第7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7回 コーパス日本語学ワークショップ

予稿集

2015年3月10日、3月11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7回 コーパス日本語学ワークショップ

予稿集

2015年3月10日(火)／3月11日(水)

Program [プログラム]

3月10日（火）

10:00～12:00　**■チュートリアル・デモ**

全文検索システム『ひまわり』を用いた既存資料の活用
既存のテキストデータを『ひまわり』にインポートし、検索（全文検索、形態素解析結果を利用した検索）する方法を説明・実習します。

12:00～13:00　**■昼食・休憩**

13:00～15:00　**■口頭発表**

コーパスシステム『Co-Chu』の開発 ―MeCab拡張データ処理機能について―
▷ラニガン・マシュー

文体指標と語彙系列の対応分析
▷浅原 正幸、加藤 祥、立花 幸子、柏野 和佳子

日本語話し言葉における情報構造と語順
▷中川 奈津子

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―
▷小磯 花絵、土屋 智行、渡部 涼子、横森 大輔、相澤 正夫、伝 康晴

15:00～16:00　**■ポスター発表　A グループ**

象は鼻が長いか ―テキストから取得される対象物情報―
▷加藤 祥

文書間距離尺度の特性
▷浅原 正幸、加藤 祥

BCCWJにおける固有表現抽出のエラー分析
▷市原 正陽、山崎 舞子、古宮 嘉那子

機械翻訳を用いた中古和文の現代語訳 ―分析と課題―
▷山田 祐実、大村 舞、岡 照晃、Kevin Duh、松本 裕治

日本語教育とコロケーション：連語の形で用法を学ぶ重要性
▷エルガ・ラウラ・ストラフェッラ

MCNコーパスにおける条件表現「たら」「れば」「ならば」のアノテーション
▷飯島 采永、佐藤 果穂、田中 リベカ、戸次 大介

代名詞・疑問詞を含む複合語の調査
▷淺尾 仁彦

16:00～17:00　**■ポスター発表　B グループ**

新しい日本語辞書定義文型の策定に向けて（第二報）
▷佐藤 理史、夏目 和子

コーパスコンコーダンサ『ChaKi.NET』のプロジェクト機能
▷浅原 正幸、森田 敏生

国語教育のための「常用漢字表」語例の検討
▷河内 昭浩

商品カテゴリの階層構造を用いた商品分類
▷中島 道幸、古宮 嘉那子

領域適応のためのサポートベクトルを用いた訓練事例の反復的選択
▷小林 優稀、古宮 嘉那子、佐々木 稔、新納 浩幸、奥村 学

会話における話者のうなずきと発話音声のプロミネンスの時間関係
▷天谷 晴香

述語項構造を意識した名詞データの構築
▷竹内 孔一、宮田 周、河村 一希

17:00　**■閉　会**

3月11日（水）

10:00～12:00　**■口頭発表**

コーパスに基づく日中副詞「絶対」と "绝对" の対照研究
▷郭 敏

中古歌合日記の品詞比率
▷冨士池 優美

BCCWJに拠る名詞別格外連体修飾形の形成傾向の分析
▷田邊 和子

代表性に配慮した『太陽コーパス』の分析法再考
▷森 秀明

12:00～13:00　**■昼食・休憩**

13:00～14:00　**■ポスター発表　A グループ**

BCCWJの接続詞の品詞情報の解析精度について
▷馬場 俊臣

『太陽コーパス』における語彙素「あう」の用字法
▷髙橋 雄太

『国民之友コーパス』に現れる一人称代名詞の計量的分析
▷近藤 明日子

『日本語話し言葉コーパス（CSJ）』の異なる講演タイプにおける外来語の質的分析
―言語外的および言語内的指標を用いた外来語分類の試み―
▷久屋 愛実

『児童・生徒作文コーパス』の設計
▷宮城 信、今田 水穂

『虎明本狂言集』のコーパスデータにおける短単位認定の諸問題
▷渡辺 由貴、市村 太郎、鴻野 知暁

否定の意志を表す「～まいとする」について
▷加藤 恵梨

14:00～15:00　**■ポスター発表　B グループ**

BCCWJに見る類義表現「～きる」「～ぬく」「～とおす」の使い分け
▷栗田 奈美

翻訳小説を資料とした品詞比率と文書間類似度による明治中期口語文体分析
▷小西 光

中古語複合形容詞の一語性 ―［名詞＋形容詞］とそれに類する複合形容詞的表現を中心に―
▷池上 尚

二字漢語名詞サ変用法の変化 ―『太陽コーパス』『BCCWJ』を用いて―
▷間淵 洋子

BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス
▷浅原 正幸、杉 真緒、柳野 祥子

上級～超級日本語学習者の作文から見た言語産出実態
▷趙 海城

医療経過記録における名詞連続の計量的特徴
▷山崎 誠、相良 かおる

15:00～16:30　**■指定討論**
▷柏野 和佳子、馬場 俊臣、小磯 花絵、冨士池 優美、古宮 嘉那子、佐藤 理史

16:30～17:00　**■全体討論**

17:00　**■閉　会**

Contents [目次]

■口頭発表

コーパスシステム『Co-Chu』の開発 ―MeCab拡張データ処理機能について― 1
ラニガン・マシュー

文体指標と語彙系列の対応分析 7
浅原 正幸、加藤 祥、立花 幸子、柏野 和佳子

日本語話し言葉における情報構造と語順 17
中川 奈津子

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― 27
小磯 花絵、土屋 智行、渡部 涼子、横森 大輔、相澤 正夫、伝 康晴

■ポスター発表 Aグループ

象は鼻が長いか ―テキストから取得される対象物情報― 35
加藤 祥

文書間距離尺度の特性 45
浅原 正幸、加藤 祥

BCCWJにおける固有表現抽出のエラー分析 55
市原 正陽、山崎 舞子、古宮 嘉那子

機械翻訳を用いた中古和文の現代語訳 ―分析と課題― 63
山田 祐実、大村 舞、岡 照晃、Kevin Duh、松本 裕治

日本語教育とコロケーション：連語の形で用法を学ぶ重要性 73
エルガ・ラウラ・ストラフェッラ

MCNコーパスにおける条件表現「たら」「れば」「ならば」のアノテーション 79
飯島 采永、佐藤 果穂、田中 リベカ、戸次 大介

代名詞・疑問詞を含む複合語の調査 89
淺尾 仁彦

■ポスター発表 Bグループ

新しい日本語辞書定義文型の策定に向けて（第二報） 95
佐藤 理史、夏目 和子

コーパスコンコーダンサ『ChaKi.NET』のプロジェクト機能 103
浅原 正幸、森田 敏生

国語教育のための「常用漢字表」語例の検討 113
河内 昭浩

商品カテゴリの階層構造を用いた商品分類 123
中島 道幸、古宮 嘉那子

領域適応のためのサポートベクトルを用いた訓練事例の反復的選択 129
小林 優稀、古宮 嘉那子、佐々木 稔、新納 浩幸、奥村 学

会話における話者のうなずきと発話音声のプロミネンスの時間関係 137
天谷 晴香

述語項構造を意識した名詞データの構築 143
竹内 孔一、宮田 周、河村 一希

■口頭発表

コーパスに基づく日中副詞「絶対」と "绝对" の対照研究 147
郭 敏

中古歌合日記の品詞比率 157
冨士池 優美

BCCWJに拠る名詞別格外連体修飾形の形成傾向の分析 165
田邊 和子

代表性に配慮した『太陽コーパス』の分析法再考 175
森 秀明

■ポスター発表 Aグループ

BCCWJの接続詞の品詞情報の解析精度について 185
馬場 俊臣

『太陽コーパス』における語彙素「あう」の用字法 195
髙橋 雄太

『国民之友コーパス』に現れる一人称代名詞の計量的分析 203
近藤 明日子

『日本語話し言葉コーパス（CSJ）』の異なる講演タイプにおける外来語の質的分析
―言語外的および言語内的指標を用いた外来語分類の試み― 213
久屋 愛実

『児童・生徒作文コーパス』の設計 223
宮城 信、今田 水穂

『虎明本狂言集』のコーパスデータにおける短単位認定の諸問題 233
渡辺 由貴、市村 太郎、鴻野 知暁

否定の意志を表す「～まいとする」について 241
加藤 恵梨

■ポスター発表 Bグループ

BCCWJに見る類義表現「～きる」「～ぬく」「～とおす」の使い分け 247
栗田 奈美

翻訳小説を資料とした品詞比率と文書間類似度による明治中期口語文体分析 257
小西 光

中古語複合形容詞の一語性 ―［名詞＋形容詞］とそれに類する複合形容詞的表現を中心に― 265
池上 尚

二字漢語名詞サ変用法の変化 ―『太陽コーパス』『BCCWJ』を用いて― 275
間淵 洋子

BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス 285
浅原 正幸、杉 真緒、柳野 祥子

上級～超級日本語学習者の作文から見た言語産出実態 293
趙 海城

医療経過記録における名詞連続の計量的特徴 303
山崎 誠、相良 かおる

口頭発表

3月10日（火）　13:00～15:00

コーパスシステム『Co-Chu』の開発
—MeCab 拡張データ処理機能について—

ラニガン　マシュー（中部大学大学院国際人間学研究科）†

Development of the Corpus System “Co-Chu”
—Regarding the MeCab Data Processing Extensions—

Matthew Lanigan (Chubu University)

要旨

本発表では、音声データを書き起こしたものを形態素解析にかける際に起こる問題点とその解決方法の一つとして、MeCab 拡張データ処理システムについて報告する。コーパスシステム『Co-Chu』は、コーパス検索だけでなく、コーパス開発のツールとして開発された。『名大会話コーパス』『日本語学習者会話データベース』『BTSJ による日本語話し言葉コーパス』をシステムに入れたところ、音声書き起こしコーパスに現れる学習者の誤用、言いよどみやフィラーなど、形態素解析のエラーを及ぼすものが様々あった。それらを排除する手段もあるが、そうすると分析対象とならないため、それらの問題点を補うシステムが必要となる。そこで、『Co-Chu』の開発の際に、特定の読みや出現形を選定するタグや辞書エントリーを一時的に導入するタグを付け、MeCab 拡張タグを開発した。本発表では、『Co-Chu』の MeCab 拡張データ処理システムとその仕組みについて報告する。

１．はじめに

話しことばコーパスの開発が困難になる原因がいくつか考えられる。

まず第 1 に、形態素解析を行う際、データがきれいでなければ、エラー率が非常に高まる可能性があり、話し言葉には様々な「きれいでない」要素が含まれている。例えば、言いよどみやフィラーなどが問題点として指摘されている(内元、野畑、山田、他 2003)。

第 2 に、コーパス開発および分析のためのツールは色々あるが、コンピューター技術に関する知識があまりなければ、使いこなすのは難しいと言えるだろう。

第 3 に、様々なツールがあっても、自分のデータで利用できる、『中納言』や『NINJAL-LWP for BCCWJ』のような強力なツールは少ない。

そこで、オープンソースソフトウェア（OSS）のコーパスシステム『Co-Chu』の開発を試みた。本発表では、『Co-Chu』の MeCab 拡張データ処理機能を中心に報告する。

２．『Co-Chu』の概要

コーパスシステムというのは、大きく分けて、コーパス開発とコーパス分析という２面で構成される。『Co-Chu』とは「中部大学コーパスシステム（Chubu University Corpus System）」の略である。現在システムは開発中であり、公開できるものになっていないが、近日中には公開予定である。

概要に入る前、本システムは『BCCWJ』や『日本語話し言葉コーパス』のような大規模コーパスの開発に利用されるためには作られていないことを注意しておきたい。『Co-

† lanigan@isc.chubu.ac.jp

Chu』に含まれている MeCab 拡張タグなどのコーパス開発ツールはほぼ必ず手作業を必要とするものであり、データの量が多ければ多いほど手におえなくなるだろう。コンピューター技術に詳しくない個人の言語研究者や小グループで開発されているコーパスを念頭に開発しようとしている。しかし、大規模コーパスのために作られていないとはいえ、データの量が非常に多くても機能するように配慮した。

２.１　システムの構造

『Co-Chu』はデータベース、アプリケーション・プログラミング・インターフェース（API）、ユーザー・インターフェース（UI）の 3 階層構造になっている。この構造を使用することにより、システムの拡張が容易になると期待できる(日本 OSS 推進フォーラム 2015)。

２.１.１　データベース

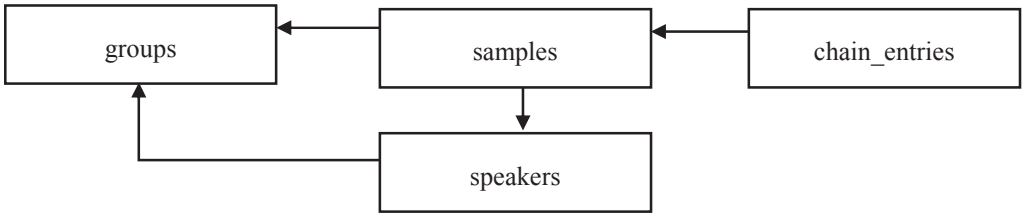


図 1 データベース構造の概要

システムの基盤となるのはデータベースである。『Co-Chu』のデータベースは PostgreSQL という OSS のリレーショナルデータベース（RDB）を利用している。他にもあるが、最も重要なテーブル（groups, samples, speakers, chain_entries の 4 つ）とその関係を図 1 に示している。

表 1 テーブルの構造

groups	
id	INTEGER
parent_id	INTEGER
name	TEXT
metadata	JSONB

samples	
id	INTEGER
group_id	INTEGER
speaker_id	INTEGER
name	TEXT
metadata	JSONB

speakers	
id	INTEGER
group_id	INTEGER
metadata	JSONB

スピーカーのテーブルがあるが、話し言葉に限られていないことを注意しておきたい。形態素解析などにおいて、話し言葉データの処理が特に困難であるため、本システムの大半の機能は話し言葉の処理のために向けられている。

サンプルというのは基本的に発話（、あるいは書き言葉の場合の文章）を示し、グループはサンプルの集まり、つまりサブコーパスにあたるものとする。BCCWJ において「サンプル」というのは作品や記事などを表すが、本システムにおいて作品や記事はグループになるのである。次に、スピーカーは発話者の関係を表し、基本的に話し言葉データでし

か利用されない。

表 1 で見られるように、以上の 3 つのテーブルは全てメタデータコラムがある。これは JSON フォーマットの非構造化データであり、PostgreSQL 9.4 の JSONB タイプによってインデックスされている。

最後に、本システムのコアとなるテーブルは chain_entries である。このテーブルには、一つのサンプルの形態素解析結果をチェインとして保存してある。つまり、MeCab による形態素解析結果に加え、表 2 に見られるように、ID 番号（id）と親 ID 番号（parent_id）があり、形態素の連鎖になる。

表 2「chain_entries」テーブルのデータ例（一部のコラム）

sample_id	id	parent_id	surface
89067	3987301	(NULL)	これ
89067	3987302	3987301	は
89067	3987303	3987302	コーパス
89067	3987304	3987303	システム
89067	3987305	3987304	で
89067	3987306	3987305	ある
89067	3987307	3987306	。
89067	3987308	3987307	(BOS/EOS)

２.１.２　API

第 2 階層になるのは Ruby で作られている REST API である。このインターフェースを通して、UI がデータベースにアクセスする。また、API には MeCab とのインターフェースがあり、そこに MeCab 拡張データ処理機能が入る。つまり、API が UI からデータ処理リクエストを受け、MeCab にデータを転送する前に拡張タグをプロセスしておく。それから、MeCab から形態素解析の結果を読み、タグに含められた指示に従い、データを処理し、データベースへ転送する。具体的なタグについては後述する。

２.１.３　UI

『Co-Chu』の第 3 階層である UI を入れ替えることができるが、グループの間での共有を考えてウェブ・インターフェースにした。未公開データなどは、セキュリティーが重要であると考えられ、ユーザー登録を必要とする。現在管理人しか新しいアカウントが作成できないが、オープンな設定にすることも考慮している。

２.２　コーパス開発機能

API／UI には様々なコーパス開発の際に役立つと思われる機能を付加しており、ここでは、開発済みもしくは開発予定の機能を簡単に紹介する。

- 組み入れ無制限のグループ構成
- WYSIWYG データ移入ツール（TXT、CSV/TSV、Excel）
- MeCab の n-best 機能に基づいたビジュアル・エラー処理ツール
- グループ特定のユーザー辞書

MeCab 拡張タグについては後述する。

２.３　コーパス分析機能

次に　簡単にコーパス分析機能を紹介する。

- 形態素連鎖の詳細検索
- いくつかのアウトプット形式（KWIC など）
- TSV ファイルへの輸出
- N-gram に基づいたコロケーション(Wei and Li 2013)
- 検索結果統計とグループ別比較

３．MeCab 拡張データ処理機能

上述したように、MeCab 拡張データ処理機能は API の一部であり、API が MeCab とインターフェースする前後に行われている。執筆時点で作成されているタグは以下の 5 種類である。フォーマットは基本的に{{<command>{<options>}}}の形をとっている。

command というは短いローマ字のタグセレクターであり、options はタグそれぞれで異なるパラメーターである。しかし、一般的にタグの最初のパラメーターはターゲットとなっている語である。

３.１　読み選定タグ {y{A|B}}

読み選定タグ（y）によって、ある文字の特定読みを選択するタグである。パラメーターは対象語と読みの 2 つである。例えば、{y{昨日|さくじつ}}で見られるように、「昨日」の中の「さくじつ」の読みを選択している。

この機能が重要なのは、書き言葉と異なり、話し言葉はもともと「字」ではなく「音」であるため、話し言葉コーパスの開発には発音が最も重要な要素である。それにもかかわらず、単に「昨日」を形態素解析に入れると、ほぼ確実に「きのう」の読みが出力される。このような例が他にも様々ある。例えば、「後」「明日」「家」などが挙げられる。

最後の「家」については、話し言葉でたまに「俺(おれ)ん家(ち)」のような例がみられるが、MeCab と UniDic で形態素解析を試みると、

出現形	出現形 発音形	語彙素 発音形	語彙素	品詞
俺	オレ	オレ	俺	代名詞
ん	ン	ノ	の	助詞-格助詞
家	イエ	イエ	家	名詞-普通名詞-一般

という結果があり、「家(いえ)」として解析されている。また、「俺んち」の形にしておいても、

出現形	出現形 発音形	語彙素 発音形	語彙素	品詞
俺	オレ	オレ	俺	代名詞
ん	ン	ンー	んー	感動詞-フィラー
ち	チ	チ	チ	記号-一般

のように出力され、「ち」が記号となっている。MeCab の n-best 機能を利用すれば、正しい読みを選択することができ、発音と合致した結果にできる。

以上の例は UniDic によるエラーであるとしても、発音を重視しながら形態素解析を行う際に必ず他の語にも現れる問題である。

３.２　語形選定タグ{w{A|B}}

読み選定タグと違い、語形選定タグはただ選択するのではなく、結果的にある語形に新しい出現形を作り上げるタグである。最初のパラメーターは読みタグと同様、対象語である。しかし、このタグの第 2 のパラメーターは語形となっている。なぜなら、日本語学習者の誤用などを表す使用の仕方が考えられる。例えば、{w{きー|来}}た では、ある学習者が発音を間違え、「来た」の「き」を長音にする。本来ならこれは「来た」に処理されるか、形態素解析後の手作業で直されるか、エラーになるかだが、このタグを利用し、「来」の新しい出現形「きー」を特定な箇所に限ってつけることができる。

つまり、このタグによって、意味も発音も保存され、以上の例からの出力は以下のようになる。

出現形	出現形 発音形	語彙素 発音形	語彙素	品詞	活用型	活用形
きー	キー	クル	来る	動詞-非自立可能	カ行変格	連用形-一般
た	タ	タ	た	助動詞	助動詞-タ	終止形-一般

API にはこのタグは MeCab にデータを転送する前に、A を B に置き換え、結果の B に相当する語の出現形を A に置き換える。

３.２.１　誤用 {err{A}}

このタグは語形制定タグに関連するショートカットであり、他のタグと違い、A は語だけではなく、他のタグを入れることができる。このタグによって、「誤用」というエントリーが対象語の用法コラムに追加される。例えば、語形選定タグの例につけることが、{err{{w{きー|来}}}}のようにできる。

３.３　辞書エントリータグ {d{A|B1,B2,…,Bn}}

辞書エントリータグによって一時的に辞書エントリーを追加することができる。第 2 のパラメーター（B1,B2,…Bn）は UniDic のユーザー辞書のフォーマットになる。基本的にこのタグを直接使う場が少なく、他のタグが利用するためにある。

３.３.１　フィラー {f{A}}

フィラータグによって、何かを 1 語のフィラーとして扱わせることができる。例えば、状況により「ん」が助詞の「の」として認識される場合があり、それを{f{ん}}にすればフィラーになる。

API には、このタグをプロセスするとき、対象語のみがフィラーに変えられるために、まず対象語にプレースホルダ―を置き換える。プレースホルダ―のエントリーを一時的に

ユーザー辞書に追加し、形態素解析を行う。それから、プレースホルダ―が結果に出たら、また対象語をそこに置き換える。

４．まとめ

以上『Co-Chu』の MeCab 拡張データ処理機能を中心に報告した。本システムはコンピューターに詳しくない研究者などが同じインターフェースを通してコーパス開発と分析ができる。また、話し言葉の形態素解析とデータ処理に役立つシステムである。現在、読み選定タグなど、MeCab のユーザー辞書と n-best 機能に基づいたいくつかの MeCab 拡張タグを利用することができる。システムのタグをさらに増やし、話し言葉データの本発表に触れていないの問題点（同時発話や相杏など）に対する解決策は今後の課題としたい。

謝辞

本研究を進めるにあたり、『Co-Chu』のテスター役を含め実際にシステムをご利用くださっている中部大学の山本裕子先生、本間妙先生の貴重なご助言に厚く御礼申し上げます。また多岐にわたるご指導を賜りました小森早江子先生に、心より感謝申し上げます。

参考文献

内元清貴、野畑周、山田篤、関根聡、井佐原均（2003）「日本語話し言葉コーパスの形態素解析」『言語処理学会　第 9 回年次大会　発表論文集』pp.113‐116.

日本 OSS 推進フォーラム（2015）「Ⅱ-3-5. OSS による Web3 層アプリケーション」2015 年 2 月 3 日参照. http://ossforum.jp/en/node/813.

Wei, Naixing, and Jingjie Li (2013). “A New Computing Method for Extracting Contiguous Phraseological Sequences from Academic Text Corpora.” *International Journal of Corpus Linguistics*. 18:4, pp.506–35. doi:10.1075/ijcl.18.4.03wei.

関連 URL

『Co-Chu』http://co-chu.org/. 執筆時点で未完成.

『名大会話コーパス』『日本語学習者会話データベース』2014 年 12 月参照. http://dbms.ninjal.ac.jp/nknet/

『BTSJ による日本語話し言葉コーパス』2015 年 2 月 3 日参照. http://www.tufs.ac.jp/ts/personal/usamiken/btsj_corpus.htm

『MeCab』2015 年 2 月 3 日参照. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html

『UniDic』2015 年 2 月 3 日参照. http://www.ninjal.ac.jp/corpus_center/unidic/

『中納言』2015 年 2 月 3 日参照. https://chunagon.ninjal.ac.jp/

『NINJAL-LWP for BCCWJ』2015 年 2 月 3 日参照. http://nlb.ninjal.ac.jp/

『PostgreSQL』2015 年 2 月 3 日参照. http://www.postgresql.org/

『Ruby』2015 年 2 月 3 日参照. https://www.ruby-lang.org/

『AngularJS』2015 年 2 月 3 日参照. https://angularjs.org/

文体指標と語彙系列の対応分析

浅原 正幸 ‡†　加藤 祥 †　立花 幸子 †　柏野 和佳子 ‡†
（国立国語研究所　† コーパス開発センター　‡ 言語資源研究系）

Correspondence Analysis between Writing Styles and n-gram/p-mer

Masayuki Asahara, Sachi Kato, Sachiko Tachibana, and Wakako Kashino
(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。本研究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う。

1. はじめに

コーパス調査において重要な要素として，利用するサンプルの文体情報がある。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 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。

本研究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) との対応分析を行う。具体的には語彙系列から高頻度の n-gram・p-mer を特徴量として，文書-特徴量行列を構成し，制約付き主成分分析を行った。分析結果について報告する。

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

表 1　文体指標と日本十進分類法 (NDC) の対応

先行研究	文体指標	高い		低い	
		NDC	タイトル例	NDC	タイトル例
柏野・奥村（2012b）	専門度	1	『大衆社会のゆくえ』『日常生活の認知行動』	9	全般
		3	『注釈会社法』『刑事司法を考える』『アジアの成長と金融』	無	全般
柏野・奥村（2012b）	硬度	1	『新・岩波講座哲学』『親鸞の思想構造序説』『日本近代思想大系』	7	『競馬探偵の憂鬱な月曜日』『松井秀喜メジャー物語』『芸能界デビュー』『二代目さん』
		2	『西洋中世都市の自由と自治』『日本近代社会成立期の民衆運動』『岩波講座近代日本と植民地』	9	『吉永さん家のガーゴイル』『耽美小説の書き方』『アンネの日記』『小説道場』
		3	『現代国際法』『フランスの社会保障』『情報ネットワーク社会の展開』		
柏野ほか（2012a）	くだけ度	7	『ほんじょの眼鏡日和。』『清水ミチコの「これ誰っ！？」』『関根勤のやるだけマルもうけ！！』	4	『日本の昆虫』『フォッサマグナ』『生体物質とエネルギー』
		9	『ヒミツの転校生』『少しだけラブストーリー』『いつでもこの世は大霊界！』	8	『暗号の数理』『大学で教える小論文の書き方』『語源をさぐる』
加藤ほか（2014）	語りかけ性度	15	『いま、四十代を生きる女へ』『サラリーマンなんか今すぐやめなさい』『敬語マニュアル』	ハウツー本でない本全般	
		49	『消化器ガン克服マニュアル』『読むだけで絶対やめられる禁煙セラピー』『妊娠レッスン』		
		59	『わかりやすいイタリア料理』『お産と育児のスーパーアドバイス』		
		8	『はがきの書き方』『人に好かれることばレッスン』『講師・講演を頼まれたら読む本』		
柏野・奥村（2012b）保田ほか（2013d）	客観度	1	『発達心理学入門』『認知とパフォーマンス』『宗教改革』『岩波講座東洋思想』	2	『インドでわしも考えた』『泣いたら負けや』『ほいほい旅団ベトナムへ行く』
		4	『超分子化学への展開』『生物科学入門』『環境化学』『ガン全種類別・最新治療法』	9	『日はまた熱血ポンちゃん』『マンボウ交友録』『パパは塾長さん』『男はオイ！女はハイ…』

NDC は日本十進分類法 (NDC 分類) の上位桁を表す。無はサンプルに NDC 分類が付与されていないもの。

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。

表 1 に先行研究で言及されている文体指標と日本十進分類法 (NDC) の対応をタイトルとともに示す。

2.2　特徴量の取得

本研究では語彙系列の n-gram (連続部分列) と p-mer (非連続部分列) を特徴量とする。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。

系列特徴量の枚挙には，系列パターンマイニングアルゴリズム prefixspan [(1)](Pei et al. (2001)) を用いる。系列特徴量の枚挙の詳細な仕様については以下の通り：

- 系列中，同じパターンが複数回出現する場合は全ての組合せを枚挙する。“ABABAB” の中に “AB” は 3 回，“A/B” は 3 回出現したと数える。
- ギャップ長については制限を課さない。
- ギャップの挿入箇所が違う p-mer は別のものとして扱う。
- 枚挙は頻度 10 以上のもの全てを展開した。

なお，文体指標付与時には台詞をのぞいた地の文のみを対象としているが，特徴量の展開は台詞・地の文の識別を行わず，可変長サンプル全体から行った。

2.3　対応分析

対応分析はクロス集計表の行と列の双方を並び替えることにより，行の項目と列の項目との相関関係を最大化するような処理を行う。基本的には主成分分析と同様にデータの分散を最大化する方向の軸（主成分）を逐次的に求め，説明変数を合成するという処理を行う。軸の選択は条件付極値問題として定式化でき，ラグランジュの未定乗数法によって解くと相関行列の固有値，固有ベクトルを求める問題に帰着する。値の大きい固有値に対応した軸から順に第 1 主成分，第 2 主成分と呼び，各軸は直交する。全固有値の総和で，各主成分に対応する固有値を割ったものを寄与率と呼び，各主成分によりどの程度説明ができているかの尺度となる。同様に第 1 主成分から第 α 主成分までの寄与率の和を第 α 主成分までの累積寄与率と呼び，当該主成分まででどの程度説明ができているかの尺度となる。

実際の計算には R の MASS パッケージを用いた。prefixspan により枚挙した特徴量を R のデータフレームのカタチで表現したあとは，R のプログラムとしては数行のものである。この数行を Latent Dirichlet Allocation (Blei et al. (2003)) のパッケージに置き換えることにより，語彙系列-文書間のトピックモデルを作成することも可能である。

[(1)] 実装は `http://prefixspan-rel.sourceforge.jp/` を用いた。

表 2　専門度と語彙系列の対応分析結果（第 1 主成分の上位 5 件・下位 5 件）

2gram 下位	3gram 下位	4gram 下位	5gram 下位
使用-者	適用-為る-れる	べし-だ-有る-と	れる-ない-ば-成る-ない
所有-権	労働-者-の	た-場合-に-は	次-の-様-だ-述べる
債権-者	規定-為る-れる	に-於く-て-は	化-為る-れる-て-いる
結合-為る	有する-て-居る	示す-れる-て-居る	構成-為る-れる-て-居る
公共-団体	た-場合-に	れる-ない-ば-成る	的-だ-物-だ-ある
2gram 上位	3gram 上位	4gram 上位	5gram 上位
御-化け	が-言う-ます	が-言う-ます-た	成る-て-居る-ます-た
姉-ちゃん	御-姉-ちゃん	御-婆-ちゃん-は	そう-だ-顔-を-為る
御-婆	婆-ちゃん-は	御-父-さん-は	て-仕舞う-た-の-です
兄-ちゃん	言う-ます-た	て-行く-ます-た	何時-も-の-様-だ
婆-ちゃん	御-婆-ちゃん	の-御-父-さん	どう-為る-た-の-だ

4mer 下位	5mer 下位	6mer 下位	7mer 下位
的 的 為る 為る	的 に て 為る 為る	は の 的 の の 為る	の の 為る 為る の 為る 為る
於く の に 為る	為る の 的 為る 為る	の 的 の の に 為る	の の 為る と の 為る 為る
的 に 的 為る	に て 的 為る 為る	は の 的 の に 為る	の に 為る 為る の 為る 為る
於く は の 為る	に 的 の 為る 為る	的 の の の を 為る	の 為る 為る の の 為る 為る
的 の 的 為る	に 的 に 為る 為る	の の 的 の に 為る	の 為る の 為る の 為る 為る
4mer 上位	5mer 上位	6mer 上位	7mer 上位
御 さ ん て た	は て を ま す-た	は の を て ま す-た	は の の の の を て ま す
さ ん-は を た	の て-居る-ま す-た	は を て を ま す-た	の の に の を て ま す
さ ん-は て た	は て て ま す-た	を て を て ま す-た	の に て の を て ま す
さ ん-は を て	を て て ま す-た	は て を て ま す-た	は の に の を て ま す
は を ま す-た	は を て ま す-た	は を て て ま す-た	の を て の を て ま す

3.　結果と考察

3.1　結果

結果のテーブルは大部なため，`http://goo.gl/4aFQTj` に公開する。参考のため表 2 に専門度と語彙系列の対応分析結果のうち第 1 主成分の下位 5 件・上位 5 件を示す。

3.2　考察

これらの文体指標の分析例には柏野・奥村 (2012b) があり，NDC 別の特徴をまとめている。先行研究に見られるサンプル群の NDC 分布の傾向から，どのような書籍が特徴的と考えられる NDC 群に含まれているのかを調べ，本手法で取得された表現が実際どのように現れているのか，特に 3gram, 4gram, 5gram, 4mer について例示する。

硬度とくだけ度については，柏野ほか (2012a) が，目視によって得られた文末・主語（一人称）・語彙（親密度が低そう・平易）などの特徴を列挙している。文末表現や主語については，本手法で得られた特徴と重なる部分がある。語りかけ性度については加藤ほか (2014) が特徴的表現を調査している。また，保田ほか (2013d) は，語りかけ性度との関連において，客観度の高低が，データ・伝聞などの表現に見る根拠の有無によって判断されているとの観点で客観度に関わる特徴的表現例を示した。先行研究で示された表現との異同を確かめ，本手法で新しく捉えられた特徴と捉えられなかった特徴を確認する。

3.2.1　専門度

最初に専門度で得られた語彙系列特徴量の第 1 主成分の偏りについて示す。

- 3gram

  「条＋数詞＋項」「適用される」「規定される」「対象とする」「裁判所の」「有している」「於いては（も）」「主義的だ」「合理的だ」「目的とする」⇔「お姉ちゃん」「おばあさん」

「たのかしら」「のだぞ」「たわね」「のかい」「なんだか」「私だって」「どうかする」

- 4gram

  「た場合には」「原則として」「たものとする」「であるとする」「べきであると」「示されている」「前提として」「明らかにされる」「することによる」「なされている」⇔「が言いました」「て行きました」「やって来ます」「たのでした」「ちゃったのだ」「って言うのは」「どうしたの」「もしかして」「あっと言う間」

- 5gram

  「れなければならない」「するのではない」「構成されている」「行われていた」「次のように述べる」「事を示している」「を意味している」「されることになる」「ということである」「であるという」⇔「していました」「てしまったのです」「ありませんでした」「のでないかな」「いつの間にか」「なんとかして」

- 4mer

  「的～的～する～する」「的～の～的～の」⇔「さんは～を～た」「さんが～て～た」「その～に～ました」

このうち 4mer の「的～的～する～する」「的～の～的～の」のような一文中に「的」が頻出するパターンは過去の研究（頻出語彙や人手による検討）で捉えられなかった特徴である。

例（LBi3_00033 『現代法社会学入門』321）[(2)]

前述のように、パレート最適という価値判断基準とコースの定理によって、法的ルールや制度について効率性の観点からの理論的分析や実証的研究、さらには規範的な問題提起や提言をすることが可能となる。

以下の例では特徴的な語彙系列を含む専門度の高い例を示す。

例（LBj1_00010『日常生活の認知行動』141）

議論の余地があるのは、もしひとつの方法だけがすべての問題に自動的に適用されるなら、さまざまな比率の比較の意味をきちんと保ったままでいるのはもっとやさしいだろうということである。しかしながら、人々は、このようなやり方で自分たちの日常生活を単純化してしまうことはない。

以下では特徴的な語彙系列を含まないが，専門度が低い例を示す。この例では語彙系列ではなく表記から専門度が低いと判定されたと考えられる。

例：小学生・幼児向き（LBcn_00024『宇宙人はほんとにいるか？』NDC 分類無）

この天の川銀河にどれくらいの数の惑星系があるのかはまだわかっていませんが、これまでの観測結果からはずいぶんあるだろうという予想がたっています。そのなかには、地球ににた、生命をやどしている星もあるにちがいありません。

以下では特徴的な語彙系列を含まないが，専門度が高い例を示す。語彙が分野的に特化している場合や，表現の共起関係が一般的でない場合は特徴が表れにくい。

例：（LBl0_00010『最終講義』041）

これを政治史でみますと、中国の国民革命軍が北伐を完成したのは一九二八年、いまま

(2) 括弧内は，サンプル ID・『タイトル』・NDC 分類。以下同様。

で分裂していた中国が北伐の完成によって統一され、軍閥、とにかく国民党が南京に都を置いて、中央政府ができてきました。そのもとで中央研究院歴史語言研究所というような研究機関が設立され、そこで安陽の発掘ができ、またいままで政治闘争にあけくれていた大学でありましたが、古典の研究などに本腰を入れてやりだした。

3.2.2　硬度

硬度について第 1 主成分の偏りのある語彙系列を確認する。

まず，2gram でレジスタ特徴の見られる名詞的表現が得やすい：

「条約の」「権の」「規定に」「著作権」「債権者」「使用者」「公共団体」「損害賠償」

さらに，3〜5gram および 4mer では名詞的表現だけでなくレジスタ特徴のあるコロケーションが得られている：

- 3gram

  「条＋数詞＋項」「労働者の」「重要な役割」「第一次」「委員会が」「裁判所の」「市町村」「我が国の」「総合的＋だ（助動詞）」「普遍的＋だ（助動詞）」「全国的＋だ（助動詞）」「象徴的＋だ（助動詞）」「有している」「提出される」「問題となる」「原則とする」「ように述べる」「化すること」「べきである」「するものと」「大きな影響を」「明らかである」「既に述べた」「結果である」⇔「なっちゃった」「って言われる」「こうやって」「思ってた」「だよね」「僕らの」「たのよ」「お姉ちゃん」「お兄ちゃん」

- 4gram

  「するものである」「重要な役割を」「すべきである」「あるとして」「と述べている」「役割をはたして」「のであるから」「これに対して」「都道府県」「第二次大戦」⇔「が好きなの」「してたの」「かなと思う」「なのですよ」「でも私は」「ありがとうございました」「そんな事を言う」

- 5gram

  「次のように述べる」「存在していた」「よく知られている」「このようにして」「なったのである」「おかなければならない」「第二次世界大戦」⇔「かもしれないけれど」「のではないです」「いませんでした」「聞いたことがある」

- 4mer

  「於いて〜する〜て」「於いて〜の〜が」⇔「私は〜ました」「僕は〜て〜た」

このうち 4mer の「於いて〜する〜て」「於いて〜の〜が」のようなパターンは過去の研究（頻出語彙や人手による検討）で捉えられなかった特徴である。

例（LBp1_00005『連続性の哲学』133）

いかなる演繹に おいても その実験に おいて 生じる全過程は、厳密にこの三つのもの、すなわち、総括、複製、消去である。それ以外には結論の観察が残るだけである。ただし、これらの三つの実験の要素は、あらゆる演繹に おいて 生じるわけではない。

以下では特徴的な語彙系列を含む硬度が高い例を示す。

例（LBi2_00076『弥生の王国』210）

復元にあたって問題になるのは尺度のことである。わが国古代には人体尺で建物を測っていたということを前提にして、このたびの復元も人体尺によることにした。そのこと

について説明して おかなければならない。

3.2.3 くだけ度

くだけ度の第 1 主成分の偏りのある語彙系列を確認する。

- 2gram
  「しちゃう」「けれどさ」「てさ」「ちゃって」「かよ」「ちゃん＋助詞」「こと言う」「訳ない」「よな」「だぜ」「ちまった」「てるって」⇔「意思決定」「連合国」「出土する」「軍事的」「東アジア」「労働組合」「安全保障」「削除する」「分布する」「承認する」「経済学」
- 3gram
  「お姉ちゃん」「つうことは」「てるのよ」「言ってた」「やってるの」「なっちゃった」「だからね」「そうだよ」「俺には」「だったな」「ですよね」「馬鹿にする」⇔「事が明らか」「例である」「総合的だ」「進められて」「形成される」「中小企業」「実質的な」「記している」「理解される」「我が国で」
- 4gram
  「のだよね」「なのだけれど」「と思ってた」「どうでも良い」「気にしない」「のだなと」「してくれない」「だとすると」⇔「示されている」「対象として」「指摘している」「とされた」「にあたっては」「原則として」「認められている」全体として」「具体的には」
- 5gram
  「かもしれないけれど」「でないかと思う」「な気がした」「見たことがない」「なっていました」「それにしても」「身につけている」⇔「と考えられている」「的なものである」「事を示している」「よく知られている」「行われていた」「られたのである」「ようなものである」
- 4mer
  「僕は～て～た」「そうな～を～た」「ような気～する」「目を～て～た」「どうして～の」「だから～のだ」⇔「於いて～に～する」「的～を～的～する」「このよう～の～する」

以下では特徴的な語彙系列を含む，くだけ度が高い例を示す：

例（LBn7_00044『ほんじょの虫干。』770）
堀辰雄（『風立ちぬ』）と並んで、当時 “四季派” と呼ばれた詩人達の中の代表的な人物。建築デザイナーでもあったんだよ。かっこいい！　初めてその詩に触れた時には、そのあまりの美しさに感激して、同時にそのあまりの儚さに（大丈夫？）って驚いたものでした。なよなよっとしててねぇ、デリケートでナイーブで神経質で、こりゃ早く死ぬのも無理ない よなぁ、って思わせるの。

以下では特徴的な語彙系列を含む，くだけ度が低い例を示す：

例（LBi4_00028『岩波講座現代の物理学』420）
われわれは, なにゆえに, 自由な Dirac 場が Fermi 統計をみたし, 自由な Klein-Gordon 場が Bose 統計に従うかをみてきた. 相対論的な自由場はこれ以外にもさまざまなスピンを記述するものが知られている. これらを逐一述べるのは本書の目的ではない. ただスピンと統計の関係として次の定理 (W.Pauli,1940) は よく知られている.

語彙系列のみからは取得しにくい特徴としては，特殊なオノマトペの多用，破格，話し言葉的な用法（例：接続詞としての「が」の用法など）がある：

例（LBk9_00244『爆れつハンター special』913）

ドクッ、ドクッ、ドクッ、ドクッ…

心音はもはや限度を超えていた。

「はっ！」

シャアアアアアアアアアアアアアアアアアアアッ！

鞭が男めがけて飛ぶ。

「！」

が、外れた。

3.2.4　語りかけ性度

語りかけ性度においては 2gram で小説以外らしき語彙系列がいくらか特徴的にみられる：

- 2gram

  「おすすめ」「患者さん」⇔「固有名詞＋は」「竦めた」「微笑んだ」「理事会」「帝国主義」「県警」「共産党」「連合軍」「首相の」

- 3gram

  「と良いです」「が大切です」「たいものです」「クリックして」「できるのです」「紹介します」「説明します」「がわかります」「方法がある」「表示される」「お客様」「見てくださる」「ないですか」「ますよね」「ていけば」「する時に」⇔「彼女は言う」「声だった」「は笑った」「歩き出した」「次の瞬間」「背を向ける」「電話を切る」「闇の中」「時計を見る」「階段を下りる」

- 4gram

  「ようにします」「こともあります」「ば良いのです」「しておきます」「ことなのです」「かと言うと」「ている人は」「おいたほうが」「てみたいと」⇔「彼は言った」「姿をあらわした」「黙っていた」「頭を下げた」「ため息をついた」「何も言わない」「逮捕された」「勤めていた」

- 5gram

  「訳ではあります」「ではないですか」「することができます」「はありませんか」「なければなりません」「ことはできません」「ようになりました」⇔「首を横に振る」「な気がした」「いたのだった」「たというのだ」「行われていた」「顔をしている」「呼ばれていた」

- 4mer

  「言う～と言う～です」「言う～と言う～ます」⇔「いたが～た」「いたが～に」

以下では特徴的な語彙系列を含む，語りかけ性度が高い例を示す：

例（LBj5_00015『すぐ役立つ家庭の電気百科』592）

金属は電波を反射してしまうので、電子レンジで加熱するときの器としては使えませんが、余分な加熱を抑えたいところ、例えば前述の徳利の上部や、魚の尾ビレ、肉の脂身などにアルミホイルをかぶせると、不必要な加熱を抑え、上手に調理することができます。

例（LBf2_00022 『石塚山古墳の謎』210）

これは先程分類しました鏡が、今度は各地でどういうふうに散らばって古墳で出てくるか という ことを示したものであります。たとえば一番古い鏡 という のは、その表で言いますと、山陽、畿内、三重、東海、関東とゴシックで書いておりますように、多いです ね。パーセントで言うと、畿内で三八%、九州にもあります。

3.2.5 客観度

- 3gram

「条一項」「をクリックする」「設定される」「原則とする」「消費者の」「定められて」「困難である」「確認される」「場合には」「理由とする」「これによる」⇔「だったな」「そうだよ」「のですって」「私のこと」「生まれて初めて」「好きになる」「なんだか」「ているのだ」「そうかも」「僕には」「私はもう」

- 4gram

「用いられている」「ことを示して」「とも呼ばれる」「たことがわかる」「重要な役割を」「数詞（千九百〜）」「例として」「発見された」「を明らかにする」⇔「のだよね」「でないかな」「言ってるのだ」「ちゃったのだ」「でも私は」「が好きなの」「かなと思う」「なんとかなる」「気がした」「私にとって」「のだそうだ」「おぼえていない」

- 5gram

「と考えられている」「事を示している」「することができます」「となっています」「れなければならない」「とされてきた」「として知られる」⇔「のでないかな」「と思ったのだ」「とは思わなかった」「いつものようだ」「たのだと思う」「聞いたことがある」「たことがあります」

- 4mer

「因る〜が〜する〜する」「因る〜の〜の〜れる」⇔「僕は〜」「私は〜」「私が〜」「私の〜」

以下では特徴的な語彙系列を含む，客観度が高い例を示す：

例（LBa3_00047『地方財政の国際比較』349）

しかもジョーンズらに よれ ば，左右を問わず地域内外 の 政治的党派的集団 の 動き が 地方団体政策決定 の 焦点となり，伝統的な委員会制度 の 機能が弱まりつつあるなど（６４），地方団体の運営にもこれら政治状況の影響が及んでいる。

例（LBo4_00040 『超分子化学への展開』431）

細胞外から の シグナル分子 が 膜受容体に結合 する と，Ｇタンパク質の働きに よっ て，細胞内部 の 酵素 が 活性化 さ れる．人工脂質膜上に人工受容体と酵素を固定化 すると，シグナル化合物に よっ て酵素活性 が 増加 する 人工システムを構築できる．

4. おわりに

本稿では，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 サンプルに対して付与された専門度，客観度，硬度，くだけ度，語りかけ性度の 5 種の分類指標に対して，品詞ごとに語彙系列を特徴量とした対応分析を行い，先行研究で言及されている定量的・定性的分析との比較調査を行った。

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

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.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2003). “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, 3, pp. 993–1022.

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.

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.

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.

日本語話し言葉における情報構造と語順

中川奈津子 (同志社大学 全学共通教養教育センター)

Information Structure and Word Order in Spoken Japanese

Natsuko Nakagawa (Doshisha University)

要旨

本発表では、日本語話し言葉コーパスの模擬講演と対話を用い、日本語話し言葉の語順と情報構造の関係について調査した。(1) まず、コーパスから名詞・代名詞とゼロ代名詞を抽出し、(2) それぞれの照応関係と項構造を同定した。(3) そして、それぞれの表現の情報の新旧を調べ、新情報 (先行詞なし) と旧情報 (先行詞あり) に分類した。(4) 次に、コーパスの語順情報を使って、述語ごとに名詞・代名詞の語順を同定し、語順と関連する属性を調べ、(5) それぞれの用例を詳しく分析した。以上の調査から、(1) 定情報を表す表現は節頭に、(2) 焦点と関わる名詞は動詞の直前に現れるという結果を得、先行研究の指摘を実証的に確かめた。また、(3) 述語の後（発話末）に現れる要素は活性化コストが低い名詞であることを明らかにし、(4) 以上の傾向の動機づけを提案した。

1　はじめに

1.1　本研究の問い

本研究では、日本語話し言葉における情報構造と語順の関係を探る。例えば、以下の会話の断片 (1) を見てみよう。

(1)　a. **A1:** 楽しいねー**音楽**

b. **B2:** うん 楽しいよー

c. **A3:** いいなーちょっと音楽部に入りたかったなー

d. **C4:** (イナ) 今からでも**合唱団**どう　　(**千葉大 3 人会話コーパス** 0332:　72.69-81.30)

(1a) では、「音楽」が述語の後に生起しているが、一方、(1d) における「合唱団」は述語の前に生起している。このような語順を決める要因は何だろうか？

さらに会話の断片 (1) を詳しく観察してみると、(1a) はすでに「音楽」の話が出たあとの発話であることに気がつくかもしれない。それに対して、(1d) における「合唱団」は、話し手によって新たに提案されている。このように、すでに以前に出たとか、新たに提案されているなどと聞き手が（そして観察者が）わかるのは、ある程度、言語にその情報が埋め込まれているからだと考えられる。筆者はこれを発話における情報構造と呼び、情報構造が言語的にどのように伝達されているかに関心がある。本研究では、情報構造と特に語順の関連について詳しく調査する。

1.2　本研究の立場

本研究では、すでに話に出たもののことを「話題」と呼び、新たに提案されたもののことを「焦点」と呼ぶ。より詳しくは、以下のように定義する。話題と焦点は「もの」であるとは限らず「命題」の可能性もあるが、本研究では「もの」を中心に扱い、「命題」は対象外とする。

(2) **話題**とは、話し手と聴き手の間ですでに共有されていて「それに関しては異存はない」という同意が得られていると、話し手によって想定されているものを指す。

(3) **焦点**とは、聞き手にとってニュースであり「聞き手が異論を唱えるかもしれない」と話し手に想定されているものを指す。

このことは例えば、以下のようなテストでも確かめることが出来る。次の話し手により「違うよ」の後で訂正されうるのは、焦点のみであり、話題を訂正すると不自然になる (Erteschik-Shir, 2007, p. 39)。このことは「焦点だけが異論を唱えうると想定されている」ことを利用している。(4) を例にして、(4a) の「太郎」を話題、「教授」を焦点と仮定すると、次の話し手 (4b) によって訂正されうるのは、焦点の「教授」のみであり、同じ手続きによって話題「太郎」を訂正することはできない。

表 1　情報構造に関わる特徴

	話題	焦点
話者の想定	前提	断定
情報の新旧	旧	新
定性	定	不定
有生性	有生	無生
項構造	動作主	被動作主
推論可能性	推論可能	推論不可能

(4)　a. 太郎って教授なんだって

　　b. 違うよ、{??次郎/助教 } だよ　　(**作例**)

このように話題と焦点をテストによって確かめることができるが、発話の中で一義的に「これが話題/焦点である」と決められるものではなく、複数の特徴が合わさって「話題らしさ」「焦点らしさ」を構成すると考える (Givón, 1983; Du Bois, 1985)。「話題らしさ」「焦点らしさ」を構成する特徴として、Givón (1976); Keenan (1976) に基づき、表 1 を想定する。

2　先行研究

日本語の語順研究は佐伯 (1998) に詳しくまとまっている。これを見る限り、伝統的な日本語学では、助詞や助動詞の相互承接の問題、係り受けの問題が中心課題のようである。プラーグ学派の伝統に基づく日本語とチェコ語の語順を情報構造の観点から分析したフィアラ (2000) や、生成文法に基づく Nemoto (1993); 遠藤 (2014) なども存在する。これらの研究は、作例に基づくか、用例収集にとどまっている。また、書き言葉が分析の中心となっている。Matsumoto (2003) はイントネーション・ユニットの観点から大量の話し言葉のデータをもとに語順を分析しており興味深い結果を報告しているが、イントネーションが単位となっているため 1 つの述語がとる項どうしの順番に関する一般化は得られていない。Ono and Suzuki (1992); 高見 (1995a,b); Ono (2007) は、後置文を分析している。相互行為の観点から語順を分析した Tanaka (2005) もある。Yamashita and Kondo (2008) は CSJ を調査し、句の長さと情報の新旧の双方が語順に影響を与えていると結論づけている。

本研究は、先行研究の成果に反論するわけではなく、機能的な観点から多角的な変数を考慮し、実際の話し言葉に基づいて語順と情報構造の関わりをより一般的に明らかにし、その傾向の背後にある言語使用に基づく動機づけを解明することを目指す。

3　調査手順

3.1　コーパス

本研究では、日本語話し言葉コーパス (CSJ: 前川, 2006) の模擬講演 12 回分を用い、語順と情報構造の相関を調べた。模擬講演は日常的な話題（「人生で一番嬉しかったこと」や「悲しかったこと」など）に

関して一般の話者が話したものである。聞き手は 3–4 名で、1 講演 10–15 分である。なお、本研究で利用したのは CSJ の RDB 版 小磯ほか (2012) である。

3.2 手順

Givón (1983); Nakagawa and Den (2012) の方法を応用して、コーパスから 4614 の名詞（ゼロ代名詞を含む）を抜き出し、それぞれの照応関係を同定した。次に、名詞の項構造、情報の新旧を半自動で同定した。各々の定義と変数の種類は表 2 のとおりである。表 1 のうちからこれらの変数を選んだのは、アノテーションが比較的客観的に行いやすいと考えたためである。他の変数に関しては、質的分析に留めた。

語順を同定するために、CSJ の `relBunsetsu2Clause` における `nth` という情報を利用した。この場合の `nth` とは、節中における文節の生起した順番である。筆者は `nth` と項構造の情報を利用して、共通の述語の項の順番だけを取り出した。例えば `nth` は (5) に示す通りフラットに節の中の文節の順番を表している。筆者はこれと項構造を利用して、ある述語のとる項だけを取り出したときの順番を作成した。少し込み入った例だが、(5) [*1]では、述語「出る」の項は「私」と「外に」であり、前者が最初に生起しているので 1、後者は後に生起しているので 2 とした。同様に、「挙げて」の項は「両親」(1) と「もろ手」(2) であり、「賛成」の項は「(私が外に出る）こと」(1) であり、それぞれの述語に関して項の生起順によって語順を決定した。

表 2　変数とその定義

名前	変数の種類	定義
a. 語順	`1, 2, 3, 4`	名詞が節の中で占める順番
b. 情報の新旧	**旧**	先行詞を持つ
	新	先行詞を持たない
c. 項構造	S	自動詞の唯一項
	A	他動詞の動作主項
	P	他動詞の被動作主項
	LOC	場所を表す項
	Ex	動詞の項ではない名詞

(5)　**発話:** [両親も [[私が 外に 出る] ことに関しましては 非常に もろ手を 挙げて] 賛成] という
`nth:` [2 [[3 4 5] 6 7 9 10] 11]
語順: [1 [[1 2] 1 2]]
(S01F0038:　183.9-188.4)

この構造の認定には問題があるかもしれないが、(5) のような込み入った例は稀であるので、今回はこの方法で遂行し、今後手法を改良していきたい。なお、(1a) のような後置文は模擬講演にはほぼ見られなかったため、これの分析は対話コーパスと比較して別に行った（5 節)。

情報の新旧 (information status) に関しては、当該の名詞が先行詞を持っていれば**旧**、持っていなければ**新**とした。項構造に関しては以下のようにコーディングを行った。ヲで標示され得る名詞を項としてとっている述語について、ガで標示され得る名詞を A、ヲで標示され得る名詞を P とした。ヲで標示され得る名詞を項としてとっていない述語について、ガで表示され得る名詞を S とした。以上の場合に関わらず、ニで表示され得る名詞を LOC とした。また、動詞の項ではない名詞に Ex を付与した。これは「象は鼻が長い」における「象」などを対象とする。定義上「項ではない」のだが、便宜的に項構造の中に含めた。

[*1] スペースの都合上、発話のフィラーを削除した。

以上の手順を終えた後、語順を独立変数、他を従属変数として一般化線形モデルで分析した。統計ソフトは R を用いた。本研究の目的は語順に関わる変数を明らかにすることであるため、名詞が 1 つだけしか生起していない例を分析対象から除外した。

4　結果と考察

一般化線形モデルの分析結果を表 3 に示す。

表 3　一般化線形モデルの分析結果

Coefficients	Estimate	p-value	
情報の新旧（新）	0.241145	0.0006	***
項構造 Ex	−0.158079	0.2067	
項構造 LOC	0.423852	< 0.001	***
項構造 P	0.824221	< 0.001	***
項構造 S	0.441074	< 0.001	***

($0 \leq$ '***' $\leq 0.001 \leq$ '**' $\leq 0.01 \leq$ '*' ≤ 0.05 '.' ≤ 0.1)

4.1　後方に生起する名詞

新情報であるほうが、すなわち先行詞がないほうが、他の項よりも後に生起するという傾向があるとも言える。このことは 高見 (1995a) の指摘する、「焦点は述語の近くに生起する」という観察と関連し、本研究の結果は、高見の観察を裏付ける結果であると考えられる。

項構造の S（自動詞の唯一項)、P（他動詞の被動作主項)、LOC（場所を表す名詞）が、他の項よりも後に出現しやすいことも確かめられた。先程述べた、新情報（焦点）が述語の近くに生起する傾向があること、S と P は新情報（焦点）になりやすいこと (DuBois, 1987) を考え合わせると、述語の近くに S と P が生起するのは自然であると考えられる。もちろん「基本語順」として日本語では APV（V は動詞）の順で発話することが一番自然なので、P が動詞の直前にあるのは当たり前の結果である。しかし本研究はこの「基本語順」が何故そのようなのかという動機付けに興味があり、そのためにこの結果を確かめておくことは重要であると考える。動機付けに関しては 6 節で議論する。

LOC に関しては、頻出するために分析の中に入れたが、時や場所を表す場合と与格の場合で全く語順が異なると考えられる。今後の課題として、ニが後続する名詞をさらに分割した後で再分析する必要がある。

4.2　前方に生起する名詞

プラーグ学派などにおける伝統的な観察 (Mathesius, 1928) に基づき、旧情報であるほうが他の項よりも先行しやすいという結果が得られることが期待されるが、このことは本研究のコーパスからは支持されなかった。旧-新の順で生起した例は 164（61%）あったのに対し、新-旧の順で生起した例は 105 (39%）あり、「旧情報は新情報に先行する」と一般化するには、39% は多すぎる。

しかし、確かに (6) のように旧情報が新情報に先行する例は存在する。特に 「菓子パン」に注目すると、初出の (6b) では動詞の直前に生起しているが、次に菓子パン（「それ」）に言及する (6d) では「お爺ちゃん」よりも前に生起している。また、再び「菓子パン」が言及される (6f) でも「犬」よりも前に生起している。

(6)　a. 更にうちの祖父っていうのがお菓子が好きなもので
　　b. よくパン屋さんで**菓子パン**を買ってくるんですが
　　c. 買い過ぎてしまいまして
　　d. え (ん) **それ**を (い) まー 要はお爺ちゃんは一生懸命食べるんですけれども
　　e. 余って

f. **それ**を犬にあげてしまうので
g. その残飯で太り
h. 菓子パンで太り
i. 味は覚えてグルメになるという
j. 最悪の育ち方をしてしまいまして　(S02M0198: 244.48-262.82)

表現のタイプ（名詞か代名詞か）と語順の関連を見ると、図 4.2 に示すように、代名詞は前に現れることが多いことがわかる（図中の NP は名詞、Pron は代名詞を指す（Yamashita and Kondo (2008) も参照））。このように、旧情報が新情報に先行する傾向は、確かに存在すると思われる。

では、新情報が旧情報に先行している例はどのような特徴を備えているのだろうか？例を観察した結果、新情報に先行される旧情報は、指示対象の同一性が不確かであるという傾向があることがわかった。例えば (7) では、「てんかん」という名詞が何度も繰り返されるが、言及の回数にかかわらず動詞の直前に生起している。てんかんは、具体的なものを指す名詞と異なり抽象的な病気の症状を指しているので、1 度目のてんかんと 2 度目のてんかんが同一かどうか、曖昧である。

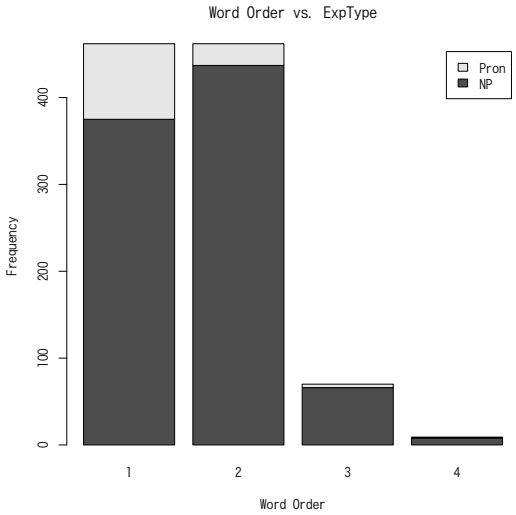


図 1　語順と表現タイプの頻度

(7) a. [飼い犬が] あと一回**てんかん**起こしたら死ぬって言ってたんですけど
b. またそそうこうしてるうちに**てんかん**起こしまして
c. ... (130.8 秒省略。このときは復活したがその後数回てんかんを繰り返し、死んでしまう。)
d. その僕が出掛ける時にもう軒下で**てんかん**起こして
e. 多分死んでいたんだろうと
f. (た) 軒下で**てんかん**起こしたが為に
g. その要は出られなくて
h. 引っ掛かっちゃって
i. 出られなくてそのまま死んじゃったんじゃないのっていう話をしたんですけど
(S02M0198: 558.7-712.8)

(8) においても同様に、(8c-f) の「水」が同一の対象を指しているかどうか、水が質量名詞であるため曖昧である。

(8) a. でこのティータイムなんですけれども
b. この標高の高いところでは高山病という非常に危険な可能性があるので

c. えー**水**が非常に重要になります
d. ですから大体一日に**ニリットルの水**を取ってくださいと言われて
e. 食事の時は必ず**マグカップで二杯分の水**を飲みますし
f. 途中途中でも必ず**水**をほあのー飲みたくなくても飲まされるという感じで
g. 水分補給を重視しておりました　(S01F0151:　339.78-341.44)

一方、(6d,f) の「それ」は明らかに祖父がある日買ってきた菓子パンと同一のものを指している。このような代名詞が他の名詞に先行する傾向と考え合わせ、本研究では、語順は定性に敏感であり、定名詞は前方に生起し不定名詞は後方に生起すると一般化しておく。

5　述語の後ろに生起する名詞

上述したように、模擬講演において述語のあとに生起する項は稀だったため、Nakagawa et al. (2008) における CSJ 対話部分の後置文の分析結果と比較した。Nakagawa et al. (2008) は、後置要素のピッチ・ピークの有無によって後置文を 2 種類に分け、それぞれの指示距離 (Referential Distance: Givón, 1983) を計測した。この場合の指示距離とは、間休止単位 (inter-pausal unit: IPU) を単位として、[*2]後置要素の名詞を含む単位と、それ以前にその指示対象を指す名詞・代名詞（ゼロ代名詞）を含む単位の間にある IPU の数である。これは指示対象の活性化コストを近似する意図で考案され、指示距離が大きいほど活性化コストが大きく、指示距離が小さいほど活性化コストも小さいと考えられる。例えば (9) は 1 行 1IPU で区切られた発話の例である。(9b) で「これ」が述語の後ろに生起している（「これ」を後置要素と呼ぶ)。「これ」の指示対象は、(9a) の「アンケート用紙」と同一であり、(9a) は (9b) よりも 1IPU 前に生起しているので、指示距離は 1 となる。

(9)　a. **L1:** えーとー 調査をするのに**アンケート用紙**を配ったっていうことなんですが
b. **L2:** どのくらいこう配ってどのくらい回収できるものなんですか**これ**は
(D04F0050:　588.8-597.0)

これと同様の手法で、CSJ 模擬講演に生起した名詞と代名詞の指示距離を、語順別に集計した。表 5 は後置要素の指示距離、表 5 は述語の前に生起した要素の指示距離の平均をまとめたものである。まず Nakagawa et al. (2008) の結果 (表 5) を確認しておくと、後置要素にピッチ・ピークがない (9b) のような場合は、指示距離が小さく、活性化コストが低い名詞が後置されていると考えられる。一方、(本研究ではこれにはあまり注目しないが）後置要素にピッチ・ピークがある場合はない場合よりも、指示距離が大きく、活性化コストが相対的に大きい名詞が後置されている。それに対して、述語の前に生起する要素の指示距離（表 5）は、ピッチ・ピークなしの後置要素の指示距離よりも全体的に大きく、活性化コストも大きいことがわかる。つまり、ある名詞の指示対象の活性化コストが小さいときは述語より後にピッチ・ピークなしで、活性化コストが大きいときは述語より前に発話される傾向があることがわかった。このことは、Givón (1983) において他の言語で指摘されていた傾向と一致する。

6　議論

結果をまとめると、述語より前に生起する名詞のうち、(1) 定情報を表す名詞はほかの名詞よりも前に、(2) 新情報、P, S など焦点と関わる名詞はそうでない名詞よりも後に発話されることがわかった。また、(3) 述語より後に生起する名詞は前に生起する名詞よりも活性化コストが低いことが明らかになった。

[*2] Givón (1983) は間休止単位ではなく節を単位としている

表 4　後置要素の指示距離

	後置要素のピッチ・ピーク	
	無	有
指示距離	6.9	39.7

(Nakagawa et al., 2008, p. 13)

表 5　述語の前要素の指示距離

	語順		
	1	2	3
指示距離	20.9	23.0	41.1

何故このような傾向があるのだろうか？この傾向を、話し手・聞き手の談話処理や発話産出という観点から説明することは出来るだろうか？以下の節では、それぞれに関して動機付けを議論する。

6.1　述語の前要素: 前方に生起する名詞

伝統的に指摘されている通り、すでに言及された要素に文頭で再び言及することは、これから言われる文の中での「いかり」の役割を果たす。それ以外にも、本研究では 2 種類の動機づけを提案したい。

まず 1 つ目の動機付けは、ゼロ代名詞の指示対象の確立である。模擬講演の発話例を多く見ていると、「X が Y なんですが、（その）X/Y（というの）は...」という例が多いことに気づく。Clancy (1980) も、日本語は同じ指示対象を指す名詞を繰り返す傾向があると指摘しているが、それが何故なのかはわからないと述べている。例えば (10) では、(10a) で「管理検査組」が導入された後、(10b) で同じ対象を指す名詞「三人」が繰り返されている。

(10)　a. で特に僕ら**検査管理組**っていうのは三人いたんですが
　　b. **その三人**は まー 非常に茂原を愛してですね
　　c. んでその後も年に一回ぐらいは えー ま ø 訪問して
　　d. で最初のうちは工場の人達も ø 訪問してたりしたんですけど
　　e. 最近は えー さすがにそういうことも ø しなくなって
　　f. あの ただ茂原を ø 通って
　　g. あ 懐かしいねってこう ø 喜んでですね
　　h. で ま ん (?ひ) 一軒ぐらいこう あの 食事をしに ø 行ったりとかですね
　　i. えー いうことで ø 楽しんだりしてます　(S05M1236: 609.5-639.4)

本研究では、これは日本語がゼロ代名詞を多用しており、明示的な代名詞とは異なり有生性や性・数の情報がないため、指示対象の確立に手間がかかるためであると提案する。その繰り返される名詞が他の名詞よりも先に生起するのは、その名詞の掛かり先が広いためであると考える。例えば、(10) では、「三人」は、(10c-i) の主語になっていることからもわかるように、掛かり先が大きい。（ゼロ代名詞をøで表し、便宜的に動詞の直前に表記している。）この「三人」の掛かり先が大きいことが、(10b) で、「茂原」よりも「三人」を先に発話する動機付けになっていると考えられる。

2 つ目の、すでに言及された要素が次の発話の文頭で繰り返される傾向の動機付けは、Den and Nakagawa (2013) で提案されている通り、発話の続きを考える時間かせぎである。(10) の例で説明すると、すでに出ている表現「三人」は比較的言いやすく、「三人」と言っている間に話し手は次の発話を考えていることが考えられる。実際、彼らの調査によれば、後に続く発話が長いほうが、名詞の繰り返しが起きやすい。[*3]

[*3] ただし Den and Nakagawa (2013) の調査は CSJ の対話を用いて行われ、話者が交替するときの名詞の繰り返しのみに注目している。

このように、定情報を表す名詞がほかの名詞よりも先に産出される、さまざまな機能的動機付けが存在する。生成文法における最近のカートグラフィー研究 (e.g., 遠藤, 2014) でも話題を最も上の階層に（そして焦点をその下の階層に）位置づけるが、話題と関連する定情報が先行するには機能的な理由があり、また下で述べるように、述語の後に産出される場合もその動機付けがある。

6.2　述語の前要素: 後方に生起する名詞

焦点に関わる特徴（新情報や S, P）が後方、典型的には述語の直前に生起するのは、述語とともに焦点という単位をなしているからだと考えられる。Lambrecht (1994) が指摘するように、述語がすでに話し手と聞き手の間で共有されていて異論をとなえる箇所でない、(11b) のような場合は、自然発話では稀にしか見られない。(「何してるの？」という質問の答えになる構造が最も一般的だと Lambrecht は述べている。)

(11)　a. **A:** [テレビを見ている B に向かって] 何見てるの？

　　b. **B:** ニュース 23（を見てるんだよ）　(**作例**)

今後、実際の発話データを見て確かめることが必要であるが、今は Lambrecht の指摘が正しいとを仮定して（筆者には直感的に正しいと思われる)、新情報、S、P などの名詞が動詞の直前に生起しやすいのは、これらの名詞が述語とともに焦点という単位を構成するからだと提案しておく。例えば、(12) において、S「感じ」は述語「する」の直前に生起している。これは「感じがする」全体で焦点という単位を構成しているからである。

(12)　そのコントラストというのは 何か とてもこうエキゾチックと言うか **不思議な感じ**がしまして

(S00F0014:　1042.9-1.47.0)

6.3　後置要素

6.1 節で議論したように、述語の前に現れる、他の名詞よりも先行する要素は、すでに言及したものにもう 1 度冒頭で言及することによって「いかり」の役割を果たしたり、以後の発話のゼロ代名詞の解釈を助けたり、話し手の時間稼ぎに使われたりする。では、すでに言及された要素に、述語の後で言及するのにはどのような動機付けがあるのだろうか？

発話は、イントネーションによりいくつかの単位に区切られており、1 つのイントネーションの単位を発話するときは、普通、高いピッチから始まり、徐々に下降していく (Liberman and Pierrehumbert, 1984; Den et al., 2010)。そして、ピッチ・ピークなしの後置要素は、活性化コストが低いため、低いピッチで言われる（例えば英語でも Halliday (1967); Bolinger (1972) などで指摘されている)。すなわち、活性化コストが低い指示対象を指す名詞は、イントネーション上の都合で述語の後ろ（発話末尾）で話し手にとって発話しやすいと言える。Siouan, Caddoan, Iroquoian などの言語における述語の後ろに現れる名詞を調べた Mithun (1995) も、同様の結論に達している。

このような後置文がなぜ独話よりも会話に多いのかに関しては、依然として謎が残る。渡辺 (1971) によれば、発話末の終助詞「ね」や「よ」は聞き手に向けられた態度を表している。後置要素は、(1a) のように後置要素は「ね」や「よ」よりもさらに後から発話されるため、もしかしたら、より聞き手への指向性が強いのかもしれない（CSJ の模擬講演では「よ」や「ね」は頻繁に生起する)。Tanaka (2005) は、会話における後置文を相互行為の観点から分析し、選好的な応答（依頼に対する承諾など）はスムーズに行うという動機付けが強いため結論部分である述語を早く産出すると述べている。一方、非選好的な

応答（依頼に対する拒否など）には遅れや非流暢性が伴うため、主語や目的語を述語の前に産出して、拒否する時間を引き伸ばす。例えば (13) は 40 代女性 3 人の会話で、服の流行が昔に戻って来ており、チカコの娘は彼女が若いときのブラウスを着ているという話をしている。これに同意したケイコ (13b) が「おんなじよ襟も」とチカコに同意する。Tanaka によればこの発言はチカコの発言 (13a) の直後に即座に行われており、チカコの発言を効果的に支持している。

(13) a. **チカコ:** 今の形とまったくおんなじ.=
b. **ケイコ:** =おんなじよ↓=[襟も.
c. **エミコ:** [あ！ほんと::. (Tanaka, 2005, p. 406)

一方、不同意などの非選好連鎖の応答部分の場合はより慎重に行われ、この場合はゼロ代名詞でも良いはずの項が述語の前に産出される。

7 おわりに

本研究は、機能的な観点から日本語話し言葉の語順と情報構造の関連を多角的に検討した。今後は Yamashita and Kondo (2008) のように長さなどを考慮に入れたり、Tanaka (2005) のような相互作用的な観点も取り入れつつ、より詳細で複合的な語順予測のモデル化を課題とする。

参考文献

Bolinger, Dwight (1972) “Accent is Predictable (If You’re a Mind Reader),” *Language*, Vol. 48, pp. 633–644.

Clancy, Patricia (1980) “Referential Choice in English and Japanese Narrative Discourse,” in Chafe, Wallace ed. *Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, Vol. 3 of Advances in Discourse Processes, New Jersey: Ablex, pp. 127-202.

Den, Yasuharu, Hanae Koiso, Takehiko Maruyama, Kikuo Maekawa, Katsuya Takanashi, Mika Enomoto, and Nao Yoshida (2010) “Two-level annotation of utterance-units in Japanese dialogs: an empirically emerged scheme,” in *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

Den, Yasuharu and Natsuko Nakagawa (2013) “Anti-Zero Pronominalization: When Japanese Speakers Overtly Express Omissible Topic Phrases,” in Eklund, Robert ed. *Proceedings of Disfluency in Spontaneous Speech (DiSS 2013)*, pp. 25–28, Stockholm.

Du Bois, John W. (1985) “Competing Motivations,” in Haiman, J. ed. *Iconicity in Syntax*, Amsterdam: John Benjamins, pp. 343-366.

DuBois, John W. (1987) “The Discourse Basis of Ergativity,” *Language*, Vol. 63, pp. 805-855.

Erteschik-Shir, Nomi (2007) *Information Structure: The Syntax-Discourse Interface*, Oxford: Oxford University Press.

フィアラ，K. (2000) 『日本語の情報構造と統語構造』，ひつじ書房，東京.

Givón, Talmy (1976) “Topic, Pronoun, and Grammatical Agreement,” in Li, Charles N. ed. *Subject and Topic*, New York: Academic Press, pp. 149-187.

Givón, Talmy ed. (1983) *Topic Continuity in Discourse*, Amsterdam/Philadelphia: John Benjamins.

Halliday, M. A. K (1967) *Intonation and Grammar in British English*, Paris: The Hague.

Keenan, Edward L. (1976) “Towards a Universal Definition of “Subject”,” in Li, Charles N. ed. *Subject and Topic*, New York: Academic Press, pp. 303-334.

小磯花絵・伝康晴・前川喜久雄 (2012) 「日本語話し言葉コーパス RDB の構築」,『第 1 回コーパス日本語学ワークショップ予稿集』, pp.393–400, 国立国語研究所, 東京. (http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_53.pdf よりダウンロード可能).

Lambrecht, Knud (1994) *Information Structure and Sentence Form: Topic, Focus and the Mental Representations of Discourse Referents*, Cambridge: Cambridge University Press.

Liberman, Mark and Janet B. Pierrehumbert (1984) “Intonational Invariance under Changes in Pitch Range and Length,” in Aronoff, Mark and Richard T. Oehrle eds. *Language, sound, structure: studies in phonology presented to Morris Halle by his teacher and students*, MA: MIT Press, p. 157233.

前川喜久雄 (2006) 「概説」,『日本語話し言葉コーパスの構築法』, 国立国語研究所, 東京, pp.1–21. (http://www.ninjal.ac.jp/corpus_center/csj/k-report-f/01.pdf よりダウンロード可能).

Mathesius, Vilém (1928) “On Linguistic Characterology with Illustrations from Modern English,” in Vachek, J ed. *A Prague School Reader in Linguistics*, IN: Indiana University Press, pp. 59–67.

Matsumoto, Kazuko (2003) *Intonation Units in Japanese Conversation: Syntactic, Informational and Functional Structures*, Amsterdam/Philadelphia: John Benjamins.

Mithun, Marianne (1995) “Morphological and Prosodic Forces Shaping Word Order,” in Downing, Pamela and Michael Noonan eds. *Word Order in Discourse*, Amsterdam/Philadelphia: John Benjamins, pp. 387–423.

Nakagawa, Natsuko and Yasuharu Den (2012) “Annotation of Anaphoric Relations and Topic Continuity in Japanese Conversation,” in *Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC’12)*, pp. 179-186, European Language Resources Association ELRA, Istanbul.

Nakagawa, Natsuko, Yoshihiko Asao, and Naonori Nagaya (2008) “Information Structure and Intonation of Right-Dislocation Sentences in Japanese,” *Kyoto University Linguistic Research*, Vol. 27, pp. 1–22.

Nemoto, Naoko (1993) “Chains and Case Positions: A Study from Scrambling in Japanese,” Ph.D. dissertation, The University of Connecticut, CT.

Ono, Tsuyoshi (2007) “An Emotively Motivated Post-Predicate Constituent Order in a ‘Strict Predicate Final’ Language: Emotion and Grammar Meet in Japanese Everyday Talk,” in Suzuki, Satoko ed. *Emotive Communication in Japanese*, Amsterdam: John Benjamins.

Ono, Tsuyoshi and Ryoko Suzuki (1992) “Word Order Variability in Japanese Conversation: Motivations and Grammaticalization,” *Text*, Vol. 12, No. 3, pp. 429-445.

佐伯哲夫（編） (1998)『要説 日本語の語順』, くろしお出版, 東京.

高見健一 (1995a)『機能的構文論による日英語比較』, くろしお出版, 東京.

高見健一 (1995b) 「日本語の後置文と情報構造」, 高見健一（編）『日英語の右方転移構文』, ひつじ書房, 東京, pp.149–165.

Tanaka, Hiroko (2005) “Grammar and the “Timing” of Social Action: Word Order and Preference Organization in Japanese,” *Language in Society*, Vol. 34, pp. 389–430.

渡辺実 (1971)『国語構文論』, 塙書房, 東京.

Yamashita, Hiroko and Tadahisa Kondo (2008) “Effects of Phrase Length and Referentiality in the Word-Order,”『電子情報通信学会技術研究報告思考と言語』, 第 108 巻, pp.125–130.

遠藤喜雄 (2014)『日本語カートグラフィー序説』, ひつじ書房, 東京.

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—

小磯 花絵，土屋 智行，渡部 涼子（国立国語研究所），横森 大輔（九州大学），
相澤 正夫（国立国語研究所），伝 康晴（千葉大学／国立国語研究所）

Interim Report on the Survey of Conversational Behavior: Towards the Design of Balanced Corpus of Conversational Japanese

Hanae Koiso, Tomoyuki Tsuchiya, Ryoko Watanabe (NINJAL),
Daisuke Yokomori (Kyushu University), Masao Aizawa (NINJAL),
Yasuharu Den (Chiba University / NINJAL)

要旨

国立国語研究所共同研究プロジェクト「均衡性を考慮した大規模日本語会話コーパス構築に向けた基盤整備」では，大規模な日本語日常会話コーパスの構築を目指し，均衡性を考慮したコーパスの設計案の策定を進めている。その準備段階として，一日の会話行動の種類と従事時間に関する調査を行っている。調査では，首都圏在住の成人 200〜250 人を対象に，任意の平日 2 日・休日 1 日（合計 3 日／人）の起床から就寝までの間に行ったそれぞれの会話について，いつ，どこで，誰と，何をしながら，どのような種類の会話を，どのくらいの長さ行ったか，などを問う調査項目に回答してもらった。本稿では，調査の中間結果について報告すると同時に，本調査に基づき日常会話コーパスをどのように設計するか，その方針について議論する。

1. はじめに

日常会話は社会生活の基盤であり，日常の話し言葉の特徴や仕組み，日常生活を円滑にするための会話コミュニケーションの有様を解明することが求められている。こうした研究を支えるものとして，実際の日常会話場面を対象とした大規模な会話コーパスの構築が求められている。また，言葉や行動様式は常に変化しているため，こうしたコーパスは，後世の人々が 21 世紀初頭の日本人の言語生活を知るための貴重な記録となる。民俗文化的価値のある日常会話を記録・保存・伝承することは，この時代に生きる我々に課された重要な課題である。

国立国語研究所共同研究プロジェクト「均衡性を考慮した大規模日本語会話コーパス構築に向けた基盤整備」（代表：小磯花絵，2014 年 7 月〜2015 年 8 月）では，21 世紀初頭の日本人の多様な会話行動を納めた日本語日常会話コーパスの構築を目指し，その基盤整備として会話コーパスの設計の策定を進めている。

我々の言語生活を正確に記述し，その本質を解明するためには，日常の言語生活の幅広いレジスターをカバーするようサンプルを選定することが求められる。『現代日本語書き言葉均衡コーパス』では，書き言葉の生産，流通，受容の各過程が書き言葉の実態を捉える上で重要とした上で，出版データと図書館収蔵図書を母集団としたランダムサンプリングを行い，生産実態と流通実態を反映したサブコーパスを設計した (Maekawa et al. 2014)。一方，日常会話コーパスの対象として予定しているのは，家庭や地域，職場，学校などで我々が直接交わす会話で

あり，いわゆるメディアを通じて受信する第三者による会話は対象としない。そこで本プロジェクトでは，我々が日常的に直接交わす会話の生産実態をとらえてコーパス設計に活かすために，現代日本人が日常，どのような種類の会話をどの程度行っているかを調査することとした。ただし，日常会話の生産実態（構成比）をそのままコーパス設計に反映させるのではなく，まずは日常会話にどのようなレジスター的多様性があるかをとらえ，多様な日常会話を網羅したコーパスを設計することを目指す。

本稿では，現在進行中の調査の概要と中間結果について報告した上で，本調査に基づき日常会話コーパスをどのように設計するか，その方針について議論する。

2. 会話行動調査

2.1 調査項目の設計

会話行動を調査するにあたり，我々が普段行っている会話をどのような視点でとらえるかが問題となる。本研究では会話行動を大きく，(1) 調査協力者（以下，協力者）の属性（性別や年代，職業など），(2) 会話の属性（会話の形式や会話の長さなど），(3) 会話状況の属性（会話の行われた時間帯や場所，活動など）の三つの軸でとらえることとした。この方針に従い表 1 から表 3 に示す調査項目を設計した。以下ではこのうち**形式**，**場所**，**活動**について簡単に補足する。

■形式　「雑談」は会話の目的や話題などがあらかじめ定められていない会話を，「用談・相談」は会話の目的はある程度決まっているが時間や場所などは定められていない会話を，「会議・会合」は「用談・相談」とは異なり時間や場所などが定められている会話を，「授業・レッスン・講演」は先生や講演者など会話の流れを導く人物がいる場での会話を指す。この設定は国立国語研究所 (1971, 1987) および畠 (1983) で述べられている話し言葉の分類を参考にしている。国立国語研究所 (1971) では，コミュニケーション上の機能にもとづき，談話を「ひとり」「あいさつ」「しらせ・用談」「おしゃべり」「あそび」「教え・さしず」「けんか」「思考」に分類している。また畠 (1983) は，計画性の程度にもとづき，言語行動の場面を「拘束場面」と「自由場面」に分類している。以上を参考に本調査では，会話の目的をもたない「おしゃべり」「あそび」を「雑談」に，目的をもつ「しらせ・用談」を拘束性の低い「用談・相談」と拘束性の高い「会議・会合」に分けた。また「教え・さしず」は「授業・レッスン・講演」とした。

■場所　「公共商業施設」には，公的に物品やサービスなどをやりとりする場として公共施設（市役所や銀行など）と商業施設（遊園地や店舗など）を含めた。「それ以外の屋内」には取引先の会社や知人，親戚の家などが，「それ以外の屋外」には公園や遊歩道などが該当する。「職場・学校」は協力者の職業から特定できるため一つにまとめた。この分類は国立国語研究所 (1980) で述べられている「会話場面」を参考に一部変更して設計した。

■活動　日本放送協会 (2010) による国民生活時間調査の行動分類（中分類）にもとづき設計した。このうち会話行動を伴わない「睡眠」や他活動と共起して現れる「マスメディア接触」は除外した。また「通勤」「通学」は家と店舗の往復などその他の移動と合わせて「移動」に，「仕事関連」「学業」は「仕事・学業」にまとめた。ただし，仕事のつきあいや部活動など仕事や学業から派生する副次的な活動は「業務外・課外活動」として新たな選択肢を設けた。

その他の選択肢として，「家事・雑事」は掃除や買物，子どもの世話などが，「身の回りの用事」は入浴や着替え，散髪などが，「療養」は通院や入院などが，「社会参加」は冠婚葬祭や町内会の行事などが，「レジャー活動」は趣味・娯楽・行楽・スポーツ・習いごとなどが，「付き

表 1　協力者の属性に関する調査項目

項目	説明	回答方式	選択肢
性別	協力者の性別	単一選択式	男性，女性
年代	協力者の年代	単一選択式	20 代，30 代，40 代，50 代，60 代以上
職業	協力者の職業	単一選択式	会社員・役員・公務員・専門職（以下，会社員等） 自営業，パート・アルバイト，学生，専業主婦， 無職・定年退職者（以下，無職等），その他
世帯員数	協力者の世帯員数	数値入力式	
居住地	協力者の居住地	単一選択式	東京都，神奈川県，千葉県，埼玉県

表 2　会話の属性に関する調査項目

項目	説明	回答方式	選択肢
形式	会話のタイプ	単一選択式	雑談，用談・相談，会議・会合， 授業・レッスン・講演
長さ	会話の長さ	単一選択式	5 分未満，5〜15 分，15〜30 分， 30 分〜1 時間，1〜2 時間，2〜5 時間， 5〜10 時間，10 時間以上
相手人数	会話相手との関係	選択式 (複数可)	家族，親戚，先生生徒，仕事学業関係， 友人知人，公共商業関係， 顔見知り・見知らぬ人
	関係ごとの人数	数値入力式	
相手属性	外国人を含む会話	オプション式	（該当する場合に選択）
モード	電話・ネットでの音声・映像会話	オプション式	（該当する場合に選択）
言語	外国語での・外国語を含む会話	オプション式	（該当する場合に選択）

表 3　会話状況の属性に関する調査項目

項目	説明	回答方式	選択肢
時間帯	会話が行われた時間帯	単一選択式	午前，午後，夜
場所	会話をした場所	単一選択式	自宅，職場・学校，公共商業施設，交通機関， それ以外の屋内，それ以外の屋外
活動	会話中にしていた活動	単一選択式	食事，家事・雑事，身の回りの用事，療養， 仕事・学業，業務外・課外活動，社会参加， レジャー活動，付き合い，移動，休息

合い」は知人との電話でのおしゃべりや同窓会など人と会うこと・話すことを主な目的とする活動が，「休息」は自宅での一家団らんや職場での休憩などが該当する。

2.2　調査の方法

言語生活の記録という意味では地理的多様性を無視することはできない。しかし，時間的・予算的な制約もあることから，第一段階として首都圏にしぼってコーパスを設計・構築することとした。このようにコーパスの対象を首都圏に限定したため，調査対象も首都圏在住者とした。また会話行動の**実態**の解明を目的とする場合，仮に丸一日会話をしない日があっても，それが生じる以上，調査の対象とすべきである。しかし本調査は，会話行動の**多様性**をとらえることを主目的としており，これを限られた期間と予算で達成するために，あまり会話しないと予想される日は調査日としないよう依頼した。その意味において，会話行動の正確な実態調査にはなっていない点に注意する必要がある。調査の概要を以下に示す。

■目的　日常会話の多様性を明らかにし，それに立脚して多様な日常会話を網羅したコーパスを設計するために，首都圏在住者を対象に 1 日の会話行動の種類や時間などを調査する。

■期間　2014 年 11 月 1 日〜2015 年 2 月末（予定）

■調査日・時間　任意の平日 2 日・休日 1 日（計 3 日／ 1 人）の起床から就寝まで
■対象　首都圏（東京・神奈川・千葉・埼玉）在住の 20 歳以上の日本語母語話者 200〜250 人 (20 代・30 代・40 代・50 代・60 代以上×男・女× 20〜25 人)。協力者はホームページおよび知人などからの紹介により募集した。
■調査項目　協力者の属性（表 1 の 5 項目），会話・会話状況に関する調査項目（表 2 の 6 項目，表 3 の 3 項目），および参考のために会話の概要（自由記述）。
■手続き　(1) 協力者に調査の手引きと調査票（1 日 1 冊，計 3 冊）など資料一式を事前に郵送した。(2) 資料が届いてから 2 週間以内を目途に，協力者本人が任意の平日 2 日・休日 1 日（計 3 日）を選択して調査を実施した。あまり会話しないと予想される日は避けるよう依頼した。(3) 調査日当日，協力者は調査票を携帯し，起床してから就寝までの間に行った全ての会話について，会話の概要を記した上で，会話と会話状況に関する調査項目（表 2・表 3）に回答した。できるだけ一まとまりの会話が終了するごとに記録するよう依頼した。(4) 調査終了後，協力者は調査票と協力者の属性（表 1 の 5 項目）を記したシートを調査者に返送した。
■謝礼　3 日間の調査に対し 6000 円。

3. 調査結果の分析

3.1 方法

2015 年 1 月 28 日現在の有効回答 219 名分（計 657 日，8296 会話）を対象に分析を行う。219 名の内訳を表 4・5 に示す。なお，1 日の平均会話数は 12.6（平日:13.2，休日:11.4），1 日の平均会話時間は 6.1 時間（平日:6.0 時間，休日:6.5 時間）である[*1]。また世帯員数は，1 人（一人暮らし）が 32 名，2 人が 62 名，3 人が 61 名，4 人が 50 名，5 人以上が 14 名である。

表 4　調査対象：**性別・年代**の内訳

	20 代	30 代	40 代	50 代	60 代以上
女性	23	24	25	24	25
男性	15	20	17	21	25

表 5　調査対象：**職業**の内訳

会社員・役員・公務員・専門職	87	自営業	10
パート・アルバイト	32	学生	28
無職・定年退職者	18	その他	12
専業主婦	32		

本研究では，表 1〜表 3 のうち，**世帯員数**，**居住地**，**相手属性**，**モード**，**言語**を除く 9 項目を分析に用いた。**相手人数**については全ての関係性の人数の合計値を用いた。

分析に際し次の方法で項目の値を一部併合した。まず 9 項目の対応関係を多重対応分析によって分析し，各値に与えられた重み係数（3 次元解）をもとにしたクラスター分析（ユークリッド距離・Ward 法) を行った。クラスター分析の結果から大きく 5 つのクラスターに分類できることが分かった（図 1)。この結果にもとづき，同じクラスターに属する値のうち類似した値を併合することとした。ただし単独で頻度の高い値は併合しない方針とした。例えば**活動**では，「休息」「食事」「付き合い」「レジャー活動」が同じクラスターに属するが，このうちいわゆる積極的レジャーに分類される「付き合い」と「レジャー活動」のみを併合し，単独で高頻度の「食事」や消極的レジャーの「休息」は併合しなかった。また**職業**の「その他」はいずれも有職者であったため「パート・アルバイト」と併合した。この方針に従い次の通り併合した。

[*1] 会話時間は，**長さ**の平均値（例：「1〜2 時間」であれば 1.5 時間）から算出した。調査依頼時にあまり会話しないと予想される日は避けるよう依頼したため，実態よりも 1 日の会話数や会話時間は多いと予想される。

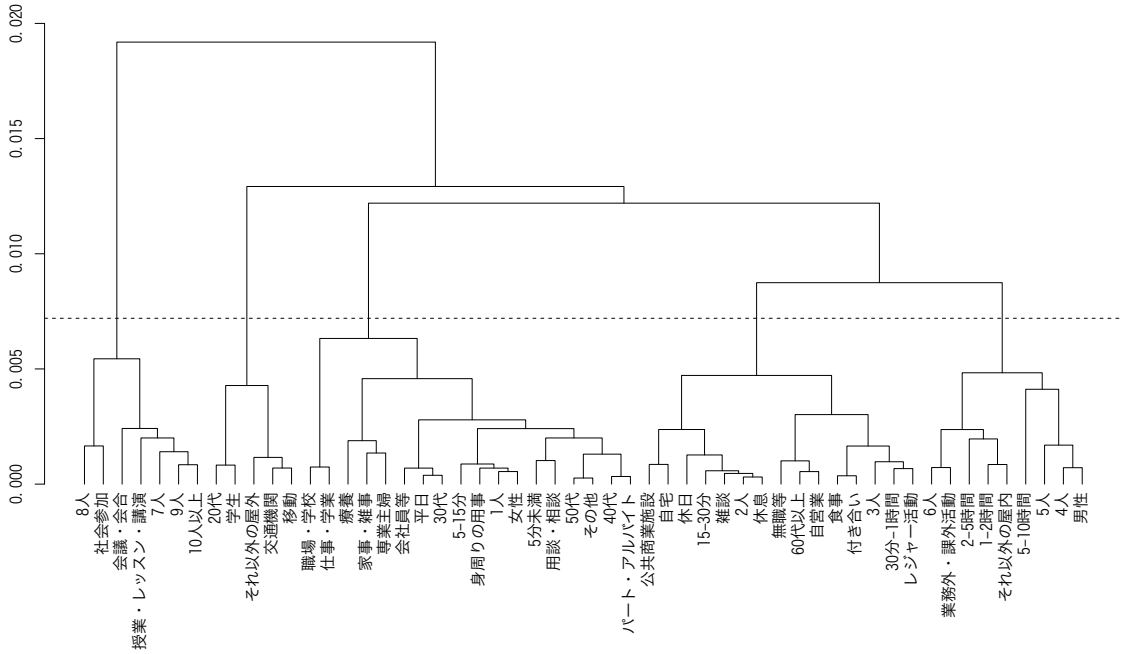


図 1　調査項目のクラスター分析の結果

職業　会社員等，自営業，パート等（パート・アルバイト + その他），学生，専業主婦，無職等
形式　雑談，用談・相談，会議・授業等（会議・会合 + 授業・レッスン・講演）
長さ　5 分未満，5〜15 分，15 分〜1 時間（15〜30 分 +30 分〜1 時間），1 時間以上（1〜2 時間 +2〜5 時間 +5〜10 時間 +10 時間以上）
相手人数　1 人，2 人，3 人，4〜6 人，7 人以上
場所　自宅，職場・学校，公共商業施設，それ以外の屋内，屋外・交通機関（それ以外の屋外 + 交通機関）
活動　食事，家事・雑事等（家事・雑事 + 身の回りの用事 + 療養），仕事・学業，社会参加等（社会参加 + 業務外・課外活動），レジャー活動等（レジャー活動 + 付き合い），移動，休息

併合した項目を含め，協力者の属性 3 項目（性別，年代，職業），会話の属性 3 項目（形式，長さ，相手人数），会話状況の属性 3 項目（時間帯，場所，活動）を分析に用いた。

3.2　結果

会話の属性 3 項目の出現傾向を職業別に見てみよう（図 2 上段）。**形式**については，いずれの職業も雑談が全体の 5 割以上（55.6〜71.1%）を占めているのに対し，会議・授業等（会合やレッスン，講演など含む）は，有職者や学生が若干多い傾向を示すものの，2.7〜7.6% と出現率は高くない。用談・相談は 24.4〜36.9% であり，どの職業でも一定数生じていることが分かる。**長さ**は，自営業を除く全ての職業で 15 分未満のごく短い会話が全体の半数以上を占めており，1 時間以上の長い会話は 10〜15% 程度に留まっていることが分かる[*2]。**相手人数**は，1 人の場合が全体の 5 割以上（54.3〜65.4%），3 人以下の場合が全体の 83.1〜91.1% を占めており，日常会話の大半が少人数の会話であることが分かる。

[*2] 自営業の場合，短時間の接客などが多いことも予想されるが，例えば途切れなく接客する場合などで調査の記録が間にあわない場合には，全体まとめて接客とし合計の時間と人数を報告しても良いとしたため，正確な値になっていない可能性がある。

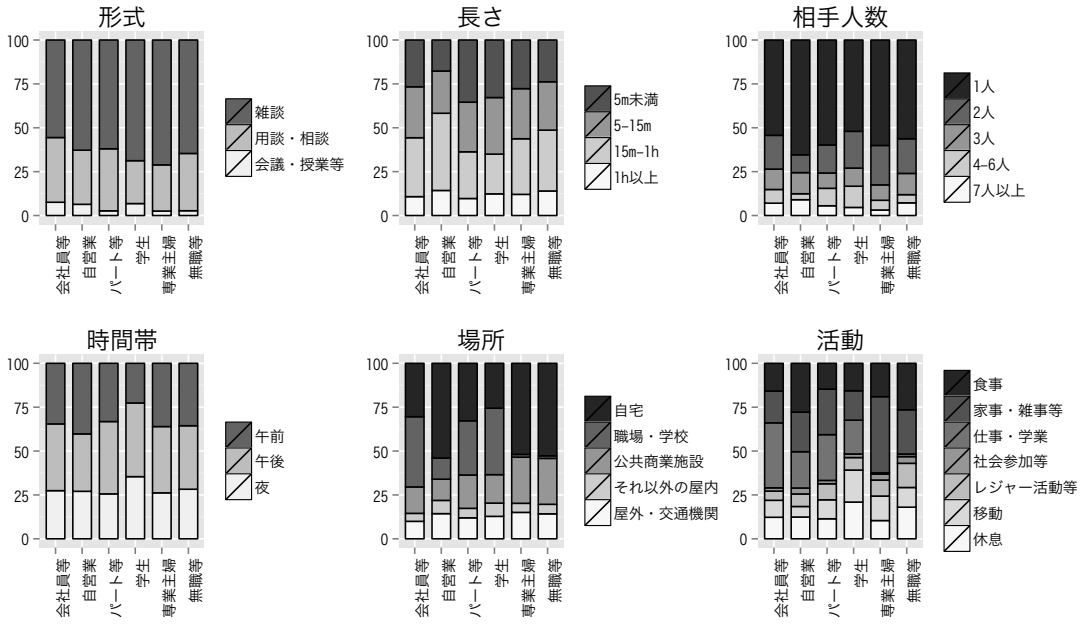


図 2　会話の属性・会話状況の属性：職業別にみた出現率（%）

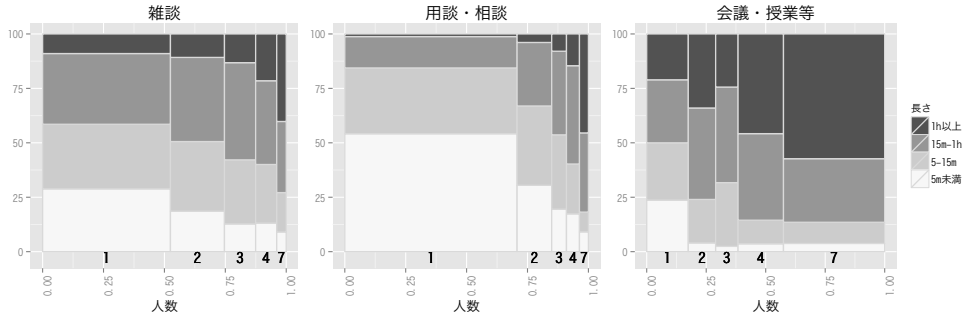


図 3　会話の属性 3 項目間の関係（横軸ラベル　**1**：1 人，**2**：2 人，**3**：3 人，**4**：4-6 人，**7**：7 人以上）

このように会話の 3 属性については，総じてどの職種においても，少人数，短時間の雑談や用談・相談が大半を占める傾向にあると言える。結果は省略するが，**性別**ごと，**年代**ごとに見ても同じ傾向がうかがえる。しかし，少ないながらも会議・授業等や 4 名以上の比較的人数の多い会話，1 時間以上の長い会話も確実に存在する。こうした頻度の少ないケースがどのような状況で生じているかを見るために，会話の 3 属性間の関係を見てみよう。

図 3 に**相手人数**と**長さ**の出現傾向を**形式**ごとに示す。スロットの幅と高さは**相手人数**と**長さ**のそれぞれの出現率を，面積は**相手人数**×**長さ**のカテゴリーの出現率を表している。図から，用談・相談では相対的に少人数・短時間の会話が頻出するのに対し，会議・授業等では多人数・長時間の会話が多く見られる。雑談はその中間の傾向を示す。またいずれの形式においても，相手の人数が増えるほど長い会話が増加する傾向が見られる。例えば，高頻度の典型的な事例として相手 1 人の 5 分未満の用談・相談を協力者の 3 属性ごとに見てみると，いずれの場合も全体の 10% 前後を占めており，相手 1 人の 5 分未満の用談・相談がどの属性の人に

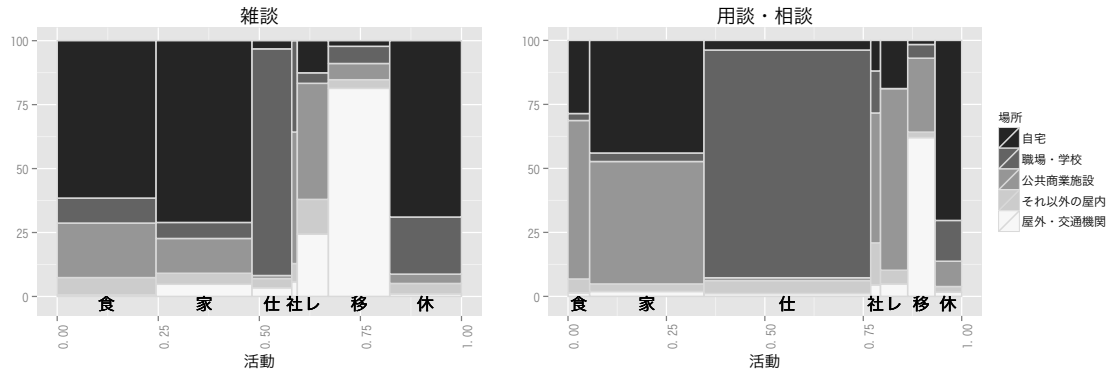


図 4　**場所**と**活動**の出現傾向（横軸ラベル　**食**:食事，**家**:家事・雑事等, **仕**:仕事・学業，**社**:社会参加等，**レ**:レジャー活動等，**移**:移動，**休**:休息）

とっても典型的な事例であることがうかがえる。

次に会話状況の属性 3 項目の出現傾向を見てみると（図 2 下段），**場所**と**活動**については，例えば自宅での会話や家事・雑事中の会話が専業主婦に，職場・学校での仕事・学業中の会話が会社員等に多いなど，職業による違いがかなり見られる。また**場所**と**活動**の関係を会話の**形式**ごとに見ると（図 4，スペースの都合で出現率の低い会議・会合等は省く），雑談は，自宅での食事や家事・雑事，休息中，職場・学校での仕事・学業中，屋外・交通機関（特に交通機関）での移動中に，用談・相談は，職場・学校での仕事・学業中や自宅・公共商業施設での家事・雑事中に多く見られるなど，**場所**と**活動**の対応関係が見られる。また，図は省略するが，例えば自宅での家事・雑事中の雑談が専業主婦に，職場・学校での仕事・学業中の用談・雑談が会社員等に多いなど，**場所・活動**と協力者の属性との関係もうかがえる。

4. コーパス設計に向けて

British National Corpus (BNC) の話し言葉のパートは，年齢・性別・社会クラス・地域に偏りがないよう選ばれた 124 人のインフォーマントが 7 日間にわたって自身で収録した日常会話を対象とするデータ群と，放送や講義など公的かつ重要な（受け手の多い）場での話し言葉を教育・教養，ビジネス，団体，レジャーの四つの領域に分けて収集したデータ群から構成される (Crowdy 1995, Burnard and Aston 1998)。後者には，授業での教師と生徒のやりとりやビジネス場面での会合など，本調査で授業・レッスン・講演や会議・会合（分析では会議・授業等）に分類されるものが含まれている。

我々も BNC と同様，インフォーマント自身に日常会話を収録してもらう方法を一つの柱としつつ，会議や授業など，この方法では収録が難しいであろう場面を個別に収録する方法も並行して行うことを計画している。前者は雑談と用談・相談が，後者は会議・授業等が中心となる。よって，まずは会話の**形式**でコーパス全体をいかに配分するかを決めることになるだろう。調査はまだ完了していないが，今回の中間結果を見ると，雑談と用談・相談は，いずれの協力者属性で見ても，前者が 6 割前後，後者が 3 割前後の出現率となっている。また会議・授業等の会話の出現率は**職業**によって偏るが，有職者や学生に限定すると 1 割弱である（図 2）。こうした割合がコーパス設計の一つの指標となりうる。ただし，これは会話の回数で見た場合であり，会議・授業等は長く用談・相談は短い傾向にあることから（図 3），長さで見ると比率

は異なる。回数と長さの両面から調査して比率を決める必要がある。

次に問題となるのが個々の**形式**をいかに分けるかということである。雑談や用談・相談と比べてより単純な構成の会議・授業等から見てみよう。図 4 ではスペースの都合で省略したが，会議・授業等では，当然のことながら有職者や学生による職場・学校での会議・授業がその大半を占めている。しかしそれに加えて，ボランティアや地域活動など社会参加中の会合や趣味・教養の教室でのレッスンなども少なからず見られ，また協力者の属性も異なることから，会話の多様性を確保する上で重要なレジスターであることが分かる。

このように**活動**と**場所**は，協力者の属性と連動する傾向にあり，結果として会話のレジスターに強く影響することが予想される。この傾向が雑談や用談・相談にも見られることは前節で言及した通りである。もう少し詳しい分析が必要だが，**形式**内の配分は，まずは**活動**か**場所**のいずれかないしは両方を参考に決めることになろう。

また，雑談や用談・相談の大半はインフォーマント（協力者）自身が収録する方法で集めることを検討しているため，多様なレジスターの会話を確保するには，インフォーマントの属性をいかに設定するかが重要となる。**性別**と**世代**のバランスは必須として，上で議論したように，**職業**は特に**場所**や**活動**に大きく影響することから，**職業**も含めてインフォーマントの配分を設定することが求められる。しかし現時点では，インフォーマントは 50 人程度を見込んでおり，性別× 5 世代ごとに 4〜5 人となると，職業をどのように組込むかが課題となる。

今回は触れなかったが，電話・ネットでの会話も全体の 1 割ほどを占めており，コーパスの設計に組み込む必要がある。また倫理的・技術的な観点からの収録の可否も加味しなければならない。小磯ほか (2015) では，各調査項目間の関連をアソシエーション分析によって抽出し，例えば 1 時間以上の長い会話には授業や会議・会合以外にも食事中や女子学生の雑談が見られるなど，複数項目間にわたって見られる傾向も明らかにしている。今後，さまざまな観点からの分析や検討を重ね，コーパス設計の策定を進めたい。

参考文献

Burnard, Lou, and Guy Aston (1998). *The BNC Handbook*. Edinburgh: Edinburgh University Press.
(北村裕 (監訳) (2004).『The BNC Handbook: コーパス言語学への誘い』　松柏社,).

Crowdy, S. (1995). “The BNC spoken corpus.” G. Leech, G. Myers, and J. Thomas (Eds.), *Spoken English on computer: Transcription, mark-up and application*. Harlow: Longman. pp. 224–235.

畠弘巳 (1983).「場面とことば」　『国語学』, 133, pp. 55–68.

小磯花絵・伝康晴・土屋智行・渡部涼子・横森大輔・相澤正夫 (2015).「一日の会話行動に関する調査とその準備的分析—均衡会話コーパス設計に向けて—」　『言語処理学会第 21 回年次大会発表論文集』.

国立国語研究所 (1971).『待遇表現の実態—松江 24 時間調査資料から—』国立国語研究所報告 41: 秀英出版.

国立国語研究所 (1980).『日本人の知識階層における話しことばの実態：「場面について」分析資料』: 国立国語研究所日本語教育センター.

国立国語研究所 (1987).『談話行動の諸相: 座談資料の分析』国立国語研究所報告 92: 三省堂.

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*, 48:2, pp. 345–371.

日本放送協会 (2010).『2010 年国民生活時間調査報告書』: NHK 放送文化研究所.

ポスター発表　グループA

3月10日（火）　15:00～16:00

象は鼻が長いか
—テキストから取得される対象物情報—

加藤　祥（国立国語研究所コーパス開発センター）†

Does an Elephant Have a Long Nose?
Features of Entities Acquired from Texts

Sachi Kato (National Institute for Japanese Language and Linguistics)

要旨

本稿は，対象物に関する情報について，コーパスから取得可能な内容・頻度と，対象物の説明文に見られる内容・頻度・順序を調査し，テキストから取得される情報の特性について考察を行う。特徴的な身体部位を有すると考えられる象をとりあげ，その調査結果を報告する。まず，現代日本語書き言葉均衡コーパス（BCCWJ）の象の用例から，取得可能な情報を調査した。また，「対象物をまったく知らない人に説明する」条件教示によりクラウドソーシング実験を行い，一般的な作文テキストを収集した。これらのテキストを分析した結果，象が大きいことと象の鼻が長いことは高頻度かつ早い順序で言及されやすいが，象の鼻の長さがどの程度かは言及されにくいとわかった。対象物認識に重要視される外観的特徴情報は，身体部位が「長い」「大きい」などの形容表現に前提的文化的知識が期待されやすく，既存のテキストのみからでは対象物のイメージが獲得しにくいといえる。

１．はじめに

テキスト情報からのみで対象物を認識するのは困難な傾向がある[1]。すなわち，我々が日常的にテキストから知識を獲得する例は多いが，正しくテキスト内容を認識できているとは限らない。知識のない読み手に対してどのような記述をすれば情報が適確に伝わるかという問題がある。

本稿は，対象物を説明するにあたり，特徴と考えられる情報がどのように言語化（記述）されるものか調査する。まず，用例としてコーパスから取得可能な特徴情報（内容・頻度）を調査することで，言及されやすい情報を整理する。次に，対象物を説明する作文を被験者実験によって収集し，対象物を効果的効率的に説明するためには，どのような情報をどのような順序で記述する傾向があるのか分析する。具体的には，象を対象とした調査を行い，象に関する記述から取得できる象についての特徴的な情報は何であるのか，また，象の鼻が長い，耳が大きいというような特徴的な情報がどのように取得できるか，あるいは取得しにくい情報は何であるか考える。

２．関連研究と本研究

国語辞書における意味は，対象物を説明するにあたって様々な内容が記述されたものと考えられる。しかし同時に，国語辞書の記述は必ずしも十分なものではないと指摘されて

† yasuda-s@ninjal.ac.jp

[1] 加藤（近刊）では，対象物についての各種テキスト（辞書語釈，被験者によって求められた情報，コーパスから取得した用例）を用いた対象物（知識率の高い動物）の同定実験を行っている。この実験結果では，いずれのテキストでも平均的に半数程度の正答率に留まっており，テキストのみから対象物を認識することの，ある種の困難さを示している。

きた（Fillmore & Atkins , 1994 など）。では，どのような記述が理想的な対象物の説明となるのか。

たとえば，國廣（1997）は「辞書の意味記述」に求める項目を示した。一般的な国語辞書の記述に現れにくいものとして，「語義的位置（語彙体系の中の位置）」「語義の対義的定義（対義語を示す）」「現象素[2]（認められる場合には図示）」「用例[3]（広く実例を観察した上で適当にまとめる）」「連想（動物名であればその動物の習性や故事来歴など（百科的知識））」が挙げられている。但し，これらの項目は国語辞書の意味記述の場合に限るため，辞書のほかのテキストからも同様に得にくい情報とは言い難いであろう。

また，辞書的意味とは異なる百科事典的知識（folk-knowledge; Wierzbicka 1996）として Natural Semantic Meta language (NSM) theory (e.g., Goddard and Wierzbicka, 2014) による記述がある。Wierzbicka (1985) の dog の例では，dog が認識可能な形や形態的な特徴を持たないため，必要十分な特性ではなく特徴的な特性のリストによって概念が定義されるとする。この際，dog の認識可能な特徴は振る舞い（特に吠える・唸る・尾を振る）であり，dog は「人とともに生き，献身的で従順，信頼し得る仲間，よき学習者，勤勉な労働者である」というような，人との関係において概念化される。しかし，人との関係が一般的に薄い動物であれば，この種の情報が記述として得にくい可能性もある。

そのほか，コーパスを用いた辞書の語釈の記述として，Sinclair が編集主幹を務めた学習者用辞書の COBUILD (1987～)では，語の意味は顕著だと見なされた最小限の細目（Sinclair 1992）とされ，コーパスに近い例文を掲載する試みが為されている（COBUILD 2009, p. xi）。

以上のような対象物に関する記述において，ある対象物を説明するにあたり特徴的な情報が適確に記述されているのかという検証は行われにくい。

加藤（近刊）は，対象物の認識に有用な情報はどのようなものかという観点で，辞書語釈やコーパスなどのテキストを用い，テキスト内の対象物認識に有用な情報を被験者実験によって調査した。この調査において対象物の認識に必要とされた記述は，主に読み手の経験や知識を喚起する情報と，提示された情報によって設定されるカテゴリに属する他メンバーとの差異に関する情報であった。記述されている情報は，予め読み手の保有している知識と合致した場合には有用な情報となる。また反対に対象物に関する知識が読み手に不足している場合には，対象物の認識に親カテゴリのプロトタイプとの差異の記述が有用であり，あるいは誤認を避けるために他メンバーとの差別化の可能な記述が有用であった。

しかし，コーパスの利用などによりテキストから取得できる情報には，その内容に限らず，頻度や記述順序という情報もある。対象物について説明するにあたり，何が特徴的な情報としてどのように記述されるかという問題が残っている。そこで本稿は，まず既存の説明文として国語辞書 10 種類の語釈を収集し，次にコーパスから対象物の用例を取得して対象物に関する情報がそれぞれどのような頻度で得られるのかを調べるとともに，同一の対象物に関する 100 以上の説明文章を作文実験によって収集し，情報内容の出現頻度と記述順序を調査することとした。

---

[2] 國廣（1994）は，現象素を「人間の認知作用を通して、ひとまとまりをなすものとして把握された現象」と呼ぶ。

[3] 「適切な用例が見付かるとは言い難いという問題がある」と指摘する。

３．調査

対象物を説明する際，辞書の語釈であれば外観に関する情報が記述されやすい[4]。そこで，Google日本語n-gramにおける動物の身体部位の用例頻度を調査したところ[5]，象（異表記を含む）については「背[6]36%（固有名詞を含む）」「鼻 21%」「耳 10%」と割合の高い部位が上位 3 種ある（図 1）という結果が得られた。象は外観的に特徴的な属性を有しているため，特徴が記述されやすいと考えられる。以上により，本稿の調査の対象として象を用いる。

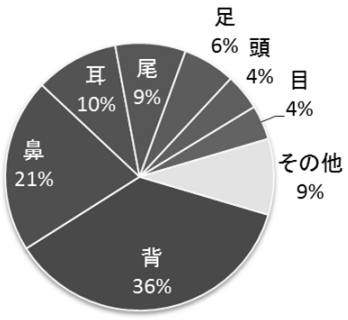


図 1　Google 日本語 n-gram における象の身体部位用例分布

調査データとして，国語辞書（3.1），コーパス（3.2），作文実験（3.3）を用いる。以下の節にそれぞれの調査結果を示す。

３．１　国語辞書

象の説明例として，まず国語辞書の語釈から得られる情報をみておきたい。

国語辞書 10 種類（表 1）の語釈における「象」項目の記述内容とその提示順序を調査した。平均 66 文字（min：14 文字，max：136 文字）を得た。

表 1　データを取得した国語辞書

辞書	三省堂国語	新明解国語	岩波国語	明鏡国語	新選国語	集英社国語	角川国語	新潮現代	大辞林	ﾃﾞｲﾘｰ国語
出版社	三省堂	三省堂	岩波書店	大修館書店	小学館	集英社	角川書店	新潮社	三省堂	三省堂
版	5 版	6 版	5 版	初版	7 版	2 版	新版	2 版	Web 版	3 版
項目数	76,000	75,000	62,000	70,000	83,000	92,000	75,000	79,000	260,000	70,000
字数(象)	65 文字	39 文字	66 文字	108 文字	80 文字	54 文字	52 文字	45 文字	136 文字	14 文字

[4] 加藤（近刊）では，国語辞書 10 種類から動物 200 種類の語釈を収集し，どのような種類の記述があるかまとめている。以下の表から，形態情報（外観に関する情報）が 9 割近くの動物で記述されており，形態情報の記述される割合が高いとわかる。語釈文においては形態情報が重要視されると考えられる。

補表　国語辞書における動物語釈の分類別記述（加藤 近刊による）

	分類	形態	生態	人間との関係	その他
当該分類の記述がある割合（200 種類中）	96.0%	87.5%	82.0%	52.5%	44.5%
各語釈における当該分類の記述割合（平均）	25.6%	36.7%	24.4%	23.3%	15.8%

[5] 身体部位の用例頻度は外観的な情報と均衡しないが，特徴的な身体部位は言及されやすい傾向がある（加藤ほか 近刊）。

[6] Google 日本語 n-gram では，「象（異表記を含む）の背」用例の 26%が「象の背に乗っ」であった。後述する 3.2 の表 3 でも「（背に）乗る」が全用例（3%）である。背が身体部位として特徴的とは言い難い。

記述内容とその提示順序を表 2 に示す。平均 5.9 種類の内容（min：2，max：9）が得られた。提示順序は内容毎に出現順を数えている。

まず，内容について，大型であることは 10 種全ての辞書で記述されていた。鼻が長いことについても 10 種全てに記述があったが，「長い」という形容詞の他に「ものをつかめる」「自由に動かせる」のような鼻についての記述があった辞書は 4 種類にとどまったため，表 2 では詳細の有無で別内容として示してある。

表 2　国語辞書における「象」項目の記述内容数とその順序（上位）

内容	記述有辞書数	1 番目	2 番目	3 番目	4 番目	5 番目
大型であること	10	3	5	1	0	0
象牙に関して	7	0	0	1	2	3
哺乳類	6	5	1	0	0	0
鼻が「長い」（詳細なし）	6	0	2	0	3	1
種類の別があるなど	5	0	0	1	1	0
生息地	4	2	0	1	0	0

次に，情報の提示順序をみると，まず 1 番目に，哺乳類であること（5 種類），大型であること（3 種類）と「アジアアフリカに」生息すること（2 種類）が記述されていた。2 番目には，大型であること（5 種類），鼻が長いこと（2 種類）が見られる。大型であることは 1〜3 番目で 9 種類，「鼻が長い」に関しては 2〜5 番目までで 10 種類と，前半に記述されやすい傾向があった。国語辞書においては，大型であることと鼻の長いことが，内容としても順序としても特徴的であると読み取れる。

３．２　コーパス

現代日本語書き言葉均衡コーパス（BCCWJ）より取得した象の用例から得られる象に関する情報を分類し，コーパスからどのような情報が取得できるのか調査した。用例の収集には中納言[7]を用い，語彙素「象」について前後 300 文字の文脈を取得した。

語彙素「象」の検索を行うと，1,323 件がヒットする。このうち，動物の「象」についての用例は 1,050 件（サンプル数では 349 件）と判断された。これらの用例の整理を，作業者の判断によって行った。同内容と考えられる例（(1)(2)のような例）を意味内容によってまとめた（(1)(2)をまとめて(3)とした例）。以下に挙げる例の下線は著者による。

(1)　しかし、与えると命がのびるので動物園の人たちは悲しみやつらさをじいっと耐え、心を鬼にして食べるものを与えなかったのです。やがて、象は何十日も食べ物を口にできず、とうとう飢えて死んでいったのでした。

（LBg9_00083：石森史郎『Once upon a time in…』[8]）

[7] 中納言 1.1.0（https://chunagon.ninjal.ac.jp/）短単位データ 1.0，長単位データ 1.0 を使用した。

[8] 用例の出典は，（サンプル ID：著者名『タイトル』（またはサブコーパス名））と記す。

(2) 私も「かわいそうなゾウ」
戦争中動物園をつぶさなくてはいけなくて動物達を毒殺したそうです。でもゾウは死ななくてしかたがなく餓死させたそうです。　（OC12_03193：Yahoo!知恵袋）

(3) 戦時中，上野動物園で餓死させられた。（意味的な用例として(1)(2)などをまとめた例）

以上のような作業により，1,314 種類の意味的な用例が取得できた。この作業にあたっては，上記(3)のように数件の用例を 1 種類にまとめた場合や，1 件の用例から 2 種類以上の意味的用例が取得される場合がある。なお，コーパスから取得した用例は，基本的に象を説明する文でないか，完結した文章でないこともあるため，内容の提示順については本調査の対象外とした。

BCCWJ における象の意味的な用例 1,314 種類を内容で分類すると，1%以上の割合で見られた内容には表 3 の種類が見られた。

表 3　BCCWJ における内容別用例分類結果出現割合上位（1%以上）

内容	出現割合	内容	出現割合	内容	出現割合
固有（象？）	20.7%	場所（国・動物園）	5.7%	歴史（祖先・来歴）	4.9%
共起（並列）	4.0%	造形（かたどったもの）	3.8%	飼育する（人が）	3.7%
大きいこと[9]	3.3%	比喩[10]	3.1%	乗る（人が）	3.0%
象牙（密猟含）	2.9%	訓練する（人が）	2.6%	種類（下位カテゴリ）	2.6%
鼻について	2.2%	伝説（英雄譚・歴史）	2.1%	共起（対照）	1.8%
重いこと[11]	1.5%	性質	1.4%	食べる（量・種類）	1.4%
例示	1.4%	メディア（経験取得）	1.2%		

まず，コーパスデータの中には，動物の象であることが擬人化などにより曖昧な固有の

[9] 以下の注 9 も同様であるが，比喩・例示と別項目に分類した例にも，大きさに関して喩える例や，大きなものとして例示している例が見られる。以下のような用例を「大きいこと」として扱うと，全体の 4.6%が大きさに関する意味的な用例であるといえる。
(補例 1) ゾウをのんだウワバミになったような、変な気分になってしまう。だから、やめよう。
（LBhn_00019：荻原規子『これは王国のかぎ』）

[10] 比喩用例として分類した用例のターゲットドメインによる細分類は以下である。

形状	大きさ	動作	耳	鼻	様態	情景	不明
1.4%	0.9%	0.4%	0.2%	0.2%	0.1%	0.1%	0.1%

比喩用例において「耳」「鼻」が着目されることからも，象は一般に「耳」と「鼻」が特徴的と考えられている可能性が考えられる。

[11] 注 7 と同様に，比喩・例示と別項目に分類した例にも，重さに関して喩える例や，重いものとして例示している例が見られる。以下のような用例を「重いこと」として扱うと，全体の 2.7%が重さに関する意味的な用例であるといえる。
(補例 2) 入ってる辞書的にはキヤノンがよかったのですが、象が踏んでも壊れない（←筆箱だって？）頑丈さと、なんと言っても電子辞書シェアＮｏ．１と言うことで、カシオになりました。
（OY05_06688：Yahoo!ブログ，原文ママ）

象用例が多く現れ，20.7%がこの種と分類された。本稿では，以下の(4)(5)のような例は固有の象と判断し，その他への細分類を行わなかった。

(4) それから白い象は大急ぎでドアに鍵をかけ、鍵はドアマットの下に押し込み、森のほうへとっとと駆けてゆきました。もちろん人の声が聞こえたのとは反対の方向へ。
（LBln_00034：C・ネストリンガー作／松島富美代訳『象さんの素敵な生活』）

(5) 大きな湖を見わたして暮らそうと、ババールがつくった「セレストビル」。学校や病院や図書館、そして映画やお芝居を楽しめる「たのしみのやかた」もある、りっぱな都です。ぞうたちが、みんな楽しく平和に暮らすババールの国。
（PM51_00768：『月刊ＭＯＥ』2005 年 9 月号）

このほかの取得可能な象に関する要素としては，見ることのできる場所（国や生息地域，動物園名，出現メディアなど），形を知ることのできるもの（模ったもの），人との関係（飼育・訓練を行うこと，乗ること，象牙をとることなど），歴史（祖先や来歴）と伝説，カテゴリ（並列・対照して共起するもの）が主となった。

上位で出現する内容を見るに，対象物そのものについては，「大きいこと」「重いこと」「鼻」が特徴的な情報として取得できている。

３．３　作文実験

「対象物をまったく知らない人に説明する」という条件提示によって，象の説明文を作文する実験を行った。クラウドソーシングを用いたタイピング入力による作文の取得を行った[12]。実験協力者は，Yahoo! クラウドソーシングに登録している 15 歳以上の男女 114 名で，150 文字以上 200 文字程度の分量を目安にするよう教示して作文を行った。

結果，平均 185 文字（max：248 文字，min：150 文字）の 114 説明文を得た。オンライン実験の特性上，Wikipedia や辞書類のコピー&ペーストも見られたが，文字数の範囲に貼り付けた部分が各々異なることや，文字数や文末表現などの調整が行われていることを鑑み，すべて調査対象とした。記述内容は 1 文あたり平均 8,2（min：4，max：13）の要素が得られた。

表 4 に記述割合が上位（25%以上）であった内容とその現れた順位を示す。形容表現については，その説明の有無に別があるため，内訳を示した。半数以上の実験協力者が記述した内容は，鼻が長いこと（96%：「鼻について 65%」，「鼻が長いことのみ（詳述なし）44%」，「長いこと＋鼻について（後述追記）47%」，「鼻の長さについて（詳述あり）4%」），大型であること（73%：「大型であることのみ（詳述なし）7%」，「大型であること（詳述あり）66%」），耳が大きいこと（65%：「耳が大きいことのみ（詳述なし）61%」，「耳の大きさについて（詳述あり）4%」）の 3 種類であった。象について説明する際，「鼻が長い」「大型」「耳が大きい」ことは重要な要素であると考えられる。

[12] クラウドソーシング実験の前に，手書き作文を取得する実験を行った。実験協力者は 3 名（20 代～50 代の男女）で， 1 回につき 5 分間の作文を行った。同様に記述を繰り返すことを 4 回行った。解答用紙は都度回収し，同内容を記述する要請などの条件提示は行っていない。得られた解答数は，3 人分×4 回の 12 説明文である。平均 299 文字（max：448 文字，min：170 文字）を得た。この結果により，200 文字程度と文字数の目安を設定した。

また，記述された順序としても，1 番目に「鼻が長い（39%)」「大型である（30%)」，2 番目に「耳が大きい（24%)」「哺乳類である（18%)」が出現しやすかったという傾向が見られる。

表 4　作文実験における「象」の記述内容とその記述順序（上位）

記述要素	記述あり	1 番目	2 番目	3 番目	4 番目	5 番目	6 番目	7 番目	8 番目	9 番目
「長い」鼻	96%	39%	18%	19%	8%	9%	1%	2%	0%	1%
（後述追記あり）	47%	20%	11%	8%	2%	5%	0%	1%	0%	1%
（詳述なし）	44%	18%	5%	10%	6%	4%	1%	1%	0%	0%
（詳述あり）	4%	2%	2%	1%	0%	0%	0%	0%	0%	0%
「大型」である	73%	30%	12%	17%	6%	2%	3%	2%	1%	0%
（詳述なし）	7%	4%	2%	1%	1%	0%	0%	0%	0%	0%
（詳述あり）	66%	27%	11%	16%	5%	2%	3%	2%	1%	0%
「大きな」耳	66%	4%	24%	12%	12%	6%	4%	1%	3%	0%
（後述追記あり）	1%	0%	0%	0%	0%	0%	1%	0%	0%	0%
（詳述なし）	61%	4%	23%	12%	12%	5%	3%	1%	3%	0%
（詳述あり）	4%	1%	1%	1%	0%	1%	0%	0%	0%	0%
鼻について	65%	0%	6%	8%	13%	9%	12%	6%	4%	4%
象牙について	47%	0%	1%	5%	6%	5%	11%	8%	7%	2%
哺乳類	35%	11%	18%	4%	4%	0%	0%	0%	0%	0%
生息地	35%	10%	5%	3%	4%	3%	5%	1%	2%	0%
重さについて	31%	0%	4%	4%	12%	8%	2%	2%	0%	0%
動物園にいる	31%	0%	1%	1%	1%	3%	3%	7%	5%	5%
草食である	27%	0%	1%	6%	4%	4%	4%	1%	4%	1%
水浴びをする	27%	0%	0%	1%	4%	11%	2%	3%	0%	1%

４．考察：象の鼻はどのように長いか

３で得たデータから，テキストに記述される情報からとくに象の鼻の長さがどのように取得されたか見ることで，象の鼻の長さがテキストからどう得られるのか考察する。

４．１　象の鼻は「長い」

象の「鼻が長い」ことについては，ほぼ全ての種類のテキストから記述が得られた。辞書においては 10 種全てで，コーパスにおいては対象物そのものについての要素として最頻出（2.2%）で，象の説明作文においては 96%で，記述があった。作文で記述される順序を見ても，1 番目であることが最も多く（39%)，3 番目までには 75%が記述される。象の「鼻が長い」ことは，象の形態的な特徴として言及されやすい要素であるといえよう。但し，作文データの詳細を見てみると，具体的な形態の説明や長さを示す記述（比喩表現，例示など）が加えられていたのは 4%（以下の(6)(7)など）のみであり，鼻についての詳細説明があった例は 47%（以下の(8)(9)など）あるが，残る 44%では，その長さの記述が全くない（以下の(10)など)。

(6) 鼻がホース状で長く牙が左右の口角にある。
(7) 鼻が長いのが特徴で、立っていても地面に届く程に長い。
(8) その長い鼻を使って器用に水を飲んだり、高いところにある果実を取る。
(9) 鼻は器用に動かすことができ、餌を口に運んだり水を飲むことも出来ます。
(10) 鼻の長い動物である。

また，コーパスから取得した用例は以下のようなものがあった。(10)に近い(11)(12)のような鼻の長さのみの例や，(8)(9)に類し(13)のように説明の加わる例も見られる。この(13)における「ニュルニュルッと、私の手元めがけて伸びて」くるという鼻の情報は，(6)(7)と同じく具体的な形態を認識することに役立つと考えられる。

(11) 校長先生に紹介されて、壇の上にあがった上野先生は、ゆっくりと、静かな声で、ぞうの話をはじめました。「ぞうさんは、食べ物をちょうだいと、長い鼻をのばしながら死にました。（後略）」　（LBkn_00031：矢崎節夫『先生のピアノが歌った』）

(12) 長い鼻がどこか象を思わせる愛敬のある顔が、のぞき込んだ。驚くほど英語がうまい。「どうせカネ目当てだろう。案内なんかいらない」と、いったんは断わったが、あまりのしつこさに根負けして、とうとう物乞いのガイドで市内の名所を見てまわるはめになった。　（LBa3_00045：五島昭『インドの大地で』）

(13) 「あなたがミッキー？　こんにちは」　息を切らしながら駆け寄る私の前に、突き出されたのは、なんと、ゾウの長〜い鼻！！　輸送用の檻の隙間からニュルニュルッと、私の手元めがけて伸びてきます。　（LBs4_00063：坂本小百合『ゾウが泣いた日』）

しかし，象の鼻は「長い」のであるが，どの程度長いのかという詳細情報がテキストからは得にくい。但し，(14)のように，比喩表現に用いられている場合などには，喩えたものの知識がある場合，具体的な情報の得られる可能性がある。

(14) だから、医者はお腹だけでなく、必ずからだ全体を診察するのだ。鼻だけを触って、ゾウは蛇のように長い動物だといった寓話もある。木を見て森を見なければ、誤診の道をたどることにもなりかねない。
（LBm4_00049：奈良信雄『名医があかす「病気のたどり方」事典』）

４．２　象の鼻はどのくらい「長い」のか

今回行った調査では，辞書・コーパス・作文のすべてのテキストで，象の鼻に関して具体的な数値（メートルなど）や比較対象などの記述があったのは(15)のみであった。

(15) 現在の大人のアフリカゾウの鼻の長さは三メートル近くあります。ゾウの鼻が、だんだん長くなってきたのは確かなのですが、どうして長くなったのかという科学的な理由は、現在でもわかっていません。
（LBqn_00035：久道健三『かがくなぜどうして』二年生）

国語辞書では 50%が，作文実験においては 44.2%が，「長い」とのみ記述しており，具体

的に詳細を示そうとする記述はなかった。これは，象の鼻が「長い」とのみいう場合，比較対象が一般的に予測されるとの前提で記述されているためと考えられる。

たとえば，象の属するあるカテゴリ（アフリカ獣上目）には，同じくハネジネズミやツチブタ（図 2）などの「鼻が長い」と評せられるメンバーが含まれている。象をはじめこれらの動物はそれぞれ鼻の長さが異なるが，どれも「長い」と評され得る。しかし，これらはその名前からもそれぞれネズミやブタのようなカテゴリが想定され，ネズミカテゴリやブタカテゴリにおいて「鼻が長い」という他メンバーと異なる特徴を有しているのであろう。

図 2　ハネジネズミとツチブタ

http://ja.wikipedia.org/wiki/ハネジネズミ より　　http://ja.wikipedia.org/wiki/ツチブタ より

しかし，辞書では「鼻が長い」と同率を占めた「大型」な動物であることが，作文の 73%で記述されていた。大きさについては，「鼻が長い」と異なり，具体的な数値や陸生動物最大であることなどの詳細情報が 66%で記述されており，「大型」であることの説明が加えられている割合が高い。「大型」は属するカテゴリ内においてもメンバーの差異として大小をいうことがあるため，一般的に「大型」というものが前提的に想定しにくい可能性が考えられる。大きさについては具体的な情報が必要と判断される場合が多いといえる [13]。

また，身体部位については，言語活動を行う人間も有している部位である場合，言及がなければ人間の部位を比較対象として想定することができるため，あえて正確な記述が必要ない可能性もある。しかし，象の「鼻が長い」ことや「耳が大きい」ことは，人間と比較するに差が大きい。テキストからのみ象の鼻の長さを明確に認識することは困難であろう。

５．まとめ

テキストから対象物に関して得られる情報として，コーパスから取得できる用例の頻度を見ると，場所情報と人間との関係情報が上位となっている（3.2 参照）。また，対象物の説明を試みた場合，特徴的と考えられる形状情報が記述されやすい。とくに形状の情報が一番目に記述されやすく，次いで場所や人間との関係が記述されるという傾向がある（3.1，3.3 参照）。

動物の象に関するテキストにおいて，全体的な大きさ（「大型」）については説明に補足的な情報が加わっていることが多く（本稿の作文実験では 66%），具体的に程度を説明しようという傾向が見られた。しかし，特徴的部位の長さや大きさは，一般的な程度認識が期待され，具体的な記述が得にくいという結果が見られた。「大型」「鼻」はコーパス・説明文ともに頻度としては上位であるが，補足的な情報は得にくく（半数以下の割合）,具体的な程度は得にくいのである。

---

13 「鼻が長い」「大型」に続いて高頻度で記述されていたのは「耳が大きい」の 65%であるが，その大きさについての詳細は 4%にとどまっていた。すなわち，特徴的な身体部位についての「大きい」という形容は，鼻についての「長い」同様，一般的な程度が前提的に期待されている可能性がある。

よって，象の鼻の長さがどの程度であるかという情報は，テキストから得にくいといえる。これは，文化的に標準と考えられる長さや大きさなどが，前提的に必要とされるためであると考えられる。今後，文化的背景の異なる相手への情報伝達において，説明文に何を記述すべきか応用可能性を考えたい。

謝　辞

本研究は JSPS 科研費 26770156 の助成を受けたものである。

文　献

Goddard, Cliff. and Wierzbicka, Anna. (2014) *Words and Meanings*. Oxford: Oxford University Press.

Fillmore, Charles. J. and Atkins, Beryl. T. Sue. (1994) “Starting where the dictionaries stop: The challenge for computational lexicography.” In B. T. S. Atkins and A. Zampolli, eds., *Computational Approaches to the Lexicon*, Oxford: Oxford University Press. pp. 349–393.

加藤祥 (近刊) 「テキストからの対象物認識に有用な記述内容―動物を例に―（仮）」『国立国語研究所論集』9

加藤祥，岡本雅史，荒牧英治 (近刊)「テキスト世界と現実世界の差異―動物の部位分布における 3 つのプロトタイプ効果」山梨正明編『認知言語学論考』12，ひつじ書房.

国広哲也 (1997)『理想の国語辞典』，大修館書店.

Maekawa, Kikuo, Yamazaki, Makoto., Ogiso, Toshinobu., Maruyama, Takehiko., Ogura, Hideki., Kashino, Wakako., Koiso, Hanae., Yamaguchi, Masaya., Tanaka, Makiro., and Den, Yasuharu.(2014) “Balanced corpus of contemporary written Japanese.” Language Resources and Evaluation 48 (2): 345-371 (DOI10.1007/s10579-013-9261-0).

Sinclair, John. (1992) “Trust the text.” In Davies, M. and L. Ravelli, eds., *Advances in Systemic Linguistics: Recent Theory and Practice*, London: Pinter. pp. 5–19.

Wierzbicka, Anna (1985) *Lexicography and Conceptual Analysis*. Ann Arbor, MI: Karoma Publishers, Inc.

Wierzbicka, Anna (1986) *Semantics: Prime and Universals*. Oxford: Oxford University Press.

資料

現代日本語書き言葉均衡コーパス（国立国語研究所）

三省堂国語辞典（5 版）,新明解国語辞典（6 版），岩波国語辞典（5 版），明鏡国語辞典（初版），新選国語辞典（7 版），集英社国語辞典（2 版），角川国語辞典（新版），新潮現代国語辞典（2 版），大辞林（3.0：Web 更新版），デイリー国語辞典（3 版），COBUILD（2009）

Kudo, Taku, and Hideto Kazawa. (2007) “Web Japanese N-gram Version 1”, Gengo Shigen Kyokai.

関連 URL

現代日本語書き言葉均衡コーパス（国立国語研究所）
http://www.ninjal.ac.jp/corpus_center/bccwj/

コーパス検索アプリケーション「中納言」1.1.0，短単位データ 1.0，長単位データ 1.0
https://chunagon.ninjal.ac.jp/

Yahoo! クラウドソーシング　http://crowdsourcing.yahoo.co.jp/

文書間距離尺度の特性

浅原 正幸 (国立国語研究所)*　加藤 祥 (国立国語研究所)

On the Document Distance Metric with n-gram and p-mer

Masayuki Asahara (NINJAL)　Sachi Kato (NINJAL)

要旨

本研究では文書間距離尺度（類似度・相関係数）の特性を様々なコーパスを用いて評価する。まず、代表的な文書間距離尺度として単純な 1-gram 形態素ベクトルから、n-gram、p-mer、順序尺度の数理的な構造を整理する。次に各文書間距離尺度を要約・語釈・再話課題コーパスを用いて評価する。多人数による課題間における尺度のふるまい、個人による再話の複数回間の尺度のふるまい、口述・タイプ入力・筆述など言語生成方法間の尺度のふるまいを分析する。

1. はじめに

まず、最初に語順に対する順序尺度を含めた距離空間・類似度・カーネル・相関係数により既存の自動評価指標の整理を行う。先行研究の文献では連続記号列を表す部分文字列 (substring) とギャップを許す部分列 (subsequence) との混同が見られ、定性的な議論が弱い。本稿では、大きく分けて一致部分文字列による尺度・一致部分列による尺度・ベクトル型順序尺度・編集型順序尺度の四つに分類し議論する。

次に言語生産・受容過程の多様性を 4 種類の尺度により評価する。複数人が同一課題を実施した場合の各尺度の分散や、同一人が同一課題繰り返し実施した場合の各尺度の分散などを検討する。生産過程においては口述・筆術・タイプ入力の 3 種類について評価し、課題においては要約・語釈・再話について評価する。

なお、本節で用いる用語や記号の定義は浅原ほか (2015) の付録を参照されたい。

2. 文書間距離

2.1 LCStr, LCS

2.1.1 記号列と文字列と部分文字列と部分列

最初に記号列と文字列と部分文字列と部分列の違いについて確認する。

何らかの全順序が付与されている記号集合のことを**記号列**と呼ぶ。本稿では記号列ベクトル $s = \langle s_1, \ldots, s_m \rangle, t = \langle t_1, \ldots, t_m \rangle$ などで表現する。文書は文字 (character) ベースの記号列もしくは形態素解析後の形態素 (morpheme) ベースの記号列とみなすことができる。

評価する記号列上の連続列のことを**文字列 (string)** と呼ぶ。記号列の要素が文字 (character) である場合を「文字ベースの文字列 (character-based string)」、記号列の要素が形態素

* masayu-a@ninjal.ac.jp

(morpheme) である場合を「形態素ベースの文字列 (morpheme-based)」と呼ぶこととする。

記号列に対して隣接性と順序を保持した部分的記号列のことを**部分文字列 (substring)** と呼ぶ。長さ n の部分文字列を特に n-gram 部分文字列と呼ぶ。記号列 s の i 番目の要素からはじまる n-gram 部分文字列を $s_{i,\ldots,i-n+1}$ で表現する。

記号列に対して順序を保持した部分的記号列のことを**部分列 (subsequence)** と呼ぶ。隣接性は保持しなくてよい。長さ p の部分列を特に p-mer 部分列と呼ぶ。記号列 s の p-mer 部分列を、インデックスベクトル $\vec{i} = \langle i_1, \ldots, i_p \rangle (1 \leq i_1 < i_2 < \cdots < i_p \leq |s|)$ を用いて、$s[\vec{i}]$ と表す。

2.1.2　最長共通部分文字列 (Longest Common String: LCStr) 長

最長共通部分文字列 (Longest Common String) の abbreviation は LCS だが、一般には 2.1.2 に示す最長共通部分列 (Longest Common Subsequence) のことを LCS と呼ぶことが多い。本稿では前者を LCStr, 後者を LCS と呼び、区別する。

記号列 s, t を与えた際の最長共通部分文字列を次式で定義する: $\mathrm{LCStr}(s,t) = \mathrm{argmax}_{s_{i,\ldots,i-n+1} | \exists j, s_{i,\ldots,i-n+1} = t_{j,\ldots,j-n+1}} n$. 記号列 s, t を与えた際の最長共通部分文字列長 (LCStr 長) を次式で定義する: $|\mathrm{LCStr}|(s,t) = \max_{\forall i, \forall j, s_{i,\ldots,i-n+1} = t_{j,\ldots,j-n+1}} n$. これを [0,1] 区間に正規化すると以下のようになる: $\mathrm{Score}_{\mathrm{LCStr}}(s,t) = \frac{2 \cdot |\mathrm{LCStr}|}{|s|+|t|}$.

2.1.3　最長共通部分列 (Longest Common Subsequence: LCS) 長と Levenshtein 距離

記号列 s, t を与えた際の最長共通部分列 (Longest Common Subsequence: LCS) を次式で定義する: $\mathrm{LCS}(s,t) = \mathrm{argmax}_{s[\vec{i}] \exists \vec{j}, s[\vec{i}] = t[\vec{j}]} |\vec{i}|$. 記号列 s, t を与えた際の最長共通部分列長 (LCS 長) を次式で定義する: $|\mathrm{LCS}(s,t)| = \max_{\forall \vec{i}, \forall \vec{j} : s[\vec{i}] = t[\vec{j}]} |\vec{i}|$. [0,1] 区間に正規化すると、以下のようになる: $\mathrm{Score}_{\mathrm{LCS}}(s,t) = \frac{2 \cdot |\mathrm{LCS}|}{|s|+|t|}$. なお、挿入のコストを 1、削除のコストを 1、代入のコストを 2(もしくは代入を禁止) した場合の Levenshtein 距離 (編集型) と LCS 長の関係は以下のようになる: $\mathrm{d}_{\mathrm{Levenshtein}}(s,t) = |s| + |t| - 2 \cdot |\mathrm{LCS}|$. さらに LCS は §2.2.2 で示すとおり、対称群上の編集型距離のうちの Ulam 距離と深く関連し、一種の順序尺度であるとも考えられる。

2.2　関連するカーネル・順序尺度

以下では、関連するカーネルおよび順序尺度について確認する。

2.2.1　カーネル・距離 (文字列の共有)

畳み込みカーネルのうち系列データに対するカーネル (Shawe-Taylor ほか (2010)) は、共通する可能な部分文字列・部分列を数え上げる。いずれも効率よく計数する方法が提案されている。また、適切に正規化することにより部分文字列・部分列の共有についての距離やスコアを規定することができる。

各カーネルの説明に入る前に、スコア化 ([0,1] 区間正規化) について示す。カーネルのスコア化は次式により行われる: $\mathrm{Score}_{K\_}(s,t) = \frac{K\_(s,t)}{\|K\_(s,s)\| \|K\_(t,t)\|}$.

■全部分文字列カーネルと文字列長加重全部分文字列カーネル　全部分文字列カーネル (All String Kernel or Exact Matching Kernel) は共通する全ての部分文字列の数を数える。長さ n の部分文字列 u を座標とする特徴量空間 $\Phi^*_{\mathrm{str}} : \sigma^* \to F_{\mathrm{all\_str}} \sim R^{|\sigma|^*}$ 但し $\Phi^*_{\mathrm{str}} = (\phi^*_u(s))_{u \in \sigma^*}$ を考える。$K_{\mathrm{n\text{-}gram}}(s,t) = \langle \Phi^*_{\mathrm{str}}(s), \Phi^*_{\mathrm{str}}(t) \rangle_{F_{\mathrm{all\_str}}} = \sum_{u \in \sigma^*} \phi^*_u(s) \phi^*_t(s), (\text{但し} \phi^*_u(s) = |\{i | s_{i\ldots *} = u\}|)$. カーネル関数を直接計算すると以下のよう

になる: $K_{\text{all\_seq}}(s,t) = \sum_{n=1}^{\min(|s|,|t|)} \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \delta(s_{i\ldots i+n-1}, t_{i\ldots i+n-1})$.

このカーネルは、提案された 2002 年ごろではバイオインフォマティクスなど特定の分野以外では有効な用途が提案されていない。言語処理の場合、得られる n-gram に対して加重をかけることが一般に行われている。例えば、文字列長に対して加重をかけたものを文字列長加重全部分文字列カーネル (Length Weighted All String Kernel or Length Weighted Exact Matching Kernel) と呼ぶ。$K_{\text{all\_seq}}(s,t) = \sum_{n=1}^{\min(|s|,|t|)} \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \omega_{|s|}\delta(s_{i\ldots i+n-1}, t_{i\ldots i+n-1})$. ここで ω_n は長さ n に対する重みを表す。このカーネルと次の n-スペクトラムカーネルは Suffix Tree を用いて効率よく計算する方法が提案されている。

■n-スペクトラムカーネル　n-gram スペクトラムカーネル (Spectrum Kernel) は共通する長さ n の部分文字列 (n-gram) の数を数える。長さ n の部分文字列 u を座標とする特徴量空間 $\Phi^n_{\text{str}} : \sigma^* \to F_{\text{n-gram}} \sim R^{|\sigma|^n}$(但し $\Phi^n_{\text{str}} = (\phi^n_u(s))_{u\in\sigma^n}$) を考える。$K_{\text{n-gram}}(s,t) = \langle\Phi^n_{\text{str}}(s), \Phi^n_{\text{str}}(t)\rangle_{F_{\text{n-gram}}} = \sum_{u\in\sigma^p} \phi^n_u(s)\phi^n_t(s)$(但し$\phi^n_u(s) = |\{i|s_{i\ldots i+n-1} = u\}|$) 直接計算すると以下のようになる: $K_{\text{n-gram}}(s,t) = \sum_{i=1}^{|s|-n+1} \sum_{j=1}^{|t|-n+1} \delta(s_{i\ldots i+n-1}, t_{j\ldots j+n-1})$.

■全部分列カーネル　全部分列カーネルは共通するすべての部分列の数を数える。任意の長さの部分列 v を座標とする特徴量空間 $\Psi^*_{\text{seq}} : \sigma^* \to F_{\text{all\_seq}} \sim R^{|\sigma|^\infty}$(但し$\Psi^*_{\text{seq}}(s) = (\psi^*_v(s))_{v\in\sigma^*}$) を考える。$K_{\text{all\_seq}}(s,t) = \langle\Psi^*_{\text{seq}}(s), \Psi^*_{\text{seq}}(t)\rangle_{F_{\text{all\_seq}}} = \sum_{v\in\sigma^*} \psi^*_v(s)\cdot\psi^*_v(t)$(但し$\psi^*_v(s) = |\{\vec{i}|s[\vec{i}] = v\}|$). $K_{\text{all\_seq}}(s,t)$ は以下のように再帰的に計算することにより $O(|s||t|)$ で計算することができる。ϵ を空記号列とすると $K_{\text{all\_seq}}(s,\epsilon) = K_{\text{all\_seq}}(t,\epsilon) = 1$ とし、$K_{\text{all\_seq}}(s,t)$ が求まると $K_{\text{all\_seq}}(s{\cdot}a,t) = K_{\text{all\_seq}}(s,t) + \sum_{1\le i\le|t|, j:t_j=a} K_{\text{all\_seq}}(s, t_{i\ldots j-1})$ と s 再帰的に定義できる。さらに $\tilde{K}_{\text{all\_seq}}(s{\cdot}a,t) = K_{\text{all\_seq}}(s,t_{i\ldots j-1})$ とすると、$\tilde{K}_{\text{all\_seq}}(s{\cdot}a,t{\cdot}b) = \tilde{K}_{\text{all\_seq}}(s{\cdot}a,t) + \delta(a,b)K(s,t)$ と t 再帰的に定義できる。

■固定長部分列カーネル　固定長部分列カーネルは共通する長さ p の部分列 (p-mer) の数を数えあげる。長さ p の部分文字列 v を座標とする特徴量空間 $\Psi^p_{\text{seq}} : \sigma^* \to F_{\text{p-mer}} \sim R^{|\sigma|^p}$(但し$\Psi^p_{\text{seq}}(s) = (\psi^p_v(s))_{v\in\sigma^*}$). を考える。$K_{\text{p-mer}}(s,t) = \langle\Psi^p_{\text{seq}}(s), \Psi^p_{\text{seq}}(t)\rangle_{F_{\text{p-mer}}} = \sum_{v\in\sigma^p} \psi^p_v(s)\cdot\psi^p_v(t)$. ここで $\psi^p_v(s) = |\{\vec{i}|s[\vec{i}] = v\}|$ とする。

■ギャップ加重部分列カーネル　ギャップ加重部分列カーネル: p-mer の部分列の数え上げの際に隣接性を考慮して重み λ を加重する。長さ p の部分列 v を座標とする特徴量空間 $F_{\text{p-mer}}$ を考える。$K_{\text{gap\_p-mer}}(s,t) = \langle\Psi^{gap\text{-}p}_{\text{seq}}(s), \Psi^{gap\text{-}p}_{\text{seq}}(t)\rangle_{F_{\text{p-mer}}} = \sum_{v\in\sigma^p} \psi^{gap\text{-}p}_v(s)\cdot\psi^{gap\text{-}p}_v(t)$ ここで $\psi^{gap\text{-}p}_v(s) = \sum_{\vec{i}:v=s[\vec{i}]} \lambda^{l(\vec{i})}$ とし、$l(i) = |s_{i_1,\ldots,i_{|v|}}|(\vec{i} = \langle i_1,\ldots,i_{|v|}\rangle)$ とする。

2.2.2　順序尺度

以下では順序尺度について考えるが、神嶌 (2009) が詳しい。基本的には同じ長さ m の二つの順位ベクトル $\mu,\nu \in S_m$ に対する 2 種類の距離を考える。

■順位ベクトル型距離　一つ目の距離は「順位ベクトル型」の距離で順位ベクトルを m 次元空間中の点を表すベクトルとみなし、ベクトル空間上の距離を定義する。ベクトル空間を θ-ノルム採用すると以下のようになる: $d_{\|\text{Rank}\|_\theta}(\mu,\nu) = (\sum_{i=1}^m |\mu(i) - \nu(i)|^\theta)^{1/\theta}$. ここで $\theta = 1$

表 1　指標・スコア・距離・カーネル・相関係数の関係まとめ

	スコア $[0,1]\uparrow$	距離 $[0,\infty]\downarrow$	カーネル $[0,\infty]\uparrow$	相関係数 $[-1,1]\uparrow$
部分文字列系	$\mathrm{Score}^{(\gamma)}_{K\mathrm{all\_str}}$		(加重) 全部分文字列 §2.2.1	
(n-gram)	$\mathrm{Score}^{(\gamma)}_{K\text{n-gram}}$		n-スペクトラム §2.2.1	
部分列系	$\mathrm{Score}^{(\gamma)}_{K\mathrm{all\_seq}}$		(加重) 全部分列 §2.2.1	
(p-mer)	$\mathrm{Score}^{(\gamma)}_{K\text{p-mer}}$		p-mer 部分列 §2.2.1	
	$\mathrm{Score}^{(\gamma)}_{K\mathrm{gap\_p\text{-}mer}}$		加重 p-mer 部分列 §2.2.1	
順序系 §2.2.2	$\mathrm{Score}_{\|\mathrm{rank}\|_\theta}$			
(ベクトル型)	$\mathrm{Score}_{\mathrm{footrule}}$	$d_{\mathrm{footrule}(\theta=1)}$		
	$\mathrm{Score}_{\mathrm{Spearman}}$	$(d_{\mathrm{Spearman}(\theta=2)^2})$		Spearman's ρ
	$\mathrm{Score}_{\mathrm{Hamming}}$	d_{Hamming}		
順序系 §2.2.2 (編集型)	$\mathrm{Score}_{\mathrm{Kendall}}$	d_{Kendall}		Kendall's τ
(最長一致部分列長)	$\mathrm{Score}_{\mathrm{LCS}}$	d_{Ulam}§2.1.3		
(加重最長一致部分列長)	$\mathrm{Score}^{(\gamma)}_{\mathrm{WLCS}}$			
(最長一致部分文字列長)	$\mathrm{Score}_{\mathrm{LCStr}}$			

の場合、特に Spearman footrule と呼ぶ。$d_{\mathrm{Footrule}}(\mu,\nu) = (\sum_{i=1}^{m} |\mu(i) - \nu(i)|)$. $\theta = 2$ の場合は通常の Euclid 距離だが、この Euclid 距離を 2 乗したものを特に Spearman 距離と呼ぶ。$d_{\mathrm{Spearman}}(\mu,\nu) = (\sum_{i=1}^{m} |\mu(i) - \nu(i)|^2)$. Spearman 距離は、距離の公理のうち対称性と正定値性を満たす。しかし、Euclid 距離を 2 乗したものなので三角不等式を満たさないが、慣習的として距離として扱われる。さらに [-1, 1] 区間に正規化したものは Spearman の順位相関係数 ρ として知られている。Spearman's $\rho = 1 - \frac{6 \cdot d_{\mathrm{Spearman}}(\mu,\nu)}{m^3 - m}$. この値は順序尺度に基づく二つの順位ベクトル μ, ν の Pearson 相関関係と等しい[(1)]。その他、順位ベクトルの同一順位のものが同じ要素である要素数を数えた Hamming 距離がある。$d_{\mathrm{Hamming}}(\mu,\nu) = \sum_{i=1}^{m} \delta(\mu(i), \nu(i))$. Hamming 距離は文字列上で代入 (コスト 1) のみを許した編集距離としても解釈できる。

■対称群上の編集型距離　二つ目の距離は「編集型」の距離である。

順序ベクトルを記号列とみなした場合、順位ベクトル μ をもうひとつの順位ベクトル ν に変換するために必要な最小操作数を Levenshtein 距離について述べた。以下では、順序ベクトルを対称群とみなした場合の編集型距離について述べる。編集に許される操作によっていくつかの距離のバリエーションがある。

Kendall 距離 d_{Kendall} は順序ベクトルを対称群とみなした際に隣接互換によって置換する最小回数によって定義される。言い換えると隣接する対象対を交換 (Swap) する操作の最小回数を用いたものである。Kendall 距離は、二つの順位ベクトル中の $\frac{m(m-1)}{2}$ 個の対象対のうち逆順になっている対の数に等しい。$d_{\mathrm{Kendall}} = \min(\mathrm{argmax}_q\, \delta((\Pi_{q=1}^{q} \pi_2(k_q, k_q + 1)) \cdot \mu, \nu)) = \sum_{i=1}^{m} \sum_{j=i+1}^{m} \chi(i,j)$. ここで χ は対象対 $\langle i, j \rangle$ が同順のとき 0、逆順のとき 1 を返す指示関数: $\chi = \begin{cases} 1 & \text{if } (\mu(i) - \mu(j))(\nu(i) - \nu(j)) < 0, \\ 0 & \text{if } (\mu(i) - \mu(j))(\nu(i) - \nu(j)) \geq 0 \end{cases}$ これをスコアとして使いやすくするために [0,1] 区間の範囲に正規化すると以下のようになる: $\mathrm{Score}_{\mathrm{Kendall}} = 1 - \frac{2 \cdot d_{\mathrm{Kendall}}(\mu,\nu)}{m^2 - m}$. こ

表 2　指標評価に使う言語資源

言語資源名	収集場所	生成過程	繰り返し	取得人数	摘要
BCCWJ-SUMM_C	クラウドソーシング	タイプ入力	なし	100-200	19 文書の要約
BCCWJ-SUMM_L	実験室	筆述	3 回	のべ 47	8 文書の要約
GROSS_C	クラウドソーシング	タイプ入力	なし	71,111,113	鶏・兎・象の語釈
GROSS_L	実験室	筆述	4 回	7,6,3	鶏・兎・象の語釈
RETELLING_I	実験室	口述	10 回	5	インタビュー
RETELLING_K	実験室	口述	3 回	3,3,3	怪談 3 種の再話
RETELLING_M	実験室	筆述	4 回	10	物語「桃太郎」の再話

れを [-1,1] 区間の範囲に正規化したものは Kendall の順位相関係数 τ として知られている。Kendall's $\tau = 1 - \frac{4 \cdot d_{\mathrm{Kendall}}(\mu,\nu)}{m^2-m}$.

Ulam 距離 d_{Ulam} は順序ベクトルを対称群とみなした際に連続した順序ベクトル部分列 $i, i+1, \ldots, j-1, j\rangle$ の巡回置換の操作のみによって置換する最小回数によって定義される。これは「本棚の本の入れ換え」で例えられる。順位ベクトル μ で並んでいる本棚の本を順位ベクトル ν に並び替えるために、ある要素を抜いて別の場所に挿入するということを行う。Ulam 距離は同じ要素が記号列に存在しないという前提のもと、最大共通部分列距離と以下の関係にあることが知られている。$d_{\mathrm{Ulam}}(\mu,\nu) = m - |\mathrm{LCS}(\mu,\nu)|$ これを [0,1] 区間の範囲に正規化すると以下のように正規化最大共通部分スコアと同じになる: $\mathrm{Score}_{\mathrm{Ulam}}(\mu,\nu) = 1 - \frac{d_{\mathrm{Ulam}}(\mu,\nu)}{m} = \frac{|\mathrm{LCS}(\mu,\nu)|}{m} = \mathrm{Score}_{\mathrm{LCS}}(\mu,\nu)$.

2.3　スコアの一般化

以上、指標・スコア・距離・カーネル・相関係数を議論してきた。まとめると表 1 のようになる。各スコアと人手の評価結果という観点からすると、平尾ほか (2007) のように、表 1 にあげたすべてのスコア $\mathrm{Score}_{\_} \in \{\mathrm{Score}_*\}$ の加重相乗平均 (下式) を考え、加重 $\omega_{\_}$ と各スコアに付随するパラメータを各指標の従属性や相関に注意しながら人手の評価指標との回帰により求めれば良い: $\overline{\mathrm{Score}_*} = \sum^{\omega} \sqrt{\Pi \mathrm{Score}_{\_}^{\omega_{\_}}}$. このスコアのあり方については議論すべき点がいくつかある。まず、substring(部分文字列: n-gram 系) と subsequence(部分系列: p-mer 系) との違いを踏まえる。次に最長一致部分文字列は対称群上の編集型距離である Ulam 距離と深く関連する。さらに順序に対する順位ベクトル型距離と編集型距離の間には様々な不等式が成り立つ。本稿では**スコアの一般化についてはこれ以上踏み込まない**。次節以降各スコアがさまざまな言語資源上でどのようなふるまいをするのかについてみていきたい。

3.　評価に用いる言語資源

ここでは研究室で有する言語資源のテキスト対のスコアを検証することにより、各スコアがとらえようとしているものが何なのかを分析する。表 2 に利用する言語資源について示す。まず言語生産の目的として、要約 (BCCWJ-SUMM) と語釈 (GROSS) と再話 (RETELLING) の 3 種類の言語資源を準備する。要約と語釈については、クラウドソーシングにより安価で大量にデータを得る手法 (タイプ入力) と実験室にて被験者に繰り返し同一課題を依頼してデータを得る手法 (筆述) の 2 種類の方法を用いた。再話のデータについては既存のデータを用いた。再

話については、言語生産形態として筆述による形態と口述による形態のデータを準備した。

以下各言語資源について解説する。

3.1 BCCWJ-SUMM_C

BCCWJ-SUMM_C は BCCWJ の新聞記事の要約を Yahoo! クラウドソーシング (15 歳以上の男女) により被験者実験的に作成したものである。

BCCWJ の 1 サンプルには複数の記事が含まれており、それを記事単位に分割したうえで元文書集合 19 文書を構築した。元文書集合は BCCWJ コアデータ PN サンプル (優先順位 A) から選択した。40 文字毎に改行した元文書を画像として提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。実験協力者の環境は PC 環境に限定した。元文書毎に約 100〜200 人の実験協力者が要約に従事した。実験実施時期は 2014 年 9 月である。

得られたデータには、文字数制限を守っていないもの・実験の趣旨を理解していないもの・既に実験を行った実験協力者から同一回答を提供されたと考えられるものなどが含まれており、これらを排除したものを有効要約とする。統計分析においてこの有効要約のみを用いる。

3.2 BCCWJ-SUMM_L

BCCWJ-SUMM_L は BCCWJ の新聞記事の要約を実験室環境で筆述により作成したものである。BCCWJ-SUMM_C で用いた元文書を印刷紙面で提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。一つの元文書に対して、3 回まで繰り返して要約文作成を行った。繰り返しに際しては、特別に「前と同じ要約文を作成してください」などといった指示は行わず、質問された場合にも「自由に要約文を作成してください」と教示した。実験協力者は原稿用紙上で筆述 (鉛筆と消しゴム利用) で要約を行い、そのデータを電子化した。現在のところデータは 8 文書のべ 47 人分に限定した。

3.3 GROSS_C

GROSS_C は語釈文を Yahoo! クラウドソーシング (15 歳以上の男女) により被験者実験的に作成したものである。

「その動物を知らない人がどのようなものかわかるように説明してください」と教示し、同意した実験協力者は兎 (単語親密度 6.6)・鶏 (6.4)・象 (同 6.0) の 3 種類から対象物を選択回答した(2) 。150 文字以上 250 文字以内で 3 文字以上の同文字連続は認めない設定とした。実験協力者 300 名を募集したところ得られた解答数は、鶏:71・兎:111・象:113(295/300) であった。

3.4 GROSS_L

GROSS_L は語釈文を実験室環境で筆述により収集したものである。

実験協力者 8 名 (20 代-50 代の男女) に、GROSS_C と同様に「その動物を全く知らない人がどのようなものかわかるように説明してください」と教示した。実験協力者は、10 分間で兎 (単語親密度 6.6)・鶏 (6.4)・象 (同 6.0) の 3 種類から 2 種類の対象物を選択回答した。目安として 5 分経過時にブザー音を鳴らした。選択した対象物について同様に記述を繰り返すことを 4 回行った。得られた解答数は、兎 7 人分× 4 回、鶏 6 人分× 4 回、象 3 人分× 4 回である。平均 145 文字 (max 227 文字, min 85 文字) を得た。

3.5 RETELLING_I

最初の再話のデータは「独話 Retelling コーパス」保田ほか (2013a,b) である。このコーパスは宮部ほか (2014) でも用いられている。

実験協力者は 5 名で、同一人が同内容をそれぞれ 10 回独話を繰り返した。就職活動を前提とした模擬面接の設定で、実験協力者は自ら予め用意した「学生生活で力を入れてきたこと (3 分間程度)」についての独話を行った。同内容を繰り返すことや何回依頼するかは知らせていない。5 人分× 10 回 (50 話分) の独話を取得した。面接官 (聴衆) は有無を交互とした。奇数回 (1・3・5・7・9 回) は聴衆なしの独話、偶数回 (2・4・6・8・10 回) は聴衆に対する独話である。聴衆には、聴いていることを表すために頷くことのみを許可しており、話者への質問や意見など、発話は一切行わなかった。収録は録音と録画を行い、音声データを書き起こした。

被験者によってインタビュー内容が異なるために、統計分析においては同一被験者の回数間のスコア (RETELLING_I(T)) のみを評価する。

3.6 RETELLING_K

次の再話のデータは怪談を繰り返し口述したものであり、保田・荒牧 (2012) によるものである。実験協力者は 3 名[(3)] で、実験は 1 名ずつ個別に行った。実験協力者は怪談を聞いたのち、その怪談について 3 回の再話を行った。怪談は 3 種類を用意したため、各人 9 回の語りを行った。語りに関しては、「怪談として他の人に伝えるよう話す」との指示をした。既存の物語では、個人の記憶による先入観の影響が予測されたため、4 分間程度の新規な怪談を 3 本作成した。実験環境はビデオカメラと録音機により、録音と録画を行った。聴衆の影響を除去するために、聴衆は設置しなかった。本稿では音声データを書き起こしたものを用いる。

3.7 RETELLING_M

最後の再話のデータは桃太郎の物語を筆述で繰り返し記述したものであり、保田 (2014) によるものである。実験協力者 10 名 (20 代-50 代の男女) に,「桃太郎の物語を全く知らない人に向けて記述してください」と教示し、実験協力者は 10 分間で記述 (筆述) した。同様に記述を繰り返すことを 4 回行った。平均延べ 284 語 (min:150 語・ max:451 語)、異なり語 107 語 (min:74 語・max:152 語) の「桃太郎」 10 人分× 4 回 (40 話分) を取得した。

4. 評価

本節では前節で述べたコーパスを用いて文書間距離がどのようにふるまうかを観察する。利用する文書間距離は以下の 30 種類である。

- n-gram スペクトラム (1,2,3,4) (char/mrph)
- n-gram 以下スペクトラム (≤2,≤3,≤4) (char/mrph)
- p-mer 部分列 (2,3,4) (char/mrph)
- p-mer 以下部分列 (≤2,≤3,≤4) (char/mrph)
- 1-gram スペクトラム +Footrule (char/mrph) (=Spearman)
- 1-gram スペクトラム +Kendall (char/mrph)

http://goo.gl/nBeMeZ にそれぞれの距離空間によるスコアの平均値 (Mean) と標準偏差 (SD) を示す。スコアについて “_char” は文字単位の記号列として評価したもの、“_mrph” は形態素単位の記号列 (MeCab-0.98+IPADIC-2.7.0 による) として評価したものである。シャピロ・ウィルク検定の結果、ほとんどの場合 p 値が 0.05 未満であり、正規分布とはいえない傾向が見られた。

unigram(n-gram(1)) を用いた場合、要約と語釈は中程度、再話はかなり高いスコアを達成している。GROSS_L(T) がほぼ再話と同程度のスコアで一方、BCCWJ-SUMM_L(T) が低いことから、要約を繰り返す際の言語生産の特殊性が見られる。要約を繰り返す際には、回数毎に文章中の重要箇所を変更するサンプル・被験者が存在し、標準偏差も高くなっている。

Bigram(n-gram(2)), skip-bigram(p-mer(2)) を用いた場合、異なる被験者間のスコアと繰り返し間のスコアとの間に差が見られるようになる。これは何らかの個々人の文体差が形態素の連接に影響を与えているのではないかと考える。

Bigram(n-gram(2)) と skip-bigram(p-mer(2)) の間の差として、語釈の場合のみ bigram のスコアが下がることがわかる。語釈という課題の都合上、物語や要約と異なり、情報の提示順が変わることも考えられる。しかし、順序尺度である Kendall のスコアでは bi-gram のスコアほど顕著な差が見られなかった。単語の隣接性が語釈のみ下がるというスコアの振る舞いについては今後検討していきたい。

クラウドソーシングと研究室内被験者実験との差 (BCCWJ-SUMM_C ⇔ BCCWJ-SUMM_L(P), GROSS_C ⇔ GROSS_L(P)) については、各スコア・各課題 (要約・語釈) で差が見られなかった。

4.1 課題間の評価

以下、課題間を比較するために、6 種類の評価軸を分析する。殆どの場合、正規分布であることも等分散であること (F 検定による) も仮定できない。ここではウィルコクソンの順位和検定 (0.05 未満で 2 群の代表値が左右にずれている) を行う。(4)

- 実験室における複数人の課題間の違いの評価
  BCCWJ-SUMM_L(P) ⇔ GROSS_L(P) ⇔ RETELLING_K(P) ⇔ RETELLING_M(P)
  - BCCWJ-SUMM_L(P) ⇔ GROSS_L(P)
    文字単位の評価の場合 n-gram(2,3,4)_char, Kendall_char に有意差が見られた。
    形態素単位の評価の場合 n-gram(2,3,4,≤2,≤3,≤4)_mrph, Footrule_mrph, Kendall_mrph に有意差が見られた。
  - BCCWJ-SUMM_L(P) ⇔ RETELLING_K(P)
    n-gram(3,4)_mrph **以外**で有意差が見られた。
  - BCCWJ-SUMM_L(P) ⇔ RETELLING_K(M), GROSS_L(P) ⇔ RETELLING_{K,M}(P)
    全てのスコアについて、有意差が見られた。
  - RETELLING_K(P) ⇔ RETELLING_M(P)
    n-gram(≤3,≤4)_mrph,p-mer(3,4,≤3,≤4) で有意差が見られた。

  要約 ⇔ 語釈間は n-gram(1) で有意差が見られなかった。同じ文字・同じ形態素を使うという観点では一致度のレベルが等しいが、語の連接や順序尺度が入ると有意差が見られることがわかった。グラフの見た目から語釈の方が語の連接や順序尺度の一致度が低い。これは語釈の目的としては情報の提示順に重要性がないことが伺える。
  要約 ⇔ 再話、語釈 ⇔ 再話の間においては有意差が見られた。再話は同じ話をするという特性から、一致度が高くなる一方、要約・語釈は目的を達成するがために同じ表現を用いなければならないという制約がなく、低くなる傾向にある。

- 実験室における単一人の回数間距離の課題間の違いの評価
  BCCWJ-SUMM_L(T) ⇔ GROSS_L(T) ⇔ RETELLING_I(T) ⇔ RETELLING_K(T) ⇔ RETELLING_M(T)
  - BCCWJ-SUMM_L(T) ⇔ GROSS_L(T)
    文字単位の評価の場合 n-gram(2,3,4)_char, Kendall_char に有意差が見られた。
    形態素単位の評価の場合 n-gram(2,3,4,≤2,≤3,≤4)_mrph, Footrule_mrph, Kendall_mrph に有意差が見られた。
  - BCCWJ-SUMM_L(T) ⇔ RETELLING_{I,K,M}(T), GROSS_L(T) ⇔ RETELLING_{I,K,M}(T)
    全てのスコアについて、有意差が見られた。
  - RETELLING_I(T) ⇔ RETELLING_K(T)
    文字単位の評価の場合 n-gram(1,4,≤2)_char, p-mer(2,≤2)_char に有意差が見られた。
    形態素単位の評価の場合、全てのスコアに有意差が見られた。
  - RETELLING_I(T) ⇔ RETELLING_M(T)
    Kendall_char **以外**について有意差が見られた。
  - RETELLING_I(T) ⇔ RETELLING_M(T)
    文字単位の評価の場合 n-gram(2,≤2,≤3,≤4)_char, p-mer(2,3,4,≤2,≤3,≤4)_char に有意差が見られた。
    形態素単位の評価の場合、n-gram(1,2,≤2,≤3,≤4)_mrph, p-mer(2,3,4,≤2,≤3,≤4)_mrph に有意差が見られた。

  複数人間の評価ではなく、複数回間の評価でも、前項と同じ傾向が見られる。
  再話課題の間については、形態素単位の評価においては、三課題のうちどの二つ組においても有意差が出る傾向にある。口述による再話 (RETELLING_{I,K}) の方が筆述による再話 (RETELLING_M) より一致度が高くなる。また口述による再話においては、自身の体験に基づく再話 (RETELLING_I) の方が、他者から聞いた話の再話 (RETELLING_K) よりも一致度が高くなることが認められた。
- クラウドソーシングにおける課題間の違いの評価
  BCCWJ-SUMM_C ⇔ GROSS_C について、全てのスコアについて、有意差が見られた。
  クラウドソーシングにおける課題間の違いについても、前項と同じ傾向が見られる。
- 要約課題においてクラウドソーシングと実験室との違いの評価 (複数人間)
  BCCWJ-SUMM_C ⇔ BCCWJ-SUMM_L(P) について、n-gram(2)_char, n-gram(3)_char, n-gram(4)_char にのみ有意差が見られた。これは、タイプ入力 (BCCWJ-SUMM_C) と筆述 (BCCWJ-SUMM_L(P)) とで、表記ゆれの統制の差が出たのではないかと考える。
- 語釈課題においてクラウドソーシングと実験室との違いを評価する (複数人間)
  GROSS_C ⇔ GROSS_L(P) について、n-gram(2,3,4)_char, n-gram(2,3,4)_mrph, Footrule_mrph, Kendall_mrph **以外**について有意差が見られた。語釈においては、クラウドソーシングの場合 wikipedia や辞書サイトからのコピーが行われる傾向にある一方、実験室の場合は参照文献なしで筆述で行うために差が出たのではないかと考える。
- 複数人間距離と単一人の回数間距離の違い
  BCCWJ-SUMM_L(P) ⇔ BCCWJ-SUMM_L(T), GROSS_L(P) ⇔ GROSS_L(T), RETELLING_K(P) ⇔ RETELLING_K(T), RETELLING_M(P) ⇔ RETELLING_M(T) について、全てのスコアについて有意差が見られた。基本的に単一人が実施したほうが一致度が高いと考えられるが、統計分析の結果からもそれが確認できる。

4.2 スコア毎の特性

前節の課題間の議論から考えられるスコア毎の特性について論じる。

文字 n-gram はタイプ入力と筆述入力の差として認められることから、表記ゆれレベルで一致度が下がる特性があると考える。形態素 n-gram は再話と繰り返しで顕著に高くなることから、個々人の言い回しや文体などを反映していると考える。

p-mer, Footrule, Kendall などは語順などを反映していると考えるが、情報の提示順が重要な

要約・再話で一致度が高い一方、語釈などにおいては低い傾向にあることがわかった。

n-gram, p-mer ともに *n*, *p* の値が高くなるにつれてスコアが低くなる。このために有意差が出にくくなる傾向にある。n-gram, p-mer ともに *n* (or *p*) 以下のスコアとして設定した場合に、より低い *n* (or *p*) の方が一致の多くなる傾向にあるために、より高い *n* (or *p*) の差異が見られなくなる傾向がある。これはスコアの自然な解釈であると考えられるが、何らかの用途で長い n-gram, p-mer を重要視する場合には加重を行う必要があるだろう。

n-gram(1)_* と Kendall_* とを比較した場合、 n-gram(1)_*では有意差が出るが、順序尺度を入れた Kendall_* では有意差が出ないスコアの組み合わせがいくつかあった。これは文字順・語順の一致度が低い場合に、順序尺度を掛けあわせたがために全体の一致度の差がなくなったことが考えられる。

5. おわりに

本稿では、文書間距離尺度の数理的構造を説明した。カーネル・距離・相関係数とどう対応しているのかを説明し、n-gram 系、p-mer 系、順序尺度の三つに抽象化した。次に様々な言語資源を用いて各指標で用いられているスコアの特性を明らかにした。要約・語釈・再話からなる 7 種類の言語資源を用いて、課題・多人数産出・複数回産出・産出手段 (口述・筆述・タイプ) の軸を用いて、どのような分散が観察されるかを確認した。

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」、科研費若手 (B)「コーパスから取得しやすい情報と取得しにくい情報の研究」、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・今田水穂 (2015).「単一文書自動要約のための言語資源構築に向けて」　情報処理学会研究報告 2015-NL-220 巻.

Shawe-Taylor，John・Nello Cristianini・大北剛 (訳) (2010).『カーネル法によるパターン解析 (Kernel Methods for Pattern Analysis)』, 第 11 章　共立出版.

宮部真衣・四方朱子・久保圭・荒牧英治 (2014).「音声認識による認知症・発達障害スクリーニングは可能か？–言語能力測定システム”言秤”の提案–」　グループウェアとネットワークサービスワークショップ 2014.

神嶌敏弘 (2009).「順序の距離と確率モデル」　人工知能学会研究会資料 SIG-DMSM-A902-07.

平尾努・奥村学・安田宣仁・磯崎秀樹 (2007).「投票型回帰モデルによる要約自動評価法」　人工知能学会論文誌, 22:2, pp. 115–126.

保田祥 (2014).「同じ話を成立させる語 –「桃太郎」を「桃太郎」として成立させる語彙–」　社会言語科学会第 33 回大会発表論文集.

保田祥・荒牧英治 (2012).「人が同じ話を何度もするとどうなるか？：繰り返しによって生じる物語独話の変化」　日本認知科学会第 29 回.

保田祥・田中弥生・荒牧英治 (2013a).「繰り返しにおける独話の変化」　社会言語科学会第 31 回大会発表論文集, pp. 190–193.

保田祥・田中弥生・荒牧英治 (2013b).「同じ話であるとはどういうことか」　社会言語科学会第 32 回大会発表論文集.

BCCWJ における固有表現抽出のエラー分析

市原　正陽（茨城大学工学部　情報工学科）
山崎　舞子（東京工業大学　大学院総合理工学研究科）
古宮　嘉那子（茨城大学工学部　情報工学科）

Error Analysis of Named Entity Extraction in BCCWJ

Masaaki Ichihara(Department of Computer and Information Sciences, Ibaraki University)
Maiko Yamazaki(InterdisciplinaryGraduate School of Science and Engineering,
Tokyo Institute of Technology)
Kanako Komiya(Department of Computer and Information Sciences, Ibaraki University)

要旨

テキスト中に含まれる固有表現を正しく認識することは，自然言語で書かれたテキストに含まれる情報を誤りなく取得するうえで必要である．よって，本研究では「現代日本語書き言葉均衡コーパス」よりランダムサンプリングをしたテキストを京都大学の「日本語構文・格・照応解析システム KNP」にかけ，その結果に含まれるエラーの分析を行った．分析結果から，KNP の固有表現抽出機能が固有表現の抽出を誤るのは，形態素解析や構文解析の誤り，辞書の知識不足が大きな要因と考えられることが分かった．

１．はじめに

固有表現抽出とは，テキストの中から人名や地名，商品名などの固有表現を自動的に抽出する処理である．しかし，誤った情報を抽出することや，本来抽出したい固有表現が抽出できないことがままある．そのため，本稿では，現在の固有表現抽出システムを使用して得られたエラーに対してエラー分析を行う．

２．使用システムおよび使用コーパス

日本語のコーパスとして「現代日本語書き言葉均衡コーパス」（BCCWJ）（Maekawa (2008)）を用いる．システムは固有表現を抽出するために「日本語構文・格・照応解析システム KNP[1]」（KNP）を使用する．KNP では CRF を用いた系列ラベリングに基づいて固有表現の解析を行っている．また KNP では，固有表現抽出を行う際の素性として形態素情報のほかに「キャッシュ素性」や「係り先素性」などを使用している（笹野ら（2008））．

また，本研究では固有表現を分類するために Information Retrieval and Extraction Exercise[2]（IREX）で定義された組織名，人名，地名，固有物名，日付表現，時間表現，金額表現，割合表現，オプショナルの 9 つの固有表現を使用した．

[1] http://nlp.ist.i.kyoto-u.ac.jp/index.php?KNP
[2] http://nlp.cs.nyu.edu/irex/index-j.html

３．BCCWJ における固有表現抽出のエラー分析手法

３．１　BCCWJ における KNP のエラー分析

今回エラーの分析をするにあたって BCCWJ のうち「YAHOO!知恵袋」「白書」「YAHOO!ブログ」「書籍」「雑誌」「新聞」の 6 つからランダムサンプリングした計 136 個のテキストに対して人手によって IREX で定義された 9 つの固有表現タグを付けた．これを正解として比較を行っていく．また， KNP の固有表現の解析を行うオプションである-ne を使うことで，それらのテキストの平文から固有表現タグの付いた平文を出力した．その後，それらの人手と KNP のタグが付けられたテキストのペアを比較することでエラーに対して分析を行った．

３．２　BCCWJ コーパスへの IREX のタグ付け

IREX の固有表現タグの人手による付与は，テキストを 5 分割したものに対して Project Next NLP の NE のタスクのメンバー5 人がそれぞれタグ付けを行った．5 分割したテキスト群のうちの一つを対象とする時にはそれぞれ「hi」「ichi」「iwa」「ko」「ta」とする．

３．３　BCCWJ コーパスにおけるエラー抽出

人の手によってタグの付けられたテキストと KNP によってタグの付けられたテキストの比較を行い，エラーの種類によって分類して分析を行った．

４．BCCWJ における固有表現抽出のエラー結果

４．１　KNP が付与したタグの正解率

表 1 に KNP の付けたタグ全体の正解していた数，不正解していた数と正解の割合を示す．

表 1 固有表現の正解不正解の内訳

	正解	不正解	総数	正解率
hi	297	194	491	60.49%
ichi	195	99	294	66.33%
iwa	303	187	490	61.84%
ko	385	385	770	50.00%
ta	452	319	771	58.63%
総数	1632	1184	2816	57.95%

KNP の付けた固有表現タグは半分以上が人手で付けたものと一致した．

４．２　タグの範囲に対する分析

タグの範囲に対する分類として，以下の 5 種類に分類を行った．

KNP なし：KNP は固有表現として抽出しなかったが，正解は固有表現だったもの
人手なし：KNP は固有表現として抽出したが，正解は固有表現ではなかったもの
範囲別：KNP は固有表現として抽出したが，正解と固有表現の範囲だけが異なっていたもの
タグ別：KNP は固有表現として抽出したが，正解と固有表現の種類だけが異なっていたもの
両方別：KNP は固有表現として抽出したが，正解と固有表現の範囲，種類がともに異なっていたもの

比較方法としては文字位置が人手で付けたタグの範囲よりも KNP が狭い範囲でタグをつけていたもの，人手で付けたタグの範囲よりも KNP が広い範囲でタグをつけていたもの，人手で付けたタグの範囲と KNP が付けたタグの範囲が一部分だけ被っているものは，それぞれ別々のエラーとしてカウントした．

そのため一方では一つの固有表現としてタグが付けられたものが，もう一方では分割されて固有表現としてタグが付けられていた場合，分割されている方の数だけエラーとしてカウントされている．その例を図 1 として以下に示す．

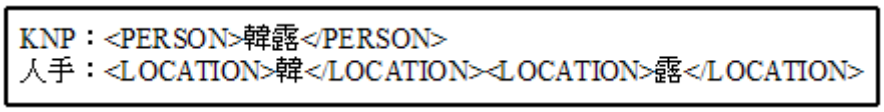


図 1 人手で付けた固有表現が KNP の出力した固有表現の中に 2 つ入っている例

図 1 と同様に KNP の出力した固有表現が人手で付けた固有表現の内側に入っていても，同じように分割されている方をカウントする．

KNP の付けたタグと人手で付けたタグの比較を行った結果を表 2 に示す．

表 2 KNP のエラーの内訳

	KNP なし	人手なし	範囲別	タグ別	両方別	エラー総数
hi	98	33	34	15	14	194
ichi	48	21	16	6	8	99
iwa	133	30	14	3	7	187
ko	212	34	38	72	29	385
ta	128	41	60	31	59	319
総数	619	159	162	127	117	1184

結果から，5 分割したすべてにおいて，KNP がタグをつけられていないエラーの数が最も多く，全体の半分以上のエラーがこれに含まれていた．次に多かったのは，タグは同様のものが付けられているが，付けられている範囲が異なっているものだった．このうち，一部分だけが被っているエラーはごく少数で，その内のほとんどは人手で付けたタグの範囲の方が広かった．

４．３　KNP が誤って付けたタグに対する分析

表 3 には KNP がタグを付けた中で，人手で付けたものと違っていたものの内訳を示す．表 3 にある 8 つの固有表現タグは，KNP によって付けられていた固有表現タグである．

ORG：ORGANIZATION，組織名，政府組織名を表す
PERS：PERSON，人名を表す
LOC：LOCATION，地名を表す
ART：ARTIFACT，固有物名を表す
DATE：DATE，日付表現を表す
TIME：TIME，時間表現を表す
MONEY：MONEY，金額表現を表す
PERC：PERCENT，割合表現を表す

表 3 タグごとの内訳

	ORG	PERS	LOC	ART	DATE	TIME	MONEY	PERC	総数
hi	27	6	19	14	30	0	0	0	96
ichi	8	34	3	3	3	0	0	0	51
iwa	22	5	16	6	1	0	2	2	54
ko	31	37	76	9	20	0	0	0	173
ta	35	52	40	35	29	0	0	0	191
総数	123	134	154	67	83	0	2	2	565

この結果から，「TIME」「MONEY」「PERCENT」に関しては，KNP は間違って固有表現タグを付けることが少ないことがわかる．また，「ARTIFACT」や「DATE」に関しても誤っているものがあるが，合わせて KNP が誤って固有表現タグを付けたもののうち 3 割に満たなかった．そして，KNP が固有表現タグを付けた誤りのうち「ORGANIZATION」「PERSON」「LOCATION」の 3 つが，誤りの大部分を占めていることが分かった．

５．KNP が固有表現タグを付与できなかったエラーに対する分析

表 2 から分かるように KNP が固有表現のタグを付ける際に出るエラーの中で最も数が多いのは，KNP が固有表現のタグを付けられないエラーだったため，それに関して分析を行った．

５．１　各タスクのエラーの割合

今回エラーを取得するために使用したテキストは BCCWJ のコアデータである「OC」「OW」「OY」「PB」「PM」「PN」の 6 つで，それぞれ「YAHOO!知恵袋」「白書」「YAHOO!ブログ」「書籍」「雑誌」「新聞」の 6 つのタスクから取得されたものである．それらのタスクごとのエラーの割合を表 4 に示す．

タグ無：KNP がタグを付けなかったエラーの数
タグ有：KNP がタグを付けたエラーの数（範囲の間違い，タグの間違いも含む）
タグ無割合：不正解の合計数に対する KNP がタグを付けなかったエラーの割合

表 4 タスクごとのエラーの割合[3]

all	正解	タグ無	タグ有	合計	不正解の合計	タグ無割合	文書数
YAHOO!知恵袋	76	84	30	190	114	73.68%	74
白書	427	150	150	727	300	50.00%	8
YAHOO!ブログ	171	94	72	337	166	56.63%	34
書籍	217	121	93	431	214	56.54%	5
雑誌	186	51	111	348	162	31.48%	2
新聞	555	119	94	768	213	55.87%	13
合計	1632	619	550	2801	1169	52.95%	136

[3] 表 3 ではタグの付けられたエラーの総数が 565 個だったものが表 4 では 550 個になっているのは，表 1 では人手と KNP 両方からみたエラーの数を表おり，表 4 では KNP のエラーに関してのみ注目しているため．

表 4 で文書数と合計数に比例関係がないのは，一つの文書内にある文字数がジャンルによって大きく異なるためである．また，それぞれのジャンルの内「YAHOO!知恵袋」が最も不正解の中でタグを付けられないエラーの割合が多く，逆に「雑誌」が一番タグを付けられないエラーの割合が低かった．

５．２　各タスクの正解率

「YAHOO!知恵袋」「書籍」「YAHOO!ブログ」「書籍」「雑誌」「新聞」それぞれの正解率と全体の合計に対するタグ無の割合を表 5 に示す．

タグ無割合：正解，不正解両方の合計数に対する KNP がタグを付けなかったエラーの割合

表 5 タスクごとの正解率とタグ無の割合

all	正解率	タグ無割合	精度	再現率	F 値
YAHOO!知恵袋	40.00%	44.21%	71.70%	43.93%	54.48%
白書	58.73%	20.63%	74.00%	63.35%	68.27%
YAHOO!ブログ	50.74%	27.89%	70.37%	55.70%	62.18%
書籍	50.35%	28.07%	70.00%	52.54%	60.03%
雑誌	53.45%	14.66%	62.63%	57.76%	60.10%
新聞	72.27%	15.49%	85.52%	73.80%	79.23%
合計	58.26%	22.10%	74.79%	61.79%	67.68%

表 5 から分かるように「新聞」の正解率が一番高かった．また「YAHOO!知恵袋」の正解率が一番低く，そのほかのタスクの正解率はその 2 つと比べると，正解率の差は少なかった．「新聞」の正解率が一番高かったのは，KNP は毎日新聞データを訓練事例としているためだと考えられる．また，「YAHOO!知恵袋」のタスクが 6 つのタスクの中で最も正解率が低いのは，新聞と文体が遠いからではないかと考えられる．また，正解，不正解の内のタグ無の割合は「雑誌」の割合が最も低く，「YAHOO!知恵袋」の割合が最も高かった．

５．３　固有表現タグの付けられなかった形態素の分析

表 5 の正解率から，最も割合の低かった「YAHOO!知恵袋」と最も割合の高かった「新聞」に含まれる形態素に対して分析を行った．

５．３．１　「YAHOO!知恵袋」内の固有表現タグの付けられなかった形態素の分析

ⅰ.商品名やキャラクター名が取れない事が多い．

実際に取れなかった商品名やキャラクター名，薬品名の一部

・サクラ大戦　・スーパーファミコン　・アクトレイザー　・バイオハザード 4

・仮面ライダー　・ウルトラマン　・ガンダム　・ミノスタシン　・アスピリン

ⅱ.略されたものが取れない．

ⅰの影響が強いのかもしれないが，略された商品名も取れていない．

・スーパーマリオワールドは取れてマリオワールドは取れない

・GC(ニンテンドーゲームキューブ)　・JNB(ジャパンネット銀行)　・LA(ロサンゼルス)

ⅲ.特殊な日付の表現が取れない．

・九十／十一／二十一

ⅳ.ひらがなで表記されていると誤って解析してしまう

”知恵ぶくらー・さとし”と記述されたファイルがあり，本来”さとし”は PERSON と取って欲しいのだが，動詞の”悟る”として解析されていた.

ⅴ.略称でなくてもアルファベットやアラビア数字と組み合わさったものが取れない

・ＰＳ２　・ＩＳＤＮ　・ＪＲ（ＪＲ西となった部分は正しく取れていた）

・Ｏｕｔｌｏｏｋ　Ｅｘｐｒｅｓｓ

５．３．２　「新聞」内の固有表現タグの付けられなかった形態素の分析

Ⅰ.基本的に取れないものがある

・半～(時間表現など様々)　・～圏(首都圏，三大都市圏)　・～地域　・～ポイント

・同～(同～年，同日，同年秋)

半日や首都圏，ユーロ地域などが誤りとして確認でき，正解には含まれていなかった.

ただし，半分は PERCENT として取得できていた.

Ⅱ.英語や日本語などを OPTIONAL として取れなかった.

本来「<OPTIONAL>英</OPTIONAL>語」「<OPTIONAL>日本</OPTIONAL>語」のように取れてほしい．しかしそもそも KNP の機能として OPTIONAL と付ける機能はない.

Ⅲ.英語表記で書かれることが少ないものが取れなかった

・KOERA　・JAPAN

Ⅳ.付近にその形態素に関する情報があっても（があると取れなかった.

・【フェニックス（<LOCATION>米アリゾナ州</LOCATION>）

Ⅴ.一般名詞やそれが組み合わさったようなものは取れないことが多かった.

ｉ（商品名やキャラクター名が取れないことが多い）の原因も同様である可能性がある

・昼寝　・ザウルス　・ファミリーマート　・シャープ　・ルネサンス

(ソフトバンクが取れている所と取れていないところがあった．取れているものはガ格に，取れていないものは文節内と解析されていた．)

６．考察

分析から，KNP の固有表現抽出機能が固有表現の抽出を誤るのは，形態素解析や構文解析の誤り，辞書の知識不足が大きな要因と考えられる．特に固有物名(ARTIFACT)は商品名などが対象となるため，他の固有表現より造語が分類されやすく，その場合一般名詞の組み合わされたパターンが分類される可能性が高いと考えられる．そのため KNP の場合先行文脈やその単語に対する係り受けの関係などからその単語が固有表現なのか推察しなければならず，正しい構文解析は重要である.

また，構文解析するにあたっても新聞などより口語的なものを扱う可能性も十分あり，そういった場合，助詞が抜けている事などが構文解析の妨げとなる事は多いと推察できる.

そのため，新聞とは書かれ方の大きく異なる文書からも学習することで，特定ジャンルでない文書から固有表現を抽出しようとする場合効果的である可能性が高い．また，取ることのできなかった固有表現の大半が wikipedia などネット上に情報があることが確認できたため，それらを辞書に取りこむことでより正確な固有表現抽出の実現が期待できる.

謝辞

本研究は，文部科学省科学研究費補助金[若手 B(No:24700138)]の助成により行われました．ここに，謹んで御礼申し上げます．

また，KNP についての質問に快く答えてくださった，東京工業大学の笹野遼平先生に謹んで御礼申し上げます．

また，Project Next NLP の NE 班の班長である岩倉友哉先生をはじめ，班員の皆様方には多くのご協力をいただきました．謹んで御礼申し上げます．

参考文献

[1]笹野遼平，黒橋禎夫(2008)「大域的情報を用いた日本語固有表現認識」情報処理学会論文誌,Vol.49No.11,pp.3765-3776

[2]笹野遼平，河原大輔，黒橋禎夫，奥村学(2013)「構文・述語項構造解析システム KNP の解析の流れと特徴」言語処理学会,第 19 回年次大会　発表論文集,pp.110-113

[3] Kikuo Maekawa(2008). Balanced corpus of contempo-rary written japanese. In ALR 2008, pp. 101-102

付録

今回対象としたBCCWJのコアデータ内の136ファイル

YAHOO! 知恵袋	OC01_00001	OC01_00002	OC01_00003	OC01_00004	OC01_00005
	OC01_00006	OC01_00007	OC02_00001	OC02_00002	OC02_00003
	OC02_00004	OC02_00006	OC02_00007	OC02_00008	OC03_00001
	OC03_00005	OC04_00001	OC04_00002	OC04_00003	OC05_00001
	OC05_00003	OC05_00004	OC05_00006	OC06_00001	OC06_00008
	OC08_00001	OC08_00002	OC08_00004	OC08_00006	OC09_00001
	OC09_00002	OC09_00003	OC09_00004	OC09_00006	OC09_00008
	OC10_00001	OC10_00003	OC10_00005	OC10_00006	OC10_00007
	OC11_00001	OC11_00002	OC11_00004	OC11_00005	OC11_00006
	OC11_00007	OC12_00002	OC12_00003	OC12_00004	OC12_00005
	OC12_00006	OC12_00007	OC12_00008	OC13_00001	OC13_00002
	OC13_00003	OC13_00004	OC13_00005	OC13_00006	OC13_00007
	OC13_00008	OC14_00001	OC14_00003	OC14_00004	OC14_00005
	OC14_00006	OC14_00007	OC14_00008	OC15_00001	OC15_00002
	OC15_00004	OC15_00006	OC15_00007	OC15_00008	
白書	OW6X_00000	OW6X_00002	OW6X_00003	OW6X_00007	OW6X_00008
	OW6X_00009	OW6X_00011	OW6X_00013		
YAHOO! ブログ	OY01_00082	OY01_00137	OY01_00148	OY01_00185	OY02_00095
	OY04_00001	OY04_00027	OY04_00173	OY06_00060	OY06_00146
	OY06_00168	OY07_00097	OY07_00135	OY07_00164	OY08_00115
	OY08_00137	OY08_00156			
書籍	PB11_00006	PB12_00001	PB22_00002	PB43_00001	PB59_00001
雑誌	PM11_00002	PM24_00003			
新聞	PN1a_00002	PN1d_00001	PN1d_00002	PN1f_00002	PN1g_00002
	PN2c_00002	PN2g_00002	PN3b_00001	PN3c_00002	PN4b_00001
	PN4c_00001	PN4c_00002	PN4f_00001		

機械翻訳を用いた中古和文の現代語訳ー分析と課題ー

山田 祐実　大村 舞　岡 照晃　Kevin Duh　松本 裕治

(奈良先端科学技術大学院大学)

Translation of Classical Japanese into Contemporary Japanese Using MT: Analysis and Future Work

Yumi Yamada, Mai Omura, Teruaki Oka, Kevin Duh, Yuji Matsumoto

(Nara Institute of Science and Technology)

要旨

国立情報学研究所による人工頭脳プロジェクト「ロボットは東大に入れるか」において，機械翻訳による古語の現代語訳が行われており，翻訳モデルの学習に平安期から江戸期にわたる古語のコーパスが使われている．しかし，時代によって用法の異なる語がある場合，他の時代の文を翻訳する際に適切な訳語が当てられない可能性がある．また，使用した小学館コーパスには他の作品と比べ敬語表現の多い『源氏物語』が約 55% 含まれるという特徴があった．そこで，学習に使用するコーパスを中古和文に絞り，『源氏物語』の文体が言語モデルへ及ぼす影響を下げるため，BCCWJ や青空文庫によるコーパスを加え翻訳を行った．その結果，翻訳性能の向上が見られた．翻訳結果を分析すると，BLEU による評価方法の見直しや訳語の対応関係の改善が今後の課題となることが分かった．

1　はじめに

現在国立情報学研究所では，現時点での人工知能の達成度と課題を測る試みとして，人工頭脳プロジェクト「ロボットは東大に入れるか」を進めている [新井ら 2012]．横野らは，国語の古文問題の解答に取り組んでおり [横野ら 2014]，内容理解に関する問いを解くために統計的機械翻訳を用いて古文から現代文への翻訳を行っている [星野ら 2014]．

統計的機械翻訳は，図 1 のように翻訳モデルと言語モデルを用いて行なわれる．星野らは翻訳モデルと言語モデルをつくるのに，本研究と同様に小学館『新編日本古典文学全集』によるコーパス（小学館コーパス）を用いている．しかしながら，星野らが用いたコーパスには平安期から江戸期にかけての幅広い作品が含まれている．このため，同じ表層形でも時代によって意味の異なる語がある場合，ある時代でよく用いられる意味に高い確率が付与されると，他の時代の文を翻訳する時に適切な訳を当てられない可能性がある．また，言語モデルの学習には小学館コーパスのみを使用し，他のコーパスを使用する試みは行っていない．

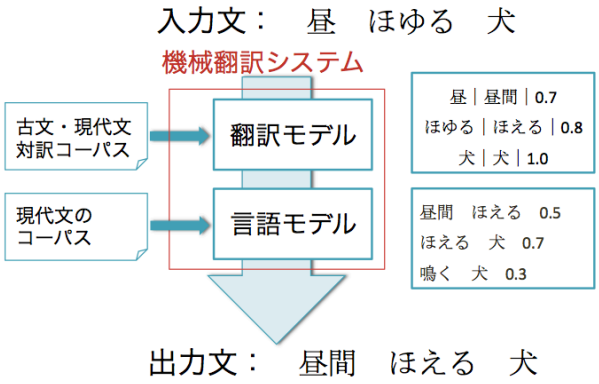


図 1: 統計的機械翻訳の概略

そこで本研究では，中古和文のコーパスを対象に翻訳を行った．さらに，言語モデルの学習用コーパスを複数試し，翻訳結果への影響を調べ，翻訳結果に見られる問題点を分析した．

以下，2 章で統計的機械翻訳について述べる．3 章では今回実験に使用したコーパスについて述べ，4 章では実験設定について説明する．5 章で翻訳の性能と実際の翻訳例を提示し，6 章では 1 文ごとの評価値を調べ，分析を行う．7 章と 8 章で翻訳例を踏まえた今後の課題について述べ，最後に 9 章で本稿のまとめを行う．

2　統計的機械翻訳について

統計的機械翻訳を古文から現代文への翻訳に使う場合，図 1 に示したように，計算機を用いてコーパスから翻訳モデルと言語モデルを生成し，古文の入力に対して適切な現代語の翻訳文を出力するシステムをつくる．翻訳モデルは，単語列間の翻訳関係に確率を付与したものである．この関係と確率が対応づいた表をフレーズテーブルとよぶ．また，言語モデルは，出力文の文としての自然さを確率で評価するものである．出力の際，翻訳候補の文の中から翻訳モデルの確率と言語モデルの確率の積が最も高いものが選ばれる．翻訳モデルは対訳コーパスを用いて作られ，言語モデルは出力言語のコーパスからコーパス内の統計情報をもとに作られる．この過程を一般に翻訳モデルの学習，及び言語モデルの学習とよぶ．

3　使用したコーパス

今回使用したコーパスは，小学館『新編日本古典文学全集』から平安期を中心とした 14 作品，現代日本語書き言葉均衡コーパス（BCCWJ）[Maekawa2008]，青空文庫与謝野晶子訳『源氏物語』[*1] の 3 種類である．

[*1] 青空文庫 与謝野晶子訳『源氏物語』http://www.aozora.gr.jp/index_pages/person52.html

日本霊異記，竹取物語，古今和歌集，土佐日記，伊勢物語，大和物語，落窪物語，平中物語，枕草子，和泉式部日記，源氏物語，紫式部日記，更級日記，讃岐内侍日記，堤中納言日記，蜻蛉日記，大鏡，今昔物語集，将門記，陸奥話記，保元物語，平治物語，方丈記，徒然草，正法眼蔵随聞記，歎異抄，平家物語，宇治拾遺物語，十訓抄，沙石集，曾我物語，近松門左衛門集，洒落本，滑稽本，人情本，俊頼髄脳，古来風躰抄，近代秀歌，詠歌大概，毎月抄，国歌八論，歌意考，新学異見

図 2: 小学館コーパス

表 1: 小学館コーパスの単語数比較

	古文	現代文	計（単語）
星野らの用いた単語数	2,837,101	3,720,257	6,557,358
本研究で用いた単語数	1,071,453	680,464	1,751,917

3.1 新編日本古典文学全集

小学館コーパスに含まれる作品を図 2 に示す．星野らは図中の全ての作品を用いて翻訳モデルと言語モデルを作成したが，平安期から江戸期までの広い時代の言葉が含まれているため，フレーズテーブルから適切な訳語が選ばれにくくなる可能性がある．そこで本研究では，図中の下線で示している中古和文で書かれた 14 作品のみを用いて翻訳モデルと言語モデルを作成した．星野ら及び本実験で使用した小学館コーパスの単語数を表 1 に示す．

本研究で用いた中古和文の 14 作品には「源氏物語」が約 55% を占めているという特徴がある．「源氏物語」の現代語訳は他の 13 作品と比べて敬語表現が多いため，この特徴が言語モデルに影響する可能性が考えられる．「源氏物語」と他の 13 作品の文体の違いを図 3 に示す．「源氏物語」の文には「なさる」や「いらっしゃる」といった敬語表現がよく用いられる．この結果，統計的機械翻訳の評価尺度である BLEU の値が悪くなると予測した．そこで，BCCWJ のコアデータ 58,355 文を言語モデルの学習用コーパスに加え，出力文体への「源氏物語」の影響力を抑えて他の 13 作品の翻訳精度を上げられるかどうか実験を行った．

3.2 現代日本語書き言葉均衡コーパス

言語モデルにおける小学館コーパスの「源氏物語」の影響を押さえるため，BCCWJ からコアデータ 58,355 文を言語モデルの学習に使用した．これは「源氏物語」9,752 文の約 6 倍の規模である．コアデータは，書籍，雑誌，新聞，白書，Yahoo!知恵袋，Yahoo!ブログから構成される．

3.3 青空文庫，与謝野晶子訳『源氏物語』

青空文庫の与謝野晶子訳「源氏物語」17,648 文も言語モデルの学習に使用した．図 3 に示したように，青空文庫の「源氏物語」の方が小学館の「源氏物語」現代語よりも他の 13 作品の文体に近いため，小学館の「源氏物語」を青空文庫の「源氏物語」に差し変えて言語モデルを学習し，翻訳を行った．

13 作品
楊貴妃が、玄宗皇帝の御使者に会って、泣いた顔にたとえて、「梨花一枝、春、雨を帯びたり」などと言ったのは、並一通りではあるまいと思うにつけて、やはりとてもすばらしい点では、他に類があるまいと感じられる。 少しお粥などをさしあげたところ、お召し上がりになりなどしたが、そのうれしさは何にたとえようもない。 耳敏川、これは、またも何をそんなに聞き耳をたてて聞きとったのだろうと、おもしろい。
源氏物語
下草のあれこれ美しく咲いている花々や紅葉などを 手折らせなさって、女二の宮の お目にかける 手土産に なさる。 大殿は廂の御簾の中に いらっしゃる ので、式部卿宮と右大臣だけが おそばにお控えになり、それ以下の上達部は簀子に居並んで、今日は正式の御賀の日ではないので、ご馳走などはそう仰々しくはなくお出ししてある。 源氏の君は、山里の人にも、久しく無沙汰のまま お過しだったことをお思い出しになり、わざわざお使者を お差し向けになった ところ、僧都の返事だけが寄せられる。
青空源氏
林の下草の美しい花や、紅葉を折らせた薫は夫人の宮にそれらを お見せした。 縁側に近い御簾の中に院のお席があって、そこにはただ式部卿の宮が 御同席され、右大臣の陪覧する座があっただけである。以下の高官たちは皆縁側に席をして、そこには形式を省いた饗応の物が出されてあった。 それで源氏の君も多忙であった。北山の寺へも久しく見舞わなかったことを思って、ある日わざわざ使いを立てた。山からは僧都の返事だけが来た。

図 3: 13 作品，源氏物語，青空源氏の文体の違い

表 2: 言語モデルの学習に使用したコーパス

言語モデル	訓練データ	開発データ	評価データ	計（文）
13 作品 + 源氏物語（ベースライン）	17,715	2,211	2,211	22,137
13 作品	7,963	996	996	9,955
源氏物語	9,752	1,215	1,219	12,186
青空源氏	17,648	-	-	17,648
13 作品 + 青空源氏	25,611	-	-	25,611
13 作品 + 源氏物語 +BCCWJ	80,292	-	-	80,292
13 作品 + 源氏物語 + 青空源氏	35,363	-	-	35,363

4　実験設定

本実験では，3 章で述べたコーパスを用いて複数通りのパターンで言語モデルを学習し，古文の現代語訳を行った．翻訳モデルの学習には，対訳になっている小学館コーパスのみを使用した．小学館コーパスは古文とその現代語訳が各作品で段落ごとに対応づいているが，統計的機械翻訳においては一文ごとに対応づいていることが望ましい．そこで，Gale らの方法を用いて一文ごとの対応づけを行った [Gale&Church1993]．実験を行った言語モデルの作成に使ったコーパスの組み合わせを表 2 に示す．

小学館コーパスは，古文・現代文ともに，訓練データ，評価データ，開発データとして 8:1:1 の割合で分割した．訓練データは翻訳モデルと言語モデルを作るのに使用した．言語モデルを学習する際に複数のコーパスを用いる場合，線形補間で複数の言語モデルを組み合わせた．評価データは古文を翻訳システムの入力とし，現代文は出力文の評価で正解データとして使用し

表 3: 実験結果 BLEU 値

	評価用コーパス			
学習用コーパス	13 作品 + 源氏物語	13 作品	源氏物語	小学館（星野ら）
13 作品 + 源氏物語（ベースライン）	22.38	24.81	20.21	-
13 作品	21.09	25.41	17.94	-
源氏物語	20.88	22.71	19.88	-
青空源氏	20.11	22.84	17.61	-
13 作品 + 青空源氏	22.46	24.98	20.24	-
13 作品 + 源氏物語 +BCCWJ	22.41	24.95	20.35	-
13 作品 + 源氏物語 + 青空源氏	21.61	25.55	18.46	-
小学館（星野ら）	-	-	-	28.02

た．開発データは翻訳システムにおける各種パラメータのチューニングに使用した．表中の「13 作品」は「源氏物語」を除いた小学館コーパス中の平安文学 13 作品を，「源氏物語」は小学館コーパスの「源氏物語」を指す．

コーパスの分かち書きには MeCab v0.98 [Kudo et al.2004]，辞書には中古和文 UniDic v1.4 [小木曽ら 2010] 及び UniDic v2.1.2[伝ら 2007]，単語アライメントには GIZA++ v1.0.7[Gao&Vogel2008] を用いた．統計的機械翻訳のツールは Moses v0.91[Koehn et al.2007] を用い, distortion limit は 0 とした．翻訳の際にはエラー最小化学習を用いてパラメータのチューニングを行った．翻訳結果の評価尺度には，翻訳結果と正解語の一致率で翻訳精度を測る BLEU[Papineni et al.2011] を使用した．

5　実験結果

5.1　BLEU スコアの評価

小学館コーパス，BCCWJ，青空文庫の 3 種類のコーパスを 6 通り組み合わせて言語モデルを学習し，古文を現代文へ翻訳した．翻訳結果および星野らの BLEU スコアを表 3 に示す．ただし，星野らは言語モデルと翻訳モデルの学習に図 2 の小学館コーパス全ての作品を用いて翻訳を行っていることに注意してほしい．

出力の評価には，正解文との比較で単語 n-gram の一致度を測る BLEU と呼ばれる評価尺度を用いた．BLEU は出力文に含まれる単語が正解文に含まれる単語と一致しているほど高いスコアを与える．言語モデルによって正解文に近い文体が出力できれば，BLEU も上がると考えられる．

表中の「学習用コーパス」は，言語モデルの学習に用いたコーパスを指す．「評価用コーパス」は，翻訳の入力に用いた評価データの古文のコーパスを指す．言語モデル学習用コーパスに「13 作品 + 源氏物語」を用いた場合をベースラインとして示す．ベースラインでは「13 作品」を翻訳した際に最も評価値が高くなった．

「13 作品」を翻訳したとき，言語モデル学習用コーパス「13 作品 + 源氏物語 + 青空文庫」で最も BLEU が高くなった．いずれの評価データを翻訳した場合も，学習用コーパスに「13 作品 + 青空文庫」や「13 作品 + 源氏物語 +BCCWJ」を用いたときにベースラインより BLEU が

古文	いといみじき心地しけり。
現代文および翻訳結果	ほんとにどうしようもない気持がした。
古文	「などてかくなくぞ」といへど、いらへもせず。
現代文および翻訳結果	「どうしてこのように泣くのか」といっても、返事もしない。
古文	今日いかにまれ、このことを定めてむ。
現代文および翻訳結果	今日どうあってもこのことを決めてしまいましょう。

図 4: 全ての言語モデルで正解データと同じ文に翻訳できた例

古文	その夜は、くろとの浜といふ所にとまる。
現代文	その夜は、黒戸の浜という所に泊った。
翻訳結果	その夜は、また、この美しい 黒戸の浜という所にとまる。
古文	雨降らぬ日、張り筵したる車。
現代文	雨の降らない日に、筵のおおいを掛けた牛車。
翻訳結果	雨は降らない日、張り筵をしている車 をしたのである。

図 5: ①古語と現代語の対応が不適切な例

高くなった．「13 作品」を翻訳した際，ベースラインと比べ「源氏物語」を除いた「13 作品」では 0.6 ポイント上がり，「13 作品 + 青空文庫」，「13 作品 + 源氏物語 +BCCWJ」，「13 作品 + 源氏物語 + 青空文庫」など「源氏物語」の影響を抑えた学習用コーパスを用いたときは 0.14〜0.74 ポイント上がるといったことから，「源氏物語」が「13 作品」の翻訳精度を下げていたと言える．いずれの結果も星野らの BLEU 値と比較して 2.47 ポイント以上低くなっているのは，表 1 で示したように翻訳モデルの学習に用いたコーパスの量が少なかったことが考えられる．ただし，6 章で示すように，BLEU 値では翻訳の性能を測りきれないため，一概に翻訳の性能が劣ったと言い切ることはできない．

5.2　翻訳のうまくいった例，うまくいかなかった例

本章では，評価データ 13 作品を翻訳した結果，正解データと同じように翻訳できた例と正解データとは違う翻訳となった例を示す．まず，どの学習用コーパスでも正解データと同じ訳に翻訳できた例を図 4 に示す．

逆に，評価用データの 13 作品を翻訳して翻訳が正解データと異なる例について①古語と現代語の対応が不適切な例，②主語や目的語など古語で省略されているが現代語では補足されている例，③ある古語に対して正解データの現代語とは表層形が異なるが同義の語が当てられている例，④同じ表層形でも違う意味（語義曖昧性）を持つ例，の 4 種類に分類し，図 5 から図 8 に示す．図 5 の上の例では，入力文である翻訳元の古文にはない「また、この美しい」という句が翻訳結果に出てきている．下の例は，文末に「をしたのである」という句が表出している．これは，フレーズテーブルに不適切な翻訳の対応が多くあることが原因である．図 6 は，古文で主語や目的語などの語が省略されているが，正解データの現代文では補われているために翻訳結果が正解データと完全には一致しない例である．図 7 の 1 つ目と 2 つ目の例は，正解データの現代文と翻訳結果とで意味はほぼ同じだが表層形が異なるものの例である．2 つ目の例では，古文の「あやしき」が正解データの「奇異な」ではなく「不思議な」に訳されている．図 7

古文	御火取に、ひと日の薫物とうでて、こころみさせたまふ。
現代文	中宮さまは、香炉に、先日の薫物を土中から取り出させてお入れになり、出来具合を ためしてごらんになる。
翻訳結果	御香炉には、一日の薫物をとうでになられて、ためしにおさせになる。
古文	それをと思ふなりけり。
現代文	その人をぜひ と思うのだった。
翻訳結果	それをと思うのであった。

図 6: ②正解データに補足語がある例

古文	むかし、二条の后に 仕うまつる 男ありけり。
現代文	昔、二条の后に お仕えする 男がいた。
翻訳結果	昔、二条の后に お仕えしている 男がいた。
古文	その花のなかに、あやしき 藤の花ありけり。
現代文	その花の中に、奇異な 藤の花があった。
翻訳結果	その花の中に、不思議な 藤の花があるのだった。
古文	河 は飛鳥川。
現代文	河 は飛鳥川。
翻訳結果	川 は飛鳥川。

図 7: ③表層形が異なる例

古文	むかし、男、狩の使よりかへり来けるに、大淀の わたり に宿りて、斎の宮の わらはべ にいひかけける。
現代文	昔、男が、狩の使いから帰ってきた時に、大淀の 渡し場 に泊って、斎宮の御殿に奉仕する 童女 に歌を詠みかけた。
翻訳結果	昔、男が、狩の使いから帰ってきたので、大淀の あたり に泊って、そのままかの 子供 に言葉をかけたのであった。

図 8: ④語義曖昧性の問題がある例

の 3 つ目の例は，異なる漢字が対応してしまった例である．図 8 は同じ語でも複数の意味を持つ場合，正解データと異なる意味の語が訳語に当てられた例である．「わたり」には「渡し場」と「あたり」の両方の意味があり、「わらはべ」は文脈により「童女」や「子供」になり得る．

6　1 文ごとの BLEU 評価

実際にどのような翻訳結果の文が BLEU を下げているのか確認するため，ベースラインで「13 作品 + 源氏物語」2211 文を翻訳し，1 文ずつ BLEU で評価した．この結果の分布を図 9 に示す．BLEU は 0 点から 100 点の値で評価を行う．この値は単純に表 3 の全体の BLEU 値と比較することはできない．表 3 に示したような通常用いられる BLEU は 1 文ごとではなく文章全体で算出するためである．BLEU を 1 文ずつ算出する場合，1 文に含まれる単語の数に対して評価データに含まれる単語がマッチする数を計算するため，1 文が短い場合，不当に BLEU が下がることがある．しかしながら，今回はどのような翻訳結果が BLEU を下げているかといった大まかな傾向を考察するためにこの方法を用いる．

図 9 で横軸は 0 点から 100 点まで 10 点ごとに刻んだ BLEU 値を表し，縦軸は各 BLEU 値における文数の分布の割合を表す．

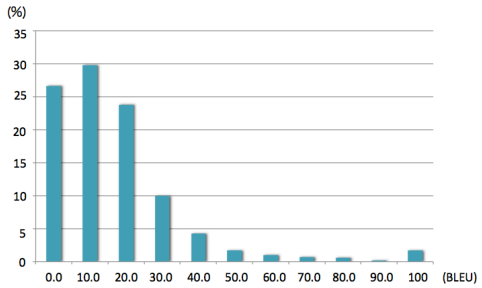


図 9: 1 文ごとの BLEU 値の分布

、 ||| 、 「 それ ごらん なさい 。 ||| 0.25 0.612571 1.231e-05 3.99231e-15 2.718 ||| ||| 4 81235
、 ||| 、 お 答え 申しあげ た 歌 。 ||| 1 0.612571 1.231e-05 2.8906e-16 2.718 ||| ||| 1 81235
、 ||| 、 こう お 答え に なる 。 ||| 0.333333 0.612571 1.231e-05 5.53684e-15 2.718 ||| ||| 3 81235
、 ||| 、 こう ロずさん だ 。 ||| 1 0.612571 1.231e-05 1.64564e-12 2.718 ||| ||| 1 81235
、 ||| 、 こう 書き送っ た 。 ||| 1 0.612571 1.231e-05 9.46122e-13 2.718 ||| ||| 1 81235
、 ||| 、 こう 言い送っ た 。 ||| 0.5 0.612571 1.231e-05 1.42811e-12 2.718 ||| ||| 2 81235
、 ||| 、 こう 言わ れ た 。 ||| 0.333333 0.612571 1.231e-05 7.64404e-14 2.718 ||| ||| 3 81235
、 ||| 、 こう 言わ れ た 。 まして ||| 0.333333 0.612571 1.231e-05 1.19935e-17 2.718 ||| ||| 3 81235
、 ||| 、 こう 詠ん だ 。 ||| 0.473684 0.612571 0.00011079 2.17605e-11 2.718 ||| ||| 19 81235
、 ||| 、 こう 詠ん だ 。 いか ||| 0.25 0.612571 1.231e-05 3.18139e-15 2.718 ||| ||| 4 81235
、 ||| 、 こう 詠ん だ 。 いか で ||| 0.25 0.612571 1.231e-05 7.738e-17 2.718 ||| ||| 4 81235
、 ||| 、 こう 詠ん だ 。 近江 ||| 1 0.612571 1.231e-05 4.04746e-16 2.718 ||| ||| 1 81235
、 ||| 、 こう 詠ん だ 。 近江 なる ||| 1 0.612571 1.231e-05 1.69815e-18 2.718 ||| ||| 1 81235

図 10: フレーズテーブル：対訳の不適切な対応例

図 9 に示した BLEU 値の分布を見ると，50 点台から 100 点台のものが少なく，0 点台から 20 点台に分布する文数が全体の約 80% を占めていることが分かる．BLEU 値ごとに翻訳結果を見ると，60 点台までは元の古文と現代文の間に対応のない語があるために訳せなかったものや，送り仮名や漢字といった表記の違いによるもの，同じ古語に正解データと異なる表層形の現代語が当てられたものが原因で BLEU が下がっている場合が多いことが分かった．対応のある語が訳せているならば翻訳自体はできていると見なせること，また，表記の違いや似た意味の語が翻訳結果に選ばれることは文の大まかな意味を知るためであれば十分な訳といえることから，BLEU による評価方法を見直す必要があると考える．

0 点台から 20 点台を見ると，上記の問題に加え，古文と現代文の評価データが 1 文ずつ正確な対応がとれていないものも多く見受けられた．他には，訳語に不必要な対応が付いているものや，文脈にふさわしくない訳語が選択されていることも BLEU を下げる原因であった．これらは翻訳として不都合であるため，翻訳の過程で改善する必要がある．

次に，学習用コーパスによって翻訳結果に文体や訳語の違いが見られた例を図 11 に示す．この例は，表 3 の BLEU 値と 1 文ごとの BLEU 値に相関の見られたものである．この例でも不適切な語の対応や異なる表層形の語など，上に挙げたような翻訳の問題が見られる．表 3 でも BLEU の低かった「源氏物語」や「青空文庫」で，文頭の「が」をはじめとした不要な対応の他に，「大人ごとに」が「そのうちの年輩ごとに」となっているなどの不適切な対応がある．表 3 で最も BLEU の高かった「13 作品 + 源氏物語 + 青空文庫」では，「這ひ来る」に不適切な訳語が対応していたり，「ほど」や「ごと」など不要な語の表出があるが，ベースラインや他の例と比較すると不適切な対応語の長さが短くなっているなど，全体的な改善が見られる．

7　評価に BLEU を用いる問題点と解決策

6 章で見たように，正解データの現代文にあって古文にない語が翻訳されないために BLEU が下がるという問題点がある．また，現代文の正解データと表層形の異なる似た意味の語が翻訳結果に選ばれた場合，翻訳文としての意味が自然であっても BLEU が下がってしまう．6 章で BLEU の大まかな傾向は妥当であるといえることが分かったが，これらの問題点に対処するためには BLEU による評価を見直す必要がある．たとえば，語の省略に頑強な評価方法として，正解データとの一致率に関する制約を緩めることが考えられる．また，同じような意味の語の評価に関しては，評価における正解を 1 つに絞らないといった対策が考えられる．

正解データ	
古文	二つ三つばかりなるちごの、いそぎて這ひ来る道に、いと小さき塵のありけるを、目ざとに見つけて、いとをかしげなる指にとらへて、大人ごとに見せたる、いとうつくし。
現代文	二歳か三歳ぐらいの幼児が、急いで這って来る道に、とても小さいごみのあったのを、目ざとく見つけて、とても愛らしげな指につかまえて、大人たちに見せているのは、とてもかわいらしい。
翻訳結果	
13+ 源氏	二つ三つぐらいの幼児が、急いでということになりてくる途中、とても小さいの塵ほどのあったのを、目ざとに目をおつけにて、とてもかわいらしげな指にとって、大人ごとにを見せているのは、とてもかわいらしい。
13	二つ三つぐらいの幼児が、急いでこそこそてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
源氏	が二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしいの指につかまえて、そのうちの年輩ごとにを見せている、それがまことにかわいらしい。
青空	が二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしいの指につかまえて、そのうちの年輩ごとにを見せているの、それがまことにかわいらしい。
13+ 青空	二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとに目をおつけにて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
13+ 源氏 +BCCWJ	二つ三つぐらいの幼児が、急いでということになりてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。
13+ 源氏 + 青空	二つ三つぐらいの幼児が、急いでそっとてくる途中、とても小さいの塵ほどのあったのを、目ざとで見つけて、とてもかわいらしげな指にとって、大人ごとに見せているのは、とてもかわいらしい。

図 11: 文体・訳語の違いと BLEU 値に相関が見られた例

8　フレーズテーブルの問題点と解決策

6 章での分析結果から，古文と現代文とで翻訳の対応が正確に取れていない例も多く見受けられた．これは，フレーズテーブルに不適切な訳語が多く発生したためと考えられる．実際にフレーズテーブルを確認したところ，図 10 に示したように，読点に読点以外の語が対応しているなど多くの不適切な対応があることを確認した．これらの不適切な対訳をフレーズテーブルから取り除く方法は，Johnson らにより提唱されている [Johnson et al.2007]．他にも，一対一の対応を強化するため対訳のコーパスに辞書を追加する方法や，GIZA++ で語の対応を学習する際に不適切な語の対応を適切な語に置き換えることで正確な対訳の確率を上げる方法も考えられる．

9　まとめと今後の課題

本稿では，統計的機械翻訳を用いて古文を現代文に翻訳する際，言語モデルと翻訳モデルの学習に使用するコーパスを中古和文に絞り，言語モデルの学習用コーパスに小学館コーパス以外のコーパスを加えることで翻訳性能の向上を図った．コーパスを加えた結果，星野らよりも評価値は低かったものの，ベースラインよりも翻訳精度は向上した．これは，言語モデルを生

成する際に，小学館作品の中古和文のコーパス内で他と文体の異なる「源氏物語」の影響が少なくなったためと考えられる．また，翻訳結果を 1 文ごとに評価し分析した結果，入力の古語がそのまま訳せている例が見られたこと，古語と現代語で不正確な対応が多くあったことから，BLEU による評価方法の見直しや訳語の対応関係の向上が今後の課題となることが分かった．

謝辞

本研究で使用したコーパス小学館『新編日本古典文学全集』は，国立国語研究所から頂いたものです．関係者各位に感謝致します．

参考文献

[Gale&Church1993] Gale, William A. and Kenneth W. Church (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational linguistics* Vol. 19.1, pp.75-102

[Gao&Vogel2008] 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

[Johnson et al.2007] Johnson,J. Howard, Joel Martin, George Foster et al. (2007). Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (EMNLP-CoNLL2007), pp. 967-975

[Maekawa2008] Maekawa, Kikuo (2008). Balanced Corpus of Contemporary Written Japanese. In *Proceeding of the 6th Workshop on Asian Language Resources* (ALR 6), pp.101-102

[Papineni et al.2011] Papineni, Kishore, Salim Roukos,Todd Ward et al. (2011). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (ACL2011), pp. 311-318

[Koehn et al.2007] Koehn, Philipp, Hieu Hoang, Alexandra Birch et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on interactive poster and demonstration sessions* (ACL2007), pp. 177-180

[Kudo et al.2004] Kudo, Taku, Kaoru Yamamoto, and Yuji Matsumoto (2004). Applying Conditional Random Fields to Japanese Morphological Analysis. *EMNLP2004*, pp. 230-237

[Stolcke2002] Stolcke, Andreas (2002). SRILM–an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pp. 901-905

[新井ら 2012] 新井紀子、松崎拓也 (2012)「ロボットは東大に入れるか？—国立情報学研究所『人工頭脳』プロジェクト—」人工知能学会誌，27:5，pp.463-469

[小木曽ら 2010] 小木曽智信、小椋秀樹、田中牧郎、他 (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，2010-CH-85:4，pp.1-8

[伝ら 2007] 伝康晴、小木曽智信、小椋秀樹、他 (2007)「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」日本語科学，22 号，pp.101-122

[星野ら 2014] 星野翔、宮尾祐介、大橋駿介、他 (2014)「対照コーパスを用いた古文の現代語機械翻訳」言語処理学会第 20 回年次大会発表論文集，pp.816-819

[横野ら 2014] 横野光、星野翔 (2014)「統計的現代語訳モデルを用いたセンター試験古文問題解答」第 5 回コーパス日本語学ワークショップ，pp.161-166

日本語教育とコロケーション：連語の形で用法を学ぶ重要性

STRAFELLA Elga Laura（国立国語研究所　日本語学術振興会特別研究員）
前川 喜久雄（国立国語研究所　コーパス開発センター　言語資源研究系）

Japanese-language Education and Collocations: The Importance of Learning Word Co-occurrences

Elga Laura Strafella　(National Institute for Japanese Language and Linguistics, JSPS Postdoctoral Fellow)
Kikuo Maekawa　(National Institute for Japanese Language and Linguistics, Department of Corpus Studies)

要旨

コロケーション研究は、現在では自然言語処理だけでなく、日本語学や日本語教育においても重要な研究領域となっている。そのような状況を踏まえて、本研究は日本語教育における慣用表現に着目している。コーパスから単語間の強い共起性に関する情報が自動抽出できる自然言語処理の分野は近年さらに進展し、狭義の言語学の研究にも適用できるようになり、文法と語彙の知識だけでは分からない表現上の慣用は特に研究の対象となった。「足を運ぶ」、「手を焼く」、「尻が重い」、などは全体の意味が個々の語の意味とは異なるので、日本語教育では意識して教えなければならないし、辞書でも一般連語句から区別して特別に扱う必要がある。しかし、コーパスからのコロケーション情報の自動的な抽出において、そうした区別は明らかに困難で、現実にコーパスの分析結果を人手で処理するしかない。本研究では、イタリア人の日本語学習者（中級者）を対象とし、BCCWJ から抽出した連語を処理するために調査を行なった結果を報告する。そして、日本語の母語話者が学校で学習する基本的な専門表現も単独で覚えるのではなく連語の形で用法を学ぶように学習者もコロケーションの意味と用法を学ぶことが重要であることを指摘する。

１．はじめに

現在、世界でコロケーション習得に関する研究が徐々に成果をあげてきている。しかし、実際の日本語教育ではその成果を生かしてないのが事実である。本稿ではヨーロッパのイタリアの実態を調べる。

２．辞書とコロケーション

コロケーションとは node[1]（共起関係にある主要語）と collocate[2]（中心語と連語する語）の習慣的な結びつきであり、典型的には名詞・動詞・形容詞および副詞からなる句である。慣用句（いわゆる「イディオム」）と比べ比較的最近、辞書記述に導入されるようになった。さらに、1995年から、コーパスに基づき編集された辞書が相次いで出版され、[3]コーパ

[1] 「中心語」.

[2] 「共起語」.

[3] 一例に、The BBI Dictionary of English Word Combinations. John Benjamins. 1997.

ス言語学の影響でコロケーション分析がコーパスと統計指標（raw frequency, *t*-score, log-likelihood ration, MI-score, など）によって行われるようになった。それにも関わらず現在でも統計的に採集されるデータは手作業で分析しなければならない。

一般的な辞書では、語の選択制限、用例、語法などが多岐にわたるため、どうしてもコロケーションの記述は不十分になる。そこでコロケーション専門の辞典が必要とされる。筆者らは、イタリア人日本語学習者のために、コーパスデータに基づく網羅的なコロケーションリストを作成することを最終目標として、先に『現代日本語書き言葉均衡コーパス』から共起語を抽出した（Strafella 2013）。本稿では、抽出されたデータを評価するための１ステップとし、第 2 言語として日本語を学ぶイタリア人（大学院の修士課程１・２）を対処としたコロケーション理解の調査研究を行った。

３．調査概要

本調査は2014年の12月に行われた。実施場所は、イタリアの「ナポリ東洋大学」である。調査対象は、大学院の修士課程の学生で、「アジア・アフリカ・地中海研究科」１年生の20名と２年生の21名、「人文社会研究科、アジア・アフリカ国際関係コース」１年生７名と２年生10名（合計58名）である。学生はコース別に授業内容が異なっているが、最終的に日本語能力試験-JLPTのN2に当たる知識を得るための教育を受ける。

調査は授業中に行われたため、四つのクラスで別々に実施した。一つの授業は２時間で行われるが、各クラスで１時間を調査のために利用させてもらった。初めに調査用紙を配布し、記入方法などの説明を行った。調査用紙には、３つの練習問題があり、次のような問題になっている。選択肢よりも翻訳の問題の方が時間を要するので、第１部と第２部の選択式問題を考えすぎないように注意を与えた。１文に対する平均的な回答時間は100秒程度であった。

第１部：文を読んでふさわしい動詞を選択してください。（11文）
第２部：文を読んでふさわしい名詞を選択してください。（11文）
第３部：次の文をイタリア語に訳してください。（10文）

問題の形式は次のようになっている：

（第１部）1）静かな待合室（まちあいしつ）で時計（とけい）の**時を＿＿＿**音だけが聞こえた。
　　a. 図る
　　b. 見る
　　c. 刻む

（第２部）1）海外旅行で一週間ほど**＿＿＿を空けます**。
　　a. 家
　　b. 穴
　　c. 間

（第３部）1）あなたの一言（ひとこと）で**目が覚めました**。
　　2）物音（ものおと）で**目が覚めた**。外はまだ暗い。

第１部と第２部の質問項目は、コロケーション辞典の見出し語としてどのような品詞が適切かを決めるために考案したものである。具体的には、名詞と動詞のどちらが学習者に

とって把握しやすいかを明らかにすることが目的である。連語[4]に含まれる語彙は *A Frequency Dictionary of Japanese* (Tono *et al.* 2013) に掲載されているもののみである。候補（太文字で示している）は ChaKi.NET という検索ツールで抽出した。それぞれのコロケーションに関する用例は NINJAL-LWP for BCCWJ（以下、NLB） を検索したものである。より難しい語彙は、ナポリ東洋大の教師と相談した上で、振り仮名をつけ、意味を説明することにした。コロケーションが含まれる文脈すら理解できなければ、慣用的な意味も把握できないことが明らかだからである。

第３部の文章には二つ以上の意味を持つ共起語が示された。それぞれの表現は文字通りの意味で使われている用例と慣用的な意味で使われている用例を一つずつ挙げている。これによって学習者が意味を区別できるかどうかを確かめた。学生には、よく理解できない文に対してもできるだけ想像を巡らして回答するよう指示を与えた。最後に調査に関するコメントも書いてもらった。個人情報としては性別、年齢、日本語能力レベルに関する情報を集めたが、氏名は匿名とした。

４．分析と結果

分析は、筆者らが手作業で行い、回答を図にまとめた。**図１**は、第１部の問題とその正答数を表したものである。

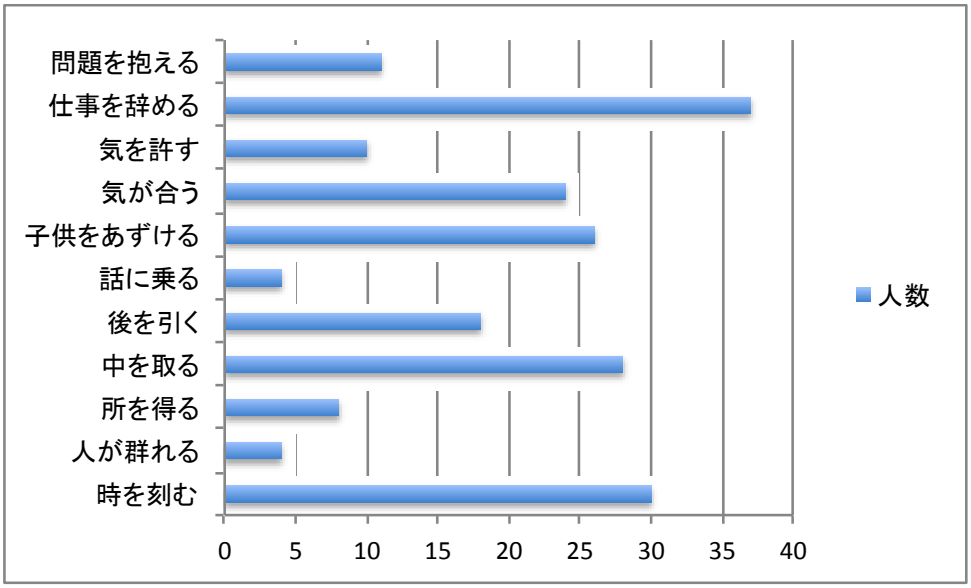


図１　動詞に関する問題とその正答数

図１から分かるように、58 名中半分以上が正しく理解できたコロケーションは、「仕事を辞める」と「時を刻む」のふたつだけであった。一方、もっとも把握しにくかった表現は「話に乗る」と「人が群れる」であり、正答数は４名であった。

図２は、第２部の問題とその正答数を表したものである。

[4] 本稿では、「連語」と「コロケーション」は同義語として使われている。

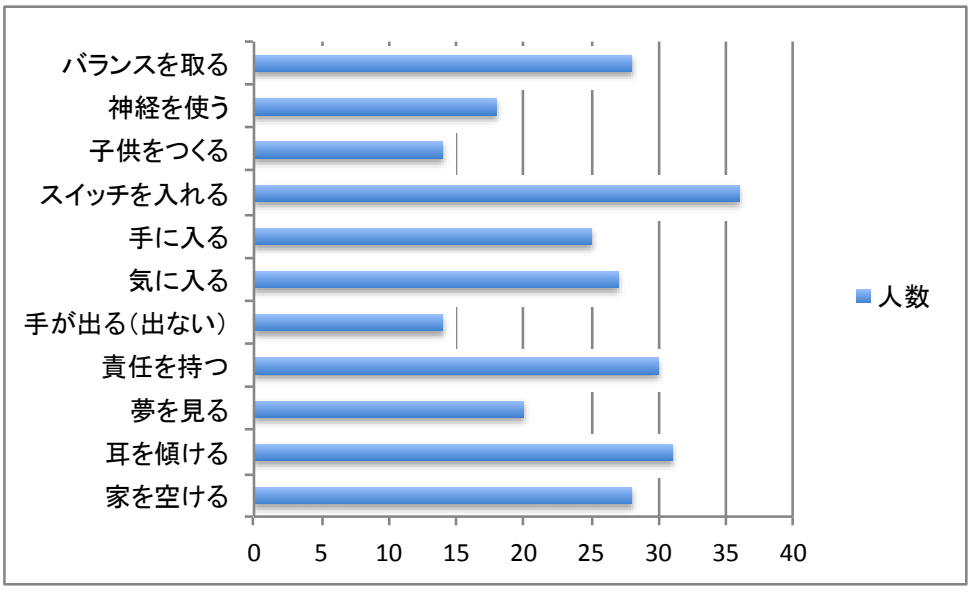


図２　名詞に関する問題とその正答数

図２から分かるように、名詞では動詞よりも正答が増える。50%以上の正答率を示した人数は少ないが、図１と比べると正答率は顕著に高い。予想に反して、「子どもをつくる」と「手が出ない」という表現の用法はあまり理解されていない。いずれも正答数は 14 人だけであった。

第３部の問題はペアごとに回答を分析した。以下のような傾向が観察されたが、そのうち i.と ii.は広く見られたものである。

i.　３章で示した用例のように、文字通りの意味で使われるコロケーション（「物音で目が覚めた」）の方が理解しやすかった。この場合、50%以上の学生が正答できた。

ii.　慣用的な表現は理解しにくく、正答率は非常に低かった。例えば、「彼は**足がある**のでピンチランナーにはうってつけだ.」と「この町は夜遅くまで**足がある**ので、便利でいい.」の場合、それぞれの正答率は 20%と 24%である。あるいは、同じ表現の複数の意味の中で一つだけがよく知られており、もう一つの意味はほとんど知られてないケースがあることも明らかとなった。例えば、「そんな大事なことを、軽々に**口にし**てはいけない.」（正答率：52%）と「こんな高級なものを、いままで**口にした**ことはありません.」（正答率：16%）。

iii.　イタリア語にも類似した表現があると正答率が 50%を上回ることがあった（例えば、「太陽が顔を出すと景色がすべて一変する.」正答率：56%）。

iv.　意外であったのは、意味が明白だと考えられる表現においても混乱が生じうることが分かった。例えば、「口を開く」の場合、「大きく口を開いてくださいと先生に言われ

ました.」という用例は 75%以上の人が理解できず、ほとんどは次のような翻訳をした：「先生にもっと大きな声で話してくださいと言われました.」。この場合、学生は先生という言葉を見ると大学の先生のことと解釈し、それに合った状況と意味を考え出したのだと思われる。

５．まとめ

本研究では、イタリア人の日本語学習者をとおして、従来から指摘されているように、コロケーションが学習者にとって非常に困難な言語現象であることを確認した。また、学習者は動詞より名詞に関する知識が深く、名詞の選択に関する問題の方が正答率が高いことを確認した。これは事前に予期したとおりであった。子供が母語を習得しはじめるとき、動詞・形容詞・副詞ではなく、最初に名詞を使えるようになる。同様に学習者も外国語で文章を作るとき名詞からスタートするのだと考えられる。この結果は、コロケーション辞典の見出し語は名詞中心にたてるべきであることが示唆していると考えられる。

また、コロケーションは母語話者の文化と言語の歴史に関わる多面的な現象であるため、辞典を編集するときには、言語外の事実に関する資料も提供しなければならない。本調査で示されたように、イタリア人と日本人が類似した言語表現を使うにも関わらず、それぞれの言語が異なる意味を持つパタンもある。

最後に、学習者が記入したコメントでも強調されていたように、日本語での文章・会話を理解するには言葉そのものの意味が分かれば、十分であるとはかぎらない。語と語が結びついて新しい表現を生み出すともともとの語の意味と微妙なニュアンスの違いを生じ、全く違う意味になることも少なくない。学生達は調査に協力したことでコロケーションの曖昧性とその難しさを知ったように思えた。

以上を要約すると、上に述べたように日本語を学ぶ学習者は語彙を単独で覚えるのではなく、連語の形で用法を学ぶことが重要である。

６．今後の課題

本研究は、日本語学習者を対象としているため、イタリアで日本語教育を行なっている大学の協力を得て調査を実施した。今後は、同様の調査を進め、最終的には日本語コロケーション辞典を編集したい。また、調査のフォローアップで学習者の意識を明らかにし、海外での日本語教育を支援するために母語話者（教師と生徒）の言語と状況をよく検討し、それに適する教材を開発したい。

謝　辞

本研究は、日本学術振興会外国人特別研究員（平成 25～27 年度）の補助によって実施した。本調査の実施にあたっては、ナポリ東洋大学日本語学科の協力を得た。 Silvana De Maio, Junichi Oue, Chiara Ghidini の各位に特に感謝申し上げる。

文　献

Maekawa, Kikuo *et al.* (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48, pp.345-371.

Shingo, Imai (2012). Development of a Learners'Dictionary of Polysemous Japanese Words and

Some Proposals for Learners'Lexicography, *Acta Linguistica Asiatica*, Vol.2, No.3, pp.63-75.
Strafella, Elga L. (2013). *Collocations in Contemporary Japanese – A Corpus-Based Language Study*. Germany: LAP LAMBERT Academic Publishing.
Tono, Yukio, Makoto Yamazaki, and Kikuo Maekawa (2013). *A Frequency Dictionary of Japanese – core vocabulary for learners*. London & New York: ROUTLEDGE.
堀正広(2011)『これからのコロケーション研究』、ひつじ書房

関連 URL

NINJAL-LWP for BCCWJ (NLB) 国立国語研究所 http:// nlb.ninjal.ac.jp/
ChaKi.NET http://sourceforge.jp/projects/chaki/

MCN コーパスにおける 条件表現「たら」「れば」「ならば」のアノテーション

飯島采永（お茶の水女子大学理学部）.
佐藤果穂（お茶の水女子大学理学部）
田中リベカ（お茶の水女子大学大学院人間文化創成科学研究科）
戸次大介（お茶の水女子大学大学院人間文化創成科学研究科
／国立情報学研究所／CREST，JST）

Annotating Japanese Conditional Expressions "Tara", "Reba", "Naraba" in MCN Corpus

Sae Iijima (Faculty of Science, Ochanomizu University)
Kaori Sato (Faculty of Science, Ochanomizu University)
Ribeka Tanaka (Graduate School of Humanities and Sciences, Ochanomizu University)
Daisuke Bekki (Graduate School of Humanities and Sciences, Ochanomizu University / National Institute of Informatics / CREST, JST)

要旨

MCN コーパスでは、命題の確実性に関わる様相・条件・否定表現に対して意味アノテーションを付与している。複数のアノテータ間で一致する判断、すなわち再現性のある言語事実を蓄積するため、ガイドラインには言語学的テストを用いている。本研究では、条件表現「たら」「れば」「なら（ば）」に対するガイドラインを作成し、『現代日本語書き言葉均衡コーパス』の新聞記事に対して計 600 件のアノテーションを行った。ガイドラインは、日本語学における先行研究の分類をコーパス上の出現例を元に分割・統合したラベル群、及びそれらに対する言語学的テストから構成される。本論文ではガイドラインの紹介に加え、多数の判断を取りうるアノテーション例についても解説する。

1. はじめに

自然言語で記述されるテキストには、事実だけでなく、推測、仮定、仮想現実などの様々な情報が含まれる。情報を識別する手がかりの一つとして、様相表現、否定表現、条件表現などによって形成される「意味的文脈」がある。人間は、自然言語で書かれた情報を読むとき、これらの文脈に基づいて情報の確実性の判断を行うことができる。機械によって情報の確実性を判断したい場合にも、これらの「意味的文脈」の認識を可能にする必要がある。MCN コーパス（川添ら(2011)）は、機械による確実性判断の基盤となるコーパスを構築するために作成されたものであり、命題の確実性に関わる「意味的文脈」に対して意味アノテーションを付与した言語データである。複数のアノテータ間で一致する判断、すなわち再現性のある言語事実を蓄積するため、言語学的テストを用いたガイドラインを作成しアノテーションを行っている。これまでに複合表現「(と) いう」「(と) する」(叢ら(2013)) や形式名詞「わけ」「はず」「つもり」(宇津木ら(2014)) のガイドラインの作成とアノテーションを行ってきたが、条件表現に対する網羅的なガイドラインは作成されていなかった。

MCN コーパスのアノテーションでは、言語学的テストを採用したガイドラインを使用して

いる。ここでいう言語学的テストとは、文や文の一部の容認性や適切性を判定するものである。たとえば、「複合機能表現『という』の分類にみる MCN コーパスの方法論検証」(叢ら(2013))におけるガイドラインでは「いう 2」は伝聞の意味を持つ分類である。「いう 2」は「そう(だ)」に置き換えることができる。

(1) a. ニュースによるとインフルエンザが流行しているという。
　　b. ニュースによるとインフルエンザが流行しているそうだ。

この置き換えは、言葉を発するという意味をもつ「いう 1」には当てはまらない。

(2) a. 花子は太郎を天才だという。
　　b. *花子は太郎を天才だそうだ。

このような分類を判定するための言語学的テストを導入した。
本研究では、条件表現「たら」「れば」「なら（ば)」に対してガイドラインを作成し、『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書籍ドメインに対して計 600 件のアノテーションを行った。各表現の分類について、条件表現について平易な文法説明を記し様々な例文を網羅した日本語教育の本である『日本語文法セルフマスターシリーズ 7 条件表現』(有田ら(2001))(以下、「セルフマスター」と呼ぶ) を参考にした。

2. 条件表現について

文(3)～文(5)に条件表現の例を挙げる。『日本語条件文と時制節性』(有田(2007)) によると条件表現とは「不確定な知識に基づく推論の明示的な言語表現」とされる。

(3) 晴れたら動物園に行く。
(4) 時間があれば本を読む。
(5) n が偶数ならば 2 で割り切れる。

代表的な条件表現としては「たら」「れば」「なら（ば)」「と」「ては」などが挙げられる。そのうち今回は「たら」「れば」「なら（ば)」に関して分析を行った。
条件表現の現れる文を「A＋条件表現＋B」としたとき、A を「前件」、B を「後件」とする。文(3)～文(5)の前件は、出来事を仮定しているもの ([仮定])、事実と反対のことを述べているもの ([偽]) に大別される。これについて「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」(川添ら (2011))では、以下のような用法の分類を与えている。

ガイドラインの分類：
分類 1：【予測的条件表現(真偽が未知、判断あり、確実性 100%)】
　　1 時間後に駅に集合したら、その足でいつもの居酒屋へ直行しよう。
分類 2：【認識的条件表現(真偽が未知、判断あり、確実性 0~99%)】
　　もしもうまくいかなかったら、別の手段を考えよう。

分類 3：【反事実条件表現(偽であることが既知)】
　　太郎が出場していたら、試合に勝てただろう。

分類 1 と 2 は前件が未来の出来事であるため、前件の真偽は未知、つまり[仮定]である。分類 1 と分類 2 の違いは前件の確実性の違いである。分類 1 では前件のおこる確率が書き手(語り手)にとって 100%であるのに対し、分類 2 の前件のおこる確率は 100%未満である。

しかし、前件の分類はこれだけでは十分ではない。たとえば、文(6)の前件は「食べてみた」であるが、これは実際に食べてみた後のため、[仮定]でも[偽]でもない。

また、条件表現を表す語が文章中に現れたとしても、常に含意を表すとは限らない。たとえば、文(7)では前件：「姉がいる」、後件：「兄がいる」となるが、前件の成立が後件の成立に寄与しないため、含意を表さない並列条件となる。文(8)では、そもそも前件が命題ではなく名詞であるために真偽での分類はできないが、文中に出現している「なら」が前方でみたような条件表現だとは考えにくい。

(6)　食べてみたら美味しかった。
(7)　私には姉もいれば兄もいる。
(8)　京都なら京都、東京なら東京の良いところがある。

このように「たら」「れば」「なら（ば)」が文章中に表れても条件表現だとは限らず、見た目だけでは条件表現かそうでないかの判断は困難である。以上のことより本研究では、先述した条件表現の定義にあてはまる例に限らず、二つの事柄を並べる並列条件の用法等も分析対象としている。また前件の分類については、出来事を仮定しているもの（[仮定])、事実と反対のことを述べているもの（[偽]）に加えて、事実を述べているもの（[真])、その他（[名詞][疑問]等）の 4 つに分けられるとしている。

3. 概要

3.1. ガイドラインの紹介

MCN コーパスのアノテーションで使用しているガイドラインは、「言語情報の確実性に 影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」(川添ら(2011))をもとにしている。もともとのガイドラインには、2 節で述べたように条件表現について用法別のカテゴリが例文とともに示されている。しかし、これらの基準だけでは、ある表現がどのカテゴリに属するかを判断できない場合があるため、「セルフマスター」にある分類を参考にガイドラインを分割・統合した（表 1，2，3)。

このガイドラインでは、新たに「前件の条件が真であり、さらに真である中にもいくつかの種類が存在している」という観点から分類をしている。

表 1:「たら」におけるガイドライン A

		前件	後件	備考	例文
1	たら	過去の事実(真)	過去の事実		食べてみたら、美味しかった
2	たら	過去の持続(真)	過去の事実		本を読んでいたら、電話が鳴った
3	たら	時間の経過(真)	過去形以外	後件が過去形なら、1	五時になったら、帰ってくるでしょう
4	たら	真	前件に基づく判断		・こんな暗い部屋で本を読んでいたら、目が悪くなりますよ
5	たら	仮定(真偽未知)	確実性0～100%		部屋が清潔だったら、病気にならないはずだ
6	たら	仮定(真偽未知)			家に帰ったら、うがいをしなければなりません
7	たら	仮定(真偽未知)	真		新聞を読みたかったら、ここにあるよ
8	たら	偽	偽		お金があったら、買えるのに
9	たら			疑問文	何を読んだら、そんなに賢くなれるの
10	たら	単なる状況	真		この道をまっすぐ行ったら、右手に白い建物があります
				慣用表現	・～したら？／どう？／どうですか(？) ・～ったら／ときたら

表 2:「れば」におけるガイドライン A

		前件	後件	備考	例
1	れば			並列	庭には梅もあれば桜もあった
2	れば	仮定			・時間があれば、本の整理を手伝ってください
3	れば	時間の経過			春になれば、芽が出る
4	れば	疑問語			どこにいけば受験案内がもらえますか
5	れば	真			ここまで来れば、あとは一人で帰れます
6	れば	偽	偽		お金があれば買えるのに
7	れば	思う・考える・言うなど		発言の前置き	思えば、楽しい人生だった。
				慣用表現	・～ば～ほど ・～ばいい(な／のに／のだが) ・～なければ[いけない／だめだ／ならない] ・～すればいい／すれば？

表 3:「なら（ば)」におけるガイドライン A

		前件	後件	備考	例文
1	なら(ば)			対比	顔もいいなら、頭もいい
2	なら(ば)	真	意思・判断		明日はプールに行くことにしたよ ーあなたがいくなら私も行くわ
3	なら(ば)	仮定	判断・働きかけ・状態		景気が回復するなら、円高になるだろう
4	なら(ば)	真	偽		海外勤務になるなら、もっと英語を勉強しておくべきだった
5	なら(ば)	偽	偽		試合に出られたなら、勝てたのに
				慣用表現	・なぜなら(ば)

そのガイドラインをもとにアノテーションを行い、コーパス上の実際の例を参考にネガティブテストを作成し、そのテストを使って再度分類を統合した（表 4, 5, 6）。

表 4:「たら」におけるガイドライン B

					例	テスト	
		前件	後件	備考		もし	
1	たら	過去の事実(真)	過去の事実		食べてみたら、美味しかった	×	「～したところ」置き換え
2	たら	過去の持続(真)	過去の事実		本を読んでいたら、電話が鳴った	×	「～していたら」
3	たら	時間の経過(真)	過去形以外	後件が過去形なら、1	五時になったら、帰ってくるでしょう	×	「ば」置換可能
4	たら	真	前件に基づく判断		・こんな暗い部屋で本を読んでいたら、目が悪くなりますよ	×	「このように」「かもしれない」挿入可能
5	たら	仮定(真偽未知)			・家に帰ったら、うがいをしなければなりません ・何を読んだら、そんなに賢くなれるの	×	「ときには」「というのなら」置き換え、前件に疑問詞あり
6	たら	仮定(真偽未知)	確実性0～100%		・部屋が清潔だったら、病気にならないはずだ ・この道をまっすぐ行ったら、右手に白い建物があります	○	「ときには」「というのなら」置き換え
7	たら	仮定(真偽未知)	真		新聞を読みたかったら、ここにあるよ	○	「前件の否定+たら+後件の否定」では文意が変わる
8	たら	偽	偽		お金があったら、買えるのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・～したら？／どう？／どうですか（？） ・～ったら／ときたら ・～といったら ・なんだったら		

表 5:「れば」におけるガイドライン B

					例	テスト	
		前件	後件			もし	
1	ば			並列	庭には梅もあれば桜もあった	×	可換
2	ば	時間の経過			春になれば、芽が出る	×	「するころ」置き換え
3	ば	真			ここまで来れば、あとは一人で帰れます	×	「ので」置き換え
4	ば	仮定(真偽未知)			・氷が溶ければ水になる ・どこにいけば受験案内がもらえますか	×	「ときには」「というのなら」置き換え、または、前件に疑問詞あり
5	ば	思う・考える・言うなど		発言の前置き	思えば楽しい人生だった	×	なくても文意が通る
6	ば	仮定(真偽未知)			時間があれば、本の整理を手伝ってください	○	「ときには」「というのなら」置き換え
7	ば	偽	偽		お金があれば買えるのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・～ば～ほど ・～ばいい(な／のに／のだが) ・～なければ／ねば[いけない／だめだ／ならない] ・～すればいい／すれば？		

表 6:「なら（ば)」におけるガイドライン B

					例	テスト	
		前件	後件	備考		もし	
1	なら(ば)			並列	顔もいいなら、頭もいい	×	可換
2	なら(ば)	名詞	前件と同様の名詞		木なら木は、そこに木があるというだけでは木ではない	×	前件と後件が同じ単語
3	なら(ば)	名詞	前件とイコールとなる名詞が入っている		だから最高の売れっ妓は遊女なら太夫、女郎なら花魁と考えればわかりやすい	×	「については」置換可能
4	なら(ば)	名詞			君のためならなんでもする	×	「なら」の前に格助詞「に」などが入る
5	なら(ば)	真	意思・判断		・明日はプールに行くことにしたよ ―あなたがいくなら私も行くわ ・海外勤務になるなら、もっと英語を勉強しておくべきだった	×	前件と後件で文章を分け、間に「それでは」が挿入可能
6	なら(ば)	仮定(真偽未知)	判断・働きかけ・状態		景気が回復するなら、円高になるだろう	○	「ときには」「というのなら」置き換え
7	なら(ば)	偽	偽		試合に出られたなら、勝てたのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・なぜなら(ば) ・なんなら		

また、更に「たら」「れば」「なら（ば)」の 3 表現間の対応を考えて改良を行った。これが最終的なガイドライン C（表 7, 8, 9）である。

表 7:「たら」におけるガイドライン C

					例	テスト	
		前件	後件	備考		もし	
1	たら	過去の事実(真)	過去の事実		食べてみたら、美味しかった	×	「～したところ」置き換え
2	たら	過去の持続(真)	過去の事実		本を読んでいたら、電話が鳴った	×	「～していたら」
3	たら	時間の経過(真)	過去形以外	後件が過去形なら、1	五時になったら、帰ってくるでしょう	×	「ば」置換可能
4	たら	真	前件に基づく判断		・こんな暗い部屋で本を読んでいたら、目が悪くなりますよ	×	「このように」「かもしれない」挿入可能
5	たら	仮定(真偽未知)			・家に帰ったら、うがいをしなければなりません ・何を読んだら、そんなに賢くなれるの	×	「ときには」「というのなら」置き換え、前件に疑問詞あり
6	たら	仮定(真偽未知)	確実性0～100%		・部屋が清潔だったら、病気にならないはずだ ・この道をまっすぐ行ったら、右手に白い建物があります	○	「ときには」「というのなら」置き換え
7	たら	仮定(真偽未知)	真		新聞を読みたかったら、ここにあるよ	○	「前件の否定+たら+後件の否定」では文意が変わる
8	たら	偽	偽		お金があったら、買えるのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・～したら？／どう？／どうですか(？) ・～ったら／ときたら ・～といったら ・なんだったら		

表 8:「れば」におけるガイドライン C

		前件	後件			もし	
1	ば			並列	庭には梅もあれば桜もあった	×	可換
2	ば	時間の経過			春になれば、芽が出る	×	「するころ」置き換え
3	ば	真			ここまで来れば、あとは一人で帰れます	×	「ので」置き換え
4	ば	仮定（真偽未知）			・氷が溶ければ水になる ・どこにいけば受験案内がもらえますか	×	「ときには」「というのなら」置き換え、または、前件に疑問詞あり
5	ば	思う・考える・言うなど		発言の前置き	思えば楽しい人生だった	×	なくても文意が通る
6	ば	仮定（真偽未知）			時間があれば、本の整理を手伝ってください	○	「ときには」「というのなら」置き換え
7	ば	仮定（真偽未知）	真		新聞を読みたければ、ここにあるよ	○	「前件の否定+ば+後件の否定」では文意が変わる
8	ば	偽	偽		お金があれば買えるのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・～ば～ほど ・～ばいい（な／のに／のだが） ・～なければ／ねば[いけない／だめだ／ならない] ・～すればいい／すれば？		

表 9:「なら（ば)」におけるガイドライン C

					例	テスト	
		前件	後件			もし	
1	なら(ば)			対比、並列	顔もいいなら、頭もいい	×	可換
2	なら(ば)	名詞	前件と同様の名詞		木なら木は、そこに木があるというだけでは木ではない	×	前件と後件が同じ単語
3	なら(ば)	名詞	前件とイコールとなる名詞が入っている		だから最高の売れっ妓は遊女なら太夫、女郎なら花魁と考えればわかりやすい	×	「については」置換可能
4	なら(ば)	名詞			君のためならなんでもする	×	「なら」の前に格助詞「に」などが入る
5	なら(ば)	真	意思・判断		・明日はプールに行くことにしたよ ーあなたがいくなら私も行くわ ・海外勤務になるなら、もっと英語を勉強しておくべきだった	×	前件と後件で文章を分け、間に「それでは」が挿入可能
6	なら(ば)	仮定	判断・働きかけ・状態		景気が回復するなら、円高になるだろう	○	「ときには」「というのなら」置き換え
7	なら(ば)	仮定（真偽未知）	真		新聞を読みたいなら、ここにあるよ	○	「前件の否定+なら+後件の否定」では文意が変わる
8	なら(ば)	偽	偽		試合に出られたなら、勝てたのに	○	「元の文＋しかし＋後件の否定」で文意が通る
				慣用表現	・なぜなら(ば) ・なんなら		

3.2. ガイドライン A と B の相違点

ガイドライン B にはアノテータ間の一致率を高めるためにテストを作成し、そのテストを用いて A の分類を再度見直した。各表現のガイドラインについて個別に行った改良を以下に解説する。（以下ではガイドライン A の分類 9 を「A9」などと表す。）

3.2.1. 「たら」ガイドラインにおける改良

「たら」ガイドラインの改良では、分類の統合を行った。『たら』ガイドライン A には、以下の分類 A9 が存在していた。

(A9) 疑問文: 何を読んだら、そんなに賢くなれるの。

しかし「何を読んだら」という前件は、疑問詞を含んでいるという違いこそあれ、文としては分類 A5〜A7 にみられるように仮定を示していると考えられる。さらにこの文には「もし」を挿入することは不可能であることから、A9 と A6 を「前件:仮定(テストで「もし」がつかない)」の B5 に統合した。

また、分類 A10 を A5 と統合し、分類 B6 としている。

(A10) 単なる状況: この道をまっすぐ行ったら、右手に白い建物があります。

分類 A10 は一見すると分類 A7 と統合されうるようにも見える。上の文に「右手に」という情報が付加されていなければ「この道をまっすぐ行ったら、白い建物があります」となり、話し手や聞き手が「この道をまっすぐ行」こうと行かまいと「白い建物」はあるので、前件の真偽に関わらず後件は真になるためである。しかし、分類 A7 を元に作成した分類 B7 の「『前件の否定+たら+後件の否定』という文を作り、元の文と比べて文意が変わらなければその文は B7 ではない」というテストに当てはめると、「この道をまっすぐいかなかったら右手に白い建物はない」となり、文意が変わらないので B7 に分類することはできない。最終的には、「たら」を「ときには」に置き換えることが可能であることから、A10 を A5 と統合し B6 とした。

3.2.2. 「なら（ば)」における改良

「なら（ば)」のガイドラインでは、分類 A4 と A2 を統合し、B5 とした。分類 A4 は前件が真、後件が偽であるような用法であり、以下のような例を含むとしていた。

(A4) 前件 真/後件 偽: 海外勤務になるなら、もっと英語を勉強しておくべきだった。

しかし後件の「もっと英語を勉強しておくべきだった」というのは前件の「海外勤務にな」ったことをふまえてのその時点での書き手にとっての反省であり、「偽」であると考えるのは不適切である。後件の反省は、書き手の前件を踏まえた感情・意思であると考えられるので、「前件 真/後件 意思・判断」である A2 と統合し B5 とした。これに伴い、前件が真、後件が偽であるとする分類は削除された。

また、新たな分類の追加も行った。新たな分類の追加は、文に対してテストを適用した結果、既存の分類のどれにも含まれないと判定された際に検討される改良である。

(9) 最高の売れっ子は遊女なら大夫、女郎なら花魁と考えればわかりやすい。

この文においては、「最高の売れっ子の遊女」＝「大夫」、「最高の売れっ子の女郎」＝「花魁」というように前件と後件の間にイコールの関係が成り立つ。この関係は既存の分類の

どこにも分類されないため、新たに分類 B3 を作成した。

(10) 木なら木はそこに木があるというだけでは木ではない。

前件と後件が同じ単語であるので分類 B3 のようにイコール関係を示しているのではなく、その単語の強調ではないかと考えられる。この関係もどこにも分類されないため、新たに分類 B2 を作成した。

(11) 君のためならなんでもする。

この例文の前件は「君のため」という名詞句であるが、B2 のような繰り返しでもなく、B3 のように後件とイコール関係を持っているわけでもないので、どちらにも分類することはできない。したがって新しく分類 B4 を作成した。

3.3. ガイドライン B と C の相違点

さらに、ガイドライン B を改良し、ガイドライン C を作成した。この改良では、「たら」「れば」「なら（ば)」各表現のガイドラインの対応を考えた。たとえば、「たら」における分類 B1「前件:過去の事実/後件:過去の事実」の用法は「たら」だけにしかない用法である。

(12) a. 食べてみたらおいしかった。
　　b. *食べてみればおいしかった。
　　c. *食べてみたならおいしかった。

しかし、「たら」の B8「前件:偽/後件:偽」の分類は、文(13)にみられるように「れば」「なら（ば)」に共通して現れている。他の用法でも対応を考慮し、更なる改良を行った。

(13) a. お金があったら買える。
　　b. お金があれば買える。
　　c. お金があったなら買えた。

また、「たら」の B7「前件:仮定/後件:真」の用法は他の表現の分類には含まれていなかったが、実際は「れば」「なら（ば)」にも対応する用法がある。そのため「れば」C7、「なら（ば)」C7 の分類を追加した。

(14) a. 新聞が読みたかったら、ここにあるよ。
　　b. 新聞が読みたければ、ここにあるよ。
　　c. 新聞が読みたいなら、ここにあるよ。

この他に、前件に名詞がくるのは「なら（ば)」特有の用法であり、更に３つの下位分類があった。このように各表現間には同じ用法もあり対応がみられるが、その一方で各表現にしかない特有の用法も見られた。

4. アノテーション作業と問題点

「たら」「れば」「なら（ば)」の３つの条件表現アノテーション作業はガイドライン設計者 2 名で行った。それぞれの表現について、多くの文章の中から該当の表現が出現する部分を抜き出し、その用法がどのカテゴリに属するかを、テストをもとに判断した。アノテーションの件数は「たら」「れば」「なら（ば)」それぞれ 200 件ずつ、計 600 件行った。アノテーションを行う中で、以下のような例に対するアノテーションが問題となった。

(15) 飴ならここにある。

文(15)の前件は一見すると名詞だが、文脈によっては「飴なら」は省略された形である可能性もあり、「飴が欲しいなら」や「飴を探しているなら」などの候補が考えられる。一方で別の文脈のもとでは、前件の名詞句と「なら」の間に格助詞を補うことも可能である。このように省略されている可能性がある場合、テストの適用が困難となり、判別ができなかったり間違った分類をしたりする恐れがある。

また、話し言葉の場合、略語が使われていてそのまま置き換えができない場合があった。たとえば、「そうしたら」を「ば」に置き換える時（分類 C3 のテスト）は「そうすれば」でいいのだが、「そうしたら」の略語である「そしたら」はそのまま置き換えようとすると「そすれば」という変な言葉になってしまう。しかし、「そうしたら」の略語であるのだから、「そうすれば」に置き換えたい。そのためには、「そしたら」を「そうしたら」に戻さなければならない。こういった省略すべてに対応表をつくることは難しい。

5. 結論

「たら」「れば」「なら（ば)」の 3 つの条件表現に関して、ガイドラインとテストを作成し、アノテーションを行った。いまだ分類が難しい例や問題点があるため更なる改良が必要である。

参考文献

宇津木舞香、佐藤未歩、青木花純、田中リベカ、戸次大介、川添愛（2014)「MCN コーパスにおける形式名詞『はず』『わけ』『つもり』のアノテーション」、言語処理学会第 20 回年次大会発表論文集、B7-1

叢悠悠、田中リベカ、中村絢子、酒向美帆、佐宗智子、清水蘭、劉月晴、川添愛、戸次大介（2013)「複合機能表現『という』の分類にみる MCN コーパスの方法論検証」、国立国語研究所第３回コーパス日本語学ワークショップ論文集、pp.71-80

川添愛、齊藤学、片岡喜代子、崔栄殊、戸次大介（2011)「言語情報の確実性に影響する表現およびそのスコープのためのアノテーションガイドライン Ver.2.4」、Technical Report of Department of Information Science, Ochanomizu University, OCHA-IS 10-4

有田節子（2007)『日本語研究叢書 20 日本語条件文と時制節性』、くろしお出版

有田節子、蓮沼昭子、前田直子（2001)『日本語文法セルフマスターシリーズ 7 条件表現』、くろしお出版

代名詞・疑問詞を含む複合語の調査

淺尾 仁彦（名古屋大学文学部）

Survey of Compounds Containing Pronouns and Interrogatives

Yoshihiko Asao (Nagoya University)

要旨

日本語には「何色」「どこ行き」のように複合語の中に代名詞・疑問詞を含むことができる。このような例は語彙的緊密性 (lexical integrity) の例外とも考えられ、理論的にも興味深い。本研究では、日本語書き言葉均衡コーパス (BCCWJ) を用いてこのような代名詞・疑問詞を含む複合語を調査してその一覧を示すとともに、「名称選択型」「句包摂型」の 2 種類への分類を提案する。

1　はじめに

日本語では「何色」「彼女任せ」のように、複合語の中に代名詞・疑問詞を含むことができる。本研究では代名詞・疑問詞を含む複合語について日本語書き言葉均衡コーパス (BCCWJ) を調査し、その生起条件について考察を加えることを目的とする。

2　研究の背景

Postal (1969) は、単語の一部だけが照応に参加するということはできないとする照応の島 (anaphoric island) の概念を提案した。実際、英語では代名詞を語の一部に用いた、**him-ite*, **who-ite*, **which-less* のような語は認可されない[*1]。一方、影山 (1993, 11), 影山、由本 (1997, 69), 影山 (1999, 8), 伊藤、杉岡 (2002, 8) などが指摘するように、日本語では「彼好み」「ここ止まり」のように、代名詞を含む語が可能である。グルジア語でも代名詞を含む複合語が可能であると主張されている (Harris, 2006)[*2]。

影山らは、「彼好み」や「ここ止まり」のように代名詞を含む複合語の例は、直示 (deixis) にあたり、照応とは区別されているとしており、原則としては「*それすくい (cf. 金魚すくい)」「*彼らげんか (cf. 夫婦げんか)」のように、照応を含む表現は容認されないという立場を取っている。しかし、例えば (1a) のような直示的用法だけではなく、(1b) のように文中で仮に導入された要素を指すような用法でも、特に容認不可能にはならない。

[*1] *forget-me-not* や *she-bear* のように照応詞としての機能を失ったものは除く。なお、*therefore*, *whatever*, *himself* のような語は例外とも考えられるが (Harris, 2006, 116)、このような例には生産性はない。

[*2] 英語で代名詞を含む複合語が許されない理由は、Sproat (1988, 297) では、最大投射が語形成に参加できないため、また、Lieber (1992, 123) では、代名詞が閉じたクラスであるためとされている。これらの説を取った場合、日本語は英語と異なり、名詞と代名詞とのあいだに統語範疇の違いがないとすれば (福井, 1989)、日本語において内向きの照応の島が生じないことは説明できることになる。影山 (1993, 336–338) の疑問詞を含む複合語についての日英の違いの議論も、基本的にこれと同じ考え方ということができる。

(1) a. この電車は [ここ止まり] です。
b. ある駅で確実に降りたい場合、[そこ止まり] の電車に乗るようにすれば、寝過ごす心配がなく安心だ。

逆に「*それすくい」や「*彼らげんか」のような表現は、直示であっても不自然であり、照応と直示の区別が容認度に関与していると言えるのかどうかは判然としない。「ここ止まり」「彼好み」のような容認可能なケースと、「*それすくい」「*彼らげんか」のように容認できないケースとでは何が異なるのかが問題になる。

3 調査

本研究では日本語書き言葉均衡コーパス (BCCWJ) を「中納言」を用いて、代名詞を含む複合語を検索した。複合語全体が辞書に登録されている場合とされていない場合の両方がありうることを考慮し、次の 2 種類の場合に分けて調査を行った。

(2) a. 代名詞のあとに名詞が後続するもの
b. 名詞のうち、初頭部分の音/表記が代名詞と一致するもの

可能性としては、名詞以外と複合する場合や、複合語の後部要素が代名詞となることも考えられるが、そのようなケースは稀であると考え、今回は除外した[*3]。上記の検索結果から、調査者の直観に基づき複合語を形成しているものを探した。その際、以下のようなものは除外された。

(3) a. 現代日本語の文とは言いにくいもの
b. 形態素解析の誤り
c. ゼロ助詞によってたまたま名詞が連続したもので、複合語ではないもの（例えば「おまえ、それ本気でやってるのか？」）
d. 音韻的語をなさないもの。該当するものには「それ全体」「彼専用」「彼女抜き」などがある。これらの例がどのようなカテゴリを成すかどうかは今後の課題となる。
e. 句の包摂ないし引用によるもの。例えば「[うそは どれ] クイズ」のようなものが該当する。
f. 重複形（「いついつ」「誰々」など）
g. 代名詞・疑問詞を含む合成語と言える可能性があるものの、代名詞・疑問詞としての機能を失っていると考えられるもの。「これ見よがし」「どっちみち」「あれこれ」「私事」「オレオレ詐欺」「彼氏」「誰彼かまわず」などが該当する。また「自分嫌い」などの「自分」を含む複合語も、照応的ないし直示的と言える可能性があるものの判断が難しく、今回は除外した。

[*3] 品詞を名詞に限ったために、実際には今回の調査対象の語と同じような性質のものであるにもかかわらず、検索結果から漏れたものに「これ系」「これ狙い」のようなものがある（「-系」「-狙い」は接辞扱いとなっている）。

h. 数詞または助数詞のついたもの（「何日」「何回」「何万」）。ただしこれらも疑問詞を含む複合語と言える可能性がある。

4　結果

検索の結果以下のような語が見つかった（表記揺れを吸収したため、実際に出現する表記と異なる場合がある）。

これ　これ目当て (2)
それ　それ目当て (4)、それ状態 (2)
あれ　あれ目当て (1)
そこ　そこ止まり (1)
あそこ　あそこ近く (1)
どこ　どこあたり (2)、どこ駅 (2)、どこ公演 (1)、どこ証券会社 (1)、どこ仕様 (1)、どこ大学 (1)、どこプロバイダ (1)、どこ方向 (1)、どこ方面 (1)、どこルート (1)
こちら　こちら側 (294)、こちら方面 (5)、こちら岸 (4)、こちら地方 (3)、こちらあたり (1)、こちら式 (1)、こちら方向 (1)、こちら任せ (1)
そちら　そちら側 (23)、そちら方面 (22)、そちらサイド (1)、そちら畑 (1)、そちら方向 (1)、そちら問題 (1)
あちら　あちら側 (38)、あちら方面 (3)、あちら関係 (1)、あちら地方 (1)、あちら持ち (1)
どちら　どちら側 (43)、どちら設定 (1)、どちら方向 (1)、どちら方面 (1)、どちら巻き (2)、どちら目線 (1)
私　私好み (4)、私色 (1)、私名義 (1)、私流 (1)、私流儀 (1)、私レベル (1)
僕　僕好み (4)、僕名義 (2)、僕譲り (1)
俺　俺好み (8)、俺ルール (3)、俺節 (2)、俺アイデア (1)、俺色 (1)、俺語 (1)、俺式 (1)、俺設定 (1)
君　君色 (1)
あなた　あなたがた (63)、あなた側 (19)、あなた好み (10)、あなた名義 (7)、あなた任せ (6)、あなた色 (2)、あなた譲り (1)、あなた目当て (1)、あなたタイプ (1)、あなた通り (1)
彼　彼好み (10)、彼側 (4)、彼名義 (2)、彼経由 (1)、彼仕様 (1)、彼目当て (1)
彼女　彼女名義 (2)、彼女好み (2)、彼女関係 (1)、彼女任せ (1)、彼女目当て (1)
何（なに）　何事 (2065)、何者 (1957)、何色（なにいろ）(180)、何語（なにご） (62)、何県 (20)、何奴（なにやつ）(14)、何味 (13)、何フェチ (10)、何人（なにじん）(7)、何区 (7)、何ゴミ (6)、何カップ (6)、何町 (6)、何部 (5)、何課 (5)、何市 (5)、何川 (5)、何星 (5)、何宗 (5)、何犬 (4)、何先生 (4)、何山 (4)、何主義 (4)、何学部 (3)、何料理 (3)、何国 (3)、何曜 (1)、何屋 (3)、何役 (3)、何新聞 (2)、何パンダ (2)、何口 (2)、何駅 (2)、何地帯 (2)、何鍋 (2)、何賞 (2)、何組 (2)、何語族 (1)、何油 (1)、何占い (1)、何おにぎり (1)、何カエル (1)、何カビ (1)、何がん (1)、何関係 (1)、何キー (1)、何景気 (1)、何高校 (1)、何サンド (1)、何痔 (1)、何石膏 (1)、何ゼミ (1)、何属性 (1)、何タイプ (1)、何罪 (1)、何トース

ト (1)、何年（なにどし）(1)、何鳥 (1)、何トンボ (1)、何なまり (1)、何猫 (1)、何パン (1)、何ふぐ (1)、何報告 (1)、何マニア (1)、何結び (1)、何指 (1)、何案件 (1)、何うどん (1)、何映画 (1)、何ガール (1)、何花粉 (1)、何カレー (1)、何球場 (1)、何銀行 (1)、何組 (1)、何現象 (1)、何公共団体 (1)、何航空 (1)、何国民学校 (1)、何婚式 (1)、何時代 (1)、何職人 (1)、何書道 (1)、何人種 (1)、何線区 (1)、何戦争 (1)、何戦隊 (1)、何ソース (1)、何地方 (1)、何出口 (1)、何都道府県 (1)、何版 (1)、何棒 (1)、何保険 (1)、何味噌 (1)、何目線 (1)、何野郎 (1)、何列車 (1)

何（なん）　何人（なんぴと）(82)、何曜日 (62)、何時（なんどき）(47)

誰　誰色 (1)、誰ファン (1)、誰情報 (1)、誰タイプ (1)

いつ　いつ頃 (990)、いつ現在 (2)、いつ時代 (1)

5　考察

代名詞・疑問詞を含む複合語について、以下のような分類を提案する。

(4) a. 語彙化しているもの（「何者」「何事」など）
　　b. 生産的なもの
　　(i) **名称選択型**：選択肢のうちから選択する性質のもので、場合によってはメタ言語的に名称を問う表現になるもの。「何語（なにご）」「どこ大学」など。
　　(ii) **句包摂型**：パターンの生産性が高く、後部要素名詞が前部要素として句を取ることも可能で、その位置に代名詞も許容されるタイプのもの。「これ目当て」「それ状態」など。

生産的なもののうち、(i) は疑問詞、(ii) は（疑問詞でない）代名詞の例が多い。ただし「そちら側」「どちら側」、「あなたタイプ」「誰タイプ」のような例は (i) と (ii) の性質を兼ね備えているように思われる。以下で (i) と (ii) の実例を挙げて具体的に論じる。

5.1　(i) 名称選択型

まず (i) には、引用の形式が明示され、明らかに名称を問うメタ言語的な使用であることが明らかなものがある (5a)。しかし、引用であることが明示されていなくても、(5b) のように実際には名称を問うていると考えられる例は多い。(5c) のようなケースは、名称よりも内容を問うている側面が強くなるが、明確な境界を定めるのが難しい。

(5) a. .. 好景気だとすると、[なに景気] と名付けるべきでしょうか？ [知恵袋 OC03_02086]
　　b. 結婚十五年目は [何婚式] ですか？ [知恵袋 OC11_01609]
　　c. 割れた陶器や鏡って、[何ゴミ] に出したらいいですか？ [知恵袋 OC08_00532]

使用される疑問詞は「何」が圧倒的に多いが、意味内容が場所名詞であったり、方向であったりする場合は、他の疑問詞が用いられる例もある。

(6) a. 台風は左右 [どちら巻き] ですか？ [知恵袋 OC12_06166]

b. ..「え、それって [いつ時代]？」ってなってみんなで考えたんだけど、..
[Yahoo!ブログ OY04_03547]

「どちら巻き」や「何ゴミ」のような例から分かるように、このような形式が用いられるためには例えば「-巻き」という形式に対して「左巻き」「右巻き」のような複数の選択肢が意識されさえすればよく、その複合パターンの生産性が特に高い必要はない。

ほとんどの例は「どこ駅」に対する「東京駅」「立川駅」のように、代名詞の部分を具体的な名詞で置き換えた表現が存在するが、わずかに例外がある。次の「どこ証券会社」「どこプロバイダ」の例は、個別の証券会社やプロバイダを指して「-証券会社」「-プロバイダ」のようにはあまり表現されないにもかかわらず出現している（ただし誤記である可能性もある）。

(7) a. インターネット取引をしようと思っていますが、手数料など考えて [どこ証券会社] がいいでしょうか？ [知恵袋 OC03_02039]
b. .. プロバイダがヤフーじゃないと知恵袋出来ないと勘違いされている人がいる見たいですが、別に [どこプロバイダ] でも出来るって知ってますよね？
[知恵袋 OC14_11504]

5.2 (ii) 句包摂型

このタイプは、複合の生産性が高く、先行文脈の内容を受けて代名詞を取るのが容易になっているものである。例えば次の「-目当て」「-状態」のような例は、前部に来ることのできる名詞に実質的に制限がない。これは (i) タイプの「-語」「-大学」のような、固有名から選択するタイプのものとは性質を異にする。

(8) a. .. 声優さんは名の知れてる方々なので、[それ目当て] じゃないと見る価値はないと思いますね。 [Yahoo! ブログ OY15_09437]
b. 左手の中指の爪の奥（腕に近いほう）を扉にぶつけて .. 放置しておいても大丈夫でしょうか .. 今まさに [それ状態] です [知恵袋 OC09_03732]

このような例は、句の包摂 (影山, 1993, 326) を許すような場合と一致するのではないかと考えられる。句の包摂を許すということは、さまざまな意味内容の表現に生産的に「-目当て」のような表現が後続できることを示しており、そのことが「それ」などによる直示ないし照応を用いた表現の利用可能性にも結びついていると考えることができる。実際、「-目当て」には次の (9a) のような例がある。さらに「-状態」のような例は、(9b) に見るように、任意の文を引用の形で取ることができる。

(9) a. 民主党の日替わり「マニフェスト」は [選挙の票目当て] の、.. [知恵袋 OC05_02551]
b. .. 高価クセモノ三兄弟などがぐるっと並んで [もうどこから攻めていったらいいかわかりません状態] になっている。 [図書館・書籍 LBs9_00004]

この (9b) のような例は、山下 (2000) で「-的」の引用機能と呼ばれているもの（「皆がやっているから私もやる的発想」のような例）や、中平 (2013) で「-疑惑」「-感」などの用法を取り

上げて「引用による文の名詞化・修飾」と分析されているものに該当する。従って (9b) のような例は、(i) の名称選択型と同じように引用機能が関与しているといえる（ただしその働きは大きく異なる）。

6 まとめ

本研究では代名詞・疑問詞を含む複合語のコーパスからの収集を通じて、その用法を分析した。代名詞・疑問詞を含む複合語が使用される条件にはいくつかのタイプがある。「名称選択型」「句包摂型」の 2 つに分類する分析を示した。

参考文献

福井 直樹 (1989). 「日・英語比較統語論：日・英語の類型論上の相違点とその理論的説明」. 井上 和子（編）,『日本文法小事典』, pp. 89–108. 大修館書店.

Harris, A. C. (2006). Revisiting anaphoric islands. *Language*, **82**: 1, pp. 114–130.

伊藤 たかね、杉岡 洋子 (2002). 『語の仕組みと語形成』. 研究社.

影山 太郎 (1993). 『文法と語形成』. ひつじ書房.

影山 太郎 (1999). 『形態論と意味』. くろしお出版.

Kageyama, T. (2001). Word plus: The intersection of words and phrases. In J. M. van de Weijer & T. Nishihara (Eds.), *Issues in Japanese phonology and morphology*, pp. 245–276. Berlin: Walter de Gruyter.

影山 太郎、由本 陽子 (1997). 『語形成と概念構造』. 研究社.

Lieber, R. (1992). *Deconstructing Morphology: Word Formation in Syntactic Theory*. Chicago: University of Chicago Press.

中平 詩織 (2013). 「引用による文の名詞化・修飾に関して」. 『筑紫日本語研究 2012』, pp. 127–135.

Postal, P. (1969). Anaphoric islands. *CLS 5*, pp. 205–239.

Sproat, R. (1988). On anaphoric Islandhood. In M. Hammond & M. Noonan (Eds.), *Theoretical Morphology*, pp. 291–301. New York: Academic Press.

山下 喜代 (2000).「漢語系接尾辞の語形成と助辞化：「的」を中心にして」.『日本語学』, **19**: 11, pp. 52–64.

ポスター発表　グループB

3月10日（火）　16:00～17:00

新しい日本語辞書定義文型の策定に向けて (第ニ報)

佐藤 理史　(名古屋大学大学院工学研究科)
夏目 和子　(名古屋大学大学院工学研究科)

Towards Full-Sentence Definitions of Japanese Words (Second Report)

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Kazuko Natsume　(Graduate School of Engineering, Nagoya University)

要旨

2014 年 3 月に開催された第 5 回コーパス日本語学ワークショップにおいて、我々は、日本語の語を定義する新しい記述法 (定義文型) として、COBUILD の full-sentence definition (FSD) に倣った日本語 FSD を提案し、そのガイドラインを示した。本稿では、その続編として、主として、副詞、形容詞、動詞に関する進捗状況を報告する。

1　日本語 FSD

我々が設計している日本語 full-sentence definition (FSD)[1] は、COBUILD[2] の FSD に倣ったものであり、それぞれの見出語に対していくつかのターゲット (表現形式) を設定し、その用法・語義を完全な文形式で提示するものである。以下に例を示す。

【そっけない】 [イ形容詞]

1. 《そっけない》〈態度・声・返事など〉とは、
   気持が入っていない〈態度・声・返事〉のこと。
   例：彼のそっけない返事に腹が立った。
2. 《そっけない》〈文章・建物など〉とは、
   工夫や飾りがなくてつまらない〈文章・建物〉のこと。
   例：この家は広いけれどそっけないね。
3. ［〈誰か〉に］《そっけなく》〈言う・答えるなど〉とは、
   ［〈誰か〉に］短いことばで感情を表さないで〈言う・答える〉ことをいう。
   例：「べつに」とそっけなく答えた。
4. ［〈誰か〉に］《そっけなくする》とは、
   ［〈誰か〉に］{親切にしない・冷たくする} ことをいう。
   例：昨日は忙しかったので、遊びに来た友達にそっけなくしたら怒って帰ってしまった。

この例の見出語は【そっけない】、定義 1 と 2 のターゲットは《そっけない》、定義 3 は《そっけなく》、定義 4 は《そっけなくする》である。各定義の FSD において、「とは、」までを**前件部**、それ以降を**後件部**と呼ぶ。前件部ではターゲットが文中でどのように用いられるか (形式と用法) を、後件部ではその場合の意味を言い換えによって提示するのが FSD の基本である。

我々は 2013 年より、このような FSD の記述法のガイドラインを定めるべく、実際の語に対して FSD を記述することを試行錯誤的に取り組んでいる。2014 年初頭の時点で、約 100 語について FSD を作成したが、2015 年 1 月の時点において、語数はそれほど増えていない。

表 1: FSD 作成の進行状況 (分数は、「作成済/作業対象」を表す)

分類	総数	か	き	く	け	こ	さ	し	す	せ	そ
イ形容詞	11/15		0/4			4/4		2/2	3/3	1/1	1/1
ナ形容詞	24/24					10/10	2/2	12/12			
形容詞的な名詞	13/13					13/13					
動詞	33/33					13/13	20/20				
副詞	32/35	10/11	5/5	2/2	1/1	3/3	3/5	3/3	4/4		1/1
サ変名詞	3/42					3/42					
カテゴリ的な名詞	0/9					0/9					
動詞由来の名詞	0/11					0/11					
その他の名詞	0/18					0/18					
総数	116/200										

表 1 に、現時点までに作成作業に着手した見出語の数 (分母) と、いちおう作業が終了したと考えている見出語の数 (分子) を示す。この表に示すように、これまでに FSD の記述法がほぼ固まってきたのは、形容詞、動詞、副詞の 3 品詞である。

実際の作業は、次のような手順で進めている。

1. 旧日本語能力試験の出題基準 [3] の語彙表の「カ行」と「サ行」の語のうち、旧 1 級に位置付けられている内容語を選ぶ。
2. その語の意味・用例を、既存の辞書[1]を参考に調べる。
3. これと並行して、コーパスの用例を調べる。
4. これらの調査結果に基づいて、採用すべきターゲット (形式) と前件部 (文中でどのように使われるか＝用法) を定め、それに対する後件部 (言い換え) を記述する。後件部は、できるだけ旧 2 級以下の語彙で記述する。
5. これらの作業を通して、適宜、FSD 作成ガイドラインを更新・修正するとともに、それまでに作成した FSD を見直して修正する。

用例は、主に『現代日本語書き言葉均衡コーパス (BCCWJ)』を、NINJAL-LWP for BCCWJ (以下 NLB[2]と略記) を用いて調べる。NLB で得られる用例が少ない (1,000 件未満) 場合は、『筑波ウェブコーパス (TWC)』を、NINJAL-LWP for TWC (以下 NLT[3]と略記) を用いて調べる。これらのコーパス検索ツールは、名詞や動詞などの内容語の共起関係や文法的振る舞いを網羅的に表示できるため、FSD 作成のための用例調査に適している。

原則として、コーパス中の出現頻度の高い形式と用法を、前件部として採用する。見出し語によっては、基本形の用法よりも他の活用形の用法の方が高い頻度で用いられるものがある。例えば、「しぶとい」について NLT で調査すると、全 908 件のなかで、出現頻度が高い形式・用法は、次のような形式である。

a. しぶとい＋名詞　187 件
b. 名詞＋ガ格の助詞＋しぶとい　113 件
c. しぶとく＋動詞　387 件

[1] 主に、岩波国語辞典 (第 7 版新版)、明鏡国語辞典 (第 2 版)、三省堂例解小学国語辞典 (第 3 版)、三省堂国語辞典 (第 7 版)
[2] 国立国語研究所、Lago 言語研究所『NINJAL-LWP for BCCWJ』(http://nlb.ninjal.ac.jp)
[3] 筑波大学、国立国語研究所、Lago 言語研究所『NINJAL-LWP for TWC』(http://corpus.tsukuba.ac.jp)

この頻度傾向に従い、「しぶとい」という基本形の形式・用法 (定義 1) 以外に、「しぶとく」という連用修飾用法 (定義 2) を採用する。定義 1 の前件部ではガ格となる名詞を、定義 2 では被修飾用言を、それぞれどのように記述するかが問題となる。これらの記述は、上記の形式・用法の b の名詞、および、c の動詞として、どのような語が現れやすいかを観察して定める。もし、MI スコア (相互情報量＝結びつきの強さの指標) が高い語があった場合は、慣用表現とみなすかどうかも判断する。以下に、「しぶとい」に対する現在の FSD を示す。

【しぶとい】 [イ形容詞]

1. 〈人・物〉が《しぶとい》とは、
   〈人・物〉がかんたんに {負けない・諦めない・死なない} 性質だということ。
   例：彼はしぶといので、なかなか負けを認めませんよ。
2. ［〈人・物〉が］《しぶとく》〈生きる・残る・続けるなど〉とは、
   ［〈人・物〉が］［困難な状況でも］諦めないで〈生きる・残る・続ける〉こと。
   例：けがや病気にも負けず、しぶとく生き残っています。

同一のターゲットに対しても、意味や用法が異なると判断すれば、複数の語義を立てる (たとえば「そっけない」の定義 1 と 2)。特に動詞の語義認定については注意を要する。これについては、5 節で述べる。

上で述べたように、FSD の後件部は、できるだけ旧 2 級以下の語彙で記述する方針を取っているが、語の定義に必要と考えられる旧 1 級・級外の語彙も存在する。具体的には、次のようなものがある。

1. FSD の記述に欠かせない語。例：「人々」、「発話」
2. 動詞 (の連用形) が名詞化したもの。例：「動き」、「頼り」、「思い」、「考え方」、「やり方」
3. 典型的な補足語・被修飾語を記述するための語。例：「行為」、「対応」、「保つ」、「乱れる」、「耐える」、「デザイン」

これは、FSD 作成ガイドラインに列挙するようにしている。

以下では、副詞、形容詞、動詞の順に、昨年度からの進捗を述べる。

2　副詞

日本語では、述語を修飾する語のうち、活用しないものを副詞に分類する。単に用言 (述語) を修飾するだけの副詞の場合、前件部には典型的な被修飾語 (用言) を示せばいい。以下に具体例を示す。

【ことごとく】 [副詞]

1. ［〈誰か〉が〈何か〉を］《ことごとく》〈行う：否定する・消し去る・変えてしまうなど〉とは、
   ［〈誰か〉が〈何か〉を］少しも残さず全部〈行う〉ことをいう。
   例：迷惑メールをことごとく無視する。

しかしながら、特定の表現や活用形と呼応する場合は、それを示すような工夫が必要となる。一方、オノマトペに分類される副詞においては、後件部の記述が問題となる。

2.1 他の表現と呼応する副詞

他の表現と呼応する副詞は、どのような表現と典型的に呼応するかを、できるだけ FSD の前件部に含めることとする。具体的には、次のような語が対象となる。

1. 述語の否定形と呼応する「程度の副詞」。例：「さほど」
2. 文末の「ムード」の表現と呼応する「陳述の副詞」
   - 概言 (真とは断定できない知識を述べるムード) と呼応するもの。例：「さぞ」
   - 比況 (ある事態を性質の類似した別の事態で特徴づけるムード) と呼応するもの。例：「さも」
   - 従属節において条件・譲歩の表現と呼応するもの。例：「仮に」

FSD の記述例を以下に示す。(一部の定義を省略した。)

【さほど】 [副詞]

1. 《さほど》〈どうではない：難しくない・大きくない・悪くないなど〉とは、
   {たいして・あまり}〈どうではない〉ということ。
   例：母はここからさほど離れていないところに住んでいます。
2. 《さほど (の)》〈差・影響・意味など〉は〈ない〉とは、
   特に大きな〈差・影響・意味〉が〈ない〉ということ。
   例：５分くらいの遅刻なら、さほど問題はないでしょう。

【さぞ】 [副詞]

1. ［〈誰か〉は］《さぞ》〈どうする・どうだ：驚く・喜ぶ・困る・疲れる・辛い・嬉しい・大変・幸せなど-｛だろう・にちがいないなど｝〉とは、
   ［〈誰か〉は］{きっと・どんなにか}〈どうする・どうだ-｛だろう・にちがいない｝〉ということ。
   例：突然の知らせに、彼はさぞ驚いたにちがいない。長い旅でさぞ疲れたでしょう。
2. ［〈何か〉は］《さぞ》〈美しい・美味しい・立派など-｛だろう・にちがいない｝〉とは、
   ［〈何か〉は］{きっと・どんなにか}〈美しい・美味しい・立派-｛だろう・にちがいない｝〉ということ。
   例：春になって桜が咲いたら、さぞきれいでしょうね。

【さも】 [副詞]

2. ［〈人〉が］《さも》〈どうである：正しい・重要だ・知っているなど-(かの) ように〉〈行う：言う・扱うなど〉とは、
   ［〈人〉が］まるで〈どうである-(かの) ように〉〈行う〉こと。
   例：この本は間違ったことをさも正しいように書いてある。彼は初めて聞いたことでも、さも知っていたかのように話す。

【仮 (に)：かり (に)】 [副詞・形容詞的な名詞]

3a. 《仮に》〈S-｛なら・だったら・だとすれば｝〉〈S〉とは、
   もし〈S-｛なら・だったら・だとすれば｝〉〈S〉ということ。
   例：仮にこの案が実現したなら、たくさんの人が喜ぶだろう。

3b. 《仮に》〈S-｛ても・でも｝〉〈S〉とは、
   たとえ〈S-｛ても・でも｝〉〈S〉ということ。
   例：この案は仮に可能だとしても、ずいぶんお金がかかるだろうね。

なお、最後の【仮 (に)】の定義 3a と 3b の「S」は、そこに文が入ることを意味する。

2.2　オノマトペ副詞

日本語の副詞には、オノマトペに分類されるものが多数ある。擬音語や擬態語であるオノマトペは、明確な意味を持ち難い。このようなオノマトペ副詞の後件部の記述では、次の原則を適用する。

1. 原則として、オノマトペ以外の表現への言い換えを示す。例：「きっぱり (定義 1)」
2. 必要に応じて、オノマトペへの言い換えを補う。例：「きっぱり (定義 1)」
3. やむを得ず、言い換えをオノマトペで示す場合でも、できるだけ他の言い換え、説明等を併記する。例：「きっぱり (定義 2)」、「がっちり (定義 1, 2)」

以下に、FSD の記述例を示す。

【きっぱり】 [副詞]

1. ［〈人〉が］［〈何か〉を］《きっぱり (と)》〈やめる・別れる・捨てるなど〉とは、
   ［〈人〉が］［〈何か〉を］{完全に・すっかり}〈やめる・別れる・捨てる〉こと。
   例：彼はタバコをきっぱりとやめた。
2. ［〈人〉が］［〈誰か〉に］《きっぱり (と)》〈言う・断る・否定するなど〉とは、
   ［〈人〉が］［〈誰か〉に］［強い意思を示して］はっきり〈言う・断る・否定する〉こと。
   例：何度も誘われたが、きっぱりと断った。

【がっちり】 [副詞]

1. ［〈人〉が〈物〉を］《がっちり (と)》〈組む・固めるなど〉とは、
   ［〈人〉が〈物〉を］動かないようにしっかり〈組む・固める〉こと。
   例：夏休みの予定をがっちり組んだ。

2a. ［〈人〉が〈何か〉を］《がっちり (と)》〈つかむ・支える・押さえるなど〉とは、
   ［〈人〉が〈何か〉を］{しっかり・確実に}〈つかむ・支える・押さえる〉こと。
   例：彼は客の心をがっちりとつかんだ。

2.3　オノマトペ発話

ターゲットが完結した発話となる用法は、次のような定義文型で記述する。

「《ターゲット》。」という発話は、　(前件部)
「〜」という気持ちを［○○的に］いう表現。　(後件部)

オノマトペの中には、このような用法を持つものがある。以下に FSD の記述例を示す。

【がっくり】 [副詞]

3. 「《がっくり》。」という発話は、
   「失望して元気がなくなった、がっかりした」という気持ちを［ひと言で］いう表現。

3　形容詞

形容詞は、大きく属性形容詞と心情 (感情) 形容詞に分けることができる [4]。

属性形容詞は人、行為、物事などの性質や特徴を表す。一般に、連体修飾用法、述語用法、連用修飾用法が可能である。用法による意味の変異が小さい場合は、代表的な用法のみを示す。以下に FSD の具体例を示す。

【すばしこい】 [イ形容詞]

1. 《すばしこい》〈動物・人・行動〉とは、
   とても速い〈動物・人・行動〉のこと。
   例:リスはすばしこいから捕まえるのがむずかしい。その子どもはとてもすばしこく逃げ回った。

これに対して、心情形容詞は、話し手 (書き手) の心情を表す表現であり、典型的には述語用法で用いられる。その主体は、原則として話し手である。心情形容詞では、以下の定義文型を基本とし、主体 (ガ格) は明示的に記述しない。

［付加的説明］《心情形容詞》とは、　(前件部)
〜と｛思う・感じる｝こと。　(後件部)

以下に FSD の具体例を示す。

【心強い：こころづよい】 [イ形容詞]

1. ［〈何か〉があって・〈誰か〉がいて］《心強い》とは、
   ［頼ることのできる｛〈何か〉があって・〈誰か〉がいて｝］安心だと｛思う・感じる｝こと。
   例：隣が病院なので、いざという時心強い。
2. 《心強い》〈味方〉とは、
   頼りになる〈味方〉のこと。
   例：友達に相談すれば、きっと心強い味方になってくれるでしょう。
3. ［〈誰か〉が〈何か〉を］《心強く》〈思う・感じる〉とは、
   ［〈誰か〉が〈何か〉を］頼もしく〈思う・感じる〉こと。
   例：母は息子たちの言葉を心強く感じた。

この例の定義 1 が、上記の文型を用いた定義である。連体修飾用法 (定義 2) や連用修飾用法 (定義 3) は、この限りではない。

4　動詞

動詞に対しては、どれだけの語義を認定して定義として独立させるかと、定義の提示順について検討した。

4.1　語義の認定

語義の認定に対しては、次のような原則を採用することとした。

1. 語義の認定では、主に、ガ格、ヲ格、ニ格の補足語の類別により語義を分ける。
   (a) 自動詞は、主に、ガ格の補足語の類別で語義を分ける。例:「冴える」、「こじれる」
   (b) 他動詞は、ヲ格やニ格の補足語の類別で語義を分ける。例:「差し出す」
2. 補足語によってターゲットの言い換え表現が (微妙に) 異なる場合は、つぎのいずれかの方法を採用する。
   (a) まとめて (抽象的な) 説明的後件部を作る。例:「こじれる (定義 1)」
   (b) 語義を分ける。例:「冴える (定義 1–4)」
   (c) 慣用語句として扱う。例:「差し出す (定義 4)」
3. 典型的な語義と発展的な語義 (比喩など) は区別する。例:「込める」

以下に FSD の記述を示す。

【冴える：さえる】 [動詞]

1. 〈色・光・音・味など〉が《冴える》とは、
〈色・光・音・味〉が｛鮮やかになる・輝く・はっきりする｝こと。
例：気温が下がると紅葉の色がいっそう冴える。

2. 〈感覚・意識：目・耳・頭・脳・勘など〉が《冴える》とは、
〈感覚・意識〉が｛鋭い・はっきりした｝状態になること。
例：夜になっても目が冴えて眠れない。今日は勘が冴えているので全問正解できそうだ。

3. 〈人〉の〈技・力：演出・推理・腕・判断力など〉が《冴える》とは、
〈人〉の〈技・力〉が｛うまく働く・発揮される｝こと。
例：職人の技が冴える日本料理。

4a. 〈状況・状態：天気・成績・株価・気分・色など〉が《冴えない》とは、
〈状況・状態〉が｛はっきりしない・良くない・鈍い｝こと。
例：高い山に登ったのに天気が冴えなくて残念だ。成績が冴えなかった人は、次は頑張ってください。日当りが悪いと葉の色が冴えない。

4b. 《冴えない人》とは、
｛ぱっとしない・鈍い感じの｝人のこと。
例：父は有名な役者ですが家では冴えないおじさんです。

4c. 〈誰か〉の《｛顔色・表情｝が冴えない》とは、
〈誰か〉が｛顔の色が良くない・元気がない｝ということ。
例：彼は表情が冴えないね、風邪でもひいたのかな。

【こじれる】 [動詞]

1. 〈話し合い・関係など〉が《こじれる》とは、
［人と人との意見が合わず］〈話し合い・関係〉が悪い方へ進んでしまうこと。
例：夫婦関係がこじれたので裁判所に相談する。

2a. 〈病気〉が《こじれる》とは、
〈病気〉がさらに重くなること。
例：風邪がこじれて肺炎になる。

2b. 〈問題〉が《こじれる》とは、
〈問題〉がさらに深くなること。
例：民族問題がこじれて争いがおこる。

【差し出す：さしだす】 [動詞]

1. ［〈人〉が］〈物・手など〉を《差し出す》とは、
［〈人〉が］〈物・手〉を相手の｛方に出す・前に置く｝こと。
例：記者が意見を求めてマイクを差し出してきた。

2. ［〈人〉が］〈何か〉を〈誰か〉に《差し出す》とは、
［〈人〉が］〈何か〉を〈誰か〉に｛与える・提供する・勧める｝こと。
例：彼は会社のために自分の財産をすべて差し出した。客にあたたかいコーヒーを差し出した。

3. ［〈人〉が］〈郵便など〉を《差し出す》とは、
［〈人〉が］〈郵便〉を送る［手続きをする］こと。
例：昨日差し出した郵便は、明日には届くと思います。

4. ［〈人〉が〈誰か〉に］《｛手・右手｝を差し出す》とは、
［〈人〉が〈誰か〉に］握手を求めるということ。
例：部屋に入ると男がにっこり笑って右手を差し出した。

【込める：こめる】 [動詞]

1. 〈人〉が〈銃など〉に〈玉〉を《込める》とは、
〈人〉が〈銃〉に〈玉〉を詰めること。
例：銃に弾丸を２発込めた。

2. 〈人〉が〈何か〉に〈心・気持ち・意味など〉を《込める》とは、
〈人〉が〈何か〉を工夫して〈心・気持ち・意味〉が伝わるようにすること。
例：この歌には恋人への愛が込められている。

この例に示したように、定義は、1, 2, 3 として区別するだけでなく、意味が似たものを、2a, 2b のようにグループ化することを認める。

4.2 語義の提示順

語義の提示順は、以下の原則を採用することとした。

1. 語義の抽象度が高いものから具体的なものへ。
2. 補足語の内容語が限定されないものから限定されるものへ。
3. 補足語が少ないものから多いものへ。
4. ターゲットの形式：見出語と同じもの、活用形、特殊なもの (否定表現など)、慣用句的表現の順

以下に FSD の記述例を示す。

【裂ける：さける】 [動詞]

1a. 〈何か：物・腹・頭・血管・膜・幕・布・壁など〉が《裂ける》とは、
〈何か〉が［線状の傷が入って］{切れる・破れる・壊れる} こと。
例：車のタイヤが裂けてしまったので交換する。脳の血管が裂けて出血する病気。強い風で幕が裂けてしまった。

1b. 〈傷・口・穴など〉が《裂ける》とは、
〈傷・口・穴〉が切れて広がること。
例：手術の後に激しく動くと傷が裂けてしまいますよ。

1c. 〈岩・地面など〉が《裂ける》とは、
〈岩・地面〉が［線状に］{割れる・分かれる} こと。
例：地震の時、ここの地面が二つに裂けたのです。

1d. 〈筋肉・木など〉が《裂ける》とは、
〈筋肉・木〉の筋に沿って {切れる・割れる} こと。

2. 《口が裂けても》〈言えない・言ってはいけない〉とは、
{何があっても・絶対に}〈言えない・言ってはいけない〉ということ。
例：病気の母にお金が欲しいとは口が裂けても言えなかった。

3. 《胸が裂ける》〈思い〉とは、
とても悲しくて辛い〈思い〉のこと。
例：事故の知らせを胸が裂ける思いで聞いた。

謝辞

本研究では、『現代日本語書き言葉均衡コーパス』と『筑波ウェブコーパス』を利用した。本研究は、JSPS 科学研究費基盤研究 (B) 「平易な日本語表現への工学的アプローチ」(課題番号 24300052) の助成を受けている。

参考文献

[1] 佐藤理史, 夏目和子. 新しい日本語辞書定義文型の策定に向けて. 第 5 回コーパス日本語学ワークショップ予稿集, pp. 153–160, 2014.
[2] John Sinclair, editor. *COBUILD Advanced Dictionary of Engish, 7th Edition.* National Geographic Learning, 2012.
[3] 財団法人日本国際教育協会国際交流基金. 日本語能力試験出題基準【改訂版】. 凡人社, 2002.
[4] 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.

コーパスコンコーダンサ 『ChaKi.NET』 のプロジェクト機能

浅原 正幸 (国立国語研究所) *
森田 敏生 (総和技研)

Project Functions on 'ChaKi.NET'

Masayuki Asahara (NINJAL)
Toshio Morita (Sowa Research Co.,Ltd.)

要旨

本稿では、ChaKi.NET の新しい機能であるプロジェクト機能を紹介する。プロジェクト機能は複数のテキストを可視化する機能である。文単位でアラインメントされたテキスト対を相互に可視化することが可能である。発表では、

- BCCWJ-Trans (BCCWJ に対する対訳付与) の複数言語の可視化
- BCCWJ 長単位・短単位の可視化
- BCCWJ の読み時間の可視化 (テキスト出現順と、視線走査順の 2 種類の分析)

についてデモを行う。

1. はじめに

本稿ではコーパスコンコーダンサ 『ChaKi.NET』 (Matsumoto et al. (2005)) の新しい機能であるプロジェクト機能について解説する。

コーパスに対して様々なレベルのアノテーションが施されている。形態論情報・係り受け構造を基本として、言語学的に多様なレベルのアノテーションを重ね合わせるためのデータ形式拡張 CaboCha フォーマット (松吉ほか (2014b)) が提案されている。このフォーマットでは、アノテーションを文字列範囲・リンク（有向・無向）・同値類に抽象化した。ChaKi.NET はこの 3 つの種類のアノテーションを可視化する機能を有している。

一方、複数のレイヤーのテキストからなるコーパスなどが整備されつつある。例えば『現代日本語書き言葉均衡コーパス』(以下 BCCWJ; Maekawa et al. (2014)) は長単位と短単位の 2 つの形態論情報を保持しており、基底となる形態論情報を 2 つ有している。自然言語処理の分野では機械翻訳のための訓練データ・評価データとして対訳コーパスが整備されている。また歴史コーパスや方言コーパスの場合、現代語訳や標準語訳が整備されることが考えられる。

ChaKi.NET のプロジェクト機能は 2 つ以上のレイヤーからなるコーパスを格納し、その 2 つのレイヤーを可視化する機能である。以下では、プロジェクト機能の概要について解説するとともに、活用例として、対訳コーパス・BCCWJ の長単位/短単位・読み時間の可視化などについて紹介する。

* masayu-a@ninjal.ac.jp

2.　プロジェクト機能の概要

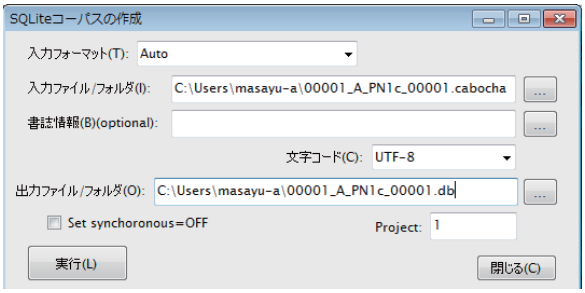
ChaKi.NET は形態素情報を格納する word テーブルや各アノテーションタグを階層化することができるプロジェクト (Project) という概念が存在する。デフォルトでは全ての要素が ID=0 のプロジェクトに存在するが、このデフォルトプロジェクト以外のプロジェクトを作成することにより、形態論情報やアノテーションをプロジェクト毎にグルーピングすることができる。

2.1　プロジェクトを指定したデータの格納・検索・可視化

以下ではプロジェクト機能の利用方法について概説する。

2.1.1　プロジェクトを指定したデータの格納

まずプロジェクトを指定したデータの格納方法について述べる。通常のデータ作成方法と同様に拡張 CaboCha フォーマット (松吉ほか (2014b)) もしくは CoNLL-U フォーマット(1)のデータを準備する。[ツール]→[SQLite コーパスの作成] よりコーパス作成の画面を立ち上げ、[出力ファイル/フォルダ] に既存の sqlite db ファイルを指定する。右下の [Project:] の値をデフォルトの 0 以外の値を指定することにより、既存のデータと別のレイヤーのデータを格納することができる。



2.1.2　プロジェクトを指定したデータの検索

メインツールバーの右端にある “Proj” 欄に Project ID を指定すると、検索時に Project ID に合致する結果のみを得ることができる。

検索結果に対して DependencyEdit を行うとき、検索に用いた Project ID、すなわちその結果が属している Project ID が DependencyEdit に伝えられる。その DependencyEdit において行われる編集 (アノテーションタグの追加・削除等）は、その Project に対して行われる。

2.1.3　2 画面モード

メニュー [表示] → [KWIC 画面を分割] により、KWIC View が上下 2 画面に分割される。

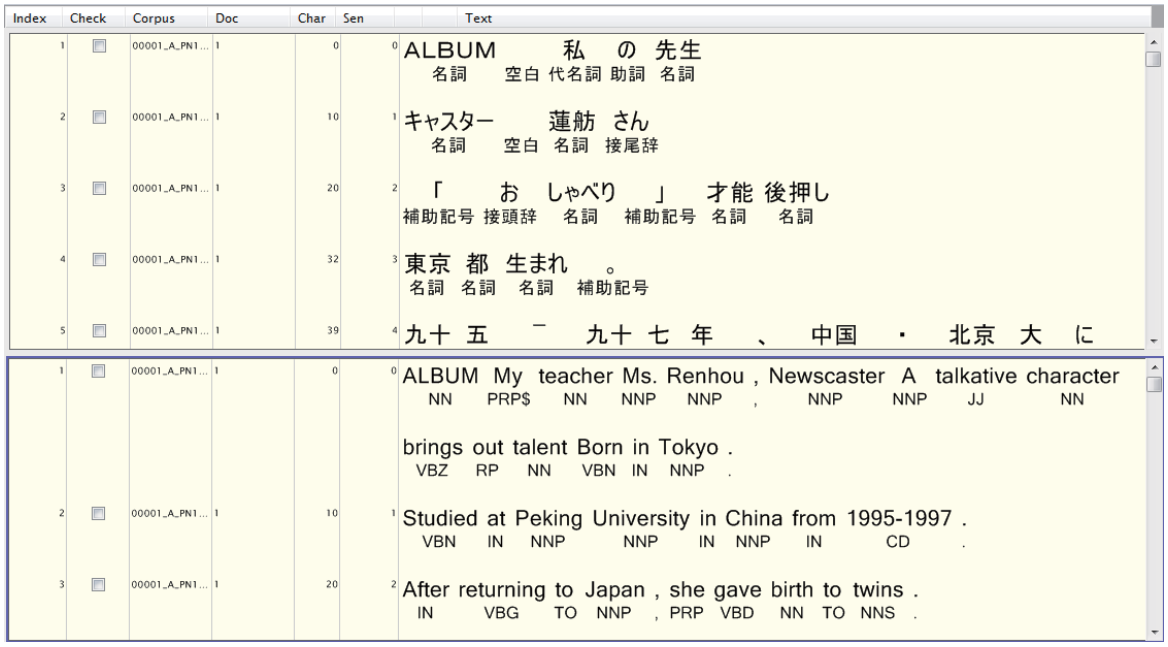
2 画面モード時に上下いずれかの KWIC View をクリックすると青い縁取りが現れる。これは、その View がカレント View であることを示す。検索コマンドの結果はカレント View に表示されるので、カレントを切り替えながら異なる検索条件で検索を行うと、検索結果同士を

(1) http://universaldependencies.github.io/docs/format.html　但し、ChaKi.NET は CoNLL-U フォーマットのうち Multiword token 表現については対応していない。

上下画面で比較することができる。

真中のスプリッター（分割線）をマウスドラッグすることにより分割位置を調整できます。元の 1 画面表示に戻すにはもう一度同じコマンドを実行する。

典型的な使い方として、上の View に Project=0 の検索結果を、下の View に Project=1 の検索結果を表示することにより複数の Project 内容を対比することなどが想定されている。



2 画面モード時に、片側の View のカレント行（1 行の全体がグレー背景になっている状態）が変更されると、その行の文番号と同一の文番号の行がもう一方の View にも存在すれば、その行が自動的にカレント行となります。この時、行が見えていない状態であれば見えるようにスクロールも行われます。Up, Down, PageUp, PageDown キーによりカレント行を変更した場合もこの自動同期が働きます。

一方、この同期機能はスクロールバーを操作するだけでは動作しません。これは、文の順序が上下の View で必ずしも一致しているとは限らないため（ソートを行った場合など）です。

2.2　形態素間マッピング

2.2.1　形態素間マッピングのインポート

ChaKi.NET には、Word と Word との間の対応関係を示すための特別なテーブル “word_word” が存在しており、対応する Word 間の対応関係を格納することができるようになっている。このテーブルは、異なる Project 間で Word と Word との対応関係を記述するのに使用することが想定されている。例えば、

- Project 0 に短単位での Word の並びが格納されていて、他の Project には長単位などそれとは異なる単位の Word の並びが格納されている
- Project 0 に日本語、Project 1 に英語というように対訳データを格納し、対応する Word をマークアップする

- Project 0 に通常の語順での Word の並びが格納されていて、他の Project には「読み順」などそれとは異なる語順の Word の並びが格納されている

などの使い方が考えられる。

Word 間マッピングをインポートするコマンドは、コマンドラインから “ImportWordRelation.exe” を実行する。下記に Usage を示す。

```
Usage: ImportWordRelation [Options] <InputFile> <Output>
Options (default):
  [-C] Do not pause on exit (false)
  [-b] Make relations bi-directional (false)
  [-a] Do not clear the mapping table; append mode (false)
InputFile - TSV File
Output    - .db file for SQLite / .def file for Others
```

入力は Project, Sentence, WordNo の 3 つ組を基本として、From-word, To-word を横に並べた Tab-separated 形式となる。すなわち、各行は、

From-word の Project・From-word の Sentence No・From-word の Word No・To-word の Project・To-word の Sentence No・To-word の Word No というカラムから成る。

関係は、デフォルトでは From-word から To-word の一方向だが、“-b” オプションを付けることで双方向とすることも可能である。この場合、1 つの入力それぞれについて、方向を逆にした 2 つのレコードが挿入される。

以下に日英対訳の場合の入力ファイルの例を示す。

日本語側入力ファイル (拡張 CaboCha フォーマット):

```
* 0 1D 0/0 0
ＡＬＢＵＭ　　名詞, 普通名詞, 一般,*,*,*,*,*,*, ＡＬＢＵＭ, ＡＬＢＵＭ
　　空白,*,*,*,*,*,*,*,*,*,*,
* 1 2D 0/0 0
私　　代名詞,*,*,*,*,*,*,*,*, 私, 私
の　　助詞, 格助詞,*,*,*,*,*,*,*, の, の
* 2 -1Z 0/0 0
先生　　名詞, 普通名詞, 一般,*,*,*,*,*,*, 先生, 先生
#! SEGMENT_S Apposition 0 5 ""
#! SEGMENT_S Apposition 6 10 ""
#! GROUP_S Apposition 0 1  ""
EOS
* 0 1D 0/0 0
キャスター　　名詞, 普通名詞, 一般,*,*,*,*,*,*, キャスター, キャスター
　　空白,*,*,*,*,*,*,*,*,*,*,
* 1 -1Z 0/0 0
```

```
蓮舫　　名詞,固有名詞,人名,一般,*,*,*,*,*,蓮舫,蓮舫
さん　　接尾辞,名詞的,一般,*,*,*,*,*,*,さん,さん
EOS
```

英語側入力ファイル (CoNLL-U フォーマット):

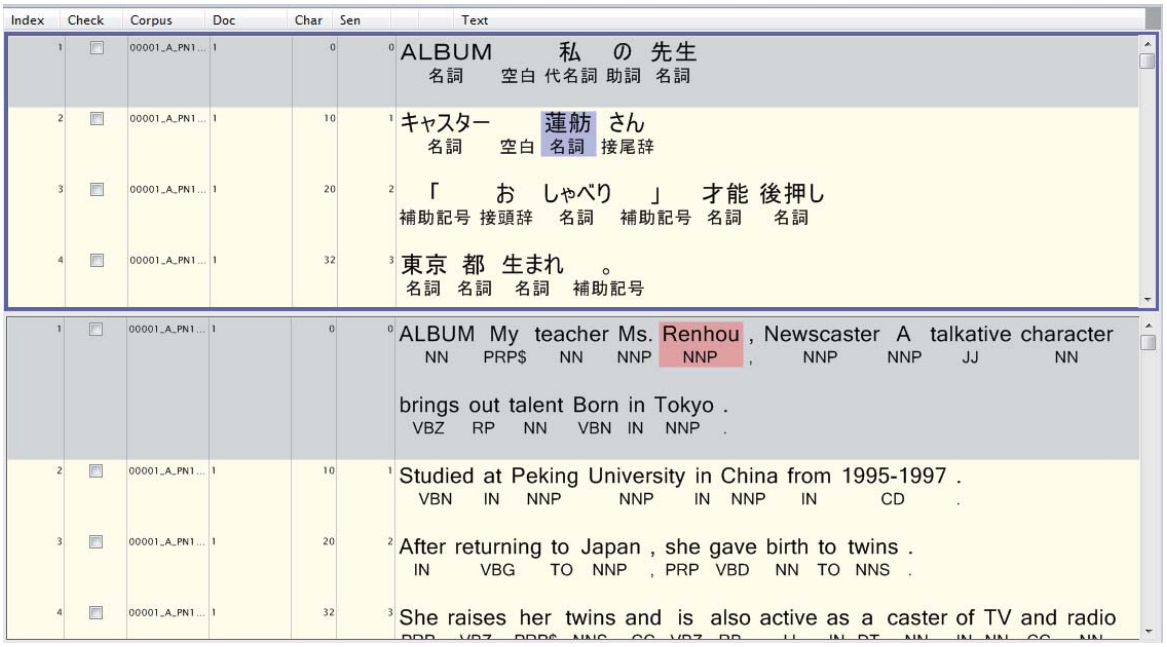
```
1   ALBUM   _   NN   NN   _   11   tmod   _   _
2   My   _   PRP$   PRP$   _   5   poss   _   _
3   teacher   _   NN   NN   _   5   nn   _   _
4   Ms.   _   NNP   NNP   _   5   nn   _   _
5   Renhou   _   NNP   NNP   _   11   nsubj   _   _
6   ,   _   ,   ,   _   5   punct   _   _
7   Newscaster   _   NNP   NNP   _   10   nn   _   _
8   A   _   NNP   NNP   _   10   nn   _   _
9   talkative   _   JJ   JJ   _   10   amod   _   _
10   character   _   NN   NN   _   5   conj   _   _
11   brings   _   VBZ   VBZ   _   0   null   _   _
12   out   _   RP   RP   _   11   prt   _   _
13   talent   _   NN   NN   _   11   dobj   _   _
14   Born   _   VBN   VBN   _   13   partmod   _   _
15   in   _   IN   IN   _   14   prep   _   _
16   Tokyo   _   NNP   NNP   _   15   pobj   _   _
17   .   _   .   .   _   11   punct   _   _
```

word_word 対応ファイル (ImportWordRelation.exe 入力ファイル):

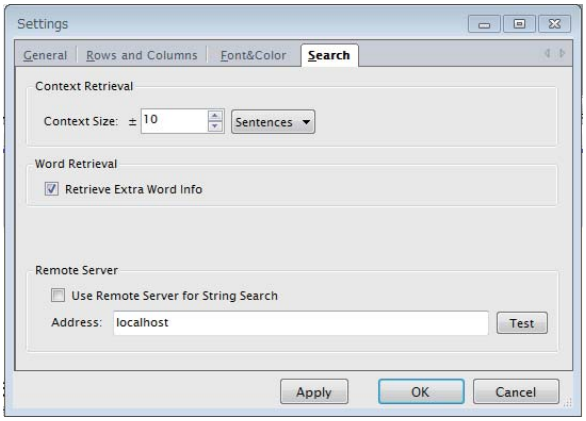
```
0       0       0       1       0       0
0       0       2       1       0       1
0       0       3       1       0       1
0       0       4       1       0       2
0       1       0       1       0       6
0       1       2       1       0       4
0       1       3       1       0       3
```

2.2.2　形態素間マッピングの可視化

形態素間マッピングの情報は KwicView の 2 画面モード上で可視化することができる。



ImportWordRelation.exe により Word-Word マッピングをコーパスにインポートしてある場合は、マッピングの From 側に一致する Word 上にマウスを置いたときに Word 背景が青色となり、同時に To 側に対応する Word の背景が自動的に赤色になる。つまり、青色背景の Word から赤色背景の Word へのマッピングが存在することを、Word 上にマウスを持っていくことにより確認することができる。 但し、word-word マッピング情報は、設定ダイアログ（メニューの [オプション]→[設定] で表示されるダイアログ）の [Search] タブにおいて、[Retrieve Extra Word Info] が ON になっていないと検索時に読み込まれないことに注意すること（デフォルトでは OFF）。



3. 活用例

3.1 BCCWJ-Trans 対訳の可視化

前節までの例では、BCCWJ に対する対訳 BCCWJ-Trans に基づいて紹介した。表 1 に BCCWJ-trans の概要について示す。今回、デモ用に形態素単位の対応を 1 サンプルにのみ付与したが、現在のところ全データに対して形態素単位の対応が付与されているわけではない。

今後、形態素単位の対応を付与していきたい。

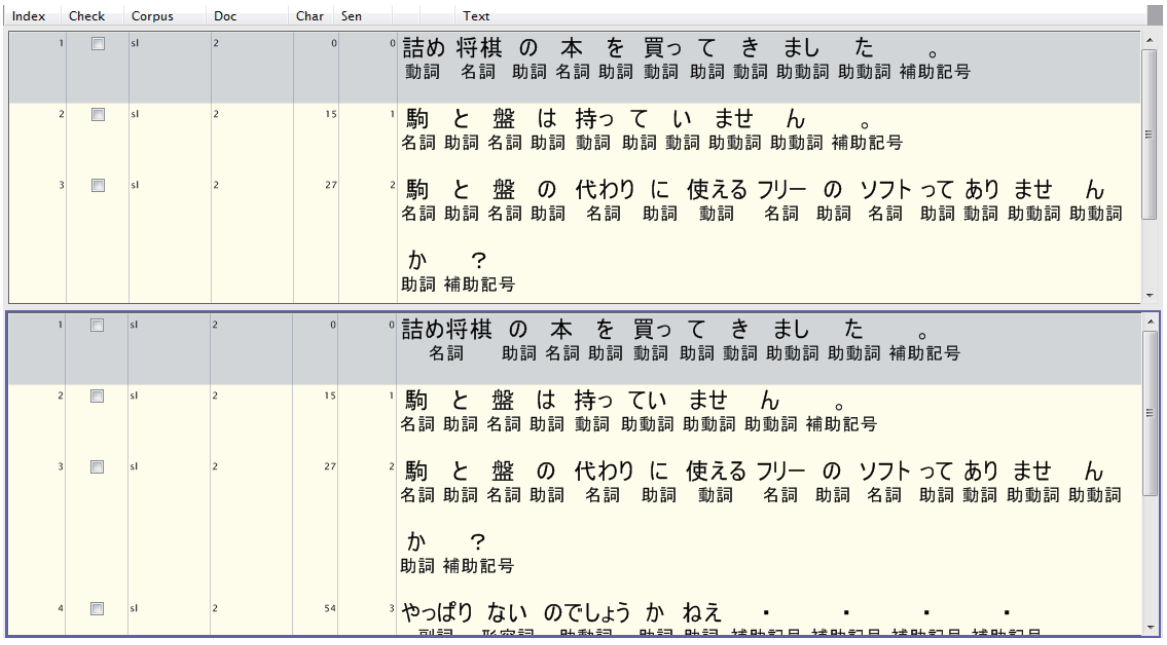
表 1　BCCWJ-Trans の概要

言語	文書数	文数	下訳	摘要
英語	6	319	有	OY 1, OC 1, PN 1, PB 1, PM 1, OW 1
中国語 (簡)	6	319	有	OY 1, OC 1, PN 1, PB 1, PM 1, OW 1
イタリア語	16	436	無	OY 6, OC 6, PN 1, PB 1, PM 1, OW 1
インドネシア語	10	337	無	OY 3, OC 3, PN 1, PB 1, PM 1, OW 1

文数は日本語側のもの。文書はアノテーションの優先順位順に選択。
OY “ブログ”, OC “知恵袋”, PN “新聞”, PB “書籍”, PM “雑誌”, OW “白書”。

3.2　BCCWJ の長単位・短単位の可視化

BCCWJ の DVD には長単位・短単位の 2 種類の形態素単位の形態論情報が付与されている。これまでの ChaKi.NET はどちらか一方の形態素単位による検索しかできなかった。プロジェクト機能を用いて、長単位・短単位を別のプロジェクトに格納することにより、それぞれの形態論情報による検索・可視化が可能になる。下図は KwicView の 2 画面モードにより 2 つの形態素単位を可視化したものである。



3.3　BCCWJ の読み時間の可視化

現在 BCCWJ に対する読み時間の付与を進めており、さまざまな可視化手法について検討している (浅原・森田 (2013), 浅原ほか (2014a))。プロジェクト機能を用いることにより、視線走査装置によって得られた読み時間の可視化をすることができる。具体的には読んでいるテキストの線形順序と視線が走査した形態素順序の 2 種類の順序を別のプロジェクトに格納する。以下では、視線走査実験結果の可視化について紹介する。

視線走査装置は、被験者がディスプレイ画面上のどの文字を注視しているのかを取得することができる機材である。この視線走査装置を用いて、視線停留箇所と停留時間を計測することにより、読文速度を取得することができる。視線走査装置として、SRResearch 社の EyeLinkCL シリーズを用いる。テキストは横書き、等幅フォントを用い、5 行単位で呈示する。自己ペース読文法と同様に、1 文書毎に内容を問う Yes/No Question に回答させる。

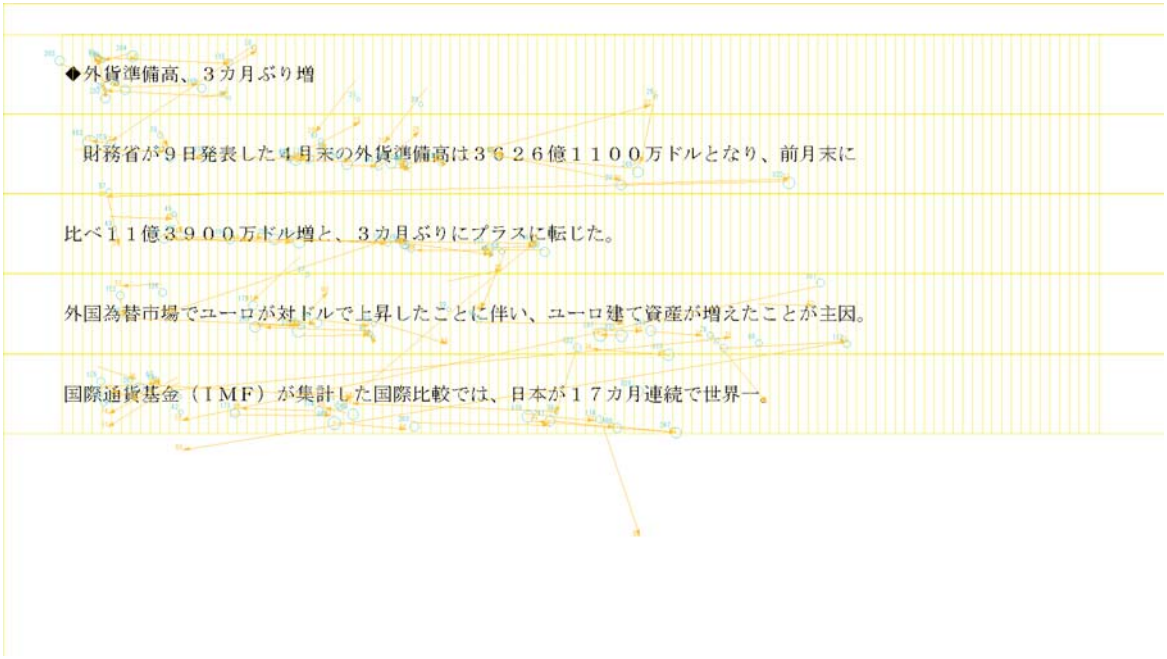


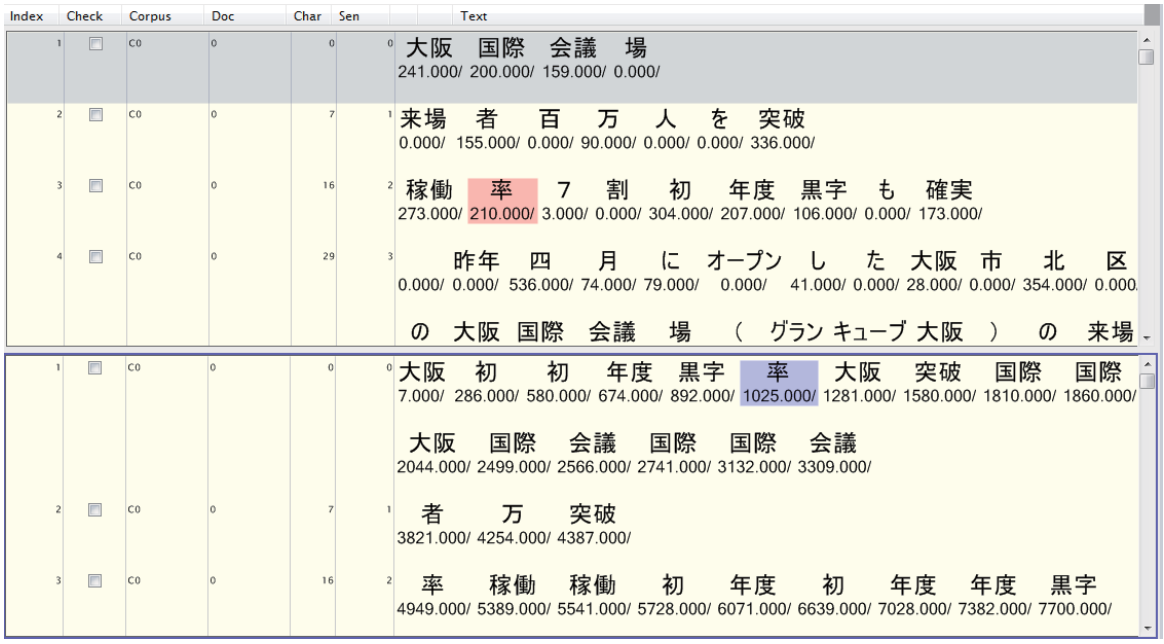
図 1　視線走査実験結果

図 1 に視線走査実験結果を示す。呈示する各文字の 1/2 幅毎に interest area (図中黄色の grid で表示) と呼ばれる領域を設定する。各 interest area 毎に視線停留箇所と停留時間、サッケード眼球運動の通過などが付与される。この interest area が設定されている半文字単位の情報に BCCWJ に付与されている短単位形態論情報、長単位形態論情報、文節境界情報を重ね合わせることにより、それぞれの単位での分析を行う。この実験法で得られるデータは読み戻しができ、かつ周辺視野により隣接する形態素・文節が読まれることもあり、全ての文節が必ず一度は読まれるわけではない。被験者は自由に読み戻し・読み飛ばしが可能である。元文書の語順に沿って分析するために次の指標が用いられる。

- First pass time
  最初に「分析単位」に視線が停留してから、他の「分析単位」に出るまでの間の視線停留時間の合計
- Total time
  「分析単位」内の視線停留時間の合計
- Regression path time
  最初に「分析単位」に視線が停留してから、より右側 (もしくは下側) の「分析単位」に

出るまでの間の視線停留時間の合計 (左側 (もしくは上側) へ停留している停留時間は累計される)

以下の図は、KwicView の上画面に読んでいるテキストの線形順序の形態素に読み時間の First pass time を付与したものを、下画面に視線走査順序の形態素に実験開始時刻を 0.000 ミリ秒とした場合の視線停留開始時刻を示したものである。下画面側の形態素（青地）にマウスカーソルを合わせることにより、当該形態素のテキスト中の位置（赤字）を示す。



4. おわりに

本稿では、コーパスコンコーダンサ ChaKi.NET のプロジェクト機能の概要と活用事例について紹介した。プロジェクト機能は ChaKi.NET Version 2.8 Revision 496 以降[(2)]で利用可能である。

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」、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・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予稿集, pp. 249–256.

浅原正幸・池本優・森田敏生 (2014a).「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予稿集, pp. 39–48.

(2) http://sourceforge.jp/projects/chaki/releases/

松吉俊・浅原正幸・飯田龍・森田敏生 (2014b).「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予稿集, pp. 223–232.

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.

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). “Chaki: An annotated corpora management and search system.” *Proc. of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

国語教育のための「常用漢字表」語例の検討

河内　昭浩（安田女子大学文学部）†

A Study of "*Joyo kanji* table" Vocabulary for Japanese Language Education

Akihiro Kawauchi (Faculty of letters, Yasuda Women's University)

要旨

常用漢字表には漢字の使用の目安として語例が掲げられている。常用漢字表は 2010 年に改定されたが，その際に語例は検討されていない。語例の，現代社会における目安としての適否を，コーパス・データをもとに検討する。使用するコーパスは「現代日本語書き言葉均衡コーパス」と「教科書コーパス」である。また，それぞれのコーパスにおける語彙のレベル判定を行う。その上で，重要度，日常度，教科書レベル，文化度といった観点で語彙を整理し，国語教育において指導すべき語彙の選定について言及した。例えば日常度が低く重要度の高い語は，学校で指導すべき語であると言える。また日常度や重要度の低い語は，現在の常用漢字表の語例としては不適切かもしれない。しかし一方で，文化の継承という観点では，学校での指導が必要な語であるとも考えられる。このように，国語教育の観点から常用漢字表の語例を考察した。

１．　はじめに

常用漢字表の「表の見方及び使い方」には，「例欄には，音訓使用の目安として，その字の当該音訓における使用例の一部を示した」とある。「目安」は，常用漢字表の性格をもっとも端的に表している語である。「一般の社会生活において，現代の国語を書き表す場合の漢字使用の目安」（常用漢字表「前書き」）として常用漢字表はある。したがって例欄に示す語が同様の「目安」という性格を有するのは当然のことである。

1981 年に制定された常用漢字表の漢字 1945 字は，書籍の出現頻度等のデータをもとに 2010 年に改定され，2136 字（196 字追加 5 字削除）となった。その際，既存の常用漢字の語例は変更されていない。約 30 年の時を経て，語例の中には現在はほとんど使用されていないものも少なくない。現代社会の使用の目安となるように，字種同様に語例も，客観的な資料をもとに検討，改定されるべきである。本発表はそうした問題意識のもとに，コーパス・データを用いて常用漢字表語例を複数の観点から検証するものである。

また常用漢字表において，学校教育における指導は「別途の教育上の適切な措置にゆだねる」（昭和 56 年国語審議会答申「常用漢字表」前文）とされている。つまり漢字の選定において教育的見地は対象外ということである。しかし，実際に常用漢字を学ぶ場は学校の国語の授業である。具体的には，常用漢字表の漢字のおよそ半数である 1006 字が，小学校学年別漢字表に示される漢字となる。さらに学習指導要領には，中学校終了までに「常用漢字の大体を読むこと」，高等学校では「常用漢字の読みに慣れ，主な常用漢字が書けるようになること」（国語総合）と記されている。またそもそも国語教育には，常に子どもの日常生活，社会生活に資する言語力を育成することが求められている。常用漢字表の，

† kawauchi@yasuda-u.ac.jp

一般の社会生活における目安という考え方と，国語教育の向かう先は方向を一としている。

そのように考えると，常用漢字表の語例は社会生活における漢字使用の目安として，同時に漢字指導のための用例として，国語教育において本来もっと注視されるべきものである。本発表では，国語教育における指導すべき語彙の選定という視点からも常用漢字表語例を眺望する。一般の社会生活における目安という観点では不適切と考えられる語例も，文化の継承という観点からは，常用漢字表に残すべき，あるいは学校で指導，継承すべきと判断できる語もある。教育の視点から見ることで，常用漢字表語例の価値の再認識ができると考えている。

２．　常用漢字表語例集の作成

語例の検討に当たり，まず常用漢字表語例集を作成した。

常用漢字表には，備考欄も含め延べ語数 9872 の語例が掲げられている。その内，「学校・社会対照語彙表」と照応が可能であった延べ語数 8544 の語例の，教科書並びに実社会での頻度等の調査を行い，表にまとめた。「学校・社会対照語彙表」は，「現代日本語書き言葉均衡コーパス」並びに「教科書コーパス」に出現する語彙の頻度やレベルの情報を一覧にしたものである。「学校・社会対照語彙表」から必要な情報を抽出し，加えて「教科書レベル」という新たな情報を付与して整理したのが本集である。「教科書レベル」は，田中（2011）にある，「現代日本語書き言葉均衡コーパス」各サブコーパスの，語彙レベルの設定方法と同じ手順で，発表者が独自に作成したものである。

順	字	種	語	語	品	級	全	級	国	数	理	社	外	技	芸	保	情	L	P	P	P	O	C	教	教
1	亜	常用	亜流	漢	名詞		0	−	0	0	0	0	0	0	0	0	0	d	e	d	e	e	e	書	
1	亜	常用	亜麻	漢	名詞	高	3	e	3	0	0	0	0	0	0	0	0	e	e	c	d	e	c		国
2	哀	常用	哀愁	漢	名詞		0	−	0	0	0	0	0	0	0	0	0	c	d	d	e	e	c	書	
2	哀	常用	哀願	漢	名詞	中	1	e	1	0	0	0	0	0	0	0	0	c	c					書	
2	哀	常用	悲哀	漢	名詞	高	3	e	1	0	0	2	0	0	0	0	0	c	c	c	c		e	書	
2	哀	常用	哀れ	和	名詞	小_前	23	e	9	0	0	13	0	0	1	0	0	a	b	d	d	c	c	書	
2	哀	常用	哀れむ	和	動詞	中	6	e	2	0	0	1	0	0	3	0	0	c	c	e		e	d		
2	哀	常用	哀れみ	和	名詞	中	6	e	1	0	0	5	0	0	0	0	0	c	d			e	e		
3	挨	新規	挨拶	漢	名詞	小_前	108	c	71	0	0	8	19	4	3	2	1	a	a	a	a	a	a	書	国外
4	愛	教育	愛情	漢	名詞	小_前	49	d	20	0	0	8	1	14	5	1	0	a	a	b	b	b	b	書	
4	愛	教育	愛読	漢	名詞	小_後	4	e	3	0	0	1	0	0	0	0	0	c	c	c	d	e	d		
4	愛	教育	恋愛	漢	名詞	中	28	d	16	0	0	6	1	5	0	0	0	a	a	a	b	a	a	書	
4	愛	教育	愛媛	固	名詞	小_後	34	d	11	0	2	17	0	0	4	0	0	c	b	b	a	c	b	教	国社
5	曖	新規	曖昧	漢	形状詞	中	48	d	31	1	0	8	1	1	1	1	4	a	a	b	b	b	b	書	国
6	悪	教育	悪事	漢	名詞	高	3	e	3	0	0	0	0	0	0	0	0	c	c	e		d	d	書	
6	悪	教育	悪意	漢	名詞	中	7	e	1	0	1	0	0	2	0	0	3	b	b	c	d	c	c	書	
6	悪	教育	醜悪	漢	名詞	高	3	e	3	0	0	0	0	0	0	0	0	d	d				e		
6	悪	教育	悪寒	漢	名詞		0	−	0	0	0	0	0	0	0	0	0	c	d			e	e	書	
6	悪	教育	好悪	漢	名詞		0	−	0	0	0	0	0	0	0	0	0	e	e						
6	悪	教育	憎悪	漢	名詞	高	9	e	8	0	0	1	0	0	0	0	0	b	c	d	d	e	c	書	
6	悪	教育	悪い	和	形容詞	小_前	232	b	74	1	46	28	15	46	2	14	6	a	a	a	a	a	a	書	
6	悪	教育	悪者	和	名詞	高	1	e	0	0	0	1	0	0	0	0	0	c	c	d		c	c	書	
7	握	常用	握手	漢	名詞	小_前	11	e	11	0	0	0	0	0	0	0	0	b	b	c	b	d	b	書	国
7	握	常用	握力	漢	名詞	中	9	e	1	3	0	0	0	0	0	5	0	e	e	e	e	e	e		数保
7	握	常用	掌握	漢	名詞	中	17	e	1	0	0	16	0	0	0	0	0	b	c	c	d	e	d		社
7	握	常用	握る	和	動詞	小_前	193	b	20	0	7	141	0	5	15	5	0	a	a	a	a	b	a		社
7	握	常用	握り	和	名詞	中	15	e	0	0	0	0	0	0	14	1	0	c	d	c	d		d		芸
7	握	常用	一握り	和	名詞	中	3	e	1	0	0	2	0	0	0	0	0	c	d	d	d	d	d	書	

３．　語例検討の観点

作成した常用漢字表語例集をもとに，本稿では以下の観点を定めて語の検討を行う。

３．１　重要度

LB（図書館書籍サブコーパス）をもとに作成された語彙のデータが，一般社会の語彙のありようをよりよく反映していると田中（2011）によって分析されている。LB の語彙レベルを重要度の指標とする。一般の社会生活の目安として掲げられている常用漢字表語例は，この LB レベルが高いことが望ましいと言える。

また社会生活に役立つことばの力の育成が求められている国語科にとっても，この重要度という指標はまさに重要である。特に重要度が高く，かつ下項で示す日常度の低い語彙は，学習語として今後，国語科語彙指導で最も扱うべき語彙であると考えている。（河内 2014）

３．２　日常度

OC（知恵袋）サブコーパスのレベルが，日常的な語彙のありようをもっともよく反映していることが，これも田中（2011）の検証によって明らかにされている。このサブコーパスの語彙レベルを日常度の指標とする。日常レベルで広く使われているという点で，LB 同様にこの OC においても，常用漢字表語例の語彙レベルは高いことが望ましいと考えられる。ただし，重要度に対して一般社会における必要性は低いと言える。

また中学校国語では，特に第 1 学年の各領域において，「日常生活」におけることばの力の育成が求められている。この中学校国語における「日常生活」と，語彙レベルにおける「日常度」は概念が完全に一致するものではない。前者は，「学校や家庭，地域など，身の回りの生活」（『中学校学習指導要領解説国語編』）を指し，後者は，書籍や新聞等と比較した上での語彙の特徴を指す。しかし日常性という性質は共通のものであり，日常において読み書きできる語の指標として，この OC レベルを用いることは可能である。

なお本発表においては，LB（図書館）と OC（知恵袋），並びに独自に設定した教科書の語彙レベルを活用している。本来，それ以外の PB（出版），PM（雑誌），PN（新聞），OY（ブログ）の語彙レベルを LB，OC の指標の補完として用いることが望ましい。参考資料としてそれらの語彙レベルの状況を示しているが，本格的な活用は今後の課題となる。

３．３　教科書レベル

教科書の特徴を示すものとしては，すでに近藤（2008）による「教科特徴度」がある。これは書籍と比較した場合の教科書における度数の特徴を示す値である。

本発表では，LB，OC そして教科書の 3 者を比較して語彙の検討を行う。したがってここでは教科特徴度は用いずに，LB，OC と同じ手法で a〜e の教科書の語彙レベルを設定した。教科書レベルと呼び，教科書の語彙の指標とする。

３．４　文化度

日常度，重要度，教科書レベルは低くても，歴史的，文化的見地から，国語科で指導，継承すべきと考えられる語がある。例えば，常用漢字表語例の中には，現代日本語の書き言葉にはほとんど見られない，次のような時季にかかわる語がある。

霧雨　観桜　盛夏　初荷　賀状　寒暑　雨季

これらの語は，現代の使用の目安という観点で考えると，常用漢字表語例としてはふさわしくない。しかし学校において，とりわけ国語科の指導で取り上げない限り，子どもたちが目にする機会はほとんどないことになる。こうした語彙を精選して文化度の高い語とする。その上で，常用漢字表から削除すべき語，学校で指導すべき語を選定していきたい。

４．　観点を用いた語例の分析

４．１　常用漢字語例　度数・割合表

前項の観点をもって，常用漢字表語例集中の，異なり語数 7514 についての分析を行った。

まず，重要度，日常度，並びに教科書レベルの度数と割合をまとめると以下の通りとなる。［表 1］

［表1］常用漢字語例　度数・割合表　LB・OC・教科書

	a	b	c	d	e	—	計
重要度（LB：図書）	2024	1705	1699	1003	1032	55	7518
	26.9%	22.7%	22.6%	13.3%	13.7%	0.7%	
日常度（OC：知恵袋）	955	953	1180	1085	2111	1234	7518
	12.7%	12.7%	15.7%	14.4%	28.1%	16.4%	
教科書	719	551	750	749	3714	1035	7518
	9.6%	7.3%	10.0%	10.0%	49.4%	13.8%	

全体の約半数（49.6%）の常用漢字表語例が，重要度の高い語（a，b）であることが分かる。一般社会の語彙のありようを反映する LB のレベルの高い語例は，「目安」としての役割を果たしていると言えるだろう。しかし一方で重要度の低い語（d，e，－）も約 1/4 にのぼる（27.7%）。

日常度については，どのレベルにも語彙が平均的に広がっている。語例と日常度との関係性が重要度と比べて低いことが分かる。

教科書レベルの高い語（a，b）が少ない（17.9%）。約半数の語の重要度が高いことと併せて考えると，ここに教科書の語彙の特徴と課題があると考えられる。子どもたちの実生活に資するために，現状の教科書の語彙がその役割を果たしているとは言えないだろう。

なお参考として，PB~OY の語彙レベルの頻度と割合の状況をまとめておく。［表 2］

［表2］常用漢字語例　度数・割合表　PB～OY

	a	b	c	d	e	—	計
PB（出版書籍）	1843	1612	1642	1084	1108	229	7518
	24.5%	21.4%	21.8%	14.4%	14.7%	3.0%	
PM（雑誌）	1664	1211	1133	1316	770	1424	7518
	22.1%	16.1%	15.1%	17.5%	10.2%	18.9%	
PN（新聞）	1461	1101	1238	992	806	1920	7518
	19.4%	14.6%	16.5%	13.2%	10.7%	25.5%	
OY（ブログ）	1437	1240	1474	1081	1857	429	7518
	19.1%	16.5%	19.6%	14.4%	24.7%	5.7%	

４．２　重要度・日常度・教科書レベル　相関表

次に 3 つのレベルの相関をまとめ，具体的な語例の検討を行う。

まず，重要度と日常度の相関関係をあらわしたのが次の表である。［表 3］

［表3］重要度（LB）＼日常度（OC）相関表

LB＼OC	a	b	c	d	e	—	計
a	912	651	347	89	24	1	2024
b	37	256	556	448	372	36	1705
c	5	42	222	382	825	223	1699
d	1	3	46	124	496	333	1003
e	0	1	8	39	379	605	1032
—	0	0	1	3	15	36	55

重要度，日常度ともに高い語については，生活上必要性の高い語であると判断することができる。ここでは重要度，日常度ともに高い語の集まりを「生活語彙」と呼ぶこととする。これらの語は常用漢字表語例として適切であると言える。国語教育においても重要性の高い語であることは言うまでもない。ただこれらの語は，後掲する一覧表を見て分かるように，実生活の中で子どもたちが自然に意味を理解し，使用できると思われる語も多い（悪い，握るなど）。取り立てて指導すべき語かどうかを精査する必要がある。

一方，重要度は高いが日常度は低いという語の集団がある。これらの語は社会生活で必要性が高い一方で，日常的には用いられることが少ないということになる。指導の必要性の高い語と言える。特にこれらの語のうち，教科書での頻度が高く，かつ各教科の専門用語以外の語は「学習語彙」として，学校で積極的に取り扱うべき語であると考えている。

また一方，重要度，日常度ともに低い語の集団がある。これらは一般社会の使用の目安という，常用漢字表語例の性質を伴っていない語である可能性が高い。常用漢字表の語例としての適否を精査する必要がある。国語教育においては，実生活に資するという観点からは指導の必要性の低い語である。しかし，前述したように文化度という観点では，学校での指導が求められる語であるとも考えられる。

以下，教科書と重要度，日常度の相関表を掲げる。［表 4］［表 5］

教科書のレベルが高く，重要度，日常度の低い語の中にも，「学習語彙」が含まれていると考えられる。また逆に教科書のレベルが低く，重要度，日常度の高い語の中には，今後学校教育に取り入れるべき語が含まれていると考えている。

［表4］重要度（LB）＼教科書　相関表

LB＼教科書	a	b	c	d	e	—	計
a	674	433	438	276	202	1	2024
b	36	91	245	333	974	26	1705
c	8	23	51	110	1361	146	1699
d	1	3	12	19	652	316	1003
e	0	1	4	10	471	546	1032
—	0	0	0	1	54	0	55

［表5］教科書＼日常度（OC） 相関表

教科書＼OC	a	b	c	d	e	—	計
a	437	180	77	14	11	0	719
b	197	165	108	56	24	1	551
c	161	201	186	107	78	17	750
d	92	150	204	158	119	26	749
e	68	254	564	669	1474	685	3714
—	0	3	41	81	405	505	1035

５．　生活語彙

［表 3］に示したように，重要度，日常度ともに高い語（a，b）が 1856 語ある。これらの語の，教科書レベル並びに PB～OY の語彙レベルとの相関は次の通りである。［表 6］

また特に，重要度，日常度，教科書レベルのいずれも高い 977 語（a，b）を一覧表にした。その一部を掲げる。［表 7］

［表6］LB(a,b)・OC(a,b)＼教科書, PB～OY　相関表

LB(a,b)・OC(a,b)	a	b	c	d	e	—	計
教科書	617	360	361	234	282	2	1856
PB	1478	359	19	0	0	0	1856
PM	1406	359	82	7	0	2	1856
PN	1155	426	217	41	10	7	1856
OY	1326	446	77	6	1	0	1856

［表7］生活語彙（重要度：a，b∧日常度：a，b∧教科書：a，b）；一部

悪い	握る	圧力	気圧	扱う	安全	不安	暗い	以上	位置	範囲	医療	委員	行為	異なる	移る	移す	椅子	意見	意味
違う	違い	維持	繊維	地域	教育	育つ	育てる	一般	統一	一日	一人	引く	印刷	印象	原因	因る	議員	左右	右
宇宙	雨	運動	運ぶ	雲	映画	栄養	経営	影響	撮影	衛生	利益	液体	血液	越える	援助	公園	喫煙	演奏	中央
反応	押す	奥	横	部屋	記憶	音楽	発音	音	気温	温める	上下	下	下がる	下さる	化学	文化	火	加える	可能
許可	何	花	価値	価格	評価	結果	科学	教科	夏	家庭	家	通過	過ぎる	過ごす	歌	歌う	我々	計画	紹介
回る	回す	会話	社会	会う	改革	海	世界	皆	機械	開始	展開	開く	解決	理解	解く	破壊	海外	外	被害
蓋	角度	三角	拡大	性格	覚える	比較	確認	正確	確かめる	獲得	学習	大学	学ぶ	楽器	楽しい	楽しむ	掛ける	掛かる	活動
生活	分割	割る	割合	完全	完成	乾燥	患者	寒い	交換	時間	人間	間	勧める	感覚	漢字	習慣	管理	関係	関する
関わる	環境	循環	簡単	観察	韓国	含む	含める	顔	企業	危険	机	希望	季節	既に	記入	記号	起きる	起こる	起こす
基礎	基準	基づく	規則	幾ら	期間	期待	最後	機会	技術	疑問	会議	九百	普及	及び	吸収	呼吸	要求	求める	研究
地球	過去	巨大	居る	根拠	許す	距離	共同	共通	公共	供給	提供	子供	協力	状況	挟む	狭い	強い	宗教	教える
競争	職業	玉	平均	近い	金属	金	細菌	勤める	筋肉	筋	禁止	緊張	銀行	区別	道具	空	繰り返す	君	直径
型	契約	計算	時計	経済	経験	景気	軽い	傾向	携帯	迎える	攻撃	激しい	穴	決める	決まる	結論	結婚	結ぶ	月
事件	条件	見る	見える	見せる	建築	建物	健康	保険	検討	派遣	権利	憲法	試験	実験	元	言う	言葉	制限	限る
限り	現象	現在	表現	減らす	厳しい	自己	古い	呼ぶ	固定	固い	事故	個人	個性	雇用	互い	前後	語る	物語	保護
人口	口	工場	加工	人工	成功	広い	広がる	広げる	交通	交ぜる	観光	光	向上	向ける	向かう	思考	参考	考える	考え
旅行	行政	行く	行う	更に	効果	厚生	厚い	天皇	学校	航空	降る	最高	高い	項目	構造	興味	購入	番号	合計
合う	試合	合わせる	報告	時刻	国際	国家	外国	国	骨	頃	今後	今日	今	困難	根	左	調査	砂糖	差別
差す	再び	採用	済む	細い	細かい	野菜	最大	最近	最も	裁判	存在	材料	財産	著作	作業	作用	動作	作る	政策
対策	皿	山	参加	産業	生産	酸素	賛成	残る	残り	残す	女子	様子	子	支持	支える	止まる	止める	仕事	歴史
市民	都市	死亡	死ぬ	糸	至る	私	使う	刺激	始める	始め	始まる	福祉	姿勢	姿	思う	思い	指示	指導	指

なお重要度，日常度ともに高いにもかかわらず，教科書レベルの低い語(d，e，－)が 518 語ある。生活語彙の 3 割近くになる（約 28%）。このことは，それらの語の全体を大まかに見る限り，常用漢字表語例の課題と言うよりも，教科書の抱える課題であるように見える。言うまでもなく各教科の学習のために教科書はある。しかし一方にある，実際の生活に資するという学校教育の目的を考えた時，教科書にかかる教育的フィルターの是非について論じていくことも必要であると考える。以下にその一部を掲げる。［表 8］

［表8］教科書不足語彙（重要度：a，b∧日常度：a，b－∧教科書：d，e，－）；一部

愛情	恋愛	曖昧	扱い	依頼	偉い	違反	違法	間違う	間違える	引退	社員	飲食	隠れる	右手	水泳	映る	越す	応援	演技
汚い	往復	殴る	下る	下手	化粧	加入	加減	何事	何十	花火	菓子	嫁	休暇	暇	靴	稼ぐ	蚊	回答	次回
改造	怪しい	後悔	悔しい	潰す	潰れる	壊す	壊れる	慾かしい	外出	外科	外れる	損害	該当	覚悟	覚え	覚める	確定	顎	割れる
若干	干す	甘い	甘える	看護	疾患	世間	勧め	関節	旅館	簡易	眼鏡	笑顔	危ない	元気	浮気	祈る	帰り	寄る	亀
偽物	詐欺	詰める	虐待	久しい	休憩	休む	休み	救急	去年	金魚	御飯	叫ぶ	狂う	恐怖	恐らく	恐ろしい	競馬	響く	曲がる
近所	金銭	金持ち	勤務	出勤	苦労	空く	理屈	熊	訓練	経費	警告	欠ける	清潔	月曜	研修	機嫌	嫌う	嫌い	賢い
懸命	玄関	限度	期限	源泉	戸籍	固まる	虎	解雇	雇う	午前	後輩	誤解	細工	向こう	好意	好み	好く	更新	今更
幸い	幸せ	荒れる	香り	控除	控える	黄色い	喉	慌てる	結構	構う	問い合わせる	告白	地獄	日頃	今朝	今年	昆布	婚約	沙汰
座席	再度	返済	詳細	交際	財布	謝罪	削除	昨日	昨年	酢	咲く	殺人	擦る	参る	土産	傘	散歩	残念	息子
支障	支店	氏名	上司	旨	必死	伺う	私立	使い	姉妹	容姿	視力	歌詞	試し	資格	飼う	次元	無事	持参	辞書
叱る	嫉妬	実力	田舎	医者	前者	感謝	謝る	風邪	借金	若しくは	寂しい	手伝う	留守	腫れる	受験	受付	優秀	拾う	臭い

６．　学習語彙

重要度が高く（a，b），日常度が低い語(d，e，－)は，国語教育において指導すべき語となり得る。以下に，教科書レベル並びに PB～OY の語彙レベルとの相関を示す。［表 9］

［表9］LB(a,b)・OC(d,e,−)＼教科書, PB～OY　相関表

LB(a,b)・OC(d,e,−)	a	b	c	d	e	—	計
教科書	18	63	150	189	534	16	970
PB	72	551	337	9	1	0	970
PM	42	261	345	222	65	35	970
PN	78	219	301	196	92	84	970
OY	6	182	496	201	83	2	970

また特に，教科書レベルの高い語（a～c）231 語を一覧表にした。［表 10］

［表10］学習語彙（重要度：a，b∧日常度：d，e，－∧教科書：a，b，c）

推移	維新	調印	繁栄	栄える	陣営	営む	炎	鉛	音色	下流	化石	果実	貨幣	通貨	余暇	画家	介入	航海	絵画
開拓	革新	格子	歌舞伎	紀元	軌道	起源	基	儀式	急速	宮殿	宮廷	宮	拠点	漁業	恐慌	強まる	強める	教え	郷土
響き	曲線	近代	接近	琴	苦しみ	君主	軍備	模型	傾斜	継承	鯨	絹	権威	孤立	後	広大	広場	広がり	交わる
光線	諸侯	洪水	耕地	農耕	鉱山	酵母	復興	振興	均衡	合同	彫刻	穀物	貧困	連鎖	鎖	星座	色彩	採集	考察
山脈	仕える	樹脂	視覚	試み	詩人	物資	諮る	活字	寺院	描写	斜面	主権	地主	狩猟	儒教	樹木	樹立	民衆	全集
成熟	順序	秩序	肖像	提唱	奨励	縄文	土壌	特色	生殖	装飾	織物	深まる	深める	森林	審議	親しむ	遂げる	中枢	北西
統制	百姓	征服	遠征	青銅	情勢	聖書	関税	脊椎	遺跡	説く	変遷	組成	組み込む	紛争	草履	創造	断層	貯蔵	一族
従属	率いる	農村	対立	帯びる	堆積	大衆	採択	石炭	団結	段落	弾圧	池	竹	家畜	蓄える	抽出	頂点	調和	追放
定まる	海底	締結	水田	粘土	電灯	東側	列島	討論	稲	稲作	道徳	神道	内外	首脳	背後	俳句	敗れる	培養	媒介
財閥	版画	運搬	批評	飛躍	標本	分布	富む	富	普遍	屋敷	武力	噴火	噴出	古墳	分別	平面	併合	歩み	連邦
倣う	鉄砲	妨げる	冒頭	膨大	北方	大木	遊牧	盆地	幕末	満ちる	同盟	連盟	滅亡	絶滅	滅ぼす	野外	羊	海洋	要点
形容	落葉	集落	反乱	流動	留意	留まる	倫理	臨む	礼拝	寒冷									

例えば「維新」，「化石」（いずれも［表 10］の 1 段目）などは，それぞれ社会，理科の専門用語であり，各教科でその意味も含めて学習することになる。

一方で「調印」（同）は，小学校から高校までの社会科の教科書で用いられるが，「調印」という言葉自体の説明は教科書にはない。本来，社会科の中で指導されるべき学習語である。また「推移」（同），「繁栄」（同）といった語は幅広い教科で用いられ，多種の物事を受ける語となっている。こうした語は国語科で指導されるべき学習語である。

なお学習語彙と離れるが，［表 8］を見て気付くことを記す。

LB と PB，OC と OY は近しいメディア・ジャンルである。［表 2］に示したように，語彙レベルの割合の傾向はほぼ同様である。しかし語彙レベルの傾向が一致しているわけではないことが，［表 9］を見るとわかる。

LB の語彙レベルが高くても，PB の語彙レベルが低い語(d，e，－)が 10 語見られる。［表 11］

OC の語彙レベルが低くても，OY の語彙レベルが高い語（a，b）は 188 語にのぼる。［表 12］

［表11］PB特徴語（LB：a，b∧OC：d，e，－∧PB：d，e，－）

余暇	在留	肉親	扇	霜	談判	銅像	陪審	土俵	陵墓

［表12］OY特徴語（LB：a，b∧OC：d，e，－∧OY：a，b）

握手	推移	偉大	一息	日陰	英雄	炎	宴会	講演	王子	屋上	観音	花壇	果実	初夏	出荷	通貨	画家	優雅	絵画
開拓	街道	覚ます	乾杯	彼岸	危うい	帰還	輝き	騎士	儀式	脚本	弓	丘	急速	宮殿	宮	救い	芝居	拠点	供
強まる	教え	境内	響き	近代	接近	襟	苦しみ	駆る	青空	群れ	模型	恵み	稽古	鯨	演劇	感激	欠く	県立	幻
源	己	後	広場	交える	向かい	洪水	慌ただしい	港	絞り	酵母	振興	合同	合戦	彫刻	貧困	示唆	夫妻	祭る	矢
師匠	詩人	寺院	鹿	描写	斜面	守り	就任	女神	乙女	秩序	小豆	沼	笑み	負傷	照らす	表彰	鐘	土壌	前進
森林	人員	遂げる	歌声	遠征	盛る	情勢	聖書	惜しむ	遺跡	足跡	絶妙	染まる	真相	捜索	創造	霜	贈り物	率いる	本尊
対立	大胆	大衆	道端	段落	池	竹	挑む	眺め	頂上	澄む	塚	弟子	庭園	転がる	藤	童話	独り	連日	断念
首脳	背後	俳句	敗れる	拍手	麦	悲しむ	漂う	分布	浮かべる	屋敷	舞う	仏像	並木	便り	歩む	暮れる	暮れ	連邦	訪ねる
冒頭	傍ら	平凡	麻	満ちる	味わい	岬	奇妙	悪夢	夢見る	霧	夜明け	迷い	同盟	連盟	悲鳴	暗闇	妖怪	海洋	踊り
着陸	留まる	林	臨む	礼拝	鈴	連れ	試練												

7．　文化語彙

重要度，日常度ともに低い語(e，－)は，常用漢字表語例としてふさわしくない語である可能性がある。一方で，文化語として国語教育で扱うべき語である可能性も有する。以下に，教科書レベル並びに PB～OY の語彙レベルとの相関を示す。［表 13］

［表13］LB(e,－)・OC(e,－)＼教科書，PB～OY　相関表

LB(d,e,－)・OC(d,e,－)	a	b	c	d	e	—	計
教科書	0	0	4	10	495	526	1035
PB	0	1	26	175	630	203	1035
PM	0	6	23	123	177	706	1035
PN	1	7	30	74	131	792	1035
OY	0	2	29	88	593	323	1035

また特に，重要度，日常度，教科書レベルのいずれも低い語(d，e，－)，1021 語を一覧表にした。その一部を掲げる。［表 14］

［表14］文化語彙（重要度：e，－∧日常度：e，－∧教科書：e，－）：一部

亜麻	好悪	握力	客扱い	砂嵐	案文	各位	位取り	胃弱	尉官	一尉	偉観	慰む	慰み	姻族	淫行	淫乱	光陰	痛飲	隠忍
雲隠れ	気宇	雨具	五月雨	霧雨	泳法	俊英	栄枯	詠草	朗詠	易者	不易	悪疫	防疫	益する	液状	悦楽	喜悦	謁する	閲歴
一円	川沿い	才媛	花園	煙霧	煙い	煙たい	野猿	鉛色	縁取り	妖艶	色艶	汚点	汚れ物	汚らしい	押韻	欧文	渡欧	殴打	桜花
桜色	葉桜	老翁	深奥	専横	恩情	謝恩	穏和	穏当	化かす	仮名	落花	果断	河岸	罪科	書架	盛夏	夏服	香華	茶菓
過つ	再嫁	嫁する	嫁ぎ先	寸暇	禍福	禍根	災禍	製靴	蚊柱	歯牙	瓦屋根	肉芽	賀状	賀する	雅趣	手回し	戒心	更改	拐帯
悔恨	海鳴り	皆勤	地階	塊状	決壊	懐手	外題	外道	外囲い	害悪	崖下	慨嘆	該博	四つ角	角笛	隔月	岳父	楽隊	喝破
渇水	割拠	葛湯	滑降	刈り入れ	甘言	甘受	汗顔	肝胆	冠詞	栄冠	一巻	看破	寒暑	寒空	敢然	閑却	鉄管	関取	交歓
緩慢	緩急	緩み	丸薬	丸洗い	含蓄	愛玩	眼力	頑健	厚顔	したり顔	顧わしい	安危	火の気	風紀	鬼才	鬼ごっこ	悲喜	棋士	棋譜
毀誉	旗色	手旗	輝石	威儀	児戯	擬音	白菊	吉例	大吉	吉報	難詰	詰み	客死	脚立	行脚	順逆	久遠	及第	及び腰
弓道	旧道	不朽	腐朽	臼歯	学究	感泣	泣き沈む	球技	牛馬	去就	挙手	壮挙	魚	煮魚	大漁	吉凶	狂おしい	享有	概況
地峡	挟撃	狭量	広狭	胸囲	胸毛	強がる	近郷	在郷	矯める	競泳	競輪	競り合う	仰角	今暁	通暁	罪業	凝り性	終極	玉石
金色	木琴	筋骨	僅差	胸襟	詩吟	苦吟	抜け駆け	愚問	愚鈍	空き巣	串刺し	串焼き	屈伸	音訓	勲功	殊勲	薫風	郡部	群居
係累	係争	黙契	契る	恵与	恵む	慶弔	慶祝	慶賀	憩う	鯨油	劇薬	橋桁	穴居	墓穴	元結	傑物	双肩	建議	節倹
勤倹	兼職	剣舞	剣	拳法	軒数	強健	圏内	圏外	絹布	遣外	分遣	金遣い	先賢	霊験	懸垂	懸想	元帳	幽玄	言行
上弦	原	目減り	人減らし	克己	下戸	古株	円弧	内股	虎穴	猛虎	枯淡	小鼓	五色	互選	互い違い	後刻	後添い	後の世	後れ毛
気後れ	悔悟	碁石	新語	正誤	功名	巧拙	広言	顔向け	江湖	廃坑	孝心	攻守	首肯	厚情	紅	大荒れ	色香	移り香	候文
就航	貢ぎ物	大降り	小康	黄	喉頭	硬度	生硬	絞り上げる	綱紀	興趣	度量衡	聴講	号外	剛健	傲然	文豪	腹黒い	脱穀	獄舎
疑獄	老骨	持ち駒	今上	紺青	紺屋	商魂	教唆	詐取	再選	良妻	砕石	砕氷	宰領	淡彩	採光	祭り上げる	秋祭り	潔斎	細腕
青菜	歳末	私財	鉄柵	圧搾	遅咲き	札入れ	増刷	靴擦れ	雑兵	山車	寺参り	蚕糸	蚕食	産み月	産湯	雨傘	辛酸	酸い	食べ残し
暫時	止宿	某氏	矢面	寸志	私腹	枝葉	雄姿	相思	施療	脂ぎる	品詞	義歯	追試	誌面	雌伏	賜暇	恩賜	示し	字画
字	末寺	療治	時候	滋味	式辞	好餌	食餌	御璽	国璽	鹿の子	車軸	地軸	室	室咲き	執心	我執	湿す	乾漆	実入り
社	捨象	喜捨	大赦	恩赦	生煮え	謝絶	平謝り	正邪	蛇の目	蛇腹	大蛇	酌量	自若	閑寂	寂然	法主	朱肉	朱筆	狩り込み

なお試みに，現勤務校の大学 4 年生 31 名に対して，［表 14］から任意に選んだ 30 語について，読みと意味についての調査を行った（平成 26 年 6 月実施）。結果は以下の通りである。今後こうした調査を大規模に行い，実態の把握に努めていく。

常用漢字語例	ふりがな	「読み」正答率	「意味」正答率	常用漢字語例	ふりがな	「読み」正答率	「意味」正答率
絹布	けんぷ	0.0%	27.0%	穀倉	こくそう	54.1%	67.6%
蚕糸	さんし	5.4%	27.0%	綱紀	こうき	56.8%	8.1%
花暦	はなごよみ	10.8%	2.7%	該博	がいはく	56.8%	5.4%
剣ヶ峰	けんがみね	10.8%	0.0%	首肯	しゅこう	62.2%	32.4%
窯業	ようぎょう	21.6%	10.8%	鯨油	げいゆ	67.6%	83.8%
逓減	ていげん	27.0%	10.8%	岳父	がくふ	73.0%	0.0%
罷業	ひぎょう	35.1%	2.7%	焦眉	しょうび	73.0%	2.7%
才媛	さいえん	40.5%	16.2%	恩赦	おんしゃ	75.7%	8.1%
懸想	けそう	43.2%	16.2%	久遠	くおん	75.7%	35.1%
辣腕	らつわん	45.9%	18.9%	葛湯	くずゆ	78.4%	40.5%
蚊柱	かばしら	45.9%	32.4%	偉観	いかん	83.8%	5.4%
蜜月	みつげつ	51.4%	13.5%	霊峰	れいほう	83.8%	37.8%
謄写	とうしゃ	51.4%	13.5%	厚顔	こうがん	89.2%	62.2%
拐帯	かいたい	54.1%	0.0%	五月雨	さみだれ	89.2%	29.7%
気宇	きう	54.1%	10.8%	漏電	ろうでん	97.3%	89.2%

８．　おわりに－今後の課題

今後は前節のデータをもとに，新しい常用漢字表語例の選定並びに文化語彙の選定の具体的提案を行っていく。　まずは［表 14］とした 1021 字の語例が検討の対象となる。字種としては 797 字が対象となる。それぞれの語のレベルやコーパス上の用例などを一覧にしていく。その上で，「学校・社会対照語彙表」の，該当字種を含む別の語彙と，頻度やレベルを比較するなどして，常用漢字表語例としての妥当性や文化度を評価していく。

一例を挙げる。［表 14］冒頭の「亜麻」は，語種「亜」の語例である。常用漢字表には，語種「亜」に「亜流」（LB：d，OC：e），「亜麻」（LB：e，OC：e），「亜熱帯」（「亜」と「熱帯」に分かれて解析されてしまうためレベル判定不能）の 3 語が語例として掲げられている。一方「学校・社会対照語彙表」には，字種「亜」を含む語が 19 語ある。その中で「亜麻」（LB：e，OC：e）より LB，OC レベルの高い語（LB，OC いずれかあるいは両方がレベル d 以上）は，「亜鉛」（LB：d，OC：d），「亜種」（LB：d，OC：e），「亜鈴」（LB：d，OC：e），「白亜」（LB：d，OC：e）である。特に「亜鉛」は教科書での頻度が高く，理科の特徴語である。「亜鉛」は「亜麻」に変わる，字種「亜」の語例候補になるだろう。しかし一方で，植物「亜麻」が，明治以降の北海道開拓の産物であるという歴史的価値を背負っていること，また「亜麻仁油」や「亜麻色」といった派生語も存在するといった事実を無視することはできない。

頻度や語彙レベルだけで判断することは難しい。用例の検討はもちろんだが，異なる年代への認知度調査や，多くの方々の識見を取り入れながら，本研究を前進させていきたい。

謝　辞

本研究は，文部科学省科学研究費基盤研究(C)一般（課題番号：25381226）の助成を受けたものです。

文　献

田中牧郎・近藤明日子（2011）「学校・社会対照語彙表」『特定領域研究「日本語コーパス」言語政策班報告書　言語政策に役立つ，コーパスを用いた語彙表・漢字表の作成と活用』，pp.69-76，JC-P-10-01

田中牧郎（2011）「語彙レベルに基づく重要語彙リストの作成－国語施策・国語教育での活用のために－」前同，pp.77-88，JC-P-10-01

近藤明日子（2008）「中学校教科書の教科別特徴度の抽出―理科を例として―」『特定領域研究「日本語コーパス」言語政策班中間報告書　言語政策に役立つ，コーパスを用いた語彙表・漢字表の作成と活用』，pp.169-174，JC-P-08-01

河内昭浩（2014）「理科教科書のことばの分析と理科学習語の選定」，『日本語学』第 33 巻第 3 号，pp.69-77，明治書院

商品カテゴリの階層構造を用いた商品分類

中島 道幸、古宮 嘉那子（茨城大学工学部情報工学科）

Product Classification Using Hierarchical Structure of Categories

Michiyuki Nakajima (Department of Computer and Information Sciences,Ibaraki University)
Kanako Komiya (Department of Computer and Information Sciences,Ibaraki University)

要旨

商品のレビュー文書から競合商品を同定する研究や商品ページの属性や属性値を用いた同一商品のクラスタリング手法の研究等、近年、同一商品の同定に関する様々な研究が行われてきている。本稿では、同一商品の同定に関する研究の足掛かりとして商品カテゴリの階層構造を用いた商品分類を行った結果を報告する。実験には、約 60 万件の楽天市場の商品データを使用した。分類器 svm を使用し、五分割交差検定でそれぞれの階層毎のカテゴリの正解率を求めた。消費者が分類することが目的なので、素性を作成する際には、商品ページから消費者が得られる情報のみを選択した。また、求めた正解率から階層毎、階層全体の重みつき平均を求め、ベースラインとの比較を行った。

1. はじめに

近年、Web 上のサービスを利用して商品を購入する“インターネットショッピング”が普及してきた。ショッピングサイトには様々な企業が出店するサイバーモールのようなタイプのものがある。このようなサイトの商品ページは出店している企業が独自に作成している場合がある。そのため、消費者は自分の求める商品を探すことが困難となっている。商品のタイトルや説明文、写真など商品ページのすべてが店舗にゆだねられている。店舗側は売り上げを上げるために商品タイトルの一部に「送料無料」や「ポイント 2 倍」などの修飾語や関連情報を付けている。このため、消費者は単純にクエリ検索を行うだけでは、望んでいる商品のページにたどり着くことができない。さらに、同一商品であるが、商品タイトルや商品説明文が異なっているものや、異なる商品であるが、用いられている商品画像が同一のものが存在する。このような現状から同一商品の同定をする手法が必要であると考え、ショッピングサイトの商品カテゴリに着目した。商品カテゴリに階層があることを利用して、階層的に分類を行った。本稿では、階層を利用していない場合との比較を行う。

2. 関連研究

カテゴリに関しての研究としては、Web 上の商品情報を利用した商品ページのカテゴリ分類という研究を佐藤らが行っていた(佐藤ら(2010))。彼らは商品ページを自動的にカテゴリ分類する手法を提案している。また（古宮ら(2013)）は既存の手法である Naïve Bayes と Complement Naïve Bayes と提案手法である Negation Naïve Bayes を比較している。分類精度が平均 67.3%とベースラインを上回る結果となり、提案手法が商品ページに対して有効であることがわかった。

分類に関する研究としては、商品ページからの属性・属性値抽出と同一商品クラスタリング手法という研究を豊橋技術科学大学の坂地らが行っていた(坂地ら(2010))。商品ページから属性・属性値を抽出し、属性のまとめ上げを行う。また、二つの商品ページを比較し、類似度スコアをつけることで、商品ページのクラスタリングを行う。

本研究では、カテゴリの階層構造を用いて、商品の分類を行っていく点で、これらの研究とは異なる。

3. 階層構造

商品には、膨大な数の商品の中から消費者の求める商品を探せるように、それぞれジャンルが付けられている。この商品ジャンルは大まかなカテゴリから細かなカテゴリまで分けられている。大まかなカテゴリの例として、インテリアを挙げてみる。インテリアには、時計やテーブル、カーテン、椅子等がある。また、テーブルと一口に言っても、ダイニングテーブル、カウンターテーブル、コーヒーテーブル等に細かい分類をすることができる。図 1 に例を示す。このように、商品ジャンルは大きいカテゴリから小さいカテゴリへと、階層構造で構成されている。消費者が欲しい商品が見つからない場合やお買い得な商品を探したいときに、大きいカテゴリから小さいカテゴリへとジャンルで絞り込んでいくことができる。

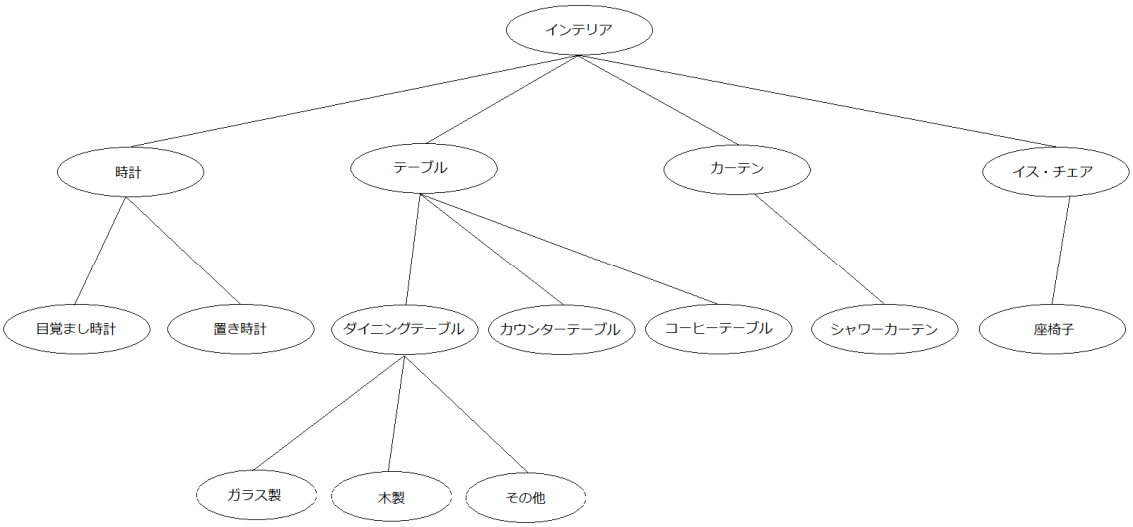


図 1：階層構造の例

本研究では、この階層構造を用いて、商品のカテゴリを機械学習による手法で絞り込んでいく手法をとる。

4. 実験データ

4.1. 実験に使用したデータ

本研究では、約 60 万件の楽天市場の商品データを使用した。商品データは 2014 年 4 月 1 日公開のものである。楽天市場の商品データは 11 個の情報で構成されている。その要素を表 1 に示す。基本的には表 1 のようなフォーマットで商品データは構成されている。実際の商品データの例を図 2 に示す。

商品コードは「店舗コード：商品 ID」と示される。販売方法別説明文とは商品説明文に入らない場合に使用される説明文である。空白となる場合もある。商品 URL はユニーク部分のみが示されている。「http://item.rakuten.co.jp/[店舗コード]/[商品 URL]/」で商品ページの URL となる。ジャンル ID は、その商品カテゴリに割り当てられた番号である。

新約「巨人の星」花形 1› guruguru2:11452005› 400›村上 よしゆき 画梶原 一騎 他原作週刊少年マガジンKC本[コミック]詳しい納期他、ご注文時は ご利用案内・返品のページをご確認ください出版社名講談社出版年月2006年11月サイズISBNコード9784063637533コミック ≫ 少年（中高生・一般） [講談社週刊マガジンKC] › 商品説明新約「巨人の星」花形 1シンヤク キヨジン ノ ホシ ハナガタ 1 シユウカン シヨウネン マガジン コミックス KC ケ-シ- 42265-53※ページ内の情報は告知なく変更になることがあります。あらかじめご了承ください登録日2013/04/08› 9784063637533› http://image.rakuten.co.jp/guruguru2/cabinet/b/7/533/9784063637533.jpg› 0› 0.00› guruguru2› 101941↓

図 2：実際の商品データの例

表 1：商品データフォーマット

順番	データ内容
1	商品名
2	商品コード
3	商品価格
4	商品説明文
5	販売方法別説明文
6	商品 URL
7	商品画像 URL
8	レビュー件数
9	レビュー平均
10	店舗コード
11	ジャンル ID

4.2. ジャンル ID

ジャンル ID は商品ジャンルに割り当てられた番号である。その商品ジャンルに当てはまる商品には、その商品ジャンルの番号であるジャンル ID がつけられる。また、その商品ジャンルには親ジャンル ID というものが割り当てられており、階層構造となっている。つまり、親ジャンル ID を辿っていくと、1 階層にある 34 種類のジャンルに辿り着く。この 34 種類のジャンルは、楽天市場のトップページから検索できる最上層のカテゴリである。階層構造の例で挙げたダイニングテーブルならば、ジャンル ID が「111346」となり、親ジャンル ID は「215476」となる。図 3 に楽天市場のトップページにあるジャンルの一部を例として示す。

図 3：1 階層のジャンルの例

5. 実験

5.1. 実験内容

次の二つの実験を行った。(1)をベースラインとし、カテゴリの階層構造を用いた実験を(2)として、(1)と(2)の重みつき平均の比較を行う。

(1)60 万件のデータを 50 分割し、svm で五分割交差検定を行う。正解ラベルは、その商品のジャンル ID(最下層)とする。

(2)階層毎に分類する手法。60 万件のデータをまず、第 1 階層カテゴリに分類し、分類されたカテゴリ中の商品をそのカテゴリの下の第 2 階層カテゴリに分類するということを最下層まで繰り返す。正解ラベルはその階層のジャンル ID とする。そして、階層毎に五分割交差検定で正解率を求めた。重みつき平均は階層毎に求め、それらを掛けることで階層全体の重みつき平均とする。

5.2. 実験設定

(1)において、60 万件のデータを 50 分割にしたのは PC のスペックの都合である。メモリが 8MB のマシンで動く最低限の分割数が 50 分割であった。

正解率を求める際は、svm のツールとして libsvm を使用する。Option に関してはカーネルのタイプを linear(線形)で行った。これは以前、カーネルタイプの比較を行った実験の結果から、本実験では線形カーネルが適切であると判断した。

(2)において、分類されたカテゴリ中の商品をそのカテゴリの下の階層に分類するとあるが、商品によっては最下層のカテゴリではなく第 2 階層から第 4 階層のカテゴリが正解のものある。そのため、2 階層まではすべてのデータが用いられるが、3、4、5 階層となっていくにつれてデータ数は減っていくということである。

素性として扱う情報については 5.1 で前述した中から商品名、商品価格、商品説明文、販売方法別説明文、商品 URL、商品画像 URL、レビュー件数、レビュー平均に絞る。これは、本研究の背景として、一般の消費者が商品分類を行うことを想定しているため、消費者が商品ページから取得できる情報に限定する必要があるからである。商品説明文に関しては、mecab で形態素解析したものを素性として使用する。また、4.1 節で説明した商品データのフォーマットにしたがってない商品データについては、素性データには含めていない。

(2)についての重みつき平均の計算方法を説明する。はじめに、それぞれの商品データの件数と svm から得られた正解率を掛け、正解数を求める。正解数を計算する際に、それぞれの階層まででおわっているものについては、それ以降の正解率を 100%として計算する。例えば、3 階層まででおわっているものについては、4、5 階層では、正解率を 100%にする。本来は最下層である 5 階層まで細かく分類したいわけだが、細かいカテゴリに属さないため、途中でおわっているものについては、それ以降の階層では、100%分類できると仮定する。次に、求めた正解数を階層毎に足し合わせる。そして、正解数の合計を用いた商品データの全件数で割ることで、階層毎の重みつき平均を求めることができる。最後に、すべての階層の重みつき平均を掛け合わせることで、階層構造全体の重みつき平均を求める。

5.3. 実験結果

表 2 に実験結果を示す。括弧内の数値は途中までで階層がおわっているジャンルを 100%で計算せずに、値として加えない場合の結果である。

表 2：実験結果

正解ラベル	重みつき平均
最下層	31.24%
1 階層	85.80%
2 階層	89.96%
3 階層	84.22%(83.48%)
4 階層	85.07%(79.95%)
5 階層	93.20%(75.77%)
階層全体	51.54%

6.　考察

5 章で行った実験の結果を考察する。まず、(2)の実験における階層毎の結果と階層全体の結果がベースラインである(1)の実験における最下層の結果を上回る結果を得られたため、本研究で提案した商品カテゴリの階層構造を用いた商品分類システムは妥当であるといえる。

(1)における実験結果は 3 割程度の結果であった。(1)は最下層のラベルということで、2 階層や 3 階層等、途中でおわるものから 5 階層にまで亘る広いカテゴリで分類したため、あまりポイントが高くならなかったのではないかと考えられる。

一方、階層毎に分類した結果では、すべて 8 割を上回った。5 階層の結果が 9 割を超えているが､途中までで階層がおわっているジャンルを加えない場合の結果は 7 割程度である。これは、途中までで階層がおわっているジャンルを正解率 100%で加えた結果が大きく関係していると考えられる。また、階層が下になるにつれて途中までの階層に当たるデータが増えてくることで、5 階層で用いるデータが減ってくる。そのため、ジャンル毎に正解率を求めている過程から、五分割交差検定での正解率が 0%になるところも増えてくる。このような理由から括弧内の結果が少し低くなっていると考えられる。

階層全体の実験結果は、5 割を超え、ベースラインを超える結果となったが、それぞれの階層のエラーの累積が全体の正解率を押し下げる結果となっている。特に階層が下った際の正解率の低下が全体の正解率の低下の原因と見て取れる。

今後の課題としては、4 階層、5 階層等の下の階層の分類精度の向上である。考えられる方法としては、末端の訓練事例数を増やすことである。今回は 60 万件で実験を行ったが、マシンのスペックがよければ、データ数を増やすことができる。また、商品データを分割する必要もない。

本研究は、商品カテゴリに関しての分類であるので、商品そのものの分類や同定ではない。なので、今後は階層構造を用いて、単一商品の分類や同定をすることを目指したい。

7. まとめ

本稿では、商品カテゴリの階層構造を用いた商品分類を行った結果を報告した。実験では、正解ラベルを階層毎に設定したものと、最下層に設定したもので重み付き平均の比較を行った。結果は提案した階層構造を用いたシステムの方が 20 ポイント高くなった。今後の課題としては、下の階層の分類精度あげることである。そのためには、訓練事例数を増やすこと等でシステムの向上を目指したい。また、将来的にはこのシステムを用いて、同一商品の同定を可能にしたい。

謝　辞

データを提供していただいた、楽天株式会社と国立情報学研究所に御礼申し上げます。また、この研究は、文部科学省科学研究費補助金[若手 B（No：24700138）]の助成により行われました。ここに、謹んで御礼申し上げます。

文　献

坂地泰紀、小林暁雄、関根聡、竹中孝真(2010)「商品ページから属性・属性値抽出と同一商品クラスタリング手法」言語処理学会第 16 回年次大会発表論文集、pp.371-374.(http://www.anlp.jp/proceedings/annual_meeting/2010/pdf_dir/PA1-27.pdf よりダウンロード可能)

佐藤直人、藤本浩司、小谷善行(2010)「ウェブ上の商品情報を利用した商品のカテゴリ分類」人工知能学会代第 87 回知識ベースシステム研究会、pp.7-10.

古宮嘉那子、伊藤裕佑、佐藤直人、小谷善行(2013)「文書分類のための Negation Naive Bayes」自然言語処理 Vol. 20、 No. 2、pp.161-182.
(https://www.jstage.jst.go.jp/article/jnlp/20/2/20_161/_pdf よりダウンロード可能)

領域適応のためのサポートベクトルを用いた訓練事例の反復的選択

小林　優稀（茨城大学工学部　情報工学科）
古宮　嘉那子（茨城大学工学部 情報工学科）
佐々木　稔（茨城大学工学部 情報工学科）
新納　浩幸（茨城大学工学部 情報工学科）
奥村　学（東京工業大学 精密工学研究所）

Iterative Selection of Training Data Using Support Vectors for Domain Adaptation

Yuma Kobayashi (Department of Computer and Information Sciences, Ibaraki University)
Kanako Komiya (Department of Computer and Information Sciences, Ibaraki University)
Minoru Sasaki (Department of Computer and Information Sciences, Ibaraki University)
Hiroyuki Shinnou (Department of Computer and Information Sciences, Ibaraki University)
Manabu Okumura (Precision and Intelligence Laboratory, Tokyo Institution of Technology)

要旨

テストの対象となるドメインではなく、異なるドメインのデータ（ソースデータ）で学習を行い、それをターゲットのドメインのデータ（ターゲットデータ）に適応することを領域適応といい、近年様々な手法が研究されている。

語義曖昧性解消のタスクについて領域適応を行った場合、ソースデータ全体を学習に用いるよりも、確信度と LOO-bound という指標を利用して、自動的に選択したソースデータの部分集合を用いたほうが、正解率が上昇することが先行研究により指摘されている。本稿では、自動的に選択したソースデータの部分集合にさらにサポートベクトルを利用して反復的にソースデータを追加することを繰り返す、という手法を試みた。その結果、ベースラインよりも正解率は劣るものの、それほど正解率を落とさずに、訓練事例の数を大幅に減らすことに成功した。

１．はじめに

テストの対象となるドメインではなく、異なるドメインのデータ（ソースデータ）で学習を行い、それをターゲットのドメインのデータ（ターゲットデータ）に適応することを領域適応といい、近年様々な手法が研究されている。

語義曖昧性解消のタスクについて領域適応を行った場合、ソースデータ全体を学習に用いるよりも、確信度と LOO-bound という指標を利用して、自動的に選択したソースデータの部分集合を用いたほうが、正解率が上昇することが先行研究により指摘されている（古宮, 小谷, 奥村(2013)）。本稿では、自動的に選択したソースデータの部分集合にさらにサポートベクトルを利用して反復的にソースデータを追加することを繰り返す、という手法を試みた。

２．関連研究

領域適応は、学習に使用する情報により、supervised，semi-supervised，unsupervised の三種に分けられる。本研究で扱うのは、semi-supervised の領域適応、つまりラベルつきのソースデータとラベルなしのターゲットデータを利用するものである。

文献(Komiya, Okumura (2012))、(古宮, 奥村, 小谷(2013))では、訓練データの選択に分類器の確信度を用いて訓練事例を自動的に選択している。用例ごとに訓練事例を自動的に選択している。

また、文献(古宮 小谷 奥村(2013))は、semi-supervised な領域適応において、あるターゲットデータに対して複数のジャンルのソースデータが混在した場合、確信度と

LOO-bound という指標を利用して、領域適応のための訓練事例の部分集合を WSD の対象単語タイプごとに自動的に選択する手法について述べている。訓練データをいくつかのグループに分け分類器を作り、分類した時の各分類器の確信度と、SVM に対し、leave-one-out-estimation を行った場合の期待値の上限である LOO-bound という指標を用いて、訓練データを選択する手法である。この研究では、確信度と LOO-bound を組み合わせたスコアを用いることで、ベースラインよりも精度が向上することを報告している。本稿でも、確信度と LOO-bound を利用した、このスコアを利用する。また、先行研究と同じくラベルなしターゲットデータが手に入ると仮定して、語義曖昧性解消についての領域適応を行った。

２．１確信度と LOO-bound

本稿では、分類器のスコアとして確信度と LOO-bound をもとにした数値を掛け合わせたスコアを使用している。

確信度とは、テストデータに対し、どの程度自信を持って分類したのかを表す。つまり、テストデータと同じドメインのコーパスをどの程度正確に分類できるかを示している。確信度は用例ごとに算出されるので、全用例の平均を分類器のスコアとした。

LOO-bound は SVM に対し、Leave-One-Out-Estimation を行った時のエラーの期待値の上限であり、サポートベクトルの数を訓練事例の数で割った値である。この値はエラー率であるため、分類器のスコアとする際に 1 からこの値を引いた。

$$\text{LOO} - \text{bound のスコア} = 1 - \frac{\text{サポートベクトルの数}}{\text{訓練事例の数}} \quad \cdots (1)$$

３．領域適応のためのサポートベクトルを用いた訓練事例の反復的選択

あるドメインのターゲットデータに対して WSD を行う。このターゲットデータのラベルは未知とする。ソースデータとして複数ドメインのコーパスが利用可能であるとし、ソースデータの全体集合から、ターゲットデータに適した訓練事例を自動的に選択することを試みる。以下で、具体的な手順を示す。

(1) ソースデータの全体集合から訓練事例をランダムに選択して、訓練事例集合を複数個作成する。

(2) それぞれの訓練事例集合で分類器を学習し、ターゲットデータに適用する。

(3) 分類器が出力する値をもとに分類器ごとにスコアを計算する。

(4) スコアの最も高い分類器を作成した訓練事例集合を選択する。

SVM では分離平面を決定する際に、サポートベクトルからの距離を最大にするという性質がある。そこで、サポートベクトルを残し、反復的に訓練事例を増加させるために、以下の処理を追加した。

(5) 選択した訓練事例集合のサポートベクターの集合（SV 集合）を作成する。

(6) SV 集合にソースデータの全体集合から訓練事例をランダムに選択して加え、訓練事例集合を複数個作成する。

(7) 有限回、(2)～(6) を繰り返す。

４．実験

４．１データセット

実験には、マルチクラス対応の分類器として SVM(libsvm)(Chih-Chung Chang, Chih-Jen Lin(2001)) を使用した。また、現代日本語書き言葉均衡コーパス(Maekawa(2008))の YAHOO！知恵袋（OC）、白書(OW)、YAHOO！ブログ(OY)、新聞(PN)、書籍(PB)、雑誌(PM)のコアデータ 6 種と YAHOO！知恵袋(YAHOO)、白書(BCCWJ)非コアデータ 2 種、RWC コーパス(Hashida, Isahara, Tokunaga, Hashimoto, Ogino, and Kashino(1998))を用いた。YAHOO 知恵袋と白書のコーパスは２種あるが、内容はほぼ同一のものなので、より用例数が少なかったコアデータの方をソースデータから除いた。

また、ソースデータにテストデータのドメインと同一のドメインのコーパスを含まないようにした。テストデータには１単語あたり 50 用例以上のものを使用した。コーパスごとの単語数とデータ数の平均値を表１に示す。

また、実験には岩波国語辞典の中分類の語義を採用した。単語の語義は、岩波国語辞典(西尾、 岩淵、 水谷 (1994))の小分類の語義を採用した。語義事の単語の内訳は、1 語義（新語義を入れると 2 語義）：可能、2 語義：生きる、一般、生まれる、書く、考える、技術、経済、現在、現場、子供、自分、情報、高い、作る、強い、電話、場合、早い・速い、文化、ほか、見せる、3 語義：相手、与える、言う、今、入れる、大きい、教える、買う、関係、聞く、市場、市民、社会、進む、地方、出来る、出る、入る、初め・始め、始める、場所、開く、前、求める、訴える、4 語義：時間、時代、出す、乗る、計る、一つ、見える、認める、持つ、進める、5 語義：やる、良い、6 語義：合う・会う、立つ・建つ、見る、もの、7 語義：手、8 語義：する、取る、上げるであった。

また、本実験で使用する素性として、次の 24 の素性を使用した。

- 対象単語と前後２つの形態素の表記　5 種類
- 対象単語と前後２つの形態素の品詞　5 種類
- 対象単語と前後２つの形態素の品詞の細分化　5 種類
- 係り受け　1 種類
- 前後２つの形態素の５桁の分類コード　4 種類
- 前後２つの形態素の４桁の分類コード　4 種類

ここで用いている分類コードとは国立国語研究所が発行している「分類語彙表」（秀英出版 (1964))に記載されている分類番号、段落番号からなる、語を意味によって分類した番号のことである。

４．２．ベースライン

本実験のベースラインとして、以下の３つの実験を行った。

・すべてのコーパス

利用できるコーパス全てを使用する

・最大のコーパス

利用できるコーパスのうち、単語ごとに用例数が最大のものを使用する

・平均的なコーパス

利用できるコーパスについて、それぞれ分類器を作成し、正解率を平均する

４．３．サポートベクトルを用いた反復的手法実験

提案手法は次の手順で行う。

(1) ソースデータの全体集合から訓練事例をすべての語義を含むようにランダムに 100 件もしくは 200 件（データ件数がこの数に満たない際にはそれ以下の件数となる）選択して、訓練事例集合を 10 個作成する

(2) それぞれの訓練事例集合で分類器を学習し、ターゲットデータに適用する

(3) 分類器が出力する値をもとに分類器ごとにスコアを計算する

(4) スコアの最も高い分類器を作成した訓練事例集合を選択する

(5) 選択した訓練事例集合のサポートベクターの集合（SV 集合）を作成する

(6) SV 集合にソースデータの全体集合から訓練事例をランダムに選択して加え、訓練事例集合を複数個作成する

(7) 10 回（10 ステージ）、(2)～(6) を繰り返す

訓練事例の部分集合は 1 単語あたり 10 個作成した。また、初期事例数を 100 件または 200 件とし、すべての語義を含むようにランダムに選択した。　予備実験の結果、繰り返し回数は 10 回程度でスコアはほぼ収束することが分かったので、本実験では(7)の繰り返し回数は 10 回とする。また、この実験はランダム性が高いので、10 セット行いそれぞれの正解率を平均した。その他、前者ではすべての語義を含むように初期訓練事例集合を作成しているが、語義数にかかわらずランダムに 100 件選択したものを用いた実験も 2 回行なった。

表 1　コーパスの単語数の内訳

	単語数	テストデータ数 平均	ソースデータ数 平均
コア　Yahoo！知恵袋	22	157.77	1630.50
コア　白書	5	79.20	508.80
コア　Yahoo！ブログ	9	245.22	10226.56
コア　書籍	35	158.91	6068.54
コア　雑誌	26	7408.00	5806.08
コア　新聞	25	92.28	7586.60
非コア　白書	38	2069.11	3564.87
非コア　Yahoo！知恵袋	42	3986.83	2901.45
RWC　新聞	66	473.79	3903.88

５．結果

ベースラインとアッパーバウンドの結果を表２に示す。Self はタグつきターゲットデータが手に入ったと仮定して、supervised の学習を 5 分割交差検定を用いて行った結果であり、アッパーバウンドである。また、表 3 に提案手法による繰り返し回数が 10 回目（ステージ 10）の 10 セット（ランダムだけ 2 セット）の平均の正解率を表す。表中の「macro」と「micro」はそれぞれマクロ平均、マイクロ平均を表している。表中では各コーパスはそれぞれコアデータの YAHOO 知恵袋（OC）、コアデータの白書(OW)、YAHOO ブログ(OY)、新聞(PN)、書籍(PB)、雑誌(PM)、 非コアデータの YAHOO 知恵袋、(YAHOO)、非コアデータの白書(BCCWJ)コアデータ２種、RWC コーパス(RWC)となっている。図 1 中の、「all_senses_100」は初期事例集合にすべての語義を含む 100 件のデータを使用したもの、「all_senses_200」は初期事例集合にすべての語義を含む 200 件のデータを使用したもの、「random_100」は初期事例集合に完全にランダムな 100 件のデータを使用したものである。図 1 は、全体のマクロ平均と訓練事例を示している。図の「average」は「平均的なコーパス」、「big」は「最大のコーパス」、「all」は「すべてのコーパス」をそれぞれ示す。

表 2　ベースラインとアッパーバウンド

	最大のコーパス		平均的なコーパス		すべてのコーパス		Self	
(%)	macro	micro	macro	micro	macro	micro	macro	micro
OC	68.42	64.22	60.47	53.81	76.09	74.13	79.80	84.02
OW	70.22	73.74	54.48	53.87	65.14	68.18	85.29	90.43
OY	77.04	75.99	67.00	57.86	82.51	86.23	77.22	82.81
PB	76.08	78.73	62.01	62.05	77.58	80.44	79.68	84.76
PM	77.45	78.70	65.16	59.07	78.32	87.61	71.98	87.67
PN	77.73	77.94	64.41	63.04	80.79	81.75	72.77	76.85
BCCWJ	83.06	86.26	64.18	70.78	84.45	86.82	90.47	95.30
YAHOO	75.26	71.22	62.18	55.17	79.28	74.70	89.73	89.23
RWC	79.08	66.12	55.29	51.31	79.92	68.05	82.34	89.20
平均	**77.26**	**75.02**	**61.12**	**59.16**	**79.58**	**77.75**	**80.27**	**90.40**

表 3　各ドメイン別正解率と全体の正解率

	all_senses_100		all_senses_200		random_100	
(%)	macro	micro	macro	micro	macro	micro
OC	72.84	70.12	74.55	72.61	76.23	73.87
OW	65.46	65.81	77.17	80.59	67.13	69.82
OY	76.41	78.20	69.07	72.32	78.23	84.75
PB	73.30	72.55	73.88	73.26	79.05	77.57
PM	74.52	82.87	77.92	76.78	76.60	86.94
PN	76.67	73.84	75.59	83.22	81.06	80.56
BCCWJ	75.57	83.24	75.43	71.64	82.14	86.78
YAHOO	73.33	68.69	77.77	84.21	80.15	78.53
RWC	76.17	71.83	76.96	66.93	77.66	67.34
平均	**74.69**	**73.39**	**76.05**	**74.89**	**78.47**	**75.14**

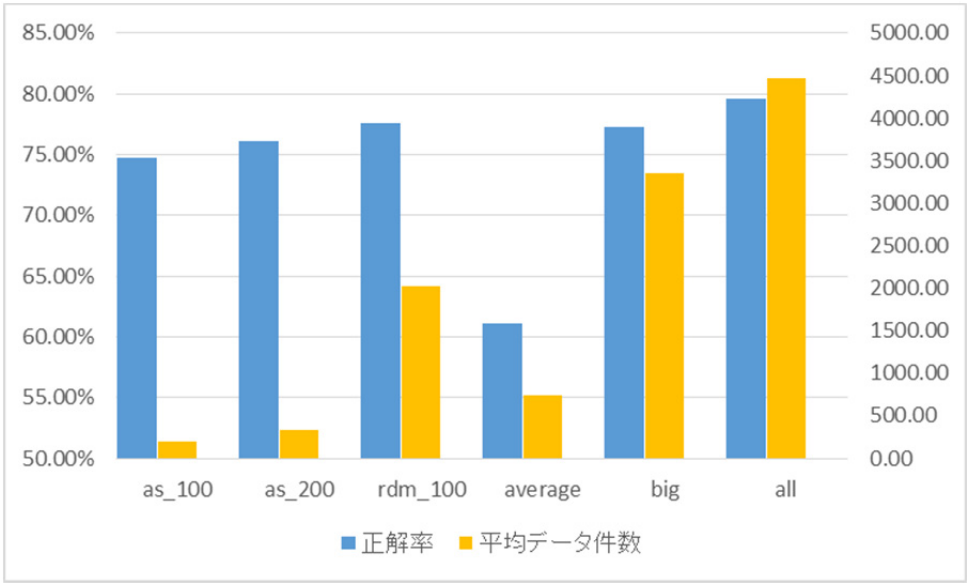


図１　正解率のマイクロ平均と訓練事例数

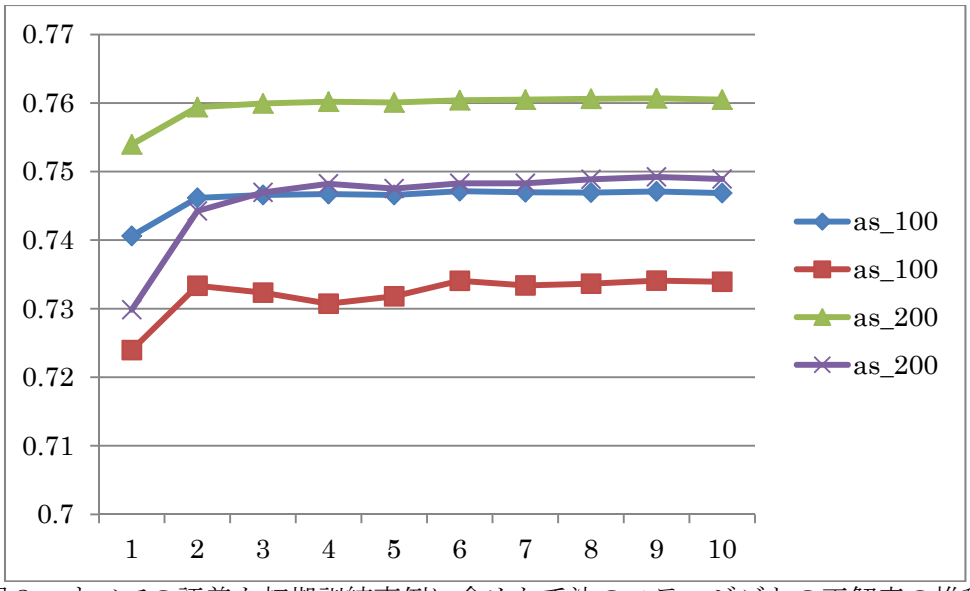


図２　すべての語義を初期訓練事例に含めた手法のステージごとの正解率の推移

６．考察

図１から、提案手法はベースラインよりも、少ないデータ数でベースラインに近い正解率を出していることが分かる。特に、「最大のコーパス」と「random_100」を比較した際、「random_100」の方が、訓練事例数が少ないのにもかかわらず、正解率はわずかながら上回っている。また、「as_100」や「as_200」、そして「random_100」を「平均的なコーパス」と比較すると、「as_100」、「as_200」、「random_100」の方が、訓練事例数が少ないのにもかかわらず、正解率が「平均的なコーパス」を上回っている。このことから、実験で用いた確信度と LOO-bound を用いたスコアが初期事例を選択する際に有効にであったと考えられる。

しかし、表２、表３からベースラインを上回ったのはドメイン別に見ると「白書」のコアデータのみで、全体の平均では、すべてのコーパスの結果に届かなかったことが読み取れる。また、図２を見ると正解率が３回目からはほとんど増加していない。そのため、サ

ポートベクトルを継承することで、分離平面の更新が起こりにくくなり、局所解に陥ってしまったと考えられる。このため、もっとサポートベクトルが入れ替わるような設定をするなどの改良をしたほうがよいと思われる。

次に、図 1 から、「all_senses_100」と「random100」を比較すると、正解率こそ「random_100」の方が優れているが、「all_senses_100」の方がより少ない事例数で分類できていることが分かる。訓練事例数は、「all_senses_100」は 189 件だったのに対し「random_100」は 2030 件であった。このことから、確信度と LOO-bound を用いたスコアが、訓練事例集合に最初から全ての語義を含むことで、より小数の訓練事例で正解率が収束することが分かる。また、「all_senses_100」や「all_senses_200」は、「平均的なコーパス」に比べ、訓練事例数を格段に少なくしながら、正解率を上昇させている。そのため、「all_senses_100」は、少量のデータを使用しながらも比較的、正解率を落とさないことが分かった。

また、「all_senses_100」の結果ステージ 10 の訓練事例が 189 件だったため、「all_senses_100」と、189 件よりも少々多めの 200 件をランダムに選択して、確信度などのスコアを使わずに分類器を作成した場合（すべての語義を含む。また、10 回の平均値）を比較した。その結果、「all_senses_100」はマイクロ平均が 73.39%、マクロ平均が 74.69%だったのに対して、ランダムの 200 件では、マイクロ平均が 72.87%、マクロ平均が 75.16%となった。このうち、マイクロ平均の結果はカイ二乗検定により有意であった。このことから、マクロ平均は、わずかに下がってしまう（有意ではない）が、マイクロ平均は確信度と LOO-bound を用いて上昇したことが分かった。このことから、局所解には陥ったものの、確信度と LOO-bound を用いたスコアにより、サポートベクトルを残して反復的に訓練事例集合を増やしていく手法は、マイクロ平均においては、語義曖昧性解消の学習に有効な訓練事例を選択するのに有効な手法であることが分かった。

7．おわりに

本稿では、semi-supervised な領域適応において、ソースデータに複数ドメインからなるデータを用いた場合に、確信度と LOO-bound を用いて部分集合を選択し、そのサポートベクトルのみを継承し反復的に訓練事例集合を選択する手法について述べた。正解率こそ全てのデータを利用するというベースラインを下回ってしまったが、正解率を大幅には落とさずに、訓練事例数を大幅に減らすことに成功した。また、その際、訓練事例数がより多かった「平均的なコーパス」の正解率を上回った。このことから、提案手法は、学習に有効な訓練事例を選択するという点において有効であることが分かった。

また、サポートベクトルの継承については局所解に陥るという問題があり、この点はもっとサポートベクトルが入れ替わるようにしたほうがよいと思われる。半面、このように反復的な訓練事例の選択を行うことで、微小ながらも正解率を上昇させるということが分かった。今後は、サポートベクトルを継承しないランダムな訓練事例集合を比較対象に含むなど、局所解に陥らないような工夫を施せば、正解率を上げることができるかもしれない。

謝辞

本研究は、文部科学省科学研究費補助金[若手 B（No : 24700138）]の助成により行われました。 ここに、謹んで御礼申し上げます。

参考文献

Chih-Chung Chang and Chih-Jen(2001),「Lin.LIBSVM: a library for support vectormachines」. http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Koichi Hashida, Hitoshi Isahara, TakenobuTokunaga, Minako Hashimoto, Shiho Ogino,and Wakako Kashino(1998).「The rwc text databases」*In LREC 1998*, pp.

457-461.
Kanako Komiya and Manabu Okumura(2012). 「Au-tomatic domain adaptation for word sense dis-ambiguation based on comparison of multipleclassiers」*In PACLIC 2012*, pp. 77-85.
Kikuo Maekawa (2008). 「Balanced corpus of contempo-rary written japanese」 *In ALR 2008*, pp. 101-102.
古宮嘉那子、奥村学、小谷善行. 「分類器の確信度を用いた合議制による語義曖昧性解消の semi-supervised な領域適応」 *第三回コーパス日本語学ワークショップ予稿集*, pp. 1-6, 2013.
古宮嘉那子、小谷善行、奥村学(2013). 「語義曖昧性解消の領域適応のための訓練事例集合の選択」 *第十九回言語処理学会年次大会予稿集*, pp.940-943
国立国語研究所(1964). *『分類語彙表』*. 秀英出版.
西尾実，岩淵悦太郎，水谷静夫(1994). *『岩波国語辞典第五版』*. 岩波書店.

会話における話者のうなずきと発話音声のプロミネンスの時間関係

天谷 晴香（東京大学大学院総合文化研究科）†

Timing Relationships between Prominences of Speaker Head Nods and Pitch Movements

Haruka Amatani (The University of Tokyo)

要旨

発話音声のプロミネンスと発話に伴うジェスチャーのストローク・ピークは一致することが多いと言われる。McNeill(1992)はこれを phonological synchrony rule によるものとした。それらの厳密な時間関係を調査した研究のひとつに Nobe(1996)がある。Nobe は英語話者の類像ジェスチャーのストローク・ピークが発話音声のピッチ・ピークと同期または先行するとしたビート・ジェスチャーもまたストローク・ピークを音声のピッチ・ピークと同期または先行させる。日本語話者の頭部ジェスチャーのうなずきには、ビート・ジェスチャーと似たふるまいを見せるものがあるが、発話のピッチ・ピークとうなずきのストローク・ピークは同期または固定した先行関係が成立しているか。アクセント語と無アクセント語を分類した上で、うなずきとピッチの各ピークの時間関係を明らかにする。

１．はじめに

話者は発話時、言語情報だけではなく非言語情報を豊富に発している。文字や音声情報に加えて、ジェスチャーなどの身体動作情報を加えたマルチモーダルな会話研究は、より包括的な記述で、会話の全体像を捉えようとするものである。

発話に伴う身体動作は、視線の動きや頭の動き、手によるジェスチャーなどがある。特に頭部動作のうなずきは日本語話者に特徴的に多く見られる動きである。メイナード(1993)によると、アメリカ英語話者の約３倍、日本語話者は会話中にうなずいている。

発話そのものに加え、うなずきや動作などが協調して会話のリズムを作っているという分析を、Erickson and Schultz(1982)は英語会話について行った。ザトラウスキー(1997)は、日本語会話のリズムは英語会話のそれとは質的に違うが、日本語会話でも非言語情報が会話リズムに貢献する可能性を示唆した。

発話音声の強弱や上昇下降調とジェスチャーの強弱や方向が一致するとしたのが Bolinger(1983)である。Bolinger のこの主張で、方向が一致するとした部分は後に否定されている(Loehr 2004)。

しかし、音声のピッチの上昇位置にジェスチャーが発現する現象は実際に見られる(Cave et al. 1996 他)。McNeill(1992)はこのような音声とジェスチャーの phonological synchrony rule(音韻共時法則)と呼んだ。Nobe(1996)は、表象ジェスチャーが英語話者によって発せられる時、そのジェスチャーの主要部分であるストロークのピークが音声のピッチ・ピークと同時かもしくは先行して起こると報告している。これと同様に、ビート・ジェスチャーのストローク・ピークが音声のピッチ・ピークと同時か先行して起こった(Loehr)。

発話に伴ううなずきと音声のピッチの関係はどのようになっているだろうか。日本語の単語には語彙アクセントがある。語彙アクセントのない言語の研究から、音声のピッチ・ピークとジェスチャーのストローク・ピークが同期しやすいことが言われている。語彙アクセントは急激なピッチ下降を生じさせ、音声的に際立っている。語彙アクセントによるピッチの動きは語彙アクセントによらないピッチの動きよりうなずきと同期しやすくある

† amatani.haruka@gmail.com

かどうか、会話音声とうなずき頭部動作を詳細に分析することで、明らかにしたい。

２． 発話に伴ううなずき

うなずきと言うと、聞き手のあいづちとしてのうなずき動作がまず思い起こされるが、話し手も発話しながらうなずき動作を相当数行っている。

メイナード(1993)は日本語話者の会話において、話し手のうなずきと聞き手のうなずきが同程度の数、出現したことを報告している。また、庵原ら(2004)は話し手のうなずきが聞き手のうなずきより多く出現したことを報告している。

３．うなずきの種類と出現位置

メイナードは話し手のうなずきの役割に、「同意」「承認」「強調」「節のマーカー」「肯定」「リズム取り」「ターンの受け継ぎに関係する機能」があるとした。

また、前田ら(2003)は、話し手のうなずきは聞き手の反応を要求するものだとしたが、金田(2007)は「対人的な機能は発話全体から見出だされるものであり」、顎刻み（話し手のうなずき）が有するものではないとしている。

金田は、話し手のうなずきの出現位置として、「発話末（句末・文末）」および「重要な箇所の最初のモーラ」を挙げている。

重要な箇所の最初のモーラに身体動作が現れるという現象は、話し手のうなずきを視聴覚韻律(audiovisual prosody)として研究されてきた身体動作のひとつとして考える根拠となる。

視聴覚韻律には、話し手のふるまいを観察した研究から、音声のピッチの動きに付随する、フランス語話者の眉の動き(Cave et al. 1996)、英語話者と日本語話者の頭部動作(Yehia et al. 2002)などが挙げられる。また Yehia et al.が出した頭部動作と音声のピッチ動作は一致しやすいという結果から Munhall et al. (2004)は日本語のデータを使って 3D アニメーションの頭部映像を作り、知覚実験を行って、頭部動作を付随させた音声は聞き取りやすくなるという結果を報告している。Krahmer & Swerts (2007)は手のビート・うなずき・眉の動きを、オランダ語の音声的強調の置かれる単語に付随させて、発話したものを、視聴者に見せる知覚実験を行っている。動きが付随した場合、強調がより強く感じられたとしている。

４．動作としてのうなずきの分析

細馬・富田(2011)は、ジェスチャー区間の観点から聞き手のうなずきを２種類に分類した。ジェスチャー区間は、Kendon(2004)が用いたジェスチャー単位の最も小さなレベルである。細馬・富田は Kendon や細馬(2008)の、主に手のジェスチャー分析で用いられてきたジェスチャー単位を援用し、頭部動作であるうなずきを分析している。

ジェスチャー単位は１つもしくは複数のジェスチャー句から形成される最も大きなレベルである。１つのジェスチャーが１つのジェスチャー句を成し、ジェスチャー句はジェスチャー区間から成り立っている。最も際立ったジェスチャー区間はストロークである。そして、ストロークの前の予備的な動きは準備区間、ストロークの後の元の定位置に戻る動きは復帰区間とされる。また、これらの区間の間に、保持と呼ばれる動きの止まる区間が存在しうる。

金田(2007)は、話し手の「うなずき」は聞き手のうなずきと異なり、顎を正面の位置から上げてからその後下げる「リズムを刻む時のような顎の動きである」ため、「顎刻み」と呼ぶとした。細馬・富田は、この金田の分析にジェスチャー区間を適用して、話し手のうなずきは PS 型と分析している。

このように、うなずきを複数のジェスチャー区間に分けて分析することで、非常に細かな時間単位においてうなずきの生起位置を特定することができる。

５．会話データと分析

５.１　データ

会話は実験室で録音・録画された。２人の参加者によるものである。それぞれ話者 A,B とする。対面で向かい合った状態で会話しており、ビデオカメラは２台で、部屋の隅からそれぞれの身体全体を一人ずつ画面におさめている。音声はマイクをヘッドセットで装着し録音した。

参加者は実験の始めに約 10 分の別々のアニメーションを視聴した。30 分の会話の中で、互いに自分の視聴した内容について説明し合った。分析に用いたデータは 30 分の会話の内、最初の 10 分である。内容は主に互いが観たアニメーションについて説明し合ったものだった。

５.２　音声の分析

会話音声のアノテーションには、X-JToBI(Maekawa et al. 2002)を用いた。また、分析に使用したソフトウェアは Praat(Boersma and Weenink 2014)である。ピッチの動きとして、基本周波数(F0)の動きを採用し、記述した。

本研究では、トーンとアクセント句の判定を重点的に行い、分析対象とした。トーンのアノテーションから、語彙アクセント位置や語彙アクセントに伴わない F0 の下降位置を抽出した。また、うなずきの共起を判断する範囲を、アクセント句とした。アクセント句は多くの場合、文節に対応する範囲である。以下で「アクセント語・無アクセント語に伴うなずき」と言った場合、その語を含むアクセント句内にうなずきのピーク位置があることを意味する。

５.３　動作の分析

動作アノテーションには、細馬・富田が聞き手のうなずきについて行ったように、Kendon のジェスチャー単位を援用した。分析に使用したソフトウェアは、ELAN(Sloetjes and Wittenburg 2008)である。ビデオをコマ送りで視聴し、頭部が動き出すタイミングや軌道が変わるタイミングを記述した。

金田の指摘のように、話者のうなずきが「顎刻み」である場合、むしろその動作のストロークは上方向の動きである可能性が考えられる。上方向の動きのピーク位置、すなわち顔・顎が一番高い位置にある点と、下方向の動きのピーク位置、すなわち顔・顎が一番低い位置にある点が、どちらも可能なうなずきのストローク・ピークになりうる。本研究では、下方向の動きを主に分析対象にしている。ただし、下方向の動きの開始位置で顔・顎は最も高い位置にあることになるため、その時点を仮に上方向の動きのストローク・ピーク位置として、分析に用いた。そのことを明記して、以下、分析結果を報告する。

また、話し手・聞き手のうなずきの区別については、聞き手のあいづちに発声のある場合、それと同時に発せられるうなずきは発話に伴ううなずきとした。

６．分析結果

６.１　うなずきと語彙アクセント

まず語彙アクセントにうなずきが伴いやすいかを調べるため、アクセント語を含むアクセント句とアクセント語を含まないアクセント句においてうなずきの出現率に差があるか測った。表１、表２にそれぞれ、話者 A、話者 B の結果を示した。

表１．アクセント句の語彙アクセントの有無とうなずきの有無の関係（話者 A）

	うなずき有り	うなずき無し
語彙アクセント有り	164	219
語彙アクセント無し	44	78

表２．アクセント句の語彙アクセントの有無とうなずきの有無の関係（話者 B）

	うなずき有り	うなずき無し
語彙アクセント有り	162	369
語彙アクセント無し	62	141

それぞれにカイ２乗検定を行った所、アクセント句の語彙アクセントの有無は、うなずきの生起率に影響していないことが分かった。アクセント語にも無アクセント語にも同様にうなずきが伴ったり伴わなかったりすることが分かった。

６.２　うなずきのストローク・ピークとピッチ・ピーク

話者 A、B それぞれに、うなずきのストローク・ピークであると考えられる下方向の頭部動作の最も低い時点と、音声のピッチ・ピークが最も高い時点の関係を以下、図に示す。また、同時に、話し手のうなずきのストロークが上方向の頭部動作である可能性をふまえて、上方向の頭部動作の最も高い時点と、音声のピッチ・ピークの時間関係も図に示す。

図１～図４が話者 A、図５～図８が話者 B の観測結果である。それぞれ、図１・２と図５・６がアクセント語に共起したうなずきの起きた回数を表しており、図３・４と図７・８が無アクセント語に共起したうなずきの個数を表している。また、図２・４と図６・８が下方向のうなずきのピークとピッチ・ピークの時間差を示しており、図１・３と図５・７が上方向のうなずき（たりえる頭部動作）とピッチ・ピークの時間差を示している。

グラフの X 軸の単位は「秒」である。この「秒」によって、動作ピークとピッチ・ピークの時間差が表されており、その差は動作ピークの起きた時間から、ピッチ・ピークの起きた時間を引くことで算出された。

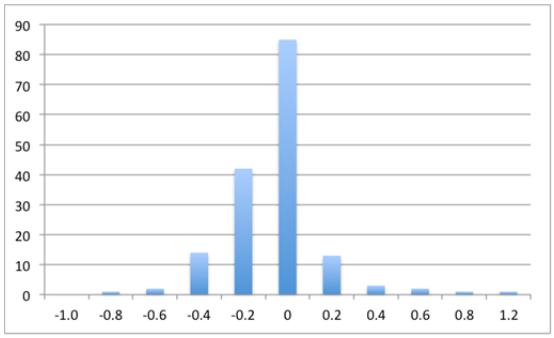


図 1. 語彙アクセントと上方頭部動作のピークの差(話者 A)

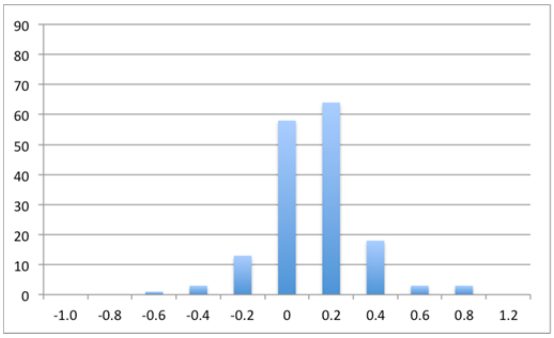


図 2.語彙アクセントと下方頭部動作のピークの差(話者 A)

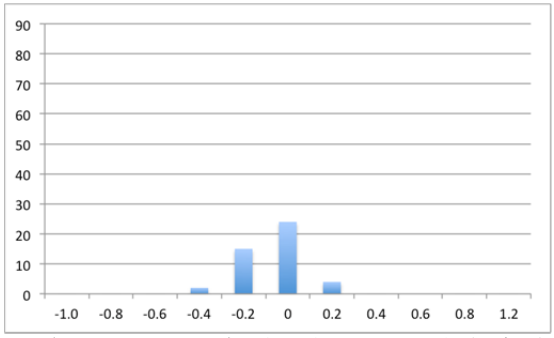


図 3.無アクセントと上方頭部動作のピークの差（話者 A）

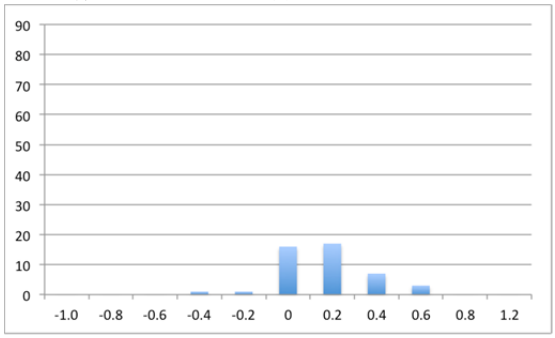


図 4.無アクセントと下方頭部動作のピークの差（話者 A）

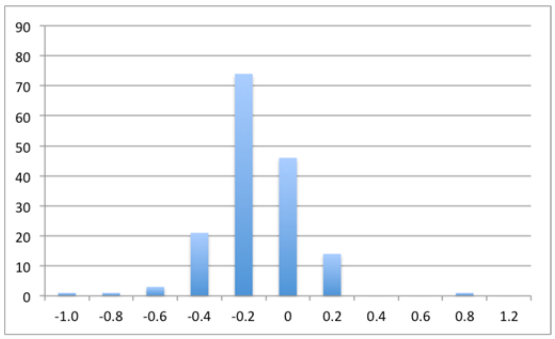


図 5.語彙アクセントと上方頭部動作のピークの差（話者 B）

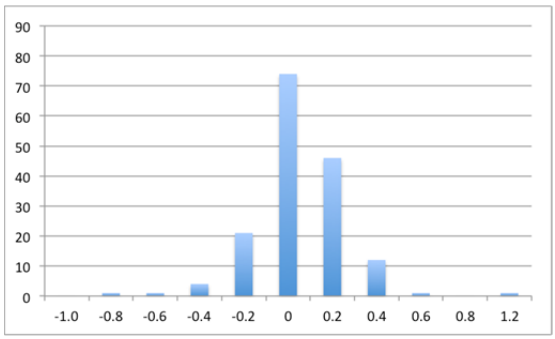


図 6.語彙アクセントと下方頭部動作のピークの差（話者 B）

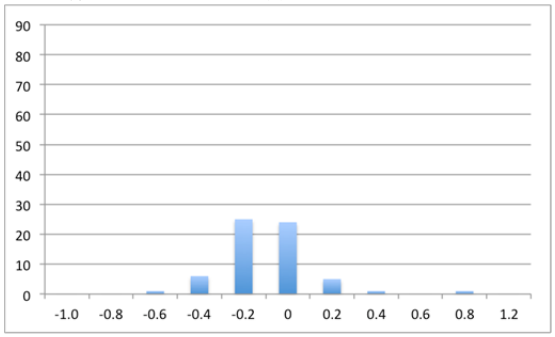


図 7.無アクセントと上方頭部動作のピークの差（話者 B）

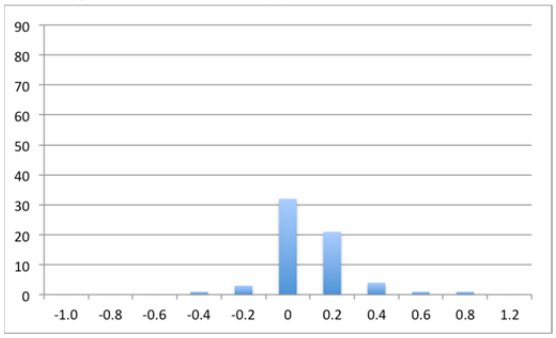


図 8.無アクセントと下方頭部動作のピークの差（話者 B）

話者 A において、上方向の頭部動作のピークは語彙アクセントの有無に関わらず、ピッチ・ピークから 0 秒～0.2 秒の間に最も多かった。また、下方向の頭部動作のピークは語彙アクセントの有無に関わらず、ピッチ・ピークから 0.2 秒～0.4 秒の間に最も多く見られた。

話者 B においては、上方向の頭部動作のピークは語彙アクセントの有無に関わらず、ピッチ・ピークから-0.2 秒～0 秒の間に最も多かった。下方向の頭部動作のピークは語彙アクセントの有無に関わらず、ピッチ・ピークから 0 秒～0.2 秒の間に最も多く見られた。

これらの結果から、うなずきの発生する音声のピッチ・ピークに対するタイミングは、語彙アクセントの有無より個人差が影響する可能性が示唆される。また、個人差はあるが、うなずきはピッチ・ピークとかなり近い位置で起こっていることも分かった。

話者のうなずきを上方向、下方向どちらの動作と捉えるかについては、結果から話者 A では上方向、話者 B では下方向と言えそうな結果になっている。しかし、バラツキもあるため、個々のうなずきを観察し判定するのが望ましく、話者のうなずきの型をひとつに決定することは難しい。

7．おわりに

うなずきのストローク・ピークと音声のピッチ・ピークは、非常に近接して起こっていることが、詳細な動作と音声の分析からわかった。ただし、語彙アクセントの有無はうなずきの発生率に影響していなかった。

音声とジェスチャーのリンクを言語的な要素に基づくものでなく、運動のメカニズムから説明しようとするのが、Rusiewicz (2012)である。言語産出の過程でなく、運動実行の過程を音声とジェスチャーは共有しており、そのために各々のプロミネンスが共起するとする。音声とジェスチャーの運動実行過程の共有を言語産出モデルに取り入れたものに、Tuite (1993)がある。

音声とジェスチャーは協調して発話リズムを作っていると考えられる。そのリズムがどこまで言語的制約に依拠し、どこから運動的なリズムによって説明され得るものかについて示唆を得られるよう、今後、頭部動作と音声の構造を詳細に分析していきたい。

謝　辞

本研究で分析に用いた会話データを収録し、筆者に使用を許可してくださった University of Victoria 博士課程の Thomas Magnuson 氏に感謝いたします。

文　献

Boersma, P. and Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.4, retrieved 4 October 2014 from http://www.praat.org/

Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harlay, F., and Espesser, R. (1996). About the relationship between eyebrow movements and f0 variations. In H.T. Bunnell and W. Isardi (eds.), *Proceedings of the 4th International Conference on Spoken Language Processing*, pp. 2175-2178.

庵原彩子、堀内靖雄、西田昌史、市川嘉(2004)「自然対話におけるうなずきの機能に関する考察」電子情報通信学会技術研究報告．HCS, ヒューマンコミュニケーション基礎 104(445), 13-18.

金田純平(2007)「発話中の話者による頭の動き–のけぞりと顎刻み–」国際シンポジウム「日本語『音声言語』の教育と基礎資料」神戸大学、2007 年 12 月

Krahmer, E. and Swerts, M. (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*, 396-414.

Loehr, D.P. (2004). *Gesture and intonation*. Doctoral dissertation, Georgetown University.

前田真季子、堀内靖雄、市川嘉(2003)「自然対話におけるジェスチャーの相互的関係の分析」情報処理学会研究報告．HI, ヒューマンインターフェース研究会報告 9, 39-46.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. (2002). X-JToBI: an Extended JToBI for spontaneous speech. In *INTERSPEECH*.

メイナード 泉子(1993)『会話分析』くろしお出版

McNeill, D. (1992). Hand and Mind. University of Chicago Press.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility: Head movement improves auditory speech perception. *Psychological Science, 15-2*, 133-137.

Nobe, S. (1996). *Representational gestures, cognitive rhythms, and acoustic aspects of speech: A network/threshold model of gesture production*. Doctoral Dissertation, University of Chicago.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*. (LREC 2008).

Rusiewicz, H.L. (2012). Synchronization of prosodic stress and gesture: a dynamic systems perspective. *Gesture and Speech in Interaction*.

ザトラウスキー ポリー(1997)「日本語の談話のリズム分析　『息の合った』会話を例に」p.101-148, 茂呂雄二(編)、『対話と知』新曜社

Yehia, H.C., Kuratate, T. and Varikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics, 30*, 555-568.

述語項構造を意識した名詞データの構築

竹内 孔一 (岡山大学大学院自然科学研究科)[1]

宮田 周 (岡山大学工学部)

河村 一希 (岡山大学工学部)

Construction of Japanese Noun Data on the Basis of Predicate-Argument Thesaurus

Koichi Takeuchi (Graduate School of Natural Science and Technology, Okayama University)
Syu Miyata (Faculty of Engineering, Okayama University)
Kazuki Kawamura (Faculty of Engineering, Okayama University)

要旨

本発表者は日本語の述語項構造辞書を構築し，公開してきた．そこでは，共通概念を約 1200 程度に定義し，意味役割を 31 種類，細分類で 72 種類定義した．これらをもとに，名詞に関する述語項構造辞書構築のための基本データを 2 種類構築している．1 つは非飽和名詞に関する辞書で最終的には，影山 (2011) が提示する Generative Lexicon の構造を予定している．現段階では，非飽和名詞に対して例文を 2500 文作成し，その全てに対して意味役割を付与した．この作業における問題点や作成された例の質について説明する．さらに「相違がある」と「異なる」が同義であるように，述語と言い換えができる名詞表現がある．これらの類語を類語辞典を参考に人手により作例を構築して作成している．人手による作業の結果,「暇を出す」など慣用句表現に近いものが多く獲得できたことを報告する．

1　はじめに

本研究グループでは日本語の述語項構造に対してシソーラス形式で語義毎に例文を作成し，意味役割と語義概念を付与した事例を構築し公開している[2]．この辞書を拡張する形で，名詞の項構造に関する 2 種類のデータを構築しているので報告する．

ひとつは，言語学において分析されている名詞の項構造 (西山 (2003, 2013); 影山 (2011); 庵 (2007); Pustejovsky (1995); Meyers et al. (2004)) である．名詞の項構造は「その芝居の主役」や「彼の上司」における「主役」や「上司」のように密接に関連する語 (ここでは「芝居」,「彼」であり項と考える) を必要とする語である．言語処理の観点からすると NTCIR の RITE-2 含意認識タスクにおいて例えば

(t1) BLT サンドイッチとは、サンドイッチの一種であり、パンに挿む食材として、ベーコン、レタス、トマト が用いられることから、それぞれの頭文字を取って名づけられた。

(t2) サンドイッチの略称として具材となるベーコン、レタス、トマトの頭文字 BLT が用いられるものがある。

の場合,「一種」「略称」「頭文字」といった言葉が項を要求し，これらの関係を解くことが含意認識を解くことに結びつく (竹内 (2014))．

もう一つのデータは名詞まわりの連語である．例えば「考案する」に対して「着想を得る」などの異品詞間での言い換えデータである．これらデータをどのように構築し，現段階でどの程度集まり，どのような問題があるか次章以降で記述する．

[1] koichi@cl.cs.okayama-u.ac.jp
[2] 述語項構造シソーラス (http://pth.cl.cs.okayama-u.ac.jp).

2　名詞の項構造データの構築

2.1　作成するデータの構造

最初の段階として文献 (竹内 (2014)) に記述したように，名詞と名詞が取る例文を作成し，述語項構造シソーラスの意味役割を付与する．例文のタイプとして現段階では「X の Y は Z」の構文をベースとする．Y が対象とする名詞であり，例えば「創立者」では

[あの図書館]【主体】の創立者は [田中さん]【対象（人)】だ

のようになる.「創立者」の項として「あの図書館」と「田中さん」があり，その意味的関係を表すラベルとして【】内に意味役割を付与する[3]．こうした例文ベースの名詞項構造のデータ構築は英語では NomLex(Meyers et al. (2004)) で行われている．一方で，先行研究として日本語における名詞格フレーム辞書 (笹野他 (2005)) では対象名詞と項の事例の大規模収集に焦点がおかれているため例文は存在しない．しかし名詞の項構造に対して例文ベースで行うことには 2 つの利点があると考えられる．一つ目の利点は項構造データ構築の際に人間が正しく関係を記述しやすいと考えられる点である．これはデータ構築の際に単語のペアを付与する場合[4]と，文として成立する表現を一度考えてから項を同定するのでは，あきらかに，後者の方が人間の言語直感を引き出せると考えられる．二つ目の利点は，名詞項構造の自動付与を視野にいれると例文は機械学習における事例として都合が良いことである．

次にこうした例文ベースのデータから最終的な名詞の項構造を表す Generative Lexicon ベースへの構造 (影山 (2011)) との比較を行っておく.「創立者」の場合には下記の様になる．

	「創立者」
外的分類	人間 (x)
目的・機能	
成り立ち	機関 [w] を創立する ‖ 創立 (x,w)

ここで機関 [w] が先ほどの【主体】にあたるもので,「創立者」 は結局，人間のことを表す部分が例文での【対象（人)】である．また「成り立ち」 の項目では動詞「創立」の項としてこれらの要素が結び付けられる.「創立」は既に述語項構造シソーラスに登録されており，概念と意味役割，さらに例文が定義されている[5]．こうした最終構造と例文を比較すると，例文から対象となる名詞のカテゴリ (先ほどの例では「人間」や「成り立ち」) での項の具現化部分が取り出せる．自動で最終構造は作成できないが，半自動で最終構造が得られる見通しである．

2.2　名詞項構造データの構築作業

上記で説明した例文ベースの事例データを構築するには，1) 対象とする名詞のリストの構築，2) 名詞に対する例文の構築，3) 例文に対する意味役割の付与を行う必要がある．以下，順に説明する．

対象とする名詞リスト

付与対象の名詞は項を持つ名詞であるが，どの名詞が項を持つかというのは前もってわからない．よってまず西山 (2003, 2013) に記載されている非飽和名詞，譲渡不可能名詞をリスト化して登録する．次に，NTCIR の RITE1 と RITE2(含意認識タスク) の開発データ例文すべてを形態素解析して，名詞に該当するものをすべて登録する．これは作成した名詞項構造データの評価として含意認識タスクを利用することを想定しているためである．優先順位としては文献から獲得した名詞リストを先にすることで，確実な非飽和名詞・譲渡不可能名詞のデータを構築する．RITE-2 から得られた名詞のリストには項構造を持たない対象外の名詞も含まれる．よって作業者は不要な名詞を分ける作業を行

[3]意味役割の全体系について簡単な説明が竹内 (2014) にある．
[4]ここで単語のペアの付与とは例えば直接項構造を作業者に記述させるような付与タスクである．
[5]Web サイトで検索して確認できる (http://pth.cl.cs.okayama-u.ac.jp)．

う必要が出てくる．

例文の構築

上記で決定した付与対象候補の名詞のリストに対して「X の Y は Z だ」の例文を作成する．各名詞に対して例文を作成し，後の意味役割付与などのデータ管理を行うためにブラウザベースの作業システムを CakePHP を利用して作成した．作業結果は MySQL に保存できるため，MySQL データを確認することで進捗を確認することが容易になる．

例文の作成において,「X の Y は Z だ」の構文には制約があり，Z は必ず名詞になるように表現する．例えば,「その演劇の主役は太郎だ」のように「太郎」など具体的に入れることで,「主役」は人間であることなどがわかる．これが Z に形容動詞などを許すと「その演劇の主役は立派だ」など表層的には適合しているが，必要とする情報が得られないためである．

しかしながら一方で，項構造がある名詞であるがこの構文では Z を具体的に表現できない場合がある．例えば譲渡不可能名詞「鼻」では「象の鼻はそれだ」となる．これは Z が具体例の名前を求めているためであり，無名のインスタンスでは表現することができず,「それ」などの指示詞でしか表現できない．非飽和名詞でも同様で例えば「理由」では「あの行動の理由はそれだ」という表現になる．現状ではこうしたインスタンスの名前が無い場合の名詞に対してどのような構文を適応すればよいか自明でないため，現段階では「それだ」ではなく例えば「美しい」など作業者が自然だと思う例文を構築している．

意味役割の付与

作成された例文に対して意味役割を付与する．CakePHP による作業システムは例文が作成されると，MeCab による形態素解析を行い，形態素単位に分割して，意味役割の付与が行えるようにする．意味役割の体系は述語項構造シソーラスに準拠するがほとんどの場合,【主体】と【対象】の付与となる．

2.3　名詞項構造データの付与作業結果と考察

対象とする名詞のリストであるが，文献から得られた名詞は 66 語，含意認識タスクから自動で獲得した名詞は 16774 語である．次に例文の付与であるが，学部学生 2 名の作業者に例文を付与していただいた．その結果 2532 事例登録できた．作業から例えば「出身」(「太郎の出身は岡山県だ」) など新たな名詞の項構造例文が付与できている．

一方で，全てが正しい例文ではない．例文を作成する段階で作業ミスがいくつか見受けられる．例えば「花」の例文で「その花はきれいだ」など「花」にかかる項の部分を全く記述せずに表層的に「X の Y は Z だ」に当てはめてしまっている．これは作業者が言語データ付与に未経験であること，また分野としても言語とは関係無かったことが原因として考えられる．また，今回の作業枠組では対応できていないことも原因である．この例ではまず「花」の語義から分類して (植物の花または職場の花など)，次に項として必須となるもの (「植物」や「職場」の具体例) を検討する必要がある．

次に意味役割付与についてであるが 3199 箇所 (約 2500 例文) 付与できている．意味役割の付与作業は例文を作成した作業者とは別で，BCCWJ の意味役割付与を行った作業者が付与した．付与した意味役割のラベルの揺れを確認するために部分的にではあるが別の付与作業者 (BCCWJ の意味役割付与を行った作業者) に付与をお願いしており，現在その結果を分析中である．基本的には意味役割の細分類，つまり【対象（人)】か【対象（生成物)】かなどどういう分類でアノテーションされているかが名詞項構造データを構築する上で重要となる．このあたりを中心に分析をすすめたい．

これに関連して，名詞の項構造の例文と意味役割付与を行うなかで問題となっているのが，名詞の概念カテゴリの必要性である．例えば,「主役」の場合には,「その演劇の主役」のように「X の Y」における X は「演劇などの名詞」がくる．こうした選択制限情報はのちの言語処理では有効と考えられるが必要とされる名詞概念の粒度の予測が立っておらず付与できていない状態である．当然，例文中に「その演劇」とインスタンスで記しているので，これらをもとに類似度計算などでの処理は可能である．

さらに名詞の基本情報として語義が必要である．京都大学名詞格フレーム辞書には国語辞典と規

則から作成した語義に相当するラベルが格スロットとして付与されている．例えば「ドリル」なら「工具」か「演習，問題」かである．ただ自動獲得であるため誤りも少なからず存在し，語義を辞書ベースで分けて付与すべきか，自動獲得ベースのデータを整理して付与すべきか方針がまだ固まっていないのが現状である．

3　名詞まわりの連語

名詞まわりの連語を獲得するために，類語辞典から述語の類語を探し，人手で例文を付与することで連語のデータを構築する．類語辞典としては角川類語辞典を選び，述語項構造シソーラスの述語と類語辞典との単語のマッチングを行い，対応する類語の分類から述語に対する類語候補を獲得した．これをもとに人手で言い換えとなっている語を抽出し，連語表現を作成した．下記の表に獲得した例を示す．

連語	シソーラスの述語	例文
違いがある	異なる	報道と事実に相違がある
着想を得る	思いつく	漫才師がネタの着想を得る
手抜かりがある	荒っぽい	仕事に手抜かりがある
焼き餅を焼く	妬ける	周囲が二人に焼き餅を焼く

アノテーション作業により現在 100 語ほど獲得できている．各例文には意味役割付与を行っている．

4　まとめ

述語項構造シソーラスの体系を利用して，名詞に関連した項構造データと連語データの構築を行っている．意味役割ラベルと語義概念を一貫して構築できるのが利点である．現段階では項構造では約 2500 の例文を構築して，意味役割付与が一人の作業者で付与できた段階である．今後，項構造のデータの評価ならびに拡張，連語データの拡張を行う予定である．

謝辞

本研究は，科研費 (26370485) の助成を受けたものである．

文献

Adam Meyers, Ruth Reeves, and Catherine Macleod (2004) “NP-External Arguments: A Study of Argument Sharing in English,” in *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pp. 96–103.

James 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．

笹野遼平，河原大輔，黒橋禎夫 (2005)「名詞格フレーム辞書の自動構築とそれを用いた名詞句の関係解析」，自然言語処理，第 12 巻，第 3 号，pp.129–144．

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．

口頭発表

3月11日（水）　10:00～12:00

コーパスに基づく日中副詞「絶対」と“绝对”の対照研究

郭 敏（北京師範大学外国語言文学学院）†

Comparison of Japanese Adverb ZETTAI and Chinese Adverb JUEDUI: A Corpus Study

Guo Min(Graduate School of Foreign Languages and Literature,Beijing Normal University)

要旨

日本語の「絶対」と中国語の“绝对”は副詞としてモダリティを表すのに重要な機能を果たしている。本稿は日中副詞「絶対」“绝对”がどのようなモダリティ表現と共起するか、どのような文類型に使用されるかを考察するものである。「現代日本語書き言葉均衡コーパス」と“北京语言大学汉语语料库（BCC）”（北京語言大学漢語コーパス）を使用し、日中副詞「絶対」“绝对”の用例を採取し、共起するモダリティ表現形式について量的分析を行った。先行研究に基づき、検索されたモダリティ表現を分類し、使用される文類型と関連付け、各文類型毎における両者の使用実態と用法の異同を考察した。

1. はじめに

日中同形語である日本語の「絶対」[1]と中国語の“绝对”はいずれも副詞として使用できるものの、相違点も指摘されている（張・楊（1995）、楊（2013））。本稿は「現代日本語書き言葉均衡コーパス」（以下、BCCWJ と呼ぶ）と“北京语言大学汉语语料库（BCC）”（北京語言大学漢語コーパス、以下 BCC と略称）を使用し、日中副詞「絶対」“绝对”に関して、共起するモダリティ表現、使用される文類型二点について調査を行い、両者の使用の実態と用法の異同を考察する。

2. 先行研究

2.1「絶対」について

副詞の「絶対」の用法について、辞書では以下のように記述されている。

「絶対」　その物事がどのような条件下でも必ず成立するという、話し手の強い気持ちを表す。例：「絶対成功させたい」「絶対君が間違っている。」等。

『明鏡国語辞典第二版』(2010)大修館書店

「絶対」とモダリティとの共起関係についての研究には、佐治（1992）と坂口（1996）がある。坂口（1996）は「絶対」「必ず」「キット」等 5 副詞を取り上げ、働きかけ文との共起関係を考察し、副詞の語彙的意味が統語的現象に与える影響を考察した。佐治（1992）は「絶対」「キット」「必ず」「どうしても」4 語の用例を作成し、13 人を対象として作例の許容度を調査した。許容度の高い「絶対」の共起対象が明らかになった。

† guomin199201@163.com

[1] 以下、日本語は「」で、中国語は“”で表す。また、「絶対」は「絶対」「絶対に」「ぜったい」「ぜったいに」のすべてを含む。

しかし、「絶対」と様々なモダリティ表現の共起頻度、「絶対」の使用実態などについてまだ研究する余地があると考えている。

2.2“绝对”について

“绝对”の用法について、以下の記述がある。

［副］1.表示对事物的肯定或否定，带有较浓的主观色彩。这个人绝对老实/这东西绝对便宜/他绝对不会失约 2.表示不受任何条件的限制，带有强调的意味。多用于祈使句。这件事你绝对要保密/今天大家绝对不能离开这里（［副詞］1.物事に対する肯定または否定の態度を表し、やや主観的な意味あいが強い。例：この人は絶対におとなしい/これは絶対に安い/彼は絶対に約束を破らない等。2. なにものにも制限拘束されないで、強調の意味を帯びている。“祈使句”[2](広義の命令文)に多用される。例：このことは絶対内緒にしなさい、今日みんなは絶対ここを離れてはいけない等）[3]

『現代漢語虚詞詞典』（2001）商務印書館

これらの記述では、「絶対」及び“绝对”2 語ともに話し手の気持ち、判断が表れる語となっている。しかし、具体的に共起頻度の高いモダリティ表現、多用される文の類型、両者の使用実態の異同については明らかではない。

2.3「絶対」と“绝对”の異同について

張・楊（1995）及び楊（2013）は「絶対」と“绝对”が使用される文脈を調査した。張・楊（1995）は「中国語の“绝对”は判断文とのみ共起し、意志・命令・依頼表現などとは共起しないが、日本語の「絶対」はそのいずれとも共起する」と述べている。楊（2013）も同意見である。

しかし、張・楊（1995）、楊（2013）は作例、限られた使用例と内省とによって考察されてきたため、使用実態と若干相違がある。たとえば、BCC コーパスから以下の例が見られた（下線部は筆者による）。

(1)　“我绝对想继续唱，”帕瓦罗蒂在意大利《新闻报》24 日刊登的访谈中说。（「わたしは絶対に歌い続けたいです。」ルチアーノ・パヴァロッティはイタリアの『新聞法』のインタビューを受けた時にそういった。）

（福建日报/2006-7-26/帕瓦罗蒂出院）

(2)　“不，乔治，这种事情你绝对别干。”（「いや、ジョージ（人の名前）、こんなことを絶対するな。」）　（布雷登/UN/奥德利夫人的秘密）

(3)　章仲箫（四下望了一望）：“还有，请你绝对保守秘密！我看见了凤鸣大哥！”（章仲箫さん(周りを見て)「それから、絶対秘密を守ってください！凤鸣さんに会ったよ！」）

（老舍/1943/谁先到了重庆）

例（1）は意志表明の文であり、例（2）は否定命令文であり、例（3）は依頼文であるが、共に“绝对”が使われている。これは張・楊（1995）の「中国語の“绝对”は意志・命令・依頼

[2] “祈使句”とは伝達機能から名付けられ、命令・依頼または制止の意味を表す文のことである。

[3]以下、本文中の翻訳は筆者によるものである。

表現などとは共起しない」、楊（2013）の「命令と意志表明の文脈では中国語の“绝对”は使えない」といった主張とは齟齬がある。より多くの使用例による精査が待たれるところである。

3. 調査の概要

3.1 調査の目的

本稿では、中国語と日本語のコーパスを用いて、副詞「絶対」・“绝对”の用例を採取し、共起するモダリティ表現について量的分析を行う。次に、「絶対」・“绝对”がどの種類のモダリティと共起しやすいか、どのような文類型に使用されるかを調査し、各文類型毎に両者の使用実態と用法の異同を考察する。

3.2 データと方法

本稿で使用した日本語のデータは、国立国語研究所が構築した『現代日本語書き言葉均衡コーパス』（BCCWJ）の「出版・書籍」サブコーパスのコア・非コアデータすべてである。書き言葉のコーパスであるが、地の文と会話文のいずれも含まれており、広範囲で多様な使用場面における言葉の使用実態を調査できることが利点である。検索にはBCCWJの検索用Webインターフェースツールであるコーパス検索アプリケーション「中納言」[4]を使用し、副詞の「絶対」[5]628[6]件、「絶対に」[7]1174件、総計1802件を採取した。

一方、本稿で使用した中国語のデータは、“北京语言大学汉语语料库（BCC）”[8]（北京語言大学漢語コーパス、以下BCCと略称）の「総合」サブコーパスである。BCCコーパスは総計150億字が含まれ、「新聞」「文学」「マイクロブログ」「科学」「総合」「古代中国語」など数多くの分野のサブコーパスが含まれ、中国の現代社会の言語生活を反映する大規模コーパスである。BCCWJの「出版・書籍」サブコーパスが総記、哲学、文学、社会科学など様々なジャンルが含まれる。それに対応するため、BCCの「総合」サブコーパスを利用した。副詞の“绝对”[9]を63311例を採取した。

また、実際の用例の分析のために、採取された「絶対」と「絶対に」の用例から500例ずつ、“绝对”の用例から1000例をランダムサンプリングし、目視により分析することとした。

---

[4] https://chunagon.ninjal.ac.jp/login

[5]検索式は次のとおりである。語彙素読み ＝"ゼッタイ" AND 品詞 LIKE "副詞%") IN (registerName="出版・書籍" AND core="true") OR (registerName="出版・書籍" AND core="false") WITH OPTIONS unit="2" AND tglWords="20" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

[6] 4 の検索式より採取された用例は730件であるが、そのうち名詞の用法、「絶対主義」等の漢字熟語を削除した数である。

[7] 検索式は次のとおりである。キー: (語彙素読み ＝"ゼッタイ" AND 品詞 LIKE "名詞%") AND 後方共起: 語彙素読み ＝"ニ" ON 1 WORDS FROM キー IN (registerName="出版・書籍" AND core="true") OR (registerName="出版・書籍" AND core="false") WITH OPTIONS unit="2" AND tglWords="20" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

[8] http://bcc.blcu.edu.cn/

[9] 検索式は「绝对/d」である。「/d」によって品詞を副詞に指定する。

4. 調査結果

まず、「絶対」“绝对”と共起するモダリティ表現について量的調査を行った。その結果が、表 1、表 2 である。紙幅の制約上、共起頻度が一番高い表現から 10 項目のモダリティ表現を表示した。

表 1「絶対」と共起頻度の高いモダリティ表現

「出版・書籍」サブコーパス / モダリティ表現	「絶対」（総 1000 件）	
	出現数（件）	使用頻度
～φ（断言）	550	55.0%
する（意志）	152	15.2%
と思う	36	3.6%
てはいけない	27	2.7%
だろう	24	2.4%
なければならない	24	2.4%
はずだ	21	2.1%
たい	21	2.1%
するな（禁止）	21	2.1%
することだ	20	2.0%

表 2“绝对”と共起頻度の高いモダリティ表現

「総合」サブコーパス / モダリティ表現	“绝对”（総 1000 件）	
	出現数（件）	使用頻度
～φ（断言）	620	62.0%
不会（はずがない）	134	13.4%
不能（てはいけない）	73	7.3%
会（はずだ）	46	4.6%
要（なければならない）	27	2.7%
不可(てはいけない)	19	1.9%
能（だろう）	13	1.3%
不要（てはいけない）	12	1.2%
可以（てもいい）	12	1.2%
V（意志）	6	0.6%

表 1 と 2 からモダリティ表現形式の詳細を比較すると、「絶対」と“绝对”と共起する上位 3 項目のモダリティ表現形式がそれぞれ全体の 75.8%、82.6%を占めており、共起するモダリティ表現に偏りがあることが明らかである。

5. 考察

本節では、検索されたモダリティ表現を分類し、「絶対」と“绝对”がどの種類のモダリティ表現と共起できるか、どのような文類型で使用されるかを考察し、「絶対」と“绝对”の用法と関連付けて考察する。

5.1 モダリティ表現との共起関係

モダリティ表現の文法研究はこれまで数多く行われているが、本稿では仁田（1991）に従って考察を進めていく。文は「言表事態」（命題）と「言表態度」（モダリティ）からなっている。モダリティは、大きく「言表事態めあてのモダリティ」と「発話・伝達のモダリティ」との二種に分かれる。「発話・伝達のモダリティ」とは、文をめぐっての発話時における話し手の発話・

伝達的態度のあり方を表す文法表現である。仁田（1991）は「文は発話・伝達のモダリティによって文に成る。発話・伝達のモダリティは文の存在様式である。従って、発話・伝達のモダリティの下位類化は、文類型の下位類化でもある」と述べている。仁田（1991）と日本語記述文法研究会編（2003）を参考に、日本語の発話・伝達のモダリティの下位分類、文類型と主な言語形式をまとめたものが、表 3 の日本語の部分である。さらに、王（2011）を参考に、対応する現代中国語の主な言語形式を書き加えたものが、表３の中国語の部分である。以上の項目に基づき、検索されたモダリティ表現を分類し、「絶对」と“绝对”が共起するモダリティ表現と文類型の用例数と使用頻度を表 3 にまとめた。

表 3　モダリティ、文類型の分類と主な言語形式

モダリティ、文類型の分類と下位分類			日本語の主な言語形式	中国語の主な言語形式	絶对（総 1000 件）	“绝对”（総 1000 件）
働きかけ（働きかけ文）	命令（命令文）	命令	命令形	必须，得 děi,	9（0.9%）	5（0.5%）
		依頼	てくれ、てください、てちょうだい	要，应该	21（2.1%）	7（0.7%）
		禁止	するな	不准，不得 dé，不许など	21（2.1%）	6（0.6%）
	誘い掛け（勧誘文）		（よ）う、ましょう	必须，要，应该	5（0.5%）	0（0%）
表出（表出文）	意志・希望（意志文）		する、（よ）う、つもりだ、まい、たい、	V，想，要，肯，愿意，乐意	182（18.2%）	9（0.9 %）
	願望		命令形	希望，想など	0（0%）	0（0%）
判断のモダリテ（判断文）	真偽判断	断定	～φ	～φ	550（55.0%）	620（62.0%）
		推量	だろう、まい、と思う	要，能，会，可能	76（7.6%）	16（1.6%）
		蓋然性	かもしれない、にちがいない、はずだ	能，会，可能	23（2.3%）	180（18%）
		証拠性	ようだ、らしい、（し）そうだ	無	3（0.3%）	0（0.0%）
	当為判断	適当	べきだ、ほうがよい	应该（应当，应，该，当）得 děi,	24（2. 4%）	8 0.8%
		必要・不	なければならない、	必须，不得不	28	28

		必要	なくてはいけない等		(2.8%)	2.8%
		許可・不許可	てもいい、てはいけない等	能，可，可以，准，许，不能，不准，不许	50 (5%)	118 (11.8%)
問いかけ（問いかけ文）			か、だろう？等	吗？等	8 (0.8%)	3 （0.3%）

5.2「絶対」と“绝对”の使用される文類型

ここでは、調査語がどのような文で使用されるのかという点から分析することにする。図1は「絶対」及び“绝对”の各文類型における使用頻度を示したものである。

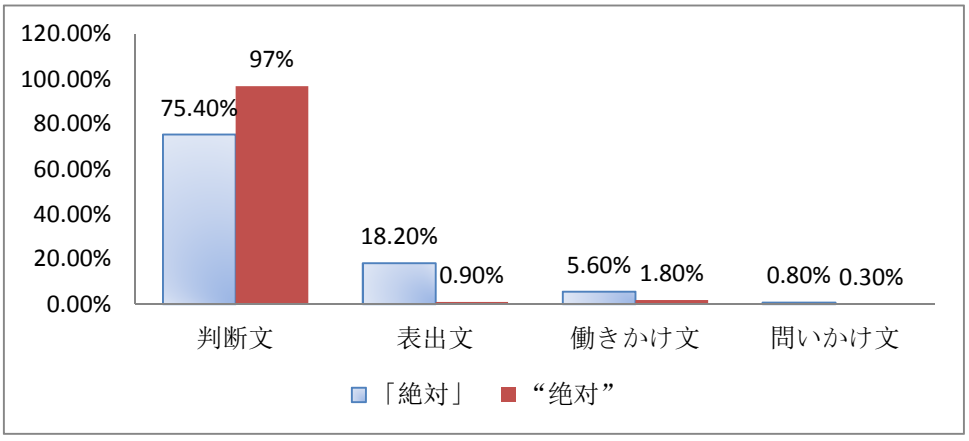


図1「絶対」及び“绝对”の各文類型における使用頻度

図1にあるように、「絶対」が最も多く使用されるのは判断文であり、全体の75.4%を占めている。次いで、表出文が18.2%を占め、三番目に働きかけ文が5.6%であり、最後に問いかけ文が0.8%を占めている。

一方、“绝对”は判断文に最も頻繁に使われ、全体の97%を占めている。次に、わずか1.8%と 0.9%がそれぞれ命令文と意志文に使用される。最後に 0.3%が問いかけ文において用いられる。

以上の結果から、「絶対」と“绝对”の主な用法は判断を表すことが分かった。このことから、「絶対」と“绝对”は「その物事がどのような条件下でも必ず成立するという、話し手の強い気持ちを表す」という意味が基底にあり、使われる文の違いによって、判断の確信度の高いこと、意志表明の強いこと、命令態度の強いこと、勧誘態度の強いことを強調することなどの意味が伴うと考えられる。しかし、“绝对”の判断の用法は絶対多数を占め、その使用頻度の割合において極端な偏りを示している一方で、「絶対」はより分散的な意味分布が見られる。

次に、各文類型毎に「絶対」及び“绝对”の使用実態を考察する。

5.2.1 判断文における「絶対」と“绝对”

判断文において、「絶対」と“绝对”がほぼ同様な使用傾向が見られる。

判断文は大きく「真偽判断」の文と「当為判断」の文に分けられる。65.2%の「絶対」と 81.6%の“绝对”は「真偽判断」の文に使用される。「真偽判断」は「絶対」と“绝对”の主な用法と言える。さらに、「真偽判断」の文が「断定」と「非断定」（推量、蓋然性判断、証拠性判断）に分けることができる。55%の「絶対」、62% の“绝对”は「断定」の文に使用されている。これは「絶対」と“绝对”の確信度が高いことを示している。

10.2%の「絶対」と 15.4% の“绝对”は「当為判断」の文に使用される。さらに、「当為判断」の文が「適当」、「必要・不必要」「許可・不許可」に分かれる。そのうち、「絶対」と“绝对”いずれも「適当」より、「てはいけない」“不能”“不可”“不能”（てはいけない）のような「不許可」のモダリティ表現と「なければならない」“要”（なければならない）のような「必要」のモダリティと共起しやすい。これも「その物事がどのような条件下でも必ず成立するという、話し手の強い気ちを表す」という意味と関わっていると考えている。

5.2.2 意志文における「絶対」と“绝对”

意志文における使用頻度において、「絶対」と“绝对”は極めて大きな差異を示している。18.2%の「絶対」は意志文に使用される一方で、わずか 0.9%の“绝对”は意志文に使用される。この点に関しては、張・楊（1995）と楊（2013）の主張とは齟齬がある。

張・楊（1995）は以下の例（4）を用い、「中国語の“绝对”は意志表現とは共起しない」、楊（2013）は例（5）を使い、「意志表明の文脈では中国語の“绝对”は使えない」と論述している。

（4）　a*我绝对去。
　　b 私は絶対行く　　張・楊（1995）

（5）　a??明天我绝对去。
　　b 明日絶対行く　　楊（2013）

しかし、BCC コーパスから採取した例の中で、中国語の“绝对”が意志文で使用される例も見られる。

（6）　a 我追问说：“为什么我不能去？如果你不解释清楚，我绝对要去！”
（雨侠/唯我独魔）
b「どうしてわたし行ってはだめなの。ちゃんと説明してくれないと、絶対行く！
（筆者による例（4a）の翻訳）

例 (4a) 及び例 (5a) は非文と非常に不自然な文と指摘されている (張・楊 (1995)、楊 (2013)) が、コーパスで例（6a）が見られる。原因を探るために、例（4a）、例（5a）と例（6a）を比較し、相違点が見られる。例（6a）で“绝对”は意志のモダリティを表す法助動詞“要”と共起し、話し手の意志を表す。一方、例（4a）と例（5a）は意志のモダリティを表す法助動詞を伴

わず、単に意志動詞“去”（動詞の無標形式）が述語になっている。中国語のモダリティは主に法助動詞によって表現されるが、法助動詞と共起しないと意志のモダリティを表せないとは言えない。表 3 のとおり、“要”などの法助動詞のほかに動詞の無標形式も意志のモダリティを表せる。例えば、

（7）　我看出蒋的用意是要我服从他，便说：“我绝对服从我们的副司令。”
（蒋さんが私を服従させたがっているのがわかったので、「絶対副司令官に服従する。」と私は言った。）
（李敖、汪荣祖\蒋介石评传)

例（7）で、“绝对”と動詞の無標形式と共起し、意志を表明する。従って、法助動詞と共起しないのは例（4a）及び例（5a）が非文と不自然な文と見なされた原因ではない。

次に音節と語感の観点から考察する。『現代漢語虚詞用法小詞典』（1984）は“绝对”は常に双音節語と共起すると記述しているが、例（4a）及び例（5a）で“绝对”は単音節語“去”と共起する。そのために、例（4a）及び例（5a）はそれぞれ非文と非常に不自然な文と見なされたと考えている。筆者からみれば、文脈がない場合に例（4a）と例（5a）はやや不自然だが、文脈があれば自然になると考える。例（4a）及び例（5a）についての語感を調べるために、筆者が簡単な調査を行った。調査対象である中国語母語話者 10 人の中で、文脈がある場合に例（4a）と例（5a）が使えるという意見を持っている人が 6 人もいた。従って、大規模コーパスを利用し、客観的で数多くのデータを採取し分析することが非常に重要だと考える。

5.2.3 働きかけ文における「絶対」と“绝对”

5.6%の「絶対」と 1.8%の“绝对”は働きかけ文に使用されている。そのうち、0.5%の「絶対」は勧誘文に使用される。それ以外すべて広義の命令文[10]（命令文・依頼文・禁止文）に使用される。「絶対」と“绝对”は話し手の強い気持ちを表すため、勧誘文に使用される場合相手への押しつけが強くなる。このようなポライトネス上の要素に制限され、日常会話では「絶対」と“绝对”いずれも頻繁に使われていないことが分かった。

「絶対」と“绝对”はいずれも命令文で使えるが、相違点がある。命令文において、「絶対」は命令のモダリティと共起するが、“绝对”は当為判断のモダリティと共起する。

（8）　“这到底是什么问题呢？”“对这件事你绝对要守口如瓶。我的年轻朋友。”
（「これはいったいどんな問題か」「このことについて絶対内緒にしなさい。私の若い友達。」）
（王永成/恐惧的总和）

例（8）は意味的に命令文であり、例（8）の“要”を日本語の「しなさい」に翻訳したほうが自然だが、“要”は中国語で「表出」のモダリティ、「判断」のモダリティ両分野にまたがる法助動詞である。日本語と違い、中国語には命令・依頼・禁止・勧誘の働きかけ専用のモダリ

[10] 以下、広義の命令文を「命令文」と呼ぶ。

ティ表現が存在しない。そのかわりに、中国語の当為判断のモダリティは特定の条件の下で、働きかけの機能を果たす。当為判断の法助動詞は、二人称主格を取り、話し手の当為判断を表した部分を非過去形にすることによって、働きかけの表現となる。

5.2.4 問いかけ文における「絶対」と“绝对”

0.8%の「絶対」と 0.3%の“绝对”は問いかけ文に使用されている。「絶対」と“绝对”の問いかけの用法は使用頻度が最も低いと言える。以下、用例を考察する。

（9）（说话人在寻找安全住所。手下金鹏为其推荐黄石镇）“金鹏，前面就是你说的黄石镇？”“是的。”“绝对安全吗？”“我们的人三个月来查过一次，全镇的人都是土生土长的，除了一个沙大户。”

（古龙/1975 /剣神）

（話手が安全な場所を探そうとしている。部下の金鵬さんが「黄石鎮」を薦めた。）「金鵬さん、この前はあなたが言った黄石鎮なのか」「はい、そうです。」「絶対安全か」「3 カ月前うちのメンバーが一度調べた。黄石鎮の人々は全部地元生まれ育ちの人だよ。沙大戸という人一人以外。」

（10）「ダッフルバッグの中にドラッグを入れてたんだ」「それは絶対に確かかな？」ボールトは訊ねた。「もしそれが空港で見た男、トラックに乗ってた男だとしたら、わ れ わ れ に と っ て は と て も 重要なことで、だから確かめておきたいんだ。

（PB29_00403）

例（9）と例（10）の問いかけ文はすべて情報要求の文である。二つの例では、「絶対」と“绝对”で問いかける前に、話し手は相手との話によって、「黄石鎮が安全かどうか」、「ダッフルバッグの中にドラッグが入っているかどうか」といった問題について既に大体判断した。しかし、それらの問題は話し手にとって非常に重要なので、確かな情報を聞こうとする。そこで、「絶対」と“绝对”を用いて、相手に最も確かな情報を要求する。これも「絶対」と“绝对”の「その物事がどのような条件下でも必ず成立するという、話し手の強い気持ちを表す」という意味に関わっていると考えられる。

6. まとめ

本稿では、中日同形語である「絶対」と“绝对”が共起できるモダリティ表現と使用される文類型について調査した。本稿は BCCWJ「出版・書籍」と BCC「総合」サブコーパスを使用し、日中副詞「絶対」“绝对”がどの種類のモダリティ表現と共起するか、どのような文類型で使用されるかを調査し、「絶対」と“绝对”の用法と関連付けて考察した、以下のような結論が得られた。

第一に、「絶対」と“绝对”と共起するモダリティ表現形式を比較すると、「絶対」と“绝对”と共起する上位 3 項目のモダリティ表現形式がそれぞれ全体の 75.8%、82.6%を占めており、共起するモダリティ表現に偏りがあることが明らかである。

第二に、使用される文類型からみれば、「絶対」と“绝对”がいずれも「判断文」「表出文」「働きかけ文」「問いかけ文」に使用されている。「絶対」が最も多く使用されるのは判断文であり、全体の 75%をも越えている。次いで、表出文が 18.2%を占め、三番目に働きかけ文が 5.6%であり、最後に問いかけ文が 0.8%を占めている。“绝对”は判断文に最も頻繁に使われ、全体の 97%を占めている。次に、わずか 1.8%と 0.9%がそれぞれ命令文と意志文に使用される。最後に、0.3%が問いかけ文において用いられる。

第三に、「絶対」と“绝对”の主な用法は判断を表すことが分かった。「絶対」と“绝对”は「その物事がどのような条件下でも必ず成立するという、話し手の強い気持ちを表す」という意味が基底にあり、使われる文の違いによって、判断の確信度の高いこと、意志表明の強いこと、命令態度の強いこと、勧誘態度の強いことを強調することなどの意味が伴うと考えられる。しかし、“绝对”の判断の用法は絶対多数を占め、その使用頻度の割合において極端な偏りを示している一方で、「絶対」はより分散的な意味分布が見られる。

本稿では、主に「絶対」と“绝对”が共起するモダリティ表現、使用される文の使用実態を考察したが、このような使用実態を引き起こす具体的な要因については次回の課題とする。

文　献

日本語関係

坂口和寛（1996）「副詞の語意的意味が統語的現象に与える影響―働きかけ文での共起関係を中心に―」『日本語教育』91、pp.1-12、日本語教育学会
佐治圭三（1992）『外国人が間違えやすい日本語の表現の研究』ひつじ書房
杉村泰（2009）『現代日本語における蓋然性を表すモダリティ副詞の研究』ひつじ書房
張麗群、楊凱栄（1995）「日本語の『絶対』と中国語の“绝对”」『教養研究』、1:3、pp.117-133、九州国際大学
仁田義雄（1991）『日本語のモダリティと人称』ひつじ書房
日本語記述文法研究会編（2003）『現代日本語文法４』
益岡隆志(1991)『モダリティの文法』くろしお出版
『明鏡国語辞典』（2010）大修館書店
楊凱栄（2013）「誤用例にみる日中表現の違い―日中対照研究の現場から―」『日本語学』、32:13、pp.54-64、明治書院

中国語関係

刘月华、潘文娱、故韡（1983）《实用现代汉语语法》外语教学与研究出版社
张斌（2001）《现代汉语虚词词典》商务印书馆
王晓华（2001）现代日汉情态对比研究
http://www.cnki.net/KCMS/detail/detailall.aspx?filename=1012251630.nh&dbcode=CDFD&dbname=CDFDLAST2012
王自强（1994）《现代汉语虚词用法小词典》上海辞书出版社
吕叔湘（1980）《现代汉语八百词》商务印书馆

中古歌合日記の品詞比率

冨士池　優美（中央大学）†

Part of Speech Ratio of "*Utaawase Nikki*" in the Heian Period

Yumi Fujiike　(Chuo University)

要旨

中古から中世にかけての歌合は、和漢混淆文が一般化する過程において、和歌の実作に基づき、和歌のあり方や歌ことばの用法について評論が加えられた資料と言える。その中でも歌合の記録である日記については、その資料性が明らかにされていない。本発表では、「天喜四年四月三十日皇后宮寛子春秋歌合」の漢文日記と仮名日記という異なる文体で書かれる 2 種類の日記を調査対象とした。調査の結果、長単位データに基づく名詞率と MVR を用い、品詞比率から見られる歌合日記のテキストの特徴は「要約的な文章」として位置づけられ、名詞率の高さが特徴的であることが明らかになった。また、名詞率と文の長さの関係について検討した結果、これまでの指摘とは異なり、文が短いほど名詞の比率が高かった。ここから、語数（音数）の制約や文の長さ以外に、名詞率の増加の要因が存在することが示唆された。

１．はじめに

歌合、中でも中古から中世にかけての歌合は、和漢混淆文が一般化する過程において、和歌の実作に基づき、和歌のあり方や歌ことばの用法について評論が加えられた資料と言える。歌合は序文・歌・判詞・日記といった多様な要素を持つが、日記については特に、これまで日本語史の資料として扱われていなかった面があり、その資料性は明らかにされていない。

本発表では、中古歌合のうち「天喜四年四月三十日皇后宮寛子春秋歌合」の日記を対象とする。「長単位」に基づく名詞率と MVR（100×相の類の比率／用の類の比率）を用い、中古歌合日記の文体的特徴を見出すことを主目的とする。特徴を明らかにするにあたり、『日本語歴史コーパス　平安時代編』の各作品との比較を行う。調査にあたっては「中古中世歌合コーパスに基づく和歌評論の語彙論的研究」（研究課題番号：25770179）で構築中の『歌合コーパス』と『日本語歴史コーパス　平安時代編』とを用いた。

† fujiike@tamacc.chuo-u.ac.jp

２．調査対象

２．１　資料

（１）歌合コーパス

発表者は現在、中古から中世初期にかけて開催された歌合を対象としたコーパス『歌合コーパス』を構築中である。この『歌合コーパス』には、歌合の中でもまとまった散文箇所と言える歌合日記を収録し、形態論情報を付している[1]。

ここで歌合日記について、説明したい。歌合日記は歌合の記録である。歌合には行事的諸要素がある。例えば、和歌の題や左右の頭、文台、員差(かずさし)[2]の州浜等の調度、衣装、楽舞の曲目等といった事前に定めおく事柄があり、当日の左右方人の集合から始まり、講師・読師・判者が召され、歌の披講があり、評定があり、楽舞の後、禄を賜り、終わる。歌合日記はこれら行事の進行に概ね沿った形で書かれ、起こった事柄[3]も併せて記録される。

本発表では、この『歌合コーパス』のうち、「天喜四年四月三十日皇后宮寛子春秋歌合」（通称「四条宮春秋歌合」、以下「春秋歌合」とする）を調査対象とする。本文は、日本古典文学大系 74『歌合集』（岩波書店）を使用した。

「春秋歌合」は天喜四（1056）年に催された歌合で、後冷泉皇后寛子が主催者であった。寛子は関白頼通の女である。天皇も密かに臨御され、頼通が後見し、盛大な歌合となった。左を春、右を秋とし、和歌のみならず、書芸・絵画・音楽・工芸・服飾を通じて春秋を競う歌合であった。この「春秋歌合」を対象としたのは、2 種類の日記が付されていることによる。2 種類とは、漢文日記[4]と仮名日記[5]である。漢文日記は記録体、仮名日記は和文体で書かれており、ほぼ同じ内容を 2 種類の文体で読み比べることができる貴重な資料と言える。ただし、「春秋歌合」の仮名日記は行事の進行上、漢文日記の半分弱のところから先が散逸している。また、歌合の行事的要素のうち、どの部分をどの程度記述するかについては差が見られ、単純な文体違いの一対の文章ではない。

漢文日記については読み下したテキストを対象に形態論情報を付与した。歌合日記には割書箇所が多い。割書は「題目〈左春／右秋〉」（〈〉が割書箇所）のような語に対する注記もあるが、詳細を文で記すものも多いため、これも形態論情報付与の対象とした。また、仮名日記については、大系のテキストに従い補読した箇所がある。例えば、「春(はるの)山べ」とあるところは「春<u>の</u>山べ」とした。

---

1 『歌合コーパス』に付した情報については、冨士池（2014a）（2014c）を参照方。

2 文台は歌を載せる台、員差は勝点計算の道具。

3 今回調査となった漢文日記には、「祝歌の左方に御製があり、是非を述べずに左の勝とした」「右方が負けたのに燈台を設けるのを忘れたので罰酒あるべしと、判者である内大臣がふざけて言った」などのエピソードが含まれている。

4 殿上日記とも言う。「春秋歌合」の漢文日記は蔵人によるもの。

5 仮名日記は甲乙の 2 種類があったとされるが、現存するのは甲の一部であり、伊勢大輔の手によるものとも言われている。

（２）日本語歴史コーパス　平安時代編

2014 年 3 月、公開された『日本語歴史コーパス　平安時代編』には、中古和文 14 作品（竹取物語、古今和歌集、伊勢物語、土佐日記、大和物語、平中物語、落窪物語、枕草子、源氏物語、紫式部日記、和泉式部日記、更級日記、堤中納言物語、讃岐典侍日記）が収録されている。このコーパスには「本文種別」として「会話」「手紙」「歌」「詞書」といった情報が付与されている。これを、歌合日記との比較対象資料として用いた。

２．２　言語単位

『日本語歴史コーパス　平安時代編』の言語単位は、『現代日本語書き言葉均衡コーパス』で採用した単位を中古和文用に修正・拡張したものであり、『歌合コーパス』の言語単位も共通の仕様とした[6]。採用した言語単位は、「短単位」、「長単位」の 2 種類であるが、このうち、構文的側面に着目して規定された「長単位」を用いた。長単位は文節を自立語と付属語に分割した言語単位である。合成語を認めており、結合回数の制限はないため、「二重織物」「思ひやる」「渡らせ給ふ」「藤少納言伊房」といった語や、接辞を含めた形式が 1 長単位となる。文脈に即して品詞を付与する方針をとっており、同じ語に対して異なる品詞を与えることがある。例えば「哀れ」の場合、「もののあはれ知りすぐし」は名詞を、「いとあはれなる句」は形状詞を付与するといった判別を行う。図 1 に長単位例を示す。

キー	語彙素	語彙素読み	品詞	活用型	活用形
去る	往ぬ	イヌ	動詞-一般	文語ナ行変格	連体形-一般
閏三月	閏三月	ウルウサンガツ	名詞-数詞		
の	の	ノ	助詞-格助詞		
比	頃	コロ	名詞-普通名詞-一般		
、	、		補助記号-読点		
恪勤	恪勤	カクゴン	名詞-普通名詞-一般		
の	の	ノ	助詞-格助詞		
女房等	女房等	ニョウボウラ	名詞-普通名詞-一般		
相議つ	相諮る	アイハカル	動詞-一般	文語四段-ラ行	連用形-促音便
て	て	テ	助詞-接続助詞		
各々	各々	オノオノ	副詞		
方人	方人	カタヒト	名詞-普通名詞-一般		
を	を	ヲ	助詞-格助詞		
取り分く	取り分く	トリワク	動詞-一般	文語四段-カ行	終止形-一般
。	。		補助記号-句点		

図１　長単位例

6 単位の概要については、コーパス検索アプリケーション「中納言」オンラインマニュアルの CHJ＞形態論情報の概要 を参照。

３．調査結果

３．１　品詞比率

樺島・寿岳（1965）は、自立語について品詞をその機能によって体（名詞）・用（動詞）・相（形容詞・形容動詞・副詞・連体詞）・他（接続詞・感動詞）の四つに分類した。この 4 分類に基づき、「春秋歌合」日記の品詞比率を図 2 に示す。『日本語歴史コーパス　平安時代編』の品詞体系では、体の類に「名詞-普通名詞-一般」「名詞-固有名詞-｛一般・人名・地名｝」「名詞-数詞」「代名詞」が、用の類に「動詞-一般」が、相の類に「形容詞-一般」「形状詞-｛一般・タリ｝」「副詞」「連体詞」が、他の類に「接続詞」「感動詞-一般」が分類される。

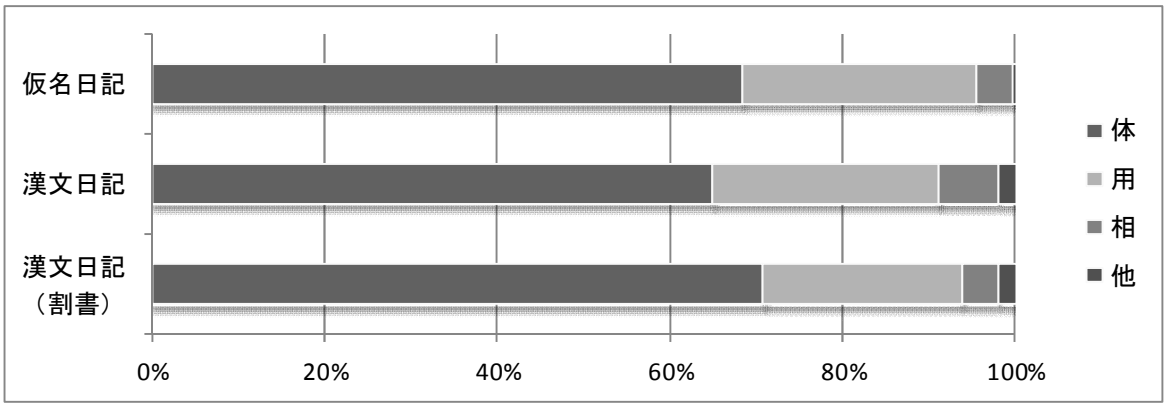


図 2　「春秋歌合」日記の品詞比率（延べ語数）

体の類の割合は漢文日記（割書）、仮名日記、漢文日記の順で高くなっている。用の類の割合は漢文日記（割書）がやや低く、相の類の割合は漢文日記がやや高い。漢文日記、漢文日記（割書）に見られる他の類は、漢文訓読によく見られる「或いは」「但し」といった接続詞である。また、表 1 に示したように、相の類の内訳が大きく異なり、仮名日記では形容詞主体、漢文日記では地の文、割書ともに副詞主体となっており、文体差が見られる。

表 1　「春秋歌合」日記における相の類の内訳（粗頻度）

品詞	仮名日記	漢文日記	漢文日記（割書）
形容詞	14	4	7
形状詞	2	2	2
副詞	5	16	12

３．２　名詞率と MVR

本発表では、品詞比率に基づきテキストの特徴を示す指標として、名詞率と MVR を用いる。名詞の比率は文章の特質を表し、名詞の比率に応じて他の品詞もある傾向を持って変化する、つまり文章のジャンルによって品詞の割合が決定されると考えられる。ここでは

延べ語数を用いて、品詞比率を求める。樺島・寿岳（1965）は、自立語について品詞をその機能によって体（名詞）・用（動詞）・相（形容詞・形容動詞・副詞・連体詞）・他（接続詞・感動詞）の四つに分類したとき、体の類と、用・相それぞれの類の関係を見るにあたり、MVR という「100×相の類の比率／用の類の比率」の式で表される指標を提案し、名詞率と MVR の組み合わせから見出せる文体的特徴として、名詞率が高く MVR が小さいものを「要約的な文章」、名詞率が低く MVR が大きいものを「ありさま描写的な文章」、名詞率が低く MVR も小さいものを「動き描写的な文章」と位置づけた。

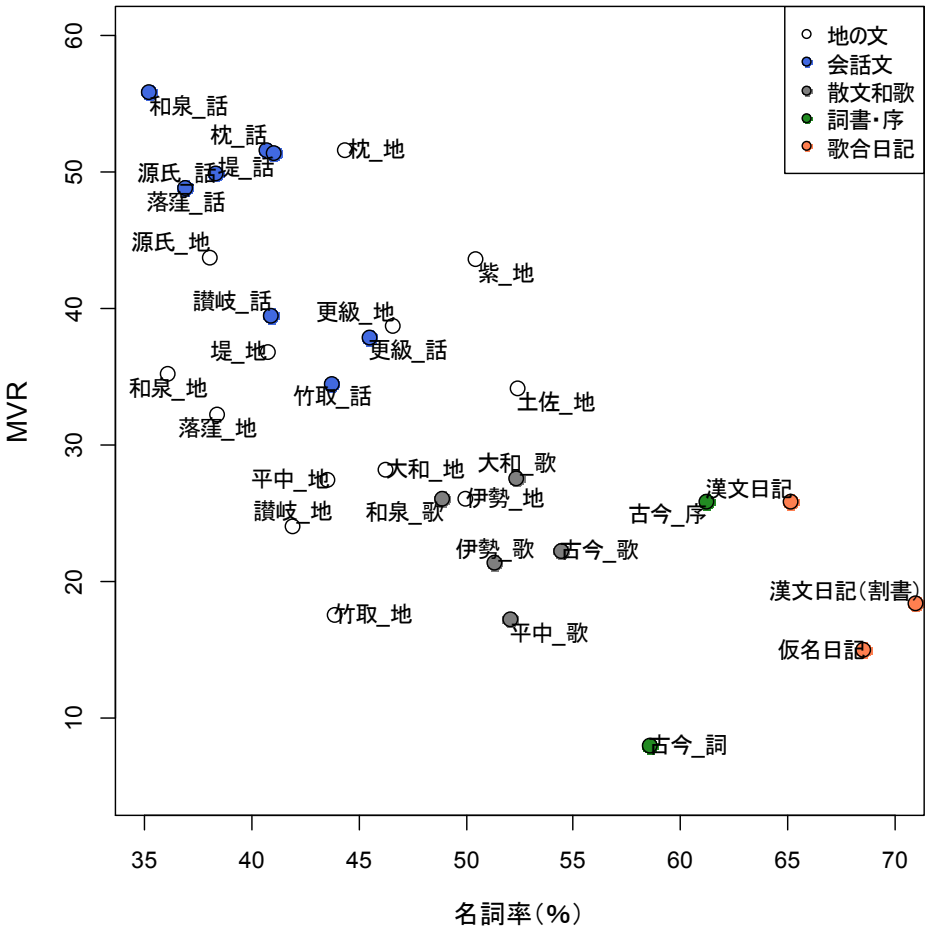


図 3　「春秋歌合」日記と中古和文 14 作品の名詞率・MVR

「春秋歌合」日記の品詞比率を中古和文の品詞比率と比較するとどのような位置付けになるのだろうか。冨士池（2014b）では『日本語歴史コーパス　平安時代編』に基づく中古和文 14 作品の名詞率と MVR[7]を示した。今回の調査結果[8]に、中古和文 14 作品の名詞率・

[7] 『古今和歌集』は歌・詞書・仮名序に、他の 13 作品は地の文・会話文・歌に分けて集計し、各作品の延べ語数の 20%以上を占める場合のみを示したもの。

[8] 図 3 では歌合日記を地の文としているが、漢文日記には 8 長単位、仮名日記には 7 長単位の会話を含む（自立語の長単位数）。会話文が 1 カ所ずつのみであったため、今回は地の文から除外しなかった。

MVR を重ね合わせた散布図を図 3 に示す。

冨士池（2014b）では「要約的な文章」として、物語・日記所収の和歌と『古今和歌集』詞書・仮名序を挙げた。しかし、図 3 から、「春秋歌合」日記の方がより名詞率が高く MVR が小さい「要約的な文章」としての特徴が強いことが明らかになった。ここから、歌合日記が物語・日記・随筆の地の文とは異なるジャンルの文章であることが見てとれる。中でも、名詞率の高さが特徴的である。漢文日記と仮名日記という文体の違いについては、名詞率より MVR、つまり相の類（形容詞・形状詞・副詞）と用の類（動詞）のバランスに現れている。

３．３　名詞率と文の長さ

「春秋歌合」の日記について、名詞率と MVR を見た結果、MVR は中古の和歌や、地の文の中で MVR が低めの資料と同程度であったが、名詞率の高さが特徴的であることが明らかになった。

文章における名詞の比率が増加する要因として、樺島（1979）では「ある内容を、限られた言葉数で述べようとするときには、凝縮化 要約化の二つが働く」とする。**凝縮**とは、意味的に重複する部分をくりこんで言葉数を減らすというもので、結果として「文の構造が複雑で」「文の長さが長い」という性格を持つとする。それに対し、**要約**は限られた言葉数の中で意味内容を表すもので、要約化が働いた文章の例として、新聞の見出し、辞典、短歌・俳句、出版目録解説、映画解説パンフレット、新聞のラジオ・テレビ案内を挙げる。

「春秋歌合」日記の名詞率の高さは、樺島（1979）に示された二つの要因で説明できるのだろうか。中古散文作品中の和歌は、現代の短歌・俳句同様に音数の制約があるために「要約」によって名詞の比率が増加していると考えられる。それに対し、『古今和歌集』仮名序・詞書や、今回の調査対象である歌合日記は語数（音数）の制限はない。「春秋歌合」日記の名詞率の高さが要約によるものでないのならば、凝縮によるものなのだろうか。凝縮に関しては、限られた言葉数という制約がなくても文章の一つのスタイルとして起こり得る現象と考える。そこで、コーパスに付与した情報のうち文境界情報[9]を利用して、「春秋歌合」日記の文の長さと名詞率との関係について、検討する。

表 2 に、「春秋歌合」日記における 1 文あたりの自立語数と名詞率を示した。1 文あたりの自立語数が文の長さを意味する。「春秋歌合」日記の仮名日記、漢文日記、漢文日記（割書）のほか、比較する材料として中古和文のうち名詞率が低いものから『源氏物語』桐壺巻の地の文を、名詞率が中程度のものから枕草子（冒頭 3 章段[10]）の地の文を、名詞率が高いものから『古今和歌集』仮名序の地の文[11]を示した。

---

9 「日本語歴史コーパス」「歌合コーパス」とも、単位ごとに文頭かそうではないかという文境界情報が付与されている。ただし、コーパス検索アプリケーション「中納言」ではこの情報は公開されていない。

10 「春はあけぼの」「ころは」「正月一日は」

11 歌、古注、古注（歌）、古注（詞書）を除いたものを地の文とした。

表 2　「春秋歌合」日記における 1 文あたりの自立語数（長単位）と名詞率

	自立語数	文の数	1 文あたりの自立語数	名詞率(%)
仮名日記	514	53	9.698	68.5
漢文日記	324	51	6.353	65.1
漢文日記（割書）	496	82	6.049	71.0
源氏物語（地の文）	2009	169	11.888	38.0
枕草子（地の文）	552	49	11.265	44.4
古今和歌集（仮名序）	805	69	11.667	61.2

表 2 から、1 文あたりの自立語数が少ない、つまり文が短いほど名詞率が高くなる様子が見てとれる。漢文日記（割書）が最も文が短いという結果になったが、1 長単位から成る文は 3、2 長単位から成る文が 1 あるほかは、極端に短い文はなかった。

1 文あたりの自立語数と名詞率の相関係数は-0.718 と負の相関が見られた。これは、樺島（1979）で指摘された、凝縮化された文章は文が長く名詞の比率が高いということと相反する結果となった。樺島（1979）では現代書き言葉を対象としているのに対し、今回の調査は平安時代の書き言葉を対象としている。今回の調査結果から、少なくとも平安時代の文章については、名詞率が増加する要因として、要約と凝縮、つまり語数（音数）の制約や文の長さ以外の第 3 の要因が存在することが示唆された。

「春秋歌合」日記の中で文体の違いについて見ると、漢文日記と仮名日記の差が大きく、漢文日記は文が短いことがわかる。また、物語・随筆の地の文や和歌集序文と「春秋歌合」仮名日記は同じ和文体であっても差があり、仮名日記の方が文が短い様子が見てとれる。これは「春秋歌合」仮名日記の特徴なのか、歌合の仮名日記に通じる文体的特徴なのか、調査対象を広げて確認する必要がある。

４．おわりに

本発表では『日本語歴史コーパス　平安時代編』『歌合コーパス』の「長単位」データを用い、品詞比率に基づきテキストの特徴を示す指標として名詞率と MVR を算出した。その結果、中古歌合日記の 1 資料である「春秋歌合」日記は、他の中古和文資料と比較して、名詞率が高く MVR が低い「要約的な文章」であることが明らかになった。また、特に名詞率の高さが特徴的であったことから、名詞率と文の長さの関係について検討したところ、文が短いほど名詞の比率が高いという、これまでの指摘とは異なる結果となった。ここから平安時代の文章については、名詞率が増加する要因として、語数（音数）の制約や文の長さ以外の第 3 の要因が存在することが示唆された。この要因の究明は今後の課題となる。「春秋歌合」日記という、漢文日記と仮名日記が揃った資料を対象としたにも関わらず、具体な描写の違いといったところまでは考察が及ばなかった。具体な描写の違いを検討し

ていく中で、第 3 の要因についても考えていきたい。

今回の調査対象は「春秋歌合」日記のみであることから、歌合日記の特徴と言い切れないところがあり、これも今後の課題となる。歌合日記は全ての歌合にあるものではなく、「春秋歌合」のように漢文日記と仮名日記が揃っているものは数少ないといった資料の制約はある。一方で、歌合日記のような行事の記録は説明的な文章であり、物語・日記といった創作とは異なる文章のジャンルである。説明的な文章の資料性についてはまだ検討が不十分な点が多く、引き続き検討していきたい。

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」、JSPS 科研費「中古中世歌合コーパスに基づく和歌評論の語彙論的研究」（研究課題番号：25770179）の成果の一部である。

文　献

樺島忠夫・寿岳章子（1965）『文体の科学』（綜芸舎）

樺島忠夫（1979）『日本語のスタイルブック』（大修館書店）

萩谷朴・谷山茂　校注・訳（1965）日本古典文学大系 74『歌合集』（岩波書店）

冨士池優美 (2014a)「中古中世歌合の構造化」『言語処理学会第 20 回年次大会発表論文集』、pp.205-208

冨士池優美（2014b）「品詞比率からみる中古和文テキストの特徴」『日本語学会 2014 年度春季大会予稿集』、pp.185-190

冨士池優美（2014c）「平安初期歌合の品詞比率」『第 6 回コーパス日本語学ワークショップ予稿集』、pp.21-30

関連 URL

日本語歴史コーパス　http://www.ninjal.ac.jp/corpus_center/chj/

コーパス検索アプリケーション「中納言」オンラインマニュアル https://maro.ninjal.ac.jp/wiki/

BCCWJ に拠る名詞別格外連体修飾形の形成傾向の分析

田邊　和子（日本女子大学文学部）†

Analysis of Japanese Noun's Inclination to Form Case-Outer Relative Clauses Based on the BCCWJ

Kazuko Tanabe (Japan Women's University)

要旨

本研究は、BCCWJ 調査に基づいた「連体修飾節を形成しやすい普通名詞の順位」に従って、名詞別に格内（内の関係）及び格外（外の関係）連体修飾形成率や修飾節の動詞の「ル形」・「タ形」別の比率を調査したものである。連体修飾形成率の頻度の高い名詞の中で、たとえば、有生名詞（animate noun）の「人」は、格内連体修飾節の主格が全体の 90%以上であり、それとは対照的に「場合」では、ほとんどが格外連体修飾節となり、時の指定の副詞節に近い役割を成す。「必要」においては、格外連体修飾で動詞「ル形」がほとんどである。「問題」は、その中間に位置し、格内連体修飾と格外連体修飾は、ほぼ半数ずつであった。さらに動詞の「ル形」（動詞連体形）が、「タ形」使用の 3 倍以上であった。このように個々の名詞の意味が、格内・格外の使用傾向、さらに格内の場合はその使用する格、格外の場合は動詞の「ル形」・「タ形」の選択に影響を与えることが明らかになった。

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でのポスター発表「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」（田辺 2014）を発展させ、格外連体修飾形のうち、共起する動詞の「ル形」と「タ形」の対比を中心に主名詞の意味と動詞の文法形式の関係について分析を試みた。

連体節の構造について確認すると、宮地（2005）は、「連体節の主名詞（底の名詞（寺村 1992））が、連体修飾節内部の用言の補語として関係を持つ「同一名詞体」（いわゆる内の関係（寺村 1992））と、そのような関係がない「付加連体」（外の関係（寺村 1992））があり、付加連体は、さらに「同格連体」と「相対連体」に整理されている（奥津 1974）。」としている。本稿での「格外連体修飾形」とは宮地の分類では「同格連体」を示す。一般的に「形式名詞」と呼ばれる「モダリティの助動詞用法」（宮地 2005）を持つ「こと」「もの」などは、本研究の対象とはしない。

言語類型論者の Comrie（1998）は、「学生が本を買った事実」という日本語の例文を挙げ、‘the fact that the student bought the book’ という英訳とともにアジア言語特有の限定修飾節として、fact-S construction という名でこの格外連体修飾節構造を紹介している。

本研究では、BCCWJ の検索結果から、「連体修飾節を形成しやすい普通名詞順位表」を作成し、その中から比較的順位の高い「人」「場合」「問題」を、また動詞の「ル形」と

† tanabeka@fc.jwu.ac.jp

「タ形」対立を論ずる材料として「事件」「動機」を取り上げ、それぞれの名詞の連体修飾節内の接続形式の特徴を明らかにしたい。そして、その結果を踏まえて、接続形式を決定付ける名詞の意味基準を提示したい。

2. 連体修飾節を形成しやすい普通名詞の順位表

下の表は BCCWJ コアデータから中納言で、①普通名詞に動詞連体形が前方共起している用例、②普通名詞に助動詞の連体形が前方共起している用例、③②の中で助動詞を「た」に特定し、その前に動詞が前方共起している用例を検索し、①から③の名詞別用例数とその割合を示したものである。（表は、①の用例で用例数の多い名詞順に並べられている。①の用例総数は 18,539、②の用例総数は 17,654、③の用例総数は 7,467）

表１　連体修飾節を形成しやすい名詞順位表

		① 動詞連体形		② 助動詞連体形		③動詞＋「た」	
	名詞	用例数	割合	用例数	割合	用例数	割合
1	こと	3528	19.03%	1564	8.86%	691	9.25%
2	ため	1058	5.71%	222	1.26%	71	0.95%
3	もの	564	3.04%	791	4.48%	341	4.57%
4	人	474	2.56%	374	2.12%	190	2.54%
5	わけ	246	1.33%	106	0.60%	54	0.72%
6	必要	190	1.02%	9	0.05%	0	0.00%
7	場合	186	1.00%	220	1.25%	115	1.54%
8	とき	177	0.95%	249	1.41%	199	2.67%
9	ところ	164	0.88%	225	1.27%	143	1.92%
10	はず	121	0.65%	61	0.35%	30	0.40%
11	事	120	0.65%	88	0.50%	31	0.42%
12	時	112	0.60%	164	0.93%	132	1.77%
13	者	107	0.58%	68	0.39%	41	0.55%
14	情報	87	0.47%	82	0.46%	21	0.28%
15	方	81	0.44%	129	0.73%	78	1.04%
16	つもり	79	0.43%	14	0.08%	10	0.13%
17	ほか	75	0.40%	45	0.25%	35	0.47%
18	一方	72	0.39%	11	0.06%	4	0.05%
19	うち	68	0.37%	25	0.14%	0	0.00%
20	前	67	0.36%	13	0.07%	5	0.07%
21	予定	63	0.34%	21	0.12%	1	0.01%
22	意味	63	0.34%	16	0.09%	8	0.11%
23	点	61	0.33%	38	0.22%	11	0.15%
24	中	60	0.32%	33	0.19%	21	0.28%
25	方法	59	0.32%	27	0.15%	4	0.05%
26	地域	59	0.32%	35	0.20%	17	0.23%

27	言葉	58	0.31%	38	0.22%	16	0.21%
28	理由	56	0.30%	50	0.28%	21	0.28%
29	方針	55	0.30%	6	0.03%	1	0.01%
30	調査	55	0.30%	20	0.11%	10	0.13%
31	際	55	0.30%	37	0.21%	34	0.46%
32	企業	49	0.26%	40	0.23%	23	0.31%
33	問題	48	0.26%	57	0.32%	13	0.17%
34	話	46	0.25%	52	0.29%	11	0.15%
35	声	46	0.25%	24	0.14%	9	0.12%
36	女性	45	0.24%	27	0.15%	13	0.17%
37	限り	45	0.24%	26	0.15%	1	0.01%
38	形	45	0.24%	59	0.33%	20	0.27%
39	気	45	0.24%	93	0.53%	7	0.09%
40	間	45	0.24%	8	0.05%	1	0.01%

【検索式】

① 動詞連体形＋名詞

キー: 品詞 LIKE "名詞-普通名詞%" AND 前方共起:(品詞 LIKE "動詞%" AND 活用形 LIKE "連体形%") ON 1 WORDS FROM キー DISPLAY WITH KEY IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

② 助動詞連体形＋名詞

キー: 品詞 LIKE "名詞-普通名詞%" AND 前方共起:(品詞 LIKE "助動詞%" AND 活用形 LIKE "連体形%") ON 1 WORDS FROM キー DISPLAY WITH KEY IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

③ 動詞＋「た」＋名詞

キー: 品詞 LIKE "名詞-普通名詞%" AND 前方共起:(品詞 LIKE "助動詞%" AND 活用形 LIKE "連体形%" AND 語彙素 = "た") ON 1 WORDS FROM キー DISPLAY WITH KEY AND 前方共起: 品詞 LIKE "動詞%" ON 2 WORDS FROM キー DISPLAY WITH KEY IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="1" AND tglWords="20" AND limitToSelfSentence="1" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-8" AND tglFixVariable="2"

表１では、格内連体修飾形と格外連体修飾形の合計数が示されている。本研究では、コアデータの例文全体を見て格内・格外を判別し、その考察に基づいて、分析を行うことにする。上位 3 位「こと」「ため」「もの」は、形式名詞としての用法と格内連体修飾形との合計数である。本稿では、第 4 位の「人」、第 7 位の「場合」、第 33 位の「問題」を、格内・格外の比率、格内使用において利用した格の種類、共起しやすい動詞・助動詞の文法形式において考察する。また、「事件」と「動機」については、動詞「ル形」と「タ形」の使用においてそれぞれ特徴的であることから、分析例として取りあげることにする。

3.動詞及び助動詞＋「人」「場合」「問題」の用例

3.1　「人」について

3.1.1　動詞＋「人」の用例

Left	Center	Right
になります。 家族の一員として犬やネコと接している	人	なら、その感覚がおわかりになると思います。それを
やチャットでも、ただの遊びで面白半分参加している	人	もいれば真面目に出会いを求めて参加している人も
り、経済的な理由からやむを得ず親と同居している	人	が最も多い（第3-2-8図、付表3-2-7）。
施しています。 また，日本語を第二言語として学ぶ	人	の中でも特に成長期にある子どもたちが，日本語や
また、いまサーフィンが人気を呼び、波乗りを楽しむ	人	たちが全国から集まり、三百人ほど定住しているとい
もいる。その一方で、当然のこととして裁判を起こす	人	もまた多い。 貴子さんは訴えられた側だから、私は
先生の授業は、型破りだった。障害や難病に苦しむ	人	の話をよく取り上げ、生徒同士で討論させた。「世の
りを心掛けている。 毎月、その月に誕生日を迎える	人	を祝う「誕生会」を実施。クリスマスパーティーやひな
彼は殉教者になれなかった」と複雑な表情を見せる	人	も。独裁者の末路に対する感慨は様々だが、市民に

図 1　動詞＋「人」

3.1.2　助動詞＋「人」の用例

Left	Center	Right
示は適法　談合疑惑を追及する住民訴訟を起こした	人	に、公正取引委員会が審判に証拠提出した事件記
に関しては、「ほぼ全面的に公的年金に頼る」とした	人	が二十九%で、千九百九十八年の前回調査より7.
るのだろうか、どこからきたのだろうか、乗っていた	人	たちはどうしたのだろうか。余裕が出てきた私に、今
くれたスタッフには、子どもを亡くしたり、家を失った	人	もいる。そんな中で手作業の復旧に全力を傾けてくれ
への準備をすること，民間企業が学校で排除された	人	達にもっと参加するように働きかけること，人々が生
はあがりっ放しだ。去年俺たちのツアーに来てくれた	人	たちも、がっかりさせることのない強力な内容で今年
し付け加えておきたいのは、死後の世界を信じない	人	ではなく、神仏を信じない人の場合についてです。
想に」まだ、子育ての本格的な苦労（笑）を知らない	人	が話題にする言葉。言った人がおかしいの！気にし
は、経済的理由から結婚しない、あるいは、できない	人	の割合が高くなっている。今後、デフレの下で経済の
した作品を地で行くような運命をたどることになった	人	だ。『裏窓』のリザという役は、いわゆる才色兼備の

図 2　助動詞＋「人」

「人」においては、動詞/助動詞共起ともに圧倒的に格内連体修飾形の用例が多く、しかも、修飾節内で主格となる用例が 90%以上である。格外連体修飾を対象としようしたが、その用例が「人」については、ほとんどなかった。また、「私が昨日見た人」というような対格使用も可能性としては有り得るが、実際の使用状況をコーパスで見ると主格使用がほとんどであった。これは、「人」が有生名詞（animate noun）であることに起因すると推察する。動詞連体形使用数は、474 であり、それに対し「タ形」使用数は、190 である。割合は、ほぼ同じで、特にどちらかに大きな偏りはない。

共起する助動詞の種類を考察すると、テンス（例：「起こした（人）」）・アスペクト表現（例：「乗っていた（人）」）・ヴォイス表現（例：「助けられる（人）」）ともに制限は見受けられず、否定表現も含まれる。

3.2「場合」について

3.2.1　動詞＋「場合」の用例

Left	Center	Right
行動1専門の会社に頼む　予算や納期に余裕がある	場合	は、この方法がよいだろう。ただし、リピートがあるた
との生活はどれに近いですか。結婚していらっしゃる	場合	は配偶者の親を含めてお答えください。（○は1つ）」
集させています。同時間帯に重なって放送されている	場合	は、両方ともちゃんと録画して目を通しておられる。I
）の関西版のこと。一人ないし数人で商売をおこなう	場合	に用いる商法だ。このとき、外部の力をいかに働かせ
ラムの再精錬のようにエネルギー消費量が減少する	場合	に大きな効果が見込まれます。ほかには廃プラスチ
ビル2階）までお送りください！！　画像データで送る	場合	はE-mailに添付してspur@skichannel. ne. jpま

図３　動詞＋「場合」

3.2.2　助動詞＋「場合」の用例

Left	Center	Right
努めるとともに、仮に海外でトラブルに巻き込まれた	場合	には、留守家族等に安否を至急連絡することなどの
の年齢別出生率で1人の女性が子どもを産むとした	場合	の平均子ども数を表すものであるが、同一出生年集
の2つの型がある。要介護など所定の状態になった	場合	には保険料の払い込みが免除される。 五十歳女性
機能が追加され、また、トラヒックが大幅に変動した	場合	には事業者間で精算を行うこととされた。 さらに、近
に時間外労働を月平均八十時間を超えて行わせた	場合	について、それぞれ具体的な措置を示している。（注
成績に基づく学校評価を重視し、改善が見られない	場合	には生徒が転校できるようにした。宗教団体による有
を切る場合もあるだろうし、条件闘争をして切らない	場合	もあるだろう」と、小島朋之・慶応大総合政策学部長
で設定ができます。もしBIOSがおわかりにならない	場合	には、残念ながら知識のある方に聞くか、メーカーの
で相当数の空家があり、将来とも需要が見込めない	場合	にあっては、当該空家部分を積極的に活用するため
支援本部と地域レベルの本部とが同時に設置される	場合	が多い。 平成十二年9月の東海豪雨災害の際には
病院として誕生した。だが、老人を病院に入院させる	場合	、患者の家族は“捨てた”という後ろめたさを感じがち

図４　助動詞＋「場合」

「場合」については、「とき」「ところ」などと類似して、格外連体修飾節の被修飾名詞というより節や句を導く副詞としての役割が大きいようである。しかし、本稿では「場合」は、「に」を伴って副詞節を導くとし、「場合」単独では格外連体修飾節として扱いたい。共起する文法形式としては、動詞「ル形」（例：「おこなう」「減少する」）、アスペクト表現（例：「放送されている」）が用いられることはもとより、助動詞でも、「タ形」（例：「状態になった」）、「受け身」（例：「設置される」「巻き込まれた」）、「使役」（例：「入院させる」）など多岐にわたる。

3.3 「問題」

3.3.1 動詞＋「問題」の用例

「問題」については、格内/格外両方に大きな偏りなく使われるので本項では、動詞・助動詞共に格内/格外別の用例をまとめて図を作成した。

3.3.1.A　動詞＋「問題」: 格内連体修飾形の用例

Left	Center	Right
ト問題や台湾との関係を含め、中国が直面している	問題	は少なくない。五輪はいやおうなく、巻き込まれてい
少年院教官調査の結果，特に，困難度が増している	問題	として，少年の資質の問題のほか，親の指導力及び
1図のとおりである。 最近，非行少年の抱えている	問題	の中身が「変化した」(「かなり変化したと思う」及び「
る。「靖国神社参拝をやめたからといって解決する	問題	ではない。教科書、尖閣諸島、東シナ海のガス田問
展を遂げた。一方で、政治体制の脆弱さから生ずる	問題	や、グローバル化の進展に伴う経済格差の拡大が顕

図５　動詞＋「問題」: 格内連体修飾形

3.3.1.B　動詞＋「問題」: 格外連体修飾形の用例

Left	Center	Right
をめざす米朝中三カ国協議に、日本や韓国を加える	問題	について「当事者間の合意があれば、柔軟な姿勢だ
合わせ、最後の1球がセンターで終わることが出来る	問題	を「詰めソリテア」としています。2つ目は「詰めタコ」(
域への努力が注目されています。地球環境に関する	問題	は、私たちの日常生活から改善すべきこと。多くの資

図６　動詞＋「問題」: 格外連体修飾形

3.3.2　助動詞＋「問題」の用例
3.3.2.A　助動詞＋「問題」: 格内連体修飾形の用例

Left	Center	Right
中野田）の芝の根付き状態に不安が指摘されていた	問題	で、日本サッカー協会の高田豊治施設委員長は十分
において、みんなで本番の紙に、みんなが考えてきた	問題	をまとめる。2かっこいいタイトルやおもしろいタイトル
代社会においては、前提条件が明確な、与えられた	問題	を解けるばかりではなく、革新すべき課題を明らかに
書の提出だけで終わらせ、事件は時効になっていた	問題	で、道警釧本監察官室は十一日、「当時の根室署の
を越えて行われており、一国のみでは解決できない	問題	であることから、サミット、国際連合等の国際的な枠
子どもは、実際の事象の説明を試みた。学習すべき	問題	は「ロウの状態変化を観察し、アトムくんでこれを説明
全を確保する必要がある。違法駐車など解決すべき	問題	は多いが、電動自転車など低速のものが安全に走れ
います。こういった問題は医者が勝手に決めるべき	問題	ではないからです。 あらかじめこういったことを話せ

図７　助動詞＋「問題」: 格内連体修飾形

3.3.2.B 助動詞＋「問題」: 格外連体修飾形の用例

Left	Center	Right
この勤務実態を偽って介護報酬を不正受給していた	問題	で、道は十五日までに、施設の短期入所療養介護と
馬全国協会の幹部が馬券を買ったとして処分された	問題	で、警視庁は9日、■■■■元常務理事（五十四）＝
ッド・マジック」で府の許可量以上の火薬を使用した	問題	で、府警保安一課と此花署は八日、火薬類取締法違
計事務所（千葉県市川市）の構造計算書が使われた	問題	で、国土交通省は二十一日、既に完成した十四棟の

図８　助動詞＋「問題」: 格外連体修飾形

「問題」は、動詞・助動詞両方の共起例を考察すると、格内修飾・格外修飾形共に大きな偏りなく両方の形式で使用される。さらに、格内使用においても「生ずる問題」（主格）、「解決する問題」（対格）、「困難度が増している問題」（所有格）というようにさまざまな格において使われている。アスペクト表現（例：「直面している」「不正受給していた」）、ヴォイス表現の受け身（例：「処分された」）もみられる。また、「～べき」との共起例が複数考察できるのも「問題」の特徴である。

4. 格外連体修飾節内の「ル形」と「タ形」の使い分けについて

丹波（2013：267）の指摘によると、「外の関係においては、基本形の持つ性格が、主名詞の性格によって制約を受けることが大きい」と述べている。岩崎（1998：33）は、従属節内のテンスの認定の問題を取り上げ、格外連体修飾節内で①「ル形」と「タ形」にテンスの対立があり表現する状況に明確な違いがある語（例：話・可能性・恐れ・噂）と、②「ル形」と「タ形」に明確なテンスの差がなく状況の違いとは結びつかない語（例：騒ぎ・事件・動機・事態）があるとしている。

①の例：地震が起きる話　　②の例：警官まで出動する騒ぎ
　　地震が起きた話　　警官まで出動した騒ぎ（丹波（2013：273））

本項では、②のグループの例から「事件」と「動機」を取り上げ、「ル形」と「タ形」の使い分けの実態を考察してみたい。

4.1 「事件」の用例

4.1.1 動詞（「ル形」）＋「事件」の用例

Left	Center	Right
融機関に百口座以上を開設し、2万～3万円で転売する	事件	も起きている。 最近では、信販会社を装って偽メール
十二年後の昭和三十六年三月、ついに五人が脱退する	事件	が起こった。 そして、その事件とともに、国語審議会の
空機が旧ソ連の携帯型地対空ミサイルの攻撃を受ける	事件	が発生した。携帯型ミサイルは安価で1人でも取り扱え
時のルール大人の目の届く場所に 子供が被害に遭う	事件	が相次いでいる。親の不安は募るばかりだが、かといっ

図９　動詞（「ル形」）＋「事件」

4.1.2 動詞＋「た」（「タ形」）＋「事件」の用例

Left	Center	Right
保護した男性（当時五十歳）を放置し、男性が死亡した	事件	で、虚偽の書類作成を指示したとして虚偽有印公文書
組長を殺人罪で逮捕した（埼玉）。（2）けん銃を使用した	事件	の発生状況 平成十五年中のけん銃（けん銃様のもの
で暴力団幹部ら男性二人が特殊警棒で殴られ死傷した	事件	で、府警捜査四課と寝屋川署は六日、殺人などの疑い
に。 毎日新聞記者の所持品が爆発して6人が死傷した	事件	は、戦場取材での軽率な振る舞いが他人の命を奪った
起こした人に、公正取引委員会が審判に証拠提出した	事件	記録を開示できるかどうかが争われた訴訟の上告審判
ん銃らしきものを突き付け、見せるなどして犯行に及んだ	事件	において、被害者、参考人等の供述等により、けん銃と
人が認められていない。「毒物中毒の患者が多数出た	事件	で、フォローアップをどうするかは大きな課題だ」 平成十

図 10　動詞＋「た」（「タ形」）＋「事件」

以下の表は、上の図９及び図 10　の内容を一つにまとめたものである。

表 2　「事件」における連体修飾節内の動詞「ル形」と「タ形」対立について

「ル形」	外	6	2～3 万で転売する事件も	起きている。
			5 人が脱退する事件が	起こった。
			ミサイルの攻撃を受ける事件が	発生した。
			子供が被害に遭う事件が	相次いでいる。
	内	3	裁定に係る事件を	職権で調停に付し
「タ形」	外	17	男性が死亡した事件で、	虚偽の書類作成
			殴られ死傷した事件で、	府警捜査四課と

	内	3	少年が起こした事件についても、	警察が捜査に準じ
			奈良県で起きた事件では、	「警察官役」の男
			父が扱った事件から、	大物プレイボーイ

「事件」という言葉を使う場合は、「事件」として認められる出来事が既に起こった後に使うことがほとんどであるから、格内・格外に関わらず、正確なテンス描写としては「タ形」であることが多いことは予想できる。これは、数値的にも「タ形」が多いことから推測できる。しかし、連体修飾節の直後あるいは、比較的近くで文が終了する場合は、その文末表現で、過去・完了時制が明確に提示される。このような時には、連体修飾節内では「ル形」が使われる傾向が窺える。これは、おそらく、時制については、主文で明示されるので従属節でいちいち表す必要もなく、内容が説明されていればいいという比較的緩慢な決定が格外連体修飾節内ではなされうる可能性があることと、それを後押しする要素として、音調的に「タ形」の重複を避けるためとも推察できる。

4.2「動機」の用例
4.2.1 動詞＋「た」（「タ形」）＋「動機」の用例

Left	Center	Right
無くてもいいはず〉 金子容疑者がWinnyを開発した	動機	は、ネット社会が到来しつつある中で、旧態依然なビ
にわたる拘束と軟禁を受ける結果となった。同行した	動機	について張氏は、西安事件によって「蔣介石の威信
を企業の壁を乗り越えて行ってきた。 発端となった	動機	は日本社会における労働組合の地位の低下と、企業
ら打ち手であるが、プロの芸能者ではない。はじめた	動機	は子供が通う保育園のお祭りの出し物で、親も参加
何とかして今までと違ったものでやろうと考え出した	動機	そのものは非常に純粋であったと思う。ところが、そ

図 11　動詞＋「た」（「タ形」）＋「動機」

コアデータからは、「動機」を被修飾名詞とする動詞接続の連体修飾節は抽出できなかった。助動詞接続として「タ形」と共起する例文 5 例が挙げられた。4.1 の「事件」の考察でも触れたが、「ル形」も「タ形」も描写する状況において違いがないといわれる語は、実際の使用状況では「タ形」使用が多いと思われる。ただし、「動機」においては、「ル形」と「タ形」の選択は、「事件」よりも話者の主観的判断が大きく左右されているようである。

「動機」については、コアデータだけでなく、検索範囲を拡げ、BCCWJ コーパス全体を対象に検索をしてみた。その結果、「殺す動機」と「殺した動機」の違いとして、「事件」と同様、文末表現が遠い時は、「タ形」が使われやすいことが明確になった。また、話者が容疑者を犯人として認めている場合は、出来事が過去のこととして判断されるので「殺した動機」という「タ形」が選択されるが、話者が、容疑者として疑われている人物が真の犯人とは認められないという気持ちを持っていたり、実際に捜査の途中であるときは、「ル形」が使用されることが考察できた（例：「香菜さんを殺す動機は、まったくない」）。

5. 格外連体修飾形を形成する名詞の具体性と抽象性

格外連体修飾の特徴は、その名詞の内容を説明することである。そこで主名詞には、抽象名詞がよく使われる。抽象名詞とは、「個体ではなく事態の集合を指示する語である」（町田 2005）ことから、現在、表現しようとしている事態がどのような事態なのか説明を受ける「余裕」のようなものが名詞の中に内包されているといえる。これに対して、「固有名詞

は、集合ではなく一人の人間や一つの場所などの単独の個体を指示する」（同上）。したがって、固有名詞では、基本的に格外連体修飾形は形成されない。格外連体修飾の主名詞となる語の特徴として、大島（2010：6）は、「連体修飾節構造を形成するにあたって名詞の持つ情報が主導するタイプ」と述べている。そして、「名詞がもつ特性が連体修飾節の統語形式に反映されているのが外の関係といえるだろう。」と結論付けている（同上：29）。本項では、格外連体修飾節を形成する名詞の特徴をより客観的に考察する目的で、格内連体修飾節を含めて、名詞の特徴について、次の①～⑤のグループに分類を試みた。

表 3　連体修飾節と被修諸語の「名詞」の特徴

連体修飾節	被修飾名詞	特徴	
格内連体修飾節	①固有名詞	個別的	基本的に格内連体修飾節のみ
	②普通名詞	具体的 抽象的	例：生命体　　「人」 例：コロケーション「生じた問題」
格外連体修飾節	普通名詞	抽象的	③テンス・アスペクト区別あり 「ル形」も「タ形」使い分けられる。 例：「問題」・「話」
		抽象的 音調的要素	④テンス・アスペクトの区別は弱く、「タ形」が多用される。 例：「事件」・「動機」
		過去・完了 話者の判断	⑤「ル形」が多用される。 例：「必要」・「予定」

①グループ

固有名詞は、基本的に格内連体修飾形のみである。その中でも地名は、連体修飾節に用いられることが多いが、格関係を考えてみると「に」格によって結びついていることが多い。（例：　昨日、富士山に登った。→　昨日登った富士山）

②グループ

普通名詞のうち、日常的な事物や出来事を示すのに使う普通名詞は、格外連体修飾節を構成しにくい。また、生命性をもつ名詞もこのグループに含まれる。そして、「人」においては、被修飾名詞は連体修飾節内では主格であることがほとんどである。

③グループ

格外連体修飾節の被修飾名詞になりやすいのは、二字漢語動名詞であり、抽象名詞であることが多い。そのうち、「問題」「話」などは、テンス・アスペクトの区別に、描写する状況の違いが反映されている。

④グループ

これらは従来、「ル形」も「タ形」も両方とも使用可能とされていた語群であるが、コーパスに拠る考察では、「タ形」が多い。主観的判断によることもある。

⑤グループ

名詞の意味上、普遍的な内容や未来に関係するものなので「ル形」使用が圧倒的に多い。

6. まとめ

本研究は、連体修飾形を形成しやすい名詞について個々にその用例を考察することによ

って、格内・格外の量的・質的比較もふまえながら、格外連体修飾節内の文法的表現形式の特徴について分析した。その結果、格外連体修飾形を形成しやすい名詞は、抽象的な二字漢語が多く、その意味によって文法形式を決定付けている特徴を持つ。したがって、主体名詞を使って表現する状況がどのようなものであるかによって、テンス・アスペクトの有効性や動詞の「ル形」か「タ形」かの選択、またはその他の助動詞連体形のいずれかと共起するかを決定することが判明した。本研究においては、コーパスから多くの具体的使用例を抽出し、焦点を絞り込んで考察できることが可能になったため、格外連体修飾形の主体名詞の意味的特性とその文法形式の繋がりを明確にすることができた。

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。また、資料制作にあたり、田和英子氏から大きな協力を得ました。深く感謝いたします。

文　献

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.

Comrie, Bernard. (2010) Japanese and the other languages of the world. *NINJAL project review1*. pp.29-45.

岩崎　卓（1998）「従属節テンス認定の問題　－外の関係の連体修飾節の場合―」『大阪大学日本学報』17　pp.27-43.

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.

宮地朝子（2005）「形式名詞に関わる文法史的展開－連体と連用の境界として－」『國文學』學燈社

中島孝幸（1995）「現代日本語の連体修飾節における動詞の形について―ル形・タ形とテイル形・テイタ形―」『人文論叢』12 号, 三重大学

丹羽哲也（2013）「連体修飾における基本形とタ形の対立」藤田保幸編『形式語研究論集』和泉書院

大島資生（2010）『日本語連体修飾節構造の研究』ひつじ書房

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

代表性に配慮した『太陽コーパス』の分析法再考

森　秀明（東北大学大学院文学研究科）†

Methodological Reconsideration on the Representativeness of "Taiyo Corpus"

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

要旨

『太陽コーパス』は、明治後期～大正期の総合雑誌『太陽』から５年分を抽出した全文コーパスである。近代日本語の確立期をカバーしているため、語や文法の経年変化分析に使用されることが多い。しかし、代表性に配慮して設計されたサンプリングコーパスではないため、用例頻度や PMW で分析しても正確な結果が得られない場合がある。このため森（2014）では PTA という調整頻度で補正する分析を試みた。しかし、PTA の効果は限定的である上、代表性も担保できない。そこで今回はより代表性を有する分析法を検討した。この結果、著者名が判明している記事の記事数や分析対象の語が出現する記事の文字量で割合分析を行う方法がより有効であると考えられた。今後『太陽コーパス』で経年変化分析を行う場合は、用例頻度だけでなく、記事数や文字量でも分析することをお勧めしたい。

１．研究の目的

皆さんは『太陽コーパス』で用例検索を行った際、その調査結果に疑問を持ったことはないだろうか。『太陽コーパス』は本当に正確な値を示しているのか。そんな疑問から、森（2014）では『太陽コーパス』におけるデータの偏りを観察した。その結果、『太陽コーパス』では、記事の長さに 27 字～51,705 字というばらつきがあり、出版年ごとにジャンルの構成比も異なるため、用例頻度や PMW（Per Million Words：百万語当たりの出現頻度）で経年変化を比較しても、正確な分析にならない場合があると考えられた。そこで森 (2014) では PTA（Per Number of the Text Average Letters：一記事平均文字数当たりの頻度） という調整頻度を考案して記事の長さによる影響を均衡化し、ロジスティック回帰分析によってジャンルの偏りを補正する方法を試みた。しかし PTA は文字数に連動して用例頻度が増加しない語の分析ではあまり効果がない。しかもその補正結果が正確かどうかは、結局、外部の指標に頼るしかない。このため今回はより代表性を持った分析法を検討する。

２．『太陽コーパス』の代表性

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。現在、コーパスの代表性を担保する方法には主に次の 2 つが用いられている。一つは、推定対象の言語をある程度反映している図書館の蔵書などを現実母集団とし、そこからデータを無作為抽出する方法。もう一つは、データを超大規模に収集することで自己均衡化させ、推定対象言語のコンパクトな相似形を作る方法である（マケナリー＆ハーディー，2014；石川，2012 など）。『太陽コーパス』は特定の雑誌の全文コーパスであるから、このような統計学的な意味での代表性は担保されていない。これまで『太陽コーパス』が代表性を持つと主張されてきた根拠は、田中（2012）で述べられている次の言葉に集約されている。

† hideaki@moriharuo.com

> コーパスの重要な要件のひとつである代表性の担保については、対象とした総合雑誌『太陽』が、分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さの四点で、当時の文献資料としては格別の価値を持っていることから、『太陽コーパス』にも「代表性」が備わっていると見ることもできる。（田中，2012）

この主張は、これまでコーパス言語学で議論されてきた統計学的な意味での代表性とは異なる観点から「代表性」を主張したものである。このため、『太陽コーパス』がこれらの「代表性」を持っていても、用例頻度が統計学的に正確な値を出すことは担保されない。例えば 1925 年に日本で出版された書籍の中でアジアという地名が使用された回数に対し、1925 年の雑誌『太陽』に出現するアジアという地名の用例頻度がその何万分の一かの縮尺になっている可能性は担保できない。その可能性を確実に担保するには、1925 年に出版された書籍から無作為サンプリングを行ってコーパスを作る以外、方法はないと考えられる。

その一方で、田中（2012）が指摘する「分量の多さ、ジャンルの広さ、執筆陣の多彩さ、読者層の厚さ」という 4 つの特徴は、図書館書籍の性格とよく似ている。図書館の蔵書はある年に出版された書籍の中で、特に流通量が多かったものを中心に、社会的な需要を考慮して幅広いジャンルの書籍が集積されたものだ。雑誌『太陽』は、博文館が当時刊行していた『日本商業雑誌』『日本大家論集』『日本農業雑誌』『日本之法律』『婦女雑誌』を廃刊して一冊に統合した総合雑誌である。その内容は「百科全書的」で、創刊号は 28 万 5 千部、創刊以後約 10 年間は 10 万部弱の発行数があったと言われている（上野，2007）。雑誌『太陽』は単一の雑誌ではあっても、そのジャンルの広さや当時を代表する執筆陣、流通規模の大きさから、図書館書籍のミニチュア版的な性格を持ち合わせていると見なすことができる[1]。雑誌『太陽』が、統計学的に図書館書籍のミニチュアになっているのなら、『太陽コーパス』は堂々たる代表性を持っていると言えるだろう。これは『現代日本語書き言葉均衡コーパス』（以下 BCCWJ と呼ぶ）の「図書館書籍」が代表性を持っているという議論と同じである。

しかし、用例レベルで考えた場合、ある年に出版され図書館に収蔵された書籍の用例に対し、同じ年に雑誌『太陽』に書かれた記事の用例が、統計学的に一定の縮尺になっている保証はない。図書館書籍でアジアという語が使用される回数と雑誌『太陽』でアジアが使用されている回数を結びつける統計学的な根拠が見出し難いからである。

だが、著者を基準に考えた場合はどうであろうか。ある年の図書館書籍の著者の多くは、雑誌『太陽』の記事を書いた著者の多くと重なっているのではないか。雑誌『太陽』には当時を代表する執筆陣が記事を書いている。図書館に収蔵される書籍も当時を代表する書籍である。その著者の多くが一致している可能性はかなり高いと考えられる。当時の平均的な図書館の蔵書目録を入手し、その著者名と雑誌『太陽』の著者名の多くが一致しているなら、『太陽コーパス』は著者レベルでは、統計学的に一定の代表性を持っていると言っても過言ではないだろう。

しかし、残念ながらこの検証は難しい。当時は図書館が未整備で、毎年一定数の書籍を

[1] 『太陽』は 1928 年（昭和 3 年）2 月に廃刊となる。廃刊当時の流通量は不明だが、その量が激減していたことは想像に難くない。この意味で、田中(2012)が指摘する 4 つの特色がどの年代まで保たれていたかは、今後十分に検討していく必要がある。

安定して購入できるような体制にはなかった。内閣統計局（1912）『日本帝国統計年鑑　第31』（p. 553）[2]によれば、1910年の図書館数は全国で374館（官立・私立の合計）、その蔵書合計は2,643,264冊で平均7,000冊程度である。しかも中には1,000冊前後しかない図書館もある。当時の平均的な図書館像を決めるのも難しく、当時の蔵書目録を入手するのはさらに困難である。このためここで著者レベルでの『太陽コーパス』の代表性を実証することは難しい。

ただし、大まかな目安ならつけられる。表1は、当時の書籍の出版数と、『太陽コーパス』で氏名が判明している著者数である。

表1 近代の出版物数[3]と『太陽コーパス』の氏名判別著者数

	1895年	1901年	1909年	1917年	1925年
著述	8,334	18,963	34,066	46,012	18,028
編集	17,712				
翻訳	124	35	57	118	
合計	26,170	18,998	34,123	46,130	
『太陽コーパス』氏名判別著者数	238	212	245	155	245

使用した統計書は年によって集計の仕方が異なるが、基本的に著述は普通出版物、編集は雑誌だと思われる。表1の「著述」の冊数がBCCWJで言えばその年に出版された全ての書籍の数＝「出版書籍」の母集団の数である。表1からごく荒く推定すれば年1、2万冊が出版書籍の母集団の数となる。ここから図書館に収蔵する書籍を選ぶとして、平均7,000冊しか蔵書のない図書館が、毎年何千冊も追加購入することは考えにくい。かといってあまりに少ない冊数では、図書館書籍自体が近代日本語の代表性を失ってしまう。いま仮に推定出版書籍数のおよそ1/10～1/20に当たる1,000冊を一年当たりに購入される図書館書籍の母集団だとしてみよう。この1,000冊を著者1,000人と読み替えるなら、その1,000人の中に『太陽コーパス』の氏名判別著者が含まれている可能性はかなり高いと言えるだろう。今、その割合が何%になるのかは分からない。しかし、重要なことは用例頻度の場合その代表性を担保する統計学的な根拠は見出し難いが、著者数で考えれば確実に何%かの代表性は担保できるということである。著者数で分析する場合、「『太陽コーパス』には代表性がない」という帰無仮説は統計学的な根拠を持って棄却されると考えられる。

３．指標としての記事数

言語の経年変化を分析する場合、用例頻度で分析するということは、例えばアジアと言う地名に対して「亜細亜」という漢字表記が何例出現し、「アジア」というカタカナ表記が何例出現しているかを調べ、その割合の変化を観察することである。一方これを著者数で観察するということは、例えば代表性を持った1,000人の中で何人が漢字で表記し、何人がカタカナで表記するかの割合の変化を見ることである。厳密に言えば用例頻度割合と著者数割合は異なる現象を観察していることになる。しかし言語変化は、つまるところそれを使用する人間の言葉遣いの変化であるから、著者数割合を使用しても言語学的に意義の

[2] http://kindai.ndl.go.jp/info:ndljp/pid/974420（2015.01.31 閲覧）

[3] 1895～1909年は『大日本帝国内務省統計報告』、1910年～1925年は『日本帝国統計年鑑』による。http://kindai.ndl.go.jp/（2015.01.31 閲覧）

ある観察をしていると考えられる。

ただし、同じ著者でも学術的な論文の場合は漢字で表記し、大衆的な読み物の場合はカタカナで表記することも考えられる。このため、一冊の書籍や一つの記事を単位とし、その書籍や記事が漢字表記、カタカナ表記、併用、未使用のどれになるかを観察した方がより実際的だと思われる。このように記事数と言う単位で観察しても、その根本は著者に根ざしているため、この記事数も一定の代表性を持っていると考えられる。

問題は、その代表性がどれぐらいあるかである。母集団 1,000 人のうち『太陽コーパス』と一致している著者が 100 人しかいない場合、代表性は 10%しかないように思える。しかし、『太陽コーパス』の 100 人が母集団のごく平均的な傾向を示しているなら、例えば 1909 年や 1925 年の著者数は 245 人であるから、100 人÷245 人＝40.8%は母集団のごく平均的な傾向を示していることになる。残りの 145 人だけが非常に偏った表記法を使用しているとは想定しにくいので、『太陽コーパス』が相当の割合で母集団の正確な姿を反映している可能性がある。その一方で母集団と一致した 100 人が平均より偏った表記法を使用していた場合、『太陽コーパス』が母集団平均と大きくかけ離れた姿をしていることも考えられる。

この問題は分析対象の言語現象にどのような要因が影響しているかに関わっている。例えば外国地名を漢字表記するかカタカナ表記するかの場合なら、学術書などの硬い文章では漢字が用いられ、大衆向けの柔らかい文章ではカタカナが用いられることなどが考えられる。これをジャンル的に見れば、社会科学などは漢字が使われやすく、文学などではカタカナが使われやすいなどの現象となって現れる可能性がある。雑誌『太陽』の編集方針が学術的な記事に偏っていたり、ジャンル構成が母集団の傾向と大きく異なっている場合、『太陽コーパス』の代表性は低い可能性がある。その逆に当時の母集団平均と同じような文章の硬軟度やジャンル構成で編集されていたとしたら、『太陽コーパス』の代表性は高い可能性がある。これ以上は想像の域を出ないが、雑誌『太陽』が百科全書的な総合雑誌であり、商業的に大きな成功をおさめた雑誌であることを考えれば、『太陽コーパス』の代表性が高い場合の方が多いのではないかと思われる。

ここまでは、『太陽コーパス』の中で著者名が判明している記事を対象に考察してきた。『太陽コーパス』の中で、著者名が判明している記事はおよそ 7 割である。残りの 3 割は無署名でその多くは雑誌記者が執筆していると考えられる。これらの無署名記事はどのように扱えばよいだろうか。これまでの代表性の議論から言えば、雑誌記者が図書館書籍の母集団に含まれている可能性は低いと思われる。また、雑誌記者の場合、編集部の方針によって表記法などの言葉遣いに一定の制約がかかっている可能性もある。このため基本的に無署名記事は除いて分析した方が正確な結果が得られると考えられる。

特に無署名記事では表 2 に見られる〈小話〉〈世界のラヂオ〉〈新刊紹介〉などのように、同じ号に同じ題名で書かれた複数の短文記事が観察される（以後、これを同号同名記事と呼ぶ）。これらは本来ならまとめて一つの記事として掲載されてもおかしくない内容だが、雑誌を読みやすくする意図からか、特に 1925 年の長文記事の間に埋め込まれるように編集されている。これらを別々の一記事と認定すると、同一の著者と思われる無署名記事を何回もカウントしてしまうため、同一著者の言葉遣いを過大に評価してしまうことになる。同号同名記事を統合して一記事と見なした上で署名記事の言葉遣いと比較し、その傾向に大きな違いがあるなら、これらを分離して観察する方法が妥当だと思われる。

表 2　1925 年 04 号の記事配列（開始から 20 記事目まで／全 78 記事）

No.	題名	文字数	No.	題名	文字数
1	昨年の今月	654	11	日米海軍勢力の比較	5,337
2	普選実施後の政党	9,408	12	〈世界のラヂオ〉	267
3	〈和田豊治氏母堂米寿に寄せられた詩歌〉	434	13	明治初年外交物語（その七）苦心の犯人捜索	7,176
4	時事漫吟	905	14	〈世界のラヂオ〉	583
5	〈小話〉	126	15	新人有馬頼寧	5,650
6	赤露印象記	6,276	16	〈冬の日に〉　丹下生	82
7	〈世界のラヂオ〉	634	17	〈小話〉	65
8	普選実施の影響と女子参政権問題	6,458	18	戦場の悪戯者―空想の兵器―運命の弾丸―	7,364
9	〈世界のラヂオ〉	329	19	〈小話〉	65
10	〈新刊紹介〉	570	20	今は我れ　丹下生	42

４．指標としての文字量

記事数という指標は、一定の統計学的な代表性を有していると考えられる。しかし、『太陽コーパス』の記事には 27 字～51,705 字というばらつきがある。記事数で分析する場合、27 字の記事も 51,705 字の記事も同じ 1 記事となるが、その扱いで良いものだろうか。

図書館書籍を日本語の代表と見なす考え方の中は、その当時、大量に流通していた書籍の方が日本語の代表としてふさわしいという前提があると思われる。短い記事しか依頼されない著者と長い記事を依頼される著者では、日本語を代表する代表度に差があると考えられる。例えば 1,000 字の記事 10 本に外国地名がカタカナ表記されていたとする。一方、10,000 字の記事では漢字表記されていたとする。その場合、カタカナ：漢字の比率は 10：1 でいいのだろうか。これが口語・文語の割合ならどうだろう。1,000 字の口語記事 10 本と 10,000 字の文語記事 1 本の場合、雑誌の口語：文語比率は本当に 10：1 でいいのだろうか。

雑誌の編集者の立場で考えた場合、記事の硬さ・柔らかさの比率や、口語・文語の比率は、当然コントロールの対象になったと思われる。これらの分量を最も読者層に受け入れられやすい比率とすることで、雑誌の販売量の最大化を図ったと考えられる。このように編集者が市場のニーズに配慮することによって反映された代表性を「市場代表性」と名付けるなら、記事数より文字量の方が市場代表性が高いと考えられる。つまり先の例でいえば、10：1 ではなく 1：1 と数える方が、より市場代表性を反映していると考えられる。

記事の硬さ・柔らかさや口語・文語の比率などは、言葉遣いの比率に大きな影響を与える。特に言語の交替現象を観察する場合、新しく使用されるようになった言葉遣いは、まず、話し言葉や柔らかい記事から使用される傾向がある。この割合がコントロールされた文字量は記事数以上に母集団の正確な姿を反映している可能性がある。また、雑誌の編集者は無署名記事も含めて様々なコントロールを行っていたと考えられるため、無署名記事を削除しない方がより市場代表性を有している可能性がある。ただし、このような市場代表性は、統計学的に立証できる類のものではないと思われる。このため、統計学的に一定の代表性を有すると考えられる記事数と併用しながら、比較検討する方法が妥当であろう。

５．ケーススタディ

ここでは 2 つの先行研究を取り上げ、記事数、文字量を指標とした割合分析の有効性と問題点を検討する。記事数、文字量を指標とするだけでなく割合分析も行うのは、『太陽

コーパス』における出版年ごとの不均衡性を平準化するためである。これまで割合分析は主に言語現象を観察する目的で使用されてきたが、出版年の影響を除く効果も高いと考えられる。例えば外国地名表記の経年変化を調べる場合、出版年ごとの文字数や記事数が異なるため、単純な頻度では比較できない。これを割合分析すればこれらの要因は相殺されて比較可能な値になると考えられる。

$$\text{カタカナ割合} = \frac{\text{カタカナの頻度} \times \cancel{\text{出版年の影響}}}{(\text{カタカナの頻度} + \text{漢字の頻度}) \times \cancel{\text{出版年の影響}}}$$

５.１井出（2005）「外国地名表記について―漢字表記からカタカナ表記へ―」の再分析

井出（2005）は、外国地名が漢字表記からカタカナ表記へ移り変わっていく経年変化を分析した研究である。この研究では、先駆的な試みとして分析の指標に記事数が使用されている。初めに井出（2005）が記事数を指標に採用した考え方を見てみよう。

> 頻度ではなく記事数を指標にしたのは，地名の場合，記事の種類によって，同一記事内に同一語が繰り返して出現している場合があり，頻度よりも記事数の方が指標としてまさっていると考えられるからである。年代別の使用の推移を見ようとするなら，一つの記事に何語出現するかということは無視し，出現した記事を１として数えた方がより正確にその推移の変化を見ることができると思われる。（井出，2005，p. 159）

井出（2005）では、地名のような特徴語[4]的性格を示す語の場合、用例頻度より記事数の方が正確だと主張されている。しかし、なぜ記事数の方が指標として優っているのかについて、理論的な考察がなされていない。このため、井出（2005）では、同号同名記事を統合する必要性や署名記事と無署名記事を分離して観察する必要性について、検討されていない。井出（2005）では、最終的に 1925 年にカタカナ表記が急激に増加したと結論づけられているが（p. 170）、その結論には疑問が残る。以下、これを再分析してみる。

井出（2005）では、21 の地名について個別に観察が行われている。しかし、21 の地名ごとに分析した結果、分析に適さないほどデータ数が少なくなっている地名が散見される。計量分析では少しでもデータ数が多い方がより正確な分析となることから、ここでは 21 の地名を合計した分析を行う。初めに用例頻度、記事数、文字量を指標とし、割合分析を行わずに経年変化を観察する。ここで使用するのは記事を統合したり無署名記事を除いたりしない、全数での観察である。

図 1 の用例頻度を観察すると、1917 年の漢字地名がそれまでの 2 倍弱使用されていることが目につく。図 2 で 1917 年の記事数を観察すると、記事数はむしろ減少していることから、この現象は一記事当たりに使用されている漢字地名が増えていることを意味している。1917 年は 1914 年に始まった第一次世界大戦や 1917 年に起きたロシア革命に関する記事などが多く、増加の原因にはそれらの記事で漢字地名が多用されたことが考えられる。問題

[4] 特徴語とは、あるテキストに頻出し、そのテキストの性格を特徴づけるような語を意味する。例えば海外の事情を紹介したテキストなどでは外国地名が頻出し、それが特徴語となる場合がある。美術・芸術、戦争・平和などのように、テキストのテーマに深くかかわる語は、特徴語となる可能性がある。

はこのような増加が雑誌『太陽』独自の現象なのか、日本語全体の現象なのかである。第 3 節で想定した例で考えれば、図書館書籍 1,000 冊から用例を抽出しても図 1 のような現象が観察されるなら、日本語全体の現象と言える。しかし、様々なジャンルの書籍 1,000 冊の合計で、なお漢字を使用した外国地名がそれまでの 2 倍弱にもなることは考えにくい。よって、この用例頻度はあくまでも雑誌『太陽』の姿を現したものと思われる。

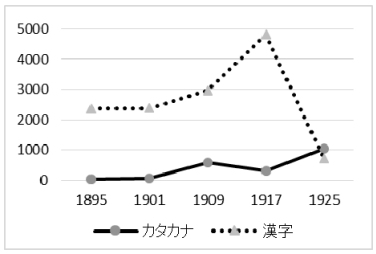


図 1 表記別外国地名用例頻度

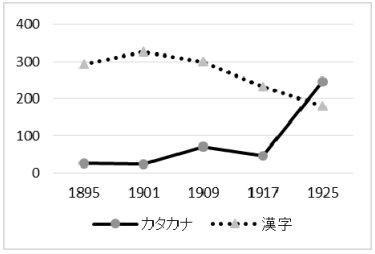


図 2 表記別外国地名記事数

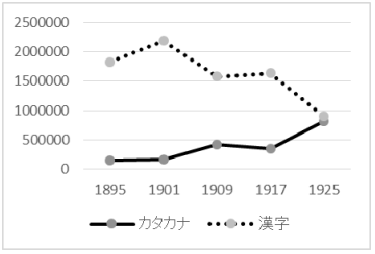


図 3 表記別外国地名記事の文字量

図 2 では、1925 年で外国地名をカタカナで表記する記事の本数が急増する現象が目につく。これと図 3 の文字量を比較すると、外国地名をカタカナで表記する記事の文字量はさほど増加していない。図 2 の現象は 1925 年のカタカナ表記をしている記事が、ごく短い文字数で書かれ、さらにその記事数が多いことを示している。これには表 2 で観察した同号同名記事の問題が反映されていると考えられる。同号同名記事は同一著者（または同一の属性を持った複数の雑誌記者）によって書かれていると思われ、これを重複してカウントすると著者を単位にした正確な分析はできない。図 3 は文字量である。文字量には、統計学的な代表性は考えにくく、読者のニーズを反映した市場代表性が推定されるだけである。しかし、図 3 を見る限り、図 1、2 に見られるような明らかな偏りは観察されない。

次に同号同名記事を統合した場合の記事数を観察する（以後これを統合記事数、統合前の記事数を単純記事数と呼び分ける）。図 4 は、統合記事数のグラフである。同号同名記事を統合した結果、1925 年の偏りは解消され、図 3 の文字量のグラフに近くなった。

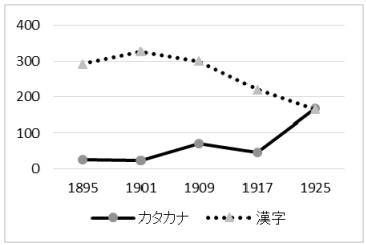


図 4 表記別外国地名統合記事数

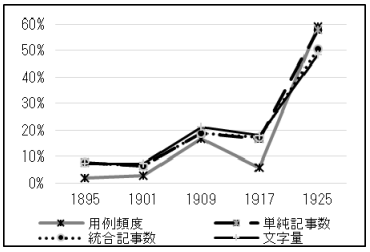
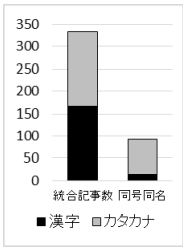
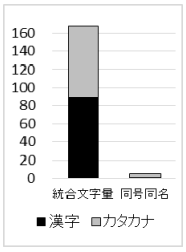


図 5 外国地名の指標別カタカナ割合



縦軸：記事数
図 6 記事数



縦軸：万字
図 7 文字量

図 5 は、用例頻度、単独記事数、統合記事数、文字量を指標として算出したカタカナ割合である。統合記事数と文字量のグラフの形状はほぼ一致し、1925 年の値が約 50%になる。一方、単純記事数は 1917 年まではこれらと同じだが、1925 年は 60%弱で、用例頻度の値と同じになる。図 6 は統合記事数と同号同名記事の本数を比較したグラフである。これを見るとカタカナを使用した同号同名記事だけで約 100 本になることが分かる。図 7 は同じものを文字量で描いたグラフである。文字量に直すと、カタカナを使用した同号同名記事は

約 1.4 万字しかなく、ほとんど影響力を持っていない。井出（2005）は、単純記事数に基づいて分析したため、1925 年のカタカナ割合を過大評価していると考えられる。

ただし、図 5 の統合記事数や文字量割合のグラフが直ちに代表性を持っているとは見なし難い。図 8 は、一記事当たりに 1、2 回しか外国地名が出現しない低頻度出現記事と、一記事当たりに 3 回～366 回出現する高頻度出現記事に分け、さらに著者名が判明しているかいないかを加味して全体を 4 つのグループに分けたグラフである。指標には文字量を使用している。今、議論を単純化するために低頻度記事を一般記事、高頻度記事を専門記事と見なすと、著者名が判明している一般記事では、カタカナ割合は一定の割合で増加していたことが分かる。著者不明の記事は、雑誌『太陽』の記者による記事と思われるため、これらのカタカナ割合は編集方針によって統制されていた可能性がある。著者名が判明している専門記事も類似の傾向を示しているが、総じてカタカナ割合が高い。

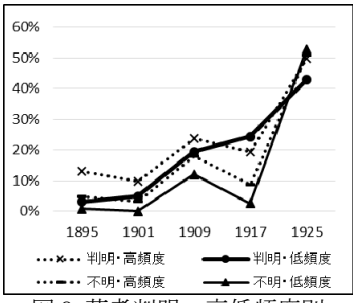


図 8 著者判明・高低頻度別カタカナ割合

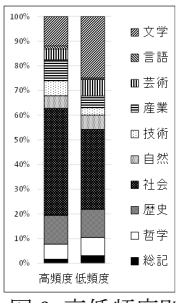


図 9 高低頻度別ジャンル

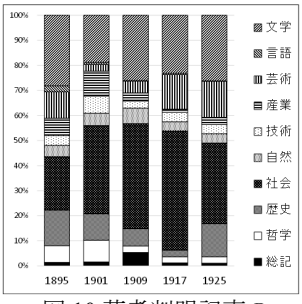


図 10 著者判明記事の出版年別ジャンル

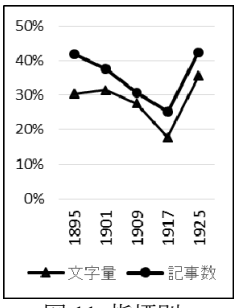


図 11 指標別低頻度記事割合

図 9 は、図 8 の著者判明記事のジャンルを高低頻度別に描いたグラフである。高頻度記事では社会のジャンルが多く、低頻度記事では社会が減って文学が増えている。図 10 は著者判明記事のジャンルを出版年ごとに描いたものである。ジャンル構成は出版年によって変化しており、特に 1909 年と 1917 年で社会のジャンルが多い。図 11 は文字量と記事数の指標別に著者判明記事の中で低頻度記事がどれぐらいの割合になるかを示したものである。特に 1909 年と 1917 年で低頻度記事が低下している。図 10 のグラフと図 11 のグラフには連動性が見られる。

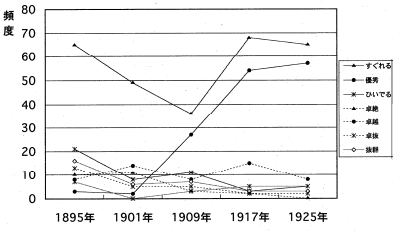
図 8 において、代表性が担保できるのは著者判明のグラフである。これらの高頻度：低頻度記事の割合は、図 11 のようにおよそ 6：4（記事数）または 7：3（文字量）となるため、そのまま合計すると高頻度記事の影響が強く出る。この結果、『太陽コーパス』の著者判明記事割合は図 5 の統合記事数のグラフに近くなる。しかし 1909 年や 1917 年にはジャンルや高低頻度割合の偏りがある。これを補正した場合、特に 1917 年の落ち込みは図 5 より少なくなると考えられる。このため、正確なカタカナ割合は図 5 の統合記事数から図 8 の判明・低頻度の形状にもう少し近づくと思われる。つまり、外国地名のカタカナ割合は 1925 年に急増するのではなく、一定の割合で徐々に増加していた可能性が考えられる。

以上の観察から、用例頻度、単純記事数、無署名記事を使用すると、分析が不正確になる例が確認された。また、著者判明記事の記事数は一定の代表性を持つと考えられるものの、ジャンル等で言葉遣いの使い分けがなされている言語現象では、『太陽コーパス』におけるジャンルの偏りを補正しないと、高い代表性は見込めないことが考えられる。

５.２　田中（2005）「漢語「優秀」の定着と語彙形成―主体を表す語の分析を通して―」の再分析

田中（2005）は明治期に新しく作られた「優秀」という漢語が、「卓越、卓絶、卓抜、抜群」といった古くからある漢語（以後「卓越類」と呼ぶ）や、「すぐれる」といった和語とのかかわりの中で、どのように定着していったのかを分析した研究である。その結果、「漢語「優秀」は，和語「すぐれる」との間に意味的な使い分けをもったことで，語彙の基本的な部分に深く浸透したものと考えられる。」（p. 139）と考察されている。これは、用例の統語的な分析を詳細に行った結果から導かれた結論だが、ここではごく単純に全体の数量的な観点から再分析してみる。

図 12 は田中（2005）に掲載されている用例頻度のグラフである。先にも述べたが、『太陽コーパス』では出版年ごとの文字数や記事数が一定でないため、用例頻度そのものでは偏りが出る。このため、用例頻度を使用して割合分析を行ったグラフが図 13 である[5]。この際、「卓越類」は合計して集計した。図 13 を見ると「優秀」と数量的に競合しているのは「卓越類」であり、「すぐれる」は数量的にはほぼ無関係であることが観察される。



田中（2005）より引用（p. 134 ）

図 12〈優秀〉語彙の年次別用例頻度

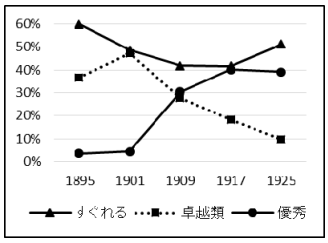


図 13〈優秀〉語彙の年次別用例頻度割合

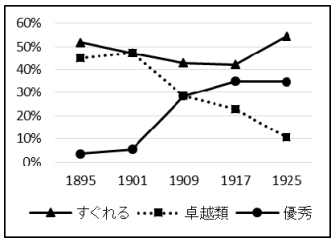


図 14〈優秀〉語彙の年次別統合記事数割合

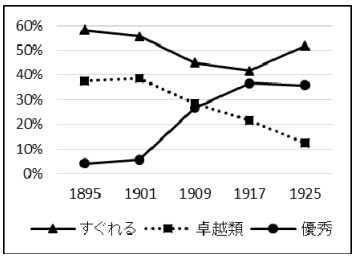


図 15〈優秀〉語彙の年次別著者判明記事数割合

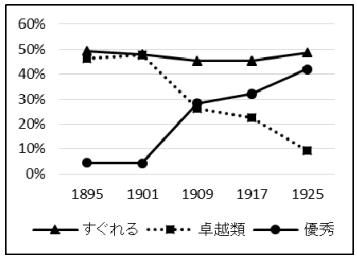


図 16〈優秀〉語彙の年次別文字量割合

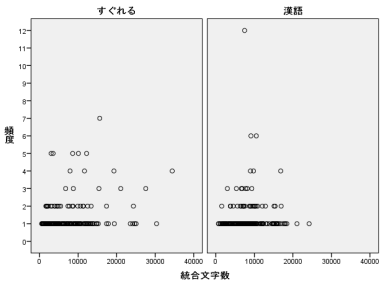


図 17 「すぐれる」と〈優秀〉漢語語彙の文字数別散布図

図 14～16 は、少しずつ形は変化するものの、基本的に図 13 と同じ形状をしている。第 2 節で行った代表性の議論からすれば、この中で統計学的な代表性を持つと考えられるのは図 15 であり、図 13 の用例頻度では代表性が担保できないはずであった。それなのになぜこれほど形状が似ているのであろうか。その理由は、図 17 の散布図にある。図 17 は、記事の文字量を横軸に、一記事当たりの使用回数を縦軸にして描いた散布図である。これを見ると、一記事に用例が 1 回しか出現しない記事が最も多く、大半は 2 回までの出現にとどまっている。この傾向はどんなに文字数が多い記事でも基本的に変わらない。用例頻度

[5] データは発表者が現行の『太陽コーパス』から抽出したものを使用している。また、1925 年 01 号阪谷芳郎「近代文明と発明」は外れ値とみなして除いてある。またこれ以後のグラフでは論点を絞り込むため「ひいでる」は描いていない。

が一記事当たり 1 回であれば、用例頻度と記事数は完全に同一になる。これが平均 2 回になったとしても、互いの出現傾向が同じであれば、割り算をすれば記事数割合と同じになる。代表性が担保できないはずの図 13 が一定の代表性を有すると考えられる図 15 とよく似たグラフになるのは、用例頻度を使用しても、その割合分析の結果が記事数割合とほぼ同様の結果となるからである。つまり、用例頻度を使用しても、割合分析の結果が記事数割合と似た値になる語の場合、概ね正確な分析結果を示すと考えられる。

これらに比べ、図 16 の文字量のグラフは「すぐれる」がほぼ直線的に推移して形状がやや異なる。この理由は「すぐれる」が和語であり、小説や雑学的な記事に現れやすいためだと思われる。小説の文字数は長いものが多く、雑学的な記事は短いものが多い。これらの割合は記事数的には出版年ごとのばらつきがあるが、文字量から見れば常に 5 割前後になっている。これは「すぐれる」と言う語が使用されるタイプの記事が、全ての出版年を通じてほぼ一定であることを示唆しているのかも知れない。第 3 節で検討した市場代表性を重く見れば、図 16 の方が正確な近代日本語の姿を示しているとも考えられる。

以上の観察から、用例頻度割合でも概ね正確な分析となる例が確認された。ただし、それは検索語がどの記事にも同程度の回数で使用され、結果的に用例頻度割合が記事数割合と同じになるからだと考えられる。

６．まとめ

これまで『太陽コーパス』の分析では、用例頻度を使用した研究が多かった。しかし、用例頻度は代表性を統計学的に担保することが難しい。その一方で著者名が判明している記事数は、統計学的に一定の代表性を担保できると考えられる。また、統計学的な証明は難しいが、用例が出現する記事の文字量は、読者のニーズを反映した市場代表性を有していると考えられる。ただし、この 3 種類の指標は、厳密には別々の現象を表していると考えられる。このため、『太陽コーパス』の分析に当たっては、これら 3 種類の指標を併用し、その振る舞いの違いを観察していく分析法が有効だと思われる。

文　献

井出順子（2005）「外国地名表記について―漢字表記からカタカナ表記へ―」国立国語研究所（編）『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，pp. 157-172.

石川慎一郎（2012）『ベーシック　コーパス言語学』ひつじ書房.

上野隆生（2007）「研究プロジェクト 日本近代化の問題点--明治国家形成期の明と暗 雑誌『太陽』の一側面について」『東西南北』2007，和光大学総合文化研究所，pp. 252-285.

田中牧郎（2005）「漢語「優秀」の定着と語彙形成―主体を表す語の分析を通して―」国立国語研究所（編）（2005）『雑誌『太陽』による確立期現代語の研究―『太陽コーパス』研究論文集―』博文館新社，pp. 115-141.

田中牧郎（2012）「近代語コーパスにおける資料選定の考え方」『近代語コーパス設計のための文献言語研究 成果報告書 』（国立国語研究所共同研究報告 12-03）.

マケナリー&ハーディー（2014）石川慎一郎（訳）『概説コーパス言語学－手法・理論・実践』ひつじ書房．[McEnery, T.& Hardie, A.（2012）*Corpus Linguistics; Method, Theory and Practice.* Cambridge University Press.]

森秀明（2014）「均衡性と代表性に配慮した『太陽コーパス』の分析法試論」『第 5 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，pp. 73-82.

ポスター発表　グループA

3月11日（水）　13:00～14:00

BCCWJ の接続詞の品詞情報の解析精度について

馬場 俊臣（北海道教育大学教育学部）

On the Precision of the POS Information: Focusing on the Conjunctions in the BCCWJ

Toshiomi Baba (Hokkaido University of Education, Sapporo Campus)

要旨

接続詞を扱った研究において BCCWJ の品詞情報を利用する際の留意点を示すために、BCCWJ で「接続詞」の品詞情報が付与された語（長単位）の解析精度の調査を行い、以下の結果を得た。(1)サンプル調査（非コアデータ各 100 件）の結果、品詞情報「接続詞」の使用頻度上位 20 語の適合率は 63.0%～100.0%の範囲にあり、特に「で」「唯」「又」の適合率が低い。(2)「又」の詳細調査（非コアデータ 1000 件）の結果、適合率は 85.8%であり、レジスター別では「特定目的・ブログ」42.4%が特に低い。(3)「で」の詳細調査（非コアデータ 1000 件）の結果（ただし 200 件の途中経過）、適合率は 62.5%であり、レジスター別では「特定目的・知恵袋」44.1%が特に低い。なお、本研究は、品詞情報付与に関する解析器改良のための参考資料を提供するものでもある。

１．はじめに

『現代日本語書き言葉均衡コーパス』（BCCWJ）を利用した接続詞研究の問題点と可能性に関する基礎的研究の一環として、本稿では、BCCWJ の接続詞に関する品詞情報の信頼性を見るために、品詞情報「接続詞」[1]の解析精度に関する調査結果を報告する。

BCCWJ の解析精度は、「長単位・短単位とも、データ全体に対して人手修正を行ったコアデータは 99%以上、データの一部に対して人手修正を行ったコアデータ以外のデータは 98%以上」（小椋、冨士池(2011):39）とされるが、品詞によって解析精度は若干異なると予想される。また、同じく接続詞であっても語により解析精度が異なると予想される。

BCCWJ を利用した重要な研究の一つに、品詞比率に基づいた文章・文体研究がある[2]。こうした巨視的な研究では、品詞の違いによる解析精度の若干の異なりは、分析結果に殆ど影響を与えず何ら問題は生じない。しかし、例えば特定の品詞に限定して、その品詞に属するいくつかの語（ないし語群）の比率を問題にする場合は対象とする語の解析精度の違いが分析結果に影響を及ぼす可能性がある。特に接続詞は、属する語の種類（異なり語）が少なく、一つ一つの語の解析精度の違いが場合によっては分析結果に大きな影響を及ぼす恐れがある。

BCCWJ を利用する際の基本としては、利用マニュアル[3]や小木曽(2014)に示されているように「解析誤り」「形態素解析の弱点」があることを前提として、研究目的・研究対象

---

1 品詞情報として「接続詞」が付与されていることを、以下「品詞情報「接続詞」」又は単に括弧を付けて「接続詞」と略記する。他の品詞についても同様である。

2 品詞比率とジャンル（レジスター）等の文体・文章構造の違いとの関連を分析した研究として、冨士池他(2011)、鯨井(2011)などの研究がある。なお、左記の二つの研究では、誤解析に対する人手修正を施したコアデータ（長単位）を使用している。

3 国立国語研究所コーパス開発センター(2011)、国立国語研究所コーパス開発センター(2013)。

に応じて人手による点検が必要になる。こうした点検を行うことによって、語による解析精度の違いの問題を避けることができる。

しかし、検索結果をそのまま利用する場合などでは特に、一つ一つの語の解析精度の違いがどの程度有りうるのかという知見を予め知っておくことが重要である。

本稿では、このような問題意識に基づいて、BCCWJ の「接続詞」の品詞情報の信頼性を見るために、「接続詞」の用例の解析精度に関する調査を行い、その結果を報告する。調査内容は次の通りである。

(1) 「接続詞」の使用頻度上位 20 語（長単位）についてサンプル調査（非コアデータ各 100 件）を行い、語ごとの適合率[4]を明らかにする。（3 節）

(2) 適合率が低い「又」（使用頻度第 1 位）について、サンプル数を増やした詳細調査（「接続詞」「副詞」各 1000 件）を行い、「接続詞」及び「副詞」の適合率を明らかにし、さらに、レジスター別での違いも明らかにする。（4 節）

(3) 適合率が最も低い「で」について、サンプル数を増やした詳細調査（「接続詞」「格助詞」「助動詞」各 1000 件）を行い、「接続詞」及び「格助詞」「助動詞」の適合率を明らかにし、さらに、レジスター別での違いも明らかにする。（5 節）

なお、本研究は、BCCWJ を利用した今後の接続詞研究[5]に対して重要な基礎的知見を提供するとともに、品詞情報付与に関する解析器の改良のための参考資料を提供するものでもある。

２．BCCWJ 全体の品詞情報の解析精度について

調査結果を示すに先立って、公表されている BCCWJ 全体の品詞情報の解析精度を示す。本稿の調査は、BCCWJ において、「接続詞」の品詞情報が付与された長単位[6]の語彙素を対象とする。検索ツールとして、品詞情報を用いた検索ができる「中納言」を利用する。

BCCWJ の形態論情報の付与では、「短単位解析には解析エンジン MeCab と形態素解析用辞書 UniDic を、長単位解析には短単位解析結果から長単位を自動構成する解析器」（小椋、冨士池(2011):44）を用いており[7]、また（短単位全体の）「1 億語のうち約 100 万語（コアデータ）については、自動解析後に人手修正を行い、解析精度 99%以上の高精度なデータとし、形態素解析システムの学習用データとして用いた」（同:64）とのことである。

接続詞に関しては、UniDic における接続詞（短単位）は 30 語であり（UniDic-mecab version 1.3.12 の接続詞辞書（Conjunction.csv）による）、さらに、長単位では 32 の「連語」（従って、そうして、其れとも、では等）が接続詞として扱われている（同:69）。

BCCWJ の形態論情報の解析精度は、コアデータは 99%以上、コアデータ以外のデータは 98%以上（同:39）とのことである。レジスター別では、「白書、書籍（文学）、書籍

---

4 本稿では解析精度として「適合率」を用いた。「適合率」は「正しく品詞情報を付与された長単位数／当該品詞情報を付与された長単位数」で求めた。本稿の調査では「再現率」は調査しておらず、従って「F 値」も求めていない。脚注 8 も参照。

5 接続詞研究においても BCCWJ を利用した研究が増えている。ただし、検索ツールや検索方法の詳細、また、検索結果に対する人手による点検の有無の詳細が示されていないものがある。コーパスを用いた研究の特徴の一つに追試可能性が挙げられる。それを保証するためには、検索及び用例確定の方法を明示することが必須となろう。

6 多くの接続詞研究において接続詞として扱われる語の単位は、「長単位」にほぼ相当する。

7 本稿での指摘は MeCab+UniDic により付与された品詞情報の問題点でもある。

（文学以外）、新聞、Web（Y!知恵袋）」の各レジスターの「品詞」の解析精度（F 値）[8]は、それぞれ 0.995693、0.9866095、0.989596、0.989116、0.984112 となっており、98%以上を実現している（同:45）。BCCWJ の利用マニュアルに記載されている解析精度は F 値のみであり、適合率及び再現率は示されていない。小木曽他(2010)では、「新聞」（毎日新聞 2007 年度版）・「文学作品」（新潮文庫の 100 冊）・「ブログ」（Yahoo!ブログ）を用いて UniDic-mecab と他の解析器との精度比較を行い「UniDic-mecab 1.3.12」での適合率、再現率、F 値を示している。新聞、文学作品、ブログの順にそれぞれ「品詞」の適合率は 0.9879、0.9772、0.9756 であり 98%前後以上である。

３．高頻度接続詞の適合率

３.１　調査の目的と方法

本節では、品詞情報「接続詞」の語のうち、使用頻度上位 20 語（長単位）（以下、「高頻度接続詞」と呼ぶ）について、サンプル調査（非コアデータ各 100 件）を行い、語ごとの適合率を明らかにする。

まず、高頻度接続詞を取り出すために、「中納言」長単位検索で「品詞　大分類　接続詞」を指定し、全レジスター対象に検索[9]を行った[10]。検索総件数は 668,836 件である。語彙素を単位として集計し、頻度合計上位 20 位までの語を選定した（表 1 参照）[11]。

次に、各接続詞からサンプルを抽出した。コアデータについては自動解析後に人手による修正を行っているため、サンプル調査の対象は非コアデータのみとする。「中納言」長単位検索で「語彙素」「品詞　大分類　接続詞」を指定し検索[12]を行い、検索結果画面上

---

8 適合率（精度）、再現率、F 値は分類の評価指標として用いられる。適合率は付与された品詞がどのくらい正しいかを表す指標である。再現率は実際にある品詞であるものをどれくらいカバーして付与できているかを表す指標である。F 値は適合率と再現率の調和平均である。接続詞を例にすると、次の式で求められる。

（適合率）＝（品詞情報「接続詞」を付与されて正しく接続詞であった件数）／（品詞情報「接続詞」を付与された件数）×100[%]

（再現率）＝（品詞情報「接続詞」を付与されて正しく接続詞であった件数）／（調査対象全体で実際に接続詞である件数）×100[%]

（F 値）＝2×（適合率）×（再現率）／（（適合率）＋（再現率））

9 検索条件式は、「キー: 品詞 LIKE "接続詞%" WITH OPTIONS unit="2" AND tglWords="10" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"」である。なお、「中納言」では 10 万件以上の一括ダウンロードができないため、いくつかのレジスターごとに分割してダウンロードを行った。

10 本稿での「中納言」検索結果は、高頻度接続詞及び「又」の詳細調査に関しては 2013 年 11 月～2014 年 2 月、「で」の詳細調査に関しては 2014 年 12 月～2015 年 1 月の期間で得られた結果である。

11 「『現代日本語書き言葉均衡コーパス』長単位語彙表 ver1.0」（DVD データに基づく語彙表）では、「だから」「だが」「所が」の頻度合計はそれぞれ 21,010、17,871、11,394 であり、本調査と比べいずれも非コアデータの頻度が 2 件、1 件、6 件低くなっている。理由は不明である。

12 検索条件式（例として「又」を挙げる）は次の通りである。

キー: (語彙素 = "又" AND 品詞 LIKE "接続詞%") IN (registerName="出版・新聞" AND core="false") OR (registerName="出版・雑誌" AND core="false") OR (registerName="出版・書籍" AND core="false") OR (registerName="図書館・書籍" AND core="false") OR (registerName="特定目的・白書" AND core="false") OR (registerName="特定目的・ベストセラー" AND core="false") OR (registerName="特定目的・知恵袋" AND core="false") OR (registerName="特定目的・ブログ" AND core="fals

で表示された 500 件の内、最初の 100 件を調査対象とした。検索結果の画面表示については、「検索ヒット数が 500 件を超える場合、検索結果からランダムで選ばれた 500 件が表示されます。」（中納言オンライン「マニュアル」更新日：2014-04-02 版）とのことであり、無作為抽出とみなした。

得られた各接続詞の用例 100 件の品詞を、前後の文脈を読み取りながら人手により確認した。副詞など接続詞以外の品詞との判別が特に問題となるものについては、次のような置き換え可能性を目安にして判断した。また、コアデータでの品詞判定も参考にした。判定に迷う場合は接続詞とした。

「又」[13]：「並びに、その上に、又は」に置き換えられるかどうか。「再び、同様に、一方、一体全体・まったく」に置き換えられる場合は副詞。

「更に」：「その上に、それに加えて」に置き換えられるかどうか。「ますます、もっと、少しも（～ない）」に置き換えられる場合は副詞。

「其れから」：「そして」に置き換えられるかどうか。「その時から」に置き換えられる場合は「代名詞＋格助詞」、両方可能な場合は接続詞扱い。

「唯」：「ただし」に置き換えられるかどうか。「単に」に置き換えられる場合は副詞。

「猶」：言い添える内容が続くかどうか。「相変わらず、やはり、一層、ちょうど（のごとし）」に置き換えられる場合は副詞。

「で」：「それで」に置き換えられるかどうか。

「其れでも」：「でも」に置き換えられるかどうか。「でも」に置き換えられず「それで」に置き換えられる場合は「それ」は代名詞。

３.２　高頻度接続詞の適合率の調査結果（語彙素別）

調査結果は、表 1 の通りである。

調査対象 20 語全体の適合率は 93.8%であり、非コアデータ全体の F 値 98%以上よりは低いが、高い適合率になっている。ただし、語ごとに見ると、適合率 90%未満の語が「又」82.0%、「更に」89.0%、「其れから」87.0%、「唯」76.0%、「猶」89.0%、「で」63.0%の 6 語ある。「又、更に、唯、猶」は副詞の誤判定[14]が目立つ。この 4 語には副詞の同形の語彙素がある。「其れから」は代名詞「其れ」との誤解析が目立つ。「で」の適合率は特に低く格助詞及び助動詞の誤判定が目立つ。

このように、語ごとに見た場合、適合率が特に低い語があり、注意が必要である。

---

e") OR (registerName="特定目的・法律" AND core="false") OR (registerName="特定目的・国会会議録" AND core="false") OR (registerName="特定目的・広報誌" AND core="false") OR (registerName="特定目的・教科書" AND core="false") OR (registerName="特定目的・韻文" AND core="false") WITH OPTIONS unit="2" AND tglWords="200" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"

13 「又」の接続詞と副詞の判別の詳細については、4 節参照。

14 本稿では、品詞分類の誤りを「誤判定」と呼び、それ以外の形態素境界の誤りや長単位の構成に関する誤りなどを「誤解析」と呼び、便宜的に呼び分ける。

表 1　高頻度接続詞（サンプル調査）の適合率（語彙素別）[15]

順位	語彙素	コアデータ頻度	非コアデータ頻度	頻度合計	調査件数	接続詞	他品詞等	適合率	他品詞等内訳
1	又	899	85,543	86,442	100	82	18	82.0%	副詞 13、誤解析「又は」5
2	然し	561	68,041	68,602	100	100	0	100.0%	
3	そして	426	62,269	62,695	100	100	0	100.0%	
4	及び	660	48,295	48,955	100	99	1	99.0%	動詞 1
5	でも*	307	36,397	36,704	100	100	0	100.0%	
6	又は*	151	29,560	29,711	100	100	0	100.0%	
7	或いは	106	26,490	26,596	100	98	2	98.0%	副詞 2
8	だから*	172	20,840	21,012	100	100	0	100.0%	
9	更に	275	18,614	18,889	100	89	11	89.0%	副詞 11
10	だが*	177	17,695	17,872	100	100	0	100.0%	
11	其れから*	54	16,570	16,624	100	87	13	87.0%	誤解析(代名詞+格助詞)13
12	唯	159	16,388	16,547	100	76	24	76.0%	副詞 23、誤解析「只松」1
13	然も	106	14,570	14,676	100	100	0	100.0%	
14	猶	89	12,272	12,361	100	89	11	89.0%	副詞 10、誤解析「尚穆王」1
15	但し	80	11,667	11,747	100	99	1	99.0%	誤解析「但一人」1
16	所が*	105	11,295	11,400	100	100	0	100.0%	
17	で	74	10,866	10,940	100	63	37	63.0%	格助詞 18、助動詞 3、誤解析（助動詞）9、誤解析(その他)5、(「て」の)誤字 2
18	即ち	38	10,717	10,755	100	100	0	100.0%	
19	従って*	36	9,900	9,936	100	100	0	100.0%	
20	其れでも*	91	9,807	9,898	100	93	7	93.0%	誤解析(代名詞+格助詞+係助詞)7
	計				2,000	1,875	125	93.8%	

3.3　高頻度接続詞の適合率の調査結果（レジスター別）

同じ調査データを用いレジスター別の適合率を集計した。表 2 に、20 語全体の数値と適合率の低い「又、唯、で」の 3 語の数値を示した。

表 2　高頻度接続詞（非コアデータ、サンプル調査）の適合率（レジスター別）

レジスター	20 語全体		又		唯		で	
	調査件数	適合率	調査件数	適合率	調査件数	適合率	調査件数	適合率
出版・書籍	589	95.4%	32	81.3%	24	75.0%	11	36.4%
出版・雑誌	76	96.1%	1	100.0%	3	66.7%	5	100.0%
出版・新聞	13	100.0%	0	0.0%	1	100.0%	0	0.0%
図書館・書籍	610	94.4%	19	89.5%	25	64.0%	21	76.2%
特定目的・白書	106	94.3%	22	77.3%	0	0.0%	0	0.0%
特定目的・教科書	11	100.0%	3	100.0%	0	0.0%	0	0.0%
特定目的・広報誌	35	97.1%	3	66.7%	0	0.0%	0	0.0%
特定目的・ベストセラー	60	93.3%	0	0.0%	3	66.7%	3	66.7%
特定目的・知恵袋	146	84.9%	6	100.0%	24	83.3%	20	45.0%
特定目的・ブログ	149	86.6%	10	60.0%	9	88.9%	38	68.4%
特定目的・韻文	2	50.0%	0	0.0%	1	0.0%	0	0.0%
特定目的・法律	60	96.7%	0	0.0%	0	0.0%	0	0.0%
特定目的・国会会議録	143	96.5%	4	100.0%	10	90.0%	2	50.0%
計	2000	93.8%	100	82.0%	100	76.0%	100	63.0%

15 「*」を付けた語彙素は、長単位で「連語」の接続詞となる語彙素である。

20 語全体では、調査件数が少ない「特定目的・韻文」を除けば、「特定目的・知恵袋」84.9%及び「特定目的・ブログ」86.6%の適合率が若干低くなってはいるが、全体的にレジスター間で大きな違いは見られない。しかし、（調査件数が少ないレジスターを除くと）「又」では「特定目的・白書」77.3%、「特定目的・ブログ」60.0%、「唯」では「図書館・書籍」64.0%、「で」では「出版・書籍」36.4%、「特定目的・知恵袋」45.0%が特に低くなっており、レジスターの違いによる適合率の大きな違いが見られる。

３.４　詳細な調査の必要性

高頻度接続詞の適合率の調査によって、調査対象 20 語全体の適合率は高いが、語ごとでは適合率の低い語があること、また、20 語全体ではレジスターの違いによる適合率の違いはほぼ見られないが、適合率の低い「又」「唯」「で」ではレジスターによる適合率の違いが見られることが明らかになった。

本節では高頻度接続詞について各 100 語を対象として調査を行ったが、サンプル数が少ないという問題点がある。サンプル数を増やしてより詳細な調査を行う必要がある。本稿では、適合率の低い語のうち「接続詞」使用頻度第 1 位の「又」及び適合率の最も低い「で」について詳細な調査を行う。

４.「又」の詳細調査

４.１　調査の目的と方法

「接続詞」使用頻度第 1 位の「又」に関してより厳密な適合率を明らかにするため、またレジスターによる適合率の違いを詳細に分析するため、「接続詞」及び「副詞」の品詞情報が付与された「又」について調査（以下、「詳細調査」と呼ぶ）を行った。

詳細調査の前に、念のために、形態素解析システムの学習用データとして用いた人手による修正済みのコアデータについて適合率を確認する調査を行った。「中納言」長単位検索で品詞情報を「接続詞」及び「副詞」と指定しコアデータ対象に検索[16]を行い、得られた用例の品詞を前後の文脈を読み取りながら人手により確認した[17]。その結果、「接続

16 検索条件式は次の通りである。「副詞」の検索では「接続詞」の箇所を「副詞」に置き換えた。

キー: (語彙素 = "又" AND 品詞 LIKE "接続詞%") IN (registerName="出版・新聞" AND core="true") OR (registerName="出版・雑誌" AND core="true") OR (registerName="出版・書籍" AND core="true") OR (registerName="特定目的・白書" AND core="true") OR (registerName="特定目的・知恵袋" AND core="true") OR (registerName="特定目的・ブログ" AND core="true") WITH OPTIONS unit="2" AND tglWords="300" AND limitToSelfSentence="0" AND endOfLine="CRLF" AND tglKugiri="" AND encoding="UTF-8" AND tglFixVariable="2"

17 「並びに、その上に、又は」（接続詞）、「再び、同様に（「～もまた」等）、一方（「秋はまた収穫の季節でもある」等）、一体全体・まったく（「どうしてまたそんなことをしたのだ」「またなんときれいな花だ」等）」（副詞）への置き換えを目安に品詞判定を行った。また、コアデータでの品詞判定も参考にした。接続詞と副詞の両方に解釈可能な用例など判定が難しい用例は、付与された品詞情報を正解として処理した。なお、「又貸し」「又聞き」等は全体で名詞とした。「又の名」「又の日」も全体で名詞（小椋、小磯、冨士池、宮内、小西、原(2011)「資料　要注意語」p.20 参照）とした。また、「山また山」「一人また一人」のような同じ名詞を繋ぐ用法は辞書により扱いが異なる。コアデータでは「一羽また一羽と死んでいきました」は接続詞としているが、詳細調査対象の非コアデータでは「足音が一歩、また一歩と大きくなった」「人また人でぎっしり埋まる」は「副詞」と判定されている。今回の調査ではコアデータに従い接続詞として扱う。

詞」の「又」899 件のうち 889 件が接続詞であり適合率 98.9%であった。また、「副詞」の「又」247 件のうち 241 件が副詞であり適合率 97.6%であった。コアデータに関しては 98%前後以上の高い適合率であることが確認された。

非コアデータを対象とした「又」の詳細調査の手順・方法を示す。まず、コアデータと同様に品詞情報を指定し非コアデータ対象に検索[18]を行い、「接続詞」の「又」の用例 85,543 件、「副詞」の「又」の用例 28,756 件を得た。これらの用例に対して、それぞれ層別無作為抽出（レジスターの 1 層）を行い、「接続詞」「副詞」各 1000 例を調査対象の用例として、前後の文脈を読み取りながら人手により品詞を確認した。なお、「接続詞」及び「副詞」の用例の抽出率は、それぞれ 1.17%、3.48%である。

４.２　「又」詳細調査での適合率の結果及び誤判定の要因

「又」の詳細調査による品詞判定の結果を表 3 に示す。

表 3　「又」詳細調査（非コアデータ）での適合率

品詞情報	人手による品詞判定				計	適合率
	接続詞	副詞	誤解析	誤字		
「接続詞」	858	117	25	0	1000	85.8%
「副詞」	160	828	11	1	1000	82.8%
計	1018	945	36	1		

「接続詞」の「又」1000 件のうち 858 件が接続詞であり適合率 85.8%であった。接続詞以外は、副詞の誤判定 117 件、誤解析 25 件（「又は」23 件、「またぐ」「三ツ又沼」各 1 件）であった。「副詞」の「又」1000 件のうち 828 件が副詞であり適合率 82.8%であった。副詞以外は、接続詞の誤判定 160 件、誤解析 11 件（「又の名」3 件、「俟つ」2 件、「尾亦、胡亦堂、興復、又七郎、又左、股」各 1 件）、誤字「復雑」（複雑）1 件であった。

「接続詞」の「又」に関しては 3 節での 100 例サンプル調査での適合率 82.0%に比べると若干高くなってはいるが、それでも 90%を下回っている。品詞情報を利用する際に十分留意する必要がある。

ただし、「接続詞」の「又」の正解 858 件と「副詞」の「又」のうち接続詞の用例 160 件とを合わせると 1,018 件となる。少なくとも「又」は、仮に「接続詞」1000 件の数値をそのまま利用したとしても大きな違いが生じないという見方もできるかもしれない[19]。

誤判定の起こる要因は断定できないが、読点（「、」及び「，」）の直後の「又」の誤判定が目立った。直前 1 文字別の適合率（調査件数 6 件以上のみ）を表 4 に示す。

表 4 の通り、「接続詞」「副詞」各 1000 件の用例のうち、ともにほぼ 4 分の 1 の用例が読点の直後の用例である。「、」の直後の「接続詞」の適合率は 73.1%であり、「，」及び「、」の直後の「副詞」の適合率はそれぞれ 21.1%、58.7%であり極めて低い。また、「接続詞」全体の副詞の誤判定 117 件のうち読点の直後の用例は 55 件(47.0%)であり、「副詞」全体の接続詞の誤判定 160 件のうち読点の直後の用例は 114 件(71.3%)であり、誤

[18] 検索条件式は、非コアデータを指定した以外は、注 13 と同様である。

[19] ただし、4.3 に示すようにレジスター別では大きな違いが生じる場合がある。特に「特定目的・ブログ」では、「接続詞」の「又」には副詞が 5 割以上含まれるのに対し「副詞」の「又」には接続詞が 144 例中 3 例あるのみであり、「接続詞」の「又」の使用頻度をそのまま用いるのは危険である。

判定の多くは読点の直後である。このように、読点の直後での誤判定の多さが、全体の適合率を下げる一つの大きな要因となっていると見られる[20]。

表 4　「又」詳細調査（非コアデータ）での直前 1 文字別適合率(調査件数 6 件以上のみ)

「接続詞」		
直前 1 文字	調査件数	適合率
、	208	73.1%
は	23	87.0%
(全角スペース)	190	90.0%
て	12	91.7%
.	19	94.7%
。	412	97.3%
,	41	97.6%
？	28	100.0%
全体	1000	85.8%

「副詞」		
直前 1 文字	調査件数	適合率
,	19	21.1%
、	242	58.7%
に	43	81.4%
て	31	83.9%
の	7	85.7%
ら	40	87.5%
を	17	88.2%
は	126	92.1%
「	17	94.1%
で	41	95.1%
が	59	96.6%
も	202	98.5%
。	9	100.0%
と	18	100.0%
ば	7	100.0%
れ	51	100.0%
全体	1000	82.8%

４.３　「又」詳細調査での適合率の結果（レジスター別）

同じ調査データを用いレジスター別の適合率を集計した（表 5 参照）。

表 5　「又」詳細調査（非コアデータ）での適合率（レジスター別）

レジスター	「接続詞」		「副詞」	
	調査件数	適合率	調査件数	適合率
出版・書籍	274	91.2%	257	77.8%
出版・雑誌	27	96.3%	25	92.0%
出版・新聞	5	100.0%	3	66.7%
図書館・書籍	236	83.9%	356	82.9%
特定目的・白書	161	86.3%	4	25.0%
特定目的・教科書	17	94.1%	3	33.3%
特定目的・広報誌	36	97.2%	3	66.7%
特定目的・ベストセラー	22	77.3%	51	86.3%
特定目的・知恵袋	86	89.5%	67	92.5%
特定目的・ブログ	66	42.4%	144	97.2%
特定目的・韻文	1	0.0%	4	100.0%
特定目的・法律	0		0	
特定目的・国会会議録	69	97.1%	83	65.1%
計	1000	85.8%	1000	82.8%

[20] コアデータの読点の直後の用例のみを取り出してみると、「接続詞」全 120 件中 4 件が副詞であり（適合率 96.7%）、「副詞」全 14 件中 1 件が誤解析（名詞「又の名」）であった（適合率 92.9%）。

レジスター別（調査件数 10 以下のレジスターは除く）に見ると、「接続詞」の「又」では、「特定目的・ブログ」42.4%（特に「、」の直後全 14 件の適合率 14.3%）、「特定目的・ベストセラー」77.3%（特に「、」の直後全 9 件の適合率 44.4%）が特に適合率が低い。「副詞」の「又」では、「特定目的・国会会議録」65.1%（特に「、」の直後全 20 件の適合率 10.0%）が特に適合率が低い[21]。

レジスター別の使用頻度に基づいた接続詞の分析を行う際には、適合率が低いレジスターがあることを十分に考慮する必要がある。

５．「で」の詳細調査

サンプル調査で適合率が最も低かった「で」に関しても、「又」と同様の方法で詳細調査（「接続詞」「格助詞」「助動詞」各 1000 件）を行っている途中である（表 6 参照）[22]。現段階（各 200 件の途中経過）では、「接続詞」に関しては、適合率が 62.5%と低く、レジスター別では「特定目的・知恵袋」44.1%が特に低くなっている。また、格助詞や助動詞の誤判定や誤解析は「で」の直前が空白（全角スペース）や記号類（、）」等）の場合、数式などを削除している場合、文頭の「であるから、でないから」等の場合に目立つ。

表 6　「で」詳細調査（非コアデータ）での適合率（途中経過）

品詞情報	人手による品詞判定						計	適合率
	接続詞	格助詞	助動詞	接続助詞	誤解析	誤字		
「接続詞」	125	31	13	1	29	1	200	62.5%
「格助詞」	0	182	15	0	3	0	200	91.0%
「助動詞」	0	46	139	0	15	0	200	69.5%
計	125	259	167	1	47	1		

６．まとめ

BCCWJ を利用した接続詞研究が増えている。接続詞研究において BCCWJ の品詞情報を利用する際の留意点を示すために、本稿では、BCCWJ で「接続詞」の品詞情報が付与された語（長単位）の解析精度の調査（非コアデータ対象）を行い、以下の結果を報告した。

① 高頻度接続詞 20 語全体の適合率は 93.8%であり、非コアデータ全体（全品詞）に比べると低いが、高い適合率になっている。しかし、語ごとに見ると、適合率は 63.0%～100.0%の範囲にあり適合率の低い語がある。適合率 90%未満の語は、「又」82.0%、「更に」89.0%、「其れから」87.0%、「唯」76.0%、「猶」89.0%、「で」63.0%の 6 語である。「又、更に、唯、猶」は副詞の誤判定が目立つ。

② 高頻度接続詞 20 語全体では、レジスターの違いによる適合率の違いはほぼ見られない。しかし、適合率の低い「又」「唯」「で」では、レジスターによる適合率の違いが見られる。

③ 「又」の詳細調査の結果、適合率は「接続詞」85.8%、「副詞」82.8%である。レ

21 「特定目的・ブログ」「特定目的・国会会議録」で適合率が特に低くなったのは、行動の叙述（時間的）、並列的な事柄の提示（非時間的）というそれぞれの内容的な特徴も関わっていると思われる。

22 「で」のコアデータの適合率は「接続詞」90.5%、「格助詞」97.0%、「助動詞」99.0%である。「接続詞」は全 74 件、「格助詞」「助動詞」は検索結果画面に表示された最初の各 100 件を対象とした。

ジスター別では「接続詞」の「特定目的・ブログ」42.4%、「副詞」の「特定目的・国会会議録」65.1%が特に低い。読点の直後の「又」の誤判定が多く、全体の適合率を下げる大きな要因となっていると見られる。

④　「で」の詳細調査の結果（ただし途中経過）、「接続詞」の適合率は 62.5%であり、レジスター別では「特定目的・知恵袋」44.1%が特に低い。

接続詞研究では、従来、コーパス検索の際、多くは文字列検索が行なわれ、また、効率的に検索するために、文頭に限定したり読点が後続する場合に限定したりすることも多かった。今後の研究において、BCCWJ での品詞情報が利用できることは極めて有益なことである。接続詞全体での品詞情報の解析精度はコーパス全体（全品詞）よりも若干劣るが、接続詞全体として他品詞と比較する場合には大きな問題は生じないであろう。しかし、異なり語の少ない接続詞内部で個々の語（語群）を分析する場合には、品詞情報の解析精度の違いが問題となる。もちろん、BCCWJ の品詞情報を利用する際には、研究の目的や方法に応じて人手による点検が不可欠である。しかし、検索結果をそのまま利用する場合では、特に分析対象とする語の解析精度の違いを十分把握しておく必要がある。

今後は、誤判定、誤解析の要因を明らかにし解析精度の向上を図ることが期待される。本稿の結果は品詞情報付与に関する解析器改良のための参考資料を提供するものでもある。

文　献

小木曽智信(2014)「第 5 章　形態素解析」山崎誠(編)『講座日本語コーパス　2. 書き言葉コーパス―設計と構築―』朝倉書店, pp.89-115.

小木曽智信、小椋秀樹、小磯花絵、宮内佐夜香、渡部涼子、伝康晴(2010)「形態素解析辞書のベンチマークテスト―IPAdic・NAIST-jdic・UniDic のジャンル別精度比較―」, 言語処理学会第 16 回年次大会発表論文集, pp.326-329.

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　（下）』国立国語研究所.

小椋秀樹、冨士池優美(2011)「第 4 章　形態論情報」, 『現代日本語書き言葉均衡コーパス』利用の手引　第 1.0 版, pp.39-73.

鯨井綾希(2011)「主成分分析を用いた文章構造の特徴抽出――品詞構成の変動に注目した分析――」, 文芸研究, 172, pp.59-48.

国立国語研究所コーパス開発センター(2011)『『現代日本語書き言葉均衡コーパス』利用の手引　第 1.0 版』国立国語研究所コーパス開発センター.

冨士池優美、小西光、小椋秀樹、小木曽智信、小磯花絵(2011)「長単位に基づく『現代日本語書き言葉均衡コーパス』の品詞比率に関する分析」, 言語処理学会第 17 回年次大会発表論文集, pp.663-666.

関連 URL

国立国語研究所コーパス開発センター(2013)『『現代日本語書き言葉均衡コーパス』マニュアル　第 1.1 版（Web 公開用）』国立国語研究所コーパス開発センター.　http://www.ninjal.ac.jp/corpus_center/bccwj/doc/manual/BCCWJ_Manual.zip

「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/

「『現代日本語書き言葉均衡コーパス』長単位語彙表 ver1.0」　http://www.ninjal.ac.jp/corpus_center/bccwj/freq-list.html

『太陽コーパス』における語彙素「あう」の用字法

髙橋　雄太（明治大学大学院国際日本学研究科）

Character Usage of the Japanese Verb “AU” in Taiyo Corpus

Yuta Takahashi (Graduate School of Global Japanese Studies, Meiji University)

要旨

語義と表記の固定が進んでいなかった明治大正時代を対象とする『太陽コーパス』を用いて、動詞「あう」に対する表記の実態と変遷を調査する。『太陽コーパス』では語彙素「会う」に対する表記としては「會」「逢」「遇」「遭」が存在するが、1895 年では現代よりも自由に表記がなされていた。さらに、「近代文語 UniDic」では語彙素認定において「合う」と「会う」に二分しているが、用例を見るとこの二つの語彙素間でも表記の通用が確認できる。本研究では語彙素「会う」と「合う」を一つの語彙素「あう」として頻度を集計し、主要な表記「會」「逢」「遇」「遭」「合」を、用例分析をした上で動作対象を分類し、明治大正期の書き分けの実態と変遷を明らかにする。また、用例分析の結果によって判明した明治大正時代の用字法と、現代語の用字法や国語政策との関連も考察する。

1. はじめに

近代においては、現代語と比較して自由に表記をしており、一つの語に対して表記が複数ある「同訓異字」が、明治大正期では現代語よりも多かった。近代語の同訓異字の研究では、京極(1998)や田島(1998)、コーパスを用いた研究では田中(2006)など、個々の語における成果が報告された。しかしながら、これら近代語の用字法の研究は数が少なく、特に資料が膨大な近代の研究に有効なコーパスはあまり活用されていない状況にある。

そこで本稿では、近代語の用字法の一つとして、『太陽コーパス』を用い、同訓異字を持つ語彙素「あう」における用字法について考えていきたい。

2. 調査

今回の調査では対象として、経年的な観察が可能な『太陽コーパス』を用いる。『太陽コーパス』に含まれる 1895 年、1901 年、1909 年、1917 年、1925 年の 5 年分のデータに「近代文語 UniDic」による形態素解析を施し、各年の表記別の頻度表を作成する。対象とする語は動詞「あう」で、「近代文語 UniDic」では「合う」と「会う」を別語彙素として認定している[1]が、これらの間でも表記に通用が見られるため語彙素「あう」として括り集計をする。また、今回の調査では、「合わす」「合わせる」のような「あう」とは別語彙素に認定された語彙素[2]、及び補助動詞用法の「あう」は全て対象外とする。

[1] 小椋ほか(2011)では「UniDic」での語彙素の認定において、「会う」「遭う」「逢う」などは「に」が前接する点で共通していることから一つの「会う」という語彙素に認定し、「合う」と区別したとしている。「近代文語 UniDic」もこれに準じていると思われる。

[2] 特に「合わせる」については、「并」「併」など別表記が関係するため調査結果が複雑化したため、調査対象から外した。

2.1　調査の前に

以下の図 1 は、「あう」の各表記の年次別表記頻度を示したもので、各表記の頻度数の増減を知ることができる。なお、平仮名表記や頻度数が 10 以下の表記については対象外とした。

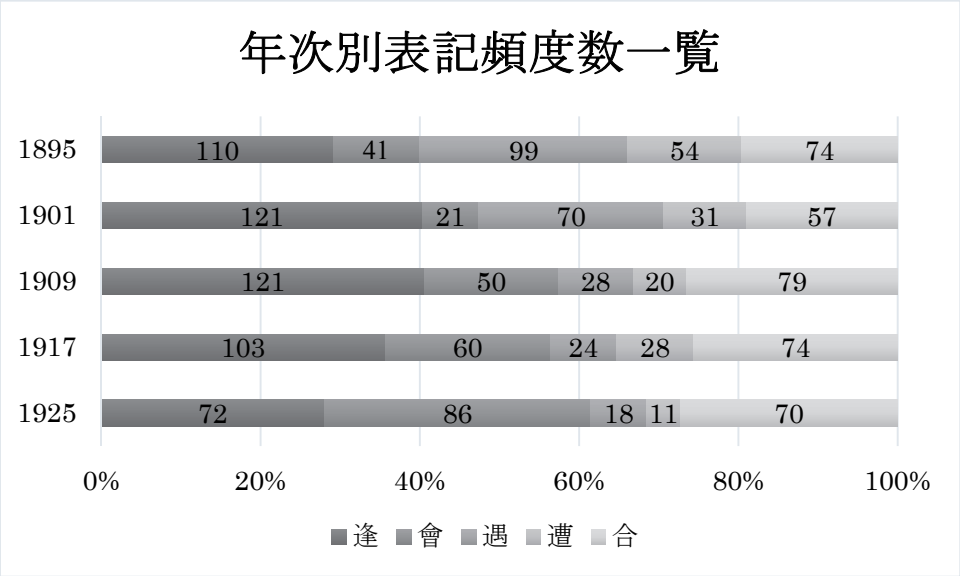


図 1　動詞「あう」の主要表記の年別頻度表

代表表記となるのは「會（会の旧字体）」であるが、5 年分の頻度数、及び 1895、1901、1909、1917 年の頻度数では「逢」が「會」を上回っている。1895 年から 1917 年まで、「逢」は大きな減少もなく最大の頻度数であったことから、動詞「あう」の表記としては「逢」が一般的であったことが分かる。その他の表記も含めて見ると、「逢」「遭」「遇」が減少しているのに対し、「會」のみ頻度数が徐々に増えていき、1925 年では「逢」と逆転している。一方で「合」は増減の幅が最も狭く、一定量使用され続けていることが分かる。

しかしながら、このような実態にある背景を考えるには、実際の用例を観察し、どの表記がどの用法と結びつくかを確認しなければならない。2.2 では動詞「あう」の対象語を分類し、各表記の性質を探る。

2.2　用例分析による動作対象の分類と統計

2.1 で述べた、動詞「あう」の対象語を分類したものが表 1 である。大分類の「人・もの」には物理的に相対することのできる対象語を、「イベント・環境」には世間や自分に起こった出来事や自身を囲む状況を表す対象語、「合用法」には現代語において通常「合」で表記する用法の対象語をそれぞれ分類した。

また、調査対象には以下の(1)、(2)にあるような「に格」に加えて、

(1)…私は後でどんな目に逢つて居るか分らぬ…　（1909 年「仏国に於ける寄宿舎生活」）

(2)…白川を固めて居つた伊治地正治に會ひまして…　（1901 年「追懐談」）

「と格」、数は少ないが「が格」や「を格」、明記していないが対象語が文脈から読み取れるもの、連体修飾節に含まれる「あう」も全て含んだ。

表 1　動詞「あう」の対象語の分類

大分類	小分類	分類基準	用例
人・もの	人・生物	一般的な人、生物。幽霊や仏も含む	母、子供、男、誰か、盗人、先生、召使、韓人、教徒、提督、大徳、幽霊、熊、獲物、蛇など
	恋人	恋人に限った人	愛人、女、二人、あなた、二つの星、など
	物体	無生物の物体	氷塊、石、船体、樹、難破船、緑林、城郭、など
イベント・環境	出来事	身の回りや世間の出来事	変化、開業、政変、故障、大赦、禁輸、質問、検査、鞭撻、神隠、ストライキ、批判、抗議など
	状況	その場全体・状況、「動詞+あう」含む	板挟み、惨状、危険、境遇、この世、逆境、来たりすぎるに、～起こるに、困難、難局、など
	戦闘・攻撃	動作により身体が傷つく行為	攻撃、砲撃、殺戮、殺害、夜討、～の変、襲撃、大戦争、不意打、乱、虐待、強盗、処刑、など
	精神・心理	～目とあるもの、または抽象的な心理的被害	酷い目、悲しい目、憂い目、好い目、苦しみ、半死半生、禍、不幸、災難、栄典、幸運など
	時期	特定の時期	正月、春、聖代、めでたい日、秋の時、時勢、など
	自然	自然災害や雨天など	俄雨、嵐、雨、暴風、旱、地震、厳寒、晴天、など
合用法	―	現代語で通常「合」で表記する用法	趣旨、時勢、意見、尺、理屈、気、辻褄、思想、性格、歩調、調子、など

(1)の例ならば「どんな目に」とあるため大分類は「イベント・環境」に、小分類は「精神・心理」に分類する。(2)ならば「伊治地正治に」とあるため大分類は「人・もの」に、小分類は「人・生物」に分類する。

対象が似ている「人・生物」と「恋人」の分類の基準は、キーの前文脈と後文脈 50 文字ずつを読んだ上で、明確に動作主と被動作主が恋人の関係にあるもののみを「恋人」に、どちらとも言えない用例は全て「人・生物」に分類をしている。

では、はじめに、表記毎の分類別の比率を示した表 2 を見る。

表 2　語彙素「あう」の表記別分類の比率

「逢」499	人・もの	338	67.7%	人・生物	304	60.9%
				恋人	27	5.4%
				物体	7	1.4%
	イベント・環境	151	30.5%	出来事	42	8.4%
				状況	21	4.2%
				戦闘・攻撃	22	4.4%
				精神・心理	53	10.6%
				時期	3	0.6%
				自然	11	2.2%
	合用法	10	2.0%	合用法	10	2.0%
「會」289	人・もの	218	75.4%	人・生物	214	74.1%
				恋人	2	0.7%
				物体	2	0.7%
	イベント・環境	68	23.5%	出来事	27	9.3%
				状況	10	3.5%
				戦闘・攻撃	4	1.4%
				精神・心理	15	5.2%
				時期	4	1.4%
				自然	8	2.8%
	合用法	3	1.0%	合用法	3	1.0%
「遇」266	人・もの	109	48.2%	人・生物	104	46.0%
				恋人	3	1.3%
				物体	2	0.9%
	イベント・環境	112	49.6%	出来事	34	15.0%
				状況	24	10.6%
				戦闘・攻撃	11	4.9%
				精神・心理	28	12.4%
				時期	4	1.8%
				自然	11	4.9%
	合用法	5	2.2%	合用法	5	2.2%
「遭」133	人・もの	17	12.8%	人・生物	13	9.8%
				恋人	1	0.8%
				物体	3	2.3%
	イベント・環境	115	86.5%	出来事	31	23.3%
				状況	13	9.8%
				戦闘・攻撃	16	12.0%
				精神・心理	40	30.1%
				時期	1	0.8%
				自然	15	11.3%
	合用法	1	0.8%	合用法	1	0.8%
「合」189	人・もの	3	1.6%	人・生物	3	1.6%
				恋人	0	0.0%
				物体	0	0.0%
	イベント・環境	4	2.1%	出来事	0	0.0%
				状況	2	1.6%
				戦闘・攻撃	0	0.0%
				精神・心理	1	0.5%
				時期	0	0.0%
				自然	1	0.5%
	合用法	182	96.3%	合用法	182	96.3%

それぞれ左には大分類、右には小分類を示し、各表記においてそれぞれの用法がどれほどの比率で使用されているかを示している。大分類をみると、「逢」や「會」は「人・もの」に「あう」ときに主に使用され、逆に「遭」は「イベント・環境」の用法で用いられやすいことが分かる。「遇」は「人・もの」「イベント・環境」のどちらにも等しく使用されている。「合」に関しては、若干の揺れがあるものの、「合用法」に分類される用例が約96%であり、明治時代・大正時代の時点で「人・もの」や「イベント・環境」で「合」を用いることがほぼ無かったことが分かる。

小分類でも同様に、「會」の「人・生物」用法への偏りが特徴的である。同様に「人・生物」の比重の大きい「逢」と比較しても、「人・生物」の比率が13%程度上回っている。これは「逢」が「人・生物」以外の用法でも頻度が高いことが原因と考えられ、「逢」はどの用法でも適切度が高かったことが言える。「遭」や「遇」に関しては、「時期」などの一部の例外を除いては、「イベント・環境」に属する小分類はほぼ全て高い比率である。

次に、語彙素「あう」の対象語別に各表記の頻度と比率をまとめると、表3になる。

表3　語彙素「あう」の対象語別の表記

大分類	小分類	「逢」		「會」		「遇」	
人・もの	人・生物	338(49.3%)	304(47.7%)	218(31.8%)	214(33.5%)	109(15.9%)	104(16.3%)
	恋人		27(81.8%)		2(6.1%)		2(9.1%)
	物体		7(50.0%)		2(14.3%)		2(14.3%)
イベント・環境	出来事	151(33.7%)	42(31.3%)	68(15.1%)	27(20.2%)	112(24.8%)	34(25.4%)
	状況		21(30.0)%		10(14.3%)		24(34.3%)
	戦闘・攻撃		22(41.5%)		4(7.6%)		11(20.8%)
	精神・心理		53(38.7%)		15(11.0%)		28(20.4%)
	時期		3(25.0%)		4(33.3%)		4(33.3%)
	自然		11(23.9%)		8(17.4%)		11(23.9%)
合用法		10(5.0%)		3(1.5%)		5(2.5%)	
全体		499(36.3%)		289(21.0%)		266(19.3%)	
大分類	小分類	「遭」		「合」		合計	
人・もの	人・生物	17(2.5%)	13(2.0%)	3(0.4%)	3(0.5%)	685	638
	恋人		1(3.0%)		0(0.0%)		32
	物体		3(21.4%)		0(0.0%)		14
イベント・環境	出来事	115(25.5%)	31(23.1%)	4(0.9%)	0(0.0%)	450	134
	状況		13(18.5%)		2(2.9%)		60
	戦闘・攻撃		16(30.2%)		0(0.0%)		53
	精神・心理		40(29.2%)		1(0.7%)		137
	時期		1(8.3%)		0(0.0%)		12
	自然		15(32.6%)		2(2.2%)		47
合用法		1(0.5%)		182(90.6%)		201	
全体		133(9.7%)		189(13.7%)		1376	

それぞれの表記の最下欄には、「表記毎の総頻度数」の「全表記の総頻度数」に対する比率が示してある。これを各表記の平均的な比率として、この数値を上回る分類については、その分類と表記が強く結びついていることを示す。

例えば小分類「恋人」における「逢」の表記は平均の36.3%を大きく上回り、81.8%にまで達している。ここから、「恋人」用法には基本的に「逢」が用いられていたことが言える。その他、「會」における「人・生物」や、「遇」における「状況」、「遭」における「戦闘・攻撃」「精神・心理」「自然」「出来事」「状況」が平均を大きく上回っている。

大分類では、「人・もの」は「逢」と「會」を合わせて 8 割を超えており、「人・もの」用法での「あう」には、基本的に「逢」か「會」が用いられていることになる。「イベント・環境」については、「逢」の総頻度数が 499、「遭」の総頻度数が 133 という違いのため「逢」が占める比率が大きくなっているが、「逢」自体は「イベント・環境」用法よりも「人・もの」に多く使用されるため、見た目の数値以上に、「遭」や「遇」の「イベント・環境」における比率は高いと言える。また、「合用法」については、「合用法」の内 9 割が「合」ので表記されていることからも、明治大正時代には「合用法」は書き分けがなされていたと言える。

2.3 「あう」の表記の変化

2.2 では、『太陽コーパス』全体の「あう」の用字法を分析したが、ここからは 1895 年から 1925 年にかけての推移を分析する。

以下の表 4 は、1895、1901、1909、1917、1925 年における各表記の大分類毎の頻度と比率を表した数値である。

表 4　表記別の対象語の比率の推移

「逢」	1895		1901		1909		1917		1925	
人・もの	53	55.8%	63	57.3%	83	69.2%	86	81.1%	53	76.8%
イベント・環境	37	38.9%	45	40.9%	36	30.0%	18	16.9%	16	23.2%
合用法	5	5.3%	2	1.8%	1	0.8%	2	1.9%	0	0.0%
「會」	1895		1901		1909		1917		1925	
人・もの	24	55.8%	14	63.6%	29	61.2%	57	82.6%	94	86.2%
イベント・環境	17	39.5%	8	36.4%	18	38.3%	12	17.4%	13	11.9%
合用法	1	2.4%	0	0.0%	0	16.1%	0	0.0%	2	1.8%
「遇」	1895		1901		1909		1917		1925	
人・もの	29	34.9%	41	57.7%	10	40.0%	18	56.3%	11	68.8%
イベント・環境	51	61.4%	27	38.0%	15	60.0%	14	43.8%	5	31.3%
合用法	3	3.6%	3	4.2%	0	0.0%	0	0.0%	0	0.0%
「遭」	1895		1901		1909		1917		1925	
人・もの	7	13.7%	2	8.0%	4	21.1%	1	3.6%	3	30.0%
イベント・環境	44	86.3%	23	92.0%	15	78.9%	27	96.4%	7	70.0%
合用法	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	10.0%
「合」	1895		1901		1909		1917		1925	
人・もの	0	0.0%	0	0.0%	2	4.7%	0	0.0%	1	2.3%
イベント・環境	1	3.1%	0	0.0%	1	2.3%	1	2.6%	0	0.0%
合用法	31	96.9%	31	100.0%	40	93.0%	38	97.4%	42	97.7%

「合」と「遭」には大きな変化はないものの、「逢」や「會」などは、1895 年や 1901 年ではあらゆる用法で使用されていたが、後年になると「人・もの」に使用が限定されてくる動きが確認できる。「遇」については年によって比率がばらついており、変化の流れを捉えることができないことから、用法が定まっていないことが考えられる。

次に、対象語別に表記の推移を表にすると、表 5 のようになる。

表 5　対象語別の表記の比率の推移

人・もの	1895		1901		1909		1917		1925	
「逢」	53	46.9%	63	52.5%	83	64.8%	86	53.1%	53	32.7%
「會」	24	21.2%	14	11.7%	29	22.7%	57	35.2%	94	58.0%
「遇」	29	25.7%	41	34.2%	10	7.8%	18	11.1%	11	6.8%
「遭」	7	6.2%	2	1.7%	4	3.1%	1	0.6%	3	1.9%
「合」	0	0.0%	0	0.0%	2	1.6%	0	0.0%	1	0.6%
イベント・環境	1895		1901		1909		1917		1925	
「逢」	37	24.7%	45	43.7%	36	42.4%	18	25.0%	16	38.1%
「會」	17	11.3%	8	7.8%	18	21.2%	12	16.7%	13	31.0%
「遇」	51	34.0%	27	26.2%	15	17.6%	14	19.4%	5	11.9%
「遭」	44	29.3%	23	22.3%	15	17.6%	27	37.5%	7	16.7%
「合」	1	0.7%	0	0.0%	1	1.2%	1	1.4%	1	2.4%
合用法	1895		1901		1909		1917		1925	
「逢」	5	12.2%	2	5.6%	1	2.4%	2	5.0%	0	0.0%
「會」	1	2.4%	0	0.0%	0	0.0%	0	0.0%	2	4.4%
「遇」	3	7.3%	3	8.3%	0	0.0%	0	0.0%	0	0.0%
「遭」	0	0.0%	0	0.0%	0	0.0%	0	0.0%	1	2.2%
「合」	32	78.0%	31	86.1%	40	97.6%	38	95.0%	42	93.3%

「人・もの」では、1901 年までは「會」よりも「遇」の占める比率が大きかったが、1909 年以降徐々に「會」の使用が増えていき、1925 年では 90%以上が「逢」もしくは「會」で表記されていることが分かる。

「イベント・環境」では、1895 年時点でも既に「遭」や「遇」の比率が大きいことが言えるが、1917 年で「遭」の頻度が全ての表記を上回っていることは特筆すべき点である。なお、1925 年は「會」や「逢」が高い比率になっているが、1925 年は「遭」や「遇」の頻度がそれぞれ 10 例・16 例と極端に少ないことが比率に影響しているため、参考にしない。

「合用法」では、いずれの年もほぼ全てが「合」で表記されているが、1895 年と 1909 年以降を比較すると、1909 年以降は、より厳密に書き分けがなされていたことが言える。

また、表 5 からは、各用法の頻度数の推移も知ることができる。表 5 を表記に関係なく集計したものが、表 6 になる。

表 6　対象語の頻度の推移

	1895		1901		1909		1917		1925		全体	
人・もの	113	37.1%	120	46.3%	128	50.2%	162	59.8%	162	65.1%	685	51.2%
イベント・環境	150	49.5%	103	39.8%	86	33.7%	69	25.5%	42	16.9%	450	33.7%
合用法	41	13.4%	36	13.9%	41	16.1%	40	14.8%	45	18.1%	203	15.2%

1895 年の時点では、頻度数では「イベント・環境」用法が最も多いが、1901 年以降は、「人・もの」用法が占める比率が徐々に大きくなっていることが分かる。「合用法」は増

減がほとんどなく、1895 年から 1925 年まで 15%前後を保っている。図 1 で、「近代文語 UniDic」による表記の頻度数の推移を示したが、「遭」や「遇」が徐々に数が減っている背景には、「遭」や「遇」と結びつきの強い「イベント・環境」用法の衰退があることが予想される。また、図 1 で頻度が後年になるほど高くなっていた「會」は、「人・もの」用法と結びつきが強いために増加していったと考えられる。

3. 国語政策と現代語における「あう」の表記について

明治大正時代の後、昭和に入ると国の政策として使用漢字やその読みに制限を与えようという方針が立てられ、揺れがあった語の表記は徐々に統一されていった。1946 年国語審議会の答申で当用漢字表が、1948 年には当用漢字音訓表が発表され、その後の公的文書や教科書、新聞などを中心に用字法が整備された。

語彙素「あう」についてどうであったかというと、当用漢字表に登録のある字は「合」「会」「遭」「遇」の 4 字で、「逢」の字はない。うち、「アウ」の音を持つのは「合」「遭」「会」の 3 字であり、「遇」は「アウ」とは読ませないとしている。これは常用漢字表でも継続されており、未だに「逢」は使用されず、「遇」は「アウ」の音を持たない。

ここで、前節の 2 における、表記と意味の結びつきと関連付けて考察をすると、対象語で分けた大分類の「人・もの」「イベント・環境」「合用法」のそれぞれの用法で、優先的に使用された「會」「遭」「合」の 3 字が当用漢字表に登録され、また「アウ」の音を持つようになったのである。このことから、当用漢字表を定める上で、それ以前に既に各用法に対する書き分けが確立されていたことが推測できる。

一方、「逢」の字は現代人ならば「アウ」と読むことが一般的に可能であるにも関わらず、常用漢字には追加されていない。これについては、『太陽コーパス』において「恋人」用法で「逢」がほぼ独占的に使用されていた状況を鑑みると、文学や歌詞など、表記に自由が利く環境で使用され続け、主に「恋人」用法を中心に現代語においても書き分けがされているのではないかと考えられる。

4. おわりに

今回は、『太陽コーパス』を用いて、動詞「あう」の表記について実態と変遷を追うことで、「合」とその他の表記が明治大正時代の時点で書き分けされていることや、表記と語義が段々に固定されていく過程を確認することはできた。しかしながら「アウ」と読む「遇」の消滅や、一度使用頻度の下がった「遭」が何故現代語で書き分けられているのかなど、明らかになっていない点もいくつか残った。

『太陽』以降の昭和時代の用字法、及び、他の語でも、同じ方法で似たような表記の現象が確認できるかの調査などが、今後の課題となるだろう。

文　献

小椋秀樹、小磯花絵、冨士池優美ほか(2011)『『現代日本語書き言葉均衡コーパス』形態論情報規定集　第 4 版　(下)』国立国語研究所

京極興一(1998)『近代日本語の研究－表記と表現―』東宛社

田島優(1998)『近代漢字表記語の研究』和泉書房

田中牧郎(2006)「「努力する」の定着と「つとめる」の意味変化―『太陽コーパス』を用いて―」倉島節尚編『日本語辞書学の構築』おうふう

『国民之友コーパス』に現れる一人称代名詞の計量的分析

近藤 明日子（国立国語研究所コーパス開発センター）†

A Quantitative Analysis of First-Person Pronouns in *Kokuminnotomo Corpus*

KONDO Asuko (National Institute for Japanese Language and Linguistics)

要旨

雑誌『国民之友』1887～1888 年刊行分をコーパス化した『国民之友コーパス』に出現する一人称代名詞の計量的分析を行った。まず、分析の前にコーパスの言語量から資料性の検討を行い、非文学の文語文が大部分を占める資料であることを確認した。次に、非文学・非翻訳記事の文語地の文を対象資料として、一人称代名詞を抽出し、各語形の頻度を集計した。そこから、「吾人」が他の語形と比較して特に高頻度に出現することが本コーパスの特徴であり、それは無署名記事での「吾人」専用とも言える実態に起因することが分かった。また、記事単位での複数語形の共起について、特に「吾人」「余」「余輩」の関係を分析し、共起の組み合わせごとに頻度上の主従関係や用法が異なることも明らかになった。

１．はじめに

近代日本語の一人称代名詞には現代語以上に種々の語形があり、語形の消長過程や語形間の用法差の解明に研究の焦点があてられてきた。その範囲は、小説・戯曲の会話部分、落語速記、口語文典などの話し言葉的性質の強い口語文を利用して当時の話し言葉での実態を明らかにする研究（岡田 1998・房 2004・祁 2006a・祁 2006b など）にはじまり、近代雑誌のコーパスを利用して書き言葉的性質の強い文章での実態を明らかにする研究（近藤 2012・2013a・2013b）へと広がりを見せている。

本稿では､2014 年 9 月に公開された新たな近代雑誌コーパスである国立国語研究所(2014)『国民之友コーパス』Ver.1.0 を利用し、そこに出現する一人称代名詞の計量的分析を試みる。『国民之友コーパス』は、雑誌『国民之友』の 1887（明治 20）～1888（明治 21）年刊行分である 1～36 号の全文をコーパス化したものである。原資料である雑誌『国民之友』は、徳富蘇峰の設立した民友社により 1887（明治 20）年から 1898（明治 31）年にかけて刊行された。主に、徳富蘇峰ら民友社社員および当時の著名知識人による政治・社会・経済・文学等の評論や文学作品を掲載する（近藤 2014、p.1)。本稿では、まずコーパスの言語量からコーパスの資料性について検討し、次に、コーパスから一人称代名詞を抽出、計量的に分析する。特に、論説・評論等の非文学かつ非翻訳記事の文語地の文に出現する一人称代名詞に注目し、記事署名の有無との対応関係や記事中での共起関係に焦点をあて、近代語の一人称代名詞の実態の一部を明らかにすることを試みる。

† kondo@ninjal.ac.jp

２．言語量から見る『国民之友コーパス』の資料性

２．１．コーパス要素別の言語量

最初に、コーパスの XML ファイルに付加された情報[1]に基づき、いくつかの観点からコーパスの言語量を計り、コーパスの資料性について概観する。まず、コーパスは記事要素（article 要素）と非記事要素（titleBlock 要素）に大きく分けることができる。それぞれの延べ語数（記号類・非日本語部分を除く）と記事数（article 要素数）を表 1 に示す。

表 1 コーパス全体の言語量（コーパス要素別）

	記事要素	非記事要素	コーパス全体
延べ語数	1005578	1402	1006980
記事数	1250	—	1256

コーパス全体の記事数は 1256 であるが、うち 6 記事は漢文からなる本文テキストが入力対象外のもので、それを除いた実質的な記事数は 1250 となる。記事要素は延べ語数 1005578 とコーパス全体のほぼ 100%占めるのに対し、雑誌タイトル・欄タイトル、欄や複数の記事に対する説明部分に相当する非記事要素は延べ語数 1402 とごくわずかである。

２．２．記事のジャンル別の言語量

次に、2.1 でコーパスのほとんどを占めた記事要素について、その内容から文学記事（小説・戯曲・詩歌）か非文学記事かの 2 ジャンルに分類しそれぞれの言語量を見ていく。記事ジャンルに関する情報はコーパスには付与されていないので、著者の判断により分類を行った[2]。各ジャンルの延べ語数と記事数を表 2 に示す。

表 2 記事要素の言語量（ジャンル別）

	文学記事	非文学記事	記事要素全体
延べ語数	26195	979383	1005578
記事数	11	1239	1250

非文学記事は延べ語数 979383 と記事全体の 97%を占める。それに対し、文学記事の延べ語数 26195 は記事全体に占める割合だけでなく絶対的な量としても少ない。記事数は 11 であるが連載記事が多く、作品数としては 3 である。3 作品中、詩歌『都の花』と小説『大東號航海日記』は文語体であり、小説『あいびき』のみが口語体である。『あいびき』の延べ語数は 4639、さらにその中の会話部分の延べ語数は 988 とごくわずかであり、当時の話し言葉の実態解明を目的とした研究に堪える言語量を本コーパスのみからは確保できないことがわかる。

２．３．文章種類別・文体別の言語量

次に、2.2 で大きな割合を占めた非文学記事について、文章種類別（地の文／引用）、地の文については文体別（文語／口語／その他）に分類し、言語量を見ていく。文章種類は、quotation 要素を「引用」、それ以外を「地の文」として分類した。文体は、該当本文テキス

[1] コーパスの XML ファイルの仕様の詳細については近藤（2014）を参照のこと。
[2] 分類の際、コーパスのコアデータのサンプリング作業に用いた記事の層別化の内部資料を参照した。

トの直上の style 属性値により「文語」「口語」「その他」に分類した。「その他」には属性値「混在」「項目」「韻文」「万葉」がすべて含まれる。各文章種類・文体の延べ語数と、該当文章種類・文体を 1 語以上含む記事数を示したものが表 3 である。1 記事に複数の文章種類・文体が含まれる場合は、各文章種類・文体で別にカウントした。

表 3　非文学記事の言語量（文章種類・文体別）

	地の文			引用	非文学記事全体
	文語	口語	その他		
延べ語数	847385	6893	2986	122119	979383
記事数	1233	4	5	563	1239

このなかで最も大きな割合を占めるのが文語地の文であり、延べ語数 847385 で非文学記事全体の 87%を占める。一方、口語地の文は延べ語数 6893 と、記事全体に占める割合だけでなく絶対的な量としても少ない。当時の口語体の書き言葉の実態解明を目的とした研究に堪える言語量は、本コーパスのみからは十分に確保できないことがわかる。引用部分は延べ語数 122119 と文語地の文に次ぐ量であるが、古い時代の典拠からの引用が含まれており、そのまますべてを近代語の資料として扱うことはできないものである。

２．４．非翻訳／翻訳別の言語量

次に、2.3 で最も大きな割合を占めた非文学記事の文語地の文について、外国語を翻訳した記事のものか、それとも翻訳でなく日本語としてはじめから書かれた記事のものかで分類し、言語量を見ていく。article タグ originalAuthor 属性に拠り、属性値が空のものを非翻訳記事、何らかの値があるものを翻訳記事として分類を行った。非翻訳／翻訳別の延べ語数と記事数を示したものが表 4 である。

表 4　非文学記事の文語地の文の言語量（非翻訳／翻訳別）

	非翻訳記事	翻訳記事	非文学記事の文語地の文全体
延べ語数	788420	58965	847385
記事数	1169	64	1233

翻訳記事の文語地の文は延べ語数 58965 と文語地の文全体の 7%を占める。翻訳の文章はその原著の言語の影響を受けている可能性があり、厳密には純粋の日本語と区別して考える必要がある。本稿では、この翻訳記事を除いた、非翻訳の非文学記事の文語地の文を調査対象として以下の調査・分析を進める。その言語量を改めてまとめて示すと表 5 のようになる。

表 5　調査対象の言語量

延べ語数（自立語・付属語）	788420
延べ語数（自立語のみ）	472428
記事数	1169

３．一人称代名詞の抽出と頻度

３．１．調査対象の頻度

2 で選定したコーパスの調査対象から一人称代名詞を抽出し、その頻度を集計する。抽出は、SUW タグ pos 属性値が「代名詞」の語を抽出し語形リストを作成、そのリストから調査対象中で主に一人称代名詞として使用されている語形を選定する方法で行った[3]。語形は接尾辞「等（ら）」の接続有無によって区別した。抽出した一人称代名詞の語形と、該当語形の粗頻度、自立語 1 万語あたりの頻度、出現記事数、出現記事率（調査対象の記事数 1169 に対する該当語形の出現記事数の割合）を表 6 に示す。

表 6　調査対象に出現する一人称代名詞

	粗頻度	自立語1万語あたりの頻度	出現記事数	出現記事率
吾人	2762	58.46	591	50.6%
余	496	10.50	119	10.2%
余輩	155	3.28	64	5.5%
我が輩	89	1.88	23	2.0%
小生	80	1.69	9	0.8%
僕	43	0.91	8	0.7%
我々	18	0.38	13	1.1%
余等	8	0.17	4	0.3%
拙者	7	0.15	3	0.3%
乃公（だいこう）	4	0.08	3	0.3%
朕	3	0.06	2	0.2%
吾人等	1	0.02	1	0.1%
乃公等	1	0.02	1	0.1%
全体	3667	77.62	738	63.1%

これによれば、もっとも高頻度の一人称代名詞は「吾人」であり、この 1 語形だけで一人称代名詞全体の頻度の 75%を占める。このように「吾人」が他の語形から突出して高頻度であることは、他の近代雑誌コーパス『明六雑誌コーパス』『太陽コーパス』『近代女性雑誌コーパス』では見られない事象であり[4]、『国民之友コーパス』に特徴的なものである。

３．２．無署名記事／署名記事別の頻度

調査対象で「吾人」が特に高頻度である背景を探るため、調査対象を無署名記事と署名記事に分けて見ていく。article タグ author 属性に拠り、属性値が「＊」のものを無署名記事、それ以外を署名記事として分類を行った。無署名／署名別の言語量を表 7 に示す。

[3] 一人称代名詞としてだけでなく反射指示代名詞としても使用される「われ」、誤解析や一人称代名詞以外の用法がほとんどの「吾曹（ごそう）」「てまえ」「わし」「わたい」「わたし」は分析対象外とした。

[4] 他の近代雑誌コーパスでの一人称代名詞の頻度については、近藤（2012・2013a・2013b）を参照のこと。

表 7 調査対象の言語量（無署名記事／署名記事別）

	無署名記事	署名記事	調査対象全体
延べ語数（自立語・付属語）	454946	333474	788420
延べ語数（自立語のみ）	273886	198542	472428
記事数	887	282	1169

それぞれに出現する一人称代名詞の語形と、その粗頻度、自立語 1 万語あたりの頻度、出現記事数、出現記事率を、無署名記事のものを表 8 に、署名記事のものを表 9 に示す。

表 8 無署名記事に出現する一人称代名詞

	粗頻度	自立語1万語あたりの頻度	出現記事数	出現記事率
吾人	2439	89.05	545	61.4%
余	29	1.06	19	2.1%
余輩	2	0.07	2	0.2%
我が輩	1	0.04	1	0.1%
小生	2	0.07	1	0.1%
僕	3	0.11	3	0.3%
我々	6	0.22	5	0.6%
余等	0	0.00	0	0.0%
拙者	5	0.18	2	0.2%
乃公	1	0.04	1	0.1%
朕	0	0.00	0	0.0%
吾人等	0	0.00	0	0.0%
乃公等	0	0.00	0	0.0%
全体	2488	90.84	558	62.9%

表 9 署名記事に出現する一人称代名詞

	粗頻度	自立語1万語あたりの頻度	出現記事数	出現記事率
吾人	323	16.27	46	16.5%
余	467	23.52	100	36.0%
余輩	153	7.71	62	22.3%
我が輩	88	4.43	22	7.9%
小生	78	3.93	8	2.9%
僕	40	2.01	5	1.8%
我々	12	0.60	8	2.9%
余等	8	0.40	4	1.4%
拙者	2	0.10	1	0.4%
乃公	3	0.15	2	0.7%
朕	3	0.15	2	0.7%
吾人等	1	0.05	1	0.4%
乃公等	1	0.05	1	0.4%
全体	1179	59.38	180	64.7%

「吾人」の粗頻度と記事の署名の有無との関係を見るために、表 10 のクロス表で χ^2 検定

（イェーツの補正あり）[5]を行った。

表 10「吾人」の別と記事署名の有無によるクロス表

	無署名記事	署名記事
「吾人」粗頻度	2439	323
「吾人」以外の一人称代名詞粗頻度	49	856

その結果、1%水準で有意差が認められた（$\chi^2(1)=1130.97$、$p=.0000$、$\varphi=0.60$）。これは「吾人」が無署名記事に多く出現していることを示す。他の近代雑誌コーパスと比較して『国民之友コーパス』で「吾人」が突出して多く出現する要因が、無署名記事での「吾人」の多用であることがわかる。さらに無署名記事の内部で見ると、「吾人」の粗頻度 2439 は無署名記事の一人称代名詞全体の粗頻度 2488 の実に 98%を占める。加えて無署名記事に出現する「吾人」以外の語形について詳細に調査すると、その多くは引用中での使用と見なせるものであったり一人称代名詞以外の用法で用いられているものであったりして、地の文の一人称代名詞と確定できるものは一層少ない。つまり、無署名記事では一部の例外を除き「吾人」が専用されていることになる。創刊当初の『国民之友』の無署名記事について、有山（1986）に「当初の「国民之友」誌上には、民友社員が署名入りで発表した文章は少ない。無署名の社説のほとんどは、殆ど蘇峰の執筆であろうし、編集企画にも彼の指導力が大きかったであろう。（略）大江義塾出身者は、論文執筆者としてよりも、編集実務担当者・無署名記事執筆者の役割を果たしていたと見ることができる。」とある。一人称代名詞「吾人」の専用は、蘇峰およびその指導下にあった民友社員による文章のありようを特徴付けるものであったと言える[6]。

一方で、署名記事では「余」が最も頻度が高く、「吾人」「余輩」「我が輩」等がそれに次ぐ。署名記事に出現する一人称代名詞全体の粗頻度 1179 に対する「余」の粗頻度 467 の割合は 40%と比較的高くはあるものの、「吾人」「余輩」「我が輩」等のその他の語形もそれなりの割合を占めており、無署名記事のように「吾人」専用といった状況は見られない。署名記事の著者数は異なりで 62 を数えるが、それらの著者の文章の個性が集合し、署名記事での一人称代名詞の多様性となって現れたと見るべきである。

４．一人称代名詞の共起

ここで、一人称代名詞の語形に多様性がある署名記事を対象として、記事単位での一人称代名詞の共起の実態について見ていく。語形ごとに、出現記事数、他の語形と共起する記事数（共起記事数）、出現記事数に対する共起記事数の割合（共起記事率）、他の語形と共起せず該当語形が専用される記事数（専用記事数）、出現記事数に対する専用記事数の割合（専用記事率）を示したものが表 11 である。1 記事に複数の語形が出現する場合、各語形で別に出現記事数をカウントした。

[5] χ^2 検定は統計分析ソフト R の chisq.test()関数に拠り、φ 係数は R の vcd ライブラリの assocstats()関数に拠った。R のスクリプトの記述では竹内・水本（編著）（2012）およびそのコンパニオン・ウェブサイト（http://mizumot.com/handbook/?page_id=515）を参照した。

[6] 「吾人」を全く用いず「拙者」を専用する例外的な無署名記事「名代役者の手紙」（22 号）が、その題名から明らかなように蘇峰・民友社員以外の人物によって（あるいは、そのように装って）執筆されたものであることもその裏付けとなる。

表 11　一人称代名詞の出現記事数（共起／専用別）

	出現記事数	共起記事数	共起記事率	専用記事数	専用記事率
吾人	46	28	61%	18	39%
余	100	45	45%	55	55%
余輩	62	46	74%	16	26%
我が輩	22	15	68%	7	32%
小生	8	2	25%	6	75%
僕	5	2	40%	3	60%
我々	8	7	88%	1	13%
余等	4	4	100%	0	0%
拙者	1	1	100%	0	0%
乃公	2	2	100%	0	0%
朕	2	2	100%	0	0%
吾人等	1	1	100%	0	0%
乃公等	1	1	100%	0	0%
全体	180	74	41%	106	59%

ここから分かるように、全体では共起記事率 41%より専用記事率 59%のほうが高い。ただし、語形によってその値には違いがある。出現記事数上位 3 語形で見ると、「余」は専用記事率のほうが高く、「吾人」「余輩」は共起記事率のほうが高い。

この 3 語形の共起についてより詳しく見ていく。調査対象には、3 語形の粗頻度合計が 5 以上でかつ 3 語形以外の一人称代名詞が出現しない記事が 50 ある。この 50 記事について、出現する語形の組み合わせごとに、記事数と、記事中の 3 語形の粗頻度合計に対して該当語形の粗頻度が 80%以上の記事数（優勢記事数）を示したものが表 12 である。

表 12 「吾人」「余」「余輩」の共起組み合わせ別記事数

	記事数	「吾人」 優勢記事数	「余」 優勢記事数	「余輩」 優勢記事数
吾人	9	9	—	—
余	11	—	11	—
余輩	2	—	—	2
吾人-余	9	3	4	—
吾人-余輩	2	1	—	1
余-余輩	13	—	8	0
吾人-余-余輩	4	0	1	0
全体	50	13	24	3

「吾人」は出現する記事数合計 24 に対する優勢記事数 13 の割合が 54%、「余」は出現する記事数合計 37 に対する優勢記事数 24 の割合が 65%であるのに対し、「余輩」は出現する記事数合計 21 に対する優勢記事数 3 の割合が 13%と低い。「余輩」は「吾人」「余」と比べて、記事中で主たる語形として用いられるよりも従たる語形として用いられる傾向にあると言える。また、「余輩」は「吾人」と共起する記事数合計が 6、「余」と共起する記事数合計が 17 であり、「余輩」は「吾人」より「余」と共起しやすいと言える。さらに、「余-余輩」の組み合わせの 13 記事中、「余」優勢記事は 8、「余輩」優勢記事は 0 であり、「余輩」は「余」と共起する場合、主たる語形となることはない。

その「余-余輩」の組み合わせの記事について、語形の用法を文脈に沿って調査すると、

「余」「余輩」ともに一人称単数として用いられており、両語形の間に明確な使い分けがあるようには見えないものがほとんどである（①②）。1 記事中に「余」だけでなく「余輩」も共起する必要がある理由は明かではない。

① 余向に聖書翻譯完成すと題する一篇の批評文を國民之友に掲げたるに該翻譯委員の一人たる松山氏より事實相違の辨駁を爲せり、（…中略…）而して兩方を知れる余輩の批評を事實相違と斷言するは氏の爲に取らざる所なり、（21 号「松山高吉氏の辨駁に答ふ」高橋五郎）

② 而して余輩が茲に之を附加するを快しとせざれども亦未だ全く之を除く能はざるは本國の利益を謀るに必要なりと思惟すれば駐在國の教會新聞議員を利用し他國より來れる同僚と合縱連衡するの權略是なり余は今之を最後に置きたれども今日外交の實勢は猶之を首要資格（クォーリティース、オフ、プライマレー、イムポータンス）の中より拔去るを許さざること余が悲む所なり（24 号「外交術及び外交家（二）」朝比奈知泉）

他の語形の組み合わせの記事についても、語形の用法を文脈に沿って見ていく。「吾人-余」の組み合わせの記事の場合、「余」優勢記事では「余」が一人称単数、「吾人」は一人称複数として使い分けられていると考えられる記事が多い（③④）。

③ 然ば則ち平民的の文明を日本に誘入し東洋古來の氣風を一變するの任は吾人平民社會を除きて他に求べきに非ず自助の精神自奮の氣象此時期に於て最も缺く可からざるなり（…中略…）予不材なりと雖願くは此精神を有するの先輩に追隨して其勞の一部を分受せんことを切望する者なり（2 号「平民社會の責任」島田三郎）

④ 然れども余の考ふる所は世人と差や異る所あり余は二十三年後の日本を以て、万事創始の日本たらしめず（…中略…）是れ吾人日本の未來を慮る者が今日に於て思慮を費すべきの一事なりと思考するなり（15 号「二十三年後の日本」肥塚龍）

一方で、「吾人-余」の組み合わせの「吾人」優勢記事は 3 記事あるが、うち 1 記事は「余」が引用中の用例と見られるもので、実質的には「吾人」専用記事である。残る 2 記事は「余」とともに「吾人」も一人称単数として用いられていると考えられる。うち 1 記事では「吾人」は本文中に、「余」は末尾注中に用いられ、文章の性質に対応した使い分けが見られる。もう 1 記事では、著者の米国での具体的な体験談を語る場面でのみ「余」が用いられており、これも文章の性質に対応した使い分けが見られる（⑤）。

⑤ 吾人が私立大學を設立せんと欲したるは一日に非ず、而して之れが爲めに經營辛苦を費したるも亦た一日に非らず、今まや計畫畧ぼ熟し、時期漸く來らんとす、吾人は今日に於て、此を全天下に訴へ、全國民の力を藉り、其の計畫を成就せずんば、再び其時期無きを信ず、是れ吾人が從來計畫したる所の顚末を陳じ、併せて之れを設立する所の目的を告白するの止む可らざる所以んなり、（…中略…）明治七年、余が米國より歸朝するに際し、適ま北米合衆國外國傳道會社の集會ありき、 米國の紳士貴女、會する者三千餘名、余の友人にして此會に集る者頗る多きにより、諸友余を要して臨會せしめ、且つ訣別の辭を求めらる、（34 号「同志社大學設立の旨意」新島襄）

つまり、「吾人」優勢記事は実質的には「吾人」「余」それぞれの専用の文章が合体して 1 記事になっているのであり、同質の文章中に「吾人」「余」が共起している例とは見なせないものである。「吾人-余」の組み合わせが同質の文章中に出現する場合は、③④で見たように「余」が一人称単数として主たる語形となり、「吾人」は一人称複数として従たる語形と

なる。

「吾人-余輩」の組み合わせの場合、「余輩」優勢の 1 記事は「吾人」が引用中の用例と見られるもので、実質的には「余輩」専用記事である。残る「吾人」優先の 1 記事では「余輩」が一人称単数、「吾人」が一人称複数として用いられていると考えられる（⑥）。

⑥　科學とは何ぞや、實際とは何ぞや予輩之を釋て曰く「科學とは天然法の解則にして實際とは社會の現状なり」と（…中略…）斯く理論家の實際世界より退けらるるや所謂る實際家なるもの恰かも強敵を千里の外に驅逐せるの思を爲し縦横己れの説を實際に試むるが故に終に吾人の社會は彼等が遊戯の舞臺と變じ私利の競爭場と化して復は如何ともする能はざるなり（8 号「理論實際の和解法」伴直之助）

以上をまとめると語形の共起関係について次のような傾向が指摘できる。「余」は一人称単数として主たる語形として用いられることが多く、その場合の従たる「吾人」は一人称複数の役割を、「余輩」は「余」と同じく一人称単数として言い換え表現的な役割を担う。一方で「吾人」は主たる語形としても用いられ、その場合は一人称単数用法となる。「余輩」も主たる語形として一人称単数として用いられる場合もあるが、その数は多くない。

５．おわりに

以上、『国民之友コーパス』を用いて一人称代名詞の計量的分析を行った。まず分析の前にコーパスの言語量から資料性の検討を行った。本コーパスは非文学の文語文が大部分を占める資料であり、口語文あるいは文学については十分な言語量がなく、他の資料と組み合わせて使う必要がある。次に、非文学・非翻訳記事の文語地の文を対象資料として一人称代名詞の抽出・分析を行った。無署名記事と署名記事では一人称代名詞の語形の分布が異なることが明らかとなった。また、記事単位での複数語形の共起関係についても分析し、「吾人」「余」「余輩」の振る舞いの傾向が明らかになった。

語形と記事署名との対応関係、語形の共起関係については本稿で新たに解明された点である。今後は他の近代雑誌コーパスについても同様の観点から調査・分析し、コーパス間の比較を行いたい。

付　記

本稿は、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」による研究成果の一部を含むものである。

文　献

有山輝雄（1986）「言論の商業化―明治 20 年代の民友社―」『コミュニケーション紀要』4、pp.1-23（http://www.seijo.ac.jp/graduate/gslit/orig/journal/communication/pdf/scom-04-01.pdf よりダウンロード可）

岡田賢二（1998）「明治期の東京語における人称代名詞の研究―明治・大正期の落語の速記本にあらわれた一、二人称代名詞―」『埼玉大学国語教育論叢』2、pp.34-58

祁福鼎（2006a）「明治時代語における自称詞の使用実態と使用規範について」『文学研究論集』24、pp.45-61

祁福鼎（2006b）「明治時代語における自称詞の推移と位相について」『明治大学日本文学』32、pp.95(1)-78(18)

国立国語研究所（2014）『国民之友コーパス』Ver.1.0、
http://www.ninjal.ac.jp/corpus_center/cmj/kokumin/
近藤明日子（2012）「明治初期論説文における一人称代名詞の分析―『明六雑誌』コーパスを用いて―」『第 1 回　コーパス日本語学ワークショップ予稿集』pp.265-272
（http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no1_papers/JCLWorkshop2012_35.pdf よりダウンロード可）
近藤明日子（2013a）「近代女性向け雑誌記事における一人称代名詞の分析―形態論情報付き『近代女性雑誌コーパス』を用いて―」『第 3 回　コーパス日本語学ワークショップ予稿集』pp.313-322
（http://www.ninjal.ac.jp/event/specialists/project-meeting/files/JCLWorkshop_no3_papers/JCLWorkshop_No3_39.pdf よりダウンロード可）
近藤明日子（2013b）「近代総合雑誌記事に出現する一人称代名詞の分析―単語情報付き『太陽コーパス』を用いて―」『近代語研究』17、pp.134-154
近藤明日子（2014）「『国民之友コーパス』解説書　第 1.1 版」
（http://www.ninjal.ac.jp/corpus_center/cmj/doc/kokumin_manual_v1_1.pdf よりダウンロード可）
竹内理・水本篤（編著）（2012）「第 11 章　頻度データ分析入門　人数や回数を比較するには」『外国語教育研究ハンドブック』松柏社
房極哲（2004）「近代語における一、二人称代名詞の変遷について」『日本文化學報』21、pp.1-15

参考 URL

R　http://www.r-project.org/

『日本語話し言葉コーパス（CSJ）』の異なる講演タイプにおける外来語の質的分析 —言語外的および言語内的指標を用いた外来語分類の試み—

久屋 愛実（オックスフォード大学）†

A Qualitative Analysis of Loanwords in Different Speech Styles in the Corpus of Spontaneous Japanese (CSJ): Classifying Loanwords Based on Extra-/Intra-Linguistic Factors

Aimi Kuya (Faculty of Linguistics, Philology and Phonetics, University of Oxford)

要旨

本稿では、レジスター横断性やジャンル横断性に留意して『日本語話し言葉コーパス(CSJ)』から「基本度」（水谷 1964）の高い外来語を抽出し、それらの語彙的特徴を記述する。分析の結果、レジスター横断的かつジャンル横断的である最も基本度の高い語群は、それ以外の語群よりも抽象的あるいは多義的な意味を表す語の割合が高く、普通名詞（一般）以外の品詞の割合が高い傾向にあった。

１．はじめに

コーパスを使った語彙研究においては、語の「基本度」（水谷 1964）を頻度により捉えるのが最も一般的であろう。通時的コーパスを使う場合は、頻度の経年的増減を追うことによって基本語化した語彙を取り出すことが可能である（金 2011、田中 2014）。しかし、共時的コーパスを扱う場合は頻度の経年的増減が捉えられないため、広範囲に分布するかどうかを示す「散らばり度」（水谷 1964）が語の基本度をはかる指標として有効である。本稿は、共時的コーパスである『日本語話し言葉コーパス（以下、CSJ）』に出現する外来語を、異なるレジスターやジャンルにまたがって分布する語かという観点から分類し、特定のレジスターやジャンルに左右されない、「いわば無性格な語群」（田中 1973）を抽出する。こうした「無性格語」は、他と比べてより基本的な語彙であると考えられるが、これらがどのような語彙的特徴をもつのかについても考察する。

２．語の散らばり度に基づいた「無性格語」の抽出

本稿では、CSJ[1]の学会講演と模擬講演部分から抽出した外来語の分析を行う。水谷（1964: 10）が指摘するように、例えば雑誌における語の散らばり度は、「あるいは一編ずつの記事、あるいは雑誌の一冊ずつ、あるいは小説・随筆・論説のような記事分類の別」によって求められる。これに倣えば、CSJ における語の散らばり度は、文章別、講演別、学会種や講演テーマ別（ジャンル別）、講演のタイプ別（レジスター別）、あるいは講演者別など、あらゆる単位からはかることが可能である。本稿では、このうち講演タイプの別（レジスター）と学会種・講演テーマの別（ジャンル）の 2 指標を用いる。

† aimi.kuya@ling-phil.ox.ac.uk

[1] CSJ の概要については国立国語研究所（2006）を参照されたい。

２．１　レジスター横断性

表 1 は、CSJ の学会講演（Academic Presentation Speech、以下 A）と模擬講演（Simulated Public Speaking、以下 S）における異なり語数・延べ語数とその比率を語種ごとに示したものである。外来語のみに関して言えば、その割合は異なり語数・延べ語数ともに模擬講演より学会講演で高い。また、外来語の異なり語数は学会講演（3555 語）よりも模擬講演（4229 語）のほうが多いものの、延べ語数でみると学会講演（100428 語）が模擬講演（67863 語）の 1.5 倍にもなり、学会講演では外来語の一語あたりの平均出現度数が高いことがわかる。

表 1：CSJ 学会講演と模擬講演における語種別の頻度と比率

		外	漢	和	混	固	記号	その他 (空白・不明等)	総計
異なり	学会講演（A）	3555	8773	5222	507	2564	759	2283	23663
		15.0%	37.1%	22.1%	2.1%	10.8%	3.2%	9.6%	100.0%
	模擬講演（S）	4229	11660	9386	1004	4407	226	2464	33376
		12.7%	34.9%	28.1%	3.0%	13.2%	0.7%	7.4%	100.0%
延べ	学会講演（A）	100428	691117	1087123	22938	20818	9990	54504	1986918
		5.1%	34.8%	54.7%	1.2%	1.0%	0.5%	2.7%	100.0%
	模擬講演（S）	67863	470064	1348909	27969	41500	1863	47014	2005182
		3.4%	23.4%	67.3%	1.4%	2.1%	0.1%	2.3%	100.0%

UniDic 短単位による[2]。品詞が「空白、記号、助詞、助動詞」となるものは含まない。

表 2：レジスター横断性

	学会講演（3555 異なり語）			総計
		模擬講演（4229 異なり語）		
	特徴語 A	共通語	特徴語 S	
外来語の異なり語数	1735	1820	2409	5964
外来語の延べ語数	20740	133382	14169	168291
一語あたりの平均度数	12	73	6	28

こうした違いは、学会講演と模擬講演という異なるレジスターで出現する外来語が完全に同質ではないことに起因すると思われる。表 1 の外来語の中には両レジスターで重複して出現するものもあればそうでないものもあり、それぞれのふるまいが異なる可能性があるからである。そこで、表 1 で抽出した外来語を、学会講演（A）にのみ出現する「特徴語 A」、模擬講演（S）にのみ出現する「特徴語 S」、どちらにも共通で出現する「共通語」の 3 種に再分類してみる。散らばり度の観点からすると、共通語は 2 つの特徴語に比べて「レジスター横断性」が高い。分類の結果、表 2 に示す通り、学会・模擬講演を統合したときの外来語の異なり語数は 5964 語で、このうち特徴語 A の 1735 語、特徴語 S の 2409 語を除くと、共通語は 1820 語にまで減少する。つまり、5964 語のうち約 7 割がどちらかひと

[2] UniDic 体系の CSJ 短単位データは、現在国立国語研究所が整備中である。今回は同研究所の許可を得て公開前のものを分析に利用したため、今後一般に公開されるデータを用いた分析とは結果が異なる可能性がある。（本データは 2014 年 11 月時点のもの）

つのレジスターにしか出現しない特徴語であることがわかる。さらに、一語あたりの平均出現度数は特徴語 A が 12 回で、特徴語 S（6 回）の 2 倍にもなる。一方、共通語の一語あたりの平均出現度数は両レジスター全体で 73 回であり、2 つの特徴語よりも圧倒的に高い。このことから見ても、この 3 つのカテゴリーは区分して論じたほうがよさそうである。

２．２　ジャンル横断性

次に、ジャンル横断的に分布する広範囲語かどうか、抽出した 5964 の外来語それぞれの「ジャンル横断性」をはかる。ここでは、学会講演における 13 の学会種、模擬講演における 12 の講演テーマをジャンル数とみなす。まず、表 3 のとおり、それぞれの外来語が講演タイプごとにいくつのジャンルに出現したかを求め、整理した。ジャンル横断性の序列は、表の色分けされた区分に従って行った。ジャンル横断性は、色なし部分が最も低く、薄い網掛け部分がその中間で、濃い網掛け部分が最も高い。

表 3：出現ジャンル数別にみた外来語 5964 語

			模擬講演				
			共通語			特徴語 A	
		出現ジャンル数	1~4 テーマ	5~8 テーマ	9~12 テーマ	なし	総計
学会講演	共通語	1~4 学会	893	306	147	1655	3001
		5~9 学会	154	105	142	79	480
		10~13 学会	7	15	51	1	74
	特徴語 S	なし	2175	206	28		2409
		総計	3229	632	368	1735	5964

散らばり度：低い（色なし）　中間　高い

表 4：ジャンル横断性と特徴語・共通語の別

ジャンル横断性	特徴語 A	共通語	特徴語 S	総計
高い	1	208	28	237
中間	79	719	206	1004
低い	1655	893	2175	4723
総計	1735	1820	2409	5964

表 4 は表 3 を色別にまとめ、先にみた特徴語・共通語の別を加えて分類しなおしたものである。その結果、ジャンル横断性が高い 237 語、中間レベルの 1004 語、ジャンル横断性が低い 4723 語に分かれた。このうち、ジャンル横断性が高い 237 語を「ジャンル横断性の高い語」または「ジャンル広範囲語」と定め、さらなる分析に利用する。ジャンル広範囲語は、特徴語 A（1 語）、共通語（208 語）、特徴語 S（28 語）の 3 つにさらに分けられる。

以下にこれら全ての語彙を示す（五十音順）。特徴語 A（1 語：「コンテキスト」）と特徴語 S（28 語：「エアロビック」～「ロープ」）は、個々のレジスターにおいてはジャンル横断性が高いが、レジスター横断的な語彙ではないため、あくまでもそれぞれのレジスターに限り広く分布している「キー・ワード」[3]（田中 1973）でしかない。これらを除いた

[3] 田中（1973）によれば、ある文章の頻度調査において頻度順位の比較的上位に来る語彙のうち、特定の文章や文献の性格に関わらず現れうる「無性格語」を排除すると「キー・ワードすなわち、『いかにも、その文章らしい単語』」が残るとする。

残りの共通語（208 語：「アイディア」～「ワールド」）が、ジャンル横断性だけでなくレジスター横断性も高いことから、特定のレジスターやジャンルに左右されない、本コーパスの「無性格語」と見ることができる。

ジャンル広範囲語全 237 語

特徴語 A：（1 語）
コンテキスト

共通語：（208 語）（＝無性格語）
アイディア、アウト、アクセス、アクセント、アップ、アドバイス、アナウンサー、アプローチ、アルバイト、アンド、イコール、イベント、イメージ、イン、インターネット、インタビュー、ウイーク、ウインドー、エネルギー、エピソード、エレベーター、エンジン、オーケー、オーバー、オープン、オフ、オブ、オフィス、オレンジ、カー、カード、ガイド、カウント、カット、カバー、カメラ、カラー、ガラス、キー、ギャップ、キャラクター、キロ、クラシック、クラス、グラフ、クリア、グループ、ケース、ゲーム、コース、コーヒー、コピー、コミュニケーション、コメント、コントロール、コンピューター、ザ、サービス、サイクル、サイズ、サイン、サポート、サン、シート、シーン、システム、ジャンル、シンボル、スーパー、スクリーン、スケジュール、スター、スタート、スタイル、ストーリー、ストップ、ストレス、スピーチ、スピード、スペース、スポーツ、スムーズ、スリー、ゼロ、センス、センター、センチ、ソフト、ターゲット、タイトル、タイプ、タイミング、タイム、ダウン、ダブル、チーム、チェック、チャンス、チャンネル、ツー（< *two*）、ツー（< *to*）、データ、データーベース、テープ、テーブル、テーマ、テキスト、デザイン、デジタル、テスト、テレビ、ドア、トップ、トラック、トラブル、ドラマ、トレーニング、ナンバー、ニュー、ニュース、ネット、ネットワーク、ノー、ノート、パーセント、ハード、ハイ、バス、パソコン、パターン、バック、バラエティー、バランス、パンフレット、ピーク、ビジネス、ヒット、ビデオ、ピンク、ヒント、ファースト、ファイブ、ファミリー、プラス、プラン、フリー、フル、ブルー、プロ、プログラム、プロジェクト、プロセス、ブロック、ペア、ページ、ベース、ペース、ペーパー、ベスト、ベッド、ポイント、ホーム、ボール、ボタン、ボックス、ボランティア、マーク、マイク、マイナス、マシン、マスコミ、マナー、マニュアル、ミス、ミリ、メーター、メートル、メール、メーン、メッセージ、メニュー、メモ、メリット、メンバー、モデル、モニター、ユニーク、ライフ、ライブ、ライン、ラジオ、ラベル、ランク、リアル、リーダー、リード、リスト、リズム、リラックス、ルーム、ルール、レコード、レストラン、レベル、ワーク、ワード、ワープロ、ワールド

特徴語 S：（28 語）
エアロビック、オーナー、クーラー、グッズ、ゴールデン、シャワー、ジャングル、ジョギング、スープ、スカート、スナック、ズボン、デザート、テント、バイク、バッグ、ハンバーグ、フルーツ、プロデューサー、マージャン、マラソン、ミネラル、メダル、リゾート、リフレッシュ、レース、レンタル、ロープ

３．無性格語の意味特性

ここでは、前節で抽出した無性格語の意味的特徴を調べるため、『分類語彙表―増補改訂版』（国立国語研究所 2004）の分類に従って意味分類を行う。手順は、各外来語に付与された UniDic の語彙素 ID を主キーとして分類語彙表から分類語彙表番号を割り出し[4]、その中の「部門」番号に基づいて 5 項目｛1 抽象的関係、2 人間活動の主体、3 人間活動-精神および行為、4 生産物および道具、5 自然物および自然現象｝に分類する、というものである。ただし、多義語の場合は、ひとつの語彙素 ID に対して複数の分類語彙表番号が割り当てられており（小木曽・中村 2011）、結果として異なる複数の「部門」番号を有することがある。そのような語彙素には、複数の意味分野を持つという意味で「多義」という 6 つ目の分類名を新たに付与した。最後に、分類語彙表において対応する語彙素 ID が見つけられない場合は、その語彙素が分類語彙表に収録されていないという意味で「未収録」とい

[4] 国立国語研究所コーパス開発センター「形態論情報データベース」（小木曽・中村 2014）上の辞書データと分類語彙表データを利用した。

う 7 つ目の分類名を付与した。なお、分類語彙表の採用語は、「現代の日常生活で普通に用いられる語を中心に、各種語彙調査の結果その他から選定」され、原版にあった語も含めて「見慣れない専門用語や古語・方言、また社会生活上使用を遠慮すべき語の類は除いている」（国立国語研究所 2004: 3）。よって、ここで「未収録」に区分された語彙は、あくまでも増補改訂版の作業時に上記条件に当てはまらないと判断されたものであり、当時から約 10 年経った現在の感覚とは異なる可能性がある。

表 5：ジャンル広範囲語（237 語）の意味分類

	1 抽象的関係	2 人間活動-主体	3 人間活動-精神・行為	4 生産物・道具	5 自然物・自然現象	多義	未収録	総計
特徴語 A							1	1
共通語（＝無性格語）	49	12	54	31	3	46	13	208
特徴語 S	1	3	5	12	4	2	1	28
総計	50	15	59	43	7	48	15	237
延べ語数（両レジスターの合計）	21290	2208	18446	8465	501	15717	5892	72519
一語あたりの平均度数	**426**	147	**313**	197	72	**327**	**393**	306

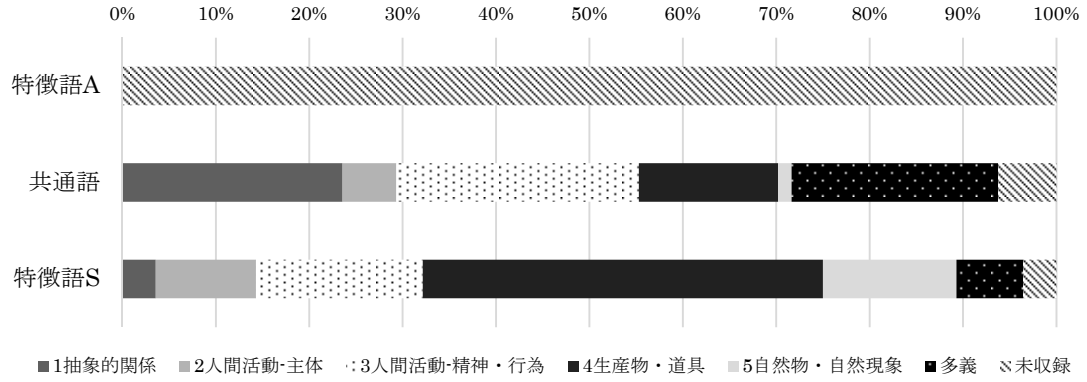


図 1：ジャンル広範囲語（237 語）の意味分類比率

表 5 は、無性格語を含むジャンル広範囲語 237 語の意味分類を示したものである。図 1 はそれを百分率に直したものである。特徴語 A は「コンテキスト」一語で、未収録語に分類されている。共通語（＝無性格語）と特徴語 S とを比較すると、特徴語 S よりも共通語で「1 抽象的関係」、「3 人間活動（精神・行為）」、「多義」の割合が高く、特に「1 抽象的関係」と「多義」は特徴語 S―共通語間の比率差が著しい。一方、「2 人間活動（主体）」、「4 生産物・道具」、「5 自然物・自然現象」の割合は特徴語 S よりも共通語で低く、特に「4 生産物・道具」と「5 自然物・自然現象」は特徴語 S―共通語間の比率差が著しい。なお、7 つの意味分類のうち、一語あたりの平均出現度数は「1 抽象的関係」、「未収録」、「多義」、「3 人間活動（精神・行為）」の順に高く、「1 抽象的関係」、「3 人間活動（精神・行為）」、「多義」の割合が高い共通語（208 語）には比較的高頻度の語彙が多く含まれていることがわかる。一方、一語あたりの平均出現度数が相対的に低いのは「5 自然物・

自然現象」、「2 人間活動（主体）」、「4 生産物・道具」であり、「4 生産物・道具」や「5 自然物・自然現象」の割合が高い特徴語 S（28 語）には、ジャンル広範囲語でありながら比較的低頻度の語彙が多く含まれていることがわかる。

４．無性格語の品詞特性

次に、無性格語の品詞的特徴を調べるため、無性格語を含むジャンル広範囲語 237 語を、UniDic の品詞分類に基づいて分類し、表 6 に示した。図 2 ではそれを百分率で示している。

表 6：ジャンル広範囲語（237 語）の品詞分類

	名-普-一般	名-普-サ変可能	名-普-サ変形状詞可能	名-普-形状詞可能	名-普-助数詞可能	名詞-数詞	形状詞-一般	総計
特徴語 A	1							1
共通語（＝無性格語）	138	46	3	10	7	1	3	208
特徴語 S	25	3						28
総計	164	49	3	10	7	1	3	237
延べ語数（両レジスターの合計）	51658	11374	335	1598	5282	2098	174	72519
一語あたりの平均度数	315	232	112	160	**755**	**2098**	58	306

*UniDic では品詞情報が語形 ID に紐づけされるため、語彙素 ID が複数の品詞情報を持つ場合がある。ここでは「オフ」と「ノート」が名-普-一般または名-普-サ変可能であった。今回は語彙素 ID でカウントするために、サ変用法が実際に確認できた前者を名-普-サ変可能、サ変用法が確認できなかった後者を名-普-一般として 1 つの品詞にまとめた。

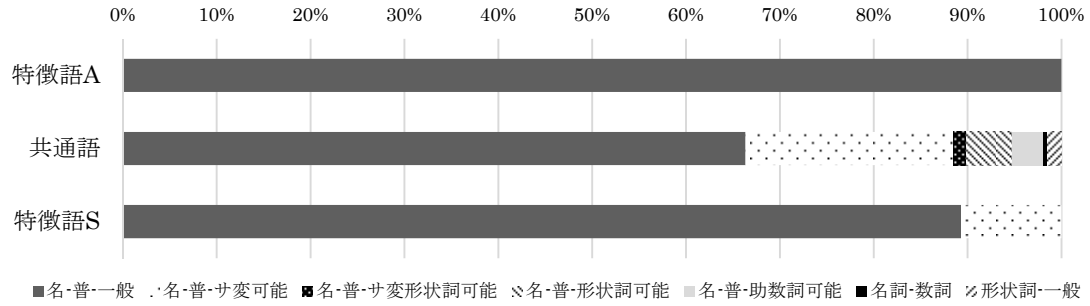


図 2：ジャンル広範囲語（237 語）の品詞比率

特徴語 A は「コンテキスト」一語で、普通名詞（一般）である。共通語と特徴語 S を比較すると、特徴語 S では「エアロビック」などの普通名詞（一般）が圧倒的多数で、「ジョギング」などの普通名詞（サ変可能）は 1 割程度である。それに対し、共通語では普通名詞（一般）が 7 割に満たず、「アクセス」などの普通名詞（サ変可能）、「オープン」などの普通名詞（サ変形状詞可能）、「イコール」などの普通名詞（形状詞可能）、「キロ」などの普通名詞（助数詞可能）、「ゼロ」などの数詞、「スムーズ」などの形状詞などが合わせて 3 割以上を占めており、普通名詞（一般）以外の品詞の割合が比較的高い。なお、7 つの品詞分類のうち、一語あたりの平均出現度数が圧倒的に高いのは「数詞」で、「ゼロ」一語で 2098 延べ語数に達する。その次に普通名詞（助数詞可能）が続き、数詞や助数詞系の語彙は少ない異なり語数がかなりの高頻度で使われていることがわかる。一方、形状詞、普通名詞（サ変形状詞可能）、普通名詞（形状詞可能）など、形状詞系は一語あたりの平均出現度数が相対的に低く、ジャンル横断的ではあるものの比較的低頻度である。

よって、頻度を基本語抽出の基準とすると、数詞や助数詞系は抽出されやすいが、形状詞系の品詞は抽出されにくい場合もあるかもしれない。

５．レジスター偏重度

最後に、無性格語 208 語についてレジスター別の出現度数を調べ、レジスターによる出現頻度の偏りのない、無性格語のなかでもさらに「無性格な」語を特定する。レジスター効果による偏りは、各語彙の「学会講演占有率（A%）」で評価する。学会講演占有率とは、学会講演における PMW（百万語当たりの出現度数）が、学会講演における PMW と模擬講演における PMW の合計の何%を占めているかを表す値である。例えば表 7 にあるように、外来語「データ」の PMW は学会講演で 4141/100428*1000000=2084、模擬講演で 85/67863*1000000=42 となり、学会講演占有率は 2084/(2084+42)=0.9801 となる。こうして求めた値をもとに、学会講演占有率が 75%より大きいものを学会講演（A）に偏って出現する「共通語（A 偏重型）」、25%より小さいものを模擬講演（S）に偏って出現する「共通語（S 偏重型）」、それ以外（25%以上 75%以下のも）を「共通語（AS 共通型）」に分類していった。その結果、表 8 に示すように A 偏重型は 48 語、AS 共通型は 109 語、S 偏重型は 51 語となった。特定の講演タイプに偏って出現している偏重型よりも、両講演タイプで同程度に出現する AS 共通型が共通語（＝無性格語）の中でもさらに「無性格な語」といえるだろう。

表 7：学会講演度数占有率（A%）に基づく共通語（＝無性格語）の下位分類作業の例

語彙素	語彙素 ID	粗頻度（学会）	粗頻度（模擬）	PMW（学会）	PMW（模擬）	学会講演占有率（A%）	語の分類名
データ	25819	4141	85	2084	42	98.01%	共通語 A 偏重型
クラス	10431	605	315	304	157	65.97%	共通語 AS 共通型
テーマ	25515	198	1432	100	714	12.25%	共通語 S 偏重型

表 8：レジスター偏重度別に見た共通語（＝無性格語）の内訳

特徴語 A	共通語（＝無性格語）			特徴語 S	総計
	A 偏重型	AS 共通型	S 偏重型		
A%=100	100>A%>75	75≧A%≧25	25>A%>0	A%=0	
	48 語	109 語	51 語		
1 語	208 語			28 語	237 語

表 9 は、本分析のまとめとして、今回扱ったジャンル広範囲語全 237 語を、これまでにみてきた意味分類、品詞分類、レジスター偏重度の 3 指標に基づいて分類したものである（五十音順、*や**は普通名詞（一般）以外の品詞であることを示す）。一方のレジスターにのみ出現する特徴語のうち、特徴語 A は「コンテキスト」1 語のみで、特徴語 S は「クーラー」、「シャワー」など具体物を示す語が多い。これ以外の、両レジスターに出現する共通語 208 語を「無性格語」と呼んだ。そのうち、レジスター偏重度の高い A 偏重型 48 語と S 偏重型 51 語を除くと、無性格語のなかでもレジスター偏重度の低い、さらに「無性格な」AS 共通型 109 語が特定できる（網掛け部分）。無性格語は総じて抽象的な語が多いが、A 偏重型では「アプローチ*」、「データ」、「データーベース」や「パーセント**」などの助数詞系など、学術分野と関連の深そうな語が目立つ。一方、S 偏重型は、「キャラク

ター」、「ファミリー」、「アルバイト*」など、より日常的な分野と関連の深そうな語が目立つ。

表 9：ジャンル広範囲語全 237 語の分類（まとめ）

	1 抽象的関係	2 人間活動-主体	3 人間活動-精神・行為	4 生産物・道具	5 自然物・自然現象	多義	未収録	計
特徴A							1 コンテキスト	1
共通A偏重型	16 アプローチ* オフ* システム ゼロ** タイミング チャンネル ツー（<*two*） データ データーベース パターン プロセス ペア ページ** ベース ランク* レベル	1 アナウンサー	11 アクセント グラフ コントロール* サポート* テキスト テスト* プログラム* プロジェクト マーク* リスト* ルール	3 キー マイク ラベル		11 イコール* カウント* カバー* グループ ターゲット ネットワーク ピーク プラス* ブロック* マイナス* モデル	6 アンド オブ ツー（<*to*） パーセント** ミリ** ワード	48
共通AS共通型	29 アップ* ギャップ サイクル* サイズ シート ジャンル シンボル スタイル ストップ* スピード スペース スムーズ* スリー タイム チャンス デジタル* ニュー ハイ* バランス* ファイブ フル* ベスト* ポイント* メーン メリット ユニーク* ライン リアル* リード*	4 オフィス ガイド* スター モニター*	27 アイディア アウト イベント イメージ* イン インタビュー* ゲーム コミュニケーション* コメント* サイン* ストレス スピーチ* センス デザイン* トレーニング* ニュース ノー ヒント プラン フリー* マナー マニュアル ミス* メール* メッセージ メモ* ワーク	15 ウインドー エンジン カー カード カメラ ガラス コンピューター スクリーン テープ テーブル ネット ビデオ ペーパー ボタン マシン	3 オレンジ ピンク ブルー*	26 エネルギー オーバー* オープン* カラー クラス クリア* ケース サービス* シーン センター ソフト* タイプ ダウン* ダブル チェック* トップ トラブル ナンバー バック* バラエティー ヒット* ファースト ボックス ライフ リーダー リズム	5 アクセス* キロ** ザ サン メートル**	109
共通S偏重型	4 キャラクター スケジュール スタート* ペース	7 チーム ファミリー プロ ボランティア メンバー レストラン	16 アドバイス* アルバイト* エピソード オーケー* クラシック* コピー*	13 エレベーター コーヒー テレビ ドア トラック バス		9 インターネット カット* コース スーパー ノート ハード*	2 ウイーク センチ**	51

		ワールド	ストーリー スポーツ* タイトル テーマ ドラマ ビジネス マスコミ メニュー ライブ リラックス*	パソコン パンフレット ベッド メーター ラジオ ルーム ワープロ		ホーム ボール レコード		
特徴S	1	3	5	12	4	2	1	28
	リフレッシュ*	オーナー スナック プロデューサー	エアロビック ジョギング* マラソン レース レンタル*	クーラー シャワー スープ スカート ズボン デザート テント バイク バッグ ハンバーグ メダル ロープ	ゴールデン ジャングル フルーツ ミネラル	マージャン リゾート	グッズ	
計	50	15	59	43	7	48	15	237

*サ変／形状詞可能名詞・形状詞系、**助数詞可能名詞・数詞系

６．まとめ

以上、本稿では、「無性格な」外来語を抽出し、その語彙的特徴についてみてきた。その際、高頻度語を特定するだけではレジスターやジャンルの影響を排除できないため、レジスター横断性・ジャンル横断性という散らばり度に留意した。さらにレジスター偏重度を調べ、無性格語のなかでもレジスターによる出現度数の偏りが少ない語を特定した。このようにして抽出した無性格語は基本度が高く、他のコーパス調査の結果とも整合性が高いのではないかと推測される。

分析の結果、ジャンル横断性もレジスター横断性も高い無性格語は、他の語群と比べて「1 抽象的関係」「3 人間活動（精神・行為）」「多義」の割合が高い反面、「4 具体物・道具」「5 自然物・自然現象」の割合は著しく低かった。表 9 を見ると、「4 生産物・道具」は主に具体語が分類されていることから、その割合が相対的に低いということは、裏を返せば、対立する抽象語の割合が高いということでもある。これは、明治後期において基本語化した漢語の 3 類型の一つとして「抽象概念を表す語」を挙げた田中（2014）の考察と共通する部分がある（ただし、「基本語」や「抽象的」の定義は完全に同じではない）。具体的な意味を持つ語よりも抽象的な意味を持つ語のほうが使用頻度や使用範囲が拡大しやすいということは直観的にも理解しやすい。金（2011）は新聞において通時的増加傾向を見せる外来語は抽象名詞に多いとし、その一例である「ケース」が意味範囲を拡大させながら類義語のなかで出現率を伸ばしていることを指摘したが、抽象的な意味を持つ語にはこうした意味範囲の拡大、あるいは変化を通じて使用頻度や使用範囲を拡大させる潜在性があるのかもしれない。

品詞に関しては、無性格語は、それ以外の語群と比べて、サ変可能名詞や形状詞可能名詞などといった普通名詞（一般）以外の品詞を多く含むことがわかった。この傾向も明治後期以降基本語化した漢語と類似している（田中 2012）。このことは、外来語が名詞だけではなく動詞系や形状詞・形容詞系といった品詞カテゴリーにおいても広がりを見せてい

ることを示唆するものである。しかし、これを確かめるには、個々の用法を吟味してサ変動詞用法や形状詞用法のみを取り出し、そうした用法が実際にどれほどあるのかをみなければならない。そうした側面を調べるために、久屋（2014）では、サ変可能名詞である「サポート」、「イメージ」、「キープ」、「マスター」、「スタート」などのサ変動詞用法だけを取り出し、これら外来語に対応する既存類義語である和語動詞や漢語サ変動詞用法との量的関係を調べた。その結果、既存語に対する外来語の使用率が若年層を中心に増加していることが明らかになった。

今回抽出した基本度の高い外来語の語彙的特徴は、明治後期以降に基本語化したかつての借用語である漢語のそれと類似する部分がある。ということは、こうした語彙的特徴は、外来語に限らずあらゆる語種にとって基本語化の重要な要素である可能性がある。いずれにせよ、こうした外来語の広がりが、同じような語彙的特徴を持つ漢語や和語にどういった影響を及ぼしているのかについては、外来語・漢語・和語の語種全体を巨視的に眺めた研究が望まれるところである。この点に関しては今後の課題とする。

謝　辞

本稿で分析に利用した CSJ および分類語彙表関連データは、筆者が国立国語研究所に特別共同利用研究員として滞在していた期間（2014 年 9 月～現在）に、同研究所の許可を得て使用させていただいたものである。ここに感謝申し上げる。

文　献

小木曽智信、中村壮範（2011）「『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装 改訂版」国立国語研究所内部報告書 LR-CCG-10-06

小木曽智信、中村壮範（2014）「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」『自然言語処理 21:2』, pp.301-332

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』

久屋愛実（2014）「外来語の共時的分布パターンの一般化に向けた予備的考察」『韓国日本語学会第 30 回国際学術発表大会予稿集』, pp.156-165

国立国語研究所（2004）『分類語彙表―増補改訂版』大日本図書

国立国語研究所（2006）『日本語話し言葉コーパスの構築法』

田中章夫（1973）「自動抄録処理におけるキー・ワードの性格」『電子計算機による国語研究 V』, pp.141-184, 国立国語研究所

田中牧郎（2012）「明治後期から大正期の語彙レベルと語種―『太陽コーパス』の形態素解析データによる」田中牧郎ほか（2012）『近代語コーパス設計のための文献言語研究 成果報告書』国立国語研究所共同研究報告 12-03

田中牧郎（2014）「明治後期における漢語の基本語化」『第 6 回コーパス日本語学ワークショップ予稿集』, pp.193-200

水谷静夫（1964）「語の基本度」『現代雑誌九十種の用語用字―第三分冊（分析）』, pp.7-51, 国立国語研究所

『児童・生徒作文コーパス』の設計

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）

Design of 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

本研究では、児童・生徒の作文能力の実態を明らかにするため、子どもたちが作成した生の作文を原本とした『児童・生徒作文コーパス』の構築を進めている。本コーパスは、協力校 4 校（小学校 2 校、中学校 2 校）9 学年の全クラスを対象に 3 年間にわたって作文資料を収集・電子化するもので、最終的に 300 万形態素規模のコーパスになる予定である。同時期に同条件（題や作文時間の指定）で作文を作成させることによって資料の均質性を保証している点、複数年の継続調査により同一児童・生徒の作文能力の変化を追跡できる点が特徴である。本コーパスを利用した研究によって、児童・生徒の学齢別の作文能力の実態や発達を明らかにし、現場の教員の作文指導の手本となる資料の作成を目指す。また、本コーパスの構築と合わせて、独自の検索システムの開発も同時に行っている。現段階の検索システムは、単純な文字列検索が行えるに留まるが、今後システムを更新して、高度な検索処理をできるようにする。本発表では、コーパスの基本的な設計方針、内容の概要、検索システムの紹介を行い、コーパスを活用した研究の展望を述べる。

１．はじめに

近年、コーパスを利用した言語研究が盛んになってきている。国語教育学研究でも子どもたちの書いた作文を資料とした作文能力の実態調査や指導法の開発などが行われている。しかしながら、後者の資料となる児童・生徒の作文でコーパスとして利用可能なものは、資料の収集や公開の難しさから質量ともに十分ではなく、十分な研究環境が整っているとは言いがたい。そのため、本研究では小中学校の児童・生徒の作文を 3 年間に亘って収集し電子化する大規模な作文コーパスの構築を進めている。本発表では現在構築中の『児童・生徒作文コーパス』（以下、「児童作文コーパス」と略す）の目的と概要を説明し、今後の研究の展望を示す。

２．児童作文コーパスの必要性

小中学校における現在の作文指導は、多くの場合子どもたちの書いた文章に教員が手を入れて書き改めさせるという方法で行われている。この指導法には次の 2 点で問題がある。

(1)　文章の修正（指導）が教師個人の語感によって主観的になされていること。
(2)　子どもによる作文の推敲が、教師による書き換え例を丸写しすることに留まり、なぜ直すのか、他にどのような表現があるのかなどの検討が行われていないこと。

† miyagi@edu.u-toyama.ac.jp

（したがって、子ども自身の作文推敲能力が育たない）。

これらの問題は、教師個人のひいては教育現場全体における経験知の不足、またそれを補い補正していく資料の不足によるものと考えられる。

作文指導には特定のマニュアルがあるわけではなく、現場依存的である。また、当然ながら子どもたちの作文能力は個々で異なっている。ベテラン教師は、勘を働かせて上手に子どもたちを誘導し、それなりの文章に推敲させることができるが、経験の浅い教師は、このような技術を持たないため、ベテラン教師の助言や手本となる用例集などの資料が必要になると考えられる。ここでいう手本となる資料は以下の要件を満たす必要がある。

(3) 子どもたちの発達段階を考慮した、相対的な基準を提示できるものであること。
(4) 文章を特定の型に揃えることを目標とするものではないこと。（言葉狩りを推奨するものではない。）
(5) 子どもたちが理解できる理由で説明がなされること。

以上の要件を満たす資料を構築するためには、まず、発達段階に応じた子どもたちの書く作文の実態（語彙や文構造、段落構成など）を知る必要がある。そのため、本研究では、子どもたちの書く作文の実態を明らかにし、それに基づいて指導資料を開発するための基礎的な研究資料として児童作文コーパスを構築する。

３．作文コーパスの設計と基本方針

３．１　作文コーパスの特徴

本コーパスは、調査協力校 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ものである。作文は「夢」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。電子化は以下の指針に従って行う。

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）

個人情報保護の理由から、収集した作文原本は非公開とし、テキストデータは範囲を限定して利用を認める。本コーパスの現在の公開範囲は限定的であるが、児童・生徒の個人情報に関する処理を施した後、学術的研究、特に学校現場への還元を目的とした研究に利用する場合での一般公開が可能になるよう協力校に交渉中である。

３．２　作文コーパスの構成

本コーパスは本文テキストとメタデータで構成される。メタデータは本文テキストには含まず、ファイル名と紐付けて別に管理する。メタデータは以下の項目を含む。

作文課題の属性　　課題 ID、実施年度、テーマ（タイトル）
執筆者の属性　　著者 ID、学校 ID、学年、クラス、性別

作文課題の実施、収集は年 2 回行い、3 年間継続する。2015 年 1 月現在、2014 年度分の課題 2 回について実施済みであり、電子化作業を進めている。

表 1　作文課題の実施計画

年度	2014		2015		2016	
課題	課題 1	課題 2	課題 3	課題 4	課題 5	課題 6
進捗状況	実施済	実施済				

最初の作文課題（課題 1）について、48 クラス分の作文原稿の収集と、23 クラス分のテキスト入力、11 クラス分のチェック作業が完了している。テキスト入力済みの 23 クラス分のデータについて、文分割と形態素解析処理を行い、文数、形態素数、文字数（改行文字を除く）を集計した結果を以下に示す。形態素解析処理には MeCab 0.996[1]と UniDic 2.1.2[2]を使用した。学年別集計は 5.1 節を参照されたい。

表 2　課題 1 の概要（23 クラス分）

学年	クラス数	作文数	文数	形態素数	文字数
小 1〜中 3	23	813	11046	237940	378652

23 クラス分のコーパスの形態素数が約 24 万なので、48 クラス分で約 50 万形態素、6 回の作文課題で最終的に 300 万形態素程度の規模のコーパスになる見込みである。

３．３　既存コーパスとの比較

児童・生徒の書き言葉を対象としたコーパスは全国の地域文集 10 年分を収集し約 47 万形態素規模のコーパスを構築した国立国語研究所(1989)などを例外として従来あまり多くなかったが、近年報告が増えている。永田他(2010)は小学 5 年生 81 人の読書ブログを 8 カ月間記録した約 4 万形態素規模のコーパスで、ブログの更新履歴を追跡できる点、一般公開されている点が特徴である。坂本(2010)は全国の小学校 265 校の Web ページで公開されている児童作文を収集した 123 万形態素規模のコーパスである。学校名、県名、学年、性別などの情報が確認できる限り付与されており、一部については著作権処理が完了しているという。鈴木他(2011)は中等教育学校の 1 年から 5 年(中 1〜高 2)の冬休みの宿題作文を記録した約 25 万語規模のコーパスである。藤田他(2012)は神奈川県内の小学校 9 校で 2 回に分けて収集した作文 672 編からなるコーパスである。表記や文法の誤りなどの指摘事項と

[1] http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[2] http://sourceforge.jp/projects/unidic/

評価が付与されている点が特徴である。

表 3　既存の作文コーパスとの比較

コーパス	国語研(1989)	永田他(2010)	坂本(2010)	鈴木他(2011)	藤田他(2012)	本コーパス
形態	作文	読書ブログ	作文	作文	作文	作文
形態素数	474,243	39,269	1,234,961	249,918	不明	3,000,000
調査対象	小 1〜小 6	小 5	小 1〜小 6	中 1〜高 2	小 4	小 1〜中 3
調査期間	10 年	8 カ月	2 年	1 カ月	1 年	3 年
収集方法	文集収集	活動記録	Web 収集	課題調査	課題調査	課題調査
備考		公開済	著作権処理済（一部）		誤用、評価情報つき	

本コーパスはコーパスの規模が約 300 万形態素と既存の作文コーパスと比べても最大規模である点、義務教育課程（小 1〜中 3）の全体をカバーしている点、同一の調査対象に対して 3 年間継続して調査を行う点などが特徴である。一方で、特定の学校のみを調査対象としているため、必ずしも全国の児童、生徒作文全体に対する代表性を保証しているわけではない点、構築したコーパスを研究目的で公開し、共有する方法が確定していない点などに課題が残る。

4．児童作文検索システム

本コーパスの構築に合わせて、「児童作文コーパス」のデータを検索するシステム（以下、「検索システム」とする）を開発する。検索システムを独自に開発する利点は、コーパスの仕様変更（5 節を参照）に合わせて、適切な検索が実行できるように検索システムを改修することができる点である。また、本コーパスは教育現場での利用も視野に入れており、現場の教員が手軽に検索を行えるインターフェイス設計を指向している。以下、検索システムの現在のバージョンにおける概要を示す。

図 1　基本操作画面

検索システム（図 1）には以下のような検索項目がある。組み合わせて、検索したい作文の条件を設定する。

検索学年　（全学年／中学校／小学校／学年指定）
検索作文　（全作文／生活作文／意見文）
検索性別　（男／女）

現在のバージョンは単純文字列検索である（正規表現には対応していない）。検索条件に文字列を指定すれば結果が得られるようになっている。

図 2　検索結果（一覧表示）

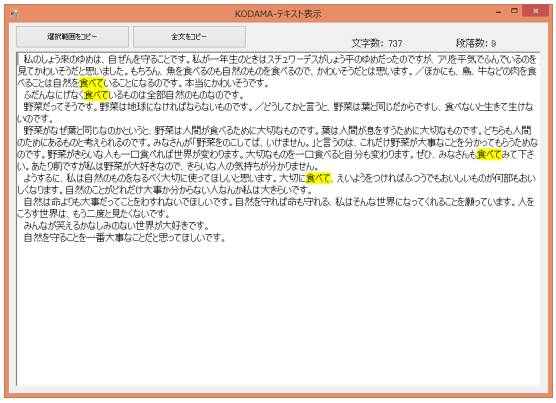


図 3　検索結果（個別・全文表示）

検索条件を「食べて」に指定して検索すると、検索された一文（句点から句点までの文字列）が一覧表で表示される（図 2）。合わせて、各文の横には作文した児童・生徒の学年も付される。また、一覧左上の「検索リストをコピー」をクリックすると、結果一覧を excel に直接貼り付けることができる。

結果一覧の任意の文をクリックすると、当該の文が検索された作文の全文が表示される（図 3）。作文全文の中の検索した文字列は、例の「食べて」のように黄色で反転表示される（複数の候補がある場合、すべての文字列が対象となる）。画面左上の「選択範囲をコピー」や「全文をコピー」をクリックすることによって、excel や word などに、当該の文章を直接貼り付けることができる。また、画面右上に、当該作文の段落数（改行記号の数）や文字数（記号も 1 字と数える）も表示される。

５．作文コーパスの展望

５．１　作文コーパスの今後の展開

本コーパスは平文テキストとメタデータの形式で構築しているが、今後、研究利用可能な言語学的情報の付与を進めたい。現時点では、自動処理による形態論情報（短単位、長単位）、文節境界情報、構文情報（係り受け情報）の付与を試行している。課題 1 の 23 クラス分のデータについて、各種情報を学年別に集計した結果を表 4 に示す。前述の MeCab 0.996、UniDic 2.1.2 の他、長単位と文節は Comainu 0.70[3]、構文解析は CaboCha 0.68[4]を使用している。

[3] http://sourceforge.jp/projects/comainu/
[4] https://code.google.com/p/cabocha/

表 4　課題 1 の学年別集計（23 クラス分）

		小 1	2	3	4	5	6	中 1	2	3	計
作文数		104	100	100	80	68	110	74	68	109	813
段落数		227	281	396	365	365	646	362	313	536	3491
文数		649	791	1263	997	1030	1919	1214	1334	1849	11046
文節数		4206	5342	9573	7577	7629	15988	10340	10360	16562	87577
長単位数		10111	13541	23051	18658	18636	38562	24487	24579	39098	210723
短単位数		11271	15149	26070	20953	21123	43680	27600	27643	44389	237878
品詞	名詞	2024	2687	5437	3880	4132	8836	5641	5412	9188	47237
	代名詞	220	393	638	537	528	1055	682	714	1136	5903
	形状詞	149	204	260	235	286	518	359	398	629	3038
	連体詞	70	76	205	195	214	452	339	331	544	2426
	副詞	284	384	557	451	479	924	554	593	813	5039
	接続詞	30	55	82	91	103	197	136	139	236	1069
	感動詞	130	63	58	90	53	75	26	31	29	555
	動詞	1400	1893	3402	2993	2888	6205	4021	4131	6559	33492
	形容詞	261	255	475	383	395	719	501	520	785	4294
	助動詞	1643	1932	3129	2494	2412	4809	3097	3154	5009	27679
	助詞	3126	4199	7450	5964	5857	12654	8078	8217	13186	68731
	その他	1934	3008	4377	3640	3776	7236	4166	4003	6275	38415
語種	和語	8545	10579	18756	15212	14740	30956	19950	20438	32392	171568
	漢語	982	1449	2645	2012	2418	5189	3352	3178	5779	27004
	外来語	167	455	833	460	472	854	423	294	545	4503
	混種語	71	106	154	109	135	238	164	180	282	1439
	固有名詞	74	93	217	91	148	224	149	84	195	1275
	記号	1428	2432	3415	3065	3164	6189	3556	3457	5188	31894
	その他	4	35	50	4	46	30	6	12	8	195
文字数		20070	26089	43987	33767	33701	67827	42470	42432	68309	378652
文字種	ひらがな	18008	20415	32518	24279	22094	43550	26523	27400	43310	258097
	カタカナ	497	1765	3313	1768	2115	3485	1980	1266	2353	18542
	漢字	128	1237	4153	4315	5924	14018	10121	10081	17144	67121
	その他	1437	2672	4003	3405	3568	6774	3846	3685	5502	34892

これらの情報を用いると、言語単位の比、品詞や語種の比、文字種の比などについて、学年別に調べることができる。例として、作文あたりの平均文数、文あたりの平均短単位数(平均文長)、MVR[5]、漢語比率、漢字比率を表 5 に示す。学年が上がるにつれて平均文数、平均文長、漢語比率、漢字比率などが増加すること、MVR が減少することなどが観察できる。

表 5　学年別の言語単位、品詞、語種、文字種比率

	小 1	2	3	4	5	6	中 1	2	3	平均
文/作文	6.24	7.91	12.63	12.46	15.15	17.45	16.41	19.62	16.96	13.59
短単位/文	17.37	19.15	20.64	21.02	20.51	22.76	22.73	20.72	24.01	21.54
MVR	0.55	0.49	0.44	0.42	0.48	0.42	0.44	0.45	0.42	0.44
漢語/短単位	0.09	0.10	0.10	0.10	0.11	0.12	0.12	0.11	0.13	0.11
漢字/文字	0.01	0.05	0.09	0.13	0.18	0.21	0.24	0.24	0.25	0.18

より高度な言語学的情報としては、文の成分（主語、述語、修飾語など）、係り受けの

[5] (形状詞+連体詞+副詞+形容詞)/動詞で計算した。

種類（並列など）、節の種類などの文法情報や、誤用情報などの付与がある。文法情報は、文の複雑さを評価するために必要となる。誤用情報は、発達段階別の誤用実態の分析や指導資料の開発のために必要となる。こうした研究の展望については、次節を参照されたい。

５．２　作文コーパスを用いた研究の展望

現時点での児童作文コーパスおよび検索システムの概要は以上である。児童作文コーパスによって明らかにされる子どもたちの作文活動の実態と研究の展望について言及する。

①　学習漢字の使用の実態

子どもたちが作文で使用する漢字は、多くの場合授業で学習済みのものであると推測される。表 6 は学年別の使用漢字を集計し、1 万文字あたりで示したものである。

表 6　学年別使用漢字（1 万文字あたり）

	小 1	2	3	4	5	6	中 1	2	3	平均
1 年配当漢字	12.5	319.3	376.5	379.1	434.7	472.2	521.3	497.7	515.2	430.7
2 年配当漢字	15.4	154.5	389.2	478.3	565.3	605.2	668.2	683.7	699.3	535.8
3 年配当漢字	9.0	0.4	147.1	258.5	382.2	419.4	461.5	409.6	435.5	326.0
4 年配当漢字	9.0	0.0	8.4	83.8	158.5	221.4	251.7	240.1	248.4	162.7
5 年配当漢字	5.0	0.0	2.3	19.0	103.3	160.6	202.3	239.7	262.9	137.1
6 年配当漢字	3.5	0.0	15.7	44.4	73.6	125.0	159.9	167.8	157.2	100.0
非配当漢字	9.5	0.0	5.0	14.8	40.4	62.8	118.2	137.2	191.2	80.4
合計	63.8	474.1	944.1	1277.9	1757.8	2066.7	2383.1	2375.8	2509.8	1772.6

表 6 を見ると、学年が上がるに連れて漢字の使用頻度が増加すること、低学年では未習漢字の使用は稀だが学年が上がるにつれて未習漢字の使用頻度が増加することなどが観察できる。児童の作文の中には、様々な外的要因によって、学習前の漢字で書くことが多い語句や、学習後でも仮名書きのままで書くことが多い語句が混在している可能性があり、非配当漢字の使用状況も併せて、詳しく調査する必要がある。また、各学年の使用漢字の比率を見ると、高学年であっても 3 年生までの配当漢字の使用比率が高い。これは使用頻度が高い語彙に使われる漢字が 3 年生までに配当されていることの帰結である可能性があり、語彙の分布と合わせて調査する必要がある。児童作文コーパスのデータと学習漢字の学年配当表を照らし合わせることによって、多くの児童に共通して観察される学習漢字の配当と使用実態のずれを明らかにすることができる。

②　接続詞の使用や文の展開の傾向性

低学年の児童の書く作文では、ある段階から「それで」や「あと」などの接続詞の使用が多くみられる（小学校中学年頃から論理的な文章を書けるようになるため「しかし」などの使用が増えるとの指摘もある）。その後、子どもたちは段階的に接続詞の種類と使用頻度を増やしていくが、ある段階から不要な接続詞の使用を控えるようになる。児童作文コーパスを使用することで、その変化を追跡する調査をすることができる。表 7 は、接続詞の学年別出現頻度を集計し、上位 10 語を 1 万形態素あたりで示したものである。接続詞は短単位では複数の語に分割されるものも多いため（表 7 の「でも」「だから」「すると」「ですが」「それから」など）、長単位で集計している。

表 7　接続詞の学年別出現頻度（長単位 1 万形態素あたり・上位 10 語）

	小 1	2	3	4	5	6	中 1	2	3	平均
ソシテ	10.9	17.7	20.8	27.9	24.1	23.1	23.7	14.2	17.6	20.5
デモ	19.8	24.4	24.7	22.5	17.2	17.6	16.7	16.7	8.4	17.4
シカシ	0.0	0.0	3.9	2.7	5.4	7.3	11.8	24.0	25.3	11.3
ダカラ	4.9	7.4	11.7	8.0	12.3	12.2	10.2	9.4	10.5	10.3
マタ	4.0	3.0	3.0	6.4	14.0	11.7	12.3	8.5	10.5	9.0
スルト	0.0	0.0	2.2	3.8	4.3	2.3	6.1	1.2	1.0	2.4
ケレド	0.0	3.0	4.3	3.8	5.4	1.6	2.0	2.0	0.8	2.4
デスガ	0.0	0.7	0.9	2.1	1.6	3.6	1.6	1.2	1.5	1.8
ソレカラ	2.0	0.0	0.9	4.3	1.1	1.8	2.5	0.8	1.3	1.6
タダ	0.0	0.7	0.0	1.1	1.1	1.0	2.5	2.0	3.1	1.5

「あと」は自動解析では接続詞ではなく名詞として解析されるため、個別に名詞用法、副詞用法、接続詞用法などの区別を判断し、集計する必要がある。参考として、それらの区別をせずに「あと」の出現頻度を集計したものを表 8 に示す。

表 8　「あと」の学年別出現頻度（長単位 1 万形態素あたり）

	小 1	2	3	4	5	6	中 1	2	3	平均
アト	44.5	10.3	9.5	4.8	3.2	4.4	2.9	2.0	1.5	6.2

表 7 と表 8 を見ると、学年が上がるにつれて「しかし」の使用頻度が増加すること、「でも」「あと」が減少すること、「そして」「だから」「また」が一度増加したのち減少することなどが確認できる。一方で、「すると」「けれど」などのように習得後もあまり定着しない（使用されない）接続詞もあり、文の展開や類似する接続詞との棲み分け意識などにも注目して分析を進める必要がある。現在のコーパスの規模では用例数が少なく、十分な分析をすることができないが、今後、コーパスの規模を拡充することによって、より詳細な分析を進めることができる。また、接続詞に限らず接続表現全体を視野に入れた（接続助詞を含む）節の複雑化に関する作文能力の変化についても実態を明らかにすることができる。

③ 文構造の複雑化に関する発達

子どもたちは発達段階に応じてどの段階でどのような複雑さの文を作文することができるのか、またどの順で文の構造を複雑化させていくのか（修飾・接続関係の習得順序）などの実態を明らかにすることができる。例えば、連体修飾と連用修飾ではどちらの方が、より早く複雑化する傾向にあるのか、また最終的にはどちらの修飾関係の文が作文されやすいかなど、子どもたちの作文表現の実態を明らかにすることができる。この研究は、いわゆる「だらだら文」（長すぎる文やくどく感じる過修飾文、主述の不対応やねじれがある文）の認定や原因の究明に寄与することも期待される。

この研究のためには、既存の構文解析器で付与可能な係り受け情報に加えて、連体、連用など係り受けの種類に関する情報や、主語、述語など文の成分に関する情報の付与が必要になる。現在、これらの情報を自動付与するスクリプトの作成を進めている。図 4 は文法情報の自動付与し、結果を可視化したものである。

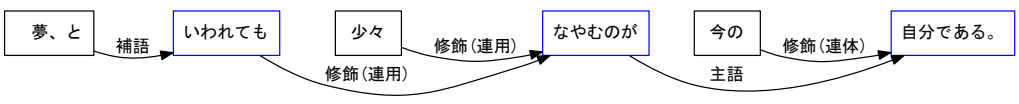


図 4　係り受けの種類と文の成分の付与

表 9 は、このスクリプトにより付与した係り受けの種類を集計し、1 万文節あたりで示したものである。

表 9　学年別の係り受け分類（１万文節あたり）

	小 1	2	3	4	5	6	中 1	2	3	平均
主語	485.0	481.1	509.8	595.2	550.5	542.9	589.9	551.2	563.3	548.3
修飾(連体)	1738.0	1576.2	1840.6	1893.9	1964.9	2056.5	2295.9	2305.0	2469.5	2102.0
修飾(連用)	2758.0	2708.7	2704.5	2636.9	2654.3	2714.5	2412.0	2643.8	2450.2	2609.1
接続	1150.7	1471.4	1166.8	1293.4	1293.7	1220.3	1184.7	1044.4	1057.2	1183.3
独立	195.0	69.3	38.7	68.6	48.5	25.0	18.4	21.2	10.9	39.3
補語	2099.4	2154.6	2309.6	2168.4	2116.9	2189.1	2298.8	2143.8	2296.8	2215.8

表 9 を見ると、学年が上がるにつれて主語や連体修飾語が増加すること、連用修飾語が減少することなどが分かる。ただし、このスクリプトはまだ試験的な段階であり、上記のデータは十分に信頼できるものではない。今後、文法情報の付与作業と検証、修正を進め、文の複雑さの評価やねじれ文の自動検出の研究へと繋げたい。

④ 誤用の実態と作文の傾向性

児童作文コーパスのデータは、多くの表記や仮名遣いの誤り（例えば低学年の児童であれば「ごはんおたべた」のような誤りがある）、語句や文法の誤用が、原本に忠実に記録されている。コーパスを使用することによって、これらの誤りが学齢の進行に伴って、質的または量的にどのように変化していくのか、具体例の提示に加えて数量的な傾向性も明らかにすることができる。また、これまでの研究は、語句レベルでの誤用の指摘が中心であり、それ以外では文の主述のねじれの提示など文レベルでの誤用に留まるものが多かった。今後は、文同士の連続の自然さや段落のつなぎ方、すなわち文の結束性の研究など、比較的大きなレベルでの誤用や不自然さの研究も進めていく必要がある。

この研究のためには、コーパスへの誤用情報の付与と数値化が必要である。今後、誤用情報付与の設計と計画を進めていきたい。

本研究で構築する児童作文コーパスは、以上のような研究課題の究明に寄与する資料として活用が期待できる。

６．まとめ

本発表では、児童・生徒の作文能力の実態を映した『児童・生徒作文コーパス』と検索システムの構築について計画と現在の状況を説明し、児童・生徒の作文能力の発達過程の数値化・視覚化など、コーパスを用いた言語研究の展望を示した。本コーパスは義務教育課程 9 学年の作文活動を 3 年間に渡って継続的に調査する 300 万形態素規模（予定）の作文コーパスであり、児童・生徒の作文を収集したコーパスとしては、データの均質性と規模において従来例のない画期的な資料である。また、本コーパスと併せて平易なインターフェイスを備えた検索システムの開発を進めている。今後は、コーパスの構築と並行して、研究利用のために必要な言語学的情報の付与と、検索システムの改良を進めたい。

本研究の最終的な目標の一つは、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。言語研究の立場から現場の教師が手軽に利用できる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、博報財団第 9 回児童教育実践についての研究助成「学校現場との協働による児童作文指導の基礎的研究」（2014 年度、研究代表者：冨士原紀絵、助成番号：2014042）、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

国立国語研究所(1989)『児童の作文使用語彙（国立国語研究所報告 98）』東京書籍.（http://www.ninjal.ac.jp/s_data/drep/report_nijla/R0098.PDF よりダウンロード可能）

坂本真樹(2010)「小学生の作文コーパスの収集とその応用の可能性」『自然言語処理』17:5、pp.75-93.（https://www.jstage.jst.go.jp/article/jnlp/17/5/17_5_5_75/_pdf よりダウンロード可能）

鈴木一史、棚橋尚子、河内昭浩(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）

永田亮、河合綾子、須田幸次、掛川淳一、森広浩一郎(2010)「作文履歴をトレース可能な子供コーパスの構築」『自然言語処理』17:2、pp.51-65.（https://www.jstage.jst.go.jp/article/jnlp/17/2/17_2_2_51/_pdf よりダウンロード可能）

藤田彬、田村直良(2012)「作文事例に基づいた児童の「書くこと」に関する学習傾向についての分析−小学四年生による紹介文・感想文を中心に−」『言語処理学会第 18 回年次大会発表論文集』pp.987-990.（http://www.anlp.jp/proceedings/annual_meeting/2012/pdf_dir/D4-3.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』のコーパスデータにおける短単位認定の諸問題

渡辺由貴・市村太郎・鴻野知暁（国立国語研究所コーパス開発センター）

Problems Concerning the Recognition of Short-Unit-Word in the Toraakira-bon Kyogensyuu Corpus

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

要旨

『虎明本狂言集』のコーパスデータの作成・整備過程で，語（短単位）の認定を行う必要があるが，その際に困難が生じる場合がある。例えば，仮名で表記された同音の語の認定，活用語尾が表記されていない語の音便形の認定，形容詞連用形の文語活用・口語活用の認定等である。

同音の語については，底本の用例の状況や校注者の傍記を元に認定を行った。音便形の認定については，四段活用動詞のうち連用形の用例数の多い語について後接語別の音便状況の調査を行い，明らかに読みが予想できる例以外については，「た」が後接するもののみを音便形とし，それ以外の語が後接するものについては無理に音便形を認めない方針とした。形容詞の活用の認定については，形容詞の終止形活用語尾と連体形活用語尾の状況を調査した結果，形容詞の口語活用化が進んでいたと判断し，明らかな文語活用の例を除き，口語活用を原則とした。

１．はじめに

国立国語研究所『日本語歴史コーパス』構築の一環として進められている『虎明本狂言集』のコーパスデータの作成・整備過程で，語（短単位）の認定を行う必要があるが，1642年に成立した『虎明本狂言集』は，その言語事象が古代語から近代語・現代語への過渡的特徴を示しており，語の認定において困難が生じる場合がある。

例えば，「異見―意見」「時宜―辞儀」のように，類似した二つ以上の語が『虎明本狂言集』の成立時期に混在し，時に混同され用いられていることがあるが，このような語についても，コーパスデータ上は，いずれかの語と認定することが必須となる。

また，活用語の音便形についても類似した問題がある。例えば，『虎明本狂言集』において，「いたいて―いたして」のように音便形と非音便形の両表記形がみられる語があるが，「致て」のような活用語尾の表記されていない語形があらわれた場合，その活用形を音便形か非音便形かのいずれかに認定する必要がある（市村 2014，pp.106-107）。

形容詞の活用型についても，『虎明本狂言集』においては文語特有の活用語尾「－し」「－き」と口語特有の活用語尾「－い」の両形が見られ，例えば形容詞「長い」の連用形「ながく」を，文語活用か口語活用かのいずれかに認定しなくてはならない。

これらの問題については，底本の注釈や索引，各種辞書の記述，研究論文等が参考になるが，これらを参照しても『虎明本狂言集』におけるそれぞれの語を確定するには至らないこともある。例えば，注釈や索引において，“二つの語のどちらの可能性もある”という示し方がされている場合があり，これは実態に即した記述であるが，コーパスデータにおいてはそのような曖昧な処理はできない。さらに，本コーパスは，『日本語歴史コーパス』の中の一つのコーパスであるため，中古語から近代語，現代語のコーパスで蓄積され

たデータの中に位置づける必要がある。

本発表では，『虎明本狂言集』のコーパスデータにおいて語の認定が難しい事例をとりあげ，注釈や索引，辞書等を参照しながら検討したい。

２．意味・用法の類似する同音の語の認定

『虎明本狂言集』においてみられる，意味・用法が類似する同音の語の認定について検討するにあたり，まず，国立国語研究所のコーパスデータにおける同語異語判別の方針を確認しておく。コーパスデータは自動形態素解析を前提としており，その精度を保つために，コーパスデータにおける同語異語の判別については次のような方針が立てられている。

> 方針１：同表記異語を生じさせるような語彙素の立て方はできる限り行わない。
> 方針２：複数の語彙素に分ける場合は，明確な基準・理由をもってし，人手で正確に区別できないような語彙素の分割は行わない。　（小椋他 2011，p.137）

現代語のコーパスデータにおいては，BCCWJ から取得した頻度情報や，『岩波国語辞典』第 6 版，『国語大辞典』，『大辞林』『広辞苑』における見出しの立て方等を考慮しながら同語異語判別を行っている。例えば，動詞「アウ」については「合う」「会う」の二つの語彙素を立て，「逢う・遭う・遇う」は「会う」の書字形としている。動詞「オサマル」については，「収まる」のみを語彙素として立て，「治まる」「納まる」「修まる」等は全てその書字形としている（小椋他 2011，pp.137-140）。

しかし，中古・近代および現代語のコーパスを作成する過程で，別の語彙素として登録されている二語が，『虎明本狂言集』においては明確に別語であると判別できない場合が存在する。資料の成立時期に，類似した二つ以上の語が混在し，時に混同され用いられているケースがあること，表記にゆれがあること，現代と中近世とでその語の意味や表記が異なっているケースがあること等がその理由で，上記の基準では語の判別・認定に迷うことがある。例えば，「意見―異見」，「憂き世―浮き世」，「辞儀―時宜」，「卑怯―比興」等がそれにあたる。以下，「意見―異見」を例に見ていく。

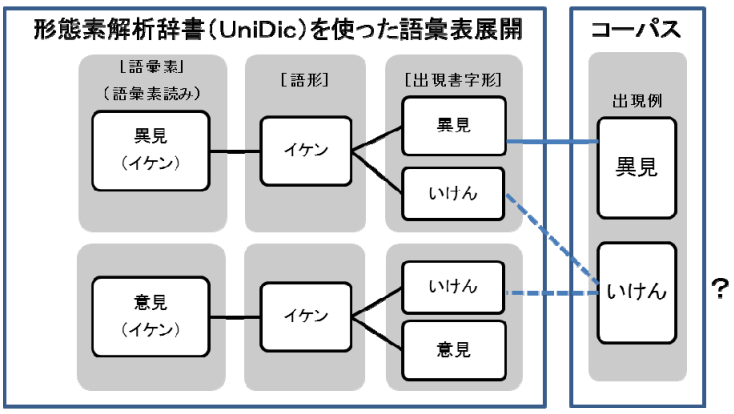


図１　同音異義語の認定　「意見―異見」

『虎明本狂言集』において見られる［語彙素読み］が「イケン」の語は，既登録の「意見」「異見」のいずれかに相当するものである。現代語において，「異見」は，特に他者

と異なった見解である場合に用い，また，その意味の場合のみ「異見」の表記をするのが一般的であると考えられるため[1]，「意見」であるか「異見」であるかは意味・表記にしたがって判別することが可能であろう。また，古くは「意見」が「本来は政務などに関する衆議の場において各人が提出する考えであった」（『日本国語大辞典　第二版』）ことから，「意見」と「異見」とは別語と認識されていたようである。しかし，中世頃にはこの両語は混同されて用いられることがあったようであり，明確に区別することが難しく，慎重な判断が必要となる。

『虎明本狂言集』における「イケン」の表記別の用例数は，「異見」表記 7 例，「いけん」表記 11 例である。「方針１：同表記異語を生じさせるような語彙素の立て方はできる限り行わない。」により，「異見」表記の 7 例については「意見」としない方が望ましく，「異見」とするのが妥当であると考えられるが，仮名表記の「いけん」については，漢字表記例に合わせて一律「異見」とするべきか，あるいは「意見」の可能性があるのか，検討が必要になる。

『日本国語大辞典　第二版』および『時代別国語大辞典』は，「異見」と「意見」を同一見出しの中に立てている[2]。また，『日本国語大辞典』の「語誌」によると，「意見」と「異見」は明治になると典拠主義の辞書編纂の立場から別の語とされるようになるが，中世後期の古辞書類や文学作品では「異見」が一般的であったとされており，『時代別国語大辞典』でも「次第に『異見』が『意見』の領域を侵して，両者の区別が失われがちであった」との記述がある。

また，『大蔵虎明本狂言集総索引』の各巻でも，「イケン」の語の見出し字が異なっており，大名狂言，女狂言，萬集類の索引が「意見」（「御意見」「御意見有る」を含む）としている一方，鬼・小名類，出家座頭類，集類の索引は「異見」（「御異見」を含む）としている。また，聟・山伏類の索引については，「ごいけん［御異見・御意見］」と，両方の表記を見出し字としている。「異見」「意見」の両語は区別しがたいものであり，結果的にその巻の担当者の判断によって見出し字が分かれることになったと想像される。

用例を確認すると，「異見」表記の例，仮名表記「いけん」の例とも，「忠告」「助言」の意味と解釈可能な例である。これらの例では，「他者と異なる見解」といったニュアンスは強くなく，現代語であれば「意見」と表記するのが一般的に思われるような例ではある。ただし，『虎明本狂言集』には漢字表記「意見」の例があらわれず，底本の校注者も仮名表記「いけん」7 例のうち 6 例に「異見」と傍記している。さらに，次の例のように，「異見」表記の例と「いけん」表記の例との間に意味の違いは認めがたく，「異見」表記の例を語彙素「異見」とするのであれば，仮名表記「いけん」の例も語彙素「異見」とするのが妥当と考えられる。

(1)（新座の者）今日よりは，かた／＼をよりおや殿とたのみまらする程に，万事よひやうに引まはされて，御【いけん】有てくだされい（鼻取りずまふ 上 p.196[3]）

1　例えば，『岩波国語辞典』（第六版）では，以下のように立項されている。
【意見】①ある問題についての考え。②自分の考えを述べて人をいましめること。
【異見】他と違った考え。

2　ただし，『日本国語大辞典』は「意見・異見」の見出しと別に「異見」も立項している。

3　以下，引用は大塚（2006）による。

(2)（親）　　しつけもなひやつで御ざる程に，今からは萬事御【異見】たのみまらする（二人袴　上 p.417）

(3)（伯蔵主＝狐）かやうにいふて又つつたと云事をきひたらば，二たびてらへもなおりやつそ，その【いけん】いたさうとぞんじて参た

（男）　　近比かたじけなふ御ざる，私をおぼしめせばこそ，さやうの御【異見】をなされてくださるれ（つりきつね 下 pp.418-419）

このように，用例や，校注者の傍記等の状況から，『虎明本狂言集』における「イケン」は全て「異見」と判断するのが妥当であろう。

３．音便形の認定

活用語の音便形の認定についても困難な例がある。市村（2014，pp.106-107）にあるように，『虎明本狂言集』においては「いたいて―いたして」のように音便形と非音便形の両表記形がみられる語があり[4]，「致て」のような活用語尾の表記されていない語形があらわれた場合，その活用形を音便形である「連用形 - イ音便」とすべきか，非音便形「連用形 - 一般」とすべきかについて判断する必要がある。『日本語歴史コーパス』のうち，中古和文のデータにおいては，基本的に活用語尾が仮名表記されているためにこのような問題は起こりにくいと考えられ，これも中世語資料のデータゆえの問題であると言える。

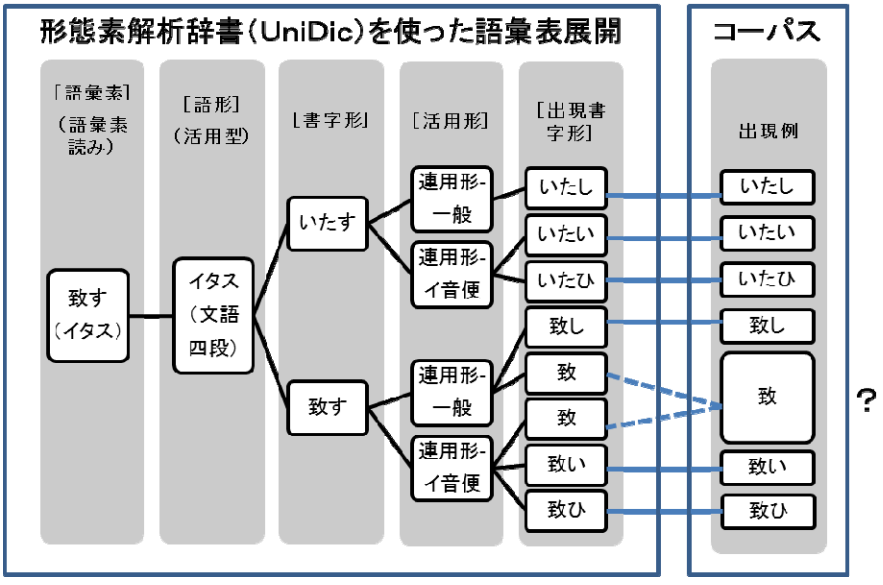


図 2　音便形の認定　「致」

『大蔵虎明本狂言集総索引』においても，例えば「い・ふ〔言ふ〕」の項目を見ると，「ーう（用）」に挙げられた例については，「＊印は『云』と漢字表記のため，音便形か

4　蜂谷(1998)も，「狂言台本では，四段活用・ナ行変格活用動詞の連用形が助動詞『た』『たり』，接続助詞『て』などに続く場合」に音便形となることが多い(p.322)とし，サ行四段動詞の音便化については，「そこには語による相違もある程度認められるが，一方，同じような場面で同じ語が原形とイ音便形とで用いられているものもあり，激しい流動の状況をうかがわせる」(p.323)と述べる。

非音便形か不明」（脇狂言）「＊印は『云』という漢字表記。そのほとんどは卜書きの部分の『云て』の形。『いひて』と読むべきものもあるかもしれない」（聟・山伏類）等の注記があり，やはり活用語尾の表記されていない例については，音便形か非音便形かの判定は難しいことがうかがえる。しかし，コーパスデータにおいては，注をつけることも判断を保留することもできないため，基準を立て，付与する情報を音便形か非音便形かに決定する必要がある。

四段活用動詞のうち，連用形の用例数が多い語（上位 10 語）について，その語の連用形の全用例数および，活用語尾無表記例の数を整理してみると，表 1 のような状況であった。「活用語尾無表記例」は，「云」「参」「申」のように活用語尾が表記されておらず，音便か非音便かの判別が必要な用例の数である。

表 1　四段活用動詞連用形の用例数（上位 10 語）

語	音便の種類	連用形用例数	活用語尾無表記例
言う	ウ音便	1356	814
参る	促音便	563	95
申す	（無）	496	484
持つ	促音便	415	77
取る	促音便	394	63
因る	促音便	358	143
致す	イ音便	343	117
成る	促音便	319	11
思う	ウ音便	298	4
急ぐ	イ音便	288	115

表 2　後接語別音便形・非音便形の用例数

語	音便形後接語						非音便形後接語					
	た		たり		て		た		たり		て	
	会話	他	会話	他	会話	他	会話	他	会話	他	会話	他
言う	63	6	31	1	148	99					7	85
参る	139	1	28		161	1	2		16		17	3
申す											1	
持つ	49	1	11		194	22			1		8	14
取る	22	3	9		118	20			3	2	10	31
因る	1				174	11						
致す	55		10		84				4	1	5	
成る	41		5		76	7			7		8	7
思う	23		42		164	2			1		2	
急ぐ	1				146					1	1	

また，この 10 語のうち，活用語尾が表記されている例が「た」「たり」「て」が後接する場合の音便形・非音便形別の用例数を示す（表 2）。なお，例えば［出現書字形］を「%

い」「%ひ」「%ゐ」として検索すると，イ音便の仮名表記の例を検索することができ，本調査においてもこのような条件で音便形の仮名表記例を抽出した。

いずれの語においても，全体的には音便形の用例数が非音便形の用例数を上回っており，音便化する例が多いことがうかがえるが，「たり」「て」が後接する場合は，音便形の例・非音便形の例のいずれも見られる。しかし，口語助動詞である「た」が後接する形では，非音便形の例は「まいりた」2 例のみで，ほぼ音便形となっていることがわかる。

なお，活用語尾の表記された四段動詞「非音便形＋た」の例として，「いだした」（出だす）「思ひ出した」（思い出だす）「かした」（貸す）「おりた」（折る）「作りた」（作る）「たちた」（立つ）等，動詞 18 種，26 例がみられたが，「音便形＋た」（動詞 251 種，1654 例）が圧倒的多数である。

このような状況から，四段活用動詞連用形のうち，音便形か非音便形かを確定できない例については，「た」が後接するものについては音便形とし，それ以外の「たり」「て」等が後接するものについては非音便形とした。

ただし，「申す」のように，活用語尾の表記された例がほぼなく，音便形の例が見られない語もある[5]。サ行四段動詞のイ音便形については多くの論考があり，早くは橘本（1962, p.28）に，「中世においても，サ行の動詞の中で，あるものは絶對に音便を起さなかつたことが知られてゐる。召スやオハス或いは申スなどがそれで，中世と言はず古今を通じてこれらの語の音便例は見當らない。」「敬語動詞であることは，音便を起しにくい條件の一つとなる」等の記述があり[6]，この「申」は非音便形と判断すべきものと考えられる。

「申す」とは反対に，「た」が後接する例以外でも非音便形の例があらわれない動詞もある。例えば，「かしこまつて御座る」の「畏まる」，機能語的な「～によつて」「～をもつて」の場合の「因る」「持つ」等である。これらについては定型的な表現として，活用語尾無表記例においても音便形と認定するのが妥当であろう。

このように，音便形もしくは非音便形の例が 1 例もない，あるいは振り仮名が付与されている等の理由から明らかに読みが予想できる例に関しては個別に読みを認定し，判断に迷う語については，「た」が後接するものについては音便形，それ以外の語が後接するものについては非音便形とすることとした。

４．活用型の認定

『虎明本狂言集』成立期は，活用体系や助動詞語彙の過渡的段階にあたり，それにともなう問題が，コーパスの語認定においても生じる。一例として，形容詞の活用型の認定の問題を挙げる。

形容詞については，文語活用の終止形である「－し」と，口語活用の終止形である「－い」の両形があらわれ，連体形についても同様に，文語活用である「－き」と，口語活用である「－い」が見られる。『日本語歴史コーパス』においては，形容詞の［解析活用型］として，「文語形容詞 - ク」「文語形容詞 - シク」および「形容詞」（口語）があり，いずれかの情報を付与する必要がある。文語活用である「－き」については「文語形容詞」，口語活用である「－い」については「形容詞」の情報を付与すればよいのだが，例えば「な

5　非音便形の活用語尾が送られている例も「よび【まし】て」の形の 1 例のみである。

6　奥村（1968, pp.44-45）でも，狂言をはじめとする中世末～近世語資料の会話文におけるサ行四段動詞の，全てが音便形の甲型の語および，音便形・非音便形が併存する乙型の語の用例数が整理されている。

がく（長い）」「すずしく（涼しい）」のような，活用語尾が「－く」となっている連用形の例については，文語形容詞，口語形容詞のいずれとするのが妥当であろうか。

表 3　形容詞終止形・連体形の活用別用例数

終止形活用語尾	用例数		
	会話	他	合計
口語活用　「－い」	725	4	729
文語活用　「－し」	116	60	176
（活用語尾無表記）	9	3	12

連体形活用語尾	用例数		
	会話	他	合計
口語活用「－い」	1711	16	1727
文語活用「－き」	340	21	361
（活用語尾無表記）	12	1	13

表 3 に，形容詞の終止形と連体形について，活用語尾を口語活用・文語活用にわけ，用例数を示した。終止形・連体形とも，口語活用の語尾の方が優勢であり，『虎明本狂言集』においては形容詞の口語活用化が進んでいたと考えられる。そこで，本コーパスにおける形容詞は，口語活用を原則とし，「－き」「－し」等の明らかな文語活用の例のみ文語活用とすることとした。

ただし，「めでたけれ」「にくけれ」のように，活用語尾が「－けれ」となるものについては，「仮定形」とするか「已然形」とするかが問題となる。室町時代には仮定条件表現は成立しており，『虎明本狂言集』においても，次の例のように，明らかに仮定条件の例があり，必ずしも已然形の已然形たる確定条件の例しか見られないわけではない。

(4)　又いそぎで【なけれ】ば，某が一細工に致すに依て，来年の今比ならではできまらせぬよ（仏師　下 p.210）

しかし，『虎明本狂言集』には(5)(6)のように「已然形＋ど・ども」の形式が残っている。また，(7)のように「こそ」による係り結びも残存しているが，仮に「仮定形」で処理すると，「こそ＋已然形」という条件で検索した場合，形容詞がヒットしないことになる。

(5)　「かほやすがたは【おそろしけれ】ど心はやさしひ（鬼のまま子　下 p.489）
(6)　いかにや／＼太郎冠者，たらされたは【にくけれ】ども，はやし物がおもしろひ（はりだこ　上 p.76）
(7)　名をとふものこそ【おほけれ】，なぜにみみをとつて引まはすぞ（腹不立　下 p.162）

このような点を勘案し，古い形に寄せた「已然形」としておくのが穏当と判断した。

また，本コーパスデータにおいては，動詞は基本的に文語活用としているため，「已然形」とすれば，形容詞に限って「仮定形」があらわれるという例外を避けることができる。さらに，「已然形」としておくことで，既存の「平安時代編」のコーパスと活用形を統一的に検索できる。このように，『虎明本狂言集』および『日本語歴史コーパス』内での統一という点からも，「已然形」とすることとした。

５．おわりに

このように，様々な面で古代語から近代語への過渡的段階である『虎明本狂言集』の言語を現代語および『日本語歴史コーパス』の既存のシステムの中で扱うにあたっては，様々な問題が生じる。蓄積された研究を反映させながら，日本語の史的研究に有用なコーパスを作成することを目指すべきであるが，一方で，語彙や文法事項について，詳細な分類や判別を行ったり，個別の例外を多く認めたりすることにより，ユーザーによる検索や形態素解析辞書の精度維持において，不都合が生じることもあろう。そのような事情から，便宜的・臨時的な判断を下さざるを得ない面もある。今後の検討や研究の成果によって塗り替えるべき箇所は多く存在するだろうが，これらの問題を解決する手がかりとなりうるのもまた，大量の事例を見渡すことのできるコーパスデータであり，コーパスデータの蓄積が，研究に寄与する部分も大きいと考える。

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」（プロジェクトリーダー：田中牧郎）による成果の一部である。

文　献

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
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
奥村三雄(1968)「サ行イ音便の消長」『國語國文』37-1，pp.34-48
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（下）」国立国語研究所内部報告書 (LR-CCG-10-05-02)
小椋秀樹・須永哲矢(2009)『中古和文 UniDic 短単位規程集』科学研究費補助金　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」研究成果報告書 2（課題番号 21520492）
橋本四郎(1962)「サ行四段活用動詞のイ音便に關する一考察」『國語國文』31-4，pp.27-43
蜂谷清人(1998)『狂言の国語史的研究―流動の諸相―』明治書院
北原保雄・村上昭子(1984)『大蔵虎明本　狂言集総索引 1　脇狂言之類』武蔵野書院
北原保雄・鬼山信行(1986)『大蔵虎明本　狂言集総索引 2　大名狂言類』武蔵野書院
北原保雄・小川栄一(1982)『大蔵虎明本　狂言集総索引 3　聟類・山伏類』武蔵野書院
北原保雄・山崎誠(1989)『大蔵虎明本　狂言集総索引 4　鬼類・小名類』武蔵野書院
北原保雄・吉見孝夫(1983)『大蔵虎明本　狂言集総索引 5　女狂言之類』武蔵野書院
北原保雄・土屋博映(1984)『大蔵虎明本　狂言集総索引 6　出家座頭類』武蔵野書院
北原保雄・大倉浩(1986)『大蔵虎明本　狂言集総索引 7　集狂言之類』武蔵野書院
北原保雄・土屋博映(1985)『大蔵虎明本　狂言集総索引 8　万集類』武蔵野書院
西尾実・岩淵悦太郎・水谷静夫(編)(2000)『岩波国語辞典　第六版』岩波書店
日本国語大辞典　「JapanKnowledge Lib」　http://japanknowledge.com/library/
室町時代語辞典編修委員会(編)(1985)『時代別国語大辞典 室町時代編一』三省堂

関連 URL

『日本語歴史コーパス』（国立国語研究所）　http://www.ninjal.ac.jp/corpus_center/chj/

否定の意志を表す「～まいとする」について

加藤 恵梨（名古屋大学）

On the Negative Volitional Expression "maitosuru"

Eri Kato (Nagoya University)

要旨

否定の意志を表す「～まいとする」がどのような表現と共起するのかを『現代日本語書き言葉均衡コーパス』の検索アプリケーション「中納言」を用いて調査し、日本語学習者が「～まいとする」を用いて文を作ったり、日本語教師が学習者に「～まいとする」の例文を提示したりする際のヒントとなるような記述を目指した。その結果、Ⅱ型（一段）動詞が「～まいとする」に前接する場合、「語幹＋まいとする」がよく用いられ、「非過去形＋まいとする」はあまり用いられないこと、不規則変化動詞「する」が「～まいとする」に前接する場合、「すまいとする」という形がよく用いられることが分かった。また、「～まいとする」に後接する表現は「～まいとしてV」が最も多く、「好ましくない事態が生じないように努力をする」という意味を表すことが多い。さらに、「～まいとする」は数は少ないが、ブログや知恵袋などでも用いられることなどを明らかにした。

１．はじめに

「～まい」には次の例(1)のように話し手の否定の意志を表す用法と、例(2)のように否定の推量を表す用法がある。

(1) あんな店には二度と行くまい。
(2) この苦しみはほかの人にはわかるまい。

（市川（2007: 219）の(1)と(2)　下線は引用者）

本稿では、話し手の否定の意志を表す「～まい」が、「～まいとする」という形で用いられる場合について考察する。「～まいとする」の例には次の例(3)～(5)のようなものがある。

(3) 銃を奪われまいとして争いになった。
(4) 夏子は泣くまいとして歯を食いしばった。
(5) 家族の者を心配させまいとする気持ちから、会社をやめたことはいわずにおいた。

（グループ・ジャマシイ（1998: 533-534）の(1)から(3)　下線は引用者）

「～まいとする」は動詞が前接し、「～ないでおこうとする」という意味を表すことが指摘されている（グループ・ジャマシイ（1998: 534））。「～まい」は話し手の否定の意志を表すが、「～まいとする」は第三者の否定の意志を表すこともできる。

以下では、「～まいとする」にどのような動詞が前接するのか、またどのような表現が後接するのか、どのような分野で用いられるのかについて、『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）の検索アプリケーション「中納言」（短単位、可変長データ）を用いて調査する。調査をもとに、日本語学習者が「～まいとする」を用いて文を

作ったり、日本語教師が学習者に例文を提示したりする際のヒントとなるような記述を目指す。

２．「～まい」に前接する表現についての先行研究の記述

「～まい」は活用のタイプによって接続の種類が異なり、一部の動詞ではゆれがあることが先行研究で指摘されている。

> Ⅰ型（五段）動詞には非過去形に接続し、Ⅱ型（一段）動詞には非過去形か、語幹に接続する。
> ・あいつには今後一切連絡をとるまい。　（Ⅰ型動詞）
> ・そんな番組、絶対｛見るまい／見まい｝。　（Ⅱ型動詞）
> 不規則変化動詞「来る」には、非過去形のほか、「来（こ）」「来（く）」にも接続する。
> ・私は二度とここには｛来るまい／来（こ）まい／来（く）まい｝。
> 不規則変化動詞「する」には、非過去形のほか、「す」「し」にも接続する。
> ・こんないたずらはもう｛するまい／すまい／しまい｝と固く決心した。
> （日本語記述文法研究会（2003: 60-61））

上の記述にあるように、「～まい」がⅡ型（一段）動詞に付く場合は二通りの言い方が可能であり、不規則変化動詞「来る」と「する」に付く場合は三通りの言い方が可能である。確かに、Ⅱ型（一段）動詞と不規則変化動詞「来る」と「する」に付く場合には複数の言い方が可能であるが、使用頻度の点から見ると、「するまい」「すまい」「しまい」が同じ頻度で用いられているとは考えにくい。よって、どの表現が良く使われているのかについて調査する必要がある。

次節では、「～まいとする」においても、Ⅱ型（一段）動詞と不規則変化動詞「来る」と「する」に付く場合、複数の言い方が用いられているのかについて調査する。

３．調査

3.1 「～まいとする」に前接する表現について

まず、「中納言」で「～まいとする」に前接する動詞の「書字形出現形」を調べると、次の表１のような結果が得られた。

表１「～まいとする」に前接する頻度の高い表現（総数 214）

順位	共起する表現	出現数	順位	共起する表現	出現数
1	考え	17	7	す	6
1	見せ	17	7	傷つけ	6
3	出す	14	9	泣く	5
4	負け	13	9	見逃す	5
5	かけ	8	9	与え	5
6	失う	7			

１位の「考える」と「見せる」、４位の「負ける」、５位の「かける」[1]、７位の「傷つ

[1] 「かける」は「心配をかける」が５例、「迷惑をかける」が３例であった。

ける」、9位の「与える」はⅡ型（一段）動詞である。これらの動詞は表1のように、BCCWJでは「語幹＋まいとする」という形が用いられており、「非過去形＋まいとする」という形は用いられていなかった。次の例(6)と(7)は最も出現頻度が高い「考える」と「見せる」の例である。

(6) （前略）むろん、せっかく気持よく酔っているときに、そこまで問い詰めることはない、という人もいるかもしれない。だがそれは一見夢見がちなロマンチストの意見で、その実、なんの答えにもなっていない。いい替えると、もともと確たる答えをもっていないから、その先のことには目をつぶって考えまいとする。
（渡辺淳一『失楽園』）

(7) 「あのとおり気丈なやつだから、弱みは見せまいとするだろ。そのじつ、俺の世話を焼くことでかろうじて自分を立たせてる、それもわかってた。本当は俺なんかよりあいつのほうがよっぽどきついってこともな。（後略）」
（村山由佳『天使の梯子』）

表1に挙げた動詞に限らず、BCCWJではⅡ型（一段）動詞は、「語幹＋まいとする」の形が用いられていることから、「非過去形＋まいとする」の形はあまり用いられていないと推測できる。しかし、「見る」は例外的で、「見るまいとする」という「非過去形＋まいとする」が3例あり、「見まいとする」という「語幹＋まいとする」は1例のみであった。

また、7位の不規則変化動詞「する」は、大部分が次の例(8)のように「すまいとする」という形で用いられている。その他は、例(9)に示したように「しまいとする」という例が1例あっただけで、「するまいとする」という形は用いられていない。

(8) フィナーレのロンド・アレグレットも、チェロのソロで開始する。これまでにないチェロの起用であるが、技法を複雑にすまいとする配慮のなかで精緻にアンサンブルさせているのは、さすが年季の入った室内楽作曲家の手になるものだ。
（高橋英郎『モーツァルト366日』）

(9) （前略）現在の段階では、これらは第三国を刺戟しまいとする政策的考慮から出た自制行為であって、必ずしも戦争の名を避けて武力行使を行う国家が交戦国としての中立法上の権利を一切行使しえないという原則が確立されているわけではない。
（山手治之『国際法概説』）

さらに、今回の調査では、「～まいとする」に不規則変化動詞「来る」が前接する例は見られなかった。

以上から、Ⅱ型（一段）動詞が「～まいとする」に前接する場合、「語幹＋まいとする」の形がよく用いられ、不規則変化動詞「する」が「～まいとする」に前接する場合は、「すまいとする」という形がよく用いられると言うことができる。

3.2 「～まいとする」に後接する表現について

次に「～まいとする」に後接する表現について見る。「～まいとする」に後接する表現を調べると、次の表2のような結果が得られた。

表 2「～まいとする」に後接する頻度の高い表現

順位	後接する表現	出現数
1	～まいとしてＶ	42
2	～まいとするＮ	37
3	～まいとした。	24
4	～まいとしている。	13
5	～まいとしていた。	8
6	～まいとする。	5

最も多いのは、次の例(10)から(13)のような「～まいとしてＶ」という形である。

(10) 〈やっぱりこいつは、鉄人２８号じゃ。球のスピードと切れが、わしとは、全然ちがう〉
　咲本は、最初は、負けまいとして懸命に投げていた。が、そのうち、無理して投げるので肩が痛くなってくる。　（大下英治『小説明治大学』）

(11) 折角ありついた地位を失うまいとして無暗に勉強したのである。
（佐々木邦『ガラマサどん』）

(12) （前略）だが、もし、地元の警察が、この日記を読んでいたら、きっと、石崎を、真っ先に疑ったろうと、思った。石崎が、堀江正彦を失うまいとして、由美を殺したのではないかと、警察は、考えたろうからである。
（西村京太郎『十津川警部の挑戦』）

(13) 目の縁から大粒の涙がいくつもこぼれ落ちた。それでも必死に泣くまいとして、ペチカの顔はぐちゃぐちゃになる。　（向山貴彦『童話物語』）

「～まいとしてＶ」という形で使われると、「好ましくない事態が生じないように努力をする」という意味を表すことが多い。例(10)の「負けまいとして懸命に投げていた」は、「相手が投げる球に負ける」という好ましくない事態が生じないように、懸命に投げる練習をしたということを表している。同様に、例(11)の「地位を失うまいとして無暗に勉強した」は、「地位を失う」という好ましくない事態が生じないように、無暗に勉強したということを表している。

一方で、「～まいとしてＶ」という表現は「好ましくない事態が生じないように努力した結果、悪い事態が生じる」という意味を表す場合がある。例(12)の「堀江正彦を失うまいとして、由美を殺した」は、「堀江正彦を失う」という好ましくない事態を避けるためにどうにかしようとして、他の人を殺すというより悪い事態が生じたことを表している。同様に、例(13)の「泣くまいとして、ペチカの顔はぐちゃぐちゃになる」は、「泣く」という悪い事態が生じないように努力した結果、「顔がぐちゃぐちゃになる」というより悪い事態が生じたことを表している。

また、「～まいとする」に後接する表現として次に多かったのが、「～まいとするＮ」である。「～まいとする」が修飾する名詞には、次の例(14)のような「責任感」、例(15)のような「配慮」、例(16)のような「意志」といった「人の気持ちや考え」を表す表現が多い。

(14) 自分の仕事が期限に遅れたり粗相をしたりすることで、顧客に、上司に、部内の他の人に、社内の他の部署の担当者に、迷惑をかけまいとする責任感に駆られて呻吟している自分の姿に気づく。　（大野正和『過労死・過労自殺の心理と職場』）

(15) 風見は少なからず驚いた。いままで紀久子が自室へ異性の社員を呼び寄せたことはなかったからである。女社長として、男の社員からなめられまいとする配慮か

らであろうが、それはそれなりに紀久子の権威を保つ効果をあげていた。
（森村誠一『新幹線殺人事件』）

(16) 一郎の手紙には、「節制」「忍耐」の言葉が頻繁に登場する。「一日中馨と一緒にいたい」「筆の運ぶままに手紙を書き綴っていたい」恋をすれば誰もが抱くこんな気持ちを抑え、薫が勉学の妨げになったと言われまいとする意志を、ここに読みとることができる。　（鳩山一郎『若き血の清く燃えて』）

3.3 「～まいとする」の使用分野について

最後に、「～まいとする」がどのような分野で多く使われているのかについて調べる。先行研究では、「～まいとする」は「書きことば的なかたい表現」（グループ・ジャマシイ（1998: 534））と指摘されている。

「～まいとする」がどのような分野で使用されているのかを「中納言」で調べると、圧倒的に書籍が多い。その他のものとして、ブログに４例、雑誌に１例、知恵袋に１例、新聞に１例用いられていた。次の例(17)はブログの例、例(18)は知恵袋の例、例(19)は新聞の例である。

(17) 忙しいところにメールが来た。N ちゃんからであった。「りゅうちゃんが熱を出して、吐き気もすると言って・・・・」娘と N ちゃんがわざわざ病院まで連れて行ったそうだ。娘と息子からはメールが無い。娘はこういう時、私に心配をかけまいとするようになった。　（Yahoo! ブログ）

(18) 仕事中、どんなに対策しても眠ってしまいます。前日にしっかり眠ってもコーヒーやドリンク剤を飲んで、眠るまいとしていても気がつけば意識が薄れ、船をこいでいます。　（Yahoo! 知恵袋）

(19) （前略）裁判中の報道について「原告の言葉を忠実に報じた。その結果、隠ぺいされていた隔離政策の実態が白日のもとにさらされ、世論を喚起した」と評価する。ただ、判決後の堰を切ったような大量の報道について「乗り遅れまいとして報道したマスコミもあったのでは」（後略）。　（中日新聞）

数は少ないが、「～まいとする」は例(17)や(18)のようにブログや知恵袋で用いられることもある。また、例(19)はある人の話を聞いて記事にしたものであることから、「～まいとする」は話しことばでも用いられているということができる。

４．まとめと今後の課題

否定の意志を表す「～まいとする」について、次のことを明らかにした。

・Ⅱ型（一段）動詞が「～まいとする」に前接する場合、「語幹＋まいとする」の形がよく用いられ、「非過去形+まいとする」はあまり用いられない。また、不規則変化動詞「する」が「～まいとする」に前接する場合、「すまいとする」という形がよく用いられる。

・「～まいとする」に後接する表現は、「～まいとして V」が最も多い。また、「～まいとして V」という形で使われると、「好ましくない事態が生じないように努力をする」という意味を表すことが多い。

・「～まいとする」は数は少ないが、ブログや知恵袋などでも用いられている。

今後の課題として、否定の意志を表す「～まい」についても調査し、「～まい」と「～まいとする」ではどのような違いがあるのかについて考察する必要がある。また、「～まいとする」の類義語である「～ないようにする」や「～ないでおこうとする」との意味の違いについても分析したいと考えている。

文　献

庵功雄、高梨信乃、中西久美子、山田敏弘（2001）『中上級を教える人のための日本語文法ハンドブック』スリーエーネットワーク
市川保子（2007）『中級日本語文法と教え方のポイント』スリーエーネットワーク
グループ・ジャマシイ（編）（1998）『教師と学習者のための日本語文型辞典』くろしお出版
友松悦子、宮本淳、和栗雅子（2010）『新装版　どんなときどう使う　日本語表現文型辞典』アルク
日本語記述文法研究会（編）（2003）『現代日本語文法④　第８部　モダリティ』くろしお出版
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

ポスター発表　グループB

3月11日（水）　14:00～15:00

BCCWJ に見る類義表現「～きる」「～ぬく」「～とおす」の使い分け

栗田　奈美（立教大学日本語教育センター）

Discriminating the Synonymous Expressions "-kiru", "-nuku", and "-toosu" Based on the BCCWJ

Nami Kurita (Center for Japanese Language Education, Rikkyo University)

要旨

本研究は、BCCWJ を用い、「行為の完遂」を表す統語的複合動詞「～きる」「～ぬく」「～とおす」の使い分けの実態を明らかにすることを目的とする。検証方法としては、3 者の前項に共通して挿入されていた動詞のうち「守る」と「走る」に注目し、それぞれの複合動詞が表す意味の相違を見た。その結果、「守る」の場合、「～きる」は最終段階が重視されるスポーツの文脈で多用され、「～ぬく」は守ることに困難が予想される抽象物（例：権利、信仰）が対象となる用例が多く見られ、「～とおす」はあらかじめ定められている抽象物（例：約束、規則）が対象となり、その状態を変えずに保ち続けることに意味を見出す文脈で多用される傾向が見られた。このことから、「～きる」は瞬時的な最終段階を、「～ぬく」は困難を伴うプロセスを経てそこから離脱するまでを、「～とおす」は一定期間継続するプロセスを、それぞれ焦点化していることが示された。

１．はじめに

本研究では、「現代日本語書き言葉均衡コーパス」（以下 BCCWJ）において、「行為の完遂」を表す統語的複合動詞「～きる」「～ぬく」「～とおす」の前項に共通して挿入されていた動詞に注目し、類義表現となっているそれぞれの複合動詞が表す意味にどのような相違があるかを分析する。

この 3 者のそれぞれ、もしくは 3 者を比較して意味分析を行った研究は、姫野（1980, 1999）、森田（1989）、石井（1988）、青木（2004）、大友（2005）、中島（2006）、杉村（2008, 2012）、許（2012）等がある。その中でも、後続する研究に多大な影響を与えたと思われる姫野（1980）は、本研究が対象とする統語的複合動詞について、以下のように意味分類している。

１）～きる
- ・完遂：行為の単なる終了ではなく、行為者の予定通り完全に行われたことを表す。
- ・極度：変化が進み、それ以上はないという究極まで達することを表す。

２）～ぬく
- ・貫徹：動作を最後まで完全に行うことを表す。
- ・極度：「非常に、とことんまで」という強い程度を表す。

３）～とおす
- ・一貫継続：継続行為もしくは反復行為として最後までし続けることを表す。

さらに、姫野は「時間性」と「意志性」という観点からそれぞれの比較を行っている。前者については、「～きる」が完遂の瞬間に重点を置いているのに対し、「～ぬく」と「～とおす」は完遂までの過程に重点を置いているとしている。後者の「意志性」については、

最終段階に至るまでに逆流が予定される「～ぬく」が最も強く、「～とおす」がそれに続き、「～きる」は最も弱いと述べている。

本研究は、姫野を始めとする先行研究の知見に考察を加え、「行為の完遂」を表す用法（姫野分類における「完遂」「貫徹」「一貫継続」に相当）のプロトタイプを以下のように整理した。

1）～きる
　継続する行為の、瞬時的な最終段階（結果）を焦点化したもの

2）～ぬく
　継続する行為のプロセスに、何らかの障害や困難さが存在するが、それを克服し、最終段階でその状態から離脱するところまでを焦点化したもの

3）～とおす
　状態（結果状態）や行為が、途切れることなく、一定不変に継続するプロセスを焦点化したもの

3 者の使い分けは、これらのプロトタイプが持つ意味的特徴を基になされていると考えられる。次節からは、この考察の妥当性を検証するとともに、新たな事実の発掘を求めて行ったコーパス調査について述べる。

２．コーパス調査の概要

２．１．目的

「～きる」「～ぬく」「～とおす」の使い分けの実態を明らかにする。特に、1 で述べたそれぞれのプロトタイプに関する考察の妥当性を検証する。

２．２．資料

BCCWJ 短単位データ 1.0 バージョン、また、コーパス検索用ツールとして「中納言」を使用する。

２．３．方法

「～きる」「～ぬく」「～とおす」のそれぞれがコーパス上に現れる件数および使用頻度の高い複合動詞にどのようなものがあるかを検索し、頻度の高いもののうち、3 者に共通する前項動詞を選択し、意味分析を行う。

３．結果と考察

まず、それぞれのデータの個数は「～きる」が 8,378、「～ぬく」が 1,311、「～とおす」が 516 であった。また、3 者それぞれにおいて使用頻度の高い複合動詞上位 50 語のうち、3 者の前項に共通して挿入されていた動詞は以下の 6 語であった。

表 1　「～きる」「～ぬく」「～とおす」共通の前項動詞 6 語のデータ数

前項動詞	「～きる」データ数	「～ぬく」データ数	「～とおす」データ数	合計
為(ス)る	954	17	25	996
遣(ヤ)る	296	59	40	395
守る	65	97	93	255
読む	55	3	34	92
信ずる	76	7	2	85
走る	40	14	8	62
合計	1486	197	202	1885

本研究では、このうち、「守る」と「走る」について分析する。この 2 語を選び、他の 4 語を除外する理由であるが、まず、「読む」「信ずる」はデータ数 5 未満のものを含むため、今回の考察対象からは外した。また、「為る」を外したのは、「～とおす」のデータ数 25 件中 17 件が「～とおし」という名詞形で現れていたことに加え、「～きる」においては「極度」[1]の意味を表す用例が多かったためである。さらに、「遣る」については「～きる」のデータ数 296 件のうち、約 9 割が「やりきれない」という辞書にも一語として記載のある語彙化した形式で現れていたため、これも除外した。以上の理由から、「守る」「走る」の 2 語について検証することにした。

3 者のいずれの前項にも共通して挿入され得るということは、3 者間での言い換えが可能であるということでもある。にもかかわらず、その文脈では 3 者のうちの 1 つが選択されているという事実に注目し、その 1 つの後項動詞が選ばれた動機づけを探りつつ、使い分けを明らかにしていく。

３．１．　前項動詞「守る」の場合

BCCWJ において「守りきる」は 65 件、「守りぬく」は 97 件、「守りとおす」は 93 件のデータが見られた。但し、「守りとおす」93 件のうち 26 件は同一ブログ内の用例で、前後文脈 50 語を確認したところ、まったく同じ内容のものがあったため、その重複分を除き、79 件を考察対象とした。これら 3 者の意味を比較対照するために、「守る」の対象に注意しながら見ていくことにする。

「守る」の対象は、「人」「場所」「具体物」「抽象物」の 4 つのカテゴリーに大別した。対象が目的語としてテキスト内に明示されていない場合には、前後文脈から判断して筆者が補った。それぞれの結果は次項からの表の通りである。なお、表の括弧内の数字は用例数を表している。

３．１．１．「守りきる」

「守りきる」の用例で特徴的なのは、表 2 が示す通り、スポーツに関する文脈で現れるものが半数近くを占めている点である。「守りぬく」の用例ではわずかに 2 件が見られるのみで、「守りとおす」では 1 件も見られなかったことを考えると、「守りきる」の用例数は突出している。これらの用例は、野球、サッカー、駅伝、アメリカンフットボール、ソフトボール等、ジャンルを問わず、様々な競技の文脈で見られた。また、対象が得点差である場合は、その大半が僅差であった。スポーツの場合、勝敗が決する最後の瞬間がハイライトとなる。したがって、典型的には、僅差である貴重なリードを試合終了のホイッスルが鳴るまで守り、それが見事に達成された瞬間を切り取って表したい場合に、「守りきる」が選ばれると考えられる。以下に BCCWJ の例を挙げる。

(1) a. 白いベールに覆われたフィールドを駆け、今泉がペナルティゴールで決めたトラの子の 3 点を守りきった。10－7 で逃げ切った早稲田は、その後日本一へと駆け上がる。（松瀬学『早稲田ラグビー再生プロジェクト』）

b. この大会で、蓮池ホワイトシャークは 1 点差を守りきったり、みんなが打って大差の勝利を収めたり、さまざまな試合をしながら、準決勝戦では北原に 2 対 0 の僅差で勝ち、決勝に進みました。（『土佐広報』2008 年 08 号）

いずれも貴重な得点差を守り、最後には勝利という結果を獲得したことがわかる。(1)a.の用例には、同様に「行為の完遂」を表す「逃げ切る」も使用されている。このような文脈で

1 「～とおす」は「極度」を表す用法を持たないため、3 者の比較とならない。

は、最終段階を焦点化する「～きる」の効果が有効に働くために、「守りきる」が好まれるものと考えられる。

表 2「守りきる」の対象（65 例の内訳）

「守りきる」の対象		用例
人（15）		父親（2）／殿／取材協力者／貴方／兄／自分／一人／喬子、大久保等の固有名詞（7）
場　所（13）	スポーツにおける防御エリア（4）	ゴール（3）／ゾーン
	組織（4）	国／一国／村／家
	城（3）	城／小山城／沼田城
	その他（2）	土地／基地
具体物（5）	貴重品（3）	相続物／村雨丸[2]（2）
	脆弱なもの（2）	胃粘膜／コンピューター
抽象物（32）	スポーツにおける得点差（25）	１点（7）／１点差（5）／リード（3）／～点（3）／～点（の）リード（2）／得点／決勝点／先制点／勝ち越し点／２ＴＤ（タッチダウン）
	定められている約束・ルール等（3）	約束／規則／ローテーション
	その他（4）	尊厳／命／１位／信託兼営

３．１．２．「守りぬく」

次に、「守りぬく」の対象と用例を見る（表 3 参照）。「守りぬく」の対象に関して特徴的な点は 2 点ある。まず 1 点は、場所が対象となる用例が多い点である。この場合の場所とは、単なる場所というより、動作主体が帰属する組織、コミュニティ、また、その構成メンバーをも含めたものとして考えた方がいいようである。3 者の中で最も多い 6 例が見られた「城を守る」では、敵の攻撃から城という建造物を物理的に防御するというだけでなく、メトニミー的にその城を所有する家や君主、家臣までを守るという文脈で使用されている。「国」が対象の場合も「国家」「祖国」「領国」等、何らかの含意を持つ語彙が使用されており、同様の傾向が伺える。つまり、これらの場所は、動作主体にとって有意味で重要性の高いものであり、場合によっては命を賭しても守るべきものであると言える。

もう 1 点の特徴は、権利、信仰、伝統等、守ることに困難が予想される抽象物が対象となる用例が多い点である。これらの抽象物には、外から脅かされる可能性があったり、強い意志がない限り、保持することが困難であったりするものが多い。以下に、これら 2 つの特徴を表す例を挙げる。

(2) a. 晴朝は落城寸前まで追いこまれたが城を守りぬき、結局、両家講和ということになった。（森好夫『松平大和守家の研究』）

b. 極論で言ってしまえば、宗教家とは神の名の下に集められた罪人であり、神とその

[2] 日本刀の名称。

教えを守り抜かんとする兵士なのだ。罪と血が、常にその傍らにある者なのである。
（渡辺水央『Trigun maximum 深層心理解析書』）

(2)a.の例は、落城寸前まで追い込まれた状況から困難を排して大切な城を守り、最後にはその苦しい状況を脱したことを表している。さらに、その結果が「両家講和」であることから、守った対象が単なる城という建造物ではなく、城を所有する家であったことがわかる。(2)b.の例からは、神とその教えを守ろうとする強い意志が感じられる。また、宗教家を兵士に喩えていることから、教えを貫くことを戦いと捉えていることがわかる。ここでは、ただ受動的に守るのではなく、武器を手にして戦うことで守るという積極的な姿勢が見られる。このように、守るというプロセスに困難が存在し、その困難を克服して守る行為を達成する場合には、「守りぬく」が選ばれるものと考えられる。

表 3「守りぬく」の対象（97 例の内訳）

「守りぬく」の対象		用例
人（18）		君（3）／この子（2）／殿／同志／家族／家内／愛するもの／これぞと思った人／主流派／相手／自分／男の身／戸田、マリア等の固有名詞(3)
場　所（25）	城（6）	城（2）／塞／小城／滝山城／鳥越城
	組織（14）	故郷（2）／国家、祖国／自分の家と領国／町／村／幕府／家庭／コミュニティ／社屋と社員／豊臣家、小県郡、そごう等の固有名詞（4）
	その他（5）	土地／この地／畑／西ベルリン／羽柴勢の背後
具体物（8）		建造物／古代超文明の遺物／市庁舎[3]／宝／崋山の絵／資産／財産／川上犬[4]
抽象物（46）	権利（5）	権利（3）／利権／独立と主権
	宗教（3）	信仰／神とその教え／学会活動[5]
	定められている約束・ルール等（5）	誓い／遺志／指示／順序／工程表[6]
	その他（33）	伝統（4）／（生）命（2）／地位（2）／大切なもの（2）／秘密（2）／治安／憲法／独自性／信頼関係／自由主義／自由貿易体制／成果／部門の誉れ／農業／言葉／一生／留守／平等論／家柄のよさ／信じるもの／立場／沈黙／形式／試合／リード／会社経営の根幹は「人間理解」にあるということ

３．１．３．「守りとおす」

最後に、「守りとおす」の対象と用例を見る（表 4 参照）。

3 延焼から市庁舎の建物を守るという文脈であったため、場所ではなく具体物に分類した。
4 小型日本犬の一種。長野県の天然記念物。
5 宗教法人創価学会の活動。
6 スケジュールの意。

表 4「守りとおす」の対象（79 例の内訳）

「守りとおす」の対象		用例
人（9）		愛する人（2）／子ども／好きになった女／皆／そなたたち／自分／胡蝶さん（2）
場所（3）		地球／区劃や広場や通り／勇猛で粗野な人のいる地
具体物（10）		コレクション（3）／現金／道具／レシピ／どんぶり／村雨丸、新田[7]等の固有名詞（3）
抽象物（57）	定められている約束・ルール等（11）	約束（4）／誓い／原則／ルール／殺生戒／日課／食べてから寝るまで２時間空けること／気が進まないことはしないということ
	操（3）	貞操／節操／女の操
	その他（43）	命（2）／沈黙（2）／秘密／信条／信義／友朋関係と信義／平和／文化／伝統／魂／真価／おのれの一分／利益／社会体制／主導権／自説／大事だと思ったもの／宮座／王座／２番目の位置／服装／涼しい顔／別姓／テンポ／設定／結婚生活／長寿食／バランス／最后の一線／ブログに書いてきたこと（12）[8]

「守りとおす」の用例で特徴的なのは、約束、ルール等、あらかじめ定められている抽象物が対象となる用例が多い点である。これらを守るための積極的、動的な活動は必要とされない。むしろ、その状態を変えずに続けることが必要であると言える。つまり「約束を守る」ことは「約束を破らない」状態を続けることであり、同様に「節操を守る」は「節義を変えない」状態を続けること、「沈黙を守る」は「口をきかない」状態を続けることである。また、「1 点差を守る」の場合は試合終了時、「城を守る」の場合は敵を打ち負かし、退散させた時点が「守る」の非明示的な完了時となるが、「約束を守る」の場合はそのような完了時は含意されない。以下に例を挙げる。

(3) a. 子育てを中心にする、という結婚する時の約束もほぼ守り通しています。（シェリー・アモテンスティーン著・月谷真紀訳『恋人と別れたくないあなたへ』）

b. それでも私は感謝しています。まず、「セルビア式のやり方でおまえをなぶり殺しにしてやる」とご親切にも予告して下さった高潔なる愛国者の皆さんに、そして沈黙を守り通した同僚や友人、知人の皆さんに。おかげで、あなた方をあてにするのは間違いだということを教わりました。（スラヴェンカ・ドラクリッチ著・三谷惠子訳『バルカン・エクスプレス』）

いずれの例も、守るための動的な活動はなされていない。また、(3)a.は進行形に結合しており、状態の継続性が顕著である。進行形との結合は、結果を焦点化する「～きる」には見られなかった特徴である。一方、(3)b.は祖国を追われた女優の書簡の一部であるが、「沈黙を守り通す」はタ形にはなっているものの、未だ完了はしておらず、その状態が続いている可能性が高い。このように、完了ではなく、その状態を保ち続けることに意味を見出す文脈の場合には、「守りとおす」が好まれるものと考えられる。

7 茶器の名称。

8 これらはすべて、前述した同一ブログ内の用例である。表現自体は多少異なるため、12 件のデータとして取り扱っているが、内容的には同一の趣旨を繰り返している。

次項では、「守る」同様、3 者の前項動詞となっていた「走る」について考察する。

３．２．前項動詞「走る」の場合

BCCWJ において「走りきる」は 40 件、「走りぬく」は 14 件、「走りとおす」は 8 件のデータが見られた。特に「走りぬく」「走りとおす」についてはデータ数も限られているため、傾向を指摘するにとどめるが、データ数が少ない分、それぞれの文脈も含め、精査することができた。3 者を比較対照した結果は以下の表の通りである。

表 5「走りきる」「走りぬく」「走りとおす」の比較対照

		～きる（データ数 40）	～ぬく（データ数 14）	～とおす（データ数 8）
走行の種類	物理的走行	38	9	8
	抽象的走行	2	5	0
経路	（中間経路／距離）を	11	3	5
	（着点）まで	2	0	1
	（起点）から（着点）まで	0	0	2
共起する副詞句	最後まで	4	2	0
	全力で	4	0	0
	～なく（例：怪我／休み）	0	0	2
文法形式	名詞形（例：～きり）	0	0	1
	可能表現（例：～きれる／～きることができる）	6	0	3
	命令形（例：～きれ）	2	0	0
	意志形（例：～きろう）	0	2	0
	希望表現（例：～きりたい）	2	2	0
	重複構文（走りに走る）	0	1	0
	～てくる	0	2	0
レースの文脈		29	7	3
困難さの含意[9]		8	10	2

３．２．１．走行の種類

3 者の比較対照に際し、まず、「走る」が物理的走行を表しているのか、あるいは抽象的走行を表しているのかに注目した。抽象的走行というのは、例えば、人生をレースに喩えて「走る」と言うような場合を指す。「走りきる」ではデータ 40 件中 2 件、「走りとおす」では 8 件中 0 件であったが、「走りぬく」では 14 件中 5 件という相対的に多い結果であった。以下に例を挙げる。

(4) 私たちはこのように多くの証人に雲のように囲まれているのであるから、一切の重荷とからみつく罪とをかなぐり捨てて、私たちの参加すべき競争を耐え忍んで走りぬこうではないか。（Yahoo! ブログ）

9 複合動詞が現れる文と同一文中に、副詞句（例：耐え忍んで）や節（例：息絶えてもいいから）により、行為の遂行の困難さが示されている場合に、含意があると判断した。

この例は聖書からの引用だそうだが、「耐え忍んで」という副詞句や「走りぬこう」という意志形の使用も特徴的である。この他、「布教のために走りぬく」「魂が走りぬく」等の例が見られたが、いずれも抽象的走行に際し、何らかの困難が予想されるものであった。「走る」をメタファー的に解釈し、比喩表現として使用した場合、最も写像されやすいのが走行中の辛さ、苦しさであるために、困難さの含意を持つ「～ぬく」に抽象的走行を表す用例が多く見られたものと考えられる。

３．２．２．経路

次に、移動経路が明示されているかどうかを観察した。「走りきる」では 40 件中 13 件、「走りぬく」では 14 件中 3 件であったが、「走りとおす」では 8 件中 8 件[10]の経路が示されており、突出して多かった。また、「（起点）から（着点）まで」という形式で表されていた用例も「走りとおす」のみに見られた。以下に例を挙げる。

(5) だいたい東京から静岡を過ぎたくらいまでの距離をオートバイで休みなく走り通せば、誰にでもその感覚を味わうことができるはずだ。（素樹文生『旅々オートバイ』）

「～とおす」は結果ではなくプロセスを焦点化するために、経路を明示する傾向が他の 2 者より強く現れたものと考えられる。

３．２．３．共起する副詞句

複数回現れた共起副詞（句）は数が少なく、「走りきる」と共起していた「全力で」が目立った程度である。移動経路が長くなればなるほど、最初から最後まで全力疾走することは難しい。そのため、「全力で」はプロセスを焦点化する「走りぬく」「走りとおす」ではなく、結果を焦点化する「走りきる」とのみ共起していたものと思われる。「走りきる」には「一気に」との共起例も見られたが、いずれも瞬時性、瞬発性が感じられる副詞である。

また、「～なく」は(5)の例に見られるように、「休みなく」や「怪我や事故もなく」という形で「走りとおす」とのみ共起していた。3.1.3.の「守りとおす」の考察で見たように、「～とおす」は積極的、動的な活動ではなく、状態を変えずに続けることを焦点化する傾向を持つ。同様に「走りとおす」では、休みや怪我のない状態を最後まで続けることに注目しているのではないか。これらの共起例を以下に挙げる。

(6) 主将の■■■■さん（6 年）は「目標は全国 3 位以上。みんなで声を出し合い、最後まで全力で走り切る」と抱負を力強く話しました。（『広報ひゅうが』2008 年 3 号）

(7) 順位やタイムなんかどうでもいいのである。とりあえず、怪我も事故もなく走りとおせるかどうかが、初体験者にとっては大問題だ。（Yahoo! ブログ）

３．２．４．文法形式

文法形式では、まず、名詞形は「走り通し」という形でしか現れなかった。同様に、動作というより状態性の能力を表す可能表現は、「走りぬく」では見られなかった。「～ぬく」は 3 者の中で最も意志性が強いために、無意志動詞となる可能表現とは共起しにくいものと思われる。前項で見た「守りぬく」でも、可能表現との共起が 97 件中 2 件（いずれも「～ぬける」という可能動詞ではなく「～ぬくことができる」の形式）で、「守りきる」の 65 件中 27 件、「守りとおす」の 79 件中 7 件と比べ、かなり少なかった。

命令形、意志形、希望表現については、「走りとおす」との共起は見られなかった。1 で見たように、姫野では「～きる」が 3 者の中で最も意志性が弱いとされていたが、実際の

10 「（中間経路／距離）を」と「（起点）から（着点）まで」の両者を含むデータ（例(5)参照）があったため、延べ 8 件となったが、データ件数は 7 件であった。

データでは命令形や希望表現と共起している用例が複数見られ、「～ぬく」ほど強くはないものの、「～とおす」より意志性が弱いとは言えない結果となった。また、強調表現である重複構文（走りに走る）は、意志性の強い「～ぬく」にのみ見られた。

最後に、あちらからこちら、あるいは過去から現在までの移動や変化を表す「～てくる（きた）」は、プロセスと結果を焦点化する「走りぬく」にしか見られなかった。これは、同様にプロセスを焦点化する「走りとおす」にも理論上は見られるものと思われるが、結果だけを焦点化する「走りきる」には多回的な場合（例：フルマラソンを何度も走りきってきた）を除き、後接しない形式である。「～ていく」についても、同様の傾向が予想される。以下に本項で取り上げたそれぞれの例を挙げる。

(8) 逃げ出せるものなら、縛り首にはなりたくありませんでした。そこでカヌーが見つかるまで、おれは走りどおしでした。（マーク・トウェイン『マーク・トウェインコレクション』）（名詞形）

(9) タイヤメーカー側の基本的な開発姿勢は、あくまで “安全に三百 km を走りきれるタイヤ” である。（柴田久仁夫『AUTO SPORT』2005 年 6 月 9 日号）（可能表現）

(10) 「小僧、後でたっぷり可愛がってやるからちゃんとゴールまで走りきれよ。もうふらふらしてんじゃないか」（斎藤純『銀輪の覇者』）（命令形）

(11) これを最後に、何を失ってもいいから走りぬきたい。足が折れてもいい、ゴールに飛び込んだ時点で息絶えてもいいから、走りぬきたい。そう思いながら、わたしは必死で足を動かしていた。（有森裕子『わたし革命』）（希望表現）

(12) ことさら かつぜんとして 秋がゆふぐれをひろげるころ たましいは 街を ひたはしりにはしりぬいて 西へ 西へと うちひびいてゆく（八木重吉『八木重吉詩集』）（重複構文）

(13) 昨夜女鬼谷を出発し、徹夜で馬をとばし、途中から道なき道を走り抜いてきた菊の乱れ髪は、勝ち気そうな美しい顔にぴったりと張り付いていた。（西谷史『ブラディー・セイント女鬼』）（～てくる）

３．２．５．レースの文脈

「走る」が用いられる文脈には、レースに関するものが多いことが予想されたが、「走りきる」では 40 件中 29 件、「走りぬく」では 14 件中 7 件、「走りとおす」では 8 件中 3 件と、出現率に差が出た。このことは、「守りきる」のスポーツの文脈における出現が突出して多かったことと並行している。つまり、一般的にレースにおいて最も重要な瞬間はゴールの瞬間であるため、結果を焦点化する「～きる」が選択されているものと考えられる。さらに、同一文中で、結果で最も重要視されるレースの到着順位にまで言及している例は、「走りきる」では 5 件、「走りぬく」では 2 件であったが、「走りとおす」には見られなかった。この傾向は、「走りとおす」が用いられている(7)の「順位やタイムなんかどうでもいいのである」という文からも明らかである。このことも「～とおす」が結果ではなく、プロセスを焦点化していることを証明している。以下に、順位にまで言及している「走りきる」の例を挙げる。

(14) たとえば同じ 1 位でも、4 分 3 時点まではクォーターごとに 300 万 800 円なのに対して、最終クォーターをトップで走り切り、チェッカーフラッグを受けると、つまり優勝すると 1599 万 6200 円になる。（城島明彦『F1 の経済学』）

３．２．６．困難さの含意

最後に、完遂表現でよく目にする「抵抗を排し、困難を乗り越えて達成する」といった

含意がどの程度見られるかに注目した。これは、予想通り、「走りぬく」が圧倒的に多く、14 件中 10 件であった。また、「走りきる」は 40 件中 8 件、「走りとおす」は 8 件中 2 件であった。以下に例を挙げる。

(15) エゴロワが迫ってくる。「もうこれ以上走れない」そう思った途端に追いつかれる、抜かれる。「足が折れてもいいから、走りぬこう」こう思った途端、エゴロワを抜く。「だめだ。限界だ」抜かれる。猛烈なデッドヒートがつづく。（有森裕子『わたし革命』）

(16) 島の暮らしのなかで、村八分にあえば、死活問題にもつながりかねなかった。しかし、悔し涙をこらえ、歯を食いしばって、広宣流布に走り抜いてきたのだ。（池田大作『新・人間革命』）

(15)はマラソン、(16)は布教活動と、文脈は全く異なるが、いずれも最後まで走ることに、下線で示したようなかなりの困難が存在し、それに対して動作主体が強い意志を持ち、克服しようとしている、あるいはしてきたことが読み取れる。

４．まとめ

本研究では、「守る」「走る」を前項に持つ複合動詞を例に、類義表現「～きる」「～ぬく」「～とおす」の使い分けを探った。BCCWJ を用いたコーパス調査の結果、「～きる」は瞬時的な最終段階を、「～ぬく」は動作主体の意志的、積極的な関与により、困難を伴うプロセスを経て、そこから離脱するまでを、「～とおす」は一定期間変化せずに継続するプロセスを焦点化していることが実際のデータの中に確認でき、それによって 3 者の使い分けがなされていることが示された。

文　献

青木博史（2004）「複合動詞「～キル」の展開」『国語国文』73:9, 35-49.

姫野昌子（1980）「複合動詞「～きる」と「～ぬく」、「～とおす」」『日本語学校論集』7, 23-46.

――――（1999）『複合動詞の構造と意味用法』ひつじ書房.

石井正彦（1988）「接辞化の一類型－複合動詞後項の補助動詞化－」『方言研究年報』30, 281-296.

許臨揚（2012）「複合動詞「～切る」の意味と用法－認知言語学の意味関連の観点から－」『日本認知言語学会論文集』12, 285-296.

栗田奈美 (2014)「視覚スキーマを用いた意味拡張動機づけの分析－完遂を表す複合動詞「～きる」「～ぬく」「～とおす」の場合－」青山学院大学大学院国際政治経済学研究科博士論文（https://www.agulin.aoyama.ac.jp/opac/repository/1000/16544/16544.pdf）

森田良行（1989）『基礎日本語辞典』角川書店.

中島紀子（2006）「複合動詞に関する一考察－「～きる」「～とおす」「～ぬく」の比較から－」『国文学踏査』18, 262-271.

大友麻子（2005）「アスペクト関数としての *cut* と「切る」」影山太郎（編）『レキシコンフォーラム No.1』ひつじ書房. pp.201-230.

杉村泰（2008）「複合動詞「－切る」の意味について」『言語文化研究叢書 7 日本語の魅力』63-79.

――――（2012）「コーパスを利用した複合動詞「Ｖ１-通す」の意味分析」『言語文化論集』34:1, 47-59.

翻訳小説を資料とした品詞比率と文書間類似度による明治中期口語文体分析

小西 光（国立国語研究所コーパス開発センター）†

The Colloquial *Genbun Itchi* Style Analysis on Translated Novels in Mid-Meiji Era by Part-of-Speech Rate and Document Similarity

Hikari KONISHI (National Institute for Japanese Language and Linguistics)

要旨

明治期の文体を論じる際、多様な文体から言文一致による口語体書き言葉成立へという変遷は指摘されているものの、その具体的な実態と詳細が明らかになっているとはいえない。本発表では明治中期に口語体で翻訳された翻訳小説を対象に「近代口語文翻訳小説コーパス」を構築し、明治 40 年代に成立したとされる口語体書き言葉への萌芽を観察する。

特徴量として名詞率に対する MVR の分布、全体の品詞比率および品詞・語彙素・出現書字形・品詞バイグラムの分布による文書間類似度を用い、『太陽コーパス』『近代女性雑誌コーパス』で「口語」とアノテーションされたデータとの比較を行った。その結果、名詞率と MVR の二次元グラフでは、『太陽』と『女性雑誌』の全データセットが翻訳小説五作品よりも近い位置にまとまって分布し、翻訳小説五作品とは異なることが明らかになった。一方、文書間類似度においては、翻訳小説五作品すべてに対して 1909（明治 42）年発行の『太陽』コアデータセットの距離が最も近いことが明らかとなった。

１．はじめに

国立国語研究所にて現在も近代語のコーパス整備が行われている。田中ほか(2012)では明治から昭和までをおよそ 15 年ごとに区切り、各時代のジャンルや文体など幅を持たせたコーパスの方向性を示している。国立国語研究所にて現在公開されているものは『明六雑誌コーパス』（明治前期）『国民之友コーパス』（明治中期）『太陽コーパス』『近代女性雑誌コーパス』（明治中期～大正期）の四つである。

一方、「近代口語文翻訳小説コーパスの構築と計量的文体研究」（研究課題番号：25770178）にて収録対象資料とした明治中期（特に明治 20 年代）の口語体翻訳小説とは、当時の文学界において初期言文一致体を試みた作家たちと密接不可分なものであり、新文体の獲得に無関係とは言えない[1]ものの、あまりその特徴が明らかにされることはなかった。口語体翻訳小説は、明治 40 年代に口語体としての書き言葉が統合・成立するその過程を捉える上で、押さえるべき資料と考える。

そこで、本発表では明治中期に口語体で翻訳された小説五作品を資料とし、その概要および品詞比率をまとめ、明治中期から大正期のコーパスである『太陽コーパス』『近代女性雑誌コーパス』（以下、『太陽』『女性雑誌』）の品詞・語彙素・出現書字形の情報を用いて文書間類似度の比較を行った。以下、２節では分析データをまとめ、３節では品詞比率と MVR、４節では各コーパスの年代別文書間類似度を比較し、５節でまとめとする。

† hkonishi@ninjal.ac.jp

1 加藤(2012)「（明治時代、）小説家は、自己の創作活動のために必要とする形式と内実を、彼の翻訳作業を通じて探索していたのだ」(pp. iv - v)

２．分析データ

２．１『太陽コーパス』『近代女性雑誌コーパス』について

2005 年に公開された『太陽コーパス』は、総合雑誌『太陽』（博文館刊）1895（明治 28）年、1901（明治 34）年、1909（明治 42）年、1917（大正 6）年、1925（大正 14）年発行の通常号全文をデータとするタグ付きコーパスである。含まれる記事数や文字数の基礎統計量については森(2014)にまとめられており、「１記事文字数、出版年ごと記事数・文字数・ジャンルにばらつきがあり（中略）非常に不均衡なコーパスである」との指摘があるなど取り扱いには注意を要する。本発表では特別な配慮は行わなかった。現在整備中の『太陽コーパス』にはコアデータと非コアデータという二種類のデータセットがあり、コアデータについては精緻な人手修正が行われ、精度の高いデータとなっている。今回の調査では発行年ごとにコアデータ（TC）と非コアデータ（TNC）を区別した。

また続いて 2006 年に公開された『近代女性雑誌コーパス』は、1894（明治 27）・1895（明治 28）年発行の『女学雑誌』31 冊（女学雑誌社）、1909（明治 42）年発行の『女学世界』6 冊（博文館）、1925（大正 14）年発行の『婦人倶楽部』3 冊（講談社）の全文をデータとするタグ付きコーパスである。『女性雑誌』には『太陽』のようなデータの区別が行われていないため、発行年ごとのデータセット（JC）としている。

両コーパスには、サンプル単位と形態素単位の両方に口語・文語（・漢文ほか）の情報が付与されており、本分析ではサンプル単位で「口語」と認定されたサンプルを利用する。サンプル単位の口語文にも、形態素単位には口語要素だけでなく文語要素（典拠・手紙ほか）が含まれるがこれらについては排除していない。

２．２「近代口語文翻訳小説コーパス」について

現在構築を進めている「近代口語文翻訳小説コーパス」の公開予定データは、表１の五作品である。このほかに現在修正中のものもあるが、資料の成立年代としては明治 20 年代を中心とした常体・口語体翻訳小説からなる形態素情報付きコーパス[2]となっている。なお、敬体の翻訳小説については、収録を予定していない。

口語体・文語体の判定については、『太陽』の文体情報付与基準と同様に「文末辞が「なり」「たり」「き」「けり」などで終わる文体は文語体、「だ」「である」「た」「です」「ます」などで終わる文体は口語体」(田中ほか 2012)とし、資料を選定した。「近代口語文翻訳小説コーパス」は基本的に全文口語文で構成されているが、『罪と罰』以外は地の文・会話文等をすべて含んだデータとなっており、『罪と罰』のみ当初地の文を分析対象としていたため、会話文や書簡文（第三回の大部分を書簡文が占める）を含んでいない（今後、品詞・形態素情報整備完了後、収録予定）。

表１に出典情報、表２に文の数、短単位の数、文の長さ、MVR[3]、名詞率[4]の値をまとめた。

『現代日本語書き言葉均衡コーパス』（以下、BCCWJ）を対象とした山崎(2014)の調査では、37 短単位数以下の文で全体の 90%をカバーしているという報告があり、五作品の文の長さが極端に長過ぎるということはなさそうではあるが、BCCWJ の文の長さの平均値よりはやや長いといえる。

また MVR については次節でも取り上げるが、小磯ほか(2010)の調査[5]によると BCCWJ 中

---

[2] 言語単位は BCCWJ を踏襲した「短単位」を採用し、品詞体系についても UniDic 品詞体系を用いた。（小椋ほか 2011）

[3] 樺島・寿岳(1965)　MVR=100*形容詞・形状詞・副詞・連体詞の数／動詞の数

[4] 樺島・寿岳(1954)では機能語を除いて名詞率を算出しているため、本稿でも同様の方法で算出した。

[5] 小磯(2010)では、分析に言語単位「長単位」を用いている。

の小説の MVR は 25～70 の間に収まり、これも文の長さ同様に大きな差異は見られず、{「玉を懐いて罪あり」「綠葉歎」} と {「洪水」『罪と罰』} の二組は近い値を示している。

表 1　「近代口語文翻訳小説コーパス」出典情報

作品名	原作者	訳者	原語	初出・刊行年	初出
あひゞき	ツルゲーネフ	二葉亭四迷	露語	明治 21(1888)年	国民之友
玉を懐(いだ)いて罪あり	ホフマン	森鴎外	独語	明治 22(1889)年	読売新聞
洪水	ブレツト、ハアト	森鴎外	独語	明治 22(1889)年	『柵草子』
綠葉歎	ドオデエ	森鴎外	独語	明治 22(1889)年	読売新聞
罪と罰	ドストエフスキー	内田魯庵	英語	明治 25(1892)年	単行本

表 2　「近代口語文体翻訳小説コーパス」文数・短単位数・文の長さ・MVR・名詞率

作品名	文数	短単位数[6]	文の長さ（短単位数／文数）	MVR	名詞率
あひゞき	159	5,557	34.95	68.46	43.06
玉を懐いて罪あり	892	25,636	28.74	47.63	53.55
洪水	124	4,429	35.72	55.90	47.51
綠葉歎	88	2,296	26.01	54.94	53.58
罪と罰	1,097	30,472	27.77	57.81	48.52
計	2,360	68,390	29.40	54.73	50.08

３.『太陽コーパス』『女性雑誌コーパス』と「近代口語文翻訳小説コーパス」の品詞比率

本節では品詞比率と MVR を用いた比較を行う。樺島・寿岳(1965)では、名詞率（以下、N 率）と MVR の関係から文章の特徴が明らかになるとした。本分析データについても、同様の手法で比較することとする。

３.１　名詞率と MVR

図 1 に「近代口語文翻訳小説コーパス」五作品と『太陽』『女性雑誌』における N 率に対する MVR の分布を示す。問題となる N 率については、「あひゞき」のみ他の四作品や『太陽』『女性雑誌』よりも値が小さく、MVR が「極めて大」とされる 56 以上の 68.5 という点から、樺島・寿岳(1965)で分類された「ありさま描写的」と言える。たしかに「あひゞき」は、語り手の視点が物陰から男女の逢引の一場面を描写するという短編であり、動作性の描写という点で、他の四作品とは異なっている。

他の四作品については、N 率は小から普通(45~54)の範囲にあり、MVR は「玉を懐いて罪あり」が普通(48～54)、「綠葉歎」「洪水」は大(54～56)、『罪と罰』は極めて大(56～)に位置している。また、「洪水」と『罪と罰』については、N＜MVR となっている。このことより、上記四作品の中では、「洪水」『罪と罰』は「ありさま描写的」、「玉を懐いて罪あり」「綠葉歎」は「動き描写的」な傾向性を持つものと考えられる。

一方、『太陽』『女性雑誌』のデータと比較をすると、N 率と MVR の関係において「あひゞき」「洪水」『罪と罰』は異なる傾向性があると言える。当然『太陽』『女性雑誌』は雑誌という性質上、小説以外の記事が含まれ、単純な比較はできない。一方で、『太陽』と『女性雑誌』というサンプリングした年代の異なるデータで、いずれも近しい値となったという点は、注目に値する。

[6] 空白・補助記号は除いた。

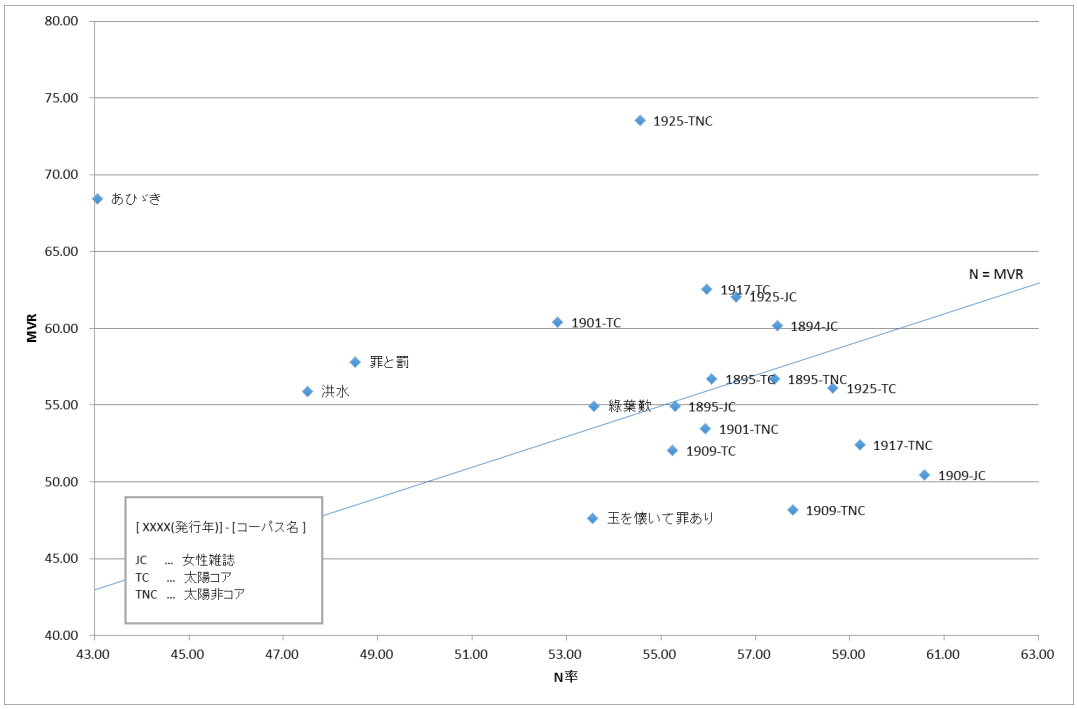


図 1　名詞率に対する MVR の分布

【例１】
此難に逢うて飾は取られたが、不思議と命を拾つた人の話に、何心なく道を行くと、突然頭を強く打たれ、其儘仆れて氣を失ひ、暫くして心付いて見れば、遙か離れた町に居て飾はなかつたといふ。家の中で殺されたものも、途で殺されたものも、撿屍の時に見ると、皆んな唯つた一つの突創が胸に在るばかり。解剖して見れば、心の臓が差し貫ぬかれてある。

（N 率：53.55　MVR:47.63 「玉を懐いて罪あり」）

【例２】
取分け自分の氣に入ツたはその面ざし、まことに柔和でしとやかで、取繕ろツた氣色は微塵もなく、さも憂はしさうで、そしてまた愛度氣なく途方に暮れた趣きも有ツた。たれをか待合はせてゐるのと見えて、何か幽かに物音がしたかと思ふと、少女はあわてゝ頭を擡げて、振り反つて見て、その大方の涼しい眼、牝鹿のものゝやうにをど／＼したのをば、薄暗い木蔭でひからせた。

（N 率：43.06　MVR:68.46 「あひゞき」）

【例３】
暫らくすると戸が少し開いて其隙間から部屋の主人が小さな眼を暗黒の中に燦つかせながら慥に猜疑の心をもて訪問者を吟味すると、溜段の上には多勢人がゐたから、やッと安神したらしく戸を排放した。少年は薄暗い前房に入ッた。壁一ト重を距てゝ奥は狭い臺所であッた。其部屋の中に黙然として屹立し不審しげにきッと少年を凝視めたは年配六十位の皺枯れて癯せこけた老婆で、鼻準透ッて鋭く尖り、陰険な色を帯びた眼光はギラ／＼人を射る様である。

（N 率：48.52　MVR:57.81 『罪と罰』）

表 3　「近代口語文翻訳小説コーパス」の品詞比率（機能語も含む）

	P	N	V	M	I	O
あひゞき	45.20	23.60	17.68	12.10	1.43	0.00
玉を懐いて罪あり	45.49	29.14	16.65	7.93	0.69	0.10
洪水	45.53	25.86	17.84	9.97	0.76	0.03
綠葉歎	45.60	29.15	15.77	8.66	0.83	0.00
罪と罰	43.55	27.39	17.87	10.33	0.85	0.00

３.２　機能語を含む作品全体の品詞比率

次に表３に「近代口語文翻訳小説コーパス」の助詞や助動詞といった機能語も含む全体の品詞比率[7]を示す。山崎(2014)の BCCWJ における品詞比率の調査（「。．！？」で終わる「通常の文」を対象とし、短単位を基準としたもの）に比べ、N の比率が 10 前後小さくなり、それ以外の V、M、I、P の値がいずれも高くなっている。山崎(2014)では「句点で終わる文に比べて疑問符、かぎ括弧で終わる文で、N の割合が低く、P の割合が多くなっているのは話し言葉的な要因が関係している可能性がある。」と指摘されている。現代語の品詞比率や考察を単純に近代語に対して適用することはできないが、BCCWJ の書籍データのうちの文学にデータを絞り、比較することを今後の課題としたい。

図２に「近代口語文翻訳小説コーパス」五作品と『太陽コーパス』『近代女性雑誌コーパス』におけるすべての品詞を対象とした品詞比率を図示する。「近代口語文体翻訳小説コーパス」では、樺島(1965)の示す通り V・M 率と N 率との間にやや相関が見られるが、『太陽』『女性雑誌』では、N 率と P 率の間に相関が見られる。これはテキストの内容（小説か評論か等）の問題と推察されるが、今後より詳細に調査していきたい。

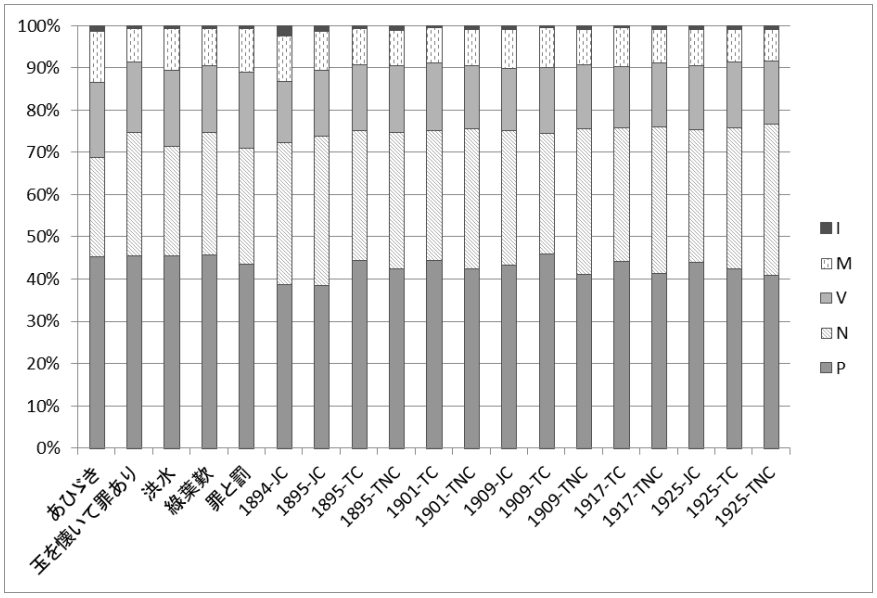


図 2　品詞比率の比較（機能語を含む）

[7] N（名詞類）：名詞、代名詞、接尾辞-名詞的、記号　V（動詞類）：動詞、接尾辞-動詞的　M（形容詞・形状詞・副詞類）：形容詞、形状詞、副詞、連体詞、接頭辞、接尾辞-形容詞的、接尾辞-形状詞的　I（接続詞・感動詞類）：接続詞、感動詞　P（助詞・助動詞類）：助詞、助動詞　O（その他）：未知語、漢文、英単語ほか（山崎 2014）

４.『太陽コーパス』『女性雑誌コーパス』と「近代口語体翻訳小説コーパス」の類似度

４.１　分析手法

以下では品詞分布・語彙素分布・出現書字形分布・品詞バイグラム分布の四種類の文体特徴量を用いた文書間類似度について検討する。各分布は頻度ベクトルの形式で保持し、頻度ベクトルのコサイン類似度を検討する。仮に比較する文書のベクトルを$\vec{s}$とし、比較される文書のベクトルを$\vec{t}$とすると、コサイン類似度は以下の式で表される：

$$\cos(\vec{s},\vec{t}) = \frac{\vec{s}\cdot\vec{t}}{|\vec{s}|\cdot|\vec{t}|}$$

通常、０から１の値をとり、文書間距離が近い（似ている）場合１に近い値を、最も文書間距離が遠い（似ていない）場合に０に近い値を取る。

品詞情報を用いた分布取得において、品詞「空白」と「補助記号-*」を排除した。UniDicの品詞体系には「名詞-普通名詞-一般」のように「[大分類]-[中分類]-[小分類]」と分類されているが、小分類まで用いている。品詞バイグラム分布において、文の先頭要素には “BOS” と当該品詞の対を特徴量として用いるが、バイグラムの前件・後件のいずれかが「空白」もしくは「補助記号-*」の場合は特徴量空間から排除してコサイン類似度の算出を行った。

４.２　各種分布による文書間類似度

表４～表７に「近代口語体翻訳小説コーパス」五作品それぞれと『太陽』『女性雑誌』の発行年別データセット（『太陽』のみコア・非コア区別あり）との文書間類似度をまとめた。

まず全体を通して共通する点を三点挙げる。一つ目は、どの特徴量においても 1894 年の『女性雑誌』データは、五作品のいずれに対しても文書間距離の値が小さく、また値の差分が、上位の値同士のそれと比較して大きい。原因を明らかにするべきであるが、次稿に

表 4　品詞分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.982	1909-TC	0.992	1909-TC	0.988	1901-TC	0.990	1909-TC	0.988	1909-TC
2	0.977	1925-JC	0.991	1895-TC	0.987	1909-TC	0.988	1917-TC	0.983	1909-JC
3	0.975	1901-TC	0.990	1901-TNC	0.985	1895-TC	0.988	1895-TC	0.983	1925-JC
4	0.971	1895-TC	0.990	1917-TC	0.984	1901-TNC	0.985	1901-TC	0.982	1895-TNC
5	0.970	1909-JC	0.989	1895-TNC	0.984	1895-TNC	0.985	1901-TNC	0.982	1895-TC
6	0.967	1901-TNC	0.989	1901-TC	0.982	1909-JC	0.985	1925-JC	0.980	1901-TC
7	0.965	1917-TC	0.988	1909-JC	0.981	1917-TC	0.984	1895-TNC	0.980	1901-TNC
8	0.964	1895-TNC	0.987	1925-JC	0.979	1909-TNC	0.983	1909-JC	0.979	1917-TC
9	0.962	1925-TC	0.986	1909-TNC	0.979	1925-JC	0.981	1909-TNC	0.979	1925-TC
10	0.961	1909-TNC	0.983	1925-TC	0.975	1895-JC	0.979	1925-TC	0.976	1909-TNC
11	0.954	1925-TNC	0.982	1917-TNC	0.971	1925-TC	0.978	1917-TNC	0.975	1917-TNC
12	0.951	1917-TNC	0.980	1925-TNC	0.970	1917-TNC	0.976	1925-TNC	0.973	1925-TNC
13	0.951	1895-JC	0.976	1895-JC	0.969	1925-TNC	0.968	1895-JC	0.970	1895-JC
14	0.917	1894-JC	0.900	1894-JC	0.897	1894-JC	0.898	1894-JC	0.897	1894-JC

表 5　語彙素分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.948	1901-TC	0.972	1925-JC	0.966	1901-TC	0.959	1909-TC	0.940	1917-TNC
2	0.945	1909-TC	0.964	1909-TC	0.960	1909-TC	0.957	1917-TC	0.938	1895-TC
3	0.933	1925-TC	0.960	1901-TC	0.956	1917-TC	0.957	1901-TC	0.932	1925-JC
4	0.928	1925-JC	0.958	1909-JC	0.950	1925-TC	0.953	1925-TC	0.931	1901-TC
5	0.925	1917-TC	0.952	1925-TC	0.948	1925-JC	0.948	1925-JC	0.924	1909-JC
6	0.916	1925-TNC	0.952	1917-TC	0.944	1925-TNC	0.946	1925-TNC	0.920	1895-JC
7	0.905	1909-JC	0.951	1901-TNC	0.937	1917-TNC	0.933	1909-JC	0.920	1909-TNC
8	0.903	1917-TNC	0.948	1925-TNC	0.933	1909-JC	0.933	1917-TNC	0.912	1925-TC
9	0.902	1895-TC	0.947	1917-TNC	0.931	1901-TNC	0.928	1895-TC	0.904	1917-TC
10	0.894	1901-TNC	0.943	1895-TC	0.929	1895-TC	0.925	1901-TNC	0.902	1901-TNC
11	0.882	1909-TNC	0.941	1895-JC	0.922	1909-TNC	0.916	1909-TNC	0.900	1895-TNC
12	0.867	1894-JC	0.937	1909-TNC	0.916	1895-JC	0.893	1895-JC	0.885	1909-TC
13	0.857	1895-TNC	0.936	1895-TNC	0.902	1895-TNC	0.891	1895-TNC	0.875	1925-TNC
14	0.845	1895-JC	0.913	1894-JC	0.881	1894-JC	0.879	1894-JC	0.858	1894-JC

表 6　出現書字形分布による文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.925	1901-TC	0.979	1925-JC	0.964	1909-TC	0.956	1909-TC	0.947	1925-TC
2	0.925	1909-TC	0.977	1909-TC	0.964	1901-TC	0.954	1925-JC	0.946	1901-TC
3	0.916	1925-JC	0.971	1901-TC	0.959	1925-JC	0.953	1925-TC	0.944	1925-JC
4	0.907	1925-TC	0.969	1925-TC	0.957	1917-TC	0.953	1901-TC	0.942	1909-TC
5	0.905	1917-TC	0.965	1909-JC	0.956	1925-TC	0.952	1917-TC	0.931	1925-TNC
6	0.893	1925-TNC	0.964	1925-TNC	0.951	1925-TNC	0.947	1925-TNC	0.925	1917-TNC
7	0.890	1917-TNC	0.963	1901-TNC	0.945	1917-TNC	0.940	1901-TNC	0.924	1917-TC
8	0.884	1909-JC	0.963	1917-TC	0.943	1909-JC	0.940	1909-JC	0.922	1901-TNC
9	0.880	1901-TNC	0.963	1917-TNC	0.940	1901-TNC	0.940	1917-TNC	0.915	1909-JC
10	0.869	1909-TNC	0.954	1909-TNC	0.933	1909-TNC	0.932	1895-TC	0.913	1895-TC
11	0.868	1895-TC	0.953	1895-TC	0.932	1895-TC	0.931	1909-TNC	0.910	1909-TNC
12	0.865	1895-JC	0.947	1895-JC	0.925	1895-JC	0.916	1895-JC	0.896	1895-JC
13	0.849	1895-TNC	0.944	1895-TNC	0.914	1895-TNC	0.914	1895-TNC	0.893	1895-TNC
14	0.847	1894-JC	0.905	1894-JC	0.887	1894-JC	0.878	1894-JC	0.867	1894-JC

表 7　品詞バイグラムによる文書間類似度

	あひゞき		玉を懐いて罪あり		洪水		綠葉歎		罪と罰	
1	0.956	1909-TC	0.983	1909-TC	0.973	1909-TC	0.971	1909-TC	0.967	1909-TC
2	0.949	1901-TC	0.973	1901-TC	0.968	1901-TC	0.962	1917-TC	0.955	1917-TC
3	0.931	1917-TC	0.972	1917-TC	0.958	1917-TC	0.955	1901-TC	0.952	1925-TC
4	0.930	1895-TC	0.968	1895-TC	0.952	1895-TC	0.949	1895-TC	0.951	1895-TC
5	0.917	1925-TC	0.956	1925-TC	0.933	1925-TC	0.940	1925-TC	0.948	1901-TC
6	0.910	1925-JC	0.949	1901-TNC	0.929	1901-TNC	0.933	1925-JC	0.940	1909-JC
7	0.907	1909-JC	0.948	1909-JC	0.927	1909-JC	0.931	1909-JC	0.939	1925-JC
8	0.902	1901-TNC	0.946	1895-TNC	0.923	1895-TNC	0.931	1901-TNC	0.936	1917-TNC
9	0.897	1895-TNC	0.946	1925-JC	0.921	1925-JC	0.925	1895-TNC	0.934	1895-TNC
10	0.887	1917-TNC	0.942	1917-TNC	0.914	1917-TNC	0.924	1917-TNC	0.932	1901-TNC
11	0.883	1895-JC	0.934	1909-TNC	0.913	1909-TNC	0.916	1925-TNC	0.929	1895-JC
12	0.883	1909-TNC	0.933	1895-JC	0.912	1895-JC	0.916	1895-JC	0.926	1925-TNC
13	0.881	1925-TNC	0.932	1925-TNC	0.905	1925-TNC	0.916	1909-TNC	0.921	1909-TNC
14	0.737	1894-JC	0.759	1894-JC	0.732	1894-JC	0.740	1894-JC	0.765	1894-JC

譲りたい。二つ目は、どの特徴量においても非コアデータである「*-TNC」の文書間距離の値が相対的に小さい[8]。これは自動解析誤りが文書間距離に影響を与えているものだと推察される。このことから、自動解析によるデータを大量に準備するよりも、少量の人手修正された翻訳小説・雑誌コーパス双方で準備することが信頼性の高い分析のためには重要であると考える。三つ目は、五作品が発表もしくは発刊された 1888 年（明治 21 年）から 1892 年（明治 25 年）に最も近いデータである『太陽』「1895-TC/TNC」と『女性雑誌』「1894/1895-JC」（明治 27、28 年）よりも 1901（明治 34）年、1909（明治 42）年との文書間距離の方が近い、つまり今回調査した特徴量においては 1894・95 年の文体よりも 1901 年・1909 年の文体の方に類似していることが読み取れる。

次に、表４～７の各分布について見ていく。

表４の品詞分布では、「洪水」以外で「1909-TC」との文書間距離が最も近い。「1909-TC」の次に文書間距離が近いデータセットは五作品すべてで異なっている。また、文書間距離の値の差分が「1894-JC」を除くと高々0.031 で抑えられ、ほぼ差がないといえる。次の表５と表６では、五作品それぞれ最も文書間距離の近いデータセットが異なっている。『罪と罰』のみ語彙素分布と出現書字形分布の文書間距離結果に差があり、他の四作品よりも値の小さい P 率（特に語彙素と出現書字形が一致する助詞）が影響しているものと推察される。最後に表７のバイグラム品詞分布だが、五作品すべてで「1909-TC」の文書間距離が最も近い。『罪と罰』と「綠葉歎」以外の三作品については、上位五データセットの文書間距離の近さが「1909-TC ＞ 1901-TC ＞ 1917-TC ＞ 1895-TC ＞ 1925-TC」の順で同じとなっている。『罪と罰』と「綠葉歎」については、上位二データセット「1909-TC ＞ 1917-TC」の順が同一である。また、他の表と比べて、文書間距離の差分が大きいことから、品詞バイ

[8] 表５「語彙素分布」の『罪と罰』のみ「1917-TNC」データセットの文書間距離が最も１に近いものとなっている。

グラム（2,495 次元）の特徴量が、データの分布を調べるのに最も適した粒度であったことが伺える（品詞・64 次元、語彙素・69,556 次元、出現書字形・106,609 次元）。

「1909-TC」にどのような記事が含まれているかというと、八サンプルすべて「文芸」の記事であり、一記事は中原青蕪による短編の翻訳である。このことから、「文芸」「小説」「文学」等のレジスタによる結果なのか、発行年代の文体による結果なのか、明確なことは指摘できないが、「翻訳小説」を「文芸」「小説」「文学」等のレジスタに含めるとすると、単純に 1909（明治 42）年前後に著された同レジスタのものに類似するという結果を重視する。

５．まとめ

本稿では、明治 20 年代の口語体翻訳小説五作品と『太陽』『女性雑誌』コーパスとの品詞比率、文書間類似度の比較を行った。３．１節では、樺島・寿岳(1965)の研究をもとに、N 率と MVR を図示化し、「あひゞき」「洪水」『罪と罰』は「ありさま描写的」、「玉を懐いて罪あり」「緑葉歎」は「動き描写的」な傾向性があることを明らかにし、『太陽』『女性雑誌』との関係があまり見られないことを示した。３．２節では、機能語を含んだ全体の品詞比率を示し、これまでの先行研究との関連性を確認したが、一方で『太陽』『女性雑誌』では N 率と P 率に相関が見られ、より詳細な調査は今後の課題とした。文書間類似度については、４．２節で五作品とも 1901 年・1909 年のデータと文書間距離が近く、品詞バイグラム分布においては五作品すべてで 1909 年のデータが最も似ているという結果が観察された。品詞の構成比率による文体的特徴（「ありさま描写的」「動き描写的」等）と文書間類似度との関連は見られなかった。

今後は、より具体的に言語現象と今回得られた結果との関連性を明らかにし、近代口語文の文体的特徴を明確に位置づけていくこととする。

謝　辞

本研究は、文部科学省科学研究費補助金若手研究(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」（平成 25～26 年度、領域代表者：小西光）による補助を得ています。

文　献

樺島忠夫(1955)「類別した品詞の比率に見られる規則性」『国語国文』24(6)、pp55-57

樺島忠夫・寿岳章子(1965)『文体の科学』綜芸舎

加藤百合(2012)『明治期露西亜文学翻訳論攷』東洋書店

小磯花絵・小椋秀樹・小木曽智信・宮内佐夜香(2010)「長単位情報に基づくジャンル間の文体に関する分析」『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（研究成果報告会）予稿集』、pp.183-190、国立国語研究所

森秀明(2014)「均衡性と代表性に配慮した『太陽コーパス』の分析法試論」『第６回コーパス日本語学ワークショップ予稿集』、pp.73-82、国立国語研究所

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』形態論情報規定集第４版（上）（下）』，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.

田中牧郎・岡島昭浩・小木曽智信・小野正弘・小島聡子・島田泰子・朱京偉・高田智和・張元哉・陳力衛・近藤明日子・須永哲矢(2012)『近代語コーパス設計のための文献言語研究成果報告書』国立国語研究所

山崎誠(2014)「言語単位と文の長さが品詞比率に与える影響」『第５回コーパス日本語学ワークショップ予稿集』、pp.233-242、国立国語研究所

中古語複合形容詞の一語性
―［名詞＋形容詞］とそれに類する複合形容詞的表現を中心に―

池上　尚（国立国語研究所コーパス開発センター）†

Compound Adjectives as One Word in Early Middle Japanese : Focusing on Noun-Adjective Compounds and the Like

Nao Ikegami (National Institute for Japanese Language and Linguistics)

要旨

名詞・評価形容詞が直接結びつく複合形容詞（候補）、名詞・評価形容詞が助詞や副詞を（複数）介して結びつく複合形容詞的表現を「日本語歴史コーパス　平安時代編」によって網羅的に抽出し、構文バリエーションの把握、コロケーション強度の数値化を行い、中古和文における複合形容詞［名詞＋評価形容詞］の一語性（名詞・形容詞の結びつきの強弱、語としての在り方）を重層的に考察した。その結果、複合形容詞［名詞＋評価形容詞］と認めるべき名詞―評価形容詞の多くが、①共時的に複合形容詞的表現にパラフレーズ可能で、語と文との境界に位置するような一語性を有していたこと、②人間のある状態についての善し悪しを表現するために産出されたと考えられること、を指摘した。

１．はじめに

ココロヨイのような名詞・形容詞の組み合わせを 1 語の複合形容詞［名詞＋形容詞］と見るか、主述関係をなす名詞―形容詞の 2 語と見るかといった語認定の問題は、内省のきかない時代の資料を扱う場合に大きな問題となる[1]。

須永(2011)は、「中古和文 UniDic」作成時の品詞情報付き中古語コーパス[2]から抽出した、名詞とヨシ／アシ／アリ／ナシとの組み合わせを対象とし、語と語とのコロケーション強度を数値化するダイス係数[3]を用いて中古語の語認定の方法を探り、“ダイス係数 0.004 以上”が一つの基準となり得ることを明らかにした。しかし、須永(2011)も指摘するように、指標の精緻化に向けては複合語候補となる 2 語の組み合わせの構文環境にも着目することが望ましい。すなわち、同じ名詞・形容詞の組み合わせでも、間に助詞を介したり（例「人の心のよきもあしきも、」紫式部日記）、連体句や副詞を伴ったりする場合があり（例「いと心よからむ人は、」同）、構文環境により 1 語としての認めやすさに差が生じるのである。

こうした観点は、コーパス開発に際しての語認定にとどまらず、古い時代を扱う複合語研究においても積極的に導入していく必要がある。これまでの先行研究や索引類では、複合語候補となり得る、前項と後項とが直接結びつくものを把握することは可能であったが、それらが有する（あるいは有しない）構文バリエーション、いわば複合語的表現までも含めた全体像については十分に知り得なかった。

---

† nikegami@ninjal.ac.jp

[1] 以下、1 語の複合形容詞であること表す場合に［名詞＋形容詞］、名詞・形容詞の 2 語が（助詞・副詞を介して）結びついていることを表す場合に名詞―形容詞と表記する。

[2] 学習用コーパス。総語数は句読点含め約 80 万語。収録作品は次の通り。伊勢物語・大和物語・土佐日記・紫式部日記・更級日記・源氏物語・竹取物語・古今和歌集仮名序・枕草子・大鏡。

[3] 中心語頻度と共起語頻度の関係から 2 語のコロケーション強度を計測する尺度である。共起頻度（組み合わさって現れた XY の語数）を中心語頻度と共起語頻度の和（組み合わせのもとになる X・Y のそれぞれの語数の和）で割って 2 倍した値である。式は次のようになる。

$$D = 2 \times \frac{\text{「XY」の語数}}{\text{X の語数＋Y の語数}}$$

本発表では如上の課題に取り組むべく、複合形容詞（的表現）と考えられる名詞―形容詞の様々な組み合わせを「日本語歴史コーパス　平安時代編」によって網羅的に抽出し、その構文パターンの観察を通して中古語における一語性（名詞・形容詞の結びつきの強弱、語としての在り方）を重層的に考察する。形容詞の中でも特に評価形容詞ヨシ／ヨロシ／アシ／ワロシ／ワルシからなるものを取り上げることで、ある複合形容詞（的表現）の類義・対義の関係にある複合形容詞（的表現）についても見ていく。

２．調査にあたって

２．１　調査対象

調査には「日本語歴史コーパス　平安時代編」（中納言 1.5.0／長単位データ 1.0）[4]を使用し、次のような検索条件式により名詞―評価形容詞のデータを抽出した。

【検索条件式の例：名詞―〔助詞／副詞〕―形容詞】

キー: ((品詞 LIKE "形容詞%" AND 語彙素読み LIKE "ヨイ")OR(品詞 LIKE "形容詞%" AND 語彙素読み LIKE "ヨロシイ")OR(品詞 LIKE "形容詞%" AND 語彙素読み LIKE "アシイ")OR(品詞 LIKE "形容詞%" AND 語彙素読み LIKE "ワロイ")OR(品詞 LIKE "形容詞%" AND 語彙素読み LIKE "ワルイ")) AND 前方共起: (品詞 LIKE "助詞%" OR 品詞 LIKE "副詞%") ON 1 WORDS FROM キー AND 前方共起: 品詞 LIKE "名詞%" ON 2 WORDS FROM キー WITH OPTIONS unit="2" AND tglWords="20" AND limitToSelfSentence="1" AND tglBunKugiri="#" AND endOfLine="CRLF" AND tglKugiri="|" AND encoding="UTF-16LE"

２．２　考察対象

データを精査する過程で、名詞―評価形容詞と見なせないもの（例「四の君によき人あはせむ」落窪物語・4）を除外し、前項名詞にかかる程度副詞・接頭辞「御」・連体句の有無についても確認した。その結果、次の表 1 に示すように、中古和文に出現する名詞―評価形容詞の構文パターンは 15 種類あることが分かった[5]（表 1 中Ⅰ・Ⅱ・Ⅲ類については後述）。

表1　中古和文における名詞―評価形容詞

類		構文					
Ⅰ	A		名詞				形容詞
	A+	程度副詞*					
Ⅲ	a	接頭辞「御」／連体句					
Ⅱ	B		名詞	助詞			形容詞
	B+	程度副詞*					
Ⅲ	b	接頭辞「御」／連体句					
Ⅱ	C		名詞	助詞	助詞		形容詞
	C+	程度副詞*					
Ⅲ	c	接頭辞「御」／連体句					
	D		名詞			副詞	形容詞
	d	接頭辞「御」／連体句					
	E		名詞	助詞		副詞	形容詞
	e	接頭辞「御」／連体句					
	F		名詞	助詞	助詞	副詞	形容詞
	f	接頭辞「御」／連体句					

*程度副詞に類する形容詞連用形イミジク・又無クを含む。

[4] 総語数（短単位）は 738153 語（空白・記号・補助記号含め 871462 語）。収録作品（その語数）は次の通り。古今和歌集（31288）・竹取物語（10317）・伊勢物語（13824）・大和物語（23090）・平中物語（12403）・土佐日記（6685）・落窪物語（54583）・堤中納言物語（15699）・枕草子（66044）・源氏物語（445675）・和泉式部日記（10891）・紫式部日記（17440）・更級日記（14659）・讃岐典侍日記（15555）。

[5] 構文のパターンとして想定されるものは他にもある（例えば、名詞が助詞 3 つを介して形容詞と結びつくもの）が、用例の得られたもののみ表 1 に掲載した。

このうち、（類を問わずに）延べ語数が 3 以上の名詞―評価形容詞を考察対象とする。

なお、「中古和文 UniDic」で 1 短単位（1 語の複合形容詞）とされている折好い・心地良い・快い・言好い・様良い・根良い・折悪しい・口悪しい・心悪しい・様悪しい・物悪しい・心悪（わろ）い・人悪（わろ）い・人悪（わる）いは、名詞・形容詞の 2 短単位に分割した上で、A／A+に分類した。

２．３　「一語性」をどのように考えるか

図 1 に示したように、名詞・形容詞が直接結びつく A・A+の場合、複合形容詞候補として十分な条件を備えていると見なせる（Ⅰ類）。しかし、名詞・形容詞が助詞を介して結びつく B・B+・C・C+の場合、1 語とは見なせない（Ⅱ類）。そして、D・d・E・e・F・f のように名詞・形容詞の間に形容詞を修飾する副詞が挟まる場合や、a・b・c・d・e・f のように名詞にかかる接頭辞や連体句が存在する場合は、2 語の隔たりは一層強く感じられる（Ⅲ類）。Ⅰ類はダイス係数の大小、Ⅱ・Ⅲ類は助詞の数の多少などを基準に、より複雑な段階を設定することもできようが、ここではひとまず図 1 のように把握する。なお、図 1 中、薄い網掛けで表したように、それぞれの構文が一語化の途中である可能性ももちろんある。

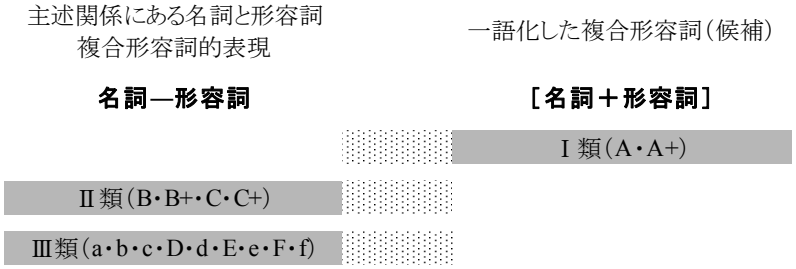


図1　一語性

実際には、ある名詞・形容詞の組み合わせがⅠ・Ⅱ・Ⅲ類のいずれかひとつに分類されることは少なく、複数の類にまたがり複雑な様相を呈する場合が多い。そうした分布状況を踏まえた上で、名詞―形容詞の一語性を検討する必要がある。

３．考察

考察対象の名詞―評価形容詞の一覧を表 2～5 としてまとめた。各構文の延べ語数と類、類それぞれの占める割合を示した。また、Ⅰ類（A・A+）についてはダイス係数（その算出に必要な名詞 X・形容詞 Y の語数）を掲げ、『日本国語大辞典（第二版）』（以下、『日国』）における立項状況についても記載した。なお、以下でコロケーション強度の強弱について触れる場合、須永(2011)の明らかにしたダイス係数 0.004 を基準としている。

３．１　名詞―ヨシ

３．１．１　複合形容詞候補

名詞―ヨシのうち、Ⅰ類としてのみ現れ、かつ、コロケーション強度の強いものに**声／折―ヨシ**がある。これらは複合形容詞としての条件を備えていると考えられる。

(1)伊勢の海ならねど、清き渚に貝や拾はむなど、声よき人にうたはせて、我も時々拍子とりて、声うち添へたまふを、　（源氏物語・明石）【Ⅰ類(A)】

(2)「さうざうしくねぶたかりつる。をりよくものしたまへるかな」　（源氏物語・常夏）【Ⅰ類(A)】

名詞―ヨシの中で注目されるのは、コロケーション強度が強いⅠ類として現れながらⅡ・Ⅲ類にわたるものが多いことである。(i)人間の姿や形といった見た目の描写に（も）用

表2 名詞―ヨシ

連体句	程度副詞	接頭辞「御」	名詞	助詞	助詞	**副詞**	形容詞	計	類			%	D係数	X+Y	X:名詞	Y:形容詞	日国
様―ヨシ								26									立項
	イト		様				ヨシ	5	A+	Ⅰ		92.3%	**0.0170**	2827	2191	636	
	ヲサヲサ		様				ヨシ	1									
			様				ヨシ	18	A								
		有	様				ヨシ	2	a	Ⅲ		7.7%	―				
心―ヨシ								25									立項
	イト		心				ヨシ	5	A+	Ⅰ		76.0%	**0.0123**	4069	3433	636	
	アマリ		心				ヨシ	1									
			心				ヨシ	13	A								
			心	ノ			ヨシ	1	B	Ⅱ		4.0%					
			心			イト	ヨシ	1	D	Ⅲ		20.0%	―				
			心	ナド	ハ	イト	ヨシ	1	F								
		有	心				ヨシ	2	a								
		有	心	ナム		イト	ヨシ	1	e								
形(カタチ)―ヨシ								22									立項
			形[カタチ]				ヨシ	14	A	Ⅰ		63.6%	**0.0301**	930	294	636	
			形[カタチ]	ナド			ヨシ	2	B	Ⅱ		13.6%					
			形[カタチ]	ハ			ヨシ	1									
			形[カタチ]			イト	ヨシ	3	D	Ⅲ		22.7%	―				
		有	形[カタチ]	ノ			ヨシ	1	a								
		有	形[カタチ]	ナド	モ	イト	ヨシ	1	f								
仲―ヨシ								8									立項
			仲				ヨシ	3	A	Ⅰ		37.5%	**0.0082**	734	98	636	
		有	仲				ヨシ	3	a	Ⅲ		62.5%	―				
		有	仲			イト	ヨシ	1	d								
		有	仲	ハ		イト	ヨシ	1	e								
顔―ヨシ								5									立項
			顔				ヨシ	2	A	Ⅰ		40.0%	**0.0042**	954	318	636	
			顔	ノ		イト	ヨシ	1	D	Ⅲ		60.0%	―				
			顔	モ		イト	ヨシ	1									
			顔	コソ		イト	ヨシ	1									
気色(ケシキ)―ヨシ								4									
			気色[ケシキ]			イト	ヨシ	2	D	Ⅲ		100.0%	―				
		有	気色[ケシキ]				ヨシ	2	a								
声―ヨシ								4									
			声				ヨシ	4	A	Ⅰ		100.0%	**0.0072**	1116	480	636	
丈立チ―ヨシ								4									
			丈立チ				ヨシ	2	A	Ⅰ		50.0%	**0.0062**	645	9	636	
			丈立チ			イト	ヨシ	2	D	Ⅲ		50.0%	―				
人柄―ヨシ								4									
			人柄	モ		イト	ヨシ	3	D	Ⅲ		100.0%	―				
			人柄	ノ		イト	ヨシ	1									
事―ヨシ								3									立項
			事				ヨシ	2	A	Ⅰ		66.7%	0.0005	7556	6920	636	
有			事	ヲ	ゾ		ヨシ	1	c	Ⅲ		33.3%	―				
人―ヨシ								3									
			人	ノ			ヨシ	1	B	Ⅱ		33.3%					
有			人				ヨシ	1	a	Ⅲ		66.7%	―				
有			人	ゾ			ヨシ	1	b								
折―ヨシ								3									立項
			折				ヨシ	3	A	Ⅰ		100.0%	**0.0050**	1204	568	636	

いられる**様／形（カタチ）／顔／丈立チ―ヨシ**、(ii)人間の気質・心身の状態を表す**心―ヨシ**、(iii)人間関係を表現する**仲―ヨシ**がこれにあたる。これらは、一語化した複合形容詞として振る舞いながらも、多様な構文を展開する複合形容詞的表現としても用いられている。

(3)涙のこぼるるさまぞ、さまよき人もなかりける。　（堤中納言物語）【Ⅰ類(A)】
(4)かたちいとよく、心もをかしき人の、　（枕草子・250）【Ⅲ類(D)】
(5)この君たち、御仲いとよし。　（源氏物語･若菜下）【Ⅲ類(d)】

なお、(ii)心―ヨシは叙述対象により表す意味が異なる。すなわち、他者の〈心が良い〉であれば評価的意味〈気立てが良い〉を表し（Ⅰ・Ⅱ・Ⅲ類：16 例）、自己の〈心が良い〉であれば感覚的意味〈気持ちが良い・快い〉を表す（Ⅰ類：9 例）[6]。興味深いのは、心―ヨシの対義表現が、評価的意味〈気立てが悪い〉では心―アシ、感覚的意味〈気持ちが悪い・不快だ〉では心―ヨシの否定表現／心地―アシである点である（後述）。

(6)もとの妻も、心いとよく、今の妻もにくき心なく、いとよく語らひてゐたりけり。…
…もとの妻、いと心よき人なれば、男にもいはでのみありわたりけれども、
（大和物語）【Ⅲ類(D)】・【Ⅰ類(A+)】
(7)［車を］いと心よう言ひて貸したるに、（枕草子・326）【Ⅰ類(A+)】

３．１．２　複合形容詞的表現

Ⅰ類として現れ得てコロケーション強度の強い 3.1.1 とは反対に、Ⅰ類として現れ得るがコロケーション強度の弱いものに**事―ヨシ**がある。この他、Ⅱ・Ⅲ類としてのみ現れる**気色（ケシキ）／人柄／人―ヨシ**がある。

(8)「よき御男ぞいで来む」とあはする［＝夢解きをする］に、この女、けしきいとよし。
（伊勢物語）【Ⅲ類(D)】
(9)人柄もいとよくおはすれば、あまた参り集まりたまふ中にもすぐれて時めきたまふ。
（源氏物語・賢木）【Ⅲ類(D)】

３．２　名詞―ヨロシ

表3　名詞―ヨロシ

	連体句	程度副詞	接頭辞「御」	名詞	助詞	助詞	副詞	形容詞	計	類		%	D係数	X+Y	X(名詞)	Y(形容詞)	日国
心地―ヨシ									5								
				心地	ハ			ヨロシ	1	B	Ⅱ	20.0%					
			有	心地				ヨロシ	2	a	Ⅲ	80.0%					
			有	心地	モ			ヨロシ	1	b			―				
			有	心地	ノ			ヨロシ	1								
気色(ケシキ)―ヨシ									3								
				気色[ケシキ]				ヨロシ	1	A	Ⅰ	33.3%	0.0017	1199	1041	158	
			有	気色[ケシキ]				ヨロシ	2	a	Ⅲ	66.7%	―				

名詞―ヨロシには、Ⅰ類として現れ得るがコロケーション強度の弱い**気色（ケシキ）―ヨロシ**、Ⅱ・Ⅲ類としてのみ現れる**心地―ヨロシ**がある。いずれも複合形容詞的表現である。また、心地―ヨロシ 5 例は全て『源氏物語』の用例であり、一般的な表現であったかは不明である。名詞―ヨロシに複合形容詞と認めるべきものはないようである[7]。

(10)「心地はよろしくなりにてはべるを、かの宮のなやましげにおはすらむに、……」
（源氏物語・若菜下）【Ⅱ類(B)】
(11)＜帰りたまはむには、御としみをぞしたまはむ。北の方けしきよろし＞と見て、
（落窪物語・1）【Ⅰ類(A)】

３．３　名詞―アシ

[6] 中古の「心」は、人間が基本的に抱き続けている思い・気持ちと人間の性質・心持ちとを表す（中尾 1999）。
[7] 『日国』においてもヨロシを後項に持つ複合形容詞は立項されておらず、小見出しとして事―ヨロシが挙げられるのみである（事―ヨロシはⅠ類(A) 1 例、Ⅲ類(a) 1 例の計 2 例のため表未掲載）。

表4 名詞―アシ

連体句	程度副詞	接頭辞「御」	名詞	助詞	助詞	**副詞**	形容詞	計	類			%	D係数	X+Y	X(名詞)	Y(形容詞)	日国
心地―アシ								35									小見出し
	イト		心地				アシ	1	A+	I	45.7%	**0.0243**	1317	1090	227	
			心地				アシ	15	A								
			心地	ノ			アシ	3	B	II	17.1%					
			心地	ハ			アシ	1								
			心地	モ			アシ	2								
			心地	ナド	ヤ		アシ	1	C		2.9%					
		有	心地				アシ	2	a	III	5.7%					
			心地	ナム		**イト**	アシ	1	E		28.6%	―				
			心地	ハ		**イト**	アシ	1								
			心地	コソ		**イト**	アシ	1								
			心地	ノ		**イト**	アシ	3								
			心地	モ		**イト**	アシ	1								
		有	心地	モ		**イト**	アシ	2								
			心地	モ		**些カ**	アシ	1								
様―アシ								24									立項
	イト		様				アシ	1	A+	I	95.8%	**0.0190**	2418	2191	227	
			様				アシ	22	A							
有			様				アシ	1	a	III	4.2%	―				
気色(ケシキ)―アシ								17									小見出し
	イト		気色[ケシキ]				アシ	1	A+	I	35.3%	**0.0095**	1268	1041	227	
			気色[ケシキ]				アシ	5	A							
	イト		気色[ケシキ]	モ			アシ	1	B+	II	5.9%					
		有	気色[ケシキ]				アシ	7	a	III	58.8%	―				
		有	気色[ケシキ]	ノ			アシ	1	b							
		有	気色[ケシキ]			**イト**	アシ	1	d							
有			気色[ケシキ]			**甚ダ**	アシ	1								
折―アシ								11									立項
			折				アシ	11	A	I	100.0%	**0.0277**	795	568	227	
乱リ心地―アシ								5									
			乱リ心地				アシ	2	A	I	40.0%	**0.0160**	250	23	227	
			乱リ心地	ノ			アシ	3	B	II	60.0%	―				
為―アシ								4									
			為				アシ	1	A	I	25.0%	**0.0045**	446	219	227	
		有	為				アシ	3	a	III	75.0%	―				
心―アシ								3									立項
	イミジク		心				アシ	1	A+	I	66.7%	0.0011	3660	3433	227	
			心				アシ	1	A							
			心	ナド			アシ	1	B	II	33.3%	―				
手―アシ								3									
有			手	ナド			アシ	1	b	III	100.0%	―				
			手	モ		**イト**	アシ	1	E							
			手	ハ		**イト**	アシ	1								
仲―アシ								3									小見出し
	少シ		仲				アシ	1	A+	I	100.0%	**0.0185**	325	98	227	
			仲				アシ	2	A							
形(ナリ)―アシ								3									
			形[ナリ]				アシ	2	A	I	66.7%	**0.0167**	239	12	227	
			形[ナリ]	ノ		**イト**	アシ	1	E	III	33.3%	―				
物―アシ								3									立項
			物				アシ	2	A	I	66.7%	0.0011	3709	3482	227	
			物	ノ			アシ	1	B	II	33.3%	―				

３．３．１　複合形容詞候補

名詞―アシのうち、Ⅰ類としてのみ現れ、かつ、コロケーション強度の強いものに**折／仲―アシ**がある。対義関係にある折／仲―ヨシ（前述）とともに、複合形容詞と言える。

(12)いつぞやも参り来てはべりしかど、折あしうてのみ帰れば、
（和泉式部日記）【Ⅰ類(A)】

(13)すこし仲あしうなりたるころ、文おこせたり。　（枕草子・80）【Ⅰ類(A+)】

名詞―アシの中で目立つのは、名詞―ヨシと同様に、コロケーション強度が強いⅠ類として現れながらⅡ・Ⅲ類にわたるものの多さである。(i)人間の姿や形といった見た目について（も）描写する**様／形（ナリ）―アシ**の他、(ii-ii)人間の心身の状態の表現に用いられる**心地／気色（ケシキ）／乱リ心地―アシ**[8]がこれにあたる。これらは、複合形容詞としての条件を十分に満たすものであるが、複合形容詞的表現としても様々な構文を展開している。

(i)人間の姿や形といった見た目の描写において、プラス評価の表現には様／形（カタチ）／顔／丈立チ―ヨシなど安定して用いられるバリエーションがある（前述）が、マイナス評価の表現において定着しているのは様／形（ナリ）―アシのみである。もちろん前項名詞の指す意味領域の相違も考慮しなければならない[9]が、人間の見た目の描写として広く捉えた場合に、評価性によって表現形式のバリエーションが異なるのは興味深い。

(14)この、いと言ふかひなく、情なく、さまあしき人なれど、ひたおもむきに二心なきを見れば、心やすくて年ごろをも過ぐしつるなり。　（源氏物語・東屋）「【Ⅰ類(A)】
(15)落窪をさしのぞいて見たまへば、なりのいとあしくて、　（落窪物語・1）【Ⅲ類(E)】

(ii-ii)人間の心身の状態の表現においては、その評価性によって一語性に差異が見られる。すなわち、マイナス評価を表す心地／気色（ケシキ）／乱リ心地―アシは複合形容詞としての性格をも有するのに対し、プラス評価を表す心地―ヨシ（Ⅱ類(B) 1例のため表未掲載）／ヨロシ、気色（ケシキ）―ヨシ／ヨロシは複合形容詞的表現である（前述）。

(16)「心地なむいとあしき」とて臥したれば、　（落窪物語・2）【Ⅲ類(E)】
(17)楫取、また鯛持て来たり。米、酒、しばしばくる。楫取、気色悪しからず。
（土佐日記）【Ⅰ類(A)】

３．３．２　複合形容詞的表現

Ⅰ類として現れ得るがコロケーション強度の弱いものに**為／心／物―アシ**[10]がある。また、Ⅱ・Ⅲ類としてのみ現れるものには**手―アシ**がある。

心―アシは(ii-i)人間の気質の描写に用いられ、他者の〈心が悪い〉つまり〈気立てが悪い〉という評価的意味を表す。前述したように、対義関係にある心―ヨシは、(ii-i)人間の気質だけでなく、心―アシには見られない(ii-ii)人間の心身の状態の描写にも用いられる[11]。

(18)かたちにくさげに心あしき人。　（枕草子・135）【Ⅰ類(A)】
(再掲)［車を］いと心よう言ひて貸したるに、　（枕草子・326）【Ⅰ類(A+)】

物―アシは『日国』に立項され、初出例として『落窪物語』が挙げられている。ただし、今回の調査によると、物―アシはコロケーション強度の弱い複合形容詞的表現と考えられ、またすべて『落窪物語』の用例であることから、広く用いられた表現とは考えにくい。

---

8 「心地」は場所や環境などにより変化する心情・気分を指す（中尾1999）のに対し、「気色（ケシキ）」は感受者が感受して初めて存在する、眼前にない個別的な人・事物の状態・動作等の現れを指す（辛島2010）という相違がある（気色（ケシキ）―アシとしては専ら人間の心理状態・機嫌を描写するようである）。
9 中世後期末～近世初期における様態を表す語彙の意味記述は、小野(1991)に詳しい。
10 物―アシのような物―形容語の「物」が接頭辞であるか名詞であるかについては諸説あるところ（東辻1997・池上印刷中 など参照）だが、ここではひとまず名詞と考えておく。
11 『日国』「こころあし」には、心身の状態を言う「(2)気分が悪い。病気である。」があり、春曙抄本『枕草子』「いささか心あしなどいへば、常よりも近く臥して、物くはせいとほしがり」を初出例として挙げる。

(19)げに今宵は三日の夜なりけるを。物のはじめに、ものあしう思ふらむ。
（落窪物語・1）【Ⅰ類(A)】

３．４　名詞―ワロシ／ワルシ

表5 名詞―ワロシ／ワルシ

連体句	程度副詞	接頭辞「御」	名詞	助詞	助詞	副詞	形容詞	計	類		%	D係数	X+Y	X(名詞)	Y(形容詞)	日国
人―ワロシ								50								立項
	イト		人				ワロシ	5	A+	I	100.0%	**0.0146**	6839	6764	75	
	イトド		人				ワロシ	2								
	少シ		人				ワロシ	2								
	一際		人				ワロシ	1								
	又無ク		人				ワロシ	1								
			人				ワロシ	39	A							
人―ワルシ								7								立項
			人				ワルシ	7	A	I	100.0%	0.0021	6783	6764	19	

延べ語数 3 以上の名詞―ワロシ／ワルシに「人」を前項とするものがある。**人―ワロシ**はⅠ類としてのみ現れ、かつ、コロケーション強度の強い複合形容詞と呼べるが、**人―ワルシ**はⅠ類として現れ得るがコロケーション強度が弱いため 1 語と認めがたい。しかし、名詞・評価形容詞を単純に足した意味でなく、〈他人に対して体裁が悪い・みっともない〉さまを表している[12]ことから、人―ワロシ／ワルシはともに 1 語として認めてよいだろう[13]。なお、こうした意味は人―アシにはない（Ⅱ類(B) 1 例のため表未掲載）。

(20)猿楽がましくわびしげに人わろげなるなど、さまざまに、げにいとなべてならず、さま異なるわざなりけり。（源氏物語・少女）【Ⅰ類(A)】
(21)都を遠ざからんも、古里おぼつかなかるべきを、人わるくぞ思し乱るる。
（源氏物語・須磨）【Ⅰ類(A)】

３．５　名詞―評価形容詞

3.1 から 3.4 までの考察を踏まえた上での全体の傾向や、補足すべき点について述べる。

３．５．１　複合形容詞候補の一語性

中古和文における複合形容詞候補の一語性の特徴として、コロケーション強度の強いⅠ類としてのみ現れる名詞―形容詞よりも、コロケーション強度の強いⅠ類として現れながらⅡ・Ⅲ類にわたる名詞―形容詞の多いことが挙げられる。このことから、中古和文における［名詞＋評価形容詞］候補の多くが、“語としてのまとまりを維持しつつも、様々な構文バリエーションを展開し得る一語性”を有していると考えられる。単に複合形容詞であると認定するだけでなく、こうした一語性についてあえて指摘するのは、複合形容詞と文との関係を考える場合に重要な観点となるためである。

言うまでもないが、複合形容詞［名詞＋形容詞］には、㋐意味変化の生じるものと㋑意味変化の生じないものがある。㋐の方が㋑よりも語としてのまとまりが強く感じられ、一語性に相違が見られる。一方で、名詞―助詞―形容詞のような文にも、㋐意味変化の生じるものと㋑意味変化の生じないものがある。㋐は一般に慣用句と呼ばれる。

前項・後項のコロケーション強度によって 1 語と認められる複合形容詞［名詞＋形容詞］の㋑が、共時的に名詞―助詞―形容詞のような文の㋑にパラフレーズ可能である場合、「複

[12] 人・ワロイ／ワルイを単純に足した〈人となりが悪い〉が専ら自己に対して用いられることで、語用論的意味である〈自分が〈人となりが悪い〉〉＝〈他人に対して体裁が悪い〉を表すようになった、という意味変化が考えられよう。なお、異なる立場に、「複合形容詞化することにより「人から～れる」というヴォイス性を持った表現になっている」（p.8）とする漆谷(2012)がある。
[13] 〈短気である〉さまを表す腹―アシ（Ⅰ類(A) 2 例のため表未掲載）も同じ条件で 1 語と認められよう。

合形容詞と文との近接現象」（山本 1996:47）が問題になる。これは、中古和文に散見される、コロケーション強度の強いⅠ類として現れながらⅡ・Ⅲ類にわたる名詞―形容詞が多いという現象を考える際の問題そのものである。“語としてのまとまりを維持しつつも、様々な構文バリエーションを展開し得る一語性”を有するものは、複合形容詞である一方で、語と文との境界に位置する言語表現と考えられるのではなかろうか。

３．５．２　複合形容詞候補の表す意味領域

中古和文における複合形容詞候補のうち、両極の評価性が描写され得る意味領域を挙げれば、(i)人間の姿や形といった見た目（様―ヨシ／アシ）、(ii-i)人間の気質（心―ヨシ／アシ）、(ii-ii)人間の心身の状態（心―ヨシ・心地―アシ）、(iii)人間関係（仲―ヨシ／アシ）、(iv)時期・機会（折―ヨシ／アシ）がある。(i)～(iii)から明らかなように、特に“人間”の描写に関わる意味領域の名詞―評価形容詞が多い。

そもそも、日本語の中で生産性のある複合形容（動）詞は、「叙述対象が、語内の名詞と部分－全体の関係にあるものに限られている」（由本 2009:219）。このことを踏まえれば、中古和文の複合形容詞候補の多くは、人間（＝全体）を描写するために、“人間の外形／内部的状態（気質・心身の状態）／他者と築く関係性”を表す名詞（＝部分）と評価形容詞とが結びつき産出された表現である言えるのではなかろうか。

３．５．３　韻文／散文の別

和歌の用例は次に挙げる延べ語数 1 のもののみであった。

(22)いで人は言（こと）のみぞよき月草のうつし心は色ことにして
（古今和歌集・14）【Ⅱ類(C)】

(23)月夜よし夜よしと人に告げやらば来てふに似たり待たずしもあらず
（古今和歌集・14）【Ⅰ類(A)】・【Ⅰ類(A)】

和歌中に複合形容詞（的表現）がないわけでなく、例えば、『日国』や「中古和文 UniDic」で複合形容詞として認められている［甲斐＋ナシ］は 11 例、複合形容詞的表現であるⅢ類(a)・(b)は 18 例ある[14]。名詞率が高く MVR（100×相の類の比率／用の類の比率）が低い「要約的な文章」と考えられる中古和歌（冨士池 2014）ゆえに、複合形容詞に限らず形容詞それ自体が地の文・会話文に比べて出現しにくいのかもしれない。

４．おわりに

本発表では、中古和文における複合形容詞［名詞＋形容詞］の一語性を探るために、名詞と評価形容詞との間に助詞や副詞を介するような複合形容詞的表現を含めた名詞―評価形容詞の調査・考察を行った。その結果、中古和文における名詞―評価形容詞それぞれの構文バリエーションの全体像を明らかにしただけでなく、この頃の複合形容詞［名詞＋評価形容詞］の候補に 2 つの特徴があることを指摘した。第一に、前項と後項とのコロケーション強度が高く複合形容詞として認められそうな名詞―評価形容詞であっても、それらの多くは共時的に複合形容詞的表現にパラフレーズ可能であり、語と文とを行き来する一語性を有していたという点である。第二に、一語化していると考えられる名詞―評価形容詞には、人間を叙述対象として、その部分・属性の善し悪しを表現するために産出されたと思われるものが目立つという点である。

今回は評価形容詞に限定したが、如上の傾向が名詞―形容詞全般に指摘し得るのかどうか確認する必要がある。調査対象を広げ考察を発展させていく中で、中古和文における複合形容詞［名詞＋形容詞］と文との関係についても検討していきたい。

14 「甲斐」が掛詞になり得ることも関係しているか。

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：田中牧郎）による成果の一部である。

文献

池上　尚(印刷中)「モノクサシの語史―嗅覚表現〈くさい〉から性向表現〈ものぐさ〉へ―」
石川慎一郎(2008)「コロケーションの強度をどう測るか―ダイス係数，ｔスコア，相互情報量を中心として―」『言語処理学会第 14 回大会チュートリアル資料』40:50
漆谷広樹(2012)「古代語・現代語の複合形容詞の比較―名詞＋形容詞の複合形容詞の場合―」『愛知大学文学論叢』146　pp.236-211
小野正弘(1991)「室町末期から江戸初期における《様態・形態》を表す語彙―「恰好」の中立的意味の成立を考えるために―」『日本近代語研究』1　pp.519-541
辛島美絵(2010)『古代の〈けしき〉の研究―古文書の資料性と語の用法―』清文堂出版
須永哲矢(2011)「コロケーション強度を用いた中古語の語認定」『国立国語研究所論集』2 pp.91-106
西尾寅弥(1972)『国立国語研究所報告 44 形容詞の意味・用法の記述的研究』秀英出版
中尾比早子(1999)「「心」と「心地」」『実践国文学』53　pp.237-254
東辻保和(1997)『もの語彙こと語彙の国語史的研究』汲古書院
飛田良文・浅田秀子(1991)『現代形容詞用法辞典』東京堂出版
冨士池優美(2014)「品詞比率からみる中古和文テキストの特徴」『日本語学会 2014 年度春季大会予稿集』pp.185-190
村田菜穂子(2005)『形容詞・形容動詞の語彙論的研究』和泉書院
山本清隆(1996)「複合語と文の境界」『日本語学』15:9　pp.41-49
由本陽子(2009)「複合形容詞形成に見る語形成のモジュール性」『語彙の意味と文法』くろしお出版　pp.209-229

関連 URL

「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

二字漢語名詞サ変用法の変化
―『太陽コーパス』『BCCWJ』を用いて―

間淵 洋子（国立国語研究所 コーパス開発センター）†

Changes in the Usage of Sino-Japanese Two-Character *Sahen* Verbs: Based on the Analysis of *Taiyo Corpus* and the *BCCWJ*

MABUCHI, Yoko (Center for Corpus Development, NINJAL)

１．はじめに

国立国語研究所コーパス開発センターでは，現在「通時コーパス」プロジェクトの一環として，形態論情報付きの近代語コーパスを構築している。これまでに，2012 年『明六雑誌コーパス』，2014 年『国民之友コーパス』が公開され，今後も資料を拡充していく計画である。その一つが雑誌『太陽』であり，2005 年に公開された『太陽コーパス』を増補改訂し，新たに形態論情報付きコーパスとして構築し直す準備を進めている。

国語研究所が中心となって開発しているコーパスでは，話し言葉（『日本語話し言葉コーパス：CSJ』），現代語（『現代日本語書き言葉均衡コーパス：BCCWJ』），古典語（『日本語歴史コーパス：CHJ』）と，収録する言語対象が変わっても，全て斉一な枠組みによる形態論情報の付与がなされている。これにより，コーパスを横断的（共時的，通時的）に分析することが可能となるという大きな利点があるが，一方で，通時的に見た時に品詞性の異なる語が存在し，コーパスへの品詞情報付与に際して問題となる場合がある。

特に，近代から現代にかけて漢語の品詞用法に変化が見られることは，池上(1953,1954)，鈴木丹士郎(1998)，鈴木英夫(2005)，永澤(2010)等，これまで多く言及されてきた。例えば，現代語においては，そのほとんどがいわゆる形容動詞語幹として用いられる漢語「複雑」は，近代において「スル」を伴うサ変動詞用法（以下「サ変用法」）を持つ。

(1)　鮪と鰹は魚類の中で最とも進歩したもので、その身體の構造が非常に【複雑】して居るのみならずいろいろな點が他の魚類と劃然たる區別を有つて居る。（『太陽』1925, 岸上鎌吉「鰹と鮪に関する新研究」）

漢語「複雑」は，コーパスの形態素解析用辞書において「名詞-普通名詞-形状詞可能」という品詞を与えられている。形状詞とはいわゆる形容動詞語幹に相当し，上記品詞は普通名詞あるいは形状詞として機能することを意味する。しかし，近代語において「複雑」は，名詞でも形状詞でもなく，サ変動詞として用いられる例があり，付与される品詞情報との間に乖離が見られる。

本発表では，このような問題を生じる漢語の把握を目的とし，二字漢語名詞のサ変用法について，『太陽コーパス』『現代日本語書き言葉均衡コーパス（以下 BCCWJ）』を用い総合的な調査を行う。その上で，近代語-現代語間の品詞性変化の有無や，サ変用法比率の変化について，実態を報告する。

† mabuchi@ninjal.ac.jp

２．調査概要

２．１　コーパス

調査には, 2005 年に公開された『太陽コーパス』, および, 2011 年に公開された『BCCWJ』を用いた。

『太陽コーパス』は，言文一致を経て口語体による書き言葉が安定し普及する時期（明治時代後期～大正時代）の書き言葉を代表できるコーパスとして作られたものであり，月刊総合雑誌『太陽』（博文館）の明治 28（1895）年，明治 34（1901）年，明治 42（1909）年，大正 6（1917）年，大正 14（1925）年について，広告や著作権処理ができなかった記事を除くほぼ全文を対象にした約 1450 万字からなるデータである。分量の多さ，ジャンル・文体・著者等の多様さから，近代における様々な言語事象を観察するのに有用な調査対象資料である。

『BCCWJ』は，現在日本において入手可能な唯一の均衡コーパスであり，書籍，雑誌，新聞，ブログ，教科書，法律といった様々なメディアから 1 億 430 万語のデータを格納する，現代語のサンプルとして好適な調査対象資料である。

『太陽コーパス』に対しては, 近代文語文を対象とする形態素解析辞書「近代文語 UniDic」（小木曽 2009）と旧仮名遣いの口語文を対象とする形態素解析辞書（小木曽 2012）を用いて形態素解析を行い，形態論情報を付与したデータが国立国語研究所の形態論情報データベース（小木曽・中村 2011）に格納されている。『BCCWJ』の形態素解析情報データも，同じデータベースに格納されているため，本発表では，このデータベースの 2013 年 12 月時点の短単位情報データを用いた[1]。データ量（自立語）は『太陽コーパス』5,034,799 語，『BCCWJ』58,823,987 語である。

２．２　調査対象表現の抽出

本研究で調査対象とするのは，二字漢語名詞のサ変用法である。

今回，調査対象を二字漢語に絞るのは，一字漢語名詞のサ変用法は，「スル」との結合度が高く文法的な振る舞いが二字漢語のそれとは異なり，また，それを反映してコーパスの単位・品詞体系においても，二字漢語＋「スル」が名詞＋動詞の 2 単語となるところ，一字漢語＋「スル」は全体で動詞 1 単語となるという大きな差があるためである。また，三字以上の漢語についても，二字漢語が元になった複合語が多く，元となる二字漢語の分析を先立って行う必要があると思われるため，今回は扱わない。

調査対象表現である二字漢語名詞のサ変用法の例を採集するために，形態論情報データベース中『太陽コーパス』『BCCWJ』の各コーパスから，以下の検索条件に合致する用例を抽出した[2]。

---

[1]データベース内の形態論情報には誤りが含まれる。また，『太陽コーパス』は整備途中のものであり，今後データの変更に伴い，本稿に挙げた数値も変動する場合がある。

[2] 検索には SQL を用いた。

```
select c.lemma, c.reading, c.pos, count(*) as 粗頻度
from corpus as c with (nolock)
inner join corpus as c2 with(nolock) on c.[close]=c2.[open] and c.[file]=c2.[file]
where c.pos like N'名詞%' and c.wType like N'漢' and len(c.lemma)=2
and c2.lemma in (N'為る', N'出来る')
and c.corpusName like N'太陽 c'
```

・ キー条件：［品詞］が“名詞”かつ［語種］が“漢語”かつ［語彙素］の文字数が２文字
・ 後文脈条件：［語彙素］が「為る」または「出来る」

これにより，『太陽コーパス』『BCCWJ』のいずれかのコーパスにおいてサ変用法を持つと思われる二字漢語として約 11,813 語を抽出することができた。次に，この検索条件により抽出した語彙素について，サ変用法を含めた全出現例数を計測し，『太陽コーパス』『BCCWJ』の両コーパスにおいて「自立語 100 万語あたりの相対頻度で 10 例以上の用例が確保できるもの[3]」を，近代語・現代語比較用の語としてリストした。この条件は，本研究においてサ変用法の有無やサ変用法比率等の分析に耐える用例を確保するために設けたものである。

更に，リスト語の抽出計測値においてサ変用法が極めて低頻度の語や複数品詞にまたがって用いられる語については，実際の用例を検討した上で，以下のものを分析の対象外として排除した。

・ 明らかに誤解析のもの
(2)　【もよう】す（催す）（「模様」;『太陽』1925, 著者表記なし「国語、字音仮名遣改定案」）
・ 複合語の構成要素となるもの，または，連体修飾を受けるもの
(3)　地方の富豪階級が替る替る立【候補】して、(『太陽』1925, 無腸公子「新長者議員の顔触」)
(4)　皆さんはどんな【対策】していますか？（『BCCWJ』特定目的・知恵袋 2005, Yahoo!知恵袋）
・ 副詞として機能しているもの
(5)　しかし、竹下は反逆したが、海部は【結局】しなかった。(『BCCWJ』図書館・書籍 2005, 岩見隆夫『角栄以後』)

その結果, 調査対象となる語彙素は 1,203 語に絞られた。このように調査対象と定めた，近代・現代のいずれかでサ変用法を持つ二字漢語名詞を，以後「サ変名詞」と呼ぶ。

３．調査結果と分析

３．１　サ変用法の有無

２節に示した調査方法により抽出したサ変名詞を，両コーパスでのサ変用法の有無によって整理すると以下の通りである。表 1 に語数を，表 2 に語例を示す。

表１　コーパス別に見た調査語のサ変用法有無

コーパス	サ変あり			サ変なし	
	語数	サ変用例数	全用例数	語数	全用例数
太陽	1,078	90,000	353,588	126	56,041
BCCWJ	1,139	1,020,918	5,271,122	64	283,080

表１，表２より，どちらかのコーパスでしかサ変用法が見られない語が，少なからず存在することが分かる。

このうち，『太陽コーパス』でのみサ変用法が見られる語について，『BCCWJ』での非サ変用法と共に例を示してみよう。

[3] この相対頻度は，太陽コーパスにおいては粗頻度で約 50 例，BCCWJ においては約 590 例に相当する。BCCWJ における相対頻度 10 の語には，例えば「生計」「好感」「特質」「忍耐」等があり，現代語において，どのようなジャンルの文章にも現れ得る一般的なレベルの語と言える。

表２　コーパス別サ変名詞例

	語数	語例 （サ変用法の相対頻度上位 20 語。括弧内の数値はサ変用法の粗頻度）
太陽のみ	64	構造(26), 一挙(18), 出来(11), 損害(8), 結局(7), 理想(7), **秩序**(6), **傾向**(6), 根底(5), 次第(5), 長寿(4), 因果(4), 生計(4), **運輸**(4), 周囲(4), 手段(4), 損益(3), 伝説(3), **服装**(3), **総裁**(3)
BCCWJ **のみ**	126	**電話**(2447), **機能**(1526), **遭難**(112), **妥当**(96), 当面(85), **冒険**(60), **哲学**(37), **工事**(36), 都合(34), **欲望**(24), 事故(19), 家事(16), 科学(16), 強盗(12), 競馬(11), 会計(9), 元気(7), **言動**(7), 思想(7), 人気(7) ＊ 太字は近世末期以降見られる漢語
共通	1,013	研究(1037,1866), 発達(960,1684), 従事(874,1649), **組織**(789,1190), 増加(1239,7200), 実行(893,3526), 輸入(554,1154), 進歩(477,467), **拡張**(459,541), 反対(610,2357), 主張(796,4686), **注意**(835,5433), **発見**(873,5928), **養成**(389,311), **希望**(538,2218), 維持(689,4061), **占領**(396,666), 観察(584,3005), 奨励(351,440), 増進(304,195) ＊ 粗頻度は(太陽, BCCWJ)

(6)　鐵煉瓦石、コンクリートの如き不燃質を以て【構造】したる建物も（『太陽』1895, 著者表記なし「工業」）

(7)　一般に生き物の【構造】は、知れば知るほど驚嘆すべき合目的性で（『BCCWJ』図書館書籍 1996, 山本健一『脳とこころ』）

(8)　其他の代議政國も十九世紀の中半以來概ね中央集權の主義に【傾向】せるの事實あるを認む（『太陽』1901, 加藤政之助「立法行政の調和（附現制度の改正）（承前）」）

(9)　住宅地価格は上昇率が高くなる【傾向】を示している。（『BCCWJ』特定目的・白書 1981, 国土庁『国土利用白書』）

(6)では「構造」は漢字の字義通り「構え造る」意で用いられているが，(7)では「造られた結果できた仕組み」を意味する。同様に，(8)では「傾向」がやはり字義通りの「かた向く」意で用いられているが，(9)は「かた向いている状態」を意味する。これらの「構造」「傾向」という語において現代語でサ変用法が見られなくなったのは，「構え造る」「かた向く」といった動作から，その結果に焦点が移行し定着したことで，元の動作性を持つ意味用法が駆逐されたものと考えられる。

『太陽コーパス』のみでサ変用法が見られる語の多くは，「構造…構え造る」に見る動詞の並立や「結局…局を結する」に見る目的語と動詞の組み合わせなど，二字漢語の構成要素となる漢字自体が動作性を持つ。大量の漢語が新たに流入し一般に多く用いられだした「漢語定着期」の近代においては，このような字面から動作性の意識できる語に「スル」を接続して簡単に動詞化するような用法が，多く行われていたものと思われる。

一方，『BCCWJ』でのみサ変用法が見られる語についても，同様に両コーパスでの用例を比較してみたい。

(10)　落葉は蘚苔と共に森林が營む所の水源涵養の【機能】をたすく、（『太陽』1901, 市島直治「落葉の効能」）

(11) 地域が解体し、親族のネットワークが【機能】しないところでは、(『BCCWJ』出版・書籍 2003, 中西正司・上野千鶴子『当事者主権』)

(12) 未だ遠い後のことであるにも拘らず、すぐ【當面】に差し迫つたことのやうによく重吉夫婦の問題となつた。（『太陽』1917, 加能作次郎「漁村賦」）

(13) しかし今日、地域福祉が【当面】している課題からみると、(『BCCWJ』図書館・書籍 1992, 真田是『地域福祉の原動力』)

(10)では「機能」は「働き」を意味するが，(11)では「働く」「作用する」意で用いられている。「機能」は，『日本国語大辞典第 2 版』によると明治中期以降訳語として広まった語であり，『太陽コーパス』においては原義の名詞用法のみが見られるが，定着する過程において原義の持つ動作性が焦点化され動詞用法が派生したものと考えられる。(12)では「当面」は「目の前」の意で用いられており，(13)では「直面する」意で用いられている。『日本国語大辞典第 2 版』によると，前者の意の「当面」は中世から見られる用法であり，後者の用法は明治末期以降に見られるものである。先に見た近代にのみ例の認められるサ変用法を持つ語と同様に，漢語構成要素の「当たる」「向き合う」と言った字義による動作性の焦点化から動詞用法が派生し，元の意味を駆逐して定着したものと思われる。

なお，上記では，一方のコーパスに用例が一例も見られなかったもののみを挙げた。『太陽コーパス』での出現度数 1 と『BCCWJ』での出現度数 1 では，元のコーパスサイズが異なるためその重みが全く異なるが，用法の有無を問題にする際に，出現度数 1 は無視できないためである。ただし，実際には『BCCWJ』のような大規模なコーパスにおいて，出現度数 1 はノイズとなる場合もある。今回の調査においても，『BCCWJ』において出現度数 1 や 2 の極めて低頻度の例については，非現代語の引用や，非現代語的文脈（史伝，歴史小説など）における用例，特殊な使用域（法律用語，文学性の高い表現など）におけるものが大半であり，これらは現代語においてサ変用法が廃れたものと判断して差し支えない。以下に，近代に見られたサ変用法が現代でほぼ失われた語とみなせる語例を示す。これらの語が持つ言語内在的な特徴は，先に見た『太陽コーパス』のみでサ変用法が見られた語と差がなく，動詞用法の衰退理由も同様のものであろう。

表３　サ変用法が廃れた二字漢語の例

複雑(31), 困難(28), 予算(24), 是非(21), 徒歩(21), 自信(16), 沙汰(14), 膨大(14), 固有(13), 教養(11), 不審(11), 悪口(9), 経歴(8), 奉行(8), 一目(7), 根拠(7), 企業(5), 通商(5), 伝統(5), 出身(4), 騒動(4), 昼食(4), 栄養(3), 現実(3), 規約(2), 疑惑(2), 集団(2), 反動(2)

* 括弧内数値は『太陽コーパス』のサ変用法粗頻度

３．２　サ変用法の比率

次に，調査対象とした語の全体の用例のうち，サ変用法がどの程度の比率を占めているか（以下「サ変率」とする），両コーパス間でその比率に差があるかを調査した。比率を求める必要があるため，どちらかのコーパスで出現度数が「0」となる語は，調査対象から除外した。

こうして求めたサ変率は，当該の漢語が動詞性の強い語なのか，名詞性（あるいは他の品詞性）の強い語なのかを計る指標となる可能性がある。以下に，『太陽コーパス』におけるサ変率上位 10 位，下位 10 位の語の各コーパスでの出現度数，100 万語あたりの相対頻

度，サ変率を例示する。

表４　コーパス別サ変率

語	太陽			BCCWJ		
	粗頻度	相対頻度	サ変率	粗頻度	相対頻度	サ変率
表明	75	14.9	98.68%	1495	25.4	68.67%
指摘	135	26.8	98.54%	5552	94.4	52.22%
無視	234	46.5	97.50%	3806	64.7	87.47%
除去	109	21.6	97.32%	684	11.6	41.56%
着目	55	10.9	96.49%	728	12.4	92.15%
発揮	438	87	96.05%	3497	59.4	88.49%
関連	126	25	95.45%	2975	50.6	26.09%
従事	874	173.6	95.41%	1649	28	62.63%
阻止	56	11.1	94.92%	769	13.1	71.14%
関与	53	10.5	94.64%	1283	21.8	60.01%
司令	1	0.2	0.18%	1	0	0.04%
費用	1	0.2	0.17%	1	0	0.01%
無理	1	0.2	0.16%	745	12.7	7.24%
総督	1	0.2	0.15%	1	0	0.16%
行政	1	0.2	0.12%	1	0	0.01%
現象	1	0.2	0.11%	35	0.6	0.61%
革命	1	0.2	0.10%	6	0.1	0.11%
結果	3	0.6	0.08%	25	0.4	0.08%
目的	1	0.2	0.04%	2	0	0.01%
必要	1	0.2	0.02%	7	0.1	0.01%

更に，サ変率によって「高頻度グループ（80%以上）」「中高頻度グループ（40%以上 80%未満）」「中頻度グループ（20%以上 40%未満）」「中低頻度グループ（5%以上 20%未満）」「程頻度グループ（5%未満）」に層別し，両コーパスにおける語の分布をクロス集計したものが表５，これを元に語を類別したものが，表６である。

表５　両コーパスのサ変率分布

太陽＼BCCWJ	80%以上	40%以上	20%以上	5%以上	5%未満	合計
80%以上	4	47	***8***	***1***	***0***	60
40%以上	7	156	139	***41***	***7***	350
20%以上	***0***	25	111	80	***14***	230
5%以上	***0***	***9***	43	90	87	229
5%未満	***0***	***1***	***7***	19	117	144
合計	11	238	308	231	225	1013

表５の合計値から，サ変用法の比率は相対的に近代で高いことが指摘できる。また，サ変用法を持つ漢語には，通時的にさほど変化せず動詞性の強い語（表６A），動作性の弱い語（同 B），どちらにも属さない語がある一方，近代から現代で動詞性が弱くなる（同 C），あるいは強くなる（同 D）といったように変化している語が存在することが分かる。

では，実際にどのような語に，どのような変化が見られるかを確認してみよう。

表６を見ると，近代から現代で動作性が下降するものは，「養成」に見られるように複合語構成要素（「教員養成」「養成所」など 50%が複合名詞用法）としての性質が強いことや，「携帯」に見られるように派生的意味用法（60%が「携帯電話」の略）の勢力が圧倒的に強いことなどに起因して，相対的にサ変用法の比率が低くなっているものである。

表６　サ変率による語の類別

サ変率	語例
A.動作性強 （50%以上)	表明, 無視, 着目, 発揮, 従事, 阻止, 関与, 遭遇, 到達, 明記, 明示, 付与, 熱中, 断言, 適合, 目撃, 断念, 否定, 計上, 接近, 躊躇, 掲載, 記入, 尊重, 排除, 付着, 獲得, 公表, 挿入, 着手, 通過, 留意, 消滅, 軽蔑, 実現, 起因, 発見, 推測, 記載, 期待, 提唱, 注目, 沸騰, 予期, 現存, 送付, 通用, 紹介, 提出, 断定, 連想, 感心, 一貫
B.動作性弱 （2%未満）	対策, 学問, 困難, 収入, 騒動, 信号, 免許, 競技, 統計, 総理, 展覧, 利益, 衝動, 保守, 懲役, 疑惑, 行為, 病気, 感覚, 収益, 電報, 規程, 客観, 直接, 栄養, 通商, 貿易, 宴会, 留守, 中立, 戦争, 出身, 信託, 殺人, 後継, 反動, 現在, 収支, 合戦, 決算, 潜水, 起源, 訴訟, 現実, 感想, 主観, 犯罪, 娯楽, 会議, 意思, 将来, 現行, 予備, 形式, 意志, 意見, 司令, 費用, 総督, 行政, 現象, 革命, 結果, 目的, 必要
C.動作性下降 近代（40%以上） 現代（10%未満）	養成, 攻撃, 増進, 記憶, 建設, 運転, 許可, 防止, 指導, 執行, 対照, 矯正, 声明, 開発, 勧告, 集合, 合併, 論議, 還付, 思考, 総合, 覚醒, 操縦, 乱用, 連続, 捜索, 携帯, 連結, 冷却, 出願, 啓発, 表彰, 償却, 虐待, 投資, 歩行, 担任, 会談, 加盟, 斡旋, 給与, 企画, 整備, 宿泊, 廃棄, 同伴, 公認, 配列, 応答
D.動作性上昇 近代（20%未満） 現代（40%以上）	油断, 考案, 応援, 即位, 発動, 由来, 登場, 参戦, 追加, 所属

一方，動作性が上昇するものは，「油断」のように，現代においても複合語構成要素としての造語力が高くない語において，現代では「油断できない」のように「スル」「デキル」と専ら連接するところを，近代で「油断がならない」「油断なし」「油断ならず」のように「スル」以外の語と連接するバリエーションがあることや，「発動」のように，固定した言い回し（37%が「○○の発動」）や雑誌『太陽』の特集に起因する特定語（35%が「発動機」）が多いことなどに起因して，サ変率が相対的に低くなっているものである。

このように，近代から現代へと，サ変率に変化のある語については，語の造語力，別義の派生による使用域の広がりや語義の限定，コーパスの性質の差（サンプルコーパスか全文コーパスか）による用法のばらつきに変化要因を求められる可能性が高く，サ変率を単純に動作性の強さを計る指標として用いることは困難であることが分かった。

３．３　近代におけるサ変用法比率の変化

次に，『太陽コーパス』と『BCCWJ』とでサ変率に大きな現象が見られるものについて，『太陽コーパス』の内部で変化が起きているかを確認するため，太陽コーパス全体で 50 例以上のサ変用法があり，かつ，『太陽』の出版年による 5 カ年の層別（1895，1901，1909，1917，1925）で，出現度数 0 になる年がない語から 12 語を対象として，サ変率の経年変化を見た（表７，図１，図２，図３）。

その結果，図１のように漸次的にサ変化率が減少するもの，図２のように大きく減少しないもの，図３のように年によるばらつきが大きいものと，複数のパターンが認められた。

このうち，図１に示した漸次的にサ変化率が減る語については，使用頻度においても年を追って極めて低頻度になっている（表１）。これらの語は，現代でもサ変用法がほぼ意識されない語であり，近代語において既にサ変用法の衰退が始まっていた語群と位置づけられる。一方で図２に示したサ変率の下降が見られない語は，やはり現代でサ変用法が意識されないものであるが，これらは近代においては保持されていたサ変用法が，現代に至る時代の流れの中で衰退した語群と考えられる。また，図３に示した年によるサ変率の変動が大きい物は，現在でもサ変用法が存在する語が多く，サ変率の変化は，３．２節で見た他用法との分布により相対的に変動しているものと位置づけられる。

表7『太陽』におけるサ変用法の変遷（粗頻度）

語	1895	1901	1909	1917	1925	合計
住居	31	21	5	4	3	64
協同	11	23	9	12	2	57
施設	13	18	8	14	1	54
同盟	64	15	10	6	2	97
携帯	34	11	6	1	4	56
合同	14	26	27	10	70	147
総合	6	10	14	19	14	63
会合	18	14	16	6	2	56
装置	13	15	5	9	11	53
一言	49	60	36	39	17	201
適当	20	20	16	10	11	77
原因	10	27	38	12	36	123

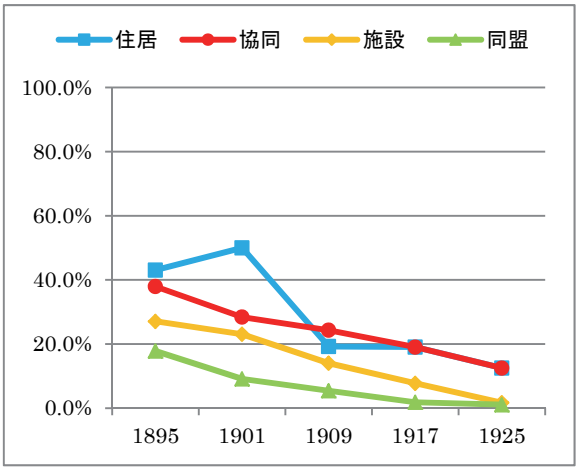


図1　サ変率の変化A

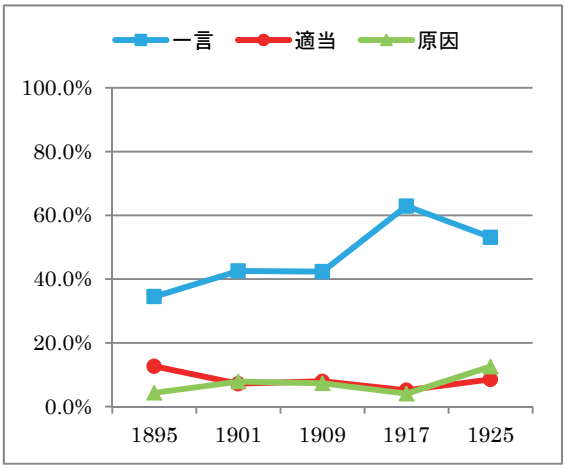


図2　サ変率の変化B

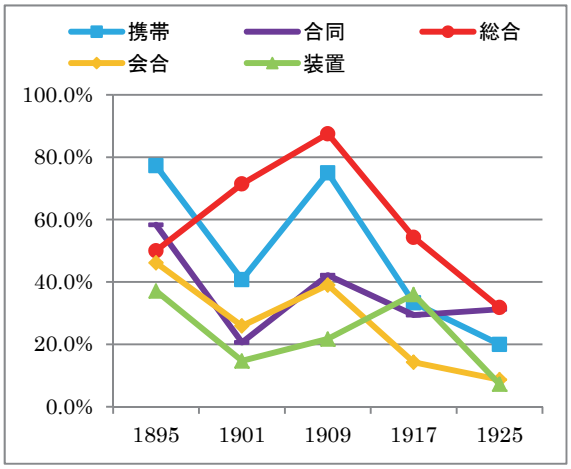


図3　サ変率の変化C

４．考察：サ変用法の有無やサ変用法比率の変化は何を表しているか？

上記調査により以下の結果を得た。

- 近代語と現代語の間で，サ変用法の有無に差のある語が存在する。これは，時代によって品詞性が変化したものと言える。
- 変化の方向性は，サ変用法が衰退・消失するものと，新たに獲得するものの両方が見

られる。

- サ変用法を持つ漢語について，当該漢語全体の用例中のサ変用法比率によって，動詞性の強い語か弱い語かに分類した結果，一部にサ変用法比率の大きな変動が見られた。その要因は，個々の語によって動作性の強さ以外の可能性が絡むものもあり，必ずしも漢語サ変名詞の動作性が現代において弱まっているとは言いがたい。
- サ変用法が近代から現代にかけて大きく減少している語について，『太陽コーパス』の内部で発行年による層別をした上で比率の変化を追うと，既に近代で衰退傾向が見られるもの，近代では保持されているがその後衰退したと思われるもの，用法の衰退とは異なる要因により変化するものがあった。

サ変用法の衰退・消失原因は，漢語定着期において語構成漢字の字義から得られる直接的な動作性のある語義から，動作の結果や状態を表す派生的意味に勢力を奪われたためだと思われる。一方，サ変用法の獲得は，訳語として出現・定着した漢語が，語義の持つ動作性から動詞用法を派生させたり，漢語の語構成パターンからの推論的な語の分解・再構築によって動作性が意識されたりすることによるものと考えられる。

調査対象とした二字漢語名詞は，個別にも，また全体的にも，近代と現代とでサ変動詞として用いられる比率に差がある。現代は近代に対してサ変用法の比率が低い。これらは，一見，サ変用法の衰退のようにも見えるが，サ変動詞以外の用法を観察すると，意味の多様化による名詞用法や形容詞・副詞用法の増加，複合名詞の増加など，語の定着に伴う用法の広がり，バリエーションの増加と見るべきであろう。

５．まとめ

本発表では，『BCCWJ』と『太陽コーパス』の形態論情報付与データを用いて，サ変用法を持つ二字漢語名詞の抽出を試み，以下の調査報告を行った。

- コーパス別に見るサ変用法の有無とその差異
- 全用法中のサ変用法の比率からみた語の分類
- 近代におけるサ変用法比率の変遷

これらの調査から，両コーパスでのサ変用法の使用状況には差があり，現代語では近代語に比してサ変用法が大きく減少していることが分かった。この減少は，サ変用法の単純な衰退ではなく，定着期の漢語が次第にバリエーション（用法や使用域）を増やして，日本語の語彙として馴染み確立されていったことを示していると考えられる。

なお，今回，手法や時間的な制約によって残された問題点のいくつかを以下に示す。

《名詞以外の品詞が割り当てられる二字漢語の品詞性変化》

今回の調査では，データベースからの対象語抽出の際に，形態素解析辞書 Unidic の大分類で「名詞」に相当するもののみをターゲットとした。しかし，二字漢語がサ変用法を持つものには，以下のような「名詞」以外の品詞が割当てられる語も存在する。今後は，これらの語も対象として，品詞性の変化を検討すべきである。

◆ 形状詞のサ変用法

(14)　租税制度として所謂體系論者の唱ふる様に組織が【完全】して居ない（『太陽』1925, 記者「財界時事小話　税制整理と日銀利下問題」）

◆ 副詞のサ変用法

(15)　世上の一部分にも漢學を廢止せんとする者少なからぬは、【畢竟】するに學ぶに困難なればぜ。(『太陽』1901, 大町桂月「教育時評」)

《サ変用法以外の品詞性変化》

今回の調査では，サ変用法の有無や比率の変遷のみを扱ったが，従来指摘・整理されてきた品詞性の変化には，以下のように名詞⇔形状詞・副詞間の変化などもある。

◆　一般名詞の形状詞用法

(16)　然るに吾が地球に於ては圜塊の表皮が既に【固形】な状態を取り、(『太陽』1909, 鶴田賢次「普通講話　宇宙開闔論」)

◆　一般名詞の副詞用法

(17)　若し構成法にも新聞の様な改正が【眞實】企られつつあらば、(『太陽』1901, 岡田三面子「法律時評」)

◆　形状詞の名詞用法

(18)　蓋し投機業者にして【豊富】の資本を有する時は、(『太陽』1901, 水島鉄也 ; 佐野善作「商業世界」)

１節で示した“実例の用法と情報付けされる品詞との間に生じる乖離の問題”を検討するためには，これらの調査・整理も欠かせない。今後の課題としたい。

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」（プロジェクトリーダー：田中牧郎）による成果の一部です。

参考文献

池上禎造(1953)「近代日本語と漢語語彙」金田一博士古稀記念論文集刊行会編『民族論叢：金田一博士古稀記念言語』三省堂

池上禎造(1954)「漢語の品詞性」京都大学国文学会『国語国文』23-11 三省堂、pp.92-101

池上禎造(1984)『漢語研究の構想』岩波書店

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

小木曽智信・中村壮範(2011)「『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装改訂版」（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（JC-U-10-01））

小木曽智信(2012)「旧仮名遣いの口語文を対象とした形態素解析辞書」『じんもんこん 2012 論文集』2012(7) 、pp.25-32

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

鈴木丹士郎(1998)「明治期漢語の品詞性と語形についての一考察」東京大学国語研究室創設百周年記念国語研究論集編集委員会編『東京大学国語研究室創立百周年記念国語研究論集』汲古書院、pp.728-750

鈴木日出男(2005)「明治時代以後の日本語　語彙・文体」近藤康弘・月本雅幸・杉浦克己編『新訂 日本語の歴史』放送大学教育振興会、pp.180-193

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

永澤済(2010)「変化パターンからみる近現代漢語の品詞用法」東京大学文学部言語学研究室『東京大学言語学論集』30、pp.115-168

BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス

浅原 正幸 (国立国語研究所) *
杉 真緒 (国立国語研究所・津田塾大学)
柳野 祥子 (国立国語研究所・津田塾大学)

BCCWJ-SUMM: A Summarization Corpus of the 'Balanced Corpus of Contemporary Written Japanese'

Masayuki Asahara (NINJAL)
Mao Sugi (NINJAL, Tsuda College)
Shoko Yanagino (NINJAL, Tsuda College)

要旨

『現代日本語書き言葉均衡コーパス』を元にした要約文書コーパスの設計について報告する。要約文書作成においては、クラウドソーシングを用いて 1 文書に対して 100 件規模で要約文書を収集する方法と、実験室において 1 人の被験者に複数回要約文書作成を依頼する方法の 2 通りを試行する。さらに作成した要約データに対する人手による主観評価情報を付与する。本稿では現在の進捗を報告するとともに今後の課題について示す。

1. はじめに

人間の文書理解過程は多様である。背景知識が異なる書き手と読み手との間には認知に乖離があり、何を伝えたいのかと何を読み取りたいのかとが必ずしも一致するとは限らない。また複数人の読み手が 1 つのテキストに対して何を重要視するかについても必ずしも一致するとは限らない。さらに 1 人の読み手の認知についても時間や回数の経過とともに変わってくるだろう。

本稿では『現代日本語書き言葉均衡コーパス』(以下 BCCWJ; Maekawa et al. (2014)) を元文書とした要約文書コーパスの設計について報告する。要約文書コーパスの分析を通して文書理解過程の多様性をとらえることを第一義的な目的とする。コーパスのその他の用途として、成人母語話者の作文能力の評価データや単一文書自動要約のためのベンチマークデータを想定している。収集した要約文書コーパスには要約文の優劣を評価し、人手による主観評価情報を付与する。5 種類の評価指針を立て、作業者 2 人により 5 段階の主観評価を行う。

以下 2 節では要約文の収集方法について述べる。3 節では収集した要約文に対する主観評価情報の付与について議論する。4 節ではまとめと今後の予定について述べる。

* masayu-a@ninjal.ac.jp

2. 要約文の収集

要約文の元文書として BCCWJ の新聞 (PN) サンプル (アノテーション優先順位 A) を用いる。BCCWJ の PN 可変長サンプルは複数記事からなるものもあり、これらについては記事単位に分割して元文書データを 19 文書作成した。

クラウドソーシングにより安価で大量にデータを得る手法 (タイプ入力:BCCWJ-SUMM_C) と実験室にて被験者に 3 回繰り返し要約作成課題を依頼してデータを得る手法 (筆述:BCCWJ-SUMM_L) の 2 種類の方法を用いた。表 1 に収集した要約文の概要について示す。

表 1　収集した要約文の概要

言語資源名	収集場所	生成過程	繰り返し	取得人数	摘要
BCCWJ-SUMM_C	クラウドソーシング	タイプ入力	なし	100-200	19 文書の要約
BCCWJ-SUMM_L	実験室	筆述	3 回	のべ 47	8 文書の要約

以下各言語資源について解説する。

2.1 BCCWJ-SUMM_C

BCCWJ-SUMM_C は BCCWJ の新聞記事の要約を Yahoo! クラウドソーシング (15 歳以上の男女) により被験者実験を行い作成したものである。

40 文字毎に改行した元文書を画像として提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。実験協力者は元文書をコピーして作業することができないために、画像を見ながらタイプ入力を行う必要がある。実験協力者の環境は PC 環境に限定した。元文書毎に約 100〜200 人の実験協力者が要約に従事した。実験実施時期は 2014 年 9 月である。

得られたデータ 19 文書の統計を表 2 に示す。収集要約数はクラウドソーシングで得られたファイルの総数である。得られたデータには、文字数制限を守っていないもの・実験の趣旨を理解していないもの・既に実験を行った実験協力者から同一回答を提供されたと考えられるものなどが含まれており、これらを排除したものを有効要約とした。

2.2 BCCWJ-SUMM_L

BCCWJ-SUMM_L は BCCWJ の新聞記事の要約を実験室環境で筆述により作成したものである。BCCWJ-SUMM_C で用いた元文書を印刷紙面で提供し、実験協力者に 50-100 文字に要約せよという指示で収集した。1 つの元文書に対して、3 回まで繰り返して要約文作成を行った。繰り返しに際しては、特別に「前と同じ要約文を作成してください」などといった指示は行わず、質問された場合にも「自由に要約文を作成してください」と教示した。被験者実験は強制ではなく被験者が拒否した時点で実験を終了するため、3 回繰り返していない事例も含めた。実験協力者は原稿用紙上で筆述 (鉛筆と消しゴム利用) で要約を行い、そのデータを電子化した。

現在のところデータは 8 文書のべ 61 人分に限定した。得られたデータの概要は表 3 のとお

表 2　BCCWJ-SUMM_C データ概要

FileID	有効要約数	収集要約数
A_01	106	198
A_02	112	195
B_02	98	149
B_03	74	100
C_01	63	100
C_02	63	99
C_03	53	100
D_01	55	100
D_02	55	100
D_03	48	99
D_05	55	99
E_01	58	99
E_02	46	98
E_03	54	100
E_04	60	99
E_05	48	100
E_06	56	98
F_01	57	100
F_02	58	100

表 3　BCCWJ-SUMM_L データ概要

FileID	有効要約数	被験者数
A_01	19	7
A_02	18	6
B_02	21	7
B_03	27	9
C_01	27	9
C_02	21	7
C_03	18	6
Q	30	10

り。本実験の実験参加者からは要約作業前に要約元文書の読み時間 (視線走査法もしくは自己ペース読文法) のデータも取得した。さらに被験者の特性 (最終学歴・語彙数・言語形成地・記憶力) などのデータについても収集した。実験実施時期は 2014 年 8 月～2015 年 1 月であるが、今後このデータは引き続き拡充していく予定である。

3.　人手による要約の主観評価

収集した要約文に対して、主に読みやすさに関して人手による要約の主観評価を付与する。

人手による要約の主観評価として DUC-2005 [(1)]で用いられた以下の 5 種類の評価指針を用いる：

- 文法性 (Grammaticality)： 誤字・文法的でない文が含まれていないか
- 非冗長性 (Non-redundancy)：全く同じ情報が繰り返されていないか
- 指示詞の明解さ (Referential clarity) : 先行詞のない指示詞（代名詞）が含まれていないか
- 焦点 (Focus)：要約全体と無関係な情報が含まれていないか
- 構造と一貫性 (Structure and Coherence)：接続詞を補ったり削除したりする必要のある箇所はないか

この 5 種類の評価指針について A (very good) - E (very poor) の評価を行う。現在主観評価付与作業は 2 人の作業者により行っている。基準の統制後、作業者を増やすことも検討する。DUC は対象言語が英語であるために、指針については DUC-2005 の quality question をそのま

[(1)] http://www-nlpir.nist.gov/projects/duc/duc2005/

ま用いず、作業者間で調整しながら基準を策定中である。現在までに得られている作業者メモから主観評価における細かい指針と論点について示す：

- 全体：
  特に問題がないものを A とし、作文として問題が軽度のものを B とする。C 以下は問題の程度に応じて付与する。
  C は欠陥が認められるがぎりぎり意味が通じる程度のものとし、程度や件数に応じて D 以下を付与する。
- 文法性 (Grammaticality)：
  問題のないものは A とする。誤字については「蓮舫」→「蓮坊」[(2)]のような単純なタイプミス、変換ミスは B とする。
  「法学部への進学し、」のような文法的な誤りが 1 件ある場合は C とし、1 件増えるごとに評価を 1 段階ずつ下げる。誤字の評価に加えて文法的でないものがあった場合、評価を 2 段階下げる。
  文法的なものについては、問題がないものには A、意味は通じるもの（読点の使い方や文のわかりやすさに改善点があるもの）には B を付与する。意味は通じるがわかりにくいもの（主語や目的語が省略されていてかつ意味が不明確なもの、コロケーションが不適切なもの）には C、日本語として不自然なもの（「たり」の使い方、助詞「の」の連続など）には D、明らかに文法的でないものには E を付与する。
  元文書にある誤用「レッテルを張る」についても漢字の誤用として評価を下げる判断を行った。
- 非冗長性 (Non-redundancy)：
  問題のないものは A とする。固有名詞や人を表す名詞（先生など）が重複しているような場合には B を付与し、普通名詞などの重複は C を付与する（喋る → しゃべりなど、品詞が変わっているものも含む）。表現の意味的な重複は D とする（才能 → 能力など）。冗長性が複数認められた場合は E とする。
  その他、言い換えられているが同じものを指す場合 C とする。
  現在のところ単語レベルの冗長性のみを検討しているが、句レベル・文レベルの基準についても事例が出現次第、随時検討する。
- 指示詞の明解さ (Referential clarity)：
  問題のないものは A とする。指すものが曖昧な場合、要約文を読むだけで曖昧性が解消できるものには B を付与し、推測はできるが書き手の指示するものが分かりにくいものには C を付与する。全く指示詞などの情報が示されていない、また明解でないものが複数ある場合、程度や件数に応じて D か E を付与する。
- 焦点 (Focus)：
  問題のないものは A を付与する。
  表現の仕方により、元文書の内容と違う読み方がされる可能性があるものは B か C を

(2) かな漢字変換ツールによっては変換が困難であるため。

付与する。要約におけるある部分要素（事例）にのみかかわる場合はBを付与し、要約全体の意味にかかわる場合はCを付与する。
要約作成者が元文書の内容理解に失敗している可能性があるものはCもしくはDを付与する。厳密には内容と合っていないものにはCを付与し、主体や語彙の意味などを取り違えているものはDを付与する。
元文書の要点とずれているものや、要約に不必要な情報が入っているものにはDを付与する。
内容と関係のない情報（原文に記述されていないことや書き手の意見）が入っているものにはEを付与する。

- 構造と一貫性 (Structure and Coherence)：
問題のないものにはAを付与する。
表記に一貫性のないものが高々1件の場合はBを付与し、複数あればCを付与する。具体的には漢字（ひらくかどうか）や呼称、記号の使用などを対象とする。
文章を通して、主語の交代が頻繁である場合はCを付与する。
接続詞の使用や、複文・重文の構成に改善点がある場合はDを付与する。具体的には接続詞の誤用、欠落など。またひとつの文を複数に切ったほうがよいものも対象とする。
文体に一貫性がないものにはD以下を付与する。具体的には語尾が一貫していないものなどを対象とする。

なお、細かい指針については今後修正される可能性がある。

表4　A_01サンプルに対する評価指標付与

	A	B	C	D	E	相関係数
文法性	9,5	7,3	3,8	3,7	1,0	0.72
非冗長性	21,9	2,5	0,4	0,5	0,0	0.07
指示詞	22,7	1,8	0,3	0,5	0,0	0.67
焦点	19,8	3,1	1,6	0,8	0,0	0.09
構造と一貫性	14,8	3,0	4,5	2,8	0,2	0.73

表4にBCCWJ-SUMM_CのA_01サンプルに対する評価指標付与結果を示す。元文書は付録A節に示す。表中カンマで区切られた2つの数字が、それぞれ2人の作業者が付与したA-Eの件数を表す。相関係数は2人の作業者の相関係数を表す。

「文法性」「指示詞」「構造と一貫性」の3つについては強い相関がみられたが、「非冗長性」と「焦点」の2つについては相関がみられなかった。表5に「文法」の、表6に「非冗長性」の、表7に「指示詞」の、表8に「焦点」の、表9に「構造と一貫性」の作業者間分割表を示す。「文法性」について対角線近くに分布しており作業者間で統制できていることがわかる。「非冗長性」・「指示詞」・「焦点」については基本的に厳しい作業者と厳しくない作業者との間に差が出ていると考える。「構造と一貫性」については評価が割れていることがうかがえる。作業者間の統制については今後検討していきたい。

表 5　文法性の作業者間分割表

	A	B	C	D	E	計
A	5	-	-	-	-	5
B	2	1	-	-	-	3
C	2	4	1	1	-	8
D	-	2	2	2	1	7
計	9	7	3	3	1	23

表 6　非冗長性の作業者間分割表

	A	B	計
A	8	1	9
B	5	-	5
C	4	-	4
D	4	1	5
計	21	2	23

表 7　指示詞の作業者間分割表

	A	B	計
A	7	-	7
B	8	-	8
C	3	-	3
D	4	1	5
計	22	2	23

表 8　焦点の作業者間分割表

	A	B	C	計
A	7	1	-	8
B	1	-	-	1
C	5	1	-	6
D	6	1	1	8
計	19	3	1	23

表 9　構造と一貫性の作業者間分割表

	A	B	C	D	計
A	4	3	1	-	8
B	-	-	-	-	-
C	3	-	2	-	5
D	7	-	-	1	8
E	-	-	1	1	2
計	14	3	4	2	23

最後に A_01 の評価事例について示す。

以下は評価が比較的高い例である：

A_01(No.18):

文法性 (A,A)・非冗長性 (A,B)・指示詞 (A,B)・焦点 (A,A)・構造と一貫性 (A,A)

> 蓮舫さんは幼いころから活発で、自分の意見をはっきり言うことができる人だった。池田弘子先生はそれを持ち前の長所だと考えて適切なアドバイスをし、蓮舫さんがキャスターになるきっかけを与えてくれた。

要約としてまとまっており、読みやすさも優れている。

以下は評価が文法性・構造と一貫性が比較的低く、指示詞・焦点の評価が一致していない例である：

A_01(No.23):

文法性 (D,C)・非冗長性 (B,A)・指示詞 (A,D)・焦点 (A,C)・構造と一貫性 (C,E)

> 蓮舫さんは思い出の先生についてこう語っている。おしゃべりだと言われていただけの自分を仕事に生かしてみたらと目を開かせてくれた。違う角度から相手の身になってくださる方だった。

以下に評価が低い理由についてのアノテータコメントを示す。

文法性：「自分を生かす」「目を開かせる」

指示詞：「仕事とは何か」、「何と違う角度からか」、「相手とはだれか」
焦点：「仕事に生かす（活かす）ことをアドバイスしたわけではない」
構造と一貫性：「『くれた』『くださる』一貫性がない」

文法性については 2 人の作業者ともに 2 文目の不自然さを指摘している。構造と一貫性については待遇表現についての指摘がある。焦点については 1 人の作業者が元文書において言及されていない点を含むことを問題視している。

以下は評価が文法性・焦点が低く、構造と一貫性の評価が一致していない例である：

A_01(No.31):
文法性 (C,D)・非冗長性 (A,A)・指示詞 (A,A)・焦点 (C,D)・構造と一貫性 (A,D)

蓮舫さんは、通っていた青山学院高等部では、ピアスをしたりしていたので、注意をする先生もいたが、二、三年時に担任だった池田弘子先生だけは、頭ごなしではなく、子どもの目線に立って聞く耳を持たせてくれた。

以下に評価が低い理由についてのアノテータコメントを示す。

文法性：「したりしていたので」「1 つの文の中で主語の違う節が多すぎる」
焦点：「先生と蓮舫さんのつながりが表わされていない」
構造と一貫性：「文を切るべき」

構造と一貫性については 1 人の作業者により 1 文中の節の多さが指摘されている。

4. おわりに

本稿では『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパスの設計について議論した。要約元文書として BCCWJ のコアデータの PN サンプルを用い、クラウドソーシングと実験室においての被験者実験により、複数人・複数回の要約作文を収集した。収集した要約作文に対して人手による主観評価を進めている。少量ではあるが、現在までに作成した主観評価結果について検討した。

引き続きデータを拡充するとともに人手による指標付与の相関の向上に努めたい。さらに複数人間・複数回間の評価の揺れを被験者属性を含めて分析することで、最終目標である文書理解過程の多様性の定量評価を行いたい。

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」、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。

付録 A.　要約元文書 A_01 サンプル

以下に要約元文書 A_01 サンプル (PN1c_00001) を示す：

ＡＬＢＵＭ　　私の先生
キャスター　蓮舫さん
「おしゃべり」才能後押し
東京都生まれ。
９５―９７年、中国・北京大に留学し、帰国後に双子を出産。
子育てのかたわらテレビ、ラジオなどで活躍中。
３３歳。
　幼稚園から大学まで通った青山学院では、とにかく活発で、目立つ生徒だったという。
高等部では自由な校風もあって、流行に乗ってかばんを薄くつぶしたり、ピアスをしたり。
呼び出して注意する先生もいたが、二、三年時に担任だった池田弘子先生（７５）は違った。
「そんな薄いかばんじゃ遊び道具も入らないよ」「体育や部活では、危ないからピアスをはずしたほうがいい」。
やんわり語りかける。
「頭ごなしでなく、子どもの目線に立って、聞く耳を持たせてくれるんですよね」
　保健の担当でスクールカウンセラーでもあった先生の授業は、型破りだった。
障害や難病に苦しむ人の話をよく取り上げ、生徒同士で討論させた。
「世の中には様々な人がいるということが、よくわかった。
ホスピスという言葉を初めて聞いたのもこの授業でした」
　台湾人の父を持ち、「家で自己主張するよう教えられていた」蓮舫さんは、いつも率先して自分の意見を言った。
「どこかみんなとは違っていたのかもしれない」。
ほかの先生たちには、「おしゃべり」のレッテルを張られていた。
　それなのに、池田先生は言ってくれたのだ。
「しゃべるのが得意なんだから、能力を生かしてみたら」と、初めて「おしゃべり」を評価してくれた。
ブラウン管の中で話すなんて、思ってもみないころだった。
　大学に進学する時も、「あなたは論理的に考えるのが得意」と、法学部に行くよう促したのは池田先生。
大学在学中にデビューし、キャスターとして活躍するその後の進路を思うにつけ、「本当によく見ていてくれた」と感謝する。
　池田先生も、蓮舫さんにアドバイスしたことを覚えていた。
「生意気という人もいたけれど、私は、彼女のようにモノをはっきり言えることがこれからは大切だと思っていました」。
ひときわ元気だった教え子に、「持ち前の才能を生かしていってほしい」とエールを送る。

参考文献

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.

上級～超級日本語学習者の作文から見た言語産出実態

趙　海城（明星大学　人文学部）

Language Production Reflected in the Composition by Advanced and Supper-Advanced Japanese Language Learner

ZHAO Haicheng　(Meisei University)

要旨：

本稿は「YNU 書き言葉コーパス」を使い、中韓の留学生と日本人学生の作文を分析し、上級～超級日本語学習者の言語産出実態、日本人との違いの一端を明らかにすることを目的とする。分析した結果、留学生の作文には挨拶語、人称代名詞、指示代名詞、連体詞「此の、其の」、助動詞「たい」が過剰使用され、様態・推量助動詞、「てる」の過少使用が見られる。中韓留学生間にも、「此の、其の、此れ、其れ」の産出数、受身・尊敬、使役助動詞の産出数の違いが見られる。留学生はレベルが上がるにつれ、作文語数が増え、品詞の大半は異なり語数、延べ語数が増え、相手との交渉表現、文構造が複雑化する。また留学生は感動詞の産出が減り、「僕・俺」が「私」に取って代わられ、形状詞・助動詞・終助詞が増加するなど、日本人の使用実態に近づいている。ただし、人称代名詞の多用、様態・推量助動詞の過少使用など、超級になっても日本人学生の産出と違う特徴が見られる。

１．はじめに

本稿は「YNU 書き言葉コーパス(以下「YNU コーパス」と略称する)」を用いて、中韓両国からの留学生と同年代の日本人学生(学部生、大学院生、研究生を一括して学生と呼ぶ)の課題作文の品詞別産出状況を考察するものである。分析する過程で、日本人学生の産出状況を参照基準値とするが、むろん日本人学生でも日本語習熟度の面においては、個体差もあり、社会人、言語熟練者と比べれば、未熟さが残る者もある(小野他 2007)ため、この参照基準値は留学生が目指す絶対基準ではないことを断っておく。

山内(2009)は、日本語学習者の OPI データ(KY コーパス)から、学習者のレベル別言語特徴を明らかにした。具体的には「レベル判定に寄与する形態素を探す」ことを試みた結果、「だ(助動詞)」、「よ(終助詞)」、「から(接続助詞)」、「やっぱり」、「と思います」は「上級以上であることを決定する形態素」とし、「こう(フィラー)」、「けれども(接続助詞)」、「っていう(複合助詞)」、「んですけ(ど)」は「超級であることを決定する形態素」としている。橋本(2011)は、山内の研究に習い、KY コーパスの上級話者 12 人、超級話者 15 分の発話データから「抽象的関係」を表す名詞を抽出して考察した。その結果、「面、風、辺、自身、状況、互い、逆」などが超級を示す実質的形態素であること、「そういう面」「そういうふう」「その辺」のように、機能形態素の超級マーカーと実質形態素の超級マーカーが密接に関連していることを明らかにした。毛(2013)は、中国日本語学習者コーパス(CJLC、4 級作文 1200 篇、8 級作文 1200 篇からなるもの)を母国語話者コーパス(会話、小説、論説からなるもの)と比較し、中国日本語学習者の高頻度語産出の特徴を考察した。その結果、中国語日本語学習者は名詞、形容詞、形容動詞、副詞などの自立語及び複合辞の産出が多く、カバー率も高く、過剰使用する傾向にあるが、接続詞、格助詞、係助詞に関しては、産出量の面においては母

語話者コーパスと多く違わないが、学習者コーパスのほうはカバー率が高いことが分かった。また、学習者が推量助動詞「そうだ、みたいだ、らしい」、準体助詞「の」、終助詞の産出が少ないことが分かった。これらのことから、学習者は文の基本的な成分を構成する単語を多用するが、ムードを表す語彙の使用が少なく、表現が単調であると指摘している。

山内(2009)、橋本(2011)の考察対象はインタビューによる学習者の発話データで、書き言葉ではない。毛(2013)は学習者の作文データを対象に分析しているが、作文は中国教育部高等学校外語専業教学指導委员会日語分委员会の主催で実施された 2007 年～2009 年度試験作文の一部であり、比較対象とする母語話者コーパスもやや古い小説、論説文、会話文となり、文体、話題等も違うため、それをもって、外国語学習環境にある中国語日本語学習者が日本語を過剰使用、過少使用と言っても統一性がないように考えられる。そこで、本稿では、第二言語学習環境にある中韓両国からの上級～超級留学生と同年代の日本人学生の同じテーマの課題作文を比較することにした。

２．調査データと調査概要

YNU コーパスは日本人学生 30 名と、日本国内にいる中国人・韓国人留学生 30 名[1]ずつに対し、状況や難易度の異なる 12 種類の作文タスクを課し、各国の学生よりそれぞれ 360 編、三カ国合わせて計 1080 編の作文データを収集したものである。12 種類のタスクは、手紙、PC メール、携帯メール、投書、レポートなどのスタイルのものとなるように配慮され、また自発型か頼まれ型か、読み手は特定の相手か不特定の相手か、読み手は特定の相手の場合、目上なのかそれとも同僚・友人なのかに分かれるように設定された。さらに、中韓両国の留学生が書かれた作文は独自の評価基準(タスクの達成、タスクの詳細さ・正確さ、読み手配慮、体裁・文体の四項目)で評価され、その達成度に応じて、下位群(10 名)、中位群(10 名)、上位群(10 名)という三つのグループに分けられる。本稿もこのグループ分けに基づく。

YNU コーパスにはオリジナルデータとオリジナルデータを補正した補正データ[2]があるが、本稿では形態素解析の利便性を考え、補正データを分析対象とした。「茶まめ」を用いて、YNU コーパスの補正データを対象に、形態素解析処理を行った。形態素解析器は「MeCab 0.996」、解析用辞書は「UniDic-mecab 2.1.2」を使用している。基本的に形態素解析して得た解析結果を使うが、「形状詞-助動詞語幹 : (そうだ(様態)、ようだ、みたいだ)」は従来「らしい」などと同じく助動詞とみなされることが多いため、助動詞に分類し直した。

３．調査結果

３．１全体の傾向

表 1 は YNU コーパスにおける延べ語数、異なり語数、文の数、【　】の数を示したものであり、表 2 はこれらの項目の 10 万語あたりの調整頻度を示すものである。

---

[1] 両国の留学生は日本の大学、大学院で講義を受けられるレベルで、一般的に言えば、上級レベル及びそれ以上のものである。旧日本語能力試験、2010 に改定された新日本語能力試験の受験結果を見ると、韓国人留学生の内訳は 1 級、N1 合わせて 19 名、2 級、N2 合わせて 3 名、未記入 8 名で、中国人留学生は 1 級、N1 合わせて 26 名、2 級、N2 合わせて 3 名、未記入 1 名である。本稿ではこれに基づいて上級～超級日本語学習者とした。

[2] 補正の主なポイントとしては、一文一行とし、不要な改行・空欄を削除する；誤漢字と送り仮名は適宜修正する；すべて平仮名書きで読みにくいものは漢字に変換して修正するというものだった(金澤 2014：16)。

表１　YNU コーパスの延べ語数・異なり語数・文数・読点数・【　】数(産出実数)

	中国				韓国				日本
	下位群	中位群	上位群	**総計**	下位群	中位群	上位群	**総計**	**総計**
延べ語数	26176	30818	31892	**88886**	21833	28288	31344	**81465**	**79337**
異なり語数	2090	2451	2453	**4024**	1521	2109	2260	**3358**	**3633**
文(句点)の数	1230	1492	1339	**4061**	1048	1242	1278	**3568**	**3165**
読点の数	1384	1723	1500	**4607**	668	1136	1326	**3130**	**3724**
【 】の数	112	120	115	**347**	87	85	115	**287**	**330**

表１の産出実数の延べ語数で言うと、中国人留学生＞韓国人留学生＞日本人学生となっている。中韓の留学生と比べれば、日本人学生はより少ない語数でタスクを達成させていることが分かる。産出実数の異なり語数(厳密には異なり形態素数)で見ると、中国人留学生の作文全体が長い分、異なり語数ももっとも多い。それに対し、韓国人留学生の作文の延べ語数は日本人学生より多いにもかかわらず、異なり語数は日本人学生より少ない。

表２　YNU コーパスの延べ語数・異なり語数・文数・読点数・【　】数（調整頻度)

	中国				韓国				日本
	下位群	中位群	上位群	**総計**	下位群	中位群	上位群	**総計**	**総計**
延べ語数	100000	100000	100000	**100000**	100000	100000	100000	**100000**	**100000**
異なり語数	7984	7953	7692	**4527**	6967	7455	7210	**4122**	**4579**
文(句点)の数	4699	4841	4199	**4569**	4800	4391	4077	**4380**	**3989**
読点の数	5287	5591	4703	**5183**	3060	4016	4230	**3842**	**4694**
【 】の数	428	389	361	**390**	398	300	367	**352**	**416**

表２の異なり語数の 10 万語あたりの調整頻度を見ると、日本人学生＞中国人留学生＞韓国人留学生の順で、日本人学生の異なり語数がもっとも多くなる。日本人学生のほうは語彙量が豊富であると予測できるため、短い作文の中でより多くの種類の語彙を産出していることが分かる。それに対し、韓国人留学生は異なり語数が少なく、同じ語が繰り返し使用されていることが示されている。

また、表１、表２から、中国人留学生は句読点をたくさん打っており、韓国人留学生は中国人留学生・日本人学生と比べれば、句点を打つわりに、読点をさほどたくさん打っていないことが分かる。日本人学生は中国人留学生ほど句読点をたくさん打っていないが、読点が句点より多いという状況は両者が似ている。表 2 の 10 万語あたりの調整頻度を見ると、日本人学生の作文には句点が一番少なく、言い換えれば文が長いことがうかがえる。文が長くなるということは、連体修飾表現をたくさん使うなど、文の構造が複雑になり、より難易度の高い文を産出していることが予測される。また、句読点の出現数については、それぞれの母語における句読点の重要さの違い、思考過程においてつい打ってしまうということも関わっている可能性がある。

【　】は「氏名」「住所」「電話番号」「メールアドレス」といった個人情報が入っている部分である。調整頻度で見ると、日本人学生(416)＞中国人留学生(390)＞韓国人留学生(352)の順となっており、日本人学生がタスクを達成させるために、一番よく個人情報を開示していることが分かる。

表１に示されたデータには形態素解析辞書「UniDic-mecab 2.1.2」の品詞分類による「記号(一般、文字)」「空白」「補助記号(一般、句点、読点、括弧等)」が含まれるが、品詞別の産出状況を分析するにあたり、これらのものを削除した。

留学生の産出された作文と日本人学生の作文との難易度を測るため、語彙密度(語彙のバラエティ)を分析する。語彙密度は文章の難易度や内容の豊富性を示す指標である。語彙密度を測定する指標として TTR、R 値が使用されることが多く、TTR、R 値が高いほど文章がバラエティに富むと言える。TTR は異なり語数(Type)を延べ語数(Token)で割る（Type/Token Ratio）ものである。R 値(Guiraud 値)は Type を Token の平方根で割った値で、データ間のサイズに差がある場合にも安定的に語彙密度を測定し、比較できると言われる(石川：2012)。表３に YNU コーパス(記号・補助記号・空白削除後)の延べ語数、異なり語数、語彙密度(TTR,R 値)を示している。

表３　YNU コーパス(記号・補助記号・空白削除)の延べ語数・異なり語数、語彙密度(TTR,R)

	中国				韓国				日本
	下位群	中位群	上位群	**総計**	下位群	中位群	上位群	**総計**	**総計**
延べ語数	22140	26339	27732	**76211**	19065	24677	27365	**71107**	**68107**
異なり語数	2026	2394	2400	**3939**	1455	2040	2197	**3249**	**3555**
ＴＴＲ	0.092	0.091	0.087	**0.052**	0.076	0.083	0.080	**0.046**	**0.052**
Ｒ値	13.62	14.75	14.41	**14.27**	10.54	12.99	13.28	**12.18**	**13.62**

表３から、中韓両国留学生の作文はともに、レベルが上がるにつれ、産出作文の異なり語数と延べ語数が増え、作文が長くなっていることが読み取れる。語彙密度(TTR、R 値)を見ると、中国人留学生の書かれた作文はどのレベルにおいても韓国人留学生より高く、より多様な語彙が使われていることが分かる。このことは注１に示したように、調査対象者の韓国人留学生に比べ、中国人留学生のほうが１級、N1 の合格者が多く（韓国人留学生は未記入が８人）、レベル的に相対的に高いことによる可能性がある。ただし、タスク完成に向け、取り組み態度といった外部要素も関わってくるため、ここでは中国人留学生のほうは語彙量が豊富と断言できない。また、Ｒ値を見ると、中国人留学生の中位群が一番高い。一方、中韓の留学生と比べれば、日本人学生は相対的に短い作文でタスクを完成させている。

表４　品詞別産出数、品詞構成比(中国人留学生のを降順基準に)

	中国		韓国		日本	
名詞	23501	*30.84%*	20852	*29.32%*	20498	*30.10%*
助詞	23033	*30.22%*	21695	*30.51%*	21010	*30.85%*
動詞	10984	*14.41%*	10312	*14.50%*	10146	*14.90%*
助動詞	8798	*11.54%*	8823	*12.41%*	8370	*12.29%*
接尾辞	2300	*3.02%*	1915	*2.69%*	1758	*2.58%*
副詞	2062	*2.71%*	1740	*2.45%*	1555	*2.28%*
形容詞	1565	*2.05%*	1437	*2.02%*	1246	*1.83%*
代名詞	1062	*1.39%*	1226	*1.72%*	984	*1.44%*
接頭辞	896	*1.18%*	794	*1.12%*	873	*1.28%*
連体詞	757	*0.99%*	748	*1.05%*	568	*0.83%*
形状詞	619	*0.81%*	616	*0.87%*	640	*0.94%*
接続詞	371	*0.49%*	346	*0.49%*	295	*0.43%*
感動詞	168	*0.22%*	224	*0.32%*	117	*0.17%*
未知語	95	*0.12%*	379	*0.53%*	47	*0.07%*
総計	76211	*100.00%*	71107	*100.00%*	68107	*100.00%*

表４に示すように、日本人学生、中韓両国留学生が書かれた作文では、品詞別産出数を見ると、中国人留学生は名詞、助詞、動詞、接尾辞などの品詞９種類において多く、韓国人留学生は助動詞、代名詞、感動詞の３種類で最も多い。日本人学生は形状詞の産出数が中韓両国留学生より多いことが分かる。一方、品詞構成比に大きな開きは見られないが、中韓両国の留学生は接尾辞、副詞、形容詞、接続詞、感動詞の構成比が日本人学生より高く、日本人

学生は接頭辞、形状詞の構成比が高いことが見て取れる。なお、未知語とはアラビア数字、英語、中国語簡体字、韓国語の固有名詞の片仮名表記したものなどを指す。

表５　中韓両国留学生の作文レベル別の品詞出現数

	中国			韓国		
行ラベル	下位群	中位群	上位群	下位群	中位群	上位群
名詞	7088	8107	8306	5879	7542	8274
助詞	6610	7892	8531	5435	7036	8381
動詞	3063	3838	4083	2717	3632	3963
助動詞	2477	3119	3202	2401	3130	3292
接尾辞	716	760	824	453	681	781
副詞	605	673	784	481	632	627
形容詞	459	547	559	398	528	511
代名詞	292	366	404	399	430	397
接頭辞	233	326	337	168	267	359
連体詞	224	275	258	258	254	236
形状詞	181	205	233	183	212	221
接続詞	93	146	132	117	144	118
感動詞	64	55	49	97	117	132
未知語	35	30	30	79	72	73
総計	22140	26339	27732	19065	24677	27365

表５は中韓両国留学生の作文レベル別の品詞出現数を示したものである。両言語とも、レベルが上がるにつれ、名詞、助詞、動詞、助動詞、接尾辞、接頭辞、形状詞の産出が増えている。副詞、形容詞、代名詞に関して、中国人留学生の作文はレベルが上がるにつれ、産出数が増えているが、韓国人留学生の作文は下位群から中位群にかけては増えるが、上位群ではやや下がっている。連体詞については中国人留学生の中位群が一番多く、上位群になるとまた下がっており、韓国人留学生の作文は下位群でもっとも多く、レベルが上がるにつれ産出数が減っている。接続詞ついては、両言語とも中位群が一番産出しており、上位群は下がっている。感動詞については、韓国人留学生の作文はレベルが上がるにつれ、産出数が増えるのに対し、中国人留学生の作文はかえって産出数が減っている。中韓両国留学生の作文における語彙は、品詞により違う産出様態を呈していることが分かる。

３．２　品詞別の使用傾向

以下、特徴的な使用傾向を示す品詞について、品詞別に見る。産出した異なり語数が 20 語を超える場合、上位 20 語に限定して提示する。

下表６は感動詞の産出状況を示すものである。中韓両国留学生の作文とも、日本人学生より感動詞を多く産出している。詳細を見ると、70%前後は「今日は」「有り難う」「御早う」といった初級で習う挨拶語であった。特に、中国人留学生は「有り難う」を、韓国人留学生は「今日は」「御早う」「あの」をよく使っており、日本人学生は「はい」「うん」といった応答用の感動詞を留学生より多く産出していることが分かる[3]。

[3] 感動詞と解析されたものの中で、誤解析されたものが一部含まれる。例えば、「あの」は感動詞「あのさ」として使われるものがある一方で、「あの本」のような連体詞の誤解析が見られた。また、「あっ」は感動詞の「あ」以外に、「あっという間」が１例見られた。

表６　感動詞の産出状況

感動詞(37個	中国			中国集計	韓国			韓国集計	日本	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
今日は	22	25	20	**67**	35	37	27	**99**	**36**	**202**
有り難う	23	11	12	**46**	5	10	18	**33**	**15**	**94**
御早う			4	**4**	14	3	4	**21**	**7**	**32**
あの	3			**3**	10	6	3	**19**	**10**	**32**
はい	1	1	1	**3**			2	**2**	**10**	**15**
おー	1	7	1	**9**		1	1	**2**	**2**	**13**
うー	2	1	1	**4**	1	3	3	**7**	**1**	**12**
まー	2			**2**	1	1	2	**4**	**4**	**10**
うん	1			**1**	1		1	**2**	**6**	**9**
あー	3	1		**4**		1		**1**	**1**	**6**
いざ		3		**3**			1	**1**	**2**	**6**
さあ		2		**2**		1	1	**2**	**1**	**5**
ほら	2		1	**3**	1			**1**		**4**
初めまして							1	**1**	**3**	**4**
否						1	2	**3**	**1**	**4**
あっ		1	1	**2**					**2**	**4**
ううん							1	**1**	**2**	**3**
ああ	1			**1**			1	**1**	**1**	**3**
ねえ	1			**1**	1			**1**		**2**
今晩は			2	**2**						**2**
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
総計	64	55	49	**168**	79	72	73	**224**	**117**	**509**

表７　形状詞の産出状況

形状詞(205個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
大事	9	7	21	**37**	20	16	20	**56**	**20**	**113**
奇麗	15	9	12	**36**	2	10	15	**27**	**39**	**102**
有名	16	11	16	**43**	7	8	16	**31**	**21**	**95**
懸命	6	4	8	**18**	5	5	8	**18**	**57**	**93**
好き	5	9	11	**25**	10	10	8	**28**	**28**	**81**
簡単	11	14	13	**38**	10	7	9	**26**	**16**	**80**
大変	4	4	7	**15**	14	10	15	**39**	**21**	**75**
大丈夫	10	5	3	**18**	9	3	6	**18**	**32**	**68**
色々	12	3	3	**18**	15	6	3	**24**	**10**	**52**
大切	7	7	4	**18**	10	4	1	**15**	**19**	**52**
可哀想	6		3	**9**	7	9	12	**28**	**12**	**49**
様々		4	2	**6**	1	6	11	**18**	**19**	**43**
真面目	1	2	4	**7**	3	8	2	**13**	**23**	**43**
重要	5	9	3	**17**	9	3	2	**14**	**11**	**42**
非常	2	12	16	**30**	3	4	1	**8**	**4**	**42**
可能	4	8	7	**19**	4	2	4	**10**	**10**	**39**
ぼろぼろ	1	6	2	**9**					**28**	**37**
沢山	4	2	10	**16**	4		2	**6**	**9**	**31**
如何		4	3	**7**	3	3	6	**12**	**10**	**29**
確か	2	4	3	**9**	3	5	3	**11**	**9**	**29**
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
総計	181	205	233	**619**	183	212	221	**616**	**640**	**1875**

表７は形状詞の産出状況を示すものである。形状詞の中には、副詞として使われることが多いと思われるもの(例：非常(に)、(一生)懸命)が混じっている。本稿ではこれを分類しなおしていない。中国人留学生は「有名」「簡単」「非常」「可能」のような二字漢語のものをたくさん産出している。韓国人留学生は「大事」「大変」「可哀想」のような感情を表す形状詞をたくさん産出している。日本人学生は「(一生)懸命」「大丈夫」「真面目」のような形状詞を多く産出している。「ぼろぼろ」に関して、タスク「七夕の物語」紹介に使われている語彙であるが、韓国人留学生の作文には１例も見られなかった。

表８　代名詞の産出状況

代名詞(29個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
私-代名詞	98	125	131	**354**	96	137	113	**346**	**322**	**1022**
其れ	45	69	99	**213**	61	106	87	**254**	**165**	**632**
何	28	21	40	**89**	19	32	34	**85**	**137**	**311**
此れ	27	30	27	**84**	37	29	48	**114**	**80**	**278**
其処	4	23	17	**44**	18	17	14	**49**	**63**	**156**
何時	9	16	19	**44**	28	18	24	**70**	**41**	**155**
僕	21	13	4	**38**	46	6	10	**62**	**10**	**110**
俺	8		2	**10**	34	18	9	**61**	**32**	**103**
君-代名詞	12	19	1	**32**	3	13	12	**28**	**5**	**65**
彼	9	6	13	**28**	3	4	8	**15**	**16**	**59**
此処	5	5	13	**23**	8	10	7	**25**	**9**	**57**
貴方	2	2	4	**8**	19	7	8	**34**	**6**	**48**
彼女	8	11	4	**23**	5	2	8	**15**	**4**	**42**
何処	6	1	5	**12**	5	3	2	**10**	**20**	**42**
御前			1	**1**	11	7		**18**	**20**	**39**
誰	2	8	4	**14**	1	7	2	**10**	**14**	**38**
我々	5	2	9	**16**	2	8	4	**14**	**5**	**35**
私			5	**5**	2			**2**	**9**	**16**
此方		1	1	**2**	1		1	**2**	**10**	**14**
何れ		2	2	**4**			3	**3**	**6**	**13**
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
総計	292	366	404	**1062**	399	430	397	**1226**	**984**	**3272**

表８は代名詞の産出状況を示している。人称代名詞を見ると、中国人留学生は「僕」をよく産出するが、「俺」はあまり使わない。また、「我々」、「君」、「彼女」「彼」もよく産出する。韓国人留学生は「僕」、「俺」、「我々」、「君」、「貴方」をよく産出する。中韓両国の留学生はレベルが上がるにつれ、「僕」、「俺」の代わりに、「私」を使う場合が多くなる。日本人学生は親しい友人に対して使うと思われるが、「僕」より「俺」をよく産出している。人称代名詞の使用を控える日本人学生に比べ、留学生、特に韓国人留学生は上級～超級になっても人称代名詞を過剰に使用していることが分かる(韓：595＞中：515＞日：429)。一方で、中国人留学生は「俺」「御前」のようなややぞんざいな言い方を控える傾向にあることが分かる。指示代名詞「此れ」「其れ」「彼れ」「何れ」の産出状況を見ると、韓国人留学生には多く、日本人学生には少ないという、人称代名詞と同じ傾向が見られる(韓：371＞中：301＞日：251)。場所を表す指示代名詞「此処」「其処」「彼処」「何処」に関しては、日本人学生の使用は多いが、あまり差が見られない(日：92＞中：84＝韓：84)。

表９　連体詞の産出状況

連体詞(23個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
其の	66	105	116	**287**	112	137	110	**359**	**206**	**852**
此の	101	105	97	**303**	92	75	77	**244**	**206**	**753**
同じ	10	13	8	**31**	10	10	10	**30**	**30**	**91**
或る	12	13	9	**34**	11	3	8	**22**	**11**	**67**
そんな	4	2	7	**13**	7	2	2	**11**	**42**	**66**
こんな	12	4	2	**18**	5	5	3	**13**	**19**	**50**
大きな	5	8	6	**19**	3	2	10	**15**	**16**	**50**
色んな	4	4	2	**10**	11	6	7	**24**	**4**	**38**
どんな	1	3	1	**5**		8		**8**	**15**	**28**
彼の	5	1	1	**7**	3	1	2	**6**	**4**	**17**
何の	1	2	4	**7**		1	1	**2**	**4**	**13**
我が		3		**3**	1	2		**3**	**3**	**9**
更なる		2		**2**			1	**1**	**3**	**6**
主な	1		1	**2**	2	1	1	**4**		**6**
小さな	2		1	**3**					**2**	**5**
本の		2	1	**3**			1	**1**	**1**	**5**
所謂		3		**3**			1	**1**		**4**
大した		2	1	**3**			1	**1**		**4**
単なる		3		**3**					**1**	**4**
あらゆる							1	**1**	**1**	**2**
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
総計	224	275	258	**757**	258	254	236	**748**	**568**	**2073**

表９は連体詞の産出状況を示すものである。中韓両国留学生の作文とも、「此の」「其の」「或る」「色んな」が日本人学生より多く産出している。特に、中国人留学生は「此の」、韓国人留学生は「其の」が多く、日本人学生は「そんな」「どんな」の使用数が多いことが分かる。産出する連体詞の中で、「此の」「其の」は三カ国とも 70%～80%と高い割合を占めている(韓：603＞中：590＞日：412)。また、頻度こそ低いが、「大した」「所謂」「主な」は中韓両国の留学生の作文にともに見られたが、日本人学生の作文には見られなかった。

表１０　助動詞の産出状況

助動詞(38個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
だ	533	643	782	**1958**	514	721	755	**1990**	**2084**	**6032**
た	497	653	758	**1908**	606	723	730	**2059**	**1857**	**5824**
ます	571	766	681	**2018**	522	699	735	**1956**	**1645**	**5619**
です	367	388	284	**1039**	311	350	324	**985**	**824**	**2848**
ない	159	198	187	**544**	157	183	178	**518**	**478**	**1540**
れる	69	88	102	**259**	35	67	86	**188**	**269**	**716**
ず	41	104	77	**222**	42	85	79	**206**	**236**	**664**
様	45	52	90	**187**	51	86	106	**243**	**221**	**651**
たい	45	54	64	**163**	50	63	75	**188**	**128**	**479**
てる	36	34	31	**101**	35	53	70	**158**	**205**	**464**
せる	30	50	64	**144**	20	26	46	**92**	**124**	**360**
られる	32	31	34	**97**	11	23	26	**60**	**84**	**241**
ちゃう	14	12	5	**31**	12	10	17	**39**	**40**	**110**
らしい	1	2	11	**14**	10	9	18	**37**	**54**	**105**
べし	10	10	5	**25**	4	9	11	**24**	**34**	**83**
みたい	6	5	9	**20**	6	4	8	**18**	**28**	**66**
させる	7	14	8	**29**		4	8	**12**	**11**	**52**
そう-様態	2		7	**9**	3	6	8	**17**	**23**	**49**
り	5	1	1	**7**	4	1	3	**8**	**4**	**19**
なり-断定		3		**3**	1	1	1	**3**	**4**	**10**
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
総計	2477	3119	3202	**8798**	2401	3130	3292	**8823**	**8370**	**25991**

表１０は助動詞の産出状況を示すものである。中国人留学生は「です」、「ます」、使役助動詞「せる」・「させる」を多く産出している。韓国人留学生は「れる」・「られる」[4]、推量助動詞「様だ」、希望助動詞「たい」の使用が多い。日本人学生は「れる」「られる」、様態・推量助動詞「らしい」「みたいだ」「そうだ」、継続・持続の意味を表す「てる」、断定助動詞「だ」を多く産出しているが、「です」「ます」「たい」が少ない。また、留学生はレベルが上がるにつれ、「れる」・「られる」、様態助動詞、推量助動詞、希望助動詞の産出数が増えている。

次の表１１は終助詞の産出状況を示している。全体的に言えば、日本人学生は終助詞を多く産出するが、中国人留学生の産出は少ない。また、レベルが上がるにつれ、終助詞の産出数が増えている。中国人留学生が「の」「わ」を多く産出し、韓国人留学生は「ね、さ、の、じゃん、(「よ」、「さ」は下位群で多いが、上・中位群で減る)」を多く産出している。日本人学生は「か、よ、な、ぞ」を多く産出している。また、「よ」「ね」の産出数の中に、「よね」は 99 例で、内訳として日本人学生が多く産出している（日：60＞韓：30＞中：9)。

[4] 受け身、可能、尊敬、自発を表す用例すべてが含まれる。

表１１　終助詞の産出状況

終助詞(19個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
か	68	77	100	**245**	51	74	87	**212**	**299**	**756**
よ	56	69	61	**186**	77	68	62	**207**	**232**	**625**
ね	45	52	70	**167**	70	66	83	**219**	**207**	**593**
な	16	16	21	**53**	15	22	39	**76**	**117**	**246**
さ		8	2	**10**	19	11	12	**42**	**25**	**77**
の	10	6	11	**27**	6	17	14	**37**	**12**	**76**
じゃん	1	1	2	**4**	4	7	5	**16**	**8**	**28**
わ		4	5	**9**	2	3		**5**	**4**	**18**
ぞ	1		3	**4**					**9**	**13**
もの	1	1	1	**3**	1	1		**2**	**5**	**10**
ぜ						5		**5**	**2**	**7**
い	1		1	**2**	2	1		**3**	**1**	**6**
け	2			**2**	1			**1**	**1**	**4**
べい			3	**3**						**3**
のう									**2**	**2**
や						1	1	**2**		**2**
ねん							1	**1**	**1**	**2**
もが									**1**	**1**
ばや		1		**1**						**1**
総計	201	235	280	**716**	248	276	304	**828**	**926**	**2470**

表１２　準体助詞の産出状況

準体助詞(1個)	中国			中国 集計	韓国			韓国 集計	日本 集計	総計
レベル	下位群	中位群	上位群		下位群	中位群	上位群			
の	196	174	319	**689**	200	241	234	**675**	**792**	**2156**
総計	196	174	319	**689**	200	241	234	**675**	**792**	**2156**

表１２は準体助詞の産出状況を示している。日本人学生が準体助詞を多く産出している。

４．考察とまとめ

以上では、中韓両国の上級～超級日本語学習者が書かれた作文を品詞の使用実態を基に量的に使用傾向を見た。考察した結果、留学生は上位群へレベルが上がるにつれ、作文語数が増え、品詞の大半は異なり語数、延べ語数が増え、語彙量が豊富になり、相手との交渉表現、文構造が複雑化していることが読み取れる。

中韓両国の留学生の作文とも、日本人学生より感動詞を多く産出している。初級の挨拶語「今日は」「有り難う」が多く見られた。特に、中国人留学生は「有り難う」を、韓国人留学生は「今日は」「御早う」「あの」をよく使っており、日本人学生は「はい」「うん」といった応答用の感動詞を留学生より多く産出している。

形状詞に関しては、中国人留学生は「有名」のような二字漢語のものをたくさん産出している。中国語母語の影響があると思われる。韓国人留学生は「大変」「可哀想」のような感情を表す形状詞を、日本人学生は「(一生)懸命」「大丈夫」「真面目」「ぼろぼろ」のような形状詞をたくさん産出している。

人称代名詞に関しては、中国人留学生は「僕」をよく産出するが、「俺」はあまり使っていない。また、「我々」、「君」、「彼女」「彼」もよく産出している。韓国人留学生は「僕」、「俺」、「我々」、「君」、「貴方」をよく産出している。中韓両国の留学生ともレベルが上がるにつれ、「僕」、「俺」の代わりに、「私」を使う場合が多くなっている。日本人学生は親しい友人に対して使うと思われる「俺」が、「僕」より多く産出している。人称代名詞の使用を控える日本人学生に比べ、留学生、特に韓国人留学生は上級～超級になっても人称代名詞を過剰に使用する。一方で、中国人留学生は「俺」「御前」のようなややぞんざいな言い方を控える傾向にある。指示代名詞「此れ」「其れ」「彼れ」「何れ」に関しては、

韓国人留学生には多く、日本人学生には少ないという、人称代名詞と同じ傾向が見られる。

連体詞を見ると、中韓両国留学生とも、「此の」「其の」「或る」「色んな」が日本人学生より多く産出している。特に、中国人留学生は「此の」、韓国人留学生は「其の」が多く、日本人学生は「そんな」「どんな」の使用数が多いことが分かる。産出する連体詞の中で、「此の」「其の」は三カ国とも、70%～80%と高い割合を占めている。

中国人留学生は「です」、「ます」、使役助動詞を多く産出している。韓国人留学生は「れる・られる」、「様だ」、「たい」の使用が多い。日本人学生は「れる・られる」、様態・推量助動詞、「てる」、断定助動詞「だ」を多く産出するが、「です」「ます」「たい」が少ない。また、留学生はレベルが上がるにつれ、「れる・られる」、様態助動詞、推量助動詞、希望助動詞の産出数が増える。

終助詞に関しては、全体的に日本人学生は終助詞を多く産出するが、中国人留学生の産出は少ない。また、レベルが上がるにつれ、留学生の終助詞の産出数が増えている。中国人留学生が「の」「わ」を多く産出し、韓国人留学生は「ね、さ、の、じゃん」を多く産出している。日本人学生は「か、よ、な、ぞ」を多く産出し、また、「よね」も多く産出している。

日本人学生は準体助詞を多く産出している。準体助詞を使うことにより、文が複雑になり、視点固定にも寄与できるという特徴を持っている。中韓両国の留学生もレベルが上がるにつれ、準体助詞の使用数が増えており、日本人学生の使用状況に近づいている。

全体的に言えば、留学生の作文には挨拶語、人称代名詞、指示代名詞、連体詞「此の、其の」、助動詞「たい」が過剰に使用され、様態・推量助動詞、継続・持続助動詞の過少使用が見られた。中韓留学生の間にも、「此の、其の、此れ、其れ」の産出数、受身・尊敬、使役助動詞の産出数などの違いが見られる。中国人留学生は「です、ます」の多用、「俺、御前」の過少使用から、中国人留学生は相手が「親」でも、相対的に改まり度の高い文体で書いていることがうかがえる。日本人学生の作文は句点が少なく、指示代名詞、連体詞「此の、其の」が少なく、また準体助詞が多いことから、文が長く、文構造が複雑だと推測される。

レベルが上がるにつれ、留学生は感動詞の産出が減り、「僕・俺」が「私」に取って代わられ、形状詞・助動詞・終助詞が増加傾向にあるなど、日本人の使用実態に近づいている。

以上の考察により、留学生が日本語のレベルが上がるにつれ、全体的傾向として日本人学生の言語産出に近づいていることが分かる。ただし、人称代名詞の多用、様態・推量助動詞の使用が少ないなど、上級～超級になっても日本人学生の産出と違う特徴が見られる。

参考文献

石川慎一郎(2012)『ベーシックコーパス言語学』ひつじ書房
小野望・田中省作・持尾弘司(2007)「母語学習者コーパスの基礎調査」『筑紫女学園大学・短期大学部人間文化研究所年報』18, 27-36,筑紫女学園大学・短期大学部人間文化研究所
金澤裕之(編)(2014)『日本語教育のためのタスク別書き言葉コーパス』ひつじ書房
橋本直幸(2011)「学習者コーパスから見る超級日本語学習者の言語特徴―２つの観点から―」『日本語教育文法研究のための多様なアプローチ』ひつじ書房
毛文偉(2013)「中国日語学習者作文詞汇量及高頻詞目研究」『外語電化教学』152,9-15
山内博之(2009)『プロフィシェンシーから見た日本語教育文法』ひつじ書房

使用データと形態素解析ツール

金澤裕之(編)(2014)『日本語教育のためのタスク別書き言葉コーパス』付属ＣＤのデータ
形態素解析処理ソフト「茶まめ」、解析器「MeCab 0.996」、解析用辞書「UniDic-mecab 2.1.2」

医療経過記録における名詞連続の計量的特徴

山崎　誠（国立国語研究所言語資源研究系）†
相良　かおる（西南女学院大学保健福祉学部）‡

Metric Characteristics of Noun Sequences in Medical Progress Notes

Makoto Yamazaki (Dept. Corpus Studies, NINJAL)
Sagara Kaoru (Faculty of Health and Welfare, Seinan Jo Gakuin University)

要旨

医療経過記録は医療の現場において作成されるメモ的な性格の強い文章である。品詞的な特徴としては名詞の比率が高く、助詞・助動詞の省略が多い。このような文章によく見られる現象として機能語を用いず、名詞等を連続して用いる臨時一語的な用法がある。本発表では､小児科における医療経過記録約 90 万短単位から抽出した名詞連続の構造を品詞、語種、意味を中心に分析したものである。比較の対象として BCCWJ を用いて、医療経過記録の特徴を明らかにした。

１．はじめに

医療記録には、専門用語に加え、略語や隠語が、そして独特な表現が含まれる。紙媒体に記録される医療記録は､限られた場所で限られた医療従事者により記録され､閲覧され､保管されてきたが、近年の電子カルテシステムの普及により、施設内での情報の共有が可能となった。

しかし、医療用語の標準化がなされないまま、電子カルテシステムが導入されていることから、医療記録データには、表記のゆれや誤字を含む同義語、類義語が含まれている。これらの自然言語処理には、機械可読のコーパスや用語辞書が必要であるが、個人情報を含む医療記録は門外不出であり、言語学的調査は容易ではない。

今回我々は、研究利用のために提供された倫理的配慮のなされた小児看護領域のプログレスノート（以後､「医療経過記録」と言う）のデータ（スペースを含め 1,355,656 文字、短単位で 906,504）について言語的調査を行った。

医療経過記録は、症状や処置などを簡潔に記録するため、短い文が多く、文を圧縮したような表現が頻繁に現れる。樺島（1979）によると、要約的な文章は名詞の比率が大きいという指摘がある。今回用いた医療経過記録のデータの品詞分布は表 1 のようになっている｡樺島(1979)の調査では名詞の比率が高いテキストとして､新聞見出し(名詞比率 74.0%)、新聞記事（同 68.3%）が挙げられているが、表 1 の名詞の比率は新聞見出しよりも高いことが分かる｡なお､表 1 は､データを Unidic-mecab 2.1.2 により形態素解析した結果から、樺島の調査結果に合わせるため、当該の品詞のみを抜き出して集計したものである。

山崎・相良（2014）では要約的な文章に出現する複合語の中に林（1982)、石井（1993）らが扱っている臨時一語が多く含まれると予想されることから、その構造分析を通して、

† yamazaki@ninjal.ac.jp
‡ sagara@seinan-jo.ac.jp

医療経過記録の特徴を明らかにしようとした。同稿ではサ変動詞になる漢字連続に限って分析したが、本稿では、名詞連続の特徴を『現代日本語書き言葉均衡コーパス』（以下、BCCWJ と略す）との比較を通して観察する。

表 1　医療経過記録の品詞分布

品詞	語数	割合（%）
名詞	362,166	77.23
動詞	72,491	15.46
形容詞	11,534	2.46
形状詞	8,863	1.89
副詞	10,157	2.17
連体詞	2,238	0.48
接続詞	941	0.20
感動詞	531	0.11
計	468,921	100.00

２．データ

２．１　医療経過記録

使用したデータは、小児科の医療経過記録の自由記載部分を抽出したものである。

医療施設での匿名化処理として、数値は“9”に置換し、固有名詞および個人名は“X”に置換されている。

また、利用者が意図的に改行を行った個所および文中の"。"の直後で分割したものを 1 行の文字列とした。同内容の文があった場合は、一方は削除されている。従って、“患児が「お腹がすいた。何か食べたい。」と言った。”というデータは、“患児が「お腹がすいた。”、“何か食べたい。”、“と言った。”の 3 行に分割されるため、構文についての分析調査には適さない。その他に以下の制限事項がある。

①検査項目である"Co2"や"HbA1c"など数値を含む固有名詞は"Co9"、"HbA9c"となっている。

②アルファベットによる固有名詞は匿名化の対象外としている。

③カタカナの固有名詞や文字長が 1 文字の固有名詞は匿名化の置換対象から除外されている。

２．２　BCCWJ

対照するデータとして BCCWJ（ver.1.0）を選んだ。BCCWJ 全体、および、医学系のサンプルがまとまって存在し、ジャンルとして抽出可能な LB（図書館書籍）、PB（出版書籍）、OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）、PM（出版雑誌）から該当するサンプルを選んだ。

３．方法

３．１　形態素解析

医療経過記録のデータは MeCab ver.0.996+unidic-mecab ver.2.1.1 で解析し、品詞の大分類が「名詞」（品詞の中分類が「接尾辞-名詞的」を含む）の連続を抽出した。抽出された名詞連続数は延べで 63,916 個、異なりで 21,874 個である。

BCCWJ のデータは既に形態素解析が施されているので、それを利用した。

３．２　医学系サンプルの抽出

BCCWJ 全体のほかに、医学系のサンプルをまとまって抽出できるレジスターとして LB、PB、OC、OY、PM がある。それぞれ、以下の方法でサンプルを抽出した。

LB、PB：NDC が 49（医学）ではじまる 1,137 サンプル（内訳：LB、346 サンプル、PB、791 サンプル）。

OC：ジャンル 3 が「健康、病気、ダイエット」および「病気、症状、ヘルスケア」である 3,705 サンプル。

OY：ジャンル 2 が「病気、症状」である 694 サンプル。

PM：ジャンル 3 が「医学」である 31 サンプル。

４．結果

４．１　BCCWJ 全体との比較

医療経過記録と BCCWJ 全体の品詞、語種の割合を比較する。この場合の品詞、語種は名詞連続を構成している各短単位をすべて数えたものである。図 1 は品詞の構成比、図 2 は語種の構成比である。品詞は名詞を中分類まで細分して示した。名詞の中分類には、普通名詞、固有名詞、数詞の 3 つがあるが、医療経過記録（図で「MD」と示した）は BCCWJ

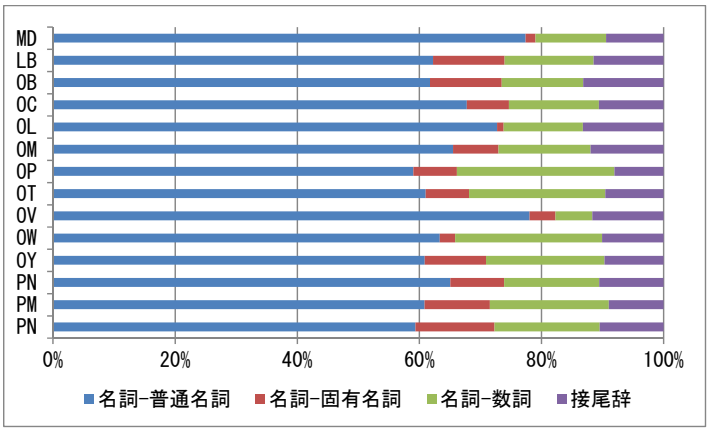


図 1　BCCWJ 全体との比較：品詞の構成比

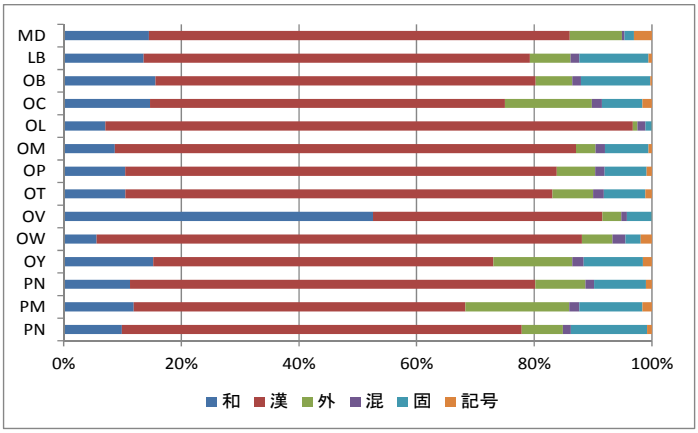


図 2　BCCWJ 全体との比較：語種の構成比

における各レジスターと比べて相対的に普通名詞の割合が高く、固有名詞が少ないことが分かる。固有名詞が少ないのはデータの匿名化のためと思われる。語種では品詞と同様に

固有名詞[1]が少ない。

４．２　医学系サンプルとの比較

図 3、図 4 は BCCWJ から医学系のサンプルを抜き出し、前節と同じ方法で比較したものである。品詞の割合では普通名詞の比率は BCCWJ 全体の場合と大きな差は見られない。

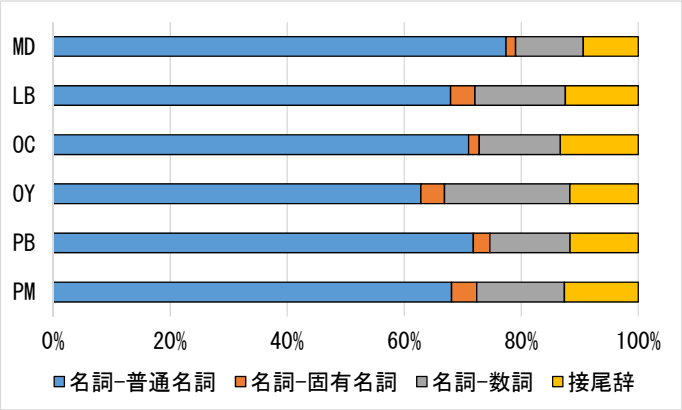


図 3　医学系サンプルとの比較：品詞の構成比

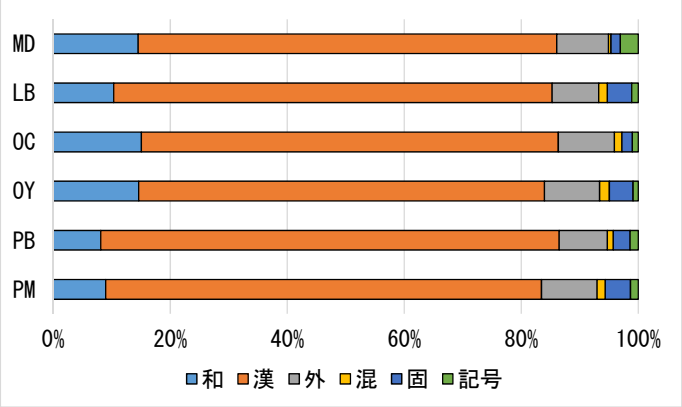


図 4　医学系サンプルとの比較：語種の構成比

次にカバー率の推移を比較する。図 5、図 6 は医療経過記録とデータの大きさが近い LB（図書館書籍の医学系サンプル）を比較したものである。LB の医学系サンプルの名詞連続数は延べで 62,028 個、異なりで 28,201 個であり、医療経過記録の延べ 63,916 個、異なり 21,874 個に近い。

図の横軸は異なり語数、縦軸は延べ語数であり、頻度の降順に並べた場合、上位何語[2]までで延べ語数の何%がカバーできるかを示している。比較のために、横軸、縦軸ともに 0%～100%を範囲とした標準化した数値で示した。

カバー率の推移は、医療経過記録（MD）の立ち上がりが早く、異なり語数の上位 20%で延べ語数の 70%、上位 40%で 80%に達する。図書館書籍（LB）の方はそれよりも 10 ポイントほど低い値となっている。

---

[1] 語種に固有名詞というカテゴリーがあるのは、UniDic の仕様による。

[2] 同順位があった場合、その順位を構成する n 語の異なりに対して、直前の順位を r とすると、r+1、r+2,…, r+n の順位の値を与えている。これはグラフのカーブを円滑化するための便宜的なものである。

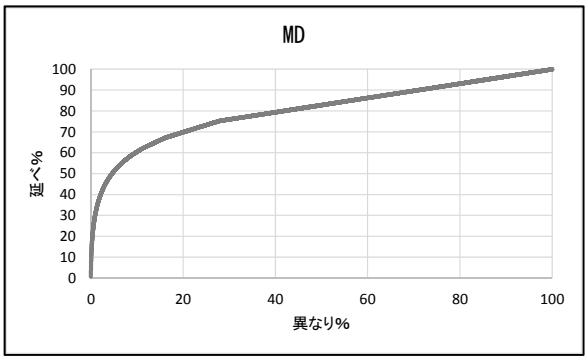


図 5 カバー率（MD)

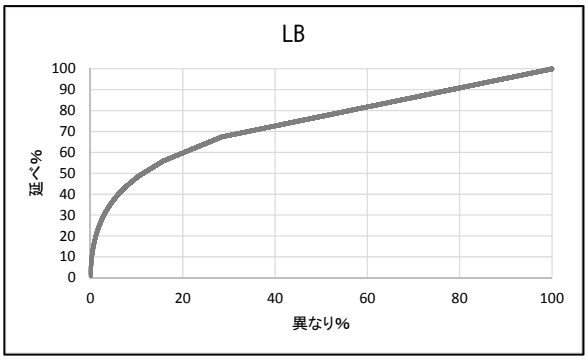


図 6　カバー率（LB)

次に使用頻度の多い語を個別に見てみよう.表 2 はそれぞれのレジスターの上位 20 語を示したものである。医療経過記録（MD）のデータは上記に数字が多く来ていることが特徴的である。これは体温などの検査の値や日付など、記録をとる上で重要な要素として記述されているものと思われる。個別の語としては、LB、OY、PB、PM に「患者さん」という

表 2　使用頻度上位 20 語[3]

	MD		LB		OC		OY		PB		PM	
順位	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素	頻度	語彙素
1	480	九. 九	512	一つ	170	一度	52	一つ	981	一つ	81	腹膜透析
2	471	九九日	431	患者さん	141	花粉症	38	日本ブログ村	822	患者さん	64	患者さん
3	455	九九	159	日本人	134	皮膚科	36	皆さん	375	高齢者	51	脂肪肝
4	398	九九九	158	人達	122	歯医者	28	花粉症	343	二つ	47	糖尿病
5	389	問題無し	155	糖尿病	119	皆さん	28	午前中	337	蛋白質	44	血液透析
6	369	九九九九グラム	142	母さん	112	整形外科	27	患者さん	314	図一	37	一つ
7	369	九. 九九	142	二つ	109	婦人科	26	一度	283	図二	26	合併症
8	328	九九. 九	139	遺伝子	96	血液検査	25	零零	266	看護師	21	コレステロール値
9	314	九回	137	癌細胞	96	耳鼻科	23	兄ちゃん	262	幾つ	21	大根番茶
10	314	九月	130	蛋白質	91	一週間	23	一日	262	図三	21	水キムチ
11	313	経過観察	126	幾つ	85	一日	23	何度	252	活性化	19	粗鬆症
12	306	圧痛無し	120	一度	75	生理痛	19	三十分	251	糖尿病	18	助中
13	300	九九九九	113	一日	75	一つ	18	子供達	221	人達	17	血液中
14	279	改善傾向	103	動脈硬化	73	健康診断	18	人達	215	遺伝子	16	高齢者
15	262	保育園	102	子供達	71	一箇月	16	一年	197	抗生物質	16	新生血管
16	240	九日	95	厚生省	67	一回	16	艶ちゃん	197	日本人	16	生活習慣病
17	214	九九時	91	治療法	66	歯医者さん	16	ヘルスブログ	195	図四	16	掌握握
18	210	九九. 九度	87	図一	64	医者さん	15	体重増加	191	表二	16	葡萄糖
19	207	全身状態	86	合併症	59	口内炎	15	皆様	190	図五		
20	196	九九度	86	神経細胞	57	産婦人科			188	十二		

[3] OY（Yahoo!ブログ）、PM（出版雑誌）は、同順位の語が複数あり、20 語を超えてしまうため内輪の範囲を挙げた。

語が現れているが、医療経過記録には見られないことである[4]。頻度 1 まで見ると「患者さん皆」が 1 例あるだけである。医療経過記録は患者の様態を記録するものなので、患者の存在が前提となっているため、「患者」という言葉を使う必要がないものと推察される。同様に「医者」「医師」の使用頻度も他のデータと比べると低い。

また、表 2 からは分からないが、症状を現す語として「～無し」が多用されていることも観察される。上位 100 位内に、「問題無し」(389)、「圧痛無し」(306)、「嘔吐無し」(129)、「異常無し」(120)、「下痢無し」(102)、「発熱無し」(99)、「変化無し」(75)、「著変無し」(69)、「咽頭発赤無し」(68)、「左右差無し」(68)、「発赤無し」(65) の 11 語が現れている[5]。逆に「～有り」は少なく、上位 100 位内には「必要有り」(64) の 1 語にとどまっている。

５．まとめと今後の課題

本稿では、医療経過記録の名詞連続を BCCWJ と比較しながら計量語彙論的な観点から概観した。医療経過記録はデータ的に制限が多く、分析に限界があるが、名詞連続の示す特徴の一端を具体的に示すことができた。本稿はケーススタディ的な考察であり、今後他の観点も交えた考察が必要である。例えば、語構成的な観点での分析や構成要素間の意味的な関係についての分析が今後の課題である。

謝　辞

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

石井正彦(1993) 臨時一語と文章の凝縮，『国語学』173，pp.104-91.
樺島忠夫(1979)『日本語のスタイルブック』大修館書店.
林四郎(1982) 臨時一語の構造，『国語学』131，pp.15-26.
山崎誠・相良かおる(2014) 医療経過記録における漢字連続複合語の計量的分析，人文科学とコンピュータシンポジウム論文集，pp.221-226.

[4] OC（Yahoo!知恵袋）には順位 29 位（頻度 47）で「患者さん」が登場する。
[5] 括弧内は頻度数。

書　名　第7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3月3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）