『日本語話し言葉コーパス』UniDic 版形態論情報の構築

渡部 涼子（国立国語研究所コーパス開発センター）†
田中 弥生（国立国語研究所理論構造研究系）
小磯 花絵（国立国語研究所理論構造研究系）

Constructing the UniDic Version of the Morphological Information of *Corpus of Spontaneous Japanese*

Ryoko Watanabe　Yayoi Tanaka　Hanae Koiso
(National Institute for Japanese Language and Linguistics)

要旨

『日本語話し言葉コーパス』（CSJ）には形態論情報として短単位と長単位の情報が付与されている。しかし，単位設計や品詞体系の点において，BCCWJ に付与されているものとは異なるため，CSJ と BCCWJ を単純に比較することができないという問題があった。そこで，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換し，中納言検索システムを通して公開することとした。本発表では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点，および UniDic 体系への変換手続きなどについて述べる。また，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスター（学会講演・模擬講演・対話）の品詞・語種の特徴を，BCCWJ の各種レジスター（書籍・新聞・行政白書・Web など）との比較を通して示す。

１．はじめに

『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese，CSJ）は，1999 年から 5 年間かけ，国立国語研究所・情報通信研究機構（旧通信総合研究所）・東京工業大学が共同で開発した，約 660 時間の日本語自発音声からなるデータベースである（国語研究所 2006）。2004 年に公開を開始して以降，音声言語情報処理，自然言語処理，日本語学，言語学，音声学，心理学，社会学，日本語教育，辞書編纂など幅広い領域で利用されてきた。

CSJ には，転記情報や文節情報，形態論情報，節単位情報，分節音情報，韻律情報，係り受け構造情報，談話境界情報，要約・重要文情報，印象評定データなど，多様な研究用付加情報（アノテーション）が付されている。このうち形態論情報については，例えば「国立国語研究所」のような複合語を一つの単位とする長単位と，これらを「国立|国語|研究|所」のように細かく分割する短単位の二種類の情報が付与されており，この点において『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）と同じであるが，単位設計について一部基準が異なる上に，品詞体系についてはかなりの相違が見られる。そのため，CSJ と BCCWJ を同一基準で検索したり，あるいは比較したりといったことができないという問題があった。そこで，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換し，BCCWJ と同じ WEB 上の検索システムを通して公開することとした。

† naberyo＠ninjal.ac.jp

本稿では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点，および UniDic 体系への変換手続きなどについて述べる。また，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスター（学会講演・模擬公演・対話）の品詞・語種の特徴を，BCCWJ の各種レジスター（書籍・新聞・行政白書・Web など）との比較を通して示す。

２．CSJ UniDic 版形態論情報の整備

２.１　CSJ オリジナル版短単位体系と UniDic 体系の設計上の主な違い

２.１.１　単位設計

CSJ オリジナル版の短単位は，現代語において意味を持つ最小の単位（最小単位と呼ぶ）二つが 1 回結合したものであり，『現代雑誌九十種』の用語用字で用いられた β 単位がもとになっている（小椋 2006）。以下に例を示す。なお，短単位の境界は「∥」，最小単位の境界は「｜」で表す。

∥話し｜言葉∥　　∥音｜声∥　　∥レーザー｜プリンター∥　　∥行こ∥う∥

コーパス日本語学への応用を志向して開発された形態素解析用辞書 UniDic（伝ほか 2007；伝ほか 2008）においても，単位設計については原則として CSJ オリジナル版の短単位基準が踏襲されたが，以下のような変更が加えられた（小椋 2008）。

- 外来語は 1 最小単位で 1 短単位とする。

  ∥レーザー∥プリンター∥　　∥オレンジ∥色∥

- 意思・推量の助動詞「う」「よう」を独立の単位とせず，活用語尾として活用語の単位に含める。

  ∥行こう∥　　∥食べよう∥

- 補助記号（「・」「,」「。」など）を独立の最小単位として認定し，1 最小単位で 1 短単位とする。

２.１.２　付加情報

単位認定基準によって認定した一つ一つの短単位は，活用変化・音の転訛・ゆれ・省略・融合等によって生じた異形態や異表記形そのままの形のものであるため，用例検索や計量研究において扱い難い。そこで CSJ オリジナル版では，転記テキストにおける短単位の出現語形（出現形，転記における基本形）とその発音（発音形）について，それぞれの単位が同じ語であるかどうか判断し，同じ語と判断した語群に対して，見出しといえる「代表形」を片仮名で付与している。また，代表形に加えて，代表形を漢字等で表記した「代表表記」という情報も与えている。代表形は片仮名で表記されているため，代表形だけでは同音異義語の区別がつかなくなってしまうが，代表表記を与えることで同音語の区別が可能となる。

UniDic ではこの点をさらに整理し，また表記の変異にも対応するべく，次のように語彙素（語彙素読み）・語形・書字形・発音形からなる階層的見出しを採用している（表 1）。

表 1　UniDic 階層的見出しの例

語彙素	語形	書字形	発音形
矢張り	ヤハリ	やはり	ヤハリ
		矢張り	
	ヤッパリ	やっぱり	ヤッパリ

２.１.３　品詞体系

CSJ のオリジナル版短単位情報は，後述するように，手作業により高精度に情報を付与した人手作業分と，それを学習データとして構築した形態素解析システムで自動解析した自動解析分の二種類がある。このうち人手作業分の品詞情報は，UniDic に比べ，詳細な分類を行なわない，粗いものとなっている。CSJ 作成時点ではコーパスを活用した研究がまだそれほど進んでおらず，どのような品詞情報が有用かの判断材料が極めて乏しい状態だった。そのため，まずは最低限必要な品詞情報を付与しておき，実際に研究に活用していく中でどのような品詞情報が望ましいか検討していく方針を取った。

具体的に名詞を例にして比較をすると，表 2 のとおり，UniDic の方が細かく下位分類まで設定されている（小椋ほか 2011）。

表 2　CSJ オリジナル版（人手作業分）と UniDic との品詞（名詞）の比較

CSJ	
品詞	その他1
名詞	
	固有名詞
	数詞

UniDic			
大分類	中分類	小分類	細分類
名詞	普通名詞	一般	
		サ変可能	
		サ変形状詞可能	
		形状詞可能	
		副詞可能	
		助数詞可能	
	固有名詞	一般	
		人名	一般
			姓
			名
		地名	一般
			国
	数詞		
	助動詞語幹		

活用語についても UniDic の方が詳細な分類となっている（小椋ほか 2011）。五段動詞を例に挙げる（表 3）。ただし，活用の種類と活用形については，同じ CSJ オリジナル版であっても，人手作業分と自動解析分では粒度が異なっており，自動解析分の方がその粒度が細かくなっている。詳細については山口ほか（2004a，2004b）を参照されたい。

表 3　CSJ オリジナル版（人手作業分）と UniDic との活用の種類（五段動詞）の比較

CSJ
カ行五段
ガ行五段
サ行五段
タ行五段
ナ行五段
バ行五段
マ行五段
ラ行五段
ワア行五段

UniDic		
五段	カ行	一般
		イク
		ユク
	ガ行	
	サ行	
	タ行	
	ナ行	
	バ行	
	マ行	一般
		済ム
	ラ行	一般
		アル
		サル
	ワア行	一般
		○ウ

また CSJ オリジナル版では，名詞のうち形状詞や副詞としても使われる語について，文脈等に基づいて名詞・形状詞・副詞の判定を行っているが，UniDic では「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」という品詞を実際の使用例に関わらず与えている。

２.２　変換手続き

CSJ のオリジナル版短単位情報は，次の二通りの方法で付与された。

- **人手作業：**約 100 万語（種々のアノテーションを人手で高精度に付与したコア 50 万語を内包）については，人手により高精度に情報を付与。
- **自動解析：**残り約 650 万語については，上記人手作業分を学習データに構築された形態素解析システム（内元ほか 2004）により自動解析した上で，部分的に人手修正。

■人手作業分のデータの変換手続き： UniDic 構築時に，学習用データとして人手で UniDic 体系に変換する作業を実施した（伝ほか 2007)。ただし，「こ これは」の「こ」のように，言いよどみに伴う語の断片は消去した上で学習用データが作成されたため，今回の整備作業で語断片を元の位置に復元した。これに伴い「言いよどみ」という品詞を新たに設けた。

■自動解析分のデータの変換手続き：次の通り変換作業を行った。

1. UniDic Ver.2.0 をもとに，CSJ オリジナル体系から UniDic 体系に自動で変換した。自動変換に先立ち，単位の粒度が異なるもののうち助動詞「う」「よう」については，活用語尾として活用語の単位にまとめる作業を自動で行った。
2. 変換候補が複数ある場合，出現確率などから，一意に自動で決定するものと，複数項目を列挙するものに分け，後者については人手で確認のうえ認定した。
3. 変換候補がない場合，次の通り対応した。
   a. UniDic に登録されていない語は，一旦保留とした。
   b. 「レーザープリンター」のように単位の粒度が異なるものは，候補を自動で抽出した上で, 分割パターンを半自動で特定した。変換候補が複数ある場合は 2 の処理を，未登録語などを含む場合は一旦保留とした。
4. 上記作業を行い，一通り UniDic 体系に変換したのち，UniDic と連動してコーパスの管理・修正作業を行うことのできるデータベースシステム（「大納言」）に搭載した。
5. 全ての未登録語を対象に, UniDic に人手で新規に語を登録した上で, 大納言上で UniDic にリンクさせる形でコーパスに情報を付与した。

■伏せ字の扱い：オリジナル版では，話者の氏名など話者を特定できる情報や差別語などについて，出現形，発音形，代表形，代表表記は伏せ字化した上で，品詞情報についてはそのまま公開している。UniDic 版を作成するにあたり，人手作業分についてはこの方針を踏襲し，品詞情報を残す形で整備した。一方，自動解析分については，品詞情報の変換はせず，品詞を一律「伏せ字」とした。この点において，人手作業分と自動解析分で扱いが異なるため，利用の際には注意が必要である。

■発音形の扱い: CSJ の転記テキストでは，実際の音声を仮名で書ける範囲で忠実に記録している。その際，「手術（シュジュツ)」を「シジツ」，「形態素（ケイタイソ)」を「ケーソタイ」と発音するなど，発音の怠けや転訛，言い間違いなどが生じた場合には，実際に発音された音と，丁寧に発音された場合に生じるであろう音を「(W シジツ;シュジュツ)」のような形で併記して表現している。オリジナル版短単位情報における発音形では，これら二つの発音情報を共に保存する形で表現しているが，UniDic 体系に変換するにあたり，コーパスと辞書の管理方法の都合などから，実際の発音情報は対象とせず，丁寧に発音された場合に生じるであろう音のみを記すこととした。UniDic 体系での実際の発音の表現については今後の課題とする。

■節単位：BCCWJ などの書き言葉では，文が認定され中納言などでの検索に利用されている。しかし話し言葉の場合，文の認定は必ずしも容易ではない。そこで CSJ では，文に代わる単位として節単位（丸山ほか 2006）が認定されている。中納言における CSJ の検索においても，この節単位を利用する。

２.３　解析精度

CSJ 自動解析分を 2.2 節の手続きに従い UniDic 体系に自動変換したデータ群に対し，ランダムに 1 万語を抽出し，①境界（単位境界が正解と一致するか否か），②品詞（境界に加え，品詞・活用型・活用形が正解と一致するか否か），③語彙素（境界・品詞・活用型・活用形に加え，語彙素が正解と一致するか否か）の三段階でその精度を評価した。結果（F 値）を図 1 に示す。

参考までに，一般的な自動解析のデータである，UniDic-mecab 1.3.12 による BCCWJ・CSJ のレジスター別自動解析精度[1]をともに示す（図 2）。なお，図 2 における CSJ とは，前節で言及した人手作業分データを UniDic の学習データ用に整備したものから一部抽出したものである。

①境界の精度は，自動変換・UniDic-mecab 1.3.12 ともほぼ同じ値を示している。②の品詞と③の語彙素の精度については，白書には及ばないものの，他のレジスターよりも高い値を示している。これは，2.2 節の自動解析分のデータの変換手続きで述べたように，全ての未登録語について，事前に登録処理を施したためである。

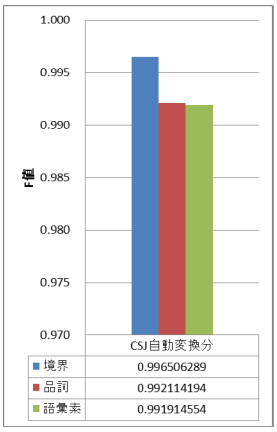

図 1　CSJ 自動変換分の精度

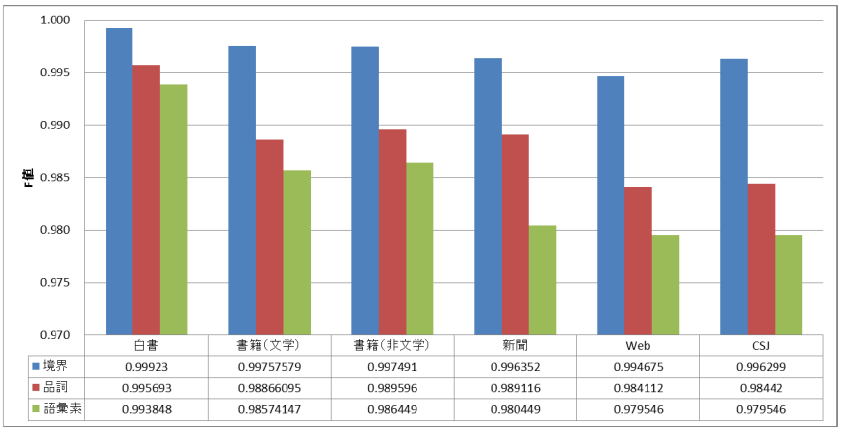

図 2　UniDic-mecab 1.3.12 による BCCWJ・CSJ のレジスター別解析精度

３．CSJ の形態論情報の特徴

３.１　CSJ の基礎統計量

表 4 に，CSJ オリジナル版と UniDic 体系変換後の短単位の語数を，人手作業・自動解析別，レジスター（学会講演＋その他の講演（以下，学会講演），模擬講演，対話，朗読）別に示す。CSJ オリジナル版と UniDic 版の語数が若干異なるのは，2.1.1 節に記した通り，単位の粒度の基準が一部異なるためである。

表 4　CSJ オリジナル版・UniDic 版の語数

	CSJ オリジナル版			UniDic 版		
	全体	人手作業	自動解析	全体	人手作業	自動解析
学会講演	3,597,474	518,024	3,079,450	3,607,546	518,798	3,088,748
模擬講演	3,637,723	436,171	3,201,552	3,640,805	436,069	3,204,736
対話	151,445	41,925	109,520	150,794	41,678	109,116
朗読	208,563	18,976	189,587	209,395	19,031	190,364
計	7,595,205	1,015,096	6,580,109	7,608,540	1,015,576	6,592,964

[1]「UniDic の解析精度」http://download.unidic.org/?page_id=12　参照

また表 5 と表 6 に，UniDic 版の各品詞，各語種の頻度を，人手作業・自動解析ごと，およびレジスターごとに示す。人手作業分と自動解析分の各品詞・語種の比率を比較すると，ほぼ同じ分布となることから，レジスターごとの頻度については，人手作業分と自動解析分に分けず，両者の合計値のみを示す。

表 5　UniDic 版の語数：品詞別

	全体	人手作業	自動解析	学会講演	模擬講演	対話	朗読
名詞	1,818,904	240,674	1,578,230	959,592	781,633	25,608	52,071
代名詞	160,478	21,377	139,101	64,142	85,442	3,957	6,937
形状詞	90,082	12,729	77,353	44,592	42,350	1,637	1,503
連体詞	94,383	12,847	81,536	50,450	41,018	1,522	1,393
副詞	219,651	29,414	190,237	73,383	132,483	8,083	5,702
接続詞	84,161	11,757	72,404	43,414	38,211	1,534	1,002
感動詞	473,527	70,234	403,293	242,661	207,356	18,759	4,751
動詞	997,482	129,836	867,646	470,295	482,220	16,335	28,632
形容詞	106,574	14,741	91,833	36,137	65,110	3,121	2,206
助動詞	886,347	119,650	766,697	386,202	455,708	19,382	25,055
助詞	2,335,347	308,060	2,027,287	1,049,007	1,172,432	45,045	68,863
格助詞	1,188,374	157,806	1,030,568	608,146	534,057	15,064	31,107
係助詞	294,909	38,675	256,234	124,248	155,684	5,493	9,484
接続助詞	405,425	53,689	351,736	176,677	212,570	5,870	10,308
終助詞	124,246	16,108	108,138	37,629	71,343	8,489	6,785
副助詞	168,841	21,635	147,206	52,112	105,152	5,670	5,907
準体助詞	153,552	20,147	133,405	50,195	93,626	4,459	5,272
接頭辞	42,080	6,079	36,001	20,747	20,131	622	580
接尾辞	160,877	20,589	140,288	84,218	67,816	2,288	6,555
記号	32,339	4,295	28,044	25,379	3,988	293	2,679
言いよどみ	96,116	13,294	82,822	47,462	44,658	2,548	1,448
その他	10,192	0	10,192	9,865	249	60	18

表 6　UniDic 版の語数：語種別

	全体	人手作業	自動解析	学会講演	模擬講演	対話	朗読
和語	5,893,040	788,933	5,104,107	2,626,644	2,979,628	127,338	159,430
漢語	1,256,168	164,910	1,091,258	733,050	471,120	14,678	37,320
外来語	178,172	24,137	154,035	104,511	68,674	1,885	3,102
混種語	55,269	7,973	47,296	25,138	28,252	904	975
固有名	72,091	10,302	61,789	25,413	42,364	3,042	1,272
その他	153,800	19,321	134,479	92,790	50,767	2,947	7,296

３.２　品詞率・語種率に見る CSJ のレジスターの特徴

本節では，品詞ごと，語種ごとの出現率から，CSJ の各レジスターの特徴を見ていく。

図 3 に，CSJ（全体）の品詞・語種の出現率を，朗読を除く三つのレジスターごとに示す。また図 4 に，BCCWJ（コア・非コア含む全体）の品詞・語種の出現率を，書籍，新聞，白書，雑誌， Yahoo!知恵袋，国会会議録に限定し，レジスターごとに示す。個々の品詞率，語種率は，サンプルごとの延べ語数に対する各品詞・語種の延べ語数の割合として求めた。ただし品詞率の算出にあたり，CSJ 固有の品詞である言いよどみと伏せ字，および CSJ に頻出する感動詞（「あのー」や「えっと」などのフィラーを含む）は集計の対象としなかった。語種については更に，助詞，助動詞，固有名詞，記号を除外した上で比率を求めた。

図には，小磯ほか（2009）など BCCWJ を主対象とする一連の文体研究で特徴的な傾向を示した品詞・語種を抜粋して示す。なお小磯ほか(2009)では，BCCWJ の構築期間中に，BCCWJ の五つのレジスターおよび CSJ 人手作業分の学会講演と模擬講演を対象に，各レジスターから 150 のサンプルを抽出して品詞率・語種率を求めた。今回の分析では，レジス

ターとして, CSJ から対話を, BCCWJ から雑誌を追加しており, また CSJ, BCCWJ ともに, サンプル数を限定せず，当該レジスターに属する全てのデータを利用している。

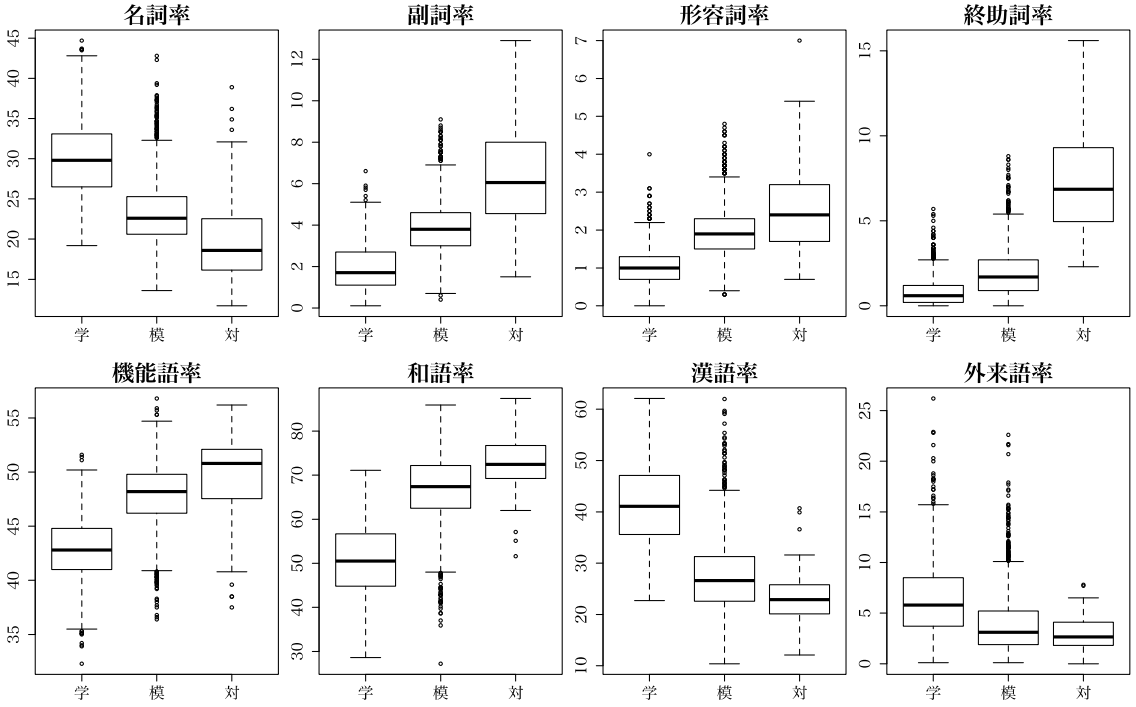

図 3　CSJ の品詞ごと・語種ごとの出現率（中央値と第 1・第 3 四分位数）
学：学会講演，**模**：模擬講演，**対**：対話

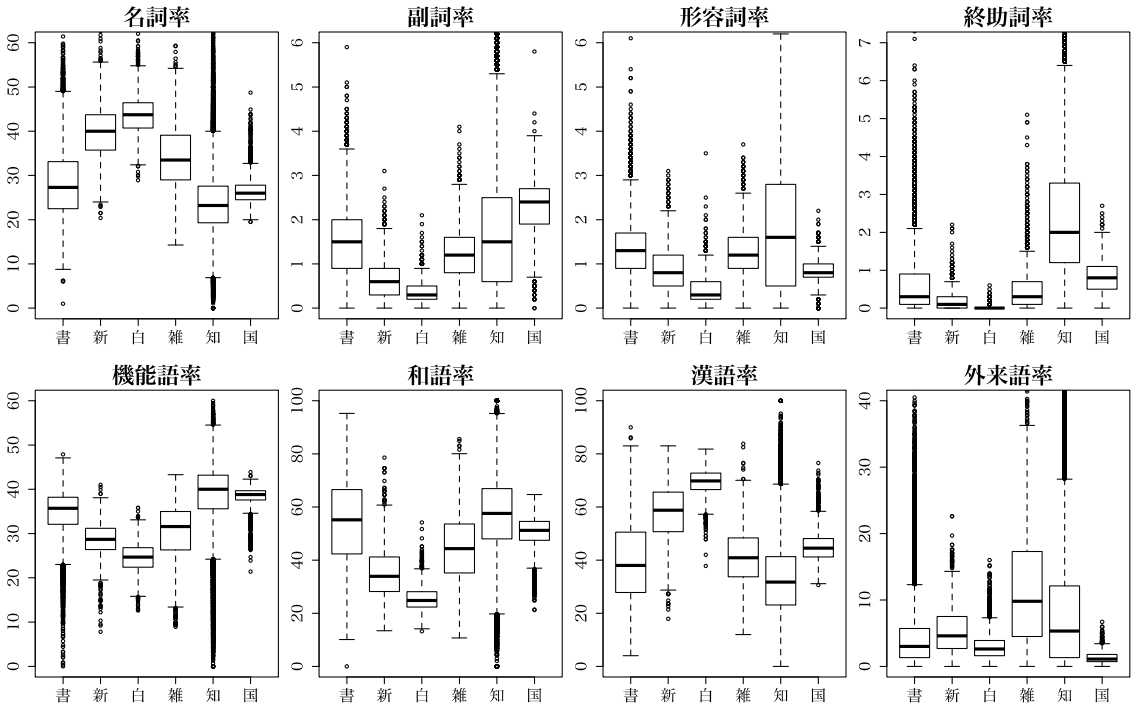

図 4　BCCWJ の品詞ごと・語種ごとの出現率（中央値と第 1・第 3 四分位数）
書：書籍，**新**：新聞，**白**：行政白書，**雑**：雑誌，**知**：Yahoo!知恵袋，**国**：国会会議録

■語種率：図 3 の CSJ の結果を見ると，漢語と名詞は「対話＜模擬講演＜学会講演」の順に多くなるのに対し，和語と機能語（助詞・助動詞）は逆の傾向を示している。こうした漢語率・名詞率と和語率・機能語率の関係は BCCWJ にも成立する。BCCWJ では，漢語や名詞は行政白書や新聞に，和語や機能語は書籍やインターネット上のテキスト，国会会議録に多く見られる。雑誌はその中間の傾向を示す。この傾向は小磯ほか(2009)とほぼ一致する。

一連の国語研究所の語彙調査や野元(1959)などから，書き言葉では和語よりも漢語が，話し言葉では逆に漢語よりも和語が多い傾向にあることが指摘されている。CSJ の各種レジスターや国会会議録，話し言葉に近い傾向を示す Web 上のテキスト(Yahoo!知恵袋)，また BCCWJ のうち小説の会話文などを含む書籍が高い和語率を示しており，上記指摘と整合的である。また丸山 (2005) は, CSJ の模擬講演を含む各種話し言葉の漢語率を比較しており，その中で，模擬講演の方が日常会話よりも漢語率が顕著に高い傾向を示すことから，敬体で改まった表現を用いる傾向の強い模擬講演のような独話では，日常会話よりも書き言葉により近い傾向を示すとしている。国語研究所(1955)でも，ニュース解説やニュースの方が日常談話よりも漢語率が高いとされる。図 3 の CSJ の結果を見ると，この傾向が顕著に観察されるのは学会講演である。学会講演では，漢語率が 4 割を越えており，新聞や白書よりは少ないものの，書籍や雑誌などの書き言葉と同じ水準となっている。国会会議録もやはり漢語率が 4 割以上であり，学会講演同様，改まりの程度の強い，書き言葉に類似した傾向を示している。また漢語の使用は硬い文体と，和語の使用は軟らかい文体と関連することが指摘されており（柏野ほか 2012），こうした各レジスターの硬軟の偏りも語種率に影響したものと考えられる。

■機能語率・名詞率：Halliday(1990)は，内容語率で定義される語彙密度という尺度を提案し，綿密に計画された，あるいはよりフォーマルな文章ほど語彙密度が高いとしている。機能語率の逆数が内容語の占める割合と考えるならば，対話よりも講演の方が，また講演の中でも模擬講演（主に個人的内容に関する一般の人によるスピーチ）よりも学会講演の方が，機能語率が低い（内容語率が高い）傾向を示しており，「対話＜模擬講演＜学会講演」の順に，より綿密に計画された，あるいはよりフォーマルなスタイルの発話であると言える。実際，学会講演では予稿集やスライドなどの発表資料を，また模擬講演では発話の流れを記したメモを準備しており，相手とのやりとりの中で発話内容を決める対話と比べて発話の計画性は高いと言える。また学会講演は，大人数の前で自身の主張を展開するものであり，2〜4 人程度の収録スタッフを前に個人的体験談などを語る模擬講演と比べ，よりフォーマルな発話であると言える。BCCWJ においても，小説などを含む書籍や Web 上のテキストよりも，行政白書や新聞の方が機能語率は低い（内容語率が高い）傾向を示しており，行政白書や新聞の方がよりフォーマルであるという直観と合致する。一方，国会会議録は，フォーマルで発話内容の計画性も高いと考えられるが，白書や新聞と比べ機能語率はかなり高い傾向を示している。国会会議録は CSJ の学会講演と同水準であることから，機能語率（内容語率）には，単に計画性やフォーマルさの程度だけでなく，話し言葉・書き言葉というモードの違いも関わる可能性がある。

また名詞率は，先述の通り機能語率と逆の傾向を示しているが，複雑な文ほど動詞群の名詞化により機能語に対する内容語の比率が高くなることから(Halliday 1985)，名詞率と内容語率（機能語率）は正（負）の相関を示すことになる。このことが上記結果につながったと考えられる。

■副詞率・形容詞率：国語研究所(1955)では，日常談話，ニュース解説，ニュースの副詞率が 6.1%，2.5%，1.3%，形容詞率が 2.7%，0.9%，0.4% と，主観的表現の多い日常談話の

副詞率，形容詞率が圧倒的に高いこと，また同じニュースでも，ある程度解説者の意見などを含むニュース解説の方がニュースよりも副詞率，形容詞率が高いことを示している。学会講演のように客観的表現の好まれるレジスターよりも，模擬講演（個人的体験談の語りなど）や対話のように主観的表現が多く含まれるレジスターの方が，副詞率，形容詞率ともに高い傾向を示しており，整合的な結果となっている。BCCWJ を見てみると，やはり客観的表現の好まれる行政白書や新聞では副詞率・形容詞率ともに低いのに対し，小説などを含む書籍では高い値を示している。

その一方で，客観的表現が好まれると予想される国会会議録において，形容詞率は確かに低いものの，副詞率については若干高い値となっている。形容詞率については，話し言葉のうち客観的表現が好まれる学会講演や国会会議録と，書き言葉で同じく客観的表現が好まれる新聞がほぼ同じ傾向を示していることから，話し言葉・書き言葉の区別なく，表現の客観性・主観性の観点がその出現に強く影響していると考えられる。一方，副詞については，書き言葉の各種レジスターよりも国会会議録は高い比率を示している。また，副詞率が最も低い行政白書と最も高い対話でその中央値が 0.3%と 6.1%となっており，形容詞の場合（0.3%と 2.4%）と比べて極端に開きがある。この傾向は，模擬講演や学会講演，国会会議録など，その他の話し言葉にも大なり小なり見られる。以上のことから，副詞については，表現の客観性・主観性に加え，話し言葉・書き言葉というモードの違いも影響している可能性が考えられる。

４．おわりに

BCCWJ との統一的な検索を目指し，CSJ の形態論情報のうち短単位情報を対象に，BCCWJ で採用されている UniDic 体系に変換する作業を実施した。2 節では，CSJ のオリジナル版短単位体系と UniDic 体系の主な相違点, および UniDic 体系への変換手続きなどについて解説した。また 3 節では，CSJ の品詞別・語種別の基礎統計量を示した上で，CSJ の各種レジスターの品詞・語種の特徴を，BCCWJ のレジスターとの比較を通して議論した。

CSJ の UniDic 版短単位情報は, 今年度中を目途に中納言検索システムを通して公開する。また，今回は短単位情報のみの公開に留まるが，今後，長単位情報についても同様に整備する予定である。

文　献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「コーパ日本語学のための言語資源－形態素解析用電子化辞書の開発とその応用－」『日本語科学』22,　pp.101-123

伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信（2008）「UniDic version1.3.9 ユーザーズマニュアル」http://chikusei.lv9.org/cms-z/zomeki-1.0.4/ext/morph/unidic/manual.pdf

Halliday, M.A.K.　（1985）*Spoken and Written Language,* Victoria: Deakin University

Halliday, M.A.K.　（1990）“Some grammatical problems in scientific English,” *Annual Review of Applied Linguistics*, 6, pp.13-37.

柏野和佳子・立花幸子・保田祥・飯田龍・丸山岳彦・奥村学・佐藤理史・徳永健伸・大塚裕子・佐渡島紗織・椿本弥生・沼田寛（2012）「書籍テキストへの文体情報付与の試み―『現代日本語書き言葉均衡コーパス』の収録書籍を対象に―」『第 2 回コーパス日本語学ワークショップ予稿集』pp.155-164

小磯花絵・小椋秀樹・小木曽智信・宮内佐夜香（2009）「コーパスに基づく多様なジャンルの文体比較－短単位情報に着目して－」『言語処理学会第 15 回年次大会発表論文集』pp.

594-597
国語研究所（1955）『談話語の実態』国立国語研究所報告 8，秀英出版
国語研究所（2006）『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』
丸山直子(2005)「話しことばにおける漢語」『東京女子大学比較文化研究所紀要』66，pp.27-38
丸山岳彦・高梨克也・内元清貴（2006）「節単位情報」『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』pp.255-322
野元菊雄（1959）「話しことばの中での漢語使用」『ことばの研究』国立国語研究所論集 1
小椋秀樹（2006）「形態論情報」『国立国語研究所報告 124：日本語話し言葉コーパスの構築法』pp.347-453
小椋秀樹（2008）「『日本語話し言葉コーパス』の言語単位」『日本語学』27 巻 5 号 pp.72-81
小椋秀樹・小磯花絵・冨士池優美・宮内左夜香・小西光・原裕（2011）国立国語研究所内部報告書『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上・下)』LR-CCG-10-05
内元清貴・高岡一馬・野畑周・山田篤・関根聡・井佐原均（2004）「『日本語話し言葉コーパス』への形態素情報付与」『第 3 回話し言葉の科学と工学ワークショップ講演予稿集』pp.39-46.
山口昌也・木村睦子・西川賢哉・石塚京子・小椋秀樹（2004a）「短単位辞書マニュアル」CSJ 同梱マニュアル　http://pj.ninjal.ac.jp/corpus_center/csj/manu-f/suwdic.pdf
山口昌也・木村睦子・西川賢哉・石塚京子・小椋秀樹（2004b）「短単位・長単位データマニュアル」CSJ 同梱マニュアル　http://pj.ninjal.ac.jp/corpus_center/csj/manu-f/wdb.pdf