られるのかを分析している。標準となるコーパスとは別に特定目的のコーパスを作成して両者を比較する応用的な研究は、現代語のコーパス研究でも行われているが、近代語コーパスにおいてもそのような研究の広がりが見込めることを教えてくれる。

最後に「日中韓対照研究」としてまとめた３編について述べたい。近代に西洋からの新概念を受容するにあたり、多くの漢語が作られたり意味を変えたりしたが、その漢語の変

容は、日本語だけに起こったのではなく、同じ漢字文化圏を形成していた中国語や韓国語にも起こり、相互に語彙の貸し借りを行っていた。この語彙交流の研究は従来から盛んであったが、三つの言語の近代語コーパスを連携して作ることができれば、この方面の研究を一層充実させることができる。そのような考えから、中国語や韓国語の近代語と日本語の近代語とを対照した研究をここにおさめた。朱論文は、日本語の漢文系の複合辞（「～に基づいて」など）とそれに相当する中国語の語彙との関係を、日中双方の近代語コーパスの調査によって明らかにしようとし、陳論文は、日中の語彙交流の記述を、多くの文献資料をもとに行う際の問題点を整理している。いずれも、両言語の多様な文献資料の性質を見きわめた上で比較することの重要性を指摘している。張論文は、近代に日本語から韓国語に訳された文語体の資料と口語体の資料の２組をそれぞれコーパス化し、語種構成や品詞構成を対照し、近代日本語と近代韓国語の比較研究を行っている。これらの研究は、近代語コーパスの構築において、東アジア言語との関わりにも留意することの重要性を示している。

３．今後に向けて

本報告書におさめた 17 編の論文は、本プロジェクトの共同研究発表会、国立国語研究所主催のコーパス日本語学ワークショップ、各種学会の口頭発表などで発表した内容に基づいているものが多いが、学術誌や著書などでは未発表のものばかりである。いずれも、現段階では、各研究者による新領域開拓の途上にあるものであり、この報告書への執筆を経て、さらに研究の段階を進めて、学術誌や著書としてより完成されたものへとまとめられるべきものである。このような性格の論文が集まったことは、コーパス設計のためという目的を共有して研究することで、共同研究者の目が自ずと新領域へと向いていった結果だと考えられる。

近代語コーパスの設計図そのものは、まだ描かれていないが、本報告書の各論文が指し示す方向のすぐ先には、その作業に着手できる場が見えているはずである。近代語コーパスの構築に本格的に着手するには、開発予算や開発体制の検討作業が不可欠であるが、そうした実務的な検討作業に際しても、この報告書が役立てられることを願うものである。

第１部　コーパスの設計

近代語コーパスにおける資料選定の考え方

田中　牧郎（国立国語研究所言語資源研究系）[1]

１．はじめに

国立国語研究所が 2011 年に公開した『現代日本語書き言葉均衡コーパス』は、代表性を担保する周到なサンプリングなされている点において（前川 2008）、これまでの日本語コーパスとは一線を画している。今後構築されるコーパスは、この代表性の担保にどのように対応するかが問われていくことになろう。一方、国立国語研究所では 2009 年から、上代から近世までの日本語の歴史をたどることのできる「通時コーパス」の設計に着手しているが、古典作品を対象とするコーパスでは、ランダムサンプリングを重視する代表性よりも、作品のアイデンティティを重視して資料の独自性を吟味する立場が重要になると見通されている（近藤 2012）。

この「通時コーパス」が対象とする近世までと、『現代日本語書き言葉均衡コーパス』が対象とする現代とをつなぐ位置にある近代における日本語を対象とした「近代語コーパス」の設計を考えるとき、資料選定において「代表性」や「独自性」はどのように考えていけばよいだろうか。この問いについて考えるには、近代語の資料のあり方を分析することを通して研究していくことが必要だろう。本稿では、近代語のコーパスを設計する際の資料選定の考え方を問題にする。

２．『太陽コーパス』から近代語コーパスへ

近代語のコーパスについて、国立国語研究所は既に『太陽コーパス』（国立国語研究所 2005a）を構築して公開している[2]。『太陽コーパス』は、言文一致を経て、口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり、月刊の総合雑誌『太陽』（博文館）の、明治 28（1895）年、明治 34（1901）年、明治 42（1909）年、大正 6（1917）年、大正 14（1925）年の 60 冊分について、その全文（著作権処理ができなかった記事を除く）を対象にしたものである。年次が６年または８年刻みとなっている点はサンプリングコーパスと言えるが、対象になった年次の全体を含んでいる点では全文コーパスとも言える。

コーパスの重要な要件のひとつである代表性の担保については､対象とした総合雑誌『太陽』が、分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さの四点で、当時の文献資料としては格別の価値を持っていることから、『太陽コーパス』にも「代表性」が備わっていると見ることもできる（田中 2005）。実際に例えば、図１は、『太陽コーパス』のジャンル（NDC）別の記事数とその比率を『現代日本語書き言葉均衡コーパス』（出版サブコーパスの書籍、図２）のサンプル数（丸山ほか 2011）と比較できる形で示したものであるが、社会科学が最も多く、文学がこれに次ぐところなど、『現代日本語書き言葉均衡コーパス』（出版サブコーパス書籍）と『太陽コーパス』は似ている面があることが分かるだろう。

しかし、大きく異なっている点として、『現代日本語書き言葉均衡コーパス』が、現代

[1] mtanaka@ninjal.ac.jp

[2] 同種のコーパスに、国立国語研究所『近代女性雑誌コーパス』があり、CD-ROM で公開している（その情報は、http://www.ninjal.ac.jp/corpus_center/）。これは、『太陽コーパス』とほぼ同時期の女性を読者とした３誌（『女学雑誌』『女学世界』『婦人倶楽部』）を対象とした約 120 万語の小規模なコーパスである。

書き言葉の種々の媒体を母集団に設定して、ランダムサンプリングが行われているのに対して、雑誌『太陽』は、そのような手続きを経て選ばれたものではないという点があげられる。むしろ、先に述べた、雑誌『太陽』が持つ、分量・ジャンル・執筆陣・読者層の四点の特徴がこの時期のコーパスの対象としてふさわしいと見た、「独自性」を重視した選定であったと言うこともできる。

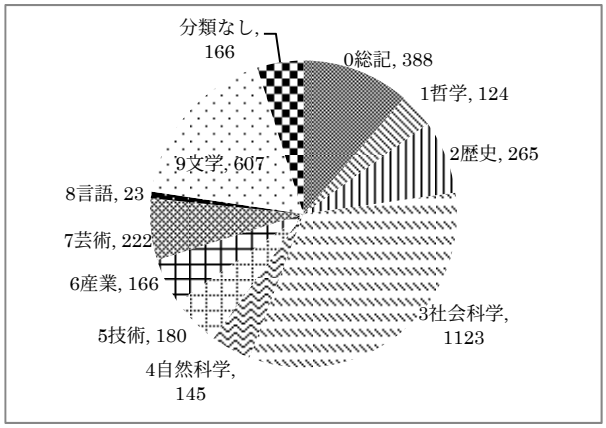


図１『太陽コーパス』のジャンル

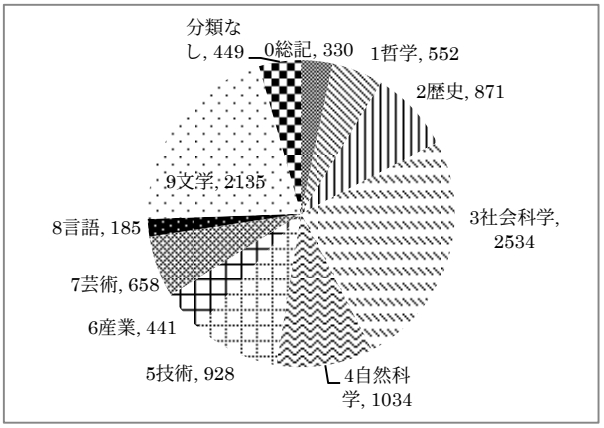


図２BCCWJ 出版サブコーパス書籍のジャンル

古典語と現代語をつなぐ位置にある近代語を対象としたコーパスに含める資料を決めていくには、「代表性」と「独自性」の両面を考慮することが望まれるのではないか。既にある近代語のコーパスとしての『太陽コーパス』を踏まえつつも、多様な近代語の資料の実態を整理した上で、コーパスの資料のあり方を考えていくことが必要である。

３．近代語の資料リストの作成

３．１　「国語辞典編集準備資料」

『太陽コーパス』は、国立国語研究所の史的国語辞典編集事業の系譜から生まれたものである。その史的国語辞典編集を行う準備研究のために設置された国語辞典編集準備室によって、用例採集の対象とすべき近代語資料をまとめた目録が、三つ作成されている。

(1)『用例採集のための主要文学作品目録』（国語辞典編集準備資料２、1980 年）
主要文学全集に収録された、明治元（1868）年～昭和 41（1966）年の 1506 作品をリスト化したもので、有識者 10 名が投票を行い得点化し、高得点の主要作品 139 点が「用語索引を作る作品」として選定されている。
(2)『用例採集のための主要雑誌目録』（国語辞典編集準備資料３、1983 年）
国立国会図書館の和雑誌目録の中から、昭和 25(1950)年以前に創刊され 20 年間以上発行されている雑誌2778件をリスト化したもので、有識者10名が投票を行い得点化し、高得点の主要誌 120 点が選定されている。
(3)『用例採集のためのベストセラー目録』（国語辞典編集準備資料４、1984 年）
ベストセラーに関する参考書に掲載された、明治元（1868）年～昭和 53（1978）年の書籍、1882 件をリスト化したもの。このリストについては得点化や主要作品の選定は、行われていない

実際の史的国語辞典編集のための用例採集事業[3]は紙媒体で開始されたが、すべての用語・用例を採集できるようにする「総索引方式」と、任意の用語・用例を選抜して採集する「スカウト式」の二段構えで着手された。総索引方式では国定国語教科書を対象とした

[3]「日本大語誌」と呼ばれるこの事業の記録は最近、飛田（2012）として公表された。

『国定読本用語総覧』（国立国語研究所 1985-1997 として完成公開）が作成され[4]、スカウト式では雑誌『太陽』の用例採集が進められた。ところが、この事業に本格的にコンピュータが導入されたことがきっかけとなって、『太陽』は途中からスカウト式を止めコーパス化の対象にされ、『太陽コーパス』が作成されたのである[5]。『太陽コーパス』の完成に先立って史的国語辞典編集のための用例採集作業は中断された形になっているが、実質的にはコーパス構築事業にその考え方は継承されており、平成 21 年度から通時コーパスと近代語コーパスの設計に関わるプロジェクトが同時に始まったことで、その側面はより色濃くなってきたと言える。近代語コーパスに含めるべき資料を検討する際に、上記の目録類は第一に参考にすべきものである。

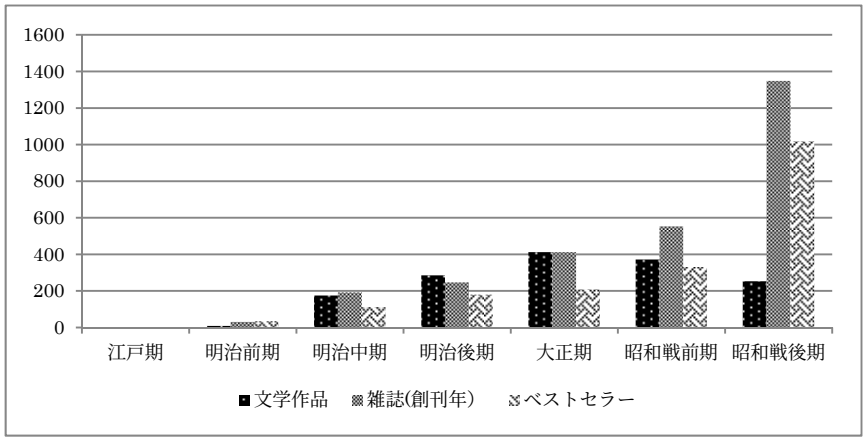


図３　国語辞典編集準備資料に掲載された資料数（時代別）

図３は、上記の三つの目録に掲載された資料の数を時代別にまとめたものである。時代区分は、明治から大正期をほぼ 15 年ごとに４つに区切り、昭和期を戦前と戦後に分けた。

明治前期：明治元～15(1868-1882)年
明治中期：明治 16～30(1883-1897)年
明治後期：明治 31～44(1898-1911)年
大正期：大正元～14(1912-1925)年
昭和戦前期：昭和元～20(1926-1945)年
昭和戦後期：昭和 21(1946)年～

明治・大正期と昭和期とで時間幅が異なっていて比較しにくい面はあるが、雑誌とベストセラーは時代を追って増加傾向にあり、文学作品は大正期まで増加し、昭和期に入って減少していると見ることができよう。こうした傾向はそれぞれの媒体が各時代にどの程度の量発行されたかという実態を反映している面もあるかもしれないが、直接的には目録作成の材料に何が使われたかということを反映しているのではないかと思われる。また、明治前期・中期が全般的に少ないのは、この目録作成が 20 世紀を主たる対象にしていたということも関係しよう。

雑誌とベストセラーは、『現代日本語書き言葉均衡コーパス』でも対象としており、文学作品は『現代日本語書き言葉均衡コーパス』では書籍の下位に NDC 分類に即して配置されている。『現代日本語書き言葉均衡コーパス』にはこのほか、新聞、教科書、白書、広報

[4]教科書については資料目録は作成されていない。国定読本の他には国定算数教科書の用語索引が作られたが、公開されてはいない（木村・加藤・田中 1999）。

[5]この間の経緯は、木村・加藤・田中（1999）参照。

誌、Yahoo!知恵袋、Yahoo!ブログ、法律、国会会議録などが含まれている。このうち、新聞、教科書、国会会議録などは、史的国語辞典編集のための資料目録作成は行われていないが、用例採集作業の対象として研究は行われており、対象資料の候補にはなっていた。一方、白書、広報誌という媒体は、昭和戦前期までは存在しておらず、Yahoo!知恵袋、yahoo!ブログのようなインターネット上の文章もまた同様である。しかし、政府や役所から国民や住民に告知する文書は戦前にもあり、知恵袋やブログを私的性格の強い文章と考えれば、手紙や日記など近代から存在していた媒体は多い。近代語コーパスの対象に含めるべき資料の候補は、さらに幅を広げて検討していくことが望まれよう。

３．２　叢書類

国語辞典編集準備資料の目録３冊は、近代語コーパスに含めるべき資料を考えるのにきわめて有益な資料であるが、不十分なところも多いため、他の材料を用いて増補していくことが必要である。特に、明治前期の資料の手薄さが目立つため、まずはこの時期の資料を豊富におさめる叢書類をもとに資料リストを増補していくことにした。用いた叢書は次の４つである。

(1) 明治文化全集　全 24 巻（1927～1932 年、日本評論社）
(2) 明治文化資料叢書　全 12 巻（1959～1963 年、風間書房）
(3) 日本近代思想大系　全 24 巻（1988～1992 年、岩波書店）
(4) 新日本古典文学大系 明治編　全 30 巻　（2001 年～刊行中、既刊 29 巻、岩波書店）

これらの叢書は、言語研究を目的として編纂されたものではないが、文化・思想・文学を中心に多様な分野の重要資料が選ばれていると考えられ、そこには、言語資料としても価値の高いものも含まれていると思われる。

表１　叢書類に収録される資料の数（時代別）

	江戸期	明治前期	明治中期	明治後期	計
明治文化全集	16	265	196	16	493
明治文化資料叢書	2	20	50	39	111
日本近代思想大系	70	959	504	7	1540
新古典大系明治編	1	26	99	14	140
計	89	1270	849	76	2284

表１は、四つの叢書に収録された資料の数を発行された時代別にまとめたものである（発行年代が大正期以後のものや不明のものは集計から除いてある）。明治前期・明治中期に集中しており、国語辞典編集準備資料の目録で不十分だった部分を補うことができよう。

この四つの叢書以外にも、資料リスト増補の材料として有用な叢書や図書目録は色々と考えられるが、まずは、上記の三つの目録と四つの叢書とから作成した資料リストの中身を分析することで、近代語史をとらえるための資料選定をどのように行っていくのがよいかを考えていきたい。

４．資料リストの分類と資料選定の考え方―明治前期・中期を例に―

４．１　文体の観点

４．１．１　文体の流れ

ここでは、明治前期・明治中期を例に取り上げたい。上記の、国語辞典編集準備資料と叢書類から作成した資料リストのうち、明治前期・明治中期の部分には、2000 点余りがお

さめられている。これについて、文体・ジャンル・媒体の三つの観点から分析を加えていこう。はじめに文体の観点から見る。

言文一致による口語体書き言葉の成立は、近代語史における最重要の出来事のひとつだが、その文体の流れを、森岡（1991）が示す図式をもとにまとめると、表２の通りである。明治初期には、文語体も口語体も多様な文体があったが、次第に統合されていき、明治 40 年代には言文一致体という口語体ひとつに統合されていく流れがあった。統合以前に多様に分かれていた文体は、研究者によって様々な分類や名付けがなされており、森岡説はそのひとつである。各文体は連続し交錯し、相互の識別が難しい場合も多い。要点は、近代の文体史は多様性から均質性へという明確な方向性をもっており、まずは文語体・口語体それぞれの内部で統合され、やがて口語体が全体に及んでいき、明治時代のうちにそれが完結するというところにある。文語体の内部、口語体の内部での文体の識別は、その指標が立てにくいが、文語体か口語体かの別については、文末辞を指標として明確に識別することが可能である[6]。

表２　近代語の文体統合の流れ（森岡 1991 に基づき作成）

<table>
<tr><td colspan="2"></td><td>明治初期</td><td>明治 10 年代</td><td>明治 20 年代</td><td>明治 30 年代</td><td>明治 40 年代</td></tr>
<tr><td rowspan="6">実用文系統</td><td rowspan="3">文語体</td><td>漢文訓読体</td><td rowspan="3">和漢折衷体</td><td colspan="2" rowspan="3">明治普通文</td><td rowspan="7">言文一致体</td></tr>
<tr><td>和漢折衷体</td></tr>
<tr><td>候文</td></tr>
<tr><td rowspan="3">口語体</td><td>問答体</td><td rowspan="3">演説体</td><td rowspan="3">演説体</td><td rowspan="3">初期言文一致体</td></tr>
<tr><td>講述体</td></tr>
<tr><td>談話体</td></tr>
<tr><td rowspan="2">文学系統</td><td>口語体</td><td>俗文体</td><td>講釈体</td><td>初期口語体</td><td>初期言文一致体</td></tr>
<tr><td>文語体</td><td>和漢折衷体</td><td colspan="2">雅俗折衷体</td><td>（雅俗折衷体）</td></tr>
</table>

４．１．２　文語体と口語体

表３　明治前期・明治中期の文体

	明治前期	明治中期
文語体	1187（93.1%）	773（91.1%）
口語体	31（2.4%）	47（5.5%）
文語体・口語体	3（0.2%）	0（0%）
その他	55（4.3%）	29（3.4%）
計	1276（100%）	849（100%）

表３は、明治前期・中期の 2000 点余りの資料について、文語体か口語体かを認定しその数と比率をまとめたものである[7]。文語体と口語体が混用されているものは、基調をなす文

[6]文末辞が「なり」「たり」「き」「けり」などで終わる文体は文語体、「だ」「である」「た」「です」「ます」などで終わる文体は口語体と識別できる。『太陽コーパス』の文体情報もこの基準で付与してある。

[7]明治前期には国語辞典編集準備資料と叢書類の両方を集計し、明治中期には叢書類のみを集計した。これは、国語辞典編集準備資料が示す資料のすべてを実際に見ることができなかったため、文体が未確認のものが残ったことによる。

体がどちらであるかによって区別した。「文語体・口語体」と記したのは、両者が同等であるもの、「その他」は漢文や英文あるいは文章でないもの（名簿など）である。明治前期では文語体がほとんどで、明治中期には口語体が数パーセント増加するものの、まだ大部分が文語体である。この時期、文語体が圧倒的に優勢であったことが確かめられる。

４．１．３　文語体

明治前期の文語体を、森岡（1991）は、漢文訓読体、和漢折衷体、候文の３種に分類するが、それぞれ、次のような文体のことを指す。上記の資料リストに含まれるものから１例ずつをあげてみよう。

○漢文訓読体

吾輩日常二三朋友ノ盍簪ニ於テ偶當時治亂盛衰ノ故政治得失ノ跡ナド凡テ世故ニ就テ談論爰ニ及ブ時ハ動モスレバカノ歐洲諸國ト比較スルコトノ多カル中ニ終ニハ彼ノ文明ヲ羨ミ我ガ不開化ヲ歎ジ果テ果テハ人民ノ愚如何トモスルナシト云フコトニ歸シテ亦欷歔長大息ニ堪ザル者アリ

（西周「洋字を以て国語を書するの論」、『明六雑誌』1、1874 年、明六雑誌原本による）

○和漢折衷体

輕重長短善惡是非等ノ字ハ相對シタル考ヨリ生ジタルモノナリ輕アラザレバ重アル可ラズ善アラザレバ惡アル可ラズ故ニ輕トハ重ヨリモ輕シ、善トハ惡ヨリモ善シト云フコトニテ此ト彼ト相對セザレバ輕重善惡ヲ論ズ可ラズ斯ノ如ク相對シテ重ト定リ善ト定リタルモノヲ議論ノ本位ト名ク諺ニ云ク腹ハ脊ニ替ヘ難シ又云ク小ノ虫ヲ殺シテ大ノ虫ヲ助クト

（福沢諭吉『文明論之概略』、1875 年、文明論之概略原本による）

○候文

浜田御預り所村々百姓共、衆訴落印と二つに相分り候に付、今度鶴田御役所より御役人様御上下拾六人、書添村へ御出張に相成、

（津山藩岡熊治郎による監察記録、1868 年、日本近代思想大系による）

候文は文末などに「候ふ」を伴うもので、文体類型として確立し、この類型に属する文章を特定していくことができるが、漢文訓読体と和漢折衷体との識別は難しい。漢文訓読体に和文や俗文の要素が交じった福沢諭吉の文章などが和漢折衷体の典型とされるが、個々の文章を漢文訓読体と和漢折衷体とに判別する明確な指標は立てることはできない。

４．１．４　口語体

森岡（1991）は、明治前期の口語体には、実用文系統に３種、文学系統に１種あったと見ているが、それぞれ、次のようなものを指すと思われる。やはり、上記の資料リストに含まれるものから例をあげよう。

○問答体の例

開化文明　サア／＼英吉君。是こそ僕が舊宅だ。

西海英吉　ホヽウ成程、茅葺の門長屋、廣庭の植ごみ、こなし部屋から牛部屋の景況、なんとなく古色を帶て、歴然たる舊家の豪農殿が兵衛が宅に來たやうだね。ソシテアノ異な歌を大勢が唱つて居るあれは何ンだね。

（横河秋濤『開化の入り口』、1873‐1874 年、明治文化全集による）

○講述体の例

世の諺にも｜不治是天福［しらぬがほとけ］と申す通りで、成程世の事國の事も自身に識らざる時は、更に心に掛［かゝ］らずして一向心配することはありますまい。だが、右の如く人間が箇［か］様［やう］に世間の物事を識らずして濟むものでありませう歟［か］。

（植木枝盛『民権自由論』、1879 年、明治文化全集による）

○談話体の例

なぐさみながら、よみあげまする。お経の文句はなにがなんだと、たずねてみたれば、作州五郡の庄屋がねんらい、あんまりおうきな盗みをしおった。そのしりだん／＼百姓がほりかけ、あちらもこちらも村々さわだち、中々ちよっこりちよっとにゃおさまりませんが、そのわけあらまし申してみふなら、ぬすんだそのかずおふひが中にも、とりわけ大きな事からあげます。

（本多応之助「鶴田騒動の阿呆陀羅経」、1868 年、日本近代思想大系による）

○俗文体の例

モシあなたヱ牛［ぎう］は至［し］極［ごく］高［かう］味［み］でござすネ此［この］肉［にく］がひらけちやアぼたんや紅葉［もみぢ］はくへやせんこんな清［せい］潔［けつ］なものをなぜいままで喰［く］はなかつたのでござウせう

（仮名垣魯文『安愚楽鍋』、1871 年、明治文学全集による）

明治前期の口語体資料は約 30 点あるが、それらが上の４種の文体のいずれであるか分類するのが難しい場合も多い。これらの種別は明確な類型としてではなく、口語体の多様な広がりの範囲を考える目安として考えるのが適切であろう。

４．１．５　資料選定における文体の扱い

以上見てきたように、明治前期に多様であった文体について、明確な類型と指標を立てて、個々の文章を分類していくことは困難である。一方、文語体と口語体の識別は文末辞を指標として明確に判別していくことが可能である。したがって、資料選定においては、文語体か口語体かの別については、これを選定の際の判断材料に用いることができるが、それぞれの中の細分類は、材料として採用しにくいと考えられる。むしろコーパスを作成した後に文体の詳細な研究が行われるべきだろう。

なお、明治前期・中期は、口語体の比率はきわめて低いが、それを理由として、当期のコーパスにおける口語体資料の構成比率をうんと低くするのは適切でないと考えられる。なぜなら、後代にすべての文体を統合していく口語体がどのように変容し発展したか、また普及し定着していったかを歴史的に把握するためには、まだ少数派だった初期段階のそれを積極的に採り、その変化の過程を研究できるようにしていくべきであるからである。このようなところは、言語史研究のためのコーパス設計における資料選定では、サンプリングによる代表性の尊重よりも、個々の資料の独自性の尊重が優先される部分だと言えるだろう。

４．２　ジャンルの観点

ジャンルの枠組みは、『現代日本語書き言葉均衡コーパス』の書籍や、『太陽コーパス』では、図書館における書籍の分類基準である NDC（日本十進分類法）が用いられている[8]。上述の資料リストに収録される資料についても図書館に収録されている書籍の場合は、

8『現代日本語書き言葉均衡コーパス』では国会図書館の書誌データに付されている NDC 番号を利用したが、『太陽コーパス』ではコーパス作成者が記事を読んで番号を付与した。

NDC 番号が取得できる場合がある。そこで、国立国会図書館の「近代デジタルライブラリー」を検索し、そこに収録されているものに NDC 番号を引き当て、明治前期・中期のジャンル分布を図４に表した。

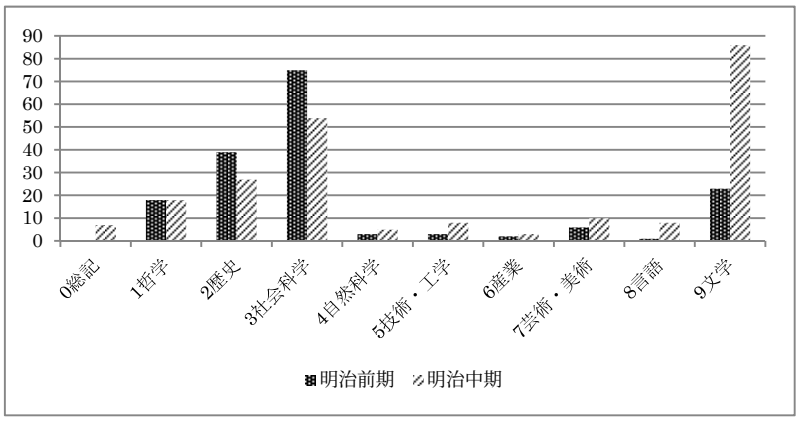


図４　明治前期・中期の資料のジャンル

明治前期は、社会科学が最も多く歴史がこれに次ぎ、さらに文学、哲学の順に多い。ところが、明治中期では文学が最も多くなっており、社会科学がこれに次ぎ、そして歴史、哲学という順となり、時代的な変容が大きい。これも、時代によるジャンルの多寡の違いが反映している面と、データ作成の典拠とした目録や叢書の性質を反映している面とがあろう。このような大きな変容があるところでは、単純に実際の構成比率にしたがってサンプルの比率を決めるだけでは適切でないように思われる。むしろまずは、資料リストの中身を見ながら、当期の当該ジャンルの資料として重要性の高いものであれば採ることを検討し、そうでなければ別に典拠とすべき叢書や目録がないか検討していくような研究段階が必要ではないだろうか。例えば、当期の自然科学や技術・工学の資料はきわめて少ないが、表４のような資料が含まれている。これらの資料を実際に見て、コーパス化の適否を考えていくことが望まれよう。このような点も代表性だけでなく個々の資料の性質への目配りが必要になるところである。

表４　明治前期の「4 自然科学」「5 技術・工学」の資料（部分）

資料	著者	NDC	文体	西暦	叢書	叢書巻
訓蒙 窮理図解	福沢諭吉	420	文語	1868	日本近代思想大系	科学と技術
物理了案	宇多健齋	420	文語	1880	明治文化全集	科学編
舎密局開講之説	三崎嘯輔	430	文語	1870	明治文化全集	科学編
天変地異	小幡篤次郎	440	文語	1868	明治文化全集	科学編
西洋時計便覧	柳河春三	535	文語	1870	明治文化全集	風俗編
男女普通家政小学	小林義則	590	文語	1880	日本近代思想大系	風俗 性
女房の心得	望月誠	590	文語	1878	日本近代思想大系	風俗 性
服製年中請負仕様書	鈴木篤右衛門	593	文語	1868	明治文化全集	風俗編
西洋料理通	仮名垣魯文	596	文語	1872	明治文化全集	風俗編
通俗男女自衛論	三宅虎太	598	文語	1878	日本近代思想大系	風俗 性

４．３　媒体の観点

資料リストを見ていくと、先に「ジャンル」として設定した NDC とは別の枠組みで分類した方がよいのではないかと思われるものが目につく。例えば、表５に示したものは、明治 8（1875）年に発行された新聞・雑誌の一群の一部である。

表５　明治 8（1875）年の新聞・雑誌（部分）

資料	著者	NDC	文体	西暦	叢書	叢書巻	出典
評論新聞	海老原穆	―	口語・文語	1875	明治文化全集	雑誌編	
仮名読新聞	―	―	口語	1875	日本近代思想大系	言論とメディア	
萬国叢話	―	―	文語	1875	明治文化全集	雑誌編	
国民気風論	西周	150	文語	1875	日本近代思想大系	天皇と華族	明六雑誌
華士族論	島地黙雷		文語	1875	日本近代思想大系	天皇と華族	共存雑誌
善良なる母を造る説	中村正直	370	文語	1875	日本近代思想大系	教育の体系	明六雑誌
真影の禁を論ず	高木登		文語	1875	日本近代思想大系	天皇と華族	朝野新聞

明治前期に次々に創刊される新聞や雑誌それ自体が叢書におさめられている場合（上の三つ）と、叢書に採られた資料の出典が新聞・雑誌である場合（下四つ）とがある。飛田（1973）は、新聞・雑誌は、近代に存在する多様な言語資料の性格をすべて合わせもっている「総合資料」という扱いをしており、雑誌『太陽』がそれ単体で代表性を持つと考えて『太陽コーパス』を設計したのも、そのような考え方に立ってのことであった。コーパス作成にあたっては、新聞・雑誌は、その総合性が生きるように、多様な資料をまとめて採集できる資料として扱うのが適切だろう。具体的には、総合性の高い新聞や雑誌をいくつか定め、その新聞や雑誌につついては、例えば、『太陽コーパス』で採ったような、等間隔の期間を置く方法などによってサンプリングを行うこが考えられる。どの雑誌・新聞を選ぶかは、資料の独自性を重視するものだが、その内部をサンプリングするのは、代表性を意識する選定方法と言うことができるだろう。

新聞・雑誌以外で目を引くのは、法令、文書、手紙・日記の類である。法令は、『現代日本語書き言葉均衡コーパス』の「特定目的サブコーパス」に「法律」として採られた枠組みに対応する。文書は、公的な文書については、同じく白書や広報誌と通じるところがあろう。手紙・日記のうち私的な性質を持っているものは、同じく Yahoo!知恵袋や Yahoo!ブログと共通する性格があろう。これらは、近代の重要資料として一群をなしているだけではなく、『現代日本語書き言葉均衡コーパス』への接続という点でも重要性の高いものである。こうした NDC によるジャンルとは別に立てることが必要だと思われる分類枠は、広い意味で「媒体」と呼ぶことができるだろう。

なお、上記の資料リストには少数しか入っていないが、近代語研究の重要資料には他に、教科書、演説や落語などの速記、日本語について記述した文典・辞書などが存在する。教科書は、『現代日本語書き言葉均衡コーパス』における教科書と対応する。速記は、同じく国会会議録や『日本語話し言葉コーパス』に対応づけられるものとしても重要であり、明治後期以後には演説や落語の録音資料も存在しており、近代語コーパスに話し言葉資料をどのように取り込むかという課題につながっていこう。また、文典・辞書などは、コーパスの直接の対象にはしにくい面もあるが、コーパスから記述できる近代語の文法や語彙の実態と対照すべき資料として重要性は高く、コーパス設計時において、その関連づけの方法を検討しておくことも有意義なことだろう。これら現段階の資料リストでは手薄な重要資料を補っていく作業も必要である。

４．４　その他の観点

上に記した、文体、ジャンル、媒体のほか、ある資料をコーパスに入れるかどうかを検討する際に考慮すべき点が、ほかにも想定される。まず、原本の参照可能性の高さという点である。文献資料に基づく日本語史研究においては、コーパスができれば原本を見なくてもよいということにはおそらくならず、コーパスのもとになった本文が原資料でどのような姿であったかを参照したいという要求が研究者には強く存在すると考えられる。そうした要求に応えられるように、コーパス作成と同時に原本の影印や画像などを作成し関連付けることも考えられるが、現実にはそこに開発コストをかけることは難しい面がある。そこで、複製本が出版されていたり、国立国会図書館などの電子図書館で画像が公開されていたりするものをコーパス化することが考えられる。同じような理由で、本文についての研究成果が反映した校訂本、注釈書、索引などが整備されている資料も、コーパス化する価値が高いであろう。

最後に指摘するのは、コーパスとして用いられる場合でなくとも、文献資料による言語研究一般において、価値が高いとされる資料は、コーパスの対象としても価値が高いという点である。例えば、振り仮名がついているものは語形が確定できる優位性があり、著者の自筆本に基づいているものは別人による改変の心配がないという優位性がある。

以上のような、コーパス化する資料そのものの優位性にかかわる情報も、資料リストに書き入れておき、選定の際の判断材料に使えるようにしておけるとよいだろう。

５．資料選定の実施に向けて

５．１　資料選定の基本的手順

以上述べてきたことを踏まえて、近代語コーパスを設計する際に、今後どのようにして資料を選定していけばよいかについて、現段階で想定される基本的な手順の見通しを記しておきたい。

(1) 時代、媒体、ジャンル、資料の四層を立て、この枠組みで分類しながら資料のリストを増補していく。利用する叢書や目録は、現在手薄となっている媒体やジャンルを中心に、範囲を広げていく。
(2) 第Ⅰ層には時代を立てる。時代区分は 5 年を一単位とし、明治・大正期は三つの単位をまとめた 15 年ごとの明治前期・明治中期・明治後期・大正期というまとまりを設定する。昭和戦前期は 20 年でひとまとまりとし、昭和戦後期も当面分割しない。
(3) 第Ⅱ層に媒体を立て、書籍（初出が雑誌・新聞等のものも含む）、新聞・雑誌、教科書、法令、文書、手紙・日記などに分類する。なお、文学作品とベストセラーの目録から収集した資料はまとめて「書籍」に入れる。
(4) 第Ⅲ層にジャンルを立て、書籍は NDC の第１階層を枠組みとし、NDC では細かすぎる場合は、部分的に統合する。書籍以外は各媒体の性質に応じて枠組みを検討するが、第Ⅲ層が不要な（直下の層が資料である）媒体もある。
(5) 第Ⅳ層は個々の資料とするが、資料リストには、各資料について、発行年、媒体、ジャンル、資料名のほか、著者名、文体、出典、複製本、注釈書、索引、所蔵図書館、表記法、底本の状態等、選定作業において有用と思われる情報をできるだけ書き加え、選定作業の判断材料とする。
(6) 四つの層による分類を見わたしながら、各資料の特質を吟味し、各層各枠の中で資料に優先順位を付けていく。
(7) 近代語コーパスの開発期間、開発予算、開発手順などが具体化してきたら、資料リストを活用して資料選定案を作成する。

上に記した作業手順は、一言で言えば、近代語資料全体のバランスと個々の資料の性質との両面を考慮した選定方法で、はじめに述べた「代表性」と「独自性」の両面を考慮し

たものである。このような作業仮説を立てて候補になる資料を実際に見ながら分類し、採否の基準やバランスの取り方を工夫していくことが重要だろう。近代語研究の最大の障壁は資料が多すぎることだと言われることもあるが（湯浅 2000）、資料論を重ねながらコーパスを設計することで、その障壁を乗り越えていく道筋も見えてくるのではないだろうか。そのような検討や工夫を議論する場を、多くの近代語研究者が参加できる形で設けていくことも大切だろう。

５．２　資料選定の実施例―明治前期を例に―

現段階では資料リストは作成途上であり、層による粗密があったり、資料の実物を見ていないために、リストに記入すべき情報が不足していたり、ジャンルや文体などの分類が不十分であったりするものも多い。資料リスト整備はさらに継続していく必要がある。ここでは、実際に資料選定を実施する場合に論点になりそうなことを、現段階の資料リストで、第Ⅰ層（時代）が明治前期（明治元～15 年）になっている、約 1300 件の資料をもとに、少し考えてみたい。

明治前期の資料の第Ⅱ層（媒体）の内訳は、書籍と新聞・雑誌がそれぞれ約 350 件、文書が 500 件弱、法令が 100 件弱で、ここまでがまとまった量があるものである。一方、手紙・日記、教科書、辞書・文典、速記、韻文等は、いずれも 10 件に満たない。これらの媒体については、書籍や文書等に分類されているものの中に、見方によってはこれらのいずれかに分類できるものがあったり、そもそも資料に関する情報収集が不十分なところがあったりするため、明治前期にあまり存在しなかった媒体だと言い切ることはできず、さらに精査していくことが求められる。文書がきわめて多くなっているのは、明治前期という社会体制が大きく変わる時期の資料を、文書から豊富に集めた叢書類の編集方針によるものである。文書における第Ⅲ層（ジャンル）をどのような枠組で分類していくかは課題であるが、例えば、叢書が立てる「宗教」「憲政」「風俗」「教育」といった内容から分類することや、典拠となっている「日本外交文書」「大久保利通文書」などのような編纂文書の種類ごとにまとめることなどが、想定できよう。

書籍の第Ⅲ層（ジャンル）は、NDC を用いるのが便利である。国会図書館等に所蔵があるなどして NDC 番号を引き当てることができた資料が 240 件ほどあり、0 番台「総記」から 9 番台「文学」までのすべてのジャンルにわたっている。そのうち、「文学」に分類されるものは、表６の 21 件である。表の中での資料の配列は刊行年順である。

第Ⅰ層の時代は、明治前期①（明治元～5 年）・明治前期②（明治 6～10 年）・明治前期③（明治 11～15 年）の三期に細分した。第Ⅱ層（媒体）、第Ⅲ層（ジャンル）はこれ以上の細分の必要はなさそうである。第Ⅳ層でどの資料を選ぶかの観点として、表６に示した「時代」「文体」「様式」「振り仮名」「日国用例数」「本の存在など」の情報を、この順で考慮したい。具体的には、時代は特定の期に偏らないようにすること、文体は口語を優先するが文語も採るようにすること、様式は多様になるようにすること、振り仮名は総ルビ・部分ルビの順に望ましいが無ルビでも排除しないようにすること、日国用例数（『日本国語大辞典第 2 版』でその資料から採られている用例の数）は多い方がよいこと、本の存在は国会図書館（近代デジタルライブラリー）や国語研に所蔵があるものが望ましいことなどを考慮するのがよいだろう。そうした考慮の結果、コーパスの対象として優先されると考えられるものから順位を付け、一番左側の列に記入した。具体的には、『安愚楽鍋』『通俗伊蘇普物語』『人間万事金世中』『西国立志編巻之貳　其粉色陶器交易』『怪化百物語』『西洋道中膝栗毛』『欧州奇事花柳春話』『近世紀聞』『鳥追阿松海上新話』『魯国奇聞烈女之疑獄』の順になり、他はさほど優先順位は高くないと考えられた。明治前期の書籍の文学では、これらがコーパス化に適切な資料ではないかと考えられ、この後は、他のジャンルや媒体などとのバランスから、さらに絞り込んでいくことになるだろう。

このような選定作業を、文学以外のジャンルや、書籍以外の媒体に対してもできるだけ行っていき、さらには、明治中期以後の時代にも行っていくことが考えられる。多くのジ

ャンル、媒体、時代について検討が進めば、それら全体を見わたした上でのバランスを取る作業も行うことができるようになるだろう。そのようにして、独自性と代表性の双方に目配りした選定作業を行っていくことが望まれよう。

表６　明治前期・書籍・文学の資料選定例

順位	Ⅰ時代	Ⅱ媒体	Ⅲジャンル	Ⅳ資料	著者	様式	刊行年	文体	振り仮名	日国用例数	本の存在など
6	明治前期①	書籍	913	西洋道中膝栗毛	仮名垣魯文	戯作	1870	口語	部分ルビ	1785 件	国会、複製あり
1	明治前期①	書籍	913	安愚楽鍋	仮名垣魯文	戯作	1871	口語	部分ルビ	1060 件	国語研蔵、索引あり
8	明治前期②	書籍	913	近世紀聞	染崎延房	記録	1875	文語	総ルビ	1954 件	国会、新古典大系
5	明治前期②	書籍	914	怪化百物語	高畠藍泉	戯作	1875	口語	総ルビ	54 件	国会、新古典大系
4	明治前期②	書籍	912	西国立志編巻之貳 其粉色陶器交易	佐藤富三郎	劇	1873	口語	部分ルビ	0 件	国会
	明治前期②	書籍	913	西国立志編巻之十 鞋補童教学	佐藤富三郎	劇	1873	口語	部分ルビ	0 件	国会
2	明治前期②	書籍	930	通俗　伊蘇普物語	渡部温	小説	1873	口語	総ルビ	46 件	国会
	明治前期②	書籍	930	開巻驚奇　爆夜物語	永峰秀樹	小説	1875	文語	部分ルビ	0 件	国会、翻訳
3	明治前期③	書籍	913	人間万事金世中	河竹黙阿弥	劇	1879	口語	総ルビ	49 件	国会、新古典大系
	明治前期③	書籍	912	日本美談	前田正名	劇	1880	口語	総ルビ	0 件	国会
	明治前期③	書籍	913	島衙月白浪	河竹黙阿弥	劇	1881	口語	総ルビ	70 件	国会、新古典大系
9	明治前期③	書籍	913	鳥追阿松海上新話	久保田彦作	小説	1878	文語	総ルビ	93 件	国会
	明治前期③	書籍	930	新説　八十日間世界 一周	川島忠之助	小説	1878	文語	無ルビ	3 件	国会、翻訳、新古典大系
7	明治前期③	書籍	930	欧州奇事花柳春話	丹羽純一郎	小説	1878	文語	無ルビ	2543 件	国会、翻訳
	明治前期③	書籍	914	高橋阿伝夜刃譚	仮名垣魯文	小説	1879	文語	総ルビ	27 件	国会、新古典大系
	明治前期③	書籍	930	哲烈禍福譚	宮島春松	小説	1879	文語	総ルビ	22 件	国会
	明治前期③	書籍	930	九十七時二十分間月 世界旅行	井上勤	小説	1880	文語	無ルビ	0 件	国会、翻訳
	明治前期③	書籍	913	民権演義　情海波瀾	戸田欽堂	小説	1880	文語	部分ルビ	0 件	国会、新古典大系
	明治前期③	書籍	930	春風情話	坪内逍遥	小説	1880	文語	総ルビ	0 件	翻訳、新古典大系
	明治前期③	書籍	930	欧州情譜　群芳綺話	大久保勘三郎	小説	1882	文語	部分ルビ	0 件	翻訳
10	明治前期③	書籍	930	魯国奇聞烈女之疑獄	柚田策太郎	小説	1882	文語	無ルビ	0 件	国会、翻訳

上に述べた書籍については、個々の資料の独自性を十分に考慮した選定を行うことがよいと思われたが、媒体やジャンルによっては、このような詳しい検討を行うことが現実的でないものもある。例えば、新聞・雑誌のような定期刊行物は、毎号の多様な記事のジャンルや文体などを逐一分析してから選定することは繁雑である。かといって、全号全文をコーパス化することもコストの面から難しい。この場合は、対象に定めた新聞や雑誌は、

4.3 に述べたように、総合資料としての性格を生かして、ランダムサンプリングを行って、コーパス化する号や記事を定めていくことが適切であろう。どの新聞・雑誌を対象とすべきかという点においては、各資料の性質を書籍の場合と同じように分類整理した上で、選定すべきである。本プロジェクトで作成するモデルコーパスには『明六雑誌』を選定したが[9]、そのような検討によって、明治前期の雑誌として、コーパス化の優先順位が最も高いと考えたことによる。ほかに、明治前期の新聞・雑誌では、『読売新聞』『東京日日新聞』『東洋学芸雑誌』『六合雑誌』などが、優先的にコーパス化されるべきだと考えられる。『明六雑誌』は、刊行された期間が短く、分量も多くないことから、モデルコーパスでは全文をコーパス化したが、刊行期間が長く分量の多い他の新聞・雑誌については、サンプリングを行ってコーパス化する号や記事を限ることが考えられる。

６．おわりに

以上本稿では、今後構築する「近代語コーパス」では、資料全体のバランスと個々の性質の双方をよく検討して、資料選定を行っていくことが重要であることを確認し、詳しい資料リストを作成して、このリストを分析しながら資料選定を実施する事例を示した。そのような資料選定を進めていくことは、「近代語コーパス」が、「通時コーパス」と『現代日本語書き言葉均衡コーパス』とをつなぐ役割を果たすためにも、欠かせないことだと考えられる。

本稿で用いた資料リストは現在作成途上であり、資料の増補や情報の整備を継続させていく必要がある。また、コーパス構築の事業規模に見通しがを立てた上での、実現性を重視した資料選定の段階に進むことも待たれよう。構築事業の立案のためには、短期・長期の両面での行程表の作成や、資料選定以外の設計にかかわる研究成果との統合などに着手することが求められよう。

文　献

小木曽智信(2005)「構造化テキストを直接利用するアプリケーション―『プリズム』と『たんぽぽ』―」（国立国語研究所 2005b 所収、pp.83-113）

木村睦子・加藤安彦・田中牧郎(1998)「国語辞典編集のための用例データベース」（『日本語科学』5、国書刊行会、pp.109-127）

国立国語研究所(1985-1997)『国定読本用語総覧』（三省堂）

国立国語研究所(2005a)『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM、博文館新社）

国立国語研究所(2005b)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』（博文館新社）

近藤泰弘（2012）「日本語通時コーパスの設計」（『NINJAL 通時コーパスプロジェクト・Oxford VSARPJ プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集』国立国語研究所、pp.1-10）

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」（国立国語研究所(2005b)、pp.1-48）

飛田良文(1973)「近代語研究の資料」（『文学・語学』66、三省堂、pp.45-60）

飛田良文(2012)『国立国語研究所「日本大語誌」構想の全記録』（港の人）

前川喜久雄(2008)「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」（『日本語の研究』4-1、pp.82-94）

丸山岳彦・柏野和佳子・田中牧郎（2011）「第 3 章 サンプリング」（『現代日本語書き言葉均衡コーパス 利用の手引　第1.0版』、国立国語研究所コーパス開発センター、pp.21-38）

[9]『明六雑誌コーパス』については、本報告書に収録した、近藤明日子・田中牧郎「『明六雑誌コーパス』の仕様」、及び近藤明日子「『明六雑誌コーパス』の概要」を参照。

森岡健二(1991)『近代語の成立 文体編』（明治書院）
湯浅茂雄(2000)「近代語研究の要点と課題」（『日本語学』19-11、明治書院、pp.138-148）

付 記

本論文は、「第１回コーパス日本語学ワークショップ」（2012 年 3 月 6 日、国立国語研究所）で発表した内容をもとに、加筆したものである。

電子化が望まれる近代語資料探索
―日本語史を研究する大学院生の報告から―

岡島　昭浩（大阪大学大学院文学研究科）[1]
森　　勇太（日本学術振興会特別研究員）
金　　曦泳（高麗大学校言語情報研究所）
竹村明日香（大阪大学大学院博士後期課程学生）
坂井　美日（大阪大学大学院博士後期課程学生）

１．趣旨

本稿の筆頭著者である岡島は、2010 年度に大阪大学大学院文学研究科における「国語史演習」[2]において、「近代語資料探索」をテーマとした。岡島が、本プロジェクトに参加することになったことを受け、これから近代語コーパスを利用することになるであろうと思われる若い研究者たちが､どのような資料がコーパスに含まれることになることを望むか、ということを知りたい、という目的があった。もちろん、日本語史研究に常に付随する資料論を意識した演習でもある。対象の学生は大阪大学大学院文学研究科に属する、博士前期課程・博士後期課程の学生である。

シラバスは、おおよそ以下のようであった。

講義題目　近代語資料探索
授業の目的　幕末・明治初期から昭和初期あたりまでの言語史資料は、英学資料などを除くと、文学作品以外への開拓はまだまだ進んでいない。また、文学作品でも、文学史上有名な一部の作品が使われているのが現状である。そこで、言語史研究に有用な資料を探し、その有効利用を探る。
講義内容　電子化されたら有用であろうと思われるものを探ることを、今回は主眼としたい。
授業計画　発表者が資料を探索し、「電子化されると大変便利であるから、電子化されるべきである」ということを主張する発表をしてもらう。どの方面(分野)に役に立つ資料か、また、底本とすべき資料、その分量(抄出ならばその度合）、必要な精度、タグの必要度なども含めること。

---

1 okajima@let.osaka-u.ac.jp　岡島以外の執筆者は、みな、2010 年度に於て大阪大学大学院文学研究科文化表現論専攻国語学専門分野の学生であった。

2 正式には、博士前期課程の学生が受講するものが「国語史演習」、博士後期課程の学生が受講するものが「国語史特殊演習」であるが、同時に行われるものである。

参考文献　『太陽コーパス』(博文館新社)

上記シラバスの他に、日本文学・国語学を専門分野とする学生を集めてのガイダンスの際や、最初の授業の際に求めたものは、次のようなことである。

膨大で実現性に乏しいものは外す（『明治文学全集』『明治文化全集』を全部、など）
資料の文字数等、言語量を大まかでよいので明示すること
どのようなジャンルの研究に有効か。電子化することによる効果[3]。（プレゼンテーション）
著作権の確認(公開可能な電子資料を目指すために)
OCR との相性チェック[4]

上記のような説明を経て、実際に受講したのは、
国語学専門分野[5]博士前期課程 4 名　博士後期課程 5 名
日本文学専門分野[6]博士前期課程 1 名
の計 10 名であった[7]。

最初の授業で岡島の例示したものは、『旧事諮問録』『史談会速記録』[8]等の歴史系の速記資料であった。「膨大で実現性に乏しいものは外す」という趣旨からは外れるが、抽出などで言語量を減らす方向で考えた。

２．提案されたもの

提案された主な資料群としては以下のようなものがあった。

---

3 電子資料のない状態で、手作業で用例数等を求め、ケーススタディを示した学生が何人もあった。

4 新字体の資料で入力の上、原本等に戻って訂正するなどの場合も考慮に入れた。現在は、版面権は認められていないので、校訂者の権利のみを気にすればよい。

5 同じ文学研究科内に、別に日本語学専門分野があることもあり、国語学専門分野では、歴史的研究・文献研究を中心にしている。

6 国語学専門分野と日本文学専門分野は、研究室を共有するに留らず、院生発表会など、多くの場面で研究を共にしている。課程在籍の間に、国語学専門分野の学生は日本文学の演習を一つ以上、日本文学専門分野の学生は国語学の演習を一つ以上取ることを求めている。

7 本稿の共著者の他に、国語学専門分野のものは、伊藤由貴・清田朗裕・目黒陽子・鈴木久恵・山本一巴。

8 原書房から復刻版が出ており、『幕末明治/研究雑誌目次集覧』古書通信社（1968）に目次がある。明治 25 年から昭和 13 年まで刊行。江戸時代から明治の初め頃までの歴史的証言を集めようとして速記したものの刊行である。史料編纂所に、出版されなかった速記録が蔵されているとのことである（『幕末明治/研究雑誌目次集覧』）

○共通語ないし東京語系

『東洋学芸雑誌』(雑誌)　　〔後述〕

『丁酉倫理会倫理講演集』(雑誌)

丁酉倫理会"が 1900(明治 33)年に発刊した雑誌(終刊は 1946(昭和 21)年。ただし継続誌有り)で、主に"丁酉倫理会"で行われた講演が収められている。口語体の講演が多く収録され、一定の言語量をもった、多数の話者によるコーパスの作成が期待される。

1)　講演録+"雑録"+"時潮"+"出版界(新刊)"+"応問[読者の質問欄]

2)　講演録は初期はすべて口語体。後世では文語体の論文も加わる。

3)　"雑録"以下は文語体・口語体の両方がみられる。

4)　表記はすべて漢字ひらがな交じり文。

縦 35 字×横 15 行　1 号あたりおよそ 70〜100 頁前後

『旧幕府』(雑誌)

明治 30〜34 年。戸川残花編。勝左衛門太郎「夢酔独言」などあり。

「帝国議会会議録」

国会図書館で現在電子テキスト化されているもの（昭和 20 年以降）よりも前のものについてのテキスト化の提案。全体では大きすぎるので、抽出して行うことを提案。

篠田鉱造『幕末百話』『明治百話』『幕末明治女百話』など

篠田鉱造（1871〜1965。報知新聞記者）による、聞書スタイルのもの。

河竹黙阿弥『狂言百種』　〔後述〕

SP レコード文句集

大空社より『大正期 SP 盤レコード/芸能・歌詞・ことば全記録』（倉田善弘・岡田則夫監修。1996〜1997）として復刊されている。金水(2001)参照。

東江学人『文明開化/内外事情』

福沢諭吉『西洋事情』との比較などで、明治初期の言語資料として。近代デジタルライブラリ所収。

松林伯円の講談「安政三組盃」

明治 18 年刊。講談速記本の濫觴とされるもの。前年刊で落語速記本の初めとされる三遊亭円朝の「怪談牡丹燈籠」と比して注目されることが少ない。リライトされた形でしか再版はなされていないようで、原刊本からの電子化が必要となる。近代デジタルライブラリには欠巻あり。

巌谷小波の言文一致もの（「初紅葉」など）

「こがね丸」の文体を換えた執筆などでは言及される巌谷小波ではあるが、その他の言文一致作品も重要である、という指摘。

○上方語系

柴田鳩翁『鳩翁道話』等。

江戸期のもの。『鳩翁遺稿』（昭和 4、柴田寅三郎[9]）による。平凡社東洋文庫の柴田実『鳩翁道話』(1970)は表記の改訂有り。『日本思想大系・石門心学』には、初編のみ。「『鳩翁道話』の文字数は、正編、続編、続々編はそれぞれ約 47,000 字、拾遺に含まれる未刊行の筆記は約 18,000 字である」との情報があった。

一荷堂半水（『諺 臍の宿替え』・『穴さがし心の内そと』）

特に『穴さがし 心のうちそと』は近世上方語資料として知られている。前田(1974)参照。『諺 臍の宿替え』は太平書屋より影印・翻刻あり(1992、武藤禎夫)。また南和男『江戸のことわざ遊び』（平凡社新書、2010)で、一部影印・現代語訳もあり）。

大阪文芸誌『なにはがた』

明治 24 年。西村天囚など大阪朝日新聞関係。また、堺利彦なども参加し、言文一致体のものを書いている。

『上方はなし』　〔後述〕

今村信雄速記『名作落語全集』

騒人社書局,1929-1930 年。紙型の流用で他の出版社からも後に刊行されている。今村信雄は 1959 年歿で、著作権保護期間終了。演者の著作権を確認する必要有り。東西の落語が収録されているが、上方落語を中心の電子化を考える。

○その他

『日本外交文書』『条約改正関係・大日本外交文書』　〔後述〕

永井荷風の小説

日本新聞歴史

番外　青空文庫のデータベース化

他にもあったが、有意義と考えられるものを示した。以下には、その数例を摘記する。

３．例

３．１　河竹黙阿弥『狂言百種』　　坂井美日

『狂言百種』は、明治二十五年四月から明治二十六年二月にかけて出版された、河竹黙阿弥脚本集である。これは、大正期に出版された『黙阿弥全集』などとは異なり、河竹黙阿弥自身が編纂したという点が特徴である。『狂言百種』におさめられた作品は、全て黙阿弥の「世話物」である。黙阿弥の自筆台帳は関東大震災で多くが消失している。

さらに多くの黙阿弥作品を扱おうとするならば、大正期に翻刻された『河竹黙阿弥脚本集』(河竹糸補修・河竹繁俊校訂、全 28 巻、1924-1926)や、その増補修正版の『河竹黙阿弥全集』(河竹糸補修・河竹繁俊校訂、全 28 巻、1924-1926、春陽堂)などがある。この中にはさらに多くの散切物が入っており、『東京日日新聞』『女書生繁』『人間萬事金世中』

9 1942 年没で、著作権保護期間終了。

などがある。

しかし、これらの資料は次の点で問題がある。

1.『河竹黙阿弥脚本集』『河竹黙阿弥全集』はともに、2017 年まで著作権期間が残っている(河竹糸(=黙阿弥の娘)の養子である河竹繁俊(1889-1967)が校訂に加わっているため。)

2.校訂が加わっているため、どの時期の言語資料として扱うべきか、検討の必要がある。

しかし二点目について、『狂言百種』、と『黙阿弥全集』には次のような有意差のある補修があるようにも思われるため、二者の異同比較は有効かもしれない。

○無生物主語、結果相、自動詞の「てある」→「ている」

(1)わずか五錢の所十五錢残ってある(『狂言百種』、木間星、12)

(2)わずか五錢の所十五銭残っている(『黙阿弥全集』木間星、672)

(無生物主語の「～ている(自動詞)」が無生物主語の「～てある(自動詞)」を圧倒して現代語と同様の体系になるのがいつなのかはまだよくわかっていないが、この異同例などからはその整理状況が伺える。)

A.散切物(明治期世話狂言)だけを抜粋した場合、文字数は約 65 万字、

B.全ての世話狂言を含む場合は 145 万字となる。

一作品あたり十三万字として、

A) 13×5 作品=65 万字

B) 13×11 作品=145 万字

OCR→ほぼ読み取れず。手打ち作業がはやいと思われる。コスト計算は次のようになる。

A)

・コピー代: 一作品あたり約 60 枚×5 作品

・手打ち作業: 30 時間×10 名

・確認作業: 20 時間×10 名

B)

・コピー代: 一作品あたり約 60 枚×11 作品

・手打ち作業: 30 時間×20 名

・確認作業: 20 時間×15 名

３．２ 『東洋学芸雑誌』 森勇太

明治 14～昭和 5 年の雑誌。567 冊。

1)　表記法としては、かなの選択(漢字片仮名交じり文か、平仮名交じり文か)、記号法・用字法の面でさまざまな表記法が混在している様相がある。

2)演説口調に近い口語文の特徴は認められるが(談話標識、「です」などの敬語表現)、演説の文章をそのまま筆録したものかどうかは確証がない。ただし、発話者(筆者)を特定するのは容易であり、個人の体系として記述を行うことには適している。

3)理化学系統の論文が収録されており、それらの分野の語彙の出現をたどったり、また、明治に入って導入された新しい概念を訳出するための語彙を探し出すこと、および訳語の変遷等に資することが期待される。

4)待遇表現の面では、「であります」「です」「でございます」などが混在している。用法の面でも、形容詞述語や動詞述語に「です」が充てられている例もあり、丁寧表現の整然とした使用がなされていない。

5)総じて、近代に入っての、表記規範や言文一致体など文体形成に至るまでの書きことば形成の過程を追う資料となりうる。特に学者などの知識層の影響を明らかにすることが期待される。

太陽コーパスからの比較の観点からいえば、以下の点で有用な資料となりうる。

1)太陽コーパス発刊以前の状況を調査する。

2)太陽コーパスにあまり掲載のない理化学系統の語彙の調査を行う。

3)太陽コーパスの並行資料として、太陽コーパスを相対的な視点から捉える。

字数は漸増傾向にあるが、約 500,000 字平均として、約 4,000,000 字程度のコーパスとなることが期待される。

OCR

1881 年と 1917 年のものを OCR にかけてみると(e.Typist　12.0 使用)、ルビが多い 1881 年のものは崩れがあるが、全く認識しない、というわけではない。1917 年のものはきれいに読めている。

著作権

『東洋学芸雑誌』そのものの著作権(東洋学芸社)は最終号発行から 79 年以上経過しており問題とならない。ただし、著作権が明確に雑誌社にあると明記された箇所がないため、論文著者の著作権が問題となる可能性がある。

コスト試算

太陽コーパスと比較検討できる材料とするためには、(タグまであれば一番良いが)少なくともルビまでを明示化した形で電子化することが必要になるだろう。しかし、以下の試算では非常にコストが高く、この点で問題となる。

『東洋学芸雑誌』電子化にかかるコスト

内訳　　個数
原本コピー代金　　3360 枚
OCR 読み込み、修正処理　　1120 時間
▼内訳
[原本コピー代金]コピー枚数の内訳:
1 号平均 70 ページ(論文部分)を見開きコピー→35 枚分×12(ヵ月)×8(年)=3360
[OCR 読み込み、修正処理]時間数の内訳:6 ページ/1 時間→1 時間に 3 枚処理。

その後、DVD-ROM 版が出た（大空社、2011）。

３．３　『上方はなし』五代目笑福亭松鶴　　竹村明日香

「上方はなし」(昭和 11〜15 発刊)は会話体中心の落語速記資料で、原稿用紙 1611 枚分の豊富な口語性を有する。資料性については矢島(2006)(2007a)(2007b)を参照。

電子化に当たってのコスト試算
コピー　　約 280 枚
OCR 読み込み・修正処理　　30 時間×8 名
原本との対照チェック　　20 時間×8 名
→すでに活字化されているので､最初から打ち込む資料よりは格段に低コストで行える。
著作権問題
原本　昭和 11-15(1936-1940)ならば、今でも自由な電子化が可能か。
図書、三田編(1971-1972)を使用すると、三田純一(別名・三田純市)氏死後の 1994+50=2044 年以降しか不可か。

一部に特殊な語彙・語法が見えるが(武士の言葉など〉、全体的に近代上方語を代表する語彙・語法を含んでおり、近代上方語資料として有用であると考えられる。

復刻版は OCR にかけてもほぼ問題なく読み込むので、電子化に大きなコストがかからず、短期間で行えるという長所がある。

３．４　『日本外交文書』　金曦泳

日本語では、「遺憾」「遺憾の意を表する」「積極的に」「前向きに検討する」「可及的速やかに」のような定型表現が存在する。（中略）これには漢文脈の名残の語形を含め、公文書の文書としての構成を背景にした言い回しも存在する。

『日本外交文書』は日本の外務省によって明治期から昭和に至るまで(進行中)の外交文書を、編年体を基本として編纂・公表された公文書集である[10]。量的な面から見ても、例

10 外務省の web ページ上で公開されている。

えば明治期や大正期の公文書における文体や語彙などの変遷を知るに有用な資料であると考えられる。

例　漢語接尾辞「-的」の用例

・「1 巻 1 冊(明治元年)」　全 942 ページ:0 例

・「22 巻(明治 22 年)」　事項 11 の 14 ページ:3 例

・「45 巻 1 冊(明治 45 年)」　事項 1 の 43 ページ:26 例

『日本外交文書』は、基本的に各項目ごとに時間順に番号が付されている。また、一定の形式に沿っているのであって、その形式情報をタグを用いて付与しておくと、より素早くまた的確に必要とする情報に接近できるようになる利点がある。また、日本語だけではなく、外国語(英語・ドイツ語など)とその和訳が同時に掲載されている場合が少なくないため、対訳文における情報をタグを使って付与しておくと、対訳コーパスとしての活用も期待される。このような点から、単に電子化することにとどまらずタグを付しておくと様々な利点があると考えられる。

ただし、

> 「日本外交文書」および日本外交文書デジタルアーカイブの著作権は外務省に帰属します。本コンテンツを著作権の保護期間内に著作権法上認められる範囲(私的使用のための複製など)を超えて使用する場合は、当省の許諾を得る必要があります。

と明記されている著作権などの問題がある。外交文書自体は、団体名義の著作物であり、公表後 50 年で著作権は切れるはずであるが、このデジタルアーカイブに新たに著作権が発生しているとすれば、許諾を得る必要がある[11]。

語数及びコスト試算

・明治期における『日本外交文書』

本巻 63 冊,その別冊 9 冊,別巻 17 冊,追補 2 冊,合計 91 冊,延べ約 73,000 ページ

・語数:約 288,373　÷　946 ページ　＝　約 305 語/頁

合計語数:73,000 ページ×305 語＝約 22,265,000 語

cf.　『現代日本語書き言葉均衡コーパス』「BCCWJ 領域内公開データ(2009 年度版のモニター公開データ)」:4,490 万語

４．まとめ

上記の注にも記したように、大阪大学大学院文学研究科の国語学専門分野では、日本語の歴史的研究を中心にしており、提案された資料については、自分の言語史研究の中で、調べるのに苦労した時代・位相などについて、それを埋めるものを探そうとしたもののように思われる。

---

http://www.mofa.go.jp/mofaj/annai/honsho/shiryo/archives/

11 その後、公開されている電子化の方式が変更されたようで、演習時のように容易に画像データを取得できるものではなくなったようである。

文法史の面で特に必要となる口語性の高い資料を求める一方、語彙的な面などから文語（ただし文芸性の強いものではなく、公用文的なもの)を求める声もあった。

口語については、近世中期頃までの上方語系資料と現代関西方言の間を埋めるような上方語資料を求めており、また、共通語・東京語系資料についても、言文一致の定着よりも前のものが主として求められている感があった。それ以後のものは、既存のもの（青空文庫や『CD-ROM 新潮の百冊』「国会会議録」など）で或る程度覆えるという感触によるものと思われる。

具体的な提案もあり､予算さえあればすぐにでも取りかかりたいと思える企画もあった。特に上方語資料などは、大阪の大学として作れないものかと思っている。

また、例として挙げたもの以外にも、小規模のテキストデータとしての存在意義は充分なものや、日本語研究者以外の手によるテキスト化を待って、日本語史資料として使うことが望まれるものもあった。

本稿は、演習時の資料および、期末のレポートによって編集したが、言語的な調査の多くについては、今後、発表されることもあるだろうということや、全体のバランスを考えて、本稿においては、岡島の責任で削除したものが多い。その他の抽出・削除・並べ替えなどについても、責任は岡島にある。

文 献

金水敏(2001)「《資料紹介》明治・大正時代 SP レコード文句集について」『語文』75,76 pp.80-88

前田勇(1974)「穴さがし心の内そと」『近代語研究』4、武蔵野書院 pp.429-484

三田純一編(1971-1972)　五代目笑福亭松鶴『上方はなし』(上・下・解説)、三一書房

矢島正浩(2006)「落語録音資料と速記本――五代目笑福亭松鶴の仮定表現の用法から――」『愛知教育大学国語国文学報』64 pp.132-116

――――(2007a)「五代目笑福亭松鶴落語における原因・理由表現の用法」『愛知教育大学大学院国語研究』15 pp.70-56

――――(2007b)「近代関西言語における条件表現の変遷原理に関する研究」平成 17-18 年度科学研究費補助金(基盤研究 C)研究成果報告書　課題番号 17520298

近代語文献を電子化するための文字セット

高田 智和　（国立国語研究所 理論・構造研究系）[1]

1．はじめに

本稿では、近代語文献を電子テキスト化する際に準拠する文字セットについて検討する。

一般的に、紙媒体の文書を電子テキストに写し取る際には、標準化された符号化文字集合に準拠することが多い。かつて、JIS X 0208 文字セットに拠って、『大正新脩大蔵経テキストデータベース』（http://21dzk.l.u-tokyo.ac.jp/SAT/）の前身、『青空文庫』（http://www.aozora.gr.jp/）、『太陽コーパス』など、学術的価値を有する電子テキスト群が作成されたが、その都度、JIS X 0208 文字セットでは表現できない文字（外字）への対処が問題となっていた（安永（1998）、下田・師（1999）、富田（2000）、池田・白井・高田（2002）、高田（2002）、田中（2005）、當山（2009）、須永・堤・高田（2011）など）。

2000 年に制定された JIS X 0213 は、外字の問題を解消し、「現代日本語を符号化するために十分な文字集合」を目指して開発された規格である。JIS X 0208 が、第 1 水準漢字、第 2 水準漢字、非漢字（X0208 非漢字と呼ぶ）の約 6,800 字であるのに対して、JIS X 0213 は、JIS X 0208 文字セットを拡張し、第 3 水準漢字、第 4 水準漢字、非漢字（X0213 非漢字）を加え、約 11,000 字の文字セットとなっている。

JIS X 0213 文字セットに準拠して作られた『現代日本語書き言葉均衡コーパス』では、のべ 99.96%の文字を符号化できることが確認されている（高田ほか（2009））。『現代日本語書き言葉均衡コーパス』での運用実験により、JIS X 0213 は「現代日本語を符号化するために十分な文字集合」であることが実証されたと言えよう。

しかし、現代から時代を遡って、歴史的な日本語文献に対して、JIS X 0213 文字セットがどの程度の有効性を持ち得るのかは、まだ試みられていない。以下に、『太陽コーパス』の JIS X 0208 外字を対象として、JIS X 0213 による再符号化結果を報告し、近代語文献の電子テキスト化における JIS X 0213 文字セットの有効性について見通しを述べる。

2．『太陽コーパス』の文字処理

『太陽コーパス』は 2005 年に公表された近代語研究向けのコーパスである。1895(明治 28)年から 1928(昭和 3)年まで発行された総合雑誌『太陽』のうち、1895 年、1901 年、1909 年、1917 年、1925 年の各 12 冊、全部で 60 冊分の記事を電子テキスト化し、引用、文、振り仮名などのタグを付けた電子化コーパスである。

『太陽コーパス』の文字処理では、JIS X 0208 文字セット（第 1 水準漢字・第 2 水準漢

[1] ttakada@ninjal.ac.jp

字・非漢字、約 6,800 字）に準拠して、『太陽』の記事の入力処理が行われている。『太陽コーパス』の開発は、JIS X 0213 が制定される以前から着手されており、開発当時の処理水準を考慮すれば、妥当な選択であったと思われる。

雑誌『太陽』では主に旧字体が使われているが、「羽」のように新字体「羽」との字体差がわずかであるものは、新字体で表現することを許容している。字体粒度の許容範囲は、原則として JIS 規格の包摂規準にしたがっている。また、JIS 規格の包摂規準を拡張させて、独自の包摂も行っている（田中（2005））。語彙研究や文法研究での活用が主に期待されているコーパスであるから、字体の忠実な再現は行っていない。

『太陽コーパス』における JIS X 0208 文字セットの使用状況をまとめたものが表 1 である。表の各セルで、上段がのべ字数、下段が異なり字数である。『現代日本語書き言葉均衡コーパス』では、第 2 水準漢字は異なり 1,311 字の使用にとどまっていることから、近代日本語書き言葉では多種多様の漢字を用いていることが裏付けられる。

表 1：『太陽コーパス』の JIS X 0208 による符号化の内訳

水準	1895	1901	1909	1917	1925	計
第 1 水準漢字	1,395,861 [2,687]	1,285,938 [2653]	1,107,698 [2,648]	1,012,692 [2,623]	848,430 [2,633]	5,650,619 [2,721]
第 2 水準漢字	221,940 [2,558]	200,767 [2,275]	172,761 [2,057]	168,536 [1,922]	124,880 [1,757]	888,884 [2,864]
非漢字	1,688,600 [291]	1,651,017 [279]	1,556,071 [278]	1,454,064 [268]	1,447,583 [287]	7,797,335 [318]
計	3,306,401 [5,536]	3,137,722 [5,207]	2,836,530 [4,983]	2,635,292 [4,813]	2,420,893 [4,677]	14,336,838 [5,903]

『太陽コーパス』では、JIS X 0208 文字セットで表現できない文字を、外字、踊字、合字、小書の 4 種のタグを用いて表現している。JIS X 0213 は JIS X 0208 を拡張したものであるから、外字、踊字、合字、小書の各タグで表現された文字類を、JIS X 0213 の拡張領域（第 3 水準漢字・第 4 水準漢字・X0213 非漢字、約 4,000 字）と突き合わせることで、『太陽コーパス』の JIS X 0213 文字セットによるカバー率をおおよそ求めることができるであろう。

JIS X 0213 の特長の一つに、JIS X 0208 では符号位置を区別せず包摂されていた字体を、符号位置を区別して表現できる点があげられる。例えば、「徳」の旧字体「德」や、「鴎」の康熙字典体「鷗」などの異体漢字が該当する。厳密さを追求すれば、JIS X 0208 では包摂するとされていた異体漢字や、『太陽コーパス』で独自に包摂した異体漢字についても、雑誌『太陽』で実際に使われている字体を確認し、JIS X 0213 の拡張領域の符号位置で表現できるかどうかを検討しなくては、JIS X 0213 文字セットによる真のカバー率を求める

ことはできない。しかし、今回の調査目的は、近代語文献の電子テキスト化における JIS X 0213 文字セットの有効性について見通しを得ることであるから、『太陽コーパス』の開発において、JIS X 0208 文字セットで表現できないとされた文字だけを調査対象としても、調査目的が達せられるものと判断した。

さて、外字、踊字、合字、小書、各タグの使用状況は表 2 のとおりである。次節以降、各タグについて、JIS X 0213 による再符号化結果を示す。

表 2：『太陽コーパス』の JIS X 0208 外字の内訳

タグ名	のべタグ数
外字	5,507
踊字	18,019
合字	8,554
小書	187
計	32,267

3. 『太陽コーパス』の JIS X 0213 による再符号化

3.1 外字タグ

『太陽コーパス』の外字タグの運用方法は 2 種類ある。一つは代用で、もう一つは本当の外字である。代用は JIS 規格の包摂規準では処理できないが、JIS X 0208 に採録された文字の異体漢字である場合に用いる。外字は JIS X 0208 の文字と異体関係が認められず、全くの別字の場合に用いる。それぞれタグの使用例を示す。なお、タグの属性「文字番号」は文字鏡番号である。

（例 1）【代用】 **凞**

<外字 文字番号="001721">凞</外字>々たる明治二十八年の新旭光は〔t189501〕

（例 2）【外字】 **磅**

内には浩然たる正氣の磅<外字 文字番号="024597">〓</外字>するところ禁ぜんと欲して能はざるあり。〔t189501〕

外字タグの再処理結果をまとめたものが表 3 である。JIS X 0208 を用いたときに外字タ

グで表現したものの 77.7%が、JIS X 0213 を用いると符号化することができる。

表 3：外字タグの再処理結果

	のべ字数	異なり字数
第 2 水準漢字	2	2
第 3 水準漢字	2,685	446
第 4 水準漢字	1,371	426
X0213 非漢字	220	38
（小計）	4,278	912
X0213 外字	1,229	579
計	5,507	1,491

第 2 水準漢字で表現できる文字は、次の 2 字（表 4 参照）である。これらはコーパス開発時の単純なバグであろう。

表 4：第 2 水準漢字で表現できるもの

文字番号	字形	面区点	度数
037880	蹴	1-77-12	1
047275	鷺	1-83-25	1

第 3 水準漢字で表現できる文字は、次の 446 字（表 5 参照）である。『太陽コーパス』での使用度数順に示す。

表 5：第 3 水準漢字で表現できるもの

文字番号	字形	面区点	度数
048824	麁	1-94-86	242
016085	欸	1-86-31	239
042124	雞	1-93-66	138
001721	濕	1-14-55	90
051106	开	1-84-17	71
056009	厲	1-14-84	64
035458	詹	1-92-08	54
050005	咤	1-14-88	37
023439	睜	1-88-85	34
032700	虛	1-91-45	29
053267	噶	1-15-20	25
040346	鉸	1-93-13	24
001244	儞	1-14-45	20
019395	燄	1-87-64	20
041315	閩	1-93-49	19
004276	嘻	1-15-18	17
019890	牖	1-87-69	17
016408	殂	1-86-38	16
024597	礪	1-89-18	16
053201	墓	1-84-88	16
011865	拘	1-84-72	15
012081	拐	1-84-77	15
011617	戝	1-84-66	14

017165	汴	1-86-52	14
027733	縈	1-90-16	14
038785	迤	1-92-52	14
038791	迨	1-92-53	14
000076	丰	1-14-06	13
005113	埈	1-15-47	13
010174	徧	1-84-34	13
015155	楣	1-85-86	13
040272	鈹	1-93-07	13
045469	鬓	1-94-27	13
000774	倘	1-14-30	12
009808	弹	1-84-22	12
018251	潢	1-87-13	12
050268	燮	1-87-67	12
008295	凿	1-47-85	11
010149	徜	1-84-33	11
015163	楨	1-85-88	11
000597	侔	1-14-22	10
001115	傲	1-14-40	10
009610	弈	1-84-19	10
010094	徉	1-84-32	10
027069	糕	1-89-86	10
028039	纍	1-90-24	10
037868	躎	1-92-39	10
045430	鬃	1-94-26	10
059830	笳	1-89-60	10
000471	你	1-14-13	9
002930	匡	1-14-82	9
014552	枘	1-85-54	9
015843	櫧	1-86-25	9
016752	毗	1-86-44	9
019012	烘	1-87-42	9
024232	碘	1-89-06	9
025635	窻	1-89-54	9
028810	翱	1-90-35	9
035370	詎	1-92-04	9
039198	邈	1-92-58	9

003528	咀	1-14-94	8
004394	噯	1-15-23	8
012181	捥	1-84-80	8
014005	晷	1-85-32	8
017526	涇	1-86-75	8
020916	珉	1-87-89	8
021253	璣	1-88-28	8
022383	癙	1-88-53	8
027247	納	1-89-90	8
034513	褰	1-91-84	8
038930	逭	1-92-56	8
003523	咖	1-14-93	7
004407	嗆	1-15-25	7
010496	怳	1-84-45	7
010771	惋	1-84-51	7
010815	惝	1-84-54	7
010949	愜	1-84-56	7
011833	扯	1-84-71	7
012105	挹	1-84-78	7
021062	琦	1-88-06	7
021270	璨	1-88-31	7
026734	籙	1-89-79	7
028952	耦	1-90-38	7
031565	蒞	1-91-13	7
032805	虬	1-91-50	7
040223	鈐	1-93-05	7
059130	溮	1-87-58	7
000418	份	1-14-09	6
004631	囉	1-15-31	6
006297	娓	1-15-81	6
007559	厊	1-47-62	6
010679	悞	1-84-50	6
018164	漪	1-87-06	6
018965	灶	1-87-40	6
024580	礹	1-89-16	6
025458	窅	1-89-50	6
026077	笄	1-89-66	6

027750	縑	1-90-17	6
043909	飂	1-94-07	6
043965	飋	1-94-08	6
001375	兗	1-14-50	5
003341	吧	1-14-86	5
003770	唵	1-15-06	5
003898	喁	1-15-09	5
007047	孽	1-47-55	5
008028	峋	1-47-74	5
010803	惕	1-84-53	5
011015	愷	1-84-59	5
011961	拖	1-84-74	5
012808	擎	1-84-92	5
015076	椵	1-85-83	5
015272	榭	1-85-92	5
017132	汙	1-86-49	5
017186	沅	1-86-54	5
017699	淼	1-86-86	5
018010	滇	1-87-01	5
018948	炫	1-87-39	5
019137	煞	1-87-48	5
021065	琨	1-88-07	5
023689	瞪	1-88-91	5
024175	硎	1-89-02	5
025556	窠	1-89-51	5
026844	敉	1-89-81	5
029700	腭	1-90-51	5
033077	蛺	1-91-54	5
036819	賡	1-92-25	5
039974	醞	1-92-88	5
040280	鉀	1-93-10	5
041383	閻	1-93-52	5
041446	閾	1-93-54	5
042940	鞮	1-93-79	5
042997	鞹	1-93-80	5
043716	顙	1-94-06	5
046597	鱷	1-94-55	5

056060	煒	1-87-54	5
000572	佾	1-14-20	4
000601	侗	1-14-23	4
004403	嘍	1-15-24	4
005179	埭	1-15-50	4
005558	墝	1-15-67	4
006002	奭	1-15-74	4
008106	峴	1-47-77	4
008488	嶠	1-47-89	4
009398	庾	1-84-13	4
012778	撾	1-84-93	4
014224	曛	1-85-42	4
015679	櫹	1-86-22	4
017319	泫	1-86-62	4
017609	涿	1-86-80	4
017639	淖	1-86-82	4
018328	淅	1-87-16	4
018833	灤	1-87-35	4
019883	牏	1-87-68	4
021865	畯	1-88-42	4
023307	眴	1-88-80	4
023310	眶	1-88-81	4
024110	砭	1-88-93	4
026313	篙	1-89-70	4
027246	紇	1-89-89	4
027371	絇	1-90-01	4
028727	翟	1-90-32	4
037452	跎	1-92-33	4
039301	邢	1-92-63	4
040365	銈	1-93-14	4
040857	鏽	1-93-39	4
041740	隄	1-93-60	4
047012	鵑	1-94-62	4
056422	鍥	1-93-25	4
056428	鏤	1-93-41	4
057094	氳	1-86-48	4
000420	仿	1-14-10	3

001895	刓	1-14-60	3
001999	剉	1-14-62	3
003422	呿	1-14-87	3
004797	圄	1-15-33	3
004889	圯	1-15-36	3
005167	埤	1-15-49	3
008548	嶸	1-47-92	3
008622	巂	1-47-93	3
009844	彀	1-84-25	3
010354	忡	1-84-40	3
011201	憍	1-84-61	3
012624	摭	1-84-91	3
013862	昱	1-85-21	3
013939	晙	1-85-27	3
013991	晢	1-85-31	3
015094	楂	1-85-84	3
017168	汶	1-86-53	3
017323	泮	1-86-63	3
017408	洮	1-86-67	3
017809	湄	1-86-89	3
018026	滎	1-87-02	3
018174	漳	1-87-08	3
018319	澈	1-87-15	3
018814	灞	1-87-33	3
020845	玕	1-87-83	3
021324	璨	1-88-35	3
021717	甯	1-88-41	3
022979	盌	1-88-72	3
026032	筠	1-89-63	3
026592	簇	1-89-75	3
027087	糙	1-89-87	3
027757	緼	1-90-18	3
028663	翎	1-90-30	3
029614	腊	1-90-47	3
030606	艷	1-90-60	3
034220	袪	1-91-73	3
034285	裊	1-91-74	3
034499	褧	1-91-83	3
037646	踠	1-92-36	3
038412	輗	1-92-46	3
038542	轔	1-92-48	3
038700	辵	1-92-51	3
040475	鋠	1-93-19	3
040640	鍪	1-93-30	3
042575	靚	1-93-75	3
042728	靳	1-93-77	3
043047	韁	1-93-81	3
045985	魦	1-94-34	3
046105	魬	1-94-40	3
047714	鱻	1-94-76	3
048480	鼳	1-94-84	3
048886	鯀	1-94-89	3
053113	儆	1-14-42	3
056019	壏	1-15-65	3
000497	佈	1-14-14	2
002497	勻	1-14-75	2
003590	哆	1-15-02	2
003908	喆	1-15-10	2
004175	嘈	1-15-16	2
004299	噉	1-15-19	2
004562	嚪	1-15-29	2
005190	埵	1-15-51	2
005592	壟	1-15-69	2
006206	姝	1-15-80	2
006469	媮	1-15-86	2
006533	媧	1-15-89	2
006941	孖	1-47-54	2
008846	帗	1-84-09	2
008851	帝	1-84-10	2
009058	帾	1-84-11	2
009079	幙	1-84-12	2
010529	恇	1-84-47	2
011940	拄	1-84-73	2
012125	掅	1-84-79	2

012494	搢	1-84-87	2
012912	擻	1-85-05	2
013040	攇	1-85-07	2
013796	昉	1-85-13	2
013903	晌	1-85-25	2
013952	晡	1-85-29	2
016275	歧	1-86-36	2
017369	泊	1-86-66	2
017412	洱	1-86-68	2
018416	澶	1-87-21	2
018939	炤	1-87-38	2
021018	琊	1-88-02	2
021067	琪	1-88-08	2
021102	琦	1-88-16	2
021361	瓚	1-88-37	2
022395	痤	1-88-54	2
022529	癌	1-88-58	2
022601	癎	1-88-59	2
022630	癭	1-88-61	2
024586	磐	1-89-17	2
027276	紓	1-89-91	2
027485	紲	1-90-06	2
027856	絲	1-90-20	2
032418	蘄	1-91-38	2
033079	蜳	1-91-55	2
033372	螈	1-91-60	2
033578	蠥	1-91-65	2
034523	褲	1-91-85	2
034544	褖	1-91-87	2
035069	觔	1-91-91	2
039658	鄱	1-92-82	2
040173	鈁	1-92-94	2
040446	鋋	1-93-16	2
040808	鏽	1-93-36	2
041052	鑩	1-93-44	2
041451	閬	1-93-55	2
044912	騽	1-94-16	2

045657	鬮	1-94-31	2
046171	鯁	1-94-42	2
046803	鶚	1-94-57	2
047034	鷂	1-94-63	2
050918	丰	1-14-05	2
053621	邝	1-92-61	2
056156	茁	1-90-76	2
056180	荇	1-90-82	2
056254	蒇	1-91-11	2
056374	藿	1-91-37	2
056390	蘸	1-91-44	2
057018	櫻	1-86-19	2
058024	檠	1-86-03	2
000382	仡	1-14-08	1
000499	佔	1-14-15	1
000789	倥	1-14-31	1
001846	刁	1-14-58	1
001899	刢	1-14-61	1
002781	卡	1-14-79	1
003302	吒	1-14-85	1
003446	呫	1-14-89	1
003540	咩	1-15-01	1
003755	唅	1-15-05	1
005419	墉	1-15-60	1
005470	墩	1-15-63	1
005747	夔	1-15-72	1
006329	娣	1-15-82	1
007098	宓	1-47-56	1
007257	寘	1-47-57	1
008010	岑	1-47-73	1
008209	崧	1-47-81	1
008502	嶧	1-47-91	1
009972	彤	1-84-29	1
009983	彧	1-84-30	1
010498	怵	1-84-46	1
012041	抃	1-84-76	1
012596	揔	1-84-90	1

012787	擄	1-85-01	1
012900	擷	1-85-04	1
013836	昺	1-85-16	1
013852	昀	1-85-19	1
013860	昱	1-85-20	1
014089	暠	1-85-36	1
014451	杈	1-85-50	1
014805	柠	1-85-67	1
015502	槼	1-86-11	1
015556	槪	1-86-15	1
016061	欻	1-86-30	1
016139	歂	1-86-32	1
016750	毖	1-86-43	1
017027	氐	1-86-47	1
017136	汜	1-86-50	1
017289	汿	1-86-60	1
017353	洄	1-86-65	1
017421	洹	1-86-69	1
017578	淬	1-86-79	1
017646	淛	1-86-84	1
017978	湊	1-86-93	1
018139	溫	1-87-04	1
018169	潔	1-87-07	1
018338	澔	1-87-18	1
018811	瀕	1-87-32	1
019174	燀	1-87-51	1
019291	熇	1-87-60	1
019304	燓	1-87-61	1
019519	燾	1-87-65	1
019521	爣	1-87-66	1
020431	狻	1-87-75	1
020643	獐	1-87-80	1
020861	玲	1-87-85	1
020869	玦	1-87-87	1
020874	玫	1-87-88	1
020962	珣	1-87-93	1
021015	琇	1-88-01	1

021071	琮	1-88-11	1
021073	琰	1-88-13	1
021122	瑗	1-88-18	1
021242	璜	1-88-26	1
022152	痎	1-88-46	1
022297	痙	1-88-49	1
022317	瘷	1-88-50	1
022634	癯	1-88-62	1
023167	盼	1-88-77	1
023392	睨	1-88-83	1
023523	睺	1-88-88	1
023541	瞀	1-88-89	1
024147	硃	1-89-01	1
024342	碭	1-89-10	1
024409	磯	1-89-11	1
024545	礨	1-89-15	1
025068	稃	1-89-44	1
025601	竊	1-89-53	1
025998	筎	1-89-62	1
026071	筋	1-89-65	1
026889	粔	1-89-83	1
027101	糝	1-89-88	1
027345	紱	1-89-94	1
027555	綮	1-90-09	1
028454	羌	1-90-28	1
029670	腠	1-90-48	1
029929	膻	1-90-52	1
032512	蘩	1-91-43	1
032820	虵	1-91-51	1
033494	螭	1-91-62	1
034292	裎	1-91-75	1
034457	褚	1-91-82	1
034816	覓	1-91-88	1
035968	譔	1-92-18	1
035976	譙	1-92-19	1
036404	豨	1-92-23	1
036878	賸	1-92-27	1

037473	跗	1-92-35	1
037887	蹐	1-92-40	1
039413	郤	1-92-69	1
039476	郷	1-92-72	1
039630	鄧	1-92-80	1
039684	鄴	1-92-83	1
040447	鋌	1-93-17	1
040940	鐳	1-93-40	1
041022	鑢	1-93-42	1
041053	鑲	1-93-45	1
041283	閧	1-93-47	1
041330	閥	1-93-50	1
041367	閶	1-93-51	1
041430	関	1-93-53	1
041467	闥	1-93-57	1
041650	陘	1-93-59	1
041849	隕	1-93-63	1
042230	雯	1-93-69	1
043191	韲	1-93-83	1
043269	師	1-93-85	1
043357	頊	1-93-87	1
043463	類	1-93-89	1
043599	顯	1-93-92	1
043600	顳	1-93-93	1
043614	顼	1-94-01	1
044779	駸	1-94-12	1
044780	駿	1-94-13	1

045597	鬻	1-94-29	1
045969	魛	1-94-32	1
046046	鯱	1-94-37	1
046071	鮭	1-94-39	1
046249	鰈	1-94-44	1
047204	鷁	1-94-66	1
048837	龔	1-94-87	1
050021	墜	1-87-25	1
053024	劑	1-91-14	1
053212	异	1-84-18	1
056028	廋	1-84-15	1
056061	燁	1-87-62	1
056068	璘	1-88-25	1
056083	磷	1-89-14	1
056138	芾	1-90-69	1
056142	苕	1-90-72	1
056206	荸	1-90-90	1
056209	莉	1-90-91	1
056220	若	1-91-03	1
056272	萌	1-91-15	1
056273	蒺	1-91-16	1
056380	蔑	1-91-40	1
056382	蘘	1-91-42	1
057035	機	1-86-21	1
058230	鯽	1-94-46	1
065501	籥	1-89-78	1

第 4 水準漢字で表現できる文字は、次の 426 字（表 6 参照）である。『太陽コーパス』での使用度数順に示す。

表 6：第 4 水準漢字で表現できるもの

文字番号	字形	面区点	度数
018019	瀛	2-79-06	252
034969	覿	2-88-42	21
025263	穌	2-83-03	17
056318	壹	2-86-81	17

057871	熳	2-80-01	13
028121	欽	2-84-66	12
056006	僊	2-03-04	12
006573	媳	2-05-70	11
013431	斃	2-13-72	11

000763	倏	2-01-57	10
008848	帕	2-08-83	10
032958	蛢	2-87-41	10
039805	酧	2-90-33	10
025324	穠	2-83-08	9
026062	筩	2-83-48	9
033703	蠲	2-87-92	9
039044	遝	2-89-93	9
007253	寖	2-08-07	8
012586	撜	2-13-41	8
012932	攇	2-13-58	8
019727	爹	2-80-13	8
025582	窯	2-83-17	8
041329	闆	2-91-56	8
050013	健	2-12-24	8
056286	蓯	2-86-65	8
056355	蘼	2-87-04	8
005805	貪	2-05-29	7
011826	扭	2-12-93	7
012694	撑	2-13-47	7
013406	斀	2-13-70	7
018777	灂	2-79-53	7
028367	屬	2-84-80	7
038473	轀	2-89-67	7
041601	陋	2-91-67	7
059756	焰	2-79-88	7
001798	凳	2-03-20	6
002689	匾	2-03-48	6
004502	嚕	2-04-45	6
011562	戕	2-12-83	6
016647	殺	2-78-04	6
018092	滹	2-79-10	6
024043	砐	2-82-28	6
028971	耬	2-85-09	6
029758	瞍	2-85-45	6
036786	賒	2-89-12	6
037149	趕	2-89-23	6

040542	錕	2-91-07	6
041362	閬	2-91-57	6
056323	蕞	2-86-82	6
000964	傖	2-01-77	5
003874	啡	2-04-08	5
004350	唒	2-04-39	5
008458	嶒	2-08-63	5
008477	嶙	2-08-66	5
011216	慦	2-12-72	5
015945	欄	2-15-85	5
018062	滙	2-79-07	5
019179	煊	2-79-84	5
023532	睽	2-82-11	5
028167	罄	2-84-70	5
032882	蚜	2-87-34	5
033873	鐲	2-88-02	5
033968	邮	2-88-04	5
037648	踢	2-89-38	5
041425	鬧	2-91-59	5
044024	飣	2-92-46	5
044179	餇	2-92-55	5
045359	髡	2-93-19	5
051023	壳	2-05-22	5
002281	劇	2-03-32	4
005746	夔	2-05-28	4
006527	媢	2-05-68	4
007811	屧	2-08-20	4
011960	挖	2-13-03	4
014483	杓	2-14-34	4
015347	橐	2-15-30	4
018239	澅	2-79-21	4
020774	獷	2-80-55	4
021777	吠	2-81-26	4
027368	紟	2-84-17	4
027532	網	2-84-33	4
027775	縝	2-84-51	4
032833	虺	2-87-29	4

034257	裀	2-88-18	4
034299	裒	2-88-19	4
035873	謗	2-88-73	4
035934	謹	2-88-74	4
037750	躅	2-89-44	4
039260	邐	2-90-04	4
040001	醨	2-90-40	4
040957	鑲	2-91-44	4
042577	靛	2-91-94	4
042957	鞘	2-92-08	4
043881	颼	2-92-35	4
043921	颶	2-92-38	4
044278	餂	2-92-63	4
044453	饜	2-92-73	4
048477	饑	2-94-69	4
053339	瞢	2-82-16	4
058529	蠒	2-87-81	4
065717	社	2-88-09	4
065718	衧	2-88-10	4
000680	俏	2-01-52	3
003579	咿	2-03-85	3
003808	啊	2-04-05	3
004974	坨	2-04-68	3
006214	姣	2-05-49	3
006215	姤	2-05-50	3
006720	嫣	2-05-81	3
006776	嬴	2-05-84	3
006932	孑	2-05-87	3
007058	宄	2-05-93	3
007804	屧	2-08-19	3
008172	崐	2-08-45	3
009125	幫	2-08-92	3
011158	慼	2-12-68	3
012096	挲	2-13-11	3
014580	枰	2-14-44	3
014759	桅	2-14-64	3
015449	椹	2-15-45	3

015660	檑	2-15-62	3
016856	搓	2-78-12	3
020286	犹	2-80-30	3
020895	玷	2-80-66	3
022453	瘳	2-81-69	3
025126	稞	2-82-92	3
025561	窣	2-83-16	3
029175	瞶	2-85-14	3
030449	艋	2-85-75	3
032583	蘼	2-87-21	3
033147	蝑	2-87-54	3
033202	蜺	2-87-58	3
033320	蝌	2-87-66	3
035660	詮	2-88-65	3
037865	蹯	2-89-49	3
044316	餼	2-92-67	3
045653	鬮	2-93-28	3
045681	虙	2-93-29	3
045861	魋	2-93-32	3
046890	鵅	2-94-14	3
047195	鸙	2-94-32	3
048261	龕	2-94-62	3
057479	鼯	2-94-68	3
057773	艉	2-85-72	3
078863	蹂	2-89-31	3
002190	劂	2-03-30	2
003701	唉	2-03-94	2
003910	喈	2-04-13	2
004069	嗒	2-04-16	2
005253	堠	2-05-01	2
006232	姮	2-05-52	2
006568	嫮	2-05-69	2
006701	嫵	2-05-80	2
007051	孿	2-05-91	2
009127	幬	2-08-93	2
010319	忒	2-12-30	2
011106	憿	2-12-67	2

011218	憘	2-12-73	2
011970	挐	2-13-04	2
012359	揜	2-13-31	2
013286	敫	2-13-68	2
014758	桄	2-14-63	2
015295	榼	2-15-27	2
017283	泱	2-78-33	2
017786	淔	2-78-82	2
018920	茨	2-79-63	2
019337	煡	2-79-92	2
019935	物	2-80-18	2
020730	獫	2-80-49	2
022631	癮	2-81-77	2
023466	罼	2-82-07	2
024392	磈	2-82-48	2
024495	磾	2-82-58	2
024713	祋	2-82-70	2
026299	篔	2-83-63	2
026314	篚	2-83-65	2
026458	簏	2-83-71	2
026801	籩	2-83-79	2
027489	綆	2-84-30	2
028006	繾	2-84-58	2
028143	缾	2-84-68	2
030107	臬	2-85-57	2
030564	艣	2-85-82	2
034945	覷	2-88-41	2
035344	訾	2-88-57	2
036168	讜	2-88-84	2
036207	谿	2-88-88	2
038099	躶	2-89-55	2
039794	酗	2-90-32	2
040296	鉎	2-90-60	2
040937	鐲	2-91-42	2
041945	雋	2-91-79	2
043812	飇	2-92-33	2
044199	餚	2-92-58	2
044480	饢	2-92-74	2
044967	驄	2-93-03	2
045535	髺	2-93-24	2
045543	鬒	2-93-25	2
047507	鸝	2-94-49	2
048274	鼁	2-94-63	2
048361	鼙	2-94-67	2
053340	慒	2-12-81	2
053365	潛	2-79-24	2
056074	瘼	2-81-68	2
056234	萹	2-86-38	2
056311	蘐	2-86-74	2
056387	藱	2-87-18	2
056394	虌	2-87-23	2
056433	轐	2-92-10	2
056462	棻	2-14-86	2
057354	箐	2-83-52	2
057486	酁	2-90-37	2
066039	鮅	2-93-77	2
000107	乀	2-01-08	1
000510	佔	2-01-36	1
000725	傛	2-01-55	1
000843	偎	2-01-65	1
001051	僇	2-01-85	1
001189	儈	2-03-01	1
001797	兢	2-03-19	1
002307	劦	2-03-35	1
002380	勅	2-03-41	1
003254	匝	2-03-67	1
003387	呃	2-03-72	1
003437	呦	2-03-74	1
003535	咦	2-03-80	1
003585	哃	2-03-86	1
003790	啁	2-04-02	1
003804	啇	2-04-04	1
003889	噂	2-04-11	1
004138	噆	2-04-27	1

004194	嘐	2-04-31	1
004625	瞲	2-04-51	1
004926	垒	2-04-65	1
004988	坳	2-04-70	1
005248	堞	2-04-94	1
005433	墔	2-05-14	1
005698	夆	2-05-25	1
006461	婺	2-05-63	1
006575	媵	2-05-71	1
006600	媿	2-05-73	1
006655	嫈	2-05-76	1
006734	嬈	2-05-82	1
006912	孌	2-05-86	1
007433	尃	2-08-13	1
008186	崛	2-08-46	1
008267	崔	2-08-53	1
008532	嵩	2-08-68	1
008549	嶂	2-08-71	1
008702	巋	2-08-77	1
008843	帑	2-08-82	1
009134	幮	2-12-01	1
009311	庶	2-12-03	1
010320	忏	2-12-31	1
010403	忼	2-12-37	1
010661	悕	2-12-46	1
010718	悰	2-12-49	1
010933	愒	2-12-59	1
011021	惮	2-12-63	1
011244	憨	2-12-75	1
012013	拽	2-13-05	1
012050	挍	2-13-07	1
012054	挐	2-13-08	1
012101	挹	2-13-12	1
012127	捄	2-13-16	1
012148	捎	2-13-17	1
012238	掄	2-13-23	1
012265	挣	2-13-24	1
012510	掐	2-13-40	1
012587	摛	2-13-42	1
012678	撇	2-13-46	1
012802	擋	2-13-50	1
012857	擃	2-13-55	1
014013	暔	2-14-09	1
014266	曬	2-14-21	1
014495	柿	2-14-36	1
014936	棖	2-14-94	1
014990	概	2-15-05	1
015378	槳	2-15-39	1
015851	櫬	2-15-75	1
015868	櫳	2-15-78	1
016517	殛	2-78-01	1
016983	穖	2-78-15	1
017085	氵	2-78-17	1
017157	汭	2-78-24	1
017320	沆	2-78-39	1
017397	泊	2-78-45	1
017398	洧	2-78-46	1
017546	涘	2-78-59	1
017593	涴	2-78-67	1
017827	湏	2-78-88	1
018177	滤	2-79-16	1
018489	澡	2-79-36	1
018709	瀹	2-79-46	1
018881	灾	2-79-59	1
019018	烜	2-79-73	1
019069	焄	2-79-75	1
019210	煞	2-79-86	1
019410	燋	2-80-03	1
019774	牉	2-80-15	1
020430	狟	2-80-36	1
020495	狰	2-80-40	1
020646	獘	2-80-47	1
020754	獯	2-80-53	1
020804	獵	2-80-56	1

020857	玦	2-80-62	1
021060	玲	2-80-78	1
021248	琉	2-81-02	1
022154	痄	2-81-46	1
023224	眴	2-81-91	1
023867	稍	2-82-20	1
024396	碘	2-82-49	1
024399	磎	2-82-50	1
024643	祊	2-82-65	1
024693	祧	2-82-69	1
025407	穸	2-83-11	1
025440	牢	2-83-13	1
026057	笂	2-83-47	1
026136	箑	2-83-53	1
027370	絀	2-84-18	1
027377	絇	2-84-19	1
027438	綑	2-84-25	1
027636	綀	2-84-41	1
027660	緦	2-84-43	1
027854	繅	2-84-55	1
027960	繳	2-84-56	1
028027	纇	2-84-60	1
028370	罋	2-84-81	1
028639	翍	2-84-90	1
028860	耄	2-85-03	1
028880	耑	2-85-06	1
029508	脞	2-85-33	1
030206	舄	2-85-62	1
030438	舿	2-85-73	1
030477	艏	2-85-77	1
032852	蚍	2-87-32	1
033137	蜙	2-87-53	1
033208	蜾	2-87-59	1
033213	蝀	2-87-60	1
033268	蝘	2-87-63	1
033570	蠴	2-87-80	1
033682	蠻	2-87-90	1

033745	蠓	2-87-93	1
034106	衱	2-88-11	1
034353	裱	2-88-25	1
035750	諼	2-88-66	1
036120	譻	2-88-81	1
036435	豭	2-89-03	1
037475	跙	2-89-28	1
037617	踔	2-89-35	1
038190	軔	2-89-59	1
038721	迁	2-89-76	1
038988	遘	2-89-92	1
039822	酡	2-90-34	1
039926	酟	2-90-39	1
040031	醮	2-90-41	1
040106	釃	2-90-44	1
040184	釨	2-90-50	1
040205	鈁	2-90-51	1
040216	鈇	2-90-52	1
040291	鉊	2-90-59	1
040351	鉼	2-90-71	1
040506	鋹	2-91-03	1
040556	錝	2-91-08	1
040642	鏊	2-91-18	1
040770	鏐	2-91-32	1
041663	陡	2-91-68	1
041895	隟	2-91-77	1
042864	鞚	2-92-05	1
043917	颿	2-92-37	1
043964	颱	2-92-41	1
043973	颺	2-92-42	1
044022	飧	2-92-45	1
044202	餛	2-92-59	1
044637	駉	2-92-82	1
044665	駔	2-92-83	1
044677	駙	2-92-84	1
044928	騶	2-93-02	1
045124	骯	2-93-07	1

046431	鰺	2-93-73	1
046432	鰌	2-93-74	1
046626	鱶	2-93-94	1
046667	鳰	2-94-02	1
047022	鵐	2-94-23	1
047208	鷯	2-94-34	1
047388	鷸	2-94-44	1
047733	麹	2-94-55	1
048278	龕	2-94-65	1
048351	鑿	2-94-66	1
048519	齣	2-94-72	1
048543	齦	2-94-73	1
048590	齟	2-94-76	1
050083	吕	2-08-79	1
050925	亹	2-01-20	1
051896	帮	2-08-86	1
053107	偶	2-01-59	1
053464	嬝	2-05-74	1
053509	槃	2-86-79	1
053616	贏	2-87-91	1
053620	鄺	2-90-29	1
054860	牖	2-80-17	1
056042	暐	2-14-11	1
056043	曙	2-14-16	1

056062	熱	2-80-07	1
056069	厩	2-81-11	1
056112	芃	2-85-90	1
056116	芋	2-85-91	1
056179	茨	2-86-16	1
056229	荏	2-86-34	1
056281	蘋	2-86-59	1
056326	蕢	2-86-83	1
056375	薜	2-87-13	1
056458	鱚	2-93-84	1
057196	耆	2-82-32	1
057313	聃	2-85-11	1
057315	聱	2-85-13	1
057368	籯	2-83-81	1
057591	虓	2-87-25	1
057637	髀	2-93-15	1
057641	骷	2-93-09	1
057850	蕨	2-86-29	1
065806	欽	2-90-48	1
067184	餈	2-92-61	1
079011	鰥	2-93-71	1
079145	路	2-89-41	1
079566	澎	2-79-39	1

X0213 非漢字で表現できる文字は、次の 38 字（表 7 参照）である。『太陽コーパス』での使用度数順に示す。

表 7：X0213 非漢字で表現できるもの

文字番号	字形	面区点	度数
063004	â	1-09-56	51
063017	é	1-09-63	30
063020	ê	1-09-64	21
063070	ǔ	1-10-68	13
063003	ä	1-09-58	10
063034	î	1-09-68	10
063068	ü	1-09-81	10

062960	Ś	1-10-05	9
063018	è	1-09-62	8
063065	ţ	1-10-55	6
063069	û	1-09-80	5
063051	ö	1-09-76	4
063060	ś	1-10-16	4
069682	ㇵ	1-03-28	4
063002	à	1-09-54	3

063014	ç	1-09-61	3
062833	æ	1-09-60	2
062845	œ	1-11-10	2
063048	ñ	1-09-71	2
063052	ô	1-09-74	2
063053	ǒ	1-08-87	2
063055	ō	1-09-94	2
063221	Ⅰ	1-13-21	2
062588	Ⴝ	1-06-57	1
062952	Ô	1-09-43	1
062968	Ǔ	1-09-50	1
063001	ά	1-09-55	1

063005	ă	1-10-41	1
063009	å	1-09-59	1
063023	ē	1-09-93	1
063030	ĥ	1-10-65	1
063031	í	1-09-67	1
063033	ï	1-09-69	1
063049	ó	1-09-73	1
063066	ú	1-09-79	1
063222	Ⅱ	1-13-22	1
063223	Ⅲ	1-13-23	1
063224	Ⅳ	1-13-24	1

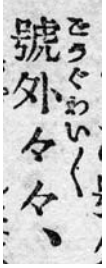
3.2　踊字タグ

JIS X 0208 には「二の字点」や「くの字点」がないため、『太陽コーパス』では「々」「ゝ」で代用し、踊字タグを使って表現している。次の例は、平仮名「る」の繰り返し符号に使われた「二の字点」を「ゝ」で代用した例である。

（例 3）　意を戦況に注がる<踊字　種類="二字点">ゝ</踊字>の致す處なりと雖も、〔t189501〕

また、繰り返し符号が複数連続する場合にも踊字タグが使われている。この場合は、JIS X 0208 で表現できないからタグを用いているわけではなく、繰り返し符号で表現された語が何であるのかを、タグの属性で示すのが目的である。

（例 4）　號外<踊字　値="號外">々々</踊字>〔t189511〕



踊字タグの再処理結果をまとめると表 8 のようになる。JIS X 0213 を用いると、『太陽コーパス』に出現するすべての繰り返し符号を表現することができる。なお、X0208 非漢字に分類されたものは、（例 4）に類するものである。

表 8：踊字タグの再処理結果

	のべ字数	異なり字数	用例
X0208 非漢字	1,444	4	々、ゝ、ゝ、ゞ
X0213 非漢字	16,575	4	二の字点、くの字点（上、上濁点、下）
X0213 外字	0	0	
計	18,019	8	

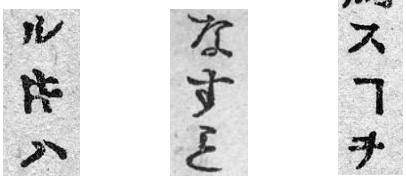
3.3　合字タグ

『太陽コーパス』では、仮名合字等に対して、合字が表わす語形を入力し、合字タグを付与している。合字タグの使用例を示す。

（例 5）　郵便切手ヲ代用スル<合字>トキ</合字>ハ五厘又ハ壹錢切手ニ限ル〔t189501〕

（例 6）　其用意をなす<合字>こと</合字>肝要なるべし。〔t189501〕

（例 7）　條約締盟國ハ日本國中何レノ地ニ於テモ通商貿易ヲ爲ス<合字>コト</合字>ヲ得領事裁判權ハ三年以内ニ之ヲ廢止シ〔t189501〕



合字タグの再処理結果をまとめると表 9 のようになる。『太陽コーパス』に出現する合字に対しては、JIS X 0213 はあまり効果がない。

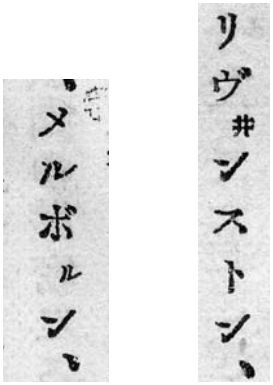
表 9：合字タグの再処理結果

	のべ字数	異なり字数	用例
X0213 非漢字	23	2	コト、より
X0213 外字	8,531	5	こと、トキ、トモ、かしこ、まゐらせ候
計	8,554	7	

3.4　小書タグ

『太陽コーパス』では、主として小書きの仮名を表わすために、小書タグを設けている。外来語表記のための片仮名が目立つ。

（例 8） メルボ<小書>ル</小書>ン〔t189510〕
（例 9） 亞弗利加の中心に深進せしリヴ<小書>ヰ</小書>ンストン〔t189510〕



小書タグの再処理結果をまとめると表 10 のようになる。小書タグのうち、JIS X 0213 を用いて符号化できるものは、のべ字数で 3 割程度にとどまる。なお、X0208 非漢字に分類した小書きのワは、コーパス開発時の単純なバグであろう。

表 10：小書タグの再処理結果

	のべ字数	異なり字数	用例
X0208 非漢字	1	1	ワ
X0213 非漢字	56	7	ハ、ヒ、フ、ヘ、ホ、ム、ル
X0213 外字	130	8	キ、コ、ヰ、ヱ、ヲ、八、忘、●
計	187	12	

なお、片仮名以外の文字について小書タグを使ったものには、次のようなものがある。（例 10）は伏せ字、（例 11）は割注に相当するものと思われる。

（例 10） 亞細亞洲にては●●、●●に在り、近南洋にては●<小書>●</小書>●<小書>●</小書>●●群島、●●大島に在り、〔t189510〕
（例 11） 大兒子、小兒子、大則以王、小則以覇、大小王<小書>忘</小書>覇<小書>八</小書>兒子〔t190114〕

4．おわりに

『太陽コーパス』の文字関連タグに対する再処理結果を加え、JIS X 0213 による符号化をまとめると表 11 のようになる。のべ字数では、JIS X 0208 によるカバー率が 99.79%であったのに対して、JIS X 0213 では 99.93%となる。カバー率の増

加はわずかに 0.15 ポイント程度である。しかし、異なり字数では、JIS X 0208 のカバー率 79.16%に対して、JIS X 0213 は 92.06%と、12.90 ポイントも増加している。

『現代日本語書き言葉均衡コーパス』の書籍（生産実態サブコーパスと流通実態サブコーパス）では、JIS X 0208→JIS X 0213 によるカバー率の増加は、のべ字数で平均 0.03 ポイント、異なり字数で平均 7.75 ポイントであるから、現代語を扱うときよりも、JIS X 0213 を用いる効果が大きい。『太陽コーパス』のような近代語文献を電子化する際には、JIS X 0208 文字セットよりも、JIS X 0213 文字セットの方が適していると言えよう。

表 11：『太陽コーパス』の JIS X 0213 による符号化

	のべ字数	異なり字数
第 1 水準漢字	5,650,619	2,721
第 2 水準漢字	888,886	2,864
X0208 非漢字	7,798,780	318
第 3 水準漢字	2,685	446
第 4 水準漢字	1,371	426
X0213 非漢字	16,874	51
（小計）	14,359,215	6,826
X0213 外字	9,890	592
計	14,369,105	7,418

最後に、JIS X 0213 文字セットで表現できない 592 字を、『太陽コーパス』での使用度数順に表 12 に示す。

表 12：JIS X 0213 外字

文字番号	字形	度数
合字こと		8,487
小書ヰ		105
T003	苍	62
047550	𪜈	55
019790	牒	32
067692	𣏌	27
058594	颺	24
合字まゐらせ候		22
020571	猪	21
012384	揪	20

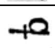
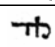
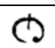
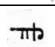
文字番号	字形	度数
073740	璨	18
057321	蚋	17
058090	圻	14
T014	纊	14
合字トキ		14
027431	紙	11
044757	騙	11
003007	厘	10
007255	竄	10
011016	惣	10
038183	軏	10

020803	玀	9
小書ヲ		9
010846	鬼	8
015158	椴	8
020483	猓	8
014585	枴	7
063059	r	7
小書ヱ		7
004658	嚩	6
021193	璀	6
065520	縩	6
074686	滀	6
T025	閧	6
013004	攪	5
020687	獞	5
059905	嶚	5
066170	厄	5
074461	誓	5
小書●		5
合字トモ		5
001706	漸	4
002225	劊	4
011737	扃	4
025000	秖	4
028693	翎	4
030766	芾	4
034535	褻	4
035801	謏	4
037152	趍	4
037993	躝	4
042302	霉	4
044479	饞	4
048306	鼉	4
048652	鮂	4
050664	麦	4
058129	耦	4
077953	晉	4

080074	洌	4
000575	佚	3
001171	僾	3
004183	嘍	3
004259	嘵	3
004981	坭	3
005403	塿	3
007191	寘	3
009463	廕	3
012241	掇	3
012712	撙	3
013751	旰	3
015675	櫆	3
018819	灡	3
019378	爦	3
020335	狯	3
020672	獎	3
024552	礉	3
030411	舺	3
033980	衅	3
034974	覷	3
039942	醬	3
043609	賴	3
044963	鷙	3
056123	芠	3
059815	碉	3
071309	垔	3
071516	粵	3
071768	蔡	3
075506	暺	3
083904	衙	3
T002	啖	3
合字かしこ		3
001134	夾	2
001584	冢	2
001953	刱	2
002041	剚	2

003478	咂	2
003673	哜	2
003806	啉	2
004202	嗎	2
004242	嘮	2
004252	嚫	2
004455	嚅	2
005133	垐	2
005261	堦	2
005404	墀	2
007285	寠	2
008263	嵉	2
008753	啻	2
008760	啻	2
011115	慦	2
011525	戇	2
012124	捂	2
012505	搪	2
012595	摟	2
012728	撦	2
012897	擰	2
012991	攙	2
013668	旐	2
013984	晾	2
014070	暯	2
014163	暟	2
015492	毃	2
017171	汸	2
017492	洚	2
017784	淯	2
019202	煞	2
019430	燇	2
020274	狉	2
020336	狹	2
020377	狗	2
020722	獷	2
020961	琅	2

022784	皤	2
023619	瞕	2
023633	瞪	2
024669	袂	2
025778	娖	2
026318	筃	2
026455	簌	2
027483	綁	2
027831	繁	2
028824	翾	2
029426	胭	2
032114	蘼	2
032411	藁	2
033770	蠹	2
035233	訐	2
035466	註	2
036843	賬	2
037787	蹏	2
039791	酗	2
040185	釭	2
040350	銘	2
040514	錁	2
041574	阱	2
044678	駃	2
044710	駧	2
045468	鬍	2
045524	鬣	2
047440	鸉	2
049191	齾	2
049534	胳	2
050020	灣	2
050031	贊	2
056037	撐	2
056114	芾	2
056412	薆	2
057125	螣	2
057128	螣	2

059535	⿰山角	2
059665	膀	2
059875	舿	2
062581	α̇	2
067039	⿱艹泣	2
067352	搉	2
077217	遲	2
081515	⿱髟亏	2
085829	⿱竹羿	2
T007		2
T016		2
T020		2
T105		2
T117		2
T125		2
T126		2
T135		2
T136		2
T137		2
T138		2
T146		2
000484	伻	1
001119	債	1
001154	偏	1
001505	顚	1
001955	封	1
002094	剮	1
002217	剝	1
002683	⿷匚旨	1
002854	卲	1
003235	另	1
003322	呷	1
003368	吵	1
003382	⿰口帀	1
003477	咁	1
003628	哎	1
003756	唬	1
003811	⿰口昏	1
003940	喔	1
004041	嗃	1
004139	嗶	1
004190	嘏	1
004195	⿰口聿	1
004348	噙	1
004979	坫	1
005080	垧	1
005208	塑	1
005323	塍	1
005326	塏	1
005425	⿰土鄉	1
005467	墦	1
005560	壖	1
006124	⿰女幼	1
006235	姱	1
006361	娬	1
006392	婇	1
006405	婑	1
006579	媸	1
006787	嬝	1
007957	岸	1
007958	岞	1
008175	崒	1
008326	⿰山冥	1
008367	嵲	1
008391	⿰山帶	1
008460	寧	1
008556	巀	1
008579	巑	1
008758	香	1
009052	幙	1
009081	幠	1
009245	庋	1
009605	弆	1
010394	忬	1

010675	悜	1
010751	愩	1
010766	悽	1
010839	悊	1
010937	愔	1
011167	憀	1
011225	憏	1
011475	憫	1
012131	捆	1
012316	摃	1
012453	摛	1
012491	搠	1
012672	摔	1
012693	撐	1
012951	攉	1
012957	攦	1
013627	旁	1
013664	旃	1
013685	旖	1
013686	旐	1
013691	旚	1
013780	旰	1
013963	晅	1
014045	暌	1
014817	棒	1
015232	橅	1
015356	槤	1
015718	檯	1
015903	欀	1
016017	欴	1
016476	殑	1
017166	汾	1
017169	汊	1
017179	沿	1
017510	浼	1
017594	浮	1
017596	涯	1
017685	淣	1
017714	淩	1
017793	淘	1
017853	湢	1
018256	溧	1
018320	溦	1
018388	澦	1
018392	澨	1
019134	煢	1
019379	燘	1
019896	牘	1
019987	牳	1
020054	犄	1
020067	犏	1
020251	犻	1
020391	狪	1
020394	猗	1
020452	狶	1
020456	狒	1
020487	猏	1
020618	猓	1
020680	猾	1
020692	獠	1
020879	玭	1
020988	琌	1
020990	琅	1
021041	琲	1
021208	瑫	1
021239	瑣	1
021564	瑩	1
021749	畀	1
021828	畟	1
022451	癡	1
022685	皂	1
022787	皾	1
022815	皽	1
023120	盱	1

023178	明	1
023373	睠	1
023555	瞅	1
024088	砢	1
024178	硐	1
024200	硜	1
024249	硢	1
024330	硜	1
024369	碩	1
024568	礧	1
024613	礸	1
024734	禍	1
024737	禍	1
024761	禋	1
025094	稕	1
025368	穱	1
025438	窕	1
025966	筎	1
026109	笄	1
026132	篚	1
026158	篳	1
026309	篛	1
026742	簻	1
026985	糕	1
027439	絛	1
027499	綍	1
027538	綞	1
027701	緼	1
028350	罿	1
028440	狌	1
028919	耟	1
028978	耮	1
029544	脰	1
029582	腎	1
029593	腄	1
029627	腒	1
029683	腠	1

029895	膴	1
030055	臞	1
030202	舄	1
030279	刮	1
030472	艎	1
030818	荼	1
031589	萩	1
031616	蒚	1
032354	蔰	1
032871	蚖	1
032941	蛆	1
033255	蜳	1
033359	蜾	1
033367	蜴	1
033415	蝖	1
033612	蟠	1
033747	蠡	1
034034	衍	1
034339	裍	1
034355	裞	1
034470	褒	1
035019	觚	1
035283	詈	1
035292	訞	1
035384	詖	1
035596	誶	1
035636	諉	1
035656	諕	1
035947	譊	1
035991	麿	1
036050	譸	1
036515	豜	1
036713	貦	1
036747	賚	1
036955	贈	1
036960	賴	1
037004	赮	1

037084	趣	1
037124	趍	1
037303	趫	1
037508	跕	1
037560	跹	1
037568	跲	1
037716	蹀	1
037806	蹚	1
037847	蹶	1
037881	躆	1
037886	蹺	1
037962	躐	1
038414	轄	1
039296	邶	1
039851	酎	1
039874	酺	1
039998	醔	1
040104	醨	1
040162	釗	1
040166	釵	1
040226	釬	1
040253	鉢	1
040262	鈰	1
040338	鈾	1
040369	鋍	1
040429	鋅	1
040437	鋑	1
040474	鋓	1
040489	鋐	1
040798	鏕	1
040865	鏟	1
040921	鏦	1
040938	鐫	1
041008	鐳	1
041045	鑐	1
041078	鑞	1
041221	閈	1
041262	閣	1
041289	閾	1
041428	闉	1
041723	隑	1
041810	隔	1
043414	頒	1
043598	顑	1
043605	顡	1
043813	颭	1
043931	飂	1
043955	飋	1
044084	詐	1
044126	飪	1
044236	餧	1
044294	餵	1
044362	饈	1
044799	騕	1
044847	驄	1
044921	驍	1
044962	驟	1
045305	髗	1
045841	魈	1
046039	鮀	1
046059	鮇	1
046095	鮋	1
046119	鮭	1
046407	鰶	1
046856	鶂	1
047937	鼓	1
048352	鼕	1
049388	耆	1
050022	琍	1
050860	緆	1
050971	嫧	1
053173	摔	1
053192	蓍	1
053438	嫫	1

053607		1
053614		1
054924		1
056017		1
056035		1
056080		1
056118		1
056132		1
056150		1
056248		1
056357		1
056359		1
056398		1
056402		1
056407		1
056410		1
057215		1
057361		1
057461		1
057530		1
057851		1
058025		1
058029		1
058660		1
059216		1
059236		1
059247		1
059271		1
059618		1
059803		1
059921		1
060001		1
060081		1
061327		1
061593		1
062583		1
062587		1
062589		1
063047		1
063084		1
065057		1
065167		1
065183		1
065417		1
065538		1
065646		1
065972		1
066007		1
066167		1
066767		1
067720		1
067843		1
068171		1
068275		1
070734		1
072382		1
072508		1
075331		1
076687		1
077626		1
078439		1
078618		1
079135		1
082432		1
082959		1
085185		1
085262		1
086272		1
086719		1
096371		1
096381		1
096461		1
096551		1
T001		1

T004		1
T005		1
T006		1
T008		1
T009		1
T010		1
T011		1
T012		1
T013		1
T015		1
T017		1
T018		1
T019		1
T021		1
T022		1
T023		1
T024		1
T102		1
T103		1
T104		1
T106		1
T107		1
T108		1
T109		1
T110		1
T111		1
T112		1

T113		1
T114		1
T115		1
T116		1
T118		1
T119		1
T120		1
T121		1
T122		1
T123		1
T124		1
T127		1
T128		1
T129		1
T130		1
T131		1
T132		1
T133		1
T134		1
T139		1
T140		1
T145		1
小書八		1
小書キ		1
小書コ		1
小書忘		1

文　献

池田証寿・白井純・高田智和(2002)「宋版漢字字体の処理」『京都大学大型計算機センター第 69 回研究セミナー報告　東洋学へのコンピュータ利用』pp.49-62

下田正弘・師茂樹(1999)「大正新脩大蔵経データベース（SAT）における外字問題」『人文学と情報処理』25, pp.35-43

須永哲矢・堤智明・高田智和(2011)「明治前期雑誌の異体漢字と文字コード」『人文科学とコンピュータシンポジウム論文集　「デジタル・アーカイブ」再考―いま改めて問う記録・保存・活用の技術―』pp.381-388

高田智和(2002)「漢字処理と『大字典』」、『訓点語と訓点資料』109, pp.99-107
高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也(2009)『JIS X 0213:2004 運用の検証（国立国語研究所内部報告書 LR-CCG-09-01)』国立国語研究所
高田智和(2011)「現代日本語コーパスにおける文字処理」『人間文化研究情報資源共有化研究会報告集』2, pp.31-40
田中牧郎(2005)「漢字の実態と処理の方法」『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―（国立国語研究所報告 122)』博文館新社, pp.271-292
當山日出夫(2009)「『内村鑑三全集』デジタル版の文字処理について」『東洋学へのコンピュータ利用第 20 回セミナー』京都大学人文科学研究所附属漢字情報センター, pp.5-18
富田倫生(2000)「青空文庫と外字」、『人文学と情報処理』26, pp.23-30
安永尚志(1998)『国文学研究とコンピュータ』勉誠社

近代語文献を電子化するための異体字処理

須永　哲矢（国立国語研究所コーパス開発センター）[1]

１．はじめに

文書の電子化にあたっては、もとの文書のどの要素をどこまで再現し、どの要素は再現できなくても良しとするかという処理方針を定めねばならないが、それはその電子化テキストの使用目的による。漢字の字体字形の問題一つをとっても、各字形差を可能な限り正確に表現した方が望ましいとは限らない。「言語研究用のコーパス作成」という場面においては、電子テキスト化はゴールではなく、あくまで研究の手段としての環境整備、という位置づけとなる。言語研究の素材として使用される電子テキストは、言語資料として「読める」こと、語彙等のサンプルが採集できることが重要となる。そのため、外字として処理された文字が多く、表示上「〓」ばかりで「読めない」テキストや、動作環境によっては適切に表示されない文字が含まれるテキスト等は望ましくない。

そこで、近代語コーパス構築の基礎研究としての本研究では、近代語コーパスの試作となる『明六雑誌』電子テキスト化の作業を通じ、言語研究の実用に適した文字処理の在り方を模索することとした。

２．JIS X0213 文字集合と包摂規準

近代語コーパスの試作としての『明六雑誌コーパス』は、JIS X 0213 文字集合に準拠して電子化することとした。『現代日本語書き言葉均衡コーパス』も JIS X 0213 を依拠する文字集合として文字処理が行われ、およそ 5,800 万字の現代日本語コーパスでは、のべ 99.96%の文字が、JIS X 0213 で表現できることが確認されている（高田ほか 2009）。

JIS 規格では、字体字形の差を処理するために「包摂規準」が定められている。JIS X 0213 では連番で１９９の包摂規準が設定されており、包摂規準の範囲内の差異であれば、同一の符号位置の文字として処理することになる。これにより、明確な規準のもとでの文字処理が可能となる。

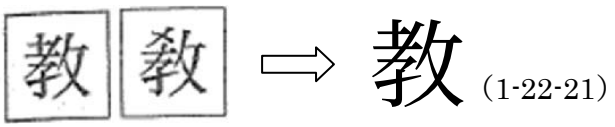


図１　包摂規準の例（連番８）

しかし時代をさかのぼって、近代以前の活字資料を対象とした場合にも JIS X0213 文字集合および包摂規準が有効であるかの検証はいまだなされていない。そこで、明治前期の雑誌である『明六雑誌』のコーパス試作を通じ、JIS X0213 の有効性と限界を見極めたうえで、近代語に適した文字処理方針を構築していくというのが今回の課題である。

３．『明六雑誌』漢字処理上の問題

『明六雑誌』を JIS X0213 に準拠して文字処理していく際、字形処理の実際において問題となるのは大きく分けて次の２つのケースである。

（A）文字集合（ここでは JIS X0213）に含まれない字

図２　『明六雑誌』に出現する JIS 規格外字

[1] tsunaga@ninjal.ac.jp

図２のような文字は、JIS X0213 では用意されておらず、表現することができない。

（B）通用字形とは（僅かな）字形差があるもの

図３『明六雑誌』に出現する「序」「万」「除」の字形（右側）

図３のように、近代の活字では、それが現在の通用字のどの字に当たるかは明らかであるが、字形差があるものが多数見受けられる。JIS 規格では包摂規準が定められているが、図に示した「序」「万」「除」の字形差に関しては、既存の包摂規準の中には明確に適用できるものがない。そのため、既存の包摂規準のみに従って処理していく場合、これらは外字となり、「〓」表示されることになる。

４.『近代語コーパス』のための文字処理方針

近代の活字においては、図３に示した『明六雑誌』での活字のように、既定の包摂規準では包摂してよいのかが明示されていない、わずかな字形の差がある場合が多く見られる。

これらを逐一外字として処理していくと、できあがった電子テキスト内の外字が増え、言語研究資料として実用に供さないものになりかねない。表１に示すとおり、『明六雑誌』の漢字字形に対し、JIS X 0213 の文字集合・包摂規準を適用した場合、その処理だけでのべ約 98.5%が表現可能となる。しかし、言語研究資料としてみた場合、200 文字のうち 3 文字が読めない電子テキストは実用に供さない。

表１　JIS X0213 文字集合・包摂規準を適用して『明六雑誌』の漢字を処理した結果

文字区分	のべ字数	異なり字数
JIS X 0213	135,797	3,218
第 1 水準漢字	117,643	2,066
第 2 水準漢字	17,953	1,061
第 3 水準漢字	118	52
第 4 水準漢字	83	39
外字	2,100	99
計	137,897	3,317
カバー率	98.48%	97.02%

また、『明六雑誌』に出現する字形は、現行の包摂規準だけを拠り所とすると、そのままでは包摂できないものが多く出現するが、その大部分は、現在の通用字体のどれに相当するかは類推でき、字形の差異もわずかなものである。

図４の「〓」表示の内実は、「時」「華」「改」の異体字である。JIS X0213 は、図中の丸囲みのような差異を包摂できる基準を持ち合わせていないため、規格以外字としての扱いとなる。

しかし、このような処理は、JIS X0213 の適用の仕方としては厳密であるが、「〓」表示になった時点で用例としては取り出せなくなってしまうため、用例検索や語彙調査といった、コーパスとしての実用面からは有用性の低い処理になってしまう。むしろ実用面からは、JIS X0213 の適用の仕方が多少ゆるくとも、これらも「時」「華」「改」に包摂し、文字として表示した方が望ましい。

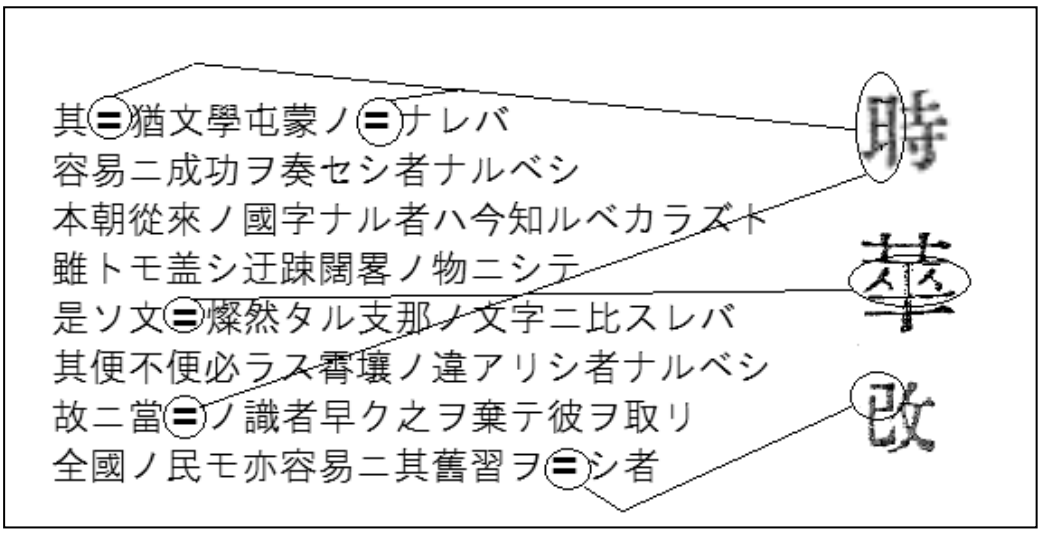


図４　JIS X0213 文字集合・包摂規準を厳密に適用した電子テキスト化の例

「言語研究用コーパス」という目的から求められる漢字処理方針とは、一言で言ってしまえば、可能な限り「〓」表示を減らすこと、つまり、可能な限り読める文字として表現すること、である。しかし、だからといって場当たり的に使える文字を当てていく、というだけでは、作業者によって処理の揺れも生じるうえに、どれが本来の JIS X0213 の範囲で処理したもので、どれが臨時的に処理したものかも後々わからなくなってしまう。そこで、本来の JIS X0213 の範囲を越えた処理をする際には、近代語用の処理基準を設けて、データ上にもタグとして記録を残しておくこととした。

本来 JIS X0213 では外字処理になってしまう文字をもなるべく読める文字として表現する、という目的のもと整備した方針は、大きく以下の２つである。

（１）既存の包摂規準に、近代語用の包摂規準を追加する。

まず、図３、４で示したような近代語特有の差異をカバーするため、既存のＪＩＳ包摂規準に加え、近代語用の追加包摂規準を新設し、その基準に従って字体包摂を行うことで、外字処理を減らす。近代語用に追加した包摂規準によって包摂処理された文字に関しては、タグの形で追加包摂規準により処理されたという情報を埋め込んでおく。

（２）包摂規準の追加では対処にくいものに関しては、別字で代用する。

差異がありすぎる等の理由で、包摂規準の追加では対処しにくい文字に対しても、類似の読みや用法がある文字がある場合、その文字で代用することでコーパス上に表現する。このような代用字に関しても、本来は外字であり別字で代用した、という情報をタグの形で埋め込んでおく。また、どの字をどの字で代用したかの一覧を作成して管理する。

この２つの処理を通して、「〓」表示を極力減らしていくことで、コーパスとしての有用性を高めていけると考える（図５参照）。以上のように追加包摂・別字代用という二つの方策で近代語資料での文字を表現していくという処理は『太陽コーパス』でも採られており、『太陽コーパス』では追加包摂により約３００字、別字代用により約２００字（ともに異なり字数）を処理したという実績がある。ただし、『太陽コーパス』では追加した包摂規準は明示されておらず、実際にどのような字形差を、どのような追加規準で包摂したのかを追跡することはできない。また、別字代用に関しては、情報抽出用アプリケーション『プリズム』を利用して外字一覧を生成することで代用字を閲覧することは可能ではあったが、異なり１０００字を越える「〓」表示の外字とあわせての表示となり、代用情報だけを得るにはやや不便であった。そこで今回の『明六雑誌コーパス』では、追加包摂、

別字代用の処理を行った文字に関してはタグ付けを行い、文字処理の情報を取り出せるようにするとともに、追加包摂規準および別字代用の一実態を一覧として公開することとした。

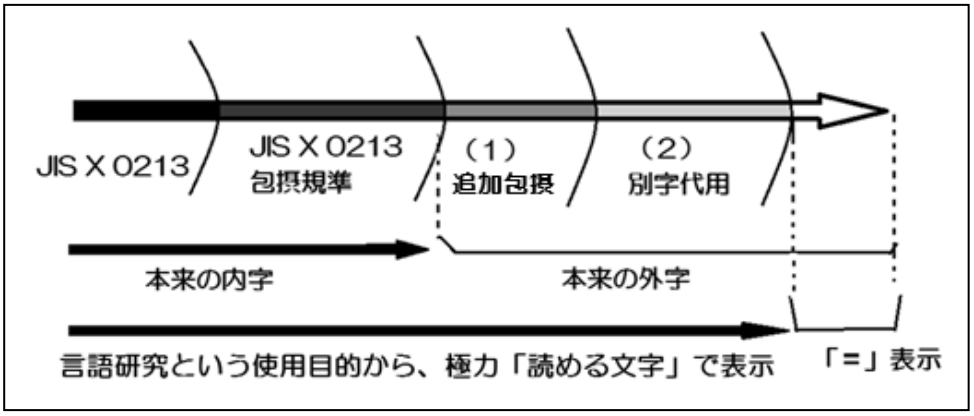


図５　「近代語コーパス」文字処理方針のイメージ

５．『明六雑誌』漢字字形処理方針

近代語文献の文字処理用に追加した包摂規準の詳細、および別字代用の一覧を本節に記す。

５．１ JIS X 0213 文字集合のうち、使用しない領域

今回、『明六雑誌』を JIS X 0213 に準拠して電子化することを試みたが、JIS X 0213 文字集合のうち、使用しない領域を３つ設けたため、ここに記しておく。

①康熙別掲字（104 字）は使用しない。

【例】

×徳(1-84-37) → 徳(1-38-33)を使用

×社(1-89-19) →社(1-28-50)を使用

②UCS 互換字（10 字）は使用しない。

【例】

×叱(1-47-52) → 叱(1-28-24)を使用

×嘘(1-84-07) → 嘘(1-17-19)を使用

康熙別掲字、UCS 互換字は、いわば JIS 包摂規準の例外であり、包摂規準に従うなら、基本的に包摂される字形差である（図６参照）。これらに関しては使用しないこととした。

図６　JIS 包摂規準連番 130、161、78、166

この方針では、本来「德」(1-84-37)で表現できる活字に対しても、包摂規準連番 130 をそのまま適用し、「徳」(1-38-33)として表現することになる。なお、仮に康熙別掲字、UCS 互換字を使用した場合、「德」(1-84-37)と「徳」(1-38-33)がさらに区別されるだけであり、この方針をとらず、康熙別掲字、UCS 互換字まで使用した場合でも、「JIS X0213 で表現される文字の総数」は変わらない。

③CJK 統合漢字拡張Ｂに符号位置が割り当てられる文字（302 字）は使用しない。

【例】

×  (1-15-44、 U+2131B) → 外字扱い

×  (1-15-91、 U+218BD) → 外字扱い

CJK 統合漢字拡張Ｂに関しては、現状では動作環境によっては適切に表示されない等の問題があるため、実用面での判断から使用しない。なお、今回の調査範囲である『明六雑誌』内では、この領域を使えば表現できる漢字は存在しなかったため、この領域を使用した場合でも、『明六雑誌』の範囲内では「JIS X0213 で表現される文字の総数」は変わらない。

５．２ 近代語用包摂基準の設定

JIS X 0213 のうち、上記３領域を除いた文字集合を用いて『明六雑誌』の字形処理を試みることにするが、前述の通り、明治前期の活字字形には、わずかな字形差の活字が多い。それらについては現行の包摂規準には明記されていないものの、感覚的には包摂したいものが多い。そこで、既存の包摂規準を文字処理の規準としたうえで、それに加える形で近代語資料用に包摂規準の拡張案（追加包摂規準）を作成し、字形処理に対応することにした。

近代語での文字処理のため、包摂規準を追加しようという場合、結局のところ、どの程度の字形差までを包摂規準として設定し、どこからを外字とするかが最後まで問題となる。

以下、追加包摂規準の設定のしかた、および追加包摂規準の設定という形では処理しない場合を、具体例と合わせて示す。

５．２．１ 包摂規準を近代語用に追加・修正するケース

（A）既存の基準の明確化

（現行字形）　　（明六雑誌）

(1-43-92)

図７　『明六雑誌』にみられる「万」の字形

このようなパターンについては、漢字字体包摂規準の「ｂ　２点画の接触交差関係の違い」のうち、「抜けるか、抜けないか」のひとつとして処理するという方法が考えられる（図８参照）が、現行の包摂規準内ではこれと完全に一致する字形は示されていない。

このような字形差は、差異の中でも特にわずかな字形差と言いたくなるだろう。漢字の字体字形処理に関しては、JIS 包摂規準以前の前提として、常用漢字表において「デザイン差」とみなされるものは字体の異なりとはしない、という方針があり、そのうち「（４）交わるか、交わらないかに関する例」という例示がなされている（図９参照）。このため、

このような字形差に関しては包摂規準を立てるまでもなく同一字体と処理される、と解釈することもできようが、常用漢字表の「デザイン差」はあくまで例示されるにとどまっており、適用範囲は明確ではない。そこで近代語の電子テキスト化にあたっては、このようなケースに関しても、新たに包摂規準を立て、明確化することとした（図 10）。

図８　包摂規準　連番６０

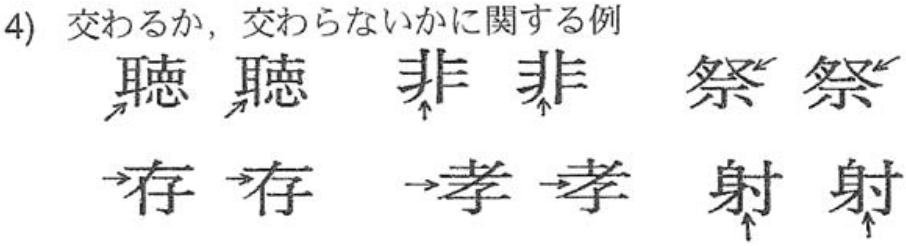


図９　常用漢字表での「デザイン差」字形例

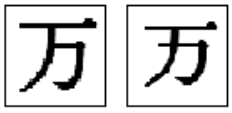


図 10　新設した包摂規準

（B）『JIS 漢字字典』個別字形例を一般化して包摂規準に格上げ

以下のような場合もある。

図 11『JIS 漢字字典』「感」「惑」の個別字形例

字形を包摂するかを判断する手引きとなる『JIS 漢字字典』には、一般規則としての包摂規準のほかに、個別の漢字字体に関して、包摂される複数の字形例が示されている場合がある。図 11 に示した通り、「感」「惑」の「心」の位置の差異に関しては、『JIS 漢字字典』に個別字形例として採られており、この２字に関しては「心」の位置の差異は包摂してよいことになる。近代ではこれ以外の字に関しても類例が見られるため、個別字形例に挙げられている字形差を一般化して包摂規準に格上げした。

図 12　新設した包摂規準

（C）類例を参考に新設

（現行字形）　　（明六雑誌）

(1-29-92)

図 13　『明六雑誌』にみられる「除」の字形

図 13 のような字形差は、現行の包摂規準には明示されていないが、既存の包摂規準「b 2 点画の接触交差関係の違い」のうち、「抜けるか、抜けないか」（図 14 参照）の類例に照らし、図 15 のような包摂規準を新設した。

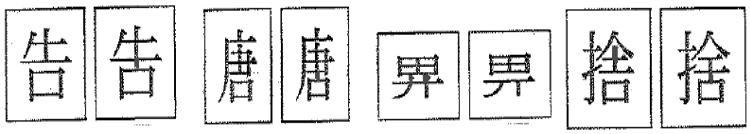


図 14　類例として参考にした既存の包摂規準

図 15　新設した包摂規準

（D）既存の包摂規準に追加

「華」などのの字形差処理のために、包摂規準連番８５が設定されている。（図 16）

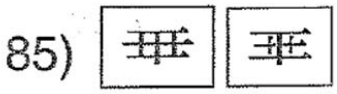


図 16　包摂規準連番８５

『明六雑誌』の「華」「樺」等は、図 17 のような字形で出現する。

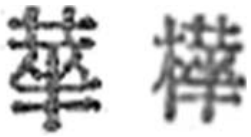


図 17　『明六雑誌』にみられる「華」「樺」の字形

このような字形も包摂するために、包摂規準連番８５の２つの字形に対し新たに図 18 の字形を追加した。

図 18　包摂規準連番８５に追加した字形

（E）既存の包摂規準を統合、より一般化

JIS X0213 では、連番６７、７０、７１のような包摂規準が設定されている。

図 19　包摂規準連番６７、７０、７１

このような字形差に類するとみられる差異で、このままの基準では処理できないものとして『明六雑誌』の「配」「改」「犯」などがある。

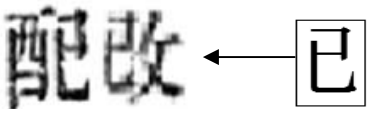


図 20　『明六雑誌』にみられる「配」「改」の字形

図 20 の「配」「改」は「己」の部分が「巳」になっているが、従来の包摂規準では「己」「巳」などの交替を認めても、「巳」は扱われていない。また、図 21 の「犯」は「巳」に交替している例だが、図 22 のとおり、包摂規準連番６７、７０ではこの交替が認められていない。ただし、包摂規準連番７１では、この三者の字形差は認められている。

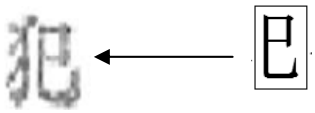


図 21　『明六雑誌』にみられる「犯」の字形

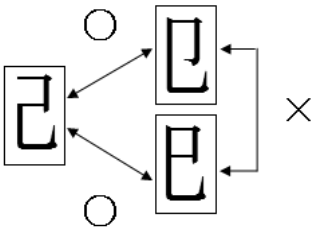


図 22　包摂規準連番６７、７０での交替関係

そこで、このような字形差も処理するために、包摂規準連番６７、７０、７１を統合し、より一般化したうえで、「巳」との交替も認めるという拡張を行った（図 23 参照）。

図 23　連番６７、７０、７１を近代語用に統合・拡張した包摂規準

５．２．２ 追加包摂規準を設定しないケース

（A）部首や部分字形が大きく異なるもの、偏の有無などの差異に関しては、包摂しない。

『明六雑誌』には、図 24、各右側に示すような異体字も出現する。これらのように部首や部分字形が大きく異なるものや偏の有無の違いなどに関しては、包摂規準の新設はせず、包摂しない。これらもコーパス上では、「派」(1-39-41)「脚」(1-21-51)「減」(1-24-26)「輩」(1-39-58)などの通用字で表現するが、包摂（＝同じ字とみなすこと）としてではなく、「別字代用」という形での処理とし、理念上は区別することとする。

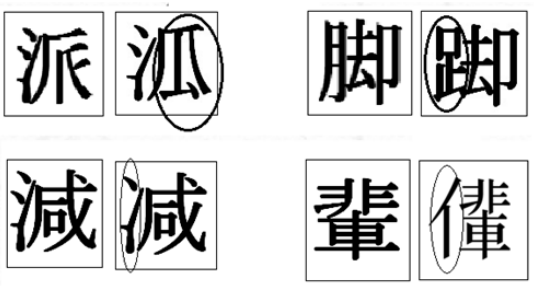


図 24　包摂 ” しない “ 字形例

追加包摂規準を設定して処理する字形差は、もとの JIS 包摂規準に掲げられている、

a）方向・曲直などの点画の性質による違い
b）２点画の接触交差関係の違い
c）２点画の結合分離の違い
d）１点画の増減の違い
e）類型の統合
f）筆法の簡化の違い

という５つに収まる範囲内とする。

（B）Unicode で表現可能な差異は、包摂しない。

将来的な Unicode 対応の可能性を考慮し、理念上は追加包摂規準の設定で処理できそうなものに対しても、Unicode で表現可能なものに対しては包摂はせず、処理上は別字とみなし、「別字代用」として処理する。なお、ここでは Unicode4.0 を参照している。

例えば『明六雑誌』での「跋」は図 25 のようのような字形で出現する。このような部分字形差は、理念上は図 26 のような追加包摂規準の設定を行う方法もありうるが、図 27 のように Unicode4.0 では両者の区別が可能である。このように Unicode では区別可能な字に対しては、将来的な Unicode 対応の可能性を考えると、同一の字にまとめるよりは別字にしておいた方がよいとの判断から、図 26 のような追加包摂規準の設定は行わず、コーパス上は「別字代用」の扱いで処理する。

図 25　『明六雑誌』にみられる「跋」の字形

図 26　追加包摂規準案

(U+8DCB)　(U+47E6)

図 27　Unicode4.0 での表現

５．３ 外字の「別字代用」

異体字のうち、追加包摂規準により同一字とみなされたもの以外は、扱いとしては外字となる。しかし言語研究資料としての使用を考慮した場合、電子テキスト上「〓」表示されていて読めない字というのは可能な限り少ないことが望ましい。そこで、外字認定されたものに対しても、極力「本来は外字であるが、言語研究資料としての使用のため別字で代用する」という手法を取ることにし、追加包摂規準とは別に、別字代用一覧を作成した（後掲）。

言語研究用という使用目的にあわせ「〓」表示は極力減らしたい。そこで、包摂規準の追加・修正で包摂できる字形は包摂し、理念上う、包摂規準レベルでの処理が難しい場合は包摂ではなく、別字で代用、という二段構えで表示上の「〓」表示を減らしていこうという試みである。

５．４ 包摂・代用する字体

本来の JIS X0213 では表現できない文字を、包摂規準の追加・修正または別字代用で表現しようとする場合、図 28 のように使用する文字の候補が複数存在する場合がある。

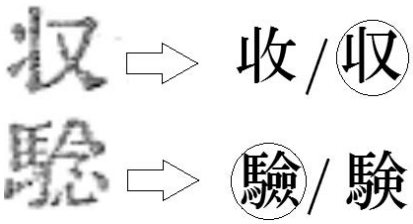


図 28　『明六雑誌』の字形と、包摂・代用候補

このような場合には、以下の方針とした。

①類似点の大きい方を使用する。
②決めかねる場合は正字を使用する。

①により「収」(1-28-93)、②により「驗」(1-81-68)が選択される。

また、別字代用に関しては、言語研究用に「読める」テキストを作成しようというところから出発しており、その目的のためにかなり思い切った代用を行った部分もある。

まず、本来は読み・意味の異なる別字に対しても、コーパス化対象テキストでの使用実態から、代用としての置き換えを認めた場合がある。

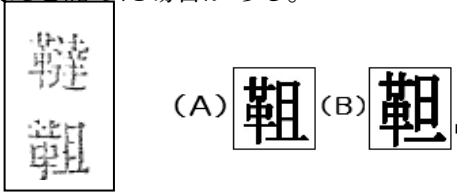


図 29　『明六雑誌』での字形差

図 29 の（A）（ソ）と（B）（タン）は本来読みも異なる別字であり、JIS X 0213 文字集合にあるのは（B）のみ（Unicode では（A）（B）ともに表現可能）である。『明六雑誌』では JIS X0213 外字となる（A）が出現するが、それは上例のように、「韃靼」の「靼」に当たる部分としてである。この使用実態を鑑みて、（A）を（B）「靼」(1-80-58)で代用することとした。

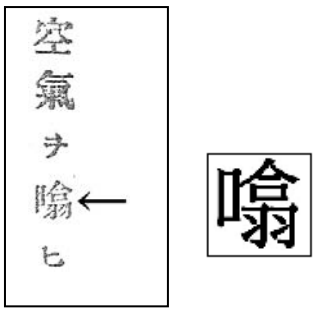


図 30　『明六雑誌』で出現字形

また、字形が大きく異なる場合でも、ほぼ同じ意味・読みの、より一般的な字への置き換えを認めた場合もある。

図 30 のとおり、『明六雑誌』においては動詞「すう」に上掲のような字が用いられる場合があるが、これは Unicode では表現できるものの、JIS X 0213 では表現できない。このような場合にも、言語研究資料として読めることを優先し、「吸」(1-21-59)で代用して処理した。

５．５ 追加包摂・代用した文字の情報管理

追加包摂、別字代用に関しては、まず規則として追加包摂規準、別字代用一覧を整備しておく。さらにコーパスのデータとしては、追加包摂規準で包摂したもの、本来外字であるが別字で代用したものにはそれぞれタグを付け、情報を記述しておく。

【例】
・追加包摂規準により包摂した文字
<包摂>序</包摂>
・代用した文字
<外字　代用=1>減</外字>
・包摂も代用もせず、残った外字
<外字　代用=0>𠮟</外字>

なお、代用としたものの中には、Unicode で表現できる字形も少なくない。そこで将来的な Unicode 対応の可能性も考慮し、Unicode で表現できる代用字に関しては、タグに Unicode 番号を記しておくこととした。

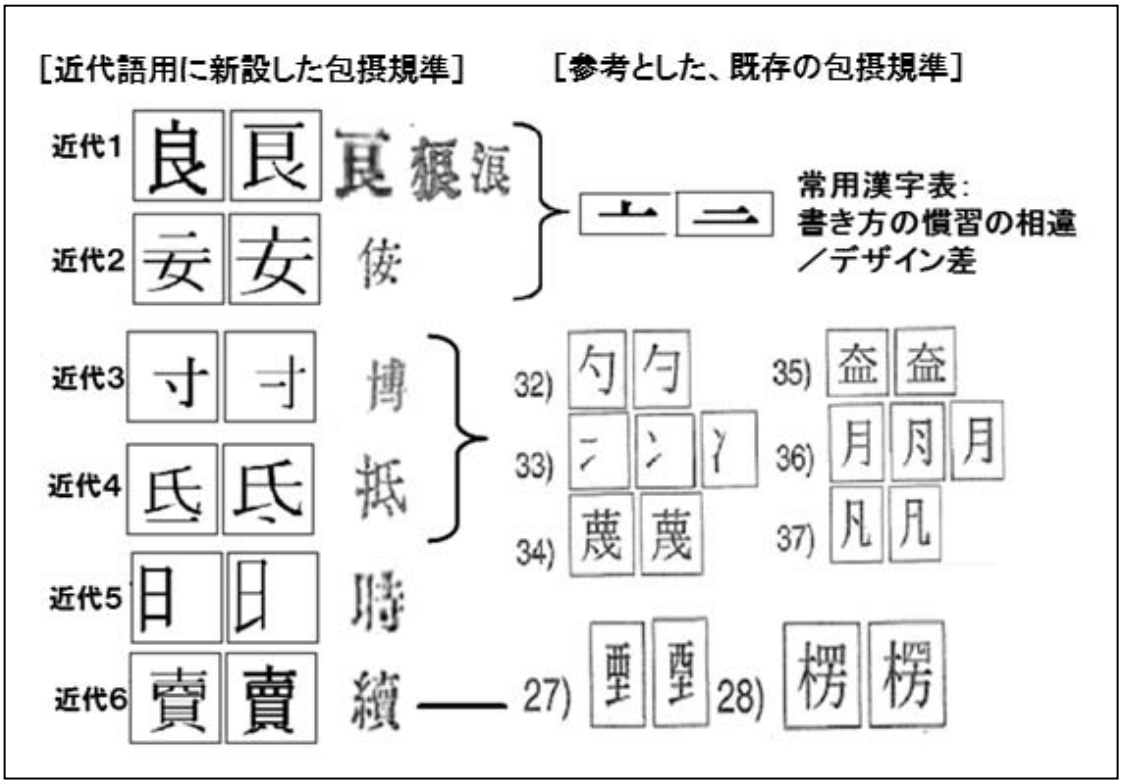
【例】
図 24 の「減」（にすい）、Unicode:51CF
→<外字　unicode="51CF" 代用="1" >減</外字>

６．追加包摂規準・別字代用一覧

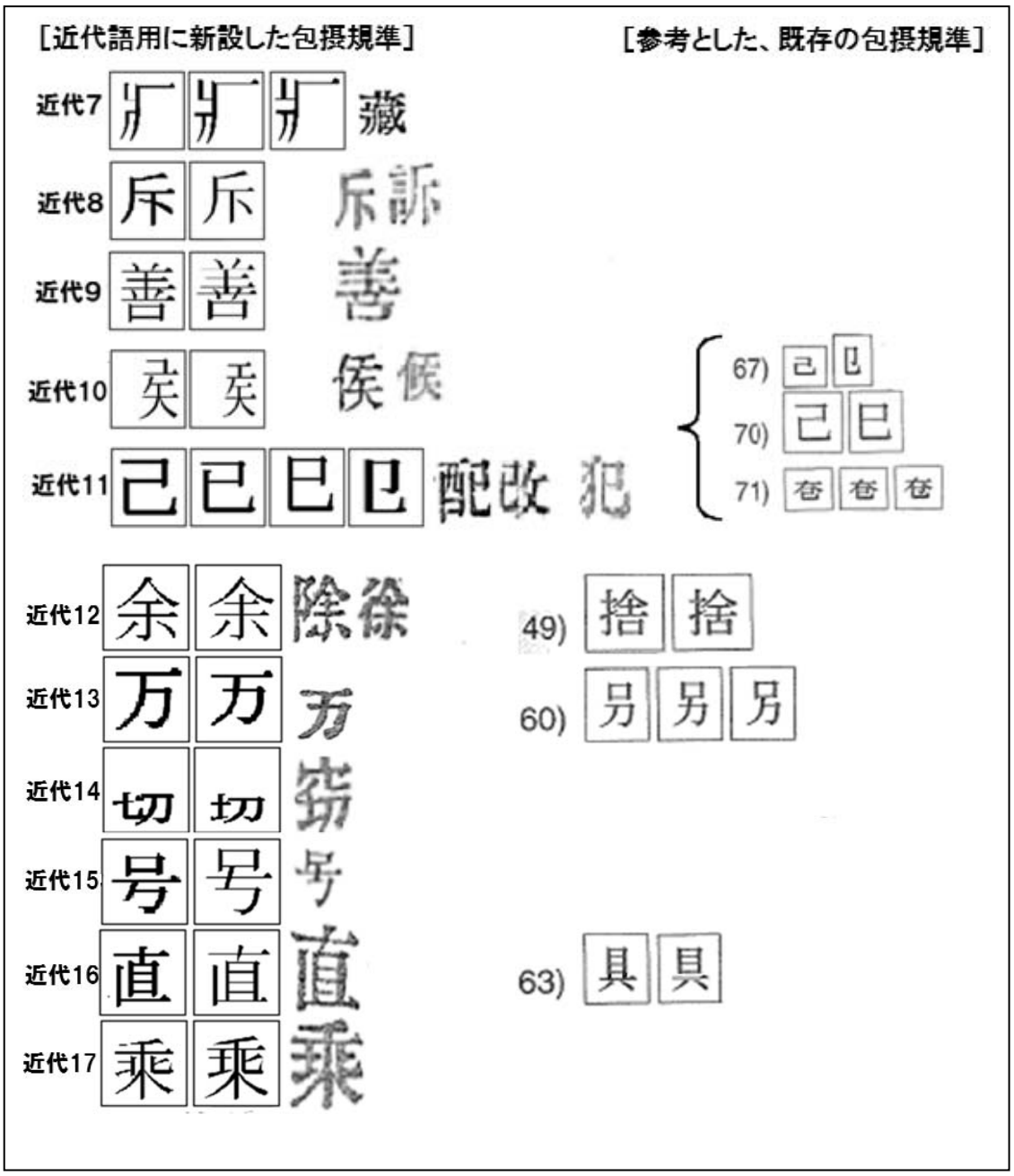
近代語用に新設した追加包摂規準、別字代用の一覧を以下に示す。

６．１　追加包摂規準

ａ）方向・曲直などの点画の性質による違い



b）２点画の接触交差関係の違い

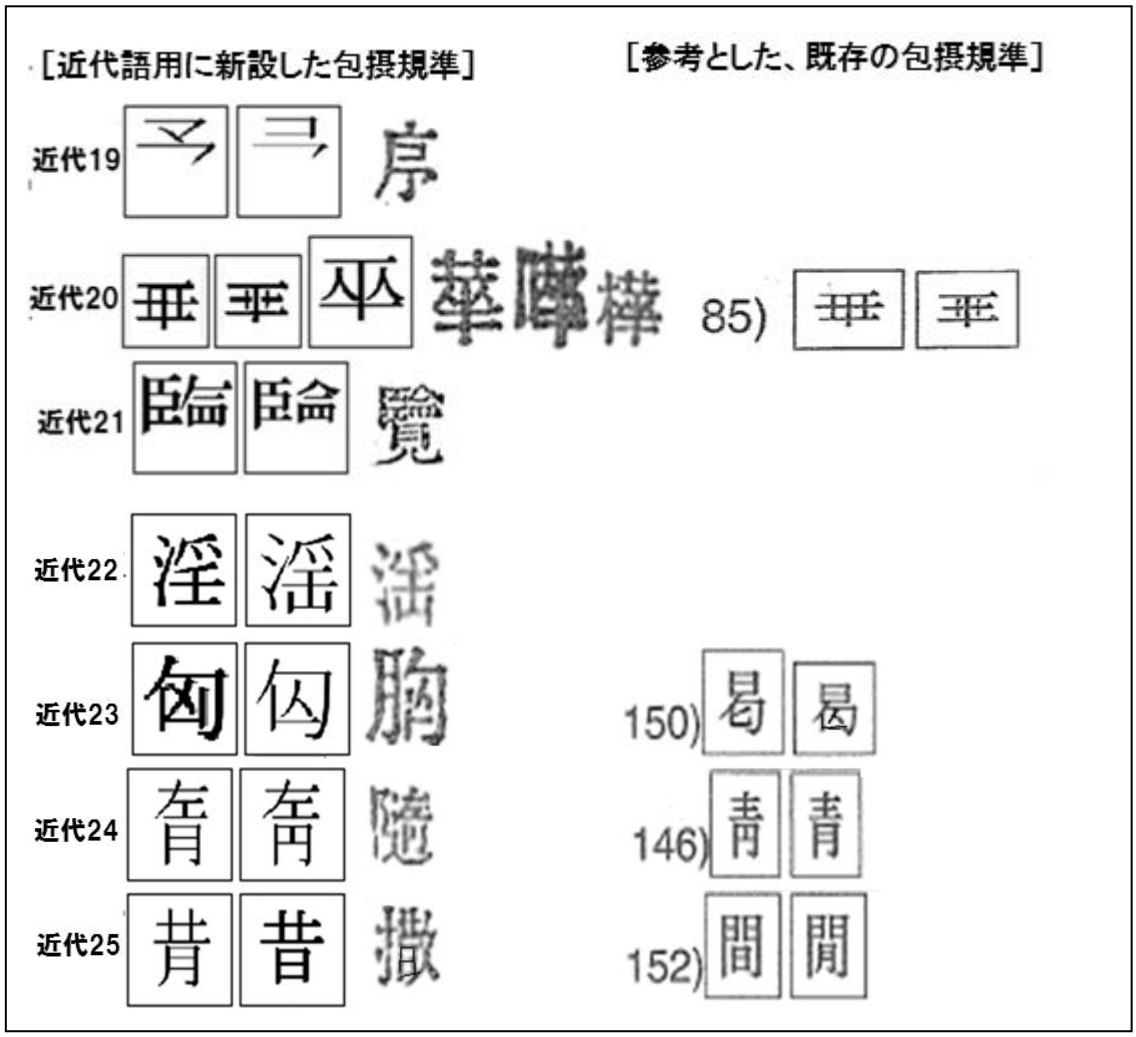


c）２点画の結合分離の違い

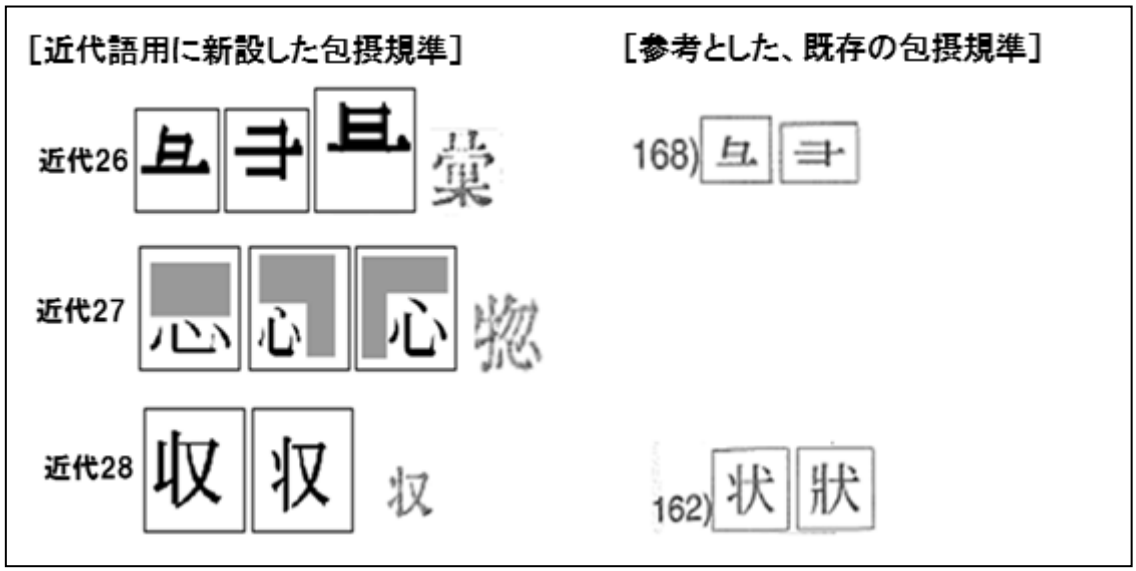
※これに関しては、追加したものはない。

d）１点画の増減の違い

e）類型の統合



f)筆法の簡化の違い



６．２ 別字代用一覧（３８字）

（１）Unicode では表現可能、JIS X0213 では外字（３３字）

（２）Unicode でも表現不可（５字）

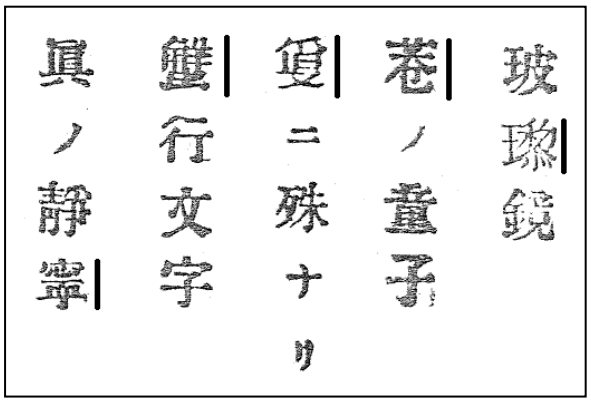


図 31 Unicode でも表現不可の字形

７．X 0213 文字集合／追加包摂／別字代用の検証

以上、包摂規準の追加・修正、外字扱いしたうえで電子テキスト上は代用字を使用する、という方法を考案したうえで、JIS X 0213 規格の包摂規準のみに依拠した場合と、今回提案した処理案を用いた場合とで、処理できる文字数の変化を検証した。

『明六雑誌』に現れる漢字の総数は 137,897、うち JIS X0213 のみで表現できるものは 135,797 字、カバー率にして 98.5%である。残る 1.5%、2,100 字が外字「〓」表示されるテキストとなるが、これは言語研究用の資料としての実用にとっては相当に多い量である。

表２　JIS X 0213 文字集合と『明六雑誌』漢字

文字区分	のべ字数
JIS X 0213	135,797
第 1 水準漢字	117,643
第 2 水準漢字	17,953
第 3 水準漢字	118
第 4 水準漢字	83
外字	2,100
計	137,897

追加包摂規準を適用すると、外字となるのべ 2,100 字のうち 1,774 字、さらに別字代用を適用すると 295 字の処理が可能になり、最終的に「〓」表示となるものは 31 字にまで減少し、99.9%の漢字を表現することができる。これらの処理を通して得られる結果は、割合からすればわずかな差でしかないが、言語研究という要請からは大きな意味を持つとも言える。

表３　各方針の適用で処理可能な文字（のべ）

	X0213 包摂	追加包摂	代用
処理可能文字総数	135,797	137,571	137,866
新たに処理できる文字総数	—	1,774	295
外字総数	2,100	326	31
カバー率	0.98478	0.99764	0.99978

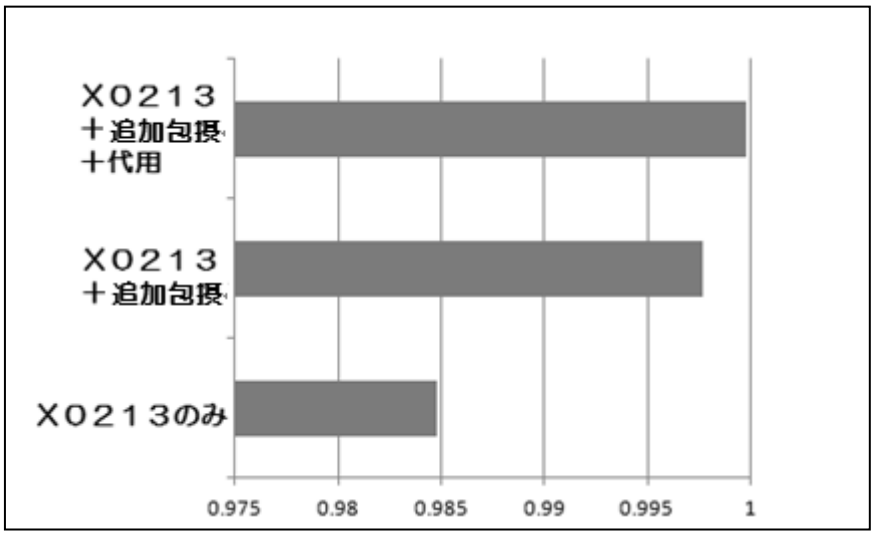


図 32　『明六雑誌』カバー率（のべ）

８．最終的に「〓」表示となる外字一覧

以上の処理を経て残る、最終的な外字、つまり「〓」表示となるものはのべ 31 字、異なりにして 25 字である。なお、これらは１字を除いて Unicode で表現可能である。

（１）Unicode では表現可能なもの（２４字）

（２）Unicode でも表現不可なもの（１字）

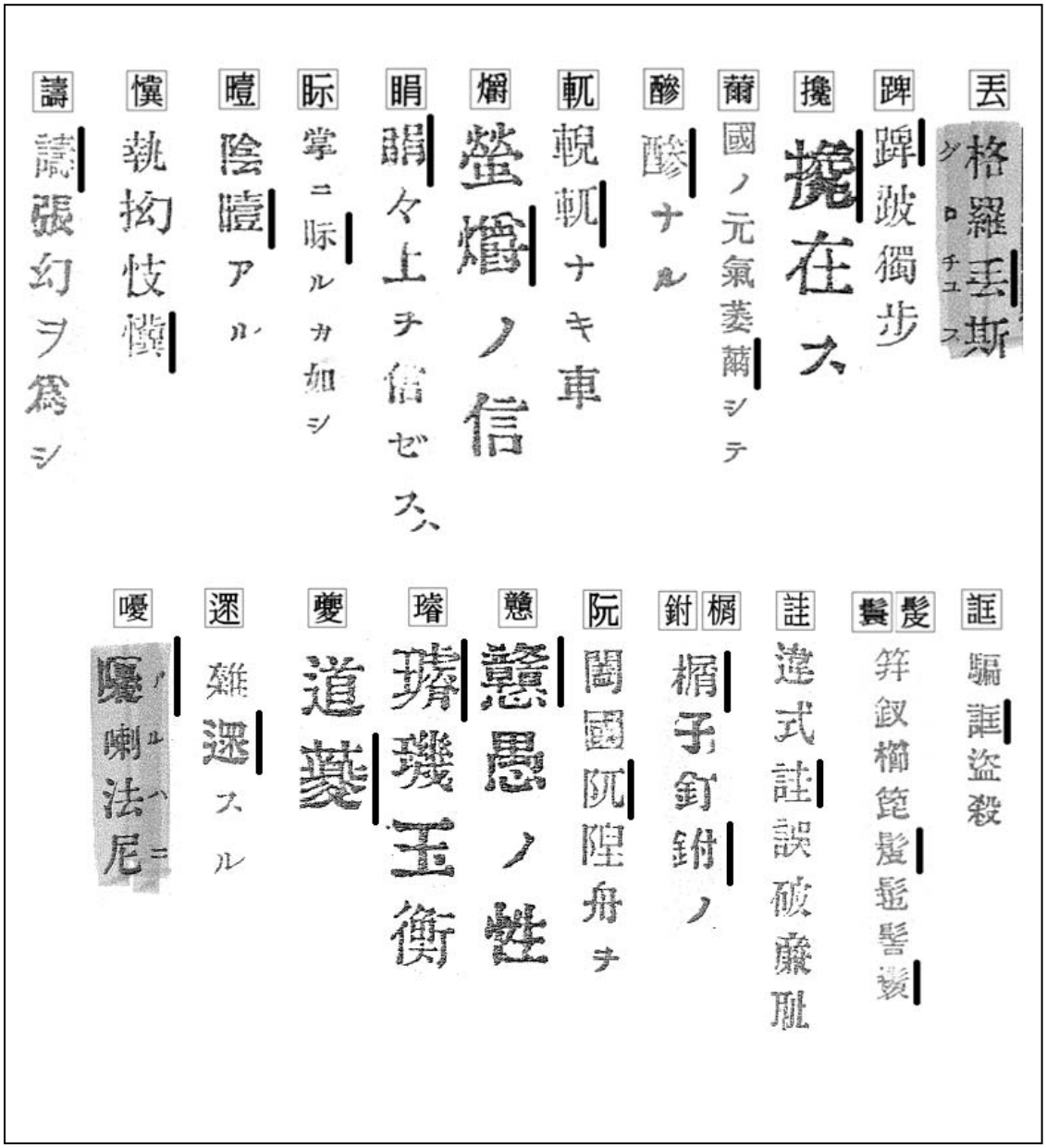


図 33　『明六雑誌』コーパスで「〓」表示となった実字形

９．今後の展望

本稿では『明六雑誌』を対象に、包摂規準の追加・修正と、別字代用の具体的方法を考案したが、『近代語コーパス』全体の文字処理を見据えた場合、今回設定した包摂規準のさらなる検証が必要となる。今後『明六雑誌』以外の活字資料を処理する場合、さらに別の包摂規準を新設する可能性や、本稿での追加包摂規準の修正を行う可能性は十分想定される。例えば『明六雑誌』の異体字処理を通して設定した「近代１９」（図 34）は「序」の異体字に対応するためだけに設けられたものである。

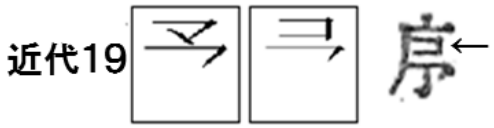


図 34　包摂規準　近代１９

しかし雑誌『太陽』(博文館、1895～1928) での活字を眺めてみると、「序」以外にも「疑」でも同様の字形差がみられ（図 35)、「近代１９」の有用性が確認できる。

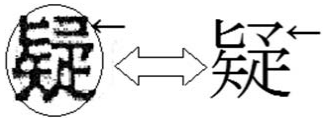


図 35　『太陽』に出現する「疑」の活字字形（左）

さらに『太陽』では図 36 のような類例も見られる。このような字形差も包摂するには、包摂規準「近代１９」を、図 37 のように修正し、さらに一般化していく方向性も考えられる。

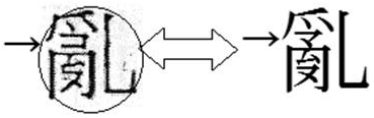


図 36　『太陽』に出現する「亂」(左)

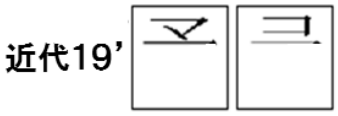


図 37　包摂規準「近代１９」修正案

このような検証を経て、近代活字用の包摂規準の整備を進めていくことが今後の課題となる。

文 献

小池和夫、府川充男、直井靖、永瀬唯(1999)『漢字問題と文字コード』(太田出版)
国立国語研究所(2005)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』(博文館新社)
柴野耕司編著(2002)『増補改訂ＪＩＳ漢字字典』(日本規格協会)
須永哲矢、堤智昭、高田智和(2011)「明治前期雑誌の異体漢字と文字コード－『明六雑誌』を事例として－」(『人文科学とコンピュータシンポジウム論文集 2011』、pp.381-388)
高田智和、小林正行、間淵洋子、大島一、西部みちる、山口昌也(2009)『JIS X0213:2004 運用の検証（国立国語研究所内部報告書 LR-CCG-09-01）』(国立国語研究所)
田中牧郎(2005)「漢字の実態と処理の方法」(国立国語研究所 2005 所収、pp.271-292)
安永尚志(1998)『国文学研究とコンピュータ』(勉誠社)

近代語テキストの形態素解析

小木曽　智信　（国立国語研究所言語資源研究系）[1]

１．はじめに

国立国語研究所により 2005 年に公開された『太陽コーパス』（2005）は単語情報を含まず、文書構造や注記等をマークアップしただけのコーパスだった。一方、同じ年に公開された現代語の『日本語話し言葉コーパス』（CSJ）や 2011 年に公開された『現代日本語書き言葉均衡コーパス』（BCCWJ）では、単語の読みや品詞などの形態論情報が付与されている。この形態論情報を用いることで、活用形や表記の違いにとらわれず語としての検索や集計が可能となり、語がもつ情報を組み合わせた高度な処理も行うことができる。

『太陽コーパス』に単語情報が付与されていないのは、当時の技術では、現代語と大きく異なる近代語のテキストに形態素解析を施すことが困難であったことによる。しかし、その後「近代文語 UniDic」が整備されたことにより、近代語のテキストであっても実用的な精度で形態素解析を行うことが可能になってきた。これにより、新たに構築される近代語コーパスでは、BCCWJ と同様の単語情報付きのコーパスとすることができる。

本稿では、新たな近代語コーパスの試作データである『明六雑誌コーパス』における処理を例に、近代語テキストの形態素解析について述べる。

２．近代語の形態素解析

２．１　日本語の形態素解析

日本語の形態素解析は 1990 年代以降にコンピューターの処理性能の向上とともに技術開発が進展し本格的な利用が可能となった。今日では、形態素解析を行うプログラム（形態素解析器）として、京都大学言語メディア研究室の JUMAN（1992～）、奈良先端科学技術大学院大学松本研究室の茶筌〔ChaSen〕（1996～）、同研究室で生まれた和布蕪〔MeCab〕（2002～）、KyTea〔京都テキスト解析ツールキット〕（2009～）などが自由に利用可能なソフトウェアとして公開されている。形態素解析は、コンピューターによる日本語処理の基盤であり、インターネット上の多くのサービスなどで活用され、欠かすことのできない技術となっている。

CSJ や BCCWJ は、国立国語研究所が中心となり新たに開発した言語研究に適した形態素解析用の電子化辞書「UniDic」（伝ほか 2007）を用いてコーパス中のテキストの形態素解析を施した。BCCWJ では、MeCab と UniDic を用いて、およそ 98%の解析精度での形態論情報のアノテーションを実現している。

２．２　近代語の形態素解析

従来、形態素解析を行うことができるのは現代語の文章だけであり、文語文の形態素解析を行うことはできなかった。たとえば、既存の形態素解析辞書（ChaSen 標準の IPADIC 2.7.0）によって文語文を解析すると図１のような結果となる（例文「こゝに漢字の利害と題するは、即ち聊か袈裟の眞價を問はんとするなり。」『太陽コーパス』「漢字の利害」より）。現代語向けの辞書によるものであるから当然の結果ではあるが、多くの誤りがあり、この解析結果を研究に利用することはできない。近代語のテキストを解析するためには、近代語向けの形態素解析辞書を作成する必要があるのである。

[1] togiso@ninjal.ac.jp

IPADIC 2.7.0/ChaSen 2.4.2

出現形	読み	品詞	活用型	活用形
こ	コ	名詞-一般		
ゝ	ゝ	記号-一般		
に	ニ	助詞-格助詞-一般		
漢字	カンジ	名詞-一般		
の	ノ	助詞-連体化		
利害	リガイ	名詞-一般		
と	ト	助詞-並立助詞		
題	ダイ	名詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
は	ハ	助詞-係助詞		
、	、	記号-読点		
即ち	スナワチ	副詞-一般		
聊か	イササカ	副詞-一般		
袈裟	ケサ	名詞-一般		
の	ノ	助詞-連体化		
眞	マコト	名詞-固有名詞-人名-名		
價		未知語		
を	ヲ	助詞-格助詞-一般		
問	トイ	名詞-一般		
はん	ハン	名詞-接尾-人名		
と	ト	助詞-格助詞-一般		
する	スル	動詞-自立	サ変・スル	基本形
なり	ナリ	名詞-一般		
。	。	記号-句点		

図 1 従来の形態素解析辞書による近代文語文の解析結果

３．近代文語 UniDic

一方、図 2 に示すのは近代語向けに新たに開発した形態素解析辞書「近代文語 UniDic」（小木曽ほか 2008, 2009）による解析結果である（近代文語 UniDic 1.2 と MeCab 0.99 で解析）。この結果からわかるように、文語の活用・歴史的仮名遣い・旧漢字・踊り字などに対応しており、文語文を正しく解析することが可能になっている。ここで、この「近代文語 UniDic」について説明する。

近代文語UniDic 1.2 /MeCab 0.99

出現形	発音形	代表形	代表表記	品詞	活用型	活用形	語種
こゝ	ココ	ココ	此処	代名詞			和
に	ニ	ニ	に	助詞-格助詞			和
漢字	カンジ	カンジ	漢字	名詞-普通名詞-一般			漢
の	ノ	ノ	の	助詞-格助詞			和
利害	リガイ	リガイ	利害	名詞-普通名詞-一般			漢
と	ト	ト	と	助詞-格助詞			和
題する	ダイスル	ダイスル	題する	動詞-一般	文語サ行変格	連体形-一般	混
は	ワ	ハ	は	助詞-係助詞			和
、			、	補助記号-読点			記号
即ち	スナワチ	スナワチ	即ち	接続詞			和
聊か	イササカ	イササカ	些か	副詞			和
袈裟	ケサ	ケサ	袈裟	名詞-普通名詞-一般			外
の	ノ	ノ	の	助詞-格助詞			和
眞價	シンカ	シンカ	真価	名詞-普通名詞-一般			漢
を	オ	ヲ	を	助詞-格助詞			和
問は	トワ	トウ	問う	動詞-一般	文語四段-ハ行	未然形-一般	和
ん	ン	ム	む	助動詞	文語助動詞-ム	連体形-撥音便	和
と	ト	ト	と	助詞-格助詞			和
する	スル	スル	為る	動詞-一般	文語サ行変格	連体形-一般	和
なり	ナリ	ナリ	なり-断定	助動詞	文語助動詞-ナリ-断定	終止形-一般	和
。			。	補助記号-句点			記号

図 2 近代文語 UniDic による解析結果

３．１　近代文語 UniDic の作成

形態素解析を行うには、解析に用いる語（見出し語）のリストに、語の出現しやすさ（生起コスト）、語・品詞間のつながりやすさ（連接コスト）の情報を付けた形態素解析用の辞書が必要である。ChaSen や MeCab などの現在使われている主な形態素解析システムでは、生起コスト・連接コストを機械学習と呼ばれる方法によって統計的に取得する。その

ため、形態素解析辞書を新たに作成するには、解析に用いる語の一覧（辞書データ）と、その辞書の内容にあわせて文章に正しく情報を付与した手本となる文章のデータ（学習用コーパス）が必要となる。辞書データと学習用コーパスから、プログラム（学習器）によって形態素解析辞書が作られる（図 3）。なお、辞書データは活用表によって各活用形に展開できるようにしておく必要がある。

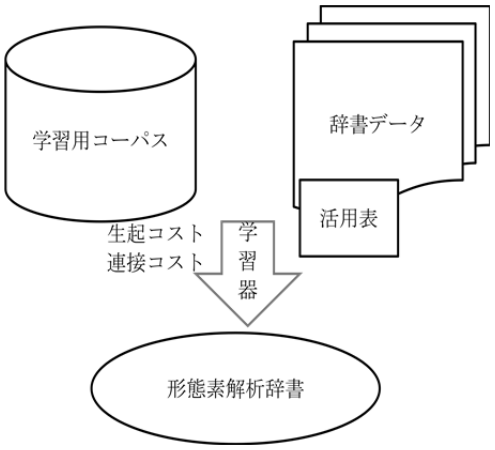


図 3 形態素解析辞書作成の流れ

３．１．１　辞書見出し語の整備

現代語とは異なるテキストを解析できるようにするためには、まず辞書データへの見出し語の追加が必要である。近代語用に追加が必要な見出し語としては、現代語では使われなくなった語、文語形、旧字・旧仮名遣いの形などさまざまなものがある。

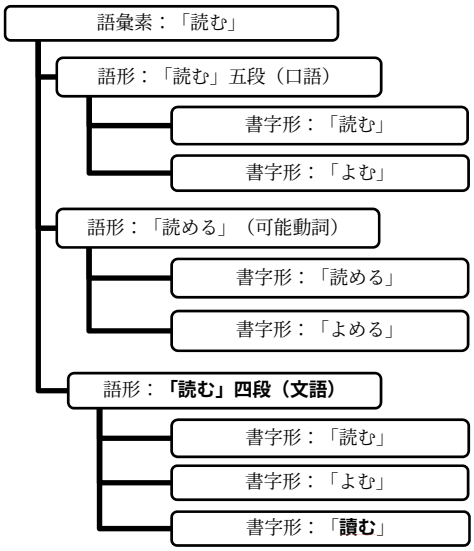


図 4 UniDic の階層（語彙素・語形・書字形）と文語形・旧字形

UniDic では見出し語を語彙素・語形・書字形・発音形の 4 段階で階層的に管理しているため、近代語解析に必要な語を各階層に整理して追加することができる。現代語としては使われなくなっている語は「語彙素」のレベルで、文語活用型の語は「語形」のレベルで、旧字形などは「書字形」のレベルで追加することになる（図 4）。これにより、現代語の語と統一的に管理することができるとともに、文語形と口語形、新字形と旧字形がそれぞれ関係を持つものであることを示すことができる。この方法により、近代語のテキストの

ためにおよそ数万語の見出し語を追加した。

この方式で、近代語のテキストのためにおよそ数万語の見出し語を辞書データベースに追加した。追加した見出し語は、当初は自動生成した文語形や旧字形を追加するところからはじめ、既存の辞書やデータ集の見出し語からも追加を行った。しかし、形態素解析辞書では、詳細な品詞や実際に現れる表記形を入力する必要があるため、単なる辞典の見出し語リストでは登録用のソースとして不十分な場合が少なくない。たとえば「名詞」といっても漢語サ変動詞の語幹としても使われるかどうかや、形容動詞の語幹や副詞としての用法を持つかどうかを区別する必要がある。また、表記の面では、辞典類の見出しに掲げられる代表的な表記ではなく、実際のテキストに現れる表記形を追加する必要がある。したがって、見出し語の追加にあたって最も効果的だったのは、実際に近代語のテキストを解析した結果から、未知語（見出し語（表記形）が形態素解析辞書にないために正しく解析されていない語）を見つけだして辞書データベースに登録することであった。

なお、辞書データでは、見出し語を追加登録してゆくとともに、活用語について活用表を整備して、必要な形を展開できるようにしておく必要がある。もともと UniDic は文語の活用型をもっていたが、近代文語 UniDic ではこれをさらに整備した。活用形の整備では、一般的な文語活用表にある活用形をそろえるほかに、特殊な表記に対応するための書字形を整備することも必要となる。たとえば、「讀て」（よみて）、「讀ず」（よまず）のように送り仮名が省略された表記等が多く用いられるためである。文法上の観点から作成される一般の活用表では問題とならないものだが、形態素解析辞書の活用表では表記上の違いについても活用表での対応が必要となる場合が少なくない。

３．１．２　学習用コーパスの整備

近代文語文の解析辞書を作るためには、辞書・活用表のほかに、機械学習を行うための学習用コーパスを整備する必要がある。近代文語文の解析のためには、辞書を拡充するとともに手本となる近代文語文の学習用コーパスを整備する必要がある。現在の近代文語 UniDic では表１に示したテキスト計約 46 万 6 千語を利用している。

表 1 近代文語 UniDic（1.2.1）の学習用コーパス

太陽	90604
女学雑誌	10802
文明論之概略	42800
法律	30868
青空文庫・論説	194364
青空文庫・小説	39294
文語詩	58377
総計	**467109**

３．１．３　テキストの解析前処理

近代語のテキストは、表記の上で、個々の語に揺れがあるにとどまらず、本文全体にわたって、仮名遣いの違い・漢字の新旧・踊り字使用の有無などのバリエーションがある。これらの問題に対処するためには、辞書に見出し（書字形）を追加して解析する方法と、あらかじめ本文の側を変換・修正してから解析する方法がある。近代文語 UniDic では、単純な置き換えが難しい仮名遣いや漢字の新旧については形態素解析辞書で対処した。一方、次の点については辞書での対応が困難であるため、解析の前に変換処理を行うことによって解析できるようにした。

漢字カタカナ交じり文

漢字カタカナ交じり文をそのまま解析できるようにするためには、仮名を含む書字形すべてについて、ひらがなとカタカナの二通りを用意する必要があり、現実的ではない。そこで、こうした本文については解析前に漢字ひらがな交じり文に変換したうえで解析することとした。「近代文語 UniDic」付属の解析用のアプリケーション「近代茶まめ」では、必要に応じて自動でカタカナをひらがなに変換させる機能を持たせている。

この処理では、漢字カタカナ交じりの文章中にカタカナとして残したい外来語等がカタカナで表れる場合、これのみを区別してカタカナのまま残すことはできない。したがって、完全な処理のためには人手による確認が必要になる。

濁点無表記

近代語のテキストでは濁点が表記されない場合も少なくないが、濁点無表記形を一々辞書登録していくことは無駄が大きい上に解析精度を低下させることにつながるため、これもあらかじめテキストを修正した後に解析を行うこととした。単純な変換処理は行えないため、原則として人手によって濁点付与を行うこととなる。

濁点付与作業を助け大量のテキストを処理するために、濁点の付与を自動で行うための研究とそのためのアプリケーション開発も行っている（岡ほか 2011）。

踊り字

「ゝ」「ゞ」などの仮名一字を単位とする踊り字については、これを含む一々の出現形を辞書登録するのではなく、解析前に踊り字を対応する文字に変換してから解析することとした。「近代茶まめ」では、この変換処理をボタン一つでできるようになっている。

しかし、くの字点（/ \）については繰り返される範囲が明瞭でないため自動変換は行わず、人手によるテキスト修正を経たのちに解析を行うか、またはそのままの形で解析することとした。そのままの形で解析される場合のために、くの字点は、「そろ / \」のように語の一部となっているものはその形を辞書に登録している。語や句を繰り返すものについては「/ \」全体を記号扱いの一単位として扱った。

なお、漢字を繰り返す「々」は今日でも「人々」のように用いられるため変換を行わず、その形を辞書登録している。しかし、近代語では「民主々義」のように語（短単位）の境界を跨いで繰り返される場合がある。近代文語 UniDic では、これらについて高い頻度で出現するものは辞書登録を行っているが、網羅的な対応は行っていない。また、漢字を繰り返す「〻」は「々」に置換している。

以上の解析前処理を完全な形で行うために、後述する『明六雑誌コーパス』の構築にあたっては、「漢字カタカナ交じり文中でカタカナをそのまま残す部分のアノテーション」や、「くの字点等の踊り字によって繰り返される範囲の明示」「濁点が期待される位置への濁点付与」の全ての作業を人手で行っている。修正を行った部分はすべてタグにより原文の状態を保持している。

３．２　解析精度

現在公開されている近代文語 UniDic（Ver.1.2.1）の解析精度は表 2（次ページ）に示す通りである。評価対象は、学習用のコーパスから約 10%を文単位でランダムサンプリングして学習対象から取り除いた人手修正済みのデータ 44587 語である。

表 2 で、「境界」とあるのは、最も基本的な評価基準で、解析結果において単語の境界が正しかったかどうかを意味する。「品詞」は境界が正しいことに加えて単語の品詞も正しく認定されていたかどうかを意味する。「語彙素」は境界と品詞に加えて語彙素（辞書見出し）としての認定も正しかったかどうかを意味する。たとえば「金」が「きん」でなく「かね」と正しく解析されているかどうかといった違いに相当する。「発音形」は、ここでは発

音というよりは語形の違いが正しく認定されているかどうかを評価するもので、境界・品詞・語彙素が正しいことに加え、さらに語形が正しいかどうかを意味する。たとえば、「言語」が文脈にあわせて「げんご」ではなく「ごんご」と正しく解析されているかどうかといった違いに相当する。表の右に行くほど評価基準が厳しくなっている。

表 2 近代文語 UniDic（1.2.1）の解析精度

	境界	品詞	語彙素	発音形
正解データ語数	44587			
出力語数	44573			
一致語数	44244	43594	43291	43162
再現率	99.23%	97.77%	97.09%	96.80%
適合率	99.26%	97.80%	97.12%	96.83%
F値	99.25%	97.79%	97.11%	96.82%

「正解データ語数」としたのは、評価データの語数である。評価データはあらかじめ人手による修正を経ているため、これが正解とみなされる。「出力語数」は形態素解析結果として出力されたデータの語数である。「一致語数」としたのは出力語数のうち評価データ（正解）に一致した語数である。たとえば、境界認定の場合、出力された 44573 語中、329 語は誤りだったことになる。

「適合率」「再現率」「F 値」は情報検索システムの性能評価でしばしば用いられる概念で、ここでは適合率（precision）は「一致語数／出力語数」（出力されたもののうちどれだけが正しかったか）に、再現率（recall）は「一致語数／正解データ語数」（正しいもののうちどれだけを出力できているか）に相当する。F 値は再現率と適合率の調和平均で「２×再現率×適合率／（再現率＋適合率）」で計算できる。一般に再現率を上げると適合率が下がり、適合率を上げると再現率が下がるため、システムの評価としては両方の値を加味する必要がある。そのため、一つの数値で精度を示す場合にはしばしば F 値が用いられる。

表 2 に示された精度は、すでに現代語の形態素解析の精度と比べても遜色ないほどのレベルに達している。しかし、これは「未知語なし」のデータに対する評価結果である。近代語のテキストでは多様な語が用いられるため、辞書に登録のない見出し語（未知語）が多く発生しがちである。近代文語 UniDic は、明治普通文と呼ばれるような比較的平易な文語論説文であれば高い精度で解析を行うことができるが、雅文調のテキストや口語的な内容を含むものではこれだけの精度は期待できない。また、もともと文語文を対象としたものであり口語文はうまく解析ができない。近代語のコーパスの中で口語文は大きな割合を占めるが、近代の口語文の解析のためには今後辞書の整備を行っていく必要がある。

４．近代語コーパスへの形態論情報付与（『明六雑誌』の場合）

『明六雑誌コーパス』の構築作業では、近代文語 UniDic で解析した結果を人手によって修正することで高い精度の形態論情報を付与した。明治初期の『明六雑誌』の語彙は、明治後期以降のデータを中心に整備してきた近代文語 UniDic の語彙とは異なる部分が大きく、登録されていない見出し語が多いため解析エラーも多くなっていた。

図 5 は『明六雑誌』の一部の修正済みデータを、公開中の近代文語 UniDic1.2.1 で解析した結果と比較して、明六雑誌コーパス構築開始時における、形態素解析の状況を示したものである（『明六雑誌』1874 年 1 号「洋字ヲ以テ国語ヲ書スルノ論」の一部で特に誤りの目立つ部分）。左側が正解となる人手修正済みのデータで、右側が 1.2.1 による自動解析結果であり、左端に「●」を付した部分が解析に誤りがあった語である。

	文境界	書字形	語彙素読み	語彙素	品詞	活用型	活用形	書字形	語彙素読み	語彙素	品詞	活用型	活用形	語種
	B	然る	シカル	シカリ	然り	動詞-一般	文語ラ行変格	然る	シカリ	然り	動詞-一般	文語ラ行変格	連体形-一般	和
	I	に	ニ	に	助詞-接続助詞			に	ニ	に	助詞-接続助詞			和
	I	如此き	カクノゴトシ	如此し	形容詞-一般	文語形容詞-ク	連体形-一般	如此き	カクノゴトシ	如此し	形容詞-一般	文語形容詞-ク	連体形-一般	和
	I	人民	ジンミン	人民	名詞-普通名詞-一般			人民	ジンミン	人民	名詞-普通名詞-一般			漢
	I	の	ノ	の	助詞-格助詞			の	ノ	の	助詞-格助詞			和
	I	愚	グ	愚	名詞-普通名詞-一般			愚	グ	愚	名詞-普通名詞-一般			漢
	I	も	モ	も	助詞-係助詞			も	モ	も	助詞-係助詞			和
●	I	左提	サテイ	左提	名詞-普通名詞-一般			左	サ	然	副詞			和
●	I	右挈	ユウケツ	右挈	名詞-普通名詞-一般			提	サゲル	下げる	動詞-一般	文語下二段-ガ行	連用形-一般	和
●	I	勞來	ロウライ	労来	名詞-普通名詞-一般			右	ミギ	右	名詞-普通名詞-一般			和
●								挈	タズサエル	携える	動詞-一般	文語下二段-ハ行	連用形-一般	和
●								勞	ロウ	労	名詞-普通名詞-一般			漢
●								來	ライ	来	接尾辞-名詞的-副詞可能			漢
	I	輔翼	ホヨク	輔翼	名詞-普通名詞-一般			輔翼	ホヨク	輔翼	名詞-普通名詞-一般			漢
	I	其	ソノ	其の	連体詞			其	ソノ	其の	連体詞			和
	I	苗	ナエ	苗	名詞-普通名詞-一般			苗	ナエ	苗	名詞-普通名詞-一般			和
	I	を	ヲ	を	助詞-格助詞			を	ヲ	を	助詞-格助詞			和
●	I	攫	ヌク	抜く	動詞-非自立可能	文語四段-カ行	連体形-一般	攫			名詞-普通名詞-一般			
●	I	冂	コト	事	名詞-普通名詞-一般			冂		冂	補助記号-一般			記号
	I	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	和
	B	去	サル	去る	動詞-非自立可能	文語四段-ラ行	連用形-一般	去	サル	去る	動詞-非自立可能	文語四段-ラ行	連用形-一般	和
	I	て	テ	て	助詞-接続助詞			て	テ	て	助詞-接続助詞			和
	I	転ら	クサギル	転る	動詞-一般	文語四段-ラ行	未然形-一般	転ら	クサギル	転る	動詞-一般	文語四段-ラ行	未然形-一般	和
	I	ざる	ズ	ず	助動詞	文語助動詞-ズ	連体形-補助	ざる	ズ	ず	助動詞	文語助動詞-ズ	連体形-補助	和
	I	冂	コト	事	名詞-普通名詞-一般			冂	コト	事	名詞-普通名詞-一般			和
	I	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	なく	ナイ	無い	形容詞-非自立可能	文語形容詞-ク	連用形-一般	和
	I	時宜	ジギ	時宜	名詞-普通名詞-一般			時宜	ジギ	時宜	名詞-普通名詞-一般			漢
	I	を	ヲ	を	助詞-格助詞			を	ヲ	を	助詞-格助詞			和
	I	制し	セイスル	制する	動詞-一般	文語サ行変格	連用形-一般	制し	セイスル	制する	動詞-一般	文語サ行変格	連用形-一般	混
	I	て	テ	て	助詞-接続助詞			て	テ	て	助詞-接続助詞			和
	I	漸次	ゼンジ	漸次	副詞			漸次	ゼンジ	漸次	副詞			漢
	I	開明	カイメイ	開明	名詞-普通名詞-一般			開明	カイメイ	開明	名詞-普通名詞-一般			漢
	I	の	ノ	の	助詞-格助詞			の	ノ	の	助詞-格助詞			和

図 5 近代文語 UniDic による解析結果

図 2 のように、『明六雑誌』の解析では、多くの未知語が発生するため、新たに辞書登録を行いながら修正作業を行った。『明六雑誌コーパス』全体の語数はのべ語数で約 180500 語・異なり語数で約 15500 語である（記号を含む）。このコーパスを整備するために新たに約 3700 語を辞書に登録する必要があった。新たに追加した語は語彙素（辞書見出し相当）のレベルから追加したものもあれば、すでにある見出し語に書字形（表記形）を新たに追加したものもある。

新規登録語のうち 2834 語は頻度が 1 であり、471 語は頻度 2 であった。つまり、新規に追加した語の大部分は非常に使用頻度の低い語であった。のべ語数では約 5600 語が未知語であり、逆に約 174900 語は既知語であった。すなわち、『明六雑誌コーパス』全体の 96.89%（174900/180500）は既存の近代文語 UniDic の語彙でカバーされていたことになる。

未知語を含まないデータで評価した近代文語 UniDic の解析精度は語彙素認定で約 97%であった（表 2）。この解析精度を加味すると、既存の近代文語 UniDic による当初の『明六雑誌コーパス』の解析精度は次のように推定できる。すなわち、未知語部分の 5600 語は全て誤りと見なし、既知語部分が 97%の精度で解析されていたとすると、正しく解析されていた語数は約 169700 語（174900*0.97）であることから、概算で全体の解析精度は約 94%（169700 / 180500）であったといえる（これは再現率ベースでの計算だが、適合率・F 値でもほぼ同じ数字である）。

表 3 に『明六雑誌コーパス』のために新たに辞書登録した、コーパスにおける頻度が 8 以上の新規追加語（60 語）を挙げる。新規追加語の中では高頻度の語だが、総じて一般的でない語や表記であることがわかる。

表 3　明六雑誌コーパスの語数と近代文語 UniDic への新規追加語数

語彙素	語形	書字形	品詞	頻度
如何	イカ	何	名詞-普通名詞-一般	10
易直	イチョク	易直	名詞-普通名詞-形状詞可能	8
曰く	イワク	云	名詞-普通名詞-副詞可能	17
置く	オク	舎く	動詞-非自立可能	13
思う	オモウ	謂ふ	動詞-一般	12
思えらく	オモエラク	以爲く	副詞	12
開交	カイコウ	開交	名詞-普通名詞-一般	15
変える	カユ	易ゆ	動詞-一般	10
関渉	カンショウ	關渉	名詞-普通名詞-サ変可能	9
気学	キガク	氣學	名詞-普通名詞-一般	9
議者	ギシャ	議者	名詞-普通名詞-一般	8
議法	ギホウ	議法	名詞-普通名詞-一般	9
下観	ゲカン	下觀	名詞-普通名詞-サ変可能	12
下民	ゲミン	下民	名詞-普通名詞-一般	13
限制	ゲンセイ	限制	名詞-普通名詞-サ変可能	12
孤陰	コイン	孤陰	名詞-普通名詞-一般	9
好和	コウワ	好和	名詞-普通名詞-一般	13
国中	コクチュウ	國中	名詞-普通名詞-一般	9
試み	ココロミ	嘗み	名詞-普通名詞-一般	8
国君	コックン	國君	名詞-普通名詞-一般	10
異	コト	特	名詞-普通名詞-一般	39
今時	コンジ	今時	名詞-普通名詞-一般	8
裁成	サイセイ	裁成	名詞-普通名詞-一般	11
三聖	サンセイ	三聖	名詞-普通名詞-一般	11
三宝	サンポウ	三寶	名詞-普通名詞-一般	50
シビリゼーション	シビリゼーション	シヴヰリゼーション	名詞-普通名詞-一般	14
者流	シャリュウ	者流	名詞-普通名詞-一般	24
習	シュウ	習	名詞-普通名詞-一般	14
上観	ジョウカン	上觀	名詞-普通名詞-サ変可能	13
シロシ	シロシ	素	名詞-固有名詞-人名-名	22
信紙	シンシ	信紙	名詞-普通名詞-一般	8
人主	ジンシュ	人主	名詞-普通名詞-一般	14
数百	スウヒャク	數百	名詞-数詞	9
少しく	スコシク	少く	副詞	8
大宝	タイホウ	大寶	名詞-普通名詞-一般	11
タバコ	タバコ	烟	名詞-普通名詞-一般	9
治刑	チケイ	治刑	名詞-普通名詞-一般	9
忠諒	チュウリョウ	忠諒	名詞-普通名詞-形状詞可能	8
蝶鋏	チョウコウ	蝶鋏	名詞-普通名詞-一般	8
つく	ツク	付く	動詞-一般	20
妻	ツマ	婦	名詞-普通名詞-一般	22
無い	ナシ	無し	形容詞-非自立可能	101
パッション	パッション	パツシヨン	名詞-普通名詞-一般	8
独り	ヒトリ	獨	名詞-普通名詞-副詞可能	28
ベーコン	ベイコン	倍根	名詞-固有名詞-人名-一般	25
邦	ホウ	邦	名詞-普通名詞-一般	8
磨する	マス	磨す	動詞-一般	8
先ず	マズ	先	副詞	11

間々	ママ	間	副詞	13
魅する	ミス	魅す	動詞-一般	17
アメリカン	メリケン	米利堅	名詞-普通名詞-一般	9
最も	モットモ	尤	副詞	12
基づく	モトヅク	本づく	動詞-一般	11
止む	ヤム	息む	動詞-一般	11
容忍	ヨウニン	容忍	名詞-普通名詞-サ変可能	11
与聞	ヨブン	與聞	名詞-普通名詞-サ変可能	9
濫出	ランシュツ	濫出	名詞-普通名詞-サ変可能	9
リバティー	リボルチー	リボルチー	名詞-普通名詞-一般	14
ルーサー	ルーサー	路惕	名詞-固有名詞-人名-一般	9
論無い	ロンナシ	論なし	形容詞-一般	11

５．おわりに

以上、近代語テキストの形態素解析について、近代文語 UniDic の解説と『明六雑誌コーパス』の構築時の解析結果修正作業を中心に述べた。

『明六雑誌』は近代文語 UniDic の主たる対象からはずれたテキストであるため、多数の未知語を追加しながら自動解析結果を修正して対処する必要があった。もっとも『明六雑誌コーパス』のように、（把握できる範囲で）誤りを全て修正したコーパスを公開するようなケースは稀であると思われる。一般的な研究利用であれば必要とされる部分についてのみ修正を行えば良いし、94%程度の解析精度があれば十分な場合も少なくないだろう。また、ここでの精度評価は、単語境界・品詞認定・語彙素認定（代表表記・読み・語種を含む）の全てが正しい場合のみを正解と見なすという、非常に厳しい評価基準によっている。読みや語種、品詞といった一部についてだけの精度であればこれを上回ることは確実である。

単に稀例を探すような場合には文字列検索で事足りるが、調査対象がテキスト全体の中でどのような位置を占めるのかを把握するためには、データ全体に対して形態素解析が施されている必要がある。形態素解析がなされたコーパスは、単に検索の手間が少なく、索引ではできなかったような組み合わせ検索ができるだけでない。テキストを、順序を持った語の集合として扱って、データベース上で自由に集計し、統計的な処理を行うことが可能になるのである。今後、近代語の研究においてもこうした本格的な語彙研究等のコーパスを活用した研究が行われることに期待したい。

なお、今回追加した『明六雑誌』の語彙を含む新しい近代文語 UniDic を近く公開する予定である。

文　献

国立国語研究所（2005）『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM、博文館新社）

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（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 号 pp.101-122.

小木曽智信・小椋秀樹・近藤明日子（2008）「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」『日本語学会 2008 年度春季大会予稿集』 pp.211-218

小木曽智信ほか（2009）『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』科研費若手研究(B) 研究成果報告書（課題番号 19720110）（http://dl.dropbox.com/u/73297026/report/unidic-MLJ_report2009.pdf）

岡照晃・小町守・小木曽智信・松本裕治（2011）「機械学習による近代文語文への濁点の自動付与」『情報処理学会 自然言語処理研究会報告』Vol.2011-NL201, No.6

URL

形態素解析辞書 UniDic ダウンロードサイト：http://download.unidic.org/
近代文語 UniDic：http://www2.ninjal.ac.jp/lrc/index.php?UniDic
形態素解析器 MeCab ホームページ：http://mecab.sourceforge.net/

近代語コーパスのための形態論情報付与規程の整備

須永　哲矢　（国立国語研究所コーパス開発センター）[1]
近藤　明日子　（国立国語研究所コーパス開発センター）[2]

１．近代語コーパスでの言語単位

１．１ 近代語コーパスでの言語単位

近代語コーパスでの言語単位には、「短単位」という言語単位を採用した。「短単位」は『現代日本語書き言葉均衡コーパス』でも採用されている言語単位で、詳細な規程のもと、揺れの少ない単語認定を実現している。実際の単位認定規程は、各ケースごとに細かく定義されているが、ここではその概要として、単位認定の大原則を紹介しておく。現代語についての規程の詳細は、小椋・小磯ほか(2011)を参照してほしい。以下、小椋・小磯ほか(2011)を『規程集』と呼ぶ。近代語については、コーパスの構築に向けて現在規程を検討中であるが、本稿では、規程を整備していくにあたっての問題点を整理する。なお、本稿で述べる問題点は網羅的なものではなく、コーパス作成準備作業において気付いた範囲でまとめたものにとどまることを、はじめにお断りしておく。

１．２「短単位」の概要

《和語》

単純語２語の結合まで、又は単純語１語と接辞１語の結合までを１短単位とする。
（“／”は短単位の切れ目を示す。以下も同様）

【例】／母／　　／母親／　　／母親／代わり／　　／真っ白／

《漢語》

2 字漢語までを１短単位とする。

【例】／大臣／　　／財務／大臣／　　／大臣／級／会合

《外来語》

原語で１語となるものを１短単位とする。

【例】　／オレンジ／　　／カラー／コピー／　　／ビタミン／剤／

[1] tsunaga@ninjal.ac.jp
[2] kondo@ninjal.ac.jp

《付属語》
付属語１語を１短単位とする。

【例】　／が／　　／だ／　　／の／で／

ここで紹介した短単位認定規程は、あくまで大原則であり、実際には、例外を含めて極めて詳細な規程が設けられている。

例えば、連体詞とされる「この」は、もとをたどれば代名詞「こ」と格助詞「の」に分割でき、そう見るならば「付属語１語を１短単位とする」という原則から、付属語である「の」を切り離し、「こ／の」と分割されることになる。しかし現代においては「こ」と「の」に分かれるという意識はもはや薄れており、代名詞「こ」が「の」以外と結合する用法も存在しない。このような事情から、現代語においては「この」に関しては「代名詞＋助詞」のように分割せず、「連体詞」と認定し１短単位とする、と規程として定めてある。
このように、単位認定の仕方が複数想定できてしまう場合や、原則通りに処理しない方が現代語の実態にとっては有用であると判断される場合などに関しても、個々に処理方針を詳細に設定したものが『規程集』となっている。

１．３　時代に合わせた規程整備の必要性

上述のとおり、言語実態に即した詳細な規程といっても、それはあくまで現代語を対象として設定されたものであり、時代が異なれば当時の言語実態と齟齬をきたす場合も少なくない。

例えば、上で例とした「この」に関しては、現代語を扱う限りには連体詞として１短単位として認定したが、平安時代までさかのぼってしまえば、当時においては、「の」以外の助詞も自由に結合でき（「こを」「こは」など）、実例上からも「こ」を単独で代名詞と認めた方が当時の実態に適合すること、などから「こ／の」と２単位に分割する方がふさわしい。

そのため、歴史的言語資料に対し「短単位」を引き続き採用する際には、その時代に合わせて規程にも拡張や変更を加えるなど、整備が必要となる。

歴史的資料を対象とした規程として整備されているのは、平安期の和文資料を対象とした規程（小椋・須永 2012）のみであり、近代語用の規程はまだ設定されていない。そこで、近代語コーパス構築のためには、近代語の言語実態に即した短単位規程を設定しなければならない。そのために、まずは『明六雑誌コーパス』の試作を通じ、作業中に生じた処理上の問題点（従来の規程どおりの処理では近代語の言語実態と齟齬をきたす場合／品詞認定等、処理の仕方が複数想定される事例に関して、従来の規程の判別基準では処理しきれない場合／処理方針を新規に定めなければならない近代語特有の問題、など）を収集し、近代語に合わせた処理方針を検討した。将来的には近代語用の詳細な規程集の作成を最終

目標としているが、本報告書での第一目標は、近代語を処理するにあたっての問題点の整理と、暫定的な処理の方向性を定めることである。

なお、一口に「近代語」と言っても、口語文と文語文で大きく様相が異なる場合があり、近代語用の規程整備という作業においても、実際には文語の場合、口語の場合と分けて設定しなければならない事例もある。そこで将来的には近代語共通の規程のほか、文語のみ、あるいは口語のみに適用される規程を細分していく必要がある。ただ、本稿の範囲内では、試作コーパスとなった『明六雑誌』の文体の多くが文語体であったこともあり、まずは近代文語を処理することを想定した際の、既存の規程の見直しから出発することとした。本稿での「近代語」は、とくに断りがない限りは近代文語を想定している。

２．近代語での単位認定の問題点と、その処理方針

２．１　辞書登録に関わる事項

近代語コーパスの形態論情報の付与にあたっては、形態素解析辞書 UniDic を用いている。近代語コーパス構築作業では、UniDic にそれまで登録されていない、近代語特有の語彙が多数出現する。そこでそれらの語彙を処理するためには、まずは辞書側にその語を登録しなければならないが、現代語から出発した UniDic に近代語特有の語を追加登録しようとすると、問題が生じる場合がある。

２．１．１　表記上、短単位分割が不可能な場合

UniDic に登録する辞書情報も、短単位ごとに登録していく、ということになるが、近代語のテキストを扱っていると、表記の都合上、短単位に分割できないという場合が多数生じる。

【例】

「非る」（あらざる）　　所謂政府なる者亦人に非るなし
「加之」（しかのみならず）　加之官職の設概するに古に簡にして後世に繁く
「不然」（しからざれば）　　不然則又豪奪の賊なり

これらはそれぞれ短単位分割の原則としては「あら／ざる」「しか／のみ／なら／ず」「しから／ざれ／ば」と分割されるべきだが、表記上、そのような分割ができないため、このような出現形に対しての処理方針を定めねばならない。
現時点では大筋において以下のような方針としている。

（１）使用頻度が高く、ある程度慣習的に認められていると見られる出現形に関しては、例外として全体を１短単位として辞書登録する。「非る」「加之」などがこれにあたる。近代語特有の表記に対し、辞書上１短単位と認めたものには、他に以下のようなものがある。

於是　ここにおいて何者　なんとなれば
而　　しこうして
遮莫　さもあらばあれ
乍併　しかしながら

（2）それほど使用頻度が高くないものに関しては、短単位に分割できるよう、テキスト側を書き換える

【例】
「不然」（しからざれば）　原文：不然則又豪奪の賊なり→コーパス：然らざれば・・・

２．１．２　語自体は既登録だが、時代によって品詞認定が異なる場合

時代とともに品詞認定が変わる語もあるので、それらについても処理方針を定めておかねばならない。中古和文では以下のように処理することとしている。

（中古和文）
例えば、現代語として副詞「夜な夜な」が UniDic には登録されているが、中古までさかのぼると、「夜な夜な」に名詞としか認められない用法が出現する。

【例】おはしましし夜な夜なのありさま

このような場合、「夜な夜な」の品詞情報を「副詞」から「名詞－副詞可能」に書き換えるという方法も考えられるが、原則として、既登録の語に対して、別の時代への対応へのために既登録語の品詞情報を書き換えるということは行わない。よって、残る方法としては（1）品詞認定の実際上のずれは多少無視して、既登録の語をそのまま使う、（2）中古和文特有の用法を、別品詞として新規登録する、という 2 つとなる。どちらの方法を採るかの目安は、以下に示すとおりである。

（1）既登録語の品詞が名詞であり、中古和文において副詞用法や形状詞用法が出現した場合→既登録の名詞をそのまま使用

（2）既登録の品詞が名詞以外であり、中古和文において名詞用法が出現した場合→別語として名詞を新規登録

※上記（1）、品詞認定の実際上のずれは多少無視して、既登録の語をそそのまま使うという処理を行うのは、あくまで意味の面でも共通性が見られる場合に限る。意味の面で大

きな違いが認められる場合は、別語として新規登録する。

【例】いかさま
現代語：名詞
中古和文：形状詞を新規登録

既登録の語が名詞であり、時代をさかのぼった結果、形状詞用法も認められた場合、基本方針としては（１）が適用され、既登録の名詞として処理される。しかし、現代語で名詞として登録された「いかさま」は「ごまかし」「いんちき」の意であるのに対し、かつての形状詞用法「いかさま」は疑問、「どのよう」「どんなふう」の意である。このように意味が異なるものに関しては、別語として形状詞「いかさま」を新規登録した。

（近代語）
近代語においても、現代語とは品詞認定が異なる場合の処理は、以下のとおりとする。

原則：既登録の語の品詞情報の書き換えは行わない。

品詞認定が異なる語の処理方針：

（１）既登録語の品詞が名詞であり、近代語において副詞用法や形状詞用法が出現した場合→新語彙素を登録することなく、既登録の名詞を使用する。（中古和文での方針と同様）

（２）既登録の品詞が名詞以外で、近代語ににおいて名詞用法が出現した場合
（２‐A）既登録語が副詞または形状詞で、近代では名詞としての用法もあるが、副詞・形状詞のとしての意味もそのまま有するものは新語彙素を登録することなく、既登録の副詞または形状詞を使用する。

【例】現 UniDic「副詞」―明六「名詞」
習ふに漸次を以てし行ふに歳月を以てし
故に我國方今の景況に於て國人の智識一層を進めば輸出入の差從て一層を加ふべし
縦令兵力今より數層を加ふも野を轉じて文と爲さず
倍氏の理學の大凡後篇に見ゆ

【例】現 UniDic「形状詞」―明六「名詞」
嘗て國家の禍を推すに歸する所は官吏人民の姑息にあり
植民地愈々廣大に至れり

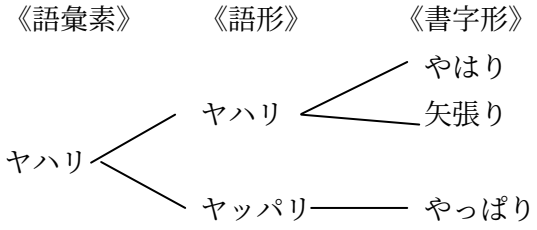
倫理の當然に従て必ず是等の惡風俗を禁止するの憲法を設定するㇵ固より緊要のㇵと云ふ可きなり
凡そ簡易明白を歓び
普通の中に特別を附せざるㇵを得ざる者あり

（２‐B）上記（２‐A）にあたらないものは、別語彙素として名詞を新規登録する。

２．１．３　外来語の同語異語判別

形態素解析辞書 UniDic の見出し語は、以下に示すように階層化された形で格納されている（小木曽・中村 2011 参照）。

【例】



新たに辞書情報を登録する際、「語彙素」のレベルから登録すべきか、「語形」レベルで登録すべきか迷う場合がある

【例】

歩く（アリク）・・・・語彙素「歩く（アルク）」とは別の「語彙素」か。
語彙素「歩く（アルク）」の一「語形」か。

このような場合の判断の揺れを抑えるために、『規程集』ではどのようなものを一つの語彙素、または語形にまとめ、どのようなものは別にするかに関しても、「同語異語判別規程」を設けている。ただし、『規程集』での同語異語判別規程はあくまで現代語を対象としたものであり、現代語では想定されていなかった差異の扱いに関しては、新規に方針を定めねばならない。近代語コーパス作成に当たっては、特に外来語において、想定されていなかった同語異語判別の問題が生じる。

【例】

レホルメルス　（reformers）

上例「レホルメルス」は reformer（今日的には「リフォーマー」が一般的なカタカナ表記か）複数形 reformers を読んだものと思われるが、語彙素「リフォーマー」の語形に「レ

ホルメルス」を登録すべきか、語彙素から「レホルメルス」を立てるべきかは、既存の規程からは判断できない。

また、出現形「レホルメルス」に関しては、注釈書その他から原語が「reformers」であることがたまたま推定できたが、そもそも、原語が何なのかわからない外来語も多数出現する。その際、出現した外来語に対し逐一原語を推定し、その上で同語異語判別をしていくというのは､近代語コーパスの試作段階においては煩雑なだけであり､非効率である。そこで、近代語コーパスの試作段階においては、外来語の同語異語判別は以下の通りに行うこととした。

外来語の同語異語判別：

『規程集』所収の規程を適用できる範囲内のものは、その規程に従って判別する。
既存の規程では判断できない語に関しては、原則として出現形をそのまま語彙素として登録する。

補則・外来語の語彙素・語形の表記に関して：

『規程集』では、外来語の語彙素・語形に用いる仮名・符号の種類が定められており、その範囲は近代語での語彙素・語形表記でも遵守するものとする。そのため、出現形の仮名表記が『規程集』に定められた使用可能範囲から外れている場合には、使用可能範囲内の表記に改める。出現形の仮名表記を『規程集』での使用可能範囲内の仮名表記に改める基準に関しては、本稿末尾の資料１を参照のこと。

２．２　単位境界認定に関わる事項

時代が変われば言語の実態も変わる。そのため、「この」の扱いを例に挙げたように、単位境界の認定が、現代語と中古和文で異なる場合がある。これから構築していく近代語の規程は、時代の上では、既存の２つの規程（現代・中古和文）の間に位置することになる。そこで、現代語と中古語とで明らかに扱いが異なるものに関しては、近代語ではどちらの時代と同じ扱いにするか、あるいはどちらとも異なる独自の扱いをするか、を定めていかねばならない。

[１]連体詞

「この」「その」などの「連体詞」の扱いは、以下の通り、現代語と中古和文で単位の切り方が異なる。

現代語：／この／　　／その／　（連体詞）
中古和文：こ／の　　そ／の　（代名詞＋格助詞）
※中古では「連体詞」という品詞を原則として認めない。

近代語では、いわゆる連体詞の扱いに関しては、中古和文に準ずるものとする。
つまり、品詞としての「連体詞」は認めず、現代語での連体詞に当たるものは「代名詞＋格助詞」に分割する。

近代語：こ／の　　そ／の

[2]「異なる」

現代語：／異なる／（動詞）
中古和文：異／なる（名詞＋断定の助動詞「なり」）

近代語では、中古和文側の処理に合わせ、名詞＋助動詞とする。

近代語：異／なる

[3]「－の～」「－つ～」「－が～」
（現代語）
短単位認定規程においては、助詞「の」「つ」「が」は、それだけで１短単位として分割するのが原則だが、現代語においては「の」「つ」「が」を含んだ全体を１短単位と認定した方が適当であるとの判断から、例外的に分割しないものがある。

【例】
「－の～」：　／日の丸／　　／床の間／　　／竹の子／
「－が～」：　／天が下／　　／雁が音／　　／剣が峰／
「－つ～」：　／国つ神／

これらについては規程集内に資料「要注意語」を設け、一覧が整備されている（小椋・小磯ほか 2011 参照）。

※「－の～」で１短単位とするものを選定するにあたっては、以下の事項をおおよその目安とする。

①「の」が読み添えとなっているもの

【例】斎宮（いつきのみや）　　対屋（たいのや）

②「の」の直前の要素が被複形のもの

【例】木の葉　　目の当たり

③「－の～」全体の品詞が名詞以外となるもの

【例】　案の定　　気の毒

④「－の～」が動植物名等を表すもの

【例】　卯の花　　竹の子　　泥の木

※「－つ～」に関しては、現代語で格助詞「つ」はもはや使用されてないため、原則として全ての「－つ～」に関して「つ」を格助詞として分割することはせず、全体を 1 短単位とする。

（中古和文）

中古和文では、当時の実態を踏まえて、「－の～」「－つ～」「－が～」全体を 1 短単位と認定する範囲を限定する。

【例】

現代語：　　／身の程／　　／天が下／

中古和文：　身／の／程　　天／が／下

中古和文において 1 短単位とするものについても、中古和文規程集に一覧が収録されている（小椋・須永 2012 参照）。

※中古和文において「－の～」「－つ～」「－が～」で 1 短単位とするものを選定するにあたっての目安は、現代語の「－の～」に関する目安①～④のうち、①③④を、中古和文の「－の～」「－つ～」「－が～」に適用する。

（近代語）

以上のように、「－の～」「－つ～」「－が～」に関しては、全体で 1 短単位とする範囲が、現代語と中古和文で異なる。近代語では、当面は現代語で定めた範囲に従って 1 短単位と認定する。

２．３　コーパスへの形態論情報付与に関わる事項

実際にコーパス上に形態論情報を付与していく際には、品詞認定等の面で判断に迷う場面がさまざまに生じる。そのような場合のためにも、『規程集』では品詞の判別基準なども設定されている。品詞等の判別に関しても、現代語および中古和文では想定されていなかった近代語特有の問題があるため、それらに対応するために近代語用の判別基準を設定しておく必要がある。以下にはその概略を示す。

以下に示すものはあくまで概略であり、実際の形態論情報付与作業においては、近代語特有の問題が個別的に生ずる場合があり、それらに対しても個別対応の詳細な規程を設けていかなければならない。しかし、現時点で扱った資料は『明六雑誌』のみであり、ここでの個別事例を一般化して規程とするにはまだ資料不足の段階である。今後の作業も含め、実例と、実際の処理方法を蓄積し、なるべく一般化した形でそのような問題に対応できる規程を細部にわたって整備していくことが、将来的な『近代語コーパス用規程集』での課題となる。

２．３．１　活用形・活用型認定

[１]活用型が上二段か四段か判じ難い場合

【例】　恨む　忍ぶ

活用型が上二段か四段か判じ難い場合があるが、その場合、『明六雑誌』の範囲内では、全体的な傾向から見て上二段としておく（近代語全体に関しては未定）。

[２]活用型が文語なのか口語なのか判じ難い場合

出現活用形が文語なのか口語なのか区別がつかない場合は文語としての処理を優先する。

【例】　斬棄てて
（文語下二段／口語下一段　　→　　文語下二段を優先）

活用形上、口語としてしか認められないもののみを口語として処理する。

【例】斬棄てる
（口語下一段）

[３]仮名遣いと活用の行

出現形の仮名遣いからすると活用の行が変わってしまうものは、別語形とする。

【例】

据へて・・・規範的な古典では「据ゑ」。

「文語下二段－ワ行」として既登録。　「文語下二段－ハ行」を新規登録

２．３．２　品詞認定

形態論情報の付与に当たって、最も判断の揺れが生じやすいのは品詞認定である。同一の出現形に対し、選択肢として辞書上に複数の品詞が用意されている場合、そのうちのどれを使うべきかを詳細に規定しておかないと、コーパスの品詞情報が揺れてしまう。例えば疑問文末の「や」に対しては、「係助詞」「終助詞」「副助詞」という 3 種類の処理がありうる。このように、揺れてしまいかねない個所を洗い出し、処理方針を確定させていく、という作業が、最終的な規程集の作成においては必要になってくる。ここでは品詞認定に悩むという事例自体の紹介として、代表的な数例を紹介するにとどめる。近代語用の、個別の判別基準に関しては次節でいくつか詳細に取り上げる。

[１]疑問・反語の助詞「や」「か」

疑問・反語の「や」「か」に関しては係助詞とする。（副助詞、終助詞は使わない）

[2]「間投助詞」と認定したい助詞の扱い

「間投助詞」と認定したい助詞に関しては、「終助詞」として処理する。（UniDic の助詞の分類には「間投助詞」が存在しないため）

「古池や蛙飛び込む水の音」の「や」のような、いわゆる切れ字についても終助詞とする。

[3]出現形「に」の判別基準

出現形「に」に関しては、助詞と認定すべきか、断定の助動詞（「だ」「なり」）の連用形と認定すべきか、判断が揺れやすい。そのため、現代語の『規程集』でも判別基準を設けたが、対象とする時代が異なると、判別基準も修正が必要になる。

現代語での判別基準の概略は、以下のとおりである。

現代語での出現形「に」の判別基準：

①形状詞　　→助動詞「だ」連用形（中古では「なり」）
②文語形容詞連体形（なきにしもあらず）→助動詞「なり」連用形
③他　　→格助詞

しかし時代をさかのぼり、中古和文を対象とする際には、この基準をそのまま適用することはできなくなる。例えば中古和文においては、上記①②以外にも助動詞「なり」の可

能性も生じるため、そのまま適用はできない。そこで中古和文では中古和文用に「に」の判別基準を修正したが、近代語ではまた別個に、現代語とも中古語とも異なる判別基準の設定が必要となったため、近代語用の「に」の判別基準を設定しなおした。近代語での「に」の判別基準は本稿末尾の資料２参照のこと。

２．３．３　読み

コーパスに形態論情報を付与していく際、漢字などに対しても読みを与えていかねばならないが、現実的には読みが不明な場合や、ひとつに確定できない場合も多い。そこで、読みを与える際に迷いそうな場合、どのように処理するかに関しても、方針を定めておかねばならない。

慣例として最も一般的、常識的な読みを採っておく、ということを原則とするが、「一般的、常識的な読み」が確定しない場合は以下の目安に従って読みを与える。

原則１　音読み・特に漢音を優先する。

【例】
一人　イチニン＞ヒトリ
給水場　キュウスイジョウ＞キュウスイバ
重複　チョウフク＞ジュウフク

原則２　音便形であるかわからないものは、元の形を優先する。

【例】
撃て　ウチテ＞ウッテ

原則３　読み添えの「の」を補って読むのは地名、姓のみとする。

【例】
藤原道長　フジワラノ　ミチナガ
中関白　　ナカ＿カンパク（「ナカノカンパク」とは読まない）
後出師表　ゴスイシ＿ヒョウ（「ゴスイシノヒョウ」とは読まない）

原則４　原文にルビがあっても、その読みに従わない場合がある。

[１]漢語に対し、外来語のルビがついている場合

（A）通常の漢語として認められるものは、原文ルビを無視して漢語として読む。

【例】「玩弄品」　原文ルビ：トイス
→コーパス上の読み：ガンロウヒン
「造鉄術」　原文ルビ：アーツオブメイキング
→コーパス上の読み：ゾウテツジュツ

（B）漢語としての用法が見当たらないものは、原文ルビに従い、原文ルビの指す外来語の書字形とする。

【例】「聖礫」　原文ルビ：クルス
「聖礫」という漢語用例見当たらず
→コーパス上の読み：クルス（語彙素「クルス」の書字形）

[2]原文ルビの読みが、熟字訓として定着しているとは言い難いものに関しては、原文ルビを無視して通常の漢語としての読みを与える。

【例】「痴漢」　原文ルビ：バカモノ
→コーパス上の読み：チカン
「吾人」　原文ルビ：ワレラ
→コーパス上の読み：ゴジン

３．今後の課題

以上、『明六雑誌コーパス』での形態論情報付与を事例として、近代語コーパスのための規程整備の必要性を確認し、現時点での近代語用規程の主要なものを概観した。規程整備にあたっては、単位境界認定から品詞認定、表記、読みの問題等、多岐にわたる事項の処理方針を詳細に設定していかねばならない。現代語を対象とした『規程集』は本稿で紹介した言語単位、「短単位」に関する部分のみで２００ページほど、これを前提とした中古和文用の追加規程集（須永・小椋 2012）で１００ページほどの分量となっており、本稿で紹介した近代語用規程の分量と比較しても明らかな通り、本稿は将来作成すべき近代語用規程集の中心原則、しかもその一部という位置づけになる。今後『明六雑誌』以外の近代語資料も見渡す過程で、今回設定した中心原則に詳細な補則を設定したり、修正を施したり、さらには原則から新設するなどして、より詳細な近代語用規程集を作成するのがこれからの課題となる。本稿で紹介した規程はおおよその原則案という位置づけであることは先に述べたが、今後この原則案を基に、近代語用規程として完成させた形のサンプルとして、細部まで整備した規程を本稿末尾に参考資料として掲載する。作業上特に問題になりやすい表記の問題、品詞認定の問題からそれぞれ「仮名表記される外来語の語形の定め方」「出

現形「に」の判別基準」を「資料１」「資料２」として示す。本稿で紹介した、他の規程ひとつひとつに関しても参考資料に掲げるような形で詳細に設定していくというのがこれからの作業となる。

『明六雑誌コーパス』構築という今回の作業に限定しても、現実の言語事実は多種多様であり、原則としての規程を設定してなお、様々な面で個別に判断を求められる場面が多くみられた。そのような「揺れ」になりやすい箇所にも対応し、統一的な処理をするためにこそ、規程は必要なものであり、今後さらなる事例の蓄積を通し、一般化した形で個々の問題に対応できる規程を細部にわたって整備していく予定である。

文　献

小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕(2011)『『現代日本語書き言葉均衡コーパス』形態論情報規程集　第 4 版（上）（下）（国立研究所内部報告書 LR-CCG-10-05-01,02）』（国立国語研究所）

小椋秀樹、須永哲矢(2012)『中古和文 UniDic 短単位規程集』（『2009-2011 年度科学研究費補助金　基礎研究 (C)「和文系資料を対象とした形態素解析辞書の開発」成果報告書 2（課題番号 21520492））

URL

UniDic	http://download.unidic.org
中古和文 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic
近代文語 UniDic	http://www2.ninjal.ac.jp/lrc/index.php?UniDic

資料１

仮名表記される外来語の語形の定め方

小椋・小磯ほか（2011）では、現代語においてカタカナ表記される外来語の出現形から語形を定める規程を定める（pp.85-87）。おおよそそれは、「外来語の表記」（平成３年内閣告示）の仮名・符号の表（表1）に基づき、外来語の語形の表記に用いることができる仮名・符号の範囲（表１中の本表・付表A）を定め、その範囲を超える出現形について、範囲内に収める方法を記したものとなっている。

表 1

本表

ア　イ　ウ　エ　オ
カ　キ　ク　ケ　コ
サ　シ　ス　セ　ソ
タ　チ　ツ　テ　ト
ナ　ニ　ヌ　ネ　ノ
ハ　ヒ　フ　ヘ　ホ
マ　ミ　ム　メ　モ
ヤ　　ユ　　ヨ
ラ　リ　ル　レ　ロ
ワ
ガ　ギ　グ　ゲ　ゴ
ザ　ジ　ズ　ゼ　ゾ
ダ　　　デ　ド
バ　ビ　ブ　ベ　ボ
パ　ピ　プ　ペ　ポ
キャ　　キュ　　キョ
シャ　　シュ　　ショ
チャ　　チュ　　チョ
ニャ　　ニュ　　ニョ
ヒャ　　ヒュ　　ヒョ
ミャ　　ミュ　　ミョ
リャ　　リュ　　リョ
ギャ　　ギュ　　ギョ
ジャ　　ジュ　　ジョ
ビャ　　ビュ　　ビョ
ピャ　　ピュ　　ピョ

ン（撥音）
ッ（促音）
ー（長音符号）

シェ　　チェ　　ツァ　ツィ　ツェ　ツォ　　ティ　トゥ　　ファ　フィ　フェ　フォ　　ジェ　　ディ　ドゥ　デュ　フュ

付表A

ウィ　ウェ　ウォ　　ツィ

付表B

イェ　　クァ　クィ　クェ　クォ　　グァ　　ヴァ　ヴィ　ヴ　ヴェ　ヴォ　　デュ　　ヴュ

近代語の短単位においても、外来語の語形の表記に用いることができる仮名・符号は現代語の規程に定める範囲に倣うこととする。ただし、近代語では出現形で用いられる仮名・符号の種類が現代語より多様であるといった、現代語にはない実態があるため、現代語の規程をそのまま近代語に当てはめることはできない。そこで、近代語用に出現形から語形を定める規程を改めてここにまとめるものである。

なお、以下の記述の中でいう「辞書」とは、『大辞林』第 2 版と『日本国語大辞典』第 2 版を指す。両辞書の記述が異なる場合、原則として『日本国語大辞典』第 2 版に従う。

（1） 長音について

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ「ア・イ・ウ・エ・オ」を用いた出現形で、「ア・イ・ウ・エ・オ」が長音を表すと考えられるものについては、原則として長音符号「ー」を用いた形を語形とする。

【例】アパアトメント→アパートメント、アンコオル→アンコール

ただし、辞書の見出し（空見出しを除く。以下同）で「ア・イ・ウ・エ・オ」が用いられている場合は、「ア・イ・ウ・エ・オ」を語形とする。

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ小書き「ァ・ィ・ゥ・ェ・ォ」を用いた出現形は、長音符号「ー」を用いた形を語形とする。

【例】オスカァ→オスカー、ロセツチィ→ロセッチー

イ・エ・オ列の仮名の後にそれぞれ「ヰ・ヱ・ヲ」を用いた出現形で、その「ヰ・ヱ・ヲ」が長音を表すと考えられるものについては、長音符号「ー」を用いた形を語形とする。

【例】ルビヰ→ルビー、ペヱトル→ペートル、レボヲリユーシヨン→レボーリューション

上記の規程を適用した結果、語形で長音符号が複数連続する場合は、一つに統合する。

【例】ウンゾォール（→ウンゾーール）→ウンゾール

（2） 本表の仮名「ツァ」「ツェ」「デュ」「フュ」を用いた出現形について

出現形と同じ仮名を用いた形を語形とする。

（3） 本表・付表 A に見られない仮名・符号について

仮名・符号ごとに以下のように語形を定める。

○「ヂ」

書字形「ヂヱ」は「ジェ」を語形とする。

【例】ブールヂヱー→ブールジェー

拗音「ジョ」の代わりに用いられていると見なせる書字形「ヂヲ」は、「ジョ」を語形とする。

【例】レリヂヲス→レリジョス

上記以外の書字形「ヂ」は「ジ」を語形とする。

【例】ラヂオ→ラジオ

○「ヅ」

書字形「ヅュ」は「ジュ」を語形とする。

【例】（未確認）

上記以外の書字形「ヅ」は「ズ」を語形とする。

【例】ヅボン→ズボン

○「ヴ」

書字形「ヴァ」「ヴヮ」「ヴハ」は「バ」を語形とする。

【例】ヴァイオレット→バイオレット、リヴヮー→リバー、ヴハンクヴアー→バンクバー

書字形「ヴィ」「ヴヰ」は「ビ」を語形とする。

【例】ヴィクトリア→ビクトリア、シヴヰリゼーション→シビリゼーション

書字形「ヴェ」「ヴヱ」「ヴヱ゛」は「ベ」を語形とする。

【例】ヴェクトル→ベクトル、ヴヱネチア→ベネチア、ヴヱ゛イ→ベイ

書字形「ヴォ」「ヴヲ」は「ボ」を語形とする。

【例】ヴォルト→ボルト、レヴヲリユーシヨン→レボリューション

書字形「ヴャ」「ヴュ」「ヴョ」は「ビャ」「ビュ」「ビョ」を語形とする。

【例】ヴュルテンベルヒ→ビュルテンベルヒ

上記以外の書字形「ヴ」は「ブ」を語形とする。

【例】ヘヴン→ヘブン

○「ワ゛」「ヷ」
書字形「ワ゛」「ヷ」は「バ」を語形とする。

【例】アワ゛ンチユール→アバンチュール、ペンシルワ゛ニア→ペンシルバニア

○小書き「ヮ」
「ウヮ」は「ワ」を語形とする。

【例】ドウヮー→ドワー

「ヴヮ」は「バ」を語形とする。

【例】リヴヮー→リバー、ドヴヮー→ドバー

上記以外の書字形「ヮ」は「ワ」を語形とする。

【例】インクヮイアラ→インクワイアラ

○「ヰ」
書字形「ウヰ」は「ウィ」または「イ」を語形とする。「ウィ」「イ」のどちらを語形とするかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ルードウヰヒ→ルードウィヒ、ダーウヰン→ダーウィン
ウヰスキー→ウイスキー

書字形「テヰ」「デヰ」「フヰ」「ツヰ」はそれぞれ「ティ」「ディ」「フィ」「ツィ」を語形とする。

【例】テヰール→ティール、グランデヰー→グランディー、フヰリツピン→フィリッピン、ツヰング→ツィング

書字形「クヰ」「グヰ」はそれぞれ「クイ」「グイ」を語形とする

【例】クヰーン→クイーン、グヰツチヨク→グイッチョク

書字形「ヴヰ」「ヰ゛」は「ビ」を語形とする。

【例】シヴヰリゼーション→シビリゼーション、スカンディナヰ゛ア→スカンディナビア

イ段の仮名に続く書字形「ヰ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】ルビヰ→ルビー

上記以外の書字形「ヰ」は「ウィ」または「イ」を語形とする。「ウィ」「イ」のどちらの語形とするかは､辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ヰーン→ウィーン、サンドヰツチ→サンドウィッチ

○「ヰ゛」「ヸ」
書字形「ヰ゛」「ヸ」は「ビ」を語形とする。

【例】スカンディナヰ゛ア→スカンディナビア

○「ヱ」
書字形「シヱ」「チヱ」「ツヱ」「フヱ」「ウヱ」はそれぞれ「シェ」「チェ」「ツェ」「フェ」「ウェ」を語形とする。

【例】シヱークスピーア→シェークスピーア、マンチヱスター→マンチェスター、カフヱ→カフェ、ツヱペリン→ツェペリン、ノルウヱー→ノルウェー

書字形「ジヱ」「ヂヱ」は「ジェ」を語形とする。

【例】サージヱン→サージェン、ブールヂヱー→ブールジェー

書字形「ヴヱ」は「ベ」を語形とする。

【例】ヴヱネチア→ベネチア、アドヴヱンテージ→アドベンテージ

エ段の仮名に続く書字形「ヱ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】ペヱトル→ペートル、メッテヱル→メッテール

上記以外の「ヱ」は「エ」または「ウェ」を語形とする。「エ」「ウェ」のどちらを語形とするかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ヱリザベス→エリザベス、サイヱンス→サイエンス
スヱーデン→スウェーデン

○「ヱ゛」「ゑ」
書字形「ヴヱ゛」「ヱ゛」「ゑ」は「ベ」を語形とする。

【例】ヴヱ゛イ→ベイ、レヱ゛ル→レベル

○「ヲ」
書字形「ツヲ」「フヲ」「ウヲ」はそれぞれ「ツォ」「フォ」「ウォ」を語形とする。

【例】ホーヘンツヲルレルン→ホーヘンツォルレルン、カリフヲルニヤ→カリフォルニヤ、ウヲートルロー→ウォートルロー

書字形「ヴヲ」は「ボ」を語形とする。

【例】レヴヲリユーシヨン→レボリューション

拗音「ョ」の代わりに臨時的に「ヲ」が用いられたと見なせるものは、「ョ」を用いた形を語形とする。

【例】ジヲルジ→ジョルジ、レリヂヲス→レリジョス

オ段の仮名に続く書字形「ヲ」で長音を表すと考えられるものは、長音符号「ー」を語形とする。

【例】レボヲリユーシヨン→レボーリューション

上記以外の書字形「ヲ」は「オ」または「ウォ」を語形とする。「オ」「ウォ」のどちらを語形とするかは､辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】ナポレヲン→ナポレオン、ガリレヲ→ガリレオ、ヲデッサ→オデッサ
コーンヲール→コーンウォール、ヲートルロー→ウォートルロー、ヲルヅヲルス→ウォルズウォルス

○「ヲ゛」「ヴ」

書字形「ヲ゛」「ヴ」は「ボ」を語形とする。

【例】ヲ゛ルト→ボルト、ヲ゛ルテール→ボルテール

○「ヽ」

前の１文字を書字形とする。

【例】クロヽホルム→クロロホルム、ダヽイズム→ダダイズム、ヒッポヽタマス→ヒッポポタマス

ただし、前の文字が濁音・半濁音の場合、清音化した文字を書字形とする場合がある。

【例】（未確認）

上記の結果、他の規程を適用する必要のある形となる場合は、適用後の形を語形とする。

【例】オヽシス（→オオシス）→オーシス

○「ゞ」

前の文字が清音の場合、濁音化した文字を書字形とする。

【例】ハヾローフスク→ハバローフスク

前の文字が濁音の場合、同じ文字を書字形とする。

【例】バヾリヤ→ババリヤ

（4） 本表・付表Aにない、小書き「ャ・ュ・ョ・ァ・ィ・ゥ・ェ・ォ」を含む出現形について

「クュ・スュ・ツュ・ヌュ・ムュ・ルュ・グュ・ズュ・ブュ・プュ」は、それぞれ「キュ・シュ・チュ・ニュ・ミュ・リュ・ギュ・ジュ・ビュ・ピュ」を語形とする。また、「ヅュ」は、「ジュ」を語形とする。

【例】アヴァンツュール→アバンチュール、クルューゲル→クリューゲル、トリブューン→トリビューン、ヂユプュイトラン→ジュピュイトラン

「ケョ・セョ・テョ・ネョ・ヘョ・メョ・レョ・ゲョ・ゼョ・ベョ・ペョ」は、それぞれ「キョ・ショ・チョ・ニョ・ヒョ・ミョ・リョ・ギョ・ジョ・ビョ・ピョ」を語形とする。

【例】テョディー→チョディー、ゲョーテ→ギョーテ、ゼョン→ジョン

「ウァ」は「ワ」を語形とする。

【例】ハーウァード→ハーワード、ショツペンハウァー→ショッペンハワー

ア・イ・ウ・エ・オ列の仮名の後にそれぞれ小書き「ァ・ィ・ゥ・ェ・ォ」を用いた出現形は、長音符号を用いた形を語形とする。

【例】オスカァ→オスカー、ロセツチィ→ロセッチー

拗音「ャ・ュ・ョ」の代わりに臨時的に小書き「ァ・ゥ・ォ」が用いられたと見なせる出現形は、「ャ・ュ・ョ」を用いた形を語形とする。

【例】ギリシァ→ギリシャ、カリゥーム→カリューム、ジォン→ジョン

上記以外のものについては、大書きの仮名に直したものを語形とする。

【例】ワィズマン→ワイズマン、ニコラゥス→ニコラウス、スクェーア→スクエーア、ロェスレル→ロエスレル

ただし、辞書の見出しや原音を参照して異なる形を語形とする場合がある。

（５）　小書き仮名が大書きされた出現形について

本表・付表Ａにある小書きの仮名を用いた形や、上記規程であげた出現形に用いられる小書きの仮名を用いた形について、小書きの仮名ではなく大書きの仮名を用いた形が出現形となる場合がある。その場合、大書きの仮名を小書きに直した上で、必要な規程を適用して語形を定めることになる。

【例】ナチユラル→ナチュラル、インテリゲンツイア→インテリゲンツィア、ヴアイオリン（→ヴァイオリン）→バイオリン、ヴワルカン（→ヴゎルカン）→バルカン、ダブルユー（→ダブルュー）→ダブリュー

大書きの仮名が小書きに直すべきものなのか大書きのままとすべきものなのかは、辞書の見出しや原音を参照して語ごとに定めるものとする。

【例】「シヤ」シヤツ→シャツ、アカシヤ→アカシヤ
「ニヤ」コニヤツク→コニャック、アンモニヤ→アンモニヤ
「テイ」ソサイテイ→ソサイティ、ステイション→ステイション
「デイ」コメデイ→コメディ、デイリー→デイリー

（６）　その他

「ファ・フィ・フェ・フォ」の代わりに「フハ・フヒ・フヘ・フホ」が用いられたと見なせる出現形は、「ファ・フィ・フェ・フォ」を語形とする。

【例】アスフハルト→アスファルト、ソフヒア→ソフィア、フヘルヂナンド→フェルジナンド、カリフホーニア→カリフォーニア

資料２
出現形「に」の品詞判別基準

出現形「に」について、断定の助動詞「なり」の連用形かそれ以外の品詞（格助詞または接続助詞）かの判断基準について、以下に記述する。

原則、次の①～⑤に該当するものは断定の助動詞「なり」とし、それ以外を格助詞または接続助詞とする。

①先行語が形状詞のもの

【例】容易に其舊習を改めし者なるべし（助動詞）
僕竊かに疑なき能はず（助動詞）

ただし、形状詞が名詞的に用いられ、かつ「に」が意味上「～で（～であって）」と解せないものは、格助詞と判断する。これは本稿２．１．２に述べたように、現状の UniDic で形状詞として登録されている語が近代語で名詞として用いられても、既登録の形状詞をそのまま用いるために起こる事象である。

【例】多少の人力財用を無用有害に費す（格助詞）

②「あり」などの存在詞が後続し、意味上「～で（～であって）」と解せるもの（「～に（は）あらず」のように「あり」が打消の助動詞「ず」を伴うことが多い）

【例】是悦ぶべくして惡むべきにあらず（助動詞）
未曾有の事を新に始るには非ず（助動詞）

意味上「～で（～であって）」と解せないものは、格助詞と判断する。

【例】耶蘇教を奉ずる一國ここにあり（格助詞）
政府の強きを致すは天下人民の同心を致すに在り（格助詞）

③「～にして」の形で用いられ、意味上「～で（～であって）」と解せるもの

【例】夫日曜は七曜の一にして、毎週の首なり、（助動詞）
倍根より以前の心靈の理學は空理のみにして實證なし（助動詞）

意味上「～で（～であって）」と解せないものは、格助詞と判断する。

【例】楊朱墨翟の相反せる兩極を合一にして之を社交の一大例規とし（格助詞）
スチプレーション箇條書を頼みにして弊害を防ぎ止めんとするの一法あるのみ（格助詞）

④係助詞「や」が後続し、「〜にやあらむ」と「あらむ」が補えそうなもの

【例】ただ艸木の幹一つにして枝八十に別るるを見て本は一つなりと誤れるにや（助動詞）

⑤やう（様）に、ごとくに

【例】我邦の制とは大に相違する所ある様に覺ゆるなり（助動詞）
正金は年々常例の如くに外出するなり（助動詞）

『明六雑誌コーパス』の仕様

近藤　明日子　（国立国語研究所コーパス開発センター）[1]
田中　牧郎　　（国立国語研究所言語資源研究系）[2]

１．はじめに

本稿では、本プロジェクトで設計している「近代語コーパス」のモデルとして構築した『明六雑誌コーパス』の仕様について説明する。

「近代語コーパス」は、明治時代から昭和時代ごろまでを対象に、近代日本語を代表でき、近世までの日本語から現代日本語への変化の過程を歴史的にたどることができるものにすることが望まれる。そのためには、多種多様な資料をコーパス化の対象にしていく必要があるが、最初に取り組む資料として『明六雑誌』を選んだ。

２．『明六雑誌』を選ぶ理由

『明六雑誌』は、明治 6（1873）年に学術啓蒙を目的に結成された明六社の機関誌で、明治 7（1874）～8（1875）年に、1 号から 43 号まで発行された。森有礼、津田真道、西周、西村茂樹、中村正直、加藤弘之、福沢諭吉、箕作麟祥ら 16 名が執筆している。西洋の近代思想を普及するために書かれた広範な論説が 155 編おさめられている。大半は漢文訓読風の文語体であるが、中には演説的な口語体も含まれている。思想史上の重要資料とされてきたが、日本語学においても、特に西洋からの新概念を取り入れるために必要とされた新漢語の資料として注目されてきたものである。明治前期の日本語研究資料として重要なものであり、複製本、注釈書、総索引なども整備され[3]、研究の蓄積もある[4]。このような特徴から、「近代語コーパス」が対象とする時代の初期の資料として、最初に取り組むのに適切なものだと判断した[5]。

本プロジェクトにおける「近代語コーパス」は、2005 年に公開した最初の近代語コーパスである『太陽コーパス』を踏まえ、これを発展させる形で設計を行っている。『太陽コーパス』は、雑誌の本文を、記事、引用、擬似的な文の単位で構造化し、マークアップ言語 XML でタグ付けをしたものであり、記事や引用については著者、話者、文体などの情報をタグの中に書き込み、校訂注記や異体字などの情報も、本文の当該箇所にタグ付けをして埋め込んだものである。日本語史研究資料におけるはじめての構造化テキストタグ付きコーパスである（田中 2005）。ところで、コーパスとして言語研究に本格的に利用して

[1] kondo@ninjal.ac.jp
[2] mtanaka@ninjal.ac.jp
[3] 『明六雑誌』の注釈書と校訂本文に山室・中野目（1999-2009）があり、複製本・総索引に高野・日向（1998）がある。
[4] 『明六雑誌』の言語研究に、神奈川大学人文学研究所（2004）や高野（2004）がある。
[5] 「近代語コーパス」の資料選定の考え方については、本報告書に収録した田中牧郎「近代語コーパスにおける資料選定の考え方」を参照。

いくためには、文章を構造化するだけにとどまらず、単語のレベルまで構造化を行うことが望まれる。『太陽コーパス』の開発段階からそれは意識されていたが、当時の技術では実現が困難であった。しかし、近年の研究の進展により、近代語テキストに対しても、単語に区切り読みや品詞を与えていく形態素解析技術が適用できるようになってきた[6]。そこで、本プロジェクトにおける「近代語コーパス」においても、形態素解析データを含めたコーパスを設計することとした。一般に形態素解析は、文法が整った書き言葉には比較的適用しやすいが、そうでない話し言葉に適用するのはやや困難が伴う。まずは、書き言葉に適用し、その後話し言葉に適用させていく手順を取ることが現実的である。特に、近代語の口語を反映した資料は、地域や階層などによる言語の多様性が大きく、形態素解析技術の実現にはいくつかの研究段階が必要とされると見通されている。また、近代語の文語体書き言葉も、言文一致以前の文体は多様であり、やはり段階を踏んだ研究が求められる。そのような背景から、論説文という等質の文章でありながら、著者やジャンルは多様である『明六雑誌』は、形態素解析技術の適用事例として最初に取り組むのに好適であると考えられる。

なお、コーパス化の対象とするのは『明六雑誌』全 43 号の全文とする。ただし、(1)表紙、(2)目次、(3)識語・奥付、(4)図表中の文字列、は対象外とする。

３．文字入力の基本仕様

３．１．基本方針

本文テキストの入力はすべて全角文字で行う。

原文の書記体が漢字片仮名交じりの場合、外来語といった一部の語を除き、片仮名を平仮名に変換して入力する。原文の書記体の種類の情報は article タグの属性として表し、片仮名のままとした文字列には span タグを付与する。

３．２．文字集合

使用する文字集合は、JIS X 0213 のうち、⑴康熙別掲字、⑵UCS 互換字、⑶CJK 統合漢字拡張Ｂに符号位置が割り当てられる文字、を除外した範囲とする。この範囲にない文字は外字として「〓」で入力し、g タグを付与し、タグの属性として文字の情報を表す。なお、⑴⑵にあたる文字は通用字形に包摂し、⑶にあたる文字は外字とする。

３．３．包摂規準

包摂規準については、JIS X 0213 のものに準拠する。

ただし、『明六雑誌』では JIS の包摂規準の適用できない字形差を持つ文字が多数出現し、それらをすべて外字として「〓」で入力するとコーパスの実用性を損ねかねない。そ

[6] 近代語テキストへの形態素解析の現状と今後の見通しについては、本報告書に収録した小木曽智信「近代語テキストの形態素解析」を参照。

こで、本コーパスでは独自に JIS の包摂規準を拡張したものを定義し、それによって字体包摂を行い、JIS 内字を用いて入力し、g タグを付与してタグの属性として拡張包摂規準の適用を表す。

また、この拡張包摂規準を適用してもなお外字となる文字についても、類似の意味・用法を持つ文字が JIS 内にある場合は、なるべくその文字で入力し、g タグを付与してタグの属性として原文の文字の情報を表す。

拡張包摂規準や別字代用の詳細については、本報告書に収録した須永哲矢「近代語文献を電子化するための異体字処理」を参照のこと。

３．４．特殊な表記

ルビは、ルビの振られた文字列に ruby タグおよび lRuby タグを付与し、その rubyText 属性値によって表す。

割書された文字列は、warigaki タグを付与しその範囲を示す。

濁点の期待される文字に濁点が付いていない場合、濁点の付いた文字で入力し、vMark タグを付与する。

踊り字は、踊り字で繰り返される文字を入力し、odoriji タグを付与する。

漢字のよみを明らかにするために漢字の前後に小さく書かれた仮名や踊り字は、通常の入力とし、特にタグは付与しない。

JIS X 0213 外字の合字は、よみに対応する複数の仮名で入力し、特にタグは付与しない。

漢文に付与された振り仮名や返り点は入力対象外とする。

３．５．空白

紙面に現れる空白は、常に空白 1 文字で入力する。ただし、レイアウト上複数行に渡って行われる字下げについては、論理行冒頭のみに空白 1 文字を入力する。

天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白（敬意欠字）は、空白を入力し、g タグを付与して敬意欠字であることを表す。

３．６．誤植

誤植と思われる文字は、適切な文字に修正して入力し、corr タグを付与しタグの属性として原文の文字の情報を表す。

３．７．判読困難な箇所

印刷のかすれや破損・抹消によって、文字の形がまったく残っておらず判読ができない場合、入力は行わず、gap タグを付与して文字の存在を表す。

文字の形が一部残り、元の文字が推測可能な場合、その文字を入力し、unclear タグを付与する。

４．XML タグセット

本コーパスは、本文テキストに XML によって文書構造・単語・文字・表記に関する情報を付与する。そのための XML タグの一覧は、次の表 1 のとおりである。各タグで表される要素について続く各節で詳説する。なお、要素詳説であげる XML 例では、説明に不要なタグを省略して示す場合がある。

表 1　XML タグセット

タグ名	説明	詳説する節番号
magazine	雑誌 1 号分を表す。	４．１．
front	雑誌中で前付けに相当する文書要素を表す。	４．２．
body	雑誌中で中心本文に相当する文書要素を表す。	４．３．
article	1 記事を表す。	４．４．
titleBlock	前付け・中心本文の中にあり、記事とは認められない文書要素を表す。	４．５．
p	記事中の段落に相当する文書要素を表す。	４．６．
block	記事中にあり、段落とは見なせない文書要素（記事タイトル・記事著者・小見出し等）を表す。	４．７．
figureBlock	図表を表す。	４．８．
warigaki	割書された文字列を表す。	４．９．
quotation	発話部分や他の文献からの引用を表す。	４．１０．
superS	引用や割書を含むため、複数の s 要素に分割された文を表す。	４．１１．
s	文を表す。	４．１２．
odoriji	踊字で表記されていたことを表す。	４．１３．
span	漢字片仮名交じり文の片仮名を平仮名に変換したテキストを作成する際、片仮名のまま残した文字列を表す。	0
gap	抹消・破損等で判読できない文字列の存在を表す。	４．１５．
pb	原本での改ページ位置を表す。	４．１６．
lb	原本での改行位置を表す。	４．１７．
SUW	語（短単位）を表す。	４．１８．
ruby	本行の右側に振られたルビを表す。	４．１９．
lRuby	本行の左側に振られたルビを表す。	４．２０．
corr	誤植を校訂したことを表す。	４．２１．
unclear	不鮮明ではあるが字体が推定できる文字を表す。	４．２２．
vMark	濁点無表記の文字を表す。	４．２３．
g	外字・敬意欠字等の特殊な文字を表す。	４．２４．
kanbun	漢文によって書かれた文章に返読･補読を行ったことを表す。	４．２５．

４．１．magazine 要素

説明

雑誌 1 号分を表す。

属性

title（必須）：雑誌名

year（必須）：発行年

issue（必須）：号番号

version（必須）：XML ファイルのバージョン

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
（…中略…）
</front>
<body>
（…中略…）
</body>
</magazine>
```

４．２．front 要素

説明

雑誌中で前付けに相当する文書要素を表す。本コーパスでは雑誌タイトルがこれに該当する。

属性

なし

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
<titleBlock>
<block>
<s>明六社雑誌第一號</s>
</block>
</titleBlock>
</front>
<body>
（…中略…）
</body>
</magazine>
```

４．３．body 要素

説明

雑誌中で中心本文となる文書要素を表す。本コーパスでは複数の記事からなる部分がこれに該当する。

属性

なし

XML 例

```
<magazine title="明六雑誌" year="1874" issue="01" version="1.0">
<front>
（…中略…）
</front>
<body>
<article title="洋字を以て国語を書するの論" author="西周" style="文語" script="漢字カタカナ">
（…中略…）
</article>
<article title="開化の度に因て改文字を発すべきの論" author="西村茂樹" style="文語" script="漢字カタカナ">
</body>
（…中略…）
</article>
</magazine>
```

４．４．article 要素

説明

中心本文中の各記事を表す。

属性

title（必須）：記事の題名

author（必須）：記事の著者名・訳者名

originalAuthor（任意）：翻訳記事の原著者名

style（必須）：記事の文体を表す。

- 文語…文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。
- 口語…口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。
- 混在…文語体と口語体が混在するもの。

script（任意）：記事の書記体

- 漢字カタカナ…漢字片仮名交じり
- 漢字ひらがな…漢字平仮名交じり

XML 例

```
<article title="人民の自由と土地の気候と互に相関するの論（一）" author="箕作麟祥" originalAuthor="モンテスキュウ" style="文語" script="漢字カタカナ">
<block>
（…中略…）
</block>
<p>
（…中略…）
</p>
<p>
（…中略…）
</p>
</article>
```

４．５．titleBlock 要素

説明

前付けまたは中心本文の中の、記事と同位の文書要素で、記事とは認められないものを表す。本コーパスでは前付け中の雑誌タイトルがこれに該当する。

属性

なし

XML 例

```
<front>
<titleBlock>
<block>
<s>明六社雑誌第一號</s>
</block>
</titleBlock>
</front>
```

４．６．p 要素

説明

記事中の 1 段落を表す。

原則として、論理改行を段落末として段落を認定する。ただし、箇条書きのように論理改行による認定がふさわしくないと考えられる部分については、人手により段落の認定を行う。

属性

なし

XML 例

```
<p>
<s>西先生の改文字論を再三熟讀するに其論説痛快精到少しも遺憾なし</s><s>果して此言の如くなるㇵを得ば實に文運の大進歩にして吾儕操觚者の最も愉快とする處なり</s>（…中略…）<s>願くは諸先生の高論を以て左の件々を議定あらんㇵを</s>
</p>
<p>
<s>　第一　會社の名</s><s>　第二　社中人員の定數</s><s>　第三　新に入社する人員を撰ぶの法</s>（…中略…）<s>　第七　書記并びに掌計者を撰ぶ事</s><s>　第八　日誌出版の法</s>
</p>
<p>
<s>〇本朝にて學術文藝の會社を結びしは今日を始めとす</s><s>而して社中の諸賢は皆天下の名士なり</s>（…中略…）<s>何とぞ諸先生の卓識高論を以て愚蒙の眠を覺し天下の摸範を立て識者の望を曠ふせざらんㇵを是祈る</s>
</p>
```

４．７．block 要素

説明

記事中、段落と同位の要素で、段落とは認められないものを表す。本コーパスでは記事

タイトル・記事著者表示・記事小見出しがそれに該当する。

属性

なし

XML 例

```
<article title="洋字を以て国語を書するの論" author="西周" style="文語" script="漢字カタカナ">
<block>
  洋字を以て國語を書するの論
</block>
<block>
  西周
</block>
<p>
（…中略…）
</p>
</article>
```

４．８．figureBlock 要素

説明

図表のある部分を表す。空要素。

属性

なし

XML 例

```
<p>
（…中略…）
</p>
<figureBlock/>
<p>
（…中略…）
</p>
```

４．９．warigaki 要素

説明

割書となっている文字列を表す。

属性

なし

XML 例

```
その教と同派のものを信ずる某宗徒の爲に此徒を管轄する他國
<warigaki>
此國も亦耶蘇教を奉ず但し別派なり
</warigaki>
の事に與聞せんと要するは則その理あり
```

４．１０．quotation 要素

説明

記事・引用中で、その記事・引用とは発話者や発話場面・文体の異なる文書要素（他文献からの引用や会話・心話など）を表す。

属性

type（必須）：引用の種類

- 会話…会話
- 心話…心話
- 手紙…手紙
- 典拠…手紙を除く他文献からの引用
- 記事説明…記事に対し編集者等が説明を加えるための文書
- 韻文…漢詩を除く韻文
- 漢文…漢文によって書かれた文書。他の引用の種類に該当するものであっても、漢文の形式で書かれていれば type 属性値は「漢文」とする。

source（必須）：引用部分の話し手や書き手、引用元の書名等

style（任意）：引用の文体が上位 article 要素または quotation 要素の文体とは異なる場合の、文体の種類。ただし、type 属性値が「韻文」「漢文」の場合は不要。

- 文語…文語体。文末辞が「なり」「たり」「つ」「ぬ」「き」「けり」のもの。
- 口語…口語体。文末辞が「だ」「ぢや」「である」「です」「ます」のもの。
- 混在…文語体と口語体が混在するもの。

XML 例

例 1　会話

```
或人曾て英公使の書記官サトウ氏に語て曰く
<quotation type="会話" source="或人">
英學頗る日本に行はる
</quotation>
とサトウ氏頭を掉て
<quotation type="会話" source="英公使書記官サトウ">
否米學なり
</quotation>
と言へりとぞ
```

例 2　心話

```
蓋拷問の苦堪ふべからず常人は乃思へらく
<quotation type="心話" source="常人">
其拷問に苦しまんよりは寧冤罪に死せん
</quotation>
と
```

例 3　典拠

```
昨年十月の布告に新聞紙發行の條目中
<quotation type="典拠" source="新聞紙発行条目">
「國體を誹り國律を議し及び外法を主張宣議して國の妨害を生ぜしむるを禁ず」「政事法律等を記載するㇵに付妄に批評を加るㇵを禁ず」「猥りに教法を記入し政法の妨害を生ぜしむるを禁ずる」
</quotation>
等の箇條あり
```

例 4　記事説明

```
<quotation type="記事説明" source="箕作麟祥">
　右佛國大學士「モンテスキュウ」所著のスピリット、ヲフ、ロウスより抄譯す尚續譯して次号に出すべし
</quotation>
```

例 5　韻文

```
本居宣長の
<quotation type="韻文" source="本居宣長">
「式島の日本心を人問ば朝日に香ふ山櫻花」
</quotation>
と詠ぜしは即ち此易直の質を以て我が國民の氣風に烙記を居ゑたる者にて流石に夫れ者だけ能名状したる者と謂ふべし
```

例 6　漢文

```
邵康節の詩
<quotation type="漢文" source="邵康節">
尋常巷陌連羅綺、到處樓臺奏管絃、天下泰平無事日、鶯花無限日高眠、
</quotation>
といへるにても想見べし
```

４．１１．superS 要素

説明

割書や引用を含むために、複数の文からなると見なされる 1 文を表す。

本コーパスでは、形態素解析での必要性から、warigaki 要素はその前後とは必ず別の s 要素と認定する。また、複数の s 要素からなる quotatin 要素も同様にその前後とは別の s 要素と認定する。よって、該当要素を含む 1 文は、1 文であるにもかかわらず複数の s 要素に分割される。これらの複数の s 要素をまとめ上げるのが superS 要素である。

属性

なし

XML 例

例 1

```
<superS>
<s type=”fragment”>先生の御論にては内養</s>
<warigaki><s>先生論ずる所即政府官吏の理治</s></warigaki>
<s type=”fragment”>外刺</s>
<warigaki><s>即人民の政府を刺衝するㇵ</s></warigaki>
<s type=”fragment”>相平均せざる可らざる内にも外刺を以て殊に緊要と被致候様に相見へ候</s>
</superS>
```

例 2

```
<superS>
<s type=”fragment”>昨年十月の布告に新聞紙發行の條目中</s>
<quotation type="典拠" source="新聞紙発行条目"><s>「國體を誹り國律を議し及び外法を主張宣議して國の妨害を生ぜしむるを禁ず」</s><s>「政事法律等を記載する㆓に付妄に批評を加る㆒を禁ず」</s><s>「猥りに教法を記入し政法の妨害を生ぜしむるを禁ずる」</s></quotation>
<s type=”fragment”>等の箇條あり</s>
</superS>
```

４．１２．s 要素

説明

1 文を表す。文の認定は人手により行う。

属性

type（任意）：

- fragment…割書や引用を含むために、1 文であるにもかかわらず複数の s 要素に分割された結果生じた、文の一部を内容とする s 要素であることを表す

XML 例

例 1　通常の s 要素

```
<s>　其他語格の若きは後日の成功を待つべし</s>
<s>右聊か愚考を陳じ諸先生の可否を請ふ</s>
<s>敢て採用を望むにあらずと雖ども諸先生幸に電覽を賜はゞ幸甚</s>
```

例 2　type 属性値が「fragment」の s 要素

```
<superS>
<s type=”fragment”>その教と同派のものを信ずる某宗徒の爲に此徒を管轄する他國</s>
<warigaki>
<s>此國も亦耶蘇教を奉ず</s><s>但し別派なり</s>
</warigaki>
<s type=”fragment”>の事に與聞せんと要するは則その理あり</s>
</superS>
```

４．１３．odoriji 要素

説明

踊り字で表記されている箇所を表す。

踊り字が繰り返す文字列を odoriji 要素の内容とし、原文の踊り字は originalText 属性として入力する。ただし、短単位中で直前の 1 字を繰り返す「々」「ゝ」は odoriji 要素とはせず、テキストを「々」「ゝ」のままとする。

属性

originalText：原文で使われている踊り字

XML 例

例 1　一字点

```
是僕尤恐る<odoriji originalText="丶">る</踊字>所なり
```

例 2　二字点

```
天下ます<odoriji originalText="ヒ">ます</odoriji>昌明の運に進み
```

例 3　同字点

```
果て<odoriji originalText="々">果て</odoriji>は
```

例 3　くの字点

```
顯國代る<odoriji originalText="／＼">代る</odoriji>興り
```

例 4　odoriji 要素としない同字点

```
奇々怪々の一案
```

４．１４．span 要素

説明

本コーパスでは本文の書記体を漢字平仮名交じりに統一するため、原文では漢字片仮名交じりの文章は片仮名を平仮名に変換してテキストを作成する。しかし外来語といった片仮名表記のままのほうがよいと判断した文字列については、span 要素としてマークアップする。

属性

type（必須）：

- カタカナ…片仮名のまま残す文字列を表す

XML 例

例 1

```
<span type="カタカナ">アベセ</span>二十六字を知り
```

例 2

```
近日<span type="カタカナ">ヘボン</span>の字書又佛人<span type="カタカナ">ロニ</span>の日本語會あり
```

４．１５．gap 要素

説明

抹消・損傷等により判読不可能な文字列の存在を表す。空要素。

属性

quantity（任意）：該当文字列の文字数がわかる場合、その文字数

XML 例

```
然る時は又天地に彌り古今<gap quantity="2"/>き人の主として方向を定むべき者唯善のみ
```

４．１６．pb 要素

説明

原文の紙面上での改ページ位置を表す。空要素。

属性

n（必須）：該当位置から始まるページの番号。原本の丁付けで一丁表となるページから順に、開始の値を「1」とする連番を振る。

originalN（必須）：該当位置から始まるページの、原本での丁番号と表裏の区別。例えば一丁表ならば属性値は「1 オ」とする。

XML 例

```
<pb n="1" originalN=”1オ”/>明六社雑誌第一號（…中略…）故に上旨は下達せず下情は上伸せずして全身不遂
<pb n="2" original=”1ウ”/>の人の如し（…中略…）然ども此弊に因て斯世の民幸福を蒙るコを得ず衰弊の極救藥すべからざるに至
（…中略…）
<pb n="24" original=”12ウ”/> 第二 社中人員の定數（…中略…）何とぞ諸先生の卓識高論を以て愚蒙の眠を覺し天下の摸範を立て識者の望を曠ふせざらんコを是祈る
```

４．１７．lb 要素

説明

原文の紙面上での改行位置を表す。空要素。

属性

なし

XML 例

```
<pb n="1" originalN=”1オ”/> <lb/>明六社雑誌第一號
<lb/> 洋字を以て國語を書するの論 西周
<lb/>吾輩日常二三朋友の盍簪に於て偶當時治亂盛衰の故政治得失の跡な
<lb/>ど凡て世故に就て談論爰に及ぶ時は動もすればかの歐洲諸國と比較
<lb/>するコの多かる中に終には彼の文明を羨み我が不開化を歎じ果て果ては
```

４．１８．SUW 要素

説明

単語（短単位）を表す。

本コーパスの SUW 要素は、近代の文語文を対象とする形態素解析辞書「近代文語 UniDic」

による解析結果を人手で修正したものである。SUW 要素の各属性の詳細については、「近代文語 UniDic」（http://www2.ninjal.ac.jp/lrc/index.php?UniDic）のユーザーズマニュアルや、「近代文語 UniDic」のもととなった現代語版「UniDic」（http://download.unidic.org）のユーザーズマニュアルを参照のこと。

属性

orthToken（必須）：書字形出現形
lForm（任意）：語彙素読み
lemma（任意）：語彙素
subLemma（任意）：語彙素細分類。区別がある場合のみ出力。
pos（必須）：品詞
form（任意）：語形
cType（任意）：活用型。活用語のみ出力。
cForm（任意）：活用形。活用語のみ出力。
pronToken（任意）：発音形出現形
kanaToken（任意）：仮名形出現形
orth（任意）：書字形基本形。活用語のみ出力。
wType（任意）：語種
start（必須）：語の始まる文字位置
end（必須）：語の終わる文字位置
originalText（任意）：原文文字列。orthToken 属性値と異なる場合のみ出力。
orderID（必須）：語の通し番号
BOS（任意）：

- True…文頭に現れる語であることを表す

XML 例

```
<SUW orthToken="洋字" lForm="ヨウジ" lemma="洋字" pos="名詞-普通名詞-一般" form="ヨウジ" pronToken="ヨージ" kanaToken="ヨウジ" orth="洋字" wType="漢" start="100" end="120" orderID="80" section="v">洋字</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" kanaToken="ヲ" orth="を" wType="和" start="120" end="130" orderID="90" section="v">を</SUW>
<SUW orthToken="以" lForm="モツ" lemma="持つ" pos="動詞-一般" form="モツ" cType="文語四段-タ行" cForm="連用形-促音便" pronToken="モッ" kanaToken="モッ" orth="以つ" wType="和" start="130" end="140" orderID="100" section="v">以</SUW>
<SUW orthToken="て" lForm="テ" lemma="て" pos="助詞-接続助詞" form="テ" pronToken="テ" kanaToken="テ" orth="て" wType="和" start="140" end="150" orderID="110" section="v">て</SUW>
<SUW orthToken="國語" lForm="コクゴ" lemma="国語" pos="名詞-普通名詞-一般" form="コクゴ" pronToken="コクゴ" kanaToken="コクゴ" orth="國語" wType="漢" start="150" end="170" orderID="120" section="v">國語</SUW>
<SUW orthToken="を" lForm="ヲ" lemma="を" pos="助詞-格助詞" form="ヲ" pronToken="オ" kanaToken="ヲ" orth="を" wType="和" start="170" end="180" orderID="130" section="v">を</SUW>
<SUW orthToken="書する" lForm="ショスル" lemma="書する" pos="動詞-一般" form="ショス" cType="文語サ行変格" cForm="連体形-一般" pronToken="ショスル" kanaToken="ショスル" orth="書す" wType="混" start="180" end="210" orderID="140" section="v">書する</SUW>
<SUW orthToken="の" lForm="ノ" lemma="の" pos="助詞-格助詞" form="ノ" pronToken="ノ" kanaToken="ノ" orth="の" wType="和" start="210" end="220" orderID="150" section="v">の</SUW>
<SUW orthToken="論" lForm="ロン" lemma="論" pos="名詞-普通名詞-一般" form="ロン" pronToken="ロン" kanaToken="ロン" orth="論" wType="漢" start="220" end="230" orderID="160" section="v">論</SUW>
```

４．１９．ruby 要素

説明

原本本行の文字列の右側に振られているルビを表す。

属性

rubyText（必須）：ルビとして振られた文字列

rubyBase（任意）：複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合、先頭の短単位を ruby 要素とする処理を行い、rubyBase 属性に実際にルビの振られている文字列を値として入力する。

XML 例

例 1

```
<ruby rubyText="サフロフ">候</ruby>文
```

例 2

```
<r rt="ケミストリ">化學</r>
```

例 3　複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合

```
<ruby rubyText="コンタラソシヤール" rubyBase="國民約束">國民</r>約束
```

４．２０．lRuby 要素

説明

原本本行の文字列の左側に振られているルビを表す。

属性

rubyText（必須）：ルビとして振られている文字列

rubyBase（任意）：複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合、先頭の短単位を ruby 要素とする処理を行い、rubyBase 属性に実際にルビの振られている文字列を値として入力する。

XML 例

例 1

```
<lRuby rubyText="ボストン">波士頓</r>
```

例 2　複数の SUW 要素により構成される文字列に 1 つのルビが振られている場合

```
<lRuby rubyText="ヂウアイン" rubyBase="上帝道">上帝</lRuby>道
```

４．２１．corr 要素

説明

原文で誤植と見られる文字を表す。

該当文字が本行にある場合、適切に修正した文字を corr 要素の内容とし、originalText 属性に原文文字を値として入力する。type 属性値が「excess」の場合は空要素。

該当文字がルビにある場合、該当ルビを表す ruby 要素・lRuby 要素の rubyText 属性値に適切に修正した文字を入力し、corr 要素の originalText 属性値でルビ全体の原文文字列を表す。

属性

type（必須）：誤植の種類

- erratum…誤字
- excess…衍字
- omission…脱字

originalText（任意）：該当文字が本行にある場合、本行の原文文字を表す。ただし、type 属性値が「omission」の場合は不要。また、該当文字がルビにある場合、ルビ全体の原文文字列を表す。

subType（任意）：

- ruby…該当文字がルビにあることを表す。

XML 例

例 1　誤字（本行）

```
<corr originalText="已" type="erratum">已</corr>に
```

例 2　衍字（本行）

```
盡く其實を<corr originalText="ヲ" type=”excess"/>明示し
```

例 3　脱字（本行）

```
改めざる可か<corr type="omission">ら</注>ず
```

例 4　衍字（ルビ）

```
<corr originalText="クヽ" type="excess" subType=”ruby”><ruby rubyText="ク">喰</ruby></corr>つて見た上で
```

例 5　脱字（ルビ）

```
<corr originalText="そ" type="omission" subType=”ruby”><ruby rubyText="ほそ">細</ruby></corr><ruby rubyText="ね">根</ruby>
```

４．２２．unclear 要素

説明

原本の損傷等により不鮮明ではあるが字体の推定は可能な文字を表す。

属性

originalText（任意）：該当文字がルビにある場合は、該当文字を「□」で表記してルビ全体の文字列を表す。該当文字が本行にある場合は不要。

type（任意）：

- ruby…該当文字がルビにあることを表す。

XML 例

例 1　本行にある場合

```
公法大學士ヒリモア<unclear>及</unclear><unclear>び</unclear>ワツテル兩家
```

例 2　ルビにある場合

```
<unclear originalText="□イロソフイカル" type="ruby"><ruby rubyText="フイロソフイカル">哲理</ruby></unclear>
```

４．２３．vMark 要素

説明

濁点の表記が期待されるにもかかわらず、原文では濁点無表記の仮名・記号が使われていることを表す。

本行に該当文字がある場合、濁点を表記した文字を vMark 要素の内容とする。

ルビに該当文字がある場合、そのルビを表す ruby 要素・lRuby 要素の rubyText 属性値には濁点を表記した文字を入力し、vMark 要素の originalText 属性値でルビ全体の原文文字列を表す。

属性

originalText（任意）：ルビに該当文字がある場合は、ルビの原文文字列を表す。本行に該当文字がある場合は不要。

type（任意）：

- ruby…ルビに該当文字があることを表す。

XML 例

例 1　本行の場合

```
談論爰に及<vMark>ぶ</vMark>時は
```

例 2　ルビの場合

```
侈其可否は<vMark originalText="トウダ" type="ruby"><ruby rubyText="ドウダ">如何</ruby></vMark>と云った時は
```

４．２４．g 要素

説明

JIS X 0213 外字や敬意欠字といった特殊な文字・記号を表す。

JIS X 0213 外字の漢字であるが拡張包摂規準により字体包摂を行う場合、包摂後の字体を入力して g 要素の内容とする。

JIS X 0213 外字で、かつ拡張包摂規準の適用外の漢字であるが、意味・用法の類似する他の漢字での代用が可能な場合、その代用字を入力して g 要素の内容とする。

字体包摂も代用字での代用もしない JIS X 0213 外字の場合、「〓」を入力して g 要素の内容とする。

天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白（敬意欠字）の場合は、空白を入力して g 要素の内容とする。

属性

type（必須）：

- 外字…JIS X 0213 外字で、かつ拡張包摂規準の適用外の文字であることを表す。
- 包摂…JIS X 0213 外字であるが、拡張包摂規準により JIS X 0213 内字に包摂した文字であることを表す。
- 敬意欠字…天皇等の高貴な人に敬意を表すために、その人に関連する語の直前に表記された空白であることを表す。

ref（任意）：type 属性値が「外字」の場合、Unicode4.0 の 16 進コードがあるものは「U+」を先頭に加えた文字列を値とし、Unicode 外字の場合は字体記述を値とする。type 属性値が「包摂」「敬意欠字」の場合は不要。

XML 例

例 1　外字（Unicode 内字）

```
<g type=”外字” ref="U+9AF2">𩦋</g>
```

例 2　外字（Unicode 外字）

```
<g type=”外字” ref="衣＋丸">𩦋</g>
```

例 3　外字（代用字による入力）

```
<g type="外字" ref="U+7FA1">羨</g>
```

例 4　包摂

```
<g type”包摂”>時</g>
```

例 5　敬意欠字

```
我大日本<g =type=”敬意欠字”>　</g>天皇陛下の特詔を垂れて
```

４．２５．kanbun 要素

説明

漢文によって日本語を書き表したと見なされる文字列を訓読するにために、返読・補読を行ったことを表す。

ただし、漢籍の引用など日本語を書き表したと見なされない漢文には返読・補読は行わず、quotation 要素とする。

type 属性が「返読前」のものは空要素。

属性

type（必須）：

- 返読前…返読の対象となる文字が、本来あった位置。
- 返読後…返読の対象となる文字が、返読後に移動した位置。
- 補読…訓読によって補読された語。

originalText（任意）：原文の文字。type 属性値が「返読前」のものは必須。type 属性値が「返読後」「補読」のものは不要。

id（任意）：返読の対象となる文字の、返読前の位置と返読後の位置を対照するために与えられた XML ファイル内固有の ID。type 属性値が「返読前」「返読後」のものは必須。type 属性値が「補読」のものは不要。

XML 例

例 1　返読

```
此段宜敷御評議を<kanbun type="返読前" originalText="可" id="00001"/><kanbun type="返読前" originalText="被" id="00002"/>遂<kanbun type="返読後" id="00002">被</kanbun><kanbun type="返読後" id="00001">可</kanbun>候<也
```

例 2　補読

```
然ども是<kanbun type="返読前" originalText="不" id="00008"/><kanbun type="返読前" originalText="得" id="00009"/>已<kanbun type="補読">を</kanbun><kanbun type="返読後" id="00009">得</kanbun><kanbun type="返読後" id="00008">不</kanbun>の時なり
```

５．コーパスの公開形式

本コーパスの公開形式は以下の 3 種類である。

５．１．XML ファイル

本文テキストに XML タグによって文書構造・形態論・文字・表記に関する情報を付与した形式で、コーパスの根幹となるデータである。

1 号 1 ファイルとし、全 43 ファイルからなる。XML ファイルの符号化形式は UTF-8（BOM なし）である。ファイル名は「m」に続く 4 桁の数字が該当号の刊行年を、次の 2 桁の数字が号番号を表す。例えばファイル名が「m187401.xml」ならば、1874 年刊行の 1 号のデータを収めた XML ファイルということになる。

５．２．形態論情報タブ区切りデータ

XML ファイルから特に SUW 要素に関する情報を抽出し、タブ区切りのデータに成形したものである。ファイル名は「merioku_suw.txt」、符号化形式は UTF-8（BOM なし）である。1 行目はフィールド名を入力した行で、2 行目以降から 1 行が 1SUW 要素に対応している。

データのフィールドリストを表 2 として示す。

表 2　フィールドリスト

フィールド名	備考
コーパス名	
ファイル名	XML ファイル名に対応
記事題名	article 要素 title 属性に対応
記事著者	article 要素 author 属性に対応
記事原著者	article 要素 originalAuthor 属性に対応
記事文体	article 要素 style 属性に対応
記事書記体	article 要素 script 属性に対応
語連番	SUW 要素 orderID 属性に対応
文字開始位置	SUW 要素 start 属性に対応
文字終了位置	SUW 要素 end 属性に対応
文頭ラベル	SUW 要素 BOS 属性に対応（B：文頭、I：文頭以外）

語彙表 ID	書字形出現形レベルで語を識別する ID
語彙素 ID	語彙素レベルで語を識別する ID
語彙素読み	SUW 要素 lForm 属性に対応
語彙素	SUW 要素 lemma 属性に対応
語彙素細分類	SUW 要素 subLemma 属性に対応
語種	SUW 要素 wType 属性に対応
品詞	SUW 要素 pos 属性に対応
活用型	SUW 要素 cType 属性に対応
活用形	SUW 要素 cForm 属性に対応
語形	SUW 要素 form 属性に対応
書字形基本形	SUW 要素 orth 属性に対応
書字形出現形	SUW 要素 orthToken 属性に対応
原文文字列	SUW 要素 originalText 属性に対応
発音形出現形	SUW 要素 pronToken 属性に対応

５．３．「ひまわり」用データ

文字列検索システム「ひまわり」用のデータである。このデータを「ひまわり」にインストールすることで、GUI による簡便なコーパスの検索が可能となる。

５．３．１．「ひまわり」へのインストール方法

データの「ひまわり」へのインストールは次の①～③の手順で行う。

① ダウンロードしたデータを解凍すると、「meiroku_himawari」フォルダが現れる。その中に次のファイルがあることを確認する。

- Corpora フォルダ…『明六雑誌コーパス』データを格納したフォルダ
- config_meiroku.xml…設定ファイル

② 「ひまわり」ver.1.3 をインストールする。国立国語研究所「言語データベースとソフトウェア」のホームページ（http://www2.ninjal.ac.jp/lrc/index.php）の画面左にある「メニュー」―「ソフトウェア」―「全文検索システム『ひまわり』」をクリックすると、「ひまわり」のページに移動する。そこに書かれた説明に従い「ひまわり」ver.1.3 のインストールを行う。

③ 「ひまわり」ver.1.3 をインストールすると「Himawari_1_3」フォルダが現れる。その中に、①の「Corpora」フォルダと「config_meiroku.xml」を移動する。その際、コンピュータ環境によっては「Corpora」フォルダの上書きの確認のメッセージが表示される場合があるが、そのまま上書きを許可してよい。

５．３．２．「ひまわり」を使ったコーパスの検索方法

「ひまわり」にインストールしたコーパスデータの基本的な検索方法を説明する。

「Himawari_1_3」フォルダ内の「himawari.exe」（アイコン  ）をダブルクリックすると「ひまわり」の起動画面が開く（図 1）。

図 1　「ひまわり」の起動画面

次に、画面上部の「ファイル」メニュー―「新規」を選択する（図 2）。すると、設定ファイルを指定するための画面が現れるので、「config_meiroku.xml」を選択する。

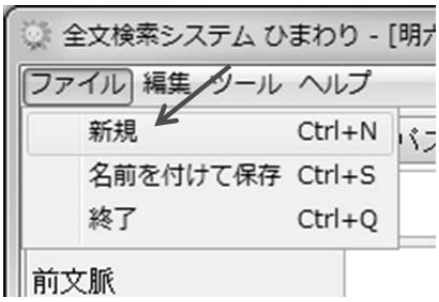


図 2　「ファイル」メニュー―「新規」

次に「プルダウンメニュー」（図 1 参照）で検索対象を指定する。検索対象のリストを表 3 としてあげる。なお、プルダウンメニューに表示される「完全一致」「部分一致」は検索対象と検索文字列との照合方法を表す。

表 3　「ひまわり」検索対象リスト

プルダウンメニュー表示	検索対象
本文	本文テキスト部分

右ルビ/完全一致	ruby 要素 rubyText 属性値
右ルビ/部分一致	
左ルビ/完全一致	lRuby 要素 rubyText 属性値
左ルビ/部分一致	
語彙素/完全一致	SUW 要素 lemma 属性値
語彙素読み/完全一致	SUW 要素 lForm 属性値
語種/完全一致	SUW 要素 wType 属性値
品詞/部分一致	SUW 要素 pos 属性値
活用型/部分一致	SUW 要素 cType 属性値
活用形/部分一致	SUW 要素 cForm 属性値
語形/完全一致	SUW 要素 form 属性値
書字形基本形/部分一致	SUW 要素 orth 属性値

次に「検索文字列」欄（図 1 参照）に検索したい文字列を入力する。「字体変換」ボタン（図 1 参照）をクリックすると、入力文字列に異体字がある場合は異体字を含めた検索ができるように「検索文字列」欄の入力が変換される。そして「検索」ボタン（図 1 参照）をクリックすると「検索結果」欄（図 1 参照）に検索結果が KWIC 形式で表示される（図 3）。

図 3　「ひまわり」での検索結果表示

「検索結果」欄に表示される列のリストを表 4 として示す。

表 4　「ひまわり」検索結果列リスト

列名	備考
前文脈	
キー	
後文脈	
語彙素読み	SUW 要素 lForm 属性に対応
語彙素	SUW 要素 lemma 属性に対応
語彙素細分類	SUW 要素 subLemma 属性に対応
語種	SUW 要素 wType 属性に対応
品詞	SUW 要素 pos 属性に対応
活用型	SUW 要素 cType 属性に対応
活用形	SUW 要素 cForm 属性に対応
語形	SUW 要素 form 属性に対応
書字形基本形	SUW 要素 orth 属性に対応
雑誌名	magazine 要素 title 属性に対応
年	magazine 要素 year 属性に対応
号	magazine 要素 issue 属性に対応
ページ	pb 要素 originalN 属性に対応
語連番	SUW 要素 orderID に対応
記事題名	article 要素 title 属性に対応
記事著者	article 要素 author 属性に対応
記事著者	article 要素 originalAuthor 属性に対応
記事文体	article 要素 style 属性に対応
引用種類	quotation 要素 type 属性に対応
引用ソース	quotation 要素 source 属性に対応
引用文体	quotation 要素 style 属性に対応

よりひろい文脈で検索結果を閲覧したい場合は、「検索結果」欄のセルをダブルクリックする。Web ブラウザが起動し、雑誌単位での閲覧ができる。閲覧表示スタイルは次の 3 種類がある。

- 本文＋付加情報（図 4）
- 本文（図 5）
- 形態論情報リスト（図 6）

閲覧表示スタイルの切り替えは「ひまわり」起動画面の「ツール」メニュー―「オプション」―「閲覧表示スタイル」から行うことができる。

図 4　「本文＋付加情報」スタイルでの文脈表示

図 5　「本文」スタイルでの文脈表示

語連番	ページ	文頭	書字形出現形	語彙素読み	語彙素	語彙素細分類	語種	品詞	活用型	活用形	語形	発音形出現形	引用種類	引用ソース	引用文体
10	1オ	B	明六	メイロク	明六		固	名詞-固有名詞-一般			メイロク	メーロク			
20	1オ	I	雑誌	ザッシ	雑誌		漢	名詞-普通名詞-一般			ザッシ	ザッシ			
30	1オ	I	第	ダイ	第		漢	接頭辞			ダイ	ダイ			
40	1オ	I	二	ニ	二		漢	名詞-数詞			ニ	ニ			
50	1オ	I	號	ゴウ	号		漢	名詞-普通名詞-助数詞可能			ゴウ	ゴー			
60	1オ	B	福澤	フクザワ	フクザワ		固	名詞-固有名詞-人名-姓			フクザワ	フクザワ			
70	1オ	I	先生	センセイ	先生		漢	名詞-普通名詞-一般			センセイ	センセー			
80	1オ	I	の	ノ	の		和	助詞-格助詞			ノ	ノ			
90	1オ	I	學者	ガクシャ	学者		漢	名詞-普通名詞-一般			ガクシャ	ガクシャ			
100	1オ	I	職分	ショクブン	職分		漢	名詞-普通名詞-一般			ショクブン	ショクブン			
110	1オ	I	論	ロン	論		漢	名詞-普通名詞-一般			ロン	ロン			
120	1オ	I	は	ハ	は		和	助詞-係助詞			ハ	ワ			
130	1オ	I	慶應	ケイオウ	慶応		固	名詞-固有名詞-一般			ケイオウ	ケーオー			
140	1オ	I	義塾	ギジュク	義塾		漢	名詞-普通名詞-一			ギジュク	ギジュク			

図 6　「形態論情報リスト」スタイルでの文脈表示

「ひまわり」の利用方法の詳細については、「ひまわり」の利用者マニュアル（「ひまわり」起動画面の「ヘルプ」メニュー―「『ひまわり』マニュアル」）を参照のこと。

文　献

神奈川大学人文学研究所（2004）『「明六雑誌」とその周辺―西洋文化の受容・思想と言語―』（御茶の水書房）

高野繁男・日向敏彦（1998）『明六雑誌語彙索引―付・復刻版「明六雑誌」』（大空社）

高野繁男（2004）『近代漢語の研究―日本語の造語法・訳語法―』（明治書院）

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」（国立国語研究所『雑誌「太陽」による確立期現代語の研究―「太陽コーパス」研究論文集―』博文館新社）

山室信一・中野目徹（1999-2009）『明六雑誌(上)(中)(下)』（岩波文庫）

『明六雑誌コーパス』の語彙量

近藤　明日子　（国立国語研究所コーパス開発センター）[1]

１．本稿の目的

本稿は、『明六雑誌コーパス』の語彙量の概要について報告するものである。

２．凡例

２．１．報告の対象

この報告では、『明六雑誌コーパス』の XML ファイルの SUW 要素（詳細は本報告書に収録した近藤明日子・田中牧郎「『明六雑誌コーパス』の仕様」を参照）1 つを 1 語として語彙量を集計する。ただし、SUW 要素のうち、形態論情報の付与の対象外としたものについては報告の対象外とする。対象外としたものは次のものである。

(1) 英語等の外国語の原語表記（SUW 要素 pos 属性値が「英単語」）
(2) 日本語のローマ字表記（SUW 要素 pos 属性値が「ローマ字文」）
(3) 漢文（SUW 要素 pos 属性値が「漢文」）
(4) 前後の文字が判読できないため形態論情報が付けられないもの（SUW 要素 pos 属性値が「読取不可」）

２．２．同語異語判別

異なり語数をカウントする際の同語異語判別には、『明六雑誌コーパス』の形態論情報付与の基盤となった、近代文語文を対象とする形態素解析辞書「近代文語 UniDic」の語彙素レベルを用いる。語彙素レベルとは辞書の見出し語に相当するもので、語形の揺れや書字形の違いを吸収し同語として扱うものである。

[1] kondo@ninjal.ac.jp

３．語彙量の報告

３．１．品詞別語彙量

品詞別に延べ語数・異なり語数を示す（表 1）。

品詞の分類は SUW 要素の pos 属性値の大分類に拠る。

表 1　品詞別語彙量

	延べ語数	異なり語数
名詞	58,428	10,823
代名詞	4,020	42
動詞	28,433	1,224
形容詞	2,298	126
形状詞	1,507	365
副詞	5,790	239
連体詞	4,626	17
接続詞	2,344	28
感動詞	52	8
接頭辞	1,062	45
接尾辞	2,070	198
助詞	52,199	62
助動詞	15,720	29
記号	43	19
補助記号	1,534	13
空白	479	1
合計	180,605	13,239

３．２．著者別語彙量

著者別に延べ語数・異なり語数を示す（表 2)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）を除いて集計する。

著者の分類はコーパスの XML の article 要素 author 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

表 2　著者別語彙量

	延べ語数	異なり語数
西周	35,424	4,549
阪谷素	31,934	4,428
津田真道	26,187	3,887
西村茂樹	15,402	1,964
中村正直	12,121	2,310
杉亨二	11,502	2,187
森有礼	9,990	1,857
神田孝平	9,306	1,717
加藤弘之	7,236	1,111
箕作麟祥	5,611	1,073
福沢諭吉	4,623	1,008
柏原孝章	3,637	827
清水卯三郎	1,597	498
柴田昌吉	1,339	421
津田仙	1,068	419
箕作秋坪	1,068	341
合計	178,045	

３．３．文体別語彙量

記事の文体別に延べ語数・異なり語数を示す（表 3)。また、延べ語数における文体比率を示す（図 1)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）を除いて集計する。

文体の分類はコーパスの XML の article 要素 style 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

表 3　文体別語彙量

	延べ語数	異なり語数
文語	167,832	12,642
口語	8,394	1,690
混在	1,819	651
合計	178,045	

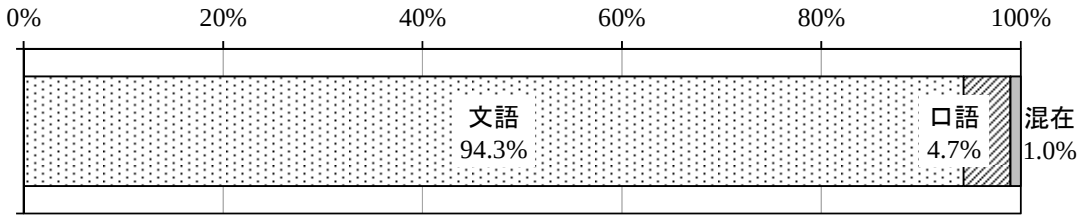


図 1　文体比率（延べ語数）

３．４．語種別語彙量

語種別に延べ語数・異なり語数を示す（表 4)。また、延べ語数および異なり語数での和語・漢語・外来語・混種語の比率を示す（図 2)。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）と助詞・助動詞を除いて集計する。

語種の分類は SUW 要素の wType 属性に拠る。wType 属性値の意味は次のとおりである。

和…和語
漢…漢語
外…外来語
混…混種語
固…固有名（品詞が「名詞-固有名詞」のもの）
記号…記号

表 4　語種別語彙量

	延べ語数	異なり語数
和	59,779	2,287
漢	43,965	9,504
外	559	250
混	3,685	378
固	2,638	682
記号	4	2
合計	110,630	13,103

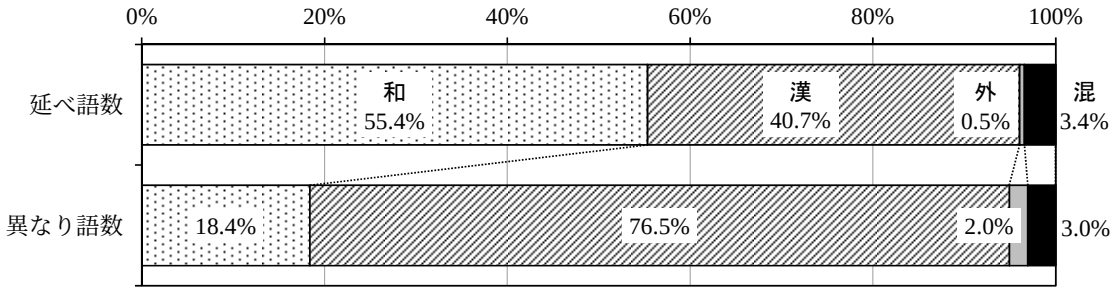


図 2　語種比率

３．５．文体・語種別語彙量

文語記事・口語記事ごとに語種別の延べ語数・異なり語数を示す（表 5・表 6）。また、文語記事・口語記事ごとに異なり語数での和語・漢語・外来語・混種語の比率を示す（図 3）。

延べ語数・異なり語数とも、記号類（品詞が「記号」「補助記号」「空白」の語）と助詞・助動詞を除いて集計する。

文体の分類はコーパスの XML の article 要素 style 属性に拠る。よって、article 要素に含まれない各号の雑誌タイトル部分は集計対象外となる。

語種の分類は SUW 要素の wType 属性に拠る。

表 5　語種別語彙量（文語）

	延べ語数	異なり語数
和	56,513	2,054
漢	41,288	9,269
外	497	227
混	3,514	360
固	2,406	652
記号	4	2
合計	104,222	12,564

表 6　語種別語彙量（口語）

	延べ語数	異なり語数
和	2,644	684
漢	1,884	778
外	60	36
混	138	53
固	109	69
記号	0	0
合計	4,835	1,620

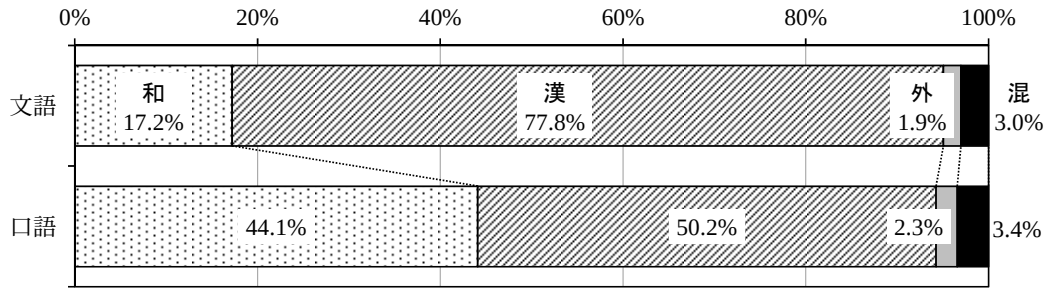


図 3　文体別語種比率（異なり語数）

第２部　コーパスの活用

明治後期から大正期の語彙のレベルと語種
―『太陽コーパス』の形態素解析データによる―

田中　牧郎　（国立国語研究所言語資源研究系）[1]

１．はじめに

明治期に多くの漢語が日本語に取り入れられたが、大正・昭和・平成と進むにつれて漢語は減少していくことが明らかにされている（国立国語研究所 1964、国立国語研究所 1987、国立国語研究所 2005a）。一方、基本語彙の中に占める漢語の比率は、次第に増加していくという報告もある（飛田 1966、宮島 1967）。また、明治前期から徐々に取り入られた外来語は、大正期から増加傾向が顕著になり、第二次大戦中にいったん減少した後、昭和時代の終わりまで増加の一途をたどる（橋本 2010）。このように、近代から現代にかけての語彙の歴史は、語種の観点からみたとき、大きな変化があることが明らかにされている。従来の研究は、漢語あるいは外来語という、ある語種に光をあててその歴史的特徴を明らかにしてきたが、和語をも含めた語種全体を見わたした歴史については、十分に解明できていないところがある。和語を含めて全体的な視点をもつことによって、漢語や外来語の歴史についても、新たな視点からその特徴を見直していくことができるのではないかと思われる。

しかし、そのような研究を行う前提となる、語彙全体の実態把握を行うことは容易でなかった。現代語については、国立国語研究所による語彙調査のデータはあったが、ある時代は雑誌、別の時代は教科書、さらに別の時代にはテレビ放送といったように特定の媒体の調査であり、日本語全体として語彙がどのように変容したのかについて考えるほどにはデータが示されてはいなかった[2]。まして、近代語については、ごく一部の資料にしか調査データが存在しておらず（国立国語研究所 1959、1985-1997 など）、その語彙の変容を記述することは難しかった。

ところが、本プロジェクトなどで、明治以後の近代日本語のコーパス構築と関連技術の整備に着手したことにより、語彙全体を射程に入れた近代語彙史の体系的な記述を行える状況に近づきつつある。本稿では、公開済みの『太陽コーパス』に対して、本プロジェクトなどで整備中の形態素解析辞書「近代文語 UniDic」を用いて形態素解析を施すことで、明治後期から大正期の語彙の体系的な変化を語種の視点からとらえる研究例を示したい。

２．『太陽コーパス』への「近代文語 UniDic」の適用

２．１　『太陽コーパス』

『太陽コーパス』（国立国語研究所 2005a）は、博文館から刊行された総合雑誌『太陽』（1985～1928 年）を対象としたコーパスである。1895（明治 28）年、1901（明治 34）年、1909（明治 42）年、1917（大正 6）年、1925（大正 14）年の５年分の全文（著作権処理ができなかった記事を除く）を対象にしている（田中 2005）。この『太陽』は、分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さなどの点で、当時の文献資料としては格別の価値を持っていることから、何か一つの資料で当時期の書き言葉を代表させるとしたら、おそらく筆頭に挙げてよい資料の一つである。

---

[1] mtanaka@ninjal.ac.jp

[2] 国立国語研究所（1987）は、雑誌という一媒体に限られるが、時間軸による語彙の変化をとらえようとしていて点で特徴的である。調査データの量は少ないが、現代語彙を通時的に扱った最初の研究として価値が高い。

２．２　「近代文語 UniDic」

分かち書きがされない日本語は単語認定の複雑さが大量のデータに対する語彙調査の制約になっていたが、近年、国立国語研究所が中心に開発を進めている形態素解析辞書「UniDic」は、従来の人手による語彙調査で実績のある言語単位に基づく斉一な単位での解析を可能にしたことと、階層構造を持たせることで品質管理や同語異語判別等の便宜を向上させたことの２点が特筆される（伝ほか 2007）。この UniDic をもとに近代語資料に対する形態素解析を可能にしたものに「近代文語 UniDic」があり（小木曽 2009 など）、本報告書におさめられている小木曽論文、須永・近藤論文[3]にも記載のある通り、実用化に見通しが立ちつつある。

２．３　『太陽コーパス』に対する「近代文語 UniDic」による自動形態素解析

この「近代文語 UniDic」によって、『太陽コーパス』に自動形態素解析を施す研究の現状は、口語体の部分では誤解析が少なくないこと、文語体の部分でも語や表記によっては誤解析が生じる場合が残されているなど、自動形態素解析の結果をそのまま無条件に利用できる段階には至っていない。

しかしながら、高精度の解析が実現されるまで待たないと、『太陽コーパス』の形態論情報を利用した研究は行えないと考えるよりも、データの完成度が低い段階でも、データに誤りが含まれる可能性には十分留意しつつも、形態論情報を使うことで可能になる新たな研究領域を開拓していくべきだと考えることの方が、建設的だろう。そこで、本稿では、『太陽コーパス』に対して、「近代文語 UniDic」による自動形態素解析を施し、その結果を用いた研究を試みることにする。具体的には、解析結果データをもとに、年次別の語彙頻度表を作成し、語彙頻度によって語彙をレベルに分け、そのレベルを指標として、語種の観点からみた語彙の変化の実態を把握することを試みる。

３．『太陽コーパス』の語種比率

「近代文語 UniDic」は、直接的には文語文を対象とするものであるが、ここでは、口語文も含めた『太陽コーパス』の全体を対象とした。文語文に比べて口語文は解析精度が悪くなるものの、決定的に劣るというわけではなく、口語文においても大部分は正しく解析できる。口語文を対象から外すと、新しい年次（1909 年以後）の分量が、かなり少なくなってしまい、経年的比較が難しくなってしまう。それよりも、多少精度が低くても全体を扱って、そのようなデータでも活用が可能な研究を展開するのがよいと考えた。

『太陽コーパス』全体に対して、「近代文語 UniDic Ver.1.2」（MeCab 版）を用いて自動形態素解析を実施した。UniDic が規定する品詞体系のうち、記号・付属語・未知語は除外した。また、UniDic の語種情報は、和語・漢語・外来語・混種語・固有名詞・記号の六種に分かれるが、このうち「記号」はアルファベット略語の類が分類されており、これは「外来語」にまとめた。その五種類の語種の年次別の語数について、延べ語数、異なり語数を集計したものが、表１・表２である。表１・表２をもとに、語種比率を見るためにグラフ化したものが、それぞれ図１・図２である。

図１・図２を一見すると、『太陽コーパス』において、語種別の比率は年次によって大きな変動はないように見える。語種から見た語彙のありようは、明治後期から大正期にかけて、大きな変化はなかったというように見ることもできそうである。しかし、よく見ていくと、わずかずつではあるが一定の方向での変化も見られ、それはこの時期の語彙の歴史として重要な側面を浮かび上がらせているのではないかと考えられる。

---

[3] 小木曽智信「近代語テキストの形態素解析」、須永哲矢・近藤明日子「近代語コーパスのための形態論情報付与規程の整備」（いずれも本報告書所収）。

表１　『太陽コーパス』の年次別・語種別語数（延べ語数）

語種	1895 年	1901 年	1909 年	1917 年	1925 年	全体
和語	639896	574523	518387	484725	452507	2670038
漢語	566709	530110	453738	421473	353699	2325729
外来語	5499	6545	4755	4452	7027	28278
混種語	32553	30537	24295	21839	16937	126161
固有名詞	66852	47491	43237	40191	36902	234673

表２　『太陽コーパス』の年次別・語種別語数（異なり語数）

	1895 年	1901 年	1909 年	1917 年	1925 年	全体
和語	11543	10026	9781	9818	10761	17878
漢語	26456	23947	20526	19485	18883	35023
外来語	1128	1133	1095	947	1321	2886
混種語	1297	1152	1092	1009	1167	2177
固有名詞	9349	6791	5889	5128	6089	16125
計	49773	43049	38383	36387	38221	74089

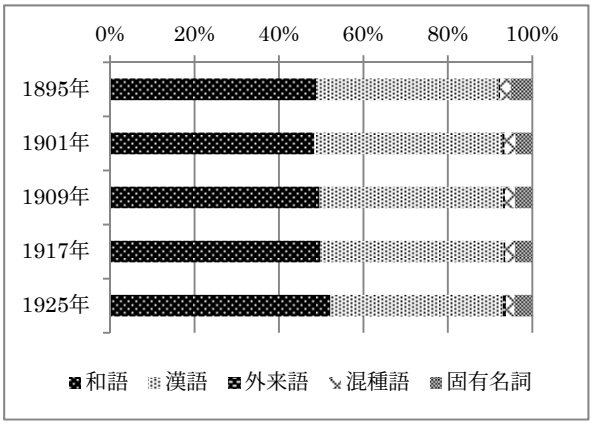


図 1『太陽コーパス』年次別語種比率（延べ語数）

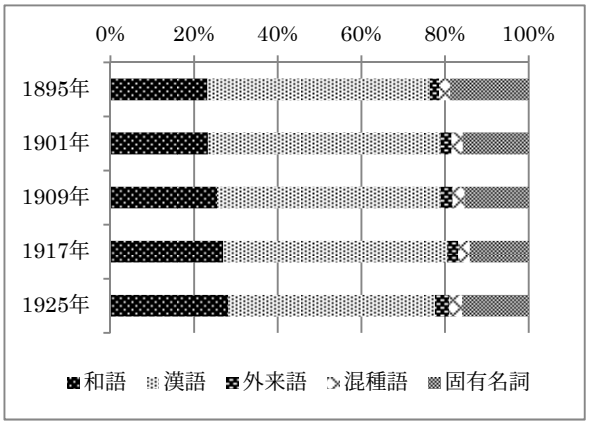


図 2『太陽コーパス』年次別語種比率（異なり語数）

まず、図１で延べ語数における語種比率を見ると、和語の比率は各年次 50%前後であるものの、よく見ると、その比率が年次を追って少しずつ増加していることに気づく。一方、漢語を見ると、各年次 40 数%であるが、わずかずつ減少していることが分かる。外来語・混種語・固有名詞は、いずれも非常に少なく、年次による変化もとらえにくい。つまり、延べ語数では、語種構成に大きな変化はないが、年次を追って少しずつ、和語が増加しその分漢語が減少していっている様子が見て取れるのである。

次に、図２で異なり語数における語種比率を見ると、やはり和語の増加と漢語の減少を確かに見て取ることができ、その増減の幅は延べ語数の場合よりもやや大きいことが分かる。また、外来語が 1925 年で比率を高めていることもとらえることができる。

このように、『太陽コーパス』における語種構成には、年次による大きな変化はないものの、和語の増加とその反面である漢語の減少が、確かな変化として認められ、大正後期に

は外来語の増加も見え始めるのである。

４.『太陽コーパス』の語彙のレベル分け

４．１　語彙のレベルの考え方

国立国語研究所（1964）は、語彙を使用頻度によって階級に分けた場合に、階級によって語種構成比に違いがあることを明らかにしている。昭和中期（1956 年）に刊行された雑誌 90 種の語彙調査で得られたデータをもとに、使用頻度別に七つの階級に分け、各階級の語種構成比率を示す、次のグラフを掲載している。

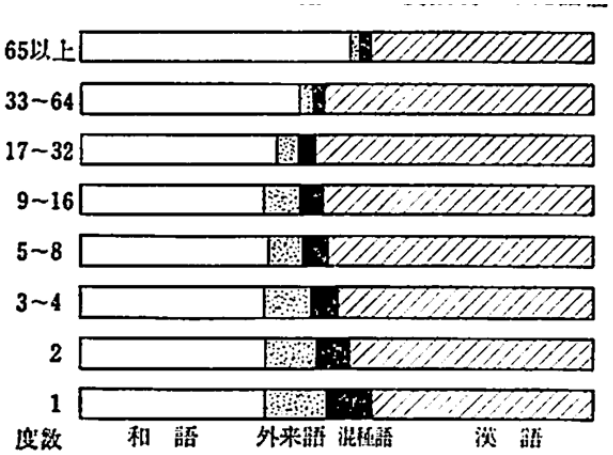


図３　雑誌90種調査における度数階級別の語種構成比（国立国語研究所1964から転載）

図３によると、最も使用頻度の高い階級では和語の比率が高く、漢語は和語よりも比率が低く、外来語・混種語の比率は極めて低い。使用頻度の低い階級に進んでいくほど、和語の比率は徐々に低くなるが、４番目の階級から後は、比率が不変となる。これに対して、漢語は、最も使用頻度の高い階級では和語よりも低い比率であるが、２番目の階級では比率を大きく高めて和語を上回り、３番目の階級で最も高くなる。その先は使用頻度の低い階級に進むほど、比率を低下させていく。そして、外来語と混種語は、使用頻度の低い階級に進むにつれて、その比率を高めていき、度数１の階級に至るまでその傾向が続いていく。以上のことから、和語は高頻度の階級で多く、漢語は中頻度の階級で多く、外来語と混種語は低頻度の階級で多いという、傾向があると言うことができよう。

このような使用頻度によって語彙を階級に分ける考え方は、語彙の特徴を研究する際に有効だと考えられ、とりわけ、語種の枠組みを用いて記述しようとする場合は、たいへんに役に立つのではないかと思われる。そこで、本稿では、使用頻度によって分ける語彙の階級を「レベル」と呼び、その概念を明確化し、この枠組みを用いて、明治後期から大正期の語彙のありようを、語種の視点で記述していきたい。

４．２　カバー率によるレベル分け

4.1 で引いた、国立国語研究所（1964）の語彙頻度に基づくレベルは、ある１時点におけるひとつの調査によるものであった。ところが、『太陽コーパス』が対象とする５つの年次それぞれで語彙レベルに分けて相互に比較しようとすると、この方法では問題が生じる。なぜなら、各年次で延べ語数が異なるために、レベルを区画する使用頻度をどこに設定するかという、線引きの困難さに直面するからである。

この問題に対応するために、使用度数の高いものから順に各語の度数を累積していき、その累積度数が延べ語数の何パーセントを占めるかという、カバー率（累積度数占有率）を算出し、この数値をもとに一定の基準を定めてレベル分けを行うことにした。五つの年

次を比較するには、あまりレベルの区画が多いと繁雑になると考え、５段階に分けることとした。その具体的なカバー率の基準と、各年次にこれを適用した頻度区間を示したのが、表３である。

表３　『太陽コーパス』年次別の語彙のレベルと頻度区間

レベル	カバー率	1895 年	1901 年	1909 年	1917 年	1925 年	全体
a	-78%	-46	-51	-54	-53	-40	202-
b	78-88%	45-17	50-19	53-19	52-19	14-39	66-201
c	88-94%	16-7	18-8	18-8	7-18	13-6	24-65
d	94-97%	6-4	7-4	7-4	6-4	5-3	11-23
e	97-100%	3-1	3-1	3-1	3-1	2-1	1-10

＊各年・全体の数字は使用頻度の区間を示す。

最も高頻度のレベルａの基準を、カバー率 78%までと定めたところ、1895 年では使用頻度 46 以上の語まででこの基準に達し、1901 年では 51 以上のところでこの基準に達した。次のレベルｂは 78%から 88%と定めると、1895 年では使用頻度 17 以上 45 以下の語がここに配された。このようにして、各年次のすべての語をいずれかのレベルに配属させた。その結果の語数をまとめたものが、表４である。

表４　『太陽コーパス』年次別・レベル別の語数

レベル	1895 年	1901 年	1909 年	1917 年	1925 年	全体
a	3915	3156	2680	2531	2928	3476
b	4760	3839	3364	3112	3889	4835
c	8083	5995	5441	5652	5869	8334
d	7314	7389	6102	5063	7544	10177
e	25701	22720	20796	20029	17991	47267
計	49773	43049	38383	36387	38221	74089

＊各年・全体の数字は配分された語数を示す。

４．３　レベル別の語種比率

次に、各年次別にレベルごとの語種構成比率を算出し、グラフに表示してみよう（図 4〜8)。どの年次も、和語はレベルａで最も高く、漢語はレベルｂ・ｃ・ｄといった中間的なレベルで高く、外来語はレベルｅで最も高いという特徴を共通してもっている。これは、4.1 で見た昭和中期の特徴と同じであり、明治後期以来、この特徴は変わらなかったと考えられる。

それでは、レベル別の語種構成比率に、時代による変化は認められないのだろうか。そのことを調べるために、各年次ごとに、レベル別の和語比率、漢語比率を算出し、その経年変化が分かるように折れ線グラフを作成した（図 9〜10)。その際、固有名詞は除外し、和語・漢語・外来語・混種語の四つの合計の中で、和語・漢語それぞれが占める比率を算出した。

図９で和語比率の変化を見ると、どのレベルでも上昇していくが、当初低かったレベルｂ・ｃ・ｄでの上昇の度合いがやや大きい。また、図 10 の漢語比率を見ると、和語の場合とは反対に低下していくが、当初高かったレベルｂ・ｃ・ｄにおいて低下の度合いがや

や目立っている。このように、中間的なレベルの語彙ほど、漢語から和語へという移行がより目立っていると見ることができよう。

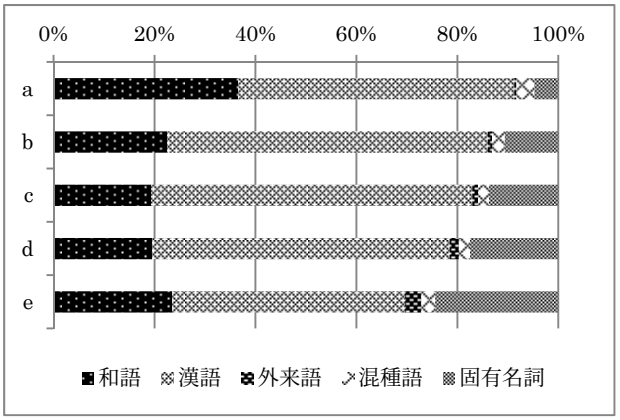


図 4　1895 年のレベル別語種比率

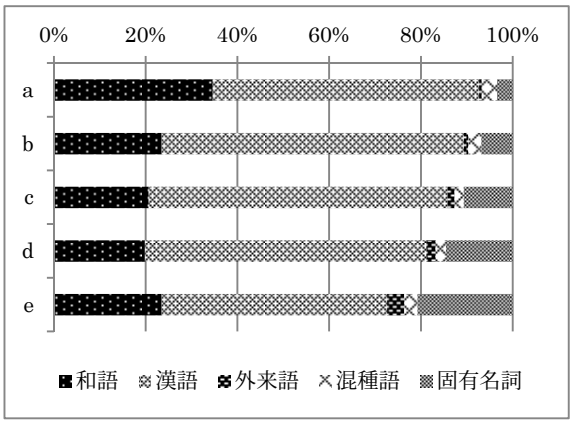


図 5　1901 年のレベル別語種比率

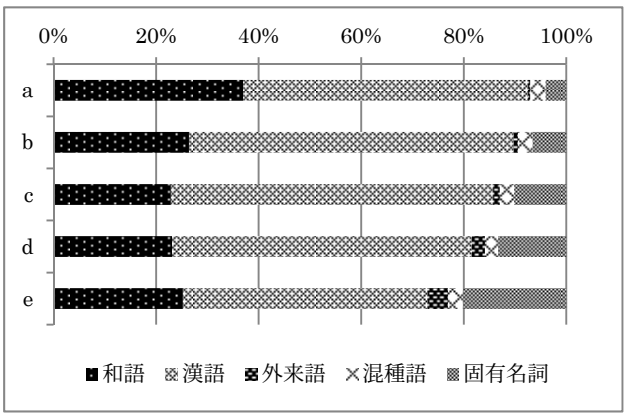


図 6　1909 年のレベル別語種比率

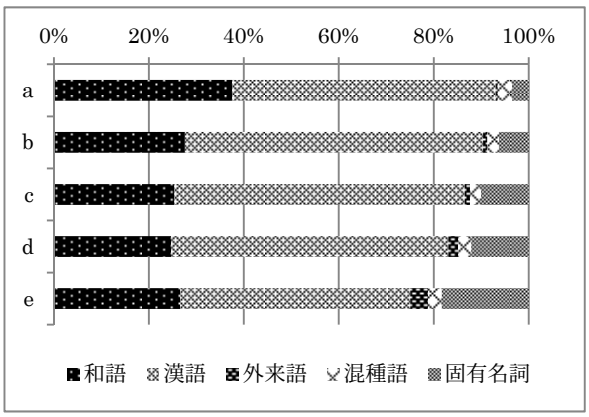


図 7　1917 年のレベル別語種比率

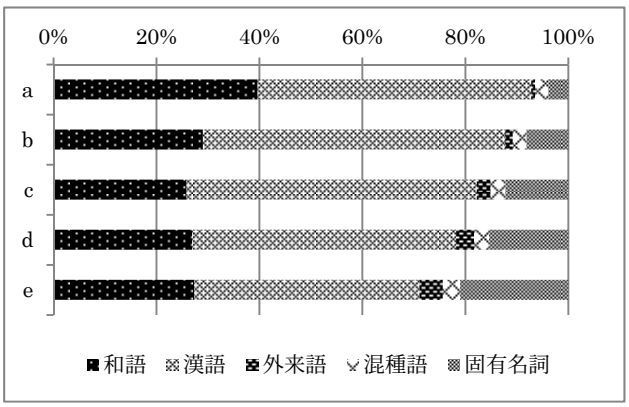


図 8　1925 年のレベル別語種比率

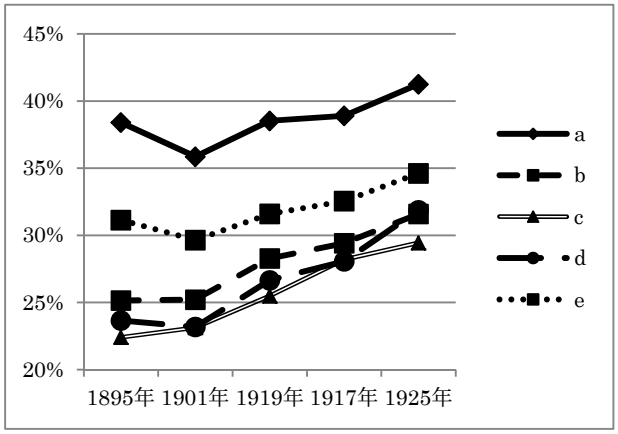


図 9　レベル別の和語率の推移

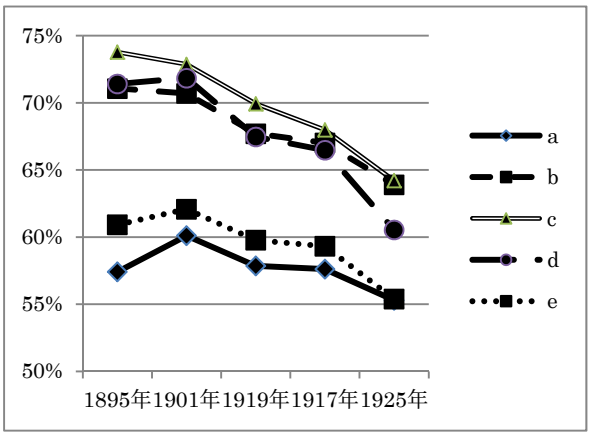


図 10　レベル別の漢語率の推移

５．レベルの変動による類型化

５．１　類型化の基準

次に、個々の語のレベルが、『太陽コーパス』が対象とする５つの年次でどのように変わる（変わらない）のかという点から、語彙の類型化を考えてみよう。年次を通じて、レベルが不変な語彙もあれば、一定の方向にレベルが変化していく語彙もある。この、レベルの不変と変化の観点から、特徴的な語彙を抽出して類型化することを考えると、表５のような五つの類型にまとめるのが分かりやすいと思う。

表５　レベルの不変と変化による類型化

類型	定義	基準例	語数
類型Ⅰ	基本的なレベルで不変	全年次を通じてレベルａまたはｂ	3523（4.8%）
類型Ⅱ	中間的なレベルで不変	全年次を通じてレベルｂまたはｃまたはｄ*	4328（5.8%）
類型Ⅲ	周辺的なレベルで不変	全年次を通じてレベルｅまたは「なし」	37522（50.6%）
類型Ⅳ	基本的なレベルに変化	図 11 の濃い網掛け部分	231（0.3%）
類型Ⅴ	周辺的なレベルに変化	図 12 の濃い網掛け部分	795（1.1%）
類型外	特定の傾向なし	上記以外	27690（37.4%）
全　体			74089（100%）

*すべての年次でレベルｂのものは、ここに入れずに類型Ⅰとする。

類型Ⅰは、基本的なレベルで安定している語彙である。最も基本的なレベルのａであり続けるものだけを基準にとれば 1650 語になるが、ここでは少し基準を緩くして、ａまたはｂのいずれかのレベルにおさまっているものという基準を立てたところ 3523 語となった。次に、類型Ⅱとして、中間的なレベルで変わらない語彙を、全年次でレベルｂからｄの間に入っているものという基準を立てたところ 4328 語となった（全年次レベルｂのものは、ここに入れずに類型Ⅰに入れる）。そして、類型Ⅲとして周辺的なレベルで不変な語彙を、全年次でレベルｅまたは「なし」のいずれかのものという基準で算出したところ、37522 語となった。「なし」とは、当該の年次には使用例がないもののことである。以上の三つの類型は、ある一定のレベルの範囲で変化しないものであったが、一定の方向のレベルに移行していくものもある。類型Ⅳは、より基本的なレベルへと変化するもので、図 11 の網掛け部分におさまるという基準を立てると、231 語となる。また、類型Ⅴは、図

12 の網掛け部分におさまるという基準で、795 語となった。

	1895 年	1901 年	1909 年	1917 年	1925 年
a				■	■
b			■	■	■
c		■	■	■	
d	■	■	■		
e	■	■			
なし	■				

図 11　類型Ⅳ（基本レベル化）とするもの

	1895 年	1901 年	1909 年	1917 年	1925 年
a	■	■			
b	■	■	■		
c		■	■	■	
d			■	■	■
e				■	■
なし					■

図 12　類型Ⅴ（周辺レベル化）とするもの

前段落で示した基準によって五つの類型に分類すると、そのいずれにもあてはまらない類型外のものが 27690 語残る。表５の基準は絶対的なものではなく、もっと緩い基準にすれば、類型外だった語をいずれかの類型に入れることもできる。ただ、あまり特徴のはっきりしなものも類型にまとめてしまうという問題が生じるため、ここでは、表５の基準で分類した。表５を見ると、類型Ⅲ（周辺的なレベルで不変）の語彙が最も多く、他の類型を圧倒している。多くの語は、周辺的なレベルにあってその位置を変えないということが分かる。類型Ⅲに比べればかなり少なくなるが、次に多いのが、類型Ⅱ（中間的なレベルで不変)、そして、類型Ⅰ（基本的なレベルで不変）である。類型Ⅳ（基本的なレベルに移行）または類型Ⅴ（周辺的なレベルに移行）の、レベルが一定の方向に変化していくものは、少数派である。

５．２　各類型の語種構成

こうして設定した類型ごとに語種構成比率を調査すると図 13 のようになる。この図から次のようなことが読み取れる。

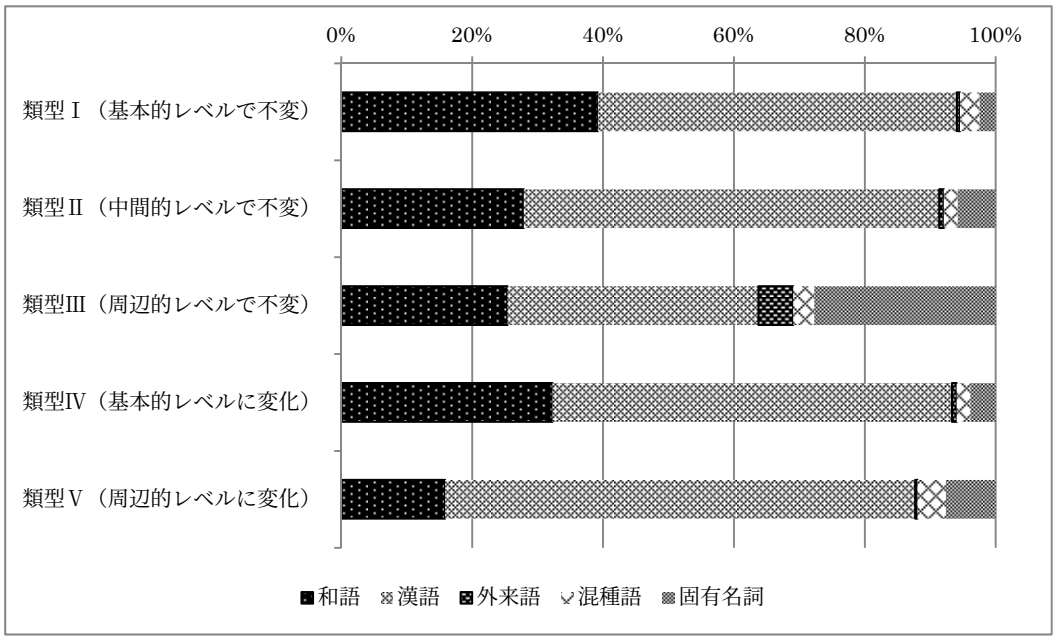


図 13　類型別の語種比率

まず、どの類型においても漢語の比率が最も高く、漢語は類型の違いにかかわらずもっとも多い語種であることが確認できる。類型別に見ていくと、まず、類型Ⅰ（基本レベルで不変）では、他の類型に比べて和語の比率の多さが目立ち、反対に、類型Ⅱ（中間レベルで不変）では、漢語の多さが目立つ。このことは、さきに４節で見た、基本的なレベルには和語が多く、中間的なレベルには漢語が多いという特徴が、通時的に見た場合も変わらずに維持されていると見ることができる。次に、類型Ⅲ（周辺レベルで不変）を見ると、他の類型と異なり、固有名詞や外来語が多くなっているが、和語と漢語の対比では、他のどの類型よりも漢語の比率が小さくなっていることに気づき、周辺的なレベルのまま変わらない語彙には、漢語は多くないと見ることができよう。そして、類型Ⅳ（基本レベル化）では、類型Ⅰ～Ⅲのいずれよりも漢語が多くなっており、基本的なレベルに進出していく語彙には、漢語がたくさんあったことが分かる。さらに、類型Ⅴ（周辺レベル化）においては、他のどの類型よりも漢語の占める割合が高く７割近くをも占めており、次第に周辺的なレベルに追いやられていく語彙には、漢語が非常に多かったことが判明するのである。

以上のように、類型化によって見えてくる語彙の特徴は、和語と漢語に顕著にうかがえ、とりわけ漢語の動きが注目される。次節では、和語と漢語それぞれについて少し詳しく見ていこう。

６．レベルから見た和語の特徴

６．１　品詞構成

図 14 は、類型ごとに和語における品詞構成をまとめたものである。「近代文語 UniDic」では、精細な品詞情報が出力されるが、ここでは、下に示す６種に統合して集計した。

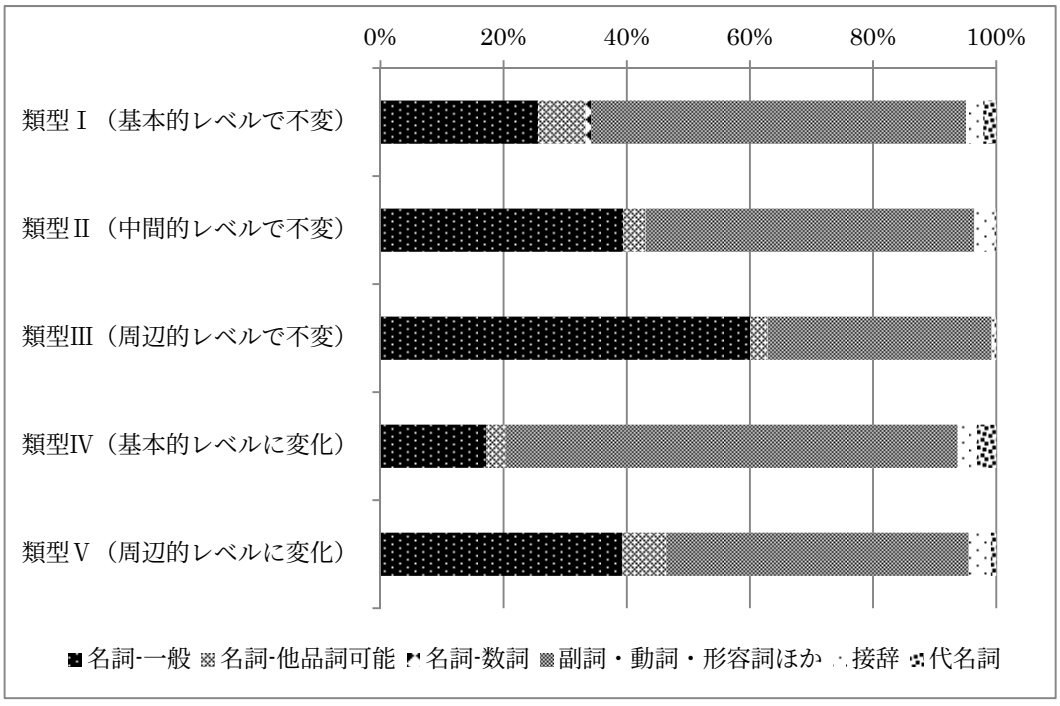


図 14　類型別の和語の品詞比率

名詞-一般：「名詞-普通名詞-一般」と出力されるもの
名詞-他品詞可能：「名詞-普通名詞-サ変可能」「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」「名詞-普通名詞-助数詞可能」などと出力されるもの
名詞-数詞：「名詞-数詞」と出力されるもの
副詞・動詞・形容詞ほか：「副詞」「形状詞-タリ」「形状詞-助動詞語幹」「接続詞」「動

詞-一般」「動詞-非自立可能」「形状詞-一般」「感動詞」「代名詞」などと出力されるもの
接辞：「接頭辞」「接尾辞-名詞的-一般」「接尾辞-名詞的-サ変可能」などと出力されるもの
代名詞：「代名詞」と出力されるもの

図 14 を見ると、類型による品詞構成が大きく異なっていることが分かる。類型Ⅰでは「副詞・動詞・形容詞・形状詞ほか」が 60%程度という多数を占め、ついで「名詞-一般」が 25%程度となっている。これが、類型Ⅱ、類型Ⅲへと進むにしたがって、「副詞・動詞・形容詞ほか」が減少していく一方で、「名詞-一般」が増加していき、類型Ⅲでは、これが逆転している。

６．２　和語における類型別の特徴

各類型にはどのような語彙が入っているのかを具体的に見ていきたい。その際、類似の語彙において類型間の比較を行うのが、各類型の特徴を見えやすくすると考えられるので、意味分野を同じくする一群の語彙を取り上げてみたい。国立国語研究所（2004）『分類語彙表増補改訂版』が示す分類項目のうち、「抽象的関係」の「様相」について、「体の類」「用の類」「相の類」のそれぞれの品詞に分類される語彙（具体的には、「1.13」「2.13」「3.13」の中項目番号の語彙）を事例に取り上げる。この類の語彙のうち、和語の各品詞について、類型別に一覧にすると、下の通りである。『分離語彙表増補改訂版』は多義語について語義ごとに別の意味番号を与えているが、「様相」の分類に複数箇所掲げられている場合は一つのみを掲げた。また、「様相」の分類以外にもその語が掲げられている場合は、より基本的と思われる語義の語と扱った。なお、レベル分けに用いた頻度は。多義語の場合、全語義を合わせたものになっており、「様相」に分類される語義でない例も頻度に含まれてしまっているという問題がある[4]。

○名詞など（「体の類」に相当）
類型Ⅰ（基本的レベルで不変）
　様（さま）、ありさま、傾き、趣、味、息、姿、身、素（もと）、根、常、差し支え
類型Ⅱ（中間的レベルで不変）
　成り行き、出来（でき）、目付き、仕掛け、仕組み、構え、国柄、傷、一通り、似合い、締まり、妨げ
類型Ⅲ（周辺的レベルで不変）
　死にざま、風向き、気配、あく抜け、上向き、口当たり、手触り、肌触り、めりはり、言葉付き、口前、見栄え、上っ面、見せ掛け、見てくれ、男前、手振り、体付き、足付き、目色、なり、身なり、見目、山並み、ばね仕掛け、成り立ち、骨組み、間取り、店構え、身構え、質（たち）、木口、肌合い、持ち前、得手、並並、持ち合い、兼ね合い、狂い、乱れ、ほつれ、人込み、もつれ、緩み、がら空き、がら明き、差し障り、当たり障り、がた、見事、あしざま
類型Ⅳ（基本的レベルに変化）
　該当なし
類型Ⅴ（周辺的レベルに変化）
　（該当なし）

[4] この点は大きな問題であるが、コーパスの用例を語義別に分類することは、現在の形態素解析技術では対応できず、そこに人手を関係づけて修正する手法についても確立されていない。今後の重要な研究課題である。

○動詞など（用の類に相当）
類型Ⅰ（基本的レベルで不変）
味わう、帯びる、装う、こしらえる、乱れる、散る、緩む、締める、絞める、張る、詰まる、込む、払う、そろう、備わる、備える、尽くす、妨げる、損なう
類型Ⅱ（中間的レベルで不変）
成り立つ、組む、組み立てる、似合う、向く、外す、乱す、かき回す、散らす、締まる、絞まる、詰める、そろえる、行き届く、障る、損ねる、傷付く、傷付ける、汚（よご）す、汚（よご）れる、汚（けが）す、壊す、廃れる、荒れる、荒らす、すさぶ、すさむ、
類型Ⅲ（周辺的レベルで不変）
身構える、そぐう、持ち合う、並外れる、狂わす、踏み外す、振り乱す、散らかる、入り組む、込み合う、立て込む、こんがらかる、こんぐらかる、こじれる、たるむ、引き締まる、締め付ける、出そろう、打ちそろう、取りそろえる、繰り回す、踏み荒らす、荒（あら）らげる、荒立つ、荒立てる、追い詰める、追い込む、切羽詰まる
類型Ⅳ（基本的レベルに変化）
作り上げる、片付ける
類型Ⅴ（周辺的レベルに変化）
（該当なし）

○形容詞・形状詞・副詞など（相の類に相当）
類型Ⅰ（基本的レベルで不変）
珍しい、まれ、殊に、おかしい、よい、よろしい、よく、悪い、うまい、甘い、寂しい、濃い、薄い、いろいろ、美しい、清い、難しい、難い、穏やか、危うい、
類型Ⅱ（中間的レベルで不変）
まずい、見事、あっぱれ、目覚ましい、うらやましい、分けて、ゆゆしい、好ましい、程よい、ふさわしい、滑らか、緩い、まばら、細か、様様、華やか、醜い、汚い、かろうじて、平たい、危ない、厳しい
類型Ⅲ（周辺的レベルで不変）
めちゃくちゃ、むちゃくちゃ、よしなに、とりどり、色よい、あつらえ向き、似つかわしい、なまなか、すっぽり、ごたごた、耳障り、まちまち、しどけない、ぎしぎし、がたがた、きつい、ばらばら、粗い、ゆったり、だらだら、きちきち、ややこしい、しげく、すっきり、手短、はでやか、はではでしい、汚らしい、易しい、やすやす、すんなり、まどか、危なっかしい、やばい
類型Ⅳ（基本的レベルに変化）
すばらしい、あたり前、ちゃんと、ゆっくり、いろんな
類型Ⅴ（周辺的レベルに変化）
（該当なし）

上記のリストから、語数の多い類型Ⅰ、類型Ⅱ、類型Ⅲの和語の特徴として、品詞を問わずに指摘できそうなことは次の通りである。まず、類型Ⅰは単純語がほとんどを占め、類型Ⅱは単純語と合成語の両方が同じ程度あり、類型Ⅲは合成語が多くを占めているという、語構成上のはっきりした差異を見て取ることができる。このことにも関連するが、類型Ⅰは、意味の抽象度が高く、幅広い意味で用いられる語が多いが、類型Ⅱ、類型Ⅲへと進むにしたがって、具体的で限定された意味を表す語が多くなる。類型Ⅰ～Ⅲは、レベルを変えない語彙であるが、ここに述べた特徴は、基本レベルにある語彙、周辺レベルにある語彙、その中間にある語彙の一般的性格と言ってよいものだと考えられる。
一方、類型Ⅳと類型Ⅴは、時代によってレベルが一定方向に変わっていく語彙であるが、語数が少ないため、＜様相＞の語群だけからでは、その特徴ははっきりと分からない。そこで、これらについては、その類型に属するすべての和語を一覧にしてみよう。誤解析が

原因でこの類型に入ってきたことが明らかな語は除外した。なお、『分類語彙表増補改訂版』では「その他」に分類される接続詞や感動詞の類や接辞は、ここでは便宜的に「相の類」に分類した。

まず、以下は、類型Ⅳの基本レベル化していく和語のすべてである。

○名詞など（「体の類」に相当）
あそこ、あちら、いたずら、苦しみ、小春、差し引き、背中、立場、手合い、飛び、羽目、振り替え、みちのり、無駄、読み物
○動詞など（「用の類」に相当）
煽る、言い替える、疑る、打ち込む、生み出す、教わる、片付ける、勝ち得る、腰掛ける、喋る、逸れる、携わる、突き込む、作り上げる、付け加える、取り去る、取り締まる、吐き出す、引っ張る、見付かる、持ち込む
○形容詞・形状詞・副詞など（「相の類」に相当）
当たり前、色んな、うん、うんと、黄色い、確り、じっと、直ぐ、すっかり、素晴らしい、そこら、そっと、たらしい、小さな、ちゃんと、ちょい、っこ、はっきり、ふむ、ほっ、まあ、真っ黒、惨め、むやみ、もう、尤も、ゆっくり

上記の一覧で、まず目に付くのは、「相の類」の中の、「うん」「ふむ」「ほっ」「まあ」などの感動詞、「うんと」「すっかり」「じっと」「はっきり」「ゆっくり」など、擬態語に由来する副詞である。これらは、話し言葉によく使われる語である。「黄色い」「素晴らしい」「真っ黒」などの形容詞・形状詞、「煽る」「疑る」「片付ける」「喋る」「逸れる」「吐き出す」「引っ張る」などといった動詞、「あそこ」「あちら」「いたずら」「羽目」「無駄」等の名詞や代名詞も、話し言葉的な語である。基本レベル化していく類型Ⅳの和語のリストからは、話し言葉に特徴的な用語が、書き言葉に進出していく流れがあったことが見て取れる。

次に、類型Ⅴの周辺レベル化していく和語を一覧にしよう。

○名詞など（体の類に相当）
錨、命、飢え、受け渡し、討ち死に、媼（おうな）、大麦、伯母、下ろし、蚕、水夫（かこ）、守（かみ）、鯨、厨（くりや）、褻（け）、薦（こも）、今宵、逆様、桟敷、柴、酢、簾、黄昏、盥（たらい）、費え、一日（ついたち）、鼓（つづみ）、科（とが）、仲立ち、法（のり）、僻（ひが）、広（ひろ）、誉（ほまれ）、麻呂、帝（みかど）、御代（みよ）、姪、設け、催し、諸（もろ）、諸共、諸人、匁（もんめ）、山羊、寡（やもめ）、行く末、忽（ゆるが）せ、葦（よし）、装い、読み売り、我が家、弁え
○動詞など（用の類に相当）
勇む、諫める、入り来たる、失せる、おわします、おわす、帰り来る、肯（がえ）んずる、くずれる、寿ぐ、差しのぼる、授かる、諭す、統べる、謗る、立ち出でる、轟く、煮る、宣う、阻む、侍る、払い込む、秀でる、ひしぐ、紐解く、臥せる、屠る、参らす、ましま す、まる、報う、愛でる、めとる、申し付ける
○形容詞・形状詞・副詞など（相の類に相当）
いずくんぞ、最も、憂（う）い、うたた、おのが、思しい、買い、くちおしい、げに、さら、ただただ、力無い、つらい、中々、生臭い、干（ひ）、偏に、一入（ひとしお）、吹き、間々（まま）、見、睦まじい、易（やす）い、ゆかしい、よし無い、

このリストを見ていくと、その多くが、古典にはよく出てくる語でありながら現代の書き言葉としては、古風な語感のする語であることが分かる。名詞では「討ち死に」「嫗」「守」「褻」「今宵」「黄昏」「誉れ」など、動詞では「いさめる」「おわします」「おわす」「寿ぐ」「秀でる」「愛でる」など、形容詞・形状詞・副詞では「いずくんぞ」「憂い」「うたた」「く

ちおしい」「げに」「ゆかしい」「よし無い」などである。これらは、もともとは話し言葉でも使われていたと思われるが、明治期までに話し言葉では使われなくなり、書き言葉として継承されてきて、明治期の文語体書き言葉には受け継がれたが、口語体書き言葉が成立するとそこでは使われなくなっていったと考えられる。書き言葉に残存した「古語」が、口語体書き言葉の成立とともに姿を消していく例である。

しかし、上記のリストの中には、「錨」「大麦」「蚕」「鯨」「酢」「煮る」などがあり、これらは古語ではない。こうした具体物や具体的行為を意味する語は、雑誌で取り上げられる話題が変わってきたことを反映したものだと考えられ、言語の変化によるものではないだろう。

７．レベルから見た漢語の特徴

７．１　品詞構成

図 15 は、類型ごとの漢語について、品詞構成をまとめたものである。

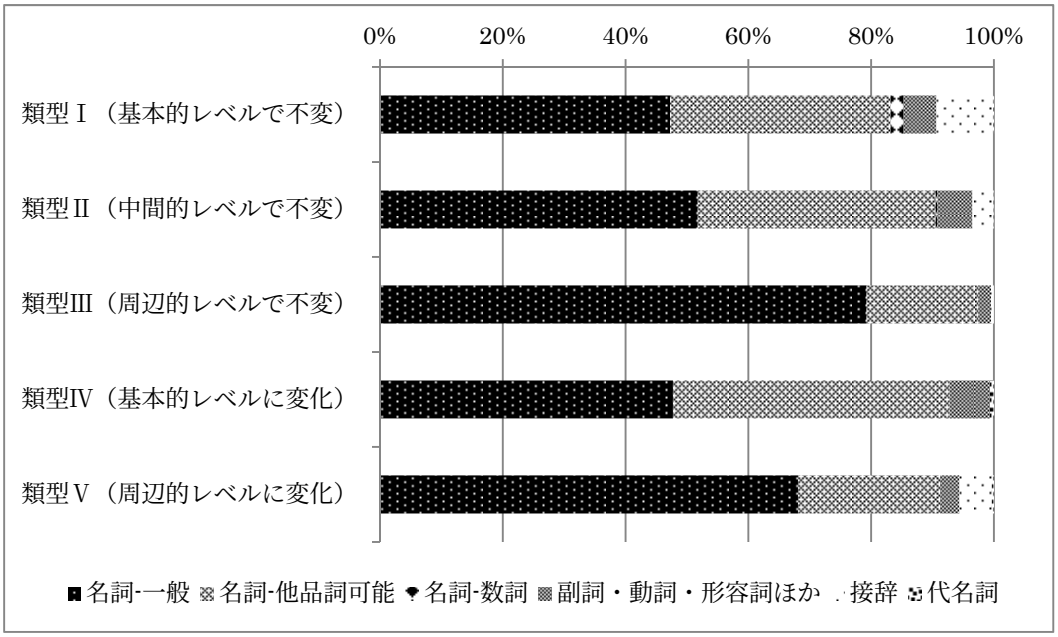


図 15　類型別の漢語の品詞比率

図 15 を見てまず目を引くこととして、類型Ⅲ（周辺レベルで不変）において名詞-一般の比率が極めて高いことがあげられ、類型Ⅴ（周辺レベル化）においても、これに準じて名詞-一般の比率が高い。これとちょうど反対に、類型Ⅰ（基本レベルで不変）において名詞-一般の比率がやや低くなっており、類型Ⅳ（基本レベル化）がこれに次ぎ、類型Ⅱ（中間レベルで不変）においても、それと大差ない。これら類型Ⅰ・Ⅱ・Ⅳにおいては、名詞-サ変等可能が、他の類型に比べて多くなっている。基本的なレベルで安定していたり、基本的なレベルに向けて変化する漢語には、普通名詞以外の品詞に転成するものが多く含まれていたが、周辺レベルで不変であったり、周辺レベルに追いやられる漢語には、普通名詞が多いという傾向を指摘することができよう。

やや細かいところに着目すると、接辞は、類型Ⅰ（基本レベルで不変）にはまとまった量が見られるが、類型Ⅳ（基本レベル化）にはほとんど見られないという違いがある、漢語接辞は、明治中期までにすでに基本的な語彙となっていたものが多く、明治後期以後に新たに基本レベル化するものはほとんどなかったということが分かる。一方、副詞・動詞・形容詞ほかは、類型Ⅳ（基本レベル化）で最も高い比率を示しており、明治後期以後に基本的な語彙となっていくものが多かったと見られるのである。

７．２　漢語における類型別の特徴

6.2 において和語について考察したのと同じ手順で漢語における類型別の特徴を概観したい。やはり「様相」の分類項目に分類される語彙のうち、漢語の各品詞について、類型別に一覧にする。

○名詞（体の類に相当）
類型Ⅰ（基本的レベルで不変）
相（そう）、真相、状態、価値、模様、様子、状況、現状、事情、裏面、形勢、大勢、景気、傾向、趨勢、気味、時勢、消息、空気、都合、調子、気風、観、相、体、像、実、要素、構造、性、性質、質、性格、人物、人格、特色、一種、長所、欠点、弱点、難、一般、精、良、不可、秩序、波瀾、美
類型Ⅱ（中間的レベルで不変）
側面、現況、実情、内情、国情、常態、旧態、事態、病状、好況、盛況、風潮、安危、動静、筆致、塩梅、口調、外観、外面、美観、奇観、格好、風采、醜態、品位、素質、本質、特質、気質、天性、本性、性情、異彩、長短、短所、遜色、空前、国粋、当否、怪、支障、事故、艱難、危急
類型Ⅲ（周辺的レベルで不変）
諸相、体様、原状、生態、病態、別状、騰勢、霊気、妖気、鬼気、俳味、力感、調、口跡、新風、和風、欧風、唐風、相好、家相、風水、吉相、仙骨、死相、人体、温容、機構、体制、材質、音質、軟質、硬質、上質、生得、獣性、神性、知性、癇、耐水、慣性、剛性、弾性、塑性、展性、延性、乾性、磁性、悪性、異体、利点、稀覯、精髄、満点、欠格、均整、変哲、波乱、粗密、繁簡、可視、可動、便宜
類型Ⅳ（基本的レベルに変化）
体質、個性、特徴、特長、欠陥、最善
類型Ⅴ（周辺的レベルに変化）
実況、惨状、学風

○名詞-サ変可能（用の類に相当）
類型Ⅰ（基本的レベルで不変）
呼吸、加減、構成、組織、相応、調和、一致、整理、始末、障害、故障、破壊、適当、円満
類型Ⅱ（中間的レベルで不変）
塩梅、適合、妥当、画一、拮抗、攪乱、混交、紊乱、紛乱、錯綜、混雑、雑踏、紛糾、緩和、整頓、整備、網羅、斟酌、配合、損傷、荒廃、退廃、難儀、重宝、邪魔
類型Ⅲ（周辺的レベルで不変）
活性、混線、混信、簡約、調律、緊迫
類型Ⅳ（基本的レベルに変化）
緊縮、調節
類型Ⅴ（周辺的レベルに変化）
輻輳

○名詞-形状詞可能・形状詞など（相の類に相当）
類型Ⅰ（基本的レベルで不変）
次第、不振、風、子細、普通、非常、特殊、珍、固有、妙、可、奇、複雑、単純、種々、純粋、困難、無理、楽、便利、便宜、不便、安全、無事、完全、危険、急、立派、大抵、尋常、一応、特別、格別、特有、変、重大、良好、善良、適切、健全、大概、一様、奇麗、密、簡単、容易

類型Ⅱ（中間的レベルで不変）
　異常、単調、皮相、壮観、不備、凡庸、非凡、破天荒、奇異、穏当、適度、適任、不当、混沌、稠密、精巧、煩、繁雑、艶、無垢、不潔、不能、平易、簡便、平安、平穏、緊急、平凡、通例、希有、独特、奇妙、奇抜、新奇、乙、珍奇、異様、深刻、粗悪、整然、粗末、粗雑、貧弱、緻密、雑多、簡明、純、流麗、至難、簡易、軽便、平易、険悪
類型Ⅲ（周辺的レベルで不変）
　好調、貧相、優性、中性、別段、最悪、適正、豊麗、典麗、凄艶、十全、火急、頓狂、大抵、特別、変梃、最適、冗漫、濃密、繚乱、安直
類型Ⅳ（基本的レベルに変化）
　結構、雑然
類型Ⅴ（周辺的レベルに変化）
　優、疎、簡、紛々、燦爛、爛漫

上記のリストで語数の多い類型Ⅰ、類型Ⅱ、類型Ⅲに分類される語を見ても、それぞれ、語構成や意味の面でこれといった特徴は指摘しにくい。和語では、類型間で語構成上の明瞭な差異が見えたのに対して、漢語では、どの類型でも二字漢語が大部分を占めていて[5]、語構成上の差異を指摘することが難しい。また、和語では、意味の抽象度・具体度の点で類型間に差異があったが、漢語については、そのような差異も指摘しにくい。確かに、現代語の語感として、類型Ⅰには最も基本的な語が多く、類型Ⅲにはなじみの薄い語が多く、類型Ⅱはその中間にある語という傾向を感じ取ることはできるが、それは現代語の語感であり、和語の場合のような、歴史を通じた確かな違いとは言えないように思われる。そのような語感に反するものの一例をあげれば、類型Ⅱの「安危」「艱難」「稠密」などは、現代語では周辺的な語という感じがし、一方、類型Ⅲの「好調」「貧相」「最悪」「適正」「大抵」などは、現代語では基本的な語と感じられる。これらの語は、『太陽コーパス』の時代以降にレベルの移行があったと思われるが、『太陽コーパス』の時代では漢語のレベルはまだ安定しておらず、各類型に属する漢語の特徴もはっきりしていない面があるのだと考えられる。

類型Ⅳ、類型Ⅴについて、語数の少ない上記のリストからは、これといった特徴は見出せない。一方、語彙全体を対象としたリストを作成して分析を試みても、和語の場合のようには、それぞれの類型に属する語彙の特徴を明確に指摘することは難しい。この時期に基本レベル化したり周辺レベル化したりする理由は、個々の漢語の語形を見ているだけでは不明である。基本レベル化や周辺レベル化する理由を探るには、漢語の使われ方を用例に即して分析していく必要がある。これについては、コーパスにおける用例分析を中心とする別の研究への展開が望まれるところである[6]。

８．おわりに

以上、本論文では、『太陽コーパス』に形態素解析を施し、その語彙全体を頻度調査の対象として、頻度に基づくレベル分けを実施し、このレベル情報を利用した語彙史研究を試みた。これは、近代語のテキストにも形態素解析が実現しつつある近年の研究状況を踏まえた新しい段階の研究例を示そうとしたものである。もっとも、開発途上の技術を利用しているためデータの精度には問題があることには留意しなければならない。

各語種が語彙全体の中でどのような位置を占め、それが時代によって変化する全体的な傾向を示すことができたところは、近代語彙史の研究の新しい展開だと言えるだろう。形

[5] 本稿では UniDic の言語単位である「短単位」によっているため、3 字漢語以上は分割されることになる。「長単位」によって調査を行えば、類型による語構成の差異が見出される可能性はある。

[6]田中（2011、2012）などは、その端緒としてまとめた研究である。

態素解析を施したコーパスを作って研究できる環境を整えていくことは、近代語の語彙史研究にとって、大きな意義を有すると考えられる。

文　献

小木曽智信（2009）『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（科学研究費補助金成果報告書）

国立国語研究所（1959）『明治初期新聞の用語』（国立国語研究所報告 5）

国立国語研究所（1964）『現代雑誌九十種の用語用字 Ⅲ分析』（国立国語研究所報告 25、秀英出版）

国立国語研究所（1987）『雑誌用語の変遷』（国立国語研究所報告 89、秀英出版）

国立国語研究所（2004）『分類語彙表―増補改訂版』（国立国語研究所資料集 14、大日本図書）

国立国語研究所（2005a）『現代雑誌の語彙調査―1994 年発行 70 誌―』（国立国語研究所報告 121）

国立国語研究所(2005b)『太陽コーパス―雑誌『太陽』日本語データベース―』（CD-ROM、国立国語研究所資料集 14、博文館新社）

国立国語研究所(2005c)『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』（国立国語研究所報告 122、博文館新社）

田中牧郎(2005)「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」（国立国語研究所(2005c)、pp.1-48）

田中牧郎（2011）「近代漢語の定着―『太陽コーパス』に見る―」（『文学』12-3、岩波書店、pp. 136-153）

田中牧郎（2012）「新漢語定着の語彙的基盤―『太陽コーパス』の「実現」「表現」「出現」と「あらわす」「あらわれる」など―」（『日本語の学習と研究』160、北京、pp.39-47）

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（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.101-122）

橋本和佳（2010）『現代日本語における外来語の量的推移に関する研究』（ひつじ書房）

飛田良文（1966）「明治以後の語彙の変遷」（『言語生活』182、筑摩書房、pp.16-24）

宮島達夫（1967）「近代語彙の形成」（『ことばの研究第３集』、秀英出版、pp.1-50）

付　記

本稿は、国立国語研究所国際学術集会「漢字漢語研究の新次元」（2010 年 7 月 30 日、国立国語研究所）および、第 104 回漢字漢語研究会（2012 年 8 月 1 日、早稲田大学）で発表した内容をもとに、大幅に書き改めたものである。特に国際学術集会のときには、『明六雑誌』『国民之友』のデータについても扱ったが、これらは『太陽』とは質の異なる面もあるため、今回は『太陽』のみを対象にした。また、本論文における頻度調査やレベルの設定は、今回新たに処理を行った形態素解析データに基づいている。

文献資料内漢語の階層化―『明六雑誌』の漢語をめぐって―

小野　正弘（明治大学文学部）[1]

１．はじめに

本稿では、『明六雑誌コーパス』を活用していく際のケーススタディーとして、近現代雑誌コーパスにおける『明六雑誌』初出となる漢語語彙について、語史（語彙史）研究を行う際の、基本的な性格付けを行なう。

具体的には、上記漢語について、『日本国語大辞典』第二版（以下、「日国第二版」と表記）の情報（初出資料、年代、意味等）と引き当てることにより、それらが、いつから使用されているものなのであるのか、また、それらが当初の意味で使用されているのか、それとも、変容された意味で使用されているのか、ということについての、展望を得る。

さらに、その分析の途上でいかなる問題が生じ、それをどう解決すべきかについても考察する。

今回は、『明六雑誌』に出現する漢語語彙（以下、「明六雑誌漢語」と略称する）について、その性格付けを試みる。その性格が明らかになれば、『明六雑誌』の漢語の位置が明らかになり、近代における語史を編んでゆく際にも、さらには、その前代の語史と接続する際にも、大きな指針となることが期待される。

２．具体的手順

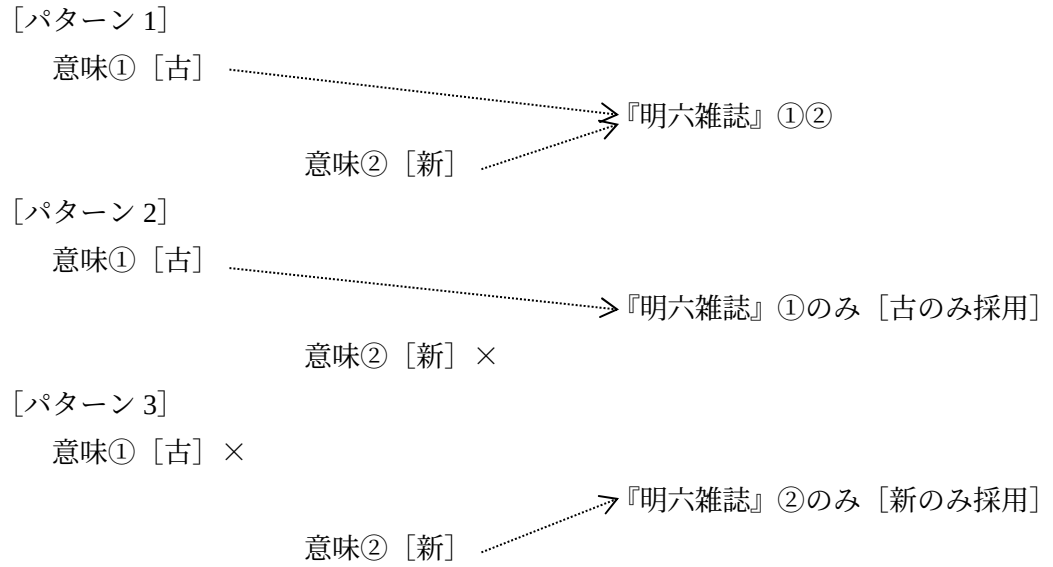
具体的な作業手順は。以下の通りである。

まず、(1)モデルコーパスとして作成中の『明六雑誌コーパス』から、漢語をすべて抽出したところ、7420 語であった[2]。(2)今回は、分析方法そのものをテストしてみたいということもあって、一時に全体をあつかうことはせず、そのうち語頭が「ア行」の漢語のみを対象として、「日国第二版」との引き当てを行った。該当するものは、496 語（全体の約 6.7%）となる。(3)上記のリストに、「初出資料」「年代」を加え、その年代が、「明治」時代と、どれほどの隔たりがあるのかを計算するスペースを設けた。(4)ところが、「初出資料」「年代」のみだと、ある漢語が、あるときに意味変化（あるいは、新しい意味の獲得を）して、その変化した意味で用いられている場合に対応できないので、「修正資料」「修正（年代）」「明治（時代との隔たり）２」という項も設けた。 (5)初出資料名は、「日国第二版」が用いているものをほぼ踏襲したが、

[1] wonomasa@kisc.meiji.ac.jp

[2] この抽出作業は、『明六雑誌コーパス』の作成途中に行ったので、公開版の『明六雑誌コーパス』とは、語数もその内訳も異なっている

「霊異記」→「日本霊異記」、「平家」→「平家物語」等のように、通行のものに改めたものもある。(6)年代については、「日国第二版」の西暦年数表記にしたがって記入した。『続日本紀』や古記録のような、記事日を年数とする表記法にも従った。ただし、「～頃」のような場合の「頃」は、計算の都合上、省いた。また、「1890-92」のように、複数年にわたるものは、最も新しい年数に統一した(上記の場合なら「1892」となる）。また、「12C初」「室町後期」などとされて、具体的な西暦年数になっていないものは、筆者が語史研究用に個人的に作成している成立年代表を用いて記入した（「平家物語」「太平記」など）。さらに、それに載っていないものは、『群書類従　正続分類総目録・文献年表』を用いて、年号付けを行なった。(7)「日国第二版」の見出しのもとに、複数の意味ブランチが分けてあるものについては、それらの中で、もっとも古い年代表示を持つものを、「初出」に記入し、「ひまわり」で当該の語を検索して、どの意味ブランチに該当するかを判断して、修正が必要な場合には、それを、「修正資料」に記入した。その場合、用例が複数あって、それらが複数の意味ブランチに該当し、なおかつ、もっとも古い年代表示を持つものにも該当する場合は、修正を加えていない[3]。



【図１】　「明六雑誌漢語」における意味の新古

さて、「明六雑誌漢語」における意味の新古のイメージは、図１のようになる。パターン１は、ある漢語の意味が①から②に変わりつつ共存し（実際には、３つ以上の場合もあるが、単純化するために、このかたちで考える）、そのどちらも『明六雑誌』で用いられている場合であり、パターン２は、意味が①から②に変化しているにも拘わらず、『明六雑誌』では①の意味で用いられている場合、パターン３は、意味が①から②に変化して、『明六雑誌』では②の新しい意味で用いられているというものである。[4]なお、(8)一字漢語で「日国

[3] 今回は、データを単純化するためにそうしたが、これをどう処理するかは、今後の課題となる
[4] このような判断を行って、修正資料の記入が必要となったものは、52 語あった。

第二版」には立項されていないものがあり、これは、全体の計数から省くことにした。また、「案」「安」などのように、「日国第二版」に立項され、一字漢語の例も示されているが、『明六雑誌』の用例が、はたして、一字漢語の用例としてよいかどうか疑わしいものがあり、そのようなものも、全体の計数から省くことにした。総じて、一字漢語の用法は、いろいろと問題がありそうであり、今後の詳細な分析が必要と考えられる。また、「以下（いか・いげ）」「王家（おうか・おうけ）」のような、読みをどう与えればよいかという問題の残るものも若干存在する（下線は、今回採用した読み）。このような観点から、全体の計数から除いたものは、49 語あり、したがって、今回の集計の際の母数は、447 語となる。[5]また、(9)「日国第二版」の掲げる初出資料よりも、『明六雑誌』の用例のほうが古そうな場合は、「ひまわり」で確認して、その旨を注記した。『明六雑誌』の用例のほうが、（たとえ１年でも）「日国第二版」よりも古いものは 45 語ある。[6]

３．分析結果

今回分析対象とする集計する、「明六雑誌漢語」の母数は、前述のように、447 語である。

３．１　年数の開きの分布

まず、「明六雑誌漢語」が、いつの時代から存在する漢語で構成されているのか、という問題を考える。

この際、従来のような、いわゆる上代・中古…からある、という考え方ではなく、明治から見て、何年前から使われている漢語を使用しているのか、という観点で考える。

まず、前記 2-(4)・(7)の修正以前の値で見てみると、次頁の【表１】のようになる。[7]

これを見ると、明治との隔たりが 1000 年以上のものが 99 語あるが、最もその差が大きいものは「意」（勝曼経義疏、611 年）で、1257 年となる。また、500 年以上前から用いられている漢語が、全体の半数（51.5%）を占める。

明治時代以降、新たにつくった（導入した）漢語は、１割を少し超えた程度である。

次に、前記 2-(4)・(7)の修正、すなわち、新しい用法で用いられている場合は、そちらの年代を採用するという操作を行なった数値で集計すると、次頁の【表 2】のようになる。

全体としてのおおまかな傾向は変わらないように見えるが、仔細に見ると、500 年以上前から用いられているものが減少し（18.1＋27.4＝45.5%、6 ポイント減）、明治以降のものが 3.6 ポイント増加している(16.8-13.2)ことが分かる。また、20 年前から用いられているものも微増している。すなわち、全体として、明治時代の方向へシフトしていることになる。[8]なお、修正前の【表 1】の数値は、『明六雑誌』のある漢語そのものが何年前から存在するかという数値であるのに対して、修正後の【表 2】数値は、『明六雑誌』のある漢

[5] 496 語［(1)参照］-49 語＝447 語、である。
[6] 3-4 節参照。
[7] 年数の区切りかたは、便宜に従ったが、49～1 あたりが幕末、199～50 あたりが江戸、499～200 あたりが室町～江戸初、999～500 あたりが平安～室町、1000 年以上以前が平安以前に対応はする。
[8] このこと自体は、予測されたことである。

【表１】「明六雑誌漢語」の時期構成(修正前)

1000 年以上前	99 語（22.2%）	明治以降	59 語（13.2）
999～500 年	131 語（29.3）		
499～200 年	84 語（18.9）	見出し語にあるが用例のないもの	2 語(0.4)
199～100 年	26 語（ 5.8）	見出し語にないもの	3 語(0.6)
99～ 50 年	18 語 （ 4.0）		
49～ 20 年	13 語 （ 2.9）		
19～ 1 年	12 語 （ 2.7）		

【表２】「明六雑誌漢語」の時期構成(修正後)

1000 年以上前	81 語（18.1%）	明治以降	75 語（16.8）
999～500 年	122 語（27.4）		
499～200 年	84 語（18.8）	見出し語にあるが用例のないもの	2 語(0.4)
199～100 年	30 語（ 6.7）	見出し語にないもの	3 語(0.6)
99～ 50 年	21 語（ 4.7）		
49～ 20 年	15 語（ 3.4）		
19～ 0 年	14 語（ 3.1）		

語の用法が、何年前から存在するかというものになっていることになる。

とはいえ、言うまでもないが、以上の結果は、あくまでも、ア行の漢語について明らかになったもので、これがカ行以降にも同様にあてはまるものであるかは、いまのところ不明であるとしか言えない。[9]

３．２ 「新しい」漢語

ここでは、「新しい」漢語というものについて考察する。「新しい」漢語には、次の二つの概念が考えられる。すなわち、

①明治時代（あるいは、その近く）になって、新たにつくられた（導入された）漢語

②語そのものは古くからあるが、明治時代（あるいは、その近く）になって、新しい用法を付与したもの

まず、①のほうから見ていく。【表１】のほうに拠って、49～ 20 年前から存在する 13 語を具体的に挙げると、

緯度,遠西観象図説,1823,-45　　委託,日本外史,1827,-41

因習（襲）,日本外史,1827,-41　　鋭意,日本外史,1827,-41

[9] ただ、ア行だから特別な傾向にあるということも言えないだろうから、ある程度、他の行にも敷衍できるものであることは推測してよいだろう。今回は、このようなかたちの分析を全体に及ぼしてはどうかという提案の意味も兼ねている。

永安,報徳記,1830,-38　　愛育,仮名文章娘節用,1834,-34
一環,江戸繁盛記,1836,-32　　演技,江戸繁盛記,1836,-32
演劇,夜航余話,1836,-32　　一世,椿山宛渡辺華山書簡,1839,-29
一隊,外国事情書,1839,-29　　塩酸,舎密開宗,1847,-21
温度,舎密開宗,1847,-21

となる[10]。次に、19〜1 年前からの 12 語を挙げると、

異性,異人恐怖伝,1850,-18　　鬱然,漂荒紀事,1850,-18
英語,外国事件書類雑纂,1856,-12　　英書,航米日録,1860,-8
英人,増訂華英通語,1860,-8　　永続,春情花の朧夜,1860,-8
英米,航米日録,1860,-8　　偉功,隣艸,1861,-7
往者,星巌先生遺稿,1865,-3　　印刷,経済小学,1867,-1
英学,財政経済資料,1867,-1　　冤罪,和英語林集成,1867,-1

となる。ここでは、「英〜」という語構成を持つものが目につく。さらに、明治時代以降のものを挙げると、

暗殺,新令字解,1868,0　　依拠,泰西国法論,1868,0
允可,明治月刊,1868,0　　英仏,新令字解,1868,0
縁由,泰西国法論,1868,0　　運輸,日誌字解,1869,1
鋭敏,漢語字類,1869,1　　愛撫,神霊を鎮祭し給へる詔,1870,2
圧伏,西洋事情,1870,2　　威力,西洋事情,1870,2
愛国,百学連環,1871,3　　圧抑,西国立志編,1871,3
安息,西国立志編,1871,3　　偉勲,西国立志編,1871,3
偉丈夫,西国立志編,1871,3　　永遠,西国立志編,1871,3
英王,西国立志編,1871,3　　英蘭,西洋聞見録,1871,3
閲歴,西国立志編,1871,3　　演繹,百学連環,1871,3
音字,百学連環,1871,3　　偉大,新撰字解,1872,4
一案,明六雑誌,1874,6　　欧亜,明六雑誌,1874,6
王党,明六雑誌,1874,6　　悪路,明六雑誌,1875,7
悪行,開花問答,1875,7　　圧制,文明論之概略,1875,7
遺利,明六雑誌,1875,7　　殷鑑,文明論之概略,1875,7
旺盛,明六雑誌,1875,7　　汚行,明六雑誌,1875,7
域内,東京新繁盛記,1876,8　　以西,米欧回覧実記,1877,9

[10] 半角カンマで区切られた部分の構造は、「漢語語彙素，資料名，日国第二版による年代，明治時代との開き」の順。以下同じ。

以南,米欧回覧実記,1877,9　衣被,米欧回覧実記,1877,9
以北,米欧回覧実記,1877,9　移民,米欧回覧実記,1877,9
愛児,花柳春話,1879,11　院長,花柳春話,1879,11
淫蕩,花柳春話,1879,11　栄進,花柳春話,1879,11
黄色,造化妙々奇談,1880,12　亜,近世紀聞,1881,13
移動,哲学字彙,1881,13　印象,哲学字彙,1881,13
異常,日本開化小史,1882,14　栄誉,日本開化小史,1882,14
威令,経国美談,1883,15　畏敬,経国美談,1884,16
汚点,狐の裁判,1884,16　一読,小説神髄,1886,18
陰険,内地雑居未来之夢,1886,18　円,工学字彙,1886,18
欧土,小説神髄,1886,18　欧文,風俗画報,1898,30
一見,思出の記,1901,33　異質,ブラリひょうたん,1950,82
圧政,城,1965,97

となる。これを見ると、『西国立志編』『百学連環』『米欧回覧実記』といった明六社関係者の資料がならぶ(『明六雑誌』は言うまでもなく)。なお、「日国第二版」の初出表示が、明らかに「明六初出」よりも後になるものがあるが、それについては後述する。

次に、②すなわち、語そのものは古くからあるが、明治時代（あるいは、その近く）になって、新しい用法を付与したものを見る。これについては、2-(7)で述べたもののパターン３に当る 52　語のうちで、その修正した資料が、新しいものを見ればよい。修正の対象となった資料が、1800　年以降のものを示すと、

一挙,延喜式,927,-941,椿説弓張月,1811,-57
唯々,太平記,1374,-494,椿説弓張月,1811,-57
悦,古事談,1215,-653,虫眼鏡,1812,-56
陰,十巻本和名抄,934,-934,都繁盛記,1837,-31
移住,紀伊続風土記付録,1194,-674,海外事情書,1839,-29
悪魔,宇津保物語,999,-869,鼠小紋東君新形,1857,-11
淫佚,令義解,718,-1150,西洋事情,1866,-2
一意,サントスの御作業,1591,-277,西国立志編,1871,3
隠匿,椿説弓張月,1811,-57,西国立志編,1871,3
悪法,日蓮遺文,1275,-593,文明論之概略,1875,7
運動,禅竹伝書,1456,-412,文明論之概略,1875,7
医学,文明本節用集,1474,-394,文明論之概略,1875,7
違式,三代格,802,-1066,音訓新聞字引,1876,8
印紙,清涼軒日録,1487,-381,西洋道中膝栗毛,1876,8
隠然,中華若木詩抄,1520,-348,音訓新聞字引,1876,8

異教,信長記,1622,-246,米欧回覧実記,1877,9
一目,往生要集,985,-883,花柳春話,1879,11
域,落葉集,1598,-270,花柳春話,1879,11
位階,続日本紀,711,-1157,刑法,1880,12
帷幕,本朝無題詩,1164,-704,五国対照兵語辞書,1881,13
一部,延喜式,927,-941,郵便報知新聞,1892,24
悪食,新吾左出放題盲牛,1781,-87,社会百面相,1902,34
悪名,梅津政景日記,1612,-256,野分,1907,39

となる。[11]

最初の「一挙,延喜式,927,-941,椿説弓張月,1811,-57」というデータを例にして説明すると、「一挙」という語は、『延喜式』(927 年)から見られる語であり、それに基づけば、『明六雑誌』との年の開きは 941 年ということになるが、『延喜式』における「一挙」の意味は《ひとたび行うこと》であり、『明六雑誌』で用いられている《ものごとが速やかにはかどること》[12]ではない。《ものごとが速やかにはかどること》の意味は、『椿説弓張月』(1811 年)から見られ、その線で修正すれば、『明六雑誌』との年の開きは 57 年ということになる。以下も同様である。

これらを見ると、かなり古くからある漢語を、新しい用法で用いたものが多いさまが見て取れよう。なお、「違式」から下にある、明治８年以降（上記修正資料でいうと、データの最後の数値８以上のもの）の資料は、言うまでもなく、『明六雑誌』の例のほうが古い。

次に、初出資料と修正資料の年代の開きが大きいものを見る。500 年以上の隔たりのあるものを示すと、以下の通りになる。

一揆,三代格,844,-1024,太平記,1374,-494,530
一流,多度神宮寺伽藍縁起資財帳,801,-1067,太平記,1374,-494,573
悦,古事談,1215,-653,虫眼鏡,1812,-56,597
悪法,日蓮遺文,1275,-593,文明論之概略,1875,7,600
悪風,今昔物語集,1120,-748,政談,1727,-141,607
運上,書陵部所蔵壬生古文書,987,-881,室町殿日記,1602,-266,615
悪道,観智院本三宝絵,984,-884,日葡辞書,1604,-264,620
有無,法華義疏,615,-1253,平家物語,1250,-618,635
一種,性霊集,835,-1033,史記抄,1477,-391,642
移住,紀伊続風土記付録,1194,-674,海外事情書,1839,-29,645

[11] 半角カンマで区切られたデータの構造は「漢語語彙素，初出資料，年代，明治時代との隔たり，修正資料名，日国第二版による年代，明治時代との隔たり」である。
[12] これらの意味表示は、「日国第二版」のものを踏襲している。ただし、表記を若干変更したものがある。

帷幕,本朝無題詩,1164,-704,五国対照兵語辞書,1881,13,717
一理,法華義疏,615,-1253,東寺百合文書,1422,-446,807
一斑,令義解,718,-1150,四河入海,1534,-334,816
悪魔,宇津保物語,999,-869,鼠小紋東君新形,1857,-11,858
一番,延喜式,927,-941,無事志有意,1798,-70,871
安居,日本書紀,720,-1148,日葡辞書,1604,-264,884
一挙,延喜式,927,-941,椿説弓張月,1811,-57,884
一目,往生要集,985,-883,花柳春話,1879,11,894
陰,十巻本和名抄,934,-934,都繁盛記,1837,-31,903
案内,続日本紀,720,-1148,応仁略記,1670,-198,950
一部,延喜式,927,-941,郵便報知新聞,1892,24,965
違式,三代格,802,-1066,音訓新聞字引,1876,8,1074
淫佚,令義解,718,-1150,西洋事情,1866,-2,1148
位階,続日本紀,711,-1157,刑法,1880,12,1169

意外に、明治時代に入ってからのものはあまりなく、24 語中、６語(修正資料名に下線を付したもの)を数えるのみである。逆に言えば、新しい意味が生みだされてのち、安定したものを用いていた、ということになろうか。年代の開きが大きいという意味で「新しい」意味を用いている漢語は、必ずしも、明治期における「新しい」漢語ではない、ということになる。

３．３　「明六雑誌漢語」と「日国第二版」

前述したように、「日国第二版」における初出例が、「明六雑誌漢語」よりも遅れるものが、45 語見つかっている。ア行に限定した 447 語のなかに 45 語も見出されるというのは、やはり多いと言って差し支えないのではないだろうか（10.1%にあたる）。『明六雑誌』は、「日国第二版」の用例採取資料ともなっているので、これは、資料全体をコーパスとしているか否かの違いによるものであろう。[13]

その具体例を挙げると、以下の通りである。

亜,近世紀聞,1881,13,,,
愛児,花柳春話,1879,11,,,
愛想,寛永刊本蒙求抄,1534,-334,,,
悪食,新吾左出放題盲牛,1781,-87,社会百面相,1902,34
悪法,日蓮遺文,1275,-593,文明論之概略,1875,7
悪名,梅津政景日記,1612,-256,野分,1907,39
圧政,城,1965,97,,,

[13] 逆に言えば、この『明六雑誌コーパス』の優秀さと、これを構築した意義もよく示すものと言える。

圧制,文明論之概略,1875,7,,,
医学,文明本節用集,1474,-394,文明論之概略,1875,7
域内,東京新繁盛記,1876,8,,,
異教,信長記,1622,-246,米欧回覧実記,1877,9
畏敬,経国美談,1884,16,,,
違式,三代格,802,-1066,音訓新聞字引,1876,8
異質,ブラリひょうたん,1950,82,,,
異常,日本開化小史,1882,14,,,
以西,米欧回覧実記,1877,9,,,
一読,小説神髄,1886,18,,,
一部,延喜式,927,-941,郵便報知新聞,1892,24
一目,往生要集,985,-883,花柳春話,1879,11
一見,思出の記,1901,33,,,
一夫,漢書列伝竺桃抄,1460,-408,,,
移動,哲学字彙,1881,13,,,
以南,米欧回覧実記,1877,9,,,
帷幕,本朝無題詩,1164,-704,五国対照兵語辞書,1881,13
衣被,米欧回覧実記,1877,9,,,
以北,米欧回覧実記,1877,9,,,
移民,米欧回覧実記,1877,9,,,
威令,経国美談,1883,15,,,
殷鑑,文明論之概略,1875,7,,,
陰険,内地雑居未来之夢,1886,18,,,
印紙,清涼軒日録,1487,-381,西洋道中膝栗毛,1876,8
印象,哲学字彙,1881,13,,,
隠然,中華若木詩抄,1520,-348,音訓新聞字引,1876,8
院長,花柳春話,1879,11,,,
淫蕩,花柳春話,1879,11,,,
運動,禅竹伝書,1456,-412,文明論之概略,1875,7
栄進,花柳春話,1879,11,,,
栄誉,日本開化小史,1882,14,,,
冤,春雨物語,1808,-60,近世紀聞,1878,10
円,工学字彙,1886,18,,,
黄色,造化妙々奇談,1880,12,,,
欧土,小説神髄,1886,18,,,
欧文,風俗画報,1898,30,,,
憶測,音訓新聞字引,1786,-82,,,

汚点,狐の裁判,1884,16,,,

以上でもっとも隔たりの大きいものは、「圧政」の 97 年で突出して大きく、つぎは「異質」の 82 年、「悪名」の 39 年（修正資料による）「悪食」の 34 年（同）、と続く。[14]

なお、「一品」、「一婦」、「姻縁」、「右顧」、「易姓」の５語は、「日国第二版」において、見出し語として立項されていないが、『明六雑誌』における用例のあるものである[15]。

３．４　「明六雑誌漢語」と初出資料

「明六雑誌漢語」が、どのような資料で初出となっているかを見てみると、以下のようになる。

『続日本紀』32　　『文明本節用集』15　　『令義解』９　『万葉集』９
『太平記』９　　『史記抄』９　　『菅家文草』８　『延喜式』７
『今昔物語集』７　『西国立志編』７　　『明六雑誌』７　『性霊集』６
『法華義疏』５　『懐風藻』５　『日本霊異記』５　　『本朝文粋』５
『正法眼蔵』５　『童子問』５　『米欧回覧実記』５　『文華秀麗集』４
『政談』４　『花柳春話』４

ここでは、頻度４以上のものを挙げたが、古い資料に例のあるものが多いことが目を引く。このことは、すなわち、「明六雑誌漢語」が、古い文献に典拠を持つもので占められていることを証し、それは、とりもなおさず、「明六雑誌漢語」を生みだした人々の、漢語についての知識のたしかさ、造詣の深さを証すものであろう。古い資料に典拠があるということは、すなわち、中国古典（あるいは、下っても、隋唐の資料）に典拠があるということだからである。

４．おわりに

以上、「明六雑誌漢語」を材料に、それがどのような性格を有しているのかについて、瞥見し、その階層化を試みた。今後は、この調査をすべての漢語に及ぼして、総合的にみるとどのような様相を示すのかを、さらに確認する必要がある。

また、新しい意味で用いられた漢語（修正資料を持つ漢語）については、古い用法も共存しているのかどうかを再確認する必要がある。そのこと自体は、修正資料を持つ漢語語彙を、「ひまわり」で検索して、実際の用例を確認することで果たせる（ただし、実際は、そこまで単純でなく、用例としてかなり微妙なものも目に付く）。

「新しい」漢語の用法についても、さらなる分析が必要である。特に、新しくつくられた（導入された）漢語については、その出所を考究する必要があるし、新しい意味で用いられるようになったものについても、その経緯を明らかにする必要がある。

---

[14] 本稿末の［補足］参照。
[15] これらが、『明六雑誌』以前に用例があるものなのかどうかは、さらなる調査を要する。

また、根本的な問題であるが、漢語の語彙素にどのような読みを与えるかも、今後の問題となろう。

さらに、同様の調査を、近現代における、他の主要な雑誌における漢語へと進めて行く必要がある。それらが完成すれば、近現代における雑誌のコーパスの漢語語彙は、この時期の漢語の語史（語彙史）をたどるための根幹として、重要な位置を占めることになろう。また、今回は、「日国第二版」を一つの巨大なデータベースと見て、そのデータとの引き当てを試みたわけであるが、そのことによって、近現代における雑誌によるコーパスの側から、「日国第二版」の側へ引き渡すべき情報が生まれた。そのことによって、「日国第二版」というデータベースを補完するという、データのキャッチボールもできよう。

［補足］

○壓政　ヤ且ヤ我國人民永ク壓政ノ下ニ屈シテ人性自由：明六雑誌：1874：12：政論ノ三：津田眞道：文語：P012A001（「ひまわり」検索）

権力や暴力などを用いて、国民の自由を奪う政治。圧制政治。

＊城〔1965〕〈水上勉〉九「若狭三十二谷に、酒井圧政怨詛の声が充満したとしても不思議でなかった」（日国第二版）

○異質　テ之ヲ性質ニテ言ハヽ異質ヲ變シ同質トナシ 之：明六雑誌：1875：42：人世三寶説四：西周：文語：P009A002

ルカ如シ 然ルニ今此異質ヲ變シテ同質トナシ抗：明六雑誌：1875：42：人世三寶説四：西周：文語：P012A003（「ひまわり」検索）

物事や人などの性質が他とちがっていること。また、そのさま。同質。

＊ブラリひょうたん〔1950〕〈高田保〉対面「同気同質の二人なら、対面して話の末に何か生れることもあるだろうが、異気異質の二人ではどうにもなるはずがない」（日国第二版）

○悪名　ザルノミナラズ 往々惡名ヲ負ヒ罪人トナリ囚獄：明六雑誌：1875：37：賞罰毀譽論：中村正直：文語：P008A002（「ひまわり」検索）

①「あくみょう（悪名）」に同じ。

＊野分〔１９０７〕〈夏目漱石〉八「結果は悪名（アクメイ）にならうと、臭名にならうと気狂にならうと仕方がない」（日国第二版）

②悪いうわさ・評判。悪評。あくめい。

＊将門記〔９４０頃か〕「前生の貧しき報いを憂へず。但し悪名の後に流（つたは）

るを吟（によ）ぶ」（日国第二版）

○悪食　跣ニテ歩行シ常ニ惡衣惡食ニ安シ開化文明ノ風地：明六雑誌：1875：34：想像鎖國説（明治八年三月十六日演説）：杉亨二：文語：P002A013（「ひまわり」検索）

①粗末な食物。粗食。あくしょく。

＊社会百面相〔１９０２〕〈内田魯庵〉代議士・下「斯ういふ悪食（アクジキ）を貪って臭きを知らざる豚の寄合ぢゃから」［中略］

②普通の人がいやがって食用としない物を食べること。いかもの食い。転じて、普通の人がいやがってしないようなことをしたり、趣味とすること。

＊洒落本・新吾左出放題盲牛〔１７８１〕折助冷飯「契りみじかき一寸の間に、かたみの瘡（かさ）の身にしみじみと、命しらずの悪食（あくシキ）者と」（日国第二版）

『明六雑誌』の一人称代名詞

近藤 明日子（国立国語研究所コーパス開発センター）[1]

１．はじめに

日本の近代の言語資料のコーパス化とそれを用いた近代語研究は今後一層の発展の期待される分野である。コーパス言語学的手法による近代語研究には、形態論情報の付与されたコーパスの開発が必須であるが、近代の文語論説文を対象とした形態素解析辞書「近代文語 UniDic」の開発により、その環境整備は飛躍的に進んだ。現在は、実際にその技術を用いた近代語の形態論情報付きコーパスの開発が始まっており、「近代語コーパス」プロジェクトにおいても、明治初期に刊行された『明六雑誌』の形態論情報付きコーパスである『明六雑誌コーパス』を開発した。

本稿は、この『明六雑誌コーパス』を用いて、そこに出現する一人称代名詞の分析を行うものである。用例の抽出や分析では、形態論情報をはじめとするコーパスに付与された情報を用い、コーパスの特長を活かした研究となることを目指す。そして、『明六雑誌』というほとんどが論説文よりなる資料を用いることで、当時の書き言葉的要素の強い資料における一人称代名詞の使用実態の一端を明らかにしたい。

２．『明六雑誌コーパス』の概要

『明六雑誌コーパス』は、明治 7（1874）年から明治 8（1875）年にかけて刊行された、明六社の機関誌である『明六雑誌』の全文コーパスである。明六社は当時の洋学者によって結成された学術団体であり、そこで行われた演説や討論を広く一般に発表する媒体として『明六雑誌』は刊行された。よって、そこに掲載された記事はほとんどすべてが、ある物事について論じ解説する論説文となっている。

この『明六雑誌』に基づく『明六雑誌コーパス』は、本文テキストに書誌・文書構造・形態論・文字等に関する情報を付与する設計となっている。付与される情報のなかで特に注目されるのは形態論情報であろう。なぜなら、これまで形態論情報の付与された近代語のコーパスはほとんど例がなく、近藤・小木曽・加藤（2010）の『高等小学読本』コーパスといったものがわずかに存在するだけだからである。本コーパスの形態論情報は、『高等小学読本』コーパス同様、近代の文語論説文（明治普通文）を対象とする形態素解析辞書「近代文語 UniDic」を用いて本文を形態素解析した後、人手修正を加えたものが付与される。それにより、語の単位として揺れのない斉一な単位である「短単位」（小椋・小磯・冨士池・他、2011）を採用し、表記の揺れや語形の変異にかかわらない見出し語を付与した、日本語研究に適した構造を持つ情報となっている。

[1] kondo@ninjal.ac.jp

本コーパスに付与された形態論情報をはじめとする情報に基づき、コーパスの規模を概観すると、全 43 号に掲載された記事の総数は 155 記事、著者（翻訳者含む）は異なりで 16 名、延べ語数は約 18 万 3 千語（記号類を除く）となる[2]。

表 1 は、著者別に記事数を示したものである。これを見ると、記事数の多い上位 3 名（津田真道・西周・阪谷素）によって著された記事が計 74 記事と、全記事数の約半分を占めていることがわかる。本コーパスの分析から導き出される実態が、当時の論説文の一般的なありようではなく、特定の著者による個別的なありようである可能性があることになり、本コーパスを言語資料として扱う際には、そのことを十分に念頭に置いておく必要があるであろう。

さらに、記事の地の文の文体について見ると、全 155 記事のうち、文語文体の記事が 150 記事、口語文体の記事が 4 記事、文語口語混合文体の記事が 1 記事となっており、ほとんどが文語文体の記事で占められ、口語文体の記事はごくわずかしかない。文体の面でもデータに偏りがあることにも留意する必要がある。

表 1　著者別記事数

著者	記事数
津田真道	29
西周	25
阪谷素	20
杉亨二	13
森有礼	12
西村茂樹	11
中村正直	11
神田孝平	9
加藤弘之	8
箕作麟祥	5
柏原孝章	4
福沢諭吉	3
清水卯三郎	2
箕作秋坪	1
津田仙	1
柴田昌吉	1
合計	155

３．分析対象とする語の抽出とその度数の概観

以下、この『明六雑誌コーパス』に出現する一人称代名詞の分析を行う。近代語の人称代名詞の研究は、これまで話し言葉的性質の強い資料（小説の会話部分、落語速記、口語文典など）を中心に行われてきた。よって、論説文といった書き言葉的性質の強い資料における実態は未だ明らかになっていない部分も多い。本稿の分析によりその実態の一端を明らかにしたい。

分析のためには、まず一人称代名詞の抽出が必要となるが、抽出作業は次にあげる手順でおこなった。

① 本コーパスの形態論情報を用い、品詞が代名詞となっている見出し語を抽出する。
② 国語辞典等を参照し、①から一人称代名詞の可能性のある見出し語を選別する。
③ 一人称代名詞と関わりの深い見出し語として本コーパスでは連体詞となっている「わが」「おのが」を②に追加する。
④ までの作業で得られた見出し語に属する用例について、文脈を確認し、実際に一人称代名詞として用いられているものを選別し分析対象とする。さらに、関連の深い用法として、人称にかかわらず対象それ自身を指す、いわゆる反射指示代名詞として用いられている用例も分析対象とした。

[2] 本稿に示すコーパスに基づく数値は 2011 年 12 月時点のデータに基づくものであり、公開中のコーパスに拠るものとは一部異なる場合がある。

この手順により、異なりで 15 語、延べで 1202 語の一人称代名詞および反射指示代名詞が抽出された。語ごとに記事の文体別の度数と表記の種類を表したものが表 2 である[3]。

表 2　表記の種類と記事文体別度数

語	表記の種類	度数			
		文語記事	口語記事	混在記事	合計
わが	我(474)、我ガ(67)、吾(19)、吾ガ(13)	538	27	8	573
よ	余(189)、予(6)	195	0	0	195
われ	我(127)、吾(20)、我レ(13)、予レ(1)	159	1	1	161
おのれ	己(101)、己レ(31)	129	3	0	132
ごじん	吾人(49)	49	0	0	49
ぼく	僕(17)	17	0	0	17
ごはい	吾輩(15)	15	0	0	15
わがはい	我輩(14)	14	0	0	14
せっしゃ	拙者(12)	0	12	0	12
ごせい	吾儕(10)	10	0	0	10
よはい	余輩(10)	10	0	0	10
それがし	某(4)、某シ(3)	4	0	3	7
わたくし	私(4)	0	4	0	4
よせい	余儕(2)	2	0	0	2
ぼくはい	僕輩(1)	1	0	0	1
合計		1143	47	12	1202

これを見ると、度数の多い上位 5 語「わが」「よ」「われ」「おのれ」「ごじん」の度数を合計すると 1110 語と全体の 90%以上を占め、これら 5 語が『明六雑誌』で主たる語であったことがわかる。

また、記事の文体別の度数を見ると、「せっしゃ」「わたくし」の 2 語はすべての用例が口語記事中に出現しており、これらの語の話し言葉的性質の強さがうかがえる。しかしながら、２．で述べたように『明六雑誌』の口語記事はごくわずかであり、その少量のデータに基づいて、語と文体との対応関係を分析し、当時の口語文体の論説文における一人称代名詞および反射指示代名詞の実態について論じるには限界がある。よって、以後は文語記事中に出現する語に限って分析を進めることとする。

４．語と後続助詞との対応関係

[3] 表 2 にあげられた表記の中には、読みの特定が困難なものもある。例えば、「我」「吾」一字の表記は、「わが」と読むのか「われ」と読むのか（それともそれ以外で読むのか）、はっきりしない場合がある。また、「吾輩」二字の表記は「ごはい」と読むのか「わがはい」と読むのか、断言することは難しい。そこで、「我」「吾」表記は、文脈から判断して「わが」「われ」いずれかに割り振り、それ以外の漢字表記はそれぞれ一種類の読みに倒して度数を数えた。よって、例えば「吾輩」表記はすべて「ごはい」と見なし、「わがはい」として数えることはしなかった。また、「己」表記は、助詞「が」が後続する場合「おの」と読むことも多分に考えられるが、「己レガ」という「おのれ」＋「が」とほぼ確定できる表記があったため、すべて「おのれ」と見なし、「おの」として数えることはしなかった。

また、『明六雑誌』では濁音を表記する仮名に濁点が用いられていない場合があり、「わが」の「が」も「カ」と表記されることがあるが、それらはすべて「ガ」に校訂した上で表記ごとの度数を数えた。

次にあげる表3は、文語記事中に出現する語ごとに代名詞としての用法別度数を示したものである。

表3　代名詞用法別度数

語	一人称	反射指示	合計
わが	456	82	538
よ	195	0	195
われ	107	52	159
おのれ	0	129	129
ごじん	49	0	49
ぼく	17	0	17
ごはい	15	0	15
わがはい	14	0	14
ごせい	10	0	10
よはい	10	0	10
それがし	4	0	4
よせい	2	0	2
ぼくはい	1	0	1
合計	880	263	1143

ここから、一人称用法を持つ語は12語、反射指示用法を持つ語は3語あることがわかる。同じ用法を持つ語が複数存在する場合、内部でさらに何らかの使い分けがなされていると考えられるが、そうした語の間の違いについて探るため、後続する助詞・助動詞ごとに度数を示した表4を用い、コレスポンデンス分析を行った。コレスポンデンス分析は、データ表の行や列に含まれる情報を少数の成分（次元）に圧縮し、それらの関係を散布図上に布置することで、行カテゴリー間の関係、列カテゴリー間の関係、および行カテゴリーと列カテゴリー間の関係を視覚的に捉えることができる分析手法で、コーパス言語学においても活用範囲が広いとされるものである（石川・前田・山崎（編）、2010、pp.245-249）。

表4　後続助詞別度数

語	が_体	ナシ	の_体	が_用	に	を	は	の_用	と	も	より	てふ	なり	など	合計
わが_一	444	0	0	12	0	0	0	0	0	0	0	0	0	0	456
わが_反	75	0	0	7	0	0	0	0	0	0	0	0	0	0	82
よ	24	100	5	26	1	3	22	3	3	7	0	0	1	0	195
われ_一	0	58	8	0	24	5	5	3	2	0	1	0	1	0	107
われ_反	0	22	9	0	7	6	0	1	4	0	1	2	0	0	52
おのれ	50	14	31	4	11	16	0	2	0	0	1	0	0	0	129
ごじん	0	28	13	0	0	2	0	6	0	0	0	0	0	0	49
ぼく	1	13	0	2	0	0	0	1	0	0	0	0	0	0	17
ごはい	0	10	2	0	0	0	1	1	0	0	0	0	0	1	15
わがはい	0	10	1	0	0	0	0	3	0	0	0	0	0	0	14
ごせい	0	6	1	0	0	0	0	3	0	0	0	0	0	0	10
よはい	0	5	2	0	0	0	1	1	0	1	0	0	0	0	10
それがし	0	4	0	0	0	0	0	0	0	0	0	0	0	0	4
よせい	0	2	0	0	0	0	0	0	0	0	0	0	0	0	2
ぼくはい	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
合計	594	273	72	51	43	32	29	24	9	8	3	2	2	1	1143

表4では、一人称と反射指示の両方の用法を持つ「わが」「われ」については、用法ごとにカテゴリー化し、一人称用法のものを「わが_一」「われ_一」、反射指示用法のものを「わが_反」「われ_反」として示した。また、後続する助詞・助動詞のうち「が」「の」につい

ては、後ろの体言にかかる連体用法をとるものと後ろの述語にかかる連用用法をとるものとを分けてカテゴリー化し、前者を「が_体」「の_体」、後者を「が_用」「の_用」として示した。助詞・助動詞の後続しないものについては「ナシ」としてカテゴリー化した。さらに、「わが」については、「わが」の「が」を後続する助詞と見なして度数をカウントした。

コレスポンデンス分析に用いたのは表 4 全体ではなく、網掛けを施した部分である。「わが」については、そもそも後続の助詞・助動詞という観点からの分析にはそぐわないため、分析対象から外し、また、外れ値の影響を考慮して、合計の度数が 10 未満のカテゴリーについては分析対象から外したものである。分析には、統計分析パッケージ R の MASS ライブラリーの corresp 関数を用いた。

分析結果から、もっとも寄与率の高い第 1 次元（47.62%）と第 2 次元（29.94%）の得点を 2 次元空間上に布置したものが図 1・図 2 で、図 1 は後続助詞の得点の散布図、図 2 は語の得点の散布図である。

まず、図 1 の第 1 次元を見ると、正の方向に「が_体」「を」「の_体」「に」が布置され、負の方向に「は」「の_用」「ナシ」「が_用」が布置されている。負の方向に布置される助詞群が受ける語は、多くの場合、述語に対し動作主や経験者といった意味的役割を担う[4]。一方、正の方向に布置される助詞群が受ける語は、述語に対し動作主や経験者といった意味的役割を担うことは「が_体」「の_体」の場合はもちろんなく、「を」「に」の場合も多くはない。つまり、第 1 次元は動作主や経験者といった意味役割を担うか否かに基づくものであることになる。これを図 2 と対応させてみると、他の語から大きく離れて正の方向に布

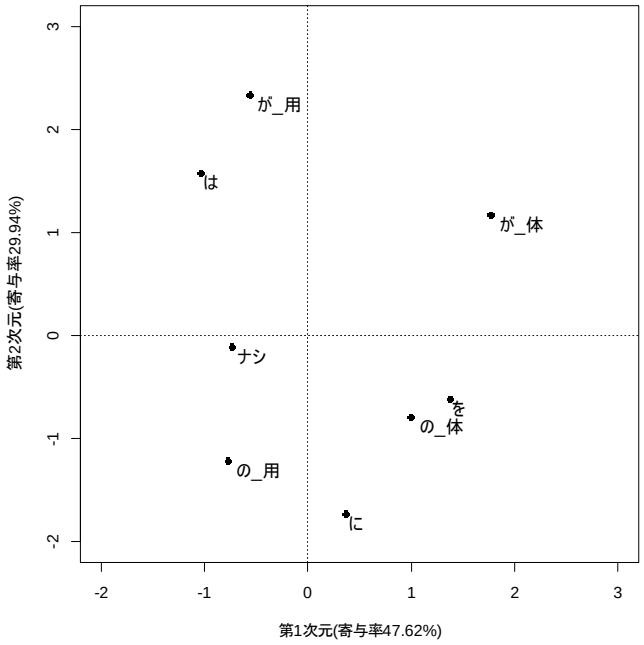


図 1　後続助詞の散布図

[4] 「は」の場合は相当する格助詞に置き換えた場合の意味的役割について言う。

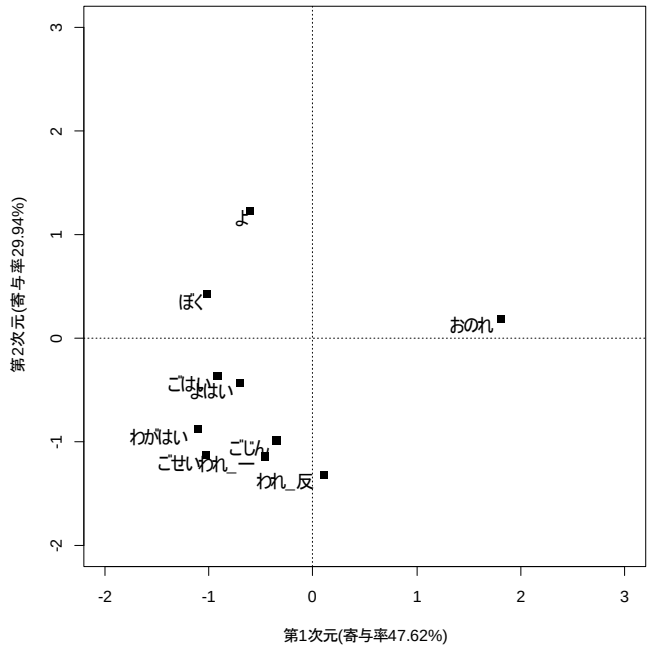


図 2　語の散布図

置されている「おのれ」は、動作主や経験者といった意味的役割を担うことが少ないといった点で特徴付けられることになる。

次に、図 1 の第 2 次元を見ると、正の方向に「が_用」「は」「が_体」が布置され、負の方向に「に」「の_用」「の_体」「を」が布置されている。この軸の解釈は難しいところがあるが、正の方向に助詞「が」が、負の方向に助詞「の」が集まっている点には留意される。「が」「の」は人を表す体言をうける場合、待遇表現上の区別が認められ、「が」の用いられる場合はその人物に対する親愛・軽蔑・憎悪・卑下等の感情を伴い、「の」が用いられる場合には敬意あるいは心理的距離があると言われている。とすると、第 2 次元は待遇の程度に基づくものであることが考えられる。図 2 と対応させて見ると、正の方向に布置される「よ」「ぼく」は、負の方向に布置される「われ_反」「われ_一」「ごせい」「ごじん」「わがはい」「よはい」「ごはい」と比較して、相対的に待遇の程度が低いことになる。

以上のように、後続助詞と語との間には明らかな対応関係があり、それにより語は大きく次の 3 つのグループに分けることができると考えられる。

A　おのれ

B　よ・ぼく

C　われ_一・われ_反・ごじん・ごはい・わがはい・ごせい・よはい

このグループ分けと語の代名詞としての用法との関係を考えると、まず A グループの「おのれ」は反射指示の用法を専らとする点で他の語と区別される。B グループの「よ」「ぼく」は一人称で、かつ書き手自身のみを指す単数用法を専らとする点で他の語と区別される。C グループの内、「われ」を除いた「ごじん」「ごはい」「わがはい」「ごせい」「よはい」は

一人称で、かつ書き手だけでなく書き手を含めた複数の人を指す複数用法を取り得る点で他の語と区別される。本コーパスでの用例を見ると、「われ」を除く C グループの中で最も度数の多い「ごじん」は専ら複数用法をとり、「ごはい」「わがはい」「ごせい」「よはい」も複数用法が認められる。

このように、後続助詞との対応関係に基づく語のグループは、代名詞としての用法に基づく語の分類とほぼ一致することがわかる。

なお「われ」は、代名詞としての用法から見ると、一人称・反射指示両方の用法を持つ点で他の語とは区別されるが、後続助詞という観点からは「ごじん」等と同じ C グループに属する結果となった。「われ」については別の観点によるさらなる分析が必要であると言える。

5．連体用法における語と被修飾体言との対応関係

次に 4．で分析の対象外とした「わが」について見てゆく。表 4 に示したように、「わが」は一人称・反射指示のどちらの用法でも連体用法をとることが多い。そこで、連体用法をとる「わが」および「の_体」「が_体」を伴う他の語について、被修飾体言との対応関係について検討し、語の間の違いについて見ていく。

表 5 は、各語が連体用法をとる場合の被修飾体言を示したものである。（　）内は各体言の度数を示す。体言の種類の多い場合は、代表的な体言のみを示し以下は省略した（「…」で表記）。また「如し」にかかるものもここに含めて示してある。

表 5　連体用法における被修飾体言

語	が_体	の_体
わが_一	国(195)、帝国(40)、人民(17)、大日本帝国(13)、国内・地球・政府(7)、邦人(6)、国民・民・国産・社(5)、日本帝国・日本・心(4)、アジア・法律・今上天皇陛下・性・東州・東方(3)…	
わが_反	国(5)、身・父(4)、為・同生同人・同人・日本・自由・父母・用・物品・子・本体(2)…	
よ	ロジック・考・言(3)、胸臆・頭脳・所見(2)…	喜び・憶説・論・意・如し(1)
われ_一		有・文章・障子ガラス・義務・民…(1)
われ_反		如し・三法・下・精神・国・父…(1)
おのれ	力(5)、為・身体(3)、用・自由・三宝・意・身・一身・利・鋭利・労(2)…	意(3)、欲・如し(2)、迷信・子・力・胸中・責・権利・国…(1)
ごじん		為・性・心裏(2)、進歩・生活・感覚・天性…(1)
ぼく	論(1)	
ごはい		雲仍(2)
わがはい		目(1)
ごせい		如し(1)
よはい		首唱・鄙見(1)

ここから、語と被修飾体言との関係を見てゆく。

まず「わが」については、特に一人称用法の「わが」は、被修飾体言が「わが」の「わ」にとっての「所属先」という関係になる場合が多いということが言える。「わが」以外の語

では、被修飾体言は各語にとっての「所有物・所属物」という関係をとることが多いのとは対照的である。典型的なのは最も度数の多い「わが国」で、「「わ」の所属する国」の意となる。「わが帝国」「わが地球」「わが社」「わがアジア」「わが東州」等も同様である。さらに、体言が「所属先の所有物・所属物」、特に「所属する国の所有物・所属物」という関係になることもある。例えば「わが人民」とは「「わ」の所属する国に所属する人民」の意で用いられている（「「わ」の統治する人民」「「わ」の所有する人民」の意ではない）。「わが政府」「わが邦人」「わが国民」「わが民」「わが法律」等も同様の関係にある。このように、一人称用法の「わが」は、被修飾体言の関係が「所属先」「所属先の所有物・所属物」となる点で特徴付けられる。なお、反射指示用法の「わが」については、その被修飾体言が「所属先」「所属先の所有物・所属物」の関係となる割合は一人称用法のものほど高くなく、「身」「父」「自由」「物品」等の「所有物・所属物」の関係となる場合も比較的多くなっている。

「わが」以外の語は、先に述べたように、被修飾体言が「所有物・所属物」の関係になることが多い。その中で、「よ」は被修飾体言が「ロジック」「考」「言」といった「所有する考え・意見」を意味する語で多く占められる点で特徴付けられる。「ぼく」「よはい」も被修飾体言に「論」や「首唱」「鄙見」をとり、「よ」と同様の傾向があるものと見られる。

以上のように、連体用法において語と被修飾体言との間にはいくつかの対応関係が見いだされることがわかった。

６．主な語の特徴

以上の分析結果に基づき、主要な語についてそれぞれの特徴をまとめる。取り上げる語は文語記事での度数の多い上位 5 語「わが」「よ」「われ」「おのれ」「ごじん」である。

まず「わが」であるが、連体用法をとる主たる語であり、(1)(2)のように被修飾体言が「所属先」「所属先の所有物・所属物」の関係となる点で特徴的である。

(1) 夫レ我ガ國ノ文字先王始メ之ヲ漢土ニ取テ之ヲ用ウ（1 号「洋字ヲ以テ国語ヲ書スルノ論」西周）[5]

(2) 目今諸省ニ於テ許多ノ洋人ヲ雇テ其學術ヲ傳取スル如ク彼尤善尤新ノ法教師ヲ雇テ公然我人民ヲ教導セシメバ奈何（3 号「開化ヲ進ル方法ヲ論ズ」津田真道）

次に、「よ」であるが、一人称単数の用法をとる主たる語である。述語に対し動作主や経験者といった意味的役割を担い、(3)のような著者の個人的な体験を語る文脈でも用いられるが、論説文という文章の性質上、(4)(5)のように著者の意見や主張を述べる文脈で用いられることが多い。連体用法をとる場合も同様で、(6)のように著者の意見や主張を意味する語が被修飾体言となる。著者個人を指し示す語ゆえに、卑下の感情を伴う助詞「が」のほうが「の」よりも後続しやすい。

(3) 余會テ歐洲ニ遊テ煉火石造ノ家屋ヲ見ル（4 号「煉火石造ノ説」西周）

[5] 本文の引用に際しては、末尾の（　）内に号数・記事題名・著者名を示す。

(4) 故ニ余敢テ謂フ我邦人倫ノ大本未ダ立ズト（8 号「妻妾論ノ一」森有礼）

(5) 余ハ思フニ政府ハ猶精神ノ如ク人民ハ猶躰骸ノ如クナリ（2 号「学者職分論ノ評」津田真道）

(6) 余ガ考ニハ狗ヲ連ルヨリモ兎ヲ輸入シテ錢ヲ取ラルヽ方遙ニ恐ル可シト思フ位ノコトナリ（26 号「内地旅行西先生ノ説ヲ駁ス」福沢諭吉）

次に「おのれ」であるが、反射指示用法をとる主たる語である。(7)(8)のように連体用法をとることが多く、述語に対して動作主・経験者といった意味的役割を担うことは少ない。

(7) 是皆個々人々日夜孜々汲々己ガ勞ヲ厭ハズ己ガ力ヲ盡シテ之ヲ求ムベキ者ニシテ（38 号「人世三寶説（一）」西周）

(8) 今日ニ至リテハ諸邦ノ君主タトヒ聰明衆ニ超タリトモ己ノ意ヲ以テ命令ヲ下スコトナシ（12 号「西学一斑（前号ノ続）」中村正直）

次に「ごじん」であるが、一人称複数の用法をとる主たる語である。著者の個人的な意見について述べる文脈で用いられやすい「よ」とは異なり、(9)(10)のように、より一般性のある説や論を述べる文脈に用いられることが多い。また、著者自身のみならず他の人も含めて指し示す語ゆえに卑下の意味を伴う助詞「が」が後続することはない。

(9) 想像ハ瞑目思想ノ間吾人覿見スル所ノ形象事歴ニシテ頗ル蜃氣樓ト相類似ス（13 号「想像論」津田真道）

(10)若夫レ吾人ノ性中情欲ヲ缺ク時ハ人類何ニ由テ生々蕃植スルコトヲ得ンヤ（34 号「情欲論」津田真道）

最後に「われ」であるが、一人称・反射指示の両用法をとる語であり、一人称用法の「われ」は「よ」「ごじん」との違いを、反射指示用法の「われ」は「おのれ」との違いを明らかにしたいところである。

一人称用法の「われ」は、後続助詞との対応関係から「ごじん」と同じグループに属し、さらに(11)のように複数用法と思われる用例が見いだされ点でも「ごじん」と共通する。

(11)米利ノ戰艦一旦江戸海ニ侵入シ請求スルニ通信ノ約ヲ以ス是ニ於テ我之ヲ託シ始テ彼ニ日本來往ノ便ヲ得シム（7 号「独立国権義」森有礼）

反射指示用法の「われ」は、(12)のように述語に対し動作主や経験者といった意味的役割を担うことが少なくなく、その点が「おのれ」とは異なる。

(12)自由ヲ伸シ羈絆ヲ脱シ租税ハ吾之ヲ増減スベシ官吏ハ吾之ヲ進退スベシ是人民ノ利ナリ（39 号「政府与人民異利害論（六月一日演説）」西村茂樹）

本稿の分析からは、このような「われ」と他の語との類似点・相違点が指摘できるが、ではさらに進んで「ごじん」との違いはどのような点にあるのか、「おのれ」との違いは何によってもたらされるのかといったことについては、明らかにはできなかった。今後の課題としたい。

７．おわりに

以上、『明六雑誌コーパス』を用いて分析を行い、当時の文語論説文における一人称代名

詞（および反射指示代名詞）について実態の解明を試みた。語と後続助詞、語と被修飾体言との間には明らかな対応関係が見いだされ、それにより一部ではあるが各語の特徴が明らかになった。今後は、別の観点からの分析を加え、一人称代名詞間の違いについてより詳細に考察したい。また、他のコーパスを用いて分析を行い、近代の一人称代名詞の通時的変遷についても考察する予定である。

文献

石川慎一郎、前田忠彦、山崎誠（編）(2010)『言語研究のための統計入門』、くろしお出版

小椋秀樹、小磯花絵、冨士池優美、宮内佐夜香、小西光、原裕（2011)『特定領域研究「日本語コーパス」平成 22 年度研究成果報告書　『現代日本語書き言葉均衡コーパス』形態論情報規程集　第 4 版（上）(下)』

近藤明日子、小木曽智信、加藤文明子（2010)「『高等小学読本』の形態論情報付きコーパス」、情報処理学会シンポジウムシリーズ Vol.2010, No.15 人文科学とコンピュータシンポジウム論文集 人文工学の可能性～異分野融合による「実質化」の方法～、pp.189-194

関連 URL

近代文語 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic

近代書き言葉における文語助動詞から口語助動詞への推移
―『太陽コーパス』の形態素解析データによる―

田中 牧郎（国立国語研究所言語資源研究系）[1]

１．はじめに

『太陽コーパス』に形態素解析を施して形態論情報付きコーパスにしていくことの有用性は、本報告書の語彙の部に収録した論文[2]でも述べたが、『太陽コーパス』の形態素解析済データは、語法・文法・文体の研究の領域にも新しい展開をもたらすことが期待できる。本稿では、そのデータを用いて、言文一致によって書き言葉が文語法から口語法に変わっていく過程の記述を試みたい。

近代日本語の書き言葉は、明治時代後半（20 世紀初頭）に進んだ言文一致により確立した。明治時代前半（19 世紀末）の書き言葉と大正時代（1910～25 年ごろ）の書き言葉を比べると、大きく異なっていることが明らかで、この時代が文体史上の画期であったことは間違いない。しかしながら、この文体の大きな変化がどのように進んだのかについては、文法や語法の記述に即して十分に明らかにされているわけではない。言文一致運動を展開した作家や啓蒙家などによる文体改革の歴史や、「だ体」「である体」「ですます体」など文末表現に注目した文体類型の消長については、多くの研究があり（山本 1965・1971・1981、木坂 1976、森岡 1991、飛田 2004 など）、それらは言文一致現象の重要な一面を照らし出しているが、文法・語法の全体を見わたすと、不明な部分が多く残されている。書き言葉の文体変化の研究には、時間軸に沿った通時的なコーパスを用いることが、きわめて有益であると考えられる。

２．『太陽コーパス』における文語体と口語体

言文一致が進んだ明治後期から大正期の書き言葉を対象としたコーパスに、『太陽コーパス』（国立国語研究所 2005a）がある。このコーパスは、約 1,450 万字（700 万語程度）の規模の、総合雑誌『太陽』一資料だけを対象としたものではあるが、この雑誌が、ジャンル、著者、読者層の諸側面でかなり広い範囲をカバーできていることから、この時期の書き言葉の実態をかなりの程度反映していると見て、コーパス化を行ったものである（国立国語研究所 2005b）。

図１は、『太陽コーパス』に含まれる 5 年分の雑誌記事約 3,400 本について、文語体の記事と口語体の記事との数をまとめたものである（田中（2005）の表 10 に基づき作図）。文体の識別は、指標とする文末辞（「なり」「たり」は文語体、「だ」「です」は口語体など）を定め、各記事ごとに中心を占める文末辞が何であるかによって行った。最初の年次である 1895（明治 28）年では、文語記事が約 95%を占めていたが、最後の年次である 1925（大正 14）年では、反対に口語記事が 95%近くに達しており、文語体から口語体へという推移をこのコーパスによって調べることができる。

---

[1] mtanaka@ninjal.ac.jp

[2] 田中牧郎「明治後期から大正期の語彙のレベルと語種―『太陽コーパス』の形態素解析データによる―」（本報告書所収）

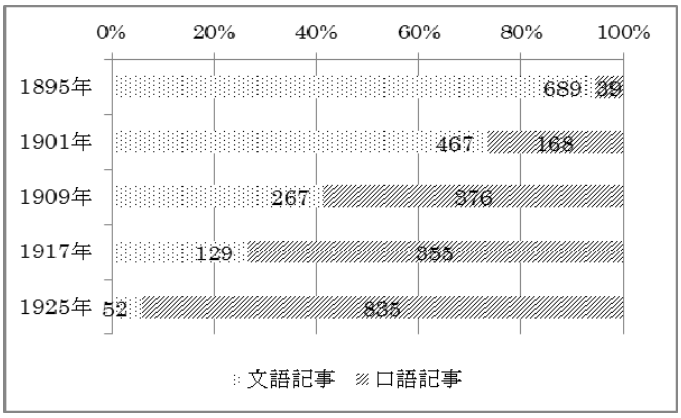


図１　『太陽コーパス』の文語記事・口語記事

『太陽コーパス』は、XML 形式によって、記事単位で本文を構造化し、加えて記事本文から引用部分を切り出す構造化を施し、記事のジャンル・著者・文体、引用の話者・文体などを、XML タグに属性として書き入れてある（田中 2005）。この仕様を生かして利用することで、文語記事中の口語引用文や、口語記事中の文語引用文を区別してデータを集計していくことも可能になる。一方、『太陽コーパス』は、『現代日本語書き言葉均衡コーパス』などと違って、形態素解析は施されていないため、単語に区切ったデータに基づく研究は行いにくい。このため、文語法や口語法の形式の用例を収集したり、その頻度を把握したりすることは容易でなく、文語法から口語法への推移の研究は行いにくかった。これは、『太陽コーパス』開発当時は、近代語テキストに対する形態素解析技術がなかったためであるが、その後、現代語に対する形態素解析辞書「UniDic」を近代語に適用できるようにした「近代文語 UniDic」の開発が進んできており（小木曽 2009）、これを用いて『太陽コーパス』の形態素解析を行うことが可能になりつつある。現段階では口語体の部分では誤解析がかなり多くなってしまうこと、文語体の部分も語や表記によっては正しく解析できない場合が少なくないなど、問題も残されている。『太陽コーパス』を形態論情報付きコーパスにしていくには、もうしばらく研究期間が必要である。

３．各年次５万レコードの調査

前節までで述べたように、『太陽コーパス』は、文語体から口語体への書き言葉の推移を具体的に記述するのに格好の資料であるが、現状では信頼できる形態論情報が取得できないために、文語法の形式と口語法の形式を数え上げるような総合的な調査には、そのままでは利用できない。そこで、本稿では「近代文語 UniDic」で自動形態素解析を行った後に、一定量について人手で誤解析を修正し、修正した範囲のみを対象に調査を実施することにした[3]。

調査データ作成の具体的な作業は、次のような手順で行った。

(1)『太陽コーパス』各年第１号の全記事に対して、「近代文語 UniDic」と MeCab を用いて自動形態素解析を実施。

(2)上記の解析結果データから、各号の冒頭５万レコードをサンプルとして抽出。ただし、1895 年第 1 号は当該部分に口語記事が多く、1901 年第 1 号は当該部分に口語記事が皆無であり、ともに『太陽コーパス』の全体的傾向と異なるため、一部のサンプルを記事ごと入れ替え、『太陽コーパス』の年次ごとの文体のバランスから大きくずれることが

[3] この態度は、本報告書に掲載した論文、田中牧郎「明治後期から大正期の語彙のレベルと語種―『太陽コーパス』の形態素解析データによる―」で、『太陽コーパス』全体の形態素解析データを対象とした態度と異なっている。この相違は、比較的誤解析が少ない自立語を対象としているか、誤解析が多い付属語を対象としているかの違いに基づいている。

表1　各年次5万レコードの調査対象の記事

年	号	記事	著者	欄	文体	ジャンル	文字数	備考
1895	1	〈扉〉	*	*	文語	***	737	
1895	1	太陽の発刊	大橋新太郎	*	文語	NDC051	3078	
1895	1	学界の大革新	久米邦武	論説	文語	NDC002	8093	
1895	1	戦勝後の教育	千頭清臣	論説	文語	NDC371	5865	
1895	1	戦争と文学	坪内逍遥	論説	文語	NDC901	7284	
1895	1	漢字の利害	三宅雪嶺	論説	文語	NDC811	5266	
1895	1	国語研究に就て	上田万年	論説	口語	NDC810	7200	
1895	1	事物変遷の研究に対する人類学的方法	坪井正五郎	論説	口語	NDC469	2567	
1895	1	経済的闘争	井上辰九郎	論説	文語	NDC333	5117	
1895	1	農業教育に就きて	横井時敬	論説	文語	NDC610	4593	
1895	1	対清政策	尾崎行雄	論説	文語	NDC329	8218	
1895	1	日本帝国の任務	中西牛郎	論説	文語	NDC311	4362	
1895	1	京都の新案内記	中川四明	地理	文語	NDC291	7354	
1895	1	紀元前の著名なる航海者	森田思軒	史伝	文語	NDC209	4458	途中
1901	1	〈扉〉	*	*	文語	***	149	
1901	1	明治三十四年	*	太陽	文語	NDC302	2171	
1901	1	昨冬の露帝	有賀長雄	論説	文語	NDC288	3509	
1901	1	学政振張と財政	久保田譲	論説	文語	NDC373	5767	
1901	1	韓国移民論	加藤増雄	論説	文語	NDC334	3198	
1901	1	欧州農業界の大勢を論じ延きて我国農業の前途に及ぶ	横井時敬	論説	文語	NDC612	5798	
1901	1	文明批評家としての文学者（本邦文壇の側面評）	高山樗牛	論説	文語	NDC901	9757	
1901	1	永遠平和の基礎を樹つるの道	国府犀東	論説	文語	NDC319	4977	
1901	1	大学派の政治的系統	*	人物月旦	文語	NDC377	5880	
1901	1	文芸時評	大町桂月	文芸時評	文語	NDC904	13782	
1901	1	政治時評	国府犀東	政治時評	文語	NDC312	9001	
1901	1	社会の腐敗救治意見	*（筆記）;清浦奎吾（談	名家談叢	口語	NDC154	4162	
1901	1	社会の腐敗救治意見	岡部長職	名家談叢	口語	NDC154	1620	
1901	1	社会の腐敗救治意見	石黒忠悳	名家談叢	口語	NDC154	1361	
1901	1	社会の腐敗救治意見	久保田譲	名家談叢	口語	NDC154	1658	
1901	1	社会の腐敗救治意見	戸水寛人	名家談叢	口語	NDC154	3404	
1901	1	宗教時評	龍山学人	宗教時評	文語	NDC162	5183	途中
1909	1	政治家の分類	*	時事評論	文語	NDC312	10387	
1909	1	大流小流	*	時事評論	口語	NDC329	1888	
1909	1	小是非	*	時事評論	文語	NDC304	1104	
1909	1	大谷光瑞法主	西湖生（筆記）;鳥谷部春	人物月旦	口語	NDC188	7006	
1909	1	新刑法に就て	鵜沢総明	論説	文語	NDC326	11075	
1909	1	大同派の威嚇	*	*	文語	NDC312	204	
1909	1	清国多難の秋	竹越三叉	論説	口語	NDC302	15281	
1909	1	列国外交機関と我外務省	望月小太郎	論説	文語	NDC319	5523	
1909	1	社会の変遷と信仰問題	姉崎嘲風	論説	口語	NDC316	11992	
1909	1	英米綽名の起原	*	*	口語	NDC832	295	
1909	1	清国の真相　清国の革命党	犬養毅（談）	論説	口語	NDC312	4669	
1909	1	清国の真相　支那政治家と支那国民	高田早苗（談）	論説	口語	NDC312	2588	
1909	1	清国の真相　清国の陸海軍	大原武慶（談）	論説	口語	NDC392	2751	途中

1917	1	挙国一致の外政策	浮田和民	*	口語	NDC319	9244	
1917	1	講和乎恒久戦乎	浅田江村	*	口語	NDC329	6636	
1917	1	海軍更迭短評	*	*	口語	NDC397	1910	
1917	1	政界の表裏　内大臣問題―新大臣月旦	無名隠士	無名隠士夜話	口語	NDC312	10191	
1917	1	法曹漫語	日東	*	口語	NDC327	2186	
1917	1	恋愛の破産時代	内田魯庵	案頭三尺	口語	NDC152	10965	
1917	1	時事俳句　その日その日	渡部霞亭	*	文語	NDC911	535	
1917	1	心頭雑草	与謝野晶子	婦人界評論	口語	NDC914	7677	
1917	1	一九一七年の国際経済	堀江帰一	経済財政時論	文語	NDC333	8000	
1917	1	戦時欧米産業界の活動	記者(文責);鶴見左右雄	*	口語	NDC333	10481	
1917	1	正貨と我が財政経済	神戸正雄	*	口語	NDC337	9421	途中
1925	1	昨年の今月	*	*	文語	NDC302	688	
1925	1	近代文明と発明	阪谷芳郎	*	口語	NDC507	7465	
1925	1	近代兵器の進歩並に将来の趨勢	大橋願四郎	*	口語	NDC559	8551	
1925	1	鼻で見、指で聞く少女	牧田環	*	口語	NDC147	4113	
1925	1	歴代の総理大臣(一)	三宅雪嶺	*	口語	NDC312	2643	
1925	1	現代の女性美	斎藤佳三	*	口語	NDC701	3428	
1925	1	政界煙話	鬼谷庵	*	口語	NDC312	4696	
1925	1	阪神船成金の今昔	乱峰子	*	口語	NDC332	2737	
1925	1	最近に於ける飛行機の発達	長岡外史	*	口語	NDC538	9060	
1925	1	官界から実業界に入りて	白仁武	*	口語	NDC335	2094	
1925	1	近世畸人伝　老鉄と鬼助	村松梢風	*	口語	NDC289	4065	
1925	1	貸金庫とはドンなものか　東京では日本興銀と三菱	記者	*	口語	NDC338	3553	
1925	1	生活上に於ける差別撤廃論	三輪田元道	*	口語	NDC361	10622	
1925	1	明治初年外交物語(その四)八太郎の虎の巻	豹子頭	*	口語	NDC210	7695	途中

＊「*」は記載がないもの。「＊＊＊」は分類不能。備考欄の「途中」は、各年次5万レコードに達したところまで。

ないよう調整した。サンプルとした記事は、表1の通りである。なお、今回のサンプルには結果的に小説類は一つも含まれなかった。

(3)サンプルとして取り出した、各年5万レコード（全体で25万レコード）について、人手で誤解析を修正。この修正によってレコードが増減することがあり、また、自動解析結果のデータには、記号や空白も1レコードとして出力されるので、実際は各年次5万語よりも少なくなる。

(4)上記の25万語弱について、同語異語判別を実施した後、品詞が助動詞と認定されたデータをもとに、年次別の助動詞頻度表を作成。

4．助動詞の頻度

4．1　助動詞全体の概観

以上の手順で作成した助動詞頻度表が、次頁の表2である。表2に至る前の段階の処理作業にあたっては、『現代日本語書き言葉均衡コーパス』の形態論情報規程集（小椋ほか2011）を参照した。この基準は、形容動詞を立てないため「勤勉だ」の「だ」などは、助動詞「だ」に扱う。また、推量の助動詞「う」はなく、活用語の意志推量形と認定される。なお、例えば「らしい」と「らし」のように口語助動詞と文語助動詞とが用例上区別しにくいものについては、筆者の判断でどちらかにまとめた（「らしい」「らし」の場合は「らしい」にまとめた）。また、断定の助動詞が「なら」の語形を取った場合は、「だ」の仮定形とはせずに、すべて「なり」の未然形と扱った。

表2を概観すると、●を付けた口語助動詞の多くは、後ろの年次で新たに出現したり、

年次が進むにしたがって増加したりする傾向を示すものが多い。一方、〇を付けた文語助動詞は、後ろの年次では姿を消したり、年次とともに減少したりするものが目立つ。全体的に見れば、文語体から口語体へという書き言葉の変化が、助動詞の消長に現れていると見ることができる。一方で、口語助動詞のすべてが同じように登場したり増加したりするわけではなく、また、文語助動詞の消失や減少の時期やペースも語によって多様である。このことから、文語体から口語体への移行にあたっては、個々の助動詞において様々な事情があったことが推測され、そのような細部に分け入った記述研究が必要とされると言えるだろう。

表２　『太陽コーパス』形態素解析データ各年次５万レコードにおける助動詞の頻度

意味	語彙素	1895(明 28)年	1901(明 34)年	1909(明 42)年	1917(大 6)年	1925(大 14)年	計
断定	●じゃ				21	1	22
	●だ	171	130	737	1019	1385	3442
	〇たり-断定	154	148	132	104	42	580
	〇なり-断定	930	1226	684	585	465	3890
丁寧	●です	3	6	1	52	8	70
	●ます	183	63		76	60	382
	●やんす	1					1
過去･完了	〇き	148	243	59	26	14	490
	〇けり	15	5		2		22
	●た	50	48	499	702	926	2225
	〇たり-完了	242	219	194	70	11	736
	〇つ	4	2	1		2	9
	●てる			1	10	11	22
	〇ぬ	22	30	9	2	1	64
	〇り	236	248	133	78	59	754
推量	〇なり-推定	2					2
	〇べし	415	406	255	167	48	1291
	〇む	258	261	106	97	30	752
	〇めり	1					1
	らしい		1	1	4	4	10
	〇らむ	1			1		2
否定	〇非ず		5				5
	〇じ	5	11				16
	〇ず	860	883	501	428	193	2865
	●ない	5	14	71	150	167	407
	●まい		3	9	23	12	47
	〇まじ	1	2	1	1	0	5
受身	られる	71	127	138	106	66	508
	れる	92	47	68	171	187	565
使役	させる			1	4	1	6
	〇しめる	104	89	73	60	27	353
	せる	9	13	6	25	26	79
比況	〇ごとし	151	204	189	146	39	729
希望	●たい	5	5	3	11	53	77
	●たがる		1			2	3

●は口語助動詞、〇は文語助動詞、●も〇もないものは口語・文語に共通する助動詞。

４．２ 断定の助動詞、過去･完了の助動詞の消長

細部を研究するための第一段階として、表２に示した助動詞のうち、文語助動詞から口語助動詞への交替の様相が明瞭に現れている、断定の助動詞、過去・完了の助動詞を取り上げてみよう。まず、頻度の低い（総頻度 30 未満）助動詞について、簡単に見ておこう。まず「じゃ」は、1909 年まで皆無で 1917 年にはじめて出現する。「てる」も、同じように前半の年次には無く 1909 から現れる。これらの口語助動詞は、口語記事が多くを占めるようになったことによって書き言葉に使われるようになったものである。文語助動詞のうち「けり」「つ」は総頻度が低いが、どちらかというと前半の年次に多く、後半の年次に少ない。これらは、文語記事が減少するにつれて、姿を消していく流れにあったものと考えられる。

断定の助動詞、過去・完了の助動詞それぞれの主要なもの（総頻度 30 以上の語）の年次別の頻度の推移をグラフにまとめたのが、図２および図３である。

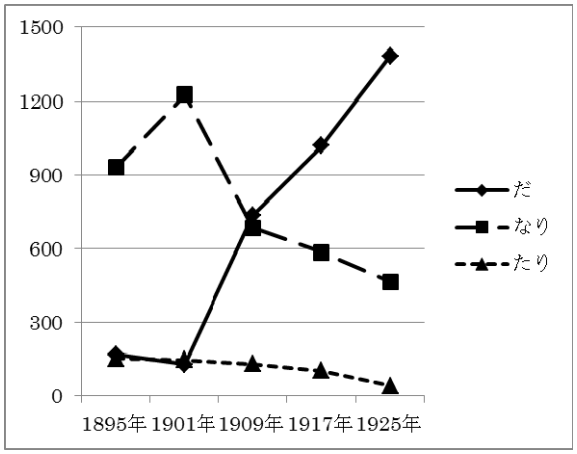


図２　主要な断定の助動詞の頻度

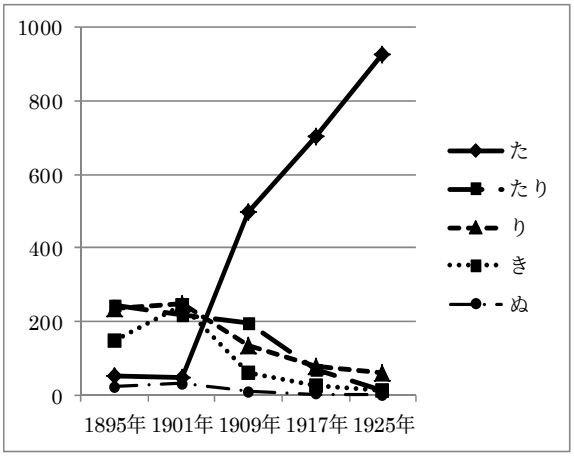


図３　主要な過去・完了の助動詞の頻度

断定の助動詞については、図２から次のようなことが読み取れる。まず、「だ」が 1901 年から 1925 年まで急速に増加し、反対に「なり」「たり」は 1901 年から 1925 年まで減少が続く。1895 年と 1901 年の間では「だ」はわずかに減少、「なり」はわずかに増加、「たり」はほとんど変化がない。1901 年以降、文語助動詞「なり」「たり」から口語助動詞「だ」への交替が進むが、口語助動詞の増加の速度に比較して、文語助動詞の減少の速度は緩やかであり、1925 年においても、「なり」は高い頻度を保っている。

次に、過去・完了の助動詞については、図３から次のようなことが読み取れる。「た」は 1901 年以後急増し、反対に「たり」「り」「き」は 1901 年以後減少傾向が続く。1895 年から 1901 年へは、「た」「り」「ぬ」は不変、「き」は微増、「たり」は微減という状況である。断定の助動詞と同じように、1901 年以降、文語助動詞から口語助動詞への交替が進むが、口語助動詞の増加の勢いに比べて文語助動詞の減少の傾きは緩やかで、「り」のように 1925 年に至っても、ある程度使われ続けるものもある。

このように、口語助動詞の増加の速度に比べて文語助動詞の減少の速度は緩やかであることは、断定の助動詞と過去・完了の助動詞とで共通している。また、断定の助動詞、過去・完了の助動詞の内部においては、個々の助動詞ごとに増え方・減り方は様々であり、こうした個別の変化を詳しく見ていく必要性が高いことが確かめられる。以下、本稿では、断定の助動詞を事例に取り上げて、さらに詳しい分析を行うことにする。

５．断定の助動詞の分析

５．１　「だ」の頻度推移

まず、「だ」の発達過程を見ていこう。図４は、「だ」の活用形ごとの年次別の頻度の変

化が分かるように、折れ線グラフに示したものである。

すべての年次で連用形「で」の頻度が最も高く、その 1901 年以降の増加傾向も著しい。「で」の後に直接続く語について、やはり年次別の頻度をグラフに示すと図５のようになる。他に「ね」「の」に続くものや、形容詞に続くものが数例あるが、グラフでは省略した。どの年次においても「ある、ござる」が最も多く、その増加傾向も顕著である。連用形「で」の増加は「である」の伸張による面が強いことが分かる[4]。「だ」の全体頻度が増加する 1909 年からは、連体形「な」と終止形「だ」の増加も目立ってくる。少数だが、未然形「だろ」[5]、連用形「だっ」も、1909 年あるいは 1917 年から見え始める。

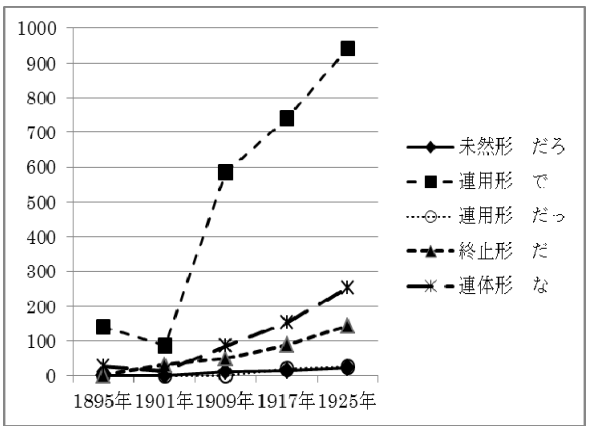


図４　「だ」の活用形別頻度

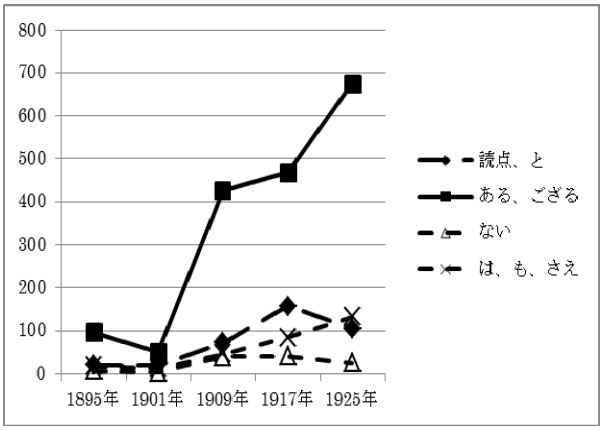


図５　連用形「で」の後接形式

大決心を有つ者でありまする。（1895 年 1 号・上田万年「国語研究に就て」）
平和の破壊であつて平和の手段でない（1917 年 1 号・浅田江村「講和乎恒久戦乎」）
或意味に於ては帝國海軍の慶事で、（1917 年 1 号・＊「海軍更迭短評」）
斯様な不都合な外交を遣るものは（1909 年 1 号・＊「大流小流」）
人生を觀じて樂觀だの悲觀だのとさわぐ（1909 年 1 号・姉崎嘲風「社会の変遷と信仰問題」）
ツマリ富谷君に關係があるのだ、（1917 年 1 号・日東「法曹漫語」）
政治的に餓死せねばならぬ時であるだらうと思ふ。（1909 年 1 号・竹越三叉「清国多難の秋」）
巳代治は桂内閣の軍師だつたが、（1917 年 1 号・無名隠士「政界の表裏 内大臣問題」）

５．２　文語文に姿を現す「だ」

上記のような「だ」の伸張は、口語記事が増加していくことによるものであり、「だ」のほとんどは口語文中で用いられている。一方、わずかではあるが、次のように文語文中に顔を出した「だ」がある。今回の調査範囲では 6 件のみである。

瑣細な質問の蒼蠅（うるさき）を嫌ひ、因て學者の群に入ることを避たり。（1895 年 1 号・久米邦武「学会の大革新」）
普通の秀才位では、到底こゝに達する能はず（1901 年 1 号・大町桂月「文芸時評」）

６件中４件が、第１例のように連体形「な」であり、多くの年次の多様な記事に点在している。全体では少数派の連体形「な」の場合に文語文中でも口語助動詞が現れやすいの

[4] 「ある、ござる」のうち、「ござる」は 1895 年にのみ多く、1901 年以後は非常に少ない。言文一致初期に多かった「でござる体」「でござります体」が次第に減少していく流れを反映するものだと考えられる。
[5] 「近代文語 UniDic」では、「だろう」で意志推量形と判定されるが、ここでは「だろ」で未然形と扱った。

は、連体形が持つ後続の名詞との連続性の強さからそこに切れが感じられなくなり、文体が意識されなくなったからだというような事情が考えられるのではないか。残り 2 件は、連用形「で」で、いずれも 1901 年の「文芸時評」（大町桂月）の例である。これは現れる記事に限定があるため、記事の文章の事情によるものかもしれない。このようにごく一部に例外的な現象があるものの、原則として、文語文中には口語助動詞は現れないと見ることが許されよう。

５．３　「なり」「たり」の頻度推移

図６は、「なり」の活用形ごとの年次別頻度の推移が分かるようにグラフに示したものである。当初 1895 年及び 1901 年では、終止形「なり」、連用形「に」、連体形「なる」が多く、未然形「なら」、已然形「なれ」、連用形「なり」は少なかった。1909 年からは、終止形「なり」、連体形「なる」が減少し、特に終止形「なり」は急速に衰退し、1925 年ではわずかになる。一方、連用形「に」はほぼ横ばいで、1925 年でも 300 件以上使われていることや、未然形「なら」も少ないながらほぼ横ばいであることは、衰退していく全体的な流れと異なる動きとして注目される。

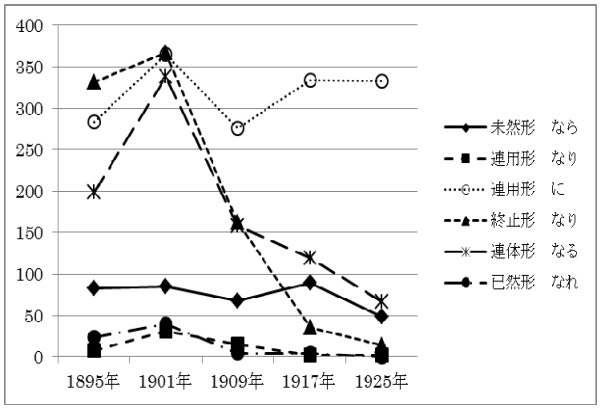


図６　「なり」の活用形別頻度

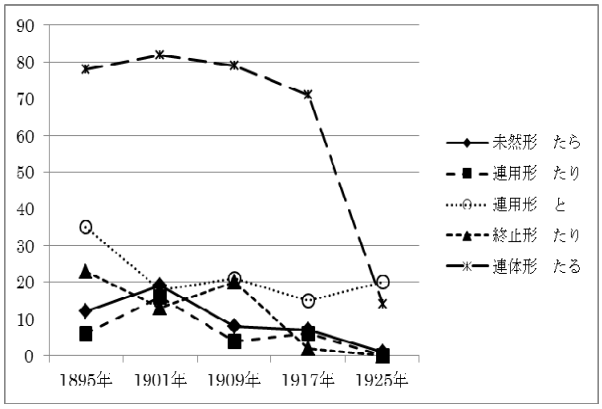


図７　「たり」の活用形別頻度

「たり」について同様にグラフにしたものが図７である。「なり」に比べて全般に頻度は低いが、連体形「たる」だけが 1917 年までよく使われ、他の活用形は当初から低頻度でいずれも次第に減少していく。連体形「たる」が 1925 年に急速に減少すること、連用形「と」が新しい年次でも比較的多く使われ続けることは、全体的な流れと異なる動きとなっている。

５．４　口語文に生きる「なり」「たり」

5.2 で見たように、文語文中に口語助動詞「だ」は原則として現れず、例外的に顔を見せた場合は何らかの特別な事情があるものであった。これに対して、口語文中に文語助動詞「なり」「たり」が現れることは多く、表３のように「なり」約 1400 件、「たり」200 件が確認できる。

表３　口語文中の「なり」「たり」の活用形別頻度

	未然	連用形-一般	連用形-に･と	終止	連体	已然形	計
なり	187(13.5%)	5(0.4%)	882(63.5%)	48(3.5%)	256(18.4%)	11(0.8%)	1389(100.1%)
たり	15(7.5%)	6(3.0%)	49(24.5%)	8(4.0%)	122(61.0%)	0(0.0%)	200(100.0%)

口語文中の「なり」のうち、63.5%が連用形「に」、次いで 18.4%が連体形「なる」、13.5%が未然形「なら」であり、他の活用形は少ない。

元老を攻撃するなどは餘りに下品ぢや。（1917 年 1 号・無名隠士「政界の表裏 内大臣問題」）

彼様なる別が有るものでござります（1895 年 1 号・坪井正五郎「事物変遷の研究に対する人類学的方法」）

其面目を一新すると云ふ意味に外ならぬのである。（1909 年 1 号・西湖生「大谷光瑞法主」）

若し是が女王でなかつたならば、（1909 年 1 号・竹越三叉「清国多難の秋」）

これらのうち、連用形「に」と未然形「なら」は、口語助動詞「だ」に、これに直接相当する用法がない。このために、文体が口語体に変わっても文語法が生き続けることになったものだと考えられる。口語文法においては、形容動詞や助動詞「だ」の連用形に「に」を、同じく未然形・仮定形に「なら」を配して、これら生き続けた文語法を口語法の中に組み入れている。口語文に生き残った語法と見ることができよう。これに対して、連体形「なる」には、同じ用法が口語助動詞「な」によっても担われている。

斯様な不都合な外交を遣るものは（1909 年 1 号・＊「大流小流」）

図 6 において、連用形「に」、未然形「なら」と違って、連体形「なる」は衰退傾向が顕著であったのは、口語助動詞「だ」の連体形「な」に直接対応する用法があったために、口語文が増えていくにしたがって、口語法の「な」に置き換わっていったからだと考えられる。

口語文中の「たり」の内訳は、表 3 のように、連体形「たる」が約 61%を占め、次いで連用形「と」が 24.5%であり、他の活用形は数%以下に止まっている。連体形の方が連用形よりも多いところは、「なり」の場合と逆である。

或は慨然として長息し（1895 年 1 号・坪内逍遙「戦争と文学」）

或は漠然と數人を愛して（1917 年 1 号・内田魯庵「恋愛の破産時代」）

決して政治家たるを得じ（1901 年 1 号・＊「大学派の政治的系統」）

帝國海軍の主力たる第一艦隊を（1917 年 1 号・＊「海軍更迭短評」）

連体形「たる」には、「なる」に対応する「な」のような口語形式がない。次のような「である」という複合形式はあるものの、「たる」と「である」の対応は、「なる」と「な」のように等価な対応とは言えない。そのことが、図 7 で見たような、「たる」が 1917 年まで衰退することなく使われ続けたことの背景にあったのではないかと思われる。

法律學者且つ外交官であるヘンリー・ホイートン氏の（1925 年 1 号・豹子頭「明治初年外交物語」）

同じく、連用形「と」にも対応する口語形式がないために、図 7 で見たように、年次が進んでも頻度を低下させずに使われ続けるのだと考えられる。この点は、「なり」の連用形「に」と同様の事情である。

6．おわりに

近代日本語の書き言葉が、言文一致を経て文語体から口語体に推移する過程を、『太陽コーパス』から取り出したサンプルにおける助動詞の分析を通して研究した。その結果、全体的には、口語助動詞が増加し文語助動詞が減少していく状況にあることが確かめられたが、その増加や減少の時期や速度は語によって様々であることも明らかになった。断定の助動詞「だ」「なり」「たり」を例に、活用形や用法の細部を分析すると、活用形や用法によって、発展や衰退が顕著なものもあれば、それがあまり目立たないものもあった。文語文中に口語助動詞が現れることは原則としてないが、口語文中で文語助動詞が使われることは多く、それは、口語助動詞にない用法を担う役割を持って口語体書き言葉にも取り入れられたものと考えられた。他の助動詞や、助動詞以外にも研究対象を広げて、近代書き言葉における文語法から口語法への推移の過程で生じた出来事を総合的に研究していくことが期待されよう。

口語体書き言葉がどのように成立していったかについての研究には、当時の話し言葉と

の対比も不可欠である。今回のサンプルには含まれていなかったが、『太陽コーパス』には小説も多く入っており、小説中の会話文の実態を分析することで、そのような方向に研究を展開させていくことが望まれよう。

『太陽コーパス』に形態論情報を加えていくことで、語法・文法・文体の研究にも大いに資することができると見込まれる。

文　献

小木曽智信（2009）『近代文語文を対象とした形態素解析のための電子化辞書の作成とその活用』（科学研究費補助金研究成果報告書、http://www2.ninjal.ac.jp/lrc からダウンロード可）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011）『「現代日本語書き言葉均衡コーパス」形態論情報規程集 第 4 版』国立国語研究所内部報告書

木坂基（1976）『近代文章の成立に関する基礎的研究』風間書房

国立国語研究所（1987）『雑誌用語の変遷』（国立国語研究所報告 89、秀英出版）

国立国語研究所（2005a）『太陽コーパス―雑誌「太陽」日本語データベース』（国立国語研究所資料集 15、博文館新社）

国立国語研究所（2005b）『雑誌「太陽」による確立期現代語の研究―「太陽コーパス」研究論文集―』（国立国語研究所報告 122）、博文館新社

田中牧郎（2005）「言語資料としての雑誌『太陽』の考察と『太陽コーパス』の設計」国立国語研究所 2005b 所収、pp.1-48

飛田良文編（2004）『国語論究 11 言文一致運動』明治書院

森岡健二（1991）『近代語の成立 文体編』明治書院

山本正秀（1965）『近代文体発生の史的研究』岩波書店

山本正秀（1971）『言文一致の歴史論考』桜楓社

山本正秀（1981）『言文一致の歴史論考 続編』桜楓社

付　記

本稿は、「国立国語研究所『通時コーパス』プロジェクト・オックスフォード大学 VSARPJ プロジェクト合同シンポジウム『通時コーパスと日本語史研究』」（2012 年 7 月 31 日、国立国語研究所）において発表した内容に基づいている。

近代語に探る〈終止形準体法〉の萌芽的要素

島田 泰子　（二松学舎大学文学部）

１．　はじめに

島田（印刷中）において私に〈終止形準体法〉と名付けた表現様式について、その新奇性を分析するとともに、歴史的位置付けや表現成立の経緯・背景に関しても解明を試みようとするとき、関連する（問題となる）ものとして注目されるのが、以下の類型を備えた表現である。

・「活用語の終止形 または 連体形」 ＋ 「格助詞 ガ または ヲ」
・「形容詞 口語終止形」 ＋ 「係助詞 ハ または モ」

近代におけるこういった類型の実例を広く採集する作業は、形態素解析を経てタグ付けされたコーパスによって、飛躍的に効率化される。本稿では、コーパス利用ならではの用例収集とその応用研究の一例として、近代語における終止・連体形の準体法的な用法及びその類例を検討し、今日的な〈終止形準体法〉の歴史的背景を探りたい。

２．　〈終止形準体法〉について

島田（印刷中）において〈終止形準体法〉と名付けたものは、次のような表現の下線部例[1]である。文法的には、「活用語の終止形が、文中において体言に準ずる扱いを受け、格助詞ヲ・ガ・ニなどを伴って、述語に対する格成分として用いられたもの」と一旦記述できるものであるが、近年、特に広告表現を中心とした特殊な位相において目立つようになってきた。

(1) 一番はじめのうまいを引き出す　（キリン一番搾り cf ナレーション、2008.6.4 録画）
(2) かわいいが好き！　（米子空港売店 どじょう掬いまんじゅうコピー、2009.11.1 撮影）
(3) 薄いに、恋して。　（携帯電話 薄型機種 pst.、2007.7.29 撮影）
(4) 祝うを、素敵に。　（紳士服の AOKI 夏の礼服 cf コピー＆ナレーション、2012.6.20 録画）
(5) 実感。剃るから、すべらせるへ。（Gillette 紳士用剃刀、cf コピー　2012.6.23capt.）

ヲやガなどの格助詞を伴うことから、これらの形容詞・動詞の終止形は、構文的には極めて名詞的な用いられ方をしているといえる。ただし、いわゆる居体言のように安定的に体言としての用法を定着させたものではなく、広告におけるキャッチコピー、書物・展示・番組名のタイトルなど、位相的にはやや偏りを伴って多用される現状が観察される。体言化にきわめて近いが、あくまでそれに準ずる準体的な用法とするのが妥当と見られる。

---

1 出典注記について。TVcf については録画日、新聞の記事・広告などは掲載日、車内広告・看板・ポスター（pst.と略記）等は撮影日、インターネット上の用例はキャプチャ（capt.と略記）した日付を示した。

早いものとしては1980年代に類似した用例が複数確認されるが、いわゆるゼロ年代以降に急増したらしく、多用される中で違和感も薄まり一般化しつつあるように見受けられる。

(6)～(8)のような感嘆符や読点、終助詞を伴うものや、(9)のように引用符が用いられたものを考えあわせるに、ある種の引用的な用法と見なすことができる。

(6) ウレシイ！をカタチに。 （消費者金融・レイク pst.、2008.5.12 撮影）

(7) 家を買う、をギャンブルにしない。 （住宅情報会社 地下鉄車内 pst.、 2012.3.3 撮影）

(8) 列車を選ぶ人は、君のいいなを願う人です。（JR 東日本 pst.、2007.1.20 撮影）

(9) 看護師の“はたらく”を応援！ （転職支援サービス会社 車内 pst.、2010.7.5 撮影）

引用された語句が文中における格成分としての自在さを持つことについては先学の指摘するとおりであるが[2]、稿者がこれらをあえて「終止形による準体法」とみなし〈終止形準体法〉と名付けたのは、以下の２つの理由による。

１つめは、(10)(12)のように、従来ならば居体言が用いられる位置に代替的に置かれる傾向があり（引用符なしでの使用も多い）、これらを(11)(13)のようなものと対比した際に認められるある種の新奇性や脱規範性に注目したこと。

(10) 天然水の力でうまいをつくる。 （サントリー 缶ビールパッケージ、2006.5.26 撮影）

(11) 本物の「うまさ」を贈りたい。 （アサヒ缶ビール広告 pst.、2007.11.20 撮影）

(12) おいしいを、日本の畑から。 （東都生活協同組合 新聞折込チラシ 2008.4）

(13) おいしさを、笑顔に。 （キリン cf、TV 画面より 2008.6.4 撮影）

２つめは、現代語における活用語の終止形が古代語の連体形に由来することから、問題の表現が「弱きヲ助け、強きヲくじく」「故きヲ温ねて新しきヲ知る」（以上、形容詞の場合）、「負けるガ勝ち」「稼ぐニ追いつく貧乏なし」（以上、動詞の場合）のような文語調に由来するもの（連体形準体法の名残り）と見ることもでき、両者の連続性を等閑視しえないこと。

以上２つの理由により、用言体言化の消長という観点から一貫して扱おうとする場合に、単なる引用に留まらない「（広義における）準体法」として認定することには、一定の意味があると考えられる。旧来の連体形準体法と今日的な〈終止形準体法〉をつなぐものとして、近代における終止・連体同一形の準体的な用法について観察する必要が、ここに生じることになる。

なお、〈終止形準体法〉と関連する類似の表現として、格助詞（ガ・ヲなど）を伴うもの以外に、次のようなハを伴うものにも目配りが必要である。

---

2 例えば、山田孝雄「引用の語句はその文中に於いては体言と同等のものとして取り扱はるゝものなるが、その取扱は大体準体句に準ぜられる。《 中略 》引用の語句も亦主格、賓格、補格として用ゐられ、又往々連体格としても用ゐらるゝことあり。」（『日本文法学概論』 第五十六章 引用の語句）、藤田保幸「引用されたコトバとは、表現されるべき対象世界において所与のコトバが、実物表示されて、つまりは、類似的に模写・写像されて出てくるものであった。いわば、模写・写像されて再現された対象世界の一断片である。 《 中略 》 端的にいえば、文中にとり込まれ、一定の分布・一定の位置をとらされることで、相対的に品詞的役割を付与されるのだと見るのがよいだろう。」（藤田 2000 総論 p.58）など。

(14) 聞くは一時の恥、聞かぬは一生の恥　(ことわざ)

(15) なんとか閉館 30 分前に着いたはいいが、身分証を忘れて入館できなかった。(作例)

(16) カワイイはつくれる!!　(エッセンシャル cf 字幕、2009.2.18 録画)

(17) すっぱいは、ハッピーのもと。　(カンロ 果汁グミ 車内 pst.、2009.7.2 撮影)

(18) 懐かしいはうれしい　(創業天和元年 カステラ元祖松翁軒 発行『よむカステラ』2008 年第 14 号)

(19) 日本人は知っている。うまいは甘い。(キリン 新・生茶 pst.、2007.10.30 撮影)

(20) 黒いは、うまい。うまいは、黒べえ。黒べえは、黒い。(黒べえ pst.、2008.6.3 撮影)

(16)などは「カワイイをつくる」のようなヲ格による表現が、(17)はガ格表現が、それぞれ前提として存在するものであり、〈終止形準体法〉と無関係ではない。また、(14)(15)は、先に示した「負けるガ勝ち」同様、特段の新奇性を持たないが、今日では同形となった終止形と連体形とのあいだに截然たる区別の意識は持たれにくく、形容詞における(16)(17)などの使用をどの程度意識の上で支えるものであるか、興味が持たれる。さらに言えば、(18)(19)のように述語部分にも形容詞が来るものは、(20)のような循環的な用例を含めて、歴史的にどこまで遡りうるかについて、慎重な検討が必要である[3]。

よって以下では、これらの類も、先に〈終止形準体法〉とした表現様式とともに取り扱うこととし、その実際的な用例を採集する。

３．　コーパスを利用した用例採集

３．１　コーパス利用の利点・その１（検索の便宜）

コーパスを利用して用例採集を行う利点の第一として、まずは検索の便宜における効率性が挙げられる。たとえば、本プロジェクトの研究のためにメンバーで利用している近代語のコーパス[4]ならびに検索ツール「大納言」（内部公開中）を用いる場合、

検索キーとして、「品詞」を「助詞-格助詞%」または「助詞-副助詞%」

その前文脈について、「活用形」を「連体形%」または「終止形%」

とそれぞれ指定すれば、この掛け合わせによる「連体形＋格助詞」、「連体形＋副助詞」、「終止形＋格助詞」、「終止形＋副助詞」の４パターンのものが、各コーパスから、現実的な用例として網羅的に収集できる。

単なる本文の電子化にとどまらず形態素解析を経て品詞や活用形などの情報がタグ付けされたコーパスは、特定の語（単語、熟語など）を調査対象とするだけでなく、こういった文法的な類型に着目した研究を行う場合にも、たいへん有効である。

想定の範囲に限らない用例の博捜が可能となるのも、現実的な表現形の揺れを超えた「語彙素」へのヒットで遺漏の少ない収集が可能となるのも、形態素解析によりタグ付け済みであるコーパスならではの利点と言える。

---

3 シェイクスピア「マクベス」坪内逍遙 1935 年の訳に、「清美(きれい・きたない)は醜穢(きたない・きれい)、醜穢は清美」とある。

4 『太陽コーパス』『近代女性雑誌コーパス』などから成り、一部を青空文庫の本文から採用したもの。

３．２　コーパス利用の利点・その２（整理・分類、分析の便宜）

検索結果を分析のためにソートを掛けて並べ替える作業が効率的に行えるのもまた、コーパスならではの利点であろう。ここでは、上記のような検索方法で得られたデータを、以下の優先順位に従ってソートした。

１，キー

２，後文脈（キーの直後の文字列頭での並び替え）

３，前文脈（キーの直前、フィールド末尾の文字列からの逆転文字列での並び替え）

これにより、類似の表現形式を持つ用例群が一ヵ所に集まって表示されることになり、パターンが把握しやすいうえに、直接関連しないレコードも一網打尽に取り除ける。

具体例を挙げると、例えば「連体形＋格助詞ノ」に該当するレコードのうち、いわゆるノダ文相当の「～のである。」「～のであるが、」「～のでした」「～のですか」が一ヵ所に集まったのを、一括して削除する（下図）などである。

	前文脈	キー	後文脈	品詞	コーパス
4389	𠮟られずして、|又平和條約議定以前に列國の間に、|更に新なる案件を生ずる	の	、|極めて不利益なるを知りて、|故らに之を回避するにも由るべしとするも、|又	助詞-格助詞	近代語
4390	列國にして、|若し清國に向て、|提出の條件を快諾し、|之れが履行の實を擧ぐる	の	、|交換條件として、|各自占領地、|若くば占領報界を厳退すべしとの約束を協定	助詞-格助詞	近代語
4391	清國をして、|列國の情誼に感激して、|大に自ら開發の道を講ずるに至らしむる	の	、|大雅量ある を得ず。|永遠なる平和の基礎を、|東邦に期望する列國にして、|は	助詞-格助詞	近代語
4393	とす。|而して一方には全力を擧げて|『太陽』に注ぎ、|傍らに合併すべからざる	の	『少年世界』を興して此旭日光を放つの新天地に立たん事を力む。|唯愛ふる處	助詞-格助詞	近代語
4394	とす。|而して一方には全力を擧げて|『太陽』に注ぎ、|傍らに合併すべからざる	の	『少年世界』を興して此旭日光を放つの新天地に立たん事を力む。|唯愛ふる處	助詞-格助詞	近代語
4395	問わず互いに相結托して強大なる団体を作し、|もって国家法度の外に超立する	の	ありさまなり。|されば仏国人民にして宗教のためその国に叛きあえて隣国のス	助詞-格助詞	近代語
4396	うまねく知ることろならが、その原因とて他にあらず、学者にして幽窓に沈思する	の	いとまを得ざるがためぬなり。|　けだし意味深遠なる著書は読者の over もまた遠く	助詞-格助詞	近代語
4397	塗うてたちまち破裂すべきや いかんせん。|而してその破裂の勢は、これを蔽むる	の	いよいよ堅固にして、時日のいよいよ久しきその割合にしたがいて、いよいよ鬱	助詞-格助詞	近代語
4398	そのひとつふたつを言はむ。|某語あり、語原つまびらかならず、外國語ならむ	の	うたがひあり、或人、偶然ご|それは、何人か、西班牙語ならむといへることあり|と	助詞-格助詞	近代語
4411	李平氏なりといえども、その後政府に奉えて実地の政務に当たり、学説を弘むる	の	ことはまったく福沢、尺の両氏に譲りたるもののごとし。|されば第三期の終りに	助詞-格助詞	近代語
4413	皆双刀を帯したるは、法律の轍む可きものなくして人々自から一身を保護する	の	たぬなりしが、習用の久しき、太平の世に至ても尚この帯刀を廃せず、常に之を	助詞-格助詞	近代語
4414	すべからずと断定せしのみにして、毫も其理由を言はず。|素より他を論議する	の	ついでに此言を附加せしものなれば、二氏も冗長をさけて其理由を言はざりしも	助詞-格助詞	近代語
4416	つけて、高く、さしあげながら、かけまはりました。|けれども、まだ、きづかはしい	の	で、|「車をとめよ」|車をとめよ。」|と、蟷のつづくだけ、さけびました。|すると、横	助詞-格助詞	近代語
4417	つかつ、この世を送ってをるやもめがありました。|べつに、これといふ職業もない	の	で、鶏を飼ったり、薪をとったりして、それを、近くの町に、賣りに、出て、わづか	助詞-格助詞	近代語
4418	つけて、寄りつかないようにする。|この雛の體色を、學問上では、警戒色といふ	の	である。|　第十八課　虎。|虎は、いんどに、もっともを多く、すめる猛獣にして、身	助詞-格助詞	近代語
4419	て、出ることを見合せ、また、航海してゐる船は、早く、港へはいって、難を避ける	の	である。|げんに、この警報があることになってから、船のこはれたり、沈んだり	助詞-格助詞	近代語
4420	と、形をかへるものである。|昆虫の変態といふのは、この、形をかへることをいふ	の	である。|たとへていへば、自いちようちょは、その卵がかへると、まづ、いもむし	助詞-格助詞	近代語
4421	らそふこともできるのである。|さて、この雛の體色を、學問上では、保護色といふ	の	であるが、なかには、周圍の色のかはるにしたがって、この保護色のかはるもの	助詞-格助詞	近代語
4422	すものもある。|すべて、昆虫は、みな、卵から、かへって、親と同じ形の虫となる	の	であるが、今くは、その間に、二三度、形をかへるものである。|昆虫の変態とい	助詞-格助詞	近代語
4423	』|『おとうさん。|私です。|今、うちから来たのです。|おかあさんがおいでなされる	の	でしたけれど、弟が病気で、手がはなせない から、私が来たのです。|おとうさん	助詞-格助詞	近代語
4424	て居ます、週んでやって下されと言ふに、正丸は大人らしう|慨りて加減が悪らい	の	ですかと底面目に問ふを、いいえ、と母親悩しき笑顔をして少し經ては癒りませ	助詞-格助詞	近代語
4425	となら、ちょうど、時計の各部分を造るものが、めいめい、隨意な形のものを造る	の	と、同じような結果におちいらなければならないのである。|　第三課　胃の説論。	助詞-格助詞	近代語
4426	て寄具なければなり。|而してそのこれを悦ぶ、これを楽しむの情は、その憐れざる	の	はなはだしきにしたがってますます切にして、往々判断の明瞭を失う者多し。|従	助詞-格助詞	近代語
4427	云々』と言えり、内義[政府の仕事]を軽しとして外刺[民間の仕事]を重しとなす	の	はなはだしきに至るときはついにこのリベラール党の論に帰する恐れなきあたま	助詞-格助詞	近代語
4428	『なる者をして益かつ愚なるものを圧せしむるの政をなすは豈にその大理に悖う	の	はなはだしきものにはあらずや〔大坂戎座板垣氏演説筆記〕。|　これによりてこ	助詞-格助詞	近代語
4429	ものはわが『日本』これなり。|当初世人はその言論のすこぶる世の風潮に逆らう	の	はなはだしきをもって、あるいはこれを猿廻論と罵り、あるいはこれを鎖国説と嘲	助詞-格助詞	近代語
4430	』|谷論太、|小学の生徒は面白がるかは知らねど、|吾等にとりては何等謹聴する	の	ふしなしと|信ず』|と、|眸りのわからず屋故、|彼のやうに通俗に、わかり易きやう	助詞-格助詞	近代語
4431	多量の水蒸気となって、空中に昇ってゐても、べつに、かわるといふことのない	の	も、つまり、この循環運動があるからである。|露。|佐山田の霧の中遠ふみ分け	助詞-格助詞	近代語
4432	話は、事實あったことか、どうかはわかりませんが、かういふ話の|衝まるといふ	の	も、もとは、忠相が裁判が上手であったことからおこったのであります。|忠相は	助詞-格助詞	近代語
4433	恐謝罪はさんすか、何とで御座んす、私の事を女郎女郎と長舌づらに言はせる	の	もお前の指圖、女郎でも買いでは無い か、鷹一本お前さんが世話にはなられな	助詞-格助詞	近代語
4434	泣き蟲こ、思はず、手を放したのだ。|その方は、もとが他人だから、子どもの泣く	の	もかまはずひっぱったのだ。|不届者め。|と、しかりつけました。|里親は、とう	助詞-格助詞	近代語
4435	ればこそ。|つつにて、母に連れられ、　外へ出でたら、その時は、　地のゆふ	の	もやみしあとにて、　四方の天は、火事のために、　すでに、まっかになりあたり	助詞-格助詞	近代語
4436	ちはえし。|恐ろ面白くない、おもしろくない、彼の人が求なければ幻燈をおじめる	の	も嫌、伯母さん此處の家に智恵の板は貰りませぬか、十六武蔵でも何でもよい	助詞-格助詞	近代語
4437	歌の方で生活主義といふ様なものがあり、此頃批評家が現實批評の詩といふ	の	も是れがためである。|　現實批評の詩といふ語は生硬の譏を免れない。|若しさ	助詞-格助詞	近代語
4438	へしを付けようと思ふよ、だから信さん友達がゐに、夫れはお前が嫌やだといふ	の	も知れてるけれども何卒我れの肩を持って、横町組の亀をすすぐのだから、ね	助詞-格助詞	近代語
4439	て、はじめて、綱の切れたことを知った。|また、父の、何か、しきりにいってゐる	の	も聞、た。|しかし、もう、そのことばを聞きとることはできなかった。|さても、不思	助詞-格助詞	近代語
4440	てたまりません。|』といった。|大助は、それをなぐさめて、|『おまへのくやしがる	の	も無理はない。|しかし、今の戦事は、昔の戦争とはちがって、ひとりで、進んで	助詞-格助詞	近代語
4441	に面白い。|文章でもさうだ。|鏡花などの作が人に印象を與へる事が深いといふ	の	も矢張りかういふ點だらうとおもふ。|一寸一刷毛でよいからその風景の中心こ	助詞-格助詞	近代語
4442	をかはふようにするので、項羽は殺すをりがなかった。|隊長は主君の身の危い	の	を見て、にはかに、座を立って、門外に居る奚嚊の所に行って、このことを告げ	助詞-格助詞	近代語
4443	前く、家の上の方まで上り、それにつないである綱をたぐってもらって、おりてくる	の	を見ては、|「にはい。」|と思ふ心は、どこへか行ってしまって、|「ひとつ乗ってみ	助詞-格助詞	近代語
4444	直からうな、己らは下駄より雪駄が好きだから、三枚裏に鐵歯の鼻緒といふ	の	を履くよ、似合ふだらうかと言へば、美登利はくすくす笑ひながら、昔の低い人	助詞-格助詞	近代語
4445	魚の社年、國民に代って兵に役し、|一旦事起らば、悠まち生命を創げんとする	の	愛國者一万數千人、|今や既に半島に霹宿せり。|彼地、|夏きはめて熱く、|ことに	助詞-格助詞	近代語

あるいは、「連体形＋格助詞ガ」のうち、主格相当ではないガによる複合辞的な表現を持つ以下のもの、

「連体形＋がため%」（ため、ために、ための等を含む。漢字表記「為」のものも）

「連体形＋がゆえ%」（ゆえに、ゆえの、ゆえなり等を含む。漢字表記「故」のものも）

「連体形＋がごと%」（ごとし、ごとき、ごとく等を含む。漢字表記「如」のものも）

などを、一括して取り除くなど。

ソート済みのデータをさらに範囲を限定して並べ替えることで、細分化した類型ごとに一ヵ所にまとめ、検討の対象から外すべきもの（俗に言う「ゴミ」）の除去を徹底することが可能となる。例えば、キーとなる助詞ハの直前に「ある」を、ハの直後に「、」を持つ、「…あるは、…」という文字列を持つレコードをソートして、「あるは、また」とな

るものを（「あるいは」に同じとみなして）一括して削除しつつ、「名詞＋あるは、」とあるものをまとめて取り出す、など。（後者は、主格相当のものと置換されたハ（先の(14)(15)の類）と見なして扱うこととなる。）

また、助詞ヲの前文脈でソートし、口語性の認められる（口語と同形の）終止形＋ヲの用例を点検するなどの作業も、同じ文字列を持つレコードが一ヵ所にまとまった状態で行えば、きわめて効率的に進めることができる（下図）。

	前文脈	キー	後文脈	品詞	コーパス
6255	いでは無いか、だから横町の野蕃漢に馬鹿にされるのだと言ひかけて我が弱い	を	恥かしさうな顔色、何心なく美登利と見合す目つきの可愛さ。｜お前の祭の姿は大	助詞-格助詞	近代語
6256	と猶ぞいひける、單衣は水色友仙の涼しげに、白茶金らんの丸帯少し幅の狭い	を	結ばせて、庭石に下駄直すまで時は移りぬ。｜まだかまだかと塀の廻りを七度び	助詞-格助詞	近代語
6257	いよいよ深き者は、いよいよ沈黙するが如し。｜而してその黙するや、これをいう	を	忘れたるに非ず、時あっていうときは、その言も亦適切にして、忌憚するところな	助詞-格助詞	近代語
6258	でなさいといい、足軽が平士に対し、徒士が大臣に対しては、直にその名をいう	を	許さず、一様に旦那様と呼て、その交際は正しく主僕の間のごとし。｜また上士の	助詞-格助詞	近代語
6259	だし意味深遠なる著書は読者の縁もまた遠くして、発兌の売買上に損益相償う	を	得ず、これを流行近浅の雑書に比すれば、著作の心労は幾倍にして、所得の利	助詞-格助詞	近代語
6260	婦女子の夜行に重大なる箱提灯を僕に持たする者もあり。｜外に出でて物を買う	を	賤しむがごとく、物を持つもまた不外聞と思い、剣術道具釣竿の外は、些細の風	助詞-格助詞	近代語
6261	諸易き道理を知らずして、結婚は唯快楽の一方のみと思い却て苦労の之に伴う	を	忘れて、是に於てか男子が老妻を捨てて妾を飼い、婦人が家の貧苦を厭うて夫	助詞-格助詞	近代語
6262	、夫婦喃々の間に決したることならんなれども、是れぞ所謂老人の口腹を養う	を	知て其情を養うの道を知らざる者なり。｜不敬不埒と言うよりも常識を失う朱愚と言	助詞-格助詞	近代語
6464	はえある心地す。｜みなこれ『コロッセウム』、『ヰクトリア』などいふ舞踏場のあく	を	待てるなるべし。」｜かく語る処へ、胸当につづけたる白前垂掛けたる下女、麦酒	助詞-格助詞	近代語
6465	うか。｜一木一草に對して、その微妙な靈に驚くことはないであらうか。｜花のさく	を	見、その外形を喜び感情を弄んだものはロマンチツクな過去の詩人であつた。｜そ	助詞-格助詞	近代語
6466	、他の獸類の水を飲み、草を食はんとて來るを待つ。｜かくて、獸類の來り近づく	を	見れば、たちまち飛びかゝりて、これを捕ふ。｜されど、もし、誤りて、捕ふることを	助詞-格助詞	近代語
6467	潮の退きたることありしが、今の海潮には満干の時刻あり。｜古は紫雲のたなびく	を	見て英雄の所在を知りたれども、今の人物は雲の中に求む可らず。｜には古今の	助詞-格助詞	近代語
6468	がこころは咽び泣けり。｜また地平よ、｜戀する眼よりも熱く――｜眞晝のかがやく	を	見れば、｜ただ悲しく、ただ戀しくただなつかしきを。｜……………………	助詞-格助詞	近代語
6469	されたるは、軽重の別を知らざる者なりと。｜此一言を聞かば印度人も又口を開く	を	得ざる可し。｜故に世の事物は唯旧きを以て価を生ずるものに非ざるなり。｜前に	助詞-格助詞	近代語
6470	として進むや、｜我國の比肩すべき所にあらずと雖ども、｜其勞力者を他國に仰ぐ	を	得ること、｜亦た我國の其望多からざると同日の論にあらず、｜是れ豈に殷鑑の近	助詞-格助詞	近代語
6471	ば思想の博大を缺く。｜是れ氏が法律出身にして、｜其の頭腦自然に保守に傾く	を	免がれざればなり。｜氏は始め單純なる壷文の目的を以て『日々新聞』に筆を執り	助詞-格助詞	近代語
6472	に殿せしものは即ち朱子の性理学也。｜吾人は今日に於て人心の哲学的に傾く	を	怪しまざる也、唯其久しくされ飢渇せしが為めに善き物と悪しき物とを撰ばずして	助詞-格助詞	近代語
6473	は、存せざるに如かず、志は飯を食はず事なければなり。｜志は欠くも、飯は欠く	を	得ざればなり。｜〇さりとも志を棄てんは惜しき時、一策あり、稍々多く志を仕入れ	助詞-格助詞	近代語
6474	眼を一方に転ずれば、彼三田翁が着々として思想界に於ける領地を拡げ行く	を	見るなり。｜文人としての彼は孳々として物質的知識の進達を助けたり、彼は泰西	助詞-格助詞	近代語
6475	はだ多くして、白人の、土人と物品を交換するには、かならず、火薬をも携へ行く	を	例とするにいたれり。｜ある時、一人の白人あり。｜例のごとく、他の物品とともに、	助詞-格助詞	近代語
6476	行きて宿りし夜、谷を隔てたるあなたの大なる森林の前を横ぎりて女の走り行く	を	見たり。｜中空を走る樣に思われたり。｜待てちやアと二声ばかり呼ばりたるを聞け	助詞-格助詞	近代語
6477	るにあらずや。｜高等學校を卒業すれば、｜洋行するを得べく、｜直に事業にも就く	を	得べき也。｜知らず、｜議會は如何に文部省の提出案を議決する乎。｜西洋の諸國	助詞-格助詞	近代語
6478	｜〇をととしの頃は三十八度以上の体熱ありて、しかも能く客と語り能く字を書く	を	自ら驚き思へり。｜今年五月よりこのかた三十九度以上の熱度を以て、能く飯し能	助詞-格助詞	近代語
6479	ながらなが／＼しき病に飽きはてて、つれ／＼のやるかたなさに書読み物書く	を	人は我を善く勉めたりといふ。｜日頃書などすさめぬ人も長き病の牀には好みて	助詞-格助詞	近代語
6480	大声叱呼して其非を鳴さざるを得ず、而して世の短視なる者詩人の斯の如く説く	を	見て直ちに是れ詩人の哲学也と曰ひ明月や池を廻つて夜もすがらと歌ひし為め	助詞-格助詞	近代語
6481	際、｜誰れか期せん東洋の天地遽然として震撼し、｜風雲暗澹雷霆轟き霹靂閃く	を	見るに至らんとは、｜英人バイルス氏言あり、｜曰く｜現時列國の人民は宛も噴火山	助詞-格助詞	近代語
6482	、｜人動もすれば見と信とを対せしめては、信の一義に宗教上千鈞の重きを措く	を	常とし、而して見の一義に至りては之れを説くもの稀也、況んや其の光輝ある意	助詞-格助詞	近代語
6483	に内附せんと企てつつある所は即ち滿州なり、｜其の滿州を清國主權の下に置く	を	名として實際上之に對し保護の權を行ひ、｜遂に埃及の英吉利に於ける如き姿と	助詞-格助詞	近代語
6484	皇固より正成の功を知らざるに非ずと雖ども、人心に戻て之を首功の列に置く	を	得ず。｜故に足利は王室を御する者にして、楠氏は王室に御せらるる者なり。｜是	助詞-格助詞	近代語
6485	掛くるを知る、｜金時計をぶらさげるを知る、｜床の間に古書畫をかけ、｜生花を置く	を	知る。｜而して未だ書物を以て室内を飾るを知らず。｜談、｜新著の事に及べば、｜彼	助詞-格助詞	近代語
6486	偏重なりしが、末年に及て人智僅に歩を進め、始て其一端に些少の分銅を置く	を	得たり。｜かの天明文化の頃より世に行はれたる著書の類は即ちこの分銅と云ふ	助詞-格助詞	近代語
6487	つと消ゆる闇の人々。｜星かげは街に沈みて｜にごり川、水音にぶし、｜芥場に動く	を	見れば｜そは飢ゑし露地の痩犬。｜街の風、旋を捲いて｜ふと起り、ひろごりのぼる	助詞-格助詞	近代語
6488	の進むや必ず直線の上に走る、｜是れ普通の理なり、｜而も一一にして物の動く	を	檢するに、｜豈一一にして直線に進むを得るとせんや。｜人を殺すこと勿れといふ、	助詞-格助詞	近代語
6489	―即ち人生に相亘るに不便なりと知るべし。ありのままに見、ありのままに描く	を	写生と言ふ。小説家たる便法は写生するに若かず。但しここに「ありのまま」と言	助詞-格助詞	近代語
6490	の如く、｜是に於て歐洲の農法は一變せざるを得ずして、｜畢竟一般に衰退に赴く	を	免れず。｜然れども文化の進歩と共に人間の嗜好は漸くに發育するを則とするが	助詞-格助詞	近代語
6491	る初には、甘をなめて自から禁ずるを知らず、ただこれを随意に任してその飽く	を	待つの外に術なしという。｜また東京にて花柳に戯れ遊冶にふけり、放蕩無頼の	助詞-格助詞	近代語
6492	｜或は云はん、｜勞力の不足は器械を以て之を補ひ、｜以て農業上の不便を防ぐ	を	得べし、｜何ぞ必しも深く憂とするに足らん｜と、｜曰く｜然り、｜豈に夫れ然らんや、｜方	助詞-格助詞	近代語
6493	又悲哀か、喜か、けぢめも分かぬ｜幽なるあひだよりして、｜これらの祈、心に湧く	を	覺ゆ。｜我望、不幸なる思想に反響して｜次第に、あかるく、｜不言のうちに波うつは	助詞-格助詞	近代語
6494	すること能はざる程の状態なり。｜整理を要する所以なり。｜本論に聲音論を缺く	を	怪むものあらむ。｜著者は聲音論と言語の内容に關する論、即、語法及句法に關	助詞-格助詞	近代語
6495	しだと振ぢ鉢巻をする男子のそばから、夫れでは私たちが詰らない、皆が騒ぐ	を	見るばかりでは美登利さんだとて面白くはあるまい、何でもお前の好い物におし。	助詞-格助詞	近代語
6496	風引かうぞと呼立てられるに、はい今行ますと大きく言ひて、其聲信如に聞えし	を	恥かしく、胸はわくわくと上氣して、何うでも明けられぬ門の際にさりとも見過しが	助詞-格助詞	近代語
6497	んとする程に、エリスの梯を駈け下るに逢ひぬ。｜彼が一聲叫びて我頸を抱きし	を	見て馭丁は呆れたる面もちにて、何やらむ髭の内にて云ひしが聞えず。｜「善くぞ	助詞-格助詞	近代語
6498	風に、さらされ、雨に、ぬれ、｜軍馬の間を往來し、｜黒きを旗のひかりとし、｜さけし	を	旗のほまれとし、｜ひとたび、えては、千代までも、｜萬代までも、つたへゆく｜聯隊	助詞-格助詞	近代語
6499	蝸牛角上の争を事とせし彼等が、｜目覺しくも一躍して、｜世界的運動を開始せし	を	壯とし、｜着々經營する所あるべきを期待せしも、｜爾來幾旬、｜其事業は杳として聞	助詞-格助詞	近代語
6500	決議を受くるありしに至ては、｜政黨の腐敗糜爛、｜殆んど其頂上の絶頂に達せし	を	見るに足らん。｜中村は既に除名せられたれど、｜星遞相起訴不起訴の問題は、	助詞-格助詞	近代語
6501	相の起訴不起訴が、｜一時の問題となりし如き、｜以て政黨腐敗の極度に達せし	を	證すべきに、｜更に久しく逆境に立ちし進歩黨中に於て、｜中村彌六が、｜亡命清客	助詞-格助詞	近代語
6502	、又静になりて戸は再び明きぬ。｜さきの老媼は慇懃におのが無禮の振舞せし	を	詫びて余を迎へ入れつ。｜戸の内は厨にて、右手の低き窗に、眞白に洗ひたる麻	助詞-格助詞	近代語
6503	という論旨の分明に解得せらるる論文の、然も大家先生等の手に成りて出でし	を	見るに至らんとは。｜若し此事にして余が所説に対して、或は余が所説に動かされ	助詞-格助詞	近代語
6504	。｜彼は東京に在りて、既に天方伯の祕書官たりしが、余が免官の官報に出でし	を	見て、某新聞紙の編輯長に説きて、余を社の通信員となし、伯林に留まりて政治	助詞-格助詞	近代語
6505	痛感慨の刺激によりて常ならずなりたる腦髓を射て、恍惚の間にここに及びし	を	奈何にせむ。｜公使に約せし日も近づき、我命はせまりぬ。｜このままにて郷にか	助詞-格助詞	近代語
6506	に來ては嫌やだよと、置きざりに一人足を早めぬ。｜お酉さまへ諸共にと言ひし	を	道引違へて我が家の方へと美登利の急ぐに、お前一處には來て呉れないのか、	助詞-格助詞	近代語
6507	彼は余に向ひて母にはこれを祕め玉へと云ひぬ。｜には母の余が學資を失ひし	を	知りて余を疎んぜんを恐れてなり。｜嗚呼、委くここに寫さんも要なけれど、余が	助詞-格助詞	近代語

４．　実例から（気付かれる点）

これらの作業を経て絞り込んだレコードを通じて、近代語における実例から気づかれることを、以下に記述する。先に２．において述べた〈終止形準体法〉との関連を探る意図から、格助詞は主格相当のガ、目的格のヲ、連体格のノについて特に注目し、係助詞は特にハ・モに絞って用例を検討した。その結果、形容詞と動詞の、今日的な終止形と同形のものが、文の（述語以外の）成分として準体的に用いられる場合には、いくつかの顕著な類型が存在することが指摘できる。

４．１　現代語形の終止形の例・形容詞の場合

４．１．１　ヲ格に立つもの（１）

まずは、形容詞について述べる。近代語資料にも、今日的な終止形と同様の語形において、以下のようなヲ格に立つ用法が見受けられる。

(21)「すし」と書ける看板よりも「壽司」とヒネつたる漢字の看板多數を占めつゝあるは爭ふことが出來ない、|判りやすいを主とする花柳界でさへ「まちあい」と書かずし

て「待合」の漢字を喜ぶ位ゐである（太陽-192801-009_東京新旧看板考_）

(22)男が泣くてへのは可笑しいでは無いか、だから横町の野蕃漢に馬鹿にされるのだと言ひかけて我が弱いを恥かしさうな顔色、何心なく美登利と見合す目つきの可愛さ。（AX_たけくらべ（樋口一葉））

(23)首筋が薄かつたと猶ぞいひける、単衣は水色友仙の涼しげに、白茶金らんの丸帯少し幅の狭いを結ばせて、庭石に下駄直すまで時は移りぬ。(AX_たけくらべ(樋口一葉))

(21)は、「判りやすさを主とする」のようにサ語尾による体言化でヲ格に立ってもよさそうな点で、一見、先の(10)(12)との近さも窺わせる。その一方で、近代にこの表現が存在することは、先に示した文語調の連体形準体法「弱きヲ助け…」「故きヲ温ねて…」などとの連続性において理解される。(22)(23)も同様に、イ音便化を経た現代語の語形ではあるもののいずれも連体形由来のものとみなされ（(22)は「我が」の連体修飾を受けての準体法、(23)は先行する名詞「丸帯」を含意し同格の用法に近い）、いずれも文語調との連続性が認められる。

４．１．２　ヲ格に立つもの（２）

こういった古語の連体形準体法に由来するものと異なり、終止形の用法に由来すると見なされ得る点で目を引くのは、次のようなものである。

(24)『どうだらうか』|小松は、|大隈と寺島の顔を見た。|『利子の高い安いを論じて居る場合ではありますまい、|一國の興亡にかゝる大問題、|構はない借りようでは厶らぬか』|（太陽-192505-031_明治初年外交物語（その八）五十万円対ゼロ_口語）

(25)風俗習慣の比較研究をする時には、|異地方住民相互の間に存する系圖的縁故の遠い近いを探り知る事が出來ませう。|言語の比較研究も亦諸種族の系圖調べの役に立ちます。（太陽-190102-044_諸種族相互の系図的関係を考へ定める方法_口語）

(26)ときに、|念を入れて調べて置かなかつたのがいけないのだ。|それで一度、|どうしても白い黒いを分けてしまはうと云ふので、|裁判に持出しかけたんだよ。|（太陽-191701-039_戯曲　生きんとすれば―二幕―_口語）

対義関係にある二語の形容詞を対比的に並べまとめてヲ格に立てるこの構文は、終止形の一用法であった「善し悪しを定むる」（枕草子・能因本三一一段）などの流れを引くと見なされ、その点で、今日的な終止形準体法（多分に引用的なもの）にきわめて近い性質を持つ。

対義語を並べ（て対比させ）ることじたいが、「高いか低いかの問題」「遠いか近いかの関係性」などといった含意を生じさせて、つまり何らかの体言的なカテゴリを与えるという意味で、おのずと準体的であろうとするのであろう。また、それは「利子が高いか低いか、という問題」「系図的縁故が遠いか近いか、という関係性」のような意味構造を担って提示される以上、おのずと引用的な性質を帯びるのであろう、と考察される。

４．１．３　助詞ハを伴うもの（１）、助詞モを伴うもの

形容詞の対義語ペアは、助詞ハ・モを伴う用法でも類型をなしていて目につく。

(27) |此故に私は貝塚の廣い狹いは住民の多寡を示し、|厚い薄いは居住年月の長短を告げるで有らうと考へて居りますが、|（太陽-189509-042_石器時代遺跡の実践は人類学上如何なる利益_口語）

(28)専門家のやうに美術の鑑定等が出來たわけではない。|　中には、|子には、|美術のいゝ惡いはわかつて居なかつた等と批難する人もある様であるが、|しかしそれはあまりに、|極端な批難であつて、（太陽-192513-028_浜尾子を追懐す_口語）

(29)先生、|小説などゝいふものは氣が乘らなければ書けないといふではありませんか。|氣が乘つたら、|苦いも辛いもございますまいに。|其所です、|其所です、|（太陽-189503-022_新聞小説（上）_口語）

(30)警部の古手ぐらゐが、|採用されてゐたもので、|郡長の職は、|さうした事務の經驗のふかい、|酸いも甘いも呑込んだものでないと滿足につとまらぬものと考へられてゐた。|（太陽-192502-053_官場の新人を評す_口語）

(31) |國民として、|決して無責任をいふことを許さなくなつて來た。|いゝも、|惡いも、|凡て國民の責任である。|（太陽-192511-026_日露国交と普選実施_口語）

ハを伴う(27)(28)が、ヲ格に立つもの同様、二語をひとまとめにして提示するのに対し、モの場合には(29)～(31)のように二語それぞれにモが伴われる点が異なる。前者（ハの例）にヲ格の例との類似性を認めるならば、(27)(28)を、終止形によって対比される二項を示す表現と見ることも出来るが、後者（モの例）に関しては、「老いも若きも」式の連体形準体法（動詞の居体言「老い」に相当する一般的な準体用法を持つものとして、形容詞では連体形が用いられる）に由来すると見ることも出来る。終止形と連体形が同一語形となった近現代語においては特に、(27)(28)の類における形容詞の用いられ方と(29)～(31)の類におけるそれとは不可分な連続性のもとで認識されるであろうから、これらは後の終止形準体法の一般化にとって無関係ではないと考えられる。

４．１．４　助詞ハを伴うもの（２）

形容詞の例では、助詞ハを伴う場合に、以下のような同語反復的な構文がまたひとつの類型として際立っている。

(32) |お前様が風雅の道に、|お嗜みが無いのはなあ！|信孝　|何と云はうと、|無いは無いのぢや。|（太陽-190901-058_三七信孝_口語）

(33) |大きいは大きいだけに影響も少ないけれど、|小さいは小さいだけに損徳の響きが甚い、|まア比喩て見ると象と蚤の喧嘩だな、|（太陽-190902-022_銅山王_口語）

(34) |積極とは初め何か儲け仕事でウンと儲け、|其れから政治を専門にすること、|ヂヨセフ・チエムバレンの如くするのであつて、|面白いは面白いが、|特殊の才がなくてはならぬ。|（太陽-190911-012_政治家と生活問題　大政治家と小政治家_口語）

(35)財政の關係等から、|遺憾乍ら中止しなければならぬこともあり、|或は又人口の關係若

は財政の關係は、|苦しいは苦しいに相違ないけれども、|其の政略關係或は對比隣列國の關係から、|これ等の苦痛を忍んで擴張を斷行しなければ（太陽-191703-027_戦後の軍備問題_口語）

(36) |じつと此大佛殿を見てゐると、|どうも均合が惡いやうに思へてならない。|大きいは大きいがどつしりとした大きさがなく、|宛然かも馬鹿に大きくて弱い力士を見るやうな感がする。|（太陽-191705-017_旧都の春を訪ねて_口語）

(37) |野田を總理大臣にしたいとは吉原盛隆でも思つて居まいね。|俊作でも、|したいはしたいだらうが到底駄目だと諦めて居るだらうぢやないか。|（太陽-192504-025_政界鬼語_口語）

例えば(32)は「無いものは無い」に、(33)は「大きいことは大きい」にほぼ等しく、それらの形式名詞を補った表現に置き換えられることから、この類型におけるハの前の形容詞は連体形由来とみなしてよいと判断される。

４．２　現代語形の終止形の例・動詞の場合

次に動詞について述べる。動詞の場合にも類型が見受けられる。

４．２．１　助詞ハを伴うもの

動詞の場合、助詞ハが伴われた例において、同一動詞の肯定－否定ペアの中に見えるものが顕著な類型をなしており、特に目を引く。

(38) |碌な事の出來ないのは初めから知れて居ります。|けれども一心は一心ですからねえ。|出來る出來ないは二の次にまア根限りに働かうと思つて居ます。』（太陽-190113-020_左巻（承前）_口語）

(39) |無論耳は聽えなくなるんだが、|聽える聽えないは別として、|あの年齡では切開した處の肉の上りが遲く、|局部に熱を持つて、|（太陽-190916-021_死んで行く人_口語）

(40) |『いけない！それがいやだから君に相談するんだよ。|實はね、|この相談に乘る乘らないは君の自由だが、|たとへ乘らないにしても、|發明の祕密だけは保證してくれるだらうね？』|（太陽-192505-049_ラヂオと犯罪_口語）

(41) |私はその態度を甚だ善いと思ふ。|術策が潜んで居る居ないは別として、|確かに善い態度だと思ふ。|（太陽-192801-024_当局者辞任せよ）

これら肯定－否定ペアの例は、先の(24)～(26)や(27)～(28)における対義語のペアに近く、やはり「出来るかどうかの問題」「乘るかどうかの判斷」などの含意を生じさせてカテゴリ（という体言的な性格）を与え引用的に示される点で、先のものと共通する性質を持つ。これらの用例にも、今日的な終止形準体法との近さを認めてよいと考えられる。

４．２．２　助詞モを伴うもの

動詞の肯定－否定ペアが助詞モを伴う例も、まとまった形で現れひとつの類型をなす。

(42) |（まさ子の手にしがみつく。|）　|　まさ子。|（強ひて平靜を裝つて）|はな子さん。|何です。|ゆるすもゆるさないもないぢやありませんか。|（終に泣聲になる。|）（太

陽-191701-039_戯曲　生きんとすれば―二幕―_口語）

(43) |『あら、|隠すも隠さないも有りませんよ。|婆アやも一處ですから、|あの人に聽いても分りまさアね。|（太陽-190107-020_東京病_口語）

(44)メーブルとの結婚は眼に見えて駄目になるの他ない。|然し、|父親の惡事を表むきにするもしないも、|ビイチの胸三寸のうちにあることで、|自分にはどうする力もない。|（太陽-192514-043_長篇探偵小説　ハートの九『第八回』_口語）

(45)我等兩人は長途の旅の勞も忘れて、|喜ぶも喜ばないもあつたものぢやない。|實に嬉しかつた。|（太陽-191713-047_岩村透君の手紙_口語）

(46) |『奈何いふ譯でせう、|何處が氣に入らないのでせう。』|『氣に入るも入らないも無い。|極氣が小さく吝に出來て、|惡く云やア吝嗇だから思切つたお金は中々出し得ないのサ。（太陽-190105-024_投機_口語）

(47) |其頃はまだ其處は誰の許可地でもなかつた、|採取人は誰の許可地であるもないも頓着せず、|唯だ砂金が多くあると云ふ故、|無暗に堀りに行きて數百人も集まつた、|（太陽-190101-031_北海道枝幸砂金地巡見_口語）

(47)「～である／ない」は厳密には助動詞とその否定のペアであるが、類例としてここに挙げた。それぞれが助詞モを伴う動詞の肯定－否定ペアは、伝統的な「行くも行かぬも別れては」「知るも知らぬも逢坂の関」（後撰集　雑一・1089）などの連体形準体法に由来するものであり、今日的な終止形準体法に対する何らかの間接的な影響については、先に４．１．３．の末尾に述べたとおり、注意深く検討するべきであろう。

なお、これら［肯定＋モ　否定＋モ］に続く述語部分に来るものは、「ない」（(42)(46)）や「ありません」（(43)）「あったものじゃない」（(45)）など、ほぼ同一の内容を表すものである。肯定・否定ともにいずれも選択肢として「ない」ことを表すこれらの類型的な表現は、(47)「頓着せず」の例を含め、「肯定か否定かというかたちで行われる問題設定じたいが、存在し得ないか、問うに値しない」ことを意味する。その点でこれらは、先に見た形容詞の(29)(30)と特に近い。

対義語のペアを用いる形容詞の場合（(24)～(31)）と、肯定－否定のペアを用いる動詞の場合（(38)～(47)）とは、対比の示し方が異なるだけで、これらはいずれも、対比的な二項に代表されるある種のカテゴリ提示の働きを持つ。大半は旧来の連体形準体法に由来するものでありながら、引用的な要素を備える点において今日的な終止形準体法との類似性を持つことは、注意される傾向であろう。

５．　おわりに

コーパスから得られた用例の検討を通じて知られる限り、近代における〈終止形準体法〉相当の（または類似する）表現は、思いのほか多様である。実際の用例の検討を通じて、本稿では、特徴的な類型の存在を指摘し、それらの意味構造と準体的であることの関連性について主に考察したが、これらに準ずるものとして、なお、次のような複合的なもの（(48)）や一部が省略されたもの（(49)）も見られる。

(48) |勝ちさへすれば宜いとするのは丁度腹さへ充つれば宜いとするやうなもので、|それでは旨いも旨くないも酸いも甘いも鹽加減も何も論は無いといふもので、|まるで滅茶苦茶である。|（太陽-192513-029_将棋のたのしみ_口語）

(49) 太陽－190911－023__表白__口語：』|と寛三は答へたが、|腹の中では|『現在自分の弟の妻が死にかゝつてると云ふに、|忙がしいもないもんだ』|と云ふやうな反感がむら〜〜と起つたのだ。|　しばらくして、|二人は病室を出て（太陽-192513-029_将棋のたのしみ_口語）

また、文語文を基調とした文脈の中で、旧来の連体形終止法が、古語の連体形ではなく現代語の語形で行われる例が、わずかながらも見受けられることを付記しておきたい。(50)は二段動詞が一段化して、(51)はナ変動詞が四段化して、それぞれ現代語と同じ終止連体形に格助詞ヲが伴われたものと見られるが、いずれも今日的な終止形準体法との区別は截然と付けづらい。

(50) |余は、|之を祝す。|左れど、|此に極めて明白なる一事あるに、|其事のあまりに明白すぎるを以て、|諸君は此際或いは失念せられたるならんか。|余は、|更ためて此一事を注意すべし。|（女学雑誌-189427-03-卒業は始業なり）

(51) あれで年は六十四、白粉をつけぬがめつけ物なれど丸髷の大きさ、猫なで聲して人の死ぬをも構はず、大方臨終は金と情死なさるやら、夫れでも此方どもの頭の上らぬは彼の物の御威光（AX_たけくらべ（樋口一葉））

これらと、今日的な〈終止形準体法〉との連関や異同など、注意深く考察すべき点は多い。本稿は、ひとまずの記述と問題提起に終わったが、コーパス利用で可能となる効率的な用例収集と分析を通じて、その歴史的な背景や位置付けなどの解明に向け分析を継続したい。

文　献

藤田保幸（2000）『国語引用構文の研究』（和泉書院）

島田泰子（印刷中）「広告表現等における〈終止形準体法〉について」（『叙説』奥村悦三先生御退休記念特別号　2013.1 刊行予定）

近代の地方出身作家の助詞の用法について—宮澤賢治と濱田廣介—

小島 聡子（岩手大学人文社会科学部）[1]

１． はじめに

近代に入って作られた「標準語」「言文一致体」は、東京語の影響が大きい。地方出身者がそのような「標準語」を話すのに苦労したことはいうまでもないが、書く場合も方言の影響を少なからず受けている可能性が考えられる。つまり近代の資料に見られる言葉の揺れのなかには、方言の影響が見られる可能性があるということである。逆にいえば、この時代の地方出身者の書いた作品の語法を細かく観察することで、「標準語」がどのように普及していたかを知る手がかりにもなりうるのではないかと考える。

２． 宮沢賢治と濱田廣介

大正末期から昭和にかけて活躍した宮澤賢治（1896（明治 29）年‐1933（昭和 8）年、岩手県）と濱田廣介（1893（明治 26）年‐1973（昭和 48）年、山形県）は、ほぼ同年代で東北地方出身、童話作品を多く手がけたことなど共通点が多い。

このうち、宮澤賢治は 1903（明治 36）年に小学校に入学しているが、この翌年 1904（明治 37）年からは、国定読本の使用が開始されている。標準語の普及には国定読本の使用がその一端を担っていることは知られており、宮澤賢治はいわば国定読本による標準語教育を受けた第一世代の一人ということになる。

ところで、「標準語」は東京の言葉と関係が深いことは先述の通りだが、近代の作家は、東京の生まれ育ちだったり、あるいは、高等教育は東京で受けるなどある程度の長期間東京の言葉にさらされていたりする人が多い。

濱田廣介も大学入学を期に上京し、その後はずっと東京で執筆活動をしている。しかし一方、宮沢賢治は、高等教育も岩手県内で受けており、何度か上京してはいるものの滞在期間は長くても一年に満たず、比較的東京の言葉との接触が少なかった。「標準語」との関連が深い東京語との接触時間の長さという点で、両者は大きく異なる。従って、基本的にあまり東京語にさらされず、教科書中心に「標準語」を習得した宮沢賢治と、東京語の只中で暮らした濱田廣介とでは、方言の影響の仕方も異なる可能性が考えられる。

そこで、彼らが書き言葉としての標準語を使用した際の言葉づかいについて、特にいくつかの助詞の用法を取り上げ、方言からの影響の有無について分析する。

一般に「東北地方」と一括りにされるが、山形・置賜（濱田廣介の出身地）の方言と、

[1] satok@iwate-u.ac.jp

岩手・花巻（宮沢賢治の出身地）の方言とでは異なる点も少なくない。しかし、標準語との差が大きかったことはどちらも共通しており、標準語と比べれば似ているところも多いので、比較することには意味があると考える。

３． コーパスの利用について

資料としては、宮沢賢治の『注文の多い料理店』（1924（大正 13）年発行　約 51,000 字程度）と、濱田廣介の『椋鳥の夢』（1921（大正 10）年発行　約 70,600 字程度）を使用する。テキストは初版本の復刻を利用し、これらを底本として電子化したものを「全文検索システム『ひまわり』」[2]によって検索可能な形にすることで、簡易コーパスを作成した。

その上で、両者の言葉の特異性を考えるための基準の一つとして、既存の『太陽コーパス』[3]、『近代女性雑誌コーパス』[4]を比較の対象とした。当時の言葉の標準的なものとして『太陽コーパス』等を用いたということである。

近代はいわば文体の模索期であり、表現の揺れも大きい。それだけに、現在の口語では用いないような表現が見られた場合でも、それがその作家に特有の語法であるのかどうか、さらに方言の影響があるかどうかなどを直ちに判断することは困難である。

しかし、簡易な形にせよコーパスにすることによって、個人の中での表現の揺れや言葉づかいの傾向を観察することが可能になる。また、これまでに作られた『国定読本用語総覧』[5]や『太陽コーパス』『女性雑誌コーパス』などのコーパスを用いれば、ある程度当時の標準的な言葉遣いの目安を得ることが可能で、それと比較することで、個人の表現の特異性について検討することもできる。特に本調査の資料は、これらの既存のコーパスが対象としている範囲とほぼ重なっており、比較の対象とするのに最適である。

４． 格助詞の用法

４．１　「へ」と「に」

４.１．１　「へ」と「に」の使用頻度

方向や帰着点・目的地等を表す助詞として、東北方言では広く「さ」や「に」が用いられるが、標準語では「へ」や「に」を用い、「さ」は用いない。

靏岡（2007）は、近代以降の作家の「へ」「に」の使用率を調査し、その違いが地域差に基づくものである可能性を指摘している。それによれば、西では「に」が多く東へ行くほど「へ」が多いという傾向があり、特に宮澤賢治の作品で「へ」の使用率が高いという。

小島（2012）では東北出身の宮沢賢治・濱田廣介の「へ」の使用について調査をした。それによると、用例数と及び「へ」を受ける動詞の異なり語数は下記の通りである。

・助詞「へ」の用例数　　宮澤賢治　99 例　／　濱田廣介　58 例

---

2 国立国語研究所の開発したソフトウェア。（http://www2.ninjal.ac.jp/lrc/よりダウンロード可能）
3 国立国語研究所（2005）
4 国立国語研究所（2006）により CD-ROM で公開されたもの。（http://www2.ninjal.ac.jp/lrc/）
5 国立国語研究所（1997）

・「へ」を受ける動詞　　宮澤賢治　48 語　／　濱田廣介　28 語

確かに宮澤賢治は「へ」が多いが、同じ東北出身の濱田廣介では「へ」の用例数も「へ」を受ける動詞の種類も多くない。

また、下の表の通り、同じ動詞で比較しても、濱田廣介よりも宮澤賢治の方が「へ」の使用率が高いことがわかる。

表 1　「行く」の場合の助詞の分布

	（行く）	へ	に[目的の例]	を
賢治	33	20 (60.6%)	13[4] (39.4%)	0 (0)
廣介	92	16 (17.4%)	23[1] (25%)	2 (2.2%)

表 2　「入る」の場合の助詞の分布

	入る	へ	に
賢治	28	8（28.6%）	16（57.1%）
廣介	10	0	8（80%）

このような違いは、宮沢賢治の方がより方言の影響下にあるということによるものではないかと考える。

先にも述べたが、標準語で「へ」や「に」を用いるような場合に、東北方言では「さ」が用いられることが知られている。この「さ」自体は東北全般で使われ、所在を言う場合に「〔場所〕さ　ある」が言えるかどうかなどで東北地方の中でも地域差がみられる。ただし、いつでも「さ」を用いるわけではなく、「に」も用いられる。いくつかの方言資料の記述を総合すると、東北方言の「に」と「さ」の傾向は次のようにまとめられそうである。

・「に」と「さ」は分布が重なる

・場所を表す場合に必ず「さ」を用いるわけではない。

・「なる」の前は「に」が多い。

・時間を表す場合は「に」が多い。

・目的を表す場合は「に」も「さ」もある。　（例「見さ行く」「遊びさ行く」）

一方、標準語の「へ」と「に」の分布は必ずしも上記の「さ」と「に」のあり方とは重ならない。例えば、標準語の「へ」は移動の方向を表すことが多く、移動が感じられない例には用いられにくい。また、「へ」は移動の目的を表す場合には用いられない。しかし東北方言の「さ」は移動を表す場合でなくとも用いられるし、移動の目的を表すのにも用いられるという点で、標準語の「へ」よりも用法が広い。

しかしながら、標準語にも東北方言にも、似たような意味で用いる助詞二つのペアがあり、そのうちの一方の「に」は共通しているとなると、もう一方の助詞「へ」と「さ」と

を対応させて用いることはありそうである。つまり、宮沢賢治は、東北方言で「さ」というところに「へ」を用いた結果、他の人々より「へ」の使用率が高く、また用法も広くなった可能性があると考えるのである。

４．１．２　「方」と共起する「へ」「に」の分布

一方、濱田廣介は宮沢賢治ほど「へ」が顕著に多いわけではないが、細かくみると標準語とやや異なる傾向がみられる。濱田廣介の場合、「へ」は「～の方」や「東西南北」の方角、「空」など漠然とした場所を表す語につく場合が多く、特に「方」の場合「へ」と共起する場合が多いのである。表 3 は、宮沢賢治・濱田廣介・太陽コーパス・近代女性雑誌コーパスの「方」のあとの「へ」と「に」の分布である。

表３　「方」に下接する助詞の分布

	賢治	廣介	太陽	女性
方へ	32	13	858	183
方に	12	9	3493	454

ここでも、賢治の「へ」の使用率の高さは歴然としているが、廣介でも「へ」の方が多くなっている。これは、『太陽コーパス』『女性雑誌コーパス』では「方に」の方が多いのとは傾向が異なる。廣介の場合、「へ」の使用そのものが多いわけではないことを考え合わせると、「へ」全体の中で「方」に下接する例の割合が非常に高いということになる。つまり廣介の場合、「へ」の上接語の範囲に明確な意識があるように感じられる。

賢治の場合はそのようなことはあまりなく、具体的な人やら場所やらにもつく。

宮澤賢治の「へ」の使用率の高さは、東北方言「さ」の言い換えとして「へ」を用いたことによるものである可能性があるのではないかと考える。

４．２　「～を好き」について

小島(2006)では、宮沢賢治がいわゆる対象語を表す場合に「が」ではなく「を」を用いることが多いことを指摘した。これにも方言の影響があると考えられる。

具体的には次のような例である。

・わたくしは、さういふきれいなたべものやきものをすきです。(序　p.7)

・あなたは黄金のどんぐり一升と、塩鮭のあたまと、どつちをおすきですか。(どんぐりと山猫　p.17)

・あれを嫌ひなくらゐなら、どうせろくなやつぢやないぜ。(山男の四月　p.60)

「が」を用いている例もあるが、「を」の方が多いのである。

一方、『国定読本用語総覧』によれば、国定読本では「好き・嫌い」の対象はすべて「が」で示されており、「を」が用いられた例はない。「どんぐりと山猫」は第 6 期国定読本に教材として採用されているが、当該箇所は「どちらがおすきですか」と改変されている。

また、『太陽コーパス』では、「が」を用いた例は皆無ではないが少なく、確認できたの

は次の例である。

・言ふまでもなく父といふ人を他の大人よりも好きであるとは思つてゐたが（1917 年 13 号「暴風雨の夜」上司小剣）
・父君には甘きを御好きかと伺はる（1895 年 2 号「鷹山公の家庭」大橋乙羽）
・一度は一度と次第に奥さまは、旦那さまをお好きにおなりでございました。（1925 年 7 号「長篇探偵小説　ハートの九『第二回』」延原謙（訳）；ビ・エル・フアルジヤン（作））
・いつも眞先に乘物を云ひ出して歩くを嫌ひであつたSが（1917 年 10 号「本田の死」豊島与志雄）

一方「が」で対象が示される例は，「すき(仮名表記，漢字表記，及び「お」を伴うものを含む)」の直前に「が」がくる例に限っても 94 例が確認できた[6]。

以上のことから，近代においても「好きだ」「嫌いだ」の対象語は「が」で提示される傾向が顕著であり、宮沢賢治の用語はやや特異であるといえる。ちなみに、濱田廣介では「が」が用いられており、「を」を用いた例はない。

岩手県に限らず、東北方言では一般的に助詞「が」「を」は用いないとされる[7]。「すき」「嫌い」の対象を表す場合も助詞を用いない形が一般的で、『方言文法全国地図』第 1 集[8]の第 5 図「酒が（すきだ）」でも、花巻も含め岩手県全般に、「酒が」という部分は助詞なしの「さけ」、あるいは「さけぁ」のような形で現れることからも確認できる。

従って、助詞を補おうとした時に「対象だから」ということで「を」を用いたという可能性がまず考えられるのである。

さらに、実は、東北の方言では「すきだ」「きらいだ」より、「すく」「すかない」という動詞の形がよく用いられることも関係している可能性がある[9]。「すく」という動詞では、対象は「が」ではなく「を」で示されるのは珍しくない。例えば『太陽コーパス』でも下記のように「～を好く」が使われている例が、10 例ほど確認できた。

　殊にウヰスキーを好くさうだ。（1925 年 1 号〈雑学〉著者不詳）

つまり「すきだ」の対象を示そうとした場合、方言では用いない助詞を補う必要があり、その際「すく」の助詞からの連想で、「を」を選択したのではないかと考えられる。方言では、助詞を用いないことがやや特異な表現をもたらしたといえよう。

6　『太陽コーパス』は、品詞の情報は付けられていないため、「すき」について、可能性のある表記で検索の上、各用例を確認した。
7　森下喜一（1982）
8　国立国語研究所編（1989）
9　前掲『方言文法全国地図』では、花巻周辺に「すかない」形は確認できないが、「すかない」という言い方は現在岩手大学の学生たちの言葉としても聞かれ、若い世代においても一般的なようである。

５．接続助詞（接続詞）

５．１　「ので」と「から」

まず、原因・理由を表す「ので」「から」の分布を下記の表に示す。なお、『太陽コーパス』『近代女性雑誌コーパス』では、接続助詞的に使われている可能性が高いものとして、それぞれ読点がついた「ので、」「から、」の形で検索した。

表 4　「ので」と「から」の用例数

	賢治	廣介	太陽	女性
ので	27	11	4558	759
から	44	85	7098	1312

これを見ると、宮沢賢治では「ので」対「から」は 2 対 3 程度の比率で分布しており、他のコーパスでの分布の傾向と大きく違わないのに対し、濱田廣介の場合は「から」の出現率が高い。

これについて、方言との関連の可能性を探る。

例えば『方言文法全国地図』第一集では、次のような場合が調査されている。

第 33 図　「雨が降っているから」の「から」

…山形・置賜、岩手・花巻ともに「kara・gara」

第 36・37 図　「子供なので」の「な/ので」

…山形・置賜「datta・namoNda/gara・de」、岩手「da・[Nda・na/gara・node」

これによれば、宮沢賢治の地元・岩手県では「ので」という形も見られるが、濱田廣介の地元・山形県置賜地方では、「だったがら」あるいは、「なもんだで」という形が用いられていて、「ので」という形は見られない。従って、濱田廣介にとっては原因・理由を表すような場合の言い方として「ので」の形はなじみが薄かった可能性が考えられる。

５．２　「けれど」と「けれども」

次に逆接に用いられる「けれど」と「けれども」の分布を示す。

表 5　「けれども」と「けれど」の用例数

	賢治	廣介	太陽	女性
けれども	18	7	2562	346
けれど	1	168	1179	357

この場合、「も」の有無という点で、宮沢賢治は「も」が付く形が殆どなのに対し、濱田廣介の方は殆どが「も」が付かない形で、両者の間に大きな差がみられる。他のコーパスで見ると、どちらかに圧倒的に偏るということはなく、この偏りはそれぞれに特異である。

こちらも方言との関係を考える。

逆接を表す場合については、『方言文法全国地図』第一集に以下の調査がある。

38「寒いけれども」の「けれども」

…山形・置賜「geNdo・geNdemo・geNdomo・geNdoN」、岩手「[Ndomo・domo」

39「だけど(行かなければならない)」の「けど」

…山形・置賜「geNdo・geNdemo」、岩手「domo」

これによると、両図ともに、山形県・岩手県ともに末尾に「も」の付いた形が広く分布している。特に、岩手県では「も」のない形は殆ど現れない。従って、宮沢賢治が「けれども」ばかりで「けれど」を用いないのは、方言で「も」のつく形しか用いられないことが影響していると考えてよいものと思われる。

しかし、山形県では、置賜地方の調査地点に濱田廣介の出身地・高畠町はないものの、高畠町に隣接する南陽市が調査地点になっており、そこでは両図とも「げんど」という「も」のない形が見られる。特に、置賜地方だけに限ってみると、「だけど」の場合は他の地点でも「も」のない形が現れる。ただし、遠藤他（1997）によれば、山形県全域で「けんども」が優勢であり、特に置賜地方は「けんども」の形を用いるという。

従って、濱田廣介が「も」のつかない形の方を用いることについては、方言との関係で二通りの可能性が考えられる。まず、濱田廣介自身が母語方言で「も」のついた形を用いていた場合、「けれども」という「も」のついた形は「方言である」という意識が強く、書く際には敢えて「も」のない形を用いたという可能性がある。しかし、『方言文法全国地図』の分布をみると、濱田廣介自身が母語として「も」のない形を使っていた可能性もあり、その場合、そのまま方言の影響で「も」のない「けれど」の方を用いたと考えることができる。

５．３　「～ないに」について

宮沢賢治の作品の中に、次のような「～ないに」という例が見られる。

もうそのころは、ぼんやり暗くなつて、まだ三時にもならないに、日が暮れるやうに思はれたのです。(「水仙月の四日」p.96)

『太陽コーパス』では、逆接とおぼしき「～ないに」の例は 4 例見られるが、下記の例を含め、いずれも、「～ないのに」という意味である。

蓋し歐洲の夏は九時十時迄も日は暮れないに、道路工夫の如きは五時以後の勞働に對しては夜間勤務として大割増を要求し(1925 年 11「死線にさまよふ日本─経済的危機の真相と救済策─」矢野恒太)

先の宮沢賢治の例も、同様に「三時にもならないのに」という解釈も可能ではある。

しかし、宮沢賢治の出身地の花巻方言を集めた花巻市教育委員会（2005）には下記のような語が収録されている。

「～ネァニ」＝「～しないうちに」

(文例）客「(来)ネァニ」片付げろ　＝　客が「来ないうちに」片付けろ

このことから、先の例は「ならないのに」ではなく、「ならないうちに」という意味で用い

ている可能性がある。

６．　副助詞等について

６．１　限定の「だけ」の用法

宮沢賢治には、「百円ぶん」のように数量の程度を限定する意で用いられている「だけ」の例が見られる。

・もう九本切るだけは、とうに山主の藤助に酒を買つてあるんだ。(「かしはばやしの夜」p.133)

・途中で十圓だけ山鳥を買つて東京に帰りました。(「注文の多い料理店」p.62)

特に、2 つ目の例のように「〈金額〉だけ」という場合、ここでは単に「十圓ぶん」購入したという意味で用いられていると思われるのだが、現代語では「だけ」は「それと限る」限定的な意味合いが強いため、つい「もっと買えるのに１０円分しか買わない」というニュアンスを感じてしまいかねない。

ただ、このような単に数量の程度を示す「だけ」の用法は、特に方言的な用法というわけではなく『太陽コーパス』でもかな書きの「だけ」では、以下のような例が見いだされる。しかし決して多くはないし、「〈金額〉だけ」という例はなかった。

・丁度其の眼の數だけ、錢を積んで主婦に渡した。（1895 年 04「ソクラテス的滑稽（続）」巌谷小波）

・豫め翌年の買入額を約束し、生産者をして安心して約束額だけを生産せしむるが如きは生産者保護の一法たるを得んか（1901 年 08「経済時評」坪谷水哉）

・二十歳の者は四十二、三十歳の者は卅四、四十歳の者は二十八、五十歳の者は二十一、六十歳の者は十五年だけの餘命が有るものと認めるやうに成つた（1901 年 12「科学雑談（一）」局外閑人）

・此會費徴集に關しては各部長十分責任を帯びて會員數だけの金高は必ず集むる事（1895 年 07「海内彙報」）

・養豚會社は一番盛んだが、五萬圓の資本だけの財産がありますまいか何歟。（1901 年 05「投機」内田魯庵）

ところで、『方言文法全国地図』第１集には、「52　みかんを百円***ぶん***ください」という表現についての調査がある。つまりこのような場合に、標準語としては「ぶん」が想定されるということであろう。この調査において、岩手県花巻周辺では「dɛɛ・dee」などの語形が採集されている。(参考までに、山形県「gana・ŋana」である)

宮沢賢治が「だけ」を用いたのは、この「でぇ」「だぇ」からの影響の可能性も考えられるのではないか。

６．２　「くらい」について

また、限定の意味を表す副助詞について宮沢賢治は「くらい」という語においても、他

とは異なる使い方をしていることを小島(2008)で指摘した。特徴的な点として、「くらい」の前に格助詞「の」を用いることがあげられる。

・栃の団子をとちの実のくらゐ残しました。(鹿踊りのはじまり　p.88)

・六疋めの鹿は、やつと豆粒のくらゐをたべただけです。(鹿踊りのはじまり　p.95)

また、先の「ぶん」のような量を示す意味で「くらい」が用いられている例もある。

・わたしたちは、氷砂糖をほしいくらゐもたないでも(序)

これは「ほしいぶんだけ」あるいは「すきなだけ」というような意味かと思われる。『太陽コーパス』では「ほしいくらい」という例はあるが、「暑くて、扇子が欲しい位だ」というような例で「くらい」の意味が異なっている。

これらの用法については、方言からの影響は直接は見いだせないが、方言では限定を表すような表現全体が標準語と異なっており、それがこのような特異な用法となって現れた可能性も考えられる。

６．３　「～ぐるみ」…濱田廣介の場合

副助詞ではないが、限定的な意味を加える接尾語の例も取り上げておきたい。

濱田廣介には次のような「～ぐるみ」という例がある。

・クツクウの青い體は葉つぱぐるみ揺れました(「青い蛙」p.193)

これは、「上接の名詞とともに全体で」の意で用いられる語で、「～ごと」と同様の意味を表す。

現代語では「地域ぐるみで取り組んでいる」や「組織ぐるみの犯行」などのように用いられるが、規模の大きい組織などについて使われることが多いように見える。

ただし、近代語としては『太陽コーパス』でも下記のような例が見いだされる。

・二分金を御札ぐるみに帶の間へ入ぬ(1895 年 01 号「従軍人夫」饗庭篁村)

・盆ぐるみ推進めた番茶の土瓶を(1909 年 14 号「実印と預金帳」柴田流星)

・地所ぐるみ借り入れたり(1917 年 13 号「暴風雨の夜」上司小剣)

・馬一頭と馬丁と三人ぐるみ一緒になつて、厄介になつてゐたのだから(1925 年 07 号「明治初年外交物語（その九）青年外交家の台頭」豹子頭)

・大雅寺を寺ぐるみ賣るから買つてはどうかと(1925 年 12 号「蕪村寺」橋本関雪)

なお、上記の例の使用者の出身地は、饗庭・柴田が江戸、上司・橋本が上方である。

このように「ぐるみ」を用いることは、近代ではさほど珍しい用語であったわけではないことがわかる。また、同様の意の「ごと」も 9 例ほど見いだされ、どちらの形も用いられていたということになる。

ところで、『方言文法全国地図』第１集には、53「みかんを皮**ごと**食べた」という調査があり、それによると、「ごと」の意は、山形県一帯で「garami・ŋarami」という形が用いられている。どちらも使われるとはいえ、「ごと」ではなく「ぐるみ」を選んだのは、方言で「ごと」を用いないことが関係している可能性も考えられるのではないか。

７． 今後の課題

今回の調査では、比較する方言の資料としては殆どを『方言文法全国地図』を利用した。これは、1989 年に刊行されたものであり、調査はそれ以前の 13 年間をかけて行われたという。従って、宮沢賢治・濱田廣介のそれぞれの作品が出された時代よりやや新しい言葉ということになる。ただし、濱田廣介の地元に近い南陽市の調査対象者の生年は 1913 年、宮沢賢治の地元・花巻市の調査対象者は 1911 年で、宮沢賢治・濱田廣介よりは若いが世代が異なるほどではないので、参考にはできるものと考える。しかし、さらに古い言葉を記録した文献を探すことは必要である。

また、標準語の例として用いた『太陽コーパス』については、やや調査が粗いところがあり、こちらもさらに精密な調査を試みたい。

文 献

国立国語研究所（1989）『方言文法全国地図』第１集

国立国語研究所（1997）『国定読本用語総覧 CD-ROM 版』三省堂

国立国語研究所（2005）『太陽コーパス: 雑誌『太陽』日本語データベース』博文館新社

国立国語研究所（2006）『近代女性雑誌コーパス』

小島聡子（2006）「『注文の多い料理店』の言葉について」アルテス リベラレス（岩手大学人文社会科学部紀要）第 78 号、pp.89-103

小島聡子（2008）「宮沢賢治の童話の語法について－副助詞「くらい」の用法を中心に－」『言語と文化・文学の諸相』（岡田仁教授・笹尾道子教授退任記念論文集）、pp.121-132

小島聡子（2012）「宮沢賢治の童話における「標準語」の語法―方言からの影響について―」『近代語研究第十六集』、pp.329-347

遠藤仁 他（1997）『日本のことばシリーズ　6　山形県のことば』（編者代表・平山輝男）明治書院

鸚岡昭夫（2007）「関西以東の「へ」と「に」の分布について―近代の小説を資料として―」『計量国語学』第 25 巻第 8 号、pp.341-351

花巻市教育委員会（2005）『花巻ことば集　せぎざくら』

森下喜一（1982）『岩手の方言』教育出版センター

『太陽コーパス』における漢文系複合辞の使われ方

朱　京偉　（北京外国語大学）

１．はじめに

中国では，1950 年代の末から借用語研究の動きが現われた。王立達氏が「現代中国語における日本語からの借用語」と題する論文で，日本語からの借用語を 8 種類に分類し，それぞれの実例をあげて説明したのが議論の始まりとなった[1]。同論文で示された日本語からの借用語は計 587 語に及び，一般語から専門語まで，しかも，二字漢語だけでなく，三字漢語・四字漢語も含まれている。当時において，質量とともに代表的な研究成果だといえる。

ここで注目したいのは，同論文で借用語の第 7 類としてあげられた「日本語を中国語に翻訳するときに，中国人によって創出された言い方」である。原文では，次のように述べられている[2]。

> 七，下面一些现代汉语词汇，是在我国人翻译日文时创造出来的。〈訳文：以下の現代中国語の言い方は日本語を中国語に翻訳するときに，中国人によって創出されたものである〉
> 1）基于（〇〇ニ基イテ）
> 2）关于（〇〇ニ関スル，〇〇ニ就イテ）
> 3）对于（〇〇ニ対シテ）
> 4）由于（〇〇ニ由ッテ）
> 以上为词尾是“于”的介词。〈訳文：以上は“于”を語尾とする介詞〉
> 5）认为（〇〇ト認メラレテ）
> 6）成为（〇〇ト成ッテ）
> 7）视为（〇〇ト視ナシテ）
> 以上为词尾是“为”的动词。〈訳文：以上は“为”を語尾とする動詞〉

ここでとりあげたのは，語彙レベルの借用ではなく，「〇〇ニ基イテ，〇〇ニ関スル」などのように，構文要素となる日本語の複合辞が中国語に影響をもたらしたということである。王立達氏の論文が発表された後，“关于，由于”などをめぐって，日本語の借用語とはいえないといった反対の意見が出てきた[3]。これに対して，王氏は，確かに“关于，由于”などの語形が日本語にはなく，中国語独自の言い方なので，「借用語」というよりも「意訳

[1] 王立達（1958a）を参照。
[2] 王立達（1958a）の p.94 を参照。本稿では，中国語と日本語の字体について，それぞれ原語のものを用いた。なお，中日双方のコーパスに出た旧字体も原文のまま取り入れた。
[3] 張応徳（1958）を参照。

語」とすべきだと認める一方，この種の言い方は，日本語の「〇〇ニ関スル，〇〇ニ由ッテ」などを踏まえて造られたものだと，自説を再確認した[4]。王氏の論文から半世紀以上経ったが，当初の問題はほとんど進展が見られず，残されたままである。そのため，電子資料の利用が便利になった現在，もう一度この課題について考えてみるのが小稿の目的である。

王氏の説を検証するには次の二つが前提条件となる。まず，中国語の“关于，由于”などの言い方が日本語からの影響を受けて成立したものであれば,「〇〇ニ関シテ, 〇〇ニ由ッテ」などは, 中国が日本語から借用語を受容し始めた 19 世紀末までに, 日本語で使用されていなければ, 王氏の説が成り立たない。これに関する調査にあたって,『太陽コーパス』のデータはちょうど 1895－1925 年の期間をカバーしているため,好都合な資料だといえよう。また，日本語側の調査と同時に，近代以前の中国語においても“关于，由于”などと同形の用法が存在したかどうかを調べる必要がある。この作業に関しては《四庫全書》（電子版）の検索によって解決できると思われる[5]。

本稿では，さしあたり，日中双方の電子資料を使って基本的な事実確認を行なっておきたい。実施にあたり，次のことに留意した。

（1）日中対応の観点からは，たとえば,「に限る」と“限于”,「に至る」と“至于”,「に属する」と“属于”,「に過ぎない」と“不过”,「に及ばない」と“不及”なども漢文系複合辞として同類扱いできそうであるが，小稿では，王立達（1958a）で言及された七つの表現だけを対象とする。

（2）『太陽コーパス』にある用例をとりあげる場合，日本語からの借用語が中国語に移入される時期を考慮して，なるべく 1895 年または 1901 年のものを選び，しかも，活用変化の様相も見られるように留意した。

（3）《四庫全書》の用例については, 時代的分布や例文の分かりやすさを中心に考えて, 適切なものを選んだ。また，筆者によって，文の区切り（句読点）と関連部分の日本語訳を付け加えた。

（4）用例中の字体は日中双方の原本のままに用いた。たとえば，現代中国語では,“基于、关于、对于、由于、认为、成为、视为”のように表記されるが,《四庫全書》では“基於、關於、對於、由於、認爲、成爲、視爲”となる。これは，新旧字体の違いだけで，意味と用法の変化とは関連がないと考えてよい。『太陽』の表記についても同様である。

[4] 王立達（1958b）を参照。原文では,“张同志认为‘关于’、‘由于’…等不应当看作是日语借词，这一点我完全接受。因为这些词从来源上说虽然与日语有直接关系，但因它们是我国人翻译日文时创制出来的，而不是从日语中借来的，所以按其性质来说，应当是意译词而不是外来词。我把它们列为日语借词的一项，是完全错误的。”と述べている。

[5] 『四庫全書』は清の乾隆帝の時，朝廷の監修によって編集された大規模の百科叢書である。秦以前の時代から清の前半期までの歴代の典籍が 3460 余種も収録されており，1781 年に完成された。紫禁城の文淵閣に所蔵されていた『四庫全書』の版本は，すなわち『文淵閣四庫全書』である。同書の電子版は検索機能が備わっているため，18 世紀前半までに存在していた漢籍語なら，手早く用例と所在の文献を見付けることができる。

２．に基づく／基於（基于）

２．１　『太陽コーパス』にある用例

『太陽コーパス』には,「に基づく・に基づき・に基づきて・に基づいて・に基づいた」などの形を含め，計 105 例が見られる。このうち，最も多い使い方は「に基づく」の 38 例と「に基づき・に基づきて」の 43 例である。現代語で用いられる「に基づいて」と「に基づいた」の形はそれぞれ 6 例しかなく，しかも，初出は 1901 年以後となっている。次は「に基づく」に関する用例である[6]。

① 人の之に求むるは自然の本性に基づくものなること明なり，（1895 年 7 号，中島力造「道徳上の權利とは何ぞ」）
② 平均壽命に基づき，掛金を定めて，營業する生命保險會社は，直接に失敗を招き，（1895 年 10 号，志田钾太郎「戰爭保險」）
③ 籬島が自らいふところに據るに此書は宮古歳時記山城名所記行の二書に基づきて作るとあり，（1895 年 8 号，饗庭篁村「都名所圖繪の板元」）
④ 尚ほ此條約に依て獲得したるものは，最惠國條款に基づける將來の利益及び清國内地に於て賣買自由の權利則ち是れなり。（1895 年 7 号，添田壽一「日清戰爭の經濟上の觀察」）
⑤ 此の外交なるものは，各國家の國是に基づいて定めらるゝものであつて，（1909 年 11 号，林董君「世界に於ける各國外交の大勢」）
⑥ 爾後棉花は成るべく米國から取り寄することを力めねばなるまいと勸告したのは右の理由に基づいたのである。（1917 年 13 号，田健治郎「世界的經濟割據の趨勢と船舶管理令」）

『太陽』にある用例を検討してみると，現代語の「に基づく」とほぼ一致する用法で使われていることがわかる。「に基づく」の用例があまり多くないとはいえ, その使い方は『太陽』雑誌が出版される以前にすでに日本語に定着していたと見られる。そのため,「に基づく」の発生時期を明らかにしようとすれば, 明治 20 年代からさかのぼって調査しなければならない。

また,『太陽』にある「に基づく」の用例を中国語に翻訳してみると，そのほとんどは現代中国語の“基于”を含む文に翻訳することが可能である。つまり，翻訳を通してみれば,「に基づく」と“基於（基于）”の間で意味上の対応関係が存在するのが一応確認できる。ただし，重要なのは近代以前の中国語には“基於”の言い方があったかどうかである。

２．２《四庫全書》にある用例

《四庫全書》（電子版）で“基於”を検索すると，1349 例がヒットした。すべての用例が日本語の「に基づく」に対応するものとは限らないが，相当数の用例が確実に見られるということで,“基於”の使い方は近代以前の中国語にはすでにあったことが確認できる。以下では“基於”の用例をあげておく。

[6] 便宜上，以下では各種の活用形を一括して「に基づく」の形で表わす。本稿でとりあげるその他の漢文系複合辞についてもこれと同様な扱い方をする。

① 惟初得陽之一，志專應四而求比於五。以陽感陰，其誠易通。臨大之治，其基於此乎。（南宋张浚撰《紫岩易传》卷 2，1158）〈訳文：陽を以って陰を感じるなら其の誠は通じ易い。大に臨むときの治まりは，これに基づくかな。〉

② 諂者本以求福，而禍常基於諂，梁竇之客是也。瀆者本以交驩，而怨常起於瀆。（南宋项安世撰《周易玩辞》卷 14，諂瀆，1198）〈訳文：こびるものは福を求めるつもりだが，禍は常にこびることに基づくもので，梁竇たちはこのような人だ。…〉

③ 夫謙卑，德也。初卑位也。養德之地未有不基於至卑之所，所養也。至則愈卑而愈不卑矣。（南宋王宗传撰《童溪易传》卷 8，12 世纪末）〈訳文：徳を養う地はいずれも至って卑しい所に基づくもので，これで養われるわけだ。…〉

④ 天下之治不生於富庶之日，而常基於經營勞苦之時。亂不肇於板蕩之秋，而常伏於宴安逸樂之際。（乾隆皇帝弘曆撰《御制日知荟说》卷 1，1735）〈訳文：天下の治まりは裕福の日から生まれるのではなく，常によく働いて苦労する時に基づくものだ。…〉

《四庫全書》にある“基於”の用例を調べてみると,“基於”の後にくる語句の長さは現代語と異なるものの,“基於”自身の用法は現代中国語のそれと大差がなく，一脈相通じていることがわかる。また，これらの用例を日本語に訳すると,「に基づく」を含む文脈になる確率が相当高いことが観察される。しかし,《四庫全書》にある用例は『太陽』雑誌の時代より全般的に古いため，常識的に考えれば，中国語の“基於”が日本語の「に基づく」に先立って成立した可能性が大きいと思われる。

３． に関する／關於（关于）

３．１ 『太陽コーパス』にある用例

『太陽コーパス』には，連体用法の「に関する」の用例が 2227 例あるほか,「に関し・に関して・に関しては・に関しても・に関してゐる・に関してをり・に関しての」などの形を持つ用例は計 1549 例数えられ，両方を合計すると，3776 例になる。これだけの用例が見られるので,『太陽』雑誌が発行された 1895 年の時点では,「に関する」の用法がすでに日本語に定着していたと考えてよい。ただし，この複合辞的用法がいつ発生したかを探るには,『太陽』雑誌より前の時期にさかのぼって調べる必要がある。次にその用例をあげておく。

① 茲に政治學術社會百般の事に關し時事に痛切なる問題に就て，朝野名流大家の卓拔精到なる議論を掲ぐ（1895 年 1 号，久米邦武「学界の大革新」）

② 人皆其の行の是非に關して，些の疑惑をも感ぜざるべく，且必ずや聲を揃へていはん是れ善なりと（1895 年 02 号，坪内逍遥「戦争と文学」）

③ 例へば農業に關しては，彼の地租輕減の如き，假令正當なりとするも現時の經濟に於ては，特に勉めて之を行はんよりは須らく力を農業信用及農業教育の如き積極的保護手段に用ゐざる可からず，（1895 年 1 号，井上辰九郎「経済的闘争」）

④ 一切の政策を此邊の寸法より割出して扨こそ朝鮮東學黨の問題より引續き，彼の國事改革の事に關しても我に對して國交際にあるまじき無禮を働きしことなるに，（1895 年 2 号，福沢諭吉「福沢翁の時事意見」）

⑤ さういふ觀察は，いろ〜〜な方面に向けることが出來ようが，それは略して英國の將來に關しての想像を少し述べて置かう。（1909 年 2 号，姉崎嘲風「名士の英吉利觀」）

⑥ 世間では未だ朝鮮の教育に關する事の外は餘り論究する人も無い様に見えます，併ながら此事は今より豫測して置く事が隨分我々日本人に取つては大切な事と思はれます，（1895 年 1 号，井上哲次郎「戰爭後の學術」）

⑦ 亞細亞に於て直接に露佛兩國の利害に關する問題に就ては露國に取りても將た佛國に取りても英國と協同一致相提携することは望ましき事なりと雖ども（1895 年 2 号，海外彙報「露西亞」）

『太陽』にある「に関する」の意味・用法を検討してみると，現代日本語のそれとほぼ変わらないようになっている。また,「に関する」の用例を中国語に翻訳すると，大体,“關於（关于）”を含む文に訳せることも確認できる。つまり，翻訳を通して見れば,「に関する」と“關於”の意味上の関連性が一応認められるわけである。

３．２　《四庫全書》にある用例

《四庫全書》（電子版）で“關於”の用例を検索してみると，2075 例がヒットした。このうち，意味を成さない文字列も多く含まれているとはいえ,“關於”の用法は近代以前からすでに中国語にあったことは事実として受け止めるべきであろう。その用例は次の通りである。

① 自序云，是傳略採經史關於好惡刑賞治道之大者，凡三百餘條，以繫於篇。（纪昀等编《钦定四库全书总目》卷 21，1789）〈訳文：自序に曰く，この伝記はおよそ経史類から好悪・刑賞・治道に関する大きいものを 300 余条採り，各篇を繋いだ。〉

② 蓋古之言政者必合於禮，言禮者必關於政，如此後世，政在俗吏，禮在書生，遂不可復合，哀哉。（南宋项安世撰《项氏家说》卷 6，12 世纪末）〈訳文：むかし政をとなえるものは必ず礼に合わせ，礼をとなえるものは必ず政に関わる。…〉

③ 容端所謂内顧，即回顧也。不端即斜視也。此等處，不但關於德容，亦且有犯忌諱。（康熙皇帝撰《圣祖仁皇帝庭训格言》，18 世纪前半）〈訳文：…これらの所は，徳容に関わるだけでなく，また忌諱を犯すことも有る。〉

④ 上古民淳事簡，事係於己，惟結繩以記之。事關於人，惟結繩以驗之。不必過為防慮，而天下已治。（清代蒋溥等编《御览经史讲义》卷 8，1749）〈訳文：…自分に係わることなら，ただ縄を結んでそれを覚え，他人に関わる事なら，ただ縄を結んでそれを験すだけだ。…〉

“關於”の用例は，時代を遡れば,“有關於”の形で出てくるものがよく見える。清以降になると，用例数の増加とともに，また，上掲の諸例のように，現代中国語の用法にかな

り近付いてくる。また,“關於”の用例を日本語に翻訳すると,「に関する」や「に関わる」などになる確率が高い。このことからは, 前述の日文中訳に続き, 中文日訳の場合でも“關於”と「に関する」の意味上の対応が一応認められる。ただし,《四庫全書》にある“關於”の用例は,『太陽』雑誌の「に関する」より時代が古いため, 中国語の“關於”が日本語の「に関する」から影響を受けたという前に, まず, 逆の方向で,「に関する」という表現が“關於”から来ている可能性について考える必要があると思われる。

４． に対する／對於（対于）

４．１ 『太陽コーパス』にある用例

「に対する」の用例だけで 3104 例見られる。また,「に対し・に対して・に対しては・に対しても・に対したる」などの形を持つ用例は 5354 例もあり, 漢文系複合辞の中でも最大級の使用量を有している。このことからは,「に対する」の用法が『太陽』雑誌の出版よりも早い時期にすでに定着していたことがわかる。その用例は次の通りである。

① 此れ正か, 彼れ邪か, 此の疑問に對して, 最後の確答を與ふるは, 一に卿等が目下の急務（1895 年 1 号, 坪内逍遥「戰爭と文學」）
② 三國より日本政府に對し, 其の遼東半島を永久に占有するは, 東洋平和のために非ずと申込めるは, 實に四月二十三日なり（1895 年 6 号, 政治「日本と三國との交渉」）
③ 而して從來歐米人に對しては, 既に開港したる重慶港を, 新たに開放すと明言するは, 無用の條件なるに似たるも, 實は然らず, （1895 年 5 号, 商業「新條約に伴ふ通商上の利益」）
④ 英國は, 支那に對しても, 亦久しく此政策を執れり。（1895 年 1 号, 尾崎行雄「對清政策」）
⑤ 通商上列國に對する關係益々頻繁を加ふべきや逆睹すること寔に易々たり, （1895 年 1 号, 井上辰九郎「經濟的鬪爭」）
⑥ 故に將來我帝國は清國に對する用意より論ずるも, 一般尚武的の精神を遍ねからしむるの必要あるのみならず, （1895 年 1 号, 千頭清臣「戰勝後の教育」）
⑦ 祖父母父母に對したる殺傷の罪は特別の宥恕及び不論罪の例を用うることを得ず（1901 年 3 号, 岡田朝太郎「法律時評」）

日本の国語辞典には「対する」の解釈として,「二つのものが向かい合う」と「対象とする」の 2 項目があるが, その区別については, 前者は具象的事物の関係を表わす場合で, 後者は抽象的事物の関係を表わす場合をさすだろうととらえられる。『太陽』雑誌では,「に対する」の用法は, ほとんど後者の意味に使われており, 現代日本語の用法とも一致する。また,「に対する」の用例を中国語に翻訳してみれば, 大体, 現代中国語の“对于”を含む文になることから, 日文中訳の場合は,「に対する」と“对于”の間で, 意味上の対応がなされていると考えてよい。ただし, その影響関係を解明するには, 近代以前の中国語に“對於”の用法があったかどうかを調べるのが前提になる。

４．２ 《四庫全書》にある用例

“對於”を《四庫全書》で検索してみると，1939 例がヒットした。これによって，近代以前の中国語には“對於”の語形が存在していたことが確認された。ただし，用例を実際に検討していくと,“對於”が種々の意味に使われていたことに気付く。その用例は次の通りである。

① 生地黄搗爛，煨熨於有瘡處，更妙用陰二陽四丹，對於有瘡處吹。（唐代孙思邈撰《银海精微》卷上，682）〈訳文：…さらに「陰二陽四丹」を巧みに用いて，腫れ物の所に向かって吹くようにする。〉

② 十二年春，亮悉大衆由斜谷出，以流馬運，據武功五丈原與司馬宣王對於渭南。亮每患糧不繼，使己志不伸。（西晋陈寿撰《三国志•蜀志》卷 5，3 世纪后半）〈訳文：武功の五丈原を占拠し，渭南で司馬宣王に対抗していた。〉

③ 凡助祭者，實皆秉持文王之德，對於文王在天之精神。（南宋宋林撰《毛诗讲义》卷 9，1190）〈訳文：凡て祭を助けるものは，実はみな文王の徳を持って，天国に在る文王の精神に応対している。〉

④ 以水火對于乾坤，成乎四象，不易之理氣也。（清代魏荔彤撰《大易通解》卷 6，17 世纪后半）〈訳文：水と火で乾坤に対応して，四象を成すのは変わらない理の気だ。〉

⑤ 惟守此中道，利害有所不計，而常奉敬謹之心，以對于神明。則憂患之至，皆所以增益其所，不能而何咎之有乎。（清代程廷祚撰《大易择言》卷 22，18 世纪前半）〈訳文：…常に敬謙な心を奉って，神様に応えている。〉

以上の用例でわかるように,《四庫全書》にある“對於”は，日本語の「に対する」と違って，抽象的事物の関係を表わすというよりも，むしろ,「向かい合う・対抗する・応える」などの具体的な動作動詞として使われていたと見られる。そのため，日本語の「に対する」に直訳できるものはたいへん少なくなっている。これに先立って,『太陽』にある「に対する」がほぼ現代中国語の“对于”に直訳できると述べたが，日文中訳と中文日訳の結果を比較すると，両者の相違が顕著に見られる。

《四庫全書》にある“對於”の用例が『太陽』雑誌の「に対する」よりも年代が古いので，中国語の“對於”が漢文訓読によって日本語の「に対する」の形になり，しだいに定着したという経緯が考えられる。ただし，日本語に定着した後の「に対する」は，漢籍の用法から徐々に離れ，具体的な動作動詞ではなく，抽象的事物の関係を表わすようになったと推測される。『太陽』にある「に対する」の用例は，まさに意味・用法の変容を遂げた後の状況をよく表わしているものといえよう。

一方，現代中国語における“对于”は，すでに《四庫全書》時代の用法から遠ざかって，どちらかといえば,『太陽』雑誌や現代日本語のそれに近付いている。この変化の裏には明治以後の日本語からの影響があったかどうかについては，20 世紀初期から増え始めた日本語から中国語に翻訳された出版物を調べることによって解明する必要がある。

５． に由る／由於（由于）

５．１ 『太陽コーパス』にある用例

日本語の「による」には昔から幾通りの漢字表記が用いられてきた[7]。中国語の“由於”との対応を考えて,「に由る・に由り」などの語形で検索してみると，連体用法が中心となる「に由る」は 269 例あるほか,「に由り・に由りて・に由りては」などの活用形を含む用例は計 308 例見られる。ちなみに，その他の漢字表記が用いられた「による」についても調べてみた。たとえば,「に依る・に依り」系の用例は 2197 例見られ,「に因る・に因り」系は 425 例で,「に據る・に據り」系は 325 例となっている。

一方，漢字表記のない「による・により」で検索すると,「による」は 622 例であるが,「により・によりて・によりての・によりては・によりても」などの活用形では，計 2675 例にも及んでいる。これを踏まえて，漢字表記のない「による・により」は『太陽』の時代から主流となっていたことが指摘できる。つまり，王立達（1958a）で言及した中国語の“由於”と日本語の「に由る」の対応は，たとえ事実であったとしても，少数の用法との対応になるわけである。以下では,「に由る・に由り」の用例のみあげておく。

① 巴里の五銀行は，露國政府の保證に由り，四分の利と九十乃至九十二の價格（券面百圓に對し九十圓乃至九十二圓）を以て，一千六百萬磅（約一億六千萬圓）を清國政府に貸與するに決せり。（1895 年 7 号，時事「日露新條約成る」）
② 蓋し人は境遇に由りて思想を變ずるものなれば境遇に由りて好尚を異にするは吾人が往々實驗する所なり然れども（1895 年 4 号，大江敬香「明治詩家評論」）
③ 例へば官吏の登庸試驗は，優者を登庸する所以に異ならずと雖も，試驗官たる淘汰者の如何に由りては，反對の結果を見ることなしとせず,（1895 年 3 号，千頭清臣「戰下・側面的觀察」）
④ 然れども通して之を論ずるに，目に由る記憶，耳に由るよりも易きは，疑を容るゝを得ず,（1895 年 1 号，三宅雪嶺「漢字の利害」）
⑤ 將士の忠勇義烈に由ると雖も，抑もまた軍隊組織，作戰計畫等の其の宜しきを得たるが故にして,（1895 年 9 号，大岡育造「新製艦案」）

上掲の用例でわかるように,『太陽』にある「に由る」は現代日本語の用法とほとんど変わらないようになっている。ただし,「に由る」の用例は現代中国語に翻訳してみると，必ずしも中国語の“由於”を含む文脈になるとは限らない。王立達（1958a）では“由於”が「に由る」を翻訳するときに中国人によって造られたものだとしているが,『太陽』の用例で検証した結果,「に由る」と“由於”の間で，一対一のような対応関係が常に存在するわけではないことが明らかになった。

５．２ 《四庫全書》にある用例

《四庫全書》（電子版）で検索すると,“由於”の用例は 7646 例も数えられている。それに，用例が上古から清の時代まで広範囲に分布していて，昔から中国語にあった表現であ

[7] 山田孝雄（1935）は第 29 項で，漢籍中の「由・因・緣・仍」などの訓読法と「よりて」の影響関係を論じているが,「由於」についての言及がない。

ることが立証できる。また,《四庫全書》にある“由於”の用例を検討してみると，現代中国語の“由于”とほぼ同じ意味・用法で使われていることから,“由於”の用法は古代から現代まで一脈通じていることがわかる。次にその用例をあげておく。

① 制度明則民用足，制度不明由於名不正。正名之道，所以明上下之稱班，爵號之制，定卿大夫之位也。(东晋袁宏撰《后汉纪》卷 16，4 世纪中期)〈訳文：制度が明らかであれば庶民の物資が足りるが,制度が明らかでないのは名が正されていないことに由る。〉
② 或因罪而致高，或處危以成名。所以天災屢降，治道未寧，皆由於此也。(东晋袁宏撰《后汉纪》卷 18，4 世纪中期)〈訳文：…天災に度々見舞われ，統治が未だに安泰でないのはみなこれに由るものだ。〉
③ 婦人之病，皆由於月病生産所致，又從胞胎所起。(唐代王焘撰《外台秘要方》卷 34，752)〈訳文：婦人の病気はみな月経や分娩に由って起きたもので，…〉
④ 古者，人稠地狹而有儲蓄，由於節也。今者，土廣人稀而患不足，由於奢也。(北宋司马光等撰《资治通鉴》卷 81，11 世纪后半)〈訳文：昔は，人口が多く土地が狭いのに貯蓄があるのは節約に由るものだが，今は，土地が広く人口が少ないのに物不足になるのは奢侈に由るものだ。〉
⑤ 然中風病不論寒多風多，大槩由于虛故，首尾不脫虛字，而淺深則自不同耳。(清代徐彬撰《金匮要略论注》卷 5，1671)〈訳文：脳卒中は寒気や風邪を問わず，だいたい虚弱に由るためだ。〉

注目したいのは,《四庫全書》にある“由於”の用例を日本語に翻訳すると，大体,「に由る」を含む文脈になることである。これに先立って，日文中訳では,「に由る」を中国語に翻訳すれば,“由於”を含む文脈になる場合もあれば，ならない場合もあることを述べたが，これに対して，中文日訳では，いま指摘したように,“由於”が日本語の「に由る」になる確率が高いので，両方のアンバランスがはっきり見受けられる。

なぜ日文中訳と中文日訳では対応関係の一致度が異なるだろうか。この点については,“由於”と「に由る」の影響関係と関連付けて考える必要がある。“由於”の用例には相当古いものが見られるので，おそらく,“由於”の言い方が漢文訓読によって日本語に移入され,「に由る」の形で定着した経緯があったかと思われる。「に由る」は漢籍の“由於”に由来しただけに，中文日訳では中国語の“由於”が日本語の「に由る」に訳されやすいことにつながったのではないか。一方,「に由る」は徐々に用法上の変容が進み，現代中国語の“由于”でカバーしきれない部分が出てきたため，日文中訳では,「に由る」の訳が“由于”になったりならなかったりする現象が現れたと考えられよう。むろん，この推論を実証するには，19 世紀末に中国で起きた日本書の翻訳ブームにおいて,「に由る」の文脈はどのように中国語に翻訳されていたかを実際に調べる必要がある。

６． と認め／認爲（认为）

６．１ 『太陽コーパス』にある用例

『太陽コーパス』には,「と認め・と認めて」を主とし,「と認めず・と認めざる・と認

めた・と認めらるる・と認められ」などの少数用法を合わせて，計 289 例の用例が見られる。ちなみに，王立達（1958a）では，中国語の“認爲（认为）”は日本語の「ト認メラレテ」の訳語にあたるといった考えを述べたが,『太陽』には「と認められて」の用例が見当たらなかったので, よく使われる表現ではないことがわかる。現代日本語では,「と認める」を複合辞として扱うことはないようだが，ここでは，中国語の“認爲”との関連を考えるという目的で,「と認める」をとりあげることにした。その用例は次の通りである。

① 愛知縣廳は右畫様を以て風俗を壞亂するの恐あるものと認め, 其發賣の自由を抑へしかば,（1895 年 7 号，海内彙報，法律界「裸婦人畫問題」）
② 本邦にては土壤を以て作物に肥料を與ふるの機關と認めて各作物に之れを施すを常とせり（1895 年 5 号，農業，矢部規矩治「本邦の氣候と施肥の關係」）
③ 是が則ち時勢の大躰經濟社會の全部に於て政府は最早や國立銀行延期の必要なしと認めたのであります,（1895 年 2 号，商業「國立銀行問題」）
④ 之を概するに彼等の眼には, 戰爭は一の兇事と認めず, 寧ろ殖民出稼の如く思惟し, 相爭ふてその役に就く。（1895 年 7 号，藤田精一「蒙古大王拔都の西歐侵掠」）
⑤ 工業の本領に着眼するの遲き，吾人は之を以て我が經濟界の不幸と認めざるを得ず。（1901 年 9 号，輿論一斑「工業政策」）

『太陽』雑誌の範囲では，文中用法の「と認め・と認めて」の形の用例が多く,「と認める」の形で文末に置かれる用例が少数に止まっている。このことからは,「と認め・と認めて」のような文中用法が典型的な形と考えられ，しかも，その用例を調べてみると，現代日本語のそれとほぼ変わらないようになっている。

また，中国語の“認爲（认为）”との対応関係について検討してみると，文中用法の「と認め，と認めて」が中国語に翻訳される場合，基本的には“認爲（认为）”を含む文脈になれるが，一方，文末用法・連体用法の「と認める」は,“認爲（认为）”に訳されるものもあれば，原文の文脈によっては“認爲（认为）”のかわりに,“承認”などで訳さなければならないものもある。その理由として,「と認め」が複合辞化しておらず，動詞「認める」の本来の意味が保たれているため，置かれた文脈に左右されやすいことが考えられる。

６．２ 《四庫全書》にある用例

《漢語大詞典》には“認爲（认为）”の見出し項目があり,“对人或事物确定某种看法，做出某种判断。”（人や物事に対して見方を示したり判断を下したりすること。）という語釈とともに，近代以前の用例も掲載されている。これは，ある意味で，中国の古典にはすでに“認爲（认为）”の語形があったことを示したものである。これを裏付けるために，さらに《四庫全書》で検索すると，663 例がヒットした。その用例は次の通りである。

① 今人只為誤将壯字認為好境，故全局皆差。（明代魏濬撰《易义古象通》卷 5，17 世紀初）〈訳文：今の人は誤って「壮」の字だけを佳境と認めているので，全般的に劣るわけだ。〉

② 父祖質産於人，子孫不能繼贖，更數十年，時事一變，皆自陳於官，認為故産。吾安得言質而復取之？（明代陈邦瞻撰《宋史纪事本末》卷 20，1605）〈訳文：…さらに数十年も経って時事が一変すると，みな自ら官庁に申し立てて，自分の古い財産だと見なしてもらおうとする。〉

③ 回民三百餘戶，懇請回家，收割田禾。恐本庄及沿途村民認為賊黨，致被擒拿。（清代纪昀等编《钦定石峰堡纪略》卷 15，1789）〈訳文：恐いのは，この村または沿道の村人に泥棒と見なされて，捕まえられてしまうことだ。〉

④ 而李時珍認為一物，亦不知玉蘭用辛夷接植而成，皆未深考又伏讀。（清代嵇璜等撰《钦定续通志》卷 176，1785）〈訳文：しかし，李時珍は同一の物だと見なし，また，玉蘭は辛夷との接木によってできたことも知らなかった。〉

“認爲”の用例を検討してみると，そのあとに続く目的語が文か語かという構文機能で，現代中国語とややずれているものの，“認爲（认为）”自身の意味に関してはほとんど差がないように思われる。つまり，近代以前の“認爲”と現代中国語の“认为”の間には，意味・用法上のギャップがなく，一脈相通じているといってよい。

ただし，日文中訳の場合では，日本語の「と認め」は大半，中国語の“認爲”に翻訳されると前述したが，それに比べて，中文日訳の場合では，中国語の“認爲”が日本語の「と認め」に訳されることもあるものの，それよりも，「と見なす，と見なされる」などの訳になるケースが案外多いようである。いわば，“認爲”と「と認め」の間では一致度の高い対応関係が見出せないということになる。一方，現代中国語の“認爲”が果たして日本語の「と認め」の影響を受けて成立したものかどうかについては，20 世紀初頭の日文中訳の実例を細かく調査し，「と認め」が“認爲”に訳されていたかどうかの実態をとらえる研究が求められる。

７． と成る／成爲（成为）

７．１ 『太陽コーパス』にある用例

検索した結果，「と成り・と成りて・成りぬ・成りける」などを合わせて 76 例があるほか，「と成る」は 53 例見られる。また，王立達（1958）には言及しなかったが，同じ漢字表記を有する「に成る」や「を成す」も中国語の“成爲”（成为）と対応する可能性もあるので，これらについても調べてみた。その結果，連体・文末用法の「に成る」だけで 198 例見られるほか，「に成り・に成りて・に成りては・に成りし・に成りたる・に成ります」などを含め，計 200 例ほど数えられる。

一方，同じ用法で，仮名表記の「となり」と「となる」はそれぞれ 4413 例と 2484 例に達しているので，漢字表記の「と成る」や「に成る」はどちらも主流的用法ではないことがわかる。王立達（1958）では，中国語の“成爲”（成为）が日本語の「と成る」を中国語に翻訳するときに造られたものだというが，そうであれば，仮名表記のない「となる」はどのように中国語に訳されていたかといった疑問に答えなければならない。

また，「を成す」は 142 例で，「を成し，を成して，を成したり，を成したる」などの活用形を持つ用例は計 199 例となっている。次に「と成る」「に成る」および「を成す」の用例をあげておく。

① 後同銀行社員と成り，廿二年辭職して東京朝日新聞社に小説の筆を執らるゝことを遺漏したれば之に附記す。（1895 年 2 号，幸堂得知「古今演劇談」）
② 余が一等書記官と成りて使節に歐米に從はん事を勸め給へり，（1895 年 4 号，史傳，福地源一郎「維新の元勳」）
③ 是等の人々の中には遠足旁々慰みを主として行くも有り，何か玩弄物と成る物を獲ようと云ふ好事心に促されて行くも有り，（1895 年 9 号，坪井正五郎「石器時代遺跡の實踐は人類學上如何なる利益有りや」）
④ 正字通，字彙，佩文韻府，康熙字典等，皆明清の間に成り，近時に至りて又語法音源の出づるあり。（1895 年 3 号，三宅雪嶺「字音の標準」）
⑤ 大使の米歐回覽實記は多く君の手に成ると云、方今職を辭して著作に從事せられ、（1895 年 1 号，久米邦武「學界の大革新」）
⑥ 盖し絶大の人物は絶大の事業を成し，絶大の事業は絶大の人物を生じ，（1895 年 1 号，中西牛郎「日本帝國の任務」）
⑦ 社會全般の事物は社會を成す所の國民の事業なれば，國民の性質智能を明かにし，（1895 年 2 号，坪井九馬三「史料の編纂は目下の急務たるを論ず」）

『太陽』にある「と成る」「に成る」や「を成す」の用例を検討してみると，1895 年の時点では，この三者の用法はすでに現代日本語のそれとほぼ一致していることがわかる。したがって，この三者の発生期を求めようとすれば，『太陽』以前あるいは明治初期に遡って調べる作業が必要になる。また，中国語の“成爲”（成为）とこの三者の対応について見ると，「と成る」の用例が中国語に翻訳されると，漢字表記の“成”に引かれるせいか，“成爲”になることが多いが，「に成る」の用例は漢字表記のない「になる」の多岐的な用法と混同しているため，その大半はむしろ“成爲”以外の訳になっている。「を成す」は他動詞的な文脈を構成するのが普通なので，中国語の“成爲”で訳されることはあまり多くないように見られる。

７．２ 《四庫全書》にある用例

《四庫全書》で“成爲”を検索してみると，3840 例がヒットした。その中で，10 世紀以前の用例も少なくないが，宋代以降の用例なら，もっと現代中国語の用法に近付いているといえる。その用例は次の通りである。

① 奇偶之策，總而言之有三百六十。是故，聖人因其乾坤奇偶之數成為一歲，凡三百六十日也。（北宋胡瑗撰《周易口义》系辞上，11 世纪前半）〈訳文：そのため，聖人はその乾坤や奇偶の数によって一歳と成り，すべてが三百六十日だ。〉
② 正月三陽既上，成為乾卦，乾體在下，三陰為坤。（南宋真德秀撰《西山读书记》卷 37，13 世纪初）〈訳文：正月には三陽が既に上がって，乾の卦と成る。…〉
③ 管子云，一農之事，必有一銍一椎，然後成為農。（明代徐光启撰《农政全书》卷 21，1639）〈訳文：管子が云わく，農民とは必ずくわとつちがあってはじめて農民と成る。〉

④ 惟君子能自强不息，亦惟自强不息，乃成為君子。（清代蒋溥等编《御览经史讲义》卷 1，1749）〈訳文：…また，自らずっと強い者になっているだけに，すなわち君子と成る。〉

以上の諸例でわかるように，中国語の“成爲”を日本語に翻訳してみると，大体,「と成る」と対応することになる。ただし,《四庫全書》にある用例が年代的に古いことを踏まえて考えれば，中国語の“成爲”が日本語の「と成る」に影響されてできたというよりも，中国語の“成爲”には日本語の「となる」の意味・用法と対応する部分があるので，漢文訓読の際,「と成る」の形で“成爲”に対応させたという可能性があるかもしれない。

８． と視る／視爲（视为）

８．１ 『太陽コーパス』にある用例

王立達（1958a）では，中国語の“視爲”は日本語の「〇〇ト視ナシテ」の影響を受けてできたものだとしている。しかし,『太陽コーパス』で検索してみると,「〇〇ト視ナシテ」の形で使われる用例が見当たらなかった。「と視て」（2 例）,「と視る」（9 例）,「と視れば」（2 例）の用例も少数にとどまっている。このほか,「と視做され・視做し」はそれぞれ 4 例,「と視為す」（1 例）のように，漢字表記「視」を持つ動詞は全般的に用例が少ないことがわかった。

一方，異なる漢字表記を持つ「とみる」のその他の用例を検索してみると,「と見る」は 273 例,「と見て，と見ては，と見ても，と見てゐる，と見ておる」は 239 例が数えられるほか,「と看る，と看て」は 7 例で,「と観る，と観て」は 13 例となっている。漢字表記のない「とみる，とみて」は 5 例しかない。つまり,「とみる」の各語形の中で,「と見る」が最もよく用いられたものだとわかる。この結果によって考えると，王立達（1958a）の言う「〇〇ト視ナシテ」はもちろん,「と視て・と視る」などを含めて，どちらも明治期によく使われた表現ではないことになる。その用例の中から一部だけ示しておく。

① 其時小説家は如何にも怨めしげに自分の顔を睨と視て, なぜですか。（1895 年 3 号, 窓下几上生「新聞小説」）
② 故に生物生活の優劣は其機關の精粗如何と視る, 生物壽命の長短も亦其機關の精粗如何と視る, 而して衆生物の其最も自己に適合する境遇を爭ふもの, 之を生存競爭と稱し,（1895 年 11 号，中西牛郎「文學界の遷流及文學者の壽命」）
③ 敵國艦隊は東西に出沒し, 前に在るかと視れば忽焉として後にあり, 先づ沿海の海運を杜絶し其の商工業を第一に妨害すべく,（1895 年 6 号, 記者「某大佐の兵事談」）
④ 彼支那の女をば男のために子を生み養育する物と視做したるとは全く風俗を異にせり。（1895 年 8 号，久米邦武「倫理の改良」）
⑤ 英國保守黨の機關と視做さるゝスタンダード新聞の論に曰く（1895 年 9 号,「英國の日英同盟論」）
⑥ 但し右年限の滿ちたるときは未だ該地方を去らざる住民を日本國の都合に因り日本國臣民と視爲すことあるべし（1895 年 6 号,「日清講和條約」）

このうち，例⑥の「視爲す」は，例④と例⑤の「視做す」と同じように,「みなす」と発音すべきもので，たまたま「視爲」の漢字表示になっただけだと推察される。その他の「と視る」の用例を調べてみると,「と見る」の意味・用法とはほぼ共通していることがわかる。また,“視”は「みる」の意を表す古典中国語の動詞で，近代以後の中国語では“看”や“看见”などにとってかわられた経緯があるので,「と視る」の用例が現代中国語に翻訳される場合,“視爲”の訳になりにくいのもあたりまえの結果ともいえよう。

８．２ 《四庫全書》にある用例

《四庫全書》で“視爲”を検索してみると，2987 例がヒットした。近代以前の中国語に“視爲”の使い方がすでにあったことは，これによってわかる。また，日本語の「と視る」は,「と見る」に比べて，少数派の用法に過ぎないことも先に述べたし，これらの理由を踏まえて考えると，出典の古い“視爲”が日本語の「と視る」に影響されてできたという王立達（1958a）の説は根拠が足りないということになる。“視爲”の用例は次の通りである。

① 上都，聖上龍飛之地，天下視為根本。（明代宋濂撰《元史》卷 126，1369）〈訳文：上都は皇帝一族の聖地で，世の中はこれを根本と見なしている。〉
② 自度無能復進，乃筆其區區之見，以與朋友講之。然視為老生常談，一覽而遂置之者多矣。（明代罗钦顺撰《困知记》附录，16 世纪前半）〈訳文：…しかし，ありふれた平凡な話と見なされ，一見してすぐ放っておくものはいかに多いか。〉
③ 至於農桑學校，王政之本，乃視為虛文而置之，將何以教養斯民哉？（清代张廷玉等撰《明史》卷 139，1739）〈訳文：農業と学校は王政の本なのに，虚文と見て放っておいたら何を以って庶民に教養を与えるのか。〉
④ 凡先王大典，皆視爲粗迹，無足怪也。（清代纪昀等编《钦定四库全书总目》卷 25，1789）〈訳文：凡そ先王の典籍なら，みな至理名言と見なすのはあたりまえのことだ。〉

《四庫全書》にある“視爲”の用例を検討してみると，その意味・用法が現代中国語のそれとほぼ一致することがわかる。また，翻訳を通して日中双方の対応関係を見てみると,“視爲”が現代日本語に翻訳される場合は,「と視る」の表記はもちろん使われないし,「と見る」になることもまれで,「と見なす・と見なされる」といった訳が多く見られる。これに先立って，日文中訳の場合では,「と視る」と“視爲”の対応関係が必ずしも固定されてはいないことを指摘したが，それに加え，中文日訳の場合でも“視爲”と「と視る」の一致度が低くなっている。結論をいうと，王立達（1958a）の前説は一種の臆測に過ぎないものである。

９． まとめ

これまでに述べてきたことについて，以下の諸点にまとめておきたい。

（1）「複合辞」はまた「複合助辞」とも呼ばれる。森田良行・松木正恵（1989）は,「どのような表現を複合辞と呼ぶのかという基準について定説はないが，本書では，単なる語の連接ではなく，表現形式全体として，個々の構成要素のプラス以上の独自の意味が生じ

ていることを一つの目安とした」と述べている[8]。また，飛田良文（2007）は，「複合助辞」について，「形式化した語や助詞・助動詞が複合して，一つの付属語のように機能する表現形式をいう」という定義を与えている[9]。つまり，「個々の構成要素のプラス以上の独自の意味が生じていること」や「形式化」は，複合辞の特徴を考えるときのポイントとなっている。これを踏まえるなら，本稿でとりあげたものの中で，「と認める・と視る」などは必ずしもその枠に入るわけではないが，中国語には対応できそうな表現があるので，一応日中比較の対象として取り入れたのである。

（2）国語史の分野において，日本語の語彙と表現にもたらした漢籍漢文の影響を対象とした研究成果が豊富に蓄積されている[10]。なかでも，山田孝雄氏はその著『漢文の訓読によりて伝えられたる語法』（1935）で，40 の項目にわたって，「ごとし・いはく・ねがはくは」など 58 種の表現と漢文訓読の関わりを論じた内容は本稿との関連がとくに強い。ただし，同著では，「よりて」と「由・因・縁・仍」などの漢字の訓読法に言及した一節があったものの，漢文の“由於”には直接触れていない。また，本稿でとりあげたその他の複合辞と漢文訓読の影響関係についての研究もないようである。

（3）複合辞研究の課題について，『日本語学』（特集・複合辞）の巻頭にある「特集の趣旨」（1989）では，「複合辞あるいは複合助辞，すなわち複合のかたちで助詞・助動詞等相当のはたらきをするものは，近代日本語の特徴的形式の一つであって，古代日本語にはごくすくないようだ。……大局からして，複合辞の発達は，文法的にも語彙的にも，近代日本語の特徴の一つだと言っていい」などと述べている[11]。また，飛田良文（2007）は「日本文法の体系の中に位置付け，一語一語の歴史を解明していく必要がある」としている[12]。これらの論述はいずれも複合辞に関する通時的研究の重要性に言及しているが，筆者の管見では，いまだにこのような課題を取り上げた論文は見当たらないようである。

（4）「漢文系複合辞」とは耳慣れない言葉であるが，筆者は次のように考えている。複合辞の中には，明治期の漢文体に由来したと思われるものがある。たとえば，本稿でとりあげたものはその一部である。このうち，「に関する・に対する」のように音読みのものもあれば，「に基づく・に由る・と成る」のように訓読みのものもあるが，さしあたり，「漢文系複合辞」と呼ぶことにする。これに対して，和文系の文学作品や口語体の文章に由来したと思われるものについては，「和文系複合辞」と名付けることができる。その例として，「するやいなや・からには・ものなら・より仕方がない」などがあげられよう。複合辞の歴史を探るような研究では，この「漢文系」と「和文系」の分け方が役に立つだろうと思われる。

（5）本稿でとりあげた複合辞は，結果的にはいずれも『太陽』雑誌が出版された時点ですでに出来上がっていた。そのため，これらの複合辞の発生を解明しようとすれば，『太陽コーパス』だけでは力が及ばず，明治初期をカバーするコーパスが必要になってくる。今度の調査でわかるように，近代語のコーパスは語彙の歴史を探るときに役立つだけでな

---

8 森田良行・松木正恵（1989）の凡例（p.xi）を参照。

9 『日本語学研究事典』の p.221 を参照。当該「複合助辞」の項目は飛田良文氏の執筆による。

10 参考文献にある山田孝雄（1935，1940），築島裕（1963），小林芳規（1967），および柏谷嘉弘（1987，1997）などは代表的なものとしてあげられる。

11 『日本語学』（特集・複合辞）1984 年 10 月号，明治書院

12 飛田氏が執筆した「複合助辞」の項目（『日本語学研究事典』の p.221）を参照。

く，飛田（2007）のご指摘の通り，複合辞の「一語一語の歴史を解明していく」上にも大いに利用されるべきものである。

（6）本稿は王立達（1958a）の仮説を検証するための調査であるが，結果から言うと，《四庫全書》にある用例が年代的に早いので，中国語の“基於、關於、對於、由於、認爲、成爲、視爲”は日本語の影響で新しく造られたという可能性が低い。ただし，こういった言い方は近代以前の中国語にはすでにあったことが明らかになったものの，20 世紀初期の日本書翻訳ブームの中で，日本語の「に基づく・に関する・に対する」などがどのように中国語に訳されていたか，日中間の対応関係が実際にあったかどうかについては，推測ではなく，当時の資料によって実証する必要がある。また，王立達（1958a）の説とは逆に，これらの日本語の複合辞はいつ，どのように形成されたのか，中国語からの影響があったかどうかについても，別の視点からの研究が必要なので，今後の課題としたい。

文　献

山田孝雄（1935）『漢文の訓読によりて伝えられたる語法』宝文館

山田孝雄（1940）『国語の中に於ける漢語の研究』宝文館

国立国語研究所報告 3（1951）『現代語の助詞・助動詞―用法と実例』秀英出版

永野賢（1953）「表現文法の問題―複合辞の認定について―」『金田一博士古稀記念言語民俗論叢』三省堂

王立达（1958a）〈现代汉语中从日语借来的词汇〉《中国语文》2 月号

郑奠（1958）〈谈现代汉语中的“日语词汇”〉《中国语文》2 月号

张应德（1958）〈现代汉语中能有这么多日语借词吗？〉《中国語文》6 月号

王立达（1958b）〈从构词法上辨别不了日语借词―和张应德同志商讨汉语里日语借词问题―〉《中国语文》9 月号

簗島裕（1963）『平安時代の漢文訓読語につきての研究』東京大学出版会

小林芳規（1967）『平安鎌倉時代に於ける漢籍訓読の国語史的研究』東京大学出版会

『日本語学』（特集・複合辞）1984 年 10 月号，明治書院

柏谷嘉弘（1987）『日本漢語の系譜―その摂取と表現』東宛社

森田良行・松木正恵（1989）『日本語表現文型―用例中心・複合辞の意味と用法』株式会社アルク

河原崎幹夫監修（1995）『辞書で引けない日本語文中表現』北星堂書店

柏谷嘉弘（1997）『続日本漢語の系譜』東宛社

砂川有里子ほか（1998）『日本語文型辞典』くろしお出版

国立国語研究所報告 122（2005）『雑誌〈太陽〉による確立期現代語の研究』博文館新社

飛田良文（2007）「複合助辞」『日本語学研究事典』明治書院

コーパス

《文渊阁四库全书电子版》香港迪志文化出版有限公司、上海人民出版社，1999

『太陽コーパス―雑誌『太陽』日本語データベース』国立国語研究所資料集 15，博文館新社，2005

日中の比較語史研究

陳　力衛（成城大学）[1]

１．問題提起

日本漢語の語史について、従来、現行辞書（『日本国語大辞典』、『大漢和辞典』、『漢語大詞典』）の初出などで跡付けようとするのが普通であった。それをふまえてさらに中国由来の漢語と和製漢語とを弁別する手順としてもよく利用されてきた。しかし、たとえば『大漢和辞典』は収録語の時代的な偏りにより、近現代の用例採集の不足が明らかであり、同じく日本漢文の用例採集が少ないのが欠点であろう。それに未登録語が多い分、それらの処理には和製漢語と看做すか、単純に漏れたかの区別もつきにくいという問題が残る。一方の『漢語大詞典』はほとんど洋学資料と英華字典を使わないし、見出し語でも用例でも近代語を無視または軽視している傾向が顕著である。したがって辞書編集のもろもろの制限からその手法の有効性に次第に疑いの目を向けるようになり、とくにデータベースやコーパスの構築により、辞書の記述をいろいろと修正しなければならなくなるころに来ている。近代語に限っていえば、辞書の不足や資料の不備がつねに課題の一つとして挙げられている。そこで最近中国で出版された『近現代辞源』（黄河清編、上海辞書出版社、2010）を材料に、いわゆる日中言語交渉を視点として、近代漢語の語史の構築を分析しようと試みる。本稿ではその際に出くわしたさまざまな問題を取り上げて、いわゆる近代的意味の不確定さと文芸中心主義による初出例の選出を指摘しつつ、近代語のコーパス利用によって修正できる範囲と可能性について考えたい。

２.中国語資料を手掛かりに

『近現代辞源』の序文によると、本辞書は「明末清初から 1949 年前後の、西洋文化の影響を受けて成立した言葉、できれば翻訳語」を 9500 語収録している。故に、近代中国の対外交流に関する言語資料を多く使用しているため、『漢語大詞典』の欠点を補う点で期待されている。これまで 2001 年に、5000 語収録の『近現代漢語新詞詞源詞典』（漢語大詞典出版社）を上梓した後、さらに語数を増やし 9500 余語を選択し編纂したのがこの『近現代辞源』である。前著との大きな違いは現代語のみを収録し、歴史語は収録していないことにあるという。

事実、この辞書の特徴といえば、何といっても用例の豊富さと、できるだけ語源をとことん追究し、中国語文献での早期の例証を列挙することが挙げられる。大塚秀明はその点を高く評価し、『近現代辞源』によっていままでの辞書の語の初出年を引き上げることができたと、下記の例で説明した[2]。

(1)前著『近現代漢語新詞詞源詞典』の不明点が『近現代辞源』では①～③の初出年を→の後の年に引き上げ、④～⑤の見出しを追加し、⑥の用例を新たに補充した。たとえば、

①打：1904→1889『遊記日本図経』、1891『格致彙編・格致釈器』
②肥料：1905→1878『格致彙編・化学衛生論』、1899『日本各地紀略』
③汽水：1917→1897『進口雑貨税則』、1901『中英新訂商約』
④商標：1889『遊記日本図経』、1902『東遊叢録・学校図表』

[1] chenliwei@seijo.ac.jp

[2]「初出、或いはそれに近い用例の記述と辞書の収録をめぐって」『国際シンポジウム「近代語の語源研究とその周辺」要旨集』漢字文化圏近代語研究会、関西大学文化交渉学教育研究拠点共催、2011.3.19

⑤銀行：1854『遐邇貫珍』、1859『資政新篇』
⑥自行車：1870『教会新報』、1891『格致彙編』

(2)『近現代漢語新詞詞源詞典』収録の「－品」10 語のうち６語が『近現代辞源』で初出年が更新されている。

①作品：豊子愷 1928　→　梁啓超 1922
②食品：清議報 1899　→　遐邇貫珍 1853
③商品：清議報 1900　→　日本国志 1890
④印刷品：近代教育 1918　→　美国視察記 1915
⑤戦利品：飲氷室合集 1920　→　日本留学参観記 1904
⑥装飾品：飲氷室合集 1910　→　博物学教科書 1906

そうしたより早くかつ確かな初出例を並べることで従来の辞書より一歩進んでいることが確認できる。となると、それをふまえていわば日中同形語の初出の比較にも役立つことになる。つまり、いわゆる日中言語交渉の視点から、当該辞書と日本の現行辞書の初出例を比較して、中国近代語にどれくらいの日本語が入って行ったか、あるいは逆に、日本近代語にどれくらいの中国語が入ってきたかを明らかにできる可能性が出てくるのではないかと思われる。事実、これまで現代中国語の 3000 語における日本借用語の調査を行い、意味分野別における日本借用語の割合をある程度解明したが、もっと語数を増やしてその全体像をつかもうとしているところへ、『近現代辞源』を手にして、資料の充実さに加えて「近代」というもっとも日中が接触しあう時代性にも魅かれ、その中の漢語の素性はどうであるかを明らかにしようと考えている。

一方、日本語から見ていわゆる新漢語の比率も気になる問題である。

a 類、中国近代語の直接借用
b 類、中国古典語の転用
c 類、日本人独自の創出

の比率は一体どうなっているかもこれを機にある程度はっきりさせたほうがいいのではなかろうと考えて、具体的な手順としてまずは日中同形語を探し出し、そしてこれらの語の筋を明らかにし、中国から日本へ入ったものか、それとも日本から中国へ入ったものかを二分する。その上にさらに c 類のような純粋な和製漢語を見つけようと目論んでいる。

調査範囲は『近現代辞源』の全般（1008 頁）を目標としているが、今回はその十分の四の１-４００ページまでの語を調査し、以下のような結果を得た。（空欄は未調査、F、J は部分的）

	A	B	C	D	E	F	G	H	J	合計
総語数	84	554	503		42	128		415	386	2112
同形語数	17	221	193		14	79		176	207	907　（42.9%）

この辞書の収録語数の同形語率は 42.9%と平均に見ることができる。この 907 語を分類可能な語とし、さらに、①『日本国語大辞典』　②『幕末･明治初期漢語辞典』　③『明治のことば辞典』　④『明治大正新語俗語辞典』　⑤『大漢和辞典』、の五冊の現行辞書によって初出の時代などを比べ、その素性明かしの作業を行う。そこでたとえば、E、F に限って基本的には次のような分類を得ることができる。

ａ類:　悪戦　児童文学　耳垢　発布　発電　発酵　発熱　発生　発音　発育　罰金　法官
ｂ類：　悪性　発表　発車　発動　発明　発射　法典
ｃ類：　悪感　悪化　兒童団　耳殻　耳膜　二年生　二審　二元論　発動机　発祥地　発行　発展　法案　法定　法規

中国近代の新語として辞書に載ってはいるものの、調査対象である３４語のうち、半数以上が日本からの同形語であるという結果になった。ただ、その過程でいろいろと問題に出くわしたので、本稿ではそのいくつかを見ながら、語史を構築する際の方法論的なものをも模索しようとする。

３.『日本国語大辞典』の初出例

３．１　文芸中心主義

本稿では、『近現代辞源』の初出例の時代的早さと比較して、『日本国語大辞典第２版』の文芸中心主義の用例採集の傾向が露呈していることを指摘し、しかも日本の国語辞典の編集方針にかかわる普遍的な問題として取り上げた。たとえば、以下の四例はいずれも「中国（『近現代辞源』）のほうが日本（『日本国語大辞典第２版』）より初出例が早いことになるが、日本語の例は文芸関係のものばかりで、⇒の後の他の文献再調査によって日本語の用例を遡ることができることが分かった。

【病例】中国 1942、日本 1959『海辺の光景』安岡章太郎「現在ではアメリカでもっとも多く見られる病例で」
⇒「父母乃務　家庭育児」三谷周策著（鍾美堂、1905）
【病原体】中国 1919、日本 1940『畸獣楽園』小栗虫太郎「この二人の医師が、睡眠病の病原体（ビョウゲンタイ）をチムバンジーに注射した」
⇒「赤痢及麻刺里亜」島珪之助編訳（誠之堂、1897）
【茶話会】中国 1901、日本「大君の目出度い誕生日は茶話会では収まらなかった」『田舎教師』田山花袋 1909
⇒珈琲会（茶話会）『処女のつとめ』（ダブイヂス著、阪田孫四郎訳博文館、1894）
【参戦】中国 1917、日本「参戦の希望がかなひましたら、男子一生の面目であり大いに頑張ります」『多甚古村』井伏鱒二 1949
⇒挙国参戦と若き民軍の力『軍事心理研究　第２巻』（下沢瑞世、武揚堂書店、1914）

同じことは、宮島達夫も「術科」という語を取り上げている。『近現代辞源』では「日本国志」(1890)、「遊歴日本観察兵制学制日記」(1899)を引用しているのに、『日本国語大辞典第２版』では阿川弘之「春の城」(1952)をひく。「これは小説以外の軍事関係文献から、もっとふるい例をさがすべきだ。日本国語大辞典の文芸偏重のあらわれである。」と指摘している[3]。

たしかに、すこし遡って調べてみると、「術科」という語はすでに 1890 年の『歩兵射撃教範』（小林又七著）に出ていることが分かる。

この調査の中で、往々にして日本語の用例の発生時代が中国語より遅れる場合に出会う。先行研究としてすでに宮島（2009）が指摘したように、幕末・明治以降を中心に『日国』の出典例を増補していたことが知られている。つまり、「文芸中心主義」という視点からその初出例は必ずしも「初出」を反映するものではないという。

【参議院】【参議員】中国 1903、日本 1946「日本国憲法」
【極限】中国 1909、日本 1925「女工哀史」「人間の技量は機械の働きの如く、世の進歩と逆比例に段々極限されて行かねばならぬ」
【基地】中国 1897、日本 1952「春の城」阿川弘之「敵が B29 の基地を造るまで何カ月かかるかだ」
【成人教育】中国 1921、日本 1932「日本はどこへ行く」土田杏村「二つの行き詰まり

[3]宮島達夫「日中同形語の発掘」『国際シンポジウム「近代語の語源研究とその周辺」要旨集』漢字文化圏近代語研究会、関西大学文化交渉学教育研究拠点共催、2011.3.19

は日本の教育が学校教育の或る体系を本幹としてとり成人教育の広汎なる部面を閉却したところより来たものである。」
【化学方程式】中国 1903、日本 1930「機械」横光利一「化学方程式さへ読めない者に実験を手伝はせたって」
【化学繊維】中国 1957、日本 1973「現代経済を考える」伊東光晴「現実は化学繊維の難点である染色がうまくいかず」
【画報】中国 1884、日本 1909『田舎教師』「荻生さんから借りた戦争画報を二三冊又借りしてやったが」
【重版】中国 1890、日本 1930
【計画経済】中国 1934、日本 1933「最新現代語辞典」
【季刊】中国 1916、日本 1955　　＊少なくとも 1925 にすでにあった。
【極光】中国 1903、日本 1921「新しき用語の泉」
【機械化】中国 1936、日本 1951「山ぴこ学校の問題」
【集約】中国 1929、日本 1946「後裔の街」金達寿「われわれの弱点はここに集約されている」
【児童文学】中国 1928、日本 1944*書物(1944)甲<森銑三>二五「小波さんのわが児童文学の上に印せられた足跡はこの上もなく大きい」
【反応】中国 1903『新爾雅』、日本 1921*新しき用語の泉(1921)〈小林花眠〉「反応(ハンオー)〈略〉又、二種若しくは二種以上の物体間に起る化学的変化」

これら例をみると、逆に中国のほうは使用が早いことになる。ただし、中国資料が証左となる場合が多い。当該辞書の引用文献目録一覧をみればわかるように、いわゆる日本関係資料が使われることが多い（後述するように、資料の位置づけが問題となる）。これはある意味では日本語の使用例を裏付けるものであろう。

３．２　日本語の使用例がない場合

『近現代辞源』には中国語の使用例があるが、『日本国語大辞典第２版』には日本語の見出しは立てたものの用例なしの場合がある。そのなかには、中国語由来の語もあれば、多くは日本語の辞書編纂の問題で、用例採集の努力が足りなかった語もある。たとえば、『近現代辞源』に出ている四語、

【華氏温度】中国 1942
【基肥】中国 1925
【不成文法】中国 1903
【児童団】中国 1942

について、『日本国語大辞典第２版』では用例を挙げていないが、近代デジタルライブラリーで検索した結果、日本語にはさらに遡って古い用例があったことがわかった。たとえば、

「摂氏華氏温度比較表」『鰛油漬缶詰製造書』伊谷以知二郎、松尾霊彦著、水産書院、1907
「肥料の種類−石灰と稲作−施肥上の原則−施肥の時期−稲の成育と施肥−基肥と補肥」『稲作改良論』　横井時敬著、博文館、1904
「成文法及不成文法より生する一般被治者の利害」『法律社会之現象』河野和三郎著、吉岡書籍店、1888
「家族兒童團規約貯金」『直江津郵便局誌　御大典記念』大正五年(1916)直江津郵便局

いずれも『近現代辞源』より早い使用例が認められる。ということは、我々には、『近現代辞源』を利用して『日本国語大辞典第２版』の初出例との時代差を捉えて、日本語辞書の問題をあぶり出して、さらに他の資料によってその初出の時代を遡ることが求められているのであろう。

ほかに、同じく中国語には使用例があって、日本語には見出しは立てたものの用例なしの場合がある。

> 加圧 1922、核反応 1942、加速器 1946、滑翔 1941、花柳病 1913、監製 1842、環節 1903、漢族 1902、基本法 1934、計時 1943、継任 1925、鍵盤楽器 1930、恒等式 1930、黒熱病 1935、財経 1948、採種 1908、彩陶 1920、奨金 1908、陳酒 1880、爆炸 1900、氷球 1949、標点 1918、部首 1920、不定根 1918、本位主義 1919、話劇 1918

この中に、「花柳病、漢族、継任、陳酒、部首、本位主義、話劇」のように明らかに中国語に由来のものもあれば、データベースによって用例確認できるものもあろう。「青空文庫」で確認したら、下記の使用例はあるものの、まだ中国の例には追い付いていない。

> 1949「思想と科学」この間に教育に関する二大法案「教育基本法」及び「学校教育法」が議会を通過した。
>
> 1925 雑誌「太陽」に「加圧鍋」の使用例があり。

この問題と関連して、宮島は日中対照研究の対象としては、従来「学校」「科学」のような日中同形語を取り扱ってきたが、日本語は「テレビ」、中国語は「電視」のように完全に分化しているものも取り上げるべきと主張した。なぜなら過去のある時期に日本で訳語としての漢語「電視」が使われていて、それが中国語に入ったかつての同形語であったからだ。その類例を「発掘」しようと、吉沢典男・石綿敏雄『外来語の語源』(1979、角川書店)を手がかりに、その日中同形の漢字語訳語を抜き出して、『近現代辞源』の初出と比較し、和製漢語かどうかを模索した結果、「『外来語の語源』は二重の意味で日中同形語の＜発掘＞に役だつわけである。第一に、現代の日本語では使われていない（が中国語にはある）「滑翔機、電視」などが見つかった。第二に、『日本国語大辞典』に見出しはあるが例文がなかったり、例文の年代が新しかったりしたもの（菜単、私刑）に、古い例文を提供した」という[4]。

４　近代資料とは何か

４．１　近代中国語資料の位置づけ

もうひとつ重要な問題をここで提起しなければならない。いわゆる近代中国語資料の位置づけはどのようにすべきかの問題である。『近現代辞源』に中国語の近代資料の一群としていわゆる日本関係資料が多く使われている点について、慎重に扱う必要があるのではないかと思う。参考文献に利用されている清末の知識人が日本で編纂した『新爾雅』や訪日記録や考察報告などに出てくる新語は中国語の新語の語源として扱われるよりは、むしろ日本語の使用の反映と見なしたほうがいいのではないかと思われる。先の「術科」の例に見られるように、「日本国志」(1890)、「遊歴日本観察兵制学制日記」(1899)などに出ていても、中国語として実際に「使用」されたのではなく、ただ日本語を「記録」しただけであるから、どちらかというと、日本語の問題点を逆に浮き彫りにさせた意味では、日本語研究のための「日本資料」と位置付けてもよかろうか。

そうした中国人の記録した日本語がままあり、翁広平の『吾妻鏡補』（1815）は千余りの日本語を採録し、近世日本語との対照資料となりうる。玉燕の『東語簡要』（1884）は『吾

[4] 同注２

妻鏡補』を参照しつつ 1017 語を収録し、『日本寄語』と同じように、意味によって 18 類に分けるほか、三字門と成句門も設けており、伊呂波歌も付している。『遊歴日本図経』(1889) は清政府の外交特使である傅雲龍が明治中期来日の際に纏めた遊歴記録であり、いわば日本紹介の百科全書のようなもので、全 30 巻のうち、巻 10 の「方言」は「星謂之保之語若火西」のように日本語 428 単語を漢字で書き表しており、また巻 20「日本文学 上」ではいろは歌や国字や五十音図をも紹介している。日本語教科書として編集されたのは『東語入門』(1895) であり、清国駐日公使陳明遠の長男陳天騏が東京で「六年挙業」の後、帰国後出版したもので、2 千近くの日本語の単語や表現を日漢対音寄語として収納した。従来、以上のように中国資料の多くが日・漢両語の音韻体系の解明に使われていたが、近代中国人の残した多くの日本訪問記にも豊かな記述とともに当時の日本語、つまり新漢語を反映するものが多い。これらは、日本新漢語の中国への伝播を裏付けるものとしてとらえる向きもあろうが、多くは見学先で目にした、耳にしたことばをそのまま記録したもので、著者が意識的に中国語として使用するつもりもないから、むしろ日本近代語を裏付ける資料と位置付けられる。つまり、日本近代漢語の初出例はこちらに多く求めることも可能であろう。

４．２中国語の出典例の追及

『近現代辞源』において、近代資料に頼るあまりに、中国語の古典との接点をおろそかにされたことがある。下記のような例はいずれも中国語側が初出でない時代例を挙げている。そうなると、単純に初出の時代から見て日本から中国へと入ったものと勘違いされやすくなる。

【倉庫】中国 1889、日本 797「続日本記」
【救急】中国 1904、日本 797「続日本記」
【草案】中国 1889、日本 835「性霊集」
【玻璃】中国 1919、日本 1876「音訓新聞字引」
【給水】中国 1913、日本 1881
【給養】中国 1890、日本 1870「西国立志編」
【記念】中国 1909、日本 1855「和蘭字彙」
【化膿】中国 1857、日本 1811-45「厚生新編」
【化石】中国 1876、日本 1763「物類品隲」
【紀元】中国 1890、日本 1890

この中国語の時代選定はいわば「近代」において使われているかどうかに重点を置いているようようである。「倉庫、救急、草案」の三語はいずれも日本語のほうは上代の初出なのに、中国語のほうは近代以降となっているのがそのためである。事実、この三語はともに中国の唐以前に出ているし、「玻璃」だって明末の『東西洋考』にすでに使われる。つまり、編者の近代という設定の枠をはみ出した使用例（使用時代）をそもそも想定していないから、初出を全部近代に想定しているようである。

４．３　洋学資料の偏り（訳語かどうか）

『近現代辞源』はいわゆる漢訳洋書と英華字典の使用において偏りを見せていて、特に英華字典の場合は１９世紀初期のモリソンだけに偏っている。中後期のロブシャイドの英華字典（1866-69）や『英華萃林韻府』（1872）の利用が少ない。

洋学資料の取り扱いには問題がある。楊少坪『増広英字指南』1879 を用例提供のソースとしているが、しかし本来は初版『英字指南』の発行年は 1879 年であり、『近現代辞源』に利用している『増広英字指南』は 1906 年の増補再版であるから、完全に版を混同していて、さまざまな誤解を招くことになる。関西大学に所蔵されている同テキストの版を調査

したところ、下記の表のように、初版と増広版はあきらかに異なるものである。

《英字指南》の版の違い（1879、1905）

	見出し語	英語	初版 1879	増広 1905	備考
日中同形語	参謀	Adviser	○	巻三	1866 英華字典
	委員＊	Deputy	○	巻三	
	反射	Semi-transparent	○	巻四	反射光
	和声	Harmony	○	巻四	
	電力	Working power of electricity	○	巻四	
	藝術	Book of arts	○	巻三	芸術類
	哲学	Philosophy	×「格致」	巻二	
	哲学家	Philosopher	×「博学人」	巻三	
	社会	Society	×「結社」	巻二	
	天文台	Observatory	×「観天台」	巻四	
	天文学＊	Astronomy	×「天学」	巻四	1874、1875
	動物学＊	Zoology	×「生物学」	巻四	1875
非同形語	白帯	Leucorrhoea	○	巻四	婦嬰新説 1858
	白蘭地	Brandy	○	巻四	1828
	白煤	Anthracite coal	○	巻五	1877
	板刷	Clothing brush	○	巻五	
	刨花	Hair shaving	○	巻五	
	被単	Sheet	○	巻四	1828
	表鏈	Watch chain	○	巻四	1873
	餅乾	Biscuit	○	巻四	1828
	除法	Division	○	巻二	
	床位	Bunk; berth	○	巻四	1856
語釈語	刻画	Engraving	○	巻四	版画 1928
	主稿	Editor	○		編輯 1916
	白鴿票	Lottry	○	巻四	彩票 1889
	笛	F lute	○	巻三	長笛 1930
	三管号筒	Trombone	○	巻三	長号 1930
	抵逐	Repulsion	○	巻四	斥力 1937
	空床	Bedstead	○	巻四	床頭櫃
	大砲	cannon	○	巻二	加農砲 1908

つまり「参謀、委員、反射、和声、電力、藝術」のように、初版も再版もまったく同じものがある一方、「哲学、哲学家、社会、天文台、天文学、動物学」などは 1879 の初版ではまったく出ていなくて、それぞれ「格致」「博学人」「結社」「観天台」「天学」「生物学」となっていたことがわかる。となると、

> “哲学”这词为日语词、于 19 世纪七八十年代传入中国、如 1879 年楊少坪『增广英字指南』卷三:“Philosophy、 哲学、性理学” （942 頁）

のように、重要概念の「哲学、社会」の中国への逆輸入の時代をこの辞書により 30 年近くも簡単に引き上げさせたという、完全に間違った情報を提供することになる。

ほかに、英華字典（1866-69）には「和音」が使われている。朱（2003）は「和声」を明

治 16-17 年の音楽関係書に見られるという[5]。ただし、こちらでは「声学」という物理学の用語であり、日本語では音楽の用語に用いられるのが異なっている。

そしてとくに問題として挙げるのは近代語としての認知の問題である。『近現代辞源』では「楽観」について『博物新編』1855 に使われる「楽観」の例を初出として採用しているが、「此乃割傘之険、人不楽観」というのはまだ字の単位での理解で「人は（それ）を見るのを楽しまず」と読むべきで、一語としてまだ確立されていない。というのは反対語の「悲観」は『新爾雅』(1903）を初出としていることからもわかるように、「楽観」ももともと英語から訳された和製漢語だからである。同じ『博物新編』に出ている「生理、水質」は中国語本来の意味（なりわい、水のように）に留まっていて、いわゆる近代語の意味に至っていないことをすでに指摘していた。

５．日中書語交流の時間的幅の設定

日中交流史の視点から整理しなおして、時代別に中国語由来の新語と日本での創出とを分けて、とくに前者に由来する新漢語を捉え、それと後者との時代的区分を検討している沈国威（1998）の捉え方があった[6]。中国における「新漢語」形成史の時期区分は 19 世紀初頭～20 世紀初頭の 100 年間を 5 期に区分し、

①準備期	1807-1840頃
②発展期	1840-1860
③官製翻訳期	1860-1880
④停滞期	1880-1895
⑤日本語導入期	1895-1919

①②③期をいわゆるａの中国語からの直接借用部分とし、⑤をｂ、ｃとするような中国語の視点から史的にさらに区分けをするものである。

今回の調査ではそれを踏まえて、時代的な修正を加えることができよう。

５．１　中国から日本へ

漢訳洋書や英華字典の日本への流入の時代を考えると、中国語から日本語へ入った言葉の時代的下限は明治十年代の後半に設定することができよう[7]。

中国から日本へ	上限：４～５世紀ころ
古典→近代→現代	
近代→現代	下限：1886

５．２　日本から中国へ

近代中国における日本語の受容について、日清戦争以降、日本から中国へ影響を及ぼすことが多くなってきた。それ以降さらに、

① 1895－1919　日清戦争からベルセイユ条約（熟成期）
② 1919－1945　ベルセイユ条約から敗戦　（決裂期）[8]
③ 1945－1972　日中国交断絶期

---

[5] 朱京偉（2003）ｐ185.

[6] 「新漢語研究に関する思考」『文林』32、1998

[7] 陳力衛「明治初期における漢訳洋書の受容」『東方學』第九九輯、2000.1

[8] 狭間（2011）の分類に従う

④ 1972－2000　国交回復以降

のように、四期に細分できよう。一般では、いままで1900年ころの留学生による翻訳の高まりを日本から中国への逆流入の上限としていたが、『近現代辞源』の取り扱う日本関係資料をそのルートの一つとして捉えるなら、二十年遡ってもよろしいかと思う。すると、

日本から中国へ　　上限：1879
古典→近代→現代
近代→現代

下限の設定もする必要があろうが、日中戦争の終結を一つの目安とも考えられるが、国交のない時代であっても専門用語辞典の流通を考えると、もっと下って 1960 年代の半ばころにしても可能であろう。そして新語などを入れると、まさに国交回復以降ともなろう。

５．３　没交渉の時代

そうなると、たとえ、下記のような比較例が見られていても、日本から中国へという設定が不可能となるわけである。「結晶、細胞」などの例のように、時代設定が難しくなる例が増えてくる。

【測定】中国 1858、日本 1823「遠西観象図説」
【花粉】中国 1858、日本 1833「植物啓原」

つまり、こうした没交渉の時代において両国が語形の同じことばをそれぞれ使っていることになる。可能性としては中国語資料の精査が足りたくていわゆるさらなる祖例を中国語に探し求めることが予想される。あるいはたまたま両国語が同じ言葉を創出した「暗合」ともとられる。

上記のような日中間の交流可能な時代幅を設定する際、どうしても 19 世紀 70 年代の使用例が両方とも出てくることに遭遇する。そこで、もう一つの交流のルートとして挙げられるのは宣教師同士の交流によって新語のお互いに利用しあう可能性である。それも以下の要因から想定される。

・『和英語林集成』の上海印刷
・江南製造局翻訳館のフライヤの訳語統一に関する演説 1890 年
・中国での新語集出版はほとんど宣教師の手によるもの

したがって、このルートでの交流がどうなっているかは未詳のままで今後の課題として大いに注目されたいと願っている。

６．終わりに

『近現代辞源』の編者自身も認めているように、この辞典を出版してから「回归线、参加、发展、管理、热情、正确、祖国」の語が漏れているという指摘を受けたという。その他、編者自身も、「根性、共用、农产、庞大、入库、世纪、水库、外货、外流、外政、腺体、消音、学术、外用、延长、账务、最惠国」等が漏れていることに気がついた。しかも「世纪」は『近現代漢語新詞詞源詞典』で収録しているにもかかわらず、『近現代辞源』では漏れてしまったと言っている。

『近現代辞源』という鏡を借りて、日本語の漢語問題に照らし合わせるという研究方向と異なって、中国語における日本語の受容、あるいは中国語における日本語借用語の問題をも照射できるものである。たとえば『近現代辞源』の収録状況を踏まえたうえで、当該

辞書は中国語における日本借用語研究に確かな実例を提供しただけでなく、「白葡萄酒、交感神経、結石、絶縁、神経錯乱、十二指腸、嗅神経、塩酸」のように、従来の中国語辞書に収録されていない語の言語接触の経路を明らかにし、あらたな借用語としての認定に役立ったとも評価できよう[9]。

いれずにしろ、データベースやコーパスの構築により、辞書編集の問題点を補完することができるだけでなく、新たな視点の提示と問題の発見にもつながることとなるから、今後の語史研究に欠かせないものとなるのであろう。

文　献

沈国威（1994）『近代日中語彙交流史』笠間書院

荒川清秀（1997）『近代日中学術用語の形成と伝播』白帝社

梶原滉太郎(1992)「天文学」の語史」『研究報告集』13、 国立国語研究所 p.77-121

朱京偉（2003）『近代日中新語の創出と交流』白帝社

宮島達夫（2008）「テレビと電視―「電視」は和製漢語か」『漢字文化圏諸言語の近代語彙の形成―創出と共有』関西大学出版部

宮島達夫（2009）「語彙史の比較（１）―日本語」『京都橘大学研究紀要』第 35 号

宮島達夫（2010）「語彙史の巨視的比較」『汉日语言对比研究论丛』第 1 辑、北京大学出版社

陳力衛（2005）「『博物新編』の日本における受容形態について」『日本近代語研究４』ひつじ書房

陳力衛（2011）「「新漢語」とは何か―漢籍出典を有する語を中心に」『言語変化の分析と理論』おうふう、p59-76

陳力衛（2012）「英華辞典と英和辞典との相互影響―２０世紀以降の英和辞書による中国語への語彙浸透を中心に－」『JunCture』03 号、名古屋大学大学院文学研究科付属日本近現代文化研究センター『国際シンポジウム「近代語の語源研究とその周辺」要旨集』漢字文化圏近代語研究会、関西大学文化交渉学教育研究拠点共催、2011.3.19

[9]何華珍「中国語にある蘭学漢字語研究－『近現代辞源』の学術価値を論ず」『国際シンポジウム「近代語の語源研究とその周辺」要旨集』漢字文化圏近代語研究会、関西大学文化交渉学教育研究拠点共催、2011.3.19

近代対訳コーパスにおける日韓語彙の諸相
-文体の異なる対訳コーパスの比較を通して-

張　元哉（韓国・啓明大学）[1]

１．はじめに

現代日韓の語彙（特に漢語）は、日中や韓中よりもその類似性（同形率）が高いことが知られている。これは、近代以降に韓国が語彙交流の相手を主に中国から日本に変えたことによるものであろう。実際、現代日韓の語彙の同形率の高さ（約90%前後）が近代にさかのぼるとそれほどの高さ(約60%)ではないこと、また日韓の語彙交流が始まった19世紀末以降、現代語に向かって同形率や日本製漢語が同様の増加曲線を描くことなどがわかっている （張2000、2003a）。

このように近代以降の新漢語の増加と輸入は現代日韓の語彙形成に影響を与えており、それは単語レベルだけではなく、語彙や表現のレベルにも及んでいたと考えられている。そうすると、これまで知られている現代日韓の語彙の量的な構成の類似点や相違点は近代語にはどうであったかという疑問が浮かぶ。ところが、これまでの日韓の近代語研究は、語彙交流の側面から、どのような語がどのように造語され、どのように交流したかについての研究に重点が置かれ、もちろん日韓の語彙交流においてさまざまなことがわかるようにはなったが、日韓の現代語より同形率（類似性）が低かった、近代日韓の語彙における量的な構成や分布についてはあまり知られていないようである。

そこで、本稿では、近現代的な語彙的要素が混在し、語彙交流の初期である、19世末・20世紀初期の文体の異なる日韓対訳コーパスの比較を通して、語彙量、語種、品詞、語構成などの観点から近代日韓の語彙の諸相を探ることを目的とする。

ただし、今回の調査は、資料の制約と限られた調査量の問題などもあって、近代日韓の語彙の様子が十分に捉えられるまでには至っていない。また計画されていた調査がまだ完了していない途中報告であることも断っておきたい。

２．調査資料と調査方法

近代の日本語と韓国語という二国間の語彙の様子を探るために、注意しなければならないのが、資料の時期、文体（ジャンル）と調査基準であろう。まず調査資料について見てみよう。

２．１　調査資料

日本語と韓国語の語彙を対照するには、資料の時期や文体という条件が同質なものでなければならないことはいうまでもない。しかし、近代という時期は、日韓ともに文体やジャンルの概念が確立しておらず､近代を文体の実験期とまで言うほどである。また、時期の問題も、たとえば日韓の新聞や雑誌の語彙を調査するにしても時期的に資料の有無やずれの問題が生じるなど、同質の調査はそう簡単ではない。これらの問題を十分に考慮しないと、調査結果が日韓の違いなのか、時期的な違いなのか、文体の違いなのかがよくわからなくなる恐れがあるからである。

それで、本稿では、近代という時期をこれまで指摘されてきた日韓の語彙研究の時期の区切りを踏まえて1910年代前後までにし、文体やジャンルの問題を解決するために日韓の対訳資料を使うことにする。近代における日韓対訳資料は、文体を考慮し、

[1] Chang_wonjae@hotmail.com

金秉喆（1975）[2]、李漢燮（1985）、李建志（2000）などの研究を参考にした。

２．１．１　文語体の資料[3]

日本の資料：『西洋事情』（福沢諭吉、1866-1870）
韓国の資料：『西遊見聞』（兪吉濬、1895）

『西遊見聞』(以下、K西遊)は、 福沢諭吉のもとで就学した韓国最初の留学生である兪吉濬の啓蒙書であり、それ以前とは違ったハングル漢字混用体として後代の文体に大きな影響を与えたものである。『西遊見聞』の構成と内容は『西洋事情』(以下、J西洋)と似ており、全20編の中には著者が著述した部分と『西洋事情』から翻訳した部分があるという(李漢燮1985)。本研究では翻訳された『西洋事情』の日本語と翻訳した『西遊見聞』の韓国語を調査対象とする。以下は、そのリストである。

『西洋事情』	『西遊見聞』
初編巻之一「政治」	第五編「政府의治制」
外編巻之二「政府の職分」	第六編「政府의職分」
初編巻之一「兵制」	第十三編「泰西軍制의来歴」
初編巻之一「病院」	第十七編「病院」
初編巻之一「博物館」	第十七編「博物館及博物園」
初編巻之一「蒸気機関」	第十八編「蒸気機関」

ただし、本稿での調査報告は、上のすべての編までは調査できておらず、「外編巻之二「政府の職分」- 第六編「政府의職分」」を除いた範囲での結果であることを言っておく。今後調査できていない編はもちろん、同じ文体に属する別の言語作品も調査する予定である。

２．１．２　口語体の資料

日本の資料：『文七元結』(幕末～明治初期？)
韓国の資料：『東閣寒梅』(1911)

『東閣寒梅』(以下、K東閣)は、当時翻訳家として活躍した玄公廉の作品であり、日本語と韓国語がパラレルに書かれている特徴がある。落語・人情本の『文七元結』(以下、J文七)を原作に翻案（地名と人名だけを変えている）したもので、日本語の部分は『文七元結』の口演速記をもとにしているが、どの速記かはまだ不明のようである（詳細は李建志2000を参照）。
本調査では、『東閣寒梅』(1911)の日本語と韓国語（『日鮮語新小説 東閣寒梅 』(玄公廉、文明社)を対象とし、発表者が電子化したものを使うことにする。

今回の調査資料における文体と時代の関係は以下のとおりである。本調査は主に近代日韓の文語体と口語体の違いを対照しているが、近代語の位置づけのために現代語の調査と比較することもある。

---

[2] 金秉喆（1975）によると1895年から1910年までに刊行された韓国翻訳文学の79作品のうち日本語からの直接翻訳や重訳された作品が54作品があるという。
[3] 『西洋事情』は、岡島昭浩氏のホームページ（http://www.ne.jp/asahi/nihongo/okajima/bungaku.htm、黄美静氏の作成）から、『西遊見聞』は李漢燮氏のホームページ（http://nihon.korea.ac.kr/、リンク切れ）からダウンロードした。

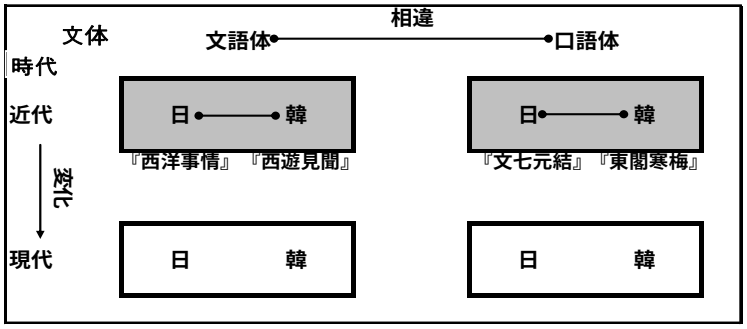


図1　文体と時代における日韓の調査対象コーパス

２．２　調査方法

２．２．１　調査単位の設定

本調査の調査単位は、日本語と韓国語において同じ基準で、比較的に調査者のゆれの少ない調査単位であること、本稿の調査目的の一つである語構成の調査ができること、本調査の結果と先行調査が比較できることという条件として設定した。長い単位としては、日本語は文節、韓国語は語節（文節と同様）単位を採用する。切り分けられた文節（語節）から日本語は助詞・助動詞・記号を、韓国語は助詞・語尾・記号を取り除いた部分を１単位語とする。

以下は本調査対象の調査単位の例示である。

J西洋
往昔｜欧羅の｜諸国は｜封建世禄の｜制度を｜以て｜臣下を｜養ひ、｜各国の｜帝王｜互に｜相攻め、｜国内の｜貴族｜互に｜相闘ひ、｜専ら｜武を｜重んじて｜文を｜勉めず、｜字を｜知る｜ものは｜唯｜僧徒のみ。

K西遊
泰西軍制의來歷
往古｜歐洲諸國이｜封建과｜世祿의｜制度로｜以·야기｜臣下·｜養·고｜各國의｜帝王이｜互相｜攻擊·며｜國中의｜貴族이｜互相｜競爭·니｜是故로｜武藝·｜重히｜·고｜文字·｜輕히｜·지라｜文字·｜解··｜者·｜宗敎의｜敎正에｜止·고｜

２．２．２　調査単位の原則

1単位の長さと幅の詳細な原則は、基本的に国立国語研究所（1987）『雑誌用語の変遷』の「長い単位」に従い、韓国語も該当するものはこれによる。この細則と異なるものや韓国語において注意が必要と思われる語は、以下のとおりである。

①文末の「ものか、ものだ、わけだ、ことだ、ところだ」などの下線の語は、1単位語として扱い、実質名詞と別見出し語にする。
②「お店、お客」などの「お～」は、１単位語とし、「店、客」と統合しない。
③「（～ては）いけない、いかない、ならない」は、１単位語とし形容詞にする。韓国語の「안된다，못하다，못되다」も1単位語とする。その他は「副詞（안，못）+動詞」に切る。
④「お～する」は、１単位語とする。これに対応する、韓国語の尊敬の接辞の시(shi)が付いた「가시다」も１単位語とし、「가다」と統合しない。
⑤「を以って、における、において、について」などは、本動詞と別の見出し語にする。韓国語も同様に扱う。
⑥「急に、楽に」などは、形容動詞の語尾「に」として扱うが、「誠に、実に」な

どは、辞書に見出し語として載っている場合、１単位語とし副詞にする。
⑦「ている、てある、てみる、てしまう」などの補助動詞は、本動詞と別見出し語にする。韓国語も同様である。
⑧韓国語の「하다」は、「する」の意味と、引用（言う）の意味として別見出し語にし、補助動詞の場合は、⑦と同様に扱う。
⑨日韓の品詞の違いによるものは、そのまま尊重する。例えば、「～たい」（助動詞）-「싶다」（形容詞）については、前者(日本語)は助動詞なので扱わないが、後者(韓国語)は扱う。
⑩単位語としての判断材料とした辞書は、日本語は『広辞苑』（第6版、2008）『新明解国語辞典』（第6版、2005）、韓国語は『ヨンセ韓国語辞典』（1998）『標準国語大辞典』（1999）である。

３． 日韓の語彙量の対照

３．１　近代における文体別の日韓の異なり語数・延べ語数

『J西洋』・『K西遊』と『J文七』・『K東閣』の異なり語数と延べ語数は以下のとおりである。

表１ 近代における日韓対訳コーパスの異なり語数・延べ語数

文語体	Ｊ西洋	Ｋ西遊
異なり	973	1194
延べ	1948	1988

口語体	Ｊ文七	Ｋ東閣
異なり	1281	1257
延べ	3628	3848

上の表１からわかることは、文語体においては異なり語数･延べ語数ともに韓国のほうが多い反面、口語体においてはそうではないということである。ただ、口語体の語数から見てわずかな違いなので、あまり日韓の相違がないかもしれない。語数の違いについては、中野洋（1976：21）では「翻訳される側とする側とでは延べ語数もかわろう。される側よりする側の方が延べ語数が多くなるかもしれない。」という指摘があり、表１の語数の違いもそのように解釈できるかもしれないが、李鐘洙（2010）の聖書の調査（韓国：旧訳1911-日本：文語訳1917）や張元哉（2004）の日韓対訳新聞の調査はそれぞれ第3国語と韓国語が原典であり、韓国語の語数が日本語より多い結果になっている。このことから考えると、語数の多少は必ずしも翻訳の方向によるものではないようである。

近代の口語体において『Ｊ文七』の異なりが若干多いが、「お～、～様(さん)、ある・御座る、する・致す、お～になる、お～する（致す）、お～なさる（下さる）」などの待遇表現が韓国語より多いと考えられる。

３．２　使用率順異なり語数の累積使用率の対照

では、文体別に使用率順異なり語数に対する累積使用率を見てみることにする。それを図示したのが以下の図2である。

図２からわかることは、今回の狭い調査範囲でのことではあるが、口語体より文語体において日韓の累積使用率の違いが目立ち、韓国語のほうが日本語より累積使用率が低いことである。累積使用率は高頻度語群の使用率が影響を与えているので、日韓の高頻度語群における文体的な違いを綿密に考察する必要があると思われる。

文語体における累積使用率が韓国のほうが低いことは、現代の日韓対訳新聞の調査でも同様な傾向であるが、近代語よりその格差は小さいこと、β単位の短い単位の調査では高頻度語群では韓国語の方が累積使用率が高くなる（張元哉2003b）ことから高頻度語群での時代的変化の様子を捉える必要もあると思われる。

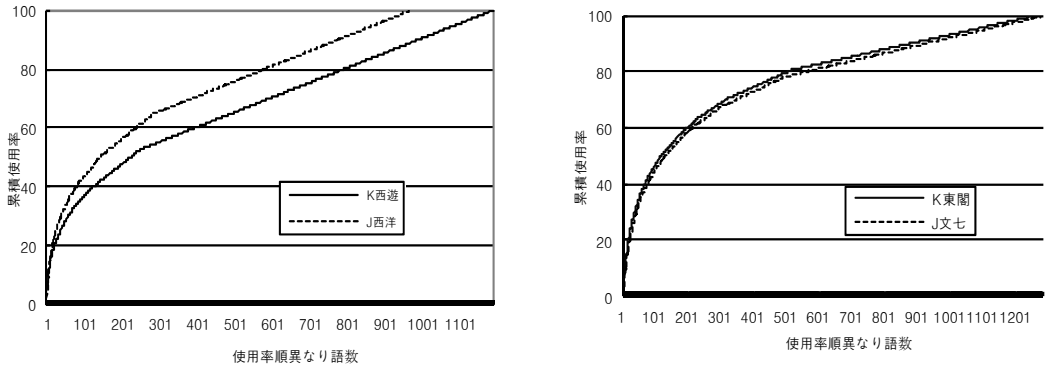


図２ 近代における文体別の日韓の累積使用率

これまでの言語間の累積使用率（カバー率）の調査では、中野洋（1976）の『星の王子様』の6ヶ国語版（日、英、独、仏、西、葡）で、日本語が一番累積使用率が低いとされており（助詞･助動詞を除く）、韓国語との比較はされていない。語彙的にも統語的にも類似している日本語と韓国語の比較こそ累積使用率の分布の違いが浮き彫りにできると思われる。現在日韓それぞれ現代語の均衡コーパスが整えられており、語彙量における日韓の位置づけや記述の精密化が期待されるところである。

４． 日韓の語種構成の対照

文体別に語種に分けて集計したのが以下の表２と図3である（延べ語数）。 語種の略称は、和語・固有語（固）、漢語（漢）、外来語（外）、混種語（混）である。

表２ 近代における文体別の日韓の語種構成

文語体	固	漢	外	混	計
J 西洋	1043	741	21	143	1948
%	53.5	38.0	1.1	7.3	100.0
K 西遊	33	1299	1	655	1988
%	1.7	65.3	0.1	32.9	100.0

口語体	固	漢	外	混	計
J 文七	2936	536	11	145	3628
%	80.9	14.8	0.3	4.0	100.0
K 東閣	3019	507	1	312	3848
%	78.5	13.2	0.0	8.1	100.0

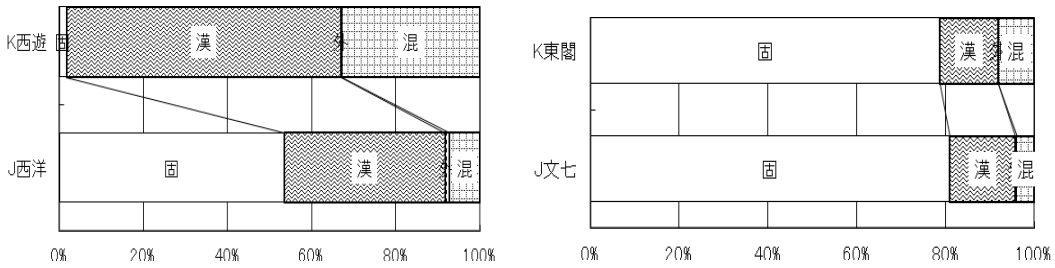


図3 近代における文体別の日韓の語種構成

表２と図３を見ると、まず、文体によって近代日韓の語種構成の相違が相当異なることがわかる。口語体では、日韓の語種構成がかなり類似していて、混種語が韓国語

に多い（2倍ほど）ぐらいであるのに対して、文語体では、日韓の語種構成の相違が激しく、日本語は固有語、韓国語は漢語と混種語が非常に多い傾向が目立つ。韓国語の固有語は、1.7%を占めるに過ぎず、固有語をできるだけ排除した様子がうかがえる。

近代の文語体における日韓の語種構成の相違が、以下の図４で示した現代語の日韓の語種の違いと類似している（張元哉2004）が、近代語のような格差はそれほど大きくないことがわかる。このことは語種における近代から現代への変化として捉えられる。

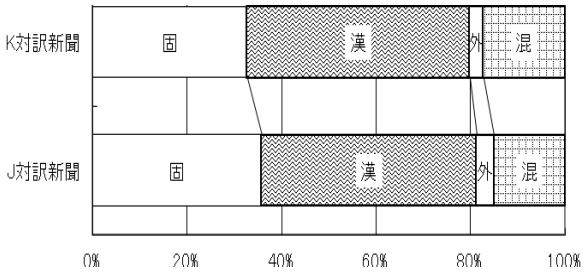


図４　現代日韓の対訳新聞における語種構成

図5は、近代と現代における日本語の語種の比率から韓国語の語種の比率を引いた数値を図示したものである。たとえば、次のようになる。

・近代の固有語：日本語53.5%-韓国語1.7%＝51.8%

・近代の漢語：日本語38.0%-韓国語65.3%＝-27.0%

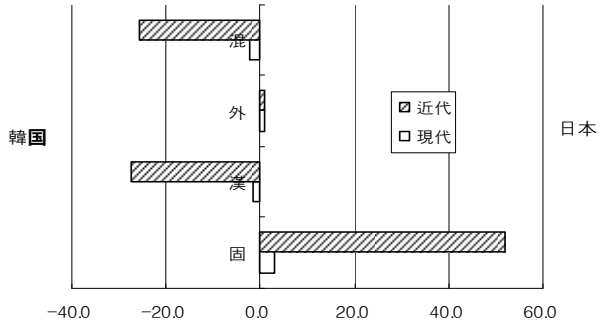


図5　近代と現代における日韓の語種の比率の差

つまり、近代語から現代語に向かって日韓の語種構成の類似性が高まったということである。このような現代語への変化の裏づけとして日本語では国立国語研究所(1987)、韓国語では韓榮均（2009）の調査結果をみると次のようになる。

まず、国立国語研究所（1987）は長い単位での「中央公論」の雑誌を調査したもので、和語と漢語は年によって増減の変動があるものの、一定の傾向は見られないこと、外来語と混種語は増えていく傾向が見られると指摘している。一方、韓榮均（2009）は文節単位での調査であり、19世紀末からの漢字ハングル混じり文の新聞の論説を調査したもので以下のような調査結果を示している。

表3　１９世紀以降の韓国新聞の論説における語種の変化（%）

語種	品詞	1890年代	1909年	1920年代	1930年代
漢語	代名詞	2.04	3.71	1.48	0.31
	連体詞	3.70	5.25	2.81	1.63
	一字名詞	11.82	9.11	3.42	4.40
混種語	一字漢語+hada	10.81	12.5	5.17	1.90

表３からいくつかの品詞の漢語と混種語が減少していることがわかる。

以上の先行論文の調査結果をまとめると、日本語は混種語が増加し、韓国語は漢語

と混種語が減少していることから、近代語から現代語への日韓の語種構成の変化の様子が推測できる。

５． 日韓の品詞構成の対照

品詞分類の基準は、国立国語研究所（1964）『分類語彙表』に従う。文体別の品詞構成は以下のとおりである。

表４　近代における文体別の日韓の品詞構成

文語体	体言	用言	相言	他	計
J 西洋	1147	537	215	48	1948
	58.9	27.6	11.0	2.5	100.0
K 西遊	1182	503	268	29	1988
	59.5	25.3	13.5	1.5	100.0

口語体	体言	用言	相言	他	計
J 文七	1651	1154	646	177	3628
	45.5	31.8	17.8	4.9	100.0
K 東閣	1600	1213	858	168	3848
	41.6	31.5	22.3	4.4	100.0

表4からわかるように口語体のほうが文語体より比較的に日韓の品詞構成の違いが目立ち、口語体では日本語は体言類が多く、韓国語は相言類が多いことがわかる。これは、日本語が名詞志向表現、韓国語が動詞志向表現を好むことの現われであろうか（林八龍1995、金恩愛2003など）。すなわち派生名詞、複合名詞、動作性名詞や動名詞が主語、目的語、修飾語、述語に立つ際に、日本語の名詞が韓国語の動詞・形容詞になりやすい傾向があるということである。

たとえば、今回の調査資料からの例を見てみると、「不思議さ→異常である、夫婦別れをしない→夫婦が別別に別れなくて、若者が→若い人、心配を→心配して、お帰りを→帰ってください」などである（日本語→韓国語の日本語直訳）。また、これらの名詞が文語体よりは口語体によく出現する(？)ために口語体により品詞構成の相違が現れるのであろうか。

以下では日韓の口語体の相違が見られる体言類、相言類を中心に考察を行うが、比較のために文語体も合わせてデータをあげることにする。

５．１　文体における日韓の体言類と相言類の内訳の対照

文体別に体言類と相言類の内訳を示すと以下のようである。

＜文語体における体言類と相言類の内訳＞

	体言類	
	J西洋	K西遊
名詞	1039	1087
代名詞	35	18
動名詞[4]	15	58

	相言類	
	J西洋	K西遊
形(形動)	69	91
副詞	89	83
連体詞	57	94

＜口語体における体言類と相言類の内訳＞

	体言類	
	J文七	K東閣
名詞	1260	1342
代名詞	248	190
動名詞	110	45

	相言類	
	J文七	K東閣
形(形動)	224	285
副詞	320	433
連体詞	102	139

[4] ここでの動名詞とは、日本語は動詞の連用形の転成名詞やそれを含む合成語、韓国語は動詞にロ(ㅁ), 기(ki)の名詞化接尾辞をつけた語である。

口語体において「J文七」に体言類が多いのは、「代名詞」と「動名詞」の多さによるものであり、「K東閣」に相言類が多いのは、「形(形動)」「副詞」「連体詞」の多さによるものであることがわかる。文語体では、口語体に比べ量的に日韓の違いは見られないものの、同様の傾向が見てとれる。但し、動名詞は「K西遊」に多く、副詞は日韓ともにほぼ同じ量である。

近代韓国における文語体の動名詞の多さは、当時のハングル漢字混じり文（漢文直訳）の特徴であり、典型的な名詞用法の語というよりも述語部分に現れる語が多く、以下の例のように「漢語+함+이라」（21回、波線)という形式のものが多い。

例：宗教信服　此･各人이其信服･･崇旨･崇奉･기許･고政府가是･勿關･야民間의軋轢･･紛爭을宰制홈이라（日本語原文：信教人々の帰依する宗旨を奉して政府より其妨をなささるを云ふ）

この使い方が韓国の現代語ではあまり使われなくなっていることを考えると、動名詞は、現代語の文語体と口語体においても日本語の方が多いのではないかと推測される。

続いて、文体別に体言･相言類における各品詞の異なり語数と用例を見てみると、表5のようになる

表5　文体における日韓の体言･相言類の異なり語数と用例

品詞		文語体		口語体	
		J西洋	K西遊	J文七	K東閣
体言	代名詞	**3語**: 此れ(30)、私(3)、此処(2)	**1語**: 此(18)	**30語**:何(40)、其れ(38)、御前(36)、此れ(26)	**24語**:나(47)、너(32)、무엇(18)、이것(13)、그것(12)
	動名詞	**11語**:競り売り(2)、妨げ(2)、妨げ(2)	**43語**:有함(3)、謂함(2)、各伸함	**75語**:使い(5)、御払い(5)、取り返し(4)	**34語**:보기(3)、하시기(3)、돌아오기(2)
相言	形容詞	**27語**:無い(27)、ごとし(9)、多い(4)	**71語**:　無하다(14)、같다(2)、過하다(2)	**88語**:様(27)、無い(20)、ない(14)	**127語**:없다(29)、같다(21)、그렇다(10)
	連体詞	**5語**: 其の(42)、此の(10)、大いなる(2)	**3語**: 其(80)、彼(1)	**19語**:此の(35)、其の(28)、彼の(7)	**15語**:그(49)、이(47)、그런(11)
	副詞	-	-	**142語**:どうも(31)、マア(19)、どう(17)	**171語**:참(48)、좀(18)、참말(14)

()の数字は頻度

表5の異なり語数からもすでに述べた傾向とほぼ一致しており、日韓の各品詞における語の種類にもその違いが現れている。

ここで注目すべきことは、文語体も口語体も日本語は代名詞（これ、それ）が多く、韓国語は連体詞（이(この)、그(その)）が多いことであり、また連体詞においては特に日本語の「その」より韓国語の「그」が多いことである。

前者は、日本語の『J文七』では代名詞（それ）で現れるものが、韓国語の『K東閣』では「이(この)、그(その)+ 名詞（人、時間、場所）」に対訳される例（14例）があるからである。

J文七：それがマアあんなに大きくなったんだものね。（八）
K東閣：그· 기가참그러케커젓스닛가응(その子供が本当にそんなに大きくなったんだものね。訳は筆者、直訳)

後者は、指示詞の用法を大きく現場指示と非現場指示に分けた場合、非現場指示における日本語はコ・ソ・ア系が現れる反面、韓国語は이(こ)、그(そ)系しか現れない（宋晩翼1991）からではないかと思われる。コ系と이(こ)系は日韓同様に対応しているのでソ・ア系とユ(そ)系の違いであろう。上の例からすると、波線の部分の「あんな」が「그렇다」（そうだ）に対応するようなものであろう。

６． 日韓の語構成の対照

近代日韓の文語体と口語体の語構成の構成は以下のとおりであり、2次結合以上のものは、最終結合を見て判断した。また、「お～する（いたす）」（23語）はここでは複合動詞に入れた。

表6　近代における文体別の日韓の語構成

文語体	単純	派生	複合	計
J西洋	1650	182	116	1948
	84.7	9.3	6.0	100.0
K西遊	1158	772	58	1988
	58.2	38.8	2.9	100.0

口語体	単純	派生	複合	計
J文七	3014	338	277	3628
	83.1	9.3	7.6	100.0
K東閣	3189	378	272	3848
	82.9	9.8	7.1	100.0

表6からわかることは、第一に、文語体に日韓の語構成の相違が認められ、日本語は単純語と複合語が多く、韓国語は派生語が多いということである。第二に、日本語は文体における語構成の比率がほぼ同じであるが、韓国語は文体によってかなり異なっていて、文語体に派生語の比率が多いということである。

では、文体ごとに複合語と派生語についてもう少し詳しく見ていくことにする。

６．１　文体別の複合語の内訳

複合語のうち、主な品詞を文体ごとにあげると以下のとおりである。

・文語体の複合語の内訳

J西洋：①名詞99回（73語）：共和政治(5)、人々(5)、蒸気機関(5)、立君独裁(3)、禽獣魚虫(2)、西洋諸国(2)、自主任意(2)、為替問屋(2)、フランス病院(2)、各々、鎧兜、見世物、工作貿易、貴族名家

K西遊：①名詞58回（46語）:蒸氣機關(4)、歐洲諸國(2)、禽獸蟲魚(2)、都下人民(2)、自由任意(2)、天下各國(2)、泰西各國(2)、熊虎獅犀、各其各目、古今物産、歐洲全幅、宮闕近處、飢寒疾苦、每兩六分

・口語体の複合語の内訳

J文七：①名詞71回（54語）： 親子(3)、見ず知らず(2)、小間物(2)、二親(2)、行く末(2)、縞柄(2)、家々
②動名詞46回（31語）：取り返し(4)、勤め奉公(3)、身寄り(3)、心持ち(3)、人通り(3)、夫婦別れ(2)
③動詞141回（107語）：受け取る(6)、申し上げる(6)、飛び込む(5)、行き立つ(3)、しやがる(3)、見捨てる(2)

K東閣：①名詞112回（73語）：계집아해(-兒孩)(8)、큰돈(5)、고공사리(雇工-)(4)、품속(4)、거짓말(3)

②動名詞：2回（2語）:다다러오기、없어졌음
③動詞128回（65語）：돌아가다(9)、돌아오다(9)、가져오다(6)、지나가다(6)、내놓다(5)、들어가다(5)

文語体における『J西洋』に複合語が多いのは、複合名詞によるもので、『K西遊』より異なる語数も多く、そのほとんどが漢語である。一方、口語体における複合語の比率は表6からだとほぼ同じであるが、その内訳を見ると、『K東閣』には複合名詞が多く、『J文七』には複合動名詞、複合動詞が多い。

複合名詞の場合は、文語体では日本語に漢語が多い傾向にあるのに対して、口語体では韓国語に固有語と混種語の複合名詞が多いことがみられて面白い。また、口語体の『J文七』に複合動名詞や複合動詞が多いことは、5節で述べた動名詞と動詞の日韓の多少の傾向と前者は一致し、後者は複合において日本語の動詞が多くなるということになる。

これまで日韓の語構成についての調査があまり行われていなかったので、今回のデータにおける日韓の特徴が当時の言語現象を反映していて一般化できるかということと、現代語との違いなどについては、まだわかっていない。今後多くの調査が必要であるゆえんである。

６．２　文体における派生語の結合パターン

まず、文語体の派生語である。『K西遊』に派生語が非常に多く、以下に派生語における品詞と語種の関係をみることにする（○：和語、●：漢語、◎：外来語）。

＜文語体の派生語＞

	J西洋	K西遊	
名詞	109	114	
(●+●)+●	30	48	J:世界中、動物園、天主教　K：世界上、動物園、基本形
●+(●+●)	5	7	J:新発明、全世界、総病院　K:大都會、大旨趣、新造物
(◎)+●	5	0	J:フランス帝
動詞	73	473	
(●+●)+○	63	269	J:一変する、発明する、改正する　K:收聚하다、發用하다
(●)+○	2	202	J:議する、害する　K:至하다、爲하다、作하다
○+(○)	3	0	J:相攻める、相戦う
副詞	0	25	
(●+●)+○	0	14	K:是故로、如此히、輕蔑히
(●)+○	0	9	K：順히、重히、輕히
形容詞	0	87	
(●+●)+○	0	56	K:口虛하다、輕便하다、過度하다
(●)+○	0	29	K:無하다、近하다、難하다

文語体における『K西遊』の派生語の多さは、動詞や形容詞の「一字漢字・二字漢字+hada」によるものであることがわかる。これは、4節の語種構成の特徴とつながるものであり、「漢字+hada」による語の表れは当時の文語体の特徴でもある。この造語法は韓国語の現代語に向けて減少する（4節）。名詞における各パターンは大差ない。

＜口語体の派生語＞

	Ｊ文七	K東閣			Ｊ文七	K東閣	
接頭辞:	144	10		接尾辞:	86	279	
名詞				**名詞**			
お(ご)+名詞	112	0		語基+さま(さん)	31	5	J:娘さん K:주인님
○+（○）	52	0	J:御金、御店	語基+さ･ki	6	46	J:可愛さ K:보기
○+（○+○）	6	0	J:お屋敷	(○)+○	7	14	J:奴等 K:그뭄께
●+（●）	8	0	J:御礼、御縁	(●+●)+●	11	7	J:年小者 K:輕薄兒
○+（●）	7	0	J:御客、御宅	(●+●)+○	6	6	J:女郎屋 K:雇工꾼
お(ご)+動名詞	28	0	J:御払い御詫び	**動詞**			
動詞				語基+するhada	21	161	J:反抗する K:생각하다
お(ご)+動詞	14	0	J:御言う御出づ	**形容詞**			
				語基+hada	0	32	K:未安하다

次は口語体の派生語の日韓の傾向である。口語体の混種語は、全体的には日韓の差が見られなかったが、詳しく見ると、『J文七』 は「接頭辞+ 語基」が多く、『K東閣』は「語基+ 接尾辞」が多い。日本語の「接頭辞+ 語基」には待遇性の接頭辞に固有語が付いたパターンが圧倒的に多く文語体ではあまり現れないパターンであり、韓国語の「語基+ 接尾辞」は文体の違いとは関係なく語基にhadaが付いた動詞と形容詞が多い。

ところで、『K東閣』に名詞化の接尾辞「기（ki）」が付いたパターンが多くみられるが、すでに述べた日本語に（複合）動名詞が多いことと、その数から考え合わせると、日本語の動名詞は、単純語の動詞の連用形とそれを含んだ複合語のパターンが多いことになる。

７． おわりに

以上のように近代語における日韓の語彙をいくつかの観点から眺めてきたのであるが、調査を行うに当たって、資料の選定や調査単位の長さと幅の問題などについて悩み、限界を多く感じた。

資料の選定については、現代語のようにジャンルや文体が確立されていないことはもちろん、当時期の日韓における資料の多少や時期のずれの問題で同質の日韓の語彙の比較に注意を要するところが多くあったからである。そこで本調査では、日韓の対訳資料を選定したのであるが､対訳資料においても翻訳の方向や量による問題があり、近代語の日韓の語彙を十分に反映しているかという疑問も残る。

調査単位については、二言語間の違いを明らかにするためには同じ調査単位の設定が必要であり、できる限り日韓の調査単位の設定の違いが調査結果に及ばないように注意を払ったつもりであるが､今後検討すべき問題もいくつかあろうかと思っている。

そのほかにも日韓の対訳コーパス構築の技術的な問題や形態素解析の可能の有無と精度の問題もあった。今後対訳コーパスとこれまで構築された近代以降の日韓のコーパスとの位置づけや日韓の近代語彙の調査方法における課題を解決しなければならない。

近代語における日韓の語彙研究は､語彙交流の研究を除けばあまり行われておらず、計量的な語彙の分析はいうまでもない。現代語より比較的日韓の類似性が低かった、近代語の日韓の語彙の特徴を明らかにすることは、近代以前の日韓の語彙の様子がわかることであり、また、現代語への変化を捉えられることにもなるのである。

これまで十分に構築されていない近代以降の日韓のコーパスを作成しつつ、近代語の共時的研究と、現代語への通時的研究を進めていきたい。

文　献

韓榮均(2009)「文体 現代性 判別의 語彙的 準拠와 그 変化-1890년대～1930년대 논설문의 한자어 사용양상을 중심으로-」『口訣研究』23

金恩愛（2003）「日本語の名詞志向構造（nominal-oriented structure）と韓国語の動詞志向構造（verbal-oriented structure）」『朝鮮学報』188

金秉喆（1975）『韓国近代翻訳文学史研究』乙酉文化社

国立国語研究所（1964）『分類語彙表』秀英出版

国立国語研究所（1987）『雑誌用語の変遷』（国立国語研究所報告89）秀英出版

宋晩翼（1991）「日本語教育のための日韓指示詞の対照研究 -「コ・ソ・ア」と「이・그・저」との用法について - 」『日本語教育』75

田中牧郎（2010）「雑誌コーパスでとらえる明治･大正期の漢語の変動」『国際学術研究集会漢字漢語研究の新次元予稿集』

張元哉（2000）「19世紀末の韓国語における日本製漢語-日韓同形漢語の視点から-」『日本語科学』8

----（2003a）「現代日韓両国語における漢語の形成と語彙交流」『国語学』54-3

----（2003b）「現代日韓語彙の対照研究-対訳コーパスを資料に-」『日本学報』(韓国日本学会)55 - 1

----（2004）「조사단위의 길이와 현대 한일어휘」『日本学報』(韓国日本学会)61-1

中野洋（1976）「「星の王子さま」6か国語版の語彙論的研究」『計量国語学』79

飛田良文（1973）「現代漢語の源流」『言語生活』259

前田富祺（1984）「語種構造の漸次相」『日本語学』3-9

李漢燮（1985）「『西遊見聞』の漢字語について-日本から入った語を中心に-」『国語学』141

李建志（2000）「海を渡った人情噺-朝鮮開化期の文学『東閣寒梅』と「文七元結」-」『江戸の文事』ぺりかん社

李鐘洙（2010）『韓・日 翻訳 聖書의 語彙 比較 研究-1900年 以後 刊行된 「로마서」를 中心으로』韓南大学大学院韓日語文学比較学科博士論文

林八龍（1995）「日本語と韓国語における表現構造の対照考察 - 日本語の名詞表現と韓国語の動詞表現を中心として - 」宮地裕・敦子先生古希記念論文刊行会（編）『宮地裕・敦子先生古希記念論集　日本語の研究』明治書院

손세모돌(1996)『국어보조용언연구』한국문화사

共同研究発表会開催記録

第１回　2009 年 12 月 13 日（日）14:30～16:00
国立国語研究所中会議室１
「研究計画の概要と国立国語研究所の近代語資源」（田中牧郎）
「近代語テキストの形態素解析」（小木曽智信）

第２回　2010 年 3 月 1 日（月）13:00～15:30
国立国語研究所多目的室
「近代語研究資料のリスト選定における諸問題」（小野正弘）
「近代語彙史をとらえるコーパスの設計」（田中牧郎）

第３回　2010 年 7 月 4 日（日）　13:30～17:00
国立国語研究所セミナー室
「コーパス設計のための近代語文献リストについて」（田中牧郎）
「使用漢字における経験的重みづけと度数調査―東京築地活版製造所四號五號活版摘要文字鑑と太陽コーパス―」（高田智和）
「蘭学資料に見える三字漢語―明治期の三字漢語とのつながりを求めて―」（朱京偉）

第４回　2010 年 12 月 27 日（月）　13:30～16:30
国立国語研究所セミナー室
「現行辞書による近代語語史の構築の危うさ―コーパス利用の可能性を兼ねて」（陳力衛）
「近代語コーパスの資料選定と性格づけ」（小野正弘）

第５回　2011 年 2 月 27 日（日）　13:30～16:30
国立国語研究所 多目的室
「電子化が望まれる近代語資料探索―日本語史を研究する大学院生の報告から―」（岡島昭浩）
「コーパス設計のための明治前期文献の分類」（田中牧郎）

第６回　2011 年 9 月 23 日（金）　13:30～17:00
2011 年 9 月 24 日（土）　10:30～12:00

岩手大学 学生センター B棟 多目的室
「『太陽コーパス』の可能表現形式―形態素解析結果を用いた記述の精密化」（小木曽智信）
「明治前期雑誌の異体漢字と文字コード」（須永哲矢・高田智和）
「近代語末期資料の探索と選定」（小野正弘）

第7回　2011年12月26日（月） 13:00～17:00
国立国語研究所　多目的室
「『明六雑誌』コーパスを用いた一人称代名詞の計量的分析」（近藤明日子）
「近代の地方出身作家の助詞の用法について ―宮沢賢治と濱田広介―」（小島聡子）
「近代語に探る〈終止形準体法〉の萌芽的要素」（島田泰子）
「近代対訳コーパスにおける日韓の語彙の諸相―文体の異なる対訳コーパスの比較を通して―」（張元哉）

※以上は、公開の研究発表会。このほか、非公開の研究会を随時開催した。

執筆者一覧

田中　牧郎　（国立国語研究所言語資源研究系准教授）
岡島　昭浩　（大阪大学大学院文学研究科教授）
小木曽智信　（国立国語研究所言語資源研究系准教授）
小野　正弘　（明治大学文学部教授）
小島　聡子　（岩手大学人文社会科学部准教授）
島田　泰子　（二松学舎大学文学部教授）
朱　京偉　　（中国・北京外国語大学教授）
高田　智和　（国立国語研究所理論・構造研究系准教授）
張　元哉　　（韓国・啓明大学副教授）
陳　力衛　　（成城大学経済学部教授）
近藤明日子　（国立国語研究所コーパス開発センタープロジェクト奨励研究員）
須永　哲矢　（国立国語研究所コーパス開発センタープロジェクト奨励研究員）
金　曦泳　　（韓国・高麗大学校言語情報研究所）
坂井　美日　（大阪大学大学院博士後期課程学生）
竹村明日香　（大阪大学大学院博士後期課程学生）
森　勇太　　（日本学術振興会特別研究員）

国立国語研究所共同研究報告 12-03

近代語コーパス設計のための文献言語研究 成果報告書

2012 年 10 月 31 日発行

著者　田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・
　　朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢
発行　大学共同利用機関法人人間文化研究機構国立国語研究所
〒190-8561 東京都立川市緑町 10-2
電話　042(540)4300（代表）
http://www.ninjal.ac.jp/

ISBN 978-4-906055-23-4　　ISSN 2185-0127

NINJAL Collaborative Research Project Reports 12-03

Study on Documents and Meta-languages for Designing a Corpus of Modern Japanese

Makiro Tanaka, Akihiro Okajima, Toshinobu Ogiso,
Masahiro Ono, Satoko Kojima, Yasuko Shimada,
Jingwei Zhu, Tomokazu Takada, Wonjae Chang,
Liwei Chen, Asuko Kondo, Tetsuya Sunaga

October 2012

NATIONAL INSTITUTE FOR JAPANESE LANGUAGE AND LINGUSITICS