第8回　コーパス日本語学ワークショップ　予稿集

国立国語研究所　言語資源研究系・コーパス開発センター

第8回 コーパス日本語学ワークショップ

予稿集

2015年9月1日、9月2日

主催　国立国語研究所 言語資源研究系・コーパス開発センター

会場　国立国語研究所

第8回 コーパス日本語学ワークショップ

予稿集

2015年9月1日(火)／9月2日(水)

Program [プログラム]

9月1日（火）

10:00～10:10　**■挨　拶**　前川 喜久雄

10:10～12:10　**■口頭発表（1）**

日中同形語の品詞の違いによる誤用について ―中国人の日本語学習者を対象として―
▷何 龍

「日中Skype会話コーパス」を用いた話題別語彙の抽出 ―「食」の場合―
▷中俣 尚己

BCCWJ図書館サブコーパスの代表性試論
▷森 秀明

「通時音声コーパス」は可能か
▷丸山 岳彦

12:10～13:10　**昼食・休憩**

13:10～14:10　**■ポスター発表（1） Aグループ**

『現代日本語書き言葉均衡コーパス』に対する時間情報表現アノテーションの再修正作業
▷浅原 正幸、坂口 智洋、渡邊 友香

『児童・生徒作文コーパス』を用いた漢字使用能力の推定
▷宮城 信、今田 水穂

『虎明本狂言集』における濁点表記状況 ―全例に濁点が付された語を中心に―
▷渡辺 由貴、市村 太郎

『今昔物語集』のコーパス化における非コアデータの精度向上作業
▷池上 尚、鴻野 知暁、河瀬 彰宏、片山 久留美

外来語における[ei]の表記のゆれ
▷小椋 秀樹

14:10～15:10　**■ポスター発表（1） Bグループ**

品詞列・係り受け部分木に基づくラベリングツールの設計と実装 ―節境界ラベリングを例に―
▷浅原 正幸、小西 光、田中 弥生、加藤 祥

形態素解析辞書「中古和文UniDic」を用いた古文単語帳作成
▷大津 千尋、三日市 綾花、須永 哲矢

二字漢語における語と漢字の意味の結びつきの特徴 ―国語辞典の語義の説明文を利用した調査―
▷本多 由美子

テキストの計量語彙論的指標はどのような条件で変化するか
▷山崎 誠

外来語「クレーム」の基本語化とその“挫折”
▷金 愛蘭

『理工学系話し言葉コーパス』における後置詞の特徴
―中級日本語教材をアカデミックなコミュニケーション能力につなげるために―
▷宮部 真由美、菅谷 有子、遠藤 直子、中村 亜美

15:10～15:20　**休憩**

15:20～17:20　**■口頭発表（2）**

中古語における意志系Yes/No疑問文の表現機能 ―日本語歴史コーパス平安時代編を利用して―
▷林 淳子

コーパスによる日本書記古訓形容詞「カシコシ、サカシ」に関する調査
▷劉 琳

漢字とその訓読みとの対応の歴史的変遷
▷芮 真慧

「…事實也。」から「。事実，…」へ ―談話機能の発達に伴う統語位置の変化―
▷柴崎 礼士郎

9月2日（水）

10:00～12:00　**■口頭発表（3）**

文体指標を特徴づける係り受け部分木の抽出
▷浅原 正幸、加藤 祥

助詞の使用実態 ―BCCWJ・CSJにみる分布―
▷丸山 直子

漢語動詞における格表示変化傾向の探索 ―ヲ格とニ格―
▷服部 匡

近代語から現代語にかけての名詞修飾表現の変化についての一考察
―1項名詞に前接する限定詞を例に―
▷庵 功雄

12:00～13:00　**昼食・休憩**

13:00～14:00　**■ポスター発表（2） Aグループ**

『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション
▷植田 禎子、飯田 龍、浅原 正幸、松本 裕治、徳永 健伸

職場における談話の修辞機能と脱文脈化の観点からの分析
▷田中 弥生

節境界認定に関する諸問題
▷佐藤 理史、丸山 岳彦

名詞の項構造データの構築
▷竹内 孔一

ディスカッション観察支援システムFishWatchrを用いた実践手法の提案
▷山口 昌也、大塚 裕子、北村 雅則

万葉集を対象とした原文と読み下し文のアライメント
▷山田 祐実、大村 舞、鴻野 知暁、Kevin Duh、小木曽 智信、松本 裕治

14:00～15:00　**■ポスター発表（2） Bグループ**

日英パラレルコーパスにみる日本語格外連体修飾形の訳され方
▷田辺 和子

コーパスコンコーダンサ『ChaKi.NET』の「文書-部分構造行列」出力機能
▷浅原 正幸、森田 敏生

現代日本語書き言葉均衡コーパス（BCCWJ）のコア・データに基づく関係節付加曖昧名詞句と
先行文脈内の結束連鎖の分析
▷中野 陽子

教科書コーパスを利用した難易度別コロケーション辞書の提案
▷李 在鎬、佐々木 馨

『日本語話し言葉コーパス』UniDic版形態論情報の構築
▷渡部 涼子、田中 弥生、小磯 花絵

アカデミック・ライティングに見られる副詞に関する分析
▷阿辺川 武、八木 豊、ホドシチェク ボル、仁科 喜久子

15:00～15:10　**休憩**

15:10～16:40　**■指定討論**
▷丸山 岳彦、金 愛蘭、丸山 直子、須永 哲矢、中俣 尚己、浅原 正幸

16:40～17:10　**■全体討論**

17:10～17:20　**■総括・閉会**　前川 喜久雄

Contents [目次]

■口頭発表（3）

■ポスター発表（2） Aグループ

■ポスター発表（2） Bグループ

口頭発表（1）

9月1日（火）　10:10～12:10

日中同形語の品詞の違いによる誤用について —中国人の日本語学習者を対象として—

何 龍（愛知淑徳大学大学院グローバルカルチャー・コミュニケーション研究科）†

Misuse of Japanese-Chinese Homographs Differing in Part of Speech: The Case of Chinese Speakers Learning Japanese

He Long(Aichi Shukutoku University,Graduate School of Global Culture and Communication)

要旨

日中同形語の学習において、中国人日本語学習者は品詞性の違いにより、母語からどのような影響を受けるのかを明らかにするため、コーパスにより例文検索を行う。その結果、中国人日本語学習者は母語に影響され誤用を起こす可能性のあることが判明した。そして、学習者作文コーパスを利用し、誤用の可能性を実証できた。本稿は関西大学が編集した『中日同形語小辞典』と曹櫻が編集した『日中常用同形語用法・作文辞典』に重なる 406 語の日中同形語を対象とし、国立国語研究所が開発した『現代日本語書き言葉均衡コーパス』と“教育部语言文字应用研究所”が開発した《国家语委现代汉语平衡语料库》の例文を用いて、研究対象の品詞性の実態を調査した。さらに、「ひのき」プロジェクトが開発した『なたね』と「自然言語処理の技術を利用したタグ付き学習者作文コーパスの開発」科研グループによる『日本語学習者作文コーパス』の例文を用いて分析を行った。

１．はじめに

日本語と中国語は同じ漢字[1]を使用することで、日本語においても、中国語においても、大量の日中同形語が存在している。一見、同じ漢字表記の日中同形語は中国人の日本語学習者にとって、簡単だと思いがちである。しかし、王（2014）の研究によると、中国人の日本語学習者は日中同形語の品詞の違いによる誤用のあることが分かった。王（2014）が事実の発見に止まった、そのような現象の原因に言及していなかった。本稿は関西大学中国語教材研究会（2011）が編集した『中日同形語小辞典』と曹（2009）が編集した『日中常用同形語用法・作文辞典』で重なる 406 語[2]の日中同形語を対象とし、コーパスによる検索の研究手法を用い、中国人の日本語学習者が日中同形語の違う品詞による誤用について検討する。

２．先行研究

２.１　日中同形語の品詞に関する先行研究-

２.１.１　侯（1997）の研究

侯（1997）は「中国人の日本語学習者が日中同形語を使用する際に、意味だけに注目し、品詞に無視してしまう傾向がある」と指摘している。そして、侯（1997）は品詞の違いに

† tcll_helong_1988[a]yahoo.co.jp

1 本稿では、日中同形語は元の漢字表記が同じであれば、同じ漢字表記と見なす。

2『中日同形語小辞典』は 150 語の日中同形語を収録し、『日中常用同形語用法・作文辞典』は 280 語の日中同形語を収録した。ここで断っておきたいのは『中日同形語小辞典』と『日中常用同形語用法・作文辞典』で重なっている 24 語については『中日同形語小辞典』の記載に従う。よって、本稿の研究対象になる日中同形語は 406 語となった。

基づいて、日中同形語を以下の 8 つのタイプに分けた。

表 1　侯（1997）の品詞パターン

タイプ	中国語	日本語
1	動詞	名詞
2	名詞	名詞、動詞
3	名詞、形容詞	名詞、動詞
4	形容詞、副詞	名詞
5	形容詞、副詞	動詞
6	他動詞	自動詞
7	他自動詞	他動詞
8	副詞	タルト形容動詞

しかし、侯（1997）は日中同形語の品詞を基に分類したが、その明確な基準に言及していなかった。そして、日中同形語が数多くあるのは言うまでもない。大量の日中同形語をただ 8 つの品詞タイプに収めるのは難しいと思われる。

２.１.２　張（2008、2009）の研究

張（2008、2009）は国際交流基金・財団法人日本国際教育協会に収録された 1 級～4 級の語彙から日中同形語を抽出した。そして、表 2 のように、張（2008、2009）は抽出した日中同形語を以下の 9 つのタイプに分けた。

表 2　張（2008、2009）の品詞パターン

タイプ	中国語	日本語
1	動詞、形容詞	形容詞
2	動詞、名詞	名詞
3	形容詞、動詞、名詞	名詞、動詞
4	形容詞、名詞	名詞
5	名詞	動詞、名詞
6	名詞	副詞、名詞
7	副詞	動詞
8	副詞	形容詞
9	副詞	名詞

しかし、張（2008、2009）はタイプごとに 1 例しか挙げていない、決して多いといえない。そして、張（2008、2009）は研究で扱われる「上級学習者」の具体的な日本語能力について言及していなかった。さらに、上記の各タイプを見ると分かるように、張（2008、2009）は動詞の自他性について、検討していなかった。よって、結果が一般化できるのは難しいだろう。

２.１.３　熊、玉岡（2014）の研究

熊、玉岡（2014）は独自のデータベースで、二字の日中同形語を検索した結果、1383 語の二字日中同形語が得られた。そして、熊、玉岡（2014）はその 1383 語の二字日中同形語を分析し、その対応関係について検討した。その結果、五つのタイプが得られた。

表 3　熊、玉岡（2014）の品詞パターン

タイプ	特徴	語数（比例）
1	日中両言語で品詞が同じもの。	802(58%)
2	日中両言語で同じ品詞もあるが、 日本語に独自の品詞があるもの。	399(29%)
3	日中両言語で品詞が全く違うもの。	79(5%)
4	日中両言語で同じ品詞もあるが、 中国語に独自の品詞があるもの。	67(5%)
5	日中両言語で同じ品詞もあるが、 中国語と日本語それぞれに独自の品詞があるもの。	36(3%)

２.２　先行研究の問題点

先行研究を調べた結果、日中同形語の品詞についての研究は少なくない。そして、そのような先行研究は基本的に辞書の記載に基づいて、日中同形語の品詞を判断した。しかし辞書に載っている品詞情報は日中同形語が実際に使われている品詞を反映しているかどうかは不明である。最近、大規模コーパスが続々と構築されている。本稿は大規模コーパスに収録した品詞情報を用いて、日中同形語が実際にどのような品詞で使われるのかについて研究する。さらに、日本語学習者作文コーパスに収録した用例文で対照分析を行う。

３．本論

３.１　本稿の目的

本稿は日中同形語の品詞の違いに注目し、その目的は以下の２つである。

1.コーパスによる検索の研究手法で、コーパスに付加している形態素解析情報に基づいて、日中同形語の実際品詞[3]を判定する。

2.実際に違う品詞を持つ日中同形語に対し、日本語学習者コーパスの用例文を用いて対照分析を行う。

３.２　研究対象

本稿は関西大学中国語教材研究会（2011）が編集した『中日同形語小辞典』と曹（2009）が編集した『日中常用同形語用法・作文辞典』に重なる 406 語の日中同形語を対象として、検討する予定である。

曹（2009）が編集した『日中常用同形語用法・作文辞典』は 150 語の日中同形語を収録している。曹（2009）は日中両言語において、同じ漢字で表していながら誤解されやすい語を集め、日本語と中国語ではどう異なるのかを用例文を用いて説明した。さらに、曹（2009）は多数の常用同形語のうち実用的な常用同形語を中心に収録している。

関西大学中国語教材研究会（2011）が編集した『中日同形語小辞典』は HSK 語彙甲級詞[4]の中にある日中同形語 324 語のうち 280 語を収録している。『中日同形語小辞典』はただある言葉の日中異同を調べるだけではなく、語の意味用法の全般について、注意すべき点や

[3] 本稿はコーパスに収録した形態素解析情報による日中同形語の品詞情報を「実際品詞」にまとめる。

[4] 『HSK 語彙大綱』に 8822 個の単語が収集されている。レベルによって、「甲」・「乙」・「丙」・「丁」の四つの級に分けられている。中には、甲級語彙 1033 語 、乙級語彙 2018 語、丙級語彙 2202 語、丁級語彙 3569 語がある。旧 HSK の試験の出題は基礎と初中等は甲・乙・丙級語彙から、高等は丁級語彙までそれぞれの比率を考えて語彙が選ばれる。

語の組み合わせ、類義語群などもできるだけ収録している。

先行研究を調べた結果、中国人の日本語学習者は母語から意味・イメージ・コロケーション・連語形式・品詞などさまざまな影響を受けることが分かった。本稿は「日中同形語の品詞の違いによる誤用」に焦点を当てて研究するため、母語による他の影響を最小限に抑えなければならない。よって、本稿は以下の基準に従い、研究対象を絞ることにする。

1.日本語コーパスにおいても、中国語コーパスにおいても、使用頻度が 50 回以上のものに限定する。

2.文化庁（1978）が収録した S 語（日中両国語における意味が同じか、または、きわめて近いもの）に属するものに限定する。

3.『中日同形語小辞典』と『日中常用同形語用法・作文辞典』は収録した日本語の品詞と中国語の品詞が違うものに限定する。

4.二字の日中同形語に限定する。

３.３　扱うコーパス

研究を進めるため、本稿は国立国語研究所が 2011 年に公開した BCCWJ[5]と中国教育部・言文字・用研究所が 2009 年に公開した《语料库》[6]を使用することにする。

BCCWJ は収録した語数が約 1 億語[7]である。この中には書籍、雑誌、新聞、白書、教科書、広報紙、Web の掲示板、ブログなど多様な日本語が含まれている。サンプルデータは公開されている各種出版データや東京都下の公共図書館の蔵書データを母集団として、そこから無作為に抽出されたものである。

《语料库》は収録した語数が 12,842,116 語である。この中には“人文与社会科学类”、“自然科学类”と“综合类”多様な中国語が含まれている。サンプルデータは主に教材、書籍、新聞、応用文から無作為に抽出されたものである。現在、日本からもアクセスできるようになった。

一方、日本語学習者コーパスを調べたところ、主に、日本語学習者話し言葉コーパスと日本語学習者作文コーパスがあることが分かった。ここで断っておきたいのは本稿が書き言葉に注目するため、日本語学習者作文コーパスを用いて検討する。日本語学習者話し言葉コーパスは研究範囲から外す。よって、本稿は東京工業大学留学生センターが開発した「なたね」[8]という学習者コーパスと「自然言語処理の技術を利用したタグ付き学習者作文コーパスの開発」科研グループが開発した「作文コーパス」[9]を利用する。

「なたね」は日本語学習者から収集した作文に対して日本語教師による添削を行った誤用タグを付与した学習者作文コーパスである。「なたね」は 192 名の日本語学習者[10]による 285 件の作文を収録した。

「作文コーパス」は日本語学習者の作文データをコーパス化したものである。初級から

---

5 本稿は『現代日本語書き言葉均衡コーパス』を BCCWJ と称する。以下は同様。

6 本稿は《国家语现代汉语语料库》を《语料库》と称する。以下は同様。

7 本稿は書き言葉に焦点をあてて研究するため、「yahoo 知恵袋」、「yahoo ブログ」と「国会会議録」という話し言葉を含む可能性が高いジャンルを研究対象から外す。よって，BCCWJ の総語数は 79,357,975 語となった。

8 本稿は『日本語学習者コーパス「なたね」』を「なたね」と称する。以下は同様。

9 本稿は『日本語学習者作文コーパス』を「作文コーパス」と称する。以下は同様。

10 本稿は日中同形語に焦点を当てるため、学習者の母語は中国語に限定する。よって、「なたね」は 115 名の中国人の日本語学習者による 152 件の作文を収録した。

上級の日本語学習者 304 名[11]の作文データが収録されている。作文のテーマは「外国語が上手になる方法について」[12]（192 名分）と「インターネット時代に新聞や雑誌は必要か」[13]（112 名分）である。

３.４　分析の手順

本稿は主に以下の手順で研究対象となる日中同形語ついて分析を行う。

1.『中日同形語小辞典』と『日中常用同形語用法・作文辞典』が収録した 406 語の日中同形語の品詞情報を収集する。そして、収集した品詞情報に基づいて、日中両言語において違う品詞を持つ日中同形語をまとめる。

2.BCCWJ と《语料库》の形態素解析情報を参照し、手順 1 でまとめた違う品詞を持つ日中同形語の実際品詞情報を収集する。そして、収集した実際品詞情報に基づいて、日中同形語の実際品詞を判断する。

3.手順 2 で実際品詞が違う日中同形語に対し、「なたね」と「作文コーパス」の用例文を用いて、対照分析を行う。

３.５　結果の分析

３.５.１　結果の概要

前にも述べた手順に従い、本稿は『中日同形語小辞典』と『日中常用同形語用法・作文辞典』が収録した 406 語の日中同形語に絞ることにした。その結果、違う実際品詞を持つ日中同形語の 38 語を得た。それを表 4 にまとめた。そして、検討しやすいように、その 38 語の日中同形語の実際品詞をまとめた結果、17 個の実際品詞パターンが得られた。それを表 5 にまとめた。

表 4　違う品詞を持つ日中同形語[14]

日本語	記載品詞[15]	実際品詞[16]	中国語	記載品詞	実際品詞
安心	n v a	n v a	安心	a v	a v
以上	n j	n adv	以上	h	n
一切	n adv	n adv	一切	t	r
一般	n a	n	一般	n a t	a
永遠	a n	n	永远	adv	d
可能	n a	a	可能	n a t	v

11 本稿は日中同形語に焦点を当てるため、学習者の母語は中国語に限定する。よって、「作文コーパス」は 160 名の中国人の日本語学習者による作文を収録した。そのうち、「外国語が上手になる方法について」（103 名分）と「インターネット時代に新聞や雑誌は必要か」（57 名分）が収録された。

12「外国語が上手になる方法について」は「自然言語処理の技術を利用したタグ付き学習者作文コーパスの開発」科研グループが収集したものである。

13「インターネット時代に新聞や雑誌は必要か」は東京外国語大学の伊集院郁子氏が収集したものである。

14 表 1 では、名詞を「n」で表記する。動詞を「v」で表記する。形容詞と形容動詞を「a」で表記する。副詞を「adv」で表記する。助詞を「j」で表記する。方位詞を「h」で表記する。代名詞を「r」で表記する。その他の品詞を「t」で表記する。以下は同様。

15 本稿は『中日同形語小辞典』と『日中常用同形語用法・作文辞典』に収録した日中同形語の品詞情報を「記載品詞」にまとめる。以下は同様。ただし、『中日同形語小辞典』と『日中常用同形語用法・作文辞典』の品詞記載が異なる場合、『中日同形語小辞典』に記載に従うことにする。

16 本稿は「品詞の違い」に注目するため、「実際品詞」が同じと判断されたものを研究対象から外す。さらに、先行研究によると、一般的にはある品詞の使用頻度は全体使用頻度の 5%未満の場合、品詞として認定しない。よって、本稿はその基準に従い、実際品詞を判断する。以下は同様。

科学	n	n	科学	n a	n a
開始	n	n v	开始	n v	v
完成	v	n v	完成	n v	v
基本	n	n	基本	a adv	a n adv
苦心	n v	n v	苦心	n a	n
結果	n adv	n adv	结果	n k	n
結局	n a	n adv	结局	n	n
結論	n v	n v	结论	n	n
健康	n a	n a	健康	ad	a
構造	n	n	构造	n v	n v
差別	n v	n v	差别	n	n
使用	v	n v	使用	n v	v
需要	n	n	需要	n v	n v
習慣	n	n	习惯	n v	n v
信用	n v	n v	信用	n	n
絶対	n adv	n adv	绝对	a	a adv
全部	n	n adv	全部	n adv	n
増加	n v	n v	增加	v	v
達成	n v	n v	达成	v	v
担当	n v	n v	担当	v	v
注意	n v	n v	注意	v	v
提出	n v	n v	提出	v	v
適当	a v	n v a	适当	a	a
電話	n v	n v	电话	n	n
努力	n v	n v	努力	n v a	v
特別	n a adv	a	特别	a adv	adv a
販売	n v	n v	贩卖	v	v
批評	v	n v	批评	n v	v
比較	v	n v	比较	v adv	adv
非常	n a	a	非常	a adv	v
変動	n v	n v	变动	v	v
友好	a	n	友好	n a	a

表 5　違う品詞を持つ日中同形語（パターン別）

番号	パターン 日本語	パターン 中国語	日中同形語[17]
1	n v a	a v	安心
2	n adv	n	以上　結果　結局　全部
3	n adv	r	一切
4	n	a	一般　友好
5	n	adv	永遠
6	a	v	可能　非常
7	n	n a	科学
8	n v	v	開始　完成　使用　増加　達成　担当

[17] 表 5 では、日中同形語はすべて日本語の漢字で表記する。

			注意　提出　努力　販売　批評　変動
9	n	a n adv	基本
10	n v	n	苦心　結論　差別　信用　電話
11	n a	a	健康
12	n	n v	構造　需要　習慣
13	n adv	a adv	絶対
14	n v a	a	適当
15	a	adv a	特別
16	n v	adv	比較

３.５.２　結果の分析

「なたね」でパターン 1 の「安心」を調べた結果、以下の用例文が見つかった。

1.お金がなくても、安心[18]に研究できることがわかるから。わざと給料を多くあげないのでしょうか。（049__a）「形容詞」[19]
2.お年寄りが安心して生活できる世の中にするためにも、全部かたかなでは無理です。（078__a）「動詞」
3.帰国して家族の安心感が得られたが、自分の国に帰ったって安全だとけっして言い切れない。（p33_a　非中国語母語話者）「名詞」
4. 以上の方法は外国語を勉強する人にとって役に立つと思うが自分の状況によってもっといい方法を探したほうがいいだ。（CN314）「名詞」

中国人の日本語学習者による用例文が 3 つ見つかった（そのうち、例 1 のような形容詞が 1 例で、例 2 のような動詞は 2 例である）。そして、「作文コーパス」を調べた結果、中国人の日本語学習者による用例文が 2 つ見つかった（2 例は動詞である）。さらに、「なたね」で非中国語母語話者の作文を調べた結果、「安心」を名詞として使われる例 3 が見つかった。よって、中国人の日本語学習者が母語の品詞に影響され、母語に存在しない品詞の使用を避ける傾向のあることが見られる。さらに、「作文コーパス」でパターン 2 の「以上」について調べた結果、上級者でも名詞の「以上」しか使わない。副詞の「以上」の用例文は一つもないことが分かった。それは上級者の学習者も母語の品詞に影響されていると言えるだろう。パターン 2、10、13、14 の用例文を分析した結果、同じ傾向が見られる。

表 5 を見ると分かるように、パターン 3、4、5、6、12、16 は日本語の品詞は中国語の品詞と全く異なっている。「作文コーパス」を調べた結果、以下の用例文が見つかった。

5.この産品に関する紹介だし、専家たちからの評論だし、似る産品の比較だし、単に一つのニュースなくて、色々知っています。（CG112　中級者）「名詞」
6.インターネットと新聞などと比較すると、何となくつめたい感じがする。（CG139　中級者）「動詞」

以上の用例文を見ると分かるように、中級者の学習者は「比較」の名詞と動詞の品詞を正しく使えるようになった。なぜ中級者は異なる品詞を持つ「比較」を正しく使用できる

18 本稿では、キーワードとなる語彙に下線をつける。以下は同様。
19 本稿では、筆者は学習者作文コーパスによる用例文の品詞認定を行った。

のか。これは中国人の日本語学習者は母語の品詞との違いに気づくからだと考えられる。

一方、「作文コーパス」で韓国人の日本語学習者の作文を検索した結果、上級者の学習者でも、名詞の「比較」を使わないことが分かった。よって、中国人の日本語学習者が母語から正の影響を受け、日本語は母語との違う品詞に気づき、正しく使用できる傾向のあることが見られる。

表 5 を見ると分かるように、パターン 8 の日本語が名詞または動詞で使われ、中国語が動詞で使われる日中同形語は他のパターンより圧倒的に多かった。熊、玉岡（2014）によると、パターン８のような日中同形語は「初級段階の中国人の日本語学習者にとっては難しいが、日本語能力が上がるにつれ、習得できるようになると予測される」ということが分かった。実際に、「作文コーパス」で「注意」を調べた結果、その中には以下の用例文が見つかった。

7.しかし、外国語が好きなら、平素でよく注意し、復述し、だんだんうまくなる。（CG035　初級者）「動詞」
8.もっといい方法を見つかることができるようこれからの日本語の勉強で注意を払うと思う。（CN308　中級者）「名詞」
9.注意しないと全くわからない場合もある。（CG025　中級者）「動詞」
10.しかし、外国語が好きなら、平素でよく注意し、復述し、だんだんうまくなる。（CG035　初級者）「動詞」

実際に、「作文コーパス」を調べた結果、初級者の作文は 7 例見つかった。7 例は全部例 7 のように動詞として使われることが分かった。さらに、中級者の作文は 9 例見つかった。9 例のうち、例 9 のように動詞として使われるのは 8 例で、例 8 のように名詞として使われるのは 1 例である。その傾向は熊、玉岡（2014）の予測と一致している。よって、パターン 8 の日中同形語の品詞の習得は初級段階の中国人の日本語学習者にとっては難しいが、日本語能力が上がるにつれ習得できる傾向が見られる。

表 5 を見ると分かるように、パターン 11 の日本語が名詞または形容詞で使われ、中国語が形容詞で使われる。実際に、「なたね」で「健康」を調べた結果、その中には以下の用例文が見つかった。

11.健康が一番だと両親に言われて、勉強をひとまずやめて帰国するしかないと彼女は言いました。（061__a）「名詞」
12.十分な家庭教育や子供との接することが出来ないため、子供の心身的に健康な成長ができるかどうか心配が増えかねない。（127__c）「形容詞」
13.大部分の高齢者は、健康に、幸せに、経済力の持つ生活を送ることがだきると思います。（159__a）「形容詞」

「なたね」で「健康」を調べた結果、全部で 7 例が見つかった。その中には、中国人の日本語学習者による用例文は例 11～13 のように名詞の 1 例と形容詞の 2 例があり、非中国語母語話者による用例文は形容詞の 4 例がある。なぜ中国人の日本語学習者だけは「健康」を名詞で使用するのか。これは母語から正の影響を受けるからだと考えられる。《・料・》で“健康”を調べた結果、形容詞の用例文は 1112 例があるのに対し、名詞の用例文は 13 例しかない。よって、中国人の日本語学習者は母語の品詞から正の影響を受け、パターン

11 の日中同形語を正しく使用できる傾向があると言えるだろう。

表 5 を見ると分かるように、パターン 9 の日本語が名詞で使われ、中国語が形容詞、名詞と副詞で使われる。実際に、「なたね」で「基本」を調べた結果、その中には以下の用例文が見つかった。

14.大量のロボットを使ったら、失業率がますます増えます。失職した人々は、生活の基本保証ができなくて、社会の不安定に導くに違いない。（043__a）「形容詞」
15.これは基本的だが、文法のような書面のものにこだわりすぎる。（CG047）
『これは基本だが、文法のような書面のものにこだわりすぎる。』（添削後）[20] 「名詞」

以上の例 14 を見ると分かるように、中国人の日本語学習者は母語の品詞から影響を受け、日本語に存在しない形容詞の「基本」を過剰に使用する恐れがある。さらに、例 15 を見て中国人の日本語学習者は中国語“基本”の形容詞の品詞から影響を受け、日本語が名詞で使用すべきものに「的」をつける誤用のあることが分かった。よって、中国人の日本語学習者は母語の品詞から負の影響を受け、日本語に存在しない品詞を過剰に使用する傾向のあることが見られる。パターン 7、パターン 15 の用例文を分析した結果、同じ傾向が見られる。

４．まとめ

本稿は日中同形語の学習において、中国人日本語学習者は品詞性の違いにより、母語からどのような影響を受けるのかを明らかにするため、コーパスにより例文検索を行う。その結果、中国人日本語学習者は母語に影響され誤用を起こす可能性のあることが判明した。その具体的な結果は以下の通りである。

1.中国人の日本語学習者が母語の品詞に影響され、母語に存在しない品詞の使用を避ける傾向がある。さらに、上級者の学習者も母語の品詞に影響されている傾向がある。
2.中国人の日本語学習者が母語から正の影響を受け、日本語は母語との異なる品詞に気づき、正しく使用できる傾向がある。
3.日中同形語の品詞を習得する際に、初級段階の中国人の日本語学習者は難しいが、日本語能力が上がるにつれ習得できる傾向がある。
4.中国人の日本語学習者は母語の品詞から正の影響を受け、日中同形語を正しく使用できる傾向がある。
5. 中国人の日本語学習者は母語の品詞から負の影響を受け、日本語に存在しない品詞を過剰に使用する傾向がある。

５．今後の課題

今回の研究は課題がいくつか残っている。それを今後の課題として検討する。

1. 本稿は『中日同形語小辞典』と『日中常用同形語用法・作文辞典』が収録した 406 語の違う記載品詞を持つ日中同形語に絞り、検討をした、今後、さらに研究対象を増やすつもりである。そして、同じ記載品詞を持つ日中同形語にも視野に入れて検討するつもりである。
2. 今回の研究では、「作文コーパス」と「なたね」を使用し、中国人の日本語学習

[20] 本稿は日本語の誤用に対する添削を行ったものに、「(添削後)」で表記する。

者の作文実例を調べたが、今後、さらに中国人の日本語学習者の作文実例を増やし、検討していきたい。

今回の研究は今まで日中同形語の意味、持つイメージ、コロケーション、連語形式と同じく、日中同形語に関する基礎研究にすぎない。これからは、このような基礎研究を数多く実施することによって、中国人の日本語学習者の日中同形語の学習に貢献できればと願っている。

文　献

王燦娟（2014）「中国人日本語学習者に見られる日中同形語の誤用について:意味、品詞、共起の誤用をめぐって」『東アジア日本語教育・日本文化研究』、17 号 pp.221-241

何龍（2013）「日中同形語の学習における母語の影響について:中国人の日本語学習者と日本人の中国語学習者を比較して」、修士論文

何龍（2014）「日中同形語の学習における母語の影響について:日本人の中国語学習者を対象として」『愛知淑徳大学論集グローバルカルチャー・コミュニケーション研究科篇』、6 号 pp.85-100、
（http://aska-r.aasa.ac.jp/dspace/bitstream/10638/5526/1/0033-006-201406-085-100.pdf よりダウンロード可能）

何龍（2015）「日中同形語の持つイメージ:「感染」を例として」『愛知淑徳大学論集グローバルカルチャー・コミュニケーション研究科篇 』、7 号 pp.57-71、
(http://aska-r.aasa.ac.jp/dspace/bitstream/10638/5681/4/0033-007-201503-057-071.pdf よりダウンロード可能）

関西大学中国語教材研究会（2011）『中日同形語小辞典』、白帝社

熊可欣、玉岡賀津雄（2014）「日中同形二字漢字語の品詞性の対応関係に関する考察」『ことばの科学』、27 号 pp.25-52
（https://www.lang.nagoya-u.ac.jp/~ktamaoka/scholarly/sadokunasi/2014/049.pdf よりダウンロード可能）

侯仁鋒(1997)「同形語の品詞の相違についての考察」『日本学研究』6 号 pp.78-89.

曹櫻（2009）『日中常用同形語用法・作文辞典』、日本僑報社

張麟声(2008)「中国語話者における日本語漢語語彙の習得について品詞性のずれに起因する習得の問題を中心に」、Linguistics of kango (Japanese words of Chinese origin),Friday 14th and Saturday 15th March 2008, Université Paris Diderot-Paris 7.

張麟声(2009)「作文語彙に見られる母語の転移―中国語話者による漢語語彙の転移を中心に―」『日本語教育』、140 号 pp.59-69

文化庁（1978）『中国語と対応する漢語』、大蔵省印刷局

関連 URL

国立国語研究所　『現代日本語書き言葉均衡コーパス』　http://chunagon.ninjal.ac.jp/

中国教育部・言文字・用研究所　《・料・在・》　http://www.cncorpus.org/index.aspx/

東京工業大学留学生センター　『学習者作文コーパス「なたね」』
https://hinoki-project.org/natane/

「自然言語処理の技術を利用したタグ付き学習者作文コーパスの開発科研グループ」『日本語学習者作文コーパス』　http://sakubun.jpn.org/

「日中 Skype 会話コーパス」を用いた話題別語彙の抽出 —「食」の場合—

中俣 尚己（京都教育大学）†

Extraction of Topic-Specialized Vocabulary from "Skype Corpus" : A Case for the Topic of 'Eating'

Naoki Nakamata(Kyoto University of Education)

要旨

本発表では、発表者が構築した「日中 Skype 会話コーパス」を用い、会話で使用される語彙について分析する。このコーパスは日本の大学生と中国の大学生が Skype で会話交流活動を行ったのを継続的に録音、文字化したもので、真正な会話であるとともに、各回の話題が指定されていることに特色がある。今回は「食」がテーマの回とそれ以外のテーマの回に分け、日本語解析システム「雪だるま」を使って単語に分割した。その後、LLR を指標として「食」関連語が抽出できるかを検証した。結果、特徴度が高かった語は基本的に「食」に関連する語であり、高い精度で抽出できた。これは、会話コーパスにおいて話題の設定が重要であることを再確認できたと言える。

１．　はじめに

この発表の目的は２つある。１つは発表者が構築し、2015 年 4 月 1 日から公開している『日中 Skype 会話コーパス』の諸特性を紹介することである。もう１つは、その特性の１つである「会話の話題が決められている」点に着目し、話題別の語彙抽出を行った結果を示すことである。結果は高い精度を示しており、会話コーパスの構築においてはごく簡単にでも話題をあらかじめ決めておくことで、語彙表の作成に役に立つデータを得ることができると言える。

２．　『日中 Skype 会話コーパス』の紹介

２．１　『日中 Skype 会話コーパス』の概要

『日中 Skype 会話コーパス』は 2012 年５月～７月に、東京・実践女子大学と長沙・湖南大学の学生間で行った Skype を利用した遠隔会話活動(中俣ほか 2013)を録音、文字化したもので、接触場面の会話コーパスに分類される。中国側の学習者は全員２年生で、日本側の母語話者は学部３年～M１の学生で日本語教育を専攻したり、関連する授業を受講していた学生である。３ヶ月の間、ペアを固定し、１週間に１度のペースで Skype を用いた会話活動を行った。実際にはビデオ通話ではあるが、行ったのは録音のみで、現時点で公開しているのはその文字化資料のみとなる。

コーパスには延べ９ペア、３８の会話を収録している。総会話時間は 46:48:35 で、１会話あたり平均 1:13:55 とまとまった長さの会話と言える。後述する日本語解析システム「雪だるま」を使って分析した結果、総語数は 204,632 語であった(記号類を除く)。

コーパスはテキストファイルで提供され、笑いや発話の重なりといった簡単な記号を含んでいるが、これらは正規表現で簡単に取り除けるようになっている。コーパスの配布は

† nakamata[at]kyokyo-u.ac.jp

http://nakamata.info/database.html で行っている。氏名とメールアドレスを登録すればすぐにダウンロードできる。

会話活動の詳細な報告は中俣ほか(2013)、Skype コーパスそのものの説明については中俣(2015)にて詳しく説明している。

２．２　『日中 Skype 会話コーパス』の特性

『日中 Skype 会話コーパス』の言語資料としての特徴として、以下の４つを挙げる。

Ａ．　真正性がある。

このコーパスの設計はもともとコーパスを作ろうとしたものではなく、まずは Skype を用いた会話活動を通し、中国の学習者には学んだ日本語を使う機会を提供するとともに学習意欲を継続させること、日本の母語話者には外国人と文化交流をしたり日本語を教えたりしながら、日本語について考えてもらうことが第一の目的であり、それにあわせて計画がデザインされている。そのため、真正性のある接触場面コーパスになっている。以下、いくつかの語について、代表的な学習者コーパスである KY コーパスと比較したものが表１である。OPI という統制された会話である KY コーパスには出現しないような語が多数出現していることがわかる。

表１　KY コーパスと日中 Skype 会話コーパスの出現数の比較[1]

語	KY コーパス	日中Skype会話コーパス
明後日	0	7
木曜	6	41
すごい	77	211
すごく	190	86
すげえ	0	4

Ｂ．　縦断的なデータである。

会話活動は１週間に１回、継続的に行った。最も多いペアで７回分の会話があり、縦断的にデータを観察することができる。

Ｃ．　一種の電話場面である。

終結部には、例えば突然食事の話題をふって、会話を終結にもっていく前終結の段階が存在するなど、電話場面と同様の構造が観察される(橋内 1999)。　また、コミュニケーション・ブレイクダウンや沈黙も多く観察される。

Ｄ．　話題が指定されている。

各回は次ページの表２のように話題が指定されており、数字はファイル名の末尾の数字

[1] 北村・冨岡・川村(2009)はコーパスの出現文書数から語の難易度を求める試みであるが、「あさって」「おととい」のような語は基本語であるものの、コーパスに出現しにくいという問題点を指摘している。また、CSJ と BCCWJ の調整頻度レベルでは一番頻度が少ない曜日は木曜である(Tono, Yamazaki and Maekawa 2013)。

に対応する。しかし、話題は必ずしも厳密に守られているわけではなく、話がそれたり日本語についての質問が行われることもある。これらの話題は事前に日中双方の学生から話してみたいことのアンケートを行い、決定した。

敬語に関しては張(2012)が、敬語について学習者で意義などについて話し合うことの効果を報告していることから採用した。

表２　日中 Skype 会話コーパスの話題

1	ポップカルチャー	6	伝統・行事
2	料理	7	夏休み・夏の予定
3	家庭・家族・子供	8	大学生活
4	故郷・今住んでいる場所	0	指定なし・トピック認定できず
5	敬語		

３．　「食」関連語彙の抽出

３．１　特徴語抽出の意義

日本語教育における教材作成において、語彙の選定は重要な作業である。中俣(2014)は文法積み上げ型シラバスを念頭に、特定の文法項目と共起する語彙をピックアップしているが、現在では話題シラバス・場面シラバスの教材も増えてきている。話題シラバス・場面シラバスの教材作成にあたっては、話題ごとにどのような語彙が用いられるかということが重要である。

話題ごとの語彙をまとめた重要な先行研究として山内(2013)『実践日本語教育スタンダード』（以下、実践 S）をあげることができる。実践 S はまず 100 の話題を選び、各話題ごとにまず文型を設定する。そしてその文型に入りうる名詞をパラディグマティックな形で提示したものであり、各名詞は難易度によって３段階に分けられている。実践 S の最初の話題は「食」であり、以下、「1.1.1.1.　食名詞：具体物」の【料理名：個体】の名詞を引用する。

表３　山内(2013)『実践日本語スタンダード』の一例

意味分類	A	B	C
【料理：固体】	カレー、パン、ごはん、サラダ、うどん、そば	サンドイッチ、ステーキ、ハンバーグ、刺身	ライス、粥、実、麺、漬物、～漬け

しかし、これらの語のピックアップや難易度判定は執筆者の主観に基づくものである。会話コーパスから機械的に話題関連語を抽出できれば、客観的かつ大規模な語彙表を作成することができ、さらに教材作成に活かすことができる言語資料となることが期待される。そこで本発表では、『日中 Skype 会話コーパス』から「食」関連語彙を機械的に抽出し、既存の語彙表である実践 S との比較を行う[2]。

2　ただし、実践Ｓの批判が目的ではない。山内(2013)は以下のように述べる。

このようなパラディグマティックに対立する語群を眺めると、語同士を直接比較できるようになるため、個々の語のレベル設定が非常に行ないやすくなる。（略）「同じ文の同じ位置に現れ得る語同士

３．２　手法

まず、コーパス全体を「料理」が話題の食コーパスとそれ以外が話題の対照コーパスに分割した(語数は食コーパスが 28,960 語、対照コーパスが 175,352 語)。一方で、学習者と母語話者の発話は分割しなかった。これは、表４に示す通り、接触場面においては学習者と母語話者の語彙に顕著な差は存在しないからである。

表４　『日中 Skype 会話コーパス』における話者別の異なり語数と延べ語数

話者	異なり語数	延べ語数	TTR
中国人学習者	5,374	103,883	0.0517
日本人母語話者	4,923	100,749	0.0489

細かく語彙を分析しても「母語話者はよく使うが、学習者はあまり使わない」あるいはその逆の語というものは一部の機能語的な語に限られていた[3]。実質語に絞って話者別に特徴語を抽出しても話題別の特徴語よりも少ない量しか抽出できない。特徴語を抽出する上では語数は多いほうが良いため、話者による語彙の違いは捨象して計算した。

次に、各コーパスを日本語解析システム「雪だるま」(http://snowman.jnlp.org/)にかけ、単語ごとに分割、品詞も付与した[4]。この「雪だるま」は長岡技術科学大学の山本和英氏が開発したシステムで、形態素ではなく「単語」に分割することを目的とし、「気が早い」のような慣用句、「かもしれない」のような複合辞、「勉強する」のようなサ変動詞、「無理だ」のような形容動詞をそれぞれ１語として出力することができる。解析は 2015 年 7 月 18 日に行った。

最後に、解析結果を元に、特徴度の指数として、田中・近藤(2011)を参考に対数尤度比(LLR)を補正した値を計算した。計算式は下記の通りである。

2(alna+blnb+clnc+dlnd-(a+b)ln(a+b)-(a+c)ln(a+c)-(b+d)ln(b+d)-(c+d)ln(c+d)+(a+b+c+d)ln(a+b+c+d))

a：当該資料での当該語の度数　b：参照資料での当該語の度数

c：当該資料の延べ語数－a　d：参照資料の延べ語数－b

ln は自然対数を表す。a または b が 0 の場合、alna または blnb を 0 として計算する。

ad-bc<0 の場合の場合、-1 を乗じる補正を行う。

教科特徴語リストに合わせ 0.1%水準で有意となる 10.83 よりも大きい語を「食」特徴語と認定する。

の比較が可能」ということに大きな意味がある。（略）また、表９(発表者注：上記表３のこと)を見ると、「パスタ」と「ラーメン」が入っていないことに気づく。「パスタ」と「ラーメン」が入っていないことに気づくことができるのも、パラディグマティックに対立する語が集められていることの賜物である。従来よく見られた五十音順の配列による語彙表では、よほどのパスタフリーク、ラーメンマニアでない限り、「パスタ」や「ラーメン」がないことには気づかないものと思われる。(p.12)

つまり、実践Ｓは話題関連語がパラディグマティックに配列されるという「枠」を示したことに大きな価値がある。本発表はその「枠」の中にさらに実際のデータから具体的な語を入れ込むことができるか、という検証であり、両者は相補的な関係にあると考える。

[3] どのような語に差異が見られるのか、またなぜ実質語には差異が見られないのかといった考察は別稿(中俣　準備中)に譲る。

[4] 2015 年７月現在、限定公開となっている。興味をお持ちの方は山本和英氏まで。

３．３　結果

発話の断片（「レタス」と言おうとして「タス」になったものなど）を誤解析したものを除くと、244 語を抽出できた。これは食コーパスのうち、異なり語数の 11.9%、延べ語数の 16.0%をカバーする。表５に品詞ごとの数を示す。また、この数字はあくまでも機械的に抽出された語数である。そこで、実際に目視でそれぞれの語が食に関連する意味で使われているかを確認した。

表５　品詞ごとの「食」特徴語の語数

名詞 (複合名詞)	動詞 （非自立含む）	形容詞 （非自立含む）	その他 （副詞、感動詞、助詞、助動詞、複合辞）
190 語　83.7%	35 語 80.0%	11 語　90.9%	8 語

感動詞や助詞（「なあ」）が特徴語とは考えられないが、助動詞「られる」、複合辞「ないで」に関しては、食の場面でよく使用される可能性は考えられる。今後の課題としたい。

＜例１＞

Ｃ：うん。なぜ日本では、このチンジャオロースはとても有名です、か。

Ｊ：家庭一でよく食べます。中華料理の中でも、＜うん＞よく作られる。

＜例２＞

Ｊ：朝ごはん食べないで会社とか学校行って、お昼食べて夜食べて、の２食っていう生活の人、が多いですね。

以下、表６、７、８はそれぞれ名詞、動詞、形容詞・副詞の語彙リストであり、実践 S にならって提示してみる。

表６　「食」特徴語名詞リスト（190 語/83.7%）

【食べ物】料理、食べ物、もの
【食事】朝ごはん、弁当、給食、朝食、夕食、間食、昼食、懐石料理、昼
【料理名・固体】年越し、刺身、煮物、餃子、パン、寿司、餅、粥、ピータン、チンジャオロース、肉じゃが、麺類、ご飯、天ぷら、麺、ワンタン、焼き魚、チャーハン、回鍋肉、お好み焼き、カレー、ハンバーガー、きりたんぽ、ハンバーグ、ピザ、焼きそば、くさや、酢豚、ダック、卵焼き、サンドイッチ、スペアリブ、天津飯、水餃子、麻婆豆腐、関東煮、天津丼、中華丼、北京ダック、ピータン豆腐、チャオピン、親子丼、卵かけごはん、ジャージャー
【料理名・液体】スープ、味噌汁
【菓子・デザート】まんじゅう、肉まん、あんまん、クレープ、菓子、アイスクリーム、綿あめ、饅頭、ホットケーキ、綿、中華まん、チョコまん
【飲み物】梅酒、牛乳、紅茶、豆乳、酒、ジャスミン茶、日本酒、緑茶
【食材】肉、パスタ、アヒル、卵、なす、トマト、玉ねぎ、野菜、小麦、じゃがいも、犬、米、魚、ピーマン、レタス、生卵、納豆、いちご、中身、パプリカ、大根、食材、ネギ、にんじん、乾物、のり、小麦粉
【調味料】醤油、塩、わさび、あんこ、つゆ、山椒、油、めんつゆ、ティエン、調味料

【調理器具】鍋
【調理の場所】台所
【食器】椀、皿、箸
【飲食店】食堂、餅屋、回転ずし
【行列】満員
【味】味、舌、バニラ、味覚
【食欲】食欲
【団らんの場所】テーブル
【量】１杯、２杯
【調理法】生、生もの、固め
【未分類】茶道、赤、つば、系統、値段、１００、黄色、中国料理、日本料理、鍋料理、家庭料理、北京料理、四川料理、比較文化、食文化、１６元、広東料理、１００種類、福建省、東北人、湖南料理
【誤抽出】平成、子供、名刺、元号、字、みず、西暦、オン、メンツ、オッケー、体面、字幕、ビデオ、何、福山、キャンパス、テスト、比較、映像、テキスト、気晴らし、新暦、学期、皇暦、1つ、岳麓山、生田斗真、１時、はなみずき、新垣結衣、聴解、声優

表７　「食」特徴語動詞リスト（35 語/80.0%）

揚げる、切る、食べる、焼く、入れる、煮る、作る、潰れる、つける、煮込む、しびれる、かける、混ぜる、点てる、開ける、食べれる、盛る、冷やす、いためる、作る、たらす、さっぱりする、くさる、溶く、保つ、つつく、練る、かぐ
【誤抽出】数える、登る、参加する、主演する、通じる、延ばす、鍛える

表８　「食」特徴語形容詞（11 語/90.9%）

甘い、おいしい、辛い、臭い、薄い、辛い、苦い、酸っぱい、安い、簡単
【誤抽出】ふさわしい

３．４　考察

３．４．１　抽出精度とカバー率

まず、誤抽出の語について考えてみたい。ここを見ると、「平成」「元号」「西暦」「皇歴」といった暦に関する語群があることに気づく。これはある会話の終わりに、突然学習者が暦に関する質問をしたためである。その他の誤抽出の語も、会話の一部の個所で集中的に出現しており、別の話題についての個所であることが明白である。

このコーパスでの話題は、前もって表２のテーマについて話すように指示しただけであり、実際に会話参与者がそれを厳密に守っているわけではない。今回、分析対象をファイル丸ごとにしたため、このような語も「食」関連語として抽出されたが、内容を仔細に観察し、話題ごとに区切ってコーパスを作れば、誤抽出の語はほぼ全て排除できる。

つまり、話し言葉であれば、規模が数万語のコーパスであっても話題の特徴語は 100%に近い精度で機械的に抽出できるということである。この精度は子供話し言葉コーパスの特徴語分析(中條ほか 2005)、Facebook と Twitter の比較(石井 2011)、twitter を用いた時制関係語の抽出（赤崎ほか 2013）といった他分野の特徴語抽出の試みよりも明らかに高い。多くの実質語は話題に従属するという山内(2013)の方針が実証されたと言えよう。また、こ

の事実は会話コーパスを作る時、緩やかにでも話題を指定しておくと、日本語教育の教材作成に非常に有益な結果が得られるということを意味している。

その一方で、本当にすべての「食」関連語がカバーできているかという問題も残る。例えば、今回の調査では「食」コーパスにのみ、1例だけ出現した「味わう」のような低頻度語は抽出できない。これはコーパスサイズを大きくすることでしか対処できないかもしれない。

３．４．２　直感では気づきにくい特徴語

次に、個々の語について見ていく。もちろん、一見して食に関連する語が多く抽出されたわけであるが、機械的に抽出を行うメリットは直感では見逃してしまうような語も発見することができる点にある。例えば、【食べ物】に分類される名詞として「もの」が抽出されている。その理由は、以下のような例が「食」コーパスに多く見られたためである。

＜例３＞

Ｊ：えーと、ハンバーグというのは、あのー、お肉とか、あの、み、ミンチのお肉とか、あの、タマネギを刻んだものとかを、えーと、ね、練り合わせて、卵とか、小麦粉とかを練り合わせて焼いたもの。これがハンバーグで、ハンバーガーというのは、パンの間にそのハンバーグとか、レタスとか、チーズとかが挟んであるものがハンバーガーです。

「辛いもの」といった単純な例も「食」コーパスに見られたが、＜例３＞のような「～を～したもの」という構文は「食」コーパスにのみ出現した。これは料理の説明をする時に頻用され、また使えると説明がスムーズにいく項目であると言える。

また、動詞では「潰れる」「しびれる」「保つ」などが出現しているが、これらはそれぞれ「お酒を飲んで潰れる」「四川の本格マーボーは舌がしびれる」「調理時、一定温度に保つ」といった文脈で使われている。

これらの構文や語は実践Ｓには収録されていない。

３．４．３　難易度をどう考えるか

実践ＳではＡ、Ｂ、Ｃの三段階で難易度が表示されているが、コーパスの出現頻度から再考できる余地がある。

表９　実践Ｓと「食」コーパスの比較

	焼く	煮る	炒める
実践Ｓ	Ａ	Ｂ	Ｂ
「食」コーパス	48 回	26 回	9 回

また、実践 S では「焼ける」「グラム」「センチメートル」といった語が難易度Ａになっていたが、これらは『日中 Skype 会話コーパス』全体を通しても出現しない。

さらに、「鍋」と「包丁」はどちらも実践ＳではＡランクにあたり、直感的にもどちらも調理に不可欠の道具であるように思えるが、『日中 Skype 会話コーパス』全体での出現数は鍋が 34 回に対し、包丁は 0 回である（フライパンは 4 回）。つまり、実際の重要度と「会話で使用するか」ということは全く別の次元の尺度であり、コーパスからわかる「会話でどのぐらい使うか」という情報が会話教材において重要になると考える。

４．　おわりに

この発表では、『日中 Skype 会話コーパス』の特性について紹介し、「食」特徴語の抽出を行った結果を発表した。会話コーパスの特徴語抽出において話題が果たす役割を極めて大きいと言える。また、機械による特徴語抽出は、直感では気づきにくい語を抽出したり、難易度を考慮することにより、日本語教材作成に貢献できることを示した。

謝　辞

本研究は、JSPS 科研費(26770180)による補助を得ました。また、LLR の計算方法については帝塚山大学の森篤嗣氏、コーパスに出現しにくい語については東京国際大学の川村よし子氏に助言を頂きました。また、単語解析は山本和英氏と長岡技術科学大学自然言語処理研究室のメンバーが作成した「雪だるま」を利用させて頂きました。お世話になった皆様に感謝申し上げます。

文　献

赤﨑優介・森田和宏・泓田正雄・青江順一(2013)「Twitter を用いた時制を表す特徴語の自動収集に関する研究」『言語処理学会第 19 回年次大会発表論文集』

石井健一(2011)「Facebook と Twitter の発言における特徴語の比較 (http://hdl.handle.net/2241/115339 よりダウンロード可能)

北村達也・冨岡洋介・川村よし子(2009)「IDF を用いた単語レベル判定システムの構築と検証」『日本語教育方法研究会誌』16(1), pp.52-53

田中牧郎・近藤明日子(2011)「教科書コーパス語彙表」『言語政策に役立つ、コーパスを用いた語彙表・漢字表等の作成と活用』pp.55-63, 2011 文部科学省科学研究費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：２１世紀の日本語研究の基盤整備」言語政策班

中條清美・西垣知佳子・内山将夫・中村隆宏・山﨑淳史(2006)「子供話し言葉コーパスの特徴語抽出に関する研究」『日本大学生産工学部研究報告 B 文系』39,pp. 65-78,日本大学生産工学部.

張贇(2012)「敬語コミュニケーション学習における「変容」に関する考察 : 上級学習者の事例分析から」『待遇コミュニケーション研究』9, 待遇コミュニケーション学会

中俣尚己・漆田彩・小野真依子・北見友香・竹原英里(2013)「Skype を活用した日中会話交流プログラム」『実践国文学』83, pp.132(25)-109(48), 実践国文学会

中俣尚己(2014)『日本語教育のための文法コロケーションハンドブック』くろしお出版

中俣尚己(2015)「日中 Skype 会話コーパスについて」(http://nakamata.info/about_skype_corpus.pdf よりダウンロード可能)

中俣尚己(準備中)「接触場面における学習者と母語話者の語彙はどこが異なるのか？―「日中 Skype 会話コーパス」の分析―」『日本語／日本語教育研究会第 7 回大会予稿集』

橋内武(1999)『ディスコース　談話の織りなす世界』くろしお出版

山内博之(2013)『実践日本語教育スタンダード』ひつじ書房

Tono, Y., Yamazaki, M., Maekawa, K. (2013) *A Frequency Dictionary of Japanese* Routledge.

関連 URL

中俣尚己のウェブサイト http://nakamata.info/

雪だるまプロジェクト http://snowman.jnlp.org/

BCCWJ 図書館サブコーパスの代表性試論

森　秀明（東北大学大学院文学研究科）†

"BCCWJ Library Sub Corpus" And Its Representativeness

Hideaki Mori (Graduate School of Arts and Letters，Tohoku University)

要旨

『現代日本語書き言葉均衡コーパス』（BCCWJ）の中で、統計分析に適するのは固定長データだと言われている。しかし固定長データのサイズはそれほど大きくない。一方、Sinclair（1991）、バイバー、コンラッド、レッペン（2003）などにより、サイズが小さいコーパスの代表性はさほど高くないことが主張されている。BCCWJ のマニュアルには、語彙の偏りを防ぐためにサンプルを短くしたとの記述が見られるが、その効果を具体的に検討した報告書類は見いだせない。このため語彙表を使用して固定長と可変長の頻度比較による検証を行った。この結果、高頻度語はデータ量に正比例して頻度が増加するが、低頻度語や特定のトピックに使用されやすい固有名詞と普通名詞などは、頻度がばらついて増加することが分かった。代表性が高ければ基本的に頻度のばらつきは生じないと考えられるため、これらの代表性はそれほど高くない可能性がある。

１．研究の目的

あるコーパスが、推定対象の言語を正確に反映していることを代表性と言う。『現代日本語書き言葉均衡コーパス』の「図書館サブコーパス」（以下 BCCWJ の図書館 SC のように表記する）は、都内公立図書館の蔵書を現実母集団とし、そこからデータを無作為抽出して製作されたコーパスであり、高い代表性を持つと考えられている。しかし田野村 (2014) など一部の研究を除けば、その代表性を検討した研究は少ない。

あるコーパスがどれほどの代表性を持つかを実証することは難しい。図書館 SC の場合、現実母集団の蔵書約 33.5 万冊の全文コーパスを作り、それと比較すれば実証できるわけだが、全文コーパスを作るのが現実的に難しいからこそサンプリングコーパスを作っているという関係になっている。このため代表性の検証は、コーパスの設計方針を検討したり、他のコーパスによる検索結果の比較を行うなどの傍証を積み重ねていくしかないと考えられる。ここでは主に設計方針の検討と語彙表の観察から図書館 SC の代表性を検証する。

以下、第 2 節では図書館 SC の設計方針を検討する。第 3 節では語彙表を概観する。第 4 節では固定長の単語の頻度が可変長で何倍になっているのかを中心に調査する。最後に第 5 節でまとめを述べる。

２．設計方針の検討

コーパスの設計で特に重要な点は、どのような方法でサンプルを抽出するかという点と、サンプルの数×サンプルの長さ＝コーパスのサイズをどれぐらいの大きさにするかという 2 点だと思われる。ここでは主にサンプルサイズの問題に絞って検討する。

図書館 SC の設計方針を検討するには、類似の方針で製作されたコーパスの設計方針と比較すると、その特徴が明確になる。このため、世界的に代表性が高いと評価されている British National Corpus（以下 BNC と言う）の設計方針を簡単に確認しておく（Burnard (ed.)，2007；

† hideaki@moriharuo.com

アシュトン、バーナード，2004）。

BNC は 1995 年にイギリスで製作されたコーパスで、総語数は約 1 億語である。そのうち書籍データは 1411 冊×平均 3.6 万語＝約 5 千万語となっている。書籍はテキストタイプを情報伝達散文（8 種類）、文芸作品、未分類の計 10 種類に独自に分類し、ベストセラーの一覧リストや図書館の貸し出し冊数を参考に選抜した。さらにそれぞれの書籍から 4 万語を目安にサンプルを取得し、4 万語に満たない書籍は全文を、4 万語以上の書籍は最大 4.5 万語を採用した。この結果、サンプル当たりの語数は平均で約 3.6 万語となっている。このような方法は世界で初めて製作された Brown コーパス（500 冊×2,000 語＝100 万語）などと類似の方法である。

次に BCCWJ の図書館 SC のサンプリング方法を概観する（国立国語研究所, 2011 ; 丸山、柏野，2014）。図書館 SC は、書き言葉の流通の実態に着目し、東京都内の公立図書館で重複所蔵されていた 1986 年～2005 年発行の書籍約 33.5 万冊分、およそ 479 億字を母集団とした。サンプルの選択に当たっては全書籍のページをランダムに並べた長大なリストを作り、これを 20 年間の出版年と日本十進分類法の 11 分類の組み合わせによって 220 層に区分した。そしてそれぞれの層から復元無作為抽出法によって 10,551 箇所を選択した。この箇所に該当した書籍からさらに無作為に場所を選んでサンプルを抽出した。

抽出に当たっては、それぞれのサンプルから記号等を除いた文字数で 1 千字に固定した固定長と、それぞれのサンプルにおける節や章などの文章のまとまりに留意し、最大 1 万字まで抽出した可変長という二種類のデータを抽出した。田野村（2014，p. 112）の表 6.3 によれば、記号等を含めた文字数の固定長平均は 1,170 字、可変長平均は 5,039 字で、可変長の文字数は固定長の約 4.3 倍になっている。語数に直してコーパスサイズを計算すると、固定長は平均 635 語×10,551 サンプル＝約 670 万語、可変長は平均 2,738 語×10,551 サンプル＝約 2,889 万語で、これも約 4.3 倍である。ただし、固定長と可変長は必ずしも重複していないため、この両者を足して重複を除いたデータが最大となる。それをここでは「両方データ」と呼ぶ。両方データのサイズは平均 2,879 語×10,551 サンプル＝約 3,038 万語である。図書館 SC の最大サイズは両方データの約 3 千万語だが、これはサンプルごとの文字数が異なるので均衡ではない。このため BCCWJ のマニュアルには、統計分析に適するのは固定長データであると記されている（国立国語研究所，2011，p. 23）。

図書館 SC は、最大サイズで言えば BNC 書籍データの 6 割あるが、統計分析に適するサイズは 13.4%しかなく、思いのほか小さなコーパスになっている。もし、固定長の文字数を可変長平均の 5 千字にしていたら、統計分析に適するデータで 3 千万語のコーパスが出来上がったはずである。仮に図書館書籍のみで 1 億語のコーパスを作るとしたら、1 サンプルから約 1 万語を抽出すればよい。これならもっと簡単に 1 億語のコーパスが作れたと思われる。様々な選択肢が考えられた中で、なぜ BCCWJ では統計分析に適するとされる固定長の長さを、約 1 千字と言うごく短い長さにしたのであろうか。これを確認するため、BCCWJ の報告書類を閲覧したが、その根拠を実証的に記述した報告は探し当てることができなかった。その代わり、その意図がくみ取れる下記のような文章が散見された。

> BCCWJ は日本語に関する初の均衡コーパスであるが、その設計にあたっては、先行する諸外国の均衡コーパスを参考にしており、いくつかの点で先行コーパスに優った設計がなされている。例えば、厳密な無作為抽出を可能なかぎり実施していること（第 3 章参照）、平均サンプル長を British National Corpus などに比べる

> と短めに抑えることによって文献による語彙の偏りを低減していることなどである。（国立国語研究所，2011，p. 1）

> より大きい範囲を抽出単位として採用すると，抽出したサンプルの中身が文脈による偏りの影響を大きく受ける可能性が出てくる．たとえば，1 冊の書籍をまるごと抽出単位にすると，サンプリング作業の負担は減るものの，たまたまその書籍に頻出していた語が大量に収録され，語彙頻度表の順位に影響する可能性がある．これでは，BCCWJ が備えるべき代表性という点に問題が生じることになる．（丸山、柏野，2014，p. 26）

これらの記述からすると、固定長の長さを短くしたのは、特定の書籍による語彙の偏りを低減させるためであったことが分かる。しかしこれとは逆に BNC のガイドブックには、語彙の偏りを解消するためにサンプルを長くしたと受け取れる次の記述が見られる。

> Sinclair（1991: 24）は、Brown コーパスと LOB コーパスについて，「この 2 つのコーパスは広い範囲のテキストに出現する比較的頻度の高い単語についてのみ信頼性の高い情報を与えてくれる」と述べています。特定のテキストタイプだけに出現するような単語については，「サンプルが短すぎるのでサンプルのバランスをとるのに必要なサブカテゴリー自体が合理的なサンプルとはなり得ていない」との理由から，「信頼性はそれほど高くない」という評価を下しています。コーパスの規模を大きくし，それぞれのサブカテゴリーにさらに大きなサンプルを収集することで，この問題はいくぶん解決できるでしょう。（アシュトン、バーナード，2004，p. 30）

また、丸山、柏野（2014）が指摘する 1 冊の書籍を丸ごと収録した場合の弊害については、Sinclair（1991）に次の記述が見える。

> The penalties to pay for including whole documents are that in the early stages of gathering, the coverage will not be as good as a collection of small samples and the peculiarities of an individual style or topic may occasionally show through into the generalities. As against these short-term difficulties, there is a positive gain in the study of collocation, which requires very large corpora to secure sufficient evidence for statistical treatment.（Sinclair，1991，p. 19）
> 丸ごとの書籍を収録する弊害は、収集の初期に現れる。この段階のカバー範囲は、小さなサンプルを集積したコーパスと同じぐらい良くないため、一般性より個別のスタイルやトピックによる特殊性がしばしば見られる。このような初期の困難を越えるに従って、コロケーションの研究では、巨大なコーパスでなければ得られないほどの統計的に安定した十分な証拠が得られる。（発表者意訳）

Sinclair は、全文採用のデータを経時的に次々と収集していくモニターコーパスの提唱者である。上記の引用で「収集の初期」のような表現があるのは、モニターコーパスが念頭にあるからだ。しかし、これは時期の問題と言うより収集量の問題と捉えることができる。

モニターコーパスの代表例には Sinclair が監修した Bank of English があるが、これも高い代表性を評価されているコーパスであり、丸山、柏野（2014）が指摘するようなサンプルの全文採用による語彙の偏りは報告されていない。

さらに、コーパスサイズと代表性については、次のような指摘もある。

> LOB Corpus による頻度一覧表によって，コーパスに基づく語彙調査の難題の 1 つも明確になってくる。具体的には，単語の意味と用法を研究するのに，非常に巨大なコーパスが必要になるという点である。つまり，100 万語のコーパスでは，多くの単語に対して，意味のある一般化を行うのに十分なデータを提供できない。頻度数と言うのは，コーパスの非常に頻度の高い単語には比較的信頼性があるが，単語の意味や連語パターンを分析するためには，生起回数が非常に多いものでなければならない。〔・・・・・〕さらに，小さなコーパスの場合，頻度がただ単に中程度の単語を含むか，それとも頻度がまれな単語を含むかどうかは，コーパス内の各テキストに描かれるトピックの違いに大きく左右される。〔・・・・・〕しかしながら，さまざまな多くのテキストを含む非常に大きなコーパスであれば，より広範なトピックが描かれているはずであり，その結果，単語の頻度が個々のテキストによって受ける影響は少なくなる。（バイバー、コンラッド、レッペン，2003，p. 36）

以上の引用からすると、丸山、柏野（2014）が指摘するサンプルを長くすることによる弊害は、確かに収集の規模が小さい場合は懸念されるが、コーパスのサイズを大きくすればその問題は解消し、より高い代表性が得られるとする考え方が存在することになる。図書館 SC の固定長データは、10,280 冊の書籍から 10,551 サンプルを取得しており、トピックの多様性は十分であるように思われるが、サンプル長が平均 635 語とごく短いため、サイズが小さいコーパスになっている。このことによって代表性が十分に高まっていない可能性も考えられる。

３．図書館 SC 語彙表の概観

コーパスのサイズが小さいことで、図書館 SC にはどんな問題が生じるのだろうか。これを確認するため、ここでは「主要コーパス語彙表」と「短単位語彙表データ」を概観する[1]。これらの語彙表はそれぞれに特色が異なる。「主要コーパス語彙表」では語彙の中から機能語が除かれているが、ある単語がいくつのサンプルに出現したかというサンプル頻度が記載されている。ただし可変長や両方データの頻度は載っていない。「短単位語彙表データ」は機能語の頻度と可変長の頻度が記載されているが、サンプル頻度や両方データの単語頻度は載っていない。サンプル頻度は単語の頻度とは質の異なる情報、例えばどれぐらい多くのサンプルに共通して使用されるかで単語の一般性を見るといった情報が得られるため、ここでは両者を併用するが、両者では収録語の対象や語数が異なり、各単語の頻度にも一部に違いが見られるため、以後の分析では必ずしもデータ数が一致しない。

表 2 は、「主要コーパス語彙表」所収の 86,002 語について、単語頻度別に単語数を数えた表、表 3 はサンプル頻度別に単語数を数えた表である。表 2 の単語頻度では、頻度 1 が

[1] これらの語彙表は http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html（国立国語研究所の HP）からダウンロードできる。

25.8%、頻度 2〜5 が 32.0%で、頻度 5 以下で 57.8%になっている。コーパスのサイズが小さいため、頻度が低い単語が大量にある。表 3 のサンプル頻度では、頻度 1 が 36.4%、頻度 2〜5 が 30.3%で、頻度 5 以下で 66.7%である。表 4 は、「短単位語彙表データ」で固定長と可変長が重複する単語 83,232 語について可変長の単語数を数えた表である。このデータには機能語が 166 語加わっているが､固定長と重複した単語で数えると総語数が少なくなる。表 4 を見ると頻度 1 が 7.1%、頻度 2〜5 が 19.3%で、頻度 5 以下で 26.4%、頻度 20 以下で 55.0%となっている（サンプル頻度はデータがないため不明である）。可変長は固定長の 4.3 倍のサイズがあるため、高頻度語の割合が高くなっている。

表 2 固定長の単語頻度

単語頻度	単語数	
1	22201	25.8%
2〜5	27523	32.0%
6〜10	11562	13.4%
11〜20	8996	10.5%
21〜50	7683	8.9%
51〜100	3355	3.9%
101以上	4682	5.4%
合計	86002	100.0%

表 3 固定長のサンプル頻度

サンプル頻度	単語数	
1	31295	36.4%
2〜5	26032	30.3%
6〜10	9427	11.0%
11〜20	7008	8.1%
21〜50	6093	7.1%
51〜100	2601	3.0%
101以上	3546	4.1%
合計	86002	100.0%

表 4 可変長の単語頻度

単語頻度	単語数	
1	5938	7.1%
2〜5	16027	19.3%
6〜10	11317	13.6%
11〜20	12484	15.0%
21〜50	15100	18.1%
51〜100	8357	10.0%
101以上	14009	16.8%
合計	83232	100.0%

これらの表を見ると、コーパスのサイズが小さいことによる最大の問題は、その代表性を云々する以前に、あまりにも頻度の少ない単語が多いことであるのが分かる。国立国語研究所（2011，p. 23）は、統計分析に適するのは固定長であるとしているが、統計分析にはデータの質だけでなくデータの量も重要である。固定長では頻度 5 以下の単語が 6 割弱あり、これらを使用して統計的に有意な分析を行うのは困難だと思われる。それならむしろ文字数のばらつきを考慮に入れながら可変長の単語頻度を使用したり、文字数のばらつきには比較的影響されにくいサンプル頻度を指標にすることを考えてみても良いだろう。分析の対象や方法によっては、可変長（正確には最もサンプル長が長い両方データ）の方が、統計分析に適していることも考えられる。「単語の意味や連語パターンを分析するためには，生起回数が非常に多いものでなければならない。」（バイバー、コンラッド、レッペン，2003，p36）という指摘は、重く受け止める必要があるだろう。

４．固定長頻度と可変長頻度の比較

図書館 SC の固定長データは、サンプル長が短くコーパスサイズが小さいため代表性が十分に高まっていない可能性が考えられる。これを検証するには、どうすれば良いだろうか。大規模な調査が可能なら、固定長データを 100 字ごとに区切ったデータを作り、コーパス文字数の増加に対する全単語の頻度増加率を観察するのが良いと思われる。文字数の増加に対して頻度が一定に増加しているなら代表性は高く、増加率が不安定なら代表性は高くないと考えられる。代表性の高いコーパスとは、どんどんサンプル長やサンプル数を増大させた結果、データ量の増加に対して頻度の増え方が正比例するようになったコーパスのことである。そのような状態に達したコーパスなら、もうそれ以上サンプル長やサンプル数を増やす必要はない。そのコーパスで得られた頻度に一定数をかければ母集団の正確な頻度が推定できる。それに対し字数が増加するたびに頻度の増加率が変わるなら、まだ母

集団を推定する準備が整っていないと言える。これは代表性が低いコーパスである。代表性とは、コーパスが母集団の正確な縮尺になっていることである。しかし、ある単語で例えば固定長の 800 字→900 字段階と 900 字→1 千字段階を比較してまだ増加率に揺れがあるなら、正確な縮尺になり切っていない可能性が高いと考えられる。

ただし、このような検証は相当に大規模な研究になる。これをもっと簡便に行うには、固定長データと可変長データの比較が考えられる。しかし、可変長は個々のサンプルごとに文字数が異なるため、統計分析には適さないとされている。例えば A という単語の頻度を可変長で調べた場合、固定長頻度の 4.3 倍になっていれば正確で、0.1 倍とか 10 倍になっていれば不正確だとは言えないとする考え方もあるだろう。A という単語が短い可変長データにのみ出現する単語であれば 0.1 倍になることもあるし、長い可変長データにのみ出現する単語であれば 10 倍になることもあり得るからである。しかし、現実的には個別の単語が可変長のサンプルの長さに連動した出現傾向を持っているとは考えにくい。機能語のような高頻度語なら、短いサンプルでも長いサンプルでも、その出現傾向はほぼ同じだと思われる。中・低頻度語の場合も、どの単語が短いサンプルに出現し、どの単語が長いサンプルに出現するかは、十分ランダムになっていると考えられる。このため固定長と可変長の比較は、厳密な正確性には欠けるかも知れないが、図書館 SC に出現する語彙の全体像を簡便に観察するための調査としては、ある程度妥当なものだと考えられる。そこでここでは、固定長と可変長の頻度を比較し、その増加率がどれほど安定しているかを調査する。データには「短単位語彙表データ」を使用する。

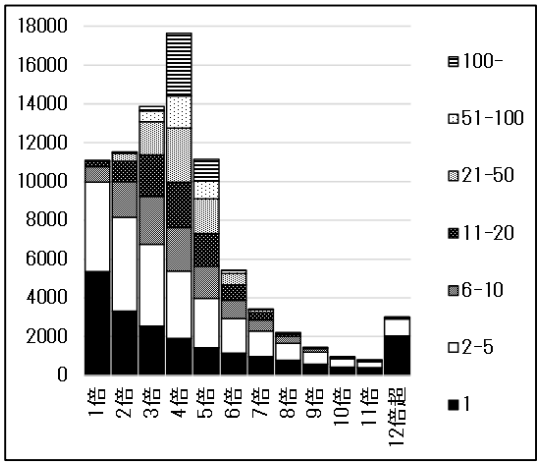


図 2 頻度別・可変長倍率ごとの単語数

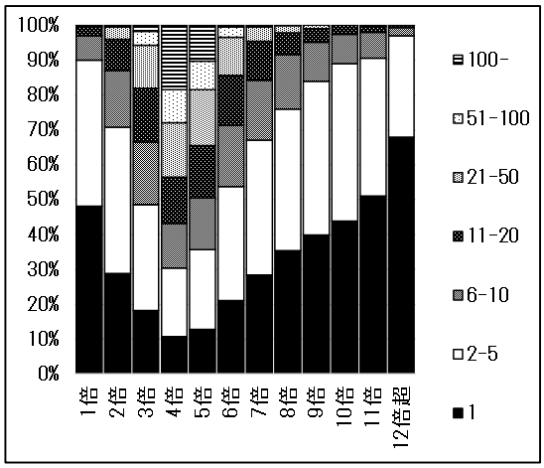


図 3 可変長倍率ごとの単語の頻度割合

図 2 は、表 2～4 の頻度区分ごとに分けた固定長の単語の数を、可変長の頻度倍率ごとに積み上げたグラフである（1 倍は 0.51 倍～1.50 倍の範囲）。この頻度倍率÷4.3 が増加率である。図 2 を見ると固定長の頻度は可変長で 4 倍になっているものが最も多い。つまりデータ量にほぼ正比例して増加している単語が最も多いということが分かる。

図 3 は、図 2 を割合で表したグラフである。高頻度の単語は 4 倍と 5 倍に多く、ここから倍率が離れるに従って低頻度の単語の割合が多くなる。頻度 100 以上の高頻度語は、4 倍が 69.8%、5 倍が 24.5%で、この二つで 94.3%になる。このことから高頻度語の頻度はデータ量の増加にほぼ正比例して増加することが分かる。その一方で低頻度語は、様々な倍率になる。この現象は、低頻度語の不安定さを示すものであり、固定長における低頻度語の

頻度が必ずしも正確だとは言い切れないことを示唆している。現在の固定長データでは頻度 1〜5 になっている単語でも、サンプリングをやり直した別バージョンの固定長データなら、頻度が 1〜15 などのように変わる可能性も考えられる。

この議論を、図 4、5 の箱ひげ図[2]を使用して整理して見よう。図 5 は図 4 の拡大図、表 5 はこれらの記述統計量である。図 4 の横軸は基本的に表 2〜4 の頻度区分と同じもので、1 は 1、2-5 は 5、6-10 は 10 のように区分の最大値で表記している。表 2 と異なり、図 4 では 101-1,000 と、1,001 以上も分けて描いた。10,000 というラベルは、固定長の頻度が 1,001 を超える超高頻度語につけている。

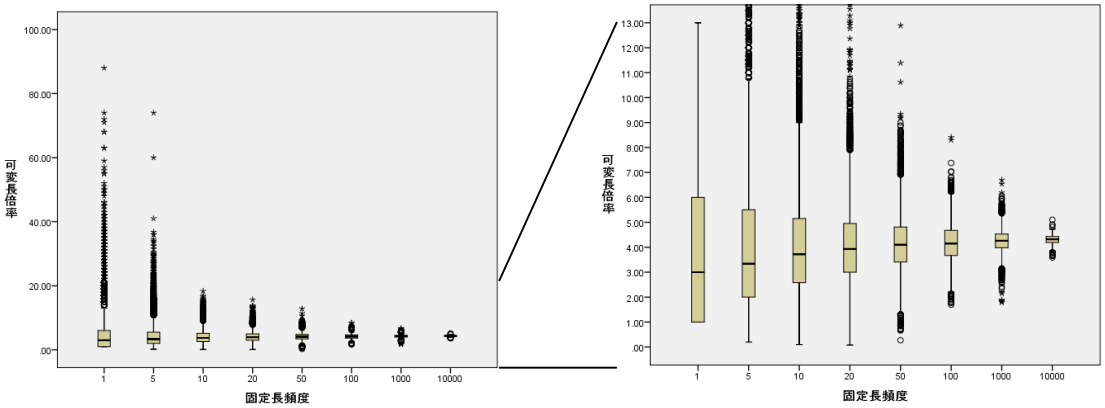


図 4 固定長頻度別可変長倍率分布（全体）　　図 5 固定長頻度別可変長倍率分布（拡大）

表 5　固定長頻度区分別における可変長倍率の記述統計量

	度数	平均	標準偏差	標準誤差	平均値の 95% 信頼区間 下限	平均値の 95% 信頼区間 上限	最小	最大
1	20826	5.0152	5.52557	.03829	4.9402	5.0903	1.00	88.00
5	26576	4.1695	3.32504	.02040	4.1295	4.2095	.20	74.00
10	11371	4.0486	2.09328	.01963	4.0102	4.0871	.10	18.34
20	8925	4.0759	1.61271	.01707	4.0424	4.1093	.08	15.64
50	7510	4.1477	1.12341	.01296	4.1223	4.1731	.27	12.89
100	3358	4.1883	.78040	.01347	4.1619	4.2147	1.70	8.41
1000	4130	4.2502	.47228	.00735	4.2357	4.2646	1.79	6.70
10000	536	4.2961	.20722	.00895	4.2785	4.3137	3.59	5.10
合計	83232	4.3582	3.51331	.01218	4.3343	4.3820	.08	88.00

表 5 で 10,000 の度数を確認するとわずか 536 しかない。これを品詞ごとに高頻度順に示せば、助詞「の・に・て」、動詞「する・いる・ある」、固有名詞「日本・アメリカ・東京」などになる。頻度 1,001 付近の単語は「働く・進む・内容・基本」などである。図 5 を見ると、10,000 の箱ひげ図は、他の箱ひげ図と比べて極めて小さいことが分かる。これはこの群に属する 536 語が可変長のデータでほとんどばらつくことなく、4.3 倍付近に集中していることを表している。表 5 で確認すると平均は 4.296、標準偏差は 0.207 である。具体的な単語で見ると助詞の「の」は固定長頻度の 342,113 が可変長では 1,473,404 と 4.31 倍に、固有名詞の「日本」が 8,846 から 37,131 と 4.20 倍に、動詞の「働く」が 1,001 から 4,397 と

[2] 箱ひげ図は、真ん中の黒い線が中央値、箱の上下が 75 パーセンタイルと 25 パーセンタイル、ひげの上下が 90 パーセンタイルと 10 パーセンタイルの位置を表す。ひげの外の○や☆は外れ値である。

4.39 倍になっている。これらの高頻度語が可変長ではそのデータ倍率とほぼ同じ 4.3 倍になっているのは、これらの頻度が極めて高く、高い代表性を持っているからだと考えられる。図書館書籍の母集団の文字数はおよそ 479 億字であるから、これらの固定長頻度を 4,790 倍にすればほぼ母集団の頻度と同じになると考えて良いだろう。

その一方で 1 の箱ひげ図は、90 パーセンタイルが可変長倍率 13 倍となるなどばらつきが大きい。図 4 を確認すると最大で 88 倍になっている。固定長で頻度 1 の単語が、可変長になると頻度 1 から頻度 88 にまでばらついて増加していることが分かる。これらの頻度を 4,790 倍にしたからと言って、母集団の正確な頻度が推定できるとは思われない。つまり、代表性は高くないと考えられる。なお、図 5 の箱ひげ図で、低頻度になるほど中央値が 3 に近づく現象が観察される。これは低頻度になるほど増加率が低くなる単語が多いためである。固定長で頻度 1 の単語には、可変長になっても頻度が 1 のままである単語も多い。これらの多くは母集団でも頻度 1 のままであることが予想される。その意味では、低頻度語の中にも代表性が高い単語が含まれていることになる。

図書館 SC の低頻度語は、可変長における頻度倍率が大きくばらつくため、その多くの代表性は高くないと考えられる。それでは低頻度語はなぜこれほどまでばらつくのであろうか。次にこの問題を調査する。

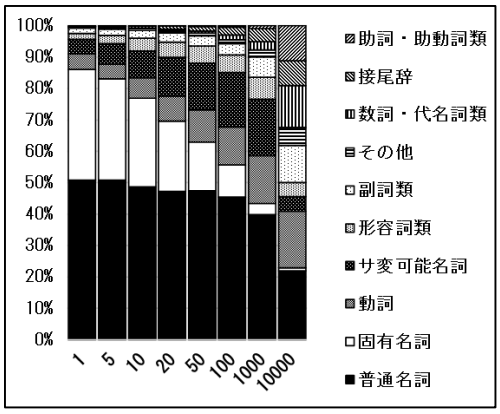


図 6 固定長頻度別品詞割合

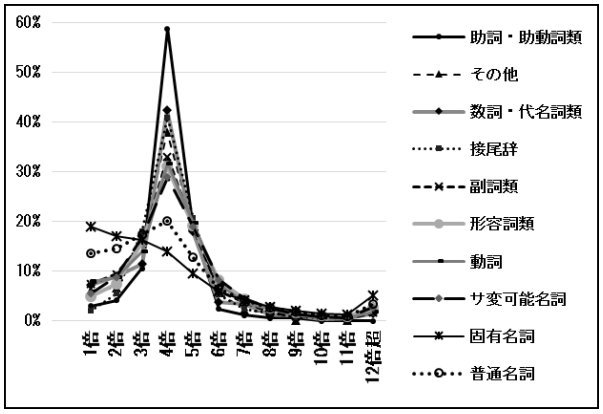


図 7 品詞ごとの可変長倍率分布

図 6 は、表 5 の頻度区分ごとに固定長データの品詞割合を示したグラフである。これを見ると低頻度語の大半は普通名詞と固有名詞であることが分かる。普通名詞は頻度区分が 1000 の場合でも 4 割程度を保つが、固有名詞は頻度区分が上がるにつれてその数を激減させる。この理由は、固有名詞の多くが特定のテキストにしか出現しない特定の単語であるためだと思われる。図 7 は、各品詞ごとに可変長で何倍になりやすいかを表したグラフである。最も高頻度語である助詞・助動詞類ではその 6 割が 4 倍、9 割以上が 3〜5 倍の範囲である。これに比べ、普通名詞と固有名詞はその多くが 1〜6 倍に散らばっている。グラフが見にくくて恐縮だが、固有名詞は 12 倍超の割合も 5%以上ある。

この二つのグラフから分かることは、固有名詞や普通名詞には低頻度の単語が多いこと、固有名詞や普通名詞は可変長になると様々な倍率で増加するということである。図 6 の普通名詞は大半の頻度区分で 5 割弱を維持するが、この普通名詞の内部でも一部のテキストでしか使われない特定の単語と多くのテキストで使われる一般的な単語の交替現象が起きていると考えられる。つまり低頻度語が大きくばらつく理由は、品詞の特性による影響、

すなわち特定のテキストに出現する特定の単語の出現パターンが原因である可能性が高い。

これを具体的な単語で観察してみよう。表 6 は「トマト」という普通名詞がどのサンプルに何個出現したかを数えた表である。固定長の頻度が多いものから順に 8 サンプルを表示している。固定長ではこの他に 66 サンプルに出現し、全体合計は 201 である。このうち上位 8 サンプルで 89 と全体の 44.2%に達するため、「トマト」の頻度ではこれら 8 サンプルの影響が強いことが分かる。書名を見ると料理関係や野菜作りのトピックが多く、「トマト」という単語は特定のトピックで多用される単語であることが確認できる。

問題は、このような単語がうまくサンプリングできているかどうかである。図 8 は、それぞれのサンプルのどの位置に「トマト」という単語が出現するのかを表している。縦軸は表 6 のNo.に対応し、整数の位置に固定長と可変長を含めた全体（両方データ）を、整数＋0.5 の位置に固定長の出現状態をプロットしている。両方データの表示にある×は、サンプルの末尾を表している。横軸は語数で、目盛りは記号等を含む固定長平均の 750 語で区切ってある。

表 6 サンプル別「トマト」の出現数

NO.	書名	固定長	可変長	倍率
8	ほんじょの虫干。	6	6	1
7	トマト弁護士被告人の甘い囁き	7	7	1
6	永田農法・驚異の野菜づくり	7	36	5.2
5	知っておきたいキッチンハーブ	10	21	2.1
4	ケンタロウの野菜がうまいッ！	10	28	2.8
3	シニアのためのライトフレンチ	14	10	0.8
2	わかりやすいイタリア料理	16	0	0
1	食べるのが大好き	19	21	1.2
	小計	89	129	1.5
	その他(固定長66冊、可変長160冊)	112	415	3.8
	合計	201	544	2.8

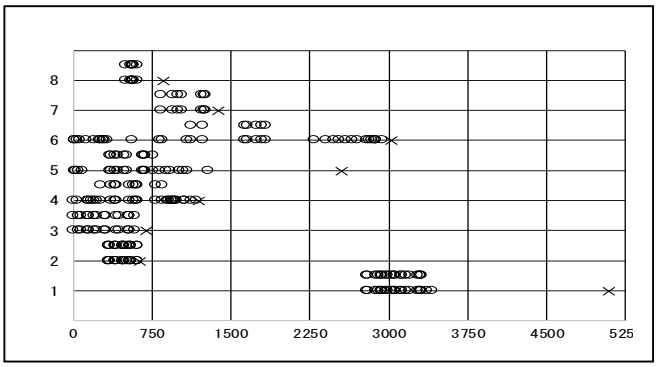


図 8 「トマト」の出現位置（上：固定長・下：全体）

No.1『食べるのが大好き』では両方データの語数は 5080 語で、そのうち真ん中から後半で集中的に「トマト」が出現する。両方データで見れば、「トマト」が出現しているのはサンプルの 1/7 に過ぎないが、固定長のサンプル長は短いため、全体に万遍なく出現していることが分かる。No.5『知っておきたいキッチンハーブ』でも、両方データでは後半には 1 語も出現しないが、固定長は前半の「トマト」が頻出する部分のみを抽出しているため、サンプル全体の平均的な頻度より多くなっている。同様の問題はNo.7『トマト弁護士被告人の甘い囁き』でも見られる。No.2、3、4、8 は両方データ自体が短いため、一見問題があるようには見えないが、サンプルを長くした場合、「トマト」と言う単語が残りの部分には全く出現しない可能性も否定できない。これらのサンプリング状況を見ると、固定長データから母集団の「トマト」の頻度を推定すれば、その頻度をかなり過大評価することになるのではないかと思われる。この理由は固定長の抽出範囲が短すぎて、テキスト全体における出現確率を正確に反映できていないためである。BCCWJ の設計方針はサンプルを無作為抽出することで各サンプルの標本誤差が均衡化されることを期待するものだが、そのような大数の法則は大量のデータでしか働かない。サンプル頻度が少ない場合は個々のサンプルが個々のテキストをある程度正確に反映している必要があると考えられる。

「トマト」は固定長のランクで 2689 位、可変長で 3862 位の高頻度語である。固有名詞や一部の普通名詞は特定のテキストに出現しやすいだけでなく、その出現の仕方も一か所に固まって出現しやすいなど特殊であるため、単語頻度 201、サンプル頻度 74 の高頻度語であっても、短いサンプル長で正確なサンプリングを行うのは困難なのだと思われる。

５．まとめ

『現代日本語書き言葉均衡コーパス』（BCCWJ）の中で、統計分析に適すると言われているのは固定長データである。しかしこれらのサイズは思いのほか小さい。一方、Sinclair（1991）、バイバー、コンラッド、レッペン（2003）などにより、サイズが小さいコーパスの代表性はさほど高くないことが主張されている。このため、本研究では図書館サブコーパスの設計方針の検討と語彙表の観察を行った。BCCWJ のマニュアル等では、語彙の偏りを防ぐためにサンプルを短くしたとの記述が見られる。そこで、サンプルを短くすれば本当に語彙の偏りが防げるのかどうかを検証するため、語彙表を使用して固定長と可変長の頻度を比較した。この結果、高頻度語はデータ量に正比例して頻度が増加するが、低頻度語は頻度がばらついて増加することが分かった。代表性が高ければ基本的にデータ量に正比例して頻度が増加するはずである。この頻度がばらつくということは、サンプル長が短い固定長の頻度が、母集団の正確な縮尺になっていないからだと考えられる。

また、低頻度語が特にばらつく理由は、固有名詞や特定のテキストに出現しやすい普通名詞が多く含まれるためだと考えられた。そこで「トマト」という普通名詞を例にサンプリング状況を観察した。「トマト」の場合、固定長では抽出範囲が短すぎ、テキスト全体における出現確率を十分に反映したサンプリングが行えていないと思われた。固有名詞や普通名詞ではこのようなサンプリングがしばしば生じていると考えられるため、高頻度語であっても一部の固有名詞や普通名詞の代表性は、それほど高くない可能性も考えられる。

ここで行った分析をさらに深める方法としては、可変長データと両方データの比較が考えられる。さらに新しい分析法としてサンプル頻度の利用も有望と思われる。現在の語彙表にはこれらのデータが不足しているため、語彙表のさらなる充実を望みたい。

文　献

Burnard, Lou（ed.）（2007）*Users' reference guide to the British National Corpus*. Oxford: Oxford University Computing Services.（http://www.natcorp.ox.ac.uk/docs/URG/を閲覧。2015.06.25）

ダグラス・バイバー、スーザン・コンラッド、ランディ・レッペン；齊藤俊雄、朝尾幸次郎、山崎俊次ほか共訳（2003）『コーパス言語学　―言語構造と用法の研究―』南雲堂.

ガイ・アシュトン、ルー・バーナード；北村裕（監訳）（2004）『The BNC Handbook　コーパス言語学への誘い』松柏社

国立国語研究所（2011）『『現代日本語書き言葉均衡コーパス』利用の手引き第 1.0 版』　国立国語研究所コーパス開発センター．http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html

丸山岳彦、柏野和佳子（2014）「サンプリング」田野村忠温（編）『講座日本語コーパス 6．コーパスと日本語学』朝倉書店，pp.21-44.

Sinclair，J. McH（1991）*Corpus, concordance, collocation*. Oxford: Oxford University Press.

田野村忠温（2014）「BCCWJ の資料的特性―コーパス理解の重要性―」田野村忠温（編）『講座日本語コーパス　6．コーパスと日本語学』朝倉書店，pp.119-151.

「通時音声コーパス」は可能か

丸山 岳彦　（国立国語研究所 言語資源研究系）†

Possibility of a Diachronic Corpus of Spoken Japanese

Takehiko Maruyama　(Dept. Corpus Studies, NINJAL)

要旨

通時コーパスとは、通常、書き言葉を対象としたものが想定される。では、話し言葉を対象とした通時コーパス、すなわち「通時音声コーパス」はどのように実現可能だろうか。本稿では、「通時」「音声」「コーパス」という 3 つの条件について検討した後、「通時音声コーパス」の実現によってどのようなことが明らかになるかについて、具体的な分析例を交えながら、その見通しを示す。

1　はじめに

2004 年に『日本語話し言葉コーパス』（CSJ）が完成・公開され、朗読音声ではない自然な発話（自発音声）の研究が飛躍的に進んだ。約 651 時間、752 万語の音声データを収録した CSJ は、音声学・音韻論・文法論などの音声言語研究に対して新しい研究データを提供しただけでなく、社会言語学におけるバリエーションの研究、音声認識・音声翻訳システムにおける音声処理・言語処理研究など、幅広い分野で利用されてきた。一方、CSJ に収録された音声の中心が独話（モノローグ）であったため、日常会話を収録した大規模コーパスの開発を求める声が以前から根強くある。これに対して現在、国立国語研究所で 2016 年度から始まる次期プロジェクトの一つとして、さまざまな場面における日常会話を大量に収録した会話コーパスの構築・公開が計画されている (小磯他, 2015)。CSJ に加えて、日常会話の音声コーパスが整備されれば、現代日本語の話し言葉を対象とする言語資源がさらに充実したものになり、話し言葉研究のさらなる拡大・深化が期待される。

上記のような現状を踏まえた上で、ここでは少し視点を変えて、「通時音声コーパス」の構築は可能か、という問題について考えてみたい。通常「通時コーパス」と言えば、書き言葉を対象としたものが想定されるだろう。これに対して、「話し言葉の通時コーパス」、すなわち、さまざまな音声資料を時代ごとに集積し、話し言葉の史的研究に利用できるようなコーパスの構築は、実現可能だろうか。それは、話し言葉の研究に何をもたらすだろうか。本稿では、「通時音声コーパス」が満たすべき条件とその問題点を示した上で、具体的な分析例を交えながら、その可能性について論じる。

2　「通時音声コーパス」の条件と制約

はじめに、「通時音声コーパス」を実現するために必要な条件について考えてみよう。ここでは、「通時」「音声」「コーパス」という 3 つに分けて、その条件を考える。

まず「通時」という点で言えば、「通時音声コーパス」は、複数の時代に録音された音声資料の時系列的な集積でなければならない。その縦断的な分析により、日本語の話し言葉の史的変遷を研究するために整えられた資料群である必要がある。次に、「音声」という観点からは、録音資料そのものが聴取できる状態になっていなければならない。音声コーパスの本質は音声データそのものであり、音声を文字化した転記テキストのみの集積は、真の意味での音声コーパスとは言えない。そしてその音声データは、可能な限り良好な音質で、特に会話の場合は話者ごとに別トラックの録音になっていることが望ましい。そして、「コーパス」であるからには、「用例を大量に偏りなく収集して電子

† maruyama @ ninjal.ac.jp

化し、検索用情報を付加したもの (前川, 2013)」というコーパスの定義を満たす必要がある。すなわち、さまざまなタイプの録音資料が大量にデジタル化され、その転記テキストや形態論情報、時間情報、話者情報、種々のメタデータなどがアノテーションされた状態であることが望ましい。

ところが実際には、当然のことながら、上記の条件を十全に満たすことは極めて困難である。例えば、「複数の時代に録音された音声資料」という点については、そもそも録音技術が開発されたのが19 世紀後半、一般に普及し始めるのが 20 世紀に入ってからという事情を考えれば、通時音声コーパスは、20 世紀以降に録音された音声のみに対象が限られるという制約がある [1]。書き言葉を対象とする通時コーパスが上代日本語（8 世紀）以降の言語資料を扱えるのに対して、通時音声コーパスは扱える範囲が極めて狭く限定されることになる。次に、「音声そのものが参照できる状態」という点については、古い時代になるほど良質な音声データが期待できないという制約がある。後にも述べるように、1950 年代に国立国語研究所が作成したさまざまな場面における会話の録音資料は、日常の会話場面をその場で実況録音したものであり、録音レベルの小ささやノイズの混入などによって聞き取りが困難な箇所も少なくない。また、デジタル化されていない音声資料は、原メディアの劣化などによって近い将来に聴取できなくなる可能性があるため、デジタル化の作業が急務となるが、そこには実務的・コスト的に大きな問題が生じることになる。さらに、「大量に偏りなく」という点については、これも当然のことながら、時代を遡るほど現存する音声資料の量は少なくなるため、そこに大規模コーパスとしての大量性や均衡性を期待することはできない。元来、話し言葉コーパスに均衡性を求めるのは非常に困難であるが (前川, 2013)、歴史的な音声資料を対象とする場合、その傾向はより顕著なものとなる。さらに、音源が残されていたとしても、著作権の問題または遺族の意向などによって、当該の音声資料を公開できない、という場合もある。

一口に「通時音声コーパス」と言っても、残されている音声資料が限られていることを考えると、その量の少なさや均衡性・多様性の偏りなどには目をつぶるしかない。言い換えれば、現存する音声資料をできるだけ幅広く収集し、それで賄うしかない。これは、例えば、上代日本語の研究者が限られた資料の中で研究せざるを得ない、という事情と同様である。古い言語資料を利用しようとする場合、このような資料の量的な制約は、必然的について回るものと言える。

そのような制約を踏まえた上で通時音声コーパスの実現を目指す場合、必要となるのは、できるだけ多様な音声資料の収集と、それを分類するメタデータの設計 (丸山, 2012) という 2 点だろう。このうち前者については、現存する音声資料をできるだけ掘り起こし、研究資料として地道に整備していくしかない。国立国会図書館がウェブ上で公開している「歴史的音源 [2] 」のうち、演説や講演などの音源資料はその有力な候補の一つになるだろう。一方、後者については、音声資料をどのような分類基準によって整理し、話し言葉の位相の中に位置づけていくか、という点を検討することが求められる。収録時期はもちろんのこと、独話と会話、発話場面、発話者（性、年齢、出身地）、聞き手との関係、スタイルの高低、自発性の度合いなど、CSJ で詳細に付与されたメタデータも参考にしながら、多様な音声資料を多角的に分類・分析するための指標の策定が必要となる。

3　「通時音声コーパス」が可能にする研究の事例

以下では、仮にある程度の規模を持った「通時音声コーパス」が実現されたと仮定し、それによってどのような言語研究が可能になるかという点について考えてみたい。ただし、20 世紀以降の話し言葉を通時的に俯瞰できる音声コーパスは現在のところ存在しないので、ここでは限られた音声資料に基づいて分析例を示し、そこから将来的な展望を述べることにする。

[1] 金澤 (2015) によれば、現存する最古の日本語録音資料は、1900 年に川上音二郎一座が欧米興行を行なった際に録音した「オッペケペー節」であるという。

[2] http://rekion.dl.ndl.go.jp/

3.1 分析対象データ

ここでは、分析対象データとして、CSJ に加えて、以下の 2 つの音声資料を用いる。

1. 「想隆社アカデミックリソースシリーズ 貴重音源コレクション 岡田コレクション I」
2. 『談話語の実態』録音資料

以下、前者を「岡田コレクション」、後者を「談話語データ」と略称する。

「岡田コレクション」とは、明治後期から昭和前期にかけて SP レコードに録音された音声資料のデジタル音源である。岡田則夫氏の収集した SP レコード 3.5 万枚のうち、165 音源、18.5 時間分の音声資料がデジタル化され、市販されている [3]。これらの音声データはすべて独話であり、演説、講演、講話、実況、法話、朗読などに分類される。音質の悪さにより、音声が不明瞭なところも散見されるが、約 100 年前の音声資料の集積として、貴重な研究データとなることは間違いない。

この音声データに対して、国立国語研究所共同研究プロジェクト「多角的アプローチによる現代日本語の動態の解明」（2009〜2015 年度、リーダー：相澤正夫）の中で、金澤裕之氏によって「岡田コレクション」の転記テキストが作成された。現在、その研究成果を収めた論文集の刊行が予定されている (相澤・金澤, 2015 予)。ここでは、このうち「演説」「講演」として分類された 109 講演、合計 14.5 時間分の音声データを用いる。異なり話者数は 76 人である。音声データの例を、表 1 に示す。

表 1: 「岡田コレクション」に収録された音声資料の例

発表年	講演者　(生年)	講演タイトル	収録時間
1915	尾崎 行雄 (1858)	司法大臣尾崎行雄君演説	0:28:10
1916	大隈 重信 (1838)	憲政ニ於ケル世論ノ勢力	0:17:14
1926	後藤 新平 (1857)	政治の倫理化	0:12:54
1931	犬養 毅　(1855)	強力内閣の必要	0:04:09
1937	林 銑十郎 (1876)	国民諸君ニ告グ	0:06:13
1941	近衛 文麿 (1891)	日独伊三国条約締結に際して	0:10:25

次に、「談話語データ」とは、国立国語研究所で 1950 年代から 1960 年代にかけて作成された録音資料である。国立国語研究所では、1948 年の設立当初から、話し言葉（東京方言および各地方言）の調査が進められていた。録音機を導入した調査の最も古いものは、記録を見る限り、1950 年 10 月に実施された福島県白河市での調査のようである。さらに、1952 年から「話しことば研究室」で開始された調査・研究では、さまざまな場面における日常談話が録音され、イントネーション、語・文節・文の長さ、文の構造、語の種類・使用度数・用法などが分析された。この研究成果は、1955 年の『談話語の実態』、1960・1963 年の『話しことばの文型 (1)(2)』という 3 冊の報告書にまとめられている (国立国語研究所, 1955, 1960, 1963)。このうち『談話語の実態』では、約 30 時間分の録音資料が作成され、うち約 10 時間分が分析に用いられた。

当時の録音資料は、現在、その大半がデジタル化されているものの、研究に用いるには未整理のままの状態になっている。筆者は以前、このうちの一部を抜き出し、音声を書き起こして新規に転記テキストを作成した。この「談話語データ」に含まれるのは、会話が 33 件で合計約 19.5 時間分、独話が 21 件で約 17 時間分である。ここに含まれる音声資料（略称）の例を、図 1 に挙げる。このうち会話は、街頭や室内でマイクで録音された一般人の雑談音声が主である。一方、独話はすべて当時国立国語研究所で行なわれた講義・講演などの音声である [4]。

[3] http://www.nichigai.co.jp/database/sp/index.html

[4] ただし、特に会話については、話し手の年齢、職業、出身地などを記録したフェイスシートや、正確な録音日、録音場所などの記録が残っていない（現時点では見つかっていない）ため、音声資料の詳細については不明なことが多い。

図 1: 「談話語データ」に収録された音声資料の例

会話： 「九段高校生」「三人の青年」「一研雑談」「鎌倉主婦」「ジイサン・バアサン」「魚屋小僧」

独話： 「助詞・助動詞」「国語講義」「日本語のアクセントほか」「新庁舎開き記念講演会」「国立国語研究所創立十周年記念講演会」

これら 2 つのデータの規模を、表 2 に示す。なお、「岡田コレクション」は収録時期によって「大正期」「昭和 1 ケタ」「昭和 10 年代」という 3 つに区分する。また「談話語データ」は独話と会話に分ける。総語数は、UniDic2.1.2+MeCab0.996 による形態素解析の結果から補助記号を除いた数で示す。

表 2: 「岡田コレクション」と「談話語データ」の規模

	岡田コレクション（1915〜1944 年）			談話語データ（1950〜1960 年代）	
	大正期	昭和 1 ケタ	昭和 10 年代	会話	独話
講演数	19	52	38	33	21
異なり話者数	16	42	30	（不明）	（不明）
収録時間	3 時間	6 時間	6 時間	19.5 時間	17 時間
総語数	23,022 語	46,998 語	49,070 語	218,497 語	182,619 語

以下では、「岡田コレクション」「談話語データ」「CSJ」という 3 種類の音声資料を用いて、いくつかの分析例を示しつつ、「通時音声コーパス」の可能性について述べる。

3.2　イントネーションの分析

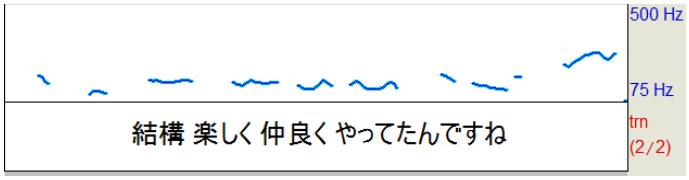
はじめに、イントネーションについて見てみることにする。ここでは、「談話語データ」に見られた、図 2 のようなイントネーションの型を取り上げよう。これは、1957 年に録音された「三人の女性」という音声資料の中に現れた、「そしてー みりんとね、卵の黄身**ね**、それ使ってね、すり鉢でする**のよ**」という発話のピッチ曲線である。図を見ると、「黄身ね」の「ね」、「するのよ」の「よ」の部分、すなわち一部の句末・発話末において、ピッチが急激に上昇していることが分かる。当然、この上昇調は聞き手に対する質問や疑問を表すものではない。

図 2: 句末の上昇イントネーション（談話語データ「三人の女性」）

図 2 の上昇調を聞いた筆者がすぐに想起したのは、古い邦画の中で女優が話している場面であった。例えば、小津安二郎監督『東京物語』の中で、原節子が発話している台詞の中に、このような急激な上昇調が多数観察される。1950 年代の若い女性は、このような発話が自然だったのだろうか。

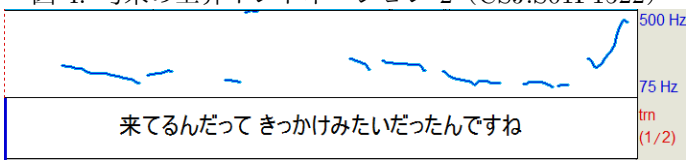
非疑問の文脈における句末の上昇調は、現代の話し言葉を収録した CSJ でも観察することができる。例えば、図 3 の発話「結構楽しく仲良くやってたんです**ね**」には、末尾に上昇調が認められる（句末境界音調は H%）。しかしながらこの上昇は、図 2 のような、急激な上昇調と同じとは認め難い。

図 3: 句末の上昇イントネーション（CSJ:S05F1600）



一方、同じく CSJ で観察された図 4 の発話「来てるんだって きっかけみたいだったんですね」の末尾に見られる上昇調は、図 2 の急激な上昇調に近いように思われる。

図 4: 句末の上昇イントネーション 2（CSJ:S01F1522）



ここで注意したいのは、図 3 の話者の生年は 1970 年代前半（収録時は 20 代後半）、図 4 の話者の生年は 1940 年代後半（収録時は 50 歳前後）で、両者の間に 25 年ほどの年齢の開きがあるという点である。考えてみると、図 2 や図 4 のような句末の急激な上昇調は、現代でも高齢の女性の発話で観察されることがある。客観的な裏付けはないが、印象として、上品な高齢の女性が少し気取って話すような場面で、図 2、4 のような上昇調が現れるように思われる。さらに、周囲の子どもに図 2 の音声を聞かせたところ、「おばあちゃんが話してるみたい」という印象が聞かれた。

図 2 の発話者である女性は、1950 年代の録音時に 20 歳代半ばだったとすれば、現在は 80 歳代になっている計算になる。ここから推測できるのは、彼女らは現在でも当時のイントネーションを（部分的に）保持しており、それを今でも使っている、ということである。それが現代の若い世代には「おばあちゃんみたい」あるいは「古い邦画の女優みたい」に聞こえることになる。若い世代の中で新しいイントネーションの型が出現し、古い型が使われなくなっていくという推移を考えれば、図 2 のような上昇調がどこかで衰退し、若年層が使わなくなった時期があるはずである。しかしその時期については、さらに多くの音声資料を準備し、縦断的・定量的に分析してみないと分からない。

3.3　文法形式の分析 (1)

次に、話し言葉の中に現れる文法形式に着目してみよう。ここでは、助動詞「まする」という例を取り上げる。「まする」は、近世初期に「ます」への移行が始まった形式と言われるが (服部, 2011)、「岡田コレクション」の中には、次のような「まする」の例が散見される。(1) は文末、(2) はト節、ガ節の述語句に「まする」が現れている。

(1) 明治 17 年、先帝陛下の御齢お五つの頃と記憶を致しておりまする。
（間部詮信「大行天皇御幼児を偲び奉りて」1927 年）

(2) 今日、新聞なぞを見まするとと、誠に嘆かわしいことがたくさんありまするが、一に良心を顧みないで悪魔の声にだまされて...　（牧野元次郎「良心運動の第一声」昭和 10 年代）

一方、「談話語データ」の中にも、次のような「まする」の例が観察された。それぞれ、カラ節、ト節、ケレドモ節の述語句に「まする」が現れている。

(3) 非常に予算の窮屈な、あー、時代でありまするから、えー、それでもって...
（山本有三「国立国語研究所十周年記念式典」1959 年）

(4) ラジオニュースの書き方というような本を見ますると、えー、ニュースには...
（波多野完治「新庁舎開き記念講演会」1962 年）

(5) 新しい字引きが二〇万語を収載すると書いてありまするけれども、その中の二万語しか...
（林大「新庁舎開き記念講演会」1962 年）

ただし、(1)〜(5) に挙げた話者の同じ講演中には、「感激致しておる次第であります。」（間部）、「論語のうちであったかと思いますが」（牧野）、「難しいんでありますから」（山本）、「聞いておりますと」（波多野）、「差もありますけれども」（林）という用例があることから、同一話者内において「ます」と「まする」の使用に揺れが生じていると言える [5]。

そこで、「岡田コレクション」と「談話語データ」の独話、CSJ の「コア」に含まれる独話（177 講演、約 41 時間分）を対象として、「まする」と「ます」の出現数を集計した。結果を表 3 に示す。

表 3: 各資料における「ます」「まする」の出現数

	岡田コレクション			談話語データ	CSJ
	大正期	昭和 1 ケタ	昭和 10 年代	独話	コア 独話
ます	271 (86.6%)	752 (89.8%)	903 (92.9%)	3,918 (98.8%)	5,604 (100%)
まする	42 (13.4%)	85 (10.2%)	69 (7.1%)	48 (1.2%)	0 (0%)

「まする」は大正期には 13.4%を占めていたが、時代が下るに従って「ます」に置き換わっていく様子が見て取れる。現代の CSJ では「まする」の用例は一つも見つからなかった。

次に、「岡田コレクション」「談話語データ」の 2 つを対象として、「まする」にどのような要素が後接しているかを集計した。上位 10 位までの結果を表 4 に示す。「助.」は助詞を表す。

表 4: 「まする」に後接する要素

岡田コレクション			談話語データ
大正期	昭和 1 ケタ	昭和 10 年代	独話
11 と（接続助.）	28 ば（接続助.）	16 句点（文末）	13 けれども
6 句点（文末）	11 が（接続助.）	13 が（接続助.）	8 から（接続助.）
5 が（接続助.）	10 名詞句	8 と（接続助.）	7 と（接続助.）
4 ゆえに	7 と（接続助.）	7 ば（接続助.）	6 し（接続助.）
3 ならば	5 に（格助.）	5 名詞句	6 名詞句
3 けれども	5 から（接続助.）	4 の（準体助詞）	6 ば（接続助.）
3 から（接続助.）	5 句点（文末）	4 から（接続助.）	4 が（接続助.）
2 の（準体助.）	4 けれども	4 か（終助詞）	2 に（助動詞）
2 に（格助詞）	3 や（終助.）	3 や（終助詞）	1 ために
1 ば（接続助.）	3 という（引用節）	3 に（格助詞）	1 ゆえに

「岡田コレクション」では文末に「まする」が現れる場合が見られるのに対して、「談話語データ」では文末位置の「まする」は皆無であった。これは、国会会議録に現れる「まする」を調査した上で

[5] なお、各話者の生年は、間部詮信が 1878 年（明治 11 年）、牧野元次郎が 1874 年（明治 7 年）、山本有三が 1887 年（明治 20 年）、波多野完治が 1905 年（明治 38 年）、林大が 1913 年（大正 2 年）である。

「主文末で用いられることがほとんどないという顕著な特徴がある」と述べた服部 (2011) の見方と符合する。一方、「岡田コレクション」では低い順位にある接続助詞ケレドモに後続する場合が、「談話語データ」ではトップとなっている。時代が下るにつれて文末で言い切る形が避けられ、並列節（ケレドモ節）でつなぐ形が好まれるように変化した、ということだろうか。

これらの観察から考えると、「まする」は、講演や講義、演説など、改まった発話スタイルの独話の中で、比較的少数の話し手により用いられていたことが予想される。ここで言う「少数の話し手」の条件には、おそらく、生年が強く影響しているだろう。すなわち、時代が進むにつれて若い世代の中で「まする」が消失していったということである。しかしながら、「まする」という形がいつごろ独話の中から消失していったのか、何年生まれの話し手までが「まする」を保持していたのか、という点については、データ量が不足しているため、現時点では分からない[6]。

3.4　文法形式の分析 (2)

さらに、終助詞の出現状況について見ておこう。ここで取り上げるのは、「談話語データ」の会話における、以下のような終助詞（の連接）の例である。

(6) a. 私だったら九州に行きたいわ。（「相模女子大生」）

b. 払うとしたら大変ですわね。（「友の会」）

c. あんたんとこのお魚、美味しいわよ。（「魚屋小僧」）

d. 先生とお話してきましたのよ。（「鎌倉主婦」）

これらはいずれも、女性の発話の末尾に現れた終助詞である。なお、この場合の「わ」「のよ」は、下降調ではなく上昇調（非疑問）である。現代の若い世代の女性の会話で、「わ」「わね」「わよ」「のよ」などの形が現れる場面は、非常に想定しにくいように思われる。もしくは、ふざけて「お嬢様」を演じているような場面（役割語としての使用）が想起される。

一方、先のイントネーションの場合と同様、高齢の女性が話者であると想定すると、これらはかなり自然に聞こえるように思われる。印象として、上品な高齢の女性が少し気取って話すような場面において、「〜するわ。」「〜したのよ。」と上昇調で話すのは、非常に自然に感じられる。

ここで、「談話語データ」と、CSJ に含まれる対話（58 対話、約 12 時間分）とを比較してみよう。である。発話末に現れる終助詞を抽出し、その一部を両者で比較した結果を表 5 に示す。

表 5: 対話の発話末に現れる終助詞

	〜わ	〜わね	〜わよ	〜のよ	〜よ	〜ね
談話語データ（会話）	153	296	116	296	1,675	5,752
CSJ （対話）	4	2	0	0	391	4,165

表 5 からは、「わ」「わね」「わよ」「のよ」の出現数が、CSJ よりも「談話語データ」の側に圧倒的に多いという事実を見て取ることができる。なお、ここでは両者ともイントネーションの型を考慮していないため、上昇調の用例は、特に CSJ でさらに少なくなるはずである。

このような発話末尾の終助詞がいつごろから若年層で用いられなくなったのか、という点についても、やはり、さらに多くの音声資料を検討してみなければ分からない。

[6] 服部 (2011) は、現代でも国会会議録の中で少数ながら「まする」の用例が観察されることを報告している。

4　まとめ

以上、本稿では、「岡田コレクション」「談話語データ」「CSJ」という 3 つの音声資料を用いて、イントネーションや文法形式に関する分析例を示してきた。実際の音声資料をもとに、話し言葉の通時的な変化を分析するという点では、いずれも興味深い事実を指摘することができたと思われる。

その一方で、多様な発話者・発話場面の違いを考慮しつつ話し言葉の変化の過程を通時的に分析しようとする観点に立つと、やはり、データの偏りや量の不足といった点が目立つ。つまり、分からないことが多い。この点を補完するためには、2 節で論じたような問題意識を持った上で、より多くの音声資料を掘り起こし、「通時音声コーパス」の整備を進めていく必要があるだろう。

2006 年、UCL（University College London）から “DCPSE”（Diachronic Corpus of Present-day Spoken English）が公開された。これは、1960 年代後半から 1990 年代前半までのイギリス英語の話し言葉を収録し、形態論情報・統語構造情報などがアノテーションされた通時音声コーパスである [7]。Aarts et al. (2015) は、助動詞 must, may, shall の使用が時代とともに大幅に減少したこと、would, could, should も減少したこと、一方で will が増加したことなどを、数量的に明らかにしている。2 節で論じたような「通時音声コーパス」の十全な仕様という点では疑問も残るが、実際に「通時音声コーパス」を作成し、話し言葉の動態を数量的に明らかにした、優れた実践例と言える。

現存する古い日本語音声の録音資料（蝋管レコードや SP レコード（落語、演説）など）については、かねてから清水康行や金澤裕之による詳細な調査・分析がある (清水, 1988, 1994, 2011; 金澤, 1991, 2000, 2015) 。今後は、現存する録音資料をより幅広く、時代縦断的に収集し、話し言葉の動態を捉えるための通時的研究に利用できるよう、「通時音声コーパス」として整備を進めていくことが重要になると思われる。

謝辞： 本研究は JSPS 科研費 24520523 の助成を受けたものです。

参考文献

Aarts, B., Bowie, J., & Wallis, S. (2015). Profiling the English verb phrase over time: modal patterns. In Taavitsainen, I., Kytö, M., Claridge, C., & Smith, J. (Eds.), *Developments in English: expanding electronic evidence*, pp. 48–76. Cambridge University Press.

相澤正夫・金澤裕之（編） (2015 予). 『（仮）戦前期 SP 盤レコードが拓く日本語研究』. 笠間書院.

服部匡 (2011). 「話者の出生年代と発話時期に基づく言語変化の研究—国会会議録を利用して—」. 『計量国語学』, **28** (2), 47–62.

金澤裕之 (1991). 「明治期大阪語資料としての落語速記本と SP レコード」. 『國語学』, **167**, 15–28.

金澤裕之 (2000). 「録音資料の歴史とその可能性」. 『日本語学』, **19** (11), 197–208.

金澤裕之 (2015). 「録音資料による近代語研究の今とこれから」. 『日本語の研究』, **11** (2), 133–140.

小磯花絵, 土屋智行, 渡部涼子, 横森大輔, 相澤正夫, 伝康晴 (2015). 「均衡会話コーパス設計のための一日の会話行動に関する調査—中間報告—」. 『第 7 回コーパス日本語学ワークショップ予稿集』, 27–34.

国立国語研究所 (1955). 『談話語の実態』. 国立国語研究所報告 8. 国立国語研究所.

国立国語研究所 (1960). 『話しことばの文型 (1) —対話資料による研究—』. 国立国語研究所報告 18. 秀英出版.

国立国語研究所 (1963). 『話しことばの文型 (2) —独話資料による研究—』. 国立国語研究所報告 23. 秀英出版.

前川喜久雄 (2013). 「コーパスの存在意義」. 前川喜久雄（編）, 『講座 日本語コーパス 1 コーパス入門』, pp. 1–31. 朝倉書店.

丸山岳彦 (2012). 「大規模コーパスの利用とメタデータの役割」. 『第 1 回 コーパス日本語学ワークショップ予稿集』, pp. 203–210. 国立国語研究所.

清水康行 (1988). 「東京語の録音資料」. 『国語と国文学』, **65** (11), 129–143.

清水康行 (1994). 「録音資料に聴く 20 世紀初めの東京語」. 『国学院大学日本文化研究所紀要』, **73**, 191–230.

清水康行 (2011). 「欧米の録音アーカイブズ: 初期日本語録音資料所蔵機関を中心に」. 『国文目白』, **50**, 29–19.

[7] 1960 年代から 1970 年代までの音声は London-Lund Corpus から、1990 年代の音声は ICE-GB から、それぞれ約 40 万語ずつが採録されている。http://www.ucl.ac.uk/english-usage/projects/dcpse/

ポスター発表（1） Aグループ

9月1日（火） 13:10～14:10

『現代日本語書き言葉均衡コーパス』に対する時間情報表現アノテーションの再修正作業

浅原 正幸 (国立国語研究所) *
坂口 智洋 (京都大学)
渡邊 友香 (統計数理研究所)

Correction of Temporal Information Annotation on 'Balanced Corpus of Contemporary Written Japanese'

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Tomohiro Sakaguchi (Kyoto University)
Yuka Watanabe (The Institute of Statistical Mathmatics)

要旨

小西ほか (2013) は，『現代日本語書き言葉均衡コーパス』(Maekawa et al. (2014)) に対してジャンル横断的に時間情報表現の正規化情報を TimeML (Pustejovsky et al. (2003)) に準ずる規定に基づき付与した。その後，同規定に基づく時間情報正規化プログラム (坂口 (2015a), 坂口・黒橋 (2015b)) が開発された。今回，アノテーションデータとプログラム出力の齟齬の対照比較を行うことにより，再修正作業を行った。さらに，基準の曖昧な点を見直し，新たな属性を導入したので報告する。

1. はじめに

情報抽出において，事象表現の生起時刻（実時間軸上の時区間）や時間的順序関係を推定するために時間情報解析が行われている。評価型国際会議 MUC-6 (the sixth in a series of Message Understanding Conference)(Grishman and Sundheim (1996)) で，アノテーション済み共有データセットが整備され，そのデータを基に各種の系列ラベリングに基づく時間表現の切り出し手法が開発されてきた。TERN (Time Expression Recognition and Normalization) (DARPA TIDES (2004)) では，時間情報の曖昧性解消・正規化がタスクとして追加され，様々な時間表現解析器が開発された。さらに，時間情報表現と事象表現とを関連づけるアノテーション基準 TimeML (Pustejovsky et al. (2003)) が検討され，TimeML に基づくタグつきコーパス TimeBank (Pustejovsky et al. (2003)) などが整備された。2007 年には，時間情報表現・事象表現間及び 2 事象表現間の時間的順序関係を推定する評価型ワークショップ SemEval-2007 のサブタスク TempEval (Verhagen et al. (2007)) が開かれ，種々の時間的順序関係推定器が開発された。後継のワークショップ SemEval-2010 のサブタスク TempEval-2 (Verhagen et al. (2010)) では，英語

* masayu-a@ninjal.ac.jp

だけでなく，イタリア語，スペイン語，中国語，韓国語を含めた 5 言語が対象となった。2013 年に開かれた SemEval-2013 のサブタスク TempEval-3 では，データを大規模化した英語，スペイン語が対象となっている。

一方，日本語においては IREX (Information Retrieval and Extraction Exercise) ワークショップ (IREX 実行委員会 (1999)) の固有表現抽出タスクの部分問題として時間情報表現抽出が定義されているのみで，時間情報の曖昧性解消・正規化に関するデータが構築されていなかった。2013 年に小西ほか (2013) は，『現代日本語書き言葉均衡コーパス』(BCCWJ)(Maekawa et al. (2014)) の一部のジャンル横断的に時間情報表現の正規化情報を TimeML に準ずる規定に基づき付与した。その後，時間情報表現と事象表現との時間的順序関係として，TimeBank の TLINK 相当の情報を被験者実験的に付与し (保田ほか (2013))，BCCWJ-TimeBank(Asahara et al. (2014)) として公開された。

小西ほか (2013) の作業開始時は，原始的な時間情報表現解析器を開発していたものの，解析器の出力を確認しながらアノテーションの修正を行う MATTER サイクル (Pustejovsky and Stubbs (2012)) を行うには至らず，作業者 1 名と教示者 1 名によりディスプレイを共有しながらペアプログラミング的にアノテーションを行う変則 MAMA サイクルを行うにとどまっていた。

その後，同規定に基づく時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) が開発された。第 2 著者である解析器開発者よりアノテーションの誤りが報告され，修正を実施した。解析器開発者より都合 3 回の誤り報告を受けながら，アノテーションデータと解析器出力の齟齬の対照比を行うことにより，ゆるい MATTER サイクルにより再修正作業を行った。再修正作業において，解析器構築や実応用の観点から，基準の修正を行った。基準の修正に際して，新たな属性を導入したので報告する。

2. アノテーション開発サイクルと修正作業

2.1 アノテーション開発サイクル

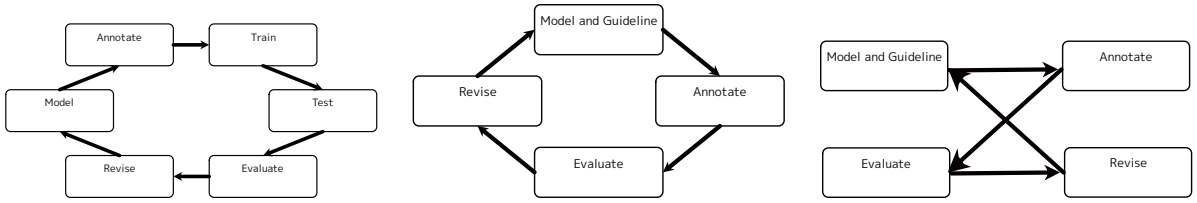
本節ではアノテーション開発サイクルについて述べる。アノテーション開発サイクルにおける各段階について Pustejovsky and Stubbs (2012) が次のように定義している(1)。

- Model: モデル化
  データを通した経験的な観察から生成される，理論的に説明可能な属性を与える構造的描写
- Annotate: アノテーション入力データの特定の構造的描写や性質をコード化する，特徴量集合を過程したアノテーション体系
- Train: 訓練
  対象となる特徴量集合がアノテーションされたコーパスを用いた解析器の訓練
- Test: 検証
  訓練データとは別に定義したテストデータ上での解析器の検証

(1) Pustejovsky and Stubbs (2012) pp.22-32 もしくは Pustejovsky (2006)

- Evaluate: 評価
  解析結果に対する標準化された評価
- Revise: 再検討
  アノテーションが機械学習アルゴリズム中で用いられる際により頑健で信頼できるものにするためにモデルとアノテーション基準を再検討する

これらの 6 つのステップを介したアノテーション開発サイクルを MATTER サイクルと呼ぶ。図 1 左に MATTER サイクルを示す。



MATTER サイクル　　MAMA サイクル　　変則 MAMA サイクル

図 1　アノテーション開発サイクル

しかしながら，アノテーション初期は適切な解析器が構成されることは少なく，MATTER サイクルの一部である MAMA サイクル (図 1 中) を用いる事が多い。「Evaluate: 評価」のステップにおいて，MATTER サイクルでは人手付与データ (GOLD) と解析器出力データ (SYS) との一致度をはかることが行われるが，MAMA サイクルではアノテーション作業者間の一致率 (Inter Annotator Agreement: IAA) をはかることが行われる。Passonneau and Carpenter (2014) はこの作業者間の一致率に対して確率モデルを導入することを提案している。

また，小西ほか (2013) は，BCCWJ-TimeBank の時間情報表現アノテーションにおいて，「Model(モデル化)」「Evaluate(評価)」を行う人と「Annotate(アノテーション)」「Revise(再検討)」を行う人と分業したうえで，同一の画面を見ながら作業を行うペアプログラミング的な方法を導入した。

保田ほか (2013) は，不定な時間的順序関係情報付与において，あらかじめ 3 人の作業者に基準を教示してから，その後相互に作業過程をフィードバックしないようにして，被験者実験のように作業者間でゆれる言語現象を明らかにするような方法を導入した。

本研究では，小西ほか (2013) のアノテーション作業が MAMA サイクルでとどまっていること，さらに同規定に基づく時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) が開発されたことから，解析器を用いた MATTER サイクルに基づく修正を行う。

尚，本稿では，上記に述べた MAMA サイクルや MATTER サイクルを介さない，人手によるパターン・規則のみに基づく解析器の出力をアノテーションとは呼ばない立場をとる。

2.2　解析器の概要

本節では，本研究で用いた時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) について簡単に示す。

解析器は時間情報表現を表す ⟨TIMEX3⟩ タグの 1. 認識， 2. TYPE 属性推定，3. VALUE 属

性推定の 3 段階からなる。時間情報表現の認識とは，⟨TIMEX3⟩ の開始タグと終了タグの挿入すべき位置を推定することである。TYPE 属性推定とは ⟨TIMEX3⟩ タグに付与された type 属性 4 種 { “DATE(日付表現)”, “TIME(時刻表現)”, “DURATION(時間表現)”, “SET(頻度集合表現)”} を推定することである。VALUE 属性推定とは時間情報表現が指し示す時刻・時間を正規化して機械可読形式に変換することである。⟨TIMEX3⟩ タグには，valueFromSurface 属性と呼ばれる表層文字列のみを用いて推定する正規化情報と，value 属性と呼ばれる前後文脈などの情報を用いて推定する正規化情報の 2 種類が定義されている。

坂口 (2015a) の時間情報正規化解析器は，「1. 認識」と「2.TYPE 属性推定」を系列ラベリング問題として同時に解き，「3.VALUE 属性推定」を正規化ルールに基づく valueFromSurface 属性推定と照応解析の手法を用いた value 属性推定の二段階の手法を用いて解いている。

「1. 認識」と「2.TYPE 属性推定」は，形態素解析器 JUMAN と係り受け解析器 KNP の出力から抽出した特徴量を用いた条件付確率場 (Lafferty et al. (2001)) による。特徴量として，見出し語・品詞・原形などの形態論情報，係り受け先の動詞，JUMAN 辞書に登録されている時相動詞か否か，記号か否か，数値の大きさを表すカテゴリ，200 近くの正規化ルールとの適合などを用いている。

「3.VALUE 属性推定」は最初に正規化ルールに基づく書き換え系により valueFromSurface 属性を復元する。valueFromSurface 属性が特定の時区間を示す定時間情報表現の場合には valueFromSurface 属性の情報がそのまま value 属性になる。曖昧性が残るような不定時間情報表現の場合に，曖昧性解消を行うことで value 属性を復元する。曖昧性解消は先に言及された定時間情報表現を参照し，復元する。参照すべき定時間表現の探索を，照応解析問題の一つである橋渡し参照問題として定式化し，SVM-Rank(Joachims (2003)) を用いたランキング学習を用いて解析する。

本研究の修正以前の解析器の性能は以下の通りである：

表 1　解析器の性能 (坂口・黒橋 (2015b))

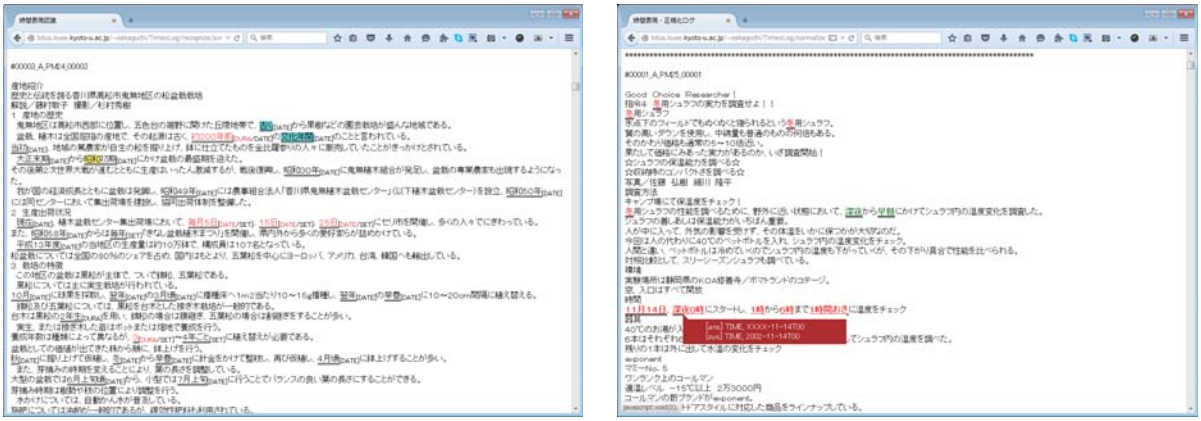
	Precision	Recall	F-value
「1. 認識」「2.TYPE 属性推定」	0.86	0.82	0.84
「3. VALUE 属性推定」	0.64	0.61	0.62

2.3　修正作業

修正作業は解析結果を図 2 のように可視化したものを，作業者に提示して行う。作業者は解析器の出力を見て，解析器が正しいか，既存のアノテーションが正しいかを確認しながら，oXygen XML Editor(図 3) 上で人手により修正を行う。

3.　基準の修正

3.1　括弧と時間情報表現範囲



認識結果　　正規化結果

図 2　修正作業に用いた解析結果の可視化

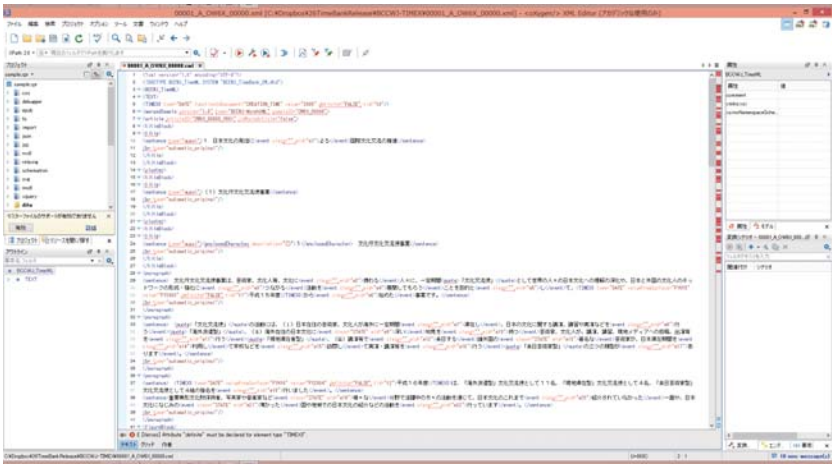


図 3　oXygen XML Editor

以前の基準 (浅原 (2012))[(2)]では「十四日午後一時半 (日本時間同七時半) 過ぎ」を「十四日午後一時半」と「 同七時半) 過ぎ」に分割して時間情報表現範囲を規定していたが，「日本時間」も時間表現の一部で value にも反映されていることから，まとめて「時間表現が連続する場合は括弧を含め，そうでない場合は括弧を含めない」という基準にする。

これによって西暦と和暦の併記の問題も同様に扱える。文書中に「1999(平成 11) 年」などの表現が出現した後は「2000(平成 12) 年 5 月」を「00(同 12) 年 5 月」などと省略することがあり，この場合，以前の基準では「00」と「(同 12) 年 5 月」に分割していたが，西暦と和暦を併記するときは両方が同じ年を指すため，まとめて「00(同 12) 年 5 月」にタグを付けることとする。

(2) 特に指定しない限り，「以前の基準」とは 2012 年 6 月 30 日現在の Version 0.10 のマニュアル相当のものとする。

3.2　type=DATE/TIME の value，valueFromSurface における DURATION の値の扱い

以前の基準では、type=DATE/TIME の value や valueFromSurface に DATE/TIME の形式の値　だけでなく DURATION の形式の値を取ることも許しており、type によって DURATION の形式を　異なる意味で用いているという問題、type=DATE/TIME の一部の時間情報表現では value や　valueFromSurface の値を DATE/TIME の形式と DURATION の形式の両方で書くことが可能なため　どちらで書くか明確に決まっていなかったという問題があった。

例えば、日付表現「３日目」「３日前」「３日後」や時刻表現「５時間前」「５時間後」は value に DURATION 形式を許している。

「３日目」
<TIMEX3 type="DATE" value="P3D">３日目</TIMEX3>
「３日前」
<TIMEX3 type="DATE" value="P3D">３日前</TIMEX3>
「５時間前」
<TIMEX3 type="TIME" value="PT5H">５時間前</TIMEX3>
「３日後」
<TIMEX3 type="DATE" value="P3D">３日後</TIMEX3>
「５時間後」
<TIMEX3 type="TIME" value="PT5H">５時間後</TIMEX3>
cf.)「３日間」
<TIMEX3 type="DURATION" value="P3D">３日間/TIMEX3>

上記の方針の場合，DURATION の値が DATE や TIME の valueFromSurface でも用いられ，DURATION の valueFromSurface と同じ記号で異なる意味を表すことになる。この混同を防ぐため，DATE や TIME の valueFromSurface 用に DURATION とは異なる，また，より表層的な情報を反映させることができる新たな記号 Q を定義する。

「３日目」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q3D">３日目</TIMEX3>
DATE のため，value の値は XXXX-XX-XX の形式となる。
「３日前」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q-3D">３日前</TIMEX3>
「前」という情報を残すために - (マイナス) を付与する。
「５時間前」
<TIMEX3 type="TIME" value="TXX" valueFromSurface="Q-T5H">５時間前</TIMEX3>
時間を表すために T を，また「前」という情報を残すために - を付与する。
「３日後」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q+3D">３日後</TIMEX3>
「後」という情報を残すために + (プラス) を付与する。
「５時間後」
<TIMEX3 type="TIME" value="TXX" value="PT5H">５時間後</TIMEX3>
時間を表すために T を，また「後」という情報を残すために + を付与する。

3.3 総称表現の識別

VALUE に含まれる X には 2 つの意味合いがある。1 つ目はある特定の時間を表すわけではない一般的な表現 (総称) で，2 つ目は特定の時間を表すがその文書での情報のみでは判別できない表現である。

例えば次の 2 つの例の「夏」は両方とも同じ VALUE となるが，X を用いる理由が異なると考えられる。

1. 特定の時間を表さない総称の例
   ・「京都の夏は暑い。」の「夏」が”XXXX-SU”
   (文書作成日時等に関わらず，XXXX となる)
2. 特定の時間を表すが，文書の情報だけでは判別できない例
   ・「その年の夏は暑かった。」の「夏」が”XXXX-SU”
   (もし「その年」が文脈から，例えば 2015 年と分かる場合は”2015-SU”となる)

これらの違いを機械で判断するのは難しい。解析器側で照応解析相当の処理を行っているが，その際に各時間情報表現が総称表現かあらかじめ情報として付与されていれば，参照すべきか否かを判別することができる。

そこで新たに属性 general を付与する。時間情報表現が総称表現である場合に general=TRUE を付与する。

```
<TIMEX3 type="DATE" value="XXXX-SU" valueFromSurface="XXXX-SU" general=TRUE>
夏</TIMEX3>
```

ここで総称表現として想定しているのは，具体的には次のようなものである。

- 「大津西小学校は，夏休み に家庭訪問を実施していた」
- 「でも，それだけだとなんなので，夏 野菜たっぷりサラダ」
- 「お問い合わせは 平日 10 時-18 時」
- 「五千万円を上限とする 年末 ローン残高の 1% を控除する」
- 「「レゴレゴ (0505)」と読めることなどから，5 月 5 日 は「レゴの日」」

4. 修正件数

表 2 に修正前件数と修正後件数を示す。約 6000 件のデータについて，修正がされなかった時間情報表現は 866 件であった。

表 2　修正前後の件数の変化

	DATE	TIME	DURATION	SET	ALL
修正前件数	4543	501	1211	151	6406
修正後件数	4755	513	1235	158	6661

表 3 に valueFromSurface に付与された Q+, Q- の件数を示す。

また，今回付与した general 属性が TRUE であった件数を表 4 に示す。

表 3　value 属性に付与された Q+, Q- の件数

	Q+	Q-	ALL
件数	13	42	55

表 4　総称表現の件数

	DATE	TIME	DURATION	SET	ALL
general=True	100	28	0	0	128

5. おわりに

本稿では，2015 年春に行った BCCWJ-TimeBank の時間情報表現アノテーションの再修正作業について報告した。

今後の課題として以下をあげる：

- TLINK 情報の再付与
  保田ほか (2013) は，時間的順序関係について 3 人の作業者に被験者実験的に行った。6688 個の関係については 3 人のアノテーションが一致したが，3011 個の関係が 2 人の一致にとどまり，545 個の関係は全く一致していない。これらの分布を評価するために，3 人の作業者で一致しなかった関係に対する情報付与を数千人規模で被験者実験的に行う。
- SLINK 情報のアノテーション
  SLINK は主節 (matrix clause)-従属節 (subordinate clause) 間の関係を規定するアノテーションである。英語以外のデータに対して付与されていることが少なく，多言語化における課題となっている。SLINK の関係ラベルは，’MODAL’, ’EVIDENTIAL’, ’NEG_EVIDENTIAL’, ’FACTIVE’, ’COUNTER_FACTIVE’, ’CONDITIONAL’ と事象の事実性に関するもので構成されている。英語においては FactBank (Saurí and Pustejovsky (2009))[(3)]として事実性解析用途に拡張している。これらのラベルを日本語に適合するために，鳥バンクの節分類を第 3 階層 (池原悟 (2007)) のレベルで付与を行い，その後，節分類の情報をもとに時制節性 (有田 (2007)) を付与することで SLINK 相当の情報を付与していきたいと考えている。

謝辞

国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

有田節子 (2007).『日本語条件文と時制節性』　くろしお出版.

(3) FactBank 1.0 https://catalog.ldc.upenn.edu/LDC2009T23

Asahara, Masayuki, Sachi Kato, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa (2014). “Bccwj-timebank temporal and event information annotation on japanese text.” *International Journal of Computational Linguistics and Chinese Language Processing*, 19:3, pp. 1–24.

DARPA TIDES (2004). *The TERN evaluation plan; time expression recognition and normalization*. Working papers, TERN Evaluation Workshop.

Grishman, R., and B. Sundheim (1996). “Message Understanding Conference-6: a brief history.” *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pp. 466–471.

池原悟 (2007).「意味類型パターン記述言語仕様書」　Technical report, 独立行政法人科学技術振興機構, 戦略的基礎研究事業, 高度メディア社会の生活情報技術.

IREX 実行委員会 (1999).『IREX ワークショップ予稿集』.

Joachims, T. (2003). “Optimizing search engines using clickthrough data.” *Proc. of the ACM Conference on Knowledge Discovery and Data Mining*.

小西光・浅原正幸・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパスに対する時間情報アノテーション」　自然言語処理, 20:2, pp. 201–222.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *Proc. of 18th International Conference on Machine Learning*, pp. 282–289.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

Passonneau, Rebecca J, and Bob Carpenter (2014). “The benefits of a model of annotation.” *Transactions of the Association for Computational Linguistics*, 2, pp. 311–326.

Pustejovsky, J. (2006). “Unifying linguistic annotations: A timeml case study.” *Proceedings of the Text, Speech, Dialogue Conference*.

Pustejovsky, J., and A. Stubbs (2012). *Natural Language Annotation*.: O’Reilly.

Pustejovsky, J., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz (2003). “TimeML: Robust Specification of Event and Temporal Expressions in Text.” *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, pp. 337–353.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, B. Sundheim, L. Ferro, M. Lazo, I. Mani, and D. Radev (2003). “The TIMEBANK Corpus.” *Proceedings of Corpus Linguistics 2003*, pp. 647–656.

坂口智洋 (2015a).「時間表現の解釈に基づく言明の抽出と整理」　修士論文, 京都大学大学院情報学研究科.

坂口智洋・黒橋禎夫 (2015b).「多様な時間表現の解釈に基づく言明の抽出と整理」　情報処理学会第 77 回全国大会, pp. 2–201–2–202.

Saurí, Roser, and James Pustejovsky (2009). “Factbank: A corpus annotated with event factuality.”

Language Resource and Evaluation, 43:3, pp. 227–269.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Kats, and J. Pustejovsky (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 75–80.

Verhagen, M., R. Saurí, T. Caselli, and J. Pustejovsky (2010). “SemEval-2010 Task 13: TempEval-2.” *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pp. 57–62.

保田祥・小西光・浅原正幸・今田水穂・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパスに対する時間情報・事象表現間の時間的順序関係アノテーション」　自然言語処理, 20:5, pp. 657–682.

浅原正幸 (2012).『BCCWJ-Timebank 日本語時間表現タグづけ基準 version 0.10』.

『児童・生徒作文コーパス』を用いた漢字使用能力の推定

宮城 信(富山大学人間発達科学部)†
今田 水穂(文部科学省初等中等教育局)

Estimation of the Ability to Use Kanji Using "A Written Composition Corpus of Japanese Elementary and Junior High School Students"

Shin Miyagi (University of Toyama)
Mizuho Imada (Ministry of Education, Culture, Sports, Science and Technology)

要　旨

本発表では、構築中の『児童・生徒作文コーパス』を用いて、児童・生徒の作文における漢字の使用実態と漢字使用能力の推定を試みる。このコーパスは小学校１年生から中学校３年生までの児童・生徒の作文を収集、電子化した 100 万語規模のコーパスで、児童・生徒の言語使用実態を縦断的に調査することができる。このコーパスを用いて、漢字種別(学年別配当漢字、常用漢字、常用外漢字)、品詞、語種などの観点から、児童・生徒の学齢別の漢字使用実態を調査する。また、『現代日本語書き言葉均衡コーパス』(BCCWJ)や、大学生の書いた作文と対照することにより、作文の文体的特徴や、大学生の漢字使用実態を到達目標とした児童・生徒の漢字使用能力の発達過程の分析を行う。

１．はじめに —本研究の目的—

これまでの諸研究では、それぞれアプローチは異なるが、学習過程で習得に適した語彙はどれかという観点から考察が進められている。河内(2015)や田中(2011)は、「国語政策・国語教育のよりどころとなるような重要語彙リストを作成する」(田中、p.86)という文言に見られるように、子ども達が優先的に学習すべき語彙の選定を念頭として、日常生活における重要語彙を検討している。また、鈴木(2011)では、中等教育課程で生徒達の語の使用を調査し、それらの重要性を検討している。

もちろん国語教育の現場において、学習に適した語彙の選定は重要である。一方で、漢字使用能力が大きく伸びていく小学校中学年から中学校にかけての発達過程の調査は、管見の限りほぼ無い。そこで本研究では、どのような語句を学習すべきかという視点ではなく、子ども達が作文する際にどのような漢字をどのように使用しているのか、すなわち児童・生徒の漢字使用能力に注目する。本研究で想定する漢字使用能力は以下のようなものである。

・ 児童・生徒が作文する際に、語の表記にどの程度漢字を使用するのか、または選択可能であるのか、という表現に関わる能力

この能力の推定のために、本発表では小学 1 年から中学 3 年までの児童・生徒の作文を収集、電子化したコーパスを用いて、次のような言語使用実態を調査する。

† miyagi@edu.u-toyama.ac.jp

・ 教育漢字や常用漢字の使用頻度は、学齢の進行に対してどのように変化していくのか。
・ 品詞別、語種別の漢字使用率は、学齢の進行に対してどのように変化していくのか。
・ 上位頻度語の漢字使用率は、学齢の進行に対してどのように変化していくのか。
・ 漢字使用の観点から見た作文の文体的特徴はどのようなものか。
・ 作文時に使用される漢字の種類と頻度は、どの時期にどの程度の水準で飽和するのか。

児童・生徒が自分の力だけで作成した文章は、彼らの漢字使用の実態を調べるために適した資料であるが、作文での漢字使用が、ただちに漢字使用能力を意味するわけではない。例えば、漢語は漢字書きが普通だが、和語は仮名書きでも違和感が少ないので、漢字を知っていても使わない、ということがあり得る。そこで、文字単位の漢字使用頻度だけではなく、品詞別、語種別、単語別の漢字使用率も併せて調査する。

また、作文における漢字使用の実態は、ただちに児童・生徒に求められる漢字能力を推定する資料とはならない。作文は書き言葉の多様な言語使用域(レジスタ)の一つに過ぎず、また児童・生徒の漢字使用能力の最終的な到達目標は学校教育の過程の先に位置する。そこで、『現代日本語書き言葉均衡コーパス』(BCCWJ)や大学生の作文と比較することにより、レジスタ横断的な観点から見た作文の文体的特徴や、学齢縦断的観点から見た漢字使用実態の飽和過程を分析する。

これらの調査により、児童・生徒の書く文章で要求される学齢別の標準的な漢字使用能力の範囲を推定する。また、それによって、現場での漢字学習や指導における重点化の判断や重要語の選出への示唆を与える。

２.『児童・生徒作文コーパス』の概要

２．１　調査の概要

国立大学附属小・中学校を調査協力校として、4 校(小学校 2 校、中学校 2 校)9 学年(小学 1 年～中学 3 年)の全児童・生徒に作文課題を課し(作成時間は小学校 40 分、中学校 45 分)、収集して電子化した。作文は「ゆめ」などのテーマ(タイトル)のみを提示し、教員は一切の事前指導を行わない。調査は、2014 年度に 2 回実施した。

第 1 回調査：「ゆめ」、2014 年 7 月実施
第 2 回調査：「ぼくの／わたしのがんばったこと」、同 12 月実施

平文テキストへの電子化は以下の指針に従って実施した。

〇電子化の指針

・できるだけ、正確に紙面を再現するよう心がける。
・段落初めの一字下げや空欄(意味不明なものも含めて)も正確に記録する。
・誤字・脱字、文字種の違いにも注意して、正確に記録する。
・入力後に入力者以外の者が原本と照合し、入力ミスを修正する。
・個人情報にかかわる部分(個人が特定される可能性のある語句や学校名、氏名・渾名など)は、当該部分を“＊”で置き換える。
・1 作文 1 ファイルで記録し、整理番号を付す。(整理番号から、課題・学年・クラス・性別などが判別できるようにする)

２．２　データの概要

2015 年 7 月現在の時点で電子化が完了しているテキストについて、構文解析を実施した結果を以下に示す。解析には、 CaboCha 0.69、UniDic 2.1.2 を使用した。

【表１】データの概要

課題	作文数	文数	文節数	短単位数	文字数
ゆめ	1,818	27,006	217,376	589,772	924,604
がんばったこと	1,599	27,829	208,208	580,151	922,914
計	3,417	54,835	425,584	1,169,923	1,847,518

データの作文数が異なるため、1 作文あたりの数も集計し、以下の結果を得た。

【表２】データの概要（1 作文あたり）

課題	文数	文節数	短単位数	文字数
ゆめ	14.9	119.6	324.4	508.6
がんばったこと	17.4	130.2	362.8	577.2
平均	16.2	124.9	343.6	542.9

1 作文あたりに換算すると、文数、文節数など今回調査した全ての項目において「がんばったこと」の方が数値が大きい。「ゆめ」は 7 月、「がんばったこと」は 12 月時点での調査である。数値の違いは、課題の違いによる可能性と、調査時期の違いによる可能性があるが、ここでは諸元の提示に留める。なお 1 作文あたりの平均的な分量は 400 字詰め原稿用紙 1.3 枚程度である。

３．『児童・生徒作文コーパス』における漢字の使用実態

３．１　学年別の漢字の使用頻度

学年別の教育漢字(小学校 6 年生までの学齢別配当漢字)、常用漢字(配当外)、常用外漢字の使用実態を以下に示す。数値は 2 課題の平均で、以下の調査も同様である。

【表３】学年別の漢字使用頻度(１万字あたり)

	小１	小２	小３	小４	小５	小６	中１	中２	中３
配当１年	113	361	438	420	477	534	581	554	588
配当２年	27	238	459	504	579	640	692	688	702
配当３年	9	10	210	290	365	414	455	415	472
配当４年	6	3	25	117	182	232	272	264	292
配当５年	3	2	5	15	77	145	190	203	222
配当６年	1	1	28	44	64	121	142	151	146
配当外	4	3	6	14	35	81	136	171	198
常用外	1	0	0	1	1	2	5	7	8
計	165	618	1171	1405	1780	2169	2474	2454	2627

表３から、小学校 1 年生の時点では、あまり漢字を用いず文章を書いているが、学齢が進むにつれて漢字の使用頻度が上がっていく様子が分かる。特に、小学校の低～中学年の

間は一定のペースで漢字の使用量が上昇する。教育漢字に関しては、小学 6 年次頃には、ほぼ変化しなくなり、一定程度定着したと見ることができそうである。

学習漢字を用いた漢字書きの発達をさらに詳細に見るため、中学３年の漢字使用状況を基準として学齢別に学習漢字の定着状況を以下に示す(中 3 の使用頻度を分母として百分率を計算した[2]。70%を超える学齢に下線を引いた)。

【表４】学年別の漢字使用頻度(中３を 100%として) [単位: %]

配当学年	小１	小２	小３	小４	小５	小６	中１	中２	中３
配当１年	17.6	61.0	**74.5**	71.6	81.1	90.4	98.9	94.5	100.0
配当２年	3.9	33.8	65.3	**71.8**	82.4	91.2	98.6	98.0	100.0
配当３年	2.0	2.3	44.0	61.9	**78.7**	88.8	97.3	88.9	100.0
配当４年	2.1	1.0	8.4	39.5	62.8	**79.9**	93.9	91.1	100.0
配当５年	1.3	0.6	2.2	7.2	34.6	64.8	**86.4**	92.3	100.0
配当６年	0.9	0.6	19.7	30.0	44.2	**83.9**	98.2	104.0	100.0
常用	2.3	1.7	2.7	7.1	17.5	39.9	66.9	**85.5**	100.0
常用外	5.2	2.8	5.6	7.1	9.4	30.2	69.5	**96.5**	100.0
計	6.3	23.5	44.6	53.5	67.8	**82.5**	94.2	93.4	100.0

教育漢字の使用頻度は、配当学年から 2 年程度で大学生の使用頻度の 70%に達する（例えば、小１配当の漢字が 70%を超えるのは、小３年次である)。中１時点では教育漢字の大半が 90%に達し、ほぼ定着したと見ることができる。

３．２　品詞別の漢字・仮名の書き分け

品詞別の漢字使用傾向を調査する。最初に、品詞別（自立語のみ）の１万語あたりの語彙頻度を示す。

【表５】学年別の品詞使用頻度（１万語あたり・自立語のみ）

品詞	小１	小２	小３	小４	小５	小６	中１	中２	中３
名詞	1882	1801	2043	1954	2015	2089	2112	2022	2182
動詞	1229	1299	1291	1376	1380	1423	1431	1461	1449
代名詞	198	247	224	221	214	229	235	266	236
副詞	283	276	236	236	244	214	206	217	187
形容詞	210	173	173	173	172	168	171	180	178
形状詞	125	118	111	108	121	126	132	135	132
連体詞	46	55	65	78	91	101	105	111	108
接続詞	28	38	33	46	47	51	53	51	53
感動詞	109	47	25	32	20	14	11	10	6

次に、これらの品詞について、漢字使用率の学年別推移を以下に示す[3]。

[2] 学習指導要領の中学３年次の文字に関する事項に「学年別漢字配当表に示されている漢字について、文や文章の中で使い慣れること。」とあるので、一応の目安とした。

[3] 漢字使用率は、品詞別の漢字書き数／当該品詞数（漢字書き＋仮名書き）で集計した。出現形が一字でも漢字を含んでいる場合、漢字書きと判定した。例えば「名まえ」のような表記も、漢字書きと判定した。以下の調査も同様に処理した。

【表６】学年別の品詞別漢字使用率［単位: %］

品詞	小１	小２	小３	小４	小５	小６	中１	中２	中３
名詞	9.5	32.9	50.9	56.5	67.1	74.0	77.8	77.8	78.6
動詞	1.5	12.5	19.6	25.7	29.0	34.4	41.2	42.1	42.7
代名詞	1.6	1.5	18.3	26.8	37.4	49.6	60.8	61.5	64.1
形容詞	2.4	4.4	9.2	10.3	12.9	19.3	22.1	23.7	24.0
形状詞	3.7	14.4	25.0	30.8	33.1	45.3	52.9	53.5	56.5
副詞	2.6	13.0	19.1	28.6	38.0	43.3	47.8	52.1	52.0
連体詞	2.5	6.1	9.0	8.4	7.1	8.1	10.7	12.5	9.2
接続詞	0.0	0.6	0.0	0.0	3.1	3.5	2.2	2.2	2.6
感動詞	0.0	0.0	0.0	0.0	0.0	0.4	0.0	0.0	0.0

品詞別に漢字書きの比率は異なり、およそ体言＞用言＞その他の語の順で漢字書きの比率が高くなる傾向がある。中３の比率を分母として百分率を計算したものを以下に示す。

【表７】学年別の品詞別漢字使用率（中３を 100%として）［単位: %］

品詞	小１	小２	小３	小４	小５	小６	中１	中２	中３
名詞	12.0	41.9	64.8	**71.9**	85.4	94.1	99.0	99.0	100.0
動詞	3.5	29.3	45.9	60.3	68.0	**80.6**	96.5	98.7	100.0
代名詞	2.4	2.4	28.6	41.8	58.3	**77.3**	94.8	95.8	100.0
形容詞	10.2	18.2	38.2	43.0	53.7	**80.2**	92.0	98.8	100.0
形状詞	6.5	25.6	44.3	54.4	58.6	**80.2**	93.7	94.7	100.0
副詞	5.0	25.0	36.7	55.1	**73.2**	83.3	91.9	100.2	100.0

漢字書きの浸透が最も早いのは名詞で、小４年次で 70%を超える。動詞、代名詞、形容詞、形状詞、副詞は小学校高学年の段階で 70%に到達する。中学１年次には、全ての品詞の漢字使用率が 90%を超えるが、この理由として中学生になれば漢字で書ける語は品詞に関わりなく漢字で書くという意識の変化(または、教師の指導)があると考えられる。

３．３　語種別の漢字・仮名の書き分け

語種別の漢字使用傾向を調査する。最初に、語種別の１万語あたりの語彙頻度を示す。

【表８】学年別の語種使用頻度（１万語あたり・記号など除く）

語種	小１	小２	小３	小４	小５	小６	中１	中２	中３
和語	7172	6946	6906	7021	6906	6981	7047	7158	7095
漢語	1010	1044	1172	1106	1215	1282	1369	1269	1463
外来語	170	222	259	237	228	195	170	151	156
混種語	147	115	102	83	97	97	102	105	101
固有名詞	47	47	60	46	57	49	35	29	29

次に、これらの語種について、漢字使用率の学年別推移を以下に示す。

【表９】学年別の語種別漢字使用率（記号など覗く）[単位: %]

語種	小１	小２	小３	小４	小５	小６	中１	中２	中３
和語	1.4	6.8	11.9	13.9	15.6	18.2	19.8	20.2	19.8
漢語	**12.3**	**36.0**	57.1	66.0	79.2	86.5	90.7	92.0	**93.1**
外来語	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0
混種語	1.6	24.7	21.7	33.0	45.9	62.3	80.9	88.1	87.9
固有名詞	9.5	22.6	28.7	42.3	49.0	55.6	62.6	55.1	60.5

漢語は漢字との関連性が高い。本調査でも、中３の段階では 93.1%の漢語が漢字書きされている。そこで、子ども達が作文時にある語(漢語)を思いついても漢語は漢字で書くものだという規範意識が働いて、当該の漢語の使用をひかえるのではないかという予想ができる。しかし実際には、学習した漢字が少ない低学年においても、仮名書きの漢語が多数使用されている(小１：12.3%、小２：36.0%)。予想とは異なり、漢字で書くべきという規範意識の語彙の選択への影響は低いと考えられる。中３の比率を分母として百分率を計算したものを以下に示す。

【表１０】学年別の語種別漢字使用率（中３を 100%として）[単位: %]

語種	小１	小２	小３	小４	小５	小６	中１	中２	中３
和語	7.2	34.4	60.3	**69.9**	78.9	91.9	100.1	102.0	100.0
漢語	13.2	38.7	61.4	**71.0**	85.1	93.0	97.4	98.8	100.0
混種語	1.9	28.1	24.7	37.5	52.2	**70.9**	92.0	100.3	100.0
固有名詞	15.7	37.3	47.4	**69.9**	80.9	91.9	103.5	91.1	100.0

漢語、和語、固有名詞は小４年次でほぼ 70%に達し、混種語は小６年次で 70%に達する。中１の段階では、いずれも 90%を超える。

３．４　高頻度語彙における漢字使用頻度

使用頻度の高い動詞 20 種について漢字の使用実態を調べる。最初に、それらの動詞の 100 万語あたりの使用頻度を示す。順位は全学年の平均頻度による。

【表１１】学年別の語彙使用頻度(動詞頻度上位 20 語／100 万語あたり)

語彙素	小１	小２	小３	小４	小５	小６	中１	中２	中３	平均
為る	13485	18657	15482	16706	17143	17538	17744	17429	18625	16979
居る	6940	9943	12435	12794	12960	14478	13181	14710	12913	12262
成る	14963	14053	11512	11101	10908	10160	9975	8780	9770	11247
言う	3817	7078	7166	7007	6980	7814	7090	8790	7799	7060
思う	2897	4967	5754	7472	7715	8345	9243	8838	7809	7005
有る	4527	4021	4893	5064	5945	7192	6636	7497	8002	5975
頑張る	9834	6657	5942	4305	4904	4560	4601	4544	4137	5498
行く	3318	3658	3845	4048	4681	4513	4436	4421	4746	4185
出来る	3586	4013	3047	3553	3481	3661	4301	3861	4498	3778
見る	4384	2689	3110	4918	3264	2807	3282	3169	3435	3451
遣る	3247	3938	3342	3277	2899	2808	2123	2756	1482	2875
来る	2486	2476	2016	2622	2548	1926	2410	2665	2402	2395
仕舞う	519	1007	1461	2011	2266	1924	1746	2176	1655	1641

作る	3573	2404	2077	1478	1119	1242	1052	677	810	1604
考える	292	287	667	634	954	1667	1442	1817	2107	1096
呉れる	540	1037	1038	1197	1536	1178	1303	1050	939	1091
貰う	1541	1401	1516	1167	925	822	815	691	500	1042
出る	916	1174	1074	1340	1101	940	934	988	902	1041
分かる	434	564	821	839	1095	1259	1272	1211	1064	951
入る	519	612	1091	950	875	1015	1216	1124	968	930

次に、これらの動詞について、漢字使用率の学年別推移を以下に示す。表の再右列は辞書形に含まれる漢字の配当学年(複数の漢字が含まれる場合は最も低い学年)である。

【表１２】学年別の語彙別漢字使用率(動詞頻度上位 20 語) [単位: %]

語彙素	小１	小２	小３	小４	小５	小６	中１	中２	中３	配当
為る	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	**0.0**	常用
居る	0.6	0.0	0.0	0.0	0.1	0.2	0.0	0.1	**0.1**	小５
成る	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	**0.1**	小４
言う	1.0	27.5	28.9	26.2	19.4	22.9	24.7	27.6	21.0	小２
思う	10.3	88.5	92.5	97.1	98.6	99.5	99.8	99.6	99.9	小２
有る	3.3	0.0	0.5	0.0	1.0	0.1	0.0	0.0	**0.2**	小３
頑張る	2.3	1.1	0.0	0.4	9.3	32.3	59.0	68.9	73.1	小５
行く	0.0	19.5	31.8	36.8	28.9	33.7	24.3	26.1	29.3	小２
出来る	0.8	2.4	11.0	11.6	7.1	13.9	13.9	12.6	10.3	小１
見る	7.5	46.8	41.9	43.9	46.1	51.6	63.4	57.1	65.2	小１
遣る	0.0	0.0	0.0	0.2	0.0	0.0	0.3	0.0	**0.0**	常用
来る	0.9	3.8	11.3	15.1	15.0	12.4	17.2	14.6	11.0	小２
仕舞う	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	小３
作る	9.4	46.5	65.5	61.3	78.6	82.9	54.9	66.2	68.7	小２
考える	0.0	79.5	91.7	96.0	99.2	99.7	99.7	100.0	100.0	小２
呉れる	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	常用
貰う	0.0	0.0	0.0	0.0	0.0	0.0	0.9	2.6	**0.0**	常用外
出る	6.8	55.0	49.7	61.4	62.1	69.4	78.4	81.4	83.4	小１
分かる	0.0	15.9	25.9	40.5	54.4	54.8	67.2	70.5	71.9	小２
入る	0.0	74.0	83.4	87.8	90.6	98.1	99.5	98.9	99.0	小１

これらの動詞の多くは中学校までに習う漢字(常用漢字)で漢字書きが可能だが、中３段階でもほとんど漢字書きされないもの、漢字と仮名の書き分けがあるもの、ほとんど漢字書きされるものがあることが分かる。ほとんど漢字書きされない語(中３年次で漢字書きが 1%未満の語に下線を付した)以外の語の多くは、使用される漢字が小１～小２に配当されており、高頻度語でありながら学習時期が遅い漢字は表の範囲では見当たらない。その意味で教育漢字の配当順は、子ども達の使用実態に即したものであると評価することができる。「頑張る」が唯一の例外だが、これは「がんばったこと」という課題の影響で頻度が高くなっているだけであり、本来はそれほど高頻度の語ではないと考えられる。

４．大人の文章との対照

４．１　横断的分析: 作文の文体的特徴

作文の文体的特徴を確認するために、大学生の作文[4]（「夢」「がんばったこと」）と BCCWJ コアデータ（知恵袋、ブログ、書籍、雑誌、新聞、白書）の漢字使用頻度を、漢字種別（小学校配当の教育漢字、教育漢字以外の常用漢字、常用外漢字）ごとに調べた。結果を以下に示す。

【表１３】レジスタ別の漢字使用頻度（１万字あたり）

レジスタ	配 1 年	配 2 年	配 3 年	配 4 年	配 5 年	配 6 年	常用	常用外	計
白書	528.4	884.0	1038.6	732.8	748.4	296.1	401.8	3.7	4633.8
新聞	779.6	912.6	745.7	544.0	453.5	231.1	414.0	32.1	4112.7
雑誌	542.9	641.6	528.2	348.3	274.9	155.7	354.2	47.0	2892.6
書籍	547.5	653.8	526.0	325.1	257.1	159.1	318.9	37.4	2825.1
作文（夢）	567.5	684.8	462.3	309.7	368.3	166.0	216.1	21.3	2796.0
作文（が）	587.8	715.3	499.6	319.6	213.5	124.8	198.4	12.3	2671.3
ブログ	458.2	633.6	453.5	303.7	201.0	125.6	285.8	36.7	2498.2
知恵袋	413.7	641.8	427.8	299.9	209.6	132.8	265.5	19.6	2410.6

大学生の作文の漢字使用頻度は１万字あたり 2671～2796 字である。4000 字以上である白書・新聞とは大きな隔たりがあるが、それ以外のレジスタとは極端な差はなく、おおよそ雑誌・書籍とブログ・知恵袋の中間程度である。

「夢」と「がんばったこと」では「夢」の漢字使用頻度が高く、特に 5 年配当漢字の使用頻度が高い。これは 5 年配当である「夢」という漢字が多く含まれている(1 万字あたり約 160 字)ためであり、それを除外すると「夢」と「がんばったこと」の差は小さくなる。それ以外の特徴としては、「夢」の方が６年配当や配当外の漢字使用頻度が高く、「がんばったこと」の方が低～中学年配当の漢字使用頻度が高い傾向がある。これは 2 つの課題で使用される語彙の違いを反映している可能性があるが、より詳細な分析は今後の課題としたい。

４．２　縦断的分析: 漢字使用能力の飽和状況

中学生までの漢字学習で、児童・生徒の漢字使用能力がどの程度まで大人の漢字使用能力に接近するかを見るために、中学３年次の漢字使用実態と大学生の漢字使用実態を対照する。配当学年別の漢字使用頻度、品詞別、語種別の漢字使用率について、中３と大学生を対照した表を以下に示す。それぞれ表３、６、９の中３の数値に、大学生の数値を並べたものである。

【表１４】配当学年別の漢字使用頻度 (１万字あたり)

配当漢字	配 1 年	配 2 年	配 3 年	配 4 年	配 5 年	配 6 年	常用	常用外	計
中３	646	705	518	320	180	127	223	6	2726
大学生	588	715	500	320	213	125	198	12	2671

[4] 大学生 1，2 年生に調査協力を依頼し、「ゆめ」、「頑張ったこと」でそれぞれ作文課題を課した。これにより、「ゆめ」108 編、「頑張ったこと」223 編の作文を収集した。なお調査に際して、A4 用紙 1 枚程度(1600 字)という目安を示したが自宅での課題としたため、条件に幅があることを断っておく。

【表１５】品詞別の漢字使用率 [単位: %]

品詞	名詞	動詞	代名詞	形容詞	形状詞	副詞	連体詞	接続詞	感動詞
中３	78.6	42.7	64.1	24.0	56.5	52.0	9.2	2.6	0.0
大学生	67.8	42.0	60.4	25.6	58.2	52.9	11.9	1.3	0.0

【表１６】語種別の漢字使用率 [単位: %]

語種	和語	漢語	外来語	混種語	固有名詞
中３	19.8	93.1	0.0	87.9	60.5
大学生	19.8	78.6	0.7	85.8	88.9

個別の項目を見ると、中３の時点でほぼ大学生と同等の水準に達しているもの、大学生の水準にやや及ばないもの、中３の時点の方がむしろ数値が高いものがある。例えば表１４は、個々の項目について前後はあるが、全体としては中３の方が大学生より漢字使用頻度が高いことを示している。表１５を見ると、名詞、動詞、代名詞、接続詞は中３の方が漢字使用率が高いが、形容詞、形状詞、連体詞は大学生の方が漢字使用率が高い。表１６を見ると、漢語や混種語は中３の方が漢字使用率が高いが、固有名詞は大学生の方が漢字使用率が高い。これらの差異の意味を分析するためには、各項目に含まれるどのような語彙が差異を生み出しているかについて、より詳しく調査する必要がある。しかし全体としては、これらの数値は概ね中学３年次の漢字使用能力が大学生の漢字使用能力に接近していることを示しており、高校以降の変化が無いとまでは言えないものの、中学修了段階でかなりの程度飽和状態に近づいていると考えられる。

５．おわりに

本発表では、作文コーパスに基づいて児童・生徒の漢字使用能力の推定を試みた。また、BCCWJ のコアデータや大学生の作文と対照することによって、子ども達の漢字使用能力が大人のそれにどの程度近づいているのかについても言及し、発達過程の概要を示した。

より詳細な分析を進めるために、現在『児童・生徒作文コーパス』の内 10 万形態素程度を目標に(全体の 1 割弱)、自動解析後に人手修正を行ったコアデータの構築を進めている。現在使用しているデータは自動解析によって形態論情報等を付与しているが、誤字脱字や仮名書きが多い低学年の作文は自動解析の精度が低く、十分な信頼性を確保できていない。人手で形態論情報、構文情報を付与したコーパスを整備することによって、本発表で得られた調査結果を再検討するとともに、今後は以下のような課題の分析を進めていきたい。

- 同一語での仮名書きと漢字書きの傾向差に関する議論
- 同一漢字を用いる異語の漢字書きの傾向差に関する議論（「下る」と「下がる」など）
- 作文文型の発達と語彙・漢字使用についての議論

本研究は『児童・生徒作文コーパス』を使用した一連の研究の一部である。これと並行して、発表者ら以外の共同研究者によって同コーパスを利用した作文能力の発達過程の推定と数値化が進められている。中でも子ども達の漢字使用能力に関する研究は、現場からの要請が強く、率先して進められるべきものの一つである。本研究の最終的な目標は、教育現場における作文教育の改善と適正化を図ることにある。研究が進み、言語研究の立場から現場の教師が手軽に利用できる漢字使用の実態の分析や作文指導の指針を提案し、有

効に活用されれば、昨今二者の乖離が叫ばれて久しい研究と教育の現場の協働の一つの形として位置づけることができる。

謝　辞

本研究は、平成 27 年度 漢字・日本語教育研究助成制度「作文コーパスを資料に児童・生徒の漢字使用・選択傾向と発達の実態を明らかにする。―語彙情報つき作文コーパスの構築と学齢別語彙・漢字使用実態調査―」(研究代表者：宮城信)、および日本学術振興会科学研究費補助金基盤研究(B)「作文を支援する語彙・文法的事項に関する研究」(平成 26～30 年度、研究代表者：矢澤真人、研究課題番号：26285196)による補助を得ています。

文　献

河内昭浩(2015)「国語教育のための「常用漢字表」語例の検討」『第 7 回コーパス日本語学ワークショップ予稿集』pp.113-122.(https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no7_papers/JCLWorkshop_No7_web.pdf よりダウンロード可能)

鈴木一史(2011)「作文コーパスからみる生徒の使用語彙」『特定領域「日本語コーパス」平成 22 年度公開ワークショップ(研究成果報告会)予稿集』pp.343-350.(http://www.ninjal.ac.jp/corpus_center/bccwj/doc/workshop/JC-G-10-02.pdf よりダウンロード可能)

田中牧郎(2011)「語彙レベルに基づく重要語彙リストの作成 ―国語政策・国語教育での活用のために―」pp.77-87.(http://www.ninjal.ac.jp/corpus_center/bccwj/doc/workshop/JC-G-10-02.pdf よりダウンロード可能)

宮城信・今田水穂(2015)「『児童・生徒作文コーパス』の設計」『第 7 回コーパス日本語学ワークショップ予稿集』pp.223-232.(https://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no7_papers/JCLWorkshop_No7_web.pdf よりダウンロード可能)

関連 URL

作文を支援する語彙文法的事項に関する研究プロジェクト　https://sites.google.com/site/sakubunshienpropject/

『虎明本狂言集』における濁点表記状況 —全例に濁点が付された語を中心に—

渡辺由貴[†]・市村太郎[‡]（国立国語研究所コーパス開発センター）

Dakuten in *Toraakira-bon Kyogen*: Focusing on Words that Appeared Always with *Dakuten*

Yuki Watanabe・Taro Ichimura (National Institute for Japanese Language and Linguistics)

要旨

『日本語歴史コーパス　室町時代編Ⅰ　狂言』（短単位データ 0.9）のデータを用い、『虎明本狂言集』における濁点の付与状況を、全例で濁点が付されている語を中心に調査した。全体としては、濁点無表記例のある語より、全例で濁点が付されている語の方が多い。また、全例で濁点が付されている語については、『虎明本狂言集』全体の語種比率と比べ和語の比率が低く、漢語の比率が高くなっていた。これは、使用頻度の高い特定の助詞・助動詞において濁点無表記率が高いためだと考えられる。さらに、全体で用例が 1 例のみの語については、9 割以上の語で濁点が表記されている一方、用例数が多くても他の語と混同される可能性があると考えられる語においては全例で濁点が付されている場合がある等の状況が確認された。『虎明本狂言集』においては、誤読を避けるべく清濁の区別を明確に示す表記が意識的に行われていたと考えられる。

１．はじめに

『虎明本狂言集』（1642）においては、他の中近世期の仮名資料と同様、濁音が想定される仮名全てに濁点が付されているわけではない。濁音で読まれる仮名に「゛」の濁点を付すという対応が定着するのは近代以降であり、中近世期には、濁音で読まれながらも濁点を付さない表記が混在していた。

沼本（1997）によれば、記号として仮名右肩に濁点を付すのが定着したのは「1600 年前後」と推定される（p.927）とのことであるが、濁音で読まれる仮名には濁点を付すという対応が定着するのは近代以降であり（近藤 2005 等）、近世期は、濁点の使用という面では、濁音で読まれながらも濁点を付さない「清濁の消極的表記」（松本 1978、p.25）が混在する時代であった。この過渡的な時代の資料における濁点付与についての調査には『玉塵抄』を対象とした出雲（1976）があり、語種、自立語・付属語の別による傾向や、用例数の多寡との関係、語の識別、「表記の経済性」（p.11）等が指摘されている。一般的な傾向、あるいは資料独自の傾向を見出すためには、さらに多くの資料を対象にデータを蓄積し、検討する必要があろう。

渡辺・市村（2014）では、このような状況をふまえ、『虎明本狂言集』における濁点の無表記箇所について述べたが[1]、濁点表記状況を明らかにするためには、一方で全例に濁点が付された語についてもあわせて考察する必要があろう。本発表では、『日本語歴史コーパス

† ywatanabe@ninjal.ac.jp

‡ tichimura@ninjal.ac.jp

1　渡辺・市村（2014）は、整備中のデータを利用したため、調査対象を「脇狂言之類」から「女狂言之類」までの各類に限定したものである。

室町時代編Ⅰ　狂言』のデータに付与したタグ情報を利用し、『虎明本狂言集』において全例に濁点が付された語を中心に検討し、中近世期の濁点表記状況を明らかにする試みの一端としたい。

２.『虎明本狂言集』コーパスデータについて

本発表では、『日本語歴史コーパス　室町時代編Ⅰ　狂言』（短単位データ 0.9）のコーパスデータを調査対象とする。このコーパスデータは、大塚光信編（2006）『大蔵虎明能狂言集　翻刻　註解』(上・下、清文堂出版)[2]を底本とし、会話(<speech>)、ト書き(<stage>)等、本文中の要素にタグを付与、XML 形式で構造化されている[3]。その過程で、濁音で読まれると推定されるものの濁点が付与されていない仮名については濁点付きの仮名に置き換え、<vMark>タグを付与している[4]。例えば、底本テキストで「さらは」となっている箇所を、コーパスデータでは「さら<vMark>ば</vMark>」としている[5][6]。

本発表では、この校訂箇所を示すタグを利用し、おもに全例に濁点が付与されている語について、もともと濁点が付されていた箇所（タグの付与されていない箇所）と濁音を表すタグが付与された箇所とを比較しつつ、計量的に検討することにより、その傾向や特徴を検討する。

３.『虎明本狂言集』における濁点の付与傾向について

３．１.濁点付与状況の概観

まず、濁音が想定される語のうち、濁点無表記例のある語および、全例に濁点が付されている語について概観する。表 1 をみると、総数・異なりの両方において、濁点無表記の例がある語よりも、全例に濁点が付されている語の方が多いことがわかる。

表 1　濁点無表記語と全例に濁点が付されている語の語数

	総数	異なり
全例に濁点が付されている語	7911	2248
濁点無表記語	4777	574

※以降、濁点無表記語の「総数」には、濁点が表記されている例を含めていない。例えば、語「合図」7 例のうち、濁点無表記（「あひつ」）の 1 例のみを「総数」に含めている。

次に、濁点無表記語例のある語と全例に濁点が付されている語について、語種別・品詞別に整理すると、次のようになる。

2　凡例に「仮名遣いや清濁・読点は原文のままとする。」（p.vi）とあり、10 曲を影印と照合し確認したところ、問題はなかった。

3　タグ仕様の詳細は市村他（2012）、市村（2014）等参照。

4　振り仮名については<vMark>タグを付与していない。

5　なお、濁点を付与すべきか判断に迷うものが現れた際は、他の曲中で底本に濁点がついている例がないか、『日本国語大辞典』『時代別国語大辞典』『日葡辞書』等における出現状況はどのようになっているか等を確認し、清音の可能性があるものには濁点を付与せず、濁音で読まれる可能性の高いもののみに付与するという方針を立てている。例えば、「ひさう」（秘蔵）という語は、仮名表記された 23 例中、「ひざう」表記の例は 1 例もなく、また『日本国語大辞典』の「ひぞう」の項に「古くは『ひそう』」とあり、『日葡辞書』でも「Fisŏ」「Fisŏna」の形で立項されているため、タグは付与せず「ひさう」のままとしている。

6　近代語資料における濁点自動付与の手法については、岡他（2013）の研究があるが、中近世語資料については、『日本語歴史コーパス　室町時代編Ⅰ　狂言』が現段階では唯一のコーパスデータであり、機械学習による濁点付与を行うには困難な点が多かった。

表 2　語種別語数（総数）

全例に濁点が付されている語（総数）			濁点無表記語（総数）		
語種	用例数	%	語種	用例数	%
和語	4791	60.6	和語	4461	93.4
漢語	2359	29.8	漢語	185	3.9
外来語	43	0.5	外来語	10	0.2
混種語	325	4.1	混種語	77	1.6
固有名詞	383	4.8	固有名詞	42	0.9
その他	10	0.1	その他	2	0.0
計	7911		計	4777	

表 3　「『日本語歴史コーパス』語彙統計」による狂言の語種比率【参考】[7]

「日本語歴史コーパス」語彙統計		
語種	用例数	%
和語	207256	88.2
漢語	18750	8.0
外来語	207	0.1
混種語	6196	2.6
固有名詞	2434	1.0
その他	251	0.1
計	235094	

語種別の語数を総数でみると、濁点無表記語は和語が 9 割以上を占めているのに対し、全例に濁点が付されている語は、和語が約 6 割、漢語が約 3 割となっている。また、固有名詞の比率も、濁点無表記語では 1%程度であるが、全例で濁点が付されている語については 5%近くとなっている。「『日本語歴史コーパス』語彙統計」による狂言全体の語種比率（表 3）と比べても、全例に濁点が付されている語の和語の比率の低さおよび、漢語・固有名詞の比率の高さがうかがえる。

表 4　語種別語数（異なり）

全例に濁点が付されている語（異なり）			濁点無表記語（異なり）		
語種	用例数	%	語種	語数	%
和語	1360	60.5	和語	395	68.8
漢語	609	27.1	漢語	103	18.0
外来語	10	0.4	外来語	5	0.9
混種語	114	5.1	混種語	33	5.7
固有名詞	147	6.5	固有名詞	36	6.3
その他	8	0.4	その他	2	0.3
計	2248		計	574	

語種別の語数を異なりでみると、表 4 のようになる。濁音で読むと想定される漢語 712 語[8]（異なり）に注目すると、85.5%にあたる 609 語において全例に濁点が付されており、漢語においては多くの場合、全例で濁点が付されていることがわかる[9]。

全例に濁点が付されている語については、総数における比率と似た傾向がみられるが、濁点無表記語については、総数と異なりとで大きく傾向が異なり、異なりでは和語の比率が総数に比べ大幅に低くなっている。これは、接続助詞「ば」や、「をば」「ごとし」等の、用例数の多い特定の機能語において、濁点無表記例が 8 割を超えているために（渡辺・市村（2014）および表 7）、総数で和語の比率が高くなっているが、異なりではその率がやや低くなっていることと関係していると考えられる。

7　「『日本語歴史コーパス』語彙統計」で示された各類の合計を整理したものである。その際、記号 42364 語（句読点等）は除いた。

8　全例に濁点が付されている 609 語と、濁点無表記例のある 103 語の合計。

9　後掲の表 8 において、全例で濁点が付されている語上位の 62 語の品詞をみてみると、和語が 40 語（64.5%）、漢語が 19 語（30.6%）、混種語が 2 例（3.2%）、固有名詞が 1 例（1.6%）となっており、表 2・4 と同様、漢語の比率が比較的高くなっている。

品詞別の用例数をみると（表 5)、全例で濁点が付されている語については、総数・異なりとも、普通名詞の比率が比較的高く、助詞・助動詞の比率は低い。一方、濁点無表記語については、総数では助詞が 5 割以上、助動詞が約 14%を占めるが、異なりではそれぞれ約 4%、約 2%となっており、表 2・4 で見られた傾向を裏付けるものである。

表 5　品詞別用例数

全例に濁点が付されている語				
品詞	総数		異なり	
	用例数	%	用例数	%
普通名詞	4639	58.6	1451	64.5
固有名詞	383	4.8	147	6.5
数詞	6	0.1	2	0.1
代名詞	115	1.5	4	0.2
動詞	1684	21.3	440	19.6
形容詞	255	3.2	49	2.2
形状詞	90	1.1	34	1.5
副詞	403	5.1	77	3.4
連体詞	3	0.0	1	0.0
接続詞	0	0.0	0	0.0
感動詞	76	1.0	6	0.3
助詞	60	0.8	7	0.3
助動詞	31	0.4	3	0.1
接尾辞	159	2.0	21	0.9
接頭辞	3	0.0	3	0.1
その他	4	0.1	3	0.1
合計	7911		2248	

濁点無表記語				
品詞	総数		異なり	
	用例数	%	用例数	%
普通名詞	513	10.7	280	48.8
固有名詞	42	0.9	36	6.3
数詞	0	0.0	0	0.0
代名詞	48	1.0	8	1.4
動詞	391	8.2	139	24.2
形容詞	48	1.0	16	2.8
形状詞	9	0.2	8	1.4
副詞	223	4.7	34	5.9
連体詞	8	0.2	2	0.3
接続詞	101	2.1	2	0.3
感動詞	30	0.6	2	0.3
助詞	2655	55.6	21	3.7
助動詞	675	14.1	12	2.1
接尾辞	30	0.6	11	1.9
接頭辞	2	0.0	1	0.2
その他	2	0.0	2	0.3
合計	4777		574	

表 6　仮名別用例数＜総数＞

仮名	当該仮名用例総数	全例に濁点が付されている語		濁点無表記語	
		総数	%	総数	%
が	7374	1126	15.3	145	2.0
ぎ	718	425	59.2	51	7.1
ぐ	492	315	64.0	19	3.9
げ	742	366	49.3	35	4.7
ご	3305	390	11.8	643	19.5
ざ	4249	445	10.5	44	1.0
じ	3930	668	17.0	75	1.9
ず	1290	93	7.2	49	3.8
ぜ	598	318	53.2	21	3.5
ぞ	2145	147	6.9	29	1.4
だ	1977	705	35.7	87	4.4
ぢ	628	280	44.6	15	2.4
づ	1234	221	17.9	368	29.8
で	4659	330	7.1	82	1.8
ど	2915	580	19.9	74	2.5
ば	4128	473	11.5	2708	65.6
び	846	476	56.3	32	3.8
ぶ	934	546	58.5	71	7.6
べ	528	269	50.9	15	2.8
ぼ	551	82	14.9	229	41.6
合計	43243	8255	19.1	4792	11.1

※一語内の二つ以上の仮名で濁点が表記／無表記されている場合は、両方の仮名の総数に含めている。

また、仮名別の用例数をみると表 6 のようになる。全例に濁点が付されている語に含まれる仮名としては、「ぐ」（64.0%）「ぎ」（59.2%）「ぶ」（58.5%）「び」（56.3%）「ぜ」（53.2%）「べ」（50.9%）が多くなっている。一方、「ぞ」「で」「ず」「ざ」「ば」「ご」等の仮名ではその比率が低くなっているが、これらの仮名は「ば」や「ごとし」等の助詞・助動詞で用

いられるため、濁点無表記の例が比較的多いことが一因であると考えられる。

３．２．助詞・助動詞について

ここで、助詞・助動詞について詳しくみていきたい。出雲（1976、pp.2-3）は、『玉塵抄』において、「もっとも濁音表記される率が低いのは、付属語、接尾語の類である。」としており、後掲の表 8 にあがっている、全例で濁点が付されている語（短単位）20 例以上の語のうち、助詞・助動詞は、副助詞「がな」および助動詞「です」の 2 語のみであるが、助詞・助動詞の濁点表記率はどのようになっているだろうか。

表 7　助詞・助動詞の濁点表記率

語	濁点表記例	語全例	濁点表記率
がな:助詞-副助詞	29	29	100
です:助動詞	21	21	100
ばし:助詞-副助詞	10	10	100
だに:助詞-副助詞	9	9	100
げな:助動詞	9	9	100
もが:助詞-終助詞	4	4	100
なんぞ:助詞-副助詞	4	4	100
がな:助詞-終助詞	3	3	100
が:助詞-準体助詞	1	1	100
べい:助動詞	1	1	100
じゃ:助動詞	1784	1790	99.7
が:助詞-接続助詞	1169	1177	99.3
なり:助動詞	2204	2225	99.1
ぞ:助詞-終助詞	1349	1363	99.0
ばかり:助詞-副助詞	100	101	99.0
が:助詞-格助詞	3736	3780	98.8
ほど:助詞-副助詞	245	248	98.8
ながら:助詞-接続助詞	248	252	98.4
など:助詞-副助詞	243	247	98.4
まで:助詞-副助詞	402	409	98.3
た:助動詞	175	178	98.3
て:助詞-接続助詞	509	519	98.1
むず:助動詞	324	332	97.6
で:助詞-格助詞	420	431	97.4
べし:助動詞	171	176	97.2
なんだ:助動詞	103	106	97.2
いで:助詞-接続助詞	258	266	97.0
ばや:助詞-終助詞	105	109	96.3
ぞ:助詞-係助詞	179	186	96.2
ども:助詞-接続助詞	298	310	96.1
ず:助動詞	488	511	95.5
たり:助動詞	20	21	95.2
たがる:助動詞	14	15	93.3
ど:助詞-接続助詞	30	33	90.9
じ:ジ:和:助動詞	36	40	90.0
で:助詞-接続助詞	36	40	90.0
ずつ:助詞-副助詞	69	80	86.3
まじ:助動詞	60	71	84.5
つ:助詞-副助詞	4	6	66.7
だ:助動詞	2	3	66.7
をば:助詞-格助詞	14	94	14.9
ば:助詞-接続助詞	178	2595	6.9
ごとし:助動詞	18	607	3.0
則ば:助詞-接続助詞	0	1	0

表 7 に示した通り、副助詞「がな」「ばし」「だに」「なんぞ」、助動詞「です」「げな」「べい」、終助詞「もが」「がな」、準体助詞「が」については、全例で濁点が付されている。また、助動詞「じゃ」「なり」 や接続助詞「が」、終助詞「ぞ」、格助詞「が」等の語は、語全体で 1000 例以上の用例があるにも関わらず、濁点表記率は 100%近くなっている。むしろ、助動詞「ごとし」、接続助詞「ば」「則ば」、格助詞「をば」のように、濁点無表記になりやすい語の方が少数である。

このように、『虎明本狂言集』においては、必ずしも全ての機能語が濁点無表記になりやすいわけではなく、特定の助詞・助動詞において濁点が付されないことが多いことがわかる。

３．３．全例で濁点が付されている語（短単位）について

ここで、濁音で読むと想定される箇所について、全例で濁点が付されている語が 20 例以上ある語を確認する。表 8 をみると、「食べる」「呼ぶ」「是非」のような用例数の多い語でも、全例に濁点が付されることがあることがわかる。用例数の多い語においては、一部濁点が無表記であっても、濁音であることを予想することが容易であるように思われるが、これらの語で、全例において濁点が付されている背景には、どのようなことが考えられるだろうか。

表 8　全例で濁点が付されている語（短単位）のうち用例数 20 例以上の語

	語（短単位）	例	用例数
1	食べる:タベル	た【べ】て、た【ぶ】れば	102
2	呼ぶ:ヨブ	よ【ば】う、よ【び】て、よ【ぶ】、よ【べ】	96
3	是非:ゼヒ	ぜ【ひ】	91
4	乍ら:ナガラ（接尾辞）	二人な【が】ら	66
5	定めて:サダメテ	さ【だ】めて御ふつきにござらふ	61
5	進ぜる:シンゼル	しん【ぜ】て	61
7	合点:ガッテン	【が】てん、【が】つてん	53
8	いで:イデ（感動詞）	い【で】くらはう	50
9	逃げる:ニゲル	に【ぐ】る、に【げ】た	49
10	山伏:ヤマブシ	山【ぶ】し	48
11	何れ:ドレ	【ど】れ	45
11	機嫌:キゲン	き【げ】ん	45
13	御:ゴ（接尾辞）	おうぢ【ご】、ちち【ご】	44
13	何方:ドチ	【ど】ちへゆくぞ	44
15	夥しい:オビタダシイ	お【び】たたしひ	42
15	出す:ダス	【だ】して	42
17	時宜:ジギ	【じぎ】、【ぢぎ】	41
18	昆布:コンブ	こ【ぶ】	40
18	座頭:ザトウ	【ざ】とう	40
20	聊爾:リョウジ	れう【じ】	38
20	成敗:セイバイ	せい【ば】い	38
20	橋懸かり:ハシガカリ	はし【が】かり	38
23	棒:ボウ	【ば】う、【ぼ】う	37
23	直ぐ:スグ	す【ぐ】	37
25	出来る:デクル	【で】きた	35
25	戯言:ザレゴト	【ざ】れ事	35
27	苦々しい:ニガニガシイ	に【が】/ \しひ	33
28	がな:ガナ（副助詞）	何と【が】なして	29
28	暇乞い:イトマゴイ	いとま【ご】ひ	29
30	舞台:ブタイ	【ぶ】たい	28
30	定まる:サダマル	さ【だ】まつた	28
30	座禅:ザゼン	【ざ】【ぜ】ん	28

	語（短単位）	例	用例数
33	床机:ショウギ	しやう【ぎ】	27
33	罪人:ザイニン	【ざ】い人	27
35	物語:モノガタリ	物【が】たり	26
35	志:ココロザシ	心【ざ】し	26
35	勝負:ショウブ	せう【ぶ】	26
38	倅:セガレ	せ【が】れ	25
38	道すがら:ミチスガラ	みちす【が】ら	25
38	出で来る:イデクル	い【で】くる	25
38	何とぞ:ナニトゾ	何と【ぞ】	25
42	騙す:ダマス	【だ】ます	24
42	しゃぎり:シャギリ	しや【ぎ】り	24
42	餓鬼:ガキ	【が】き、【が】つき	24
45	恥:ハジ	は【じ】、は【ぢ】	23
45	自然:シゼン	し【ぜ】ん	23
45	雁:ガン	【が】ん	23
48	前廉:マエカド	まへか【ど】	22
48	座敷:ザシキ	【ざ】しき	22
50	博労:バクロウ	【ば】くらう	21
50	苦る:ニガル	に【が】つた	21
50	です:デス	大名【で】す	21
50	何処許:ドコモト	【ど】こもと	21
54	詫び言:ワビゴト	わ【び】事	20
54	被く:カズク	か【づ】く	20
54	流石:サスガ	さす【が】	20
54	互い:タガイ	た【が】ひ	20
54	脅す:オドス	お【ど】す	20
54	首:クビ	く【び】	20
54	ブアク:ブアク	【ぶ】あく	20
54	楽屋:ガクヤ	【が】くや	20
54	慰み:ナグサミ	な【ぐ】さみ	20

表 9　濁点無表記の場合に別の語と表記が重なる語の例

	語（短単位）	濁点無表記の場合に表記が重なる語の例	狂言内の表記
1	食べる:タベル	耐える	たへ
2	呼ぶ:ヨブ	酔う、用、様	よふ、よへ
8	いで:イデ（感動詞）	行く	い（て）
11	何れ:ドレ	取る	とれ
18	昆布:コンブ	請う	こふ
18	座頭:ザトウ	砂糖	さたう
20	聊爾:リョウジ	漁師	れうし
23	棒:ボウ	方、法、箔	はう、ほう
28	がな:ガナ（副助詞）	哉	かな
42	餓鬼:ガキ	柿、垣	かき
45	恥:ハジ	橋、端、箸、嘴	はし
45	雁:ガン	感、羹、燗、漢	かん
50	博労:バクロウ	白浪	はくらう
54	互い:タガイ	高い	たかひ
54	脅す:オドス	落とす	おとい、おとさ、おとし、おとす、おとひ
54	首:クビ	杭	くひ
54	楽屋:ガクヤ	隔夜	かくや

〇濁点無表記の場合に別の語と表記が重なる語について

「食べる」全例に濁点が付されていることの一因に、「耐える」との混同を避けることが考えられる。「食べる」のうち、83 例が「たべ」表記であるが、「耐える」6 例のうち 4 例が「たへ」（あとの 2 例は「たえ」）表記であり、仮に「食べ」を「たへ」と表記すると、両者の表記が重なってしまう。このような混同を避けるために、「食べる」において濁点が明示された可能性がある。なお、「食べ」を含む複合語である「食べ酔う」10 例、「食べ過ごす」1 例についても、全例で濁点が付されていた。

「呼ぶ」については、仮に「よふ」と表記すると、「酔う」や「用」「様」等の語と表記が重なる。この他、濁点を表記しなかった場合に別の語と表記が重なる語の例を表 9 に示したが、このように、これらの語において用例数が多いにも関わらずそれぞれに濁点が明示された背景には、表記が類似する語との混同を避けることがあると考えられる。

また、「棒」全例に濁点が付されている点についても、「はう」「ほう」と表記した場合に起こりうる、「方」等の語との混同の回避が考えられる。ただし、同じく「ボウ」と読む「坊」については濁点無表記例があり、仮名表記の 28 例中、濁点無表記例が 8 例となっているが、「坊」の例を見ると、「きたい【は】う」（希代坊）4 例、「ふしやう【は】う」（不請坊）3 例、「てらのご【は】う」（寺の御坊）1 例のいずれも、「方」との混同が起こりにくい。さらに「希代坊」「不請坊」については、次の例のように、同曲内で直前に「坊」の表記がなされており、「はう」表記であっても、誤読の可能性が低いと考えられる。

（1）きたひ坊にふしやう坊、ふしやう坊にきたい【は】う、／＼、／＼（名取川）

なお、語という単位に限らず、誤読を招きやすい文字列については濁点が付されやすい傾向も見られ、例えば濁点無表記の場合に「アフ」と誤読しやすいと推測される「アブ」を含む語をみてみると、「アブクマ（川）」（固有名詞）1 例は濁点無表記であるが、他の「危ない」18 例、「鐙」4 例、「炙る」3 例は全例で濁点が付されている。また同様に、「オビ」を含む語をみてみると、「帯」16 例、「オビクロウ」（固有名詞）1 例、「夥しい」42 例、「帯びる」2 例、「腰帯」3 例、「細帯」1 例で「び」に濁点が付されている。

○出現頻度 1 の語について

他方、誤読を避けるという観点で言えば、出現頻度の低い語については、濁点を付す傾向にあると推測される。そこで、出現頻度 1 の語（短単位）について、濁点が表記されているか否かを調査したところ、濁点が表記されているものが 1172 語、濁点無表記のものが 97 語であった。これらを合計すると、濁音で読むと推定される出現頻度 1 の語は 1269 語ということになるが、このうち、9 割以上にあたる語で濁点が表記されていることになる。また、全例で濁点が表記されている語は、異なりで 2248 語あるが（表 1）、出現頻度 1 の語がそのうちの 52.1%を占めていることになる。一方、濁点無表記の語は、異なりで 574 語あるが、出現頻度 1 の語は、そのうちの約 17%となっている。

なお、濁点無表記の 97 語のうち 24 語は、同一の形態素を使った語の用例があるため、純粋に出現頻度 1 とは言い難い語である。例えば、出現頻度 1 である「梅壺」「伏し沈む」の語については、それぞれ「壺」「沈む」の用例が他箇所にある。これらの語を、出現頻度 1 の語から除外すると、出現頻度 1 の語の濁点無表記率はさらに低くなる。このように、出現頻度の低い語では、濁点が付されることが多いようである。

４．まとめ

『虎明本狂言集』において、全例で濁点が付されている語を中心に、濁点の付与状況を調査した。全体として、濁点無表記例のある語より、全例で濁点が付されている語の方が多い。また、全例で濁点が付されている語については、『虎明本狂言集』全体の語種比率と比べて和語の比率が低く、漢語の比率が高い。これは、和語には使用頻度が大きく濁点無表記率が高い特定の助詞・助動詞が含まれることが大きい。さらに、表記用例数が多くとも誤読の可能性のある語については全例で濁点が付されている、狂言全体で用例が 1 例のみの語については、9 割以上の語で濁点が表記されている等、誤読を避けるために清濁の区別を明確に示す表記が行われていたと考えられる。

付　記

本研究は、国立国語研究所共同研究プロジェクト「通時コーパスの設計」による成果の一部である。

資料・文献

大塚光信編（2006）『大蔵虎明能狂言集　翻刻　註解』上・下　清文堂出版
土井忠生・森田武・長南実編訳（1980）『邦訳日葡辞書』岩波書店
日本国語大辞典　ジャパンナレッジ Lib　http://japanknowledge.com/library/
室町時代語辞典編修委員会編（1989、1991、1994、2000、2001）『時代別国語大辞典　室町時代編』一～五　三省堂
市村太郎・河瀬彰宏・小木曽智信（2012）「近世口語テキストの構造化とその課題」『情報処理学会研究報告. 人文科学とコンピュータ研究会報告』CH96（1）
市村太郎（2014）「近世口語資料のコーパス化―狂言・洒落本のコーパス化の過程と課題―」『日本語学』33-14、pp.96-109
出雲朝子（1976）「玉塵抄の濁音表記について」『國語學』104
岡照晃・小町守・小木曽智信・松本裕治（2013）「統計的機械学習を用いた歴史的資料への濁点付与の自動化」『情報処理学会論文誌』54-4
近藤明日子（2005）「濁点文字使用率から見る濁音表記」国立国語研究所編『国立国語研究所報告 122　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
沼本克明（1997）『日本漢字音の歴史的研究―體系と表記をめぐって―』汲古書院
松本宙（1978）「表記論覚え書き・4　清濁の書きわけと音韻史の記述」『弘前学院大学国語国文学会　学会誌』4
渡辺由貴・市村太郎（2014）「『虎明本狂言集』における濁点無表記箇所について―コーパス整備の過程から―」　日本語学会 2014 年度秋季大会発表予稿集

関連 URL

国立国語研究所コーパス開発センター（市村太郎・渡辺由貴ほか）編（2015）『日本語歴史コーパス　室町時代編 I　狂言』（短単位データ 0.9、中納言バージョン 1.5）
https://maro.ninjal.ac.jp
「日本語歴史コーパス」語彙統計
https://maro.ninjal.ac.jp/wiki/index.php?CHJ%2F%E8%AA%9E%E5%BD%99%E7%B5%B1%E8%A8%88

『今昔物語集』のコーパス化における非コアデータの精度向上作業

池上　尚[†]・鴻野知暁・河瀬彰宏・片山久留美（国立国語研究所コーパス開発センター）

Morphological Analysis for the Konjaku-Monogatarishū Corpus Non-core data

Nao Ikegami　Tomoaki Kouno　Akihiro Kawase　Kurumi Katayama
(National Institute for Japanese Language and Linguistics)

要旨

『今昔物語集』のコーパス化における形態論情報の付与作業、特に非コアデータに対する精度向上作業の方針を示した。発表者らは、まず、コアデータとして 5 つの巻を選定し、これについては「中古和文 UniDic」による形態素解析の結果すべてに目を通し人手修正を加えた。残る非コアデータについては、はじめに、コアデータを学習用データとして作成した「和漢混淆文 UniDic」を用いて形態素解析を行い、約 94%の精度を得た。次に、非コアデータのサンプリングチェックによる誤解析結果から、コーパス公開までの短期間で精度を効果的に向上させる方針を打ち出した。すなわち、「漢字一字表記、かつ、活用語尾（一部）非明示の用言」、「助動詞の前接用言」、「欠字欠文・破損の前後」などのチェックである。上記の作業により精度は約 99%まで向上している。

１．はじめに

国立国語研究所コーパス開発センターでは、共同研究プロジェクト「通時コーパスの設計」と連携し、『日本語歴史コーパス』（Corpus of Historical Japanese, CHJ）[1]の開発を進めている。江戸時代以前の口語性の強い資料群から優先してコーパス化を進め、2014 年 3 月には中古和文 14 作品を収録した平安時代編、2015 年 3 月には『虎明本狂言集』を収録した室町時代編 I 狂言を公開してきた。

一方で、日本語史研究において重要な文語性の強い資料群のコーパス化にも着手しており、現在、和漢混淆文資料を中心に収録した鎌倉時代編 I （説話・随筆など）の構築を進めている。中でも、このコーパスに収録予定の『今昔物語集』[2]は規模が大きく、技術的な問題点を多くはらむため、形態素解析を施す研究に特に注力してきた（冨士池・田中 2012、冨士池ほか 2013 など）。本発表では、これまでの研究を踏まえた上で、『今昔物語集』のコーパス化の全体的な方針と作業の過程を示す。そして、形態論情報の付与作業、特に非コアデータに対する精度向上作業の方針と進捗について報告する。

２．『日本語歴史コーパス』の資料選定方針

２．１　代表性の担保

『日本語歴史コーパス』においてコーパス化の対象とする主な資料群は、日本語史研究において重要な位置を占めてきた文学作品である。『日本語歴史コーパス』の嚆矢となった

[†] nikegami@ninjal.ac.jp

[1] http://www.ninjal.ac.jp/corpus_center/chj/

[2] 平安時代末成立とされるが、『今昔物語集』から始まる説話の一群が鎌倉時代に集中するため、便宜的に鎌倉時代編に収録する。

平安時代編も、「日本語史研究の源流となった、藤原定家や本居宣長などに始まる古典学の主たる対象になってきた作品群がその中心をなしており、古典のコーパス化の対象として最初に取り組むのに妥当なもの」（田中 2014）として選定された中古和文 14 作品の全文がコーパス化されている。平安時代編収録の作品とその語数（短単位）[3]をまとめた表 1 から分かるように、ジャンルは歌集・作り物語・歌物語・日記・随筆にわたり、約 74 万語（短単位）規模のコーパスである[4]。

表 1　平安時代編の作品・語数

ジャンル	作品名	語数
歌集	古今和歌集	31,288
作り物語	竹取物語	10,317
歌物語	伊勢物語	13,824
歌物語	大和物語	23,090
歌物語	平中物語	12,403
日記	土佐日記	6,685
作り物語	落窪物語	54,583
作り物語	堤中納言物語	15,699
随筆	枕草子	66,044
作り物語	源氏物語	445,675
日記	和泉式部日記	10,891
日記	紫式部日記	17,440
日記	更級日記	14,659
日記	讃岐典侍日記	15,555
計		738,153

２．２　鎌倉時代編の構築

平安時代編に後続する鎌倉時代編の収録作品候補としては、和漢混淆文資料として重要な軍記・説話・随筆が挙げられる（田中 2014）。そこで、まずは鎌倉時代編 I として説話・随筆のコーパスの作成に着手し、2016 年 3 月の公開を目指して現在作業中である。このコーパスが鎌倉時代の説話・随筆の実態の縮図となり得るよう、収録作品は当代の代表的な説話・随筆 5 作品とした。すなわち、説話は『今昔物語集』（1120 頃か）本朝部[5]、『宇治拾遺物語』（1220）、『十訓抄』（1252）の 3 作品、随筆は『方丈記』（1212）、『徒然草』（1336）の 2 作品である。表 2 は、上記の作品の語数（短単位）[6]をまとめたものである。全体で約 71 万語（短単位）となり、規模としては平安時代編とほぼ同等となる。

ただし、表 2 の語数から明らかなように、『今昔物語集』（本朝部）が量的に大きな割合を占めている。文学作品の場合、一作品の全文をコーパス化することが前提であり[7]、『今昔

[3] 空白・記号・補助記号は含まない。語（短単位）の認定基準については小椋・須永(2012)を参照。

[4] 2016 年 3 月には『蜻蛉日記』『大鏡』の 2 作品を追加する予定である。

[5] 天竺部・震旦部を含まない理由については 3 節を参照。

[6] 空白・記号・補助記号は含まない。語（短単位）の認定基準については小椋・須永(2012)に従うが、鎌倉時代編収録の作品に適用するにあたり一部変更したところがある。

[7] 文学作品をコーパス化する場合、一ジャンルから一部の作品を収めるという意味でのサンプリングはあっても、作品の一部を収めるという意味でのサンプリングは望ましくなく、一作品の全文をコーパス化する必要がある（近藤 2014）。

物語集』（本朝部）のように規模の大きな作品であってもそれに変わりはない。しかしながら、限られた時間・人手の中にあっては、コーパス総語数の約 70%を占めるような一作品の全文をコーパス化することに専心するよりも、それ以外の複数の説話作品を収めるコーパスへと拡張していく方が、『日本語歴史コーパス』としての代表性は担保されよう。そこで、発表者らは、『今昔物語集』（本朝部）の全文コーパス化・公開を目標とした上で、巻ごとにコアデータ・非コアデータの区別（3 節）を設け、それぞれ異なる作業方法により形態論情報の付与を行うことにした（4 節）。

表 2　鎌倉時代編 I の作品・語数

ジャンル	作品名	語数
説話	今昔物語集（本朝部）	499,712
説話	宇治拾遺物語	101,250
説話	十訓抄	73,514
随筆	方丈記	4,605
随筆	徒然草	33,767
計		712,848

『今昔物語集』は全 31 巻（うち巻 8・18・21 は欠巻のため、現存するのは 28 巻）、1000 話あまりの説話から構成され、一つ一つの説話は基本的に「今昔」という書き出しに始まり「トナム語リ伝ヘタルトヤ」と結んで終わる形式をとる。つまり、一話完結の説話を集めた説話集である。一話一話、一巻一巻の繋がりが希薄である一話完結の説話集だからこそ、作品の一部分をコアデータとして選定することが可能になるという側面もある。

３.『今昔物語集』（本朝部）におけるコアデータ・非コアデータ

コーパス化の対象とする『今昔物語集』の本文は、小学館の「新編日本古典文学全集」の『今昔物語集 1～4』（馬淵和夫・国東文麿・稲垣泰一校注）により、コーパス構築のために小学館から国立国語研究所に提供された電子テキストを利用している。『今昔物語集 1～4』には巻 1～10 の天竺部・震旦部は収録されておらず、巻 11～31 の本朝部のみが収録されている。よって、コーパス化の対象もこの範囲となる。底本は、巻 12・17・27・29 が『今昔物語集』最古の写本である鈴鹿本（現在は京都大学図書館蔵）、巻 11・13～16・19・20・22・24 は実践女子大学本、巻 23・25・26・28・30・31 は東京大学国語研究室本である。

このうち、まず、鈴鹿本を底本とする巻 12・17・27・29 をコアデータに選定した。『今昔物語集』は、最初の方の巻は漢文訓読体としての性格が強く、後ろの巻に進むにつれ和文体としての性格が強まるという性質を有し、その境は巻 20 前後と言われている[8]。よって、上記 4 巻は、漢文訓読体の性格が強い 2 巻（巻 12・17）、和文体の性格が強い 2 巻（巻 27・29）ということになる。この 4 巻に、文体から見た場合に中間的な巻となる巻 20 を加え、計 5 巻（本朝部の約 30.0%・約 15 万短単位）をコアデータとした。コアデータである 5 巻を除いた残りの 14 巻（本朝部の約 70.0%・約 35 万短単位）が非コアデータとなる。

[8] 佐藤(1984)の序章に研究史が詳細にまとめられている。

４.『今昔物語集』（本朝部）のデータ整備

前述のコアデータ・非コアデータの区別を踏まえた上で、以下、『今昔物語集』（本朝部）のデータ整備の手順(1)～(7)について詳述する。はじめに概要を示し、次に詳細を述べる。

(1) テキスト整形　……　全データ
(2) 「中古和文 UniDic」による全文の形態素解析　……　〃
(3) コアデータの整備　……　コアデータ
(4) 「和漢混淆文 UniDic」による非コアデータの形態素解析　……非コアデータ
(5) サンプリングチェック　……　〃
(6) 非コアデータの精度向上作業　……　〃
(7) 現在の精度　……　〃

(1) テキスト整形

冨士池ほか(2013)で述べたように、漢字片仮名交じりの和漢混淆文である『今昔物語集』のテキストは、形態素解析を施す前処理としてテキストを整形する必要があった[9]。その理由として、第一に、和漢混淆文ゆえに語順の転換、形態素の重複、形態素の不足があり、上から順に文字と形態素との対応がとれないテキストであったこと、第二に、「中古和文 UniDic」では非対応であった片仮名活用語尾・万葉仮名を含んでいたことが挙げられる。以下、データ整備の手順(5)・(6)に関わるものを中心に具体例をいくつか紹介する。

まず、語順の転換、形態素の重複が問題となる①返読文字がある[10]。返読文字とは、「不」「令」といった助詞・助動詞・接尾辞等と意味が対応する漢文の助辞に当たるものを指す。代表的な処理例として、「不知ズ→知ズ」（シラズ）のように返読文字を除外するタイプ、「不知リ→知**ザ**リ」（シラザリ）、「不知→知**ヌ**」（シラヌ）のように返読文字を除外し対応する語（の一部）を挿入するタイプなどがあった（□は返読文字、**太文字**は挿入箇所）。

次に、形態素の不足が問題となる②助詞・助動詞等の省略表記がある。これについては、「今昔（いまはむかし）→今**ハ**昔」「此二（このふたり）→此**ノ**二**人**」のようにルビに基づき補読処理を施した（**太文字**は挿入箇所）。ただし、「畢（をはり）テ」のように活用語尾が非明示のものについては、語彙素「終わる」―語形「オワル」―書字形「畢る」の連用形として「畢」が登録されていれば UniDic でも対応が可能なため、補読処理の対象としなかった。

同じく形態素の不足が問題となるものに、空格で示される④欠字欠文・破損がある。これは、「破損による欠字」「意識的欠字」を指す。後者には、「綿厚ク␣タル」のように、漢字で表記することを意図しながらもその表記を保留した欠字や、「磐田ノ郡、⇦␣⇨ノ郡ニ」のように固有名などの具体表記を保留した欠字がある。

テキスト整形が必要だったもののうち、形態素の不足については平安時代編を構築していた段階では特に問題とならず、『今昔物語集』のコーパス化に着手して初めて直面した課題であった。平安時代編のコーパス化の対象となった「新編日本古典文学全集」所収の中

[9] テキスト整形前の原文の状態は XML タグに記録してある。

[10] 『今昔物語集』の返読文字の詳細は冨士池・田中(2012)を参照。なお、本文中の丸数字①・②・④は冨士池ほか(2013)をそのまま引用する。

古和文 14 作品においては、読解の便をはかり、送り仮名などを適宜補入するという校訂方針がとられていたためである[11]。

(2) 「中古和文 UniDic」による全文の形態素解析

(1)の整形を経たテキストに対し「中古和文 UniDic」を用いて自動形態素解析を施した（解析器：MeCab 0.993）。

(3) コアデータの整備

(2)の解析結果のうち、コアデータとして選定した 5 巻について目視で確認し、誤解析の修正や揺れの統一、未知語の辞書登録を手作業で行い、短単位データを整備した。

(4) 「和漢混淆文 UniDic」による非コアデータの形態素解析

(3)の人手修正が完了したコアデータを学習用コーパスとして利用し、和漢混淆文を対象とした辞書「和漢混淆文 UniDic」を作成した[12]。さらに、この「和漢混淆文 UniDic」を用いて、人手修正の入っていない非コアデータ 14 巻の再解析を行った（解析器：MeCab 0.993）。結果は次の表 3 に示す通りである[13]。

表 3　「和漢混淆文 UniDic」による『今昔物語集』（本朝部）非コアデータの解析精度

評価レベル	Level 1 単語境界	Level 2 品詞認定	Level 3 語彙素認定	Level 4 発音形認定
解析精度（F値）	0.9889	0.9585	0.9479	0.9449

(5) サンプリングチェック

35 万短単位の規模になる(4)の解析結果から、2000 語を無作為に抽出するサンプリングチェックを行い、誤解析の傾向を確認した。

(6) 非コアデータの精度向上作業

(5)で確認した誤解析の結果からその要因を検討し、コーパス公開までの短期間で精度を効果的に向上させる方針を打ち出した。以下、特に重点的に行った作業の内容を述べる。

a. 漢字一字表記、かつ、活用語尾（一部）非明示の用言

誤解析の中でも特に目立ったのが、漢字一字で表記され、活用語尾が（一部）明示されない用言の語彙素・発音形の誤りである。テキストにルビが振られていればそれを参考に語彙素・発音形を決定する[14]が、機械解析ではテキストのルビを参照しないため、正しい語彙素・発音形を認定できない可能性が高くなる。「新編日本古典文学全集」の『今昔物語集』

[11] 作品ごとの校訂方針については「新編日本古典文学全集」当該巻の「凡例」を参照。

[12] 今後公開する予定である。なお、コアデータ 5 巻は約 15 万短単位あり、学習用コーパスに必要な 5 万～10 万語という目安（小木曽 2014）をクリアしている。

[13] 解析精度は 4 つのレベルで評価される。すなわち、「単語境界」（単語の境界の正しさ）、「品詞認定」（「単語境界」＋単語の品詞・活用型・活用形の正しさ）、「語彙素認定」（「品詞認定」＋UniDic の見出し語である語彙素認定の正しさ）、「発音形認定」（「語彙素認定」＋読み方の正しさ）の 4 つである。

[14] 小椋・須永(2012)に従い、ルビよりも「中古基本読み」を優先する場合は、ルビと発音形は一致しない。

は校注者によって漢字表記語ほぼ全てにルビが振られており[15]、このルビを尊重しつつ語彙素・発音形を決定しようとすると、機械解析の結果とずれが生じやすい（表 4）。

表 4 “漢字一字表記、かつ、活用語尾（一部）非明示の用言”誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_14c_S037_令誦方広経知父成牛語第三十七	家の主悲で、牛の辺に寄て、藁の座を敷て云く、「牛、実の我が父に	在さ	ば、此の座に登り給へ」と。	ましま	オワサ	オワス	おわす	動詞-一般	文語四段-サ行	未然形-一般
2	38_今昔物語集 04_30c_S003_近江守娘通浄蔵大徳語第三	持来べき便も思ず。奇異き事かな」と思て、「今は此の事	止め	て、偏に行ひをせむ」と思けれども、尚愛欲の思ひに勝ずして、	とど	ヤメ	ヤメル	止める	動詞-一般	文語下二段-マ行	連用形-一般
3	35_今昔物語集 01_13c_S042_六波羅僧講仙聞説法花得益語第四十二	愛執の過に依て、小蛇の身を受て、彼の木の下に住す。	願く	は、我が為に法花経を書写供養じて、此の苦を抜て	ねがは	ネガワシク	ネガワシイ	願わしい	形容詞-一般	文語形容詞-シク	連用形-一般
4	35_今昔物語集 01_11c_S015_聖武天皇始造元興寺語第十五	「東西二町に外閣を廻す事は、菩提涅槃の二果を證ずる相を	表す	。南北四町なる事は、生老病死の四苦を離れむ事を表す。	あらは	ヒョース	ヒョウスル	表する	動詞-一般	文語サ行変格	終止形-一般
5	37_今昔物語集 03_26c_S008_飛弾国猿神止生贄語第八	衣は思に随て着す、食物は	無	物無く食すれば、有しにも似ず、引替たる様に太りたり。	なき	ム	ム	無	名詞-普通名詞-一般		

こうした誤解析は、テキストの校訂方針、和漢混淆文である『今昔物語集』本来の表記の在り方に加え、出来る限り原文を尊重するという(1)テキスト整形の方針も影響している。

(1)テキスト整形における①返読文字の処理では、返読文字を除外（し意味の対応する助動詞（の一部）を挿入）しても、動詞の活用語尾を送り仮名として補入しなかった（「不知ズ→知ズ」、「不知リ→知**ザ**リ」など）。その結果、動詞の活用語尾が正しく解析されず、誤解析に繋がりやすくなった。

これと同様のことが、(1)テキスト整形における②助詞・助動詞等の省略表記に対する処理についても指摘できる。用言の活用語尾が非明示の場合は、UniDic に登録された活用形によって対応可能であると考え、ルビに基づく補読処理を施さなかった（「畢（をはり）テ」など）。しかし、実際には、非コアデータを扱う中で初めて出現したもの（新たに活用形として登録すべきもの）も多く、それらが結果として誤解析に繋がった。

発表者らは、まず、誤解析の大きな割合を占める“漢字一字表記、かつ、活用語尾（一部）非明示の用言”について、集中的に修正作業を行うことにした。そのためには、誤解析の可能性をもつ“漢字一字表記、かつ、活用語尾（一部）非明示の用言”の全例を洗い出す必要がある。そこで、非コアデータ中、ルビと発音形が不一致となっているキーに着目し、【ルビ 1 文字目と発音形 1 文字目が一致しないもの】、【ルビ 1 文字目と発音形 1 文字目は一致するが、ルビ 2 文字目と発音形 2 文字目が一致しないもの】の 2 パターンのリスト[16]を作成した上で、特に頻度の高いものから修正を施していった。表 5 には、活用語尾が明示されない漢字一字表記のもの[17]の中で、頻度・修正率ともに高かったものを示す。

別語彙素でありながら同一表記となりうるものが誤解析を起こしやすいのは、容易に想像がつく。表 5 で言えば、6「焼（ヤケル）」—9「焼（タク）」、17「行（オコナウ）」—22「行（アリク）」などである。このタイプには、7「畢（オエル）」—19「畢（オワル）」、29

[15] ルビは、「もし当時、仮名で書くとしたならばこう書いたであろうと校訂者が再構した仮名づかいで付してある（ただし、これには「平安仮名づかい」［発表者注：いわゆる「古典仮名づかい」とは違う、平安時代に行われた仮名づかい］は採用しなかった）。いわば校訂者の試論ともいうべきものである。」「新編日本古典文学全集」『今昔物語集 1』凡例

[16] ルビが歴史的仮名遣い、発音形が現代仮名遣いであることからリストに挙がってくるキーも多く（「可咲（をかし）」など）、目視での確認が必要であった。また、このリストは全ての品詞を対象とし作成したため、これを基に用言以外の修正も行っている。

[17] 活用語尾が（一部）明示される場合もあるため、語彙素自体の頻度とは必ずしも一致しない。

「下（クダス）」—30「下（クダル）」のように、動詞の自他で別語彙素となるものも含まれる。また、28「来（キタル）」のような漢文訓読体に特徴的な語が頻出する一方で、和文体に特徴的な「来（クル）」も使用されるため、類義語で文体差のある語彙素の対にも注意して修正作業を進める必要がある。

活用形ごとに見てみると、未然形・連用形の修正件数が多い。これには、その活用形自体の頻度が高いことに加え、未然形・連用形接続の助動詞の頻度が高い（後述）ことも関係していよう。漢字一字表記用言の発音形と関連する活用形については、次に述べる「助動詞の前接用言」の処理によって正しく修正されたものも多いことを補足しておく。

表 5　“漢字一字表記、かつ、活用語尾非明示の用言”修正例

№	表記	語彙素読み	頻度	誤解析	修正率	活用形別修正件数					
						未然形	連用形	終止形	連体形	已然形	命令形
1	開	ヒラク	84	84	100.0	3	79	2	0	0	0
2	咲	ワラウ	66	66	100.0	11	51	1	3	0	0
3	寄	ヨセル	41	41	100.0	8	32	1	0	0	0
4	合	アワセル	38	38	100.0	6	30	2	0	0	0
5	生	ウマレル	31	31	100.0	1	19	11	0	0	0
6	焼	ヤケル	22	22	100.0	1	21	0	0	0	0
7	畢	オエル	14	14	100.0	3	11	0	0	0	0
8	遣	オコセル	13	13	100.0	4	9	0	0	0	0
9	焼	タク	11	11	100.0	1	10	0	0	0	0
10	聞	キコエル	10	10	100.0	3	6	1	0	0	0
11	勝	スグレル	10	10	100.0	0	10	0	0	0	0
12	小	チイサイ	31	30	96.8	0	0	0	30	0	0
13	通	カヨウ	21	20	95.2	2	15	2	1	0	0
14	下	オロス	14	13	92.9	8	4	1	0	0	0
15	上	アガル	41	37	90.2	0	35	2	0	0	0
16	御	オワシマス	17	15	88.2	1	14	0	0	0	0
17	行	オコナウ	67	58	86.6	20	29	4	5	0	0
18	生	イキル	88	76	86.4	0	67	9	0	0	0
19	畢	オワル	42	36	85.7	2	34	0	0	0	0
20	遣	ツカワス	30	25	83.3	3	19	2	1	0	0
21	出	イダス	51	38	74.5	15	22	0	1	0	0
22	行	アリク	21	15	71.4	1	12	0	2	0	0
23	替	カワル	26	18	69.2	7	11	0	0	0	0
24	悪	アシイ	27	18	66.7	2	0	0	16	0	0
25	見	ミエル	82	53	64.6	34	18	1	0	0	0
26	入	イレル	157	100	63.7	15	84	1	0	0	0
27	立	タテル	138	80	58.0	9	69	2	0	0	0
28	来	キタル	466	265	56.9	33	215	2	9	1	5
29	下	クダス	22	12	54.5	9	2	1	0	0	0
30	下	クダル	103	54	52.4	4	47	2	1	0	0

b. 助動詞の前接用言

非コアデータに出現する助動詞のうち、用言を前接するものを抽出し、前接語の活用形や発音形について確認した。対象となったのは以下の助動詞である（語彙素で示す）。併せて、接続する活用形ごとのおよその頻度、括弧内には前接用言の修正件数を示した。

未然形接続：れる・られる・せる・させる・しむ・ず・じ・む・むず・まし・まほし
……約 8500(1730)
連用形接続：き・けり・つ・ぬ・たり（完了）・たし・けむ　……約 17000(1692)
終止形接続：べし・まじ・らむ・めり・なり　……約 1500(425)
連体形接続：なり（断定）　……約 8000(216)
命令形接続：り　……約 800(57)

また、助動詞として抽出されたキーそれ自体が正しい語彙素・活用形であるかについても確認している。特に、次のような、全体で 1 短単位とすべき他動詞「輝かす」「動かす」が「輝か｜す」「動か｜す」のように分割されていないか確認した（表 6）。

表 6　1 短単位とする他動詞例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_11c_S004_道照和尚亘唐伝法相還来語第四	|其の後夜に至て、其の光房より出て寺の庭の樹を	曜かす	。|久く有て、光西を指て飛び行ぬ。	かかや	カカヤカス	カガヤカス	輝かす	動詞-一般	文語四段-サ行	終止形-一般
2	35_今昔物語集 01_14c_S009_美作国鐵堀入穴依法花力出穴語第九	|底の人此れを引て	動す	。|然れば、|「人の有る也けり」|と知て、忽に葛を以て籠を造て、	うごか	ウゴカス	ウゴカス	動かす	動詞-一般	文語四段-サ行	終止形-一般

c. 欠字欠文・破損の前後

(1)テキスト整形で述べたように、『今昔物語集』に見られる欠字欠文・破損は空格を示す記号「␣」「⇦␣⇨」で置き換えている。これらの前後の文字列は誤解析が生じやすい（表 7）。

表 7 欠字欠文・破損前後の誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	35_今昔物語集 01_13c_S038_盗人誦法花四要品免難語第三十八	二つの手をば、上に大なる木を渡して、其れを␣	か	せて縛り付けつ。		カ	カ	か	助詞-係助詞		
2	36_今昔物語集 02_19c_S018_三条大皇大后宮出家語第十八	簾の内の女房␣て泣事糸⇦␣⇨	し	。|挟み畢奉て、聖人居去かむと為る時に、聖人音を高くして云く、		シ	スル	為る	動詞-非自立可能	文語サ行変格	連用形-一般
3	35_今昔物語集 01_13c_S015_東大寺僧仁鏡誘誦法花語第十五	|或時には、夢の中に白象来て随ひ␣	ふ	。|「此れ定て普賢文殊の護り給ふ也」|と知ぬ。		フ	フ	符	名詞-普通名詞-一般		
4	36_今昔物語集 02_16c_S038_紀伊国人邪見不信蒙現罰語第三十八	て、大きに嘖て、即ち、往きて妻を喚ぶ彼の導師此れを見て、慈の心を発して教へて	導	␣す。|而るに、夫此れを⇦␣⇨|「汝は此れ我が妻を婚むと為る盗人の法師也。|速に、	だう	ドー	ドウ	ドウ	名詞-固有名詞-人名-一般		

例 1 は「␣か」で 1 語の動詞・未然形、例 2 は「⇦␣⇨し」で 1 語の形容詞・終止形、例 3 は「␣ふ」で 1 語の動詞・終止形とそれぞれ推測される。例 4 は「導␣す」のどこで短単位が切れるのか不明である。例 1・2 は意識的欠字（漢字表記保留）に後続する文字列、例 3･4 は破損の前後に位置する文字列であったために誤解析となった例である。このように、語の一部が「␣」「⇦␣⇨」となっているとほぼ誤解析になる。もちろん、語がそのまま欠字欠文・破損である場合も、その前後では誤解析の生じる場合がある。

欠字欠文・破損は計 705 箇所（欠字・欠文：479 箇所、破損 226 箇所）あり、これらについては空格を表す記号「␣」「⇦␣⇨」を抽出した上で、その前後の修正を行った。例えば、例 1「␣か」・例 2「⇦␣⇨し」・例 3「␣ふ」であれば、空格直後の「か」「し」「ふ」にそれぞれ「解釈不明」という品詞を付与した。例 4「導␣す」であれば、空格前後の「導」「す」にそれぞれ「解釈不明」という品詞を付与した。

d. 題

一つ一つの説話冒頭には、その説話の題と当該巻中で第何話にあたるかが示されている。コアデータではこの「題＋第〇」のまとまりに対して、人手で「題」という品詞を付与していった。そのため、「和漢混淆文 UniDic」を用いたとしても、非コアデータの「題＋第〇」部分は本文同様に解析されてしまい、誤解析となっていた（表 8）。計 477 箇所あるこれらは、コアデータと同様に人手で品詞を付与した。

表 8　題の誤解析例

No.	ファイル名	前文脈	キー	後文脈	ルビ	出現発音形	語彙素読み	語彙素	品詞	解析活用型	活用形
1	38_今昔物語集 04_31c_S029_蔵人式部拯貞高於殿上俄死語第二十九	蔵人式部	拯	貞高於殿上俄死語第二十九｜ 今は昔、円融院の天皇の御時に、	くらうどしきぶのじやうさだたかてんじやうにしてにはかにしぬることだいにじふく	スクイ	スクウ	救う	動詞-一般	文語四段-ハ行	連用形-一般
2	37_今昔物語集 03_24c_S056_播磨国郡司家女読和歌語第五十六	播磨国郡司家女読和歌語第	五十	六｜ 今は昔、高階の為家の朝臣の幡磨の守にて有ける時、指せる事無き侍有けり。	はりまのくにのぐんじのいへのをむなわかをよむことだいごじふろく	ゴジュー	ゴジュウ	五十	名詞-数詞		

(7) 現在の精度

(6)の精度向上作業を経て、2000 語のサンプリングチェックを再度行った。非コアデータの現在の精度は Level 4（発音形認定）で 99.1%まで上昇している。

５．おわりに

『今昔物語集』のコーパス化は、テキスト整形、コアデータ整備と「和漢混淆文 UniDic」の作成、非コアデータの精度向上作業の 3 つの柱からなる。本発表では、その 3 つ目の柱について、作業方針・作業内容を明らかにし、精度が約 94%から約 99%まで向上したという結果をもってその方針の妥当性を示した。『日本語歴史コーパス』鎌倉時代編 I には、コアデータに準ずる精度となった非コアデータも含め、『今昔物語集』（本朝部）全文の収録を予定している。

また、『今昔物語集』非コアデータの精度向上作業によって、今後のコーパス開発、『今昔物語集』研究に次のような展開が期待されよう。まず、コーパス開発においては、今回、特に注力した(6)a「漢字一字表記、かつ、活用語尾（一部）非明示の用言」の誤解析処理によって新たに辞書登録した活用形も多く、他の和漢混淆文資料のコーパス化におけるコスト軽減に繋がると期待される。研究面においては、(6)a で散見された“同一漢字表記でありながら別語彙素の語”に着目することで、語から表記、表記から語へと往還しながらの網羅的な調査が可能になる。これまでの先行研究では『今昔物語集』の用字法が一語一表記で安定しているとされてきたが、語によって表記の安定性が異なる点については慎重に検討する必要がある（田中 1988）。表記の安定性を考察するにあたっては、語から表記、表記から語へといった双方向の検索が瞬時に可能な『今昔物語集』コーパスにより、示唆的なデータが提供されるのではなかろうか。

付記

本発表は、国立国語研究所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／田中牧郎）の成果の一部である。

参考文献

小木曽智信(2014)「歴史コーパスにおける形態素解析と辞書整備」『日本語学』33:14, pp.83-95

小椋秀樹・須永哲矢(2012)『中古和文 UniDic 短単位規程集』科研費　基盤研究(C)「和文系資料を対象とした形態素解析辞書の開発」（課題番号 21520492）研究成果報告書 2（中古和文 UniDic HP からダウンロード可）

近藤泰弘(2014)「歴史コーパスとは何か」『日本語学』33:14, pp.6-15

佐藤武義(1984)『今昔物語集の語彙と語法』明治書院

田中牧郎(1988)「仮名交じり文 3『今昔物語集』」『漢字講座 5 古代の漢字とことば』明治書院

田中牧郎(2014)「『日本語歴史コーパス』の構築」『日本語学』33:14, pp.56-67

冨士池優美・岩崎瑠莉恵(2014)「『今昔物語集』の捨て仮名」『第 5 回コーパス日本語学ワークショップ予稿集』pp.261-270

冨士池優美・河瀬彰宏・野田高広・岩崎瑠莉恵(2013)「『今昔物語集』のテキスト整形」『第 4 回コーパス日本語学ワークショップ予稿集』pp.125-134

冨士池優美・田中牧郎(2012)「今昔物語集の返読文字について―形態素解析の前処理を通して―」『日本語学会 2012 年度春季大会予稿集』pp.223-228

関連 URL

「通時コーパスの設計」プロジェクト　http://historicalcorpus.jp/

『日本語歴史コーパス　平安時代編』　http://www.ninjal.ac.jp/corpus_center/chj/

「中古和文 UniDic」http://www2.ninjal.ac.jp/lrc/index.php?UniDic

「MeCab: Yet Another Part-of-Speech and Morphological Analyzer」http://code.google.com/p/mecab/

外来語における[ei]の表記のゆれ

小椋秀樹（立命館大学文学部）

Orthographic Variation of [ei] in Loanwords

Hideki Ogura (College of Letters, Ritsumeikan University)

要旨

本稿の目的は、原語で二重母音[ei]を含む外来語を取り上げ、その二重母音が長音として長音符号で表記されるか、連母音で表記されるかという表記のゆれの実態を明らかにすることである。

『現代日本語書き言葉均衡コーパス』の出版サブコーパスの書籍、雑誌、新聞、特定目的サブコーパスの知恵袋、ブログを資料とし、それぞれのサブコーパスで頻度 100 以上の語を対象に表記のゆれの実態を調査した。その結果、両サブコーパスとも長音符号による表記が約 9 割を占めること、表記のゆれにはレジスター差が見られること、長音符号による表記と連母音による表記とで意味・用法に違いの見られる語があることなどを明らかにした。

１．はじめに

本稿は、小椋(2013、2014)に続き、大規模コーパスを活用して外来語表記のゆれの実態を解明しようとするものである。

小椋(2013)は、『現代日本語書き言葉均衡コーパス』(以下、BCCWJ とする。)のコアデータ[1]を資料として、外来語表記にどのようなゆれがあるか見通しを立てようとしたものである。この調査では、外来語表記のゆれの割合には、レジスターによる差異が見られることを明らかにした上で、各レジスターにおいて、具体的にどのような外来語表記のゆれが見られるのかなどについても調査を行った。その結果、長音に関する表記のゆれが最も多く、全てのレジスターに見られることを明らかにした。

小椋(2013)で指摘した長音に関する表記のゆれには、大きく分けて二つの種類がある。一つは、語中・語末長音を長音符号で書くか省くかというゆれである。例えば、「コンピューターーコンピュータ」「マネージャー・マネージャーマネジャー」が挙げられる。もう一つは、長音符号で書くか連母音で書くかというゆれである。例えば、「プレーヤーープレイヤー」が挙げられる。

前者については、小椋(2014)で BCCWJ の出版 SC、特定目的 SC・知恵袋、同・ブログを資料として実態調査を行った。そこで本稿では、長音符号で書くか連母音で書くかというゆれを取り上げることとし、その中でも特に、原語で二重母音[ei]を含む外来語に着目する。

原語の二重母音[ei]をエ段長音として長音符号で書くか、連母音で書くかについては、外来語の表記の基準を考える際に問題となることが多い。この表記の問題は、そもそも原語の[ei]を、日本語の音韻体系に合わせて長音[e:]で取り入れるか、原語の発音に基づいて母音連続[ei]で取り入れるかという発音のゆれに起因するものである。

国語審議会は、1952 年に術語・表記合同部会の報告として「外来語の表記について」を

[1] BCCWJ の設計等については、前川(2008)、山崎(2011)を参照。

公表した。ここでは、原語の二重母音[ei]について、

なお、原語における二重母音「エイ」「オウ」は長音とみなす。

ショー(show)　メーデー(May Day)

〔例外〕　エイト(eight)　ペイント(paint)

とあり、長音として取り入れられているという立場から、表記の基準を示している。現在の外来語表記の基準である『外来語の表記』(1991 年、内閣告示第 2 号、同訓令第 1 号)でも、

3　長音は、原則として長音符号「ー」を用いて書く。

〔例〕　(前略)　ゲーム　ショー　テーブル　パーティー (以下略)

注 2　「エー」「オー」と書かず、「エイ」「オウ」と書くような慣用のある場合は、それによる。

〔例〕　エイト　ペイント　レイアウト　(以下略)

とあり、「外来語の表記について」の考え方が継承されている。

しかし近年、原音に基づいて連母音で書こうとする傾向が見られ、表記の基準を改定したものもある。例えば、読売新聞社(2011)では「メーンイベント」(main event)と、『外来語の表記』の原則に基づき長音符号で書き表していたのを、読売新聞社(2014)では「メインイベント」と、連母音による表記に改めた。また、NHK 放送用語委員会における議論の概要をまとめた塩田(2006)によると、NHK では原語の二重母音[ei]について長音表記を本則としているが、近年、一般社会において連母音による表記が増えているため、この本則を再検討する必要があるとして検討事項に上がっている。

このような現代における外来語表記の問題を踏まえ、本稿では、原語で二重母音[ei]を持つ外来語を取り上げ、BCCWJ を資料として表記のゆれの実態を明らかにする。

以下、2 節で先行研究を概観した後、3 節で調査資料とするレジスター、調査対象とする語の範囲について述べる。4 節で調査結果を報告し、最後に 5 節で本稿をまとめる。

なお、本稿では、語の表記を示す際には、「プレーヤー」のようにかぎ括弧を付けて示し、語を示す際には《プレーヤー》のように二重山括弧を付けて示す。また、長音符号による表記を長音表記と、連母音による表記を連母音表記と呼ぶ。

２．先行研究

ここでは、本稿の調査に関連する先行研究を見ていくこととする。まず実態調査に基づくものとして宮島、高木(1984)、佐竹(1986)、荻野(2014)を取り上げる。また、長音表記か連母音表記かという表記のゆれには、長音で発音しているか、連母音で発音しているかという語形のゆれの問題が関係する。そこで、外来語における[ei]の発音のバリエーションを調査した岡田(2004)を取り上げる。

宮島、高木(1984)は 1956 年発行の雑誌 90 種を対象とした外来語表記のゆれに関する調査報告である。佐竹(1986)は、当時、国の基準が示されていなかった外来語の表記の問題点について、国立国語研究所(1983)[2]を手がかりにしながら述べたものである。宮島・高木(1984:55)は、「2 重母音という意識があるとき」に連母音表記が取られるとし、佐竹(1986:417)は、長音表記ではなく連母音表記が取られるのは「長音でないという意識が強いことの証明であり」、「そのような意識が強いというならば、長音符号と母音表記との対立は、もはや長音表記のしかたのゆれではなく、発音のゆれの問題である」と述べる。

[2] 1966 年発行の朝日・毎日・読売 3 紙を対象とした語表記のゆれに関する調査である。

荻野(2014)は、Web をコーパスとして利用した研究で、《テークアウト》《クラスメート》など 20 語を対象に、外来語における[ei]が長音表記されるか、連母音表記されるか調査している。その結果、長音符号による表記が圧倒的に多いこと、長音表記か連母音表記かは語ごとに決まっており、同程度で表記がゆれている語は見られないこと、古い時代に日本語に入ってきた語は、長音表記される傾向にあることを述べる。また、「ネーム」と「ネイム」とを取り上げ、前者は会社名、商品名に使われることが多く、後者は全体的に曲名での使用が多いことを示し、長音表記と連母音表記とで意味・用法に差異のあることを明らかにしている。

次に、外来語における母音連続[ei]の発音に関する岡田(2004)を見ていく。岡田(2004)は、『日本語話し言葉コーパス』を資料として、原語で二重母音[ei]を持つ語が外来語として日本語に取り入れられる際に、二重母音を長音[e:]で取り入れるのか、母音連続[ei]で取り入れられるのかを調査したものである。その結果、[ei]で発話されるのは約 7%にとどまり、「長母音[e:]で実現される傾向を認めることができる」(p.37)と述べる。また、どのような場合に[ei]となるのかについても調査し、/ei/＋/N/という音節構造の場合に[ei]で実現される傾向にあることを明らかにしている。また、語のなじみ度も緩やかに関係している可能性があると指摘している。

以上、本稿に関連する先行研究を概観した。原語の[ei]について、発音の面では長音で実現される傾向にあり、表記の面では長音表記が圧倒的に多く、長音表記か連母音表記かは語ごとに決まっているという指摘は、重要なものである。ただ、荻野(2014)の調査対象は 20 語と少なく、Web を利用しているためレジスターによる差異の有無についても明らかにはされていない。宮島、高木(1984)、佐竹(1986)は、大規模言語調査に基づく研究ではあるが、いずれも単一のレジスターを対象としたものであり、そもそも現在から約 50 年～60 年前の言語調査を基にしているという問題もある。

このような研究の現状から、原語で二重母音[ei]を持つ外来語の表記については、多様なレジスターを資料にして、より現在に近い時期の実態を明らかにする必要がある。そこで本稿では、多様なレジスターを収録している BCCWJ から出版・書籍、同・雑誌、同・新聞、及び特定目的・知恵袋、同・ブログの各レジスターを資料として、外来語における[ei]の表記のゆれの実態を計量的な手法によって明らかにしていく。具体的には、外来語における[ei]の表記が長音表記か連母音表記かを調査し、レジスターによる差異を明らかにする。さらに、意味・用法の面からも表記のゆれの傾向を見ていくこととする。

３．調査資料・調査対象

３．１　調査資料

表記の問題を取り上げる際、注意しなければならないのは、表記の基準や校閲の存在である。1 節で述べたとおり、外来語の表記には、国が定めた基準である『外来語の表記』がある。この基準に従って表記の統一を図った場合、本稿で取り上げている外来語の[ei]という音については、長音表記で統一されることとなる。また、著者のほかに編集者等による校閲があれば、ゆれが抑制される可能性もある。

このような点を踏まえて、本稿では、BCCWJ に収録されたレジスターの中から、出版・書籍、同・雑誌、同・新聞と特定目的・知恵袋、同・ブログとを資料とすることとした。出版 SC の各レジスターは程度の差はあるものの、編集者の校閲が想定される。新聞については、『外来語の表記』を基に各社が表記の基準を設け、表記の統一を図っている。それに

対して、特定目的 SC の知恵袋、ブログ(以下、まとめて呼ぶ場合は Web とする。)は、どのような表記を取るかは著者の自由である。

BCCWJ は、言語単位として長単位と短単位の 2 種類を採用している[3]。今回の調査には、そのうち短単位を用いた。各レジスターの延べ語数を表 1 に示した(短単位の語数。記号、補助記号、空白は除く。)。

表 1：各レジスターの延べ語数

レジスター	延べ語数	レジスター	延べ語数
出版・書籍	28,552,283	特定目的・知恵袋	10,256,877
出版・雑誌	4,444,492	特定目的・ブログ	10,194,143
出版・新聞	1,370,233		

３．２　調査対象

本稿では、原語で[ei]という音を含む外来語から、次のように調査対象を絞り込んだ。

出版 SC と Web とでは、出現する語に違いが見られることが予想される。そこで、出版 SC と Web とを別々に集計した上で、それぞれで頻度 100 以上の語を対象とすることとした。ここで頻度 100 以上としたのは、語別に表記のゆれの状況を把握するため、偏りが生じやすい生起頻度の低い語は除くのが適切だと判断したことによる。また、固有名詞を除く一般語を対象とすることとした。

用例の収集に当たっては、短単位データ 1.0.0 を対象に、『中納言』1.1.0 で、語彙素に片仮名表記のエ段長音を含むもの(検索条件：%[エケセテネヘメレゲゼデベ]ー%)を検索した。検索結果を基に、頻度 100 以上の語(固有名詞を除く。)に絞り込んだ上で、更に原語で[ei]という音を含むものを抽出した。その結果、出版 SC では 101 語、Web では 71 語が対象となった。

４．調査結果

４．１　[ei]の表記のゆれ

本節では、原語における二重母音[ei]の表記の実態について、レジスター別に見ていく。

原語の二重母音[ei]について、長音表記、連母音表記がそれぞれどの程度用いられているのかを、表 2 にまとめた。表 2 では、長音符号による表記、連母音による表記の度数と、それぞれの表記が占める割合とを示した。

出版 SC 全体では、長音表記が 89.2%、連母音表記が 10.8%で、長音表記が圧倒的に多い。この傾向は、Web でも同様であり、長音表記が 90.9%、連母音表記が 9.1%となっている。原語における二重母音[ei]は、長音表記で定着しているといえる。岡田(2014)で明らかにされているとおり、話し言葉では原語の[ei]は長音で実現される傾向にある。長音表記が圧倒的に多いのは、話し言葉において長音が圧倒的に多いことによると考えられる。

レジスター別に見ても、長音表記が圧倒的に多いことに変わりはないが、若干の差異を認めることができる。連母音表記の割合を見ると、出版 SC では、雑誌が 13.7%で最も高く、次いで書籍が 10.0%である。一方、新聞は最も低く 5.6%にとどまる。特定目的 SC では、ブログが 11.6%で 1 割台であるが、知恵袋は 7.0%と低い。新聞において連母音表記の割合が

[3] BCCWJ における言語単位の概要、単位認定基準については、小椋、小磯、冨士池他(2011)を参照。

低いのは、『外来語の表記』に基づき長音表記で統一を図っていることによると考えられる。

表 2：外来語における[ei]の表記(延べ)

	長音	連母音	総計		長音	連母音	総計
出版	47283	5739	53022	Web	36975	3700	40675
	89.2%	10.8%	100.0%		90.9%	9.1%	100.0%
出版・書籍	33143	3667	36810	特定・知恵袋	20645	1553	22198
	90.0%	10.0%	100.0%		93.0%	7.0%	100.0%
出版・雑誌	12402	1969	14371	特定・ブログ	16330	2147	18477
	86.3%	13.7%	100.0%		88.4%	11.6%	100.0%
出版・新聞	1738	103	1841				
	94.4%	5.6%	100.0%				

語別に見た場合、ゆれの見られない語もあれば、長音表記、連母音表記のいずれかに偏る語や、二つの表記が同程度に用いられている語が見られる。そこで、ゆれの程度に応じた分類を試みることとする。まず、ゆれの見られない語を「固定」、一方の表記が 8 割以上を占めている語を「独占」、それ以外を「ゆれ」と呼ぶこととする[4]。それぞれの分類に属する語数(異なり)を出版 SC、Web ごとに集計したのが表 3 である。

表 3：「固定」「独占」「ゆれ」と語数(異なり)

	固定	独占	ゆれ	総計
出版	52(3)	38(6)	11	101
	51.5%	37.6%	10.9%	100.0%
Web	46(1)	22(6)	3	71
	64.8%	31.0%	4.2%	100.0%

「固定」「独占」の括弧内の数字は、連母音表記で固定している(連母音表記が 80%以上を占める)語の数である。出版 SC では「固定」に分類される 52 語のうち 3 語が連母音表記で固定している。

出版 SC、Web とも表記にゆれのみられない「固定」が最も多いことがわかる。出版 SC では 52 語(51.5%)、Web では 46 語(64.8%)といずれも過半数を占めている。「独占」が共に 3 割台で続いており、異なりで見た場合、9 割前後の語がほとんど表記にゆれが見られず、また長音表記が圧倒的に優勢であることが分かる。

「ゆれ」に分類される語、「独占」に分類される語のうち連母音表記に偏る語、「固定」に分類される語のうち連母音表記で固定している語を連母音表記の割合とともに示したのが、表 4 である。出版 SC では 20 語、Web では 10 語となっている。

表 4 を見ると、Web で「ゆれ」に分類される《デー》《プレーヤー》《プレー》の 3 語は、出版 SC でも「ゆれ」に分類されている。表記の基準や校閲の有無といったレジスターの性格にかかわらず、現代においてまさに表記のゆれている語といえる。

連母音表記で固定している語、及び連母音表記が 8 割を超える語は、出版 SC と Web と

[4] この 3 区分は、1956 年発行の雑誌 90 種を対象に、語表記のゆれを調査した宮島(1997)を参考にしたものである。ただし宮島(1997)は、「独占」を「特定の形式が 9 割以上をしめているもの」(p.103)としており、本稿と異なる。

で共通するものがある。《ディスプレー》《メーク》《ネール》《リメーク》《メーン》《メード》《ブレーク》の 7 語が挙げられる。今回の調査では頻度 100 以上の比較的高頻度の語を対象としていることも関係していると思われるが、専門用語というよりは一般語に属する語が多く見られる。これらは、現代において、『外来語の表記』の原則とは異なる表記で定着している語群ということになる。

表 4：「ゆれ」に分類される語、連母音表記が優勢である語

出版SC			Web		
語彙素	原語	連母音率	語彙素	原語	連母音率
プレー	play	40.1%	デー	day	28.8%
プレーヤー	player	43.2%	プレーヤー	player	43.6%
クラスメート	classmate	53.8%	プレー	play	50.8%
テーク	take	56.3%	ディスプレー	display	88.4%
メーク	make	57.6%	メーク	make	91.1%
デー	day	58.4%	ネール	nail	93.6%
ディスプレー	display	71.3%	リメーク	remake	94.1%
エッセー	essay	75.5%	メーン	main	98.4%
トレー	tray	75.7%	メード	made	98.9%
ウエート	weight	79.4%	ブレーク	break	100.0%
ハイウエー	highway	79.6%			
ウエートレス	waitress	87.6%			
テースト	taste	89.5%			
メーン	main	91.8%			
ネール	nail	96.7%			
ウエー	way	97.1%			
ネービー	navy	99.3%			
ネーティブ	native	100.0%			
ブレーク	break	100.0%			
メード	made	100.0%			

４．２　意味・用法と[ei]の表記

荻野(2014)では、長音表記と連母音表記とで意味・用法に差異のあることが指摘されている。本節では、この指摘を受け、出版 SC で「ゆれ」に属する語の中から、《ディスプレー》《メーク》の 2 語を取り上げ、意味・用法と表記との関係などについて検討する。なお、適宜、Web の調査結果と対照して見ていく。

（１）ディスプレー

《ディスプレー》は、[1] 展示すること、陳列すること、 [2]コンピューターの出力表示装置(モニター)という二つの語義を持つ。その例を次に示す。

(1) あんまり綺麗にディスプレイできないので(OC14_08488)

(2) コンピューターのディスプレイから目を離さずに(PB29_00337)

そこで、これらの語義と[ei]の表記との間に関係があるか否かを見ることとする。その結果を表 5 にまとめた。表 5 では、各語義における長音表記、連母音表記の頻度(割合)を示した。出版 SC だけではなく、Web も併せて示した。

表 5 を見ると、出版 SC、Web とも、どちらの意味においても連母音による表記の割合が高いことが分かる。しかし、「陳列・展示」の意味よりも「モニター」の意味の方が連母音

による表記が用いられる割合が高い。出版 SC では約 8 割が、Web では約 9 割が連母音による表記である。

両語義とも連母音表記の割合が高いが、特に「モニター」の意味で用いられた場合に、連母音表記となる傾向が強い。

表 5：《ディスプレー》の意味と表記

		長音		連母音		総計
出版	モニター	55	20.8%	210	79.2%	265
	展示・陳列	63	42.9%	84	57.1%	147
Web	モニター	6	9.8%	55	90.2%	61
	展示・陳列	9	27.3%	24	72.7%	33

（２）メーク

《メーク》は、出版 SC に 813 例用いられており、そのうち 763 例が美容・ファッション関係での用例であった。例えば、次のような例である。

(3) そんなわけでふだんはノーメークに近いのだとか。(PB4n_00148)

(4) 今年はちょっと大人っぽく見せるメイクがイチオシ。(PM21_00527)

その他の例は、《メークドラマ》《スコアメーク》《チャンスメーク》のような用法である。

美容・ファッション関係での用例を対象に長音表記、連母音表記の頻度(割合)を調査した結果を表 6 に示した。

表 6：《メーク》の表記(ファッション・美容関係)

	長音		連母音		総計
出版	332	43.5%	431	56.5%	763
Web	28	5.6%	472	94.4%	500

表 6 を見ると、出版 SC では、長音表記が 43.5%、連母音表記が 56.5%であり、連母音表記が優勢ではあるものの、その差は余り大きくない。まさに表記がゆれているといえる。なお、《メーク》は出版・新聞に 8 例(いずれも長音表記)しか出現しないので、出版・新聞の影響により、長音表記の頻度が高くなっているわけではない。

一方、Web では連母音表記が 94.4%を占めている。出版物ではゆれが生じているが、Web のような個人が自由に表記を選択できるレジスターでは連母音表記が定着していると考えられる。

５．終わりに

本稿では、BCCWJ の出版・書籍、同・雑誌、同・新聞と特定目的・知恵袋、同・ブログを資料として、原語で二重母音[ei]を含む外来語を対象に、[ei]が長音表記されるか連母音表記されるかについて実態調査を行った。その結果、次のことが明らかとなった。

(5)・[ei]の表記は、長音表記が圧倒的に多く、出版 SC、Web とも長音表記が約 9 割を占める。ただし、長音表記、連母音表記のゆれには、レジスター差も若干認められる。

・《ディスプレー》は、意味・用法によって連母音表記の割合に差がある。また美容・

ファッション関係で用いられる《メーク》は、表記のゆれにレジスター差がある。

本稿では、上に述べたように長音表記が圧倒的に多いという結果が得られたが、これには、調査対象を頻度 100 以上の語に限定したことが関わっている可能性も考えられる。つまり、既に一般語化しているため、原語の二重母音[ei]が日本語の音韻体系に合わせて長音として取り入れられ、長音符号による表記が取られているとも考えられるのである。佐竹(1986)には、最近使われ出した語に連母音表記が見られるという指摘がある。今後、低頻度も含めて[ei]の表記の実態を調査する必要がある。

謝　辞

本研究は、国立国語研究所共同研究プロジェクト(基幹型)「コーパス日本語学の創成」(リーダー：前川喜久雄)、同「多角的アプローチによる現代日本語の動態の解明」(リーダー：相澤正夫)、JSPS 科研費「大規模コーパスに基づく現代語表記のゆれの実態解明」(代表者：小椋秀樹)による補助を得た。

参考文献

岡田祥平(2004)「『日本語話し言葉コーパス』に観察される母音連続/ei/のバリエーション ―外来語の場合―」『電子情報通信学会技術研究報告〔音声〕』104-148、pp.35-40.

荻野綱男(2014)『ウェブ検索による日本語研究』、朝倉書店.

小椋秀樹(2013)「現代日本語における外来語表記のゆれ」相澤正夫(編)『現代日本語の動態研究』、おうふう、pp.151-171.

小椋秀樹(2014)「外来語語末長音の表記のゆれについて」『論究日本文学』100、pp.195-208.

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)』(国立国語研究所内部報告書 LR-CCG-10-05-01、LR-CCG-10-05-02).

佐竹秀雄(1986)「外来語表記法の問題点」宮地裕(編)『論集 日本語研究(1) 現代編』、明示書院、pp.407-422.

塩田雄大(2006)「外来語の発音とカタカナ表記 ～ [エイ・ケイ・セイ]などを中心に ～」『疱瘡研究と調査』56-3、pp.74-75.

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.82-95.

宮島達夫(1997)「雑誌九十種表記表の統計」、『日本語科学』1、pp.92-103.

宮島達夫、高木翠(1984)「雑誌九十種資料の外来語表記」『研究報告集』5(国立国語研究所報告 79)、pp.43-76.

山崎誠(2011)「第 2 章 『現代日本語書き言葉均衡コーパス』の設計」、国立国語研究所コーパス開発センター『『現代日本語書き言葉均衡コーパス』利用の手引き』第 1.0 版、pp.113-20.

読売新聞社(2011)『読売新聞用字用語の手引き　第 3 版』、中央公論新社.

読売新聞社(2014)『読売新聞用字用語の手引き　第 4 版』、中央公論新社.

関連 URL

「国語施策情報」　http://kokugo.bunka.go.jp/kokugo_nihongo/joho/index.html

ポスター発表（1）Bグループ

9月1日（火）　14:10～15:10

品詞列・係り受け部分木に基づくラベリングツールの設計と実装 –節境界ラベリングを例に–

浅原 正幸 (国立国語研究所) *
小西 光 (国立国語研究所)
田中 弥生 (神奈川大学・国立国語研究所)
加藤 祥 (国立国語研究所)

Design and Implementation of a Labeling Tool Based on Morpheme Subsequences and Dependency Subtrees – a Use Case in Clause Boundary Labeling –

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Hikari Konishi (National Institute for Japanese Language and Linguistics)
Yayoi Tanaka (Kanagawa University, National Institute for Japanese Language and Linguistics)
Sachi Kato (National Institute for Japanese Language and Linguistics)

要旨

コーパスに対する情報付与の方法として、悉皆的に人手で行うアノテーション以外に、パターンに基づくラベリング手法がある。後者のラベリング手法は、人手で行うアノテーションの前段階で用いられるほか、形態論情報や係り受け部分木に基づくパターンで記述可能な手がかり句がわかっている際に広く用いられる。しかしながら、パターンに基づくラベリングツールは様々な研究者により開発されているが、再利用性が乏しく、パターンを記述するために汎用的な記述方法が求められている。本研究では、コーパスの検索系で用いられるクエリ言語を基にした形態論情報や係り受け部分木に基づくラベリングツールを設計し、実装した。また、節境界ラベリングを例にした応用について紹介する。

1. はじめに

日本語のコーパス分析において、形態素解析器によって単語分かち書きと同時に形態論情報を付与した形態素解析結果を、人手で修正して分析するという手法が多く用いられてきた。係り受け解析器の精度が高くなるについてれ係り受け情報を用いた分析も少しずつ増えてきている。さらに、個々の研究者によってアノテーションされた統語的・意味論的情報をもとに分析する研究も行われている。

このアノテーションを行う際に、アノテーション対象が手がかり句 (cue phrase)[(1)]によって

* masayu-a@ninjal.ac.jp

(1) アノテーション対象の存在を示す語句。

ある程度表現可能な場合がある。品詞体系や係り受け構造が広く利用されているものであれば、形態論情報や係り受け部分木によって共有可能な手がかり句が表現され、手がかり句を再利用することが可能になる。(2)

ほとんどの手がかり句は各種検索系に基づく、検索クエリ相当の表現で記述することが可能である。そこで本稿では、各種検索系で可能なクエリとクエリ言語について整理し、JSON (JavaScript Object Notation) 形式での統一的なパターン記述言語を設計する。さらに、「鳥バンク」(池原 (2007)) で採用されている節境界認定基準を中心に、節境界ラベリングツールの試作を行う。

2. 形態論情報・係り受け部分木に基づく検索系

2.1 中納言

コーパス検索アプリケーション「中納言」(国立国語研究所 (2015a)) は、形態論情報に基づくコーパス検索が可能な Web アプリケーションである。現在のところ『現代日本語書き言葉均衡コーパス』(Maekawa et al. (2014)) や『日本語歴史コーパス』(国立国語研究所 (2015b)) が検索できるようになっている。

図 1 は、連体形が前置する手がかり句「かたわら、」（鳥バンク：「副詞句：二者関係：：対比：FUj001」）を中納言で検索した例である。

図 1　「中納言」による手がかり句「かたわら、」の検索フォーム

「中納言」では、クエリ言語として、SQL に似た記述言語を用いている。図 2 は上記検索フォームの内容を「中納言」のクエリ言語に変換したものである。

(2) 本稿では、規則やパターンに基づいて自動的に情報を付与する作業を「アノテーション」とは呼ばない立場をとり、ラベリングと呼ぶこととする。

```
キー: 書字形出現形 = "かたわら" AND 前方共起: 活用形 LIKE "連体形%" ON 1 WORDS % FROM キー
AND 後方共起: 書字形出現形 = "、" ON 1 WORDS FROM キー WITH OPTIONS unit="1"
AND tglBunKugiri="#" AND tglWords="20" AND limitToSelfSentence="1" AND tglKugiri="|"
AND endOfLine="CRLF" AND encoding="UTF-16LE" AND tglFixVariable="2"
```

図 2　「中納言」による手がかり句「かたわら、」の検索クエリ

2.2　ChaKi.NET Tag Search

コーパスコンコーダンサ「ChaKi.NET」(Matsumoto et al. (2006)) の Tag Search 機能でも、同様の検索が可能である。検索フォームは図 3 のようになる。

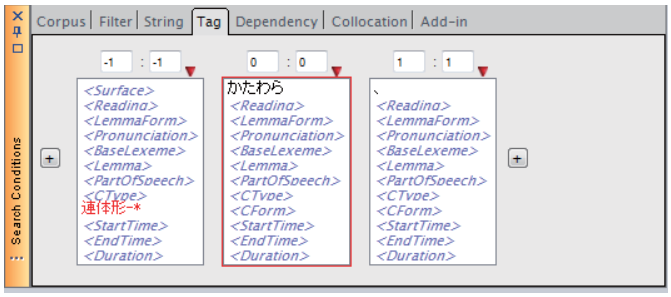


図 3　「ChaKi.NET」による手がかり句「かたわら、」の検索フォーム

この検索フォームは外部 XML ファイルに保存可能である。図 4 に Tag Search の検索条件<TagCond>の冒頭の形態素を指定している箇所を示す。

```
<TagCond><LexemeConds><LexemeCondition><PropertyPairs>
  <PropertyPair>
    <Key>CForm</Key>
      <Value xsi:type="CForm"><StrVal>連体形-*</StrVal><IsRegEx>true</IsRegEx>
      <IsCaseSensitive>true</IsCaseSensitive><ID>0</ID><Name>連体形-*</Name></Value>
  </PropertyPair>
 </PropertyPairs>
 <RelativePosition><Start>-1</Start><End>-1</End></RelativePosition>
  ...
```

図 4　「ChaKi.NET」による手がかり句「かたわら、」の検索クエリ

2.3　超大規模コーパス 検索系 (形態論情報検索)

図 5 は現在開発中の超大規模コーパス検索系のうち形態論情報を検索するための試作 UI である。「ChaKi.NET」の Tag Search を参考にした UI になっている。この UI は 3.2 節に示す JSON によるクエリ言語を発行することができる。

2.4　MREP

MREP [(3)] は MeCab の出力をベースとしたパターンマッチャーで、形態論情報に対して、品詞と表層文字列をアルファベットとする正規言語相当のパターン (図 6) にマッチすることがで

[(3)] http://www.slideshare.net/unnonouno/miura-dsirnlp6

図 5　超大規模コーパス検索系 (形態論情報検索) による手がかり句「かたわら、」の検索フォーム

きる。

- .
  任意の形態素にマッチ
- <pos=x>
  品詞が x の形態素にマッチ
- <surface=x>
  表記が x の形態素にマッチ
- X*
  X の 1 回以上 の繰り返しにマッチ
- X|Y
  X か Y にマッチ

図 6　MREP のクエリ言語仕様

2.5　ChaKi.NET Dependency Search

図 1,3,5 のフォームは主に形態素列に基づく検索フォームであった。コーパスコンコーダンサ「ChaKi.NET」の Dependency Search 機能では、係り受け部分木に基づく検索が可能である。

図 7 は、手がかり句「をいいことに」（鳥バンク：「副詞句：その他：慣用的表現：状況の悪利用：FUp202」）を ChaKi.NET Dependency Search で検索した例である。

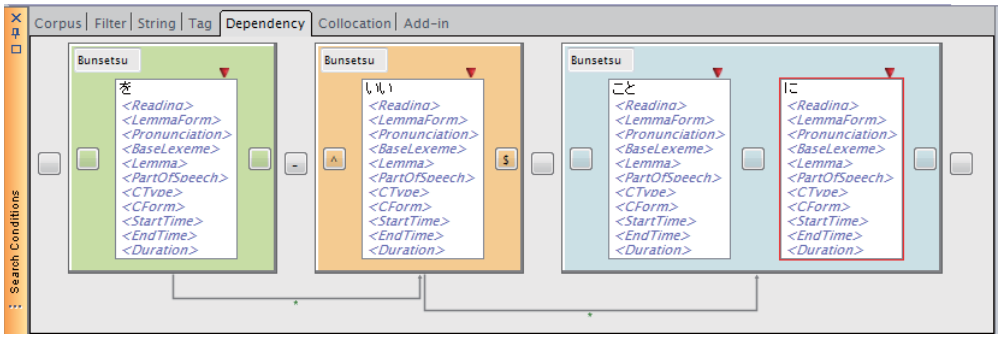


図 7　「ChaKi.NET」による手がかり句「をいいことに」の検索フォーム

この検索フォームも外部 XML ファイルに保存可能である。図 8 のように Tag Search の条件を文節内に格納するような形式で記述する。

```
<DepCond><BunsetsuConds><TagSearchCondition>
 <LexemeConds><LexemeCondition><PropertyPairs><PropertyPair>
  <Key>Surface</Key>
   <Value><StrVal>を</StrVal><IsRegEx>false</IsRegEx><IsCaseSensitive>true</IsCaseSensitive></Value>
   </PropertyPair></PropertyPairs>
   <RelativePosition><Start>0</Start><End>0</End></RelativePosition>
   <LeftConnection>32</LeftConnection><RightConnection>32</RightConnection><IsPivot>false</IsPivot>
 </LexemeCondition></LexemeConds>
 <LeftConnection>32</LeftConnection><RightConnection>45</RightConnection>
 <SegmentTag>Bunsetsu</SegmentTag>
 </TagSearchCondition>
  ...
```

図 8　「ChaKi.NET」による手がかり句「をいいことに」の検索クエリ

2.6　超大規模コーパス 検索系 (係り受け部分木検索)

図 9 は現在開発中の超大規模コーパス検索系のうち係り受け部分木を検索するための試作 UI である。「ChaKi.NET」の Dependency Search を参考にした UI になっている。この UI は 3.3 節に示す JSON によるクエリ言語を発行することができる。

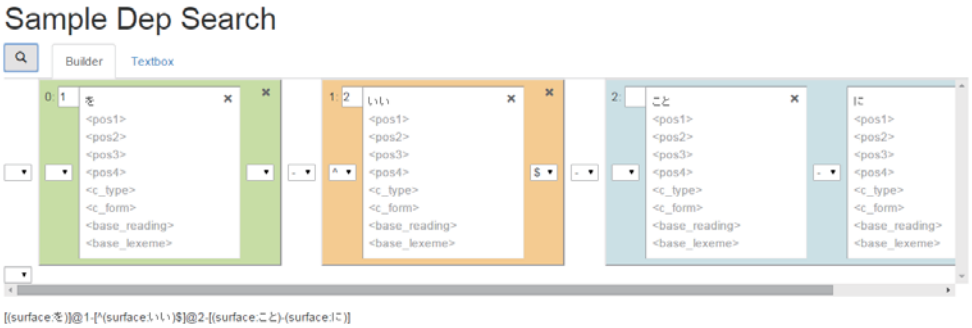


図 9　超大規模コーパス検索系 (係り受け部分木検索) による手がかり句「をいいことに」の検索フォーム

2.7　検索系のちがい

各検索系で細かな点で機能に違いがある。表 1 に各検索系機能のまとめを示す。各項目の意味は以下の通りである：

- 「形態論情報正規表現」: 形態論情報を指定する際に正規表現が指定できるか否か
- 「Ignore Case」: 形態論情報を指定する際に大文字・小文字の違いを無視できるか否か
- 「文頭・文末指定」: 形態素位置を文頭・文末からの相対位置で指定できるか否か
- 「文境界またぎ」: 検索クエリを文境界を越えて指定できるか否か
- 「語彙表・n-gram 出力」: 検索クエリに適合する結果の異なりを度数とともに出力できるか否か
- 「文節境界相対位置」は、文節境界との相対位置で形態素位置を指定できるか否か
- 「係り受け部分木」は、（文節）係り受け関係に基づく検索クエリが発行できるか否か
- 「中心位置（キー）指定」は KWIC 表示時に出力する中心位置を指定できるか否か

3.　ラベリングツールの仕様

3.1　仕様の概要

今回開発したラベリングツールは、形態素解析器 MeCab および 係り受け解析器 CaboCha の出力を入力とする。ラベリングツールの出力形式は、各形態素の右列にラベルを付与するか、

表 1　各検索系機能まとめ

検索系	中納言	ChaKi.NET Tag Search	超大規模コーパス 形態論情報	MREP	ChaKi.NET Dep. Search	超大規模コーパス 係り受け部分木
形態論情報正規表現	○	○	×	×	○	×
Ignore Case	×	○	×	×	○	×
文頭・文末指定	○	×	○	×	○	○
文境界またぎ	○	×	×	×	×	×
語彙表・n-gram 出力	×	○	×	×	○	×
文節境界相対位置	×	×	×	×	○	○
係り受け部分木	×	×	×	×	○	○
中心位置（キー）指定	○	○	○	×	○	×

拡張 CaboCha 形式の SEGMENT_S 相当のラベル (松吉ほか (2014)) を付与することによる。

パターン記述言語は、超大規模コーパス検索系で利用している JSON 形式のクエリ言語を拡張したものを用いる。図 10 にパターン記述言語の仕様について示す。

- パターン記述言語 := {"patterns": [ラベル付きパターン JSON+]}
- ラベル付きパターン JSON := {"pattern": {パターン JSON}, "label": 文字列}
- パターン JSON := 形態素列パターン JSON|係り受け部分木パターン JSON

図 10　パターン記述言語仕様（概要）

「パターン記述言語」は複数の「ラベル付きパターン JSON」からなる。ラベル付きパターンは、検索系で用いられるクエリ言語を流用した「パターン JSON」と付与するラベル (label) からなる。「パターン JSON」は「形態素列パターン JSON」と「係り受け部分木パターン JSON」からなり、それぞれ 3.2 節、3.3 節で説明する。

図 11 の例は、形態素列に基づくパターンを記述したものである。左が「補足節:名詞節:コト型:HSa100」、右が「補足節:名詞節:ノ型:HSa200」のパターンの例である。

図 12 の例は、係り受け部分木に基づくパターンを記述したものである。格要素を持つ用言がなす名詞修飾節（「名詞修飾節:補足語修飾節:限定的:MSa100」）にマッチするパターンを示す。

3.2　形態論情報系

図 13 に形態論情報の指定に利用する形態素列パターン JSON の仕様を示す。

形態素列パターン JSON は、指定する形態素の数からなる形態素 JSON 列 ("morphemes") と、各形態素 JSON の出現位置 ("positions") を指定する形態素出現位置 JSON からなる。形態素列による指定の場合には中心位置（マッチする形態素）を形態素出現位置 “0” により指定するが、係り受け部分木による指定の場合には形態素 JSON の "is_target"を True にして中心位置を指定する。

形態素 JSON は、形態素を指定するためのものである。文字列で形態論情報を完全一致で指定するため、現状では ChaKi.NET などで可能な正規表現による形態素指定ができない。今後、正規表現指定のフラグを入れることを検討する必要がある。

```
{
  "patterns": [
    {
      "pattern": {
        "morphemes": [
          {
            "base_lexeme": "事",
            "pos1": "名詞",
            "pos2": "普通名詞",
            "pos3": "一般"
          },
          {
            "pos1": "助詞"
          }
        ],
        "positions": {
          "0": {
            "min": 0,
            "max": 0
          },
          "1": {
            "min": 1,
            "max": 1
          }
        }
      },
      "label": "補足節:名詞節:コト型:HSa100"
    },
    {
      "pattern": {
        "morphemes": [
          {
            "surface": "の",
            "pos1": "助詞",
            "pos2": "準体助詞"
          }
        ],
        "positions": {
          "0": {
            "min": 0,
            "max": 0
          }
        }
      },
      "label": "補足節:名詞節:ノ型:HSa200"
    },
    ...
```

図 11　形態素列に基づくパターン例

```
{
  "patterns": [
    {
      "pattern": {
        "segments": [
          {
            "morphemes": [
              {
                "pos1": "助詞",
                "pos2": "格助詞"
              }
            ],
            "relations": {},
            "prefix_match": false,
            "suffix_match": false
          },
          {
            "morphemes": [
              {
                "c_form": "連体形-*",
"is_target": true
              }
            ],
            "relations": {},
            "prefix_match": false,
            "suffix_match": true
          },
          {
            "morphemes": [
              {
                "pos1": "名詞"
              }
            ],
            "relations": {},
            "prefix_match": false,
            "suffix_match": false
          }
        ],
        "relations": {},
        "dependencies": {
          "0": 1,
          "1": 2
        },
        "prefix_match": false,
        "suffix_match": true
      },
      "label": "名詞修飾節:補足語修飾節:限定的:MSa100"
    }
  ]
}
```

図 12　係り受け部分木に基づくパターン例

3.3　係り受け部分木系

図 14 に係り受け部分木の指定に利用する係り受け部分木パターン JSON の仕様を示す。

係り受け部分木パターン JSON は文節を指定する文節 JSON 列 ("segments") と、文節間係り受けを表す係り受け JSON 列 ("dependencies") と、文節の相対位置を表す情報 (隣接関

```
形態素列パターン JSON := {
  "morphemes": [ 形態素 JSON+],
  "positions": {
    形態素インデックス番号: 形態素出現位置 JSON+
  }
}

形態素出現位置 JSON := { "min": 出現下限位置, "max": 出現上限位置 }
```

```
形態素 JSON := {
  "surface": 文字列,
  "pos1": 文字列,
  "pos2": 文字列,
  "pos3": 文字列,
  "pos4": 文字列,
  "c_type": 文字列,
  "c_form": 文字列,
  "base_reading": 文字列,
  "base_lexeme": 文字列,
  "is_target": 真偽値
}
```

図 13　パターン記述言語仕様（形態素列パターン JSON）

```
係り受け部分木パターン JSON := {
  "segments": [ 文節 JSON+ ],
  "relations": { 隣接関係 JSON+ },
  "dependencies": { 係り受け JSON+ },
  "prefix_match": 真偽値,
  "suffix_match": 真偽値
}

隣接関係 JSON := 隣接関係ラベル
```

```
文節 JSON := {
  "morphemes": [ 形態素 JSON+ ],
  "relations": { 隣接関係 JSON+ },
  "prefix_match": 真偽値,
  "suffix_match": 真偽値
}

係り受け JSON :=
  文節インデックス番号: 係り先インデックス番号
```

図 14　パターン記述言語仕様（係り受け部分木 JSON）

係 JSON "relations", "prefix_match", "suffix_match") からなる。係り受け JSON は文節インデックス番号により指定する文節の係り先をインデックス番号で指定する。隣接関係 JSON("relations") は、隣接関係ラベルとして、文節間が隣接しているか "-"、出現順を保存するか "<"、何も規定しないか " "が指定できる。"prefix_match", "suffix_match"は文頭・文末に隣接するかを指定する。文節 JSON は形態素 JSON 列と、形態素の相対位置を表す情報 (隣接関係 JSON"relations","prefix_match", "suffix_match") からなる。文節 JSON 内の隣接関係 JSON("relations") は、隣接関係ラベルとして、文節と同様に形態素間の関係を"-"、"<"、" "により指定でき、"prefix_match", "suffix_match"は文節頭・文節末に隣接するかを指定できる。

3.4　クエリ言語をメンテナンスする UI

2.3 節・2.6 節で示した超大規模コーパス検索系の Web UI はクエリ言語を発行することが可能である。一方、JSON 形式のファイルを編集するエディタが各種開発されており、それを用いてメンテナンスすることも可能である。図 15 は XML Editor の一つである oXygen XML Editor [(4)]によりクエリ言語を編集している画面である。図 16 は Google Chrome の拡張機能で

[(4)] http://www.oxygenxml.com/

ある JSON Editor [5]によりクエリ言語を編集している画面である。

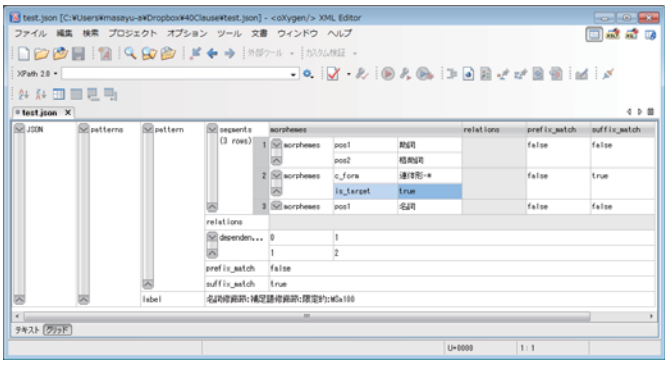


図 15　oXygen XML Editor

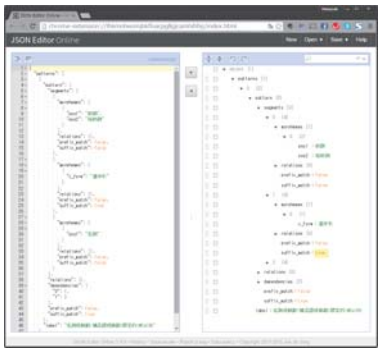


図 16　Google Chrome JSON Editor

4.　節境界ラベリングツールの試作

形態論情報基準[6]・係り受け関係基準[7]については、公開されているコーパスや解析器の出力が事実上の標準として用いられ、共有されている。

一方、節境界については、益岡・田窪品詞体系 (益岡・田窪 (1992)) に基づく丸山ほか (2004) の CBAP や「鳥バンク」(池原 (2007)) で採用されている節境界認定基準などがある。後者の節境界認定基準が公開されているパターンでもっとも細かい記述がされており[8]、解析器が公開されている一方、以下のような問題点がある。

- 品詞体系 IPADIC 品詞体系に基づいており、ChaSen 出力形式である必要がある。
- 文字コードが EUC-JP でなくてはならない。
- 意味属性コードなど、形態論情報や係り受け部分木を超える情報を参照する規則がある。

そこで、UniDic 品詞体系・CaboCha の係り受け構造に基づき、池原 (2007) に規定されている節間意味コードの第 4 段階のパターンの、UTF-8 による再記述を進めている。

5.　おわりに

本研究ではアノテーションの前段階で手がかり句がわかっている場合に用いるラベリングツールの設計と実装について説明した。手がかり句を検索系のクエリで記述可能なことから、各種検索系で発行可能なクエリを示し、各クエリ言語を整理した。整理した結果に基づき、JSON に基づくパターン記述言語を設計し、実装した。

今後の課題として、まず各種検索系との結合が考えられる。パターン記述時には、パターンが実際にコーパス中にマッチする事例を見ながら検討することが多い。検索系と結合すること

---

[5] https://chrome.google.com/webstore/detail/json-editor/lhkmoheomjbkfloacpgllgjcamhihfaj

[6] 益岡・田窪品詞体系益岡・田窪 (1992) に基づく JUMAN・IPADIC 品詞体系に基づく IPADIC/NAIST-jdic/MeCab・UniDic 小木曽・伝 (2013) など。

[7] 京都大学テキストコーパス・KNP・CaboCha など。詳細は浅原 (2013) を参照。

[8] 大分類（第 1 段階）で補足節 28093 パターン、名詞修飾節 40450 パターン、副詞節 66548 パターン、並列節 36321 パターンからなる。

で、シームレスなラベリングツールの開発が行えると考える。手始めとして、「ChaKi.NET」の一機能として実装することを検討している。

次に検索系の整理を行う。形態論情報や係り受け部分木に基づく検索系が複数開発され、検索可能なパターンが少しずつ異なっている。統一的なクエリ言語で対照分析して整理を行う。

最後に節境界ラベリングツールの展開について述べる。第一の動機として、既存のラベリング規則の通時適応がある。現代語の節境界ラベリング規則が固まり次第、近代語への拡張を行う。次に、BCCWJ に対するアノテーションを行う。新聞記事サンプルを中心に鳥バンクの節分類の第 3 段階レベルのアノテーションを進める。さらに、BCCWJ-TimeBank (Asahara et al. (2013)) に対する時制節性 (有田 (2007)) 情報付与があげられる。英語の TimeML (Pustejovsky et al. (2003)) では、SLINK として従属節-主節間の事象構造の関係を付与しているが、これに相当する情報として時制節性のアノテーションを行う。また、鳥バンク節分類と「益岡・田窪体系」(益岡・田窪 (1992)) との節ラベルの対応づけを行う。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」(25284083)、科研費萌芽「近代語コーパスに対する統語情報アノテーション基準策定」(15K12888)、科研費基盤 (C)「「修辞機能」と「脱文脈化程度」の観点からのテキスト分析手法確立と自動化の検討」(15K02535)、科研費若手 (B)「近代口語文翻訳小説コーパスの構築と計量的文体研究」(25770178)、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

有田節子 (2007).『日本語条件文と時制節性』　くろしお出版.

浅原正幸 (2013).「係り受けアノテーション基準の比較」　第 3 回コーパス日本語学ワークショップ予稿集, pp. 81–90.

Asahara, M., S. Yasuda, H. Konishi, M. Imada, and K. Maekawa (2013). “BCCWJ-TimeBank: Temporal and Event Information Annotation on Japanese Text.” *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*.

池原悟 (2007).「意味類型パターン記述言語仕様書」　Technical report, 独立行政法人科学技術振興機構, 戦略的基礎研究事業, 高度メディア社会の生活情報技術.

国立国語研究所コーパス開発センター (2015a).『コーパス検索アプリケーション「中納言」』, https://chunagon.ninjal.ac.jp/.

国立国語研究所コーパス開発センター編 (2015b).『日本語歴史コーパス』.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

益岡隆志・田窪行則 (1992).『基礎日本語文法–改訂版–』　くろしお出版.

Matsumoto, Yuji, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Otani, and Toshio Morita (2006). “An Annotated Corpus Management Tool: ChaKi.” *Proc. of LREC-2006*, pp. 1418–1421.

小木曽智信・伝康晴 (2013).「UniDic2: 拡張性と応用可能性にとんだ電子化辞書」　言語処理学会第 19 回年次大会発表論文集, pp. 912–915.

Pustejovsky, J., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz (2003). “TimeML: Robust Specification of Event and Temporal Expressions in Text.” *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, pp. 337–353.

丸山岳彦・柏岡秀紀・熊野正・田中英輝 (2004).「日本語節境界検出プログラム CBAP の開発と評価」　自然言語処理, 11:3, pp. 39–68.

松吉俊・浅原正幸・飯田龍・森田敏生 (2014).「拡張 CaboCha フォーマットの仕様拡張」　第 5 回コーパス日本語学ワークショップ, pp. 223–232.

形態素解析辞書「中古和文 UniDic」を用いた古文単語帳作成

大津　千尋，　三日市　綾花，　須永　哲矢（昭和女子大学）†

Compilation of Classical Literature Wordbooks Using an Electrical Dictionary for Morphological Analysis "Chuko-Wabun UniDic"

Chihiro Ohtsu,　Ayaka Mikkaichi,　Tetsuya Sunaga　(Showa Women's University)

要旨

形態素解析辞書「中古和文 UniDic」の教育転用の一例として、古文単語帳の作成を試み、作成方法の紹介と、作成結果から読み取れる言語事実の報告を行う。作成方法の概要は以下の通り。１）高校の古典教科書をテキストデータ化し、「中古和文 UniDic」により形態素解析、解析結果を Excel に出力する。２）解析結果をもとに高校の教科書に使用されている語の語彙頻度表を作成する。３）頻度表をもとに、単語帳に収録すべき古文単語を選定し、実例に基づいた訳語を充てる。今回の研究では、まずは特定の教科書 1 冊を元に単語帳の作成を目指し、「教科書に載るテキストの高頻度語」を明らかにした。教科書に出現する自立語延べ約 6500 語、異なり約 1500 語を対象に調査したところ、異なり語数にして全体の 2 割程度、300 語強でテキスト全体の約 7 割をカバーできることが明らかになった。ここで作成した単語リストを別の教科書テキストに対しても適用したところ、ほぼ同等のカバー率を得ることができ、有効性が確認できた。

１．はじめに

国立国語研究所「中古和文 UniDic」の公開により、特に機械処理の知識を持たない一般ユーザーであっても、歴史的資料に対して機械処理を行った研究が可能になっている。「中古和文 UniDic」は、現代語を対象とした従来の解析辞書では無力であった古典資料に対し、高精度で解析することを可能にした画期的な形態素解析辞書であり、実際これを利用したデータとして国立国語研究所「日本語歴史コーパス」の公開も始まっている。古典語のみならず、近年さまざまなコーパスが公開され、研究環境は充実しているが、コーパスを利用するという場合には、調査対象は自動的にコーパス化されているもののみに限られる。しかし研究目的によっては、コーパス化されている範囲と調査したい範囲が異なるという場合も十分ありうることで、そのような場合には自分でデータを作るということになる。その際には、特別な知識がない一般ユーザーにとっても使用しやすい UniDic は非常に有用である。形態素解析辞書「中古和文 UniDic」の利用の可能性は研究利用にとどまらず、須永(2014)のように教育面においても、主に高等学校での古典学習教材等、さまざまな活用法がありうる。本稿では、形態素解析辞書「中古和文 UniDic」の教育転用の一つとして、古典教科書本文をもとに形態素解析を行ったデータをもとに古文単語帳の作成を試み、その手順の紹介、および有効性の検証を行う。

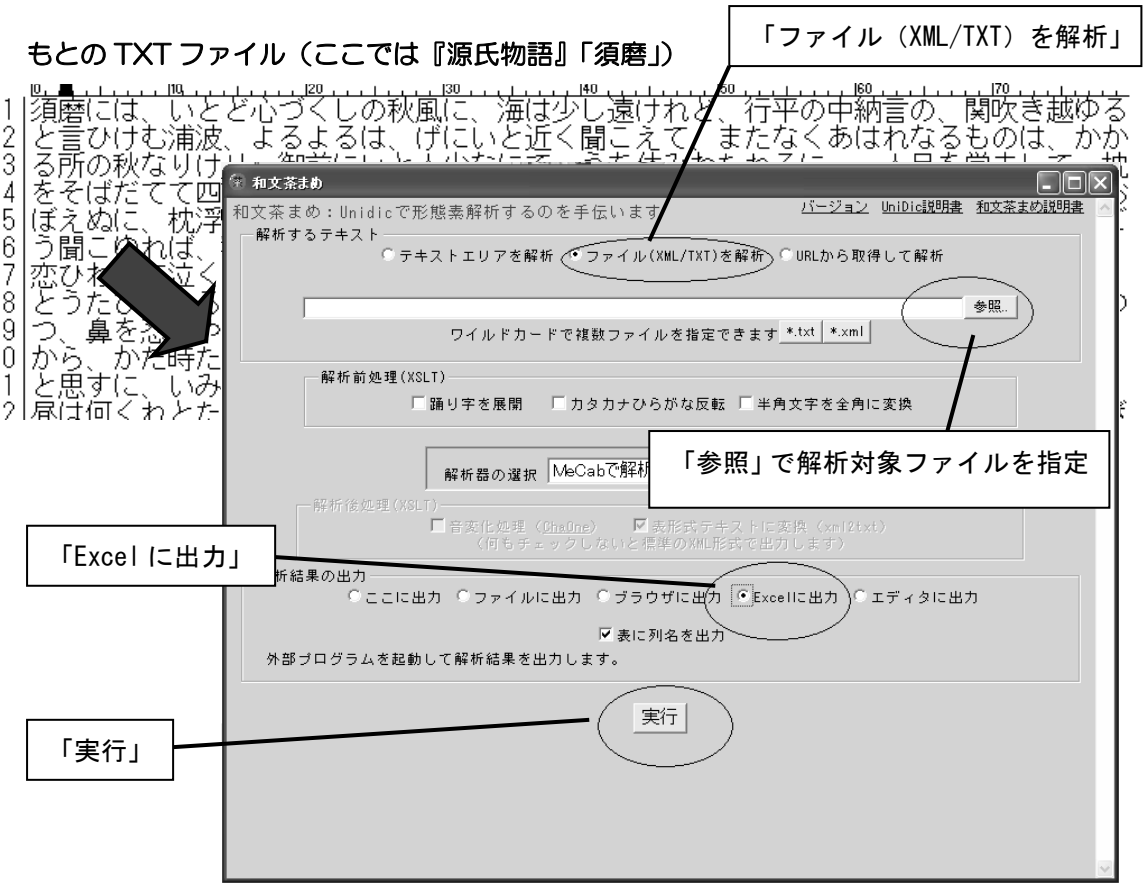
２．形態素解析辞書「中古和文 UniDic」とその利用

形態素解析とは、簡単に言えば「機械が自動で品詞分解して、活用の種類や活用形を書き出してくれる」というものである。公開されている「中古和文 UniDic」は中古和文 UniDic ホームページより無償でダウンロードできる。利用するには「MeCab0.96」以降（こちらも

†21112069@st.swu.ac.jp　21112318@st.swu.ac.jp　tsunaga@swu.ac.jp

無償）がインストールされていることが前提となるが、それも含め、ホームページでの指示に従ってダウンロード・インストールを行えば、特に機械処理に関する詳しい知識がなくとも、誰でも手軽に形態素解析を行う環境を手に入れることができる。

実際の操作にあたっては操作用ツール「和文茶まめ」が用意されており、ユーザはマウス操作で簡単に解析が行えるようになっている。古典本文を txt 形式で用意しておけば、あとはこの操作画面でファイルを指定してやれば、自動で品詞分解が完了する。（おおよそのイメージは図 1 参照)。



	1	2	3	4	5	6	7	8	9	10	11
1	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
2	源氏須磨	B	須磨	スマ	スマ	スマ	名詞-固有名詞-地名-一般			スマ	固
3	源氏須磨	I	に	ニ	ニ	に	助詞-格助詞			ニ	和
4	源氏須磨	I	は	ワ	ハ	は	助詞-係助詞			ハ	和
5	源氏須磨	I	、			、	補助記号-読点				記号
6	源氏須磨	I	いとど	イトド	イトド	いとど	副詞			イトド	和
7	源氏須磨	I	心づくし	ココロズクシ	ココロヅクシ	心尽くし	名詞-普通名詞-一般			ココロヅクシ	和
8	源氏須磨	I	の	ノ	ノ	の	助詞-格助詞			ノ	和
9	源氏須磨	I	秋風	アキカゼ	アキカゼ	秋風	名詞-普通名詞-一般			アキカゼ	和
10	源氏須磨	I	に	ニ	ニ	に	助詞-格助詞			ニ	和
11	源氏須磨	I	、			、	補助記号-読点				記号

図 1　操作画面「和文茶まめ」での操作と、出力される Excel ファイル

形態素解析を通し、機械が品詞分解をした結果、さまざまな情報が付与されるが、その中に「語彙素」という情報がある。「語彙素」とはいわば辞書見出し形であり、実際の表記・

活用形がどうであれ、辞書形・代表表記に戻したうえで語を表示する列であり、たとえば本文内の出現形が「はしる」であろうと「走ら」であろうと、語彙素レベルでは「走る」に統一される（図 2)。そこで、この「語彙素」列を利用することで、日本語で語を数える際の難関である、表記や活用形などの語形のゆれを乗り越えて、単語の数を自動で、正確に数え上げることが可能になる。

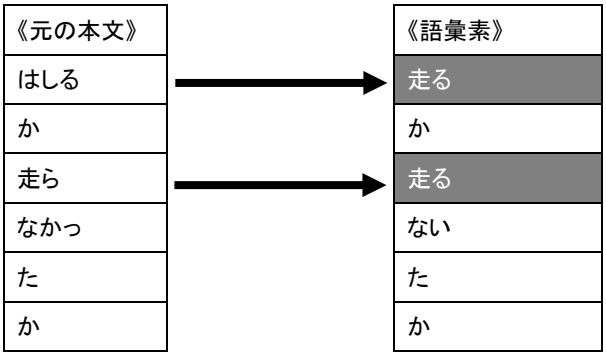


図 2　「語彙素」列のイメージ

３．古文単語帳の作成

上述の「語彙素」列を利用することにより、頻出語を抽出することが可能となる。「中古和文 UniDic」以前は、表記や活用の問題があり、古文テキストから単語を自動的に取り出すことは困難であった。表記や活用の問題が深刻でない英単語においては、機械処理をもとにした学習参考書・英単語帳が数多く見られるのに対し、古文単語帳の方ではそのような客観的根拠をもとにしたものがさほど見られなかったのは、このような事情によると思われる。そこで今回は、「中古和文 UniDic」での形態素解析を利用し、出現頻度という客観的根拠をもとにした単語帳の作成を試みたいと考えた。収録語数・レベルなどによって目標設定は変わりうるが、今回は第一回めの試作ということもあり、教科書に出現する単語を対象とし、必要最低限の入門的な単語帳、というレベルを想定している。

３．１　作成元となるテキスト

今回の単語帳作成元となる古文テキストは、高校の教科書 1 冊分とした。対象とした教科書は第一学習社『古典Ｂ』(2015 年度版)。『古典Ｂ』の中には中世以降、近世までのテキストも収録されており、中古のいわゆる「古典」とは毛色の違う作品も多い。「中古和文 UniDic」は中古語を対象としていること、また、学校教育において中世以降の作品に触れることはあっても、文法教育や単語教育の面においては実際のところ中古語に照準が合わせられていることを考え併せ、調査対象は中古のものに限定した。今回の試作で元とした古典作品は表 1 に示す 8 作品 26 話、総語数は 1 万 2860 語である。

表 1　単語抽出元とした作品（第一研究社『古典Ｂ』収録部分）

作品名	収録タイトル	語数
枕草子	「宮に初めて参りたるころ」「古今の草子を」「二月つごもりごろに」「ふと心劣りとかするものは」「この草子、目に見え心に思ふこと」	1764
源氏物語	「須磨の秋」「住吉参詣」「明石の姫君の入内」「紫の上の死」「薫と宇治の姫君」	3360
紫式部日記	「若宮誕生」「日本紀の御局」	585
更級日記	「門出」「源氏の五十余巻」「大納言殿の姫君」	1227
大鏡	「雲林院の菩提講」「花山院の出家」「道長と伊周―弓争ひ―」「時平と道真」	3692

	「兼通と兼家の不和」「道隆と福足君」「三舟の才」「道長と隆家」	
堤中納言物語	「このついで」	814
とりかへばや物語	「父大納言の苦悩」	659
しのびね物語	「偽りの別れ」	759
	計	12860

３．２　UniDic の単位認定と、単語帳作成面での精度

「中古和文 UniDic」はあくまで機械プログラムによって自動で品詞分解しているのであり、自動解析結果にはエラーも生じる。中古和文 UniDic は、平安仮名文学作品に対しては高い解析精度を実現しており、中古和文 UniDic　Ver0.5 の段階で、単位境界（品詞の切れ目が正しいか）で 99.3%、品詞認定で 97.8%という解析精度が報告されている（中古和文 UniDic ホームページほか）が、教科書のテキストに対してはどの程度の精度をもって解析が可能なのかは検証しなければならない。実際の作業においては、データの正確さのためには自動解析結果を人の目で確認、エラーを修正する必要がある。今回は自動解析に加え、人手による確認・修正作業も行った。今回解析に使用した「中古和文 UniDic」は Ver1.4(2014 年 3 月公開)である。

また、「中古和文 UniDic」が自動で「単語に分ける」という際の言語単位についても補足しておかねばならない。「中古和文 UniDic」は、国立国語研究所のデータ共通の言語単位として「短単位」という単位を採用しており、表 1 の語数もこの「短単位」の数による。「短単位」認定の詳細については規程集が公開されているためそちらを参照されたいが、一般的な高校教育での単語認定と、形態素解析結果の「短単位」としての語認定での相違点として注意せねばならないのは、以下の 2 点である。

①解析結果の 1 語は、一般的な高校教育での 1 語より小さい場合がある。

例えば高校教育では「吹き越ゆ」「大納言」などで 1 語とする方が一般的であるが、「中古和文 UniDic」では「吹く」＋「越ゆ」、「大」＋「納言」の 2 単位として解析される。

②解析結果の品詞・活用形認定は、一般的な高校教科書と異なる場合がある。

大きく異なるのは以下の 2 点である。（1）形容動詞の認定：UniDic の品詞体系では「形容動詞」はなく、いわゆる形容動詞語幹を「形状詞」、続く「なり」は断定の助動詞と認定する。例えば「きよらなり」は学校教育では形容動詞 1 語という認定だが、UniDic では形状詞「きよら」＋助動詞「なり」となる。（2）完了の助動詞「り」が接続する活用形は、学校教育では已然形が一般的であるのに対し、UniDic では命令形と認定する。高校の古典教材作成の用途・目的によっては、以上 2 点に注意し、修正が必要となる。

しかし、今回の目的は単語帳の作成であり、単語帳のための頻出語洗い出しという目的からは、上記①②はさほど問題にならない。まず①についてであるが、学校教育に倣って「吹く」「越ゆ」とは別個に動詞「吹き越ゆ」を認定し、別動詞として新たに指導するよりも、「吹き越ゆ」も分割して「吹く」と「越ゆ」の中に解消して処理する方が、一般性が高く、効率的である。このような複合語については、複合によって、元の語の足し算からは導けないような意味が生じる場合のみ、注意せねばならないが、大部分の、意味の足し算で複合語の意味も導けるような場合に関しては、むしろ UniDic のように分割して元の語だけを意識させる方が効率的である。①および②（1）に関しては、品詞認定と品詞分解の切れ目を示す教材を作成する、というような用途にとっては致命的だが、古文が読めるように、よく出る語を洗い出す、という用途にとっては問題は生じない。①に関しては可能な限り基本的な語に分解しておいた方が複合語として項目を立てるよりも一般性が高く有用であるし、②（1）の「形容動詞」／「形状詞＋なり」という認定の差についても、UniDic での「形状詞」を形容動詞として数え上げればよいだけの話であり、問題はない。②（2）に関しても、単語帳作成という範囲では、代表形としての「語彙素」が取り出せればよい

のであって、活用形の認定の違いは問題にならない。

以上のような観点から、単語帳作成のために単語抽出を行うという目的において、「中古和文 UniDic」が高校古典教科書に対してどの程度の精度を実現しているのか、エラーチェック作業を通して検証したところ、1 万 2860 語のうち、「語彙素」「品詞」レベルで語認定が誤っていたのはわずか 1 か所であった。高校の古典教科書に収録されるテキストは、高校生に読みやすいよう、表記、仮名遣いが統一された整ったテキストになっており、このようなテキストに対しては、「中古和文 UniDic」は通常以上の精度を達成できることが実証された。活用形などの認定込みで、別の学習教材を作成する場合、活用形レベルでのエラーを拾うとなるとエラーはもう少し増えるが、それとてたいした量ではなく、作業面において十分実用に足る精度と言える。極端な話、単語帳のための語彙頻度表を作成するだけなら、自動解析のままエラーチェックをしなくてもさして問題がないほどであると見てよかろう。

表 2　単語抽出目的における誤解析状況

作品名	語数	エラー
枕草子	1764	なし
源氏物語	3360	なし
紫式部日記	585	なし
更級日記	1227	なし
大鏡	3692	「さいつごろ」→接頭辞「さ」+「いつ頃」　（本来は先／つ／頃）
堤中納言物語	814	なし
とりかへばや物語	659	なし
しのびね物語	759	なし
計	12860	1か所

３．３　解析結果をもとにした語彙頻度表の作成

「中古和文 UniDic」では、解析結果を Excel に出力することができるので、解析結果をそのまま Excel データとして利用し、簡単に語彙頻度表を作成することができる。方法は人によってさまざまであるが、ここでは作業の中心となる手順の一例を紹介する。

（1）「語彙素」列をコピーする。

	A	B	C	D	E	F	G	H	I	J	K
1	出典	文境界	書字形	発音形	語彙素読み	語彙素	品詞	活用型	活用形	語形	語種
2	紫式部日記	B	十月	カミナズキ	カミナヅキ	神無月	名詞-普通名詞-一般			カミナヅキ	和
3	紫式部日記	I	十	ジュー	ジュウ	十	名詞-数詞			ジュウ	漢
4	紫式部日記	I	余	ヨ	ヨ	余	名詞-数詞			ヨ	漢
5	紫式部日記	I	日	ニチ	ニチ	日	名詞-普通名詞-助数詞可能			ニチ	漢
6	紫式部日記	I	まで	マデ	マデ	まで	助詞-副助詞			マデ	和
7	紫式部日記	I	も	モ	モ	も	助詞-係助詞			モ	和

図 3　「語彙素」列を利用

（2）新しいシートにコピーした「語彙素」列を 1 列あけて 2 列コピーする。一方の列（図 4 ではC列）に対し、「データ」＞「重複の削除」で重複の削除を行う。A列がテキスト出現順に単語が並んでいるのに対し、C列は重複を削除したことにより、そのテキストの異なり語のリストとなる。この時点で、A列に並んでいる語の総数が延べ語数、C列の語の総数が異なり語数ということになる。

図 4　「重複の削除」を利用し、延べ語・異なり語リストを作成

（3）異なり語リストをもとに、延べ語の列における各語の出現数を計算する。ここでは COUNTIF 関数を使用する。COUNTIF 関数とは、指定した条件に一致するセルの個数を計測する関数で、図 5 のとおり、結果を表示させたいセルに直接「=countif」と入力する。（範囲,検索条件）の「範囲」は計測する範囲、「検索条件」は、ここでは計測対象とする語となる。図 5 では、「=countif(A:A,C2)」と指定しているが、これは「A:A」（A列全て、つまりテキスト上に出現した延べ語リスト）から、「C2」のセルにある文字列「昔」と一致するセルの数をカウントするよう指定していることになる。範囲や検索条件の指定は、直接入力せずとも、マウスのカーソル移動・指定でも可能である。

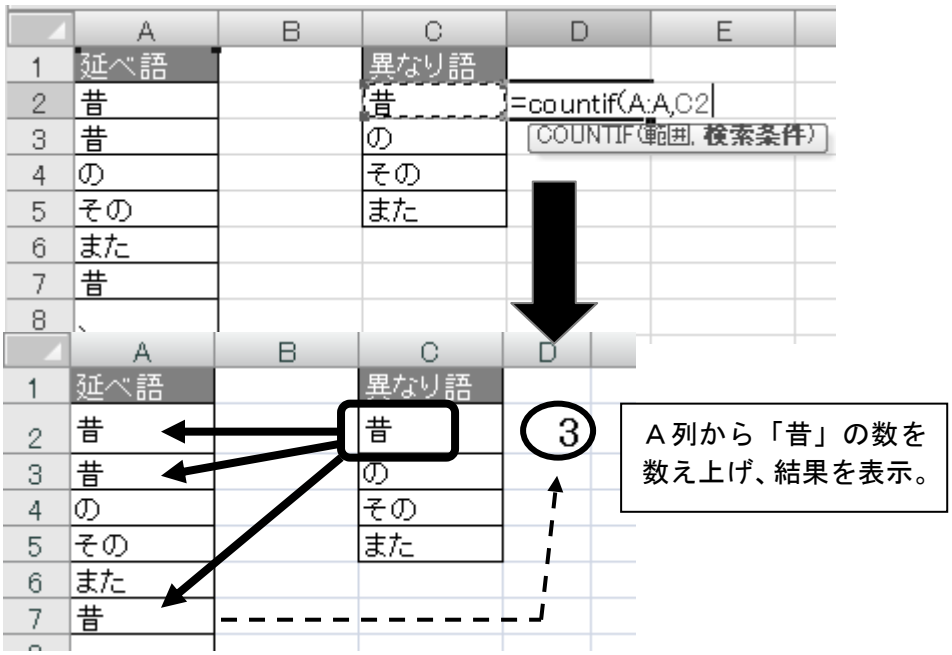


図 5　COUNTIF 関数の利用

（4）以上の操作で、単語の出現頻度を算出することが可能となる。この後は「並べ換え」などを利用し、高頻度順に並べ直したりすればよい。

４．古文単語帳の試作

以上の手順を利用して作成した語彙頻度表をもとに、高頻度語を抽出し、古文単語帳の作成を試みる。まず古文単語帳に収録する品詞の範囲であるが、助詞・助動詞といった付属語はむしろ「文法」の要点であり、数の上でも有限で、文法教育の側でカバーされる。このため、「単語帳」の収録対象は自立語に限定し、さらに固有名を排除することとした（頻出の固有名も将来的には収録すべきかと考えられるが、今回の試作では除外）。この時点で、元になるテキストの総語数は延べ 6488 語となる。

表 3　調査対象となる自立語（固有名除く）の延べ語数・異なり語数

延べ	異なり
6488	1485

４．１　「よく出る単語」の抽出方法

さて、各テキストの語のリストから「よく出る単語」を抽出するわけだが、何をもって「よく出る」とするかについては幾つか別の考え方がありうる。一つは素直に、教科書の対象テキスト全体から、出現数の高い順に語を取りだしていく、という方式であるが、この場合、ある作品のある箇所にのみ多数登場するが、他の作品ではほとんど登場しない、という語があった場合、たまたま教科書に載った箇所の特殊性ゆえに、高頻度語に位置づけられてしまう可能性もある。そこで別の方法として、「その語が何作品にまたがって出現するか」という尺度も導入することとする。今回対象となる古典作品は表 1 に示した 8 作品であり、総数は問わず、複数作品に出現した語を「よく出る」とする見方である。作品は問わず、全体の総数順で「よく出る」と認定した「総語数方式」と、総数は問わず、出現した作品数で「よく出る」と認定した「作品数方式」の２種を試し、有効性に差があるのかを以下で検証する。

４．２　総語数方式・作品数方式による単語抽出とカバー率

まず総語数方式で 4 回以上出現する語を抽出したところ、345 語であった。調査対象テキスト全体の異なり語数が 1485 であるため上位 23%を切り出したことになる。この 345 語で、実際のテキスト全体の自立語のうちどの程度がカバーできるかを算出したところ、72%がカバーできることが明らかになった。

続いて作品数方式であるが、作品数方式では出現作品数を 4 回以上とすると 325 語がこれに該当し、総語数方式で 4 回以上出現した語の語数とほぼ同じ規模になる。この場合のカバー率も 70%と、総語数方式とさほど差は出なかった。実際、両方式で抽出した 345 語・325 語のうち、278 語が共通であった。参考までに表 5，6 に各方式の上位 10 語を挙げるが、その大部分がどちらの方式で抽出しても取りだせるものであることがわかる。

総語数方式であれ、作品数方式であれ、よく出る単語の上位 2 割、300 語程度で、実際のテキストの 7 割ほどがカバーできるのである。

表 4　総語数方式・作品数方式のカバー率

	語数	作品全体の異なり語数に対するカバー率	作品全体の延べ語数に対するカバー率
総語数方式、4 回以上	345	23.2%	72.4%
作品数方式、4 作品以上	325	21.9%	70.6%

表 5　総語数方式による上位語 10 位（数字は出現語数）

形容詞・形容動詞		動詞		副詞		名詞	
なし	51	給ふ	371	いと	91	事	135
いみじ	43	す	129	かく	24	人	95
あはれなり	30	あり	114	然	20	程	59
をかし	21	思ふ	97	ただ	17	物	56
めでたし	18	見る	75	少し	16	様	49
怪し	15	言ふ	74	げに	16	心	45
あさまし	14	出づ	72	いかに	16	方	37
近し	12	侍り	67	なほ	16	世	36
とし	11	成る	51	え	14	一	27
悲し	11	申す	51	しばし	13	前	26

表 6　作品数方式による上位語 10 位（数字は作品数）

形容詞・形容動詞		動詞		副詞		名詞	
なし	8	給ふ	8	いと	8	一	8
いみじ	8	す	8	ただ	8	物	8
あはれなり	8	あり	8	いかで	8	方	8
近し	7	思ふ	8	かく	8	内	8
をかし	7	見る	8	え	7	世	8
口惜し	7	言ふ	8	しばし	6	程	8
怪し	6	出づ	8	少し	6	様	8
あさまし	6	侍り	8	なほ	6	人	7
心苦しい	5	成る	8	げに	5	事	7
悲し	5	覚ゆ	8	しばし	5	心	7
（他にも 5 作品出現語多数）		（他にも 8 作品出現語多数）		（他にも 5 作品出現語多数）		（他にも 7 作品出現語多数）	

※白抜きは総語数方式・作品数方式ともに出現

４．３　人による単語選定と、意味記述

以上、実数にして 300 語ほどでテキストの 7 割をカバーできる単語リストを得ることができるが、ここから人手の作業が残されており、この人手作業を経てこそ、単語帳の実用性は高まると考える。第一に意味記述の問題がある。形態素解析から作れるのは単語リストまでであり、教科書に合わせて必要十分な意味を記述していくのは人間の仕事ということになる。また、単語リストから覚える必要のない語を、人間の目で排除していくことで、単語数はさらに減らすことができる。たとえば表 5 の頻出名詞を見ると 1 位は「事」2 位は「人」3 位は「程」…となっており、これらは現代語にも共通する基本語彙であって、「古文単語」としてとりたてて覚える必要はない。「300 語ほど」とした語数の中にはこのような語も多数含まれるため、人間の目で選定していけば「カバー率 7 割の入門用の単語帳」は、より少ない語数で実現することが可能となる。現代では使わない古文特有の単語、および現代でも使う語ではあるが古文特有の意味・用法をもつ語を重点的に洗い出して記述

していくことで、より効率的な単語帳が作成できるはずである。

以上の手順で作成した語彙表をもとに、単語を予備的に選定したところ、この約 300 語から、実際覚える必要のある語は 120 語ほどという見通しを得た。「いみじ」や「具す」などに代表される、現代で使わない古文特有の単語としては 56 単語、「めでたし（→古典語では「すばらしい」）」や「驚く（→古典語では「目が覚める、気付く」）」のように、現代でも形式自体は使うが、古文特有の意味・用法をもつ単語として、64 単語というのがその内訳である。選定基準や、選定語そのものについては今後とも検討を要すると考えているため、今回のここでの報告はあくまで予備調査としての見通しにとどまるが、実用面を考慮し、人間の目で単語選定をすることによって、今回の語彙リストにおいては「古文単語」として覚えるべき基本語彙は半数以下になることが確認された。

５．実用性の検証

今回の語彙リスト作成の段階で、頻出語上位 300 語ほどで教科書の 7 割がカバーできることが明らかになった。ただしこれはあくまで１つの教科書をもとにした結果である。データを取る元となったテキストに対し、カバー率を測定したのであるから、この時点でカバー率が高くなるのはある意味当然といえる。ここで作成した単語リストが、他の同レベルのテキストでも有効なのか、あるいはあくまで今回対象とした教科書限定の単語帳なのかを明らかにせねば、このような単語帳の作成法が本当に有効なのかは判断ができない。そこで今回は検証実験として、作成した単語リストを、別の教科書の、今回採られていない話に対して適用し、その場合のカバー率を測定することとした。対象としたのは『大和物語』より「旅寝の夢」、今回データ採取対象の教科書には収録されていないが、教科書一般の定番である『源氏物語』より「葵の上と物の怪」「藤壺の里下がり」、および、後の時代の作品として『徒然草』より「あだし野の露消ゆるときなく」である。教科書に収録されている分量ということもあり、各話の総語数はさほど大きくない規模での検証実験である。

表 7　効果の検証に用いた別教科書のテキストと、その自立語総語数

	大和物語	源氏「物の怪」	源氏「藤壺」	徒然草	計
自立語総語数	102	435	409	97	1043

カバー率の検証結果は表 8 のとおりで、別教科書に適用しても、同時代の作品であればデータ採集元となった教科書とほぼ変わらない効力を発揮することが明らかになった。また、時代の異なる『徒然草』に対しては、やはりカバー率がやや下がることも確認された。

以上の検証から、教科書 1 冊をもとにした入門用の単語リストが、別教科書に対しても適用できる、一般性の高いものであると判断してよかろう。

表 8　別教科書に適用した際のカバー率の検証

	（元データ教科書）	大和物語	源氏「物の怪」	源氏「藤壺」	徒然草
総語数方式	72.4%	71.6%	69.0%	70.0%	64.0%
作品数方式	70.6%	70.6%	67.9%	68.0%	57.8%

また、今回試作した単語リストに収録された語が、これら別教科書において異なり語としてどの程度出現するのかという、稼働率の算出も試みた。表 7 のとおり、テキスト量がさほど大きくないため、検証に用いた 4 話を統合した上で、総語数方式・作品数方式の双方のリストと突き合わせ、稼働率を測定したところ、1000 語ほどのテキストを相手に 56%ほどの稼働率を見せ、汎用性の高さが証明された。なお、参考までに作品別にも稼働率を

算出したが、検証対象となる自立語総数が 100 語ほどの『大和物語』や『徒然草』は、当然稼働率は低く 1 割程度であり、「葵の上と物の怪」、「藤壺の里下がり」といった自立語総数 400 語程度のテキストになると、3 割台の稼働率を見せるようになる。これが 1000 語ほどのテキストに対しては稼働率 5 割半ばとなる。

表 9　別教科書を対象にした際の稼働率の検証

	徒然草	大和物語	源氏「藤壺」	源氏「物の怪」	4 話統合
	（97 語）	（102 語）	（409 語）	（435 語）	（1043 語）
総語数方式	11.6%	11.6%	33.6%	35.4%	56.5%
作品数方式	10.5%	11.4%	34.9%	36.7%	56.2%

以上の検証により、これらの単語リストは、カバー率の面でも、稼働率の面でも高成績と評価してよく、この単語リストは利用に際して、効率の良いものであると言えよう。

６．おわりに

以上、「中古和文 UniDic」を利用した学習教材開発の一環として、本稿では解析結果をもとにした単語帳作成の流れと、実効性の検証を行った。今回の研究で頻出語上位 300 語ほどで、古典教科書の 7 割ほどがカバーできること、また、語彙採集元とは別の教科書に対しても同様の有効性が見込めることが明らかになった。今後の作業としては、今回の単語リストをもとに実際に覚えるべき語の選定と、意味記述が待っているが、予備調査を通して得た見通しとしては、上位 300 語のうち、覚えるべき語は 120 語に減らせる見込みである。120 語覚えれば 7 割カバーできる、というのは非常に効率的であると考えられる上に、実際の学習上コストとしては、覚える語は 120 語より増やして、200 語、300 語程度にしてもまだまだ現実的な語数といえる。よって今後は、意味記述の精密化など、これに続く作業を継続するのはもちろんであるが、並行して、語彙リストをさらに拡充し、8 割程度をカバーできる単語帳作成なども目指していきたい。

文　献

小木曽智信・小椋秀樹・田中牧郎・近藤明日子・伝康晴（2010）「中古和文を対象とした形態素解析辞書の開発」『情報処理学会研究報告 人文科学とコンピュータ』Vol.2010-CH-85(No.4) pp.1-8

小木曽智信・小椋秀樹・近藤明日子・須永哲矢(2010)「形態素解析辞書「中古和文 UniDic」とその活用例」『日本語学会 2010 年度秋季大会予稿集』 pp.243-248

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』短単位規程集第 4 版』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　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』平成 21（2009）- 平成 23（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」研究成果報告書 2（課題番号 21520492、代表者　小木曽智信）

須永哲矢(2014)「形態素解析辞書「中古和文 UniDic」を利用した古典学習教材の作成」『第 6 回コーパス日本語学ワークショップ予稿集』pp.11-20

関連 URL

日本語歴史コーパス「中納言」 http://maro.ninjal.ac.jp/

中古和文 UniDic http://www2.ninjal.ac.jp/lrc/index.php?UniDic

MeCab　http://taku910.github.io/mecab/

二字漢語における語と漢字の意味の結びつきの特徴

—国語辞典の語義の説明文を利用した調査—

本多　由美子（一橋大学大学院言語社会研究科）[1]

Features of Meaning-Kanji Association in Two-character Sino-Japanese Words: Survey of Dictionary Texts

Yumiko Honda (Hitotsubashi University Graduate School of Language and Society)

要旨

漢字二字から成る漢語（以下、二字漢語）とその漢語を構成する各漢字の意味の結びつきについて、BCCWJ の高頻度語を対象に分析を行った。漢語と 1 字ごとの漢字の意味の結びつきに注目し、「語と漢字の意味が 2 字とも結びつく語」、「2 字とも結びつきにくい語」、「1 字のみ結びつく語」に 3 分類し分析した。その結果、高頻度語を頻度順にグループ分けすると、最上位 100 語以外では、3 分類の割合はほぼ一定であることが明らかになった。また、語と漢字の結びつきは、1 字目と 2 字目の漢字では、品詞による違いがあり、語構成との関係が示唆された。

本調査では結びつきを判断する際、国語辞典の語義の説明文を用いた。この結果を、日本人大学生を対象にした調査結果（桑原（2013)）の透明度の数値と比較したところ、結びつきについて同様の傾向が見られた。

１．はじめに

漢字はそれぞれの字が意味をもち、また、漢字が組み合わさった熟語は語としての意味を持つ。二字漢語には、漢字 2 字とも語の意味と結びつく語、2 字とも結びつきにくい語、2 字のうち 1 字は結びつくが 1 字は結びつきにくい語がある（表 1)。

表 1　語と漢字の意味の結びつき　「国外」「人口」「条件」

二字漢語	国語辞典の説明文	1 字目の漢字		2 字目の漢字		二字漢語と漢字の結びつき
		説明との一致部分	語との結びつき	説明との一致部分	語との結びつき	
国外	国のそと。	国	結びつく	そと	結びつく	2 字とも結びつく
人口	人の数。	人	結びつく	なし	結びつきにくい	1 字のみ結びつく
条件	物事を決定したり約束したりするときに，前提あるいは制約となる事柄。	なし	結びつきにくい	なし	結びつきにくい	2 字とも結びつきにくい

母語話者や漢字に慣れた日本語学習者は、よく目にする語であれば、1 字ごとの漢字の意味を考えることなく、語の意味を思い浮かべることができるだろう。また、語の語源や漢字

[1] nihonda@hotmail.com

の字義に関する知識が豊富であれば、意味が結びつく語もある。しかし、現代において一般的に使われている語や漢字の意味で漢語を捉えた場合には、結びつくものと結びつきにくいものがあるのではないだろうか。

そこで、本研究では、よく目にする漢語について、語と漢字の意味の結びつきという視点から、どのような傾向や特徴が見られるかを調査し考察することにした。

語と漢字の意味の結びつきは、日本語教育でも活用できる可能性がある。筆者自身、非漢字圏の初級学習者から「親切」の漢字表記は、その学習者が知っている漢字の１字ごとの意味では、語の意味と結びつかないと言われた経験がある。語や漢字の知識が十分でない学習者は、日々、学んだことのない漢語を目にする。中には、知っている漢字の組み合わせでも語としては初めて見るものもあるだろう。日本語学習者への教育を考える際に、語と漢字の意味の結びつきは、利用できる情報の一つであると考える。

２．先行研究

国語教育の観点から漢語と漢字の意味の結びつきについて述べられているものに宮島（1968）がある。宮島（1968）では、漢語には、1字ごとの漢字の意味が語の意味と結びつく語と1字ごとに分解しても語の意味に結びつかない語があること、また、それらの語の特徴によって、1字ごとに分解する方法や2字まとめる方法など、教え方を変える必要があることが指摘されている。漢字を音訓漢字や字音漢字などの機能から分類したものに森岡（2004）がある。森岡（2004）では、JIS 漢字表の各漢字について、現代語の和語や漢語を表記する際に用いられているか否かによって漢字が分類されている。語構成の観点から二字漢語を漢字二字の結合パターンによって分類したものに野村 (1988)、張 (2014) がある。

３．本研究の目的とリサーチクエスチョン

本研究の目的は、二字漢語と漢字の意味の結びつきについて、よく目にする語の傾向や特徴を明らかにすることとする。語の特徴、漢字の特徴、結びつき方の特徴について考察するために、以下のリサーチクエスチョン（以下、RQ）を立てた。

二字漢語を、語とその語を構成する漢字の意味の結びつきにより、「語と漢字の意味が 2字とも結びつく語」、「1 字のみ結びつく語」、「2 字とも結びつきにくい語」に分類した場合

RQ1．語の頻度により、結びつき方に違いがあるか。

RQ2．語の品詞により、結びつき方に違いがあるか。

RQ3．語を構成する漢字について、1 字目、2 字目の漢字の結びつき方に違いがあるか。

４．調査

４．１　調査方法の検討

本調査では、語の意味については、国語辞典の語義の説明文を用いることにした。国語辞典の語義の説明は、漢字を説明するために書かれているものではないと思われるが、語と漢字の意味に結びつきがあれば、ある程度語義の説明に表れるのではないかと考えたからである。当初、筆者は周りの日本語母語話者数名に聞きながら、読み下し文をつけることを試みたが、語と漢字の意味の結びつきの判断には個人差があり、客観性に欠けると判断した。

国語辞典の語義は、辞典によって説明の仕方が偏る可能性があるため、複数の辞典の語義の説明を用いることにした。辞典については４．５．３で述べる。以下、調査に用いる国語辞典の語義の説明文を「国語辞典説明文」と呼ぶ。

本調査の辞典を用いて結びつきを調べる方法については、桑原（2013）が日本人大学生を対象に調査した「熟語の意味の『透明性』」の数値と比較をした。桑原（2013）が調査対象とした語について、本調査と同じ手順で結びつきを調べたところ、結びつきについて「透明性」の数値と、本調査の 3 種類の分類の傾向に類似が見られた。詳細は 6 で述べる。

４．２　語と漢字の意味の結びつきの分類

本調査では、語と漢字の意味の結びつきを、二字漢語を構成する各漢字ごとに判断して分類する。以下、二字漢語の 1 字目の漢字を「前漢字」、2 字目の漢字を「後漢字」とよぶ。

表 4.1 は「国外」、「提出」と「精神」の例である。「国外」の国語辞典説明文は「国のそと。」である。二字漢語の前漢字である「国」は国語辞典説明文に書かれている。したがって、「国」という漢字と「国外」という漢語は意味が「結びつく」と判断する。後漢字の「外」は国語辞典説明文に「そと」と書かれているので、「外」という漢字と「国外」という漢語は意味が「結びつく」と判断する。1 字ずつの漢字と語の結びつきは「結びつく」、「結びつきにくい」の 2 種類である。同様に見ると、「提出」は後漢字のみ結びつく。

これらの前漢字の結びつき、後漢字の結びつきを 2 字組み合わせて、二字漢語における語と漢字の意味の結びつきを判断する。前漢字、後漢字の 2 字とも漢字と語が結びつく場合、二字漢語における語と漢字 2 字の意味の結びつきを「2 字とも結びつく」とした。どちらの漢字も語と結びつきにくければ、「2 字とも結びつきにくい」、1 字のみ結びつく場合は、語と漢字 2 字の意味の結びつきを「1 字のみ結びつく」とした。

語と各漢字の意味の結びつき		二字漢語における語と漢字の意味の結びつき
・2 字とも、語と漢字が結びつく	→	「2 字とも結びつく」
・2 字とも、語と漢字が結びつきにくい	→	「2 字とも結びつきにくい」
・どちらか 1 字のみ、語と漢字が結びつく	→	「1 字のみ結びつく」

表 4.1　語と漢字の意味の結びつきの例 1「国外」「提出」「精神」

二字漢語	国語辞典説明文『大辞林』より	前漢字		後漢字		二字漢語と漢字の結びつき
		説明との一致部分	語との結びつき	説明との一致部分	語との結びつき	
国外	国のそと。	国	結びつく	そと	結びつく	2 字とも結びつく
提出	文書などをしかるべきところに差し出すこと。	なし	結びつきにくい	（差し）出す	結びつく	1 字のみ結びつく
精神	人間の心。	なし	結びつきにくい	なし	結びつきにくい	2 字とも結びつきにくい

４．３　語と漢字の意味が「結びつく」ときのパターン

語と各漢字が結びつくと判断するのは、大きく分けて 3 パターンである（表 4.1、表 4.2）。

1. 構成する漢字の訓読みが国語辞典説明文に書かれている場合

例）「国外」と「国」、「外（そと）」、「提出」と「出（出す）」、「入院」と「入（はいる）」、「購入」と「入（入れる）」、「重視」と「重（重く）」

2. 構成する漢字を使った漢語が国語辞典説明文に書かれている場合（1 字漢語を含む）
例）「入院」と「院（病院）」
3. 国語辞典説明文には直接書かれていないが、漢字の意味が結びつく場合
例）「購入」と「購（買う）」、「重視」と「視（見る）」、「最高」と「最（いちばん）」

3.に当てはまる漢字は、主に常用漢字表では訓読みがない漢字（例：視、購）である。常用漢字表については、４．６で述べる。その他、「最高」の「最」の訓読みには「もっとも」がある。「一番」は「もっとも」を簡単に言い換えた言葉と考え、語と意味が「結びつく」と考える。

表 4.2　語と漢字の意味の結びつきの例 2「入院」「重視」「購入」「最高」

二字漢語	国語辞典説明文『大辞林』より	前漢字		後漢字		二字漢語と漢字の結びつき
		説明との一致部分	語との結びつき	説明との一致部分	語との結びつき	
入院	治療のために, ある期間病院にはいること。	はいる	結びつく	病院	結びつく	2 字とも結びつく
重視	重く見ること。	重く	結びつく	見る	結びつく	2 字とも結びつく
購入	買い入れること。	買い	結びつく	入れる	結びつく	2 字とも結びつく
最高	高さが一番高いこと。	一番	結びつく	高い	結びつく	2 字とも結びつく

４．４　調査方法

まず、調査方法について述べる。調査対象の語や辞典などについての詳細は次項で述べる。

1. 調査対象の語について、3 冊の辞典からそれぞれ説明文を 1 文取りだし、語義がそろっているか目視で確認する。
2. 前漢字について、1.の説明文と漢字が一致している部分を抜き出す。漢字の一致する部分を含む数文字を検索、抽出し、目視で一致部分を確認する。辞典 3 冊のうち 1 冊以上に４．３で述べた結びつきが見られれば、語義の説明文には、その漢字の意味が含まれており、語と漢字の意味が結びついていると判断する。
3. 上記 2.で 3 冊の辞典いずれにも結びつきが見られなかった語と漢字について、漢字辞典の字義を用いて、字義が一致するかどうかを確認する。４．３のパターンの 3 の「国語辞典説明文」の「一番」を「最も」で言い換えたように、字義の言い換えも確認する。
4. 上記 2 と 3 の結果を合わせて、語と前漢字の意味の結びつきを判断する。辞典 3 冊のうち 1 冊以上に結びつきが見られれば、語と漢字の意味が結びついていると判断する。
5. 後漢字についても、2~4 を同様に行う。
6. 前漢字と後漢字の漢字の結びつきを合わせ、二字漢語と漢字の意味の結びつきを、４．２に従って「2 字とも結びつく」「2 字とも結びつきにくい」「1 字のみ結びつく」に分類する。

４．５　データ

４．５．１　調査対象とする漢字

漢字の表記は、常用漢字表の範囲とした。訓読みも常用漢字表を範囲とした。新聞など一般的な表記の目安にされているためである。

４．５．２　調査対象の語

国立国語研究所「『現代日本語書き言葉均衡コーパス』語彙表[2]」の「短単位語彙表データ」（以下、BCCWJ 語彙表）から、語種が漢語である語を高頻度順に並べ、上位 1000 位までを対象とした。ここから、以下の語は調査対象から除外したため、調査対象の語数は 958 語となった。調査対象から除外した語 42 語は、数詞 18 語（三十、二千など）、語彙素で示された漢字での表記の割合が少ない語[3]8 語（箇月、所為など）、調査に用いた 3 冊の辞典のうち、1 冊以上に見出し語がなかった語 15 語（男女、前年など）、常用漢字表外の漢字を含む語 1 語（勿論）である。3 冊の辞典については、次項で述べる。調査対象の語 958 語で使用されている漢字は、延べ字数 1916 字、異なり字数 743 字である。前漢字は延べ字数 958 字、異なり字数 493 字、後漢字は延べ字数 958 字、異なり字数 482 字である。また、前漢字と後漢字で重複する漢字は延べ字数 1093 字、異なり字数 232 字である。

４．５．３　語義の説明文

本調査では、国語辞典の語義の説明文を利用した。国語辞典の語義の説明文を使用したのは、４．１で述べたように、客観性が保てると考えたからである。しかし、国語辞典 1 冊では、説明に偏りがあると考え、3 冊の辞典を使用した。『大辞林第三版』(三省堂、以下、『大辞林』)、『岩波国語辞典第七版新版』(岩波書店、以下、『岩波』)、『チャレンジ小学漢字辞典第五版コンパクト版』（ベネッセ、以下、『チャレンジ』）である。3 冊は『大辞林』が中型辞典、『岩波』が小型辞典、『チャレンジ』が小学生向けであり、出版社とタイプが異なる辞典であるため、語義の説明の偏りを減らせるのではないかと考えた。

語義の説明の仕方は、辞典によって、様々である。そのため、辞典の語義の説明から取り出す文（「国語辞典説明文」）は原則として、1 文とした。文がなければ、句、語を用いた。文が複数書かれている場合は、1 文目の説明が中心的な意味に近いと判断し、原則として、最初に書かれている文をとるようにした。

４．５．４　品詞、複数の語義、複数の字義の扱い

品詞は、BCCWJ 語彙表の品詞情報に合わせた。語義が複数ある語については、項目番号が小さい、つまり最初のほうに載っている意味がより一般的な意味に近い[4]と判断し、原則として項目番号が「1」の語義を用いた。同じ語でも辞典によって、語義の順番が異なる場合がある。その場合は、原則として、3 冊のうち 2 冊が同じ語義であれば、その語義を用い、3 冊とも異なる場合は、原則として『岩波』の語義を用いた。このように、多義語について

---

[2] 『現代日本語書き言葉均衡コーパス』語彙表（http://pj.ninjal.ac.jp/corpus_center/bccwj/freq-list.html）

[3] 語彙素の漢字での表記が少ない語（8 語）
下記の語は「BCCWJ 中納言」の原文文字列のデータにおいて語彙素での表記の割合が 20%未満であったため、調査の対象から除外した。（　）内は語彙素での表記の割合である。「箇月（0.7%）」、「所為（1.4%）」、「奇麗（1.7%）」、「御免（5.7%）」、「丁度（10.9%）」、「一杯（副詞、11.1%）」、「沢山（副詞、16.7%）」、「沢山（形状詞-一般、16.9%）」語彙素以外での表記とは、例えば、「一杯（副詞）」の場合は、「いっぱい」、「イッパイ」、「一ぱい」、「丁度」の場合は、「ちょうど」、「恰度」、「丁ど」など、ひらがな、カタカナ、漢字とひらがなが混ざったもの、語彙素の表記以外の漢字を使用したものがあった。「奇麗」は、「綺麗」が 28.7%であった。「綺」は常用漢字ではないため調査対象外とした。

[4] 「大辞林第三版」web 版の凡例に以下の記述がある。

「1. 語義解説現代語(1) 意味の記述順序は次のようにした。(ア) 現代語として用いられている意味・用法を先にし、古語としての意味・用法をあとに記述した。(イ) 現代語は一般的な語義を先にし、特殊な語義や専門的な語義をあとに記述した。

（http://www.excite.co.jp/dictionary/japanese/?menu=true）

は、語義を 1 つに決めて調査をした。

「４．４　調査方法 3」での漢字の字義は、漢字辞典『例解学習漢字辞典』（小学館）の字義の説明を使った。字義が複数ある場合は、原則として項目番号が「1」の字義を用いた。

５．結果と考察

５．１　語全体の傾向

二字漢語と漢字の意味の結びつきについて、調査対象語（以下、BCCWJ 高頻度語）全体の割合を表 5.1 に示す。「2 字とも結びつく」と「1 字のみ結びつく」がそれぞれ約 40%、「2 字とも結びつきにくい」が約 20%である。

表 5.1　BCCWJ 高頻度語（958 語）における結びつき（全体）

	2 字とも結びつく	1 字のみ結びつく	2 字とも結びつきにくい	計
語数(%)	396(41.3%)	380(39.7%)	182(19.0%)	958

５．２　頻度の傾向

頻度順上位から 100 語ごとの結びつきの割合を図 5.1 に示す。1-100 の語（以下、最上位 100 語）においては、「1 字のみ結びつく」語の割合がやや高い。101 から 900 までの 100 語ごとの結びつきの割合は、全体の割合とほぼ同様の傾向を示している。このことから、本調査の範囲では、最上位 100 語を除くと、頻度と結びつき方には、大きな違いはないと思われる。901-958 の 58 語は、語数が少ないため、考察の対象としない。

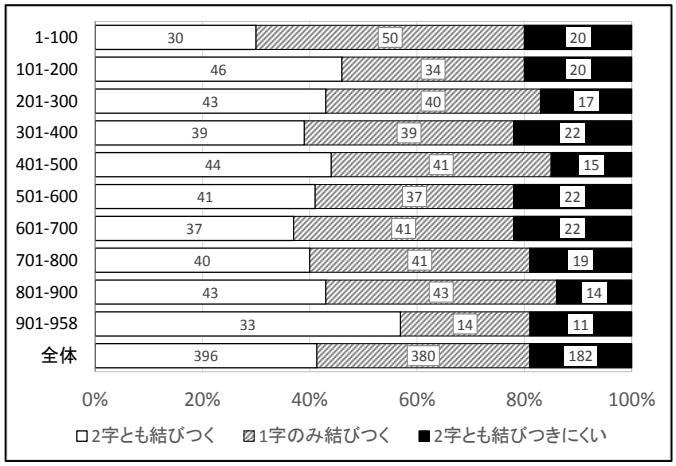


図 5.1　BCCWJ 高頻度語における結びつき（100 語ごと）

５．３　品詞の傾向

次に、品詞別の結びつきについて、全体の結びつきを表 5.2 に示す。全体の語数に占める割合が高い「名詞-普通名詞-一般」と「名詞-普通名詞-サ変可能」における結びつきは、全体の割合とほぼ同様の傾向を示している。「2 字とも結びつく」、「1 字のみ結びつく」の割合は、いずれも 40%程度、「2 字とも結びつきにくい」が 20%程度である。

副詞は、語数が少ないが、「2 字とも結びつく」の割合が低く、「2 字とも結びつきにくい」の割合が高い。副詞 13 語の語と漢字の結びつきは以下の通りである。

・2 字とも結びつく（1 語）是非

・1 字のみ結びつく（3 語）直接、全然、当然
・2 字とも結びつきにくい（9 語）結構、多分、突然、十分、大変、一層、一体、一番、随分
「是非」は「是が非でも。(『岩波』)」という説明文から結びつくに分類した。これらの語は、意味が語源から次第に離れてきた語であると思われる。これらの語がひらがなで表記されることもあるのは、漢字に語の意味が表れていないため、ひらがなで表記したほうが意味を適切に表すことができるという意識が働いているからではないだろうか。上記以外に、4．5．2 で調査対象の語を取り出すとき、語彙素での表記の割合が低い 9 語を対象外としたが、その語の中に、「一杯」や「沢山」などの副詞が含まれている。

また、「2 字とも結びつく」語の割合が比較的高い品詞に「名詞-普通名詞-副詞可能」がある。BCCWJ 高頻度語の範囲では、この品詞には時間の関係や量の関係を表す語が多い。
・2 字とも結びつく語の例　以後、今後、午前、最初、以内、以下、多数
BCCWJ 高頻度語の「名詞-普通名詞-副詞可能」では、語数と比べると異なり字数が少なく、「前」「後」「今」「多」「以」など、同じ漢字が複数回用いられている。これらの字は意味がはっきりしており、語の意味と結びつきやすいため、2 字とも結びつく語の割合が高いと考えられる。

表 5.2　BCCWJ 高頻度語における結びつき（品詞別）

品詞	2字とも結びつく		1字のみ結びつく		2字とも結びつきにくい		計	
	語数	%	語数	%	語数	%	語数	%
名詞-普通名詞-一般	178	42.8%	163	39.2%	75	18.0%	416	43.4%
名詞-普通名詞-サ変可能	155	39.6%	158	40.4%	78	19.9%	391	40.8%
名詞-普通名詞-副詞可能	35	56.5%	20	32.3%	7	11.3%	62	6.5%
形状詞-一般	7	22.6%	19	61.3%	5	16.1%	31	3.2%
名詞-普通名詞-形状詞可能	11	42.3%	9	34.6%	6	23.1%	26	2.7%
副詞	1	7.7%	3	23.1%	9	69.2%	13	1.4%
名詞-普通名詞-サ変形状詞可能	5	38.5%	6	46.2%	2	15.4%	13	1.4%
名詞-普通名詞-助数詞可能	3	75.0%	1	25.0%	0	0.0%	4	0.4%
接続詞	1	100.0%	0	0.0%	0	0.0%	1	0.1%
接尾辞-名詞的-一般	0	0.0%	1	100.0%	0	0.0%	1	0.1%
全体	396	41.3%	380	39.7%	182	19.0%	958	100.0%

5．4　「1 字のみ結びつく」における前漢字と後漢字（品詞別）

結びつきが前漢字に見られるか後漢字に見られるかを見るために、表 5.3 で表 5.2 の「1 字のみ結びつく」について、前漢字と後漢字に分けて結びつきを示す。

「名詞-普通名詞-サ変可能」は後漢字のみ結びつく語が多い。さらに「名詞-普通名詞-サ変可能」158 語について、張（2014）の語構成[5]をもとに分類すると、後漢字のみ結びつく（前漢字は結びつきにくい）語において、前漢字も後漢字も動態類である語の割合が高い（表 5.4）。語を見ると、後漢字に比較的基本的な漢字が使われており、前漢字の意味が明確にわからなくても、何となく語の意味がわかると思われる語が少なくない。詳細な分析は今後の課題としたい。
・後漢字のみ結びつく語の例　参加、追加、提出、輸出、輸入、通知、参考、放送、輸送

[5] 張（2014）は、二字漢語動詞を漢字部分の構成要素の品詞性と構成要素間の関係に従って、AV 型、VN 型、VV 型、MV 型、接辞型に分類している（A=様相類、V=動態類、M=副用類、N=事物類）。

表 5.3　「1 字のみ結びつく語」における前漢字の結びつきと後漢字の結びつき

品詞	1字のみ結びつく(語数)	
	前漢字のみ結びつく	後漢字のみ結びつく
名詞-普通名詞-一般	98	65
名詞-普通名詞-サ変可能	60	98
名詞-普通名詞-副詞可能	9	11
形状詞-一般	11	8
名詞-普通名詞-形状詞可能	4	5
副詞	2	1
名詞-普通名詞-サ変形状詞可能	2	4
名詞-普通名詞-助数詞可能	0	1
接続詞	0	0
接尾辞-名詞的-一般	1	0
計	187	193

表 5.4　「名詞-普通名詞－サ変可能」の語構成の型（張（2014）より）

「名詞-普通名詞-サ変可能」の型(張(2014))	前漢字のみ結びつく		後漢字のみ結びつく	
	語数	%	語数	%
VV型(動態類の組み合わせ)	40	66.7%	77	78.6%
VN型(動態類・事物類の組み合わせ)	10	16.7%	12	12.2%
その他	10	16.6%	10	10.2%
計	60	100.0%	98	100.0%

６．日本人大学生を対象とした調査との比較

4.の調査では、二字漢語の語と構成要素である各漢字の意味の結びつきについて、辞典の語義の説明文を用いて調べた。この方法を桑原（2013）の結果を用いて、日本人大学生を対象にした意味の結びつきについての調査結果と比較した。

６．１　桑原（2013）の調査

桑原（2013）は、2 字の漢字から成る熟語について、「熟語を構成する個々の漢字が熟語の意味とどの程度容易に結びつけられるかを示す指標を、熟語の意味の「透明性」（transparency)」とし、500 語について、日本人大学生 51 名を対象に調査を行い、透明性を数値化した。この調査は「非漢字系学習者の漢字指導に有用なデータ（桑原（2013))」を得るためのものであり、調査対象語の 500 語は桑原が日本語学習者に対する意味の推測過程の調査で用いた語と、語構成や頻度調査の先行研究の中から抽出した語である。

桑原（2013）の調査では、日本人大学生に語のみを提示し、語と漢字の意味の結びつけやすさを 5 段階の尺度評定によって、「1（まったく結びつかない)」から「5（非常に結びつけやすい)」まで、調査票を用い調査した。調査後、調査協力者が回答した 5 段階の数字を平均して、透明性を表す数値「透明度」としている。この調査に際し、桑原（2013）は被調査者に対して、調査の目的が漢字 2 字熟語と各漢字の意味の結びつけやすさを測ることであることを伝え、「登山」と「皮肉」を例に出して説明している[6]。

桑原（2013）は語を提示し、漢字 2 字を合わせて語の意味と結びつけられるかを質問している。漢字の表す意味は質問していない。また辞書を見ないで答えるよう指示している。

４．で行った本研究の調査では、漢字ごとに別々に結びつきをみた点、語義の説明に書か

[6] 具体的な文面は以下の通りである。
「この調査は、漢字 2 字熟語を構成する漢字のそれぞれの意味と、その漢字熟語の意味とがどのぐらい容易に結び付けられるかを調べることを目的としています。たとえば、「登山」は「登」と「山」の 2 つの漢字からできています。「登」と「山」のそれぞれの意味の組み合わせと、「登山」登山の意味は非常に結びつけやすいのではないでしょうか。それに対して、「皮肉」は、「皮」と「肉」からできていますが、「皮」と「肉」のそれぞれの意味の組み合わせと、「皮肉」の意味は結び付けにくいでしょう。
（中略）それぞれの漢字熟語の意味について、その熟語を構成する漢字の意味と「まったく結びつかない」と思ったら「１」、「非常に結びつけやすい」と思ったら「5」として、1 から 5 までの間で適当な数字に○をつけてください。」（下線は桑原による）

れている説明文そのものを、結びつきを判断する際の元のデータにしているという点で、見方や方法に違いがある。しかし、語と漢字の結びつきをみるという点では、目的が重なっており、それを異なる方法で調査したものだと考え、比較を行った。なお、桑原（2013）を比較の対象としたのは、目的が重なっており、被験者数と調査した語数が多く、傾向を比較しやすいと考えたからである。

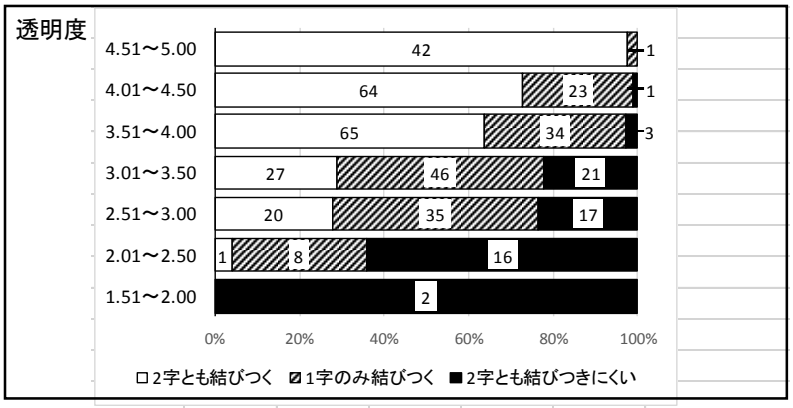
６．２　比較方法

桑原（2013）が調査を行った 500 語について、「茶まめ（unidic-mecab 2.1.2 使用)」で語種を調べたところ、漢語は 453 語であった。そのうち、『大辞林』『岩波』『チャレンジ』の中の 1 冊以上に見出し語として掲載されていなかった語が 26 語（院生、社風、破断、病欠、連泊など)、常用漢字表外の漢字を含む語が 1 語（綺麗）あり、これら 27 語は比較調査の対象外とし、426 語を用いて、比較調査を行った。この 426 語のうち、本調査で調査対象とした BCCWJ 高頻度語に含まれる語は 166 語であった。それ以外の 260 語については、４.で行った調査と同じ手順で語と漢字の結びつきを調べた。

６．３　比較結果

図 6、表 6 は、比較の結果である。桑原（2013）の調査は、1 が「まったく結びつかない」、5 が「非常に結びつけやすい」の 5 段階尺度である。透明度の数値を 0.5 ごとの範囲で区切り、その範囲に含まれる漢語について、４.で行った調査方法による結びつきの割合を示した。

図 6　桑原（2013）の透明度における「2 字とも結びつく」「1 字のみ結びつく」「2 字とも結びつきにくい」の割合



透明度	2字とも結びつく	1字のみ結びつく	2字とも結びつきにくい	計
4.51～5.00	42	1	0	43
4.01～4.50	64	23	1	88
3.51～4.00	65	34	3	102
3.01～3.50	27	46	21	94
2.51～3.00	20	35	17	72
2.01～2.50	1	8	16	25
1.51～2.00	0	0	2	2
1.00～1.50	0	0	0	0
計	219	147	60	426

表 6　桑原（2013）の透明度における「2 字とも結びつく」「1 字のみ結びつく」「2 字とも結びつきにくい」の割合

このグラフを見ると、桑原（2013）の調査では、透明度が高い語には、本研究の調査でも「2 字とも結びつく」に分類される語の割合が高いこと、数値が低くなるにしたがって、「2 字とも結びつく」の割合が減り、「1 字のみ結びつく」と「2 字とも結びつきにくい」の割合が増え、2.50 以下の範囲で、「1 字のみ結びつく」と「2 字とも結びつきにくい」の割合が逆転するという傾向があることがわかる。

このように、漢語と漢字の意味の結びつきを見る方法として、桑原（2013）の日本人大学生への調査結果から透明度を数値化する方法と、本調査で行った辞典を用いて語と漢字の結びつきを分類する方法とで、結果を比較すると、傾向に類似が見られると思われる。

7．まとめ、今後の課題

二字漢語と漢字の意味の結びつきについて、BCCWJ の高頻度語を対象に分析を行った。本調査では、辞典の語義の説明文を用いて結びつきを判断し、分類したが、日本人大学生を対象にした調査結果（桑原（2013））と、同様の傾向が見られることがわかった。

二字漢語と漢字の結びつきについては、以下の特徴が見られた。

(1) BCCWJ 高頻度語の範囲において「語と漢字の意味が 2 字とも結びつく語」、「1 字のみ結びつく語」、「2 字とも結びつきにくい語」の 3 分類の割合は約 2:2:1 の割合であった。

(2) 語の頻度については、高頻度語を頻度順にグループ分けすると、最上位 100 語以外では、3 分類の割合はほぼ一定であった。

(3) 品詞による違いは、「副詞」と「名詞-普通名詞-副詞可能」について結びつきに特徴が見られた。

(4) 全体における語数の割合が高い「名詞-普通名詞-一般」と「名詞-普通名詞-サ変可能」は語の 3 分類の割合は全体の割合とほぼ同様であるが、前漢字と後漢字に分けて結びつきを見ると、結びつきに違いがあることがわかった。特に「名詞-普通名詞-サ変可能」は語構成との関係が示唆された。

今後は、(4) の点から、前漢字と後漢字に分けた結びつきについての詳細な分析を行う。語の意味分野による結びつき方の違いや、1 字ごとの漢字に注目した分析も今後の課題である。また、本研究の日本語教育への活用も模索していきたい。

文献

桑原陽子（2013）「漢字 2 字熟語の意味の透明性の調査」，『福井大学留学生センター紀要』，8, pp. 1-13.

張志剛（2014）『現代日本語の二字漢語動詞の自他』くろしお出版.

野村雅昭（1988）「二字漢語の構造」，『日本語学』7:5, 44-55.

宮島達夫（1968）『単語指導ノート』，むぎ書房.

森岡健二（2004）「現代の漢字調査」，『日本語と漢字』第 4 部, 明治書院, pp. 221-287.

調査資料

『岩波国語辞典第七版新版』，岩波書店（LogoVista 電子辞典シリーズ）

『大辞林第三版』，三省堂　（電子版、検索エンジン「excite 辞書」より取得 http://www.excite.co.jp/dictionary/japanese/）

『チャレンジ 小学漢字辞典 第五版』，ベネッセ.

『学習例解漢字辞典 第七版』，小学館.

テキストの計量語彙論的指標はどのような条件で変化するか

山崎　誠（国立国語研究所言語資源研究系）[1]

Under What Conditions does the Textual Index of Quantitative Lexicology Change?

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)

要旨

テキストにおける TTR(Type/Token Ratio)の値は､そこに使われている普通名詞の使用状況に大きな影響を受けているとされる（山崎：2012）。本稿は、その続編として、テキストの特徴を表す計量語彙論的な指標の一つである TTR がテキストの一貫性という観点から、どのような条件で変動するかを調査した。『現代日本語書き言葉均衡コーパス』（BCCWJ）から抽出したテキストを利用して、語順のランダム化、テキストの合成、テキストのｎ分割などの方法を用い、それぞれの場合に TTR がどのような変動を見せるかを調査した。これらの観察結果から、テキストの一貫性と TTR との関係を考察した。

１．はじめに

テキストを成立させる条件として一貫性と結束性という概念が提唱されている。Halliday＆Hassan（1976）によると、結束性は文法的結束性（指示、代用、省略、接続）と語彙的結束性（繰り返し、関連語）とに分かれるとされる。結束性は文法的結束性を中心に言語学や言語処理の分野で研究が行われているが、一貫性についてはまだ十分に研究が進んでいるとは言えない。とくに一貫性を計量的言語学的に把握する研究が少ないようである。

ところで、結束性と一貫性の関係について、Widdowson（1978）では以下のように述べている。

結束性が関係するのは，さまざまな文構造上の操作によって命題を結びつけ，テクストを形成するところまでである。それに対し，一貫性は，こうした命題の発語内的機能，つまり，報告・描写・説明などのさまざまな種類のディスコースを作り出すために命題がどのように用いられるかということに関係している．（邦訳『コミュニケーションのための言語教育』p.66）

また、結束性と一貫性の関係について、Widdwoson（1978）は、以下の例を示して説明している。

1. A:　What are the police doing?
　　（警察は何をしているのですか.）
　B:　They are arresting the demonstrators.

[1] yamazaki [at] ninjal.ac.jp

（デモの参加者を逮捕しています.）

2. A: What are the police doing?
   B: The fascists are arresting the demonstrators.
   （ファシストらはデモの参加者を逮捕している.）
3. A: What are the police doing?
   B: I have just arrived. （今来たばかりです.）

（前掲書 p.34）

発語内行為のいかんにかかわらず，文と文の間の命題関係が統語的にも意味的にもはっきりと形態上で示されていれば，そこには結束性(cohesion)があることがわかる．したがって，結束性とは文を通して表現された命題間の明らかな関係のことである．一方，命題そのもののつながり具合は必ずしもあきらかでないにしても，その命題そのものが行っている発語内行為の間に何らかの関連を見い出すことができれば，その発話には一貫性(coherence)があると言える．上にあげたやりとりを，これらの用語を用いて説明してみると，1 と 2 には結束性と一貫性の両方があり，3 には結束性はないが，一貫性はあるということになる．（前掲書 p.35）

結束性は個々の言語要素間の関係としてとらえられるため、比較的計量的測定が行いやすいが、一貫性はテキスト内のどの要素を測定すればよいのだろうか。そのためには一貫性がテキスト内のどこに存在するのかを把握する必要がある。上述の 3.A、3.B の例で考えると、一貫性は 3.A と 3.B との間、すなわち文と文との意味的な関係としてとらえることができる。また、テキストは文の連続体であるので、当該のテキスト全体にわたる属性としてとらえることもできるだろう。

本稿では、一貫性が生じる条件として言語要素の出現順序という性質に注目してそれを客観的にとらえる方法を考える。例えば、出現順序を操作した結果の指標の測定値を、もとの測定値と比べるという方法である。

２．一貫性のタイプ

一貫性は当該のテキスト全体にわたって、それを統括する働きを有すると考えられるが、その分布のあり方に応じて 2 つのタイプに大別することができるだろう。そのための準備的考え方としてテキストの構造をトピック（話題）の集まりとしてとらえる。トピックは形式的には段落の形で実現することが多いだろうが、意味的なまとまりであるので必ずしも段落と対応するとは限らないと考えられる。このような考え方のもとに、一貫性のあり方は次の 2 つのタイプを認めることができる。

A トピック内部の一貫性
B トピックを超えた一貫性

A のトピック内部の一貫性とは、あるトピックの中でその内容に関係するものである。例えば、トピックに合った適切な語を選択することや、ある文の次にその文の内容に関連した文をつなげることなどである。B のトピックを超えた一貫性とは、あるトピック全体をと

らえてそれに関連する別のトピックを次に配置することなど、テキストの構造に関係するものである。一般的には、テキスト全体のテーマに従って適切に構成単位を配列することがトピックを超えた一貫性の表れである。いわば、トピックをメタ的に扱う一貫性と言える。

A のトピック内部の一貫性は、トピックのまとまりということへの関与ということから、語の集合である語彙の計量的な特性、例えば語彙の集中度などに現れるのではないかと推測される。一方、B のトピックを超えた一貫性は、構成単位の順序性を測ることによってその一端が測定できるのではないかと期待できる。

B のトピックを超えた一貫性について 2 つ例を挙げる。

(1)吾輩は猫である。うとうととして目がさめると女はいつのまにか、隣のじいさんと話を始めている。私はその人を常に先生と呼んでいた。こんな夢を見た。

(2)『明鏡国語辞典　第二版』より
みつ - ど【密度】〔名〕❶一定の面積・体積などの中にある量が含まれる割合。「人口の―」
❷内容の充実している度合。「―の濃い議論」❸物質の単位体積あたりの質量。
ミッドナイト［midnight］〔名〕真夜中。深夜。
ミッドフィルダー［midfielder］〔名〕サッカーで、ハーフバックのこと。MF。
（原文は縦書き）

(1)は夏目漱石の小説「我が輩は猫である」「三四郎」「こころ」「夢十夜」の冒頭の文を並べた人工的なテキストである。無関係なトピックが連続するため、一貫性は存在しないと考えられるが、仮に最後の文「こんな夢を見た」をそれ以前の文を統括するものと考えれば、やや牽強付会ではあるがトピックを超えた一貫性があるとも解釈できる[2]。また、(1)の末尾に「これらは夏目漱石の作品の冒頭文をつなげたものである。」を付け加えれば、そのことで、トピックを超えた一貫性があると解釈できる。

(2)は国語辞典の一部であるが、連続する見出しは五十音順に並べられているため、それらの間には一貫性はないのが普通である。ただし、その五十音順に並べるという配列規則がここでは、トピックを超えた一貫性であると考えることができる。(2)のような一定の配列のもとに、並べられたテキストを本稿ではリストタイプのテキストと呼ぶことにする。リストタイプのテキストは、辞書がその典型であるが、箇条書きなども含まれる。例えば、『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）では次の表 1 のような例が挙げられる（山崎 2010）。表 1 は短単位で計った 1 語あたりの平均使用度数（n/k 値）の低いサンプルを挙げたものであるが、それらはリストタイプのサンプルであったことが指摘されている。このことからトピックを超えた一貫性は語彙の計量的指標に反映される可能性があることが示唆される。

[2] 3 文目の「その人」が 2 文目の「隣のじいさん」を指すと解釈すればそこに語彙的結束性が存在するとも考えられる。

表１　1語あたりの平均使用度数（n/k 値）が低いサンプル

n/k 値	サンプル ID	NDC	出典名	著編者	文章のタイプ
1.5198	PB17_00159	7 芸術・美術	淡路人形浄瑠璃　伝統芸能　国宝重要文化財等保存事業		リスト（用語集）
1.5771	PB18_00010	8 言語	漢字・仮名・記号テキスト	佐々木光朗	リスト
1.5906	PB2n_00001	分類なし	日本を伝える	梅澤実（監修）	リスト（図録）
1.6018	LBe2_00037	2 歴史	昭和家庭史年表 1926～1989	家庭総合研究会	リスト
1.6683	LBj8_00006	8 言語	日本語キーワード英語表現辞典 日本語の発想で引けて英語表現が豊かになる辞典 名詞編	三省堂編修所	リスト
1.6814	LBo2_00009	2 歴史	1946-1999 売れたものアルバム	Media View	リスト

３．方法とデータ

前節で一貫性は 2 つのタイプに分けることができ、その特徴を利用して一貫性の測定の方法が考えられることを示した。そのことを実現するために、一貫性のないテキストを 2 種類の方法で人工的に作り、それと元のテキストを比べるという方法をとる。その際の比較のための指標は異なり語数の延べ語数に対する比である TTR（Type/Token Ratio）を用いる。TTR は 1 語あたりに平均使用度数の逆数であり、語彙の多様性の指標とされ、コーパス言語学では TTR がよく用いられる。具体的な方法は次の 2 つである。

(3)トピック内部の一貫性については、語をランダムに入れ替え、n-gram による組み合わせを比べる。

(4)トピックを超えた一貫性については、テキストの前半と後半とをそれぞれ別のテキストから選び、トピックを合成して人工的に一貫性を低下させたテキストの TTR 値を元のテキストの TTR 値と比較する。

データは BCCWJ の図書館サブコーパス（LB）から無作為に選んだ 22 テキストである。ただし、TTR 値は延べ語数に影響を受けるため、本発表では短単位・可変長部分が延べ語数で 2,000～2,100 語の範囲に限定している。なお、選択の際は、分野を考慮して各 NDC（図書分類）と分類なしとから 2 テキストずつを選んでいる。

４．考察１

４．１　語順のランダム化

テキスト内に現れる語が一定の順序で現れる通常のテキストと、語順をランダムに並べ替えて一貫性を低下させたテキストとについて、2-gram（＝2 語の連続。但し記号は除外する）の TTR 値を比較した。語順のランダム化の例を(5)(6)に挙げる。(5)のテキストをランダム化したのが(6)である。

(5)吾輩は猫である。名前はまだ無い。どこで生れたかとんと見当が付かぬ。

(6)。見当吾輩。はである生れ名前かどこたぬ付かが。は無いとんとまだで猫

結果を図 1 に示す。ランダム化したテキストでは，元のテキストに比べて 2-gram の TTR

値が有意に高くなることが確認された（t=-20.93,df=21,p<0.001）。

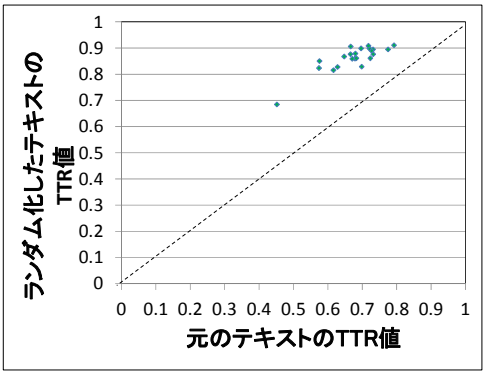


図 1 ランダム化したテキストの TTR 値の増減

４．２　テキストの合成

22 サンプルについて、それぞれのサンプルの前半と別のサンプルの後半を合併した人工的なテキストを作り、その TTR を計測した。全部で 462 のテキストが作成される[3]が、そのテキストの TTR を元となった 2 つのサンプルの TTR の平均値と比較する。そうすると、全 462 テキスト中、元となった 2 つのそれぞれの TTR の値と比べると値が増加しているものが多いが、減少しているものも見られた。ただし、元となったテキストの TTR の平均値と比べると 462 テキスト中 461 テキストで人工的に作成したテキストの TTR の値が増加していることが分かった（平均で 0.028 増加）。結果を表 2 に示す。

表 2　合成テキストの TTR 値

比較する対象	TTR 値が増加	TTR 値が減少
テキスト 1 の TTR 値	359	103
テキスト 2 の TTR 値	370	92
上記 2 つの平均	461	1[4]

実際の分布の様子を図 2 に示す。図 2 の横軸は、1 つめ（前半）のファイルにおける、元の TTR の値と合併したファイルの TTR の値との差であり、縦軸は、2 つめ（後半）のファイルにおける、元の TTR の値と合併したファイルの TTR の値との差である。元のテキストと□□的に作成したテキストの TTR との差には負の相関があることが分かる。

なお、テキストを 3 分割した場合は全 9,241 例の合成テキストのすべてにおいて人工的に合成したテキストの TTR 値がそれを構成する 3 つのテキストの TTR 値の平均を上回った。（平均 0.043 増加）。

[3] 同じテキスト同士の合成は除外したので、22×21 ファイルが対象となる。

[4] NDC8(LBs8_00014)と NDC6(LBb6_00012) の組み合わせである。

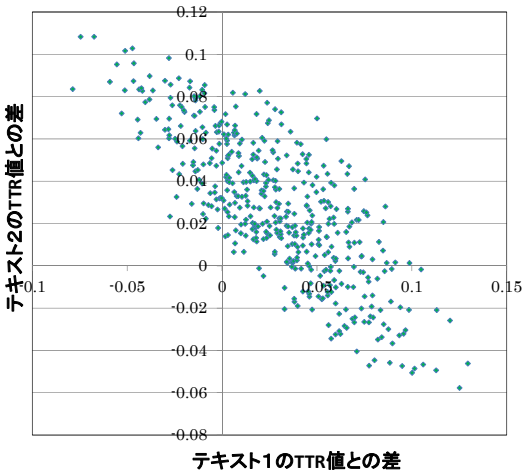


図 2 合成テキストの TTR 値の増減の分布

以上 2 つの事例の結果により、一貫性が低くなると語彙的指標である TTR の値にその影響が現れる場合があることが確認された。しかし、その逆である TTR の値が低くなれば、一貫性が低くなるかこの方法では把握できない。

５．考察２

本節では､テキストをいくつかの区間に分割した場合の TTR 値の変化の様子を観察する。単純に n 分割したもの、n の剰余系により分割[5]したもの、ランダムに n 分割したものの 3 つの人工的テキストについて TTR を計測する。

データは、図書館サブコーパス（LB）の可変長部分の延べ語数（空白・補助記号・記号を除く）が 5,000～5,100 語である 252 ファイルである[6]。

分割数に応じた TTR の値の変化を図 3 に示す。図 3 から、単純に分割した場合よりも

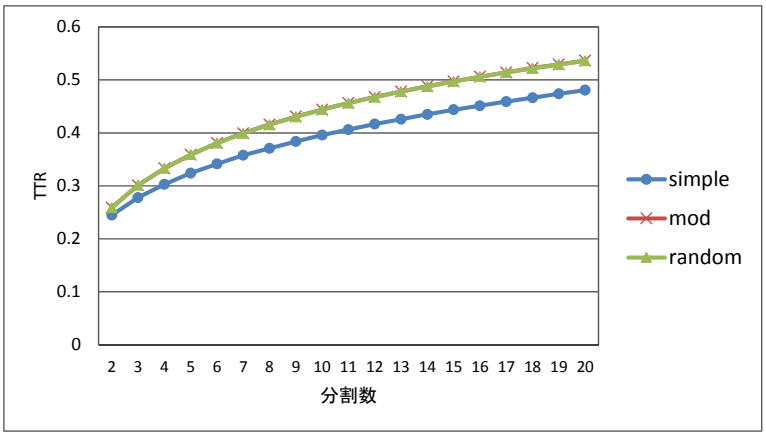


図 3　分割数に応じた TTR の値

[5] テキストを構成する語に先頭から番号を付け、それらを n で割った余りが同じものを一つの語彙として分割したもの。たとえば、2 分割の場合は、偶数番目の語の集合と奇数番目の語の集合とに分かれる。

[6] 各レジスターの内訳は、LB93 個、OB17 個、OL7 個、OP4 個、OT1 個、OW19 個、PB99 個、PM11 個、PN1 個である。

剰余系による分割およびランダムに分割した場合のほうが TTR が高いことが分かる。また、剰余系による分割とランダム分割とには差がないことも見て取れる。図からは、単純な分割と剰余系・ランダム分割との TTR の差[7]は 0.05 くらいに収束しているように見える。

次に n 分割した n 番目の区間の TTR の特徴を見よう。

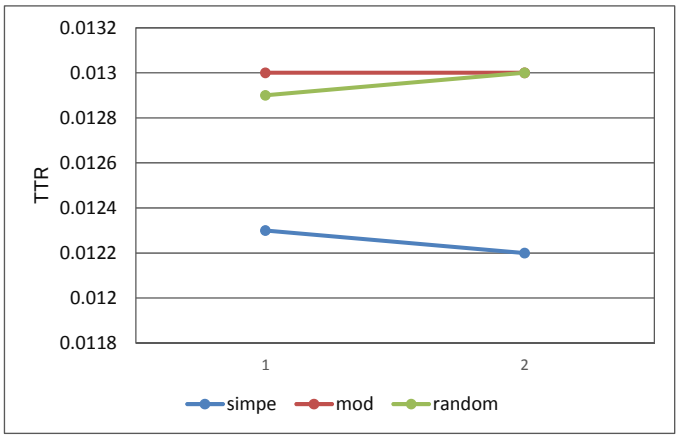


図 4　分割区間ごとの TTR の値（2 分割）

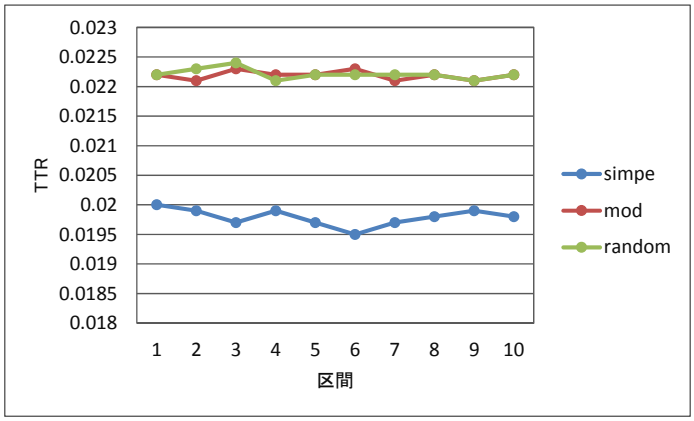


図 5　分割区間ごとの TTR の値（20 分割）

図 4 は 2 分割、図 5 は 20 分割の例である。ここでも、単純な n 分割の場合と剰余系による n 分割、ランダムな n 分割との関係は図 3 と同様である。各区間の TTR の値はランダムに上下しているようであり、特定の傾向は見出しにくい。ただし、区間 1 と区間 2 との関係だけを見てみると、単純な n 分割は、2～20 分割のすべての例において、区間 1 よりも区間 2 の TTR の値が低かったのに対して、そのような傾向を見せるのは、剰余系による n 分割では 9 個、ランダムな n 分割では 11 個であった。このことは、文脈が維持されている場合、冒頭部分から一定の分量の区間は、語の繰り返しが多いことを示唆しているものと思われる。

7 シンプルな分割の TTR から、剰余系による分割の TTR＋ランダムに分割による TTR÷2 を引いた値。

６．まとめと今後の課題

本稿ではテキストの計量語彙論的指標である TTR の値がどのような条件で変化するかを考察した。とくにテキストの一貫性という観点から、文脈がそのまま維持されている場合と文脈が破壊されている場合を比較するという手法で TTR の値を観察した。その結果、文脈を維持せずに人工的に合成したテキストは総じて TTR の値が高くなることが確認された。今回の考察では、剰余系による分割とランダムな分割との間には TTR の差が見いだされなかった（見込みでは幾分かの差があると想定した）。今後の課題としては、文脈がどの程度維持されていれば、TTR の値が維持されるのか、新たな条件を模索することが挙げられる。

謝　辞

本稿は 2013 年 7 月 21 日に行われた、国立国語研究所基幹型プロジェクト「コーパス日本語学の創成」の共同研究発表会で行った発表「テキストの一貫性と計量語彙論的属性との関係」および山崎（印刷中）に加筆・修正したものである。

本研究は国立国語研究所の共同研究プロジェクト「コーパス日本語学の創成」による研究成果の一部である。データとして利用した BCCWJ は、国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度、領域代表者：前川喜久雄）による補助を得て構築したものである。

参考文献

Halliday, M.A.K. and Hasan, R.（1976）Cohesion in English. London:Longman.（邦訳『テクストはどのように構成されるか』、大修館書店、1997 年刊）

Widdowson, H. G.（1978）Teaching Language as Communication. Oxford:Oxford University Press（邦訳『コミュニケーションのための言語教育』研究社出版、1991 年刊）

山崎誠（2010）語の平均使用頻度に現れるテキストの特徴、特定領域研究「日本語コーパス」平成 21 年度公開ワークショップ（研究成果発表会）予稿集 pp.5-14.

山崎誠（2012）Type/Token Ratio と品詞との相関,修剛（編）『新時代的世界日語教育研究』pp.59-64、北京：高等教育出版社

山崎誠（印刷中）テキストの一貫性を表す語彙的指標について、『日語研究』10、北京：商務印書館

外来語「クレーム」の基本語化とその“挫折”

金　愛蘭（広島大学大学院教育学研究科・国立国語研究所共同研究者）†

Failure of Inclusion of the Loanword "Kurêmu" into Japanese Core Vocabulary

Eran Kim (Hiroshima University, NINJAL)

要旨

発表者は，これまで 20 世紀後半の新聞コーパスを用いて，現代日本語語彙における「外来語の基本語化」現象の記述とその理論化を試みてきた。本発表では，その一環として外来語「クレーム」に注目する。自作の 20 世紀後半の通時的新聞コーパスを調査したところ，「クレーム」は 1970 年以降使われるようになり，1991 年ごろまではその使用量を増加させて基本語化に向かうように思われたが，その間も類義語「苦情」「文句」を上回ることはなく，また 2000 年から 2010 年にかけては使用量を大きく減らし，結局，その基本語化は“挫折”したように見える。発表では，その要因・背景として，「クレームをつける」という動詞句を媒介としてマイナスの感情的意味が付着した可能性を指摘し，外来語の基本語化をそれに“挫折”した語によってより多角的に把握し得る可能性について述べる。

1.　はじめに

日本語の，とくに書きことばの基本語彙については，近代以降のマクロな変化の動向が，ある程度明らかにされている。宮島達夫（1967）は，国立国語研究所の「雑誌 90 種の語彙調査」（1956 年）で得られた上位 1000 語が歴史上いつごろから使われているかを調べる中で，明治時代には抽象名詞の漢語が，大正・昭和時代には具体名詞の外来語が現れ，増えた可能性があるとした。また，石井正彦（2013）は，上の 90 種調査と，同じ国語研究所の「月刊雑誌 70 誌の語彙調査」の結果とを比較し，現在は，それに次ぐ第三の段階として，外来語の抽象名詞が増え，基本語彙の中に進出している時期と考えられるとしている。

こうした基本語彙のマクロな変化は，個々の語が新たに基本語彙の仲間入りをする「基本語化」と，逆に基本語彙から外れる「周辺語化」というミクロな変化をその内実としている。しかし，近現代日本語の大規模な通時コーパスが整備されていない状況では，個別の語の使用の変化動向を明らかにすることは容易ではなく，当然，基本語化・周辺語化した語を特定することも困難であった。基本語化・周辺語化は，基本語彙の変化から当然想定される現象であるが，それを実証することはできなかったのである。

そこで，発表者は，現代語の通時的なコーパスを自ら構築して，個別語の「基本語化」現象を実証的に把握・記述する研究を構想・実践してきた。金愛蘭（2011）は，1950 年から 2000 年までの『毎日新聞』について，10 年おきに各年平均 200 万字を超える大規模な「通時的新聞コーパス」を作成し，その語彙調査に基づいてすべての外来語についてその「増加傾向係数」を算出して，20 世紀後半の新聞において基本語化した可能性の高い（抽象的な）外来語を取り出した。また，「トラブル」「ケース」をはじめとするいくつかの外来語について，それぞれの基本語化の過程を，類義語となる和語・漢語との関係をも明ら

† kimeran【at】hiroshima-u.ac.jp

かにしながら記述するとともに，それらの基本語化の背景に，現代の新聞文章の概略化傾向がこうした外来語を基本語として必要としているという見方を提示した。

本発表では，上記研究の一環として，外来語「クレーム」に注目する。具体的には，自作の通時的新聞コーパスを資料に，20 世紀後半の新聞における「クレーム」とその類義語の使用状況を調査し，得られた用例を検討することによって，「クレーム」の基本語化が“挫折”したことを述べる。また，その“挫折”の要因・背景として，「クレームをつける」という動詞句を媒介としてマイナスの感情的意味が付着した可能性について検討する。

2. 資料―「20 世紀後半の通時的新聞コーパス」

調査には，発表者自らが作成した「通時的新聞コーパス」（各年 36 日分増補版）[注1] を用いる。同コーパスは，1950 年から 2010 年までの『毎日新聞』から，ほぼ 10 年おきに，毎月 3 日分（5 日・15 日・25 日），各年 36 日分（全体では 252 日分）の朝刊全紙面の記事（見出しと本文）を，1950～80 年は『縮刷版』からテキスト入力し，1991～2010 年については『ＣＤ－毎日新聞データ集』から抽出して作成したものである（抽出比率は，約 10 分の 1）。コーパスの規模は，表 1（空白は除く）の通り。全体で 2,000 万字近くとなり，ページ数の極端に少なかった 1950 年，やや少なかった 1960 年を除けば，各年ほぼ 300 万字程度の，20 世紀後半（から 21 世紀初頭）の通時コーパスとしては，個別の語の分析にも耐え得るような規模のコーパスを構築することができた。コーパス設計・作成の詳細については，金愛蘭（2011）を参照されたい。

表1　各年の文字数

年	文字数
1950	793,692
1960	2,208,396
1970	3,183,297
1980	3,218,737
1991	3,265,786
2000	3,994,933
2010	3,119,875
計	19,784,716

3. 外来語「クレーム」とその類義語の量的変動

3.1 類義語の範囲

はじめに，「クレーム」の使用量の変動を調査するが，その際，比較のための類義語として，「苦情」と「文句」の使用量も同時に調査する。金愛蘭（2011）で述べたように，類義語の特定は必ずしも容易ではないが，今回は用例数の多いこの 2 語に限定し，他の類義語の可能性[注2] については今後の課題とする。

---

[1] 「通時的新聞コーパス」の作成にあたっては，（財）博報児童教育振興会「第 3 回ことばと教育研究助成」と，文部科学省科学研究費補助金「20 世紀後半の新聞における外来語の基本語化に関する調査研究」（平成 22～23 年度・若手研究Ｂ・課題番号 21720168）および「基本外来語の談話構成機能に関するコーパス言語学的研究」（平成 24～26 年度・若手研究Ｂ・課題番号 23720241）の交付を受けた。本発表では，金愛蘭（2011）の毎月 2 日分を 3 日分に増補し，さらに 2010 年分も加えたものを用いる。

[2] たとえば，国語研究所（2004）『分類語彙表　増補改訂版』の「クレーム」と同じ分類・段落番号（1.3135「批評・弁解」の 06 段落）には，他に「苦情，言い分，申し分，物言い，異議，難癖［～を付ける］，けち，文句，言葉とがめ，ブーイング」がある。

3.2　通時コーパスにおける出現状況

表 2 に，外来語「クレーム」と類義語「苦情」「文句」の，「通時的新聞コーパス」における出現頻度を示す[注3]。これからわかるように，「クレーム」は 1970 年以降使われるようになり，2000 年ごろまではその使用量を増加させて基本語化に向かうように見えるが，その間も類義語「苦情」「文句」を上回ることはなく，また 2010 年には使用量を大きく減らしている（2010 年には「苦情」「文句」も減少するが，その理由は不明）。

図 1 は，表 1 の数値を相対頻度（使用率）として構成比棒グラフに表したものであるが，これを見ると，「クレーム」は，1970 年から 91 年にかけてその勢力（類義語に対する割合）を大きくして基本語化する勢いを見せたものの，2000 年から 2010 年にかけてはその割合を減らし，結局，その基本語化は“挫折”したように見える。

表2　通時コーパスにおける「クレーム」と類義語の出現頻度

	50 年	60 年	70 年	80 年	91 年	00 年	10 年	計
クレーム	0	0	8 (2.5)	11 (3.4)	9 (2.8)	18 (4.5)	2 (0.6)	48
苦情	5 (6.3)	10 (4.5)	27 (8.5)	30 (9.3)	19 (5.8)	55 (13.8)	19 (6.1)	165
文句	2 (2.5)	11 (5.0)	19 (6.0)	15 (4.7)	7 (2.1)	18 (4.5)	8 (2.6)	80

（上段は実数，下段は 100 万字当たりの出現率（換算値，小数点第二位で四捨五入））

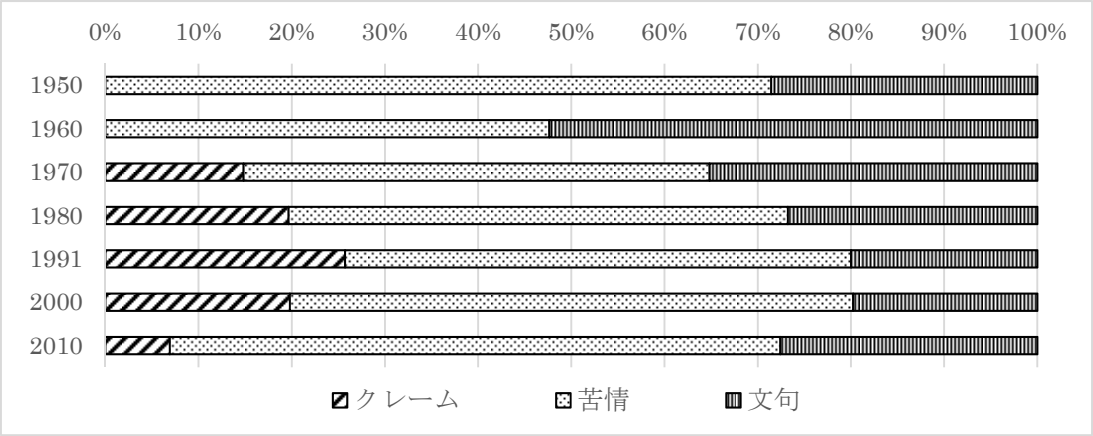


図 1　通時コーパスにおける「クレーム」と類義語の出現頻度

「クレーム」が基本語化に“挫折”したことは，それが使われた紙面の範囲がいったん広がったものの結局狭まったように見えること（表 3），調査期間を通してほとんど自立用法ばかりで，結合用法すなわち造語成分としてはたらくことが広まらなかったこと（表 4）

[3] 「文句」の分析には，「文句なしに，うたい文句，脅し文句，決まり文句」といった慣用句と類意をなさない用例（例：ベストセラーのクリスマス・カードの文句が「ラブ」）は対象外とした。

からも，うかがうことができる（2000 年の結合用法 8 例は，すべて同じ話題の記事におけるもの）。

表 3　「クレーム」の紙面別出現頻度

	50 年	60 年	70 年	80 年	91 年	00 年	10 年	計
社会			4	2	2	7	1	16
経済			2	5	1	3		11
総合				1		7	1	9
第一面			1		2			3
スポーツ				1	1			2
第二面					2			2
第三面						1		1
家庭				1				1
特集			1					1
社説					1			1
政治				1				1

表 4　「クレーム」の自立用法・結合用法の頻度

用法	50 年	60 年	70 年	80 年	91 年	00 年	10 年	計
自立			8	11	9	10	2	42
結合						8		6

4.　“挫折”の背景・要因

用例数が十分ではないため，「クレーム」の基本語化がほんとうに“挫折”したかどうかについては，なお検証の必要がある。ここでは，それを仮説として認めたうえで，その背景ないし要因を考えてみる。

4.1　〈経済〉から〈非経済〉への意味の拡大

『日本国語大辞典』（第二版）には，次のようにある。

> クレーム（英　claim）①貿易などの商品取引で、取引の相手が品質不完全、着荷不足、損傷その他の契約違反をした場合、相手方に対して損害賠償の請求や苦情を申し立てること。＊第２ブラリひょうたん(1950)〈高田保〉商法「通商白書によると、クレイムの四八パーセントが品質不良だとある」②一般に、商品、相手の行為や処置などに対する苦情。＊鏡子の家(1959)〈三島由紀夫〉二「うちの品物はまだクレームをつけられたことがないんだから」③公的団体の立案に対する他の公的団体からの異議申し立て。

これによると，「クレーム」は，主に「商取引などの経済活動上の苦情」という意味合いで 1950 年代から使われているらしい。そこで，「クレーム」の自立用法の使用例を，経

済活動にかかわるもの〈経済〉とかかわらないもの〈非経済〉とに分けて集計すると，表 5 のようになる。

表 5　「クレーム」の意味

用法	50 年	60 年	70 年	80 年	91 年	00 年	10 年	計
〈経済〉			4	3	5	5		17
〈非経済〉			4	8	4	5	2	23

これを見ると，1970 年以降，〈経済〉と〈非経済〉とがほぼ互角に使われ，新聞で使われはじめたころにはすでに，「クレーム」の意味（語義）は，「経済活動上の苦情」から「経済にかかわらない事柄についての苦情」へと拡大していたことがわかる。（1）は〈経済〉の，（2）は〈非経済〉の用例である。（3）は，商取引ではなく貿易全体にかかわる苦情だが，〈経済〉としてよいだろう。

（1）また某商社は、昨年輸入したソ連材が契約した量に足りないとクレームをつけたところ、その後の木材輸入商談ではピシャリと締出しを食うという報復を受けた。
（2）七四パーセントに及ぶ民主主義肯定の中で、その実践について、問９、１０にみられるほど多くの人々がクレームをつけるのはなぜだろう。
（3）第二は輸出の二割を占める欧州で、日本からの輸出急増をめぐって、欧州工作機械工業連合委員会代表者がさきごろ来日し、クレームをつけるなど貿易摩擦が持ちあがっている点である。

このような意味の拡大は，「クレーム」の基本語化にかなう変化である。すなわち，抽象名詞の外来語の基本語化は，意味がより抽象化・概括化して類義語の上位語の位置に立つことにより，その使用量を増大させるからである。しかし，「クレーム」は意味が拡大しているにもかかわらず，基本語化しなかった。それは，なぜだろうか。

4.2　マイナスの感情的意味の付着

「クレーム」の自立用法を，前後の語との共起関係という観点から分けると，表 6 のように，後続の動詞と結びついて動詞句を構成するものが 40 例中 31 例と圧倒的に多い。その中でも，他動詞句「クレームをつける」と自動詞句「クレームがつく」が明らかに多い（前者には受け身の例も含める）。このうち，「クレームをつける」は，1970 年から 91 年まで使われるが，それ以降は見られない。

さらに，この「クレームをつける」は，1970 年・80 年あたりでは，先の用例（1）～（3）のように，〈経済〉であれ〈非経済〉であれ，「クレーム」の持ち主（仕手）が組織や集団あるいはその代表者であるためか，個人が「文句をつける」といった意味合いは感じられない。しかし，1991 年の次の例（4）では，持ち主が個人[注4] であるために，そのような

[4] 個人が個人へ向けたものとして，次のような例があった。
（例）インフルエンザで１週間も休園している孫が「退屈だからビデオを借りてきて」と夫に電話で頼んできました。３本で５００円とのこと。指定されたビデオを届けたのですが、あとで

ニュアンスがあるようにも感じられる。

表 6　「クレーム」の用法

	50 年	60 年	70 年	80 年	91 年	00 年	10 年	計
名詞句ほか			1	3	2	2		8
動詞句								
～をつける			4	6	2			12
～がつく			1	2	2	3		8
～を出す			1					1
～が通る			1					1
～を送る[注5]					2			2
～が入る						1		1
～がない						2		2
～がある						1		1
～が来る						1		1
～が相次ぐ							1	1
～がつながる							1	1
中止用法					1			1

(4) 今回の組閣は、宮沢新総裁が決まった十月二十七日から、臨時国会初日の首相指名の五日まで「間（ま）がありすぎる」（斎藤氏）ことが特徴だが、もうひとつ、閣僚人事をめぐるヤマのような情報の中に、宮沢氏の肉声がないことだ。すでに、渡辺美智雄氏の副総理兼外相、羽田孜氏の蔵相起用が内定。他の主要閣僚ポストも党内各派への割り振りと派閥推薦閣僚候補をあてはめる作業が進んでいるが、調整の中で、宮沢氏が拒否したり、クレームをつけたなどのうわさもない。

こうした見方は，もちろん，「クレームをつける」が，「文句をつける」「言いがかりをつける」「いちゃもんをつける」などと同じ「～をつける」という形式を持ち，そのために，これらが持っているマイナスの感情的意味を付着させてしまったのではないかと解釈できる，ということである。「クレームをつける」も，70 年・80 年あたりはまだそうしたマイナス語感の付着はなかったのかもしれないが，91 年にはそうした傾向が現れつつあったものと思われる。

もしそうだとすると，こうしたマイナスの感情的意味は，当然，「クレーム」という名詞そのものにも付着することになるだろう。以下の例で，「厳しいクレーム」「激しいクレーム」という表現は，そうしたことを間接的に示しているように思われる。

孫からクレームがつきました。「バンビと言ったのに、じいちゃん、ゾンビを借りてきた」（大分市・６０歳）［2000 年 3 月 5 日総合］

[5] 「請求書を送る」の例。

（例）エネルギア側は今年一月三十日付で契約代金全額支払いを求める請求書（クレーム）をパ社に送っているが、未払いのまま。［1991 年 5 月 5 日第一面］

（5）これに先輩の政治記者から厳しいクレームが相次いだことを紹介した。当時から“変人”扱いだった小泉さんにも厳しかったが、何といってもベスト３の鳩山、船田、谷垣３氏には「記者まで一緒に素人では困る」「彼らに激動期を乗り切る資質があるとは思えない」など、要するに「頼りない」という批評が相次いだ。
（6）学校や保育園など子どもを預かる施設が気に掛けるのが、親との関係だ。親の激しいクレームにつながることもある。

4.3　他の動詞句やサ変動詞化の可能性

要するに，「クレーム」は，〈経済〉から〈非経済〉へと意味を拡大し，それに伴って使用量を増やして基本語化の方向に向かいかけたが，その多くが「クレームをつける」という動詞句であったために，「文句をつける」などが持つマイナスの感情的意味を付着させてしまい，より抽象的な意味を持つ（類義語の）上位語として基本語化することができなくなってしまったのではないか，ということである。

では，なぜ，「クレーム」の動詞句に「～をつける」という形式が選ばれたのだろうか。もし，「クレームを言う」など別の動詞との結びつきを採用していたら，あるいはまた，「クレームする」というサ変動詞を成立させていたら，「クレーム」は基本語化していたかもしれない。このうち，サ変動詞については，BCCWJ を検索すると 26 例[注6] が得られ，そのほとんどが特許関係の専門語ないしジャーゴンとして使われている。前後関係は明らかではないが，「クレームする」が専門分野で使われてしまえば，それが一般語として採用される可能性は少なくなるだろう。

5.　“挫折語”からみる基本語化

以上，本発表では，外来語「クレーム」が基本語化に“挫折”した要因・背景として，「クレームをつける」という動詞句が，「文句をつける」などと共起動詞を同じくする形式であったことから，それらが持つマイナスの感情的意味を付着させてしまい，その結果，「クレーム」そのものにも同じ感情的意味が付着してしまったために，より抽象的で広い意味を持つ（類義語の）上位語として基本語化することができなかったのではないかと推測した。もちろん，これは仮説であり，今後，別に検証していく必要がある。ただ，そうではあっても，基本語化に“挫折”した外来語が，基本語化の条件や要因を検討するうえで，有用な手がかりを提供してくれることは間違いないように思われる。

付　記

本研究は，文部科学省科学研究費補助金「近現代日本語彙における『基本語化』現象の記述と類型化」（2014 年度～2016 年度，基盤研究 C，研究代表者：金愛蘭）および国立国語研究所「多角的アプローチによる現代日本語の動態の解明」（基幹型プロジェクト，2009 年

[6] 今回の新聞データでは出現しなかったが，国立国語研究所の BCCWJ（検索ツールは，中納言を利用）にはサ変動詞の用例があった。なお，「クレイム」という表記をするものも 5 例あった。
（例）既に述べたように，多項制のメリットは１つの発明を多面的な観点からクレームして保護できるところにある。#明細書の作成にあたっては，このことを十分に活用すべきであろう。
［LBs5_00009，竹田和彦（2004）『特許の知識』ダイヤモンド社］

度～2015 年度予定，研究代表者：相澤正夫）による研究成果の一部である。

文　献

石井正彦（2013）「和語・漢語・外来語―基本語彙に見る攻防―」『日本語学』32-11
金愛蘭（2006a）「外来語『トラブル』の基本語化－20 世紀後半の新聞記事における－」『日本語の研究』2 巻 2 号
金愛蘭（2006b）「新聞の基本外来語『ケース』の意味・用法－類義語『事例』『例』『場合』との比較－」『計量国語学』25 巻 4 号
金愛蘭（2011）『20 世紀後半の新聞語彙における外来語の基本語化』『阪大日本語研究』別冊 3 号
金愛蘭（2013）「外来語動名詞「チェック」の基本語化－通時的新聞コーパス調査と意識調査の結果から－」相澤正夫編『現代日本語の動態研究』おうふう
金愛蘭（2015）「基本語彙構造における外来語の進出領域」斎藤倫明・石井正彦『日本語語彙へのアプローチ―形態・統語・計量・歴史・対照―』おうふう
国立国語研究所（2004）『分類語彙表　増補改訂版』大日本図書
田中牧郎（2013）『近代書き言葉はこうしてできた』岩波書店
宮島達夫（1967）「現代語いの形成」『ことばの研究　第３集』国立国語研究所

関連 URL

現代日本語書き言葉均衡コーパス 中納言 1.1.0　https://chunagon.ninjal.ac.jp/

『理工学系話し言葉コーパス』における後置詞の特徴
―中級日本語教材をアカデミックなコミュニケーション能力につなげるために―

宮部 真由美（文教大学文学部･東京大学大学院工学系研究科）†
菅谷 有子（文教大学文学部･東京大学大学院工学系研究科）
遠藤 直子（広島工業大学工学部）
中村 亜美（東京大学大学院工学系研究科）

A Study of the Characteristics of Postpositions in "The Science and Engineering Spoken Japanese Corpus": Connecting Intermediate Japanese Teaching Materials to Academic Communication Skills

Mayumi Miyabe, Yuko Sugaya (Bunkyo University･The University of Tokyo)
Naoko Endo (Hiroshima Institute of Technology)
Ami Nakamura (The University of Tokyo)

要旨

本発表は，東京大学大学院の理工学系のゼミにおける研究発表と質疑応答などの自然発話を資源として構築した『理工学系話し言葉コーパス』について分析を行なったものである。具体的には，中級レベルの学習者にとって，学習優先度が高いと思われる後置詞(複合辞)が，上記コーパスにどのように現われているかを，量･質の両面において調査し，その結果を踏まえ，市販の中級レベルの教科書，親しいもの同士の雑談が採集されている『名大会話コーパス』との比較を行なった。そして，日本語の学習途上にある留学生が，少しでも早い段階からゼミでの発表や質疑応答を含むディスカッションにおける日本語の理解と使用が可能となるよう，後置詞の学習・指導に関してどのような視点が必要であるか，また，既存の教科書をどのように補完すればいいのかを考察した。

１．はじめに

東京大学大学院工学系研究科コーパスチームでは『理工学系話し言葉コーパス』を構築している。このコーパスは 7 分野[1]の研究室のゼミにおける会話を 5 年にわたって収録したなかから，主に母語話者の自然発話を収録したものである。7 分野の収録時間は 153 時間で，テキスト化したコーパスの延べ形態素数は 1,550,954，異なり形態素数は 16,485 である。

この発表では，理工学系の学生に対する日本語指導を考えた場合の観点の一つとして，後置詞をとりあげ，『理工学系話し言葉コーパス』での実際の使用の状況と中級の日本語教科書での扱われかたをみながら考察していくことにする。

２．研究の目的

発表者が担当する日本語クラスは理工学系の学生を対象とするもので，クラスを受講する学生のほぼすべてが，自分の専門に関して，英語で授業を履修することができ，また論文も英語で執筆することが可能である。こうした環境ではあるが，日本語クラスを受講す

† z5000926@k.bunkyo.ac.jp

1 7 分野とは，電気系工学，都市環境工学，都市計画，建築学，社会基盤学，化学システム工学，情報理工学系である。

る学生は，生活のための日本語以外に研究のための日本語も理解したいと感じており，具体的には同じ研究室の学生とのコミュニケーションや，日本語が用いられる研究場面(例えば，研究室やゼミでの会話)で情報を得，コミュニケーションに参加したいという願望を持っている。しかし，研究場面での日本語は話しことばとはいえ，アカデミックな場面における日本語であるため，中級以前のレベルの日本語の力では理解することもままならないということが，学生たちへのインタビュー調査からあきらかになった。しかしながら，日本語で話されているすべてがわからなくても，何の話題について話しているかということだけでもわかれば，自分の専門分野の話であれば，予測しながら理解することができるということもわかった。

そこで，今回，「(に)ついて」，「(に)関して」，「(に)対して」などのような後置詞を分析対象とし，量的に多く用いられているものや，談話のトピックやテーマを表わすものを中心に，後置詞が『理工学系話し言葉コーパス』にどのように用いられているかを調べ，教育現場へのフィードバックを探ることとした。

３．分析の方法

後置詞とは「単独では文の部分とはならず，名詞の格の形(およびその他の単語の名詞相当の形式)とくみあわさって，その名詞のほかの単語に対する関係を表わすために発達した補助的な単語である」(鈴木重幸(1972:499))。本発表では，『理工学系話し言葉コーパス』から，下記 20 個の後置詞を抽出する[2]。そして，抽出した後置詞のうち，数の多い上位の後置詞について分析を行なう。また，『理工学系話し言葉コーパス』に出現する後置詞との比較のために，親しいもの同士の雑談が採集されている『名大会話コーパス』(名古屋大学)および，中級レベルの日本語教科書 7 冊[3]に出現した後置詞についてもみてみることにする[4]。

表 1 抽出した後置詞

(に)	おいて	(と)	して	(を)	おいて	(の)	おかげで
	ついて		いっしょに		もって		ために
	つき		ともに		めぐって		くせに
	とって				とおして		
	むけて						
	むかって						
	よって						
	対して						
	関して						
	つれて						

[2] 研究の対象とした後置詞は，高橋太郎ほか(2005:185)に挙げられている連用形式の 20 個の後置詞とした。高橋太郎ほか(2005)では，そのほかに連体形式のもの(「(に)おける」，「(に)おいての」など)や，「とりたて的なはたらきをもつ後置詞」(「(から)いえば」，「(から)みれば」など)があげられている。

[3] 次の 7 冊である。『テーマ別 中級から学ぶ日本語』研究社(1～23 課)，『科学技術基礎日本語 留学生･技術研修生のための使える日本語―読解編―』金沢工業大学 (1～13 課)，『新中級から上級の日本語』The Japan Times，『中級を学ぼう(前期)』スリーエーネットワーク(1～8 課)，『中級を学ぼう(後期)』スリーエーネットワーク(1～10 課)，『中･上級のための日本語読解』文教大学出版事業部(1～12 課)，『大学･大学院 留学生の日本語①読解編Ⅰ』アルク(1～14 課)

[4] この二つのコーパスと比較を試みる理由は，『理工学系話し言葉コーパス』がゼミでの発表を含む質疑応答のセミフォーマルな自然発話であるのに対して，『名大会話コーパス』は日常的なインフォーマルな会話であり，中級レベルの日本語教科書は規範的な日本語の書きことばであることより，典型的に種類の異なるコーパスとで比較が可能であると考えたからである。

４．調査結果・分析結果

４．１．各コーパスにおける後置詞の現れかた

３節であげた 20 個の後置詞は，各コーパスに，表 2 にあげるように現れていた。表 2 では『理工学系話し言葉コーパス』での出現数が多い順にあげることにする。

表 2 後置詞の現れかた[5]

		理工学系話し言葉コーパス	名大会話コーパス	中級教科書(7 冊)
1	(と)して	2,476(111)	243	78(1)
2	(に)ついて	1,216(25)	83	47
3	(に)よって	1,178(5)	96	53
4	(に)対して	587	50	22
5	(に)関して	549(37)	27(1)	8
6	(に)おいて	285(17)	8	9
7	(の)ために	120	67	18
8	(に)とって	79	59(4)	22
9	(と)ともに	30	1	8
10	(と)いっしょに	29	83	9
11	(を)もって	27	1	0
12	(を)とおして	14	1	6
13	(に)つき	6	0	0
14	(の)おかげで	4	4	3
15	(に)つれて	2	0	1
16	(を)めぐって	1	0	1
17	(に)むけて	0	0	0
18	(に)むかって	0	0	0
19	(を)おいて	0	0	0
20	(の)くせに	0	4	1
各コーパスの総形態素数		1,550,954	1,924,289	62,068

各コーパスの大きさが異なるため，表 2 に提示した数値で単純に比較はできないものの，『理工学系話し言葉コーパス』の上位の後置詞は，ほかのコーパスと比較して明らかに数量が多いことがわかるだろう[6]。

『理工学系話し言葉コーパス』の上位の後置詞についてみてみると，「(に)ついて」，「(に)対して」，「(に)関して」の後置詞は，これで示される文の部分が，述語に対する広い意味で対象をさしだしている。また，ある場合には，その文を含む談話におけるテーマやトピ

[5] 表内のカッコ内の数値は，「(に)つきまして」などのように，丁寧な形で現われていたものの数である。なお，この数値はカッコ外の数値に含まれている。

[6] 教科書は学習のためにコントロールされた日本語であるといえ，いくつかの後置詞が一通り現れるような構成となっていることから，本来は量的な分析には向いていないといえる。

ックをさしだすこともあり，この後置詞を含む文の部分の情報が取得できるかどうかは，ゼミで話されている内容が何であるかということの理解に重要なポイントとなるといえる。さらに，表2の『理工学系話し言葉コーパス』でもっとも多く用いられている「(と)して」は，その文の述語で述べられることがらに対する立場･役割をさしだすもので，話されている内容のより正確な理解という点を考えると，この部分の情報の取得ができることも重要であるといえる。

次からの節では，『理工学系話し言葉コーパス』の上位の6つの後置詞「(と)して」，「(に)ついて」，「(に)関して」，「(に)対して」，「(に)よって」，「(に)おいて」について，個別に行なった分析の結果を述べていく。

４．２．「(と)して」

「(と)して」は，『理工学系話し言葉コーパス』においてもっとも多く用いられていた後置詞である。また，どのコーパスにおいてもこの後置詞はみられ，量的な点でもほかの後置詞よりも，多く用いられていることがわかる。

そして，これら3つのコーパスを比較した際，「(と)して」はいずれのコーパスでも立場や役割としての用法が中心であったが，『理工学系話し言葉コーパス』では「結果として」(74例)，「方法として」(52例)，「研究として」(34例)，「目的として」(31例)，「特徴として」(28例)，「例として」(27例)，「前提として」(15例)など[7]のように，繰り返し用いられるものがあった。

「(と)して」は，日本語記述文法研究会(2009:99)によると，「役割(述語で表わされる事態の成立にあたっての，主体や対象が担う働きのこと)」を表わすものであると述べられており，「留学生として日本に来た」，「豚をペットとして飼っている」など，そのほかにもさまざまな用例があげられているが[8]，日本語記述文法研究会(2009:99)にあげられているさまざまな「(と)して」の一つ一つの意味をとらえることは難しい。そのため，上であげたようなまとまった表現となっているものを，そのかたまり(慣用的な言いまわし)として，この後置詞を示すことは指導の一つとして有効ではないかと思われる。また，丁寧な形である「(と)しまして」(111例)もみられた。

４．３．「(に)ついて」

いずれのコーパスにも，「～について」の部分が，言語活動や思考活動を表わす述語に対する対象をさしだしている用例が多くみられた。

(1) 「・・・えーと，現在用いてる，えー，ウィルス濃縮方法の概要について述べさせていただきますと・・・」(都市環境工学)
(2) 「・・・このバッテリー側から UPFC に供給されている有効電力についても考えなければならないので，・・・」(電気系工学)

『理工学系話し言葉コーパス』で際立っていたのは，上記の例を含め，(3)，(4)のように，ニ格部分に文相当の句がくる用例が多かった点である(48例)。

---

7 と格の名詞が修飾をうけて，名詞句となっている用例も多くある。
8 例えば，「校長は監督責任者としてつらい状況に置かれている」，「お礼として手紙を書く」，「緊急の対策として予防注射を実施した」など。

(3) 「・・・なんでこのように，新しい位置にピークが出てきたかというのについて，えーと，ちょっと考察をしてみたのですが，・・・」(化学システム工学)
(4) 「・・・その，そういった手法がどうしたら今後広がっていくかっていうことについて仮説していこうという，えー，ことになりそうです」(建築学)

また，ニ格に「そこ，このこと，そのこと，こちら，そちら，これ，これら」(33 例)，「それ」(27 例)のような代名詞となっている用例も多かった。

これら代名詞の用例や(3)，(4)の用例などは，読解文などのような書きことばで提示されれば，その前の文･段落について時間をかけて確認することが可能であるが，話しことばの場合はそういうわけにはいかない。また，ゼミのようなアカデミックな場面では，内容も抽象的であるため，こうした場合の音声的に長い名詞句を含む後置詞部分の理解は難しいだろう。また，丁寧な形である「(に)つきまして」(25 例)も用いられていた。

４．４．「(に)関して」

「(に)関して」は，「(に)ついて」と同様に，言語活動や思考活動を表わす述語に対する対象をさしだす後置詞である。『理工学系話し言葉コーパス』の「(に)関して」の用例は，その 70%弱が「～に関しては」のように，取り立てられた形で用いられており，(5)のように「～に関しては」の部分はその時点での話題･主題であるものとしてさしだしている。

(5) 「この調査対象，この調査に関しては主に 2 つの点を，えー，ちょう，調査目的としました」(建築学)

そして，「(に)ついて」ではなく，「(に)関して」を用いることで，その話題･主題を，ニ格に表わされる名詞に関連･関係するものとしてさしだしている。多くの場合，「(に)関して」は「(に)ついて」と置き換えが可能であるようにと思われるのであるが，「(に)ついて」を用いると，ニ格に表わされるものが言語活動や思考活動の対象そのものであり，「(に)関して」を用いた場合の対象周辺のことがらも含むというような広がりは感じられない。

また，ニ格部分に文相当の句がくる用例(33 例)，ニ格に「ここ，そのこと，こちら，あれ，これ，これら，それ」(80 例) のような代名詞となっている用例も多かった。丁寧な形である「(に)関しまして」(37 例)も用いられていた。

４．５．「(に)対して」

上の二つの後置詞とは異なり，「(に)対して」の対象とは「働きかけの目当てとして」(日本語記述文法研究会(2009:45))の対象である。(6)のように，述語に表わされる動詞などがはたらきかけていく対象を表わす。

(6) 「居住履歴っていうものも，住環境に対して要求する，その個人的な，価値観であったり，えー，理想とする住環境であったり，そういうものに影響を及ぼす」(都市計画)

「～に対して」の部分がこのような対象を表わす用例は，いずれのコーパスにおいても，

この後置詞の用法としてもっとも多く用いられている。ただし，『理工学系話し言葉コーパス』では，ニ格部分に文相当の句がくる用例(53 例)，ニ格に代名詞がくる用例(118 例)も多かった。

(7) 「で，その，け，環境，景観を保全するっていうことに対して支払っていう名目がたっているんですけど，・・・」(都市計画)
(8) 「で，これに対して，最後に海浜モデルの推定モデルを適用します」(社会基盤学)

また，「～に対して」の部分が次のように割合や対比を表わす用例が，ほかのコーパスより比較的多く用いられていた。

(9) 「このようにひとつの送電線に対して複数の TCSC が影響をもつ場合・・・」(電気系工学)
(10) 「・・・現状の問題点として計画移転世帯 5000 世帯に対して，移転世帯が 385 世帯にとどまっている」(都市計画)
(11) 「で，自然由来の godolinium は主にコロイドに付着しているのに対して，人為起源の godolinium は安定の錯体でありまして，えー，通常の下水処理過程では除去されないという報告があります」(都市環境工学)

４．６．「(に)よって」

教科書には，「(に)よって」が原因･理由を表わすものや手段を表わすもの，対応を表わすものが用いられている。『理工学系話し言葉コーパス』でも，原因･理由を表わすもの(用例(12))，手段を表わすもの(用例(13))，対応を表わすもの(用例(14))が，それぞれみられた。

(12) 「電力網においては，電力が遠回りに送電されることによって，余計な損失が生じたり，過負荷送電線が生じるという現状があります。」(電気系工学)
(13) 「ファジイ理論は広く知られていますように，数学的なモデルを必要とせず経験的知識によって入出力の関係を調整することができるという特徴があります。」(電気系工学)
(14) 「衛星画像を用いて海岸線の変化を見るっていうのを中心に考えていて，ただ，その中でも，プリズムとパルサーによって，ま，見える色が違う。」(社会基盤学)

初級レベルの日本語学習の段階では，「(に)よって」は受け身文と一緒に学習する。この場合の「～によって」は，基本的には受け身文の述語が表わす動作の動作主をさしだす。『理工学系話し言葉コーパス』でも，「～によって」が受け身文の動作主を表わす用例もあったが，(15)や(16)のように，受け身文であっても，「～によって」が原因や手段を表わしているものの方が多くあった。

(15) 「・・・その，低い堤防は，えー，まあ，津波によって多くが壊されて，で，その後ろの・・・」(社会基盤学)
(16) 「だから，それは何らかの手段によって，そこの地域はこういうふうに保全されるべきとか，こういうふうに活用されるべきっていう・・・」(都市計画)

４．７．「(に)おいて」

日本語記述文法研究会(2009)には，「動きの場所を表わす」(p.55)場合と「事態の成立する領域」(p.94)を表わす場合とがあると述べられているが，どちらの用例も調査対象とした三つのコーパスにおいてみられた。『理工学系話し言葉コーパス』では，「事態の成立する領域」を表わす場合，(17)，(18)，(19)のように，ニ格の名詞にはさまざまな抽象名詞が用いられていた。

(17) 「短時間フーリエ変換，フーリエ分析においては，えー，時間窓のとり方が重要になるので・・・」(化学システム工学)
(18) 「実際にサンプリングした期間においては，えー、大腸菌群濃度っていうのは，10の１から10の４乗。」(都市環境工学)
(19) 「図の 1-12 の通常軸で表したグラフにおいては，えー，ＴＮと近い挙動を示していました。」(都市環境工学)

５．分析のまとめと日本語教育の現場への応用

『理工学系話し言葉コーパス』の後置詞の特徴は，４節に述べたとおりである。『名大会話コーパス』や中級レベルの教科書と比べると，後置詞の種類やそれぞれの後置詞がもつ用法の種類に大きな違いはなかったといえるが，４節に述べたように『理工学系話し言葉コーパス』に特徴的なこともあった。以下では，その特徴について，日本語教育との関連において述べていくことにする。

後置詞は，その後置詞を含む文の部分が，ほかの文の部分に対してどのような関係にあるかを明確にする機能があり，読解を中心に学習が進められる中級レベルの日本語教育では必須の学習項目であるといえる。中級レベルの総合クラスでは，このレベルの教科書のつくりの多くが読解本文を提示し，それを軸にして学習が進められる。このような書きことばの文章では，読み手は幾度となく読み返すことができるため，文脈指示の代名詞や後置詞を含む文の部分が示す内容をとらえることは，時間をかければ可能である。

そして，今回の『理工学系話し言葉コーパス』における調査･分析で明らかになったことは，アカデミックな現場での発話では，①「～ということ／もの／の／ところ」など文相当の句がニ格名詞句にあらわれ，その場面での話題に関連するマーカーとしてはたらく場合があること。したがって，中級レベルの話しことばの学習には，このような点により重点を置いた指導や教材開発が必要であるといえる。また，②話しことばの指示詞が文脈指示として用いられていること，③「結果として，目的として」のように繰り返し用いられ，研究場面で使用される談話構成のキーワードとなる論理的な表現であるものはひとまとまりの表現として学ぶという方法を取るのが有効ではないかということがみえてきた。こうした点を考慮し，学習者に音声レベル(話しことば)における理解をうながすような学習･指導も必要ではないかと考える。

また，こうした指導では，学習者に身近な専門的な語彙をあわせて提示するような配慮も必要であり，教育現場では汎用的なアカデミックな用語･表現のみならず，個別の専門分野に対応できる教材の開発が求められるだろう。

６．おわりに

『理工学系話し言葉コーパス』は，『名大会話コーパス』と同様，話しことばのデータ

でありながら，後置詞の使用頻度が高い。これは『理工学系話し言葉コーパス』で扱われているトピックがアカデミックな内容であるためだと考えられる。後置詞を用いることで，その後置詞を含む文の部分が，ほかの文の部分に対してどのような関係にあるかを明確にしているからであろう。

一方で，『理工学系話し言葉コーパス』では，「名詞＋後置詞」という単純な構造ではなく，ニ格部分に文相当の句がくる用例も多かった。このことは，アカデミックな場面の話しことばにおいて，後置詞が聞き手の頭の中の情報をいったん保留させ，整理しなおす機会を与えている可能性もある。この点に関しては十分な分析ができなかったが，このような後置詞を含む句の談話的な機能の点にも意識させながら，中級レベルの日本語学習者に後置詞を含む長い文を理解し，産出させることも今後の教育方法の一つとして考えられるのではないだろうか。

付　記

本研究は平成 23 年度科学研究費補助金挑戦的萌芽研究（課題番号 23652113）「研究支援を目指した『理工学系基本口頭表現用例学習辞典』の開発」を基に行っている。

文　献

鈴木重幸(1972)『日本語文法･形態論』むぎ書房.
高橋太郎，金子尚一，金田章宏，齋美智子，鈴木泰，須田淳一，松本泰丈(2005)『日本語の文法』ひつじ書房.
日本語記述文法研究会編(2009)『現代日本語文法２』くろしお出版.

調査資料

『理工学系話し言葉コーパス』東京大学大学院工学系研究科
『名大会話コーパス』名古屋大学
『テーマ別 中級から学ぶ日本語』研究社
『科学技術基礎日本語 留学生･技術研修生のための使える日本語―読解編―』金沢工業大学
『新中級から上級の日本語』The Japan Times
『中級を学ぼう(前期・後期)』スリーエーネットワーク
『中･上級のための日本語読解』文教大学出版事業部
『大学･大学院 留学生の日本語①読解編Ⅰ』アルク

口頭発表（2）

9月1日（火）　15:20～17:20

中古語における意志系 Yes/No 疑問文の表現機能
—日本語歴史コーパス平安時代編を利用して—

林　淳子（東京大学大学院人文社会系研究科）[1]

Functions of Intention-expressing Yes-No Interrogative Sentences in Early Middle Japanese.

Hayashi Junko (Graduate School, the University of Tokyo)

要旨

本発表は、現代日本語の「シヨウカ」疑問文による質問（「そろそろ行こうか？」「荷物持ちましょうか？」など）の特殊性への関心から、中古語において話し手の意志あるいは相手の意志をめぐる Yes/No 疑問文がどのような表現として存在していたかを明らかにすることを目的とする。そこで、日本語歴史コーパス平安時代編を利用し、中古語において意志を表すのに用いられる助動詞「ム」「マシ」と疑問の係助詞「ヤ」「カ」との組み合わせからなる 8 文型の疑問文を対象に調査を行った。その上で、各文型の意志系 Yes/No 疑問文については、前後文脈を参考に各例の表現機能を判断した。その結果、8 文型の中でも特に意志系 Yes/No 疑問文の例が多く見られる「…ムヤ」「～ヤ…ム」「～ヤ…マシ」について、現代語のシヨウカ疑問文とは異なる範囲へ表現機能が広がることが分かった。

1.　はじめに

1.1　現代日本語シヨウカ疑問文による質問の特殊性

現代日本語の Yes/No 疑問文のうち、文末が「シヨウカ／シマショウカ」の形式をとるものを、本発表ではシヨウカ疑問文と呼ぶ。シヨウカ疑問文は話し手の意志や相手の意志をめぐる疑問を表す文であり[2]、具体的には次のような表現に用いられる。

■意志をめぐる躊躇感の表明　「行こうか？やめておこうか？」
■申し出　「その荷物、持とうか？」
■相談　（BBQ をしながら）「このお肉、もう裏返そうか？」
■誘い　（デートの帰り道）「次は映画を見に行こうか？」
■共同行為のもちかけ　（一緒に出かける相手に）「そろそろ行こうか？」
■提案　「待ち合わせは 8 時にしようか？」
■行為の誘導　「黙ってないで、そろそろ話そうか？」

意志をめぐる躊躇感の表明・申し出・相談は話し手の意志、誘い・共同行為のもちかけ・提案は話し手と相手の意志、行為の誘導は相手の意志をめぐる疑問文により実現される表

[1] jhayashi52[at]gmail.com　※[at]を@に置き換えてください。
[2]「明日こそは晴れようか？」のように推量系のシヨウカ疑問文も存在するが、意志系の「シヨウカ」と推量系の「ダロウカ」（「明日こそは晴れるだろうか？」）との棲み分けが進んだ結果、現在ではほとんど用いられなくなっている。

現である。

これらは、意志をめぐる躊躇感の表明を除けばすべて対人的な質問の表現でもあるが、シヨウカ疑問文による質問は、「～デスカ？」「～マスカ？」「～ノデスカ？」など他の文末形式をとる Yes/No 疑問文の質問と異なり、厳密な意味での解答を求めているとは言えない（林(2014b)）。上記の例から明らかなように、シヨウカ疑問文は、疑問の内実が事態実現にまつわる相手の意向が分からないというところにあり、質問によって求める答えが話し手の事態実現意向に対する相手の意向（応じるか否か）である点で特殊なのである[3]。

1.2　発表の目的

現代日本語シヨウカ疑問文による質問の特殊性は、シヨウカ疑問文が話し手や相手の意志をめぐる Yes/No 疑問文であることから自然に導かれるものであろうか。意志をめぐる Yes/No 疑問文は通時的に見ていつでも、話し手の意向に対して相手から応諾の意向を求めるという特殊な表現であり続けてきたのか。本発表は、このような問題関心から、意志をめぐる Yes/No 疑問文（以下、「意志系 Yes/No 疑問文」と呼ぶ）の中古語における表現機能を確認することを目的とする。

結論を先に述べれば、中古語における意志系 Yes/No 疑問文の表現機能は、現代語のそれとは相当に異なるものであり、中古語の状況から現代語における意志系 Yes/No 疑問文の表現機能の成立過程を探ることはできない。しかし、資料が韻文に偏る上代語を除けば、疑問文の表現機能を確認することが可能な最も古い時代である中古語の様相を確認しておくことは、意志系 Yes/No 疑問文の表現機能のありうる広がりを把握する上でも必要であろう。

1.3　方法

表 1　検索対象の文型と検索方法

文型		検索方法	
係り結び	「～カ…ム」「～ヤ…ム」 「～カ…マシ」「～ヤ…マシ」	キー設定	語彙素が「む」／「まし」
		前方共起条件	キーから 10 語以内 語彙素が「か」／「や」
承接	「…ムカ」「…ムヤ」 「…マシカ」「…マシヤ」	キー設定	語彙素が「む」／「まし」
		後方共起条件	キーから 1 語 語彙素が「か」／「や」

具体的な方法としては、日本語歴史コーパス平安時代編を利用して意志系 Yes/No 疑問文の用例を検索し[4]、小学館『新編日本古典文学全集』の本文を参考に各用例の表現機能を確認するという手順を踏んだ。意志を表すのに用いられる助動詞「ム」「マシ」[5]と疑問の係助詞「ヤ」「カ」が、係り結びあるいは承接によって連動してはたらく文を中古語の意志系

[3] この違いを反映して、シヨウカ疑問文による質問とその他の文型の疑問文による質問では、終助詞＜ね＞＜な＞の付加に伴う表現機能の変化の様相が異なる（林(2014a)）。

[4] 国立国語研究所（2014）『日本語歴史コーパス　平安時代編』https://maro.ninjal.ac.jp（2015 年 6 月 12 日確認）

[5] ただし、平叙文で話し手の意志を表す用法を持つ「ム」と異なり、「マシ」は疑問文の述語に用いられたときにのみ、意志を表す（川村(2014)）。

Yes/No 疑問文とみなし、表 1 に挙げる 8 つの文型を検索対象とした。

係り結び文型を検索する際に前方共起条件を「キーから 10 語以内」と設定したのは、これが設定しうる最も広い範囲であったためである。したがって、助詞「ヤ」「カ」と助動詞「ム」「マシ」の係り結びによって構成される疑問文であっても、両者が 11 語以上離れている例は検索結果に含まれないという点で、この検索方法には限界がある。しかし、本発表の目的は中古語における意志系 Yes/No 疑問文の表現機能の広がりを確認することであり、11 語以上離れて係り結びを構成する文があったとしても、結果に大きな影響を与えるものではないと判断した。

2.　中古語の意志系 Yes/No 疑問文

2.1　意志系 Yes/No 疑問文の文型

上記の方法で検索を行った結果、8 つの文型で合わせて 1,692 例を得た。この 1,692 例を Yes/No 疑問文と Wh 疑問文に分けた上で、Yes/No 疑問文についてはさらに、述語「～ム」「～マシ」が推量系（推量、妥当性、可能性など）の意を表すものと意志系の意を表すものに分けた。表 2 にそれぞれの例数を挙げる（「呼応なし」は 10 語以内に共起した係助詞「ヤ」「カ」と助動詞「ム」「マシ」が係り結びを構成していないことを指す）。

表 2　文型別の例数

	Yes/No 疑問文		Wh 疑問文	呼応なし	その他	合計
	推量系	意志系				
～カ…ム	6	3	658	27	1	695
～ヤ…ム	549	18	10	91	2	670
…ムカ	3	0	0		0	3
…ムヤ	121	54	0		5	180
～カ…マシ	0	0	25	6	0	31
～ヤ…マシ	22	53	0	8	0	83
…マシカ	0	0	0		0	0
…マシヤ	25	5	0		0	30

一定数の意志系 Yes/No 疑問文が見られるのは「～ヤ…ム」「…ムヤ」「～ヤ…マシ」の 3 文型においてである。そこで、以下ではこの 3 つの文型の意志系 Yes/No 疑問文がどのような表現機能を持つかを見ていく。

2.2　意志系 Yes/No 疑問文の表現機能

2.2.1　本文種別

意志系 Yes/No 疑問文「～ヤ…ム」「…ムヤ」「～ヤ…マシ」が現れる本文の種別[6]は表 3 の通りである。

---

6 日本語歴史コーパス平安時代編の検索結果に本文種別が記載されていない例については、発表者が調査・判断した。また、検索結果においては本文種別が「会話」となっている例の中でも、会話のなかで「～しようかと思って、…した」のように語られる思考内容である場合には、「心内語」と判断した。

表 3　本文種別

	会話	歌	心内語	その他	合計
～ヤ…ム	1	16	0	1	18
…ムヤ	51	1	2	0	54
～ヤ…マシ	7	13	33	0	53

「～ヤ…ム」は歌、「…ムヤ」は会話、「～ヤ…マシ」は心内語と、よく現れる文種を棲み分けている様子が伺える。そこで、まずは現代語シヨウカ疑問文と同様に会話で多用される「…ムヤ」の表現機能から見ていきたい。

2.2.2　「…ムヤ」

「…ムヤ」文型の意志系 Yes/No 疑問文には、「…ム」の形で表される行為の主体すなわち主語が 1 人称（話し手）であるものと 2 人称（相手）であるものとがある。古典文法において「「…ム」は意志を表す」と言うときの「意志」は通常話し手の意志を指す（小田(2007)）ため、2 人称者が主語である「…ムヤ」疑問文の「…ム」を厳密な意味で「意志」とは言うことはできないかもしれない。しかしながら、

- 「…ム」が話し手の意志を表すという前提は、平叙文を基準にしたものである。
- 現代日本語では、平叙文と異なり、事態を述べきるわけではない疑問文においては、相手の心の内（意志もこれに含まれる）を話し手が言語化してしまう場面がある（林(2015)）。

の 2 点を考慮し、疑問文を考察する本発表では 2 人称主語であっても意志系 Yes/No 疑問文であると考えたい[7]。その上で「…ムヤ」の表現機能別例数を一覧にすれば次のようである。

表 4　「…ムヤ」の表現機能別例数

主語	表現機能		例数
1 人称	対人的	宣言（意志表明）	4
		反語による意志不在表明	6
	非対人的	意志をめぐる躊躇感表明	2
2 人称	実現意向伺い		13
	依頼		23
	勧め		4
	誘い		2

A　1 人称主語

1 人称主語の意志系 Yes/No 疑問文のもっとも基本的な表現機能は、自らの意志をめぐる躊躇感表明である。しかし、「…ムヤ」疑問文の場合は、意志をめぐる躊躇感表明は非対人的な場面でしか見られず、対人的な表明の場面では躊躇感がほとんど感じられない単なる意志表明か、あるいは自らその意志の存在を否定する反語しかない。

[7] 野村(2014)は、「ム」の用法として「⑥意志」とは別に「⑧聞き手の意志」を挙げ、「…ムヤ」疑問文をその例としている。

■宣言（対人的意志表明）…4 例

(1)（末摘花から送られた元日の装束について源氏が）「とり隠さむや。かかるわざは人のするものにやあらむ」　（源氏物語 1 末摘花　p.301[8]）

■反語による意志不在表明…6 例

(2)（浮気しないよう忠告されて）少将、「あなゆゆし、よし、聞きたまへ。文をだにものしはべりてむや。『御用意あり』とうけたまはりしよりなむ、限りなく頼みきこえし」とのたまひて、　（落窪物語　p.180）

■非対人的・意志をめぐる躊躇感表明…2 例

(3) 容貌（かたち）はしもいと心につきて、つらき人の慰めにも、見るわざしてんやと思ふ。　（源氏物語 3 少女　p.64）

B　2 人称主語

一方、2 人称主語の意志系 Yes/No 疑問文の表現機能は、基本的には事態実現に関する相手の意向を伺うことであり、「…ムヤ」疑問文にもこれに当たる例が多い。

■実現意向伺い…13 例

(4) むかし、女をぬすみてゆく道に、水のある所にて、「飲まむや」と問ふに、うなづきければ、　（伊勢物語　p.217）

相手の意向伺いである「…ムヤ」疑問文の中でも、特に話し手がその事態の実現を希望している場面では、依頼・勧め・誘いの表現となる。すなわち、話し手の受益を前提としていれば「依頼」、話し手の受益を前提としない場合のうち、聞き手のみが行う行為についての実現意向を問うのが「勧め」、話し手自身も行おうとしている行為について相手の実現意向を問うのが「誘い」である[9]。

■依頼…23 例

(5)（弁の少将が中納言邸の女房に対して）『我いと思ふさまにおはすなるを、必ず、御文つたへてむや』とのたまひしかば、　（落窪物語　p.91）

■勧め…4 例

(6) 主の侍従は、故大臣に似たてまつりたまへるにや、かやうの方は後れて、盃のみすすむれば、「寿詞をだにせんや」と辱められて、竹河を同じ声に出だして、まだ若けれどをかしううたふ。　（源氏物語 5 竹河　p.72）

■誘い…2 例

(7)（僧都が妹の尼君に、源氏への挨拶に誘う場面）「この世にののしりたまふ光る源氏、かかるついでに見たてまつりたまはんや[10]。世を棄てたる法師の心地にも、いみじう世の愁へ忘れ、齢のぶる人の御ありさまなり。いで御消息聞こえん」　（源氏物語 1 若紫　p.209）

---

[8] 巻数・頁数は小学館『新編日本古典文学全集』による。ただし、古今和歌集には頁数ではなく、歌番号を記す。

[9] 勧めと誘いのこのような区別は、小田(2015)（p.222）に従うものである。

[10] 述語が尊敬語「〜たまふ」であることから、主語は 2 人称である。この点で、同じ「誘い」といっても、1 人称複数主語である現代語の誘い（「次は映画を見に行こうか？」）とは異なる。

このように、依頼の例が多いことから、「…ムヤ」を「…ム」と一括して「「～む」型の行為指示表現」と見る先行研究（藤原(2014)など）もある。藤原(2014)では「…ムヤ」の「ヤ」は命令形の文末に接続する「ヤ」と同様に「行為のうながしとして用いられる」と説明する。しかし、(4)のように相手の意志の有無をたずねる例がある以上、やはり「…ムヤ」の「ヤ」は疑問の助詞と見るべきである。小柳(2014)の述べる通り、依頼表現が確立していない時代には「「～むや」という相手の意向を尋ねる疑問表現を使って」「間接的に要求」していたと見る方が適切であろう。

また、そもそも川上(2005)のように、この種の「…ムヤ」を「推量＋疑問」と見る研究もあるが、依頼だけならともかく、勧めや誘いの例も存在することを考慮すればやはり、意志をめぐる疑問と見るべきであろう。

2.2.3　「～ヤ…ム」「～ヤ…マシ」

「…ムヤ」が 1 人称者（話し手）の意志をめぐる疑問を表す場合もあれば 2 人称者（相手）の事態実現に対する意向をたずねる場合もあったのに対し、「～ヤ…ム」「～ヤ…マシ」が扱うのは、1 人称者の意志に限られる。また、「…ムヤ」疑問文は対人的表現がほとんどであったのに対し、「～ヤ…ム」「～ヤ…マシ」はともに、非対人的すなわち独り言的に話し手の意志あるいは意志をめぐる躊躇感を表明する表現が多い。「～ヤ…ム」「～ヤ…マシ」の表現機能別例数は表 5 の通りである。

表 5　「～ヤ…ム」「～ヤ…マシ」の表現機能別例数

主語	表現機能			～ヤ…ム	～ヤ…マシ
1 人称	対人的	宣言（意志表明）		4	1
		申し出	意志をめぐる	0	1
		提案	躊躇感表明	1	1
	非対人的	意志表明		6	1
		躊躇感表明		6	49
	その他			1	0
合計				18	53

A　1 人称主語・非対人的

■意志表明…「～ヤ…ム」6 例・「～ヤ…マシ」1 例

(8) 三千歳になるてふ桃の花ざかり折りてやかざさむ君がたぐひに（落窪物語　p.271）

(9) ともかくも御覧ずる世にや思ひ定めましと思しよるには、（源氏物語 5 宿木　p.377）

■意志をめぐる躊躇感の表明…「～ヤ…ム」6 例・「～ヤ…マシ」49 例

(10)（ちゃんとした衣装を持たない母北の方が）すくよかなる衣のなきぞいといとほしき。「隠しの方にやあらむ」とのたまふ。（落窪物語　p.324）

(11)（源氏が末摘花の琴の音を聞きながら）ものや言ひ寄らましと思せど、うちつけにや思さむと心恥づかしくて、やすらひたまふ。（源氏物語 1 末摘花　p.269）

中心的な表現機能である「非対人的な、意志をめぐる躊躇感表明」において、「～ヤ…ム」

と「～ヤ…マシ」には次の 2 点の違いが認められる[11]。

①扱う事態の重大さ・躊躇の度合い

「～ヤ…ム」：身近な単発の動作を行うか否かを問題にする。軽い迷い。

(12)（ちゃんとした衣装を持たない母北の方が）すくよかなる衣のなきぞいといとほしき。「隠しの方にやあらむ」とのたまふ。　（(10)再掲）（落窪物語 p.324）

(13) 散るをまたこきや散らさむ袖ひろげひろひやとめむ山の紅葉を
（平中物語　p.512）

「～ヤ…マシ」：今後の方針として何を選ぶかを問題にする。深い逡巡。

(14)（源氏が玉鬘への恋情を抑えられなくなり）わが御心にも、すくよかに親がりはつまじき御心や添ふらむ、父大臣にも知らせやしてましなど、思しよるをりをりもあり、　（源氏物語 3 胡蝶　p.174）

(15)（明石の君が姫君を引き取るべきか思案する）いかにせまし、迎へやせまし、と思し乱る。　（源氏物語 2 松風　p.424）

したがって、次の 2 例のように同じ「言ふ」という行為でも、「～ヤ…ム」と「～ヤ…マシ」では事態の重大さが異なる。

(16) 世の中にいづらわが身のありてなしあはれとや言はむあな憂とや言はむ
（古今和歌集 943）

(17) この男、苦しうなりて、かういへるとて、げに、たち返り来ぬべきことをやいはましと思へど、　（平中物語　p.528）

②「ツ」「ヌ」の参加による意味合いの違い

助動詞「ツ」「ヌ」が「～ヤ…ム」の「ム」に上接する例はあまり見られないのに対し、「～ヤ…マシ」の「マシ」には「ツ」「ヌ」がしばしば上接する。

表 6　「ツ」「ヌ」が上接する用例の数

	ム／マシ	テム／テマシ	ナム／ナマシ	その他	合計
～ヤ…ム	17	1	2	0	20
～ヤ…マシ	29	10	13	1	53

この内、「～ヤ…テマシ」「～ヤ…ナマシ」は用いられる場面状況に一定の傾向が見られる[12]。「～ヤ…テマシ」は、好機のついでに、一見大胆に見える方向へ舵を切ろうとする前向きな方針転換に伴う躊躇感表明の場面で用いられる。

(18)（玉鬘の裳着の機会に、内大臣に玉鬘引き取りの経緯を説明しようと思案する）まして、内大臣にも、やがてこのついでにや知らせたてまつりてましと思しよれば、いとめでたうところせきまでなむ。　（源氏物語 3 行幸　p.295）

「～ヤ…ナマシ」は、状況の悪さに投げやりな気持ちになり、これまで続けてきたことを終

11 「ム」と「マシ」の違いについて、山口(1968)は「非事実性をそなえた意味領域の中で、「まし」の領域はより非現実的であり、「む」の領域はより現実的である」と述べている。また、高山(2002)は連体ナリとの承接関係の有無を根拠に「マシは〈非現実〉面だけに関与し、ムは〈現実〉〈非現実〉の両面に関与する」と論じている。

12 意志系 Yes/No 疑問文において「ツ」「ヌ」の上接がもたらすニュアンスの違いについては、岡崎(1996)の「…ムヤ」「…テムヤ」「…ナムヤ」に見られる違いの分析がある。

えてしまおうとする後ろ向きの方針転換に伴う躊躇感表明の場面で用いられる。

(19)（六条御息所が娘とともに伊勢に下ろうかと思案する）大将の御心ばへもいと頼もしげなきを、幼き御ありさまのうしろめたさにことつけて下りやしなまし、とかねてより思しけり。（源氏物語 2 葵　p.18）

B　1 人称主語・対人的

「～ヤ…ム」「～ヤ…マシ」には、少数ながら、話し手の意志あるいは意志をめぐる躊躇感を対人的に表明するものもある。

■宣言（対人的・意志表明）…「～ヤ…ム」4 例・「～ヤ…マシ」1 例

(20) 今はとて返す言の葉拾ひおきておのがものから形見とや見む　（古今和歌集 737）

(21) 折すぎてさてもこそやめさみだれて今宵あやめの根をやかけまし（和泉式部日記 p.26）

■申し出（対人的・意志をめぐる躊躇感表明）…「～ヤ…ム」0 例・「～ヤ…マシ」1 例

(22) かくのみしゆくへまどはばわが魂をたぐへやせまし道のしるべに（平中物語　p.495）

■提案（対人的・意志をめぐる躊躇感表明）…「～ヤ…ム」1 例・「～ヤ…マシ」1 例

(23) ふみわけてさらにやとはむもみぢ葉のふりかくしてし道と見ながら（古今和歌集 288）

(24) 片岡にわらびもえずはたづねつつ心やりにや若菜つまましん　（大和物語 p.310）

対人的といっても、これらはすべて、問答歌や文のやりとりの中で詠まれた歌であり、前後の歌との関係から臨時的に、話し手の意志表明が宣言に、意志をめぐる躊躇感表明が申し出や提案に解されるに過ぎない。すべて歌の例であることを考えれば、文自体の表現機能を申し出や提案と言うことはできないであろう。しかし一方で、現代語シヨウカ疑問文のように 1 人称主語の意志系 Yes/No 疑問文が申し出や相談のような相手の意向をたずねる質問になる可能性自体は、中古語の意志系 Yes/No 疑問文にも潜在していたと言えよう。

3.　現代語シヨウカ疑問文との比較

現代語の「シヨウ」が古代語の「セム」の現代的な姿であるとはいっても、「セム」から「シヨウ」に至る間にこの形式の性質は当然変質している（尾上(2012)）。係助詞「ヤ」と現代語の終助詞「カ」も同様であろう（阪倉(1993)）。しかし、それぞれの時代に「…ムヤ」「～ヤ…ム」「～ヤ…マシ」および「シヨウカ」が意志をめぐる Yes/No 疑問文の文型であったことを重視し、あえて両者を比較検討すれば、表 7 のようになる（△は限定的に存在することを示す）。

表 7　各文型の表現機能

主語		1 人称				1 人称複数	2 人称
		対人的		非対人的			
		宣言	躊躇感表明	意志表明	躊躇感表明		
中古語	…ムヤ	○	×	×	○	×	○
	～ヤ…ム	△	△	○	○	×	×
	～ヤ…マシ	△	△	△	○	×	×
現代語	シヨウカ	×	○	×	○	○	△

表 7 から明らかなように、意志をめぐる躊躇感を非対人的に表明する機能は時代や文型の別を問わず見られるが、その他の点では相違点が多く、そこから現代語シヨウカ疑問文の特殊性を考えるにあたって問うべき問題が見えてくる。

①中古語の意志系 Yes/No 疑問文は、表現機能の傾向に基づいて「…ムヤ」タイプと「～ヤ…ム」「～ヤ…マシ」タイプに分けることができる。「ヤ」の位置の違いによってこの差が生まれるとすれば、文末で疑問の意を添える「ヤ」[13]と文中で係り結びを構成する「ヤ」とでは疑問のあり方が異なると見ることができる。

⇒現代語シヨウカ疑問文は冒頭に挙げた通り幅広い表現機能を有するが、「シヨウカ」の「カ」はすべて同じようにはたらいていると言えるのか。

②意志系 Yes/No 疑問文の文型はすべて係助詞「カ」ではなく「ヤ」によって構成されるものであることから、「ヤ」による疑問のあり方と意志をめぐる疑問文に何らかの関係があったと見ることができる[14]。

⇒現代語シヨウカ疑問文の文末の助詞「カ」は何をどのように疑問することにはたらいているのか。

③中古語「…ムヤ」には 2 人称主語の例が多いのに対し、現代語シヨウカ疑問文では 2 人称主語の例は相手の行為を誘導する場合（「黙ってないで、そろそろ話そうか？」）に限られる。現代語では、2 人称主語の意志系 Yes/No 疑問文は「スルカ／シマスカ」や否定疑問文が担う。

(25)「これ、食べますか？」＜実現意向伺い＞

(26)「お塩取ってくれますか？」＜依頼＞

(27)「良かったら、いらっしゃいませんか？」＜誘い＞

⇒「シヨウカ」「スルカ」の機能分担はいつから発生したのか。現代語でも限定的に「黙ってないで、そろそろ話そうか？」のような 2 人称主語の例があるのはなぜか。

④現代語シヨウカ疑問文には 1 人称複数主語のものが多く見られるが、中古語の意志系 Yes/No 疑問文には、1 人称複数を主語とするものは存在しない。

⇒1 人称複数主語の意志系 Yes/No 疑問文はいつ頃から見られるのか。

4. まとめ

本発表では、中古語の意志系 Yes/No 疑問文として「…ムヤ」「～ヤ…ム」「～ヤ…マシ」の 3 つの文型の疑問文に注目し、日本語歴史コーパス平安時代編を利用して、各文型の表現機能の広がりを調査した結果、以下の考察を得た。

・中古語の意志系 Yes/No 疑問文は、現代語シヨウカ疑問文と同じく話し手の意志をめぐる躊躇感表明の機能を有する。

・しかし一方で、現代語シヨウカ疑問文にはほとんど見られない 2 人称主語の例が「…ム

---

[13] 阪倉(1993)によれば、文末に「ヤ」を添える「―ヤ。」タイプの疑問文は、「文の叙述が終止形でいちおう完了したところに「や」を添えて、これをそのまま相手に持ちかけるかたちをとる」疑問文であり、それゆえに鎌倉時代以降、「問いかけ」の語気が薄れ、反語など情意的な方向へ傾くという。

[14] これに関連して、野村(2001)の「ヤによる問い掛けは価値的」であり、「真偽性とは直接関わらない」という指摘は、上代語に関するものであるとはいえ、本稿で論じた意志系 Yes/No 疑問文と「ヤ」の関係を考える上で示唆に富む。

ヤ」疑問文には多く見られ、現代語シヨウカ疑問文の大部分を占める 1 人称複数主語の例が見られないなど、両者の違いも認められる。

この結果を通して、意志系 Yes/No 疑問文が持ちうる表現機能の広がりを確認するとともに、中古語と現代語ではその広がりが重なりつつ異なることが明らかになった。

この考察を踏まえ、今後は、意志系 Yes/No 疑問文が現代語特有の表現機能を持つに至る過程を調査・分析していきたい。

参考文献

岡崎正継(1996)『国語助詞論攷』おうふう.

小田勝(2007)『古代日本語文法』おうふう.

小田勝(2015)『実例詳解古典文法総覧』和泉書院.

尾上圭介(2012)「不変化助動詞とは何か―叙法論と主観表現要素論の分岐点―」『国語と国文学』89 巻 3 号，pp.3-18.

川上徳明(2005)『命令・勧誘表現の体系的研究』おうふう.

川村大(2014)「マシ」日本語文法学会編『日本語文法事典』，pp.587-588.

小柳智一(2014)「奈良時代の配慮表現」野田尚史、高山善行、小林隆『日本語の配慮表現の多様性―歴史的変化と地理的・社会的変異』，pp.57-74.

阪倉篤義(1993)『日本語表現の流れ』岩波書店.

高山善行(2002)『日本語モダリティの史的研究』ひつじ書房.

野田尚史、高山善行、小林隆(2014)『日本語の配慮表現の多様性―歴史的変化と地理的・社会的変異』くろしお出版.

野村剛史(2001)「ヤによる係り結びの展開」『国語国文』，70 巻 1 号，pp.1-34.

野村剛史(2014)「ム」日本語文法学会編『日本語文法事典』，pp.601-602.

林淳子(2014a)「疑問文における終助詞＜ね＞と＜な＞」『日本語学論集』，10 号，pp.152-167.（http://hdl.handle.net/2261/55750 よりダウンロード可能）

林淳子(2014b)「「返事をさせる表現」の全体像―解答要求表現の位置づけを求めて―」『日本語文法学会第 15 回大会予稿集』，pp141-148.

林淳子(2015)「Yes/No ノ無し疑問文と代弁的質問」『日本語学会 2015 年度春季大会予稿集』，pp.41-48.

藤原浩史(2014)「平安・鎌倉時代の依頼・禁止表現に見られる配慮表現」野田尚史、高山善行、小林隆『日本語の配慮表現の多様性―歴史的変化と地理的・社会的変異』，pp.75-92.

山口堯二(1968)「「まし」の意味領域」『国語国文』，37 巻 5 号，pp.21-35.

山口堯二(1990)『日本語疑問表現通史』明治書院.

コーパスによる日本書記古訓形容詞「カシコシ、サカシ」に関する調査

劉 琳(北海道大学大学院文学研究科)

Corpus-based Study of Adjectives "kashikoshi" and "sakashi"in Old Manuscripts of Nihon Shoki

Liu Lin (Graduate School of Letters Hokkaido University)

要旨

形容詞「カシコシ、サカシ」は『日本書紀』において漢字・漢語の解釈である和訓として多く使われた。一方、和文の文学作品においてもこの二語の使用が多く見られる。本稿では、『日本書紀』における漢字「賢」に関わる古訓形容詞「カシコシ、サカシ」の二語を取り上げ、まず日本書紀古訓としての意味用法を中心に検討する。次は「カシコシ、サカシ」が上代から現代への意味変化の実態を明らかにするための考察の一階梯として、上代、中古の文学作品に使用された「カシコシ、サカシ」の用例を抽出し、日本書紀古訓と平安仮名文学における意味的特徴を明らかにした上で、上代以降の歴史的な変遷の実態を記述する。用例の収集にあたっては、「日本語歴史コーパス」(国立国語研究所)、『新編日本古典文学全集』(Japan Knowledge Lib)などを利用した。

1. はじめに

『日本書紀』において形容詞「カシコシ」は一般に「畏、懼」に、「サカシ」が「賢、哲」などの漢字に附された和訓として用いられている。漢字の字義を考えると、『日本書紀』における「カシコシ」は主に「畏怖、畏敬」の意味、「サカシ」は「賢明」という意味を表すと推測される。『古事記』、『万葉集』における「カシコシ、サカシ」の和訓を充てられた漢字を見ると、万葉仮名以外に、『日本書紀』とは変わらない漢字を用いた。一方、『日本書紀』の各古写本において、「カシコシ」は次のような漢語の和訓として使われた用例も見られる。

①賢愚－カシコクオロカナルコト(岩崎本)、②智謀－カシコキ(北野本)

③英才賢徳－カシコクサカシクマシマス(圖書寮本)」

更に、1540 年に書写した兼右本日本書紀における「賢哲（才智のある）」の和訓には「左訓：カシコキヒト　右訓：サカシヒト」の二種が見られる。

上記の用例をみると、日本書紀古訓の「カシコシ」は「畏怖」の意味以外に、「才智のある」という意味も表し、「サカシ」とは意味的に共通な面があると思われる。

ここから、古代において「カシコシ」は主に「畏怖、畏敬」、「サカシ」は「賢明、才能がある」の意味として使われ、二語は意味的に共通な面があることが分かる。

次は、現代語の「かしこい、さかしい」の意味用法について国語辞書を用いて調べると、「カシコイ」は主に「頭がいい、利口だ」の意味として使われている。「さかしい」は現代語において方言として生き残る言葉[1]であり、「かしこい、利口だ」の意味を持つが、現代においてほとんど用いられず、「こざかしい」のようなマイナス的な意味は普通に用いられる[2]。また、「さかしい」の使用状況について、「web データに基づく形容詞用例デー

[1]『新明解国語辞典(第 7 版)』

[2]『現代形容詞用法辞典』

タベース」を用いて調査し、一例も見つからないが、「こざかしい」は 1366,461 件の用例がヒットした。

このように、「カシコシ、サカシ」の意味用法が変遷したことが分かった。この二語は現代語に至るまでどのような変遷を経てきたのか、どのような理由によって意味変化が生じたのか、取り組むべき課題が多くある。本稿はこれらの問題を解決するための考察一階梯として、まず上代、中古における「カシコシ、サカシ」の意味用法を確認し、意味的特徴を明らかにする。そして、この二語が上代以降の歴史的な意味変遷の実態を記述する。

2. 国語辞書における記述及び先行研究

「カシコシ、サカシ」の「語誌」について、松浦(1983)は次のように説かれている。[3]

「カシコシ」は記紀、万葉の時代から多く用いられたが、畏怖、畏敬の念を表す心情表現の語であった。その意味は現代語の「頭がよい、利口だ」といった、知恵、才覚についてのものではなかった。

「サカシ」は上代において、知恵や才覚の優れた意味を持つ語として使われ、高い評価を伴う語であった。平安時代から意味が変遷し、現代語の「コザカシイ」に通じる低い評価を与えられている語になった。

上記二語について、上代から中世までの意味用法を『時代別国語大辞典』を利用して確認し、次のように語義を記述されている。

『時代別国語大辞典(上代編)』

■　カシコシ【恐・畏】(形ク)

①恐ろしい。②恐れ多い。③驚くべきである。ただごとではない。

■　サカシ【賢】(形シク)

賢明である。

『時代別国語大辞典(室町時代編)』

■　カシコシ【畏し・賢し】(形ク)

㈠すぐれた絶対的な力に対して、おそれ、敬う気持ちである。(畏敬の対象：①神仏などの霊力、②天皇などの権威、③卓越のもの)

㈡人のすぐれた知的能力が、感心させられるほど適切に機能するさまである。(①知恵、適切な判断力、②優れた能力、③適切な対処、④思いもよらずめはしが利く)

■　サカシ【賢し】(形シク)

①才気をたのみ、ぬけめなく、すばやい判断を下すさまである。

②丈夫で、無病息災である。

この記述内容をみると、上代以降この二語の意味用法が拡大し、「カシコシ」は「才知、能力がある」という意味を持ち、「サカシ」と共通な意味を持つようになった。「サカシ」の「丈夫で、無病だ」という意味は上代では見えない。そして、松浦説の低い評価の用法が中世までは見えない。

「カシコシ」について、『源氏物語』における用例を分析し、論考したのは東辻(1967)である。山崎(1977)は「サカシ・サガシ」といった二つの形容詞についての論考である。そ

[3]佐藤喜代治編『講座日本語の語彙 9 語誌Ⅰ』ｐ199－203

して、土居(2001)は『土佐日記』にある「さかしきもなかるべし」」をめぐって、平安時代和文における「サカシ」の意味用法を論述した。本稿では、以上のことをふまえて、上代・中古の文学作品に使用された「カシコシ、サカシ」の意味用法を分析し、意味的特徴を明らかにする。

3.『日本書紀』における「カシコシ、サカシ」

『日本書紀』古写本[4]を利用し、「カシコシ、サカシ」の訓を持つ漢字・漢語を収集し、次のように示す。

◆　カシコシ

畏、懼、威、稜威、賢、智謀、英才、貴、重

上記「カシコシ」の訓を持つ漢字・漢語を見ると、「畏、懼」の二字は意味的には近いと推測される。このように上記の漢字を大きく①「畏、懼」、②「威、稜威」、③「賢、智謀、英才」、④「貴、重」の四組に分類できる。これから原文において、文脈に基づき各用例の意味用法を確認する。ここでは、用例の一部を示す。

①「畏、懼」

(1) 原文：仍奏表之曰「天上有神、地有天皇。除是二神、何亦有**畏**(カシコキコト)乎。」
(岩崎本訓)

訳文：そして、上表文を奉って、「天上に神がおいでになり、地には天皇がおいでになります。この二神のほかに、どこに**畏敬する**ものがありましょうか。…」[5]

(2)原文： 於是天皇詔之曰「是陵自本空、故、欲除其陵守而甫差役丁。今視是怪者、甚**懼**(カシコシ)之。無動陵守者。」則且、授土師連等。　(前田本訓)

訳文：そこで天皇は詔して、「この陵はもともと空である。そのため陵守を廃止しようと思って、初めて役丁に徴発したのだ。今この不吉な前兆を見ると、はなはだおそれ**恐れ多い**。陵守を廃止してはならない」と仰せられ、すぐにまた陵守を土師連らの管掌下に置かれた。

②「威、稜威」

(3)原文：則謂夫曰「汝祖等、渡蒼海跨萬里平水表政、以**威**武(カシコクタケキ)傳於後葉…」
(圖書寮本訓)

訳文：そこで夫に語って、「あなたの先祖たちは、蒼海原を渡り万里を超えて、**畏敬すべき**武力をもって後世に名を伝えてきました…」

③「賢、智謀、英才」

(4)原文：相共**賢**(カシコク)愚、如鐶无端。　(岩崎本訓)

訳文：お互いが**賢**であり愚でもあって、鐶に端がないようなもので区別はつかない。

(5)原文：億計王曰「弟**英才**(カシコク)賢徳(サカシクマシマス)、爰無以過。」
(圖書寮本訓)

---

[4] 古写本の岩崎本、圖書寮本、前田本を利用した。兼右本と寛文九年版本について筆者が以前収集した 22 と 24 巻のデータも利用した。神代巻に関して、『六種対照日本書紀神代巻和訓研究索引』を利用した。また、『訓点語彙集成』も参照した。

[5]用例の現代語訳は『新編日本古典文学全集(小学館)による

訳文：億計王は、「弟は**才能があって**賢く徳もある。これに勝る人はいない」と仰せられた。

(6)原文：既而天皇謂高市皇子曰「其近江朝左右大臣及**智謀**(カシコキ)群臣共定議…」（北野本訓）

訳文：やがて天皇は高市皇子に語って、「いったい近江朝では、左右大臣と**智略にたけた**群臣が協議して事を決定している。…」

④「貴、重」

(7)原文：顙搶地叩頭曰「臣之罪實當萬死。然當其日、不知**貴者**(カシコキヒト)。」（圖書寮本訓）

訳文：額を地面につけて叩頭して、「私の罪は実に死に当たります。しかしながら、あの日は、**貴い人**だとは存じあげませんでした」と申し上げた。

(8)原文：「愛之叔父、勞思、非一介之使遣**重臣**(カシコキマチキムタチ)等而教覺、是大恩也。（北野本訓）

訳文：「親愛なる親父は私を労わって、使者一人だけではなく**重臣**たちを遣わして教え諭された。これは大いなる恩愛である。

例文(1)は神、天皇のような権威のある者に対する恐れ敬うことを表す意味であり、『古事記』にも同じ用法が見られる。例(2)は霊力のあるものに対する恐れる気持ちである。例(3)は威力のあり、すぐれる人に対する畏敬の気持ちを表す。例(4)(5)(6)は訓と対応する漢字が異なるが、意味的には共通する部分がある。いずれも才能、思慮を意味している。例(7)(8)は身分が高い意味を表す。従って、『日本書紀』における「カシコシ」には①霊力、権威に対する恐れる気持ち、②才能のある、身分の高い者をおそれる。敬うべきだ」などの意味をしている。そのうち、①の意味を表す用例が最も多い。

『訓点語彙集成』において「カシコシ」の訓を持つ漢字を確認すると、「尊、貴、賢」以外、ほかは全部「畏怖」の意味を持つ漢字である。「英才、貴者、貴国」に附される訓として、「才能のある。身分・国が優れる。あがめ敬うべきだ」という意味を持つ用例は『日本書紀』にしか見えないのである。

◆　サカシ

「サカシ」の訓を持つ漢字・漢語は「賢、賢哲、賢徳、賢聖、哲、明哲、師、叡智」などが挙げられる。

(9)原文：所寶惟**賢**「(サカ)シク、サカシキヒト」、爲善最樂。（前田本訓）

訳文：宝とすべきは**賢人**であり、善を行うことを最大の喜びとする。

(10)原文：及乎繼體之君、欲立中興之功者、曷嘗不賴**賢哲**「(サカシ)ク」之謨謀乎。（前田本訓）

訳文：皇位継承の君主として、中興の功を立てようとすれば、昔からどうしても**賢哲**の策謀に頼らなければならない。

(11)原文：天皇、以心爲**師**(サカシ)、誤殺人衆、天下誹謗言「太惡天皇也。」（前田本訓）

訳文：天皇はご**自分の判断をただしい**とされたため、誤って人を殺すことが多かった。天下の人々は誹謗して、「大悪の天皇である」と言った。

(12)原文：天皇、幼而聰明**叡智**(サカシクマシマス)、貌容美麗、及壯仁寬慈惠。（前田本訓）

訳文：天皇は幼少の頃から聡明で**叡智**があり、容貌も美麗でいらっしゃった。成年に及んでは、大そう思いやりがあり情け深くていらっしゃった。

「サカシ」の意味にはプラス評価とマイナス評価の両方ある。『日本書紀』においては、「サカシ」は上記例文のように「賢」、あるいは「賢」字で構成する漢字熟語、「賢」と近似的意味を持つ「哲、叡智」などの訓として使われている。これらの漢字・漢語は、いずれもプラスの評価を持つものである。当然それと対応する訓としての「サカシ」は、マイナスの意味用法が見られない。

4. 平安時代文学作品における「カシコシ、サカシ」の意味

「カシコシ、サカシ」の中古における意味用法について、国立国語研究所が開発した「日本語歴史コーパス」の平安時代編を利用して用例を収集した。平安時代編には『古今和歌集』、『土佐日記』、『竹取物語』、『源氏物語』、『枕草子』のように和歌、日記、物語、随筆の各ジャンル全14の作品が収録された。検索された用例数からみると、「カシコシ」は『源氏物語』が最も多く、136 例があり、その次は『枕草子』の 34 例である。「カシコシ」に対し、「サカシ」の用例は少ない。同じく用例数が最も多いのは『源氏物語』で、30 例あり、『枕草子』は５例ある。本稿では、『源氏物語』及び「枕草子」の用例を中心に検討する。

◆　カシコシ

『萬葉集』、『古事記』、『日本書紀』には「カシコシ」は主に「畏怖、畏敬」の意味を表す。『日本書紀』において、「カシコシ」は「才能あり、能力がすぐれている」の意味を持つ「英才」などの漢語に充てられた和訓として使われる用例も見られる。これに対し、平安仮名文学の『源氏物語』、『枕草子』の用例を見ると、上記の意味以外に、独特の意味用法が見られる。

平安仮名文学では、次に示す用例のように、「カシコシ」が表す「畏敬」の意味が軽くなった。また、大切にする、慎重の意味を持つようになった。

(13)などてか、それをもおろかにはもてなしはべらん。**かしこけれど**、御ありさまどもにてもおしはからせ給へ。　『源氏物語・夕霧』
(14)とみのもの縫ふに、**かしこう**縫ひつと思ふに、針を引抜きつれば、はやく後をむすばざりけり。　『枕草子 91 段』

◆　サカシ

前節で述べたように、「サカシ」の意味にはプラス評価とマイナス評価の両方ある。上代の文学作品の用例や『日本書紀』古訓としての意味はプラス評価である。平安時代の「サカシ」は、判断がしっかりしていて物に動じないことをいった。自分自身の内に蔵する力、判断力によって事を決めて、その結果に自信をもっていることを表す。[6]

『枕草子』の「さかしきもの」の段は短い内容であるが、「サカシ」は四回使われ、そのうちの三例が「身分の卑しい者の小ざかしいこと」についてのマイナス評価である。同じ

[6]土居(2001：36)

意味用法は「源氏物語」にも見られる。

5. おわりに

本稿では、日本書紀の訓点本及び平安時代文学作品における「カシコシ、サカシ」の意味用法について、収集した用例を用いて考察を行った。平安時代以降「カシコシ、サカシ」の意味用法は拡大し、上代や現代よりはるかに意味用法が広い。「カシコシ、サカシ」の関係、中世以降の意味用法の実態、どのように現代語の意味用法に移行していったのかについての考察を今後の課題とする。

文　献

著作

石塚晴通(2006)『宮内庁書陵部影印集成・日本書紀』八木書店
内田貞徳(2005)『上代日本語表現と訓詁』塙書房
小島憲之ほか(1994-1998)『新編日本文学全集 1-3 日本書紀』小学館
杉浦克己(1995)『六種対照日本書紀神代巻和訓研究索引』武蔵野書院
築島裕(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
築島裕・石塚晴通(1978)『東洋文庫蔵岩崎本日本書紀本文と索引』日本古典文学会
佐藤喜代治編(1983)『講座日本語の語彙 9 語誌Ⅰ』明治書院

論文

土居裕美子 2001「平安時代和文における『さかし』の意味用法について」『高知大国文』(32) 高知大学
東辻保和 1967「源氏物語＜畏敬＞語彙の研究－＜かたじけなし＞＜かしこし＞考」『国語学』71
山崎馨 1977「形容詞さかし・さがし考」『松村明教授還暦紀念国語学と国語史』明治書院

辞書

大槻文彦(1907)『言海』吉川弘文館
石川孝ほか編(2011)『三省堂現代新国語辞典』三省堂
土井忠生・森田武・長南実編訳(1980)『日葡辞書：邦訳』岩波書店
中田祝夫編(1983)『古語大辞典』小学館
西尾実ほか編(2011)『岩波国語辞典(第 7 版)』岩波書店
山田忠雄ほか編(2012)『新明解国語辞典(第 7 版)』三省堂
日本大辞典刊行会(2001)『日本国語大辞典(第 2 版)』小学館

関連 URL

日本語歴史コーパス　https://chunagon.ninjal.ac.jp/chj
新編日本古典文学全集　http://japanknowledge.com.ezoris
web データに基づく形容詞用例データベース http://csd.ninjal.ac.jp/adj/

漢字とその訓読みとの対応の歴史的変遷

芮真慧（中国遼寧大学外国語学院日本語学科）

Historical Changes of the Correspondence between Kanji Characters and their Readings

Zhenhui Rui(The Japanese Department of College of Foreign Studies of Liaoning University)

要旨

中国における日本漢字研究を見てみると、音読み或いは国字に関する研究が多く、訓読みに関する研究はほとんどない。そこで、本研究は現在一般に行われている漢字とその訓読みの対応関係がどのように出来上がったのかを考察し、言語情報学的な研究手法を用いて考察することで、その歴史的変遷を明らかする。平安時代を中心に各時代における資料を介して「常用漢字表」（1981）を基準に一般の社会生活で最もよく使われる漢字とその訓読みを調査範囲としてその歴史的変遷を調べた結果、平安時代、室町時代、江戸時代、明治時代以降において、それぞれ常用字と常用訓というものがあり、時代により多少の相違はあるが、共通の部分が存在すること確かである。その共通部分は時代が進むとともに拡大していくことを実証的に論じた。

1. はじめに

本論文は現在一般に行われている漢字とその訓読みの対応関係がどのように出来上がったのかを平安時代以降の辞書を資料として考察し、その歴史的変遷を明らかにしたものである。「常用漢字表」（1981）を基準として一般の社会生活で最もよく使われる漢字とその訓読みを取り上げて調査の範囲を設定する。研究方法は、平安時代を中心にして、鎌倉室町時代、江戸時代、明治時代から昭和時代初頭（以下、明治時代以降）まで過去の文献資料と比較しながら「常用漢字表」の漢字とその訓読みについて検討することによって行う。①「常用漢字表」の漢字とその訓読みとの対応関係が平安時代以降においてどうなっているか、②確認できた漢字とその訓読みが各時代において一般的な読み方であったかどうかを中心に考察する。ここで言う一般的な読み方は「定訓」と呼ばれてきたものである。

2. 漢字の定訓

漢字・漢文の訓読が始まった当初、その訓は一つの漢字に対して複数存在し、固定的ではない。なお、訓読の方法が発達するとともに、1 義 1 訓の形に次第に訓が限定されていき、室町時代から江戸時代にかけて訓がかなり固定化される。明治時代以降、特に、戦後になってからは当用漢字の設け[1]や本論文で取り扱う「常用漢字表」など様々な漢字政策も行われ、漢字の数はもちろん読み方などもかなり整理される。こうして一つの漢字に対して固定的な読み方が定着し、一般化されるが、ここで言う一般的な読み方は「定訓」と呼ばれてきたものである。漢字の「定訓」について、今まで種々の研究が行われており、本節では定訓に関する先行研究と本論文で取り上げる「常用漢字表」について簡単に述べる。

2.1　定訓に関する先行研究

定訓に関する研究として取りあげられるのは小林（1970）、峰岸（1984a）、峰岸（1984b）、峰岸（1984c）、山田（1971）などである。

小林（1970）では、訓字[2]という用語を用いて上代における書記用漢字[3]の訓の体系につい

[1] 当用漢字表（1946）、当用漢字別表（1948）、当用漢字音訓表（1948）、当用漢字字体表（1949）および当用漢字改定音訓表（1973）など一連の法令によって定められた漢字政策全般を指す。

[2] 訓字とは、訓読の記入に際して、仮名やヲコト点とは別に、同訓異字の漢字を使って、「某也」或は「某」と傍記したり欄外に摘記したりするものを指す。この訓字には二つの場合が考えられる。第一は、原漢文

て研究を行っている。平安初期訓点資料[4]を用いて「平安初期訓点資料における読添え用の訓字一覧」を作成し、平安初期の訓点資料における訓字（例：令（シム）、如（ゴトシ）、申（モウス）、奉（マツル）など）は単に訓読を記入する一つの方式として、訓点の世界で工夫され、その世界に使用されただけではなく、上代から書記用漢字の体系が存在しており、それが平安初期の訓字にも現われているという点については奈良時代の文献を検討することで証明している。訓字の歴史的変遷の研究においては、ほかに小林（1974）と小林（1978）が挙げられるが、前者では、『新撰字鏡』の中の字訓の漢字を割り出し、その字訓の漢字は一字一訓が大多数を占めていることを証明している。また、これらの漢字は字種としては平易なものが多く、その訓も基本的なものが主となっており、一対一のものが多い。

峰岸氏は上代文献の漢字にはすでに「定訓」というものが存在しており、平安時代の文献においてもこの「定訓」は存在しているとする。峰岸（1984a）では、上代文献に使用された漢字について、『古事記』上表文の本文表記に関わる記事などを手掛りに定訓の存在を推定し、峰岸（1984b）は峰岸（1984a）に掲載できなかった、その論述に関わる基本資料の提示を中心に、そこに述べ残したところを補足したものであるが、前半で上代における漢字の定訓についてその語形を根拠となる資料とともに提示し、後半で上代における常用の漢字をその使用例と共に提示することで、上代に使用された漢字において定訓が存在したということを証明した。

また、峰岸（1984c）では、平安時代における漢字の定訓について詳細に記述している。真仮名文・漢字文・漢字仮名交じり文など漢字表記を有する文章における借字表記に注目し、『新撰万葉集』『日本紀竟宴和歌』（平安初期）、『将門記』と古記録（平安中期）から和訓に基づく借字表記を取り出し、分析することで当時期における漢字の定訓の存在を検証している。峰岸（1984c）での漢字の定訓に関する検証は、平安時代における漢字の定訓の存在を証明しただけではなく、三巻本『色葉字類抄』所収各項目の掲出最上位漢字に注目することによって、当代における日常常用の漢字の定訓についてもその全貌を多少知る手掛かりをえることができたのである。例えば、峰岸（1984c）で取り上げている「借」の場合、「借」と「カル・カス・カリ」の関係は常用漢字とその訓の関係と同様であって、これは現在まで残っている。「借」は「常用漢字表」に収録されている漢字であり、それに「かりる」という字訓が定義されている。つまり、「借」に対する「かりる」という訓は平安時代から定着していたわけである。

一方、山田（1971）は、「訓が複数もしくは多数認められる時、その諸訓の中でどんな関係が見られるのか」という主題をめぐって、キリシタン版『落葉集小玉篇』を資料にして漢字の定訓の存在を証明し、「定訓」について次のように述べている。

> 某一字について、その呼称を考へる時に、直ちに喚起される字訓を、先づ第一にその字の定訓（又はその一つ）に擬することが許されるであらうと考へる。それは又、一般に、漢字の三要素といはれる形音義の、音とならんですでに、その字固有の呼称となったものと考へてもよいであらう。しかしながら、その定訓は訓である以上、字義と全く無関係には成立しない。（中略）このやうな意味で、その字を指し示すに援用できて、十分その機能がみとめられるレベルに達してゐる語を、その字の定訓ということができよう。

の漢字と、その訓を表すために注記された漢字とに対応関係のある場合である（例：「盛（入也）」「造（至也）」）。第二は、原漢文にはそれに対応する漢字がないが、訓読に当たって、読添える必要のあるテニヲハ[2]を、そのテニヲハの訓に当たる漢字で記入する場合である（例「[令（シム）] 一未（は）信者（の）信（人を）、々不（令むる）退故（せる）（に）（山田本観弥勒上生経賛平安初期朱点）」）。

[3] 書記用漢字とは漢字に対する「訓」を背景として、日本語をその漢字によって書記するものの、漢字を用いて日本語として文章を書記したものを指している。具体的には和化漢文・訓仮名に依る万葉仮名表記・宣命体などであると述べている。

[4] 『持人菩薩経』『願経四分律古点』『中観論古点』『東大寺諷誦文』『妙法蓮華経化城喩品古点』など計 26 点の訓点資料を扱っている。

つまり、「定訓」とはある時代・ある地域で一般的に用いられ、その字にある程度定着されたものである。小林芳規の訓字研究をはじめ、峰岸明の上代文献における借用表記を用いた定訓に関する研究はもちろん山田俊雄の『落葉集』を資料とした研究は全て定訓というものが存在したということを証明している。

2.2　現在の定訓

本研究では「常用漢字表」(1981) 1945 字を基準として一般の社会生活で最もよく使われる漢字とその訓読みを取り上げて調査を行っているが、現在はそれを改訂した「常用漢字表」（2010）2,136 字が行われているため、追加されている 196 字については別に調査を行うことにしている。

1981 年、日本内閣訓令告示によって公布された「常用漢字表」はその字種と音訓[5]の選定に当たって「語や文書を書き表すという観点から、現代の国語で使用される字種や音訓の実態に基づいて総合的に判断する」という原則を取っており、法令・公用文書・新聞・雑誌・放送など、一般の社会生活で用いる場合の効率的で共通性の高い漢字を収めることにしている。しかし、常用漢字表」には「遵」「勺」「逓」[6]のようなあまり使われていないものが収録されており、「誰」「奈」「頃」「阪」「岡」[7]のような普段よく使われているものは収録されていない。このような問題が原因で「常用漢字表」の見直しに関する議論が始まり、2010 年 11 月 30 日、内閣告示第 2 号によって新しい「常用漢字表」が公布されるが、「改定常用漢字表」の字種選定のために行われた「漢字出現頻度数調査」[8]を用いて「常用漢字表」(1981) 所載の漢字を見てみると 1,945 字のうち、60 字を除いて他のものは出現頻度数順位が 2,500 位以内のものである。したがって、漢字数は別にして漢字が常用度の高いものであれば本論文の一般の社会生活でよく使われている漢字を取り上げようとする趣旨に反しない。

そこで、「常用漢字表」（1981）における漢字の音訓状況を分析し、整理すると、1,945 字のうち、音読みのみ定義されている漢字が 737 字、訓読みのみ定義されている漢字が 40 字、音訓ともに定義されている漢字が 1,168 字である。本論文では、訓読みの定義されている漢字 1,208 字を研究対象の候補とし、さらに常用訓の数によって分類すると、複数の常用訓を持つ常用字が 445 字、一つの常用訓を持つ常用漢字が 763 字である。漢字とその訓読みとの対応と定着度を見ることが目的であるから、まず常用訓が一つの常用字を検討し、その後常用訓が複数の漢字を検討する。なお、便宜上、「常用漢字表」における漢字は常用字と呼び、それに対応する訓読みは常用訓と呼ぶ。

なお、先行研究においては、主に「訓字」と「定訓」という用語が出てくるが、常用字とそれに対応する常用訓は漢字と訓の関係を示す点においては、訓字や定訓と同様である。従って、本論文では統一して常用字、常用訓という用語を用いることにする。一方、各資料における常用字と常用訓については「常用漢字表」の常用字・常用訓と区別するために、「」を用いて「常用字」「常用訓」と示す。

3. 研究方法と調査資料

[5] 音訓については、当用漢字音訓表（1948）を原則として受け継ぎ、新しく加わった漢字については、当表にあげたものに準じて新たに音訓を選定した。

[6] 文化庁の平成 18 年度世論調査によると、「遵」「勺」「逓」は「よく使われていると思う」「時々使われていると思う」を合わせると 3 割台半ば，「余り使われていないと思う」「全く使われていないと思う」を合わせると約 6 割となっている。

[7] 文化庁の平成 18 年度世論調査によると、「誰」「奈」「頃」「阪」「岡」は「よく使われていると思う」だけで 8〜9 割である。一方、「余り使われていないと思う」「全く使われていないと思う」を合わせても，1 割に満たない。

[8] この調査は「教育等の様々な要素はいったん外して、日常生活でよく使われている漢字を出現頻度調査の結果によって機械的に選ぶ」という考え方に基づいて実施されている。

従来の定訓に関する研究をまとめてみると大きく三つに分けられる。一つは上代文献を利用した借用表記による定訓の確認であり、もう一つは訓字を用いて漢字とその和訓の関係を証明したものである。最後に取上げられるのは『類聚名義抄』『色葉字類抄』『落葉集』など辞書を利用して定訓の存在を証明している研究である。そのうち、借用表記を利用した研究方法は上代文献に限られ、訓字による研究も訓点資料の膨大さなどを考えると実行するには困難が大きい。そのため本論文では、峰岸（1984）や山田（1971）などの研究成果を踏まえて、各時代の代表的な辞書を取り上げて調査を行うことにする。

平安時代においては『類聚名義抄』『色葉字類抄』の 2 種類の辞書を取り上げて調査を行い、さらに参考資料として『訓点語彙集成』を取り上げることにする。平安時代以降においては、大きく鎌倉室町時代（中世)、江戸時代（近世)、明治時代以降（近現代）に分けて調査を行い、取り扱う資料は次のとおりである。

室町時代：『節用集』『倭玉篇』『落葉集』
江戸時代：『書言字考節用集』『増続大広益会玉篇大全』『和英語林集成』
明治時代以降：『大言海』『大字典』『和英袖珍新字彙』

これらの辞書は各時代の日本語表記の基準を反映した規範性の高い文献である。言葉の世界で規範性の高いものと言えば辞書が代表的であり、新しい言葉が出現してきてもある程度社会に定着しない限り、辞書には収録されない。逆に言うと辞書に収録されているということはその語が社会的に認知されていることを示している。一方「常用漢字表」は現代の日本語表記の基準として行われる規範そのものである。各時代の実際の日本語表記の実態とは差があると考えられるがまずは規範的文献の内容を整理・分析し、次の段落で通常の文章における「常用字」「常用訓（定訓）」の実態を記述していくのがよいだろう。本論文で辞書を中心に検討するのはこのような理由によるものである。

また、本研究の研究対象となる 763 字についてはその常用訓を品詞によって分類してから調査し、大きく名詞 393 字 393 語（以下、393 字と略。他の品詞も同様。）、動詞 293 字、形容詞 57 字、その他 20 字に分ける。

4. 各時代における常用字と常用訓の対応関係

4.1　平安時代における常用字と常用訓の対応関係

平安時代においては三巻本『色葉字類抄』観智院本『類聚名義抄』及び『訓点語彙集成』を取り上げて調査を行う。『色葉字類抄』と『類聚名義抄』はそれぞれ平安時代の国語辞書と漢和辞書である。『訓点語彙集成』は平安時代の実際の文献における使用例を集めたものであり、平安時代の訓点資料を中心に複数の訓点本における和訓語彙が収集されている。

この三つの資料において確認できる常用字と常用訓（名詞）を示すと【表 1】の通りである。「○」は対応あり、「×」は対応なしを示す。

【表 1】平安時代の資料における常用字と常用訓の対応（名詞）

分類	色葉字類抄	類聚名義抄	訓点語彙集成	合　計
A	○	○	○	270（68.7%）
B	○	○	×	11（ 2.8%）
C	○	×	○	16（ 4%）
D	×	○	○	17（ 4%）
E	○	×	×	5（ 1%）
F	×	○	×	2（0.5%）
G	×	×	○	25（ 6%）
H	×	×	×	47（12%）
合計	302（76.8% ）	298（75. 8% ）	328（83.5% ）	393（100%）

紙幅の関係上動詞（293 字)、形容詞（57 字）とその他（20 字）については表を取り上げ

ないが、数字を見てみると A 類つまり『色葉字類抄』『類聚名義抄』『訓点語彙集成』全ての資料に収録されているものに属するのがそれぞれ動詞 182 字、形容詞 39 字(68.4%)、その他 7 字（35.0%）である。

以上から分かるように、名詞の場合は68.7%、動詞の場合は62.1%、形容詞の場合は68.4%、その他の場合は 35.0%がすべての資料において確認できる。その他を除いて品詞別の差はあまり見られず、どちらも 6 割を超えている。すなわち、全 763 字のうち、498 字（65.3%）は『色葉字類抄』『類聚名義抄』『訓点語彙集成』全ての資料に収録されている。そこで、各資料における常用字と常用訓の対応を見てみると、『色葉字類抄』が 74.7%、『類聚名義抄』が 74.1%、『訓点語彙集成』が 81.0%を占めている。これは大多数の常用字と常用訓において、平安時代から現在に至るまでその対応関係に変化が生じてないことを示している。

なお、ここで問題となるのはこれらの常用字と常用訓が平安時代においても一般的なものであったかどうかという点である。この問題を解決するために、本研究では研究資料として取り上げている『色葉字類抄』『類聚名義抄』『訓点語彙集成』における「常用字」と「常用訓」を確認し、両者を比較している。『色葉字類抄』『類聚名義抄』『訓点語彙集成』における「常用字」と「常用訓」の判断は次のように行う。『色葉字類抄』は漢字に対する合点の有無と配列順位、『類聚名義抄』は和訓に対する声点の有無と配列順位、『訓点語彙集成』はその用例漢字と用例数を分析する。この方法により、各資料における「常用字」と「常用訓」(定訓と考えられるもの)を確認する。これは芮（2011）によって発表されたものであり、その結果によると「Ⅰ．常用字が『訓点語彙集』で用例数の一番多い用例漢字である。Ⅱ．常用訓の『類聚名義抄』での掲出順位が最上位である。Ⅲ．常用字の『色葉字類抄』での掲出順位が最上位である。」という三つの条件を全部満たすものは、A の分類に属する 270 字のうち、174（64.4 %）字である。これら 174 字は平安時代において常用字と常用訓が安定した対応関係を成していたと判断してよいであろう。一方、三つの条件のうち、二つを満たしているのは 79 字、一つを満たしているのは 12 字、三つとも満たしていないのは 5 字である。また、各資料における常用字と常用訓の対応を見てみると『色葉字類抄』が 76.8%、『類聚名義抄』が 75.8%、『訓点語彙集成』が 83.5%を占めている。この結果は常用字とその常用訓の対応関係が平安時代から定着していたということを示している。

4.2　室町時代における常用字と常用訓の対応関係

室町時代においても平安時代と同じく国語辞典の一種である『節用集』と漢和辞典『倭玉篇』及び参考として漢字辞典『落葉集』の三つの資料を取り扱う。まず、この三つの資料において確認できる常用字と常用訓がどれぐらいあるかを確認するが、調査対象は名詞の常用訓とそれに対応する常用字を取り上げて分析を行う。調査対象を名詞に限定したのは、名詞には活用形がなく判定が容易であるからである。また、古辞書には名詞が優先的に登載される。それに、すでに述べたように「常用漢字表」所載の常用訓が一つの漢字には同訓異字のものがあり、名詞（6.4%）と比べて動詞（14.0%）と形容詞（12.3%）はその数が多い。従って、同訓異字の影響が少ない名詞から調査を行うことにする。本論文での品詞分類によると調査対象となる名詞の常用字・常用訓は計 393 字であり、その結果を示すと【表 2】のとおりである。

【表 2】室町時代における常用字と常用訓の対応

節用集	倭玉篇	落葉集	合計
○	○	○	263（66.9%）
○	○	×	35（ 8.9%）
○	×	○	5（ 1.3%）
×	○	○	24（ 6.1%）
○	×	×	5（ 1.3%）

×	○	×	7（ 1.8%）
×	×	○	11（ 2.8%）
×	×	×	43（10.9%）
308（78.3%）	328（83.5% ）	303（77.1% ）	393（ 100%）

【表 2】から分かるように 393 字のうち 263 字（66.9％）は常用字と常用訓との対応が三資料に確認できるものである。これは『節用集』（易林本）のみ用いた場合の数字で他の写刊本も使って調べると三資料すべて確認できるのは 277 字（70.5%）になる。『節用集』諸本の総計は 308 字（78.3%）が 325（82.7%）となる。

なお、ここで説明したいのは、『節用集』において本論文では易林本を取り上げているが、平安時代の資料とは異なって『倭玉篇』『節用集』『落葉集』の三つの資料においては『落葉集』以外、各資料における「常用字」と「常用訓」を判定する先行研究はない。キリシタン『落葉集』に関する研究としては先行研究で紹介した山田（1971）が取り上げられるが、それによると『落葉集』に収録されている単字の右側もしくは左側に位置する訓は、いわゆる定訓（標準的な訓）として示されている。つまり、漢字の左右に示されている訓は『落葉集』における「常用訓」であり、訓の位置からそれが「常用訓」であるかどうかを判断することができる。

そこで、『落葉集』の「常用字・常用訓」264 字と『節用集』『倭玉篇』を比べてみると、共通しているものは 237 字（89.7%）であり、名詞全体の（393 字）60．1％を占めている。これに比べて平安時代において「常用字」「常用訓」と思われるものは 174 字であり、名詞全体の約 44.3%を占めているにすぎない。そこで、平安時代における調査において「常用字」「常用訓」と思われるもの 174 字と『落葉集』における「常用字」「常用訓」の 264 字を比較してみると一致しているものは計 139 字あり、平安時代の「常用字」「常用訓」の 79.9%を占めている。これは、平安時代において「常用字」「常用訓」であったものが室町時代においてもその対応関係は変わらず非常に安定しているということを示している。

4.3　江戸時代における常用字と常用訓の対応関係

江戸時代においては『書言字考節用集』『増続大広益会玉篇大全』『和英語林集成』を扱っているが」、『書言字考節用集』は、1717 年に刊行された分類体辞書であり、イロハ順に配列されており、その部門は 『節用集』（易林本）に大きく影響されている。『増続大広益会玉篇大全』は毛利貞斎が中国南北朝の『玉篇』を校正・増補した漢和辞典である。『和英語林集成』（*A Japanese-English and English-Japanese dictionary*）は、19 世紀後半にジェームス・カーティス・ヘボン（*James Curtis Hepburn*）が収集した日常語を中心に編纂した日本最初の和英、英和辞典である。その結果は【表３】のとおりである。

【表 3】江戸時代の資料における常用字・常用訓の対応

書言字考節用集	増続大広益会玉篇大全	和英語林集成	合計
○	○	○	297（75.6%）
○	○	×	14（ 3.6%）
○	×	○	19（ 4.8%）
×	○	○	10（ 2.5%）
○	×	×	8（ 2.0%）
×	○	×	4（ 1.0%）
×	×	○	21（ 5.3%）
×	×	×	20（5.1%）
338（86.0 %）	325 （82.6% ）	347（88.3% ）	393（100%）

近世においても確認できる常用字・常用訓はその数が多く、393 字のうち 297 字（75.6%）

であり、平安時代の 272 字（69.2%）と中世の 263 字（66.9%）を上回っている。なお、平安時代に比べて中世においてその数字があまり変わっていないのは『訓点語彙集成』と『落葉集』の資料の性格が異なっているのが原因であろう。仮に、『訓点語彙集成』と『落葉集』を除いて、国語辞書と漢和辞書による結果をみると、『色葉字類抄』と『類聚名義抄』の共通の常用字・常用訓は 393 字のうち 280 字（71.2%）であり、『節用集』と『倭玉篇』の共通の常用字・常用訓は 293 字（74.6%）である。『書言字考節用集』と『増続大広益会玉篇大全』の共通の常用字・常用訓は 311 字（79.1%）であり、その数字は徐々に上がっている。さらに、平安時代から江戸時代までの九つの資料における共通の常用字・常用訓を見てみると 393 字のうち 208 字（52.9%）が一致している。

4.4　明治時代以降における常用字と常用訓の対応関係

明治以降においては常用字と常用訓の対応が前の時代より遥かに上回っていくことが予想されるが、その結果は【表 4】のとおりである。

【表 4】明治時代以降における常用字・常用訓の対応

大言海	大字典	和英袖珍新字彙	合計
○	○	○	346 (88.0%)
○	○	×	22 (5.6%)
○	×	○	3 (0.7%)
×	○	○	4 (1.0%)
○	×	×	4 (1.0%)
×	○	×	10 (2.5%)
×	×	○	1 (0.2%)
×	×	×	3 (0.7%)
375 (95.4%)	382 (97.2 %)	354 (90.1%)	393 (100%)

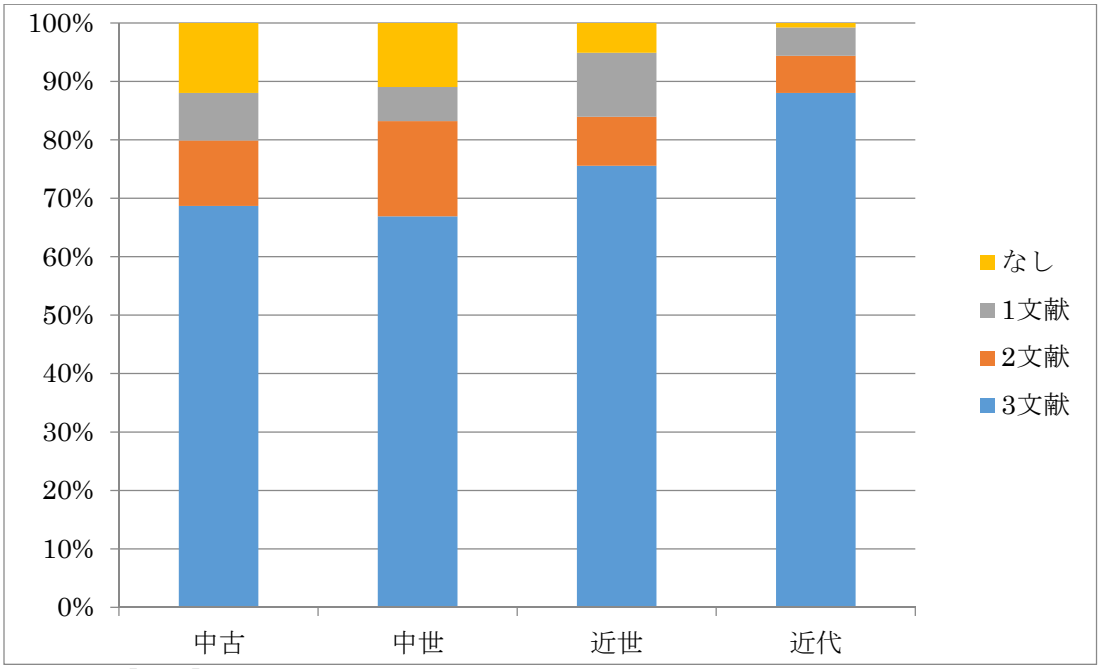
表【4】から分かるように近代においては常用字と常用訓の対応の割合は非常に高い。三つの資料に共通しているのが 346 字で全体の 9 割近くの比率を占めている。各資料においてもそれぞれ『大言海』が 95.4%、『大字典』が 97.2%、『和英袖珍新字彙』が 90.1%を占めており、対応してないのは約 1 割程度のものである。そのうち、『和英袖珍新字彙』のみで対応を成していない常用字と常用訓の数（22 字）が他に比べて少し多い。これは、国語・漢和辞書に比べて和英辞書に収録されている語彙の数が少ないからである。

5. 常用字と常用訓の対応関係の歴史的変遷

本研究の調査範囲である常用字・常用訓（名詞）393 字について平安時代の三つの資料において全部確認できるのは 272 字であり、室町時代においては 393 字のうち、263 字である。江戸時代と明治時代においてはそれぞれ 294 字と 346 字がその時代の全ての資料において確認でき、208 字が 12 点の資料において対応関係を認めることができる。一方、確認できなかった常用字と常用訓について見ると平安時代は 47 字、室町時代は 43 字、江戸時代は 15 字、明治時代以降は 3 字である。これは、平安時代において安定していなかった常用字と常用訓が室町時代からはますます安定するようになったということを証明していると理解できる。そこで、平安時代から明治までの調査結果を示すと【図 1】のようになる。例えば 3 文献は各時代について三つの文献に出てくることを示す。

時代	3 文献	2 文献	1 文献	なし	計
中古	270	44	32	47	393
中世	263	64	23	43	393
近世	297	33	43	20	393

近代	346	25	19	3	393
時代	3 文献	2 文献	1 文献	なし	計
中古	68.7%	11.2%	8.1%	12.0%	100.0%
中世	66.9%	16.3%	5.9%	10.9%	100.0%
近世	75.6%	8.4%	10.9%	5.1%	100.0%
近代	88.0%	6.4%	4.8%	0.8%	100.0%



【図 1】中古から近代までの常用字と常用訓の対応関係の変遷

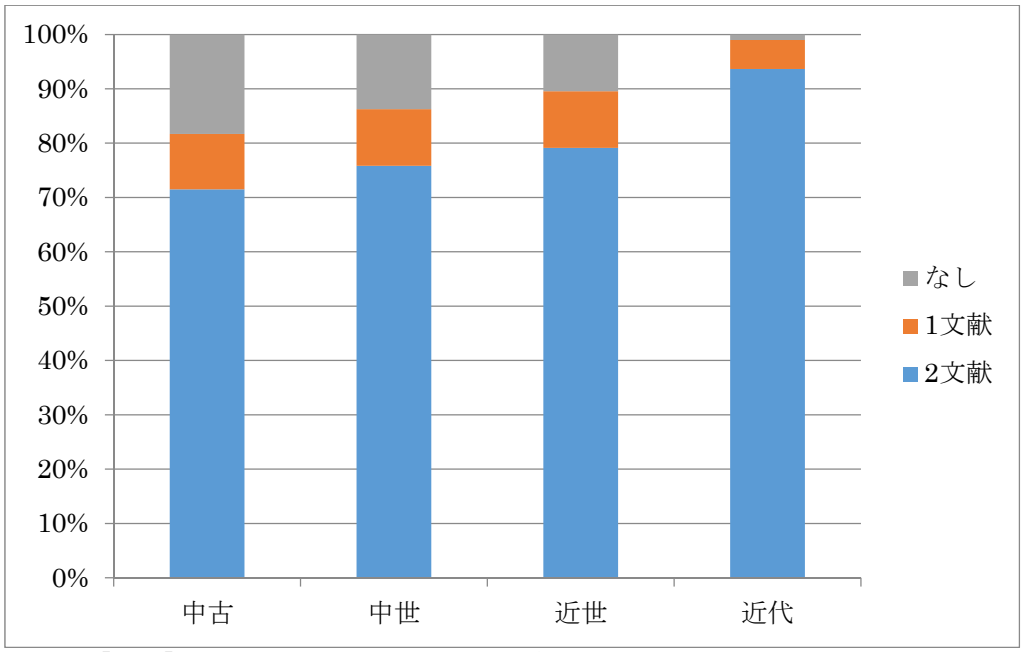
【図１】から分かるように中古より中世の常用字と常用訓が対応する比率が低い。これはおそらく『訓点語彙集成』と『落葉集』の性格が異なっているからである。すでに紹介したように膨大な訓点資料の和訓語彙を集めた『訓点語彙集成』に対して、『落葉集』は先達が用いた文字と言葉の今に残存しているものを広く収集したものである。「今に残存しているもの」という内容からも分かるように『訓点語彙集成』に比べて『落葉集』に収録されている語彙が少ないのは明らかである[9]。また、見出し語に対応する漢字の数も大きく異なる[10]。そこで、各時代の国語辞書と漢和辞書による結果を示すと【図 2】のようになる。

時代	2 文献	1 文献	なし	計
中古	281	40	72	393
中世	298	41	54	393
近世	311	41	41	393
近代	368	21	4	393
時代	2 文献	1 文献	なし	計
中古	71.5%	10.2%	18.3%	100.0%

[9] 名詞の常用訓・常用字 393 字のうち、『訓点語彙集成』において確認できたが『落葉集』において確認できなかったものは 47 字であり、『訓点語彙集成』においては確認できなかったが、『落葉集』において確認できたものは 21 字である。なお、この 21 字のうち 10 字は平安時代において確認できなかったものである。

[10] 「盾／たて」の場合『落葉集』においては「楯／たて」の対応関係であるが、『訓点語彙集成』においては「たて／楯 14・干 6・盾 3……」の対応関係である。

中世	75.8%	10.4%	13.7%	100.0%
近世	79.1%	10.4%	10.4%	100.0%
近代	93.6%	5.3%	1.0%	100.0%



【図 4】国語・漢和辞書による常用字・常用訓の対応関係の変遷

【図 4】から分かるように時代が進んでいくとともに対応関係を成す常用字とその常用訓の数は多くなる。対応関係だけではなくその定着度もますます高くなっている。平安時代の「常用字・常用訓」と思われるものが 174 字であるに対して、室町時代は 244 字である。対応している漢字とその訓の数に差が見られなくても定着度は大きく異なっている。

6. おわりに

本研究では、現代日本語における漢字とその訓読みとの対応関係について、平安時代、室町時代、江戸時代、明治時代以降の資料を三つずつ取り上げて分析した。現代日本語における漢字とその訓読みの対応は、「常用漢字表」（1981）所載の漢字の常用訓が一つのもの(763 字/語)とし、その考察内容をまとめると、次のようになる。

(1) 平安時代においては、名詞 393 字、動詞 293 字、形容詞 57 字、その他 20 字に分けて調査したところ、名詞 68.7%、動詞 62.1%、形容詞 68.4%、その他 35.0%という結果を得た。これによって、「常用漢字表」(1981)の常用字と常用訓との対応が見られるものが多いことが明らかになった（その他 20 字はもともと例が少ないので除外）。平安時代における定着度が高いと判定される「常用字」「常用訓(定訓)」との対応を見ると、四割以上（名詞 44%、動詞 52.7%、形容詞 46.2%）が一致していることが明らかになった。

(2) 室町時代においては、「常用漢字表」(1981)の常用字と常用訓(名詞のみ)の対応関係が確認できるものは、6 割以上（393 字中 263 字、66.9%）を占め、平安時代と比べてあまり変化していない。次に室町資料における「常用字」「常用訓(定訓)」と思われるものは平安時代より多く、237 字であり、名詞全体の（393 字）60.1%を占めている。平安時代より室町時代のほうが常用字と常用訓の対応関係が定着・安定している。

(3) 江戸時代以降になると、常用字と常用訓が対応しているものが大多数であり、88.0%を占める。また、室町時代の「常用字」「常用訓(定訓)」と思われる 237 字のうち、224 字は江戸時代以降の六つの資料にてその対応関係が見られる。このような結果は、

漢字とその訓読みの対応関係は平安時代から変化していないものが多く、それが定着するようになるのは主に室町時代以降であるということを示している。

このように漢字とその訓読みとの対応関係の全体像を把握するため、本論文では各時代の資料を三つずつ取り上げて調査を行った。今まで『色葉字類抄』や『類聚名義抄』などの資料を用いて「定訓」の存在を考察した研究はあったが、三つの資料を同時に用いて常用字と常用訓との対応関係を考察したものはない。なお、資料の性格差による相違や資料ごとの分析については言及しなかったため、検討において不十分なところがある。しかし、平安時代、室町時代、江戸時代、明治時代以降において、それぞれ常用字と常用訓というものがあり、時代の流れによって多少その範囲は異なってくるが、共通の部分が存在することは確かである。

文　献

小林芳規（1970）「上代における書記用漢字の訓の体系」『国語と国文学』47-10　東京大学国語国文学会　pp. 50-80
小林芳規（1974）「新撰字鏡における和訓表記の漢字について--字訓史研究の一作業」『文学 』42-6　岩波書店　pp. 80-99
小林芳規（1978）「漢字とその訓との対応及び変遷についての一考察」『国語学』112　武蔵野書院　pp. 11-28
小松英雄（1963）「語調資料としての類聚名義抄－図書寮本および観智院本にみえる和訓の声点の均質性の検討－」『国文学漢文学論業』9　東京教育大学文学部　pp. 1-37
小松英雄（1966）「声点の分布とその機能（1）」－前田家蔵三巻本『色葉字類抄』における差声訓の分布の分析－」『国語国文』35-7　京都帝国大学国文学会　pp. 1-34
芮真慧（2011）「平安時代における常用字と常用訓」『国語国文研究』139　北海道大学国語国文学会　pp. 70-80
舩城俊太郎（1976）「三巻本色葉字類抄につけられた朱の合点について」『二松学舎大学論集』51　二松学舎大学論集　pp. 59-89
舩城俊太郎（2011）『院政時代文章様式史論考』勉誠出版
峰岸明（1971）「今昔物語集における漢字の用法に関する一試論[一]―副詞の漢字表記を中心に―」『国語学』85　国語学会　pp. 18-35
峰岸明（1984a）「上代における漢字の定訓について」『横浜国大国語研究』2　横浜国立大学国語国文学会　pp. 1-13
峰岸明（1984b）「上代漢字の定訓考証：『万葉集』を資料として」『横浜国立大学人文紀要　第二類　語学・文学』31　横浜国立大学教育学部　pp. 85-106
峰岸明（1984c）「平安時代における漢字の定訓について」『国語と国文学』61　東京大学国語国文学会　pp. 44-60
宮澤俊雅(1992)「図書寮本類聚名義抄の注文の配列について」『小林芳規博士退官記念国語学論集』汲古書院
山田俊雄（1971）「漢字の定訓についての試論：キリシタン版落葉集小玉篇を資料として」『成城国文学論集』4　成城大学大学院文学研究科　pp. 1-256

調査資料

イーストレーキ・神田乃武（1891）『和英袖珍新字彙』三省堂
上田万年・岡田正之[ほか]（1917）『大字典』啓成社
大槻文彦（1932－1935）『大言海』冨山房
J．Cヘボン著、飛田良文・李漢燮編集（2001）『和英語林集成：初版・再版・三版対象総索引』港の人
築島裕（2007）『訓点語彙集成』（第一巻、第二巻、第三巻）汲古書院
築島裕（2008）『訓点語彙集成』（第四巻、第五巻、第六巻）汲古書院
築島裕（2009）『訓点語彙集成』（第七巻、第八巻、別巻）汲古書院
中田祝夫・峰岸明（1964）『色葉字類抄―研究及び索引本文索引篇』風間書房
中田祝夫（1968）『古本節用集六種研究並びに総合索引』風間書房
中田祝夫・小林祥一郎（1973）『書言字考節用集研究並びに索引』風間書房
中田祝夫・北恭昭編纂（1976）『倭玉篇研究並びに索引』風間書房
福島邦道解説（1973）『キリシタン版落葉集』勉誠社
正宗敦夫（1962）『類聚名義抄』風間書房
毛利貞斎（1692）『増続大広益会玉篇大全』京都・沢村昌益

「. . . 事實也。」から「。事実. . . 」へ
－談話機能の発達に伴う統語位置の変化－

柴﨑礼士郎（明治大学）†

From Predicate Use to Adverbial Use: Syntactic Changes in Tandem with Discourse-Functional Development

Reijirou Shibasaki (Meiji University)

要旨

本稿は文頭・節頭（以下文頭と略記）に使用される「事実（事實），. . . 」に注目し、特に明治期以降の史的発達を考察する。北原・他（2006）によれば、文末・節末（以下文末と略記）に使用される「事實也」（名詞＋繋辞）のような述部用法は平安期から確認可能であるが、文頭に使用される副詞用法は 20 世紀初頭からと記述されている。そこで本稿では、『国民の友コーパス』、『明六雑誌コーパス』、『近代女性雑誌コーパス』および『太陽コーパス』を使用し、明治大正期における「事実」の文頭副詞機能の発達経緯を詳細に分析する。更に、『現代日本語書き言葉均衡コーパス』（特に書籍ジャンル）を用いて 1970 年代から 2000 年代初頭における直近の変化を捉える。調査結果から「文末用法＞文中用法＞文頭用法」という史的発達が確認できるものの、現代日本語においては文頭用法（「事実」）と文末用法（「事実である、事実です」）に特化した分布が見て取れる。

１．はじめに

2010 年代に入り、名詞構文が新たな注目を集めている印象を受ける。例えば、角田(2012)の提示する「人魚構文」（角田(1996)で提示された「体言締め文」の新展開）は、その命名からだけでも目を引くものであるし、鳴海(2015)による漢語名詞の副詞化に関する研究も既存の国語学の枠を超える質感を伴う。他方、ニュース報道で使用されている名詞構文に正面から取り組む田中(2012)などもある。

対照言語学的色合いの濃い新屋(2014: 第 1 章)によれば、これまで翻訳研究を中心に指摘されてきた「英語＝名詞中心、日本語＝述語中心」という見解は、どうも再考の余地があるとのことである。例えば以下の例文の下線部に注目したい。

(1) 何かあった模様だ。
(2) どうやら無事におさまった気配だ。（新屋 2014: 8）

「わけ、ところ、つもり、もの、こと」などの形式名詞と異なり、実質的な意味を有する名詞が文末詞的な働きをすることに新屋(2014)は注目し、こうした表現を含むものを「文末名詞文」と呼んでいる。日本語の形態統語構造に注目した形式名詞の文法化なども注目すべき現象であるが（e.g. Shibasaki 2011）、実質名詞の多機能性に注目することにより、日本語の名詞句・名詞構文を対照言語学的あるいは通言語学的に再解釈する意義が見いだせると思われる。本稿では、実質的意味を保持する「事実」に注目し考察を進める。

北原・他(2006)に従い極簡単な史的変遷を以下に示す。(3)に示すように、「事実」は名詞

† reijiro（at）meiji.ac.jp　*(at)の部分を@に変えて御使用ください。

として述部の一部に組み込まれて用いられていたが、現代の日本語では(4)のように副詞的機能を果たす場合も多い。(3)は名詞として、(4)は副詞としての初出例である。

(3) 摂政被来云、今夜斉院盗人入云々、仍奉遣奉云々、右大弁来云、斉院事実也。
（寛仁元年（1017）七月二日『御堂関白記』; 北原・他 2006）
(4) 兄さんは誰よりも今の若い人達の心をよく知ってゐる。そして事実、東京で若い多くの女のお友達もおありの事であったらうし。
（1914『田舎医師の子』＜相馬泰三＞五；北原・他 2006）

(3)では「事実なり」のように述部の一部として使用されているが「事実」は実質的意味を保持しており、(4)では接続詞を伴った形で文副詞的機能を果たしている。また、副詞機能が 20 世紀初頭頃に生起し始めた可能性も(4)から分かる。これら以外にも指示詞や節を伴う用法もあるが、「事実」は提題助詞なども伴わない独立用法を特に発達させている。そこで本稿では、「事実」の使用を文レヴェルで捉え、述語の一部としての「文末用法」から副詞としての「文頭用法」への拡張過程をコーパスを用いて考察する。

本稿の構成は以下の通りである。第 2 節では研究の背景を簡潔に提示し、第 3 節ではコーパスを用いた調査結果を提示する。第 4 節では調査結果の意義を例示する。第 5 節はまとめである。使用するコーパスは表 1 の通りである。尚、『現代日本語書き言葉均衡コーパス』については、近年の史的変遷を見るため、および、他のコーパスとの整合性（ジャンル）を揃えるために、今回の調査では「書籍」ジャンルに限定してある。

表 1 使用コーパス[1]

コーパス	語彙数	時期	備考
『明六雑誌コーパス』	約 18 万語	1874-1875 年(明治 7-8 年)	
『国民之友コーパス』	約 101 万語	1887-1888 年(明治 20-21 年)	
『近代女性雑誌コーパス』	約 210 万字	1894-1895 年(明治 27-28 年) 1909 年(明治 42 年) 1925 年(大正 14 年)	『女学雑誌』(1894-1895 年) 『女学世界』(1909 年) 『婦人倶楽部』(1925 年)
『太陽コーパス』	約 1450 万字[2]	1895 年(明治 28 年) 1901 年(明治 34 年) 1909 年(明治 42 年) 1917 年(大正 6 年) 1925 年(大正 14 年)	
『現代日本語書き言葉均衡コーパス』(BCCWJ)	約 6,270 万語	1971-2005 (昭和 46 年-平成 17 年)	「書籍」ジャンルのみ使用

２．研究の背景

前節で紹介した名詞研究に加え、高橋・東泉(2013, 2014)や東泉・高橋(2013)の取り組みは注目に値する。以下の例文で確認してみる。

[1] 国立国語研究所のホームページを参考に作成してある。

[2] 『太陽コーパス』の収録語彙数については近藤(2013)にヒントがある。同論文を紹介して下さった東泉裕子先生へ御礼申し述べます。

(5) 人民の情と合和して、かかる結菓となりしなり。
（1872『自由之理』＜中村正直訳＞; 北原・他 2006; 高橋・東泉 2014: 104）
(6) 親戚朋友度々相往來し、相共に飲食談笑せし結果、流れ〜〜て、果ては、多くの虚禮が、うるさき迄に出來しならんか。(1895 HM 生「歳暮」;『太陽コーパス』)
(7) 女にはなぜ作曲家がいない？「そこで、女のものの考え方について非作曲家的なところを考えてみた。結果、女の考え方というのは、1+1 は 2 であるということだ」
（1974-75 藤本義一『男の遠吠え』; 北原・他 2006; 高橋・東泉 2014: 107）

北原・他(2006)によれば、実質名詞としての「結果」は(5)のように述部の一部として使用されはじめ、徐々に(6)に示す連体修飾を受けて接続詞的に用いられる用法が発達している。その後 20 世紀の後半に入り、(7)のような「前文を受けて副詞的に用いる」談話機能に至っている。

ここまでの調査報告であれば、既存の国語学と言語学の成果に基づく亜流とみなされる可能性もある。しかし、高橋・東泉(2013, 2014)と東泉・高橋(2013)から読み取ることができる下記の点は、今後の言語変化を俯瞰的に捉えられる可能性を含んでいる。つまり、実質的内容を持つ名詞として生起した「結果」が、述部の一部として用法を発達させ、節接続機能を創発し、最終的には文頭の副詞機能に至っている点である。換言すると、文レヴェルの構文として考えた場合、述部という「文末用法」としての機能から（上述の新屋 2014 に詳しい)、節接続用法という「文中用法」、そして、前文を受けて後述の情報を導入する「文頭用法」という機能拡張は、談話機能の発達に伴う統語変化として「文末＞文中＞文頭」のようにまとめられる。大局的に見れば、形式言語学のアプローチによる Roberts and Roussou (2003)の研究成果と一致している。一方、機能言語学の視点から英語の *then* に注目し、歴史的に「文頭＞文中＞文尾」という談話機能と統語位置の拡張が確認できるとする Haselow (2012)の研究成果とは逆方向の変化となる。こうした文（あるいは発話）の周辺から周辺へという変化は近年注目を集めていることからも（e.g. Beeching and Detges 2014)、高橋・東泉の研究は示唆に富んでいる（Higashiizumi 2015 も参照)。[3]

もう一点付け加えるとすれば、高橋・東泉の一連の研究成果は、北原・他(2006)で例示されている「初出」例よりも早い事例を紹介できている点である。これは、高橋・東泉の入念な調査もさることながら、「コーパス」というツールの効用と見るべきであろう。

本稿では、高橋・東泉(2013, 2014)および東泉・高橋(2013)の研究成果が、果たして「事実」という異なる実質名詞の機能拡張に応用可能かどうかも確認したい。[4]

３．考察手順と結果

紙幅制限上、分析手順を(8)に示す論点に絞り込む。

[3]「周辺部」という考察点は Onodera (2011)、小野寺(2014)に詳しい。左右の周辺部に生起する表現が融合する現象に取り組む柴﨑(2015a)、Shibasaki (forthcoming)も関連現象である。

[4] ただし、本稿の内容は、Shibasaki (2014a,b)および柴﨑(2015b)で提示した英語を中心とした西欧語における「周辺部」の研究に根差しており、高橋・東泉による一連の研究とは異なる出発点から始まっている点を明記しておく。

(8) a.　文頭用法（副詞用法）：　…。事実／事実上／事実は（、）…
　　b.　文中用法（節接続用法）：…事実なるが／であるが／ですが（、）…
　　c.　文末用法（述語用法）：　…事実なり／である／です。

勿論、(4)のような異形態も多数存在するが（e.g. そして事実、事実上、事実なるが如し、事実なりとす、etc.）、網羅的に一覧を作成して各々を論じる紙幅の余裕はない。予備的研究として柴﨑(2015c)で示した通り、文頭用法の「事実」は比較的安定した頻度を示しており、文中用法と文末用法とで比較対象し易い点もある。本稿の新しい点は、柴﨑(2015c)で調査した文頭用法と文末用法の更なる精査に加え、文中用法という節接続用法の調査結果を加えることにより、談話機能の発達と統語変化を俯瞰することである。

表 2 文頭用法「事実（事實）／事実（事實）は／事実上（事實上）」（『太陽コーパス』）

	1895	1901	1909	1917	1925	合計
事実（、）	0	2	1	5	10 (1)	18
事実上（、）	0	0	1	2	2	5
事実は（、）	0	0	0	0	2	2
合計	0	2	2	7	14	25

表 3 文中用法「事実（事實）なるが／であるが／ですが」（『太陽コーパス』）

	1895	1901	1909	1917	1925	合計
事実なるが（、）	1	2	1*	3	0	7
事実であるが（、）	0	2	4	6	14	26
事実ですが（、）	0	0	0	0	2	2
合計	1	4	5	9	16	35

*「～事實なるが故に」

表 4 文末用法「事実（事實）なり／である／です」（『太陽コーパス』）

		1895	1901	1909	1917	1925	合計
事実なり	①　事実なり、（読点）[5]	24	18	3	3	0	48
	②　事実なり。（句点）	10	25	24	12	1	72
	小計	34	43	27	15	1	120
事実である	①　事実である、（読点）	0	8	5	2	1	16
	②　事実である。（句点）	0	0	30	71	72	173
	小計	0	8	35	73	73	189
事実です	①　事実です、（読点）	0	0	0	0	1	1
	②　事実です。（句点）	0	0	1	4	5	10
	小計	0	0	1	4	6	11

[5] 渡辺・村石・加部(1993)によれば、今日のような句読法が普及し始めたのは明治 20 年代から 30 年代頃とある。例えば、坪内逍遥の『小説神髄』（明治 18 年刊行）には句点および読点も使用されていなかったという（渡部 1995: 3-4 に詳しい）。表 4 には、句点の意味で読点を用いていると読めるものを提示した。

表 2〜4 に各用法の発達経緯を提示する。数値は素頻度を表している。尚、括弧内の数値は曖昧事例数を意味し、全体の素頻度にも含めてある。注意すべき点は、表 1 に示した近代語コーパスのうち、『太陽コーパス』を除く 3 コーパスは(8)に提示した事例を殆ど確認できないことである。そこで、表 2〜4 には『太陽コーパス』からの検索結果を提示し、その他のコーパスからの検索結果は必要に応じて記すこととする。文頭用法（副詞用法）が 20 世紀初頭頃から使用され始めたことは第 1 節で確認した（北原・他 2006）。その上で、(8)の用法がいつ頃から使用され始めたのかを更に精査し、談話機能の発達経緯を統語位置から再考することが本考察のポイントである。収録語彙数の異なるコーパスを用いて素頻度を標準化頻度に均して計量化することは、本考察の域を超えるものであることを記しておく。

４．分析

４.１　『太陽コーパス』の場合

表 2〜4 から以下の点を読み取ることができる。一点目は、1895 年（明治 28 年）時点では文頭用法（副詞用法）が確認できず、20 世紀に入って徐々に散見し始める点である。二点目は、「文末用法＞文中用法＞文頭用法」という機能拡張過程が読み取れる点である。つまり、1895 年（明治 28 年）時点で見ると、文末に生起する述部用法の使用例が相対的に高く、節接続機能としての文中用法は低頻度で確認できる程度である。[6] 三点目は、繋辞の変化が見て取れることである。明治大正期における大きな変化として、「なり型」から「である型」への過渡期を数値から読み取ることが可能である。更に、「です型」の文末用法が 1909 年（明治 42 年）から確認可能であるが、頻度面から黎明期と判断できそうである。「です型」の文中用法が 1925 年（大正 14 年）から確認できる点は、「文末用法＞文中用法」という流れを確認できることも見逃せない。

４.２　『現代日本語書き言葉均衡コーパス』（書籍ジャンル）の場合

『太陽コーパス』に基づく調査結果と分析が妥当であるかを、『現代日本語書き言葉均衡コーパス』(BCCWJ)の書籍ジャンルを検索することで確認してみたい。第 3 節と同じ手順による考察結果は表 5 にまとめてある。現代日本語の書籍ジャンルに限定してはあるが、文頭用法の「事実」と文末用法の「事実である、事実です」に特化した発達が確認できる。一方で、文中用法は全体的に伸び悩んでいる感も見て取れる。こうした分布上の違いは何を意味しているのであろうか。

一つの解釈として、繋辞と共に生起する述語用法（文末用法）の場合は、各時代で好まれる繋辞の違いはあれども、「事実＋繋辞」としての構文が時代を超えて固定化する方向に進んでいることを示唆していると判断できる。一方、20 世紀初頭頃より使用例が確認で

---

[6] 『明六雑誌コーパス』では、1875 年（明治 8 年）の段階で「…事實なり｜即（ち）…」という文末用法が 3 例確認できるが、「事實なるが」という文中用法は皆無である。尚、「｜」はコーパス作成段階で、作成者が「文の切れ目」と判断したことを示す記号である（ワークショップ当日の個人談話：田中牧郎先生、近藤明日子先生）。『国民之友コーパス』でも文末用法と判断できる読点付き「…事實なり、」が 12 例確認できる（1888 年[明治 21 年]）が、「事實なるが」という文中用法は皆無である。『近代女性雑誌コーパス』でも、文末用法と判断できる事例が 5 件確認できる一方（1894 年[明治 27 年]に 2 件、1895 年[明治 28 年]に 3 件）、文中用法は 1 件のみである（1895 年[明治 28 年]）。大局的に見て、文末用法が徐々に接続機能を発達させたことで文中用法が創発されたことが窺いしれる。

きる副詞用法（文頭用法）は後続する主情報を導入する談話機能を担っている。つまり、文頭は対話機能を担う表現が創発されやすい統語位置と考えられうる。[7]

表 5 文頭／文中／文末用法の分布と変遷（『BCCWJ』の書籍ジャンル）

		1970-74	1975-79	1980-84	1985-89	1990-94	1995-99	2000-05	合計
文頭用法	事実（、）	0	7	3	50	101	101	351	613
	事実上（、）	0	0	0	0	8	2	22	32
	事実は（、）	0	0	0	2	2	1	7	12
文中用法	事実なるが（、）	0	0	0	0	0	0	0	0
	事実であるが（、）	0	0	0	1	4	2	19	26
	事実ですが（、）	0	0	0	2	2	3	11	18
文末用法	事実なり。	0	1	0	0	0	0	0	1
	事実である。	0	7	8	58	87	96	309	565
	事実です。	0	0	4	9	26	43	158	240

４.３　機能拡張の方向と分布

第 4.1 節で指摘したように、機能拡張の方向は「文末用法＞文中用法＞文頭用法」で間違いなさそうである。この点は、高橋・東泉(2013, 2014)および東泉・高橋(2013)の研究成果を支持できる考察結果と言える。一方で、20 世紀初頭から始まる機能拡張は、各用法に均等に進行しているとは言えない。つまり、「文頭用法」と「文末用法」に特化した分布が表 5 から明らかである。節接続機能である「文中用法」は、「文末用法」から「文頭用法」へという機能拡張の橋渡しとして創発したが、20 世紀後半での使用頻度からは伸びが確認できない。この点は、高橋・東泉の一連の研究からは明確な見解が得られないことからも、今後の課題として取り組む価値のある事象である。

本節を締め括るにあたり、他言語における関連研究を一つだけ紹介しておく。節と節を接合する機能を担う接続副詞（linking adverbials; *then, however, though*, etc.）の最新の研究報告として Lenker (2015)がある。Lenker (2015)は接続副詞の発達を古英語から後期近代英語まで俯瞰している。仮に本稿と同じ「文頭・文中・文末」という基準で Lenker (2015)の報告を見た場合、先行情報を後行情報へ繋げる節接続機能を果たす文中用法の発達が、初期近代英語期（Lenker のデータでは 1570 年代）以降着実に増加している事実が明らかとなる。構造的に異なる英語と日本語を俄かに比較することはできない。しかし、英語では文中用法が近年発達しているのに対して、日本語では文頭用法と文末用法の発達が著しい点は注意すべきであろう。言語構造と文体的ヴァリエーションには相関性があると考えられるからである。

５．まとめ

本稿では、近代語コーパスと『現代日本語書き言葉均衡コーパス』（書籍ジャンル）を用いて、「事実」の「文頭用法・文中用法・文末用法」を考察した。19 世紀末あるいは 20 世

[7] 相互行為言語学（interactional linguistics）では、こうした機能を担う表現群を「投射構文」（projector constructions）と呼び慣わしている。関連研究として Shibasaki (2014a,b)、柴﨑(2015b)および柴﨑(近刊)などがある。

紀初頭頃より拡張の兆しが見え始め、「文頭用法＞文中用法＞文末用法」という方向で変化拡張が確認できた。一方で、20 世紀後半における分布状況は「文頭用法」と「文末用法」に特化してきており、節接続機能を果たす文中用法は相対的に衰退しつつあるようにも見えた。今後の展望としては、高橋・東泉(2013, 2014)および東泉・高橋(2013)などで報告されている漢語副詞なども含めた包括的な言語変化研究に取り組む点、および、Shibasaki (forthcoming)などで報告される他言語における関連事例の研究を進める点が挙げられる。

謝　辞

本研究は、日本学術振興会科学研究費基盤研究(C)「英語史に見る主要部と依存部の競合関係について」（研究代表：柴﨑礼士郎; 課題番号：25370569）による補助を一部得ています。また、本科研費プロジェクトは、英語史における同現象の詳細な研究成果を対照言語学的あるいは通言語学的研究へ応用させることにも主眼の置かれている点を付記しておく。尚、本稿の一部は『文法化：日本語研究と類型論的研究』（国立国語研究所 国際シンポジウム、2015 年 7 月 3－5 日）での発表とも関連している。発表当日、貴重な助言を下さった先生方へこの場を借りて感謝申し上げます（敬称略・五十音順：大野剛、大堀壽夫、古賀裕章、鈴木亮子、高橋圭子、Bernd Heine、東泉裕子、堀江薫）。

文　献

小野寺典子(2014)「談話標識の文法化をめぐる議論と「周辺部」という考え方」、金水敏･高田博之･椎名美智（編）『歴史語用論の世界』、3-27、ひつじ書房.

北原保雄、他（編）(2006)『日本国語大辞典』第二版、小学館.

近藤明日子(2013)「『近代女性雑誌コーパス』小説会話部分に現れる一・二人称代名詞の計量的分析」『第 4 回コーパス日本語学ワークショップ予稿集』、pp.135-144、国立国語研究所.

柴﨑礼士郎(2015a)「共有構文（*ἀπὸ κοινοῦ*）の創発と談話構造－現代アメリカ英語を中心に－」『ことばと人間』第 10 号、pp. 17-37.「言語と人間」研究会.

柴﨑礼士郎(2015b)「直近のアメリカ英語史における *the problem is (that)* の分析－構文の談話基盤性を中心に－」『語用論研究』第 16 号、pp.1-19. 日本語用論学会.

柴﨑礼士郎(2015c)「文副詞的機能を担う名詞の史的発達と文法化の方向性について－「事実」と「問題」を中心に－」『文法化：日本語研究と類型論的研究』国立国語研究所 国際シンポジウム、2015 年 7 月 3 日－5 日.

柴﨑礼士郎(近刊)「現代アメリカ英語の二重コピュラ構文」秋元実治、青木博史、前田満（編）『日英語の文法化と構文化』ひつじ書房.

新屋映子(2014)『日本語の名詞指向性の研究』ひつじ書房.

田中伊式(2012)「ニュース報道における「名詞＋です」表現について」『放送研究と課題』October 2012、pp.16-29.

角田太作(1996)「体言締め文」鈴木泰、角田太作（編）『日本語文法の諸問題：高橋太郎先生古希記念論文集』、pp.39-161、ひつじ書房.

角田太作(2012)「人魚構文と名詞の文法化」『国語研プロジェクトレビュー NINJAL Project Review』No. 7, pp.3-11.

高橋圭子、東泉裕子(2013)「漢語名詞の副詞用法～『現代日本語書き言葉均衡コーパス』『太陽コーパス』を用いて～」『第 4 回コーパス日本語学ワークショップ予稿集』、pp.195-202、

国立国語研究所.
高橋圭子、東泉裕子(2014)「近代語コーパスにみる「結果」の用法」『第 6 回コーパス日本語学ワークショップ予稿集』、pp.103-112、国立国語研究所.
鳴海伸一(2015)『日本語における漢語の変容の研究ー副詞化を中心として』ひつじ書房.
東泉裕子、高橋圭子(2013)「「結果、こういうことが言えそうです。」～コーパスにみる名詞の文副詞的用法～」『第 3 回コーパス日本語学ワークショップ予稿集』、pp.91-96、国立国語研究所.
渡辺富美雄、村石昭三、加部佐助(1993)『日本語解釈活用事典』ぎょうせい.
渡部善隆(1995)「横書き句読点の謎」九州大学情報基盤研究開発センター.
（http://yebisu.cc.kyushu-u.ac.jp/~watanabe/RESERCH/MANUSCRIPT/OTHERS/YOKO/ten.pdf）
Beeching, Kate and Ulrich Detges. (eds.) (2014) *Discourse Functions at the Left and Right Periphery*. Leiden: Brill.
Haselow, Alexander. (2012) “Discourse Organization and the Rise of Final *then* in the History of English.” In Irén Hegedűs and Alexandra Fodor (eds.), *English Historical Linguistics 2010: Selected Papers from the Sixteenth International Conference on English Historical Linguistics (ICEHL 16), Pécs, 23-27 August 2010*, pp.153–176. Amsterdam: John Benjamins.
Higashiizumi, Yuko. (2015) “Periphery of Utterance and (Inter)subjectification in Modern Japanese: A Case Study of Competing Causal Conjunctions and Connective Particles.” In Andrew D. M. Smith, Graeme Trousdale and Richard Waltereit (eds.), *New Directions in Grammaticalization Research*, pp.135-156. Amsterdam: John Benjamins.
Lenker, Ursula. (2015) “Knitting and Splitting Information: Medial Placement of Linking Adverbials in the History of English.” In Simone E. Pfenninger, Olga Timofeeva, Anne-Christine Gardner, Alpo Honkapohja, Marianne Hundt and Daniel Schreier (eds.), *Contact, Variation, and Change in the History of English*, pp.11-38. Amsterdam: John Benjamins.
Onodera, Noriko O. (2011) “The Grammaticalization of Discourse Markers.” In Heiko Narrog and Bernd Heine (eds.), *The Oxford Handbook of Grammaticalization*, pp.614-624. Oxford: Oxford University Press.
Roberts, Ian and Anna Roussou (2003) *Syntactic Change: A Minimalist Approach to Grammaticalization*. Cambridge: Cambridge University Press.
Shibasaki, Reijirou. (2011) “From Nominalizer to Stance Marker in the History of Okinawan.” In Marcel den Dikken and William McClure (eds.), *Japanese/Korean Linguistics 18*, pp.101-113. Stanford: CSLI Publications.
Shibasaki, Reijirou. (2014a) “On the Development of *the point is* and Related Issues in the History of American English.” *English Linguistics* 31 (1), pp.79–113.
Shibasaki, Reijirou. (2014b) “On the Grammaticalization of *the thing is* and Related Issues in the History of American English.” In Adams, M., Fulk, R. D. & Brinton, L. J. (eds.), *Studies in the History of the English Language: Evidence and Method in Histories of English*, pp.99–121. Berlin: De Gruyter Mouton.
Shibasaki, Reijirou. (forthcoming) “Sequentiality and the Emergence of New Constructions: *That's the bottom line is (that)* in American English.” In Hubert Cuyckens, Hendrik De Smet, Frauke D’hoedt, Liesbet Heyvaert, Charlotte Maekelberghe and Peter Petré (eds.), *ICEHL-18 Volume (provisional title)*. Amsterdam: John Benjamins.

口頭発表（3）

9月2日（水）　10:00～12:00

文体指標を特徴づける係り受け部分木の抽出

浅原 正幸 (国立国語研究所) *
加藤 祥 (国立国語研究所)

Extraction of Dependency Subtree Features for Writing Style Indexing

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Sachi Kato (National Institute for Japanese Language and Linguistics)

要旨

柏野 (2013), 柏野・奥村 (2012b) は文体を計量する指標として，専門度，客観度，硬度，くだけ度，語りかけ性の 5 種の分類指標を提案し，現代 日本語書き言葉均衡コーパス (BCCWJ) の図書館サブコーパス 10,551 サンプルに対して悉皆的に付与を行った。浅原ほか (2014) では，この分類指標に対して語彙素を特徴量とした制約付き主成分分析を行い，各指標と特徴的な語彙分 布の対応を品詞ごとに定量的に評価した。浅原ほか (2015b) では語彙素を語彙の系列 (n-gram, p-mer) に拡張し対応分析を行い，既存の定性的な分析結果との比較を行った。今回は，係り受け解析結果の部分木を特徴量とした決定株とブースティングに基づく分類器を用い，文体指標に対して代表的な係り受け部分木の評価を行った。

1. はじめに

柏野 (2013), 柏野・奥村 (2012b) は文体を計量する指標として，専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標を提案し，『現代日本語書き言葉均衡コーパス』(BCCWJ) の図書館サブコーパス (LB サンプル)10,551 サンプルに対して悉皆的に付与を行った。このデータに対して，硬度・語りかけ性度を中心に，定量的・定性的な分析が進められてきた (柏野ほか (2012a), 保田ほか (2012b,a,c, 2013d,a,c,b), 加藤ほか (2014))。

また，浅原ほか (2014) では，この分類指標に対して語彙素を特徴量とした制約付き主成分分析を行い，各指標と特徴的な語彙分布との対応を品詞ごとに定量的に評価した。さらに，浅原ほか (2015b) ではこの手法を拡張し，各指標と語彙系列 (語彙素の連続・非連続列) との制約付き主成分分析を行った。本予稿集の別の発表 (浅原・森田 (2015a)) では，主成分分析に必要な「文書-単語行列」・「文書-部分構造行列」をプログラミングすることなしに生成する方法について紹介している。

本稿では，文体を評価する特徴量として係り受け解析結果に基づく単語単位係り受け部分木を用いた識別学習を行い，文体分析を行う。識別学習器として，与えられた木構造の部分木を特徴量とした決定株 (Decision Stumps) を弱学習器とした Boosting アルゴリズムによる機械学

* masayu-a@ninjal.ac.jp

習器 bact (Kudo and Matsumoto (2004)) を用いる。BCCWJ LB サンプル (10,511 サンプル) に対する識別分析のほか約 14 億文からなる Web コーパス (Asahara et al. (2014)) に適用し，識別性能を確認した。

2. 分析手法

2.1 文体指標

柏野 (2013) は文体指標として以下の 5 種類を規定した：

- 【専門度】: 1 専門家向き，2 やや専門的な一般向き，3 一般向き，4 中高生向き，5 小学生・幼児向きの 5 段階指標
- 【客観度】: 1 とても客観的，2 どちらかといえば客観的，3 どちらかといえば主観的，4 とても主観的の 4 段階指標
- 【硬度】: 1 とても硬い，2 どちらかといえば硬い，3 どちらかといえば軟らかい，4 とても軟らかいの 4 段階指標
- 【くだけ度】: 1 とてもくだけている，2 どちらかといえばくだけている，3 くだけていないの 3 段階指標
- 【語りかけ性度】: 1 とても語りかけ性がある，2 どちらかといえば語りかけ性がある，3 特に語りかけ性はないの 3 段階指標

対象は BCCWJ に収録されている図書館サブコーパス 10,551 サンプル（書籍サンプル）とし，20〜50 代女性作業者延べ 9 名に可変長サンプルを呈示して文体指標付与を行った。作業において，インタビューなどのテキスト構造が文体付与に適さないものや外国語や数式などが多いサンプルなど内容や表現が文体付与に適さないものなど 1,664 サンプルを，文体指標付与対象から除外している。本研究ではこれらのサンプルもラベルなしデータとして利用した。

2.2 識別学習器 bact

識別学習器として，ラベル付き順序木の部分木を特徴量とした決定株 (Decision Stumps) を弱学習器とした Boosting アルゴリズムによる機械学習器 bact (Kudo and Matsumoto (2004)) を用いる[(1)]。

決定株は深さ 1 の決定木と同一で単一の特徴量に基づく分類器である。bact では，ラベル付き順序木の部分木を特徴量として考慮した決定株を逐次生成し，これを弱学習器とした Boosting (重み付き多数決) を行う。最右拡張などに基づく部分木構造マイニングアルゴリズムを適応することで，効率的に最適な弱学習器（に対応するラベル付き部分木）を枚挙する実装になっている。

Support Vector Machines などの Large Margin Classifier が事例スパースの解（Support Vector となる事例）を導出するのに対し，Boosting が特徴量スパース（弱学習器に対応する特徴量＝ラベル付き部分木）を導出する。このことは人文系研究者にとって，「どのような特徴量を用いて分析しているのか」を陽に示すだけでなく，SVM と比べて解析速度が高速であるという

(1) http://chasen.org/~taku/software/bact/

利点がある。

2.3　係り受け解析結果に基づくラベル付き部分木の与え方

本研究ではラベル付き部分木として文節係り受け解析器 CaboCha-0.69 の UniDic 主辞規則[2]により生成したものを用いる。文節係り受け解析結果を単語単位係り受け解析に変換する手法として，(1) 文節内最右要素を主辞として残りの要素を鎖状にかける手法と，(2)CaboCha-0.69 の UniDic 主辞規則に基づく内容語主辞[3]を主辞として残り要素を左右から鎖状にかける手法があるが，事前の実験の結果，(1) より (2) の方が良い性能が得られたために (2) を用いる。

他の手法として，Mori et al. (2014) のような単語係り受け木 (上記 (1) に近い木) や Universal Dependencies (UD)(McDonald et al. (2013), Universal-Dependencies-contributors (2015)) のような単語係り受け木 (金山ほか (2015))(上記 (2) に近い木) が考えられる。さらに田中・永田 (2015) は Stanford typed dependency (SD)(Marneffe and Manning (2008)) に基づくラベル付き単語係り受け木において，3 種類の主辞決定規則[4]を定義している。

3.　BCCWJ によるモデリング実験

3.1　手法

BCCWJ LB サンプル (10,511 サンプル) を文単位 (1,651,084 文) に分割し，CaboCha-0.69 (UniDic 主辞規則) により文節係り受け解析を行う。文節係り受け解析結果は 2.3 節に述べた手法で単語係り受け解析結果に変換する。以下に変換事例を示す：

```
(˜EOS(居る (。(た (て (支える))(顎 (を))(両手 (で))(伸ばす (，(床 (に))(体 (を (しなやか (だ (スリム (だ))))))))(オクタビアン (は (　(˜BOS))))))))
(˜EOS(感ずる (。(其れ (を))(横顔 (に (˜BOS))))))
(˜EOS(分かる (。(ない (私 (は (に)))(行く (，(か (の (た (何処 (へ))(人 (が (メンネンカルト (˜BOS))))))))))))
```

文体指標は n 段階評価によりレーティングラベルである。一方，今回用いる識別学習器は二値分類器である。二値分類器をレーティングラベルに対して適用する手法として，順序ラベルのようにレーティングの上位下位に基づく手法[5]が考えられるが，指標によってはラベルが規定されていないものもあり，単純な one-vs-others 法を用いることとした。評価において，全ての二値分類器が負の値を返した場合には「ラベルなし」として認定することとした。

---

(2) `./configure --with-posset=unidic`

(3) 以下のような CaboCha の出力において，1 行目の `* 0 1D 2/4 0.000000` の `2/4` が主辞を表す。/左の `2` が内容語主辞で，`4` が機能語主辞：

```
* 0 1D 2/4 0.000000
”　　補助記号,括弧開,*,*,*,*,,” ,*,,*,*,*,*,*,*,*,*
警察　　名詞,普通名詞,一般,*,*,*,ケイサツ,警察,*,ケーサツ,*,*,*,*,*,*,*
メディア　　名詞,普通名詞,一般,*,*,*,メディア,メディア,*,メディア,*,*,*,*,*,*,*
”　　補助記号,括弧閉,*,*,*,*,,” ,*,,*,*,*,*,*,*,*,*
が　　助詞,格助詞,*,*,*,*,ガ,が,*,ガ,*,*,*,*,*,*,*
```

(4) 主辞後置型 1：内容語と格要素となる後置詞句の間で先に構造を作る句構造を作り，格構造で最右要素を主辞とする；
主辞後置型 2：接続助詞を除いた述部の文節相当の単位で先に構造を作る句構造を想定し，格構造で最右要素を主辞とする；
述部内容語主辞型：述部の文節相当の単位で先に構造を作る句構造を想定し，述部において最左要素を主辞とする。

(5) {1,2,3,4} というレーティングラベルに対して，{1}vs.{2,3,4}・{1,2}vs.{3,4}・{1,2,3}vs.{4} の 3 種類の二値分類器を構成し，これらの多数決により分類する手法。

本実験では bact の iteration 回数を 10,000 回として，BCCWJ LB サンプル全てでモデルを学習し，bact によって得られた連続部分木を分析する。

3.2　得られた規則

得られた規則数を本稿末尾の表 2 の「規則数」の列に示す。およそ，各ラベルごとに数百 (min. 157, max. 411)，文体指標ごとに千前後 (min. 558, max. 1683) 程度の特徴量に基づく規則が得られている。

得られた規則の一例として，表 1 に客観度に対して得られた特徴量（上位 10 位・下位 10 位まで）を示す。客観的なものの例として，引用表現（“言う 。”, “」。” など)，法律用語（“法”, “条” など）などが上位に来る傾向がある。一方，主観的なものの例として，一人称表現（“私”, “僕”, “俺” など）や感嘆符・疑問符などが上位に来る傾向がある。

客観度	1 とても客観的	2 どちらかといえば客観的	3 どちらかといえば主観的	4 とても主観的
デフォルト	−0.0024571855	−0.0006028928	−0.001862472	−0.0022119238
上位1位	0.0012419398 ，	0.0027332267 ユダヤ	0.0032610654 オウ	0.0019894564 ちゃう
上位2位	0.0010513154 権	0.0027150333 ［ ~BOS	0.0020614907 ~EOS 無い 。 か	0.0018879718 僕
上位3位	0.0010498933 」 。	0.0018783132 食品	0.0006848079 ，	0.0014030064 有る 。 だ
上位4位	0.0010301565 有る 。 だ	0.0015717303 寺	0.0006361754 。 だ	0.0013411624 私
上位5位	0.0010125919 ） （ ~BOS	0.0010845271 居る 。	0.0005789535 子供	0.0012610379 ！
上位6位	0.0008768302 図	0.0006354240 ~EOS 有る 。 だ	0.0005688861 作品	0.0012515885 ？
上位7位	0.0008472139 細胞	0.0005610703 ~EOS 有る 。 た だ	0.0005660776 私	0.0008426375 。
上位8位	0.0007825627 ）	0.0004399209 。 ます	0.0005103392 君	0.0007344837 よ
上位9位	0.0006846557 法	0.0003861222 言う 。	0.0005022981 ~EOS 有る 。 だ	0.0006952075 俺
上位10位	0.0005550320 条	0.0002879612 は	0.0004832747 （ （	0.0005798904 ね
下位10位	−0.0006896641 子供	−0.0005263574 思う	−0.0005475509 場合	−0.0006841687 因る
下位9位	−0.0006971665 自分	−0.0005590740 。 た 私	−0.0005766390 銀行	−0.0006940692 ）
下位8位	−0.0007551097 ね	−0.0005657794 ね	−0.0006162622 ．	−0.0007068332 化
下位7位	−0.0008653166 御	−0.0005927404 な	−0.0006934120 発生	−0.0007170490 千
下位6位	−0.0008783045 。 だ	−0.0008522566 ！	−0.0007463329 条	−0.0007193994 企業
下位5位	−0.0010509838 ！	−0.0009144568 ~EOS 居る 。	−0.0007609778 図 ~BOS	−0.0007560440 図
下位4位	−0.0010796309 ？	−0.0009766754 よ	−0.0008078806 ）	−0.0008436630 的
下位3位	−0.0012399752 さん	−0.0015972556 私	−0.0010902414 有る 。 だ	−0.0009392407 ．
下位2位	−0.0012867395 僕	−0.0020705533 」 ？	−0.0012004897 無い 。 か	−0.0017862343 ~EOS 有る 。 だ
下位1位	−0.0015122806 私	−0.0021344768 僕	−0.0020124672 ，	−0.0028774426 ，

表 1　分類指標「客観度」に対する規則（連続部分木パターン）

4.　交差検定による評価

次に，構成した識別学習器の性能を評価するために，BCCWJ LB データ (10,551 サンプル) 上での 5 分割交差検定を行う。ファイル名がサンプルの属性の情報を含んでいるために，乱数を発行することによりサンプル単位で LB データを 5 分割した。識別学習は文単位で行い，文単位評価 (4.1 節)・サンプル単位評価 (4.2 節)・サンプル全体における文の位置に対する正答率 (4.3 節) の 3 種類の評価を実施した。

4.1　文単位評価

文単位の評価結果を本稿末尾の表 2 の「文単位評価」の列に示す。OK は左に示すラベルをシステムが正答した件数，SYS は左に示すラベルをシステムが出力した件数（右に全体における割合を % で表示），GOLD は左に示すラベルを人手により付与された件数（右に全体における割合を % で表示），PREC が精度 (precision) で OK/SYS，REC が再現率 (recall) で OK/GOLD を意味する。

全体の傾向として，GOLD における分布の大きいものが，SYS において大量に生成されるように尖度が高くなる傾向にある。言い換えると，学習データにおいて多数のものの再現率が

高くなる傾向にあり，学習データにおいて少数のものの精度が高くなる傾向にある。

さらに，識別学習器の出力は必ずしも元のサンプルの分布を保存するようなものではなく，識別学習器の分布を用いて，コーパスの文体分布の計量的な調査を行うことは不適切であることがわかる。

一方，低頻度のラベルについて識別結果の精度の高いことは，稀な文体ラベルの事例に似た事例を大量のコーパスから抽出するのには適していると考える。

4.2　サンプル単位評価

サンプルを構成する文単位の評価の重みなし多数決を用いて，サンプル単位の評価を行った。サンプル単位の評価結果を本稿末尾の表 2 の「サンプル単位評価」の列に示す。各列の意味は「文単位評価」と同じである。

重みなし多数決を行う結果，より一層 GOLD における分布の大きいものが SYS において大量に生成される傾向が強くなり，分布の小さいものの判定の出現する確率が下がる傾向にある。例えば，専門度においては 98.3% が「3 一般向き」と出力される。客観度においては 94.3% が「ラベルなし」と出力される。

4.3　サンプル全体における文の位置に対する正答率

評価は 10 文以上のサンプルのみについて行った。表中 (n)-(n-1)% はサンプル全体における評価対象文の位置を表す。

どの指標も 80-90%, 90-100% で正答率の下がる傾向が見られた。サンプリングにおいて 90-100% に位置するデータが少なくなる傾向にある [(6)]にしても，有意に差がある。

これはサンプルの末尾にプロフィールやまとめなどの本文と異なる文体が出現しているためではないかと考える。

5.　超大規模コーパスへの適用

最後に現在国語研コーパス開発センターで開発している超大規模コーパス (Asahara et al. (2014))（2014 年 10-12 月収集分）全文（1,463,142,939 文=14.6 億文，23,836,100,595 語=238 億語，EOS・句点を含まず）に 3 節で構成した識別学習器を適用した結果を表 2 右の「超大規模コーパス分布」に示す。

基本的に SYS （システムが出力した件数）とその割合のみを示す。4 節に示した通り，学習元データにおける分布の大きいラベルがより多く出力される傾向にある。しかしながら，分布の小さいラベルでも出力されることがあり，これらの出力結果は精度の高いものであると考える。

[(6)] 例えば，サンプル中 19 文の場合，90-100% の文数が 1 に対し，他の箇所は文数 2 となる。

6. おわりに

6.1 本研究のまとめ

本研究では，BCCWJ LB サンプルに付与された文体指標を単語係り受け部分木を特徴量とした識別学習器によりモデル化し，分析を行った。

特徴量の抽出 (3 節) においては，単語・単語列の特徴的な表現を抽出しているが，単語係り受け木を使う有効性までは確認することができなかった。交差検定による性能評価 (4 節) においては，チャンスレベルと比較するとよい性能が得られたが，学習元データの分布をそのままシステム出力することがないことが確認された。

今回学習した識別学習器を Web から収集した 14 億文規模のテキストコーパスに適用した (5 節)。識別学習器の出力の分布を用いて文体指標の分布を分析することが困難である一方，少ないラベルの精度が高いことから，大量のテキストから似た文体の事例を高精度で収集することが可能であると考えられる。

6.2 今後の検討課題

今後検討すべき課題は以下のとおりである：

- 特徴量スパース vs. 事例スパース
  2.2 節に述べた通り，今回用いた識別学習器は特徴量を抽出することによる二値分類器である。計算量は大きくなるが Tree Kernel に基づく Large Margin Classifier などを用いることで事例スパースな解を与えて，境界事例を分析することを考えたい。
- bact に与える単語係り受け木
  日本語においては文節係り受け木に基づく自然言語処理の研究が進んでいる一方，文節係り受け木に基づいてどのような単語係り受け木を与えるかを検討する必要がある。2.3 節に示した通り，ここ数年日本語の単語係り受け木の規定について様々な提案がなされており，文の識別問題として定式化した文体指標分析に最適な単語係り受け木を調査する必要がある。
- 二値分類器のレーティングラベル適応
  本研究では浅原ほか (2014, 2015b) の対応分析の結果から，レーティングラベルが必ずしも線形上に分布していないとし，one-vs-others 法を用いた。その結果，分布の偏りを増長するような識別学習器を構成することになった。どのようにラベル空間を構造学習器に反映させるかを検討する必要がある。

謝辞

本研究の一部は国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2014).「文体指標と語彙の対応分析」　第 6 回コーパス日本語学ワークショップ, pp. 11–20.

浅原正幸・森田敏生 (2015a).「コーパスコンコーダンサ『ChaKi.NET』の「文書-部分構造行列」出力機能」　第 8 回コーパス日本語学ワークショップ.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2015b).「文体指標と語彙系列の対応分析」　第 7 回コーパス日本語学ワークショップ, pp. 7–16.

Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi (2014). “Archiving and analysing techniques of the ultra-large-scale web-based corpus project of ninjal, japan.” *Alexandria*, 25:1-2, pp. 129–148.

de Marneffe, Marie-Catherine, and Christopher D. Manning (2008). “The stanford typed dependencies representation.” *Prof. of COLING-2008: Workshop on Cross-framework and Cross-domain Parser Evaluation*.

金山博・宮尾祐介・田中貴秋・森信介・浅原正幸・植松すみれ (2015).「日本語 Universal Dependencies の試案」　言語処理学会第 21 回年次大会発表論文集, pp. 505–508.

柏野和佳子・立花幸子・保田祥・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織 (2012a).「テキストの硬さと軟らかさの考察–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　第 1 回コーパス日本語学ワークショップ, pp. 131–138.

柏野和佳子・奥村学 (2012b).「書籍テキストへの分類指標人手付与の試み–『現代日本語書き言葉均衡コーパス』の収録書籍を対象に–」　言語処理学会第 18 回年次大会, pp. 1260–1263.

柏野和佳子 (2013).「書籍サンプルの文体を分類する」　国語研プロジェクトレビュー, 4:1, pp. 43–53.

加藤祥・柏野和佳子・立花幸子・丸山岳彦 (2014).「語りかける書きことばの表現」　国立国語研究所論集, 8, pp. 85–108.

Kudo, Taku, and Yuji Matsumoto (2004). “A boosting algorithm for classification of semi-structured text.” *Proc. of EMNLP-2004*, pp. 301–308.

McDonald, Ryan T., Joakim Nivre, Yvonnne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, and Slav Petrov Hao Zhang Oscar Täckström Kuzman Täckström, Keith B. Hall (2013). “Universal dependency annotation for multilingual parsing.” *Prof. ACL-2013(2) 92-97*.

Mori, Shinsuke, Hideki Ogura, and Tetsuro Sasada (2014). “A japanese word dependency corpus.” *Proc. of LREC-2014*, pp. 1631–1636.

田中貴秋・永田昌明 (2015).「日本語のラベル付き依存構造解析の検討」　言語処理学会第 21 回年次大会発表論文集, pp. 1044–1047.

Universal-Dependencies-contributors (2015). *Universal Dependencies*. `https://universaldependenceis.github.io/docs/`.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012a).「「語りかけ性」を有すると判断される書きことばの表現」　第 2 回コーパス日本語学ワークショップ, pp. 43–50.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2012b).「「語り性」を有する書きことばの典型例の分析」　第 1 回コーパス日本語学ワークショップ, pp. 139–146.

保田祥・柏野和佳子・立花幸子 (2012c).「総体として印象を与える表現：「語りかけ性」を有すると判断する根拠」　人工知能学会第 41 回ことば工学研究会.

保田祥・立花幸子・柏野和佳子・丸山岳彦 (2013a).「「ベテランは足を保護する」が語りかけるとき」　第 4 回コーパス日本語学ワークショップ, pp. 345–354.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013b).「アノテーターコメントを用いた「語りかけ性」分析の試み–頻度情報から捉え難いテキスト性質の解明に向けて–」　言語処理学会第 19 回年次大会発表論文集, pp. 358–361.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013c).「語りかけると判断される文体–大規模コーパスを用いた特徴的表現の分析–」　日本文体論学会第 104 回大会.

保田祥・柏野和佳子・立花幸子・丸山岳彦 (2013d).「書きことばにおける「語りかけ」は何のために用いられるのか」　第 3 回コーパス日本語学ワークショップ, pp. 143–152.

表 2　文体分析結果

専門度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		3,193	17,595	1.1%	284,947	17.3%	0.181	0.011	10	40	0.4%	1,730	16.4%	0.250	0.005	4,399,250	0.3%
1 専門家向き	391	28	367	0.0%	15,992	1.0%	0.076	0.001	0	0	0.0%	141	1.3%	0.000	0.000	13,833	0.0%
2 やや専門的な一般向き	411	7,961	21,998	1.3%	106,685	6.5%	0.361	0.074	81	138	1.3%	929	8.8%	0.586	0.087	2,042,510	0.1%
3 一般向き	238	1,076,748	1,602,737	97.1%	1,093,905	66.3%	0.671	0.984	7,036	10372	98.3%	7,065	67.0%	0.678	0.995	1,456,613,821	99.6%
4 中高生向き	395	280	621	0.0%	91,866	5.6%	0.450	0.003	0	0	0.0%	384	3.6%	0.000	0.000	9,238	0.0%
5 小学生・幼児向き	248	4,699	7,766	0.5%	57,689	3.5%	0.605	0.081	1	1	0.0%	302	2.9%	1.000	0.003	64,287	0.0%
計	1,683	1,092,909	1,651,084		1,651,084				7,128	10,551		10,551				1,463,142,939	

硬度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		53,846	276,337	16.7%	284,947	17.3%	0.194	0.188	282	1,548	14.7%	1,730	16.4%	0.182	0.163	526,050,369	36.0%
1 とても硬い	389	1,775	4,741	0.3%	68,194	4.1%	0.374	0.026	10	13	0.1%	619	5.9%	0.769	0.016	464,034	0.0%
2 どちらかといえば硬い	252	78,699	175,765	10.6%	420,734	25.5%	0.447	0.187	622	1,112	10.5%	3,065	29.0%	0.559	0.202	47,655,505	3.3%
3 どちらかといえば軟らかい	182	604,739	1,192,266	72.2%	753,383	45.6%	0.507	0.802	4,087	7,877	74.7%	4,440	42.1%	0.518	0.920	888,566,194	60.7%
4 とても軟らかい	214	819	1,975	0.1%	123,826	7.5%	0.414	0.006	0	1	0.0%	697	6.6%	0.000	0.000	406,837	0.0%
計	1,037	739,878	1,651,084		1,651,084				5,001	10,551		10,551				1,463,142,939	

くだけ度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		43,721	251,880	15.3%	284,947	17.3%	0.173	0.153	55	357	3.4%	1,730	16.4%	0.154	0.031	179,609,675	12.3%
1 とてもくだけている	157	328	1,052	0.1%	89,419	5.4%	0.311	0.003	0	0	0.0%	473	4.5%	0.000	0.000	347,197	0.0%
2 どちらかといえばくだけている	163	158,122	353,419	21.4%	511,680	31.0%	0.447	0.309	813	1,526	14.5%	2,696	25.6%	0.532	0.301	18,614,436	1.3%
3 くだけていない	287	550,230	1,044,733	63.3%	765,038	46.3%	0.526	0.719	5,188	8,668	82.2%	5,652	53.6%	0.598	0.917	1,264,571,631	86.4%
計	607	752,401	1,651,084		1,651,084				6,056	10,551		10,551				1,463,142,939	

語りかけ性度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		31,947	125,616	7.6%	284,947	17.3%	0.254	0.112	288	1,216	11.5%	1,730	16.4%	0.187	0.131	46,693,906	3.2%
1 とても語りかけ性がある	206	13,182	45,451	2.8%	112,441	6.8%	0.290	0.117	9	14	0.1%	833	7.9%	0.642	0.010	1,297,863	0.1%
2 どちらかといえば語りかけ性がある	179	57	416	0.0%	197,646	12.0%	0.137	0.000	0	1	0.0%	1,379	13.1%	0.000	0.000	805	0.0%
3 特に語りかけ性はない	173	1,003,963	1,479,601	89.6%	1,056,050	64.0%	0.678	0.950	6,409	9,320	88.3%	6,609	62.6%	0.687	0.969	1,415,150,365	96.7%
計	558	1,049,149	1,651,084		1,651,084				6,706	10,551		10,551				1,463,142,939	

客観度

ラベル	規則数	文単位評価 OK	SYS		GOLD		PREC	REC	サンプル単位評価 OK	SYS		GOLD		PREC	REC	超大規模コーパス分布 SYS	
ラベルなし		652,752	1,084,518	65.7%	937,252	56.8%	0.592	0.685	4,609	9,948	94.3%	4,650	44.1%	0.463	0.991	1,302,522,743	89.0%
1 とても客観的	340	7,039	18,627	1.1%	102,858	6.2%	0.377	0.068	73	112	1.1%	950	9.0%	0.651	0.076	3,761,009	0.3%
2 どちらかといえば客観的	198	99,538	374,577	22.7%	299,282	18.1%	0.265	0.332	241	487	4.6%	2,523	23.9%	0.494	0.095	142,993,449	9.8%
3 どちらかといえば主観的	276	8,332	60,169	3.6%	200,814	12.2%	0.138	0.041	1	3	0.0%	1,566	14.8%	0.333	0.000	516,827	0.0%
4 とても主観的	278	11,416	113,193	6.9%	110,878	6.7%	0.100	0.102	1	1	0.0%	862	8.2%	1.000	0.000	13,348,911	0.9%
計	1092	779,077	1,651,084		1,651,084				4,925	10,551		10,551				1,463,142,939	

全体における位置に対する正答率

(10文以上)	0–10%	10–20%	20–30%	30–40%	40–50%	50–60%	60–70%	70–80%	80–90%	90–100%	90–100%
専門度	0.675	0.677	0.675	0.675	0.675	0.675	0.675	0.673	0.627	0.401	0.401
硬度	0.463	0.467	0.469	0.467	0.470	0.468	0.467	0.465	0.442	0.331	0.331
くだけ度	0.548	0.550	0.550	0.551	0.549	0.548	0.544	0.546	0.510	0.365	0.365
語りかけ性度	0.640	0.641	0.641	0.640	0.640	0.639	0.639	0.638	0.607	0.450	0.450
客観度	0.493	0.495	0.494	0.492	0.491	0.495	0.491	0.493	0.477	0.358	0.358

サンプル単位5分割交差検定(ランダム順分割)

文単位評価:　サンプルを文単位に分割して、各文がどのラベルに属するかを推定。どれにも分類されないばあいには「ラベルなし」をつける。
サンプル単位評価:　文単位で評価したものの多数決。
超大規模コーパス分布:　超大規模コーパスを解析したもののシステム出力の分布。

OK は正当した数、SYSはシステムが出力した数で右に総数に対する割合(%)を付与、GOLDは人手で付与した数で右に総数に対する割合(%)を付与。
PREC は精度 (OK/SYS)、REC は再現率 (OK/GOLD)。

全体における位置に対する正答率:　サンプルが10文以上のものに対し、全体における位置に対する正答率。

助詞の使用実態　－BCCWJ・CSJにみる分布－

丸山 直子（東京女子大学現代教養学部）†

Usage of Postpositional Particles in BCCWJ and CSJ

Naoko Maruyama (Tokyo Woman's Christian University)

要旨

現代日本語の助詞について、現代日本語書き言葉均衡コーパス（BCCWJ）及び話し言葉コーパス（CSJ）における用いられ方を観察し、書き言葉と話し言葉の違い、及びそれぞれのサブコーパス（レジスター）ごとの違いを明らかにした。BCCWJはコアのみ（新聞、雑誌、書籍、白書、知恵袋、ブログ）を調査対象とし、CSJは、同一話者による独話（学会講演）と対話（自由会話）4件ずつを対象として調査を行った。コレスポンデンス分析も行った。

BCCWJもCSJも、全語数の約30%が助詞であり、助詞の中では格助詞が最も多い。BCCWJにおいては、白書と知恵袋・ブログは、助詞の使用法に関して、様々な点で対極にある。白書はかなり特殊で、格助詞相当の複合辞が多く、短単位と長単位で大きく分布が異なる。新聞は多少白書に似た性質を持つ。知恵袋とブログは、終助詞が多い等の話し言葉的な性質を帯びているが、相互に異なる性質も持つ。CSJは、講演の方が格助詞が多く、対話には副助詞・終助詞が多い。融合・縮約の多さも話し言葉特有の現象として指摘できる。

１．はじめに

現代日本語の助詞について、現代日本語書き言葉均衡コーパス（BCCWJ）及び話し言葉コーパス（CSJ）における用いられ方を観察することで、書き言葉と話し言葉の違い、及びそれぞれのサブコーパス（レジスター）ごとの違いを明らかにする。BCCWJはコアのみ（新聞、雑誌、書籍、白書、知恵袋、ブログ）を調査対象とし、CSJは、同一話者による独話（学会講演）と対話（自由会話）4件ずつを対象とする。

２．調査対象

BCCWJ、CSJの、調査対象としたものを表1、表2に記す。BCCWJは、コアすべてで、短単位で約100万語、長単位で80万語である。CSJは、4名の学会講演・自由会話1件ずつで、計8件である。こちらは短単位で計3万語という小さなサンプルである。

表１　BCCWJの調査対象

	短単位総数	長単位総数
出版・新聞コア	308,504	224,140
出版・雑誌コア	202,268	159,883
出版・書籍コア	204,050	169,730
特定目的・白書コア	197,011	129,646
特定目的・知恵袋コア	93,932	78,770
特定目的・ブログコア	92,746	75,242
計	1,098,511	837,411

† maruyama@lab.twcu.ac.jp

表 2　CSJ の調査対象

講演者 ID	性別	生年代	基にした学会講演 ID	短単位数	長単位数	自由会話 ID	短単位数	長単位数
1185	女	70to74	A11F0703	5,634	4,697	D03F0034	3,021	2,699
19	女	65to69	A05F0043	3,512	2,655	D03F0058	2,330	2,039
471	男	75to79	A11M0369	3,119	2,246	D03M0004	2,491	2,178
373	男	45to49	A11M0469	6,763	5,379	D03M0038	3,638	3,278
計				19,028	14,977		11,480	10,194

３．助詞の分類

本稿では、BCCWJ は中納言オンライン版の短単位・長単位分割及び品詞分類に基づき、CSJ は、DVD に収められている、短単位・長単位データに基づき集計した。BCCWJ・CSJ とも、格助詞・副助詞・係助詞・接続助詞・終助詞・準体助詞の六分類である。

４．調査で得られた助詞

以下に、それぞれのコーパスに含まれていた助詞の一覧を表にして示す。

表 3　コーパス中の助詞一覧

	BCCWJ	CSJ
格助詞（短単位）	ガ、ヲ、ニ、ト、デ、ヘ、ヨリ、カラ、ノ、トテ、ニテ、サ	ガ、ヲ、ニ、ト、デ、ヘ、ヨリ、カラ、ノ、デハ（じゃ）
格助詞（長単位）	ヲ通ジテ、ヲハジメ、ヲメグル、ヲモッテ、ニアタッテ、ニアタリ、ニイタルマデ、ニオイテ、ニオケル、ニ関シテ、ニ関スル、ニ際シ、ニ際シテ、ニシテ、ニ対シ、ニ対シテ、ニ対スル、ニツイテ、ニツキ、ニトッテ、ニヨッテ、ニヨリ、ニヨル、ニヨルト、ニヨレバ、ニワタッテ、ニワタリ、ニワタル、際ニ、トイウ、トイッタ、トシテ、カラシテ、カラスルト、カラスレバ、タメノ	ヲモトニシタ、ヲモトニシテ、ニオイテ、ニオケル、ニ関シテ、ニ関シマシテ、ニ関スル、ニ比ベテ、ニ従ッテ、ニ対シテ、ニ対シマシテ、ニ対スル、ニツイテ、ニツキマシテ、ニトッテ、ニ伴ウ、ニ基ヅイタ、ニ基ヅイテ、ニ基ヅク、ニヨッテ、ニヨル、ニヨリマス、ニヨリマスト、トイウ、トイッタ、トシテ、トイタシマシテ
副助詞（短単位）	ダケ、ノミ、バカリ、キリ、マデ、クライ、ナド、ナンカ、ナンテ、カ、ヤ、ヤラ、ホド、シカ、サエ、スラ、ッテ、タリ、シ、カシラ、ガニ、シモ、ズツ、ゾ、ダニ、タラ、ツ、デン、ドコロ、ナリ、ナンゾ、ナント	ダケ、ノミ、マデ、クライ、ナド、ナンカ、カ、ヤ、ホド、シカ、スラ、ッテ、タリ、シモ、ズツ、タッテ、モ、コソ
副助詞（長単位）	ダケデナク、ノミナラズ、ツウ、ニ限ラズ	トカ
係助詞（短単位）	ハ、モ、コソ、ゾ、バ、ヤ	ハ
係助詞（長単位）	トイエドモ、トイッテモ、トキタラ、ニイタッテハ	なし
接続助詞（短単位）	シ、テ、ト、バ、カラ、ガ、ケレド、トモ、ニ、タッテ、ツツ、ナガラ、ケン、サカイ、ド、トテ、ナリ	シ、テ、ト、バ、カラ、ガ、ケレド、ツツ、ナガラ、テハ（ちゃ）
接続助詞（長単位）	カラトイッテ、カラニハ、ウエデ、ウエニ、カト思ウト、タトコロ、タトコロデ、タメニ、トシタラ、トシテモ、トスレバ、トテ、ト同時ニ、トトモニ、トハイエ、ニ関ワラズ、ニシタガイ、ニシタガッテ、ニシテハ、ニシテモ、ニシロ、ニセヨ、ニツレ、ニツレテ、ニモカカワラズ、モノノ、ヤイナヤ、ワリニ	テハ、テモ、ノデ、ノニ

終助詞(短単位)	カ、サ、ナ、ネ、ヨ、ゼ、ゾ、ワ、ノ、イ、カシラ、ヤ、ケ、モノ、ジャン、エ、カナ、クサ、チョ、デ、テン、ド、ネン、ノウ、バイ、ベイ、モガ	カ、ナ、ネ、ヨ、ゾ、ワ、カシラ、ヤ、ケ、モノ
終助詞(長単位)	なし	なし
準体助詞(短単位)	ノ	ノ
準体助詞(長単位)	なし	なし

長単位の欄は、短単位にない形のものを載せている。それぞれ、出現形が異なるものも含んでいる。特に話し言葉には、縮約・融合の形が多く含まれる。

５．BCCWJ における助詞

５.１　全語数における助詞の割合と助詞内における各助詞の割合

助詞の数を以下に示す。

表 4　BCCWJ 全語数における助詞の割合（短単位）

	全語数	格助詞	副助詞	係助詞	接続助詞	終助詞	準体助詞	助詞総数 個数	%
新聞コア	308504	57052	4186	11838	8876	554	1207	83713	27.14
雑誌コア	202268	35722	3015	9108	8491	1483	2047	59866	29.60
書籍コア	204050	38083	3303	10426	10744	1369	2681	66606	32.64
白書コア	197011	36619	1985	4290	6084	59	120	49157	24.95
知恵袋コア	93932	14274	2086	4215	5457	2384	1976	30392	32.36
ブログコア	92746	14104	1533	4137	4405	1607	1309	27095	29.21
計	1098511	195854	16108	44014	44057	7456	9340	316829	28.84

この調査から、以下のことがわかる。

１）　全語数の約 30%が助詞である。

２）　助詞の中では格助詞が最も多い。助詞のうち 47%～74%が格助詞。

３）　知恵袋・ブログは、他に比べて、格助詞が少なく、終助詞が多い。

４）　白書と、知恵袋・ブログは、対極にある。白書はかなり特殊である。新聞は多少白書に似た性質を持つ。

全体：　格助詞＞接続助詞・係助詞＞副助詞＞準体助詞＞終助詞
新聞：　格助詞＞係助詞＞接続助詞＞副助詞＞準体助詞＞終助詞
雑誌：　格助詞＞係助詞＞接続助詞＞副助詞＞準体助詞＞終助詞
書籍：　格助詞＞接続助詞＞係助詞＞副助詞＞準体助詞＞終助詞
白書：　格助詞＞接続助詞＞係助詞＞副助詞＞準体助詞＞終助詞
知恵袋：格助詞＞接続助詞＞係助詞＞終助詞＞副助詞＞準体助詞
ブログ：格助詞＞接続助詞＞係助詞＞終助詞＞副助詞＞準体助詞

（上記二重下線は、他のレジスターに比べて相対的に多いもの、一重下線は少ないもの。以下同様。）

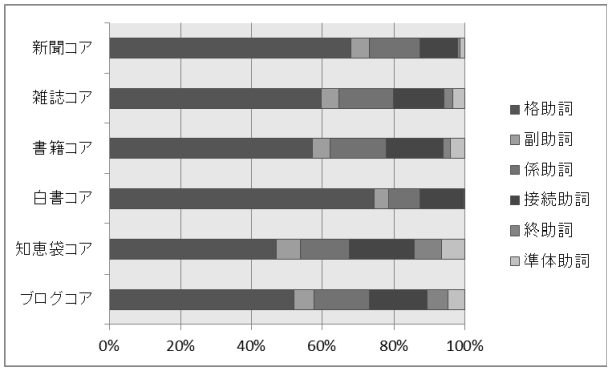


図 1　BCCWJ レジスターごとの助詞の割合（短単位）

長単位でも、全体の傾向は変わらない。格助詞が最も多い。
短単位の場合の１万語当たりの数は、表５の通りである。

表５　BCCWJ　１万語当たりの助詞の数（短単位）

	格助詞	副助詞	係助詞	接続助詞	終助詞	準体助詞	助詞全体
新聞コア	1849	136	384	288	18	39	2714
雑誌コア	1766	149	450	420	73	101	2960
書籍コア	1866	162	511	527	67	131	3264
白書コア	1859	101	218	309	3	6	2495
知恵袋コア	1520	222	449	581	254	210	3236
ブログコア	1521	165	446	475	173	141	2921
計	1783	147	401	401	68	85	2884

このクロス表の内容をもとに、ジャンルと助詞の関係をより詳細に把握するため、助詞タイプを第１アイテム、コーパス種別を第２アイテムとしてコレスポンデンス分析を行った。その結果、下記の散布図を得た（図２）。なお、第１次元の寄与率は90.64%、第２次元の寄与率は6.95%、２つの次元による累計寄与率は97.59%であるため、２つの次元に基づく解釈に一定の妥当性があると判断した。第１次元の寄与率が圧倒的である。軸解釈を行うと、第１次元はブログや知恵袋などのくだけた話し言葉的ジャンル（＋）と、新聞・白書のようなかたい書き言葉的ジャンル（－）を区分している軸と考えられる。また、第２次元は書籍・雑誌のような一般的内容を扱ったジャンル（＋）と白書のような特定内容を扱ったジャンル（－）を区分する軸と考えられる。このことから考えると、第１象限、つまり、くだけた言語と一般的内容を特徴とする領域には係助詞、準体助詞が多く、第２象限、つまり、くだけた言語と特定内容のジャンルには終助詞が多い。第３象限、つまり、かたくて一般的なジャンルに特徴的な助詞は存在せず、第４象限、つまり、かたくて特定内容のものには格助詞が多い。

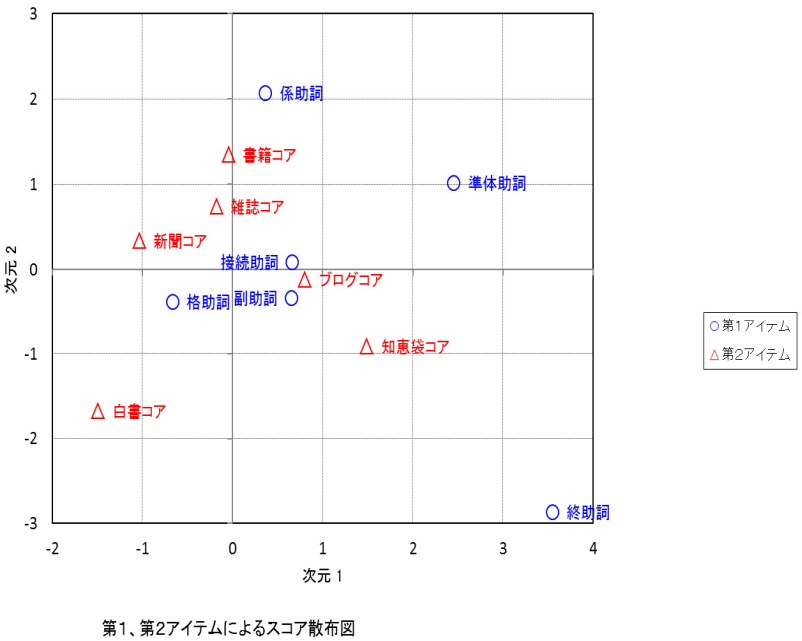


図２　BCCWJ コレスポンデンス分析の散布図（短単位）

５．２　それぞれの助詞における語の割合

５．２．１　格助詞

BCCWJ コアにおける格助詞の内訳は以下の通りである。

表 6　BCCWJ 格助詞の数（短単位）

	ガ	ヲ	ニ	ト	デ	へ	ヨリ	カラ	ノ	その他
新聞コア	7235	10294	9826	6165	5102	522	141	1362	16390	15
雑誌コア	5000	6192	6645	4168	2728	228	137	900	9714	10
書籍コア	5372	6576	7361	5079	2530	277	140	861	9880	7
白書コア	3545	6265	7948	3615	1456	341	129	675	12636	9
知恵袋コア	2375	2029	2641	1931	1552	57	97	326	3261	5
ブログコア	2018	1965	2643	1710	1371	118	82	410	3773	14
計	25545	33321	37064	22668	14739	1543	726	4534	55654	60

新聞：　ノ＞ヲ＞ニ＞ガ＞ト＞デ＞カラ＞へ＞ヨリ
雑誌：　ノ＞ニ＞ヲ＞ガ＞ト＞デ＞カラ＞へ＞ヨリ
書籍：　ノ＞ニ＞ヲ＞ガ＞ト＞デ＞カラ＞へ＞ヨリ
白書：　ノ＞ニ＞ヲ＞ト＞ガ＞デ＞カラ＞へ＞ヨリ
知恵袋：ノ＞ニ＞ガ＞ヲ＞ト＞デ＞カラ＞ヨリ＞へ
ブログ：ノ＞ニ＞ガ＞ヲ＞ト＞デ＞カラ＞へ＞ヨリ

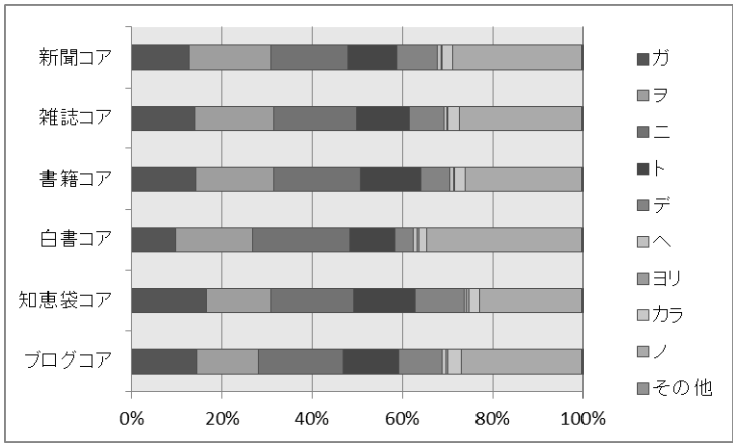


図 3　BCCWJ レジスターごとの格助詞の割合（短単位）

長単位で調査すると、だいぶ値が異なる。ニを伴う複合辞、トを伴う複合辞の数が多いことがわかる。特に白書には、「により」「における」等、ニを伴う複合辞が多い。

表 8　BCCWJ 格助詞の数（長単位）

	ガ	ヲ	ニ	ニを伴う複合辞	ト	トを伴う複合辞	デ	へ	ヨリ	ノ
新聞コア	7168	10177	8402	1180	5160	910	4998	522	133	16290
雑誌コア	4895	6156	5860	479	3126	882	2579	228	128	9633
書籍コア	5155	6534	6297	658	3651	1187	2375	277	126	9826
白書コア	3473	6148	3843	3676	2688	636	1417	341	128	12450
知恵袋コア	2310	2027	2390	130	1586	261	1450	57	93	3247
ブログコア	1967	1957	2386	136	1375	257	1197	118	77	3756
計	24968	32999	29178	6259	17586	4133	14016	1543	685	55202

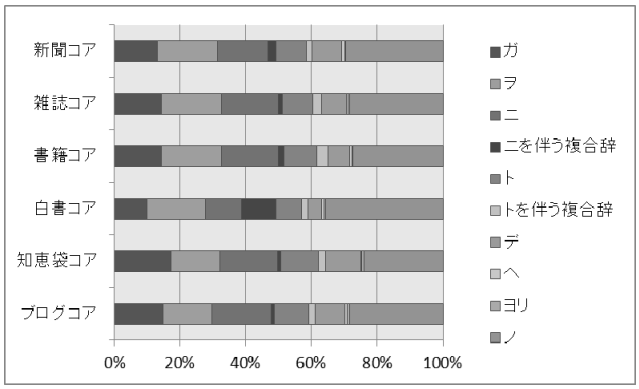


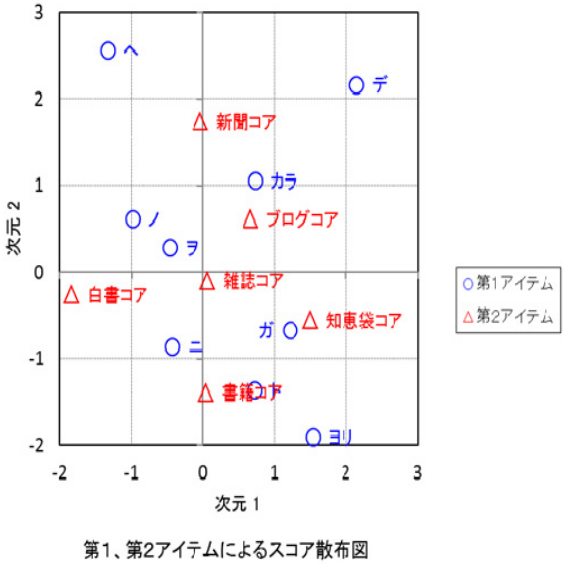
図 4　BCCWJ レジスターごとの格助詞の割合（長単位）

短単位の 1 万語当たりの数は、表 7 の通り。

表 7　BCCWJ　1 万語当たりの格助詞の数（短単位）

	ガ	ヲ	ニ	ト	デ	ヘ	ヨリ	カラ	ノ
新聞コア	235	334	319	200	165	17	5	44	531
雑誌コア	247	306	329	206	135	11	7	44	480
書籍コア	263	322	361	249	124	14	7	42	484
白書コア	180	318	403	183	74	17	7	34	641
知恵袋コア	253	216	281	206	165	6	10	35	347
ブログコア	218	212	285	184	148	13	9	44	407

このクロス表の内容をもとに、ジャンルと助詞の関係をより詳細に把握するため、助詞タイプを第 1 アイテム、コーパス種別を第 2 アイテムとしてコレスポンデンス分析を行った。その結果、下記の散布図を得た（図 5）。なお、第 1 次元の寄与率は 80.21%、第 2 次元の寄与率は 10.03%、2 つの次元による累計寄与率は 90.23%であるため、2 つの次元に基づく解釈に一定の妥当性があると判断した。長単位についても同様の分析を行い、散布図を得た（図 6）。軸解釈は、図 5,6 とも図 2 と同様でよいかと思われるが、長単位の方が、より、白書及びニ＋α（ニを伴う複合辞）の位置が特徴的となっている。



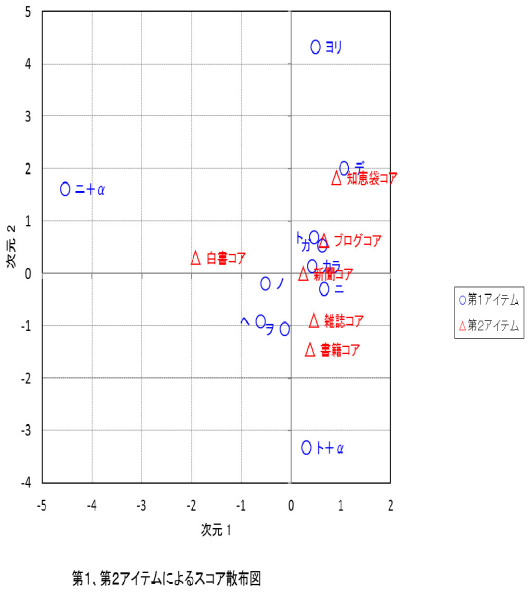


図 5 コレスポンデンス分析の散布図(短単位)　　図 6 コレスポンデンス分析の散布図(長単位)

格助詞に関しては、以下のことがわかった。

・格助詞の分布は、短単位と長単位でだいぶ異なる。特に、白書において違いが顕著である。
・短単位で白書にニが多い理由はニを伴う複合辞が多いからである。このことは、長単位の調査を行うとわかる。白書は、格助詞ニの 46.25%が複合辞である。新聞が 12%、あとのレジスターは一ケタである。「により」「における」「において」が多い。「により」「によって」は、白書以外は「により」より「によって」が多い。「に対し」「に対して」は、新聞のみ「に対し」が多い。複合辞に関わる格助詞はニとトが主である。接続助詞テが格助詞相当の複合辞を作ることが多いため、白書は短単位で調べると接続助詞のテが多い。
・デは白書には少ない。デは話し言葉的であり、デの代わりに複合辞を用いるためであると思われる。
・知恵袋にはガ・デが多い。

５．２．２　副助詞

BCCWJ コアにおける副助詞の数は以下の通りである。合計数が多い順に並べた。副助詞以降は、紙幅の関係で図を省略する。

表 9　BCCWJ 副助詞の数（短単位）

	ヤ	ナド	カ	マデ	ダケ	ッテ	タリ	ホド	クライ	シカ	バカリ	ナンテ	ノミ	サエ	その他	計
新聞コア	1106	1358	400	552	260	26	142	73	34	66	51	17	19	20	62	4186
雑誌コア	574	420	574	338	232	156	136	134	89	67	62	48	43	36	106	3015
書籍コア	506	387	830	354	313	132	143	144	101	81	59	37	24	48	144	3303
白書コア	881	613	70	224	42	0	53	26	3	3	3	0	45	2	20	1985
知恵袋コア	169	170	613	148	139	318	121	55	160	51	22	44	12	7	57	2086
ブログコア	129	82	474	172	127	169	74	45	88	49	23	32	17	6	46	1533
計	3365	3030	2961	1788	1113	801	669	477	475	317	220	178	160	119	435	16108

・新聞と白書にはヤ・ナドが多い。
・知恵袋・ブログにはカ・ッテが多い。

５．２．３　係助詞

BCCWJ コアにおける係助詞の数は以下の通りである。

表 10　BCCWJ 係助詞の数（短単位）

	ハ	モ	コソ	その他	計
新聞コア	8956	2840	41	1	11838
雑誌コア	6350	2706	47	5	9108
書籍コア	7316	3069	38	3	10426
白書コア	3577	711	2	0	4290
知恵袋コア	2780	1425	9	1	4215
ブログコア	2655	1466	13	3	4137
計	31634	12217	150	13	44014

・どのレジスターも、ハ＞モ＞コソの順である。
・白書は係助詞が全体的に少ないが、特にモが少ない。

５．２．４　接続助詞

BCCWJ コアにおける接続助詞の数は以下の通りである。合計数が多い順に並べた。

表 11　BCCWJ 接続助詞の数（短単位）

	テ	ガ	ト	バ	カラ	ケレド	ナガラ	シ	ツツ	その他	計
新聞コア	6652	922	473	380	172	36	145	43	35	18	8876
雑誌コア	6104	692	483	398	343	150	148	114	28	31	8491
書籍コア	7803	811	540	607	514	123	159	126	23	38	10744
白書コア	5416	207	193	87	37	1	62	1	74	6	6084
知恵袋コア	3304	890	361	307	265	150	35	132	6	7	5457
ブログコア	2893	437	237	184	219	217	57	127	19	15	4405
計	32172	3959	2287	1963	1550	677	606	543	185	115	44057

・白書は、テを用いた複合辞が多いので、テが多い。
・白書は、テ以外の接続助詞は少ない。

５．２．５　終助詞

BCCWJ コアにおける終助詞の数は以下の通りである。合計数が多い順に並べた。

表 12　BCCWJ 終助詞の数（短単位）

	カ	ネ	ヨ	ナ	ノ	ワ	サ	ゾ	その他	計
新聞コア	323	77	71	42	8	4	1	10	18	554
雑誌コア	508	237	330	165	75	47	53	26	42	1483
書籍コア	626	183	206	125	67	61	25	20	56	1369
白書コア	58	0	0	0	1	0	0	0	0	59
知恵袋コア	1286	435	449	136	36	7	3	5	27	2384
ブログコア	417	468	288	265	40	22	30	13	64	1607
計	3218	1400	1344	733	227	141	112	74	207	7456

・白書は、カ以外はノが１件あったのみ。
・知恵袋は、カが多い。
・ブログは、出現形の種類が多い。
例）ナ：265 例中「なぁ」67 例、「な～」20 例、「なー」5 例、「なァ」2 例、「ナー」1 例
　　ヨ：288 例中「よ～」14 例、「よぉ」10 例、「よー」8 例、「ヨ」5 例、「よん」2 例、
　　「よう」1 例、「よ～ん」1 例

６．CSJ における助詞

６.１　全語数における助詞の割合と助詞内における各助詞の割合

表 13　CSJ 全語数における助詞の割合（短単位）

		全語数	格助詞	副助詞	係助詞	接続助詞	終助詞	準体助詞	助詞総数 個数	助詞総数 %
学会講演	A11F0703	5634	924	199	126	299	57	95	1700	30.17
	A05F0043	3512	567	39	76	141	9	34	866	24.66
	A11M0369	3119	556	28	89	97	0	7	777	24.91
	A11M0469	6763	1111	123	138	411	43	62	1888	27.92
	A グループ計	19028	3158	389	429	948	109	198	5231	27.49
自由会話	D03F0034	3021	266	201	56	129	206	65	923	30.55
	D03F0058	2330	217	236	37	99	138	68	795	34.12
	D03M0004	2491	266	106	44	107	96	90	709	28.46
	D03M0038	3638	422	175	71	178	177	88	1111	30.54
	D グループ計	11480	1171	718	208	513	617	311	3538	30.82
総計		30508	4329	1107	637	1461	726	509	8769	28.74

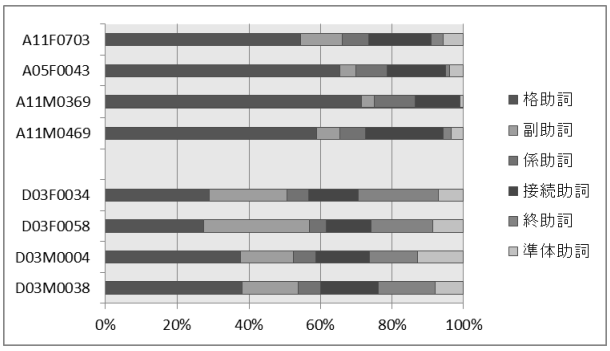


図 7　CSJ 独話（学会講演）・対話（自由会話）の助詞の割合（短単位）

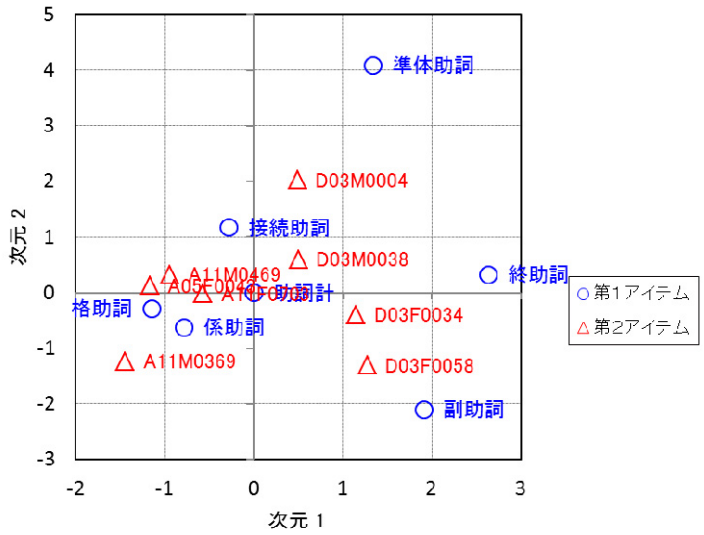


図 8　CSJ コレスポンデンス分析の散布図（短単位）

Aグループ（独話）とDグループ（対話）は、かなり異なる。図 8 を見ると、横軸の＋－できれいに分かれているのがわかる。独話（学会講演）には格助詞が多く、対話（自由会話）には終助詞・副助詞が多い。

６．２　それぞれの助詞における語の割合

６．２．１　格助詞

表 14　CSJ 格助詞の数（短単位）

		ガ	ヲ	ニ	ト	デ	ヘ	ヨリ	カラ	ノ	その他	計
学会講演	A11F0703	114	135	157	196	75	0	0	42	205	0	924
	A05F0043	110	60	94	92	20	0	3	7	180	1	567
	A11M0369	88	82	59	93	31	5	0	20	176	2	556
	A11M0469	144	191	154	210	60	2	0	79	271	0	1111
	A グループ計	456	468	464	591	186	7	3	148	832	3	3158
自由会話	D03F0034	47	4	34	71	50	0	1	6	53	0	266
	D03F0058	36	19	28	60	38	0	5	1	30	0	217
	D03M0004	51	20	51	41	35	0	0	7	61	0	266
	D03M0038	100	60	63	87	50	3	0	15	43	1	422
	D グループ計	234	103	176	259	173	3	6	29	187	1	1171
総計		690	571	640	850	359	10	9	177	1019	4	4329

６．２．２　副助詞

表 15　CSJ 副助詞の数（短単位）

		カ	モ	ッテ	ナド	ダケ	クライ	マデ	タリ	ヤ	ナンカ	その他	計
学会講演	A11F0703	43	72	48	3	5	10	1	4	0	8	5	199
	A05F0043	4	20	1	1	2	0	3	0	1	0	7	39
	A11M0369	1	6	0	10	0	0	0	0	10	0	1	28
	A11M0469	25	52	1	22	4	4	8	0	2	0	5	123
	A グループ計	73	150	50	36	11	14	12	4	13	8	18	389
自由会話	D03F0034	76	54	42	0	12	7	3	4		1	2	201
	D03F0058	127	44	49	0	3	4	4	4	0	0	1	236
	D03M0004	40	23	32	0	2	3	2	3	0	1	0	106
	D03M0038	59	42	63	0	2	2	4	2	0	0	1	175
	D グループ計	302	163	186	0	19	16	13	13	0	2	4	718
総計		375	313	236	36	30	30	25	17	13	10	22	1107

・独話（学会講演）にナドを多く使う人がいる。個人差がある。
・対話（自由会話）にはカ・ッテが多い。

６．２．３　係助詞

・CSJ ではハのみを係助詞としている。会話より講演の方が使用している。

６．２．４　接続助詞

・講演にテが多い。会話は縮約形「てる」が助動詞とされていることも影響していると思われる。
・講演にガが多く、会話にケレドが多い。（講演にケレドを用いている人も一人あり。）

６．２．５　終助詞

・会話の方が終助詞が多い。

７．複合辞について

長単位として扱う場合と、長単位にしない場合がある。例えば、「ので」「のに」は、BCCWJ では、短単位分割でも長単位分割でも準体助詞「の」＋助動詞「だ」、準体助詞「の」＋格助詞「に」として扱うが、CSJ では、長単位分割では接続助詞として扱っている。複合辞の扱いは今後の課題である。

８．複数の分類にまたがるものの扱いについて

例えば「って」は、少なくとも三種に分けられる。「手術って聞いてびっくりした」は格助詞、「~なんですって。」は終助詞、「人生って楽しいことばかりじゃないよ」は係助詞。しかし、BCCWJ 及び CSJ においては、すべて副助詞として扱っている。形態素解析としては副助詞として扱うとしても、その働きの違いを明らかにする必要がある。

９．まとめと今後の課題

・全語数の約 30%が助詞である。助詞の中では格助詞が最も多い。
・BCCWJ において、知恵袋・ブログは、他に比べて、格助詞が少なく、終助詞が多い。
・白書と、知恵袋・ブログは、対極にある。白書はかなり特殊である。新聞は多少白書に似た性質を持つ。知恵袋とブログは、ともに話し言葉的な性質を帯びているが、両者間で異なる性質も持つ。知恵袋の方がより独特である。
・CSJ においては、同一話者においても、学会講演と自由会話における違いが見られた。
・助詞の分類の仕方、認定の仕方には課題も残る。

漢語動詞における格表示変化傾向の探索 —ヲ格とニ格—

An Exploratory Study of Changes in Case Marking with Sino-Japanese Verbs: Shifts between *o* and *ni*

服部匡(同志社女子大学表象文化学部)
Tadasu Hattori (Doshisha Women's College of Liberal Arts)

要旨

二字漢語動詞のうち、その意味的な項となるニ格とヲ格の成分が、大きく意味役割を換えることなく交替する例のあるものについて、主に 60 年間の国会会議録のデータを用い、格助詞の選択傾向の変化を探索した。先行研究で主張された一般傾向とは反対にヲ格からニ格への推移が見られる動詞が少なくとも 6 語あり、ニ格からヲ格への推移が見られる動詞などもある。

1. 目的と方法

下記のデータを用い、二字漢語動詞のうちニ格とヲ格に交替の余地があるものについて、格助詞の選択傾向の変化を探索的に研究する。

・1947～2007 年の国会会議録　　約 35 億字
(国会図書館のサイトからダウンロードしたもの)

より早い時期の用例を知るため一部の語では補助的なデータとして下記のものを用いる。

・1911～1944 年の新聞記事　　約 0.5 億字
(神戸大学付属図書館『新聞記事文庫』の 37,776 記事(2015.7.10 取得))
・青空文庫収録の作品[1]　　約 1.5 億字
(ひまわり用『青空文庫』パッケージ(国語研究所 2015.4.2)に含まれる 12,279 作品)

2. 先行研究

コーパスに基いて漢語動詞の統語的性質の通時変化を扱った研究には永澤(2007)があるが、動詞の自他という観点からのものである。コーパスデータから漢語動詞でのニ格とヲ格の入れ替わりの傾向を探索した包括的な研究が従来なかった。

「～{ニ/ヲ} 怖づ」のように動詞の項に関してニ格とヲ格が交替する現象は古くから見られることで、また「背く、慣れる、祈る」など、現代語では主にニ格をとる動詞が古典語ではヲ格をとった例がある（「極める」のようにその逆の例もある。山田(1980)、信太(1981)、坂梨(1981)、小田(2010)などによる)。

現代日本語では、「触る、頼る、耐える、(宿{に/を})当たる」などでニ格とヲ格が交替しうることを塚本(1991)が指摘している。影山・高橋(2011)は「触る、頼る」などで、「「を」＝全体的・直接的作用、「に」＝部分的・間接的作用」という意味的な差異があるという。

漢語動詞については、丸山(2011)が「複数の格助詞を殆ど同じように使うことができるも

[1] 大部分は 1850 年代から 1910 年代までの生まれの著者によるものである(服部 2014)。

の」の例として「欠席する、信頼する、納得する」を挙げている。

通時変化に関しては、工藤(2012)が、漢語動詞の格支配で「（カラ/ニ/デ/ヲから)『を』一つに収斂する」という変化が進行中であると主張する。例えば「～ニ配慮する→～ヲ配慮する」、「(医院等）デ受診する→（医院等）ヲ受診する」、「(人）カラ聴取する→（人）ヲ聴取する」のような変化が進行中であるという。「受診する」「聴取する」については新聞記事データベースの用例数変化が根拠としてあげられている。

また、島田(2014)は近年の若年層でニ格から他の格への移行が見られるといい、漢語動詞の「言及する」「暴行する」、和語の「鑑みる」「心がける」などでニからヲへの移行が進んできているという。また、塩田(2006)はウェブでの質問調査に基づき「参拝する」という動詞で若年層ほどニよりヲを用いる傾向があると指摘している。

このように、個別の動詞の格表示変化の指摘やそれに基づいた一般的変化傾向の仮説提示は行われているが、潜在的にはニとヲが交替しうる動詞の全体の中でどれだけのものにどのような方向の変化が起こっているのか、という観点からの定量的研究が不足していた。理想的にはニ格・ヲ格の例があるすべての動詞について均しくデータを分析する必要がある。今回は、形式的条件により網羅することが容易な二字漢語動詞を対象とする。

3. コーパスに見るニ/ヲの交替・変化

国会会議録(1947～2006)から、「{に/を}+漢字二字+{する/致す(いたす)}」の部分を抽出し、形式的に特定しうるゴミを排除した(動詞は諸活用形を含むが受身・使役は除く。格助詞と漢語が隣接するものに限る)。この段階でニ格とヲ格両方の用例があり両者合わせて 100 例に達する動詞の中で、意味的にヲ格とニ格が交替する余地がありそうな動詞を選び用例を精査・選別した。その結果、実際にヲ格とニ格が通時的または共時的に交替していると思われる動詞を以下に取り上げる。このような手順によるため、ヲとニの交替がある動詞の一部をまだ見落としている可能性がある。

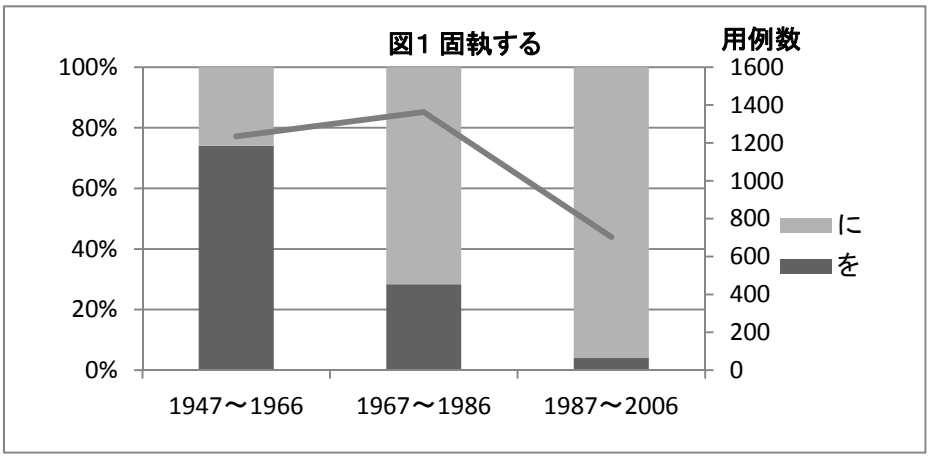
1947 年から 20 年ごとの 3 期間にわけ、ニとヲの比率、および合計の用例数 [2] を示す。また、なるべく形式的・意味的に似た性質のニ格/ヲ格成分をとっている例を並べて示す。

片方の助詞を伴う用例がわずかな数しかない動詞[3]、ヲとニ合わせての用例数が 0 に近い期間のあるものなどはとりあげない。

3.1. ニ格の比率が増大しているもの

「固執する、反撃する」の 2 語で特にニ格の増加傾向が顕著である。他にも、程度はともあれ､相対的にヲ格よりもニ格が優勢になる方向への推移の見られる動詞が 4 語存在する。以下にそれぞれ数値をグラフで示し観察する。

[2] 期間によって会議録の総文字数が異なるので、動詞の頻度変動の指標としてはこの数値は不適である。
[3] 「楽観する、懸念する、考慮する」などではニ格の用例が、「賛成する」「反対する」などではヲ格の用例が(比率として)少数ある。



固執する

(1) 国が国という立場で、国の訴訟代理という立場だけに固執するならば（佐々木静子,1974）

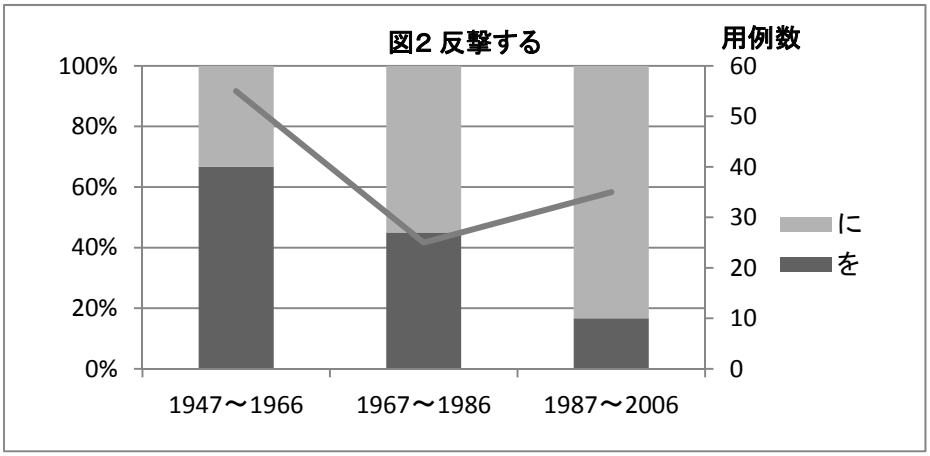
(2) 一つ電電公社にも在来の方式だけを固執するようなことのないように、技術的な進歩というものに対してもう少し目を開いて、（田中角榮,1957）

(3) 私はただいまの案でよろしいと思っておりますが、よりよき修正案がありますならば、あえて原案に固執するものではございません。（藤枝泉介,1967）

(4) 提案者としてはあえて原案を固執するものでございません。（井手以誠,1955）

(5) ひとりわが国のみが古典的な自由資本主義に固執し、やがて動脈硬化の経済体制に追いやろうとしております。（多賀谷委員,1961）

(6) いたずらに経済理論に走り、資本主義を固執する吉田内閣の欠陥は、万人認めるところの民主不安定政策であります。（堤（ツ）委員,1953）



反撃する

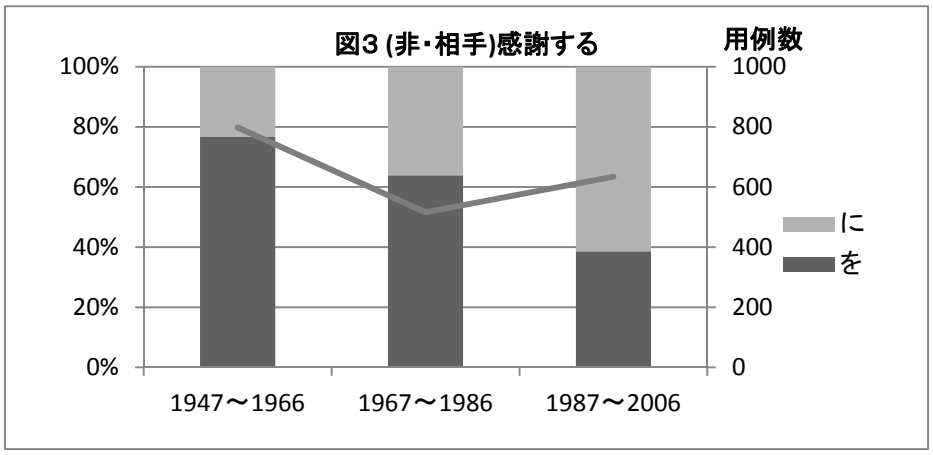
(7) これは朝鮮戦争に派遣された大国中心の国連軍というのが、力によって北朝鮮からの攻撃に反撃するということを目的にして出された。（芹田健太郎(公述人),1992）

(8) それに対するオレンジ軍の攻撃を反撃して、つまりエンタープライズを護衛する訓練をしたということになりますね。（上田耕一郎,1984）

(9) 保守政権のもとで、重税に苦しみぬいた国民層か政府に反撃した結果、しぶしぶ実施した国民世論の勝利であります。（平林剛,1957）
(10) われわれはこの点で大いに政府を反撃して選挙演説をやるのに都合がいい、この点はまことに感謝にたえないことです。（坂本昭,1957）

「感謝する」「反論する」「配慮する」でもヲ格からニ格への推移が観察されるが、これらの動詞では、格成分の意味役割への考慮が必要になる。

「感謝する」には、大別して、「{国民/アメリカ/英霊}に感謝する」のように<相手>の項を取るものと、「{協力/お答え/好意}に感謝する」のように<事柄>の項を取るものがある。<相手>では、ニ格の例はあるがヲ格の例がない。そこで、明確に<相手>の項とみなせる例を除いた場合(<相手>かどうか判定しにくい例も含む)の数値[4]をあげると次のようであり、ニ格の優勢化が見て取れる。



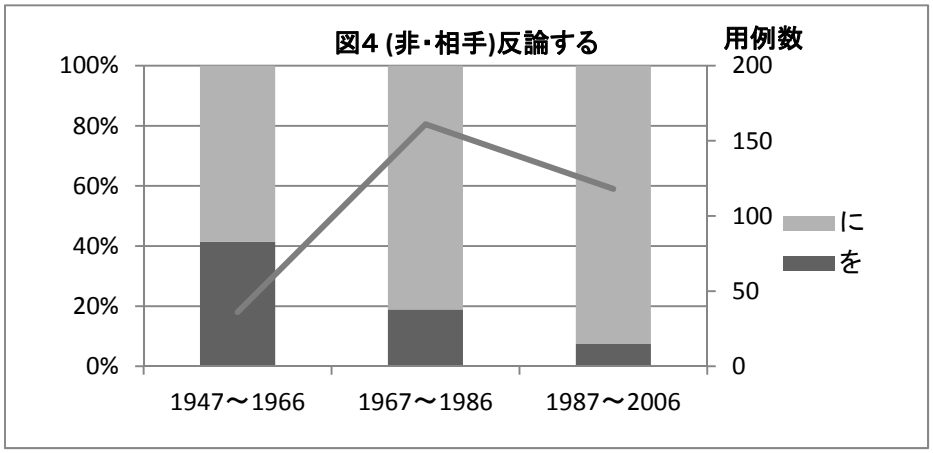
<事柄>感謝する
(11) それで私はもう時間がございませんから、委員長の御好意に感謝してこれで私の質問を終りたいと思います。（須藤五郎,1952）
(12) そこで、本問題について、貴国が従来示されたご好意を感謝すると共に、今後一層のご援助を得て（廣瀬小委員長,1957）

(13) コーエン長官からは、日本政府の協力に感謝するとともに、これらの措置が実施されることを期待しているという発言がございました。（渋谷政府参考人,2000）
(14) 総理から従来の協力を感謝するとともに、いまおっしゃったような証言の問題についても、一層のアメリカ側の協力を得られるよう（稲葉国務大臣,1976）
(15) 我が国に対する、また我が国の国民に対する皆様の御支援に感謝しております。（マイケル・トーマス・ソマレ(参考人),2006）

[4] 「<相手>に<事柄>を感謝する」の形の例が 2 例ある。下記に示す。
(i) また総理は、昨日、我が党村議員からの戦犯の軍神扱いはやめよという立場から、合祀している戦犯に何を感謝するのかと問われたのに対して、まともに答えず、冷たい言葉をかける人は正常な人間の心を持っているか甚だ疑問に思うと言われました。（安武洋子,1985）
(ii) 右報告を終るに当り、今回の出張に際し福岡県当局関係の労使双方及びスト規制法案の懇談会に出席された公益代表の各位に御協力を感謝する次第であります。（専門員（高戸義太郎）,1953）

(16) 数値目標は若干上回る形で達成させていただきまして、大変皆さんの御支援を感謝しております。（生田参考人,2005）

「反論する」でも「感謝する」と同様、<相手>の場合にはニ格の例しかない。そのため、明確に<相手>の項であるとみなせる例を除く。さらに、「ありのままを反論して～」「そうでないということを反論して～」「これだけ巨額でいいのかということを反論しようと思えばできます」のようにヲ格が<反論内容>であるものはニ格と交替しないため、その明確な例は除く。その結果の数値を見ると、「感謝する」の場合と同じ傾向が認められる。



<事柄>反論する

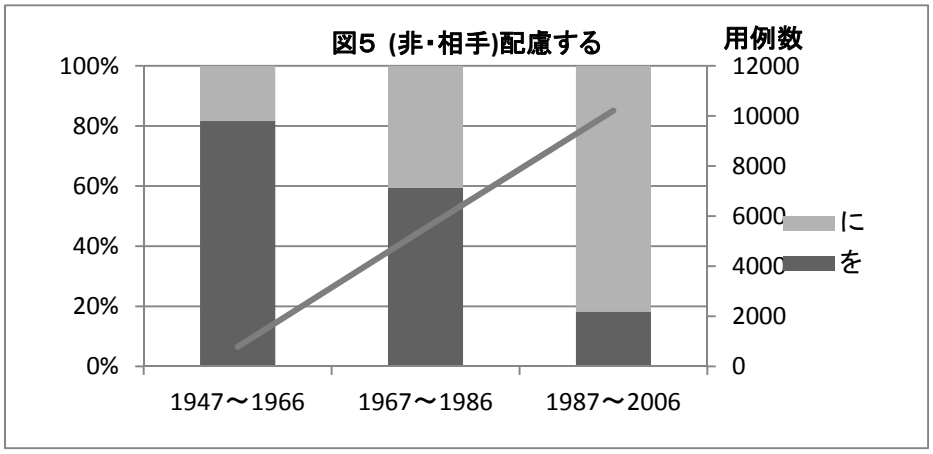
(17) このことに反論する一部の論拠といたしまして、サンフランシスコ条約における直接占領軍事費に同資金が含まれていないこと、あるいは同条約の第十四条。（田中幾三郎,1962）

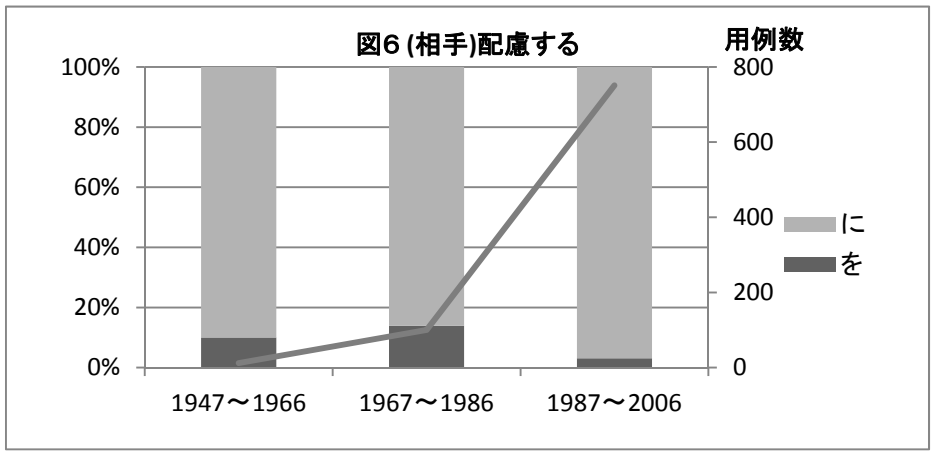
(18) 視聴した結果、ほかの学者がほかの公開されておりますいろいろな機関でそのことを反論すること自身については、もちろん、これは学問的に自由でございますけれども、（宮地政府委員,1980）

(19) 積極的な位置づけとして申し上げたわけではなくて、この提案理由の説明に反論するといいますか、（工藤公述人,1972）

(20) これでは、国防会議自体において制服の説明を反論し、あるいはこれを補佐し、修正する実際の資料を作成することは不可能となり、（石橋政嗣,1956）

「配慮する」にも、「消費者に配慮する」のように<相手>の項を取るものと、「{趣旨/融資/プライバシー}に配慮する」のように<事柄>の項を取るものとある。「感謝する」「反論する」と異なるのは、「配慮する」では、<相手>の場合でもニ格と並んでヲ格の例もあることである。そこで明確に<相手>の場合とそれ以外の場合に分けてそれぞれ数値を示すと次のようになる。どちらの場合にも長期的にはニ格が伸張しているが、<相手>の場合は<事柄>の場合に比べて早くからニ格優勢であったことが分かる。ただ後者では初期の例がごく少ない。





<事柄>配慮する

(21) プライバシーに配慮した運用のルールなどはどのようになされているか教えていただけますでしょうか。(福島瑞穂,2001)

(22) 先ほど塩川証人は、血友病患者さんの独特のプライバシーなどを配慮して安部先生は出さなかったというふうにおっしゃいますけれども、(土肥委員,1996)

(23) 私ども各金融機関に対しましては特に中小企業向けの融資に配慮するようにという指導を加えておるのでございますが、(森永貞一郎(参考人),1975)

(24) 十一億五千万円の融資を配慮しておられるというので、聞いておると非常に大変な心配をしておられるように聞えるのです。(兼岩傳一,1949)

(25) 特別徴収となる年金の範囲については、公租公課禁止規定の趣旨等に配慮し、遺族年金、障害年金、老齢福祉年金は含まれてないというふうに言っているんですよ。(小池晃,2005)

(26) 肥料工業の構造改善に当たっては、産業構造審議会の答申の趣旨を配慮しつつ、生産コストの低減が進められるよう指導すること。(竹内（猛）委員,1979)

<相手>配慮する

(27) これらの利用者に配慮しました、例えばエレベーターつきの横断歩道橋の設置に当たりましては、(藤田忠夫(説明員),1990)

(28) 二番、敷地内及び館内における誘導ブロックの設置や車いす利用者を配慮した動線の整備。（小川榮一(参考人),2006）

下記の数値も合わせて考えると、工藤(2012)が「～を配慮する」を格支配の変化により近年生じた言い方であるとするのは疑わしい。新聞記事と青空文庫でのヲ格の早い用例をあげておく。ただし新聞でのニ格の 1 例は 1914 年と早く、それ以前の状況は不明である(『国民之共』32 号、1888 年、に「～を配慮する」の例がある)。

「青空文庫」の用例数（すべて非・相手）

ニ配慮する 0　ヲ配慮する　1

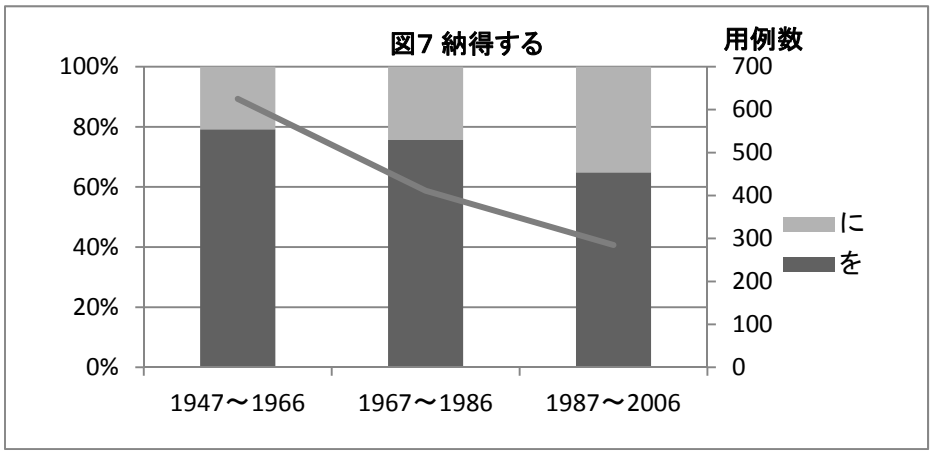
「新聞記事文庫(1911～1944)」の用例数（すべて非・相手）

ニ配慮する 1　ヲ配慮する　4

(29) 会社は自利一点張の為に彼等坑夫の保健並に生活状態を配慮するの遑あらず（台湾日日新報 1920.3.13-25）

(30) 甥の将来の安定を配慮するためにした冬の旅（ベートーヴェンの生涯 ロマン・ロラン　片山敏彦(1898～)訳 1944）

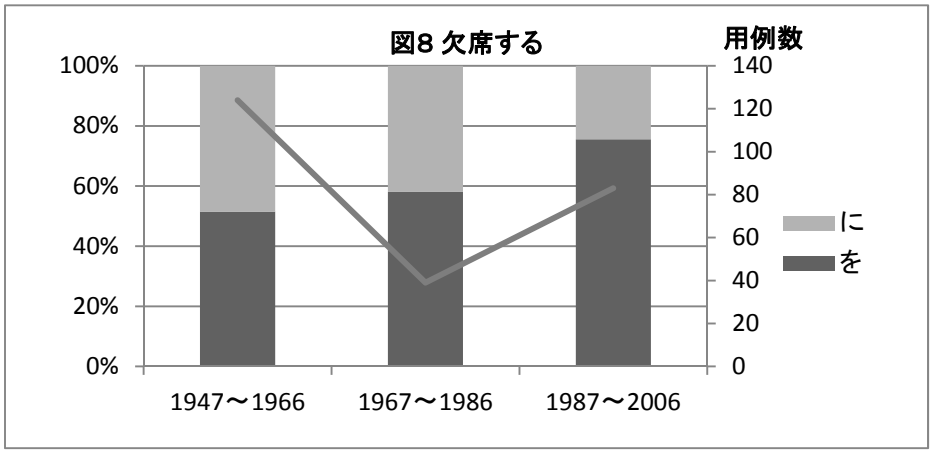
これらの語ほど明瞭ではないが、「納得する」でも若干ニ格の比率が増加している。「人を納得する」のような使役的な意味のものは、当然、用例数から除いた。

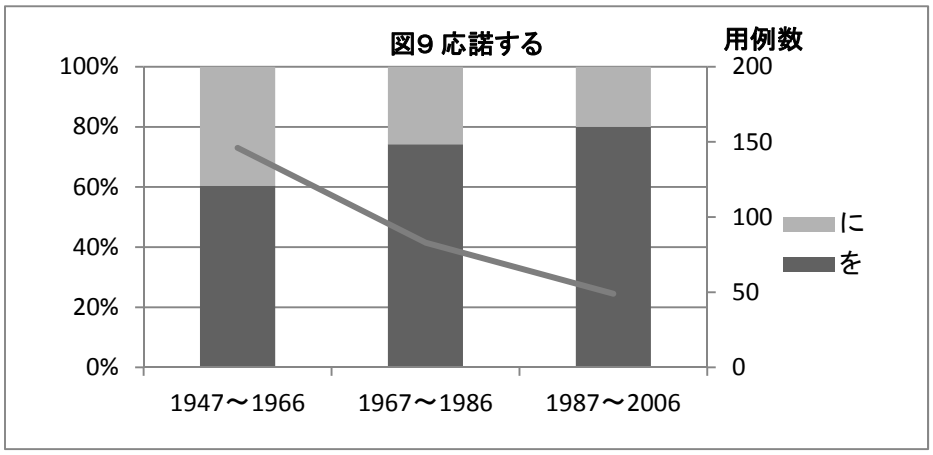


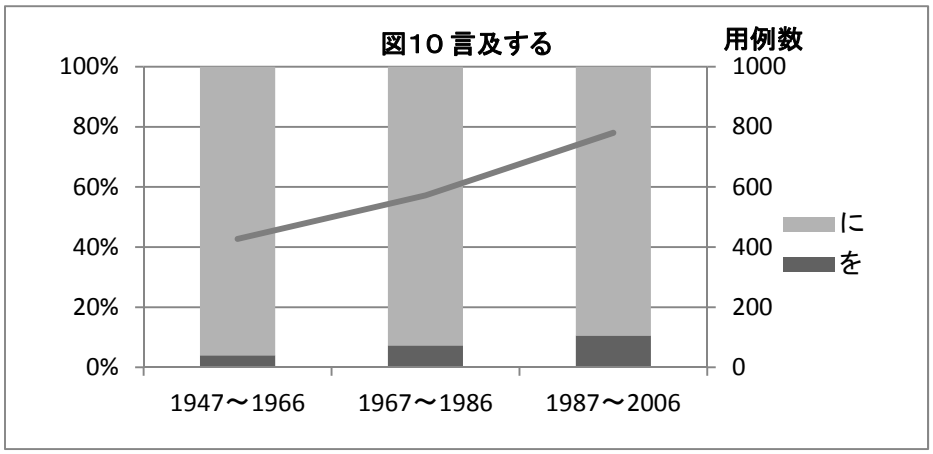
以上の 6 語でヲ格からニ格への推移傾向が観察された。その要因としては、語により、類義動詞の格表示の影響(例:「固執する」に対する「こだわる」「執着する」)、<相手>のニ格への類推などが考えられるが、説得的な説明はない。また、語によって、動詞の意味そのものの変化の可能性を検討する必要がある。これは今後の課題である。

3.2.　ヲ格の比率が増大しているもの、その他

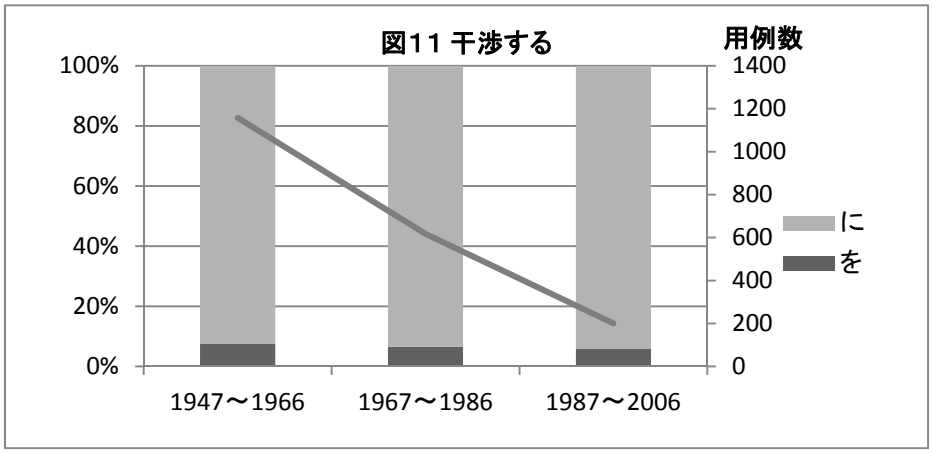
「欠席する」「応諾する」「言及する」ではヲ格の比率が上昇しているように見える。上昇したとすれば、類義動詞の格選択との関連(例：「欠席する」に対する「休む」)などが要因として考えられるが、やはり説得的な説明はない。紙数の都合で例は省略する

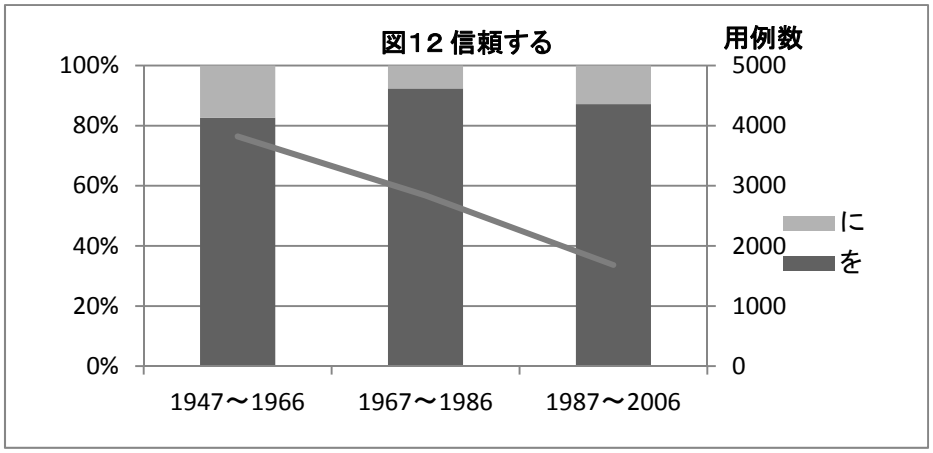






他に、ニ格とヲ格の比率に明確な変化傾向を見出しにくい語がいくつかある。以下に数値をあげる。グラフを省略するが「参拝する」も同様である。





「信頼する」では、憲法前文にある「(平和を愛する諸国民の)公正と信義に信頼(して)」の語句を含むものが 413 例あり、それらを除外した場合は 1987-2006 年のニ格の比率がやや低くなる。「～に信頼する」は、(31)(32)のように、古くからある言い方であるが、下記の二つの数値を考え合わせると、長期的にはヲ格が勢力を伸ばしてきた可能性がある。

(31) 長が荏苒として愈えなかつたことと、榛軒が清川玄道の技倆に信頼してゐたこととが知られる。(森鴎外(1862～) 伊沢蘭軒)

(32)「無論なら安心して、僕に信頼したらよかろう。」(夏目漱石(1867～) 二百十日)

「新聞記事文庫(1911～1944)」の用例数

ニ信頼する 108 ヲ信頼する 61

「青空文庫」の用例数

ニ信頼する 111 ヲ信頼する 154

4. 結語

本研究では、二字漢語動詞におけるニ格とヲ格の使用傾向の推移を調査検討した。相対的に、ヲ格よりもニ格が優勢になる方向への推移を示す動詞が少なくとも 6 語あり、その反対方向の傾向を示す語も見られる。工藤(2012) のいうようなヲ格への収斂、あるいは島田(2014)のいうようなニ格の衰微といった一般化は、少なくとも過去の数十年単位での全般

的な変化傾向の記述としては、裏付けることができない。もっとも、今回扱ったよりも後の世代の話者、あるいは、国会会議録には現れにくいようなスタイルでは別の傾向が見られる可能性はある。

本論では動詞の意味用法の幅(ヲ・ニによる相違)やその変化の面は、十分な観察分析をなしえていない。これらを含めた現象の総合的な記述は今後の課題である。

参考文献

小田勝(2010) 『古典文法詳説』 おうふう.

影山太郎・高橋勝忠（2011）直接目的語と前置詞付き目的語　影山太郎（編）『日英対照名詞の意味と構文』119-147. 大修館書店.

工藤力男(2012) 『日本語に関する十二章　詫びる？詫びない？日本人』 和泉書院.

塩田雄大(2006) インターネットを用いた言語調査の一試論　公開型ウェブ調査の結果から 『NHK 放送文化研究所年報 2006』 93-122.

島田泰子(2014) 現代日本語におけるニ格表現の衰微と交替 『二本松学舎大学論集』 57: 45-65.

信太知子(1981) 「～をそむく」から「～にそむく」へ ―動作の対象を示す格表示の交替― 『国語語彙史の研究 二』 和泉書院.

坂梨隆三(1982) 近代の文法Ⅱ(上方篇) 築島裕(編)『講座国語史　第 4 巻　文法史』 大修館書店.

塚本秀樹(1991) 日本語における格助詞の交替現象について 『愛媛大学法文学部論集 文学科編』 24: 103-127.

永澤済(2007) 漢語動詞の自他体系の近代から現代への変化 『日本語の研究』3/4: 17-32.

服部匡(2014) 現代日本語の通時変化 『講座日本語コーパス 6 コーパスと日本語学』 朝倉書店.

丸山直子(2011) 動詞の格情報 ―国語辞書の記述とコーパス― 『日本文学』107: 227-245 東京女子大学.

山田みどり(1980) 「～をそむく」と「～にそむく」 『成蹊国文』 14: 24-85.

近代語から現代語にかけての名詞修飾表現の変化についての一考察 —1 項名詞に前接する限定詞を例に—

庵　功雄（一橋大学国際教育センター）†

Remarks on the Change in Noun-Modifying Expressions between Early-Modern and Modern Japanese: In Case of Determiners Heading One-Place Nouns

Isao Iori (Hitotsubashi University)

要旨

文脈指示用法で使われる限定詞「この」と「その」に後接する名詞を見ると、「その」の場合は「1 項名詞（「〜の」を統語的に要求する名詞）」が占める割合が高い。しかし、現代語では「その」は文体的な理由からあまり使われず、「ゼロ」が使われることが多い。一方、近代の文語文では同じ環境で「その」が使われる割合が高い。本発表では、この点を現代日本語書き言葉均衡コーパス（BCCWJ）と太陽コーパスを用いて検証した。そして、そうした変化が起こった原因として、近代語では、「そ」が独立語としての用法を持っていることを挙げた。

１．はじめに

動詞の研究に比べ、名詞の研究は大きく立ち後れている。発表者はこれまで、名詞についていくつかの考察を行ってきた（cf. 庵 1995a, 1999, 2007, Iori 2013）が、本発表では、近代語と現代語の相違という観点からこの問題について少し考えてみたい。

２．２種類の名詞

庵(1995a, 2007, Iori 2013)などで指摘してきているように、日本語の名詞は 2 種類に大別できる。すなわち、「〜の」を義務的に取る名詞と、それを義務的には取らない名詞である。前者を「1 項名詞（one-place noun）」、後者を「0 項名詞（zero-place noun）」と呼ぶ。両者は次の統語的テストによって区別できる（庵 1995a, 2007, Iori 2013）。

(1) **「そうですかテスト」**

A と B の対話の始発文で話し手 A が φ N[1]（N は名詞）を含む文を発したとき、協調的な聞き手 B が「ああそうですか」（に相当する表現）で答えて談話を閉じられるとき、その名詞 N を「0 項名詞」と称し、そのように答えることができず、必ず「X の？」（X は疑問詞）という疑問を誘発するとき、その N を「1 項名詞」と称する。1 項名詞は統語的に項を必須的にとるのに対し、0 項名詞は項を必須的にはとらない。

例えば、次の(2)(3)から「著書」は 1 項名詞であるのに対し、「本」は 0 項名詞であることがわかる[2]（(3B)で「ああそうですか。」が可能であるのに対し、(3A)ではそれが不可能であることに注意されたい）。

(2) A：先週、著書を読んだんですよ。

---

† isaoiori AT courante.plala.or.jp

1 φ はそこに音形を持つ要素がないことを表す。

2 「1 項名詞－0 項名詞」の区別は、西山(2003)の「非飽和名詞－飽和名詞」の区別と似ているが、両者は別の概念である。両者の関係については、庵(2007)を参照されたい。

B：??ああそうですか。／えっ、だれの？
(3) A：先週、本を読んだんですよ。
B：ok ああそうですか。／えっ、だれの？

3．名詞の種類と限定詞の使い分け

このように、日本語の名詞は2種類に大別されるが[3]、この違いが文脈指示（anaphoric）用法の限定詞（determiner）「この」と「その」の使い分けに反映している（庵 1995b, 2007）[4]。

「この」と「その」の文脈指示用法は「指定指示」と「代行指示」に分かれる。「指定指示」は「この／その NP」全体で先行詞（antecedent）と照応するものであるのに対し、「代行指示」は「こ／そ」の部分だけが先行詞と照応するものである。

(4) この間築地で寿司を食べたんだが、｛この寿司／その寿司｝はうまかった。（指定指示）
(5) この間築地で寿司を食べたんだが、｛この味／その味｝はよかった。　（代行指示）

ここで、現代日本語書き言葉均衡コーパス（BCCWJ）における、「その」と「この」に後接する名詞の分布を見ると次のようになる[5]。

表1　「その」と「この」の頻度（BCCWJ）

順位	その		この		順位	その		この	
1	事	2727	事	2694	26	辺	393	地	457
2	物	2478	点	1713	27	一つ	391	作品	435
3	時	2476	法律	1670	28	下	381	時代	422
4	中	2437	場合	1592	29	手	375	二人	418
5	後	2417	問題	1443	30	姿	373	町	385
6	他	2043	時	1431	31	結果	372	前	384
7	為	2031	人	1377	32	度	357	方法	370
8	人	1463	本	896	33	内容	336	地域	367
9	日	1288	辺り	810	34	話	326	他	356
10	場	1127	二つ	778	35	年	317	部屋	355
11	上	1063	国	757	36	顔	314	方	352
12	間	832	時期	754	36	目	314	条	344
13	頃	815	日	748	38	先	298	法案	343
14	内	798	辺	704	39	辺り	287	手	343
15	意味	793	男	702	40	夜	272	件	335
16	前	723	中	683	41	一方	264	章	330
17	侭	688	間	670	42	気	258	世界	307
18	点	621	侭	637	43	場合	256	場	302
19	言葉	614	子	634	43	次	256	時点	284
20	男	601	言葉	604	45	家	247	程度	269
21	方	584	事件	572	46	原因	246	店	266

[3] 固有名詞は全て0項名詞である。
[4] ア系統には結束性（cohesion）に関わる意味の文脈指示用法は存在しないため、「あの」は考察対象に含まれない（cf. 庵 1994, 2007）。
[5] 検索には中納言を用い、長単位検索を行った（キー：語彙素＝其の／此の、キーから1語：品詞＝名詞）。

22	理由	563	種	568	47	国	207	頃	252
23	子	474	家	547	48	通り	206	仕事	244
24	*声*	440	年	518	49	場所	205	質問	242
25	*名*	411	話	490	50	*数*	197	項	242

表 1 において□で囲んだものは 1 項名詞の用法でしか使われないものであり、斜字体にしたものは 1 項名詞としての用法を持ちうるものである。これを見ると明らかなように、「その」は 1 項名詞と結びつきやすい[6]。

4．現代語と近代語の違い

以上見たように、「その」は 1 項名詞と結びつきやすい。本発表では代行指示について考えるが[7]、そこで考えるべき問題点がある。例えば、(6) では「その」をつけないのが普通である。

(6) 実験は｛その／*この／φ｝結果が重要だ。[8]

つまり、現代語の場合、「この」は多くの場合、統語的に排除されるが[9]、「その」も文体的理由で避けられることが多く、実際には「ゼロ」が最も普通に使われるということである。

ところが、近代の文語文を見ていると、現代語よりも「その」の使用が多いことに気づく。現代語なら「ゼロ」が想定される場合に「その」が使われていることが多いということである。

本発表では、こうした直感を確かめるべく、コーパスを用いて調査を行った。具体的には、現代語としては BCCWJ を、近代語としては太陽コーパスを用いて調査を行った。

5．コーパスによる調査

本節では、今回の調査の結果を報告する。なお、今回の検索対象語は、BCCWJ における「その」の前接頻度上位 100 位の中で、「この」も「ゼロ」も可能であるものを選んだ。その結果、検索対象語は以下の 10 語となった[10]。

(7) 一部、影響、結果、原因、内容、背景、表情、方法、目的、理由（50 音順）

5.1　現代語の調査

現代語については BCCWJ を用いて検索を行った。まず、可能な限り名詞が連結されるように長単位で検索を行った。また、「ゼロ」の用例を適切に採集するために、キーの指定を行わなかった。中納言における検索条件は以下の通りである。

(8) キー：指定せず、キーから 1 語：品詞＝名詞、キーから 2 語：品詞＝助詞

その後、「キー」の部分が「の」および「連体形」であるものを Excel 2010 のフィルター機

6 このことは、「その」と「この」を比べた場合、代行指示では「その」が無標であり、指定指示では「この」が無標であることの帰結である（cf. 庵 2002, 2007, 2012）。
7 指定指示について詳しくは庵(2007, 2012)を参照されたい。
8 *はその文が非文法的であることを表す。
9 代行指示において「この」が使えるためには一定の条件を満たす必要がある。これについて詳しくは、庵(1995a, 2007)、Iori (2013)を参照されたい。
10 これ以外に、「過程、可能性、気持ち、周辺」もあるが、これらはいずれも、太陽コーパスでの用例がないか、非常に少ない（「可能性」以外は用例ゼロ）ため、調査対象外とした。

能で排除し[11]、残ったものを目視で「ゼロ」かそうでないかに振り分けた。

以上の基準で検索した結果は次の通りである。

表 2　「ゼロ、その、この」の分布（BCCWJ）

	ゼロ		その		この		合計[12]
目的	6321	91.86[13]	442	6.42	118	1.71	6881
理由	2227	55.15	1695	**41.98**	116	2.87	4038
内容	2514	63.25	1340	**33.71**	121	3.04	3975
結果	2282	57.52	1430	**36.05**	255	6.43	3967
影響	3302	87.35	433	11.46	45	1.19	3780
原因	2015	72.80	658	23.77	95	3.43	2768
背景	2187	81.09	399	14.79	111	4.12	2697
方法	1061	45.01	407	17.27	889	**37.72**	2357
一部	1529	81.20	348	18.48	6	0.32	1883
表情	1095	81.41	242	17.99	8	0.59	1345
合計	24533	**72.82**	7394	21.95	1764	5.24	33691

５.２　近代語の調査

一方、近代語の調査は太陽コーパスを用いて行った。(7)の 10 語をそれぞれ「ひまわり」で検索し、全例について、「ゼロ、その、この」の用例数を数えた。「ゼロ」の認定基準は現代語の場合と同様である。また、「その」には「その、其の、其、そが、其れが」を含め、「この」には「この、此の、此」を含めた。

表 3　「ゼロ、その、この」の分布（太陽コーパス）

	ゼロ		その		この		合計
目的	529	53.33	385	38.81	78	7.86	992
結果	114	12.06	779	**82.43**	52	5.50	945
方法	93	23.02	156	38.61	155	**38.37**	404
理由	155	41.44	189	50.53	30	8.02	374
影響	150	50.68	134	45.27	12	4.05	296
原因	94	31.86	166	56.27	35	11.86	295
内容	139	55.60	110	44.00	1	0.40	250
一部	111	49.33	106	47.11	8	3.56	225
背景	49	89.09	5	9.09	1	1.82	55
表情	17	85.00	3	15.00	0	0.00	20
合計	1451	37.63	2033	**52.72**	372	9.65	3856

11 つまり、名詞修飾成分が前接したものは「ゼロ」と見なさないということである。

12 「ゼロ、その、この」が前接するものの合計であり、「当該の名詞＋助詞」の全数ではない。

13 同じ名詞における「ゼロ、その、この」それぞれの％を表す。ここで言えば、「6881」に対する「6321」の％を表している。

6．考察

表 2、表 3 からわかるように、現代語と近代語で、1 項名詞に前接する「ゼロ」と「その」の分布には差が見られる[14]。

表 4　「ゼロ、その、この」の分布（左：BCCWJ、右：太陽コーパス）

	ゼロ		その		この		ゼロ		その		この	
目的	6321	91.86	442	6.42	118	1.71	529	53.33	385	38.81	78	7.86
結果	2282	57.52	1430	**36.05**	255	6.43	114	12.06	779	**82.43**	52	5.50
方法	1061	45.01	407	17.27	889	**37.72**	93	23.02	156	38.61	155	**38.37**
理由	2227	55.15	1695	**41.98**	116	2.87	155	41.44	189	50.53	30	8.02
影響	3302	87.35	433	11.46	45	1.19	150	50.68	134	45.27	12	4.05
原因	2015	72.80	658	23.77	95	3.43	94	31.86	166	56.27	35	11.86
内容	2514	63.25	1340	**33.71**	121	3.04	139	55.60	110	44.00	1	0.40
一部	1529	81.20	348	18.48	6	0.32	111	49.33	106	47.11	8	3.56
背景	2187	81.09	399	14.79	111	4.12	49	89.09	5	9.09	1	1.82
表情	1095	81.41	242	17.99	8	0.59	17	85.00	3	15.00	0	0.00
合計	24533	**72.82**	7394	**21.95**	1764	**5.24**	1451	**37.63**	2033	**52.72**	372	**9.65**

このことの理由は複数考えられると思われるが、その 1 つは「そ」が単独で語として使えるということである。

この点を確かめるべく、太陽コーパスで「そ」が指示詞として使われている例を数えたところ、次のようになり、「が、は、を」以外の助詞との共起例はなかった（「の」を除く）[15]。

表 5　「そ」の分布（太陽コーパス）

が	142
は	353
を	92
合計	587

それぞれの例は次の通りである。

(9) 隨て露は東洋の外交上に斟酌する所なかるべからず。例へば彼れ朝鮮に野心を逞うし、そが公使等の手を經て朝鮮の王室に畫策し、～
（『太陽』1895 年 8 号、稲垣満次郎「一大外交」）

(10) 討幕派はその意外なるに驚きぬ。そは、討幕の密勅を乞ふも、今は幕府を討つべき名なきにくるしめばなり。（『太陽』1895 年 2 号、落合直文「しら雪物語」）

(11) たとへその身吐瀉する事なくとも、傍人若しこれある時は、そを見たる人、必ずや又

---

14 表 4 の全ての名詞について、BCCWJ と太陽コーパスの「ゼロ」と「その」の値に関する 2×2 のカイ二乗検定を行った結果は次の通りである（「方法」のみ「この」を含む 2×3）。
目的：$\chi^2(1)=1057.17$、結果：$\chi^2(1)=621.79$、理由：$\chi^2(1)=17.16$、影響：$\chi^2(1)=272.96$、原因：$\chi^2(1)=178.58$、内容：$\chi^2(1)=8.65$、一部：$\chi^2(1)=103.45$、背景：$\chi^2(1)=1.11$、表情：$\chi^2(1)=0.004$（「背景、表情」は有意差なし、その他は全て 0.1%水準で有意）
方法：$\chi^2(2)=117.03$、$p<.001$ で、BCCWJ の「ゼロ」と太陽コーパスの「その」が有意に多く、前者の「その」と後者の「ゼロ」は有意に少なく、「この」には有意差がなかった。

15 「そが」の「が」は主格ではなく、属格であり、「そが」は「その」と同義で使われている。

胸わろくなりて、嘔吐をせんもはかりがたし。
（『太陽』1895 年 7 号「青山白水と旅行」）

このように、『太陽』の当時はまだ、「そ」が単独で語としての用法を持っており、そのことが、（代行指示の）「その」の使用を容易にしていたと考えられるのである。

7．おわりに

本発表では、近代語から現代語にかけての変化の例として、代行指示における限定詞の選択の問題を取り上げた。発表者は先に、大規模コーパスを用いて、近代語と現代語における漢語サ変動詞の自他の変化について考察したが（庵・張 2015）、こうした考察は、言うまでもなく、これらのコーパスが整備されてきたおかげである。改めてこれら諸コーパスの作成に携わられた関係各位に心よりお礼申し上げる。その上で、今後こうした形での実証的な研究が多く行われ、これまでの、現代語のみ、近代語のみの研究では明らかになっていないさまざまな言語事実が解明されることを期待したい。

文　献

庵　功雄(1994)「結束性の観点から見た文脈指示」『日本学報』13、pp.31-42、大阪大学
庵　功雄(1995a)「語彙的意味に基づく結束性について」『現代日本語研究』2、pp.85-102、大阪大学
庵　功雄(1995b)「コノとソノ」宮島達夫・仁田義雄編『日本語類義表現の文法（下）』pp.619-631、くろしお出版
庵　功雄(1999)「名詞句における助詞の有無と名詞句のステータスの相関についての一考察」『言語文化』35、pp.21-32、一橋大学
庵　功雄(2002)「「この」と「その」の文脈指示用法再考」『一橋大学留学生センター紀要』5、pp.5-16、一橋大学
庵　功雄(2007)『日本語研究叢書 21　日本語におけるテキストの結束性の研究』くろしお出版
庵　功雄(2012)「指示表現と結束性」澤田治美編『ひつじ意味論講座 6　意味とコンテクスト』pp.183-198、ひつじ書房
Iori, Isao (2013) “Remarks on some characteristics of nouns in Japanese”, *Hitotsubashi journal of arts and sciences*. 54-1、pp.5-18、一橋大学
庵　功雄・張　志剛(2015)「漢語サ変動詞に見る近代語と現代語」『日本語の研究』11-2、pp.86-100.
西山佑司(2003)『日本語名詞句の意味論と語用論』ひつじ書房

使用したコーパス

現代日本語書き言葉均衡コーパス（BCCWJ）、太陽コーパス

ポスター発表（2） Aグループ

9月2日（水） 13:00～14:00

『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション

植田 禎子 (日本システムアプリケーション)
飯田 龍 (情報通信研究機構)
浅原 正幸 (国立国語研究所) *
松本 裕治 (奈良先端科学技術大学院大学)
徳永 健伸 (東京工業大学)

Predicate-Argument Structure and Coreference Relation Annotation on 'Balanced Corpus of Contemporary Written Japanese'

Yoshiko Ueda (Japan System Applications Co., Ltd.)
Ryu Iida (National Institute of Information and Communications Techonology)
Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Yuji Matsumoto (Nara Institute of Science and Technology)
Takenobu Tokunaga (Tokyo Institute of Technology)

要旨

述語項構造・共参照関係は、形態論情報・係り受け構造などの基本的な情報とともに重要であり、京都大学テキストコーパスや NAIST テキストコーパスなど、さまざまな基準に基づいたアノテーションが共有され、それに基づく解析器が整備されてきた。しかしながら、アノテーション先のテキストのほとんどが新聞記事であり、ジャンル横断的な述語項構造・共参照関係アノテーションは整備されていなかった。本発表ではアノテーション基準として NAIST テキストコーパス互換の基準を採用し、『現代日本語書き言葉均衡コーパス』コアデータ全体に対して行った、述語項構造・共参照関係アノテーションについて報告する。2015 年 7 月末時点のアノテーションの統計量を示すとともに、ポスター発表においてアノテーション単位・基準や公開データ形式について紹介する。本アノテーションデータは 2015 年 9 月末に公開予定である。

1. はじめに

言語処理の分野において日本語の述語項構造や共参照関係の研究が盛んに進められ、アノテーション基準やそれに基づくアノテーションが整備されている。京都大学テキストコーパス (KTC) は、新聞記事に対して、益岡・田窪品詞体系 (益岡・田窪 (1992)) に基づく形態論情報や

* masayu-a@ninjal.ac.jp

文節係り受けとともに、補助動詞や助動詞を含めた述部出現形に対する名詞句の格関係を、格助詞をラベルとして付与している (河原ほか (2002))。さらに名詞句間の属格の情報や、共参照関係・上位/下位関係・総称/非総称関係が付与されている。同様の情報が解析済みブログコーパス (Kyoto University and NTT Blog Corpus: KNBC) に付与されている (橋本ほか (2011))。NAIST テキストコーパス (NTC) は、KTC と同じ新聞記事に述語基本形に対する格関係を格助詞をラベルとして付与している (飯田ほか (2010))。京都大学ウェブ文書リードコーパス (萩行ほか (2014)) は、ウェブ上にある多様な文書の書き始め先頭 3 文に対して、形態論情報・係り受け関係・固有表現のほか、述語項構造・共参照関係を付与している。特に新聞記事などにはあまり出現しない外界ゼロ照応の細分類について規定している。さらに、松林ほか (2014) は NTC の基準を中心に、述語項構造アノテーションの 2014 年現在の論点について整理して考察を行っている。

本稿では、2008 年より進められてきた『現代日本語書き言葉均衡コーパス』(Maekawa et al. (2014)) コアデータ全体に対する述語項構造・共参照関係アノテーション (BCCWJ-PAS)(小町・飯田 (2011)) について報告する。BCCWJ-PAS はアノテーション基準として NTC の基準を継承している。松林ほか (2014) で示されている論点について全てを反映しているわけではないが、雑誌・書籍などを含む均衡コーパスに対するジャンル横断的なアノテーションとして、2015 年 9 月に公開予定である。

以下、2 節では BCCWJ-PAS のアノテーション基準の概略を示す。3 節では 2015 年 7 月現在のアノテーションの統計を示す。4 節では 公開の形態について示し、5 節にまとめと今後の課題について示す。

2. アノテーション基準

2.1 NAIST テキストコーパス基準

表 1　ラベル一覧

	ラベル
述部	「述語」「事態」「助動詞」(「機能語相当」)
格要素	「ガ」「ヲ」「ニ」「ハ」「追加ガ/ニ」
ゼロ照応	「外界一人称」「外界二人称」「外界一般」「節照応」

BCCWJ-PAS のアノテーションコーパスのアノテーション基準は基本的に NAIST テキストコーパスの基準に準拠する。しかしながら、新聞記事以外のジャンルのテキストを扱うほか、前提となっている形態論情報が異なるために細かい点で異なっている。

まず、アノテーション対象の述部について述べる。BCCWJ に付与されている短単位形態素 1 単位の基本形を基本単位として、UniDic 品詞体系が動詞・形容詞・名詞がアノテーション対象である。動詞・形容詞に対しては「述語」タグを付与する。名詞は述語をなす事態性名詞で「名詞句＋助動詞 (“だ”)」や「節末の名詞句」を対象とし、「事態」タグを付与する。さらに、後で述べるように、格が増えるような助動詞・補助動詞には、「助動詞」タグを付与する。なお、動詞・形容詞などが機能語・機能表現をなす場合には述語と認定せず、別途「機能語相当」と

いうタグを付与する。

格関係は、述語の基本形に対する「ガ格」、「ヲ格」、「ニ格」の格助詞をラベルとして付与する。「好き」などの形状詞や「できる」などの可能動詞の場合、対象となっている要素を表す「ガ格」とは別に、提題されている要素に対して「ハ格」を付与する。

PM41_00219

私 ハ の好き 述語 な街 ガ 1

格交代によって、主動詞のガ格、ヲ格、ニ格以外に格が増える場合には、対象となる助動詞や補助動詞に「助動詞」タグを付与し、追加で導入される格に「追加ガ/ニ」を付与する。

OY01_00137

自分 追加ガ/ニ が４０歳になった頃から、税理士会・ＴＫＣ活動に参加、以後幾多の会議に出席させて頂く機会 ヲ に恵ま 述語 [a]れ 助動詞 ました。

[a] 「恵ま」のガ格は「外界一般」。

共参照関係は文節の主辞をなす名詞句を対象とする。談話内に出現した名詞句（言及）が指示的で、現実世界の実体に写像可能な表現である際に、直接照応可能な先行詞が存在する場合に付与する。総称名詞は照応詞・先行詞として考えない。

PN3b_00001 (共参照をインデックス番号で表示)

★プレゼントはビートルズ 1
米ＣＮＮテレビのプロデューサー、ウェンディ・ウィットワースさん 2 の５０歳の誕生日、夫のラルフさん 3 が“プレゼント”したのは何と、元ビートルズ 1 のポール・マッカートニー 4 のプライベートコンサート 5 だった。
誕生パーティーで夫 3 がポール 4 の登場を告げると、彼 4 の大ファンのウェンディさん 2 は思わず涙。
約１時間半のコンサート 5 の終幕近くには、ポール 4 はウェンディさん 2 をステージに上げ、ビートルズ 1 の「バースデイ」を演奏したという。
ポール 4 は、報酬の１００万ドルを反地雷の慈善団体に寄付するとしている。

ゼロ代名詞の参照の実体が談話における一人称（著者）や二人称（読者）を指す場合、それぞれ「外界一人称」「外界二人称」のタグを付与する。一人称（著者）や二人称（読者）はともに単数であることを前提とする。Yahoo! 知恵袋の場合、質問者-回答者の関係についても「外界二人称」とみなす。その他の談話中に出現する外界照応については「外界一般」のタグを付与する。ゼロ代名詞が節・文・文章などを指す場合、節末の文節の主辞に「節照応」としてタグ付けする。

OY08_00189 (外界のみ表示)

今日は「後ろのヤツ」とはしゃべら ガ：外界一人称 んかった。
まあ授業に集中し ガ：外界一人称 てましたからね。
っていうよりは英語とプログラミングが移動なんでね。
プロは考え ガ：外界一人称 てる暇すら与え ニ：外界一人称 てくれないよ。
でも今日は指の調子が良かったので、早く打ち終わる ガ：外界一人称, ニ：外界一般 ことが出来 ガ：外界一人称 ましてん。
今日は部活があったけど１時半で帰れた ガ：外界一人称, ニ：外界一般 のよ〜ん。
そいじゃ、まったあしたぁ。

表 1 にラベル一覧を示す。詳細については https://sites.google.com/site/ryuiida/ntc-annotation-scheme を参照すること。

2.2　実アノテーション作業

BCCWJ-PAS のアノテーション作業は図 Tagrin (高橋・乾 (2006))[1] を用いて行った。

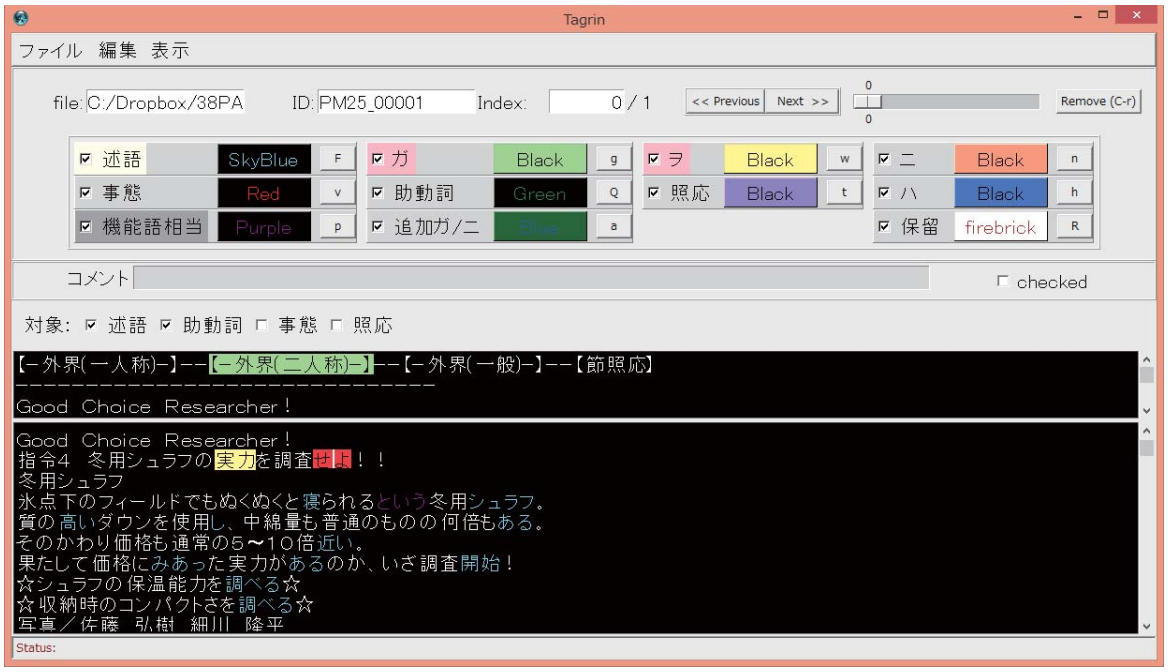


図 1　アノテーションツール Tagrin

作業においては以下の点に気をつけながら作業を進めた。

アノテーション誤りよりもアノテーション漏れが多くなるタイプのタスクであり、アノテーション漏れがないように作業者に教示した。

タグづけの範囲の例外として、複合動詞の扱いがある。複合動詞の格は基本的に前の動詞につけるが、後ろの動詞によって格が変化する場合には複合動詞全体を述語とみなし述語タグを付与する。

格要素が直接の係り受け関係にない場合は、作業者間の揺れが大きくなる傾向にある。この場合、格要素のうち直近のものを選択するように教示した。

3.　統計

本節では、2015 年 7 月時点でのアノテーションの統計を示す。残念ながら現時点で、BCCWJ DVD 1.1 版や BCCWJ-DepPara との重ね合わせが実現しておらず、形態論情報や係り受け構造との組み合わせによる集計ができていないため、限定的な情報になる。

[1] http://kagonma.org/tagrin/

3.1　述語項構造アノテーションの統計

表 2 に述部と格要素の頻度を示す。

表 2　述語と格要素の頻度

ジャンル	述部	述語数	ガ	ヲ	ニ	ハ	追加ガ/ニ
PB	述語	27027	26904	9652	5264	796	9
		100.00 %	99.54 %	35.71 %	19.48 %	2.95 %	0.03 %
PB	事態	5813	5802	2722	656	5	3
		100.00 %	99.81 %	46.83 %	11.29 %	0.09 %	0.05 %
PB	助動詞	478	8	1	3	0	468
		100.00 %	1.67 %	0.21 %	0.63 %	0.00 %	97.91 %
PN	述語	32434	32423	13035	5498	200	1
		100.00 %	99.97 %	40.19 %	16.95 %	0.62 %	0.00 %
PN	事態	14949	14947	6084	1320	1	1
		100.00 %	99.99 %	40.70 %	8.83 %	0.01 %	0.01 %
PN	助動詞	426	2	0	1	0	423
		100.00 %	0.47 %	0.00 %	0.23 %	0.00 %	99.30 %
PM	述語	25275	25272	9206	4140	56	0
		100.00 %	99.99 %	36.42 %	16.38 %	0.22 %	0.00 %
PM	事態	5039	5039	2131	410	1	0
		100.00 %	100.00 %	42.29 %	8.14 %	0.02 %	0.00 %
PM	助動詞	347	0	0	4	0	343
		100.00 %	0.00 %	0.00 %	1.15 %	0.00 %	98.85 %
OW	述語	15436	15384	6941	2388	471	0
		100.00 %	99.66 %	44.97 %	15.47 %	3.05 %	0.00 %
OW	事態	15439	15432	9048	1314	1	1
		100.00 %	99.95 %	58.60 %	8.51 %	0.01 %	0.01 %
OW	助動詞	158	0	0	0	0	158
		100.00 %	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %
OC	述語	13466	13343	4293	2501	606	1
		100.00 %	99.09 %	31.88 %	18.57 %	4.50 %	0.01 %
OC	事態	3472	3472	1756	664	4	0
		100.00 %	100.00 %	50.58 %	19.12 %	0.12 %	0.00 %
OC	助動詞	269	2	0	0	0	268
		100.00 %	0.74 %	0.00 %	0.00 %	0.00 %	99.63 %
OY	述語	12706	12703	3740	1761	138	0
		100.00 %	99.98 %	29.43 %	13.86 %	1.09 %	0.00 %
OY	事態	2997	2997	1043	222	0	0
		100.00 %	100.00 %	34.80 %	7.41 %	0.00 %	0.00 %
OY	助動詞	146	0	0	0	0	146
		100.00 %	0.00 %	0.00 %	0.00 %	0.00 %	100.00 %

ほとんどの述語・事態に対してガ格が規定されていることがわかる。述語については、PN（新聞）・OW（白書）についてはヲ格が多く、OC（Yahoo! 知恵袋）・OY（Yahoo! ブログ）はヲ格が少ない。事態は述語に比してヲ格が多い傾向にある。また、OW は事態が多い傾向にある。ほとんどの助動詞に対して追加ガ/ニが規定されている。

表 3 に述語のジャンルごとの項の組み合わせの分布について示す。3 項（ガヲニ）は OC が若干多い傾向が見られた。2 項（ガヲ）は、PN・OW が多く、OC・OY が少ない。1 項（ガの

表 3　項の組み合わせの分布 (述語)

ジャンル	ガヲニ	ガヲ	ガニ	ヲニ	ガ	ヲ	ニ
PB	1516	8075	3643	6	12904	39	47
	5.61 %	29.88 %	13.48 %	0.02 %	47.74 %	0.14 %	0.17 %
PN	1641	11373	3840	3	15369	6	2
	5.06 %	35.07 %	11.84 %	0.01 %	47.39 %	0.02 %	0.01 %
PM	1324	7878	2812	1	13202	2	
	5.24 %	31.17 %	11.13 %	0.00 %	52.23 %	0.01 %	0.00 %
OW	649	6271	1697		6310	17	21
	4.20 %	40.63 %	10.99 %	0.00 %	40.88 %	0.11 %	0.14 %
OC	890	3382	1553		6986	19	29
	6.61 %	25.12 %	11.53 %	0.00 %	51.88 %	0.14 %	0.22 %
OY	404	3328	1351	1	7482	1	1
	3.16 %	26.01 %	10.56 %	0.01 %	58.47 %	0.01 %	0.01 %
ジャンル	ガヲニハ	ガヲハ	ガニハ	ガハ	ニハ	ハ	*
PB	8	2	35	713	1	29	9
	0.03 %	0.01 %	0.13 %	2.64 %	0.00 %	0.11 %	0.03 %
PN	4	7	7	181			1
	0.01 %	0.02 %	0.02 %	0.56 %	0.00 %	0.00 %	0.00 %
PM		1	3	52			
	0.00 %	0.00 %	0.01 %	0.21 %	0.00 %	0.00 %	0.00 %
OW	1	3	19	434	1	13	
	0.01 %	0.02 %	0.12 %	2.81 %	0.01 %	0.08 %	0.00 %
OC		2	27	502	2	73	1
	0.00 %	0.01 %	0.20 %	3.73 %	0.01 %	0.54 %	0.01 %
OY		6	4	128			
	0.00 %	0.05 %	0.03 %	1.00 %	0.00 %	0.00 %	0.00 %

* は「追加ガ/ニを含むその他の組合せ」。

み）は OY が多い。また二重主語である 2 項（ガハ）は PB・OW・OC が多く、PN・PM が少ない。

表 4　項の組み合わせの分布 (事態)

ジャンル	ガヲニ	ガヲ	ガニ	ガ	ヲ	ニ	ガハ	*
PB	234	2479	414	2670	6	5	2	3
	4.03 %	42.65 %	7.12 %	45.93 %	0.10 %	0.09 %	0.03 %	0.05 %
PN	203	5879	1117	7746	2		1	1
	1.36 %	39.33 %	7.47 %	51.82 %	0.01 %		0.01 %	0.01 %
PM	90	2041	320	2587			1	
	1.79 %	40.50 %	6.35 %	51.34 %			0.02 %	
OW	511	8531	802	5586	6	1	1	1
	3.31 %	55.26 %	5.19 %	36.18 %	0.04 %	0.01 %	0.01 %	0.01 %
OC	243	1512	420	1293			3	
	7.00 %	43.55 %	12.10 %	37.24 %			0.09 %	
OY	49	994	173	1781				
	1.63 %	33.17 %	5.77 %	59.43 %				

* は「追加ガ/ニを含むその他の組合せ」。

表 4 に事態のジャンルごとの項の組み合わせの分布について示す。3 項（ガヲニ）は OC・PB が多い傾向にある。OC は他に 2 項（ガニ）が多く、1 項（ガ）が少ない。OW は 2 項（ガ

ヲ）が多く、1 項（ガ）が少ない。一方、PN・PM・OY は 2 項（ガヲ）が多く、1 項（ガ）が多い。

表 5　項の組み合わせの分布 (助動詞)

ジャンル	ガヲニ	ガ	ニ	追加ガ/ニ	*
PB	1	7	2	468	3
	0.21 %	1.46 %	0.42 %	97.91 %	0.63 %
PN		2	1	423	
			0.23 %	99.30 %	
PM			4	343	
			1.15 %	98.85 %	
OW				158	
				100.00 %	
OC		1		267	1
		0.37 %		99.26 %	0.37 %
OY				146	
				100.00 %	

* は「追加ガ/ニを含むその他の組合せ」。

表 5 に助動詞のジャンルごとの項の組み合わせの分布について示す。ほとんどが「追加ガ/ニ」である。

3.2　外界ゼロ照応・節照応の分布

表 6 にガ格の外界ゼロ照応・節照応の分布について示す。

まず、紙媒体 (PB・PN・PM・OW) には「外界一人称」・「外界二人称」があまり出現しない。紙媒体の中では、OW には「外界一人称」・「外界二人称」がともにほとんど出現しない一方、PM は読者を意識した「外界二人称」が出現することがわかった。Web 媒体 (OC・OY) は「外界一人称」・「外界二人称」が多く出現することがわかった。質問者-回答者間のインタラクションが発生する OC は「外界二人称」が多く、個人の経験を記述する OY は特に「外界一人称」が多かった。

OW には「外界一般」が多く出現した。これは、OW において、組織・集団を意識した記述が多く、個人を意識した記述が少ないためだと考えられる。報道を目的とする PN は外界ゼロ照応が少ないことがわかった。基本的に文書内に閉じた空間になっていることがわかる。

3.3　共参照関係

本節では共参照関係の統計について示す。

同一の実体を参照する言及間の共参照関係を言及の数[(2)]と実体の数[(3)]で集計する。

表 7 に集計結果を示す。「言及/実体」は実体の言及数の平均値である。

サンプル中の形態素数が大きい PB・OW は 1 つの実体の言及数が多くなる傾向になる一方、サンプル中の形態素数が小さい OC は 1 つの実体の言及数が少なくなる。PN・PM は 1 つのサンプルに複数の記事が含まれていることから、実体の言及数が中程度になっていると考える。

(2) 共参照関係を同値関係とみなした場合の同値類の要素ののべ。
(3) 共参照関係を同値関係とみなした場合の同値類の異なり。

表 6　外界ゼロの分布 (ガ格)

		述語数	ガ	外界一人称	外界二人称	外界一般	節照応
PB	述語	27270	26904	212	88	3384	4
			100.0 %	0.8 %	0.3 %	12.6 %	0.0 %
PB	事態	5835	5802	16	17	1955	2
			100.0 %	0.3 %	0.3 %	33.7 %	0.0 %
PB	助動詞	1865	8	0	0	1	0
PN	述語	32712	32423	298	178	4013	4
			100.0 %	0.9 %	0.5 %	12.4 %	0.0 %
PN	事態	15204	14947	31	95	3921	15
			100.0 %	0.2 %	0.6 %	26.2 %	0.1 %
PN	助動詞	426	2	1	0	0	0
PM	述語	25421	25272	452	289	3686	17
			100.0 %	1.8 %	1.1 %	14.6 %	0.1 %
PM	事態	5141	5039	38	118	1410	6
			100.0 %	0.8 %	2.3 %	28.0 %	0.1 %
PM	助動詞	347	0	0	0	0	0
OW	述語	15654	15384	75	9	4438	0
			100.0 %	0.5 %	0.1 %	28.8 %	0.0 %
OW	事態	15475	15432	12	0	7383	2
			100.0 %	0.1 %	0.0 %	47.8 %	0.0 %
OW	助動詞	260	0	0	0	0	0
OC	述語	13754	13343	1724	1137	1274	3
			100.0 %	12.9 %	8.5 %	9.5 %	0.0 %
OC	事態	3482	3472	406	289	1070	0
			100.0 %	11.7 %	8.3 %	30.8 %	0.0 %
OC	助動詞	1131	2	1	0	0	0
OY	述語	12807	12703	2528	314	1462	5
			100.0 %	19.9 %	2.5 %	11.5 %	0.0 %
OY	事態	3035	2997	534	97	932	0
			100.0 %	17.8 %	3.2 %	31.1 %	0.0 %
OY	助動詞	146	0	0	0	0	0

表 7　共参照関係の集計（言及と実体）

ジャンル	言及	実体	言及/実体
PB	9459	2237	4.33
PN	16476	4445	3.76
PM	9866	2658	3.82
OW	7169	1661	4.47
OC	2468	994	2.49
OY	3506	1106	3.23

表 8 に実体の言及数[(4)]の分布を示す。OC において、実体の言及数 2 が全体の 72% を占めている。一方、どのサンプルでも 11 回以上言及される実体が数 % あることが確認された。

[(4)] 実体がはる同値類を構成する要素数。

表 8　共参照関係の集計（実体の言及数）

ジャンル	実体の言及数分布							
	2	3	4	5	6..10	11..15	16..20	21..
PB	1278	270	135	92	170	55	50	73
	57.13 %	12.07 %	6.03 %	4.11 %	7.60 %	2.46 %	2.24 %	3.26 %
PN	2232	579	617	257	337	70	36	52
	50.21 %	13.03 %	13.88 %	5.78 %	7.58 %	1.57 %	0.81 %	1.17 %
PM	1362	426	251	153	173	57	36	46
	51.24 %	16.03 %	9.44 %	5.76 %	6.51 %	2.14 %	1.35 %	1.73 %
OW	822	228	163	81	124	66	35	54
	49.49 %	13.73 %	9.81 %	4.88 %	7.47 %	3.97 %	2.11 %	3.25 %
OC	717	166	51	30	16	11	16	21
	72.13 %	16.70 %	5.13 %	3.02 %	1.61 %	1.11 %	1.61 %	2.11 %
OY	638	139	133	69	53	23	16	25
	57.69 %	12.57 %	12.03 %	6.24 %	4.79 %	2.08 %	1.45 %	2.26 %

3.4　機能語相当の統計

表 9 に「機能語相当」タグの統計を示す。機能語相当の表現は OW・PB が多く、PM・OY が少ないことがわかる。

表 9　機能語相当表現の統計

ジャンル	PB	PN	PM	OW	OC	OY
機能語相当	3031	1889	933	5179	1000	428

4.　公開形態

本節では BCCWJ-PAS の公開形態について示す。

他データとの重ね合わせとして、まず BCCWJ-DVD 版 (Version 1.1)(国立国語研究所 (2015)) の形態論情報との統合を行う。さらに係り受け・並列構造アノテーションである BCCWJ-DepPara (浅原・松本 (2013)) との統合を行い、同データに付与されている文境界情報 (小西ほか (2013)) を統合する。さらに、受身に関する格交代の情報を統合するために、れる・られるの用法 (小山田ほか (2012)) との統合を行う。

公開するファイル形式は、2015 年 9 月の時点では NTC に相当するファイル形式での配布を検討する。形態論情報、係り受け、並列構造、れる・られるの用法は、拡張 CaboCha フォーマット (松吉ほか (2014)) に準拠するが、述語項構造、共参照関係は NTC 互換の形式にとどめる。

5.　おわりに

本稿では 2015 年 9 月に公開を予定している『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション BCCWJ-PAS について紹介した。

本稿執筆時点 (2015 年 7 月) では、形態論情報や係り受けなどのアノテーションと統合がさ

れていないために、アノテーションデータの分析に制限がある。今後の課題として、4 節に示した公開形態への整形作業とともに、述語の品詞ごとの分析や係り受け・文節境界・文境界との位置関係による分析がある。また、並行して進めている節境界との統合も行いたい。

謝辞

8 年間にわたる述語項構造アノテーションに携わったすべてのみなさまに感謝の意を表します。本研究の一部は科研費特定研究「書き言葉コーパスの自動アノテーションの研究」(18061005)、科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」(25284083)、科研費萌芽「近代語コーパスに対する統語情報アノテーション基準策定」(15K12888)、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

浅原正幸・松本裕治 (2013).「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」　言語処理学会第 19 回年次大会発表論文集.

萩行正嗣・河原大輔・黒橋禎夫 (2014).「多様な文書の書き始めに対する意味関係タグ付きコーパスの構築とその分析」　自然言語処理, 21:2, pp. 213–247.

橋本力・黒橋禎夫・河原大輔・新里圭司・永田昌明 (2011).「構文・照応・評価情報つきブログコーパスの構築」　自然言語処理, 18:2, pp. 175–201.

飯田龍・小町守・井之上直也・乾健太郎・松本裕治 (2010).「述語項構造と照応関係のアノテーション: NAIST テキストコーパス構築の経験から」　自然言語処理, 17:2, pp. 25–50.

河原大輔・黒橋禎夫・橋田浩一 (2002).「「関係」タグ付きコーパスの作成」　言語処理学会第 8 回年次大会発表 論文集, pp. 495–498.

国立国語研究所コーパス開発センター (2015).『『現代日本語書き言葉均衡コーパス』利用の手引き 第 1.1 版』, http://pj.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual_01.pdf.

小町守・飯田龍 (2011).「BCCWJ に対する述語項構造と照応関係のアノテーション」　『現代日本語書き言葉均衡コーパス』完成記念講演会.

小西光・小山田由紀・浅原正幸・柏野和佳子・前川喜久雄 (2013).「BCCWJ 係り受け関係アノテーション付与のための文境界再認定」　第 4 回コーパス日本語学ワークショップ予稿集, pp. 135–142.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

益岡隆志・田窪行則 (1992).『基礎日本語文法–改訂版–』　くろしお出版.

松林優一郎・飯田龍・笹野遼平・横野光・松吉俊・藤田篤・宮尾祐介・乾健太郎 (2014).「日本語文章に対する述語項構造アノテーション仕様の考察」　自然言語処理, 21:2, pp. 333–377.

松吉俊・浅原正幸・飯田龍・森田敏生 (2014).「拡張 CaboCha フォーマットの仕様拡張」　第 5 回コーパス日本語学ワークショップ, pp. 223–232.

小山田由紀・柏野和佳子・前川喜久雄 (2012).「助動詞レル・ラレルへの意味アノテーション作業経過報告」　第 2 回コーパス日本語学ワークショップ予稿集.

高橋哲郎・乾健太郎 (2006).「アノテーションツール”Tagrin”の紹介」　言語処理学会第 12 回年次大会発表論文集, pp. 228–231.

職場における談話の修辞機能と脱文脈化の観点からの分析

田中弥生（神奈川大学外国語学部・国立国語研究所理論・構造研究系）†

Discourse Analysis of Business Communication in Terms of Rhetorical Functions and the Degree of De-contexturisation

Yayoi TANAKA (Kanagawa University, National Institute for Japanese Language and Linguistics)

要旨

選択体系機能言語理論における談話分析手法の一つである修辞ユニット分析（Rhetorical Unit Analysis）によって、職場における談話の分析を試みた。『合本　女性のことば・男性のことば（職場編）』を資料として、「会議」場面における談話を修辞機能と脱文脈化程度の観点から、その出現および展開の様子を確認し、「会議」における談話の特徴をとらえることを試みた。その結果、「会議」の下位分類である「打合せ」と「雑談」の特徴をとらえられることがわかり、「会議」の下位分類を設定する際の指標となりえる可能性がうかがえた。また、話し言葉に RUA を適用する際の課題についても検討した。

１．はじめに

談話の分析に用いられる手法には様々なものがある。修辞ユニット分析（Rhetorical Unit Analysis　以下、RUA）は選択体系機能言語理論において用いられる談話分析手法のひとつで、バフチンの chronotope の概念（1981）である空間と時間の融合が言語テクストにどのように示されているかをとらえ、脱文脈化言語(de-contextualised language)・文脈化言語(contextualised language)[1]の相違を捉える枠組みとして知られている(Cloran, 1994, 1999, 2010)。テクストの意味単位を特定するための手法（佐野 2010b）だが、その過程において発話機能(speech function)、中核要素(central entity)、現象定位(event orientation)の 3 つをメッセージ単位で認定することで、修辞機能(rhetorical function)の種類を特定し、その結果として脱文脈化の程度(degree of de-contextualisation)を知ることができる。母子会話の他、学校における教師と生徒の説明的な談話の様相を示し（Cloran 1999，2010）、日本語への適用については佐野・小磯（2011）によって検討され、英語と日本語の言語の違いに関わる修正が加えられている。その後、専門性の低い作文を高い作文に修正する RUA を用いた指導の説明（佐野 2010b）、インターネット上の Q&A サイト「Yahoo!知恵袋」やクチコミサイトを対象とした分析（田中・佐野 2011a、2011b、2011c、田中 2011、2013a、2013b）などが進められている。しかし、日本語話し言葉への RUA の適用はまだ進んでいない。本研究では、RUA の分析手法を用いて日本語の話し言葉を分析する試みとして、「会議」における談話の修辞機能と脱文脈化程度の特徴を明らかにし、また日本語話し言葉における RUA 分析適用における課題について検討する。以下、２で分析方法、３で分析結果と考察、４でまとめと今後の課題を述べる。

† yayoi@ninjal.ac.jp

[1] Cloran（1999）に基づき、脱文脈化言語を「一般化された要素の習慣的・恒久的な行動や状態について表現する言語」、文脈化言語を「物質的状況に存在する要素の現在の行動や状況について表現する言語」とする。

２．分析方法

２．１．分析対象

本研究で使用する談話資料『合本　女性のことば・男性のことば（職場編)』では、「場面１」（収録が行われた場面。「朝」「会議」「休憩」）と、「場面２」（談話の場面や具体的な場面情報）が付与されている。本研究では、「会議」の中で「場面２」に「打合せ」と「雑談」の両方をもつ協力者のデータを用いて、話し言葉への RUA の適用を検討するとともに、「打合せ」と「雑談」の修辞機能と脱文脈化程度の特徴を明らかにすることを試みる。

表 1 「女性のことば」　「会議」の「場面２」内訳

協力者 / 場面2	03	04	05	06	07	08	09	10	11	12	13	14	15	17	総計
《その他》				1											1
《不明》								13							13
挨拶						3		4							7
挨拶(電話)								5							5
休憩時雑談						1									1
検討会										44					44
雑談			14					107							121
取引先との電話折衝											108				108
小会議		322			54		219						257	68	920
相談						43		12							55
打合せ			189	137		141		267	87	58	89	151			1,119
打合せ(電話)								5							5
大会議	249														249
電話引き継ぎ											11				11
電話取り次ぎ											1				1
電話取り次ぎ(電話)											3				3
総計	249	322	203	138	54	188	219	413	87	102	212	151	257	68	2,663

表 2　「男性のことば」　「会議」の「場面２」内訳

協力者 / 場面2	01	02	03	04	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	総計
コンピュータの操作方法の相談と説明												259									259
応対																	10	28			38
会議		129				167				192						73					561
客との応対	11																				11
研究室会議					182																182
雑談	63								79					17	105					117	381
仕事(応対)																		97			97
仕事(打合せ)																		34			34
指導																			11		11
出張報告											89										89
打合せ	87		124	156				12						158	56			24	202	198	1,017
打合せ(商談)									140												140
打合せ(説明)																	194				194
電話																		40			40
反省会													64								64
報告							80	117													197
総計	161	129	124	156	182	167	80	129	219	192	89	259	64	175	161	73	204	223	213	315	3,315

表１に網掛けで示した「女性のことば」の「協力者１０」、「男性のことば」の「協力者０１」「協力者２１」を分析対象とする。「女性」の「協力者０５」と「男性」の「協力者１５」は「雑談」が少ないため除外した。また、当該資料は、文字起こしデータが提供さ

れており、発話内容が不明瞭な部分は「＃」によって示されているが、「男性」の「協力者１６」は「＃」出現率が 21.1%（161 行中 34 行）で分析不能な部分が多いため、除外した[2]。

当該資料では、「朝、職場についてから 1 時間、会議打ち合わせなどの時、１時間、休憩時間 1 時間、の計 3 時間の録音をお願いした。そのうち、資料としては、処理の際の量を考えて、それぞれ 1 時間の録音の中の、まとまった談話のある 10 分前後を取り扱うことにした」（「女性のことば」p.9、「男性のことば」p.9）とあり、必ずしも「会議」の開始から終了までが提供されているわけではなく、「会議」の開始から終了までの展開をとらえることはできない。しかし、実際の談話の場面である「会議」の修辞機能と脱文脈化程度をどのように分析できるかを検討する。当該話資料では、「基本的に１文を１レコード（＝１行）とし」「ただしここでは、「あっ。」とだけ言って直後に沈黙を伴ったり、発話者の交代が生じるものなども 1 文扱いにしている。」（「女性のことば」p.20、「男性のことば」p.20,21）とされている。しかし、｛うん　Inf(女)｝のような形で他者の発話に埋め込まれている部分もあり、談話資料の行数を文数や発話単位として扱う場合には配慮が必要であると考えられる。

２．２．分析対象のメッセージの認定と種類の認定

RUA では、「メッセージ」を基本的な分析の単位とする。メッセージは原則として節を最小単位として表わされるものと捉える。RUA による修辞機能の特定と脱文脈化程度の確認の手順は、１．メッセージとその種類の認定、２．発話機能・中核要素・現象定位の認定、３．修辞機能の特定と脱文脈化指数の確認、である[3]。まず、分析対象であるテクストをメッセージ単位に分割(segment)する。対話をデータとする場合、ポーズ等や他者のあいづち、あるいは共話のために分割された行を、統合して 1 つのメッセージと認定する場合もある。主部や述部が省略されていると考えられる場合には補足してメッセージへの分割、統合を行う。メッセージは、「位置づけ positioning」、「拘束 bound」、「自由 free」に分類する。「位置づけ」は挨拶・定型句・フィラーなど述部を含まない節のみによって構成されるもので、この後の認定対象とはしない。「自由」は独立して時制やムードなどを表わすもので認定対象となる。(1)ではメッセージ単位で(a)から記号付けをし、メッセージの種類を「　」内に付与している。

(1) (a)　今日、議事担当課長会があるから。　10A5588[4]
　　３時からね。　10A5589「自由」
　(b)　ここに、予定がはいってるけど。　10A5590「自由」
　(c)　予定表もらってあるーんでしょ↑　10A5591「自由」
　(d)　え↑　10C5592「位置づけ」
　(e)　ある↑、★新しいの。　10A5593「自由」
　(f)　→どっかいっちゃった。←　10C5594「自由」
　(g)　えっ。＜笑い＞　10A5595「位置づけ」

---

[2] 分析対象資料の「＃」出現率は、「女性」の「協力者１０」3.4%、「男性」の「協力者０１」6.8%「協力者２１」10.2%である。

[3] 各種認定および用語は原則として佐野(2010a)、佐野・小磯（2011）に依った。

[4] 行末に、協力者番号、発話者記号、行番号の順に示している。

(h) このへん、どっか、おいといたはず。　10A5596「自由」
(i) ううん。　10A5597「位置づけ」
(j) なるべくねー、転記するようにしてんだよ、　「自由」
(k) ｛うん　Inf(女)｝　「位置づけ」
(l) ああ、書いてある、　「自由」
(m) ★議事課長会、書いてある。　10C5598「自由」

「拘束」は「拘束;意味的従属」と「拘束;形式的従属」に分類する。「拘束;意味的従属」は従属するメッセージの状況（時間・場所・原因・結果・条件等）を説明するもので、従属するメッセージの一部と考えられる。(2)の(a)(c)が該当し、単独ではこの後の認定は行わないが、従属するメッセージ (d) ((b)の「位置づけ」は除外するため) とともに認定を行う。「拘束;形式的従属」は意味的には並列の関係だが時制（過去）などの側面で従属するメッセージに形式的に依存するもので、(3)の(b)(c)が該当する。「拘束;形式的従属」はこの後の認定を行う。

(2) (a) 頭数（あたまかず）増やせばー　「拘束;意味的従属」
(b) ｛そうねー　（２１D)｝、　「位置づけ」
(c) あんまり今の値段と変わらず＜笑いながら＞　「拘束;意味的従属」
(d) でかい部屋が使えるんじゃないかなー、ってゆうのが。21A10931「自由」

(3) (a)で、最初にーお茶ーをだしてー、　「拘束;形式的従属」
(b) でー、もう少ししたら、　「拘束;意味的従属」
(c) そう、４０分か５０分たったら、　「拘束;意味的従属」
(d) 珈琲と、あと、ケーキかなんかで。＜笑い・複＞　10A5548「自由」

２．３．発話機能の認定

発話機能は、「提言 proposal」か「命題 proposition」に分類する。「提言」は表 8 の(a)の品物・行為の交換（提供あるいは要求）に関するメッセージ、「命題」は(b)の情報の交換に関するメッセージが該当する。前掲の(2)及び(3)で取り上げたメッセージはすべて情報の交換で「命題」である。

表 3 発話機能(Halliday & Matthiessen 2004：107)

role in exchange	commodity exchanged	
	(a)goods&service	(b)information
(i)giving	“offer” would you like this teapot?	“statement” he’s giving her the teapot
(ii)demanding	“command” give me that teapot!	“question” what is he giving her?
	提言	命題

(4) とりあえず、曲、ある人は持ってきてくださーい。　21B10786「自由」

(4)は「持ってくる」という行為を要求しており発話機能は「提言」である。発話機能が「提言」のメッセージは、この後の中核要素および現象定位の認定を待たず、修辞機能は「行動」、脱文脈化指数は[1]と特定される。発話機能が「命題」のメッセージについて、この後、中核要素と現象定位の認定を行う。

２．４．中核要素の認定

中核要素はメッセージの中心となるものがコミュニケーションの場面に存在するか否かによって特定する。基本的には主語によって表現されるが、照応など前後のメッセージを用いて判断する場合もある。また、「このカレーは野菜がたっぷりだ」のように、述部「野菜がたっぷりだ」が「このカレー」の性質を表している場合には、「このカレー」を中核要素と認定する。中核要素の分類を図１に示す。

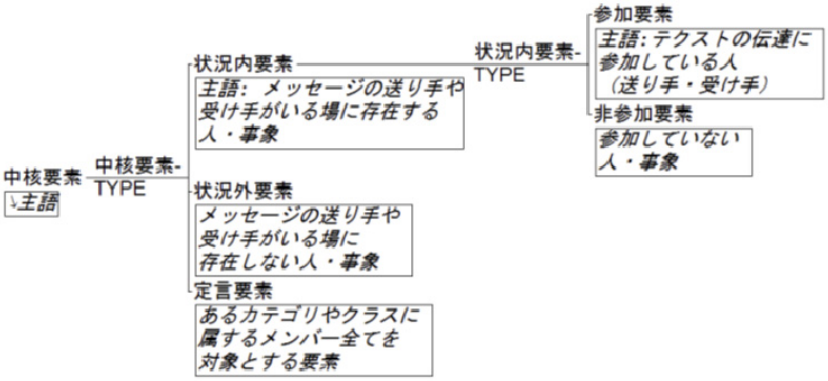


図 1 中核要素の分類（佐野・小磯 2011）

中核要素はまず「状況内要素 co-present entity」「状況外要素 absent entity」「定言要素 generalised entity」のいずれかに分類し、「状況内要素」はさらに「参加要素 interactants」「非参加要素 non-interacting entity」に分類する。なお、(5)(6)(7)にメッセージ単位で、中核要素及び現象定位の認定と、その修辞機能、脱文脈化指数の特定を示した。

２．４．１．状況内要素

主語が「メッセージの送り手や受け手がいる場に存在する人・事象」である場合に「状況内要素」と認定され、さらにその伝達に参加している人を「参加要素」、伝達には参加していない人・事象を「非参加要素」と認定する「参加要素」は、基本的には一人称、二人称が該当し、典型的なものは「私は」である。(5)では、(c)で「あなたは」、(h)と(j)で「私は」がそれぞれ省略されていると考え、「状況内;参加要素」と認定する。 (a)の「議事担当課長会が」や(b)の「予定が」は、その打ち合わせの場にある予定表に記載されている事象で、尚且つ発話主体ではないため「状況内;非参加要素」と認定する。

２．４．２．状況外要素

(6)では、「あたしのいとこが」が、その場に存在しない人であるため、【状況外要素】と認定する。

２．４．３．定言要素

「定言要素」は、「あるカテゴリやクラスに属するメンバー全てを対象とする要素」で、例えば「醤油は大豆からできている」の「醤油は」は【定言要素】である。

２．５．現象定位の認定

現象定位は，メッセージによって表現されている出来事がいつ起こったかを、メッセージが伝達されている時（Time of speaking 以下、Ts）を基準とした時間的な位置を特定して示す要素である。副詞や述部から判断する。現象定位の分類を図 2 に示す。

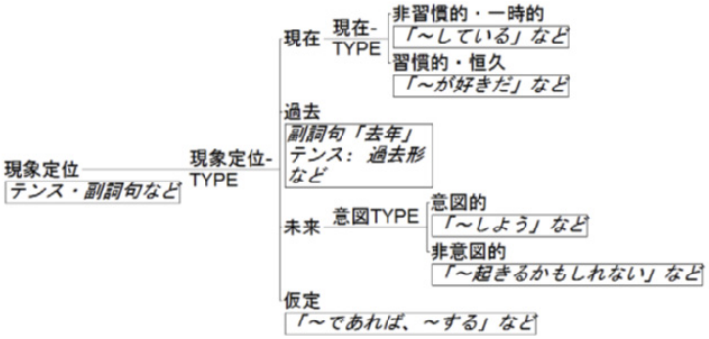


図 2 現象定位の分類（佐野・小磯 2011）

２．５．１．現在

メッセージで述べていることが Ts において起こっていて、習慣性や恒久性について述べている場合には、「現在;習慣的・恒久」と認定する。(5)の(j)は「～することにしている」と習慣を述べている。一方、メッセージで述べていることが Ts において起こっていて、一時的なもの、非習慣的なものは「現在;非習慣的・一時的」と認定する。(7)の(b)などが該当する。

２．５．２．過去

Ts より前に起こったことを述べているメッセージの現象定位は「過去」と認定する。(6)の(f)や(h)が該当する。

２．５．３．未来

Ts では起こっていないことを述べるメッセージの現象定位は「未来」あるいは「仮定」である。「未来」はその行動・現象が意図できるかできないかによって「意図的」と「非意図的」の 2 つに分類される。(6)の「上京する」は主語である「いとこ」が意図できることであるため、「意図的」、(5)(a)は「3 時」という未来に起こる会議はすでに決まった予定であり「非意図的」と認定する。

２．５．４．仮定

「仮定」は、「A が生じた場合、B が起こる」という因果関係を持つものが該当する。(7)では、(a)の「頭数ふやす」ということが生じれば、(c)(d)が起こる、という因果関係にある。

(5) (a) 今日、**議事担当課長会が**[5]*ある*[6]から。　10A5588
*3時から*ね。　10A5589「自由」
【命題＋状況内;非参加＋未来;非意図的⇒状況内予想[5]】
(b) ここに、予定がはいってるけど。　10A5590「自由」
【命題＋状況内;非参加＋現在;非習慣・一時的⇒実況[2]】
(c) （φ[7]＝あなたは）予定表もらってあるーんでしょ↑　10A5591「自由」
【命題＋状況内;参加＋現在;非習慣・一時的⇒実況[2]】
(d) え↑　10C5592「位置づけ」
(e) ある↑、★新しいの。　10A5593「自由」
【命題＋状況内;非参加＋現在;非習慣・一時的⇒実況[2]】
(f) →どっかいっちゃった。←　10C5594「自由」
【命題＋状況内;非参加＋過去⇒状況内回想[3]】
(g) えっ。＜笑い＞　10A5595「位置づけ」
(h) このへん、（φ＝私は）どっか、おいといたはず。　10A5596「自由」
【命題＋状況内;参加＋過去⇒状況内回想[3]】
(i) ううん。　10A5597「位置づけ」
(j) なるべくねー、（φ＝私は）転記するようにしてんだよ、　「自由」
【命題＋状況内;参加＋現在;習慣的・恒久⇒自己記述[7]】
(k) ｛うん　Inf(女)｝　「位置づけ」
(l) ああ、書いてある、　「自由」
【命題＋状況内;非参加＋現在;非習慣・一時的⇒実況[2]】
(m) ★議事課長会、書いてある。　10C5598「自由」
【命題＋状況内;非参加＋現在;非習慣・一時的⇒実況[2]】

(6) あのねー、今度ねー、**あたしのいとこが**ねー、*今度上京すんのねー*。
21B10983「自由」
【命題＋状況外＋未来;意図的⇒予測[11]】

(7) (a) 頭数（あたまかず）増やせばー　「拘束;意味的従属」
(b) ｛そうねー　（２１D)｝、　「位置づけ」
(c) あんまり今の値段と変わらず＜笑いながら＞　「拘束;意味的従属」
(d) でかい部屋が使えるんじゃないかなー、ってゆうのが。21A10931「自由」
【命題＋状況内;参加＋仮定⇒状況内推測[6]】

２．６．修辞機能の特定と脱文脈化指数の確認

表 4 に示したように、発話機能と中核要素と現象定位の組み合わせによって修辞機能が特定される。脱文脈化指数とは、中核要素の here（発話地点との空間的な距離）の程度と現象定位の now（発話時点との時間的な距離）の程度によって、近いものから遠いものまで修

5 中核要素は太字で示す。
6 現象定位の根拠となる部分をイタリックで示す。
7 省略されているものを復元するときはφ＝　で示す。

辞機能を線上に示した際の指数で、1 から 14 まである（図 3）。脱文脈化指数の数値が大きいものほど脱文脈化の程度が高く一般的・汎用的で、小さいものほど脱文脈化の程度が低く個人的・限定的であることを示す。

表 4. 修辞機能の特定と脱文脈化指数[8]

<table>
<tr><td colspan="4"></td><td colspan="6">発話機能</td></tr>
<tr><td colspan="4" rowspan="3"></td><td rowspan="3">提言</td><td colspan="5">命題</td></tr>
<tr><td colspan="5">現象定位</td></tr>
<tr><td colspan="2">現在</td><td rowspan="2">過去</td><td colspan="2">未来</td><td rowspan="2">仮定</td></tr>
<tr><td>非習慣的
一時的</td><td>習慣的
恒久</td><td>意図</td><td>非意図</td></tr>
<tr><td rowspan="4">中核要素</td><td rowspan="2">状況内</td><td>参加</td><td>[1]行動</td><td rowspan="2">[2]実況</td><td>[7]自己記述</td><td rowspan="2">[3]状況内
回想</td><td>[4]計画</td><td rowspan="2" colspan="2">[6]状況内
推測</td></tr>
<tr><td>非参加</td><td rowspan="3">n/a</td><td>[8]観測</td><td colspan="2">[5]状況内予想</td></tr>
<tr><td colspan="2">状況外</td><td>[9]報告</td><td>[13]説明</td><td rowspan="2">[10]状況外
回想</td><td colspan="2" rowspan="2">[11]予測</td><td rowspan="2">[12]推量</td></tr>
<tr><td colspan="2">定言</td><td>n/a</td><td>[14]一般化</td></tr>
</table>

「n/a」は該当なし／背景が灰色の部分が修辞機能の種類／[ ]内は脱文脈化指数

図 3　修辞機能と脱文脈化程度

３．分析結果と考察

前掲の(5)は、「場面２」が「打合せ」の談話の一部である。スケジュール確認【状況内予想[5]】が行われ、付随して、【実況[2]】【状況内回想[3]】【自己記述[7]】などが表れていた。同じ協力者の「打合せ」では、脱文脈化程度の高いものから低いものまで修辞機能が用いられているのに対して、「雑談」では、【報告[9]】が 5 割以上を占め、他には【実況[2]】、【状況外回想[10]】などが用いられている。「打合せ」で幅広い修辞機能が用いられ、「雑談」ではいくつかに限定される傾向は、「男性のことば」の「協力者０１」及び「協力者２１」のデータでも同様に見られた。これは、「打合せ」はその目的によって「伝達」や「報告」など、主となる修辞機能があり、そこからその場のやりとりの中でさまざまな修辞機能が用いられるのに対し、「雑談」では限定的になるためではないかと考えられる。

対話データに RUA を適用するにあたり、共話を考慮する必要があると考えるが、その判断がつきにくい部分の扱いについて、検討が必要である。たとえば、(8)の(e)は音声があれば判断がつく可能性もあるが、「嫌だ」と述べているのか、フィラーの「いや」なのか、(h)へ続く発話なのか、文字と文脈からは判断がつきにくい。話し言葉を分析する際の基準を明確にしていく必要があるだろう。

(8) (a)ライブ参加、みたいな話はしたのね。　２１Ｂ10890

[8] 佐野（2010b）および佐野・小磯（2011）の修辞機能の特定表に脱文脈化指数を合わせて示したもの

(b) いいなー、混ざりたいなー、とかって。　２１Ｂ10891
(c) え、１回ぐらい出れば、みたいな。　２１Ｂ10892
(d) ぜんぜん、かまわない。　２１Ｂ10893
(e) ただ、なんとなくー、いや＜笑い＞。　２１Ａ10894
(f) てゆうか、い、いんだよ別に。　２１Ｂ10895
(g) だって、別にー、そうゆうなんか、いやだとかじゃなくてー、★やり、［名前］がやりやすいほうのがいいんだからー。　２１Ｂ10896
(h) →やりにくいなー。←　２１Ａ10897
(i) 気分的にちょっとなー、ってんだったらもー。　２１Ａ10898
(j) ぜんぜん、それはそれでなし。　２１Ｂ10899

４．まとめと今後の課題

本研究では、RUA を用いた日本語話し言葉の談話分析の試みとして、職場における「会議」の談話資料を分析対象として、検討を行った。２節では、分析資料の性質と分析対象の選定基準を述べ、RUA の認定手順を例をで示しながら解説した。３節では、分析の過程で明らかになった問題点と、現状の分析からうかがえた下位分類「場面２」の「打合せ」と「雑談」の特徴を議論した。今回は「会議」の下位分類「場面２」の中から「打合せ」と「雑談」の２つのみを取り上げたが、他の場面でも同様に類型化ができるのか、検討していきたい。

今後の課題として、修辞機能と脱文脈化指数の展開パターンと使用される語彙との組み合わせから、「場面２」のような、具体的な場面の分類認定に使用できる可能性を検討していきたい。また、将来的な自動分類に向けて、話し言葉、特に対話の場合に見られる、言いさし、共話等の扱いの検討、また、話者交代と、修辞機能及び脱文脈化程度の関連についても、検討を行っていきたいと考えている。また、話し言葉を文字化したものを RUA の分析対象とする場合、表面に現れていない情報をいかに解釈するかが問題となることも明らかになった。コーパスの構築についても検討していきたいと考える。

謝　辞

本研究は、文部科学省科学研究費補助金基盤研究（C）「「修辞機能」と「脱文脈化程度」の観点からのテキスト分析手法確立と自動化の検討」（平成 27 年度～29 年度、代表者：田中弥生）による補助を得ています。

文　献

Cloran, C. (1994) *Rhetorical Units and Decontextualisation: an Enquiry into some Relations of Context, Meaning and grammar.* Nottingham: University of Nottingham.

———— (1999) Contexts for learning. In Christie, F. (ed.) *Pedagogy and the Shaping of Consciousness,*London: Cassell, 31-65.

———— (2010). Rhetorical unit analysis and Bakhtin's chronotype. *Functions of Language* 17:1, 29–70.

Halliday, M. A. K. & Matthiessen. C. (2004) *An Introduction to Functional Grammar (3rd ed.)* London: Arnold.

現代日本語研究会　編(2011)『合本 女性のことば・男性のことば(職場編)』ひつじ書房

佐野大樹(2010a)「日本語における修辞ユニット分析の方法と手順 ver.0.1.1－選択体系機能言語理論（システミック理論）における談話分析－（修辞機能編）」（http://researchmap.jp/systemists/資料公開/ （RUA の方法と手順 ver0.1.1）よりダウンロード可能）

――――(2010b)「選択体系機能言語理論を基底とする 特定目的のための作文指導方法について ―修辞ユニットの概念から見たテクストの専門性―」『専門日本語教育研究 12』pp.19-26.

佐野大樹、小磯花絵(2011)「現代日本語書き言葉における修辞ユニット分析の適用性の検証－「書き言葉らしさ・話し言葉らしさ」と脱文脈化言語・文脈化言語の関係－」『機能言語学研究』第 6 巻、pp.59-81.

田中弥生(2011) 修辞ユニット分析を用いた Q&A サイトの質問と回答における修辞機能の展開の検討」『社会言語科学会第 28 回大会発表論文集』 pp.226-229.

――――(2013a)「評価の高低によるクチコミサイト「アットコスメ」における談話構造の特徴 ―修辞ユニット分析を用いて―」『神奈川大学 言語研究』35、pp.1-23

――――(2013b)「クチコミサイトにおける修辞機能の商品評価の高低による違い –修辞ユニット分析による検討–」『機能言語学』7、59-74

田中弥生、佐野大樹(2011a)「Yahoo!知恵袋における質問の修辞ユニット分析－脱文脈化-文脈化の程度による分類－」『信学技報』110(400)、 NLC2010-32、pp.13-18.

――――――――― (2011b)「修辞ユニット分析からみた Q&A サイトの言語的特徴」『言語処理学会第 17 回年次大会(NLP2011)論文集』

――――――――― (2011c)「Yahoo!知恵袋における質問と回答の分類－修辞ユニット分析を用いた脱文脈化-文脈化の程度による検討－」『社会言語科学会第 27 回大会発表論文集』pp.208-211.

節境界認定に関する諸問題

佐藤 理史　(名古屋大学 大学院工学研究科)
丸山 岳彦　(国立国語研究所 言語資源研究系)

Issues of Clause-Boundary Detection

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Takehiko Maruyama　(National Institute for Japanese Language and Linguistics)

要旨

本稿では、文中の節境界を認定するために必要な処理について議論する。我々は、既存のCBAP と異なり、まず、文節境界を認定したのち、節境界を認定する方法を採用する。この方法の採用により、節境界認定問題を、(1) 文節境界 (および文節) をどのように定義するか、(2) 文節にどのような属性を認定するのか、(3) どこを節境界と定義するか、の 3 つの部分問題に分割できる。これらの問題に対する現時点での方針を述べ、BCCWJ のコアデータの一部への節境界付与の見通しについて述べる。

1　はじめに

日本語の文を構成する単位の一つに、「複文を構成するところの、述語を中心とした各まとまり [1]」と定義される**節**という単位がある。これまでの日本語の言語処理において、節という単位はそれほど重要視されてこなかった。しかしながら、我々は、センター試験の国語問題を解くシステムを開発する過程で、長い文をいくつかの部分に分割する必要性に遭遇し、そのための基礎となる節境界の認定が不可欠であるという認識に至った [2, 3]。

節境界検出プログラムには、丸山らの CBAP [4] がすでに存在する。しかしながら、CBAP は、特定の形態素解析システム (ChaSen/IPAdic) に依存していること、および、形態素解析結果の文字列を書き換える方式で実装されているため、保守性・拡張性に難がある。このため、CBAP を改良するのではなく、完全に新しいシステムを作成する方針を採用し、節境界認定システム Rainbow を試作した [2]。

Rainbow の特徴は、(1) 文節境界の認定、(2) 文節属性の認定、(3) 節境界の認定、という 3 つの段階を踏んで、節境界を認定する点にある。このような段階を踏むことにより、節境界認定問題を、3 つの部分問題に分割して解く。もちろん、これらの部分問題は相互に関連しているが、それぞれにおいて解くべき問題はかなり明確であり、保守性・拡張性の高いシステムを作成することができる。

我々は、現在、「現代日本語書き言葉均衡コーパス (BCCWJ)」に節境界を付与するために、Rainbow の新しいバージョン (Rainbow3) を実装している。これに合わせて、上記の (1)(2)(3) に対し、より明確な基準を定めようとしている。本稿では、その内容について報告する。

なお、Rainbow3 の内部は、原則として、益岡・田窪文法 [1] に準拠した文法体系を採用している。

2　3ステップによる節境界認定

節境界認定とは、より正確には、**節の末尾の境界**を認定することを意味する。たとえば、以下の文では、-C-が節境界となる。

(1) 太郎が荷物を軽々と運んだので-C-花子は驚いた-C-

この境界を認定するために、まず、文中の文節境界を認定する。文節境界は-B-で表す。

(2) -S-太郎が-B-荷物を-B-軽々と-B-運んだので-B-花子は-B-驚いた-S-

文節境界で区切られたものを文節と認定する。なお、文の先頭と末尾には、それぞれ文境界 (-S-) があるものとみなす。

次に、それぞれの文節の属性を認定する。ここで最も重要な属性は、その文節が文中で述語として働いているかどうかを表す属性である。この例では、「運んだので」と「驚いた」の2つの文節が述語として働いていると認定する。

(3) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-B-花子は-B-**驚いた**-S-

最後に、この結果に基づいて、節境界 (-C-) を認定する。なお、文末の文境界も節境界であるが、これは、そのまま-S-と表記する。

(4) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-C-花子は-B-**驚いた**-S-

3　境界と句読点

前節に示したように、ここでの解析は、単位を中心とした解析ではなく、**境界を中心とした解析**である [5, 6]。多くの場合、境界は文字と文字の間に存在するが、句読点や括弧などの補助記号は、それ自身が境界を表すとみなす。つまり、これらは実体を持った境界である。句点は「文境界を表す記号」であり、読点は「文中の比較的大きな境界を表す記号」である。実体を持った境界は、以下のように角括弧付きで表す。

(5) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-C[、]-花子は-B-**驚いた**-S[。]-

4　文節境界認定

文節 [7] は、文を構成する単位の中で、おそらく最も合意が取りやすい (個人差が少ない) 単位であろう。以下のように文を文節に区切ることは、多くの人々にとって自然である。

(6) -S-太郎が-B-荷物を-B-軽々と-B-運んだので-B-花子は-B-驚いた-S-

しかしながら、文節境界の認定にも、合意が取りにくい (すなわち、明確な定義が必要な) 場合もある。

Rainbow3 では、**助詞を中心とした**文節境界認定を採用する。すなわち、複合語や派生語を含む**長い単位の語**を助詞と助詞以外 (W) に分け、次のような境界を認定する。これが、長い単位の語の間の境界のすべてである。

助詞-j-助詞
助詞-B-W
W-A-助詞
W-B-W

この結果、文節は、次のいずれかの形式をとることになる。

-B-W-B-
-B-W-A-助詞の列-B-

ここで、文節の W の部分を**主要部**、助詞の列を**機能部**と呼び、主要部と機能部の境界をA境界 (-A-) と呼ぶ。「助詞は文節の機能部のみに現れ、文節の機能部は助詞のみで構成される」ことと、「文節の主要部は、長い単位の語1語で構成される」ことの2点が、Rainbow3の文節モデルの根幹をなす大原則である。このような文節モデルにおいては、助詞の集合を定義すれば、文節境界がほぼ定まる。(より正確には、どんな単位を長い単位の語と認定するか—すなわち、長い単位の語の中に存在しうる全ての境界—を明確に定義する必要がある。)

我々は、文節境界のみが、節境界となりうる境界と考える。

5　節境界と節末形式

日本語の文や節では、述語が他の要素より後ろに配置されるという制約がある。このため、述語を含む文節 (以下、述語文節と呼ぶ) の末尾の境界が、節境界の第一候補となる。

[述語文節]-C-

我々は、これをオーバーライトする場合を規定することにより、節境界の位置を定める。

5.1　拡大述語文節

述語文節の直後に、助動詞を主要部とする文節 (助動詞文節) が後続する場合、助動詞文節を含めて、拡大述語文節と考える。まれに、助動詞文節が連続することがあるが、この場合も、連続する助動詞文節を含め、拡大述語文節とみなす。

(拡大述語文節)
[[述語文節]-B-[助動詞文節]]-C-　　例：書く-B-らしいが-C-

以下の説明では、特に断らないかぎり、述語文節は拡大述語文節を含むものとする。

5.2　節末機能文節

述語文節の直後に、ある特定の文節が後続するとき、これを節に含め、後続文節の末尾境界を節境界と認定する。このような認定を行なう特定の文節を**節末機能文節**と名付ける。節末機能文節が後続する場合、述語文節は原則として連体修飾の形式をとる。

(7) a. 太郎が荷物を軽々と-B-運んだ-B-**こと**-A-**が**-C-花子を驚かせた。
　　b. 太郎が荷物を軽々と-B-運んだ-B-**とき**-C[、]-...
　　c. 太郎が荷物を軽々と-B-運ぶ-B-**あいだ**-A-**に**-C[、]-...

5.3　節境界の決定

節末機能文節が後続する場合にかぎり、「節境界は述語文節の末尾境界」という原則をオーバーライトする。以上により、節末形式は、次のいずれかとなる。

I. 述語文節-C-
II. 述語文節-B-節末機能文節-C-

すでに述べたように、節末機能文節が後続する場合、述語文節は原則として連体修飾の形式をとる。すなわち、若干の例外を除き、述語文節が連体修飾不可能な形式であれば、I 型となる。述語文節が連体修飾可能な形式の場合、後続文節が節末機能文節であれば II 型、そうでなければ I 型となる。

以上の議論から明らかなように、我々のモデルでは、次の情報があれば、節境界を定めることができる。

1. 文節境界 (文節境界だけが、節境界となる可能性がある)
2. 文節が述語文節か否か
3. 文節が助動詞文節か否か (拡大述語文節の判定に必要)
4. 文節が連体修飾可能か否か (述語文節と助動詞文節のみ)
5. 文節が節末機能文節か否か

これらのうち、2–5 の 4 つは、**文節の属性**と考えることができる。つまり、先に述べたように、

Step1 文節境界を認定し、
Step2 その結果認定される各文節の 4 つの属性の値を決定すれば、
Step3 それらの情報のみを用いて、節境界を認定する

ことができる[1]。

残された問題は、Step1 と Step2 の詳細化である。これらのうち、以下では、述語文節の認定と、節末機能文節の認定について議論する。

6　述語文節の認定

節は述語を中心としたまとまりである。つまり、述語があって、初めて節が構成される。そのため、述語を認定することが、節境界を認定する前に必要である。この段階では、文は文節に分割されているので、述語として働いている文節 (述語文節) を認定する問題となる。

述語文節の認定のために、文節の主要部 W の品詞を定義する。Rainbow3 では、次の 10 品詞を採用する。これは、益岡・田窪文法 [1] の品詞体系を踏襲している (指示詞は設けない)。

名詞 (代名詞、数詞を含む)、動詞、形容詞 (ナ形容詞とイ形容詞)、副詞、
連体詞、接続詞、感動詞、助動詞、判定詞、助詞

このうち、助詞を除く 9 品詞が文節の主要部となるが、文節が述語文節となりうるのは、原則として、主要部が

動詞、形容詞、判定詞、助動詞

[1] 述語文節の後ろに、節末機能文節が連続して接続する III 型を考える必要があるかどうかは、現時点では保留とする。ただし、III 型を導入しても、3 ステップ節境界認定法は堅持できる。

の 4 品詞の場合である。助動詞は、ほとんどの場合、動詞、形容詞、判定詞に後続するが、名詞に直接接続する場合があるので、便宜上、述語文節を構成するとみなす。同様に、「か」「かしら」などの一部の終助詞は、名詞に直接接続する場合があるので、「学生かしら」のような文節は、述語文節と認める必要がある。

述語文節の認定の難しさは、文節内の情報のみからでは、その文節が述語文節かどうか判定できない場合があることにある。これには、節の定義の問題も関連する。

問題となるのは、主に、形容詞の連用形と連体形である。イ形容詞[2]は、単体で述語を副詞的に修飾している場合 (連用修飾) は述語として働いていない (節を構成しない) 一方で、補足語 (格要素) を支配する場合は述語として働いている (節を構成する) とみなすのが一般的である。ただし、これらを区別せず、両者とも節とみなす立場もある。

(8) a. 花が-B-美しく-B-咲いた-S[。]-　(連用修飾語)
　　b. 花が-B-美しく-C-香りも-B-いい-S[。]-　(並列節)

形容詞の連体形も同様である。

(9) a. きれいな-B-女性-B-　(連体修飾語)
　　b. 声が-B-きれいな-C-女性-B-　(連体節)

以上の例からわかるように、これらの区別は、当該文節の範囲内では決定できないのは明らかである。

文節の属性を計算するという立場から考えれば、ここで行うべきことは、文節内の情報に基づいて、その文節が述語文節となりうる可能性を持つか否かの判定である。この段階では、可能性を持つものには、すべて「述語性」という属性を付与する。

その先の処理には、いくつかの選択肢がある。

1. 一般的な述語および節の定義に従うように、述語文節を認定するルール (ヒューリスティック) を実装する。
2. 境界ラベルあるいは節ラベルを工夫することにより、格要素を伴わない場合には節とならないという情報を明示する。
3. 格要素を伴わない場合でも節と認める。

最終的な決定は保留しているが、現時点では、2 番目の選択肢を中心に検討している。

7　節末機能文節の認定

節末機能文節が関与する節は、主に、副詞節と補足節である。これらの節では、連体節との区別が問題となる。

7.1　副詞節と連体節

副詞節か連体節かの判定は、後続文節を節末機能文節と認定するか否かに帰着させる。

(10) a. わたしが 16 だった-B-**とき**-C[、]-彼女はまだ 7 つでした。 (副詞節)
　　b. わたしが 16 だった-C-年、彼女はまだ 7 つでした。 (連体節)

[2] ナ形容詞の並列節はテ形をとるので、ナ形容詞の連用形は連用修飾語とみなしてよい。

節末機能文節が「に」や「で」を伴う場合、「に」や「で」を節に含める。これは、「に」や「で」を助詞とみなすことに相当する。以下の最後の例に示すように、「で」は助詞ではなく、判定詞「だ」のテ形の場合もあるが、この場合の「で」も、特別な助詞 (判定詞由来の助詞) として助詞扱いとする。

(11) a. 採決が終わった-B-**後**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
b. 採決が終わった-B-**後に**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
c. 採決が終わった-B-**後で**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
d. 大勢の人が反対意見を言い始めたのは、採決が終わった-B-**後で**-C[、]-それが問題を引き起こした。　(並列節)

これに対して、副詞節の形式に判定詞「だ」の基本形が後続する場合は、判定詞の前を節境界とする。(判定詞は助詞ではないので、文節の主要部となる。)

(12) 大勢の人が反対意見を言い始めたのは、採決が終わった-B-**後**-C-だ。 (副詞節)

「〜せいで」は副詞節を作れるが、「〜せい」は作れないので、次のような扱いとする。

(13) a. 電車が止まった-B-**せいで**-C[、]-会議に行けなかった。 (副詞節)
b. 会議に行けなかったのは、電車が止まった-C-せいだ。 (連体節)
c. 会議に行けなかった。電車が止まった-B-**せいで**-C-だ。 (副詞節)

なお、「電車が止まったせいで (も) ある」の扱いは、現時点では保留である。いずれにしても、判定詞が後続する場合はすっきりしないことは免れない。

7.2　補足節と連体節

補足節か連体節かの判定も、後続文節を節末機能文節と認定するか否かに帰着させる。補足節を作る節末機能文節の主要部は、「の」「こと」「ところ」に限られるため、どのような助詞を伴うかが焦点となる。

(14) a. 花子は太郎がその店に入る-B-**ところを**-C-見かけた。 (補足節)
b. 太郎はその店に入る-C-ところで、花子はそれを見かけた。 (連体節)
c. 太郎は、店の勝手口に入る-C-ところで、花子と会った。 (連体節)
d. 太郎はその店に入る-C-ところだ。 (連体節)
e. あの勝手口が、太郎がお店に入った-C-ところだ。 (連体節)
f. 結婚する-B-**ことに**-C[、]-母が反対した。 (補足節)
g. 母が反対したのは、結婚する-B-**ことに**-C-だ。 (補足節)
h. 君にあげた-B-**のは**-C[、]-この指輪だ。 (補足節)
g. この指輪は、君にあげた-の-だ。 (見分けがつかない)

原則として、格助詞あるいは係助詞「は」「も」を伴う場合は補足節とみなし[3]、それ以外の場合は連体節とみなす。

[3] 理想的には、述語と格関係にある場合は補足節とみなすべきであるが、高い精度で機械的にそれが判定できるかどうかは不明である。

7.3 「という」「ような」

「という」と「ような」は、後続が形式名詞であっても、そこで区切る。ただし、これらの後ろが「の」の場合は、そこで切らない。

(15) a. 彼が書いたらしいという-C-ことが、 (連体節)
b. 彼が書いたような-C-ことは、 (連体節)
c. 論文を書くというのが-C-望ましい。 (補足節)
d. 似たようなのが-C-他にも 2 個以上ある。 (補足節)

「という」に接続助詞が後続する場合は、接続助詞までを節とする。

(16) a. 彼は書いたというし-C[、]- (連用節)
b. 彼は書いたというが-C[、]- (連用節)
c. 彼が書くらしいというので-C[、]-期待が高まった。 (副詞節)

8 BCCWJ コアデータへの節境界付与

Rainbow3 が前提としている文法体系は、益岡・田窪文法 [1] を文節文法に焼き直したものである。一方、BCCWJ の解析済データは、これとは異なる文法体系を前提としている。そのため、BCCWJ の解析済データに節境界を付与するためには、その不整合を吸収する必要がある。

現時点では、まず、長単位 (LUW) の TSV データを入力とし、それに節境界を付与することを先行させている。これは、LUW の TSV データには、文節区切りの情報が含まれていること、長単位認識のための解析が不要であることの 2 つの理由による。

LUW の TSV データに対して節境界を認定するための前処理 (不整合の吸収) は、おおよそ次の 2 種類に分類できる。

1. 品詞の付け替え
2. 単位の調整 (LUW と Rainbow3 の長い単位の語の不整合を調整)

現時点では、BCCWJ 体系から Rainbow 体系への変換は必要最小限に止めているが、その中で最も煩雑なのは、BCCWJ の助動詞の変換である。BCCWJ の助動詞の多くは、Rainbow 体系では、活用語尾、接尾辞、動詞 (複合動詞後件) 扱いとなる。

BCCWJ の文節と Rainbow3 の文節は、大体の場合は一致する。一致しないのは、主に、判定詞、助動詞、形式名詞が関わる場合と、複合辞が関わる場合である。これ以外に、テ形複合動詞の扱いが一部異なる。たとえば、「引返してゆく」を BCCWJ は 2 文節とみなすが、Rainbow3 では 1 文節とみなす。

残された問題は、節境界ラベルの設計である。応用の立場からは、意味的な節境界ラベルが望ましい。一方、認定処理の立場からは、ほぼ一意に決定できる境界ラベルが望ましい。現時点では、形式的な境界ラベル集合 (たとえば、「とき節」) と意味的な境界ラベル集合 (たとえば、「副詞節-時間」) の両方を設計し、それらの両方を付与する (意味的境界ラベルは可能な候補を付与する) 方針を立てている。

図 1 にサンプル PB12_00001 の冒頭部分の節境界認定例を示す。この図では、節境界を認定した部分で改行している。最右欄に示す節境界ラベルは仮のものである。

-S[ ]-パソコン-A-の-B-画面-A-や-B-本-A-など-j-に-B-集中し-A-ながら-C[、]-　ながら節
-C[、]-自分-A-の-B-入れ-k-た-C-　動詞-タ形連体形
-C-飲み物-A-に-B-手-A-を-B-伸ばし-C[、]-　動詞-連用形
-C[、]-飲み物-A-に-j-は-B-まったく-B-目-A-を-B-遣ら-n-ない-B-まま-C-　ママ節
-C-飲む-A[、]-という-B-の-A-は-C-　という-ノ節=は
-C-だれ-A-で-w-も-B-やる-B-こと-C-　コト節
-C-だろう-S[。]-　判定詞-意志推量形
-S-自分-A-で-B-入れ-k-た-B-のだ-A-から-C[、]-　から節
-C[、]-それ-A-が-B-なん-B-な-B-の-A-か-j-は-C-　ノ節=か-は
-C-見-n-なく-k-て-A-も-C-　イ形容詞-テ形=も
-C-わかる-S[。]-　動詞-終止形
-S-だから-J-たいがい-B[、]-なん-A-の-B-問題-A-も-B-ない-S[。]-　イ形容詞-終止形
-S[ ]-ところが-J[、]-ごくごく-B-稀-k-に-B[、]-変-k-な-B-こと-A-が-C-　コト節=が
-C-起こる-S[。]-　動詞-終止形
-S-たとえば-B[、]-紅茶-A-を-B-入れ-k-た-A-の-w-に-C[、]-　のに節
-C[、]-どう-w-いう-B-わけ-A-か-C[、]-　ワケ節=か
-C[、]-コーヒー-A-を-B-入れ-k-た-A-と-C-　と節
-C-勘違いし-t-てしまう-S[。]-　動詞-終止形
-S[ ]-手-A-を-B-カップ-A-に-B-伸ばす-S[。]-　動詞-終止形
-S-頭-A-は-B-コーヒー-A-を-B-入れ-k-た-A-と-C-　と節
-C-思い込ん-t-でいる-A-から-C[、]-　から節
-C[、]-口-A-は-B-すっかり-B-コーヒー-A-を-B-受け容れる-C-　動詞-連体形
-C-態勢-A-に-B-なっ-t-ている-S[。]-　動詞-終止形

図 1: 節境界認定例 (PB12_00001 の冒頭部分)

-j-を除く小文字の境界記号は、語内の境界を表す。-J-は接続詞の末尾の境界を表す。

謝辞

本研究では、『現代日本語書き言葉均衡コーパス』を利用した。本研究は、JSPS 科学研究費基盤研究 (B) 「文章の読解と産出のための言語処理技術」(課題番号 15H02748) の助成を受けている。

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法—改訂版—. くろしお出版, 1992.
[2] 加納隼人, 佐藤理史. 日本語節境界検出プログラム rainbow の作成と評価. 第 13 回情報科学技術フォーラム (FIT2014), E-005, 第 2 分冊, pp. 215–216, 2014.
[3] 加納隼人, 佐藤理史, 松崎拓也. 節境界検出を用いたセンター試験『国語』評論傍線部問題ソルバー. 情報処理学会自然言語研究会, NL–220–8, 2015.
[4] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.
[5] 佐藤理史. 境界認定の提案: (1) コンセプトと実現法. 情報処理学会自然言語研究会, NL–164, pp. 25–32, 2004.
[6] 佐藤理史. 境界認定の提案: (2) 背景と思想. 情報処理学会自然言語研究会, NL–164, pp. 33–44, 2004.
[7] 橋本進吉. 国文法体系論. 岩波書店, 1959.

名詞の項構造データの構築

竹内 孔一 (岡山大学大学院自然科学研究科)[1]

Construction of Japanese Noun Argument Structure Data

Koichi Takeuchi (Graduate School of Natural Science and Technology, Okayama University)

要旨

本研究では言語処理に利用可能な名詞の項構造データの構築を行っている．名詞の持っている情報を記述する方法として，述語と意味的関係の強い名詞や述語を関係タイプを付与して整理する素性構造が提案されており，本研究でも最も単純な構造を仮定して，テキストデータに付与を行っている．しかしながら,「以外」や「頭文字」など参照先の概念に対して操作を要求する語 (ここでは抽象名詞とする) があり，従来の素性構造でもよくわからない．そこで本稿では，既に構築している名詞項構造データについて整理した上で，集められた抽象名詞に対して分析を行う．どの程度のタイプ分けが可能かについて言語処理における含意認識タスクを意識した分析を提示する．

1　背景

本研究者は既に述語の項の関係を整理し意味役割を付与した述語項構造シソーラスを構築し，規則ベースによる意味役割付与システム公開してきた．しかしながら一方で含意認識タスクなど，言語処理において，ある表現が他の表現に含まれているかどうかを判定するには，述語の情報だけで無く，名詞の情報が必要である (竹内 (2014))．名詞の持つ意味における先行研究として近年では，項構造を基にした分析 (庵 (2007))，非飽和名詞に基づく分析 (西山 (2003, 2013))，GL (Pustejovsky (1995); 影山 (2011))，データ構築 (A. Meyers and R. Reeves and C. Macleod and R. Szekely and V. Zielinska and B. Young and R. Grishman (2004)) などがあげられる．

本研究ではこれらの成果を受けて，西山 (2003) から非飽和名詞を集めて例文の作成と項構造の付与を行っている (竹内他 (2015))．さらに素性構造に基づく名詞の意味構造の付与 (竹内他 (2014)) を検討している．これにより,「著者」や「作者」といった表現の異なりと関連する名詞 (著者名や作品名) との関係付けを行うことが可能であるが，一方で,「以外」 「一種」など項構造の関係だけでなく，意味的な操作を要求する抽象名詞があり，扱うことができていない．また同じ抽象名詞でも「こと」や「場合」などは既存の項構造の枠組で捉えることで問題が無いように見える．つまり言語処理を踏まえた上での抽象名詞まで含めた名詞の意味構造の記述枠組がまだ確定できていない．

そこで本論文では NTCIR の含意認識タスク RITE-2 [2] を言語処理の応用例として設定して，名詞や抽象名詞の意味構造について考察を行う．まず，既に提案している名詞の項構造によってどのような意味的関係が記述できるかをまず整理した後，抽象名詞の分類とその意味記述の試案を行う．抽象名詞の収集方法は明らかでは無かったが，近年，田邊 (2008) による連体修飾の方法で一部収集できることが明らかになってきた．よって収集された抽象名詞ならびに含意認識タスクの分析から集めた抽象名詞を分類しどのような意味構造記述が可能か検討する

2　構築中の名詞項構造データと含意認識タスク

本節では，現在構築中の名詞データを整理すると共に，含意認識タスクにどのように寄与するかについて明らかにする．構築する名詞項構造データは下記のものがある．

1 名詞項構造事例データ

2 名詞項構造アノテーションデータ

3 名詞辞書データ

[1] koichi [at] cl.cs.okayama-u.ac.jp
[2] http://www.cl.ecei.tohoku.ac.jp/rite2/.

まず **1** 名詞項構造事例データであるが，普通名詞に対して「X の Y は Z だ」という形式で例文を人手で作成し，X や Y に対する意味的な関係を意味役割で付与したデータである．現在 2500 例文を作成している．例えば,「上司」ではその部署の名前と人そのものが現れるので下記のようになる．

・ [主体 私の勤める会社] の上司は [対象（人）田中さん] だ

【主体】や【対象（人)】は意味役割を表しており，述語シソーラスで 72 種類定義されている[3]．この事例から,「上司」の周辺には 2 つの項が出現すること，またその具体的な例が提示されていることから後に名詞の項構造に対する機械学習の適用が期待できる．また，意味役割は今のところ述語から定義した意味的関係を付与しているが，それだけでは関係が捉えられないことが予測される．そこで，PropBank や NomBank 同様，ARG0, ARG1 など数字として項を固定して，これに対して【主体】や【対象（人)】を付与することにする．これにより異なる理論が現れても，まずある名詞の ARG0, ARG1 などのどれかを固定することができるため，同じ土台で議論や実験が行える見通しである．

一方で，項構造事例データはどのように役立つであろうか．例えば下記のような含意認識の事例の際に役立つと考えられる．

t1 サファリジャケットの特徴は、一般的に軽量コットン生地または、より軽いポプリンで縫製されたジャケットで、伝統的にカーキ色、そして付属ベルトや、肩章と 4 つ以上のプリーツ付きポケット（ベローズポケット）が付いている。

t2 サファリジャケットは、4 つ以上あるポケットが特徴だ。

含意認識タスクとは 2 つ目の文の内容が 1 つ目の文に含まれているかどうかを判定するタスク (最も単純な場合) である．テスト事例として含意される場合とされない場合の両方があり，含意認識システムはこれらを正確に分類する必要がある．上記の例は含意する例であり,「特徴」という名詞のもつ意味的関係が解くポイントである．この例文では [t1] では「X の特徴は Y」という構文 (construction) であり，[t2] の方は「X は Y が特徴だ」という構文である．ここで X は「特徴」を持つ主体であり，Y はその特徴そのもの (つまり意味役割では【対象】) となる．よって「特徴」という項構造の事例と X，Y といった項との関係が同定されれば，これらの含意関係は正しく取り出せる可能性が高くなる．

次に **2** 名詞項構造アノテーションデータは普通名詞に対して意味的な関係のある名詞が文中に出た場合，項として意味関係を記述するタスクである．対象とする文は RITE-2 タスクの文であり，これにより RITE-2 の名詞の項構造付与システムを作成した場合に，直接精度評価を行うことができる．また作例では無いため，気がつかない項や語義を見つける可能性がある．基本的にはタスク **1** と同じだがタスク **2** では文に合わせて項を付与する．基本的には NomBank と同様で，例文に対して名詞の項に意味関係を付与する．

最後に **3** 名詞辞書データは上記の項構造データに加えてさらに影山 (2011) が提案している素性構造を付与する．全ての名詞について記述するのは不必要に難しくなるため含意認識に必要だと思われる名詞に対して付与する．例えば「著者」ならば項構造例文と合わせて表 1 のような構造となる．表 1 では「書く」は動詞の語義 [生成] といった述語項構造シソーラスの語義概念を付与する．

一方，こうした素性構造は「著者」と著作物との関係を掴むことができる．どう役立つか下記の RITE-2 の含意関係にある 2 文で説明する．

t1 『世界征服者の歴史』とは、『集史』などと並び、モンゴル帝国の政治家・歴史家のアラーウッディーン・アターマリク・ジュヴァイニーによって 1260 年に完成されたモンゴル帝国を語る上で重要な歴史書である。

t2 アラーウッディーン・アターマリク・ジュヴァイニーは歴史書『世界征服者の歴史』の著者として知られている。

[3]以降ここでは【】は意味役割を表す.

表 1: 「著者」の素性構造

	「著者」
項構造	(ARG0, ARG1)
項構造例文	[ARG1/対象 重力ピエロ] の著者は
	[ARG0/動作主 伊坂幸太郎] だ
外的分類 (Formal)	人間 (ARG0)
成り立ち (Agentive)	書く [生成](ARG0, ARG1)

この例文では [t1] では「X によって完成された Y」という構文があり，[t2] では「X は Y の著者」という構文が記述されている．上記の素性構造では,「著者」と「書く [生成]」という動詞が結び付けられており，述語シソーラスの分類では「完成」も 3 層目で「生成」と分類されている．こうした情報から項構造が同定できれば，素性構造を通して含意関係にあることが予測されやすくなる．

このように現在構築中の名詞項構造データは含意認識タスクでの貢献が期待されるが一方でこれら意味構造では捉えられない名詞が存在する．次節では取り上げて有効な構造について論じる．

3　抽象名詞の事例と名詞の意味構造の検討

抽象名詞にどのような種類があるか言語学的な視点というよりも，言語処理の観点から意味構造を検討してみたい．まず田邊 (2008) が抽出した抽象名詞を分類した後，意味構造が素性構造だけでは捉えられない抽象名詞について検討する．

本研究での言語処理の視点は項構造に基づく意味構造に分解することである．述語が取る項の意味的関係の分類 (竹内 (2014)) から，大きくわけて構文類 (【補語相当】など構文的な関係をしめすもの)，対象類 (【対象】などイベントの状態変化や状態に関わるもの)，動作種類 (【動作主】や【原因】といった述語が指すイベントを発生させる要因)，条件周辺類 (【時間】や【場所】,【条件】といったイベントの存在条件に関係するもの) の 4 つに分類される．こうした構造と上記の名詞の項構造データにより文同士の意味的関係を捉えるとする．この視点から，田邊 (2008) の抽象名詞を分類すると表 2 のようになる．

表 2: 連体修飾で得られた名詞の分類

意味役割	時空間	「中」「際」「地域」「ところ」「とき」「うち」「間」
	手段	「方法」
	原因	「わけ」「ため」「理由」
	条件	「場合」「限り」
	逆接	「一方」
分類 (名詞の属性)		「人」「もの」「声」「情報」「点」「こと」「事」
		「言葉」「方針」「女性」「気」「形」「調査」「意味」
モダリティ		「必要」「つもり」「はず」「予定」
機能的		「ほか」「前」

表 2 における意味役割に分類されているものは抽象名詞であり，述語に対するその名詞句の意味役割を指定する働きがあると考えられる．また名詞の属性では「人」や「こと」といった言葉でその名詞句が人間か出来事なのか分類を明確にして述語との関係を明確にしようとしている．また，モダリティは確信度や必須か義務かなどを表すが，英語の名詞項構造の分析 (NomBank) で既にモダリティ情報を付与することが提案されており日本語でも同様であることがわかる．

一方で，機能的な名詞であるが表 2 で示した「ほか」「前」のほかに RITE-2 データから「以外」，「頭文字」「略称」「一種」「候補」などがある．これらの機能的な言葉は，なにか元の概念に対して論理的な操作を加えると考えられる．論理的な操作の観点から分類すると，(1) 集合操作 (「以外」「ほか」「一種」)，(2) 順序操作 (「前」「後」「候補」)，(3) 対象と操作 (「略称」「頭文字」) の 3 つに分けられる．(1) は集合を仮定する必要があり，含意認識タスクでは「日本・ドイツ以外のほぼすべての国」のように国の集合が必要になる．よって (1) を解くには文で求められる集合を獲得する手法が必要であると考えられる．また (2) から順序構造を扱う必要があることがわかる．時間や場所以外にも，必要に応じた順序の計算が求められる．最後の (3) であるが両方とも文字に関する操作であるが，どういうことが「略する」ことかは人には理解できるが記述は容易ではない．このあたりは事例を収集してより分析したい．

4　おわりに

構築中の名詞の項構造データの含意認識タスクにおける効果について検討し，見通しを明らかにした．また抽象名詞について分析を行い，ほとんどのものは既存の意味役割関係や名詞の分類であることを示した．さらに，抽象名詞について論理的な操作の観点から分類を行い，集合的な操作や順序操作だけでなく自由度の高い操作を指定する抽象名詞があることを明らかにした．

謝辞

本研究は，基盤研究 (C) 課題番号 26370485 (研究代表者: 竹内孔一) の補助を得ている．ここに記して深く感謝する．

文献

A. Meyers and R. Reeves and C. Macleod and R. Szekely and V. Zielinska and B. Young and R. Grishman (2004) “Annotating Noun Argument Structure for NomBank,” in *Proceedings of LREC2004*, pp. 803–806.

J. Pustejovsky (1995) *The Generative Lexicon*: MIT Press.

庵功雄 (2007) 日本語におけるテキストの結束性の研究，くろしお出版．

影山太郎 (2011) 日英対照 名詞の意味と構文，大修館書店．

西山佑司 (2003) 日本語名詞句の意味論と語用論，ひつじ書房．

西山佑司（編） (2013) 名詞句の世界，ひつじ書房．

竹内孔一 (2014)「述語項構造シソーラスを意識した名詞の意味構造アノテーションのための名詞意味構造の検討」，第 6 回コーパスワークショップ予稿集，pp.51–56．

竹内孔一，竹内奈央，石原靖弘 (2014)「述語項構造シソーラスによる述語と名詞の構造化」，人工知能学会全国大会, 2I5-OS-08b-1．

竹内孔一，宮田周，河村一希 (2015)「述語項構造シソーラスを意識した名詞データの構築」，第 7 回コーパスワークショップ予稿集，pp.143–146．

田邊和子 (2008)「BCCWJ に拠る名詞別格外連体修飾形の形成傾向の分析」，第 7 回コーパス日本語ワークショップ予稿集，pp.165–174．

ディスカッション観察支援システム FishWatchr を用いた実践手法の提案

山口昌也 (国立国語研究所言語資源研究系) †
大塚裕子（公立はこだて未来大），北村雅則（南山大学）

Proposal of Methods of Discussion Training Using Discussion Observation Support System "FishWatchr"

Masaya YAMAGUCHI (Dept. Corpus Studies, NINJAL)
Hiroko OTSUKA (Future University Hakodate), Masanori KITAMURA (Nanzan University)

要旨

本稿では，筆者らが開発している，ディスカッション観察システム FishWatchr を実践に適用する二つの方法を提案する。FishWatchr は，学習者がディスカッションなどの言語活動を観察したり，観察結果の評価を行うのを支援するために開発された。主な機能は，ビデオや音声データなどのメディアデータに対して，リアルタイムでアノテーションし，その結果をリフレクションなどで活用しやすいように表示することである。本稿では，対象とする実践を (a) グループ・ディスカッションを録画した教材用ビデオに対して，各学習者が個別にアノテーションし，グループで評価活動を行うもの，(b) 実際のグループ・ディスカッションを録音し，当事者がそれぞれリフレクション活動を行うものとし，それぞれのタイプの実践に FishWatchr を導入する方法を示す。

1　はじめに

言語活動の観察は，ディスカッション教育におけるフィッシュボール，ロールプレイ，スピーチ練習などに導入されている。観察の結果は，グループでの評価活動やピアでのコメント活動を伴うリフレクションといった協同学習的な手法で活用される (大塚・森本 2009; Douglas et al. 2014 など)。このような手法の利点は，他人から自分では気づかない点のフィードバックを得られることのほか，他者のよい点を取り入れたり，他者への教授による自発的な学びが期待される点である (Barkley et al. 2009)。

上記のような観察に基づく教育活動を行う場合，言語活動を記録し，適宜参照できるビデオや音声データ（以後,「メディアデータ」）は有用であり，その教育的有効性はさまざまな形で検証されている (Yousef et al. 2014)。また，メディアデータに対するアノテーションツールの開発も盛んに行われており (Rich and Hannafin 2009)，例えば，Driver[1]，ELAN (Brugman and Russel 2004)，STUDIOCODE[2]，Transana[3] などがある。その一方で，アノテーションツールが適用されている分野は，教師による授業のリフレクションなど (大倉 2009，小川ら 2012 など)，一種の専門家に利用されるのが主流であり，学習者自身によるアノテーションは広く行われているわけではない。

そこで，筆者らが開発中のディスカッション観察支援システム FishWatchr[4] を用いて，学習者自身によるアノテーションを取り入れた，二種類の実践方法を提案する。ここで言う「アノテーション」とは，メディアデータの特定の位置に対して，コメントやコード（利用者によって定義されるラベル）を付与することを指す。本稿では，学習者によるアノテーションが導入されにくい理由として，アノテーションツール導入が授業に与える影響の大きさを挙げ，既存の授業との差異を少なくするような，アノテーションツールの運用方法を考える。また，既存授業との差異という観点から，提案した実践方法を考察する。

†http://www2.ninjal.ac.jp/masaya
[1]http://diver.stanford.edu/
[2]http://studiocodegroup.jp/
[3]http://www.transana.org/index.htm
[4]http://www2.ninjal.ac.jp/lrc で無償公開している。

2　FishWatchr

2.1　システムの概要

FishWatchr は，ディスカッションや発表練習などの言語活動を観察・評価するのを支援するためのアノテーションツールである。筆者らは，リアルタイムで進行している言語活動に対するアノテーションや，アノテーション結果を利用したグループ活動への活用を目指して，開発を進めている。FishWatchr は，スタンドアロンのデスクトップアプリケーションである。Java で記述されていることから，Window, MacOS X，Linux など，ほとんどの PC 上で動作する [5]。図 1 は，FishWatchr を使って，フィッシュボウルに対してアノテーションを行っている例である [6]。

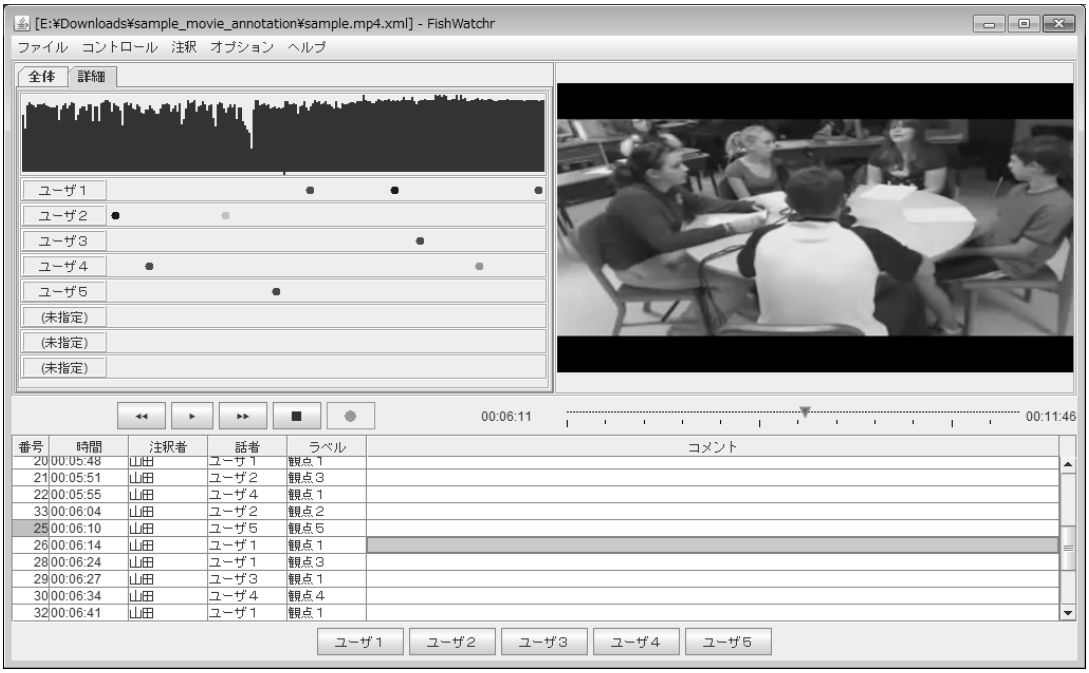


図 1: FishWatchr の動作例

FishWatchr の主要な機能であるアノテーション機能は，メディアデータに対して，注釈をつける機能である。図 1 では，再生中のビデオに対して，アノテーションを行っている。ウィンドウ下部のボタンを押すと，対応付けられた情報がその時点の再生位置に付与される。図の例では，各ボタンがディスカッションの発話者名と対応している。「ユーザ 1 」を押すと，次のように，再生位置の時間情報，事前に設定した注釈者名（「山田」）とともに，話者名（「ユーザ 1 」）が付与される。

時間	注釈者	話者	ラベル	コメント
00:06:14	山田	ユーザ 1	観点 1	（自由記述のコメント）

ボタンには，２種類のコード（話者，任意ラベル）の値を自由に定義できるようになっている。図の場合，話者として，ディスカッション参加者５人分の名前が「ユーザ 1 」～「ユーザ 5 」と定義されている。各アノテーションには，コードの他，自由形式のコメントを書き込める。

付与された情報は，ウィンドウ下部の「アノテーション一覧」に追加される。加えて，ウィンドウ左には，アノテーション一覧が話者ごとに時系列順にプロットされる。また,「アノテーション一覧」中の特定のアノテーションをダブルクリックすると，その位置を再生することができる。

[5]マルチメディアファイルの再生には内部的に VLC を利用しているため，それぞれのプラットフォーム用に VLC のインストールが必要である。

[6]再生中のビデオは，次のビデオからフィッシュボールの場面を引用している。
Paul Bogush: Middle School Fish Bowl Discussion (http://www.youtube.com/watch?v=RwxnBv-dNBI)

FishWatchr の詳細については，別稿にゆずるとして，以下２節では，本稿に関係する機能として，メディアデータとアノテーションとの関係，および，グループでの評価活動の支援機能について説明する。

2.2　アノテーションとメディアデータとの関係

FishWatchr は，次の三つの状況でアノテーションできるように設計されている。これにより，収録済みのメディアデータに対するアノテーションだけでなく，実際に目の前で実施されている言語活動へのアノテーションにも対応できる。

- ファイル・アノテーション： FishWatchr 上で再生中のメディアファイルに対して，アノテーションする。これは，図 1 のように，収録済みのメディアファイルに対してアノテーションする状況である。
- リアルタイム・アノテーション（音声録音）: リアルタイムで進行中の活動に対して，アノテーションすると同時に，FishWatchr で録音する。
- リアルタイム・アノテーション（別機器収録）: リアルタイムで進行中の活動に対して，アノテーションする。活動の収録は FishWatchr とは別の機器で行い，アノテーション結果と別途同期する。

アノテーション結果は,（メディアデータに記述するのではなく）独立したファイル（アノテーション結果ファイル）として保存される。このような保存方法の場合，メディアデータとアノテーション結果ファイルは，何らかの形で時間情報を同期させる必要がある。上記三つの状況のうち，上から二つの状況では，アノテーション時に記録される時間は，再生開始，もしくは，収録開始からの経過時間と一致するので，メディアファイルの先頭からの経過時刻とそのまま同期できる。一方,「リアルタイム・アノテーション（別機器収録）」だけは，収録機器と FishWatchr（を実行する PC）で個々に時間情報を計測するので，精度を考慮し，手動で同期させる [7]。

2.3　グループ活動の支援

各学生が行ったアノテーション結果は，その後のグループ活動や教師の指導で活用することを想定している。そのため，複数の学生が行ったアノテーション結果を統合することが求められる。このことを実現するためには，FishWatchr では，アノテーション結果のインポート・エクスポート機能を用いる。具体的には，統合時，エクスポートしたアノテーション結果ファイルを一つの PC に集め [8]，一括してインポートする。

統合結果の利用方法として，アノテーション結果の閲覧支援機能 (図 2) を用意している。図 2 のように，ラベルごと（左図），話者ごと（右図）にアノテーションを時系列に表示できるほか，注釈者ごとの表示も可能である。この機能を用いれば，例えば，話者や注釈者ごとの傾向を簡易的に分析することに利用できる。

3　実践案

3.1　想定する授業

実践案を示す前に，それぞれの実践の背景として，実践案を導入する授業について説明する。本稿で提案する実践案は，次の二つのディスカッション教育スタイル (高垣 2010）に基づいた授業を想定している。実施場所は，大学である。

- 気づき支援型：この授業では，議論の中で学習者が自らの気づきをとおして，スキルや態度を促す。ディスカッション教育を専門に扱う授業の中で使うことを想定している。

[7] 原始的な方法だが，収録時に同期用音声に対してアノテーションする方法，正確な時計の表示を録画しておく方法などを考えている。
[8] 学内の共用ファイルサーバや Dropbox などのネットワーク上のサービスを利用することを考えている。

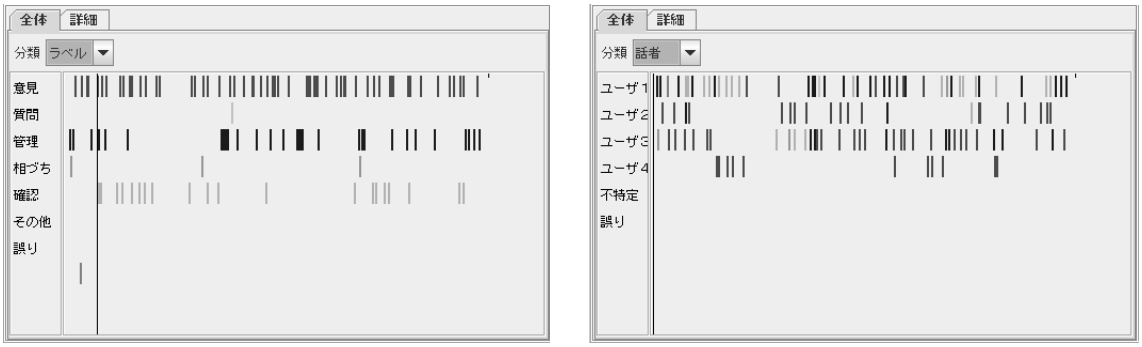


図 2: FishWatchr の動作例

- ルール提示型：教師が事前に話し合いのルールを提示し，それらを守ることによってディスカッション活動を円滑に進める。想定するのは，ディスカッション専門の授業ではなく，授業の中でグループ活動があり，ディスカッションを道具として利用する授業である。

3.2　実践案作成の方針

本稿では，1 節で述べたように，学習者によるアノテーションが観察に基づいた教育活動に導入されにくい理由として，既存の授業との「差異」があると考える。そこで，その「差異」を少なくすることを目標に実践案を作成する。

本稿では，既存の授業との「差異」を生じさせる，二つの要因に焦点を当てる。一つは，ディスカッションへのアノテーションを行うために，追加的な時間が必要になるということである。大学の授業は半期 15 回であり，カリキュラム全体への影響を考慮すると，導入に伴う時間増加は避けなければならない。そこで，本実践案では授業外の時間を利用することにより，この問題を解決する。

もう一つの要因は，PC，収録機器，メディアデータなど，通常のディスカッション活動には必要とされない，追加的な機器類が必要になることである。また，機器類の運用方法も考慮する必要がある。例えば，メディアデータ，特にビデオファイルは容量も大きく，その配布方法によっては実践上の問題が発生しかねない[9]。本実践案では，多くの大学で利用可能だと考えられる PC 教室の使用を前提とし，可能な限り，既存の機器類を活用する方法を模索する。

以下 2 節では，「気づき優先型」「ルール提示型」の二つのタイプの実践案を示す。

3.3　気づき優先型の実践案

気づきを促すタイプのディスカッション練習として，「フィッシュボウル」がある。本実践案では，フィッシュボウルを実際に行う前の練習として，グループ・ディスカッションを収録したビデオを観察する。この実践の目的は，見本となるような，よいディスカッションのやり方を見つけて，グループで共有することである。したがって，使用するビデオは，見本となるディスカッションを収録したものとなる。実践の手順は，次のとおりである。

(1) グループ・ディスカッションのビデオを観察し，主として，参考になる箇所にアノテーションする。この段階は，授業前に自習として行う。アノテーションの内容としては，大塚・森本 (2011) の評価の観点を参考にした任意ラベルを付与するとともに，アノテーションの詳細説明を自由記述で付与する。この段階の準備として，FishWatchr のセットアップと観察用ビデオを学生全員に配布する必要がある。ディスカッションの長さは 30 分程度とする。この長さだと，ファ

[9]例えば，ネットワーク上においたビデオファイルに対して，多数の学生が一斉にアクセスした場合，ネットワーク環境によっては，コマ落ちなどの再生上の問題が発生する可能性がある。

イルサイズが 1GB 以上になることがあるため [10]，ダウンロード時間に対する配慮が必要である。特に，学生数が多い場合は，グループごとに USB メモリなどで配布し，ネットワークでの配布は避けたほうがよいと思われる。

(2) 授業の際に，各自のアノテーション結果を持ち寄り，FishWatchr 上で統合する。グループの人数は，３～４名とする。

(3) アノテーション結果を共有するためのグループ活動を行う。基本的に各自がアノテーションした場所を再生しつつ，アノテーションした理由を説明する。この際，他のメンバーのアノテーションと近接している場合は，意見を交換する。

(4) 教師が全グループのアノテーション結果をすべて統合し，全員で共有できるようにフィードバックする。

3.4　ルール提示型の実践案

ルール提示型の実践案は，グループ・ディスカッションとそのリフレクションで構成する。気づき優先型のディスカッション練習と異なり，ディスカッション自体が授業の最終目的ではなく，授業内で行うグループ活動の準備として，ディスカッションを学ぶものである。そのため，定められたルールに基づくディスカッションを短期間で習得することを目的とする。実践の手順は，次のとおりである。

(1) ルールにしたがって，グループ・ディスカッションを行う。グループの人数は，３～４名とする。ディスカッションの長さは最長 30 分とし，音声の収録は FishWatchr で行うものとする（2.2 節で示した「リアルタイム・アノテーション（音声録音）」を用いる）。グループは複数あり，互いの活動がノイズとなりうるので，リフレクションに耐えうる品質で録音できるよう配慮する。なお，この段階では，追加機器として，マイクが必要になる。

(2) リフレクションの前に，ディスカッションの音声データをグループのメンバーに配布する。音声データの配布方法は，音声データのファイルサイズが mp3 形式でおおむね数十 MB 程度なので [11]，USB メモリなどで容易に可能であると考えられる。

(3) 各自が FishWatchr を用いてリフレクションする。リフレクション時は，自分の発言が規定のルールに従っているか評価し，評価結果をアノテーションする。この活動は，宿題などの授業外で行うものとする。ディスカッション中の発話の聞き取りには，聞き取りやすさや，周囲への配慮のことを考慮すると，追加機器として，ヘッドホンが必要になる。

(4) リフレクションで行ったアノテーション結果を教師に提出する。教師は，全グループのアノテーション結果をもとに，次回の授業で全員に対して，フィードバックする。学習者が教師に提出するのは，アノテーション結果（全学生分）と音声データ（グループ分）である。前述のとおり，アノテーション結果の統合は，グループごとに行う必要があるので，グループ数が多いと，教師の統合の手間は大きくなる。したがって，学生に提出させる場合は，音声データと全メンバー分のアノテーション結果を統合させてから提出させるのがよいだろう。

4　考察

本節では，アノテーションツール導入が授業に与える影響を，(a) 時間的観点，(b) 追加的機器類の観点から考察する。

まず，時間的な面での授業への影響について見てみると，アノテーションは「気づき優先型」「ルール提示型」の両方とも授業外で行うので，アノテーションによる授業自体への時間的影響は少ない。ただ,「気づき優先型」(2) のアノテーション結果の統合と (3) のグループ学習で，学習者が FishWatchr

[10] 使用予定のビデオは，約 26 分で約 1.9GB である。ビデオファイルの設定は，コーデック MPEG4-H264, 解像度 1440x1080, フレームレート 30fps，サンプリングレート 48kHz，ステレオである。

[11] CD と同等の品質で録音し，圧縮率が 1/10 とした場合

を円滑に使用できるかは，実践をやってみないとわからない。また，(2) については，(3) の前提になるので，授業でやるのではなく，事前にやっておいたほうが安全である。

次に，(b) の追加的機器類の観点から授業への影響を見てみる。二つの実践案において，必要となる追加的機器類は，次のとおりである。

- 「ルール提示型」(1) のマイク・収録環境，(3) のヘッドホン
- 「気づき優先型」のメディアファイルの配布手段

マイクとヘッドホンについては，収録時の周辺ノイズなどの影響や，聴取時の周囲への配慮を考慮すると，新規に用意するのは必要不可欠である。収録環境については，事前に収録データの品質を確認し，グループやマイクの配置を検討しなければならない。

メディアファイルの配布手段については,「気づき優先型」(1) で述べた方法を用いることができる。ただし，どのくらいの手間がかかるのかは，PC やネットワーク環境に依存するため，実環境において事前の確認が必要である。なお，もし，ネットワーク環境が整っているようであれば，メディアファイルの再生は，ストリーミングで行うことも考えられる。ただし，現時点では，FishWatchr はストリーミングでの再生に対応していないため，今後の課題である。

5　終わりに

本稿では，筆者らが開発している，ディスカッション観察システム FishWatchr をディスカション教育の実践に適用する二つの方法を提案した。提案した実践案では，学習者によるアノテーションが導入されにくい理由として，アノテーションツール導入が授業に与える影響の大きさを挙げ，既存の授業との差異を少なくするよう設計した。現在のところ，実際に実践は行っていないが，アノテーションツール導入が授業に与える影響を考察した。考察において示した未解決の課題については，今後，予備的な実験を行うことにより，明らかにする予定である。

参考文献

小川修史，小川弘判，掛川淳一，石田翼，森広浩一郎（2012）「動画アノテーションシステム VISCO を用いた協調的授業改善のケーススタディ」，日本教育工学会論文誌 35(4)，pp.321-329

大倉孝昭（2009）「授業ビデオ評価学習支援システムの開発と評価」，日本教育工学会論文誌 32(4)，pp.359-367

大塚裕子・森本郁代（2011）『話し合いトレーニング　伝える力・聴く力・問う力を育てる自律的対話入門』，ナカニシヤ出版

高垣マユミ（2010）『授業デザインの最前線 II』，北大路書房

Ahmed Mohamed Fahmy Yousef, Mohamed Amine Chatti, Ulrik Schroeder (2014), The State of Video-Based Learning: A Review and Future Perspectives, *International Journal On Advances in Life Sciences* 6(3/4), pp.122–135

Brugman, H., Russel, A. (2004). Annotating Multimedia/ Multi-modal resources with ELAN, *Proceedings of LREC 2004*

Elizabeth F. Barkley, K. Patricia Cross, Claire Howell Major 著，安永 悟 監訳 (2009)「協同学習の技法」，ナカニシヤ出版

Kathy A. Douglas, Josephine Lang, Meg Colasante (2014), The Challenges of Blended Learning Using a Media Annotation Tool, *Journal of University Teaching and Learning Practice* 11(2)

Peter Rich, Michael Hannafin (2009), Video Annotation Tools: Technologies to Scaffold, Structure, and Transform Teacher Reflection, *Journal of teacher education* 60(1), pp. 52–67

万葉集を対象とした原文と読み下し文のアライメント

山田 祐実 †　大村 舞 †　鴻野 知暁 ‡　Kevin Duh †
小木曽 智信 ‡　松本 裕治 †
(† 奈良先端科学技術大学院大学　‡ 国立国語研究所)

Word Alignment between Original Text and Its Reading in *Man'yōshū*

Yumi Yamada †, Mai Omura †, Tomoaki Kouno ‡, Kevin Duh †,
Toshinobu Ogiso ‡, Yuji Matsumoto †
(†Nara Institute of Science and Technology
‡ National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所で開発中の『日本語歴史コーパス』（CHJ）では，『萬葉集』の歌の原文を古文の読み下し文（訓読文）と関連付けて扱えるようにする予定である．しかし，人手でこの作業を行うには量が膨大であるため，自動で行えることが望ましい．本稿では，IBM モデルを用いた原文と訓読文の自動対応付け（アライメント）を行った．IBM モデルによる自動アライメントの結果，概ね正しい対応結果が得られることが分かった．これらの不適切な対応関係を自動で修正するために，読み仮名の情報を用いる手法を用いて対応のない訓読文側の文字を原文側の文字へ対応させた．また，品詞の情報を用いる手法により不要な対応が付いている訓読文側の格助詞の対応を除去した．これにより，アライメントの改善が見られた．今後の課題として，誤った対応の付いたものを正解の対応へ修正することが必要であることが分かった．

1　はじめに

日本語の歴史研究において校訂作業が行われた資料を扱う場合，校訂作業前の原文でどのように表記されていたかという情報は重要である．何故なら多くの場合，校訂作業後の古文の読み下し文（訓読文）は原文で書かれた資料から一段離れたものとなるためである．特に，『萬葉集』のように原文と訓読文の表層形が大きく離れている場合，原文を参照することは必須となる．たとえば，小学館『日本国語大辞典』などの研究用の辞典では，『萬葉集』の原文と訓読文の情報が両方明記されている．

現在国立国語研究所で開発中の『日本語歴史コーパス』（CHJ）[小木曽ら 2013] では，訓読

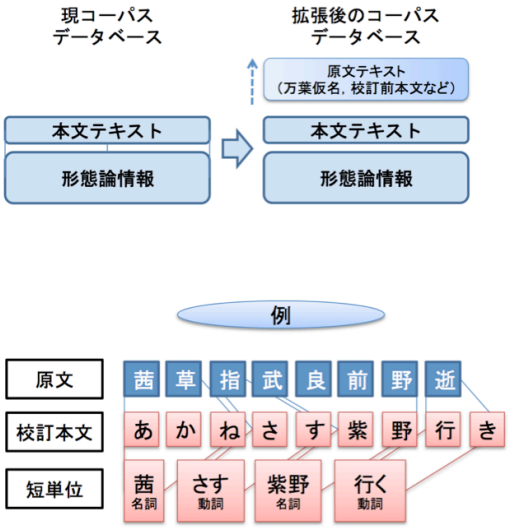


図 1: 日本語歴史コーパスの拡張

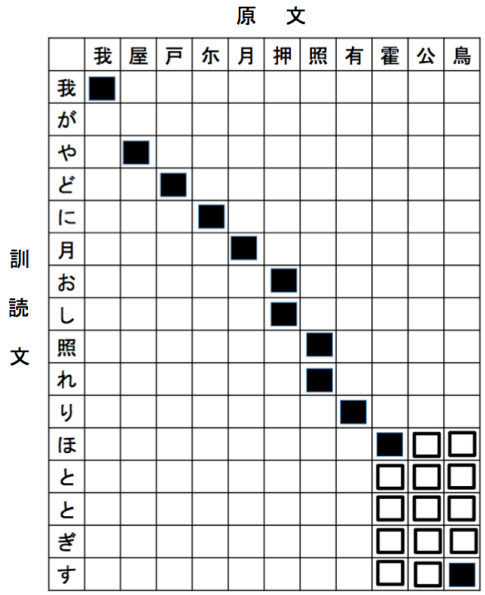


図 2: アライメントの例
黒四角と白抜きの四角がそれぞれ S アライメントと P アライメントを表す

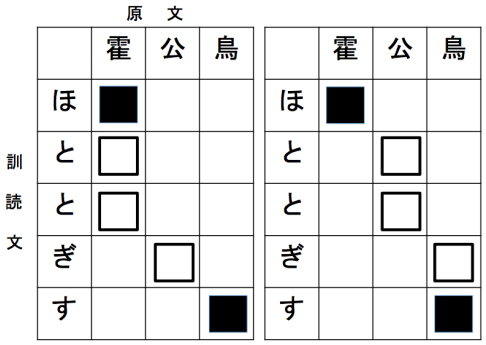


図 3: P アライメントによる評価

文を扱うことができる一方で，原文の情報は訓読文と関連付けて扱えていない．今後，図 1 のようにコーパス内に原文レイヤーを設け，『萬葉集』の原文の情報も同時に扱えるようにする予定である．しかし人手でアライメントを付与するには量が膨大であるため，自動で行えることが望ましい．そこで，本研究では訓読文と原文を文字単位で対応付け（アライメント）することによって，訓読文を原文と結びつけて扱えるようにする．

本稿では，『萬葉集』の歌の原文と古文による訓読文で語の対応付けを自動で行う手法を提案する．具体的には自動で語同士のアライメントを付与する手法である IBM モデル [Brown et al. 1993] を用いて語の対応付けを行う．

提案手法によって対応付けを行った結果，概ね正しい対応付けを行うことが分かった．しかし一方で，一部不適切な対応関係が見られた．本稿ではさらに，不適切な対応関係を自動で修正する手法について検討し追加実験を行った．

2　原文と読み下し文の対応付け

ここでは歌の原文と訓読文の対応付けについて概略を述べる．『萬葉集』における原文と訓読文の対応関係の例を図 2 に示す．図中に黒い四角で示した通り，基本的には原文一文字に対して訓読文一文字以上のアライメントを付与する．黒い四角で表したアライメントのように，対応に曖昧性がなく一つに決まるものを本稿では S アライメントとして表す．「我が」の「が」など，補読の助詞は対応する漢字がないものと見なす．「照れ-照」，「鳴き-鳴」のように，送り仮名の対応は原文の漢字一文字に含める．また，「ほととぎす-霍公鳥」，「とよもせ－響令」のように，文字同士の対応付けが正確には難しく，曖昧になるもの（熟字訓）も存在する．そのような曖昧な対応関係を本稿では P アライメントと表現する．図 2 の白抜きの四角のマスと黒い四角のマスの合計が P アライメントである．P アライメントを用いることで，図 3 の「ほととぎず-霍公鳥」の対応はいずれも正解であるとみなす．

3　IBM モデルを用いたアライメント（提案手法 1）

本章では，使用したコーパス及び IBM モデルを用いて語の対応付けを行う方法（提案手法 1）と，その結果について述べる．

3.1　使用したコーパス

『日本語歴史コーパス』の一部として収録予定の小学館『新編日本古典文学全集』の『萬葉集』（以下、CHJ 万葉集データ）を用いた．本研究では，CHJ 万葉集データの万葉仮名による原文と古文の訓読文を用いた．原文には，漢字の文字列に加え，書き下し文に戻すことができるように，レ点，上下点，一二点といった返り点がそれぞれの漢字に付けられている．

3.2　手法

ベースラインとして，自動アライメントの方法である IBM モデルを用いて実験を行った．IBM モデルは Brown ら (1993) が提案した機械翻訳のための手法である．IBM モデルは統計分布を基にして統計値を計算し自動的にアライメントを求める手法である．訓読文と原文のペアを入力として与えることで自動的にアライメント結果を得ることができる．

原文と訓読文は語順が違うため，語順の違いがアライメント結果に悪影響を与える可能性がある．そのため原文は IBM モデルに与える前に，コーパス中に記載されている返り点を基に語順を入れ替えておく．提案手法 1 では原文の語順を入れ替えた後に，IBM モデルに原文と訓読文を与えることで自動的にアライメント結果を得る．提案手法 1 のイメージを図 4 に示す．

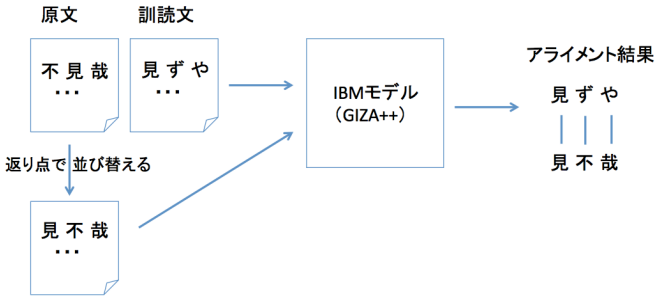


図 4: 提案手法 1 の流れ

表 1: 実験 1 評価値

	マイクロ F 値	適合率	再現率	出力アライメント数
並び替えなし	0.9478	0.9422	0.9522	1,384
並び替えあり	0.9589	0.9524	0.9654	1,386

3.3 実験設定

テストデータとして，CHJ 万葉集データからランダムに 50 首選んだ．本稿では選んだ 50 首のテストデータを対象にして実験，評価を行った．このテストデータには，正解の S アライメントの数は 1 ,360 箇所，P アライメントの数は 1,407 箇所存在する．IBM モデルの実装として，GIZA++ v1.0.7[Gao et al. 2008] を用いた．GIZA++ のパラメータは，デフォルト値に設定した．

IBM モデルでは，計算量を減らすため，翻訳される元の文 (原言語) から翻訳後の言語 (目的言語) への対応は最大 1 単語までという仮定を置いて統計値を計算している．原文から訓読文へは一文字以上の対応が付けられるが，その逆は殆ど起こらないことが分かった．そこで本稿では，原言語を訓読文，目的言語を原文として実験を行っている．

3.4 評価方法

評価のために，ランダムに選んだ 50 首の歌に対して人手で正解データを作成した．単語アライメントの評価値として F 値を用いる．F 値は以下の式のように適合率と再現率のマクロ平均として計算される．適合率と再現率は S アライメントの数 a_s と P アライメントの数 a_p に基づいて求められる．適合率はアライメント出力全体数 a に対する P アライメントの正解出力数 ($|a \cap a_p|$) の割合を表す．再現率は S アライメント出力全体数 a_s に対する S アライメントの正解出力数 ($|a \cap a_s|$) の割合を表す．P アライメントと S アライメントを用いることで，アライメントの曖昧性を考慮した評価をすることになる．

$$\text{F 値} = \frac{2 \times \text{適合率} \times \text{再現率}}{\text{適合率} + \text{再現率}}, \quad \text{適合率} = \frac{|a \cap a_p|}{|a|}, \quad \text{再現率} = \frac{|a \cap a_s|}{|a_s|}$$

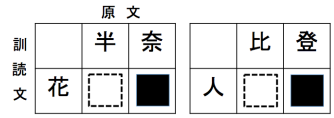


図 5: エラー 1
破線の四角は実際には対応が取れなかったアライメントを表す

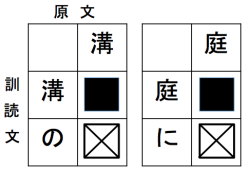


図 6: エラー 2
クロス記号の四角は誤って出力したアライメントを表す

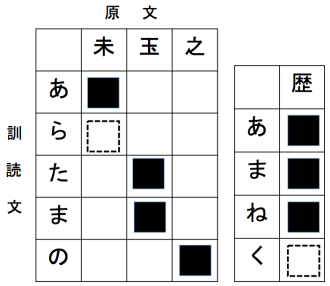


図 7: エラー 3

3.5　結果・考察（提案手法 1）

提案手法 1 の実験結果の評価値を表 1 に示す．比較のため，原文を並び替えずに IBM モデルを適応した手法 (並び替えなし) と原文を並び替えた後 IBM モデルを適応した手法 (並び替えあり) の結果を載せている．「並び替えあり」の方が，「並び替えなし」よりもアライメントの評価値が高いことが分かった．

ほとんどの場合で適切にアライメントができていることが分かったが，一部で不適切なアライメントが存在することが分かった．「並び替えあり」について，対応付けが正確にできなかったものは大きく分けて 3 種類に分類できた．一つ目のエラー（エラー 1）は，図 5 における「花—半奈」の「半」，「人—比登」の「比」のように，原文側から訓読文に対応するべき語が対応していないものである．全アライメント中，エラー 1 は 13 箇所あった．逆に，2 つ目のエラー（エラー 2）は，訓読文から原文へ対応がないにも関わらず，他の文字に誤った対応が付いているものである．エラー 2 は 40 箇所見つかった．図 6 に示したように，訓読文の補読語である「の」や「に」まで「溝」や「庭」に対応が付いてしまっている．3 つ目のエラー（エラー 3）は，図 7 のように，訓読文から原文へ対応するべき文字が誤った文字に対応している，もしくはどの文字にも対応していないものである．エラー 3 は 33 箇所あった．

4 節以降はエラー 1，エラー 2 について対処するための手法について説明する．エラー 1 を改善するために，読み仮名の情報を用いてアライメントを修正した．エラー 2 に対しては，品詞の情報を用いてアライメントの修正を行った．エラー 3 については，今回は対処方法を考案していないため，今後の課題とする．

4　読み仮名の情報を用いた手法（提案手法 2）

ここでは，3.5 節のエラー 1 で示した，読み仮名の情報を用いたアライメントの修正方法を提案し，実験結果について述べる．

4.1　手順（提案手法 2）

原文側から訓読文側へ対応すべき語の対応がないエラー 1 には読みの情報が有効であると考える．たとえば，図 5 で示したエラーの例を見ると，原文の「半奈」も訓読文の「花」も

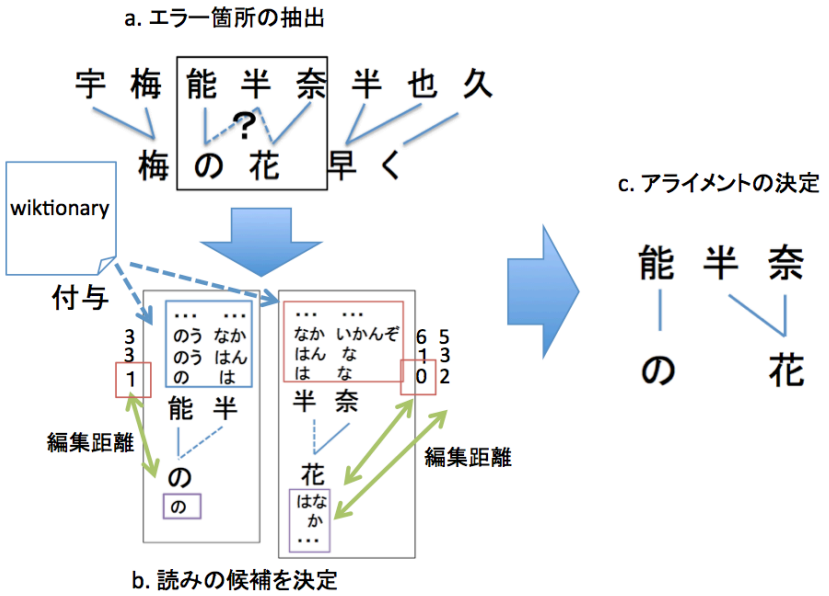


図 8: 読み仮名の情報を用いた手法

「はな」と読むことができる．「半」は「ハン」の他に「は」という読み方もできるためである．同様に，「比登」と「人」においても，どちらも「ひと」と読むことができる．このように，原文と訓読文に共通する読み仮名の情報を用いることで，訓読文の「花」に対応していない「半」も「はな」の一部であることが判断できる．そこで，エラー 1 に対しては読み仮名の情報を用いて対応付けの修正を行う．漢字に対応する読み仮名の情報は，ウェブ上で利用可能な Wiktionary[*1]から取得した．

エラー 1 を修正するためには，まず始めにエラーの箇所を自動で検出する必要がある．置き字（「焉」，「也」）の例外を除き，原文側の語から訓読文側の語へ必ず一文字以上の対応があることが分かっている．そこで，提案手法 1 の結果の中で原文側から訓読文へ一文字も対応を持たない原文の語があれば，そこにエラーがあるとみなした．この仮定に基づいて，対応語を持たない原文の語とその前後の語，及びそれにアライメントが付与されている訓読文の語を検出した．例えば図 8 の a では，原文側の「半」は訓読文側のいずれにも対応が存在していないため，「半」を修正対象として検出する．

次に，対応を持たない原文の文字がその前後一文字のどちらと単語を成すのかを決定する．これを決定することで，対応を持たない原文の文字を訓読文のどの文字へ対応付ければ良いかを決める．たとえば，図 8 の場合，「半」が前後の「能」と「奈」のどちらに付くかを判断することにより，訓読文側の「の」と「花」のどちらに対応付けられるかを決定する．つまり，「能半—の」と「半奈—花」のどちらの対応関係が適切かという問題になる．これを判断するために，各文字に付与した読み仮名の類似度をレーベンシュタイン距離 (以降編集距離) を用いて計算する．

編集距離は二つの文字列がどのくらい離れたものであるかを数値（コスト）を用いて表す指

[*1] Wiktionary：漢字索引 音訓 https://ja.wiktionary.org/wiki/Wiktionary:漢字索引_音訓

標である．二つの文字列のうち片方の文字列について，もう一方の文字列と等しくなるまで文字の挿入・削除・置換のいずれかの操作を繰り返す．一度の操作のコストは，挿入・削除では 1 点，置換は 2 点とし，コストの合計を二つの文字列間の編集距離とする．原文と訓読文の読みの候補の全ての組み合わせについて編集距離を計算する．

図 8 の b に示したように，Wiktionary を用いて原文と訓読文の漢字に全ての読みの候補を付与した．読みの全ての組み合わせについて編集距離を計算し，最小のものを前後それぞれの読みの候補として決定する．最後に前後について，最も小さいコストをとる対応関係を選ぶ．図 8 の b において，「の は」と「の」の編集距離と「はな」と「はな」の編集距離を比較すると 1 ＞ 0 となるので，編集距離のより小さい後者を対応関係として選ぶ．以上の手順を踏まえることで，図 8 の c のように「半 奈」と「花」が適切な対応として選ばれる[*2]．このように読みの情報を用いることで対応が取れなかった原文側の漢字のアライメントを得ることができる．

4.2 結果・考察（提案手法 2）

3.3 節で述べたランダムに選んだ 50 首について，原文側で対応がとれていなかった例は 13 個存在した．この対応が取れていなかった漢字に対して読みの編集距離を用いて適切なアライメントを付与した結果を表 2 に示す．二重山括弧で示したものが対応の取れていなかった漢字であり，山括弧で示したものが上述の方法で選ばれた適切なアライメントである．表 2 に示したように，13 個あるエラーのうち 11 個は改善された．改善されなかったものは，「真田葛―まくず」と「古保志―こほし」の 2 つである．この 2 つは前後の編集距離のスコアが同点になったため，この方法では適切な対応を選択することができなかった．「真田葛―まくず」のように，文字ごとに読みを対応させることが難しい熟字訓の場合，読み仮名を手掛かりに対応を取ることができないことがある．また，「古保志―こほし」は，訓読文の「恋し（こほし）」に対応する．「古保―恋」の対応が取れなかった理由は，現代語では「恋し」を「こほし」と読まないためと考えられる．Wiktionary の読み仮名の情報にも「レン，こい，こ」しか存在しない．このように，歴史的仮名遣いに応じた漢字によって原文と訓読文が対応している場合にも，この手法では限界がある．

5 品詞の情報を用いた手法（提案手法 3）

次に，3.5 節のエラー 2 で示した，原文へ対応のないはずの補読語が他の文字に対応付いてしまう誤りを修正するための手法について説明し，実験結果について述べる．

5.1 手順（提案手法 3）

修正するエラーの対象は，訓読文側の格助詞に原文側へ対応語があるとき，その原文の対応語が訓読文側に複数の対応語を持つ場合である．たとえば 3.5 節の図 6 に示した「溝の」の「の」や「庭に」の「に」などのように，訓読文側の格助詞から原文へ不適切な対応のある場合

[*2] スコアが同点の場合はどちらが適切か選択できないためアライメントを追加しない．

表 2: 読みの編集距離を用いた結果

	エラー	（前）編集距離のスコア	（後）編集距離のスコア
1	能《半》奈	の，のは　1	〈はな，はな　0〉
2	将《尓》焼	む，すすむそ　3	〈をや，そや　2〉
3	真《田》葛	ま，まや　1	つずら，やつずら　1
4	比《登》能	〈ひと，ひと　0〉	の，みの　1
5	伊《波》毛	〈いわ，いわ　0〉	も，わも　1
6	安《伎》也	〈あき，あき　0〉	やま，ぎやま　1
7	許呂《母》弖	〈ころも，ころも　0〉	て，もて　1
8	伊《麻》佐	〈いま，いま　0〉	さか，まさか　1
9	毛ゝ《等》利	も，もゆすりら　4	〈とり，とり　0〉
10	古《保》志	こ，こほ　1	し，ほし　1
11	岐《多》流	〈きた，きた　0〉	る，なる　1
12	知《可》豆	〈ちか，ちか　0〉	づ，べづ　1
13	芸《可》久	ぎ，ぎべ　1	〈かく，かく　0〉

が多いためである．このような格助詞の不適切な対応を修正するためには，はじめに訓読文側の格助詞を見分ける必要がある．

まず訓読文側で品詞の情報を得るために，MeCab v0.98 [Kudo et al. 2004] と中古和文 UniDic v1.4 [小木曽 2013] を用いて品詞の情報を付与した．その後，訓読文の文字の中で「格助詞」と判定された文字に対応しているアライメントを検出する (アライメント 1 とする)．その格助詞からアライメント 1 で対応している原文側のアライメントについて，複数のアライメントが付与されているか確認する．もし複数のアライメントがあった場合，アライメント 1 を除く．例えば，図 6 の「庭に―庭」の対応関係の場合，訓読文側の「に」は格助詞であり，原文側の「庭」に対応が付いている（アライメント 1)．次に，原文側の「庭」を見ると，訓読文側の「庭」にもアライメントが付与されていることが分かるため，アライメント 1 を除く．

また，実験設定として，提案手法 2 の読み仮名の情報を用いた対応付けを行った後に，提案手法 3 の品詞の情報を用いた対応付けを行った．逆に，提案手法 3 の後に提案手法 2 を行うことも試みた．

5.2　結果・考察（提案手法 3)

この節では，提案手法 3 の結果について述べる．また，提案手法 2 と 3 の二つの追加実験を行った後に適切な対応の取れていないエラーについて考察を行う．

まず提案手法 3 で品詞の情報を用いた結果，改善されたエラーの例を図 9 に示す．補読語の格助詞が原文の漢字に対応付いていた不適切な対応がなくなった．

読み仮名の情報を用いたものと，品詞の情報を用いたもの，その両方を用いたものの評価値の比較を表 3 に示す．F 値は読み仮名の情報を用いた後に品詞の情報を用いたものが最も高

表 3: 提案手法の比較
「並び替えあり」は返り点によって原文の漢字の文字列を並び替えたもの．「読みの情報のみ」は提案手法２のみを適用したもの．「品詞の情報のみ」は提案手法３のみを適用したもの．「品詞 → 読み」は提案手法３の後に提案手法２を適用したもの．「読み → 品詞」は提案手法２の後に提案手法３を適用したもの．

	マイクロ F 値	適合率	再現率	出力アライメント数
並び替えあり	0.9589	0.9524	0.9654	1,386
読みの情報のみ	0.9630	0.9528	0.9735	1,397
品詞の情報のみ	0.9665	0.9712	0.9618	1,354
品詞 → 読み	0.9688	0.9706	0.9669	1,362
読み → 品詞	0.9706	0.9714	0.9699	1,366

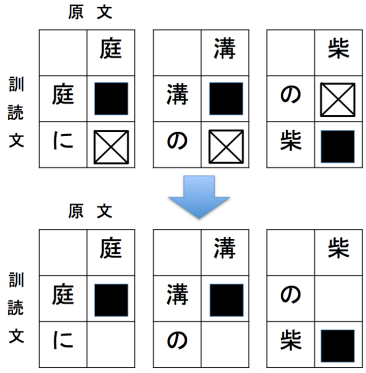


図 9: 提案手法３による改善例（上が改善前，下が改善後）

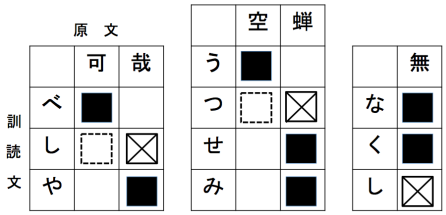


図 10: 今後改善の必要なアライメントの例

い．読みの情報のみを用いた場合，返り点による並び替えのみを行った結果よりも出力の S アライメント・P アライメントの数は 11 箇所増えた．これは，4.2 節で述べたように，原文側へ対応を持たない例の 13 箇所のうち 11 箇所が改善されたためである．

次に，品詞の情報を用いた場合と並び替えのみの結果を比較する．出力アライメント数は 32 箇所減り，S アライメント・P アライメントの数は 5 箇所減った．32 箇所の格助詞を除いたが，そのうちの 5 箇所は誤って除いてしまったことを意味している．格助詞の中でも除いてはいけない対応が 5 箇所存在していたためである．しかし，提案手法 2 と組み合わせることにより，対応が修正できたため，あまり最終的な結果に影響がでなかった．

また，以上の試みの後に，アライメントが改善されていない部分について分析を行った．エラーの例を図 10 に示す．これらの例は 3.5 節で触れたエラー 3 に分類でき，訓読文から原文へ誤った対応を持つものである．図 10 の「べしや-可哉」と「空蝉-うつせみ」は，対応する文字が不適切であるものを示す．「べしや-可哉」の例では，「可-べし」「哉-や」と対応すべきところが，「可-べ」「哉-しや」と対応付いている．「無-なく」は，助詞以外の語で不適切な対応を持つものである．これらの対応関係は複数の歌で出現する語であることから，IBM モデルに対して頻度が高い対応のペアを再度辞書として追加して制約を与えることにより，アライメントを改善する方法が考えられる．

6 まとめ

本稿では，IBM モデルの実装である GIZA++ を用いて『萬葉集』の原文と訓読文の文字単位での対応付けを行った．GIZA++ を用いる際，事前に返り点で原文の並び替えを行った方が並び替えを行わないものよりも評価値が高くなった．次に，原文側から訓読文側へ対応する語が対応付いていないエラーに対し，原文と訓読文それぞれの文字に読み仮名を割り当てて類似度を測ることにより，不適切な対応付けの修正を行った．最後に，訓読文側に品詞の情報を用いて格助詞の誤った対応付けを修正した．それぞれの修正により，アライメントは改善されたが，まだ不適切な対応関係が残っていることが分かった．今後は，対応関係に誤りのある文字を正しい対応関係へ修正する方法について考える必要がある．

参考文献

[Brown et al. 1993] Brown, Peter F., Vincent J. Della Pietra, and Stephen A. Della Pietra et al. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational linguistics* Vol. 19.2, pp.263-311

[Gao et al. 2008] Gao, Qin and Stephan Vogel (2008). Parallel Implementations of Word Alignment Tool. In *Proceedings of Software Engineering, Testing, and Quality Assurance for Natural Language Processing* (ACL2008), pp.49-57

[Kudo et al. 2004] Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing* (EMNLP2004), pp. 230-237

[小木曽 2013] 小木曽智信 (2013)「中古仮名文学作品の形態素解析」日本語の研究，9:4，pp.49-62

[小木曽ら 2013] 小木曽智信、須永哲矢、冨士池優美、他 (2013)「『日本語歴史コーパス 平安時代編』先行公開版について」第 3 回 コーパス日本語学ワークショップ予稿集，pp.269-276

ポスター発表（2） Bグループ

9月2日（水） 14:00～15:00

日英パラレルコーパスにみる日本語格外連体修飾形の訳され方

田辺　和子（日本女子大学文学部）†

Variation in Japanese-English Translation of Case-Outer Relative Clauses ～In the Case of Japanese-English Pararel Corpus～

Kazuko Tanabe (Japan Women's University)

要旨

本研究は、日英パラレルコーパス（中條・アンソニー：2013）を使って、日本語の格外連体修飾形がどのように英語に訳されるか分析したものである。その訳され方は、被修飾語（いわゆる底の名詞）と修飾節に格関係がないので、意味解釈によってさまざまな様式を採る。現在のところ、大きく分けて次の 5 つのタイプが抽出されている。例えば、①「(～する）事態」に対して、動詞を用いる。②「(～する）必要」に対して、助動詞および形容詞を用いる。③「(～する）動機」に対して、分詞構文で説明を加える。④名詞修飾節を作る。⑤まったく、該当する表現がなく文全体で状況を描写する。つまり、日本語の被修飾名詞に相当する英語の抽象名詞を用いるのではなく、何らかの動詞を用いて活動として表現する傾向が見られた。これは、Cassirer（1989）が述べるように、「日本語は、名詞的な型を厳密に形成して対象的な見方をする」特徴を表している。

1. はじめに

本研究は、第 6 回コーパス日本語学ワークショップ、ポスター発表「BCCWJ と日英パラレル新聞コーパスに基づいた格外連体修飾形の研究」、及び第 7 回口頭発表「BCCWJ に拠る名詞別格外連体修飾形の形成傾向の分析」の考察をふまえて、今回は、日英パラレルコーパス WebParaNews（中條・アンソニー：2013）を使って、格外連体修飾形の英訳のヴァリエーションの分類を試みたものである。

日本語の格外連体修飾形（いわゆる寺村（1992）のいう「外の関係」、すなわち「さんまを焼くにおい」という例のように、連体節の主名詞（底の名詞）「におい」が、連体修飾節内部の用言の「焼く」の補語としての格関係を持たない形）は、インド・ヨーロッパ言語と比較して、その特異性を指摘されている。

言語類型論者の Comrie（1996）は、「誰かがドアをたたく音」という日本語の例文を挙げ、‘the noise of someone knocking at the door’という英訳を示しながら、“Asian type”の名詞修飾形であると述べている。また、日本語全体の特徴としてドイツの言語哲学者 Cassirer は、『シンボル形式の哲学』（生松・木田訳、1989：pp.378 - 379）において、H．ヴィンクラーを引用し「(日本語は)・・・動詞的名詞をともなう単一の支配的な本名詞によって明確に表現されていることになる。」と記し、また、アルタイ語圏の諸言語の特徴として「文章構造の全体が、一つの対象的表現を単純に他の対象表現と並べ、付加語的にそれと結合する、

† tanabeka@fc.jwu.ac.jp

というように組み立てられる。」と述べている。

このような記述から、日本語の格外連体修飾形は、比較言語学的観点からその意味論的・語用論的結びつきを考察するに値するテーマとして意義のあるものと判断し、取りあげることにした。

2. パラレルコーパス画面

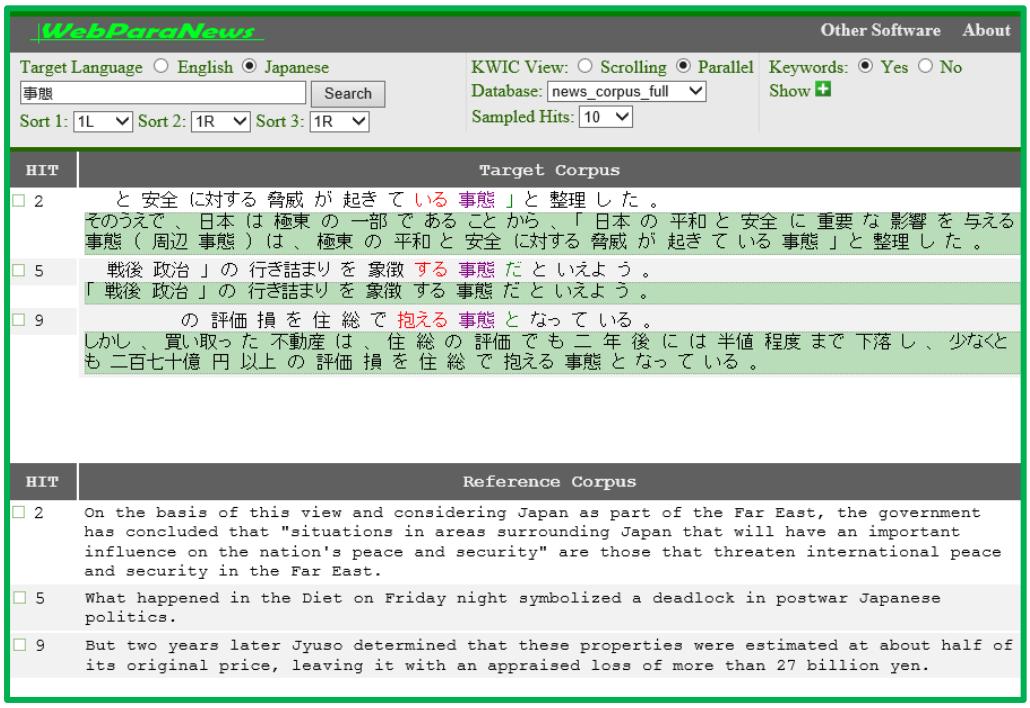


図１　WebParaNews による「事態」の画面

3. 日本語格外連体修飾節の英訳のヴァリエーション

本項で取り上げる「名詞」の選択は、寺村（1992）、大島（2010）の例文の中で取り上げられている名詞や、「連体修飾を形成しやすい普通名詞の順位表」（田邊：2015 p. 166）を参考にして選んだ。

3.1　動詞を用いる

日本語	英語
被害に遭った社員のうち、額が多い八人が詐欺容疑で元幹部を今月中にも捜査当局に告訴する**方針**で、金融監督庁と証券取引等監視委員会も損失補てんや詐欺行為の経緯について、同証券に報告を求める方向で検討を始めた。	The Financial Supervisory Agency and the Securities and Exchange Surveillance Commission, which have begun investigations, **plan** to ask Nikko how the former employee, a division chief, carried out allegedly fraudulent practices while at the securities company.
しかし、国内の販売不振と輸出の低迷で九八年の国内生産は百五十五万台まで落ち込み、昨年十二月には、能力を百七十万台まで15%減らす**方針**を打ち出していた。	However, the groups real output for 1998 fell to 1.55 million units due to sluggish domestic sales and decreased exports and the auto manufacturer last December **decided** to cut its production capacity by 15 percent to 1.7 million units.
警察庁は、これら捜査で教団の実像に迫る**方針**だ。	The NPA has already **ordered** local police authorities to pursue their investigations.
このうち、半数近い46%が九八年十月以降に初めて買い物をしており、一年あまりで急速にネットショッピングが普及している**様子**が浮き彫りになった。	The survey also **showed** 46 percent of them had placed their first orders on the Internet since October 1998, indicating a rapid increase in the number of Net shoppers over the year.
右下腹に痛みが残っているが歩行に苦痛はないため、炎症範囲が拡大している**様子**はない、というのが担当医師の所見。	Although he had some pain in his lower abdomen, it **did not impede** his walking and the inflammation had not spread, the doctors said.

図２　動詞を用いるもの

「方針」の名詞の英訳として‘decide’ ‘intend to’ 等の動詞が用いられる。「方針」を「方向性を定める」という意味解釈において、「決定する」という動詞が適切との判断からであろう。その他の例としては、「意見」では、‘favor’ ‘suggest’などの動詞が使われ、「事実」において

は、「事実を明らかにする」は、‘claim that～’「事実をかみしめる」は、‘consider that～’と訳されている。「様子」においては、‘appear to be~’が用いられている。

3.2　助動詞および形容詞を用いる

日本語	英語
これを補い、より完全なものにするためには、できるだけ早い時期に、しかも何度も繰り返し、見直し作業をする**必要**がある。	To reinforce the agreement, the protocol **must** be reviewed repeatedly.
一方、今回の派遣が、国際社会の日本への期待からすれば「小さな一歩」に過ぎない、ことを認識する**必要**がある。	On the other hand, Japan **must** recognize that participation in this program is only a small step towards realizing the expectations of the international community.
政府依存の姿勢からの脱皮を急ぐ**必要**がある。	They **must** do away with their mentality of depending on the government.
これが国会での安保論議を低調にし、コンセンサスづくりを遅らせている**原因**だ。	Their discord is partially **responsible for** the languid Diet debate on security matters, preventing a national consensus.

図３　助動詞および形容詞を用いるもの

「必要」の英訳例の多くに、助動詞‘must’が用いられている。日本語の「強い必要性がある状況」表現を、英語において「人間の行動の義務化」表現と転換するところが、日英二カ国語のそれぞれの特徴が表されている。また、「原因」の英訳として‘responsible for~’が用いられ、人間中心の問題の根源の「ありか」を明示する表現と転換されているのも同様な判断だと考える。また、人間に責任を負わせない場合でも「原因」は、‘due to~’と訳されている。

3.3　分詞構文を用いる

日本語	英語
しかし、買い取った不動産は、住総の評価でも二年後には半値程度まで下落し、少なくとも二百七十億円以上の評価損を住総で抱える**事態**となっている。	But two years later Jyuso determined that these properties were estimated at about half of its original price, **leaving** it with an appraised loss of more than 27 billion yen.
家族の要請を本人意思と推定できるとした被告・弁護側の主張に対しては「治療中止を求める**動機**となった患者の苦痛の性質などについて、家族は正確に把握しておらず、被告人も患者や家族との意思疎通がなかったため、患者の意思を推定することはできない」とした。	The defense claimed the family's request for euthanasia could be assumed to represent the desires of the patient, but the court ruled that "because the family did not accurately understand the nature of the patient's pain, **prompting** it to ask Tokunaga to terminate treatment, and the defendant did not communicate adequately with the patient and his family, the family could not have known the patient's true wishes."
大阪市などが今月六日に開いたフーリガン説明会では、パブの窓からイスを投げ出したり、火をつけたりして暴れ回る**様子**を、約四十人の商店主がビデオで見て言葉を失った。	After watching tapes of hooligans **throwing** chairs out of pub windows and setting fires, the 40 shop owners who attended the meeting were at a loss for words.

図４　分詞構文を用いるもの

表現形式の選択として、英訳では特定の動詞を分詞構文として用いるパターンもしばしば見受けられる。日本語における名詞修飾形の持つ状態的表現要素と、英語における行動的表現指向の折衷案として適当であるためだと推察する。図４では、「事態となっている」に‘leaving ~’、「動機となった」に対しては、‘prompting ~’、ここでの「様子」は、「フーリガン」の乱暴ぶりを表す目的で‘throwing ~’が用いられているのがわかる。

3.4　名詞修飾節を用いる

日本語	英語
営利企業と業務内容が競合する公益法人は営利法人への転換を指導するとともに、転換不可能な場合は三年以内に設立許可を取り消す**方針**を打ち出している。	It **suggests that** permits issued to nonprofit corporations be canceled after three years if the firms cannot become profit-making.
自民党山崎派会長の山崎拓・前政調会長は二十八日、読売新聞のインタビューに対し、九月の党総裁選で、集団的自衛権の行使を禁じた憲法九条の改正を公約として掲げる**方針**を明らかにした。	Taku Yamasaki, leader of the Liberal Democratic Party's Yamasaki faction, **said** Wednesday that his platform for the LDP presidential election, scheduled for September, will include a pledge to amend Article 9 of the Constitution, which prohibits Japan from exercising the right of collective self-defense.
警視庁は供述を始めた幹部から、爆発物を作った場所や時期、青島知事を狙った**動機**などについて、さらに事情を聞いている。	They plan to question him on the place and date the explosive was made and the **motives** for targeting Aoshima.

図５　名詞修飾節を用いるもの

このグループは、特に英訳において日本語原文と構文的にも、意味的にも大きな差異が見られない例である。「動機」においては、‘motives’とほぼ「動機」に相当する名詞で処理する例文もあった。「命令」においては、ほぼ直訳の語である‘order’が名詞または動詞かのいずれかで訳されていることが多かった。「案」においては、名詞では‘plan’が頻繁に用いられていた。

3.5　該当する部分が特にないもの

日本語	英語
「戦後政治」の行き詰まりを象徴する**事態**だといえよう。	What happened in the Diet on Friday night symbolized a deadlock in postwar Japanese politics.
国連PKOが単独で活動する**事態**もあるが、それと前後して、あるいは並行的に有志の国からなる多国籍軍が行動することがある。	In some cases U.N. peacekeeping operation units act alone in the countries concerned, but there are also cases in which multinational forces from volunteered countries work in parallel with the U.N. peacekeepers.
現在、湾岸の米軍兵力は約二十三万人だが、来年初めまでに四十万人前後に増強する**方針**である。	This would increase the strength of the U.S. forces, currently about 230,000, to as many as 400,000 by early next year.
米国での簿外取引で、約千百三十億円の巨額損失を出した**事件**をめぐり、米検察当局との司法取引で、罰金約三百五十八億円を支払った大和銀行（本店・大阪市）が、この罰金全額を課税対象とならない「損金」として処理、税務申告していたことが十二日、わかった。	Daiwa Bank deducted 35.8 billion yen in fines paid to the U.S. government from its taxable income by declaring the fines as a loss when it filed a tax return with the Osaka Regional Taxation Bureau, it was learned Monday.
ホテル前で客待ちをしていたタクシーの男性運転手（54）は「車がロビーに入っていったので、一瞬、目を疑った。映画の撮影かと思った」と興奮した**様子**で話していた。	"I could not believe my eyes when I saw the car drive into the lobby," said a 54-year-old taxi driver who had been waiting for a guest in front of the hotel at the time of the incident.

図６　該当する部分が特にないもの

文面上は、特に外の関係の底の名詞の該当部分の訳と思われる表現が認められない例文も少なくはない。「様子」「事態」「事件」などの名詞は、特に訳さないでも、その文全体が表現している状況を描写することができるからである。

4. まとめ

格外連体修飾形の英訳のされ方は、名詞修飾という枠組みを超えて、動詞および助動詞、形容詞などの用言に類するものに訳されることが多い。その名詞によって、同一の訳語や表現を用いられることが多いものと、訳のヴァリエーションが広いものとある。これらを全体的に考察すると、日本語が名詞を用いて、付帯的状況説明として表現する傾向があることに対して、ヨーロッパ諸語においては、動詞の持つ動的意味を中心に据える傾向があることが判明した。

謝　辞

本研究は、文部科学省科学研究費補助金、基盤（C）課題番号 25370496 (研究代表者：田辺和子) による補助を得ています。

文　献

Cassirer,Ernst.(1989)『シンボル形式の哲学（一）』岩波文庫

Chujo, K., K. Oghigian and S. Akasegawa, A Corpus and Grammatical Browsing System for Remedial EFL Learners. In Leńko-Szymańska, A. and A. Boulton (eds.), *Multiple Affordances of Language Corpora for Data-driven Learning*. pp. 109-128, Amsterdam: John Benjamins, 2015.

Comrie, Bernard. (1996) The unity of noun modifying clauses in Asian languages. *Pan-Asiatic Linguistics*: Proceedings of the Fourthe International Symposium on Languages and Linguistics, January 8-10, 1996, Volume 3, pp.1077-1088.

Comrie, Bernard. (1998) Rethinking the typology of relative clauses. *Language design*.　pp.59-86.

Kawaguchi, Yuji(eds.). (2007) *Corpus-Based Perspectives in Linguistics.* John Benjamins. Amsterdam/Philadelphia.

Matsumoto, Yoshiko. (1988) Semantics and pragmatics of noun-modifying constructions in Japanese. *Berkeley Linguistics Society* 14, pp.166-175.

大島資生（2010）『日本語連体修飾節構造の研究』ひつじ書房

田窪行則編（1994）『日本語の名詞修飾表現』くろしお出版

田邊和子（2015）「BCCWJ に拠る名詞別格外連体修飾形の成傾向の分析」『第 7 回コーパス日本語学ワークショップ予稿集』

寺村秀夫（1975-1978）「連体修飾のシンタクスと意味(1)-(4)」寺村(1992)所収

寺村秀夫（1992）『寺村秀夫論文集 I―日本語文法編―』くろしお出版

コーパスコンコーダンサ『ChaKi.NET』の「文書-部分構造行列」出力機能

浅原 正幸 (国立国語研究所) *
森田 敏生 (総和技研)

Document-Substructure Matrix Output Function on 'ChaKi.NET'

Masayuki Asahara (NINJAL)
Toshio Morita (Sowa Research Co., Ltd.)

要旨

コーパスを用いて統計処理を行う上で、「文書-単語行列」を作成をすることが多い。コーパスコンコーダンサ『ChaKi.NET』は従来より形態論情報に基づくクエリ Tag Search の Wordlist 機能を用いることにより、「文書-単語行列」を作成することが可能であった。今回この機能を拡張することにより、n-gram データや係り受け構造上の部分木などの「文書-部分構造行列」出力機能を実装した。さらに、既存の出力形式である Excel, CSV に加えて、R の dataframe 形式を出力できるようにした。ポスター発表では、当該機能のデモを行う。

1. はじめに

複数文書コーパスを用いて主成分分析や対応分析などの統計処理を行う際に「文書-単語行列」を作成をすることが多い（浅原ほか (2014)）。コーパスコンコーダンサ『ChaKi.NET』(Matsumoto et al. (2006))[(1)]は、Wordlist 機能を用いることにより文書-単語行列を容易に生成することができる[(2)]。特徴量空間として単一の単語表層形や語彙素のみならず、形態素系列（浅原ほか (2015)）や係り受け部分木（浅原・加藤 (2015)）などの部分構造データを用いることにより、より深い分析が行うことができる。しかしながら、部分構造データの枚挙においては、順列・組み合わせの枚挙といった煩雑な作業が伴う。プログラミングに不得手な研究者にとって、この作業が一つの障壁となっている。

今回『ChaKi.NET』の Wordlist 機能を拡張して、n-gram などの連続部分系列や連続部分木などを特徴量空間とする「文書-部分構造行列」を出力する機能を追加した。[(3)] 既存の出力形式である Excel 形式や CSV 形式に加えて、R の dataframe 形式を出力できるようにした。本稿では、これらの新機能を解説するとともに、非連続部分構造を枚挙する際の注意点について示す。

* masayu-a@ninjal.ac.jp

(1) http://osdn.jp/projects/chaki/

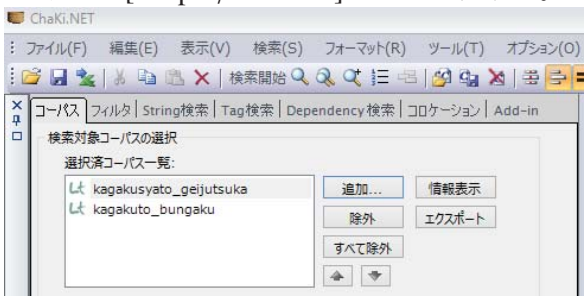
(2) http://qiita.com/masayu-a/items/66285bcb8d40c6bbb494

(3) ChaKi.NET 3.00β Revision 500

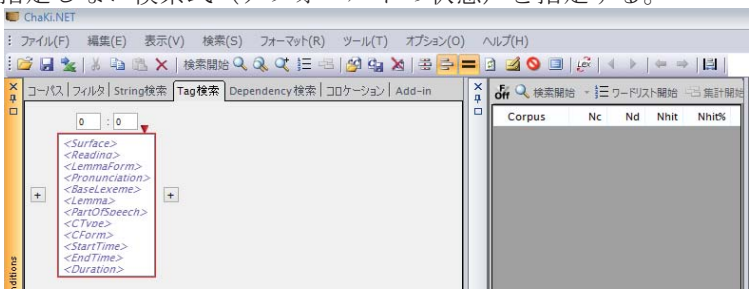
2. 『ChaKi.NET』の Wordlist 機能

最初に『ChaKi.NET』の Wordlist 機能を用いた「文書-単語行列」作成機能について解説する。あらかじめ分析対象のテキストを形態素解析器 MeCab などで解析して、ChaKi.NET 用の sqlite db ファイルを作成してあることを前提とする。後に述べる係り受け部分木に基づく分析を行う場合には、最初から係り受け解析器 CaboCha などで解析してあることが望ましい。(4)

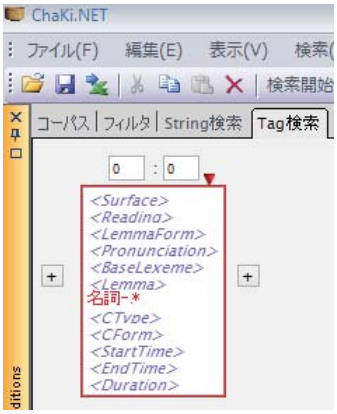
まず最初にコーパスを ChaKi.NET にコーパスを読み込ませる。sqlite db 化した複数ファイルを Search Conditions パネルの [Corpus/コーパス] タブに展開する。



Search Conditions パネルに [Tag Search/Tag 検索] タブを選択し、以下の図のように 1 形態素に対して何も指定しない検索式（デフォールトの状態）を指定する。



特徴量空間として、名詞しか定義しない場合には以下の図のように [PartOfSpeech] に名詞-* を選択する。検索窓が赤字の場合、当該検索クエリが正規表現であることを表す。



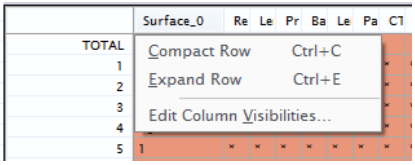
この状態で [Wordlist/ワードリスト開始] ボタンを押すと下図のように「文書-単語行列」が展開される。表中 1 列目から 9 列目が形態論情報を表す。10 列目、11 列目に選択したコーパ

(4) 複数のテキストファイルをバッチで係り受け解析を行い、sqlite db ファイルをに格納する方法については http://qiita.com/masayu-a/items/5e61dcf0ed7068c01f62 を参照すること。

スの頻度が示される。12 列目の [All] の列に全コーパスの頻度が示される。

	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL										1365	4978	6343	100
1	–	×	×	×	×	×	×	×	×	2	2	4	0.0630...
2	｜※[#「	×	×	×	×	×	×	×	×	1	0	1	0.0157...
3	2	×	×	×	×	×	×	×	×	1	0	1	0.0157...
4	4	×	×	×	×	×	×	×	×	1	0	1	0.0157...
5	5	×	×	×	×	×	×	×	×	1	0	1	0.0157...
6	68	×	×	×	×	×	×	×	×	1	0	1	0.0157...
7	あまり	×	×	×	×	×	×	×	×	1	0	1	0.0157...
8	アルキメーデス	×	×	×	×	×	×	×	×	1	0	1	0.0157...
9	いかん	×	×	×	×	×	×	×	×	1	3	4	0.0630...
10	インスピレーシ...	×	×	×	×	×	×	×	×	1	0	1	0.0157...
11	ヴォルテーア	×	×	×	×	×	×	×	×	1	0	1	0.0157...
12	うるか	×	×	×	×	×	×	×	×	1	2	3	0.0472...
13	エネルギー	×	×	×	×	×	×	×	×	1	3	4	0.0630...
14	かく	×	×	×	×	×	×	×	×	1	0	1	0.0157...
15	がら	×	×	×	×	×	×	×	×	1	6	7	0.1103...
16	キュービズム	×	×	×	×	×	×	×	×	1	0	1	0.0157...
17	こと	×	×	×	×	×	×	×	×	3	129	132	2.0810...
18	これ	×	×	×	×	×	×	×	×	20	41	61	0.9616...
19	これら	×	×	×	×	×	×	×	×	2	4	6	0.0945...
20	コンジェニアル	×	×	×	×	×	×	×	×	1	0	1	0.0157...

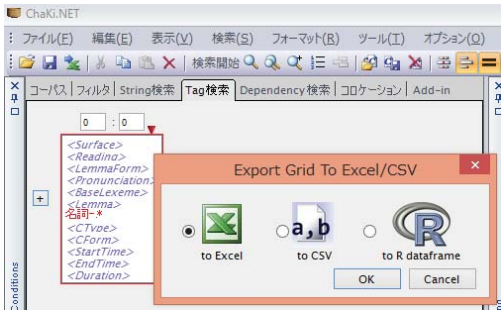
デフォールトの設定では形態素表層形のみが展開されている。各列のヘッダ部を右クリックすることにより、以下の図のように畳み込む [Compact Row Ctrl+C] か、展開する [Expand Row Ctrl+E] かが選択できる。



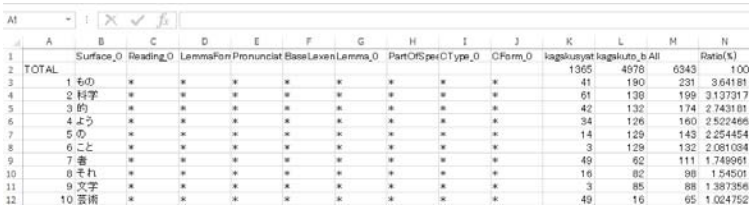
各列のヘッダ部を左ダブルクリックすることにより、当該列で昇順 → 降順にソートされる。以下の図は [All] 列 (全コーパス中の頻度) で降順ソートしたものである。

	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL										1365	4978	6343	100
1	もの	×	×	×	×	×	×	×	×	41	190	231	3.6418...
2	科学	×	×	×	×	×	×	×	×	61	138	199	3.1373...
3	的	×	×	×	×	×	×	×	×	42	132	174	2.7431...
4	よう	×	×	×	×	×	×	×	×	34	126	160	2.5224...
5	の	×	×	×	×	×	×	×	×	14	129	143	2.2544...
6	こと	×	×	×	×	×	×	×	×	3	129	132	2.0810...
7	者	×	×	×	×	×	×	×	×	49	62	111	1.7499...
8	それ	×	×	×	×	×	×	×	×	16	82	98	1.5450...
9	文学	×	×	×	×	×	×	×	×	3	85	88	1.3873...
10	芸術	×	×	×	×	×	×	×	×	49	16	65	1.0247...
11	場合	×	×	×	×	×	×	×	×	10	53	63	0.9932...
12	事	×	×	×	×	×	×	×	×	49	13	62	0.9774...
13	これ	×	×	×	×	×	×	×	×	20	41	61	0.9616...

この状態で [File/ファイル (E)] → [Send To Excel/CSV] を選択し、[to Excel] を選択するとと展開された「文書-単語行列」を保存することができる。尚、Microsoft Excel がインストールされていない機材の場合はこの機能が利用できない。



保存された Excel ファイルは以下のようになる。



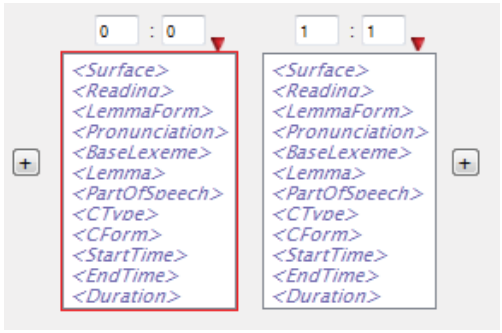
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1		Surface_0	Reading_0	LemmaForm	Pronunciat	BaseLexem	Lemma_0	PartOfSpee	CType_0	CForm_0	kagakusyat	kagakuto_b	All	Ratio(%)
2	TOTAL										1365	4978	6343	100
3	1	もの	*	*	*	*	*	*	*	*	41	190	231	3.64181
4	2	科学	*	*	*	*	*	*	*	*	61	138	199	3.137317
5	3	的	*	*	*	*	*	*	*	*	42	132	174	2.743181
6	4	よう	*	*	*	*	*	*	*	*	34	126	160	2.522466
7	5	の	*	*	*	*	*	*	*	*	14	129	143	2.254454
8	6	こと	*	*	*	*	*	*	*	*	3	129	132	2.081084
9	7	者	*	*	*	*	*	*	*	*	49	62	111	1.749961
10	8	それ	*	*	*	*	*	*	*	*	16	82	98	1.54501
11	9	文学	*	*	*	*	*	*	*	*	3	85	88	1.387356
12	10	芸術	*	*	*	*	*	*	*	*	49	16	65	1.024752

前の画面で [to CSV] を選択すると csv 形式のファイルが、[to R dataframe] を選択すると R 言語の dataframe 形式のファイルが出力される。

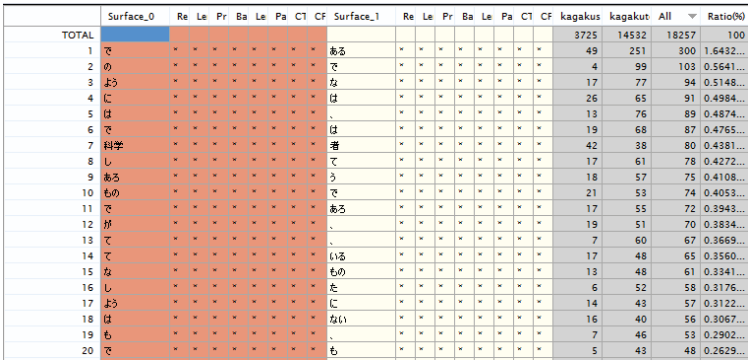
3.　文書-連続部分系列行列

以下では、文書-部分系列行列の展開方法について説明する。

Search Conditions パネルに [Tag Search/Tag 検索] タブを選択し、以下の図のように 2 形態素に対して何も指定しない検索式を指定することにより bigram 特徴量空間を考慮した文書-部分系列行列が展開できる。

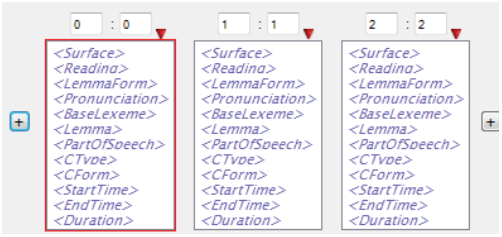


1 列目から 9 列目までが前件の形態論情報で、10 列目から 18 列目が後件の形態論情報である。19 列目以降に頻度情報が格納される。

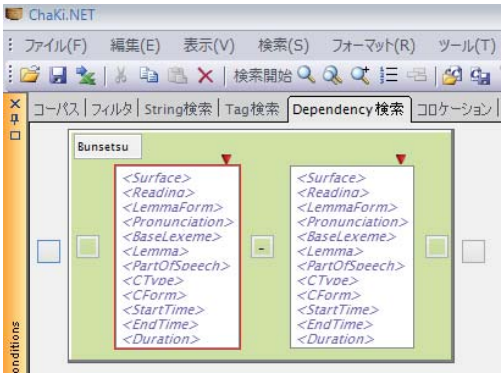


	Surface_0	Re	Le	Pr	Ba	Le	Pa	CT	CF	Surface_1	Re	Le	Pr	Ba	Le	Pa	CT	CF	kagakus	kagakut	All	Ratio(%)
TOTAL																			3725	14532	18257	100
1	で	*	*	*	*	*	*	*	*	ある	*	*	*	*	*	*	*	*	49	251	300	1.6432...
2	の	*	*	*	*	*	*	*	*	で	*	*	*	*	*	*	*	*	4	99	103	0.5641...
3	よう	*	*	*	*	*	*	*	*	な	*	*	*	*	*	*	*	*	17	77	94	0.5148...
4	に	*	*	*	*	*	*	*	*	は	*	*	*	*	*	*	*	*	26	65	91	0.4984...
5	は	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	13	76	89	0.4874...
6	で	*	*	*	*	*	*	*	*	は	*	*	*	*	*	*	*	*	19	68	87	0.4765...
7	科学	*	*	*	*	*	*	*	*	者	*	*	*	*	*	*	*	*	42	38	80	0.4381...
8	し	*	*	*	*	*	*	*	*	て	*	*	*	*	*	*	*	*	17	61	78	0.4272...
9	ある	*	*	*	*	*	*	*	*	う	*	*	*	*	*	*	*	*	18	57	75	0.4108...
10	もの	*	*	*	*	*	*	*	*	で	*	*	*	*	*	*	*	*	21	53	74	0.4053...
11	で	*	*	*	*	*	*	*	*	ある	*	*	*	*	*	*	*	*	17	55	72	0.3943...
12	が	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	19	51	70	0.3834...
13	て	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	7	60	67	0.3669...
14	て	*	*	*	*	*	*	*	*	いる	*	*	*	*	*	*	*	*	17	48	65	0.3560...
15	な	*	*	*	*	*	*	*	*	もの	*	*	*	*	*	*	*	*	13	48	61	0.3341...
16	し	*	*	*	*	*	*	*	*	た	*	*	*	*	*	*	*	*	6	52	58	0.3176...
17	よう	*	*	*	*	*	*	*	*	に	*	*	*	*	*	*	*	*	14	43	57	0.3122...
18	は	*	*	*	*	*	*	*	*	ない	*	*	*	*	*	*	*	*	16	40	56	0.3067...
19	も	*	*	*	*	*	*	*	*	、	*	*	*	*	*	*	*	*	7	46	53	0.2902...
20	で	*	*	*	*	*	*	*	*	も	*	*	*	*	*	*	*	*	5	43	48	0.2629...

trigram 以上の特徴量空間を規定するためには以下のように形態素の box を増やせばよい。



係り受け解析結果を格納することにより、文節境界の情報がデータベースに格納される。[Dependency Search/Dependency 検索] 機能を用いることにより、文節を越えない部分系列のみを展開することができる。以下の図は、文節内 bigram のみを特徴量とした文書-部分系列行列を展開するための式である。内側の形態素の boxes 間に - を入れることにより、2 形態素が隣接していることを表している。



4.　文書-非連続部分系列行列作成時の重複枚挙の問題

4.1　連続部分系列と非連続部分系列

前節では連続部分系列 (n-gram) を特徴量空間にした場合の「文書-部分系列行列」を展開する方法を述べた。本節では非連続部分系列 (p-mer) を特徴量空間にした場合の「文書-部分系列行列」の展開する方法と注意点について述べる。

非連続部分系列 (p-mer) とは、連続していないとびとびの部分列のことである。特に言及しない場合、非連続部分系列 (p-mer) は連続部分系列 (n-gram) を含むものとする。n-gram とは系列に対する長さ n の連続部分列 (substring) のことをいい，p-mer とは系列に対する長さ p の部分列 (subsequence) のことをいう。

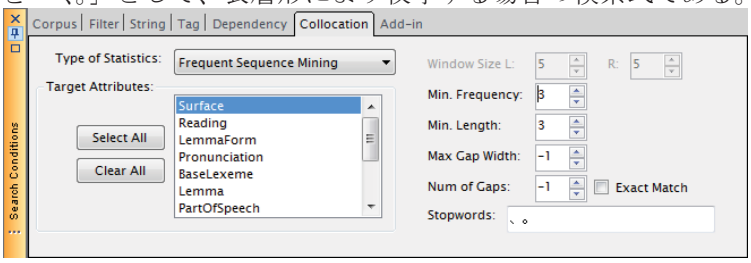
例えば “ABCDE” という系列に対して，3-gram は “ABC”, “BCD”, “CDE” の 3 種類あり，3-mer は “ABC”, “AB/D”, “AB/E”, “A/CD”, “A/C/E”, “A/DE”, “BCD”, “BC/E”, “B/DE”, “CDE” の 10 種類あり，それぞれ頻度は 1 である。p-mer の “/” は，そこにギャップがあることを意味している。

文全体にわたって非連続部分系列を枚挙する方法として、系列パターンマイニングアルゴリズム (Pei et al. (2001)) が知られている。ChaKi.NET には検索した文に対して、頻出系列パターンを枚挙する機能が実装されている。

4.2　既存の非連続部分系列枚挙機能

1 文書に対する非連続部分系列枚挙機能は以前から ChaKi.NET に実装されている。

[Search Condition] パネルから [Collocation/コロケーション] タブを選択し、[Type of Statistics] に "Frequent Sequence Mining"を選択することによって、頻出系列パターンの枚挙が行われる。以下の例では、最小頻度 3、最小系列長 3、最大ギャップ長 ∞、最大ギャップ数 ∞、ストップワードを 「、。」として、表層形により枚挙する場合の検索式である。

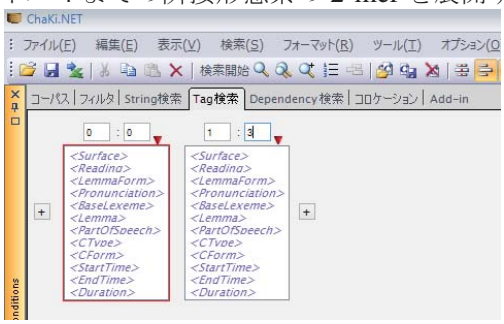


この手法では、1 文書毎に同じ作業を行う必要がある。

4.3　Wordlist 機能を用いた非連続部分系列枚挙

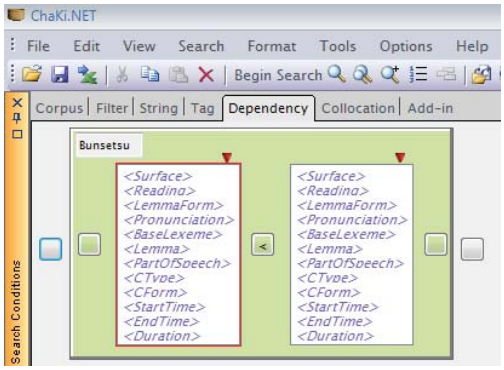
以下 Wordlist 機能を用いて、非連続部分系列を枚挙する方法について述べる。[Tag Search/Tag 検索] では、形態素の box の上についている index により、形態素の隣接性を規定することができる。

以下の例は Windows サイズ 4 までの隣接形態素の 2-mer を展開する検索式である。



Window サイズ n を広げると、各形態素位置に対して ${}_nC_p$ の組合せが展開されるので注意すること。

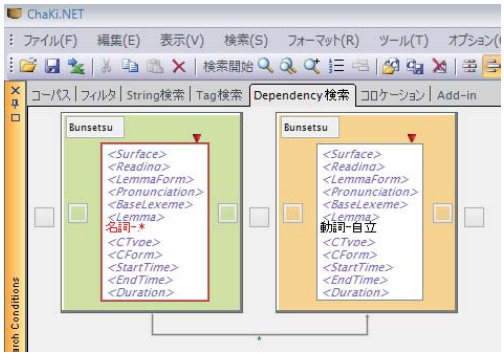
Window サイズを制限する他の方法として、文節境界により p-mer の枚挙を制限する方法がある。[Dependency Search/Dependency 検索] で以下の検索式を指定すると、文節内 2-mer を枚挙する。2 形態素 boxes 間の <は形態素の順序を規定する。この記号がない場合は、逆順についても枚挙してしまうので注意すること。



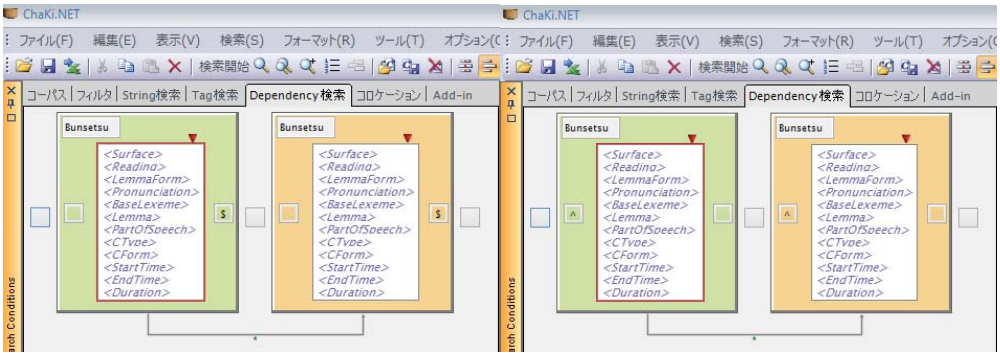
5. 文書-部分木行列

係り受け部分木を特徴量空間にする場合、[Dependency Search/Dependency 検索] を用いて Wordlist 機能を用いればよい。

以下の例では「動詞-自立」に係る「名詞」を枚挙する。しかし、文節内の形態素の位置を規定していないため、1 文節内に複数の名詞が存在する場合には、それぞれ別のものとして枚挙される。



残念ながら、文節内の形態素位置については先頭位置か末尾位置しか指定することができない。以下の例は各文節内形態素の出現位置を先頭位置にしたもの（左図）と末尾位置にしたもの（右図）である。



6. おわりに

本発表では、コーパスコンコーダンサ ChaKi.NET の「文書-部分構造行列」出力機能について紹介した。ChaKi.NET は他にも様々な機能がある (浅原・森田 (2013, 2014, 2015)) ので組み

合わせて利用されたい。

謝辞

本研究の一部は科研費基盤 (B)「言語コーパスに対する読文時間付与とその利用」(25284083)、科研費萌芽「近代語コーパスに対する統語情報アノテーション基準策定」(15K12888)、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

Matsumoto, Yuji, Masayuki Asahara, Kiyota Hashimoto, Yukio Tono, Akira Otani, and Toshio Morita (2006). “An annotated corpus management tool: Chaki.” *Proc. of LREC-2006*, pp. 1418–1421.

Pei, Jian, Jiawei Han, Behzad Mortazavi-Asi, Helen Pinto, Qiming Chen, Umeshwar Dayal, and Mei-Chun Hsu (2001). “PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth.” *Proceedings of the 17th International Conference on Data Engineering*, pp. 215–224.

浅原正幸・加藤祥 (2015).「文体指標を特徴づける係り受け部分木の抽出」 第 8 回コーパス日本語学ワークショップ.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2014).「文体指標と語彙の対応分析」 第 6 回コーパス日本語学ワークショップ, pp. 11–20.

浅原正幸・加藤祥・立花幸子・柏野和佳子 (2015).「文体指標と語彙系列の対応分析」 第 7 回コーパス日本語学ワークショップ, pp. 7–16.

浅原正幸・森田敏生 (2013).「コーパスコンコーダンサ『ChaKi.NET』の連続値データ型」 第 4 回コーパス日本語学ワークショップ, pp. 223–232.

浅原正幸・森田敏生 (2014).「コーパスコンコーダンサ『ChaKi.NET』の連続値データ型 (2)―読み時間の表示―」 第 5 回コーパス日本語学ワークショップ, pp. 39–48.

浅原正幸・森田敏生 (2015).「コーパスコンコーダンサ『ChaKi.NET』のプロジェクト機能」第 7 回コーパス日本語学ワークショップ, pp. 103–112.

現代日本語書き言葉均衡コーパス（BCCWJ）のコア・データに基づく関係節付加曖昧名詞句と先行文脈内の結束連鎖の分析

中野 陽子（関西学院大学）†

Cohesive Chains Formed between Noun Phrases Including Ambiguous Relative-Clause Attachments and the Preceding Context—Analyses of the Core Data of the Balanced Corpus of Contemporary Written Japanese

Yoko Nakano (Kwansei Gakuin University)

要旨

「黄色い服を着た少女の母親」のように関係節（下線部）が２つの名詞句（少女、少女の母親）のうち、どちらを修飾するのか曖昧な名詞句を関係節付加曖昧名詞句（関係節＋名詞句１の名詞句２）と呼ぶ。関係節付加曖昧名詞句とその先行文脈とのあいだの関係について、英語の関係節の非制限用法に基づいた想定はできるがコーパスに基づいた研究はされていない。そこで現代日本語書き言葉均衡コーパスのコア・データから関係節付加曖昧名詞句を含む分を抽出し、個々の事例毎に日本語母語話者２名に名詞句１、２と先行する談話とのあいだに形成される語彙的結束について、その種類を判定してもらった。判定結果を集計して関係節付加曖昧名詞句と先行文脈の関係を分類した。その結果、従来の先行研究では理論に基づき一種類しか仮定されていなかったが、この分析によって日本語では先行する談話と関係節付加曖昧名詞句との関係のパターンには数種あることが分かった。

１．はじめに

心理言語学の実験では文が単独で提示されることが多いが、日常生活で使われている文はテキストを構成している複数の文の１つとなっており、先行する他の文からの情報を参考に理解される。関係節付加曖昧名詞句を含んだ文の処理に関する心理言語学的研究も同じことが言える。関係節付加曖昧構文は実験の中では単独で提示されることが多い。関係節付加曖昧名詞句には構造的に曖昧な部分があるが、もし先行する他の文の情報があれば、その曖昧性を解消することができる。例えば、英語の関係節の制限用法は先行文脈内に関係節が修飾している名詞句の指示物と同じ種類のものが複数あることが前提となっているとき、その中のどれを指すのか特定するときに使われる。関係節付加曖昧名詞句で使われている関係節の用法は制限用法なので、テキストの中にあれば、先行する文の中に関係節の先行詞となっている名詞句と同じものまたは同等の語句があり、それが関係節の付加に関する曖昧性を解消すると考えられる。

下記の例（1）では下線部が関係節付加曖昧名詞句となっており、下線部のみを単独で読んでも、関係節の that liked swimming in the river が dog と boy のどちらを修飾しているのか曖昧である。しかし先行する文脈に 2 匹の犬がおり、その内の１匹が川で泳ぐのが好きであることが述べられている。先行する文と関係節付加曖昧名詞句との照応関係に整合性を持たせるために、関係節の that liked swimming in the river は boy ではなく、dog を修飾しているという解釈の方が自然である。

† y-k.nakanoATkwansei.ac.jp

(1) A boy had two dogs[a-1&b-1]. One dog liked swimming in the river and the other dog[b-2] liked running along the river bank. The boy's father walked the dog[b-3] of the boy that liked swimming in the river.

例 1 では先行する文内の名詞句(dog)が関係節付加曖昧名詞句内に繰り返し現れることで曖昧性が解消したのである。

１つのテキストの中に同じ名詞句、あるいは同等の語句が繰り返し現れると、それらの語句を含む文がお互いに関連付けられ、複数の文からなるテキストができる。このような関連付けを結束(cohesion)と呼び、語句の連なりは文の繋ぎの役割を果たしており結束連鎖(cohesive chain)と呼ばれる（Halliday & Hassan, 1976）。例（1）では同じ dog という語が繰り返されて文同士が関連付けられテキストを構成している。また dogs[a-1&b-1] と dog[b-2] と dog[b-3] とで結束連鎖が形成されている。結束連鎖を形成する語句と語句の関係は、同じ語句同士の関係に限らず、複数の種類に分類される（詳細は２．２を見てください）。

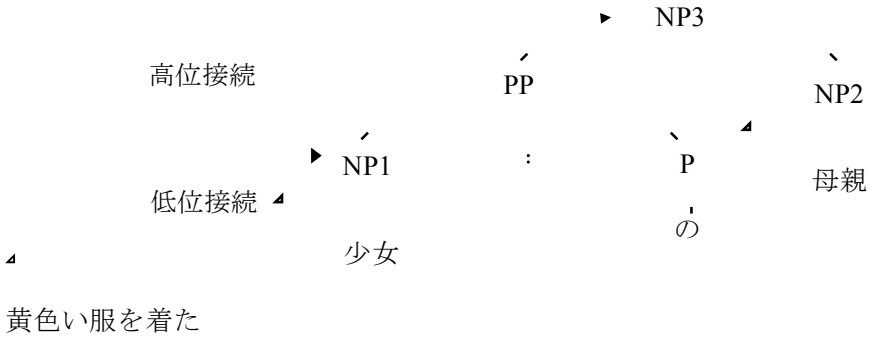
例 1 で見たように、関係節付加曖昧名詞句を含むテキストでは、関係節の先行詞である名詞句(dog)と、先行する文内に現れた同じ名詞句(dog)または同等の語句が含まれている。そこで本研究では BCCWJ のコア・データから関係節付加曖昧名詞句を含むテキストを抽出し、関係節付加曖昧名詞句がテキスト内で先行する文とどのような関係を結んでいるのか、またそれが関係節付加の曖昧性の解消に役立っているのかについて、結束連鎖の種類を分析することによって検討した。

２．背景

２．１　関係節付加曖昧名詞句

「黄色い服を着た少女の母親」のように関係節（下線部）が２つの名詞句（少女、少女の母親）のうち、どちらを修飾するのか曖昧な名詞句を関係節付加曖昧名詞句（関係節＋NP2 の NP1）と呼ぶ。関係節付加曖昧名詞句を樹形図にすると下記の例 2 のようになる。

(2) 黄色い服を着た少女の母親



階層的な樹形図における NP1 と NP2 の高さが異なることから、位置の低い方の NP1 への接続を低位接続（low attachment, LA）、高い方の NP2 への接続を高位接続（high attachment, HA）と呼ぶ。

関係節の接続に関する好みは言語によって異なるという先行研究がある。スペイン語と英語の関係節付加曖昧名詞句に関する調査ではスペイン語母語話者は高位接続を好む傾向が見られ、英語母語話者では低位接続を好む傾向が見られたと報告されている（Cuetos &

Mitchel, 1988）。この研究をきっかけにさまざまな言語で関係節付加曖昧名詞句における関係節の接続に関する好みの調査が行われた。日本語は高位接続が好まれることが示唆されている（Kamide & Mitchell, 1997；中野、早野、西内、井本, 2007）。日本語では関係節のあとに NP1 と NP2 が現れる。それと同じように中国語でも関係節のあとに NP1 と NP2 が出現するが中国語母語話者は高位接続を好むことが報告されている（Shen, 2006）。先行文脈の影響を調べた研究もいくつかある（フランス語：Zagar, et al. 2010; オランダ語：Desmet et al. 2002; ギリシャ語：Papadopoulou & Clahsen, 2006）。どの研究も文処理中の様子を調べる実験と関係節の接続に関する最終判断を調べる課題を実施している。先行文脈の影響があるかどうか文処理中の様子を調べる実験の結果は一致していない。これらの研究はさまざまな点で異なっており、オンラインの文処理の研究結果が異なる理由を特定するのは難しい。一方、関係節の接続に関する最終判断を調べる課題の結果は一致している。どの研究でも先行文脈の影響を受けて関係節の接続が選択される結果となっている。

(3) 低位接続文脈（複数のNP1、単数のNP2）

"L'audience allait débuter et on attendait le juge. Le public nombreux bavardait bruyamment et commentait l'affaire. La chanteuse [a-1] et ses avocats [b-1] se tenaient dans un coin du prétoire. Un journaliste a borda l'avocat [b-2] [N1] de la chanteuse [a-1] [N2] qui paraissait plus confiant(e) que les autres." (The hearing was about to begin and everyone was waiting for the judge. The audience was chatting noisily and talking about the case. The singer [female] and her barristers [male] were standing in a corner of the courtroom. A journalist approached the barrister [male N1] of the singer [female N2] who seemed more confident [feminine or masculine gender] than the others.)

(Zagar et al. 2010; p. 427)

Zagar らの実験で使われた例を見てみると、複数の弁護士（avocats[b-1])が先行文脈に登場するが、歌手は(La chanteuse)1 人だけである。一方、関係節付加曖昧名詞句（二重下線部）では歌手（la chanteuse [a-1]）と弁護士(l'avocat [b-2])が 1 人ずつ登場している。歌手は 1 人しかいないので関係節で特定しなくても指示対象が明確であるが、弁護士は複数いるので関係節の制限用法を用いて特定するとどの弁護士について言及しているのか明確になるため、文脈は低位接続を支持する文脈となっている。実際の実験では歌手を複数形にして弁護士を単数形にすることによって高位接続を支持する文脈条件も作られた。視線計測の実験では二重下線部のような完全な関係節付加曖昧構文が提示されたが、文完成課題では関係節の部分が空欄となっており、被験者が文を完成させるようになっていた。

上記の例 3 では先行文脈内の名詞句と関係節の先行詞が同じ名詞句であり、同じ名詞句の繰り返しで結束性連鎖が形成されている。ただし、文脈内の名詞の複数形であるのに対し、関係節の先行詞は同じ名詞の単数形であり、関係節の先行詞は意味上、文脈内の名詞の複数形に含まれる。

２．２　結束の種類と結束連鎖

結束には２つの種類がある―文法的結束と語彙的結束である。文法的結束は照応、置換、省略、接続などによって形成される。語彙的結束は繰り返しやコロケーションによって形成される（Halliday & Hassan, 1976）。本研究では関係節付加曖昧名詞句内の関係節の先行詞

と、先行するテキスト内にある先行詞と同じまたは同等の語句との間の関係を調べる。同じ語句または同等の語句の繰り返しを扱うため、本研究では語彙的結束の中の繰り返しを扱う。下記の例 4 a〜d のそれぞれで下線部の語が if 節の主語の he と同じものを指している。このように繰り返される語句は結束連鎖（cohesive chain）を形成する。4 つの例は結束連鎖を形成している語彙の種類という点で異なっている—同じ語の繰り返し（4a）、同意語または同意語に近いもの（4b）、上位語（4c）、一般的な用語（4d）。

(4) There is a boy climbing that tree.

a. The boy's going to fall if he doesn't take care.　（同一語）
b. The lad's going to fall if he doesn't take care.　（類義語）
c. The child's going to fall if he doesn't take care.　（上位語）
d. The idiot's going to fall if he doesn't take care.　（一般的な語）

(Halliday and Hassan, 1976; pp. 280-281)

本研究ではコーパスから「関係節＋NP1 の NP2」の名詞句を含むテキストを抽出し、テキスト内で NP1 と NP2 と結束連鎖を形成する語句が、どのような種類の結束連鎖を形成しているのか分類することによって、先行文脈の関係節の接続の曖昧性の解消への影響を調べた。

３．本研究

３．１　材料のサンプリング

現代日本語書き言葉均衡コーパス（BCCWJ）のコア・データから検索エンジンの中納言を用いて、「関係節＋NP1 の NP2」の名詞句を含むテキストを抽出し、各ジャンル毎の数を算出した。そのあと、分析に必要な適正サンプル数を計算し[1]、総数における各ジャンルの比率を変えないようにランダムに抽出した（表１）。

表１：抽出されたテキスト数と分析対象にしたテキスト数

ジャンル	新聞	雑誌	書籍	白書	Yahoo 知恵袋	Yahoo ブログ	合計
抽出数	572	133	133	89	173	169	1269
比率(%)	45	10	10	7	14	14	100
分析対象のテキスト数	134	30	30	21	42	42	299

本研究は「関係節＋NP1 の NP2」の名詞句とその前にあるテキストとの関係を調査対象としているため、「関係節＋NP1 の NP2」の名詞句の前にテキストが無い事例は分析の対象外として、その数のテキストを、残りのテキストからランダムに抽出した。また同じテキストが複数回サンプルに入った場合は１回と数え、分析適正数を満たせるように残りのテキストからランダムに抽出した。

[1] 下記の計算式が９５％信頼区間内に入るテキスト数の計算に用いられた。

$$n >= \frac{N}{\frac{N-1}{P(1-P)} \times \left(\frac{e}{Z}\right)^2 + 1}$$

N=the number of samples, P=0.5, e=0.05, Z=1.96

Yahoo 知恵袋と Yahoo ブログにも「関係節＋NP1 の NP2」の名詞句を含むテキストが入っていたが、テキストとして意味を成さない事例もあり、本研究では分析しないことにした。従って、新聞、雑誌、書籍、白書から抽出した事例を分析対象とした。

３．２　分析方法

日本語母語話者２名の判定者に、抽出されたテキストについて、関係節の接続傾向のほか、関係節付加曖昧名詞句「関係節+NP1 の NP2」に先行するテキストの中に、NP1、NP2、関係節の内容が記述されているかについて分野ごとに判定してもらった。また、先行するテキストにこれらの要素が記述されている場合は、これらの要素と関係節付加曖昧名詞句とのあいだの関係についても分類してもらった。判定者間の信頼度は各分野ごとに Cohen's Kappa が 0.8 以上であった（新聞：0.823、雑誌：0.871, 書籍：0.830, 白書：0.937）。表 4 以降の結束連鎖を形成する語の種類の分類には統計ソフトのエクセルを用いた。例えば、NP1 または NP2 の名詞と先行するテキストの繰り返されている語が同じかどうかコマンドを入力して検出し同一語を抽出して数を算出するようにした。

４　結果

４．１　NP1 及び NP2 に関する先行するテキスト内での言及

判定者に下記の３点について分析してもらった。

(1)「関係節＋NP1 の NP2」の名詞句に先行するテキスト内で NP1 についての言及があるか。
(2)「関係節＋NP1 の NP2」の名詞句に先行するテキスト内で NP2 についての言及があるか。
(3)関係節の先行詞は NP1(低位接続)または NP2(高位接続)のどちらか。

表 2：先行するテキスト内での NP1 と NP2 の言及と関係節の接続の比率（数）

NP1 と NP2 に関する言及	関係節の接続の選択		
	低位接続	高位接続	合計
どちらについても言及がない。	43(25)	57(33)	100(58)
NP1 についてのみ	44(27)	56(35)	100(62)
NP2 についてのみ	37(11)	63(19)	100(30)
NP1 と NP2 の両方	48(32)	52(35)	100(67)
合計	44(95)	56(122)	100(217)

日本語では限定用法または非限定用法であるかどうかは表記から判断することが難しく、NP1 と NP2 のどちらにも言及がなかった事例は非限定用法に該当する可能性がある。また中納言では先行文脈の語数が 500 字と限られている。この範囲外で言及があった可能性もある。「関係節＋NP1 の NP2」の名詞句に先行するテキスト内で NP1 にも NP2 にも言及がなかった事例では高位接続の方が低位接続よりもやや多かったが、上記のような点を考慮すると接続の傾向について断定することはできない。NP1 と NP2 のどちらか、または両方について言及がある事例では高位接続を選択する事例が多くなっているがどの場合もあまり大きな差はない。

４．２　関係節に関する先行するテキスト内での言及

判定者に「関係節＋NP1 の NP2」の名詞句に先行するテキスト内で関係節についての言及があるかどうかについて判定してもらい、その合計を算出した（表 3)。関係節についての言及がない場合の方が言及がある場合よりも多かった。

表 3：先行テキスト内での関係節に関する言及の比率（数）

関係節に関する言及なし		関係節に関する言及あり		合計
低位接続	高位接続	低位接続	高位接続	
27(59)	48 (104)	17(36)	8(18)	100(217)

４．３　繰り返しによる語彙結束を形成する NP1、NP2 及び先行するテキスト内名詞句

語彙的結束の繰り返しを同一語、類義語、上位語、一般的な語に分類した。「関係節＋NP1 の NP2」の名詞句内（例 5 の下線部:[黒潮が育てた$_{関係節}$ [漁船$^{a\text{-}8}{}_{NP1}$] の[民俗文化$_{NP2}$]]）で、NP1 または NP2 と同じ語が先行する文脈内にある場合は**同一語**（NP1=漁船$^{a\text{-}8}$と漁船$^{a\text{-}1}$)、NP1 または NP2 の類義語が先行する文脈内にある場合は**類義語**（NP1=漁船$^{a\text{-}8}$と船$^{a\text{-}2}$)、NP1 または NP2 の上位語が先行する文脈内にある場合は**上位語**（NP1=漁船$^{a\text{-}8}$と船舶（例 6 には含まれていない))、NP1 または NP2 の一般的な語が先行する文脈内にある場合は**一般的な語**（NP1=漁船$^{a\text{-}8}$と海生丸$^{a\text{-}3}$、漁生丸$^{a\text{-}4}$、正丸$^{a\text{-}5}$、直美丸$^{a\text{-}6}$、美衣丸$^{a\text{-}7}$）とした。

(5) 岸壁につながれた漁船$^{a\text{-}1}$は、よく見ると、どれもこれも「眼のある船$^{a\text{-}2}$」だった。海生丸$^{a\text{-}3}$、漁生丸$^{a\text{-}4}$、正丸$^{a\text{-}5}$、直美丸$^{a\text{-}6}$、美衣丸$^{a\text{-}7}$…、みんな舳に可愛い眼が付いていた。　種子島にはこれまで何度も訪れていたが、気が付かなかった。[黒潮が育てた$_{関係節}$]漁船$^{a\text{-}8}$ (NP1) の民俗文化(NP2)が、

語彙的結束連鎖が形成されている事例について、その種類を分類したところ（表 4)、同じ語、類義語、上位語、一般的な語の 4 種類は、それぞれ 38.39%, 4.52%, 15.81%, 41.29%の比率となり、NP1 と NP2 の同じ語を繰り返す、または一般的な語に言い換える比率が高いことが分かった。更に種類毎に関係節の接続が高位接続か低位接続かについて分類した(表 5)。

表 4：繰り返しの語彙の種類の比率（数）

繰り返しの語	NP1	NP2	合計
同一語	32(38)	68(81)	100(119)
類義語	43(6)	57(8)	100(14)
上位語	45(22)	55(27)	100(49)
一般的な語	56(72)	44(56)	100(128)
合計	45(138)	55(172)	100(310)

表 5：繰り返しの語彙の種類と関係節の接続の比率（数）

繰り返しの語	繰り返されている語句	関係節の接続		合計
		低位接続	高位接続	
同一語	NP1	46(13)	54(15)	100(28)
	NP2	44(35)	56(45)	100(80)
類義語	NP1	39(7)	61(11)	100(18)

	NP2	100(2)	0(0)	100(2)
上位語	NP1	43(9)	57(12)	100(21)
	NP2	37(10)	63(17)	100(27)
一般的な語	NP1	43(20)	57(26)	100(46)
	NP2	45(37)	55(45)	100(82)
合計		37(113)	44(133)	56(171)

先行研究では関係節の制限用法は先行文脈に同じ種類のものが２つ以上あり、そのうちどれを指しているのか明示するために使われることが前提となっているが、先行文脈内の語と NP1 または NP2 の関係を分析したところ当てはまらない事例も多くあった。例えば、下記の例 6 では、先行文脈は過去から現在の日本の農業の様子を記述しており、農業界 b-19 は上位語として先行文脈全体を指し、[[逆風の吹く a-6 関係節] [[日本 b-15 NP1]の[農業界 c-19 NP2] NP3] NP4]は現在の日本の農業の様子を総括している。農業界の一部を指すのではなく、全体を総括する表現として関係節付加曖昧名詞句が使われている。このような例から関係節の先行詞が上位語、先行文脈内の語が下位語の事例もあり、先行研究で想定されている以外の語彙的結束性の連鎖が形成されていることがわかった。そこで表 6 のように先行文脈の語句が一般的な語を、NP1 または NP2 の下位語となっている場合と同じレベルの語である場合とに分類した。

(6) 社説　二千一・２・二十一　【中日 b-1 農業 c-1 賞】危機 a-1 突破に若者の力　中日 b-2 農業 c-2 賞が第六十回を機に衣替えし、若い農家 c-3 に絞って顕彰することになった。日本 b-3 の農業 c-4 危機 a-2 突破の力となることを期待する。　三十数年前、ちゃぶ台にこぼれたわずかなご飯粒 c-5 を「もったいない」と言いつつ口に運んだ時代、農業 c-6 はまだ国 b-4 の基幹的な産業 c-7 であった。が「飽食の時代」と呼ばれる今、その c-8 存在は、とかく軽く見られがちである。　そんな時代に、中部地方 b-5 の農業者 c-9 を顕彰する中日 b-6 農業 c-10 賞は審査対象年齢を四十歳以下に絞り、二十一世紀を担う若い農家 c-11 を励ますことになった。　背景に、日本 b-7 の農業 c-12 に対する危機 a-3 感がある。何よりも、国際 b-8 競争の激化 a-4 が日本 b-9 の農業 c-13 を揺さぶっている。安い労賃や、広大で安価な土地で生み出される海外 b-10 の農作物 c-14 が輸入解禁となり、宿命的な悪条件下 a-5 で作られる国産 b-11 農作物 c-15 を駆逐しつつある。　とくに国際 b-12 分業論を信奉する人々は、生産性の低い日本 b-13 の農業 c-16 そのもの a-17 を経済発展の足手まといととらえ a-6「日本 b-14 に農業 c-18 はいらない」とまで述べている。まさに[[逆風の吹く a-6 関係節] [[日本 b-15 NP1] の[農業界 c-19 NP2] NP3] NP4]であり、。。。

NP1 または NP2 が先行する文脈内で繰り返されている語にとって、どのような関係にあたるかを分類し、その数を算出した（表 6）。先行文脈の語が例 6 の海生丸 a-3 で NP1 がその総称で「船」や「漁船」なら**下位語**とした。

表 6：結束連鎖を形成する繰り返される名詞句の種類の数

繰り返しの語	語彙の種類	関係節の接続	低位接続	高位接続	合計	合計
一般的な語	同レベルの語	NP1	18	13	31	67
		NP2	7	29	36	

	下位語	NP1	19	22	41	61
		NP2	17	3	20	
上位語		NP1	9	12	21	48
		NP2	10	17	27	
合計			80	96	176	176

５．まとめ

本研究は現代日本語書き言葉均衡コーパスのコア・データから関係節付加曖昧名詞句を含むテキストを抽出し、関係節付加曖昧名詞句とそれに先行するテキストの部分とで形成されている結束連鎖を分析した。その結果、心理言語学の先行研究で想定していた結束は同一語の繰り返しで成立されるもののみだったが、多くの種類の結束連鎖があることが分かった。表 5 を見ると同一語では高位接続の方が低位接続より多くなっており、関係節付加曖昧構文を単独で提示している研究の結果と一致する。一方、表 5 や表 6 で NP1 や NP2 と結束連鎖を形成している他の種類の語を見ると、必ずしも高位接続が低位接続より多くなってはいない。したがって、文脈情報が関係節の接続の選択にどのように影響するか、心理言語学的研究を行った場合、従来よりも複雑な仕組みが明らかになる可能性がある。コーパスから得られるデータに基づいた研究と心理言語学的な実験から得られるデータに基づいた研究の成果を合わせていくとより発展的な研究ができる可能性がある。

謝　辞

本研究は、喜田桃世さん、近藤眞樹さん、西本優さんにご協力をいただきました。また、科学研究費補助金基盤（C）（代表者：中野陽子　No. 24520484）による補助を得ています。ここに記して感謝の意を表します。

文　献

Cuetos, F., and Mitchell, D.C. (1985). Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish. *Cognition, 30*, 73-105.

Desmet, T. Baecke, C. D., and Brysbert, M. (2002). The influence of referential discourse context on modifier attachment in Dutch *Memory & Cognition, 30*, 150-157.

Halliday, M. A. K., and Ruqaiya Hasan. 1976.*Cohesion in English.* London: Longman.

Kamide, Y., & Mitchell, D.C. (1997). Relative clause attachment: Non-determinism in Japanese parsing. *Journal of Psycholinguistic Research, 26*, 247-254.

Papadopoulou, D., and Clahsen, H. (2006). Ambiguity resolution in sentence processing: the role of lexical and contextual information. *Journal of Linguistics, 42*, 109−138.

Zagar, D., Pynte, J., and Rativeau, S. (1997). Evidence for Early closure Attachment on First pass Reading Times in French. T*he Quarterly Journal of Experimental Psychology Section A, 50*, 421-438.

中野陽子、早野賢譲、西内万貴、井本智子（2007）中国人留学生の第二言語としての日本語における関係節付加曖昧構文の処理について　国際社会文化研究第 8 号 109-126.

教科書コーパスを利用した難易度別コロケーション辞書の提案

李在鎬（筑波大学）†
佐々木馨（国際交流基金）

Proposal of Collocation Dictionary Based on the Textbook Corpus Analysis

Jae-ho Lee (University of Tsukuba)
Kaori Sasaki (Japan Foundation)

要旨

近年，コミュニケーション能力を重視した言語教育の必要性が指摘されているが，形態素解析などで使用する言語単位（短単位）は，言語教育における単位としては不十分と言わざるを得ない。コミュニケーション能力の育成をはかるためには，実質的な意味機能が担える単位が必要であり，また，学習者の習熟度に応じた網羅的な表現のリストが必要であるが，こうしたリストは存在しない。そこで，本研究では，日本語リーダビリティシステムの構築のために利用した「レベル別コーパス」（文章の難易度がアノテーションされたコーパス，60 万語規模）をもとに，N-gram データを作成したあと，コロケーション表現を抽出した。抽出の結果として，8,121 項目のリストが完成した。各項目は，「レベル別コーパス」での出現頻度を差異係数で処理し，初級レベルとして 3,903 項目，初中級レベルとして 1,472 項目，中級レベルとして 2,746 項目を抽出した，現在，人手で確認作業をすすめており，来年度の春に公開する予定である。本発表はその中間報告である。

１．研究背景と目的

日本語教育研究においてコーパスを利用する意味は，次のように要約できる。コーパスは，個人単位の言語直感では得られない一般的レベルの言語の使用実態を明らかにできる。そのため，コーパスを利用することで，汎用性のある言語教育コンテンツが作成できる。

コーパスの利用範囲は非常に広く，日頃の教育活動での利用はもちろんのこと，教材開発や辞書開発などの汎用的な教育コンテンツの作成において，重要な資料になり得る（具体的な利用例は李・石川・砂川 2012，中俣 2014，本田(他)編 2014，庵・山内 2015 参照）。

しかし，コーパスは，生の言語使用データであるため，そのままの形では言語教育の場に持ち込めない。とりわけ，語彙や文法表現などの言語的素材が持つ潜在的な難易度に対する配慮が必要である。学習者の理解度や習熟度に応じた難易度の調整がなされてこそ，十分な教育効果が期待できる(李 2011)。こうしたことから，学習者に提示する学習コンテンツに関しては難易度に関する調整が常に必要になる。例えば，「日本語教育語彙表」（http://jhlee.sakura.ne.jp/JEV.html，Sunakawa et al.(2012)）では，均衡コーパスと日本語教材コーパスをもとに 17,920 語の語彙表を作成しているが，それには，日本語教師の主観判定に基づく難易度情報が入っており，すべての単語が初級前半，初級後半，中級前半，中級後半，上級前半，上級後半のいずれかにカテゴリー化されている。

さて，本研究は，「日本語教育語彙表」の拡張として，日本語のコロケーション辞書構築

† jhlee.n@gmail.com

を目的とする。具体的な課題としては，1）日本語教科書コーパスをもとに共起語（機能語，内容語問わず）に関する網羅的調査を行うこと，2）語形に関する網羅的調査を行うことを目的とする。

2．データと方法

日本語学習における学習効果を考えた場合，難易度に関するアノテーションは不可欠と言える。しかし，コロケーション表現の難易度を決めるのは，容易ではない。その一番の理由として，コロケーション表現の難易度は単語の難易度から直接予測することができない。例えば，「歌」と「読む」は，「日本語教育語彙表」で調べるといずれも初級前半の語彙である。しかし，この２つがコロケーションを作り，「歌を読む（一般的には「詠む」と表記する）」となった場合，初級の表現としては明らかに違和感がある。同じことが，「日記」と「つける」は中級前半の単語であるが，「日記をつける」になると，さらに難易度があがる。こうした問題を考えた場合，コロケーション表現そのものに対して，何らかの難易度を付与すべきと考える。しかし，その作業には膨大な労力を要する。

これを踏まえ，本研究では，日本語教科書コーパスをもとに構築した「レベル別コーパス」（Lee et al. 2015 in press）を利用することで作業の効率化をはかった。具体的には，難易度判別に代わるものとして，「レベル別コーパス」での出現頻度をもとに，差異係数を計算し，差異係数の値をもとに難易度を決めるという方法論を使用した。なお，「レベル別コーパス」とは，リーダビリティシステムを構築するためのトレーニングデータであり，日本語の教科書データと BCCWJ を利用して構築したものである。コーパスサイズは，以下のとおりである。

表 1.　「レベル別コーパス」のコーパスサイズ

	初級前半	初級後半	中級前半	中級後半	上級前半	上級後半
異なり語	3,178	2,858	5,156	10,291	6,833	4,712
延べ語	72,691	68,746	87,433	174,953	69,268	122,269

単位：UniDic に基づく短単位

表１における６スケールのレベルイメージは，以下のとおりである。

表 2.　６スケールのレベルイメージ

レベル	レベルイメージ
初級前半	単文を中心とする基礎的日本語表現に関して理解できる．複文や連体修飾構造などの複雑な文構造は理解できない。
初級後半	基本的な語彙や文法項目について理解できる．テ形による基本的な複文なども理解できる。
中級前半	比較的平易な文章に対する理解力があり，ある程度まとまった文章でも内容が把握できる。
中級後半	やや専門的な文章でも大まかな内容理解ができ，日常生活レベルの文章理解においてはほぼ不自由がなく遂行できる。
上級前半	専門的な文章に関してもほぼ理解できる．文芸作品などに見られる複雑な構造についても理解できる。
上級後半	高度に専門的な文章に関しても不自由なく，理解できる．日本語のあらゆるテキストに対して困難を感じない。

本研究が目指すコロケーション表現の抽出も，最終的には表 2 のレベルイメージに準拠することを目指すが，現時点では，初級，初中級，中級の 3 レベルのものとして整理している。

さて，本研究では，とりわけニーズが高いと思われる初級と中級レベルのコロケーション辞書を作成する目的で，表 1 の初級前半～中級後半のデータを利用し，N-gram によるコロケーション表現の抽出を試みた。具体的には，以下の手順で作業を行った。

- ステップ 1.「レベル別コーパ」スの中から初級前半～中級後半のデータを MeCab 0.996 ＋UniDic 2.2.0.1 で解析する。
- ステップ 2. 形態素解析済みデータに対して 3gram～6gram の連結データを作成する。
- ステップ 3. 連結データを集計し，サブコーパス別および合計出現頻度を計算する。
- ステップ 4. 合計出現頻度 5 以上のものを絞り込む
- ステップ 5. サブコーパスによる差異係数を計算し，レベルを決める。

３．結果

ステップ 1 の結果，403,823 語のデータが得られた。ステップ 2 の結果，75,668 項目のデータが得られた。ステップ 3・4 の結果，8,121 項目のデータが得られた。見出し語の例と見出し語の数を表 3 に示す。

表 3. N-gram による見出し語の数と実例

	見出し語数	見出し語例
3gram	4994	ています/ありません/と思います/ても良い/た事が/になった
4gram	2117	というのは/しています/かもしれない/がありますか/ことができます
5gram	752	てしまったんです/ことが分かりました/だと思いますか
6gram	258	と言われています/といっていました/はどこにありますか
総計	**8121**	

３つの短単位で構成された 3gram の見出し語は，4994 項目が得られた。具体例としては，「～ています」などの初級の学習項目に相当するものが多い。次に，４つの短単位で構成された 4gram の見出し語は，2117 項目，5gram の見出し語は 752 項目，6gram の見出し語は 258 項目が得られた。7gram 以上のデータも作成してみたものの，コーパスサイズが小さいこともあって，頻度 5 以上のものは少ない上に，表現として不完全なものが多いため，対象から外した。

次に，得られた見出し語の特徴分析のため，品詞単位で調べてみた。表 4 に 3gram から 6gram で高頻度パターン上位５位を報告する。

表 4. 品詞の組み合わせの高頻度パターン

	品詞の組み合わせ	具体例
3gram	［助詞-格助詞/名詞-普通名詞-一般/助詞-格助詞］	の方が
3gram	［助詞-格助詞/動詞-一般/助詞-接続助詞］	によって

3gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能]	思っている
3gram	[助詞-格助詞/動詞-一般/助動詞]	と思います
3gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般]	事が分かる
4gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能]	と思っている
4gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	思っています
4gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般/助詞-接続助詞]	文章を読んで
4gram	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	しています
4gram	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	をしている
5gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	と思っています
5gram	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	をしています
5gram	[助詞-接続助詞/動詞-非自立可能/助動詞/助詞-準体助詞/助動詞]	ていたのだ
5gram	[動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	言っていました
5gram	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	していました
6gram	[助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	といっていました
6gram	[名詞-普通名詞-一般/助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	会社に勤めています
6gram	[助詞-格助詞/代名詞/助詞-格助詞/動詞-非自立可能/助動詞/助詞-終助詞]	に何がありますか
6gram	[助動詞/助詞-格助詞/動詞-一般/助詞-接続助詞/動詞-非自立可能/助動詞]	たいと思っています
6gram	[助動詞/助詞-格助詞/動詞-一般/助動詞/助詞-接続助詞/動詞-非自立可能]	だと言われている

次に，難易度判別のために，初級教科書での使用頻度と中級教科書での使用頻度をもとに差異係数を使用し，どちらの（レベルの）教科書でよりたくさん使用されているかを調べた。差異係数がマイナス値のものを初級，差異係数が 0～0.49 のものは初中級，0.50～1.0 のものを中級とし，集計してみた。

表 5. Ngram×レベルのクラス集計表

	初級レベル	初中級レベル	中級レベル
3gram	2156	991	1847
4gram	1084	352	681
5gram	467	99	186
6gram	196	30	32
総計	**3903**	**1472**	**2746**

以上の方法で，完成したデータは，以下の通りである。

■ 初級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	て居ります	(ています)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初級	373	363	0.013586957
3	有りますず	(ありません)	[動詞-非自立可能/助動詞/助動詞]	初級	135	180	-0.142857143
3	が有ります	(があります)	[助詞-格助詞/動詞-非自立可能/助動詞]	初級	118	195	-0.29600639
3	たのです	(たんです)	[助動詞/助詞-準体助詞/助動詞]	初級	80	160	-0.333333333
3	と思います	(とおもいます)	[助詞-格助詞/動詞-一般/助動詞]	初級	117	122	-0.020920502
3	成りますた	(なりました)	[動詞-非自立可能/助動詞/助動詞]	初級	105	118	-0.058295964
3	ますたか	(ましたか)	[助動詞/助動詞/助詞-終助詞]	初級	7	198	-0.931707317
3	のですか	(んですか)	[助詞-準体助詞/助動詞/助詞-終助詞]	初級	59	143	-0.415841584
4	為るて居ります	(しています)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能/助動詞]	初級	102	98	0.02
3	ますずか	(ませんか)	[助動詞/助動詞/助詞-終助詞]	初級	31	158	-0.671957672
3	と思うて	(とおもって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初級	92	81	0.063583815
3	済みますず	(すみません)	[動詞-一般/助動詞/助動詞]	初級	16	151	-0.808383234
3	ても良い	(てもいー)	[助詞-接続助詞/助詞-係助詞/形容詞-非自立可能]	初級	31	136	-0.628742515

■ 初中級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	為るて居る	(してい)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	初中級	239	138	0.267904509
3	を為るて	(おして)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	初中級	209	123	0.259036145
3	為るますた	(しました)	[動詞-非自立可能/助動詞/助動詞]	初中級	168	121	0.162629758
3	居るますた	(いました)	[動詞-非自立可能/助動詞/助動詞]	初中級	148	76	0.321428571
3	の中に	(のなかに)	[助詞-格助詞/名詞-普通名詞-副詞可能/助詞-格助詞]	初中級	134	56	0.416526316
3	て居ります	(ていまし)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	119	52	0.391812865
4	て居るますた	(ていました)	[助詞-接続助詞/動詞-非自立可能/助動詞/助動詞]	初中級	119	52	0.391812865
3	につくて	(について)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初中級	121	45	0.457831325
3	かも知れる	(かもしれ)	[助詞-副助詞/助詞-係助詞/動詞-一般]	初中級	88	55	0.230769231
3	を持つて	(おもって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	初中級	81	49	0.246153846
3	て仕舞うた	(てしまった)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	82	42	0.322580645
3	て居らない	(ていない)	[助詞-接続助詞/動詞-非自立可能/助動詞]	初中級	68	49	0.162393162
3	だのだ	(なので)	[助動詞/助詞-準体助詞/助動詞]	初中級	81	36	0.384615385

■ 中級レベルのコロケーション

Gram	語彙素(基本形)	発音(出現形)	品詞	難易度	中級合計	初級合計	差異計数
3	て居るた	(ていた)	[助詞-接続助詞/動詞-非自立可能/助動詞]	中級	374	55	0.743509744
3	為るて居る	(している)	[動詞-非自立可能/助詞-接続助詞/動詞-非自立可能]	中級	309	75	0.608375
3	と為るて	(として)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	中級	297	12	0.922330097
3	て来るた	(てきた)	[助詞-接続助詞/動詞-非自立可能/助動詞]	中級	219	43	0.671755725
3	て居るの	(ているの)	[助詞-接続助詞/動詞-非自立可能/助詞-準体助詞]	中級	215	45	0.653046154
3	たのだ	(たので)	[助動詞/助詞-準体助詞/助動詞]	中級	190	51	0.576763485
3	れるて居る	(れている)	[助動詞/助詞-接続助詞/動詞-非自立可能]	中級	212	26	0.781512605
3	と言う事	(とゆーこと)	[助詞-格助詞/動詞-一般/名詞-普通名詞-一般]	中級	206	16	0.855855856
3	に成るて	(になって)	[助詞-格助詞/動詞-非自立可能/助詞-接続助詞]	中級	164	35	0.649241206
3	に困るて	(によって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	中級	160	12	0.866666667
3	のだ有る	(のである)	[助詞-準体助詞/助動詞/動詞-非自立可能]	中級	172	2	0.977011494
3	と言うて	(といって)	[助詞-格助詞/動詞-一般/助詞-接続助詞]	中級	137	26	0.689981595
3	為るれるて	(されて)	[動詞-非自立可能/助動詞/助詞-接続助詞]	中級	138	21	0.735849057

４．まとめと今後の課題

本発表では，日本語教科書データを利用したコロケーション辞書作成について紹介した。3gram から 6gram の見出し語として 8,121 項目のリストが構築できた。全体的に機能語に対するリスト化については，ある程度成功しているが，コーパスサイズが小さい問題があり，内容語に対するリストとしてはまだまだ不十分な状態である。今後の予定として，均衡コーパスに対するリーダビリティ値を計算し，「レベル別コーパス」を大きくした上で，内容語も含めたコロケーション辞書の拡張を行いたい。また人手によるチェック作業を継続し，数などを踏まえた上で，初級前半，初級後半，中級前半，中級後半のコロケーション表現のリストとして完成させたい。

謝　辞

本研究は、文部科学省科学研究費補助金「読解教育支援を目的とする文章難易度判別システムの開発（課題番号：25370573，代表者：李在鎬）による補助を得ています。

文　献

庵 功雄 ， 山内 博之 (2015)『データに基づく文法シラバス (現場に役立つ日本語教育研究 1)』くろしお出版

中俣尚己 (2014)『日本語教育のための文法コロケーションハンドブック』くろしお出版

本田 弘之，岩田 一成，義永 美央子（2014)『日本語教育学の歩き方―初学者のための研究ガイド』大阪大学出版会

李在鎬（2011）「大規模テストの読解問題作成過程へのコーパス利用の可能性」,『日本語教育』148, pp.84-98.

Lee, Jae-ho &Yoichiro Hasebe(2015 in press)　“Readability Measurement for Japanese Text Based on Leveled Corpora”

李在鎬，石川慎一郎，砂川有里子（2012）　『日本語教育のためのコーパス調査入門』くろしお出版

『日本語話し言葉コーパス』UniDic 版形態論情報の構築

渡部 涼子（国立国語研究所コーパス開発センター）†
田中 弥生（国立国語研究所理論構造研究系）
小磯 花絵（国立国語研究所理論構造研究系）

Constructing the UniDic Version of the Morphological Information of *Corpus of Spontaneous Japanese*

Ryoko Watanabe　Yayoi Tanaka　Hanae Koiso
(National Institute for Japanese Language and Linguistics)

要旨

『日本語話し言葉コーパス』（CSJ）には形態論情報として短単位と長単位の情報が付与されている。しかし，単位設計や品詞体系の点において，BCCWJ に付与されているものとは異なるため，CSJ と BCCWJ を単純に比較することができないという問題があった。そこで，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換し，中納言検索システムを通して公開することとした。本発表では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点，および UniDic 体系への変換手続きなどについて述べる。また，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスター（学会講演・模擬講演・対話）の品詞・語種の特徴を，BCCWJ の各種レジスター（書籍・新聞・行政白書・Web など）との比較を通して示す。

１．はじめに

『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese，CSJ）は，1999 年から 5 年間かけ，国立国語研究所・情報通信研究機構（旧通信総合研究所）・東京工業大学が共同で開発した，約 660 時間の日本語自発音声からなるデータベースである（国語研究所 2006）。2004 年に公開を開始して以降，音声言語情報処理，自然言語処理，日本語学，言語学，音声学，心理学，社会学，日本語教育，辞書編纂など幅広い領域で利用されてきた。

CSJ には，転記情報や文節情報，形態論情報，節単位情報，分節音情報，韻律情報，係り受け構造情報，談話境界情報，要約・重要文情報，印象評定データなど，多様な研究用付加情報（アノテーション）が付されている。このうち形態論情報については，例えば「国立国語研究所」のような複合語を一つの単位とする長単位と，これらを「国立|国語|研究|所」のように細かく分割する短単位の二種類の情報が付与されており，この点において『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）と同じであるが，単位設計について一部基準が異なる上に，品詞体系についてはかなりの相違が見られる。そのため，CSJ と BCCWJ を同一基準で検索したり，あるいは比較したりといったことができないという問題があった。そこで，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換し，BCCWJ と同じ WEB 上の検索システムを通して公開することとした。

† naberyo＠ninjal.ac.jp

本稿では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点，および UniDic 体系への変換手続きなどについて述べる。また，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスター（学会講演・模擬公演・対話）の品詞・語種の特徴を，BCCWJ の各種レジスター（書籍・新聞・行政白書・Web など）との比較を通して示す。

２．CSJ UniDic 版形態論情報の整備

２.１　CSJ オリジナル版短単位体系と UniDic 体系の設計上の主な違い

２.１.１　単位設計

CSJ オリジナル版の短単位は，現代語において意味を持つ最小の単位（最小単位と呼ぶ）二つが 1 回結合したものであり，『現代雑誌九十種』の用語用字で用いられた β 単位がもとになっている（小椋 2006）。以下に例を示す。なお，短単位の境界は「∥」，最小単位の境界は「｜」で表す。

∥話し｜言葉∥　　∥音｜声∥　　∥レーザー｜プリンター∥　　∥行こ∥う∥

コーパス日本語学への応用を志向して開発された形態素解析用辞書 UniDic（伝ほか 2007；伝ほか 2008）においても，単位設計については原則として CSJ オリジナル版の短単位基準が踏襲されたが，以下のような変更が加えられた（小椋 2008）。

- 外来語は 1 最小単位で 1 短単位とする。

  ∥レーザー∥プリンター∥　　∥オレンジ∥色∥

- 意思・推量の助動詞「う」「よう」を独立の単位とせず，活用語尾として活用語の単位に含める。

  ∥行こう∥　　∥食べよう∥

- 補助記号（「・」「,」「。」など）を独立の最小単位として認定し，1 最小単位で 1 短単位とする。

２.１.２　付加情報

単位認定基準によって認定した一つ一つの短単位は，活用変化・音の転訛・ゆれ・省略・融合等によって生じた異形態や異表記形そのままの形のものであるため，用例検索や計量研究において扱い難い。そこで CSJ オリジナル版では，転記テキストにおける短単位の出現語形（出現形，転記における基本形）とその発音（発音形）について，それぞれの単位が同じ語であるかどうか判断し，同じ語と判断した語群に対して，見出しといえる「代表形」を片仮名で付与している。また，代表形に加えて，代表形を漢字等で表記した「代表表記」という情報も与えている。代表形は片仮名で表記されているため，代表形だけでは同音異義語の区別がつかなくなってしまうが，代表表記を与えることで同音語の区別が可能となる。

UniDic ではこの点をさらに整理し，また表記の変異にも対応するべく，次のように語彙素（語彙素読み）・語形・書字形・発音形からなる階層的見出しを採用している（表 1）。

表 1　UniDic 階層的見出しの例

語彙素	語形	書字形	発音形
矢張り	ヤハリ	やはり	ヤハリ
		矢張り	
	ヤッパリ	やっぱり	ヤッパリ

２.１.３　品詞体系

CSJ のオリジナル版短単位情報は，後述するように，手作業により高精度に情報を付与した人手作業分と，それを学習データとして構築した形態素解析システムで自動解析した自動解析分の二種類がある。このうち人手作業分の品詞情報は，UniDic に比べ，詳細な分類を行なわない，粗いものとなっている。CSJ 作成時点ではコーパスを活用した研究がまだそれほど進んでおらず，どのような品詞情報が有用かの判断材料が極めて乏しい状態だった。そのため，まずは最低限必要な品詞情報を付与しておき，実際に研究に活用していく中でどのような品詞情報が望ましいか検討していく方針を取った。

具体的に名詞を例にして比較をすると，表 2 のとおり，UniDic の方が細かく下位分類まで設定されている（小椋ほか 2011）。

表 2　CSJ オリジナル版（人手作業分）と UniDic との品詞（名詞）の比較

CSJ	
品詞	その他1
名詞	
	固有名詞
	数詞

UniDic			
大分類	中分類	小分類	細分類
名詞	普通名詞	一般	
		サ変可能	
		サ変形状詞可能	
		形状詞可能	
		副詞可能	
		助数詞可能	
	固有名詞	一般	
		人名	一般
			姓
			名
		地名	一般
			国
	数詞		
	助動詞語幹		

活用語についても UniDic の方が詳細な分類となっている（小椋ほか 2011）。五段動詞を例に挙げる（表 3）。ただし，活用の種類と活用形については，同じ CSJ オリジナル版であっても，人手作業分と自動解析分では粒度が異なっており，自動解析分の方がその粒度が細かくなっている。詳細については山口ほか（2004a，2004b）を参照されたい。

表 3　CSJ オリジナル版（人手作業分）と UniDic との活用の種類（五段動詞）の比較

CSJ
カ行五段
ガ行五段
サ行五段
タ行五段
ナ行五段
バ行五段
マ行五段
ラ行五段
ワア行五段

UniDic		
五段	カ行	一般
		イク
		ユク
	ガ行	
	サ行	
	タ行	
	ナ行	
	バ行	
	マ行	一般
		済ム
	ラ行	一般
		アル
		サル
	ワア行	一般
		○ウ

また CSJ オリジナル版では，名詞のうち形状詞や副詞としても使われる語について，文脈等に基づいて名詞・形状詞・副詞の判定を行っているが，UniDic では「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」という品詞を実際の使用例に関わらず与えている。

２.２　変換手続き

CSJ のオリジナル版短単位情報は，次の二通りの方法で付与された。

- **人手作業：**約 100 万語（種々のアノテーションを人手で高精度に付与したコア 50 万語を内包）については，人手により高精度に情報を付与。
- **自動解析：**残り約 650 万語については，上記人手作業分を学習データに構築された形態素解析システム（内元ほか 2004）により自動解析した上で，部分的に人手修正。

■人手作業分のデータの変換手続き： UniDic 構築時に，学習用データとして人手で UniDic 体系に変換する作業を実施した（伝ほか 2007)。ただし，「こ これは」の「こ」のように，言いよどみに伴う語の断片は消去した上で学習用データが作成されたため，今回の整備作業で語断片を元の位置に復元した。これに伴い「言いよどみ」という品詞を新たに設けた。

■自動解析分のデータの変換手続き：次の通り変換作業を行った。

1. UniDic Ver.2.0 をもとに，CSJ オリジナル体系から UniDic 体系に自動で変換した。自動変換に先立ち，単位の粒度が異なるもののうち助動詞「う」「よう」については，活用語尾として活用語の単位にまとめる作業を自動で行った。
2. 変換候補が複数ある場合，出現確率などから，一意に自動で決定するものと，複数項目を列挙するものに分け，後者については人手で確認のうえ認定した。
3. 変換候補がない場合，次の通り対応した。
   a. UniDic に登録されていない語は，一旦保留とした。
   b. 「レーザープリンター」のように単位の粒度が異なるものは，候補を自動で抽出した上で, 分割パターンを半自動で特定した。変換候補が複数ある場合は 2 の処理を，未登録語などを含む場合は一旦保留とした。
4. 上記作業を行い，一通り UniDic 体系に変換したのち，UniDic と連動してコーパスの管理・修正作業を行うことのできるデータベースシステム（「大納言」）に搭載した。
5. 全ての未登録語を対象に, UniDic に人手で新規に語を登録した上で, 大納言上で UniDic にリンクさせる形でコーパスに情報を付与した。

■伏せ字の扱い：オリジナル版では，話者の氏名など話者を特定できる情報や差別語などについて，出現形，発音形，代表形，代表表記は伏せ字化した上で，品詞情報についてはそのまま公開している。UniDic 版を作成するにあたり，人手作業分についてはこの方針を踏襲し，品詞情報を残す形で整備した。一方，自動解析分については，品詞情報の変換はせず，品詞を一律「伏せ字」とした。この点において，人手作業分と自動解析分で扱いが異なるため，利用の際には注意が必要である。

■発音形の扱い: CSJ の転記テキストでは，実際の音声を仮名で書ける範囲で忠実に記録している。その際，「手術（シュジュツ)」を「シジツ」，「形態素（ケイタイソ)」を「ケーソタイ」と発音するなど，発音の怠けや転訛，言い間違いなどが生じた場合には，実際に発音された音と，丁寧に発音された場合に生じるであろう音を「(W シジツ;シュジュツ)」のような形で併記して表現している。オリジナル版短単位情報における発音形では，これら二つの発音情報を共に保存する形で表現しているが，UniDic 体系に変換するにあたり，コーパスと辞書の管理方法の都合などから，実際の発音情報は対象とせず，丁寧に発音された場合に生じるであろう音のみを記すこととした。UniDic 体系での実際の発音の表現については今後の課題とする。

■節単位：BCCWJ などの書き言葉では，文が認定され中納言などでの検索に利用されている。しかし話し言葉の場合，文の認定は必ずしも容易ではない。そこで CSJ では，文に代わる単位として節単位（丸山ほか 2006）が認定されている。中納言における CSJ の検索においても，この節単位を利用する。

２.３　解析精度

CSJ 自動解析分を 2.2 節の手続きに従い UniDic 体系に自動変換したデータ群に対し，ランダムに 1 万語を抽出し，①境界（単位境界が正解と一致するか否か），②品詞（境界に加え，品詞・活用型・活用形が正解と一致するか否か），③語彙素（境界・品詞・活用型・活用形に加え，語彙素が正解と一致するか否か）の三段階でその精度を評価した。結果（F 値）を図 1 に示す。

参考までに，一般的な自動解析のデータである，UniDic-mecab 1.3.12 による BCCWJ・CSJ のレジスター別自動解析精度[1]をともに示す（図 2）。なお，図 2 における CSJ とは，前節で言及した人手作業分データを UniDic の学習データ用に整備したものから一部抽出したものである。

①境界の精度は，自動変換・UniDic-mecab 1.3.12 ともほぼ同じ値を示している。②の品詞と③の語彙素の精度については，白書には及ばないものの，他のレジスターよりも高い値を示している。これは，2.2 節の自動解析分のデータの変換手続きで述べたように，全ての未登録語について，事前に登録処理を施したためである。

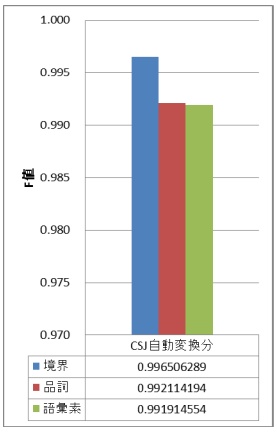


図 1　CSJ 自動変換分の精度

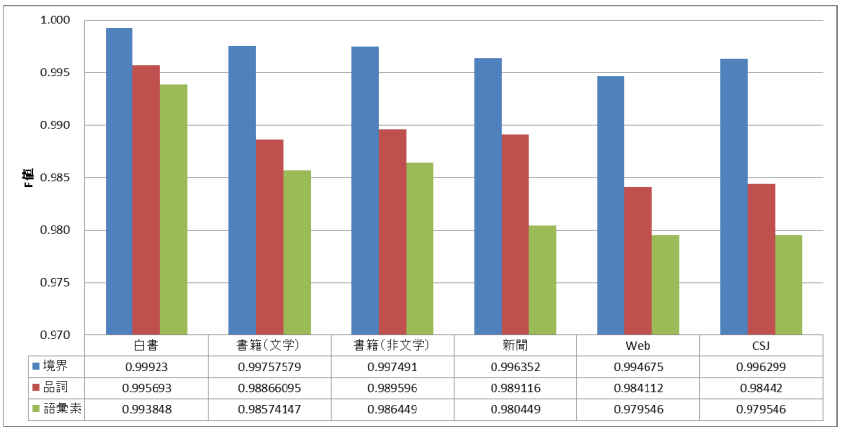


図 2　UniDic-mecab 1.3.12 による BCCWJ・CSJ のレジスター別解析精度

３．CSJ の形態論情報の特徴

３.１　CSJ の基礎統計量

表 4 に，CSJ オリジナル版と UniDic 体系変換後の短単位の語数を，人手作業・自動解析別，レジスター（学会講演＋その他の講演（以下，学会講演），模擬講演，対話，朗読）別に示す。CSJ オリジナル版と UniDic 版の語数が若干異なるのは，2.1.1 節に記した通り，単位の粒度の基準が一部異なるためである。

表 4　CSJ オリジナル版・UniDic 版の語数

	CSJ オリジナル版			UniDic 版		
	全体	人手作業	自動解析	全体	人手作業	自動解析
学会講演	3,597,474	518,024	3,079,450	3,607,546	518,798	3,088,748
模擬講演	3,637,723	436,171	3,201,552	3,640,805	436,069	3,204,736
対話	151,445	41,925	109,520	150,794	41,678	109,116
朗読	208,563	18,976	189,587	209,395	19,031	190,364
計	7,595,205	1,015,096	6,580,109	7,608,540	1,015,576	6,592,964

[1]「UniDic の解析精度」http://download.unidic.org/?page_id=12　参照

また表 5 と表 6 に，UniDic 版の各品詞，各語種の頻度を，人手作業・自動解析ごと，およびレジスターごとに示す。人手作業分と自動解析分の各品詞・語種の比率を比較すると，ほぼ同じ分布となることから，レジスターごとの頻度については，人手作業分と自動解析分に分けず，両者の合計値のみを示す。

表 5　UniDic 版の語数：品詞別

	全体	人手作業	自動解析	学会講演	模擬講演	対話	朗読
名詞	1,818,904	240,674	1,578,230	959,592	781,633	25,608	52,071
代名詞	160,478	21,377	139,101	64,142	85,442	3,957	6,937
形状詞	90,082	12,729	77,353	44,592	42,350	1,637	1,503
連体詞	94,383	12,847	81,536	50,450	41,018	1,522	1,393
副詞	219,651	29,414	190,237	73,383	132,483	8,083	5,702
接続詞	84,161	11,757	72,404	43,414	38,211	1,534	1,002
感動詞	473,527	70,234	403,293	242,661	207,356	18,759	4,751
動詞	997,482	129,836	867,646	470,295	482,220	16,335	28,632
形容詞	106,574	14,741	91,833	36,137	65,110	3,121	2,206
助動詞	886,347	119,650	766,697	386,202	455,708	19,382	25,055
助詞	2,335,347	308,060	2,027,287	1,049,007	1,172,432	45,045	68,863
格助詞	1,188,374	157,806	1,030,568	608,146	534,057	15,064	31,107
係助詞	294,909	38,675	256,234	124,248	155,684	5,493	9,484
接続助詞	405,425	53,689	351,736	176,677	212,570	5,870	10,308
終助詞	124,246	16,108	108,138	37,629	71,343	8,489	6,785
副助詞	168,841	21,635	147,206	52,112	105,152	5,670	5,907
準体助詞	153,552	20,147	133,405	50,195	93,626	4,459	5,272
接頭辞	42,080	6,079	36,001	20,747	20,131	622	580
接尾辞	160,877	20,589	140,288	84,218	67,816	2,288	6,555
記号	32,339	4,295	28,044	25,379	3,988	293	2,679
言いよどみ	96,116	13,294	82,822	47,462	44,658	2,548	1,448
その他	10,192	0	10,192	9,865	249	60	18

表 6　UniDic 版の語数：語種別

	全体	人手作業	自動解析	学会講演	模擬講演	対話	朗読
和語	5,893,040	788,933	5,104,107	2,626,644	2,979,628	127,338	159,430
漢語	1,256,168	164,910	1,091,258	733,050	471,120	14,678	37,320
外来語	178,172	24,137	154,035	104,511	68,674	1,885	3,102
混種語	55,269	7,973	47,296	25,138	28,252	904	975
固有名	72,091	10,302	61,789	25,413	42,364	3,042	1,272
その他	153,800	19,321	134,479	92,790	50,767	2,947	7,296

３.２　品詞率・語種率に見る CSJ のレジスターの特徴

本節では，品詞ごと，語種ごとの出現率から，CSJ の各レジスターの特徴を見ていく。

図 3 に，CSJ（全体）の品詞・語種の出現率を，朗読を除く三つのレジスターごとに示す。また図 4 に，BCCWJ（コア・非コア含む全体）の品詞・語種の出現率を，書籍，新聞，白書，雑誌， Yahoo!知恵袋，国会会議録に限定し，レジスターごとに示す。個々の品詞率，語種率は，サンプルごとの延べ語数に対する各品詞・語種の延べ語数の割合として求めた。ただし品詞率の算出にあたり，CSJ 固有の品詞である言いよどみと伏せ字，および CSJ に頻出する感動詞（「あのー」や「えっと」などのフィラーを含む）は集計の対象としなかった。語種については更に，助詞，助動詞，固有名詞，記号を除外した上で比率を求めた。

図には，小磯ほか（2009）など BCCWJ を主対象とする一連の文体研究で特徴的な傾向を示した品詞・語種を抜粋して示す。なお小磯ほか(2009)では，BCCWJ の構築期間中に，BCCWJ の五つのレジスターおよび CSJ 人手作業分の学会講演と模擬講演を対象に，各レジスターから 150 のサンプルを抽出して品詞率・語種率を求めた。今回の分析では，レジス

ターとして, CSJ から対話を, BCCWJ から雑誌を追加しており, また CSJ, BCCWJ ともに, サンプル数を限定せず，当該レジスターに属する全てのデータを利用している。

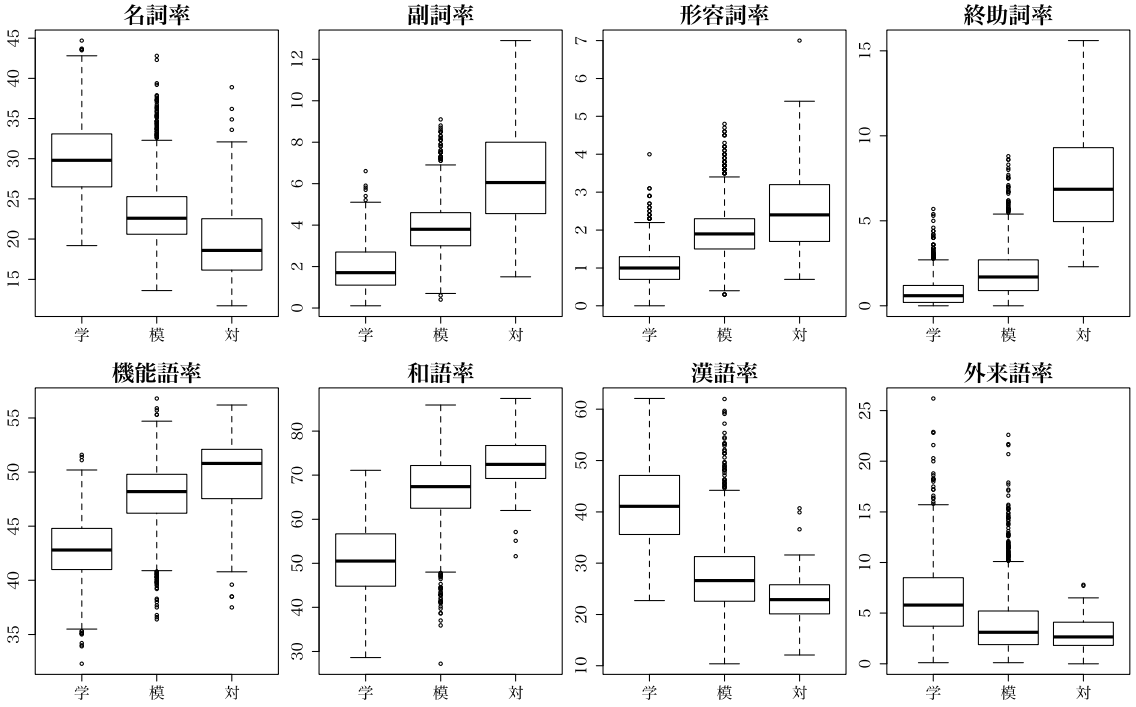


図 3　CSJ の品詞ごと・語種ごとの出現率（中央値と第 1・第 3 四分位数）
学：学会講演，**模**：模擬講演，**対**：対話

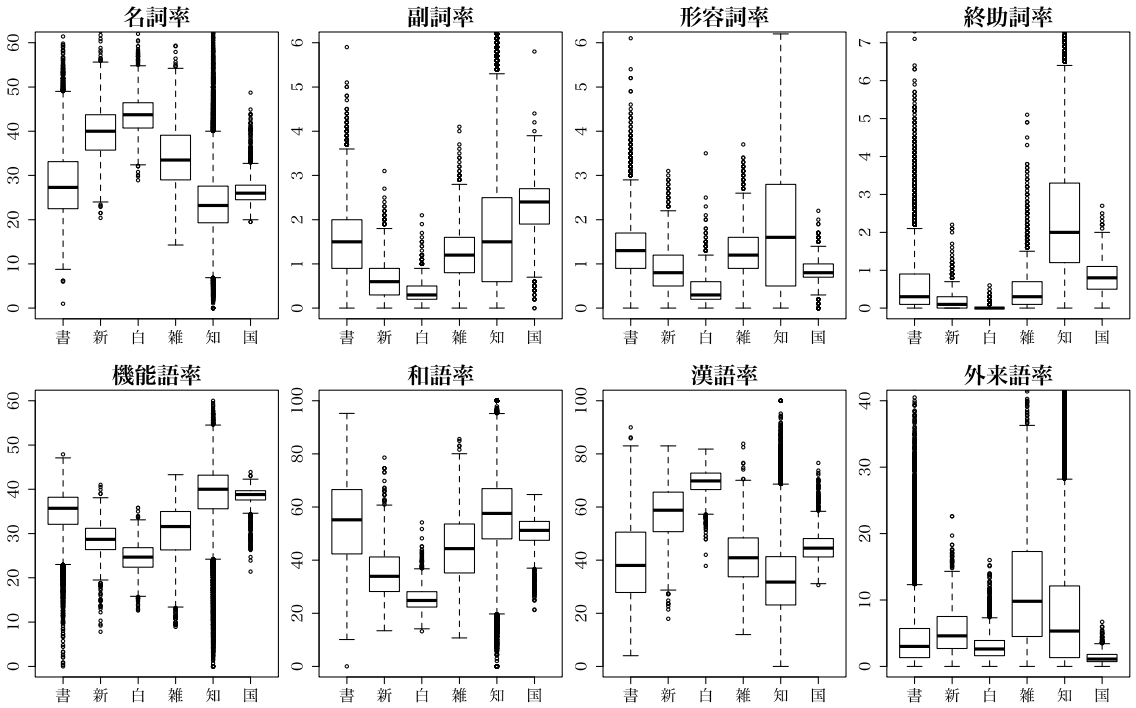


図 4　BCCWJ の品詞ごと・語種ごとの出現率（中央値と第 1・第 3 四分位数）
書：書籍，**新**：新聞，**白**：行政白書，**雑**：雑誌，**知**：Yahoo!知恵袋，**国**：国会会議録

■語種率：図 3 の CSJ の結果を見ると，漢語と名詞は「対話＜模擬講演＜学会講演」の順に多くなるのに対し，和語と機能語（助詞・助動詞）は逆の傾向を示している。こうした漢語率・名詞率と和語率・機能語率の関係は BCCWJ にも成立する。BCCWJ では，漢語や名詞は行政白書や新聞に，和語や機能語は書籍やインターネット上のテキスト，国会会議録に多く見られる。雑誌はその中間の傾向を示す。この傾向は小磯ほか(2009)とほぼ一致する。

一連の国語研究所の語彙調査や野元(1959)などから，書き言葉では和語よりも漢語が，話し言葉では逆に漢語よりも和語が多い傾向にあることが指摘されている。CSJ の各種レジスターや国会会議録，話し言葉に近い傾向を示す Web 上のテキスト(Yahoo!知恵袋)，また BCCWJ のうち小説の会話文などを含む書籍が高い和語率を示しており，上記指摘と整合的である。また丸山 (2005) は, CSJ の模擬講演を含む各種話し言葉の漢語率を比較しており，その中で，模擬講演の方が日常会話よりも漢語率が顕著に高い傾向を示すことから，敬体で改まった表現を用いる傾向の強い模擬講演のような独話では，日常会話よりも書き言葉により近い傾向を示すとしている。国語研究所(1955)でも，ニュース解説やニュースの方が日常談話よりも漢語率が高いとされる。図 3 の CSJ の結果を見ると，この傾向が顕著に観察されるのは学会講演である。学会講演では，漢語率が 4 割を越えており，新聞や白書よりは少ないものの，書籍や雑誌などの書き言葉と同じ水準となっている。国会会議録もやはり漢語率が 4 割以上であり，学会講演同様，改まりの程度の強い，書き言葉に類似した傾向を示している。また漢語の使用は硬い文体と，和語の使用は軟らかい文体と関連することが指摘されており（柏野ほか 2012），こうした各レジスターの硬軟の偏りも語種率に影響したものと考えられる。

■機能語率・名詞率：Halliday(1990)は，内容語率で定義される語彙密度という尺度を提案し，綿密に計画された，あるいはよりフォーマルな文章ほど語彙密度が高いとしている。機能語率の逆数が内容語の占める割合と考えるならば，対話よりも講演の方が，また講演の中でも模擬講演（主に個人的内容に関する一般の人によるスピーチ）よりも学会講演の方が，機能語率が低い（内容語率が高い）傾向を示しており，「対話＜模擬講演＜学会講演」の順に，より綿密に計画された，あるいはよりフォーマルなスタイルの発話であると言える。実際，学会講演では予稿集やスライドなどの発表資料を，また模擬講演では発話の流れを記したメモを準備しており，相手とのやりとりの中で発話内容を決める対話と比べて発話の計画性は高いと言える。また学会講演は，大人数の前で自身の主張を展開するものであり，2〜4 人程度の収録スタッフを前に個人的体験談などを語る模擬講演と比べ，よりフォーマルな発話であると言える。BCCWJ においても，小説などを含む書籍や Web 上のテキストよりも，行政白書や新聞の方が機能語率は低い（内容語率が高い）傾向を示しており，行政白書や新聞の方がよりフォーマルであるという直観と合致する。一方，国会会議録は，フォーマルで発話内容の計画性も高いと考えられるが，白書や新聞と比べ機能語率はかなり高い傾向を示している。国会会議録は CSJ の学会講演と同水準であることから，機能語率（内容語率）には，単に計画性やフォーマルさの程度だけでなく，話し言葉・書き言葉というモードの違いも関わる可能性がある。

また名詞率は，先述の通り機能語率と逆の傾向を示しているが，複雑な文ほど動詞群の名詞化により機能語に対する内容語の比率が高くなることから(Halliday 1985)，名詞率と内容語率（機能語率）は正（負）の相関を示すことになる。このことが上記結果につながったと考えられる。

■副詞率・形容詞率：国語研究所(1955)では，日常談話，ニュース解説，ニュースの副詞率が 6.1%，2.5%，1.3%，形容詞率が 2.7%，0.9%，0.4% と，主観的表現の多い日常談話の

副詞率，形容詞率が圧倒的に高いこと，また同じニュースでも，ある程度解説者の意見などを含むニュース解説の方がニュースよりも副詞率，形容詞率が高いことを示している。学会講演のように客観的表現の好まれるレジスターよりも，模擬講演（個人的体験談の語りなど）や対話のように主観的表現が多く含まれるレジスターの方が，副詞率，形容詞率ともに高い傾向を示しており，整合的な結果となっている。BCCWJ を見てみると，やはり客観的表現の好まれる行政白書や新聞では副詞率・形容詞率ともに低いのに対し，小説などを含む書籍では高い値を示している。

その一方で，客観的表現が好まれると予想される国会会議録において，形容詞率は確かに低いものの，副詞率については若干高い値となっている。形容詞率については，話し言葉のうち客観的表現が好まれる学会講演や国会会議録と，書き言葉で同じく客観的表現が好まれる新聞がほぼ同じ傾向を示していることから，話し言葉・書き言葉の区別なく，表現の客観性・主観性の観点がその出現に強く影響していると考えられる。一方，副詞については，書き言葉の各種レジスターよりも国会会議録は高い比率を示している。また，副詞率が最も低い行政白書と最も高い対話でその中央値が 0.3%と 6.1%となっており，形容詞の場合（0.3%と 2.4%）と比べて極端に開きがある。この傾向は，模擬講演や学会講演，国会会議録など，その他の話し言葉にも大なり小なり見られる。以上のことから，副詞については，表現の客観性・主観性に加え，話し言葉・書き言葉というモードの違いも影響している可能性が考えられる。

４．おわりに

BCCWJ との統一的な検索を目指し，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換する作業を実施した。2 節では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点, および UniDic 体系への変換手続きなどについて解説した。また 3 節では，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスターの品詞・語種の特徴を，BCCWJ のレジスターとの比較を通して議論した。

CSJ の UniDic 版短単位情報は, 今年度中を目途に中納言検索システムを通して公開する。また，今回は短単位情報のみの公開に留まるが，今後，長単位情報についても同様に整備する予定である。

文　献

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）「コーパ日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－」『日本語科学』22,　pp.101-123

伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信（2008）「UniDic version1.3.9 ユーザーズマニュアル」http://chikusei.lv9.org/cms-z/zomeki-1.0.4/ext/morph/unidic/manual.pdf

Halliday, M.A.K.　（1985）*Spoken and Written Language,* Victoria: Deakin University

Halliday, M.A.K.　（1990）“Some grammatical problems in scientific English,” *Annual Review of Applied Linguistics*, 6, pp.13-37.

柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛（2012）「書籍テキストへの文体情報付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『第 2 回コーパス日本語学ワークショップ予稿集』pp.155-164

小磯花絵・小椋秀樹・小木曽智信・宮内佐夜香（2009）「コーパスに基づく多様なジャンルの文体比較－短単位情報に着目して－」『言語処理学会第 15 回年次大会発表論文集』pp.

594-597
国語研究所（1955）『談話語の実態』国立国語研究所報告 8，秀英出版
国語研究所（2006）『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』
丸山直子(2005)「話しことばにおける漢語」『東京女子大学比較文化研究所紀要』66，pp.27-38
丸山岳彦・高梨克也・内元清貴（2006）「節単位情報」『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』pp.255-322
野元菊雄（1959）「話しことばの中での漢語使用」『ことばの研究』国立国語研究所論集 1
小椋秀樹（2006）「形態論情報」『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』pp.347-453
小椋秀樹（2008）「『日本語話し言葉コーパス』の言語単位」『日本語学』27 巻 5 号 pp.72-81
小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書『『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)』LR-CCG-10-05
内元清貴・高岡一馬・野畑周・山田篤・関根聡・井佐原均（2004）「『日本語話し言葉コーパス』への形態素情報付与」『第 3 回話し言葉の科学と工学ワークショップ講演予稿集』pp.39-46.
山口昌也・木村睦子・西川賢哉・石塚京子・小椋秀樹（2004a）「短単位辞書マニュアル」CSJ 同梱マニュアル　http://pj.ninjal.ac.jp/corpus_center/csj/manu-f/suwdic.pdf
山口昌也・木村睦子・西川賢哉・石塚京子・小椋秀樹（2004b）「短単位・長単位データマニュアル」CSJ 同梱マニュアル　http://pj.ninjal.ac.jp/corpus_center/csj/manu-f/wdb.pdf

アカデミック・ライティングに見られる副詞に関する分析

阿辺川 武（国立情報学研究所）†
八木 豊（株式会社ピコラボ）
ホドシチェク・ボル（大阪大学言語文化研究科）
仁科 喜久子（東京工業大学名誉教授）

Analysis of Adverb in Japanese Academic Writing

Takeshi Abekawa (National Institute of Informatics)
Yutaka Yagi (Picolab Co., Ltd.)
Hodošček Bor (Osaka University)
Kikuko Nishina (Tokyo Institute of Technology)

要旨

我々は BCCWJ に科学技術論文を加えたコーパスを使用してレジスター誤り検出を行う日本語作文推敲支援システム「ナツメグ」を開発した。システムでは、アカデミック・ライティングの文体に近い準正用コーパスと、話し言葉を多く含む準誤用コーパスでの使用頻度の比を利用して、レジスター誤りと思われる表現を検出しているが、準正用コーパスでの頻度が高いにもかかわらず、システムが誤用と判定してしまう表現が存在する。本発表ではシステムの検出精度の向上をめざし、誤検出となる表現の中から、話し言葉と書き言葉のレジスターの異なりが顕著に見られる副詞に着目し、分析をおこなった。準正用コーパス中で頻度上位の副詞について、実際に用いられている文脈を参照し、書字形および語彙素別にまとめあげ、日本語教育の専門家の意見を参考にしながら、アカデミック・ライティングとしてふさわしい表現であるかを分析した。

１．はじめに

日本の大学で学ぶ理工系留学生は日本語での実験レポート、授業での課題レポート、卒業論文、学位論文、投稿論文が必要になることが想定される。これらをアカデミック・ライティングというジャンルの一部と考え、このジャンルの作文支援をすることを目的に作文支援システム「ナツメグ」の開発を進めている。「ナツメグ」は学習者が論文などの文章を入力すると、システムが入力された表現が適切か否かを判定し、不適切な表現の場合は、適切なヒントを提示することを目指している(八木ら 2014a)。

学生たちは初級から中級に至るまで、主として話し言葉を中心に学んでいるため、上級になって「である体」あるいは「だ体」の書き言葉による文章を学んでも、いざ書く場合になって、どのような用語を用いるかを習得できていないことがある。次の例文は我々が作成した学習者作文コーパス「なたね」の中にある理系学部１年生による１文である。

例 1：今日本では片仮名で書くのは ちょっと多いと聞いたことがある。意味は同じだが、片仮名で書き直したらなんだか新鮮でファッションな、おしゃれな 感じがするようになる。もし先生という言葉 は平仮名で書くとすぐ親切な先生が思い出す。ほんとに器用な言語と思う。

この文中で「ちょっと」「なんだか」「すぐ」「ほんとに」は話し言葉であり、アカデミックな文章では用いられない。「ちょっと」は「やや」に、「ほんとに」は「実に」などで言い換えることができる。

† abekawa [a] nii.ac.jp

本稿では作文推敲支援システムの開発にあたり学習者の文章を観察した結果、このような不適切な表現が見られる中で特に副詞に注目した。副詞を取り上げた理由として、他の品詞と比較すると論文などで用いられる副詞の数はかなり限られていること、また話し言葉と書き言葉のレジスターの異なりが顕著に見られること、そしてシステムの誤用判定と教育者の誤用判定結果が異なる表現が少なからず存在することからである。文末表現や句と句、文と文の接続などの機能語にも不適切な表現が見られるが、これらは共起関係や他の語との意味的関係を考慮しなければならないことも多く、定量的な分析が困難である。それ対して、副詞は独立した品詞として抽出しやすく、分析の緒としては適切だと判断した。

２．使用するコーパスと誤用判定の仕組み

話し言葉と書き言葉という対立、砕けた文章と堅い文章という対立、小説やエッセイなど主観や感性を重視する文章に対する学術的な客観性を重視する文章などのジャンルは多様であり、そこで用いられる言語表現も異なっている。このようにジャンルによって異なる表現のヴァリエーション（言語変異）を語のレジスターと呼ぶ(Halliday 1976)。本研究では理系留学生に必要とされるアカデミックなレポート・論文のための日本語表現をアカデミック・レジスターと定義し、開発中のシステムがその条件にふさわしい表現か否かの判定をすることで、目標とする文章を向上させることを我々は目指している。

システムのために用意するコーパスは国立国語研究所で開発した「現代日本語書き言葉均衡コーパス」（以下 BCCWJ と呼ぶ）および独自に収集した科学技術論文である。このコーパス中の副詞を分析対象とする。コーパスの中でアカデミックな文章に近いものを準正用データ、アカデミックな文章から遠いものを準誤用データとし、アカデミック・ライティングに適合した表現か否かを判定し、適切な表現に導くという手続きを取る。準正用データに含まれるデータは「科学技術論文」データと BCCWJ の中の「白書、法律」データである。これらの文書は、論文に準じる語彙と文体からなると判断した。一方、準誤用データは、同じく BCCWJ の中の「Yahoo!ブログ、Yahoo!知恵袋、国会会議録」である。Yahoo!ブログと知恵袋は、書き言葉であるが情緒的で口語的な表現が多い。国会会議録は、話し言葉を書き起したものであるため、話し言葉の要素が大きく、この 3 データはアカデミックな文章とは対称的なものであると判断し、準誤用データと位置づけた。その他の一般的な「書籍、雑誌、広報誌、新聞」など、どちらにも属していない中立のデータ群も有意差を決定するために用いている。これらのコーパスは、UniDic に基づいてデータが構成されおり、語は語彙素の下に語形があり、その下に書字形・発音形がある（伝ら 2007）。語彙素の下はさまざまな表記のヴァリエーションとしての書字形からなり、1 語彙素に対して、1 から十数個までの書字形が存在する。したがって、システムにおける語の頻度を計算するに当たっては、語彙素と書字形の関係に注意を払わなければならない。語彙素は意味用語を同一にする語形の集合で見る方が良い場合、語形はテキスト上でその語がどのような用字法で記載されているかを見分ける方が良い場合というそれぞれの観点で必要な単位であり、それぞれ分析時に使い分ける必要がある。日本語表記については、英語などのような一国の言語としての正書法が存在しないが、その補佐的なものとして文部科学省が公示した「公文書要領」があり、国の公文書はその指針に従って作成している(文部省 1960)。しかし、新聞、雑誌、その他の出版物は、それぞれの会社や機関が定めた文書作成規則に従って作成しており、強い拘束力はない。

ここで､我々が注目する副詞は書字形で約 7,400 項目存在する｡これらの書字形ごとに(ホ

ドシチェク 2011)の判定式を施すことで各項目の語についてレジスターとしての可否を判定する。例えば「良く」という語彙素は「よく、良く、ヨク、よーく」などの 15 種の書字形からなっている。全コーパスの語に対して頻度計算をした後、準正用データと準誤用データ間の使用頻度の差および有意差の有無によってアカデミック・レジスターとしての可否を示すことになる。

システムでは学習者によって入力された語の妥当性を判定式によって統計的に処理し、その語が有意に誤用と判定されれば、その語はアカデミックな文章としては適切でないため、学習者に注意が喚起される。学習者はこの喚起によって、不適切な用法に気づき、自ら適切な用法を検討するように導かれる(八木ら 2014b)。

表 1: 各コーパスで頻出する副詞（語彙素別、PPM:100 万形態素あたりの相対頻度）

	全体		準正用		準誤用			全体		準正用		準誤用	
順	語彙素	PPM	語彙素	PPM	語彙素	PPM	順	語彙素	PPM	語彙素	PPM	語彙素	PPM
1	そう	910.4	例えば	408.3	どう	1,537.4	15	最も	161.8	良く	62.6	直ぐ	245.8
2	どう	827.9	最も	262.7	そう	1,373.9	16	何故	161.6	予め	57.1	迚も	244.6
3	もう	434.2	特に	250.2	もう	690.8	17	全く	161.1	一層	55.8	可成	237.2
4	こう	411.3	先ず	249.3	こう	685.7	18	更に	157.9	極めて	54.6	何故	235.5
5	良く	300.5	より	227.5	矢張り	572.9	19	詰まり	157.4	余り	52.2	特に	227.6
6	未だ	259.4	どう	162.1	一寸	490.6	20	一番	153.1	可成	51.5	全く	203.4
7	例えば	251.1	更に	131.9	良く	404.0	21	余り	148.9	主に	51.5	宜しく	195.2
8	少し	243.0	略	119.2	少し	395.7	22	若し	145.6	未だ	47.9	勿論	194.7
9	先ず	230.6	こう	97.8	未だ	379.1	23	既に	145.4	もう	46.3	例えば	187.8
10	矢張り	227.3	詰まり	94.2	一番	300.0	24	勿論	143.7	全く	44.0	中々	186.0
11	特に	212.0	そう	84.7	又	282.7	25	迚も	130.3	十分	34.5	結構	176.9
12	又	208.0	必ず	79.5	余り	278.8	27	初めて	129.6	次いで	32.0	もっと	171.8
13	一寸	207.8	既に	78.7	色々	257.5	28	より	119.9	やや	31.9	初めて	163.2
14	直ぐ	186.1	直接	66.9	先ず	251.7	29	可成	117.6	若し	30.1	ずっと	130.7
15	最も	161.8	良く	62.6	直ぐ	245.8	30	もっと	115.5	何故	26.4	必ず	121.0

3．準正用データと準誤用データの比較

本システムが用いるコーパス全体および準正用データと準誤用データにおける語彙の構成について、その様相を概説する。表 1 はコーパス全体、準正用、準誤用データの語彙素別の副詞上位 30 位までを示している。全コーパスでは上位 30 位までで 53.29%をカバーしている。準正用データでは、30 位までで 71.54%、100 位では 91.26%をカバーしている。全コーパスにおけるカバー率と比較すると、テキスト中での副詞の使用が限られた高頻度語に集中していることがわかる。一方、準誤用データの上位 30 位までのカバー率は 58.70%であった。これにより、アカデミック・レジスターでは、他のグループより限られた副詞で文章が構成されていることがわかる。準正用コーパスと準誤用データの頻出副詞の異同を見ると、不一致語の中で準正用には存在せず、準誤用のみに見られる語は「一番、もっと、一寸、勿論、矢張り」など 17 語あり、これらの語が学習者コーパスの中でしばしば見られ、論文として違和感を与える一因になっている。

４．アカデミック・レジスターとして不適切とされた副詞の分析

システムの判定結果の妥当性を検証するために人手による判定と比較する観察実験を行った。その結果、システムが誤用と判定したものの中に日本語教育の専門家が科学技術論文のレジスターとして適切であると評価したものが少なからず存在した。両者の不一致の原因を知るために 1) 複数の書字形を有する副詞、2)高頻度副詞「こう」「そう」「どう」についての分析をおこなった。

４.１ 複数の書字形を有する副詞「矢張り」

先に述べたようにシステムが利用する語彙データは、BCCWJ で用いられている UniDic に依拠している。語彙素は書字形の異なる形を一つの概念としてまとめる語の抽象的な集合と言える。書字形を多く有し、システムが誤用であると判定した語として「矢張り」を例に問題点を述べる。

語彙素「矢張り」は語形「ヤハリ」「ヤッパリ」「ヤッパ」に分かれ、更に書字形「矢張り」「やはり」「やっぱり」「ヤッパリ」「矢っ張り」「やっぱ」などの話し言葉の発音に近い形として出現する。各コーパスにおける相対出現頻度を表 2 に示す。それぞれの書字形についてのシステムの判定は、「やはり」「やっぱり」が誤用となっている他は、低頻度のため判定不可(NA)となっている。「公文書要領」によると、語彙素「矢張り」は平仮名の「やはり」が推奨されているが、準正用データにおいては 78.5%、コーパス全体では 60.3%、準誤用データでは 58.2%であり、準正用データにおける表記法が他に比べて規範に沿っていることがわかる。なお準正用データでも「やっぱり」「ヤッパリ」のような砕けた口語を含んでいるが、これは論文中に引用した文芸作品などの引用と推測される。

表 2: 語彙素「矢張り」の相対頻度 (単位 PPM)

書字形	システムの判定	全体	準正用	準誤用
やはり	誤用	137.1	11.3	323.5
やっぱり	誤用	77.5	1.6	205.7
やっぱ	N/A	10.5	0.3	39.2
矢張り	N/A	0.8	0.5	0.6
やっぱし	N/A	0.7	0.1	1.7
ヤッパリ	N/A	0.3	0.4	1.0
やぱ	N/A	0.3	0.0	1.1
やつぱり	N/A	0.2	0.0	0.0
矢っ張り	N/A	0.1	0.1	0.0
矢つ張り、矢ッ張り、やッぱり、矢っ張、矢つ張	N/A	0.0	0.0	0.0

４.２「こそあど」語彙からなる副詞

「こそあど」語彙からなる副詞「こう、そう、ああ、どう」の占める割合は全ジャンルを通して非常に多く、準正用データにおいても「ああ」を除いて高頻度語に位置している。全体コーパス、準正用、準誤用の順で「そう」(1 位、11 位、2 位)「どう」(2 位、6 位、1 位)「こう」(4 位、9 位、4 位)である。科学技術論文では、「このように、そのように、どのように」という書き言葉の表現が併用されるため、「こう、そう、どう」の頻度が相対的に低くなっていると考えられる。しかし、システムによるレジスター判定では準正用データで高頻度であるにもかかわらず、これらの副詞が誤用となっている。このような様相

を科学技術に論文におけるレジスターの問題として検討する。なお、「ああ」については、準誤用データにおいて用例が存在するが、準正用データの中では、「ああ」が使用される例は極めて少なく、論文中に言語分析のための例文が入っているテキスト以外には見られない。一方、学習者コーパスにおける作文では「ああ」の使用がしばしば見られる。

４.２.１「こう」

全体コーパスで第 4 位、準正用データで第 9 位、準誤用データで第 4 位とどのコーパスにおいても高頻度であるが、3 データの比から計算すると判定式は誤用となる。しかしながら、準正用データにおける使用頻度は少なくはない。準正用データ中でどのような用法があるのか見るために、「こう」に続く連語をみると、「こうした」(74.8PPM)「こうして」(8.7PPM)が高頻度で出現し、これらの連語が準正用データにおける「こう」の 85.3%を占めている。これらの連語は文章中の前方照応の機能を果たしていることが多い。

例 2:こうして収集された日本語の用例文を翻訳家に英訳してもらう。(科学技術論文.自然言語処理.言語処理学会予稿集)

副詞句「こうして」、連体詞句「こうした」は話し言葉や砕けた文章にも見られる「こう」から派生した連語であり、改まった文章では「このようにして」「このような」という論文などでよく見られる形態に置き換えることができる。また、更に砕けた表現として「こんな（に）」との対照があるが、すべて準誤用での用法が多く、準正用ではほとんど見られない。これらの観察の結果として、アカデミック・レジスターとしては「こういう」は用いられることが少なく、「こうした」は準正用が準誤用より多いことがわかる。この観察から「このような／に」をアカデミック・レジスターとして認め、「こうした」もこれに準じて許容しでもよさそうである。

表 3: 副詞「こう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
こう	形態素	411.4	97.8	685.7
こうして	複合語	44.6	8.7	19.1
こうした	複合語	100.2	74.8	36.2
こういう	複合語	130.8	1.4	391.4
こう言う	複合語	2.7	0.1	4.5
こう云う	複合語	0.2	0.0	0.2
このような	複合語	148.9	255.1	71.1
この様な	複合語	1.6	1.6	4.0
このように	複合語	79.4	111.2	45.5
この様に	複合語	0.4	0.6	0.6
このようにして	複合語	7.3	12.6	0.8
こうやって	複合語	5.5	0.1	8.7
こんな	形態素	200.9	3.7	333.4

４.２.２「そう」

「そう」はコーパス全体で頻度 911PPM であり、副詞頻度の最高値である。準正用データでは 84.7PPM、準誤用データ 1,370PPM であり、システムの判定では誤用となる(表 4)。「こ

う」と同様に、判定結果が「誤用」であるにもかかわらず、準正用での出現頻度は低くない。そこで、「こう」の場合と同様に、後に続く語をみると、「A は B。そういう X は～」「A は B。そういった X（状態、状況）は～」などのような表現であり、科学技術文章の中では、慣用的な文型といえる。また、「A が B である場合～、一方 A がそうでない場合」というような前方照応の定型的な表現も多く見られる。（例：「ペアが含まれるなら真、そうでないなら偽である」）。これは、前文の内容を言い換えた代言（パラフレーズ）表現と言える。「そのよう」との対応を考えると、「そのようでないなら」という言い換えはできない。肯定表現では「そのような場合には」はとなり、「そう」は出現しない。

一方、「そう解釈できる」は「そのように解釈できる」と書き換えることが可能であり、「そういう」「そういった」「そう解釈できる」は、前記の用法より、科学技術文章の一般的な表現からやや遠い表現だと思われる。実例をみると、「そうして、そうした、そういった、そのように、そのような、そんなに、そんな」の連語において、準誤用データに圧倒的に多く用例があり、準正用データの例は少ない。「そうした」「そのような」は正用データ中でやや多く見られるが、いずれも全データ中の 10%以下である。結論として、「そう」の用法からすべての「そう」を科学技術論文レジスターから排除するのではなく、「A は B である。そうでない場合、A は～」「そういった」「そういう」などのように文脈上、前方照応の機能を担ったフレーズを正用として認めるなどの措置は有用であろう。

表 4: 副詞「そう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
そう	形態素	910.6	84.7	1,373.7
そうして	複合語	20.2	0.7	19.5
そうした	複合語	57.0	10.2	38.4
そういう	複合語	229.0	2.7	555.8
そう言う	複合語	8.7	0.2	9.5
そう云う	複合語	0.5	0.0	0.3
そのような	複合語	54.6	48.2	53.0
その様な	複合語	0.8	0.3	3.0
そのように	複合語	13.3	2.6	19.6
その様に	複合語	0.2	0.0	0.5
そのようにして	複合語	1.3	0.2	0.8
そうやって	複合語	6.6	0.1	6.4
そんな	形態素	316.2	5.1	443.0

４.２.３「どう」

「どう」は全コーパスで 828PPM(2 位)、準正用データ 162PPM(6 位)、準誤用データ 1,540PPM(１位)であり、準誤用データの中では最も多用される副詞である。その中で準正用データ中に顕著な句構造をみると、連語としての「かどうか」(148PPM)の頻度は高く、他の連語「どういう」、「どうすれば」、「どうしても」などと比較しても抜群に高頻度である。また、「どう考えるか」などのように「どう」の後に動詞が来て、「か」で結ぶ係り受け構造となるものがある。

書き言葉では「どう」は「どのように」とする方が、フォーマルな表現とされているた

め、その差を見ることにする(表 5)。準正用データでは「どのように」が「どう」の約 2.6 倍、一方、準誤用データでは「どう」の使用が「どのように」の約 21.3 倍となり、準正用では「どのように」の使用割合が高いことがわかる。「どういう」も「どのような」とフォーマルな表現に書き換えられるが、同様に相対頻度の比をみると、準正用では「どういう」:「どのような」を語彙素で比較すると、1:18、準誤用ではほぼ 2.6:1 と全く逆の使用頻度となる。従って、「どういう」を「どのような」へと書き換えすることを推奨すべきである。

さらに、「かどうか」と「か否か」の対比をみると、準正用では「かどうか」:「か否か」がほぼ 2.3:1、準誤用ではほぼ 27.4:1 となり、準正用では,準誤用のほぼ 12 倍になる。「かどうか」についても「か否か」への書き換えを推奨することが考えられる。また「どういった」は、やや書き言葉的な傾向があるが、これも「どのような」に書き換えられるものである。「どういった」と「どういうような」の用法は、例 3 のように執筆者個人の嗜好によることが多いように思われる。

例 3: このような箇所を読むことで，著者がどういった目的でその論文を参照したのかがわかる.（科学技術論文.自然言語処理）

以上「どう」についてまとめると、「どう」「どういう」「どんな/に」はアカデミックな分野で多用される「どのように/な」に置き換える指示を出し、アカデミック・レジスターとして書き換えを認めるべきであろう。また、「かどうか」は「か否か」への書き換える方が適切であるが、実際は「かどうか」が多用されているので、その許容の程度は検討する必要がある。

表 5: 副詞「どう」と関連する連語の相対頻度（単位 PPM）

表現	種別	全体	準正用	準誤用
どう	形態素	828.25	162.10	1,537.40
どうして	複合語	120.44	4.27	174.91
どうした	複合語	48.62	1.78	92.58
どういう	複合語	77.66	5.14	177.60
どう言う	複合語	0.78	0.00	2.78
どう云う	複合語	0.22	0.00	0.06
どのような	複合語	67.38	91.54	66.40
どの様な	複合語	1.02	0.71	3.69
どのように	複合語	59.48	61.18	72.91
どの様に	複合語	1.12	0.44	4.21
どのようにして	複合語	6.03	3.06	5.92
どうやって	複合語	22.02	1.18	47.21
どんな	形態素	158.63	12.57	251.99
かどうか	複合語	112.22	108.25	147.73
か否か	複合語	18.99	47.43	5.44

５．おわりに

高頻度副詞群の中にあって、システムの判定は誤用とされている語が存在し、その中で日本語教育の専門家の判定が正用となる語が少なからず存在した。専門家の判定は論文指

導者とも近いと考えられ、学習者がシステムを使用する際に、専門家が可とする語をシステムが誤用と判定すると、学習者に混乱を招く可能性が予測される。矛盾と思われる要因は、1）判別式の欠陥、2）データの偏りなどが考えられる。この矛盾を解消するために 1）については、頻度の閾値を人手の判断も加味しながら再検討し、比較的頻度の高いものは、論文執筆において使用が認められる語であるとすることも可能である。準正用の頻度が一定の値を超えていれば、正用とするという条件を加えることも検討の余地がある。ある程度高頻度であり、かつ専門家が可とする場合は、システム判定式に対して追加条件を設けることも一策であろう。

2）については、現時点で使用している各データに考慮すべき問題がある。準正用の論文データ中に言語処理、言語を扱うものがあり、その論文の中にかなりの割合で、話し言葉を含む例文が存在している。そのため、誤用データに属するような語が出現している。これを解消するためには、言語学・言語処理以外のさらに多くの論文データを投入することが考えられる。

以上、今回の副詞の分析を通して、判別式の問題、データ構造の問題とともに、語解析の問題も見えてきた。形態素を超えた連語・イディオムの扱い方、語彙素と書字形の問題などであり、副詞以外の語彙についても発展できる可能性が見られた。例えば、機能語、形容詞、形容動詞においても、同様の分析をすることで、判定式の精度をあげることも考えられる。また、学習者に対する対策として、混乱を防ぐためにも規範的な規則も導入し、リスト化したデータからヒントを提示する可能性があることを示した。これらをシステムに反映することで、精度の向上に努めることを今後の課題とする。

謝　辞

本研究は、文部科学省科学研究費補助金基盤研究 C「日本語作文支援システムにおける誤用の検出及び添削に有用な情報の提示法の研究」（平成 27～29 年度、代表者：阿辺川武）による補助を得ています。

文　献

八木豊、ホドシチェク・ボル、阿辺川武、仁科喜久子、室田真男（2014b）「作文推敲支援システムによる誤り指摘への学習者の対処に関する調査」日本教育工学会研究報告集 No.14(5)、pp.151-156.

八木豊、ホドシチェク・ボル、阿辺川武、仁科喜久子（2014a）「日本語作文推敲支援システム「ナツメグ」における誤用検出手法の評価」第 5 回コーパス日本語学ワークショップ予稿集、pp.167-170.

ホドシチェク・ボル、仁科喜久子（2011）「作文支援システムにおけるレジスターの扱い」世界日本語教育研究大会 異文化コミュニケーションのための日本語教育 2、pp.522-523.

伝康晴、小木曽智信、小椋秀樹、山田篤、他（2007）「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」日本語科学 22 号、pp.101-122.

Halliday, M. and Matthiessen, C. (2004) An Introduction to Functional Grammar (3rd Edition), Routledge

文部省(1960)、公用文の書き方―資料集―:
http://kokugo.bunka.go.jp/kokugo_nihongo/joho/series/21/21.html

関連 URL

日本語学習者作文コーパス「なたね」: https://hinoki-project.org/natane/

日本語作文推敲支援システム「ナツメグ」: https://hinoki-project.org/nutmeg/

書　名　第8回 コーパス日本語学ワークショップ予稿集
発行日　平成27年8月25日
発行者　国立国語研究所 言語資源研究系・コーパス開発センター
http://www.ninjal.ac.jp/organization/chart/03/
http://www.ninjal.ac.jp/organization/chart/06/
連絡先　〒190-8561　東京都立川市緑町10-2
大学共同利用機関法人 人間文化研究機構 国立国語研究所コーパス開発センター内
電　話　042-540-4300（代表）