大規模コーパスの利用とメタデータの役割

丸山 岳彦　（国立国語研究所 言語資源研究系）†

The Role of Metadata in the Analysis of Large-scale Corpora

Takehiko Maruyama　(Dept. Corpus Studies, NINJAL)

1　はじめに

大規模コーパスを言語分析に利用するためには、メタデータを参照することが欠かせない。また、コーパスを評価する際にも、メタデータの参照が必須となる。本稿では、大規模コーパスの利用・評価にとってメタデータがどのような役割を果たすかについて、『現代日本語書き言葉均衡コーパス(BCCWJ)』の「書誌情報データベース」を例に論じる。具体的には、メディア・ジャンル情報を利用したモダリティ形式の分析（4 節）、初出情報を利用した書籍サンプルの評価（5 節）を行なう。

2　メタデータの種類と役割

コーパスの本体となるテキストデータや音声データに対して、そのデータの出自（書誌情報、収録内容）、メディア、ジャンル、書き手・話し手の属性、社会的な位相などの情報を記録したデータを、「メタデータ」と呼ぶことにする。さまざまな種類のテキストデータ・音声データ（の転記テキスト）を検索して得られたコンコーダンスは、言語表現の断片の集積でしかないため、それぞれの言語表現が本来使われていた使用文脈や発話場面から切り離された状態にある。そこでメタデータを参照することにより、例えば、検索結果を書き手・話し手の性別によって分類したり、使用傾向の違いをジャンルごとに分析したりすることができる。大規模コーパスを用いて言語の使用実態を多様な観点から実証的に明らかにするためには、メタデータの存在が必須である。

ここでは、コーパスを構成する個別のサンプルの中身に対して付与（注釈付け）されたデータを「アノテーション情報」と呼び、メタデータと区別する。アノテーション情報は、文章・談話を構成する言語的階層（音素 < 語 < 文節 < 節 < 文 < 文章・談話）の各階層に対して付与される。これに対してメタデータは、サンプル全体に対して付与されるものとする。テキスト全体から自動的に算出される種々の統計値もメタデータの一種と考える。アノテーション情報とメタデータの関係の例を、図 1 に示す。これらはいずれも、「データのためのデータ（data about data)」として位置づけられる。

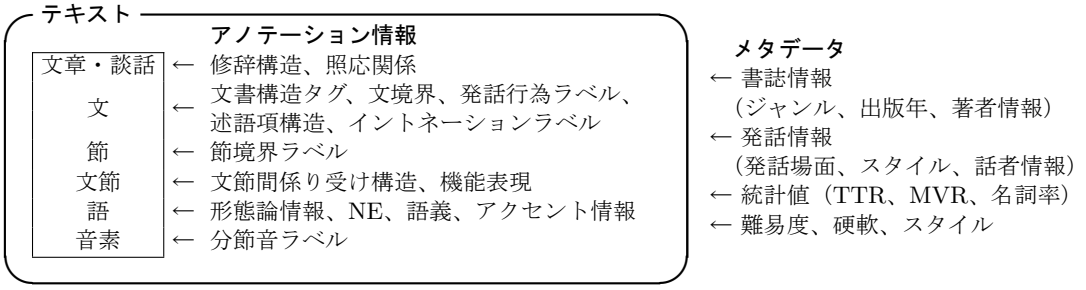


図 1: アノテーション情報とメタデータの例

Burnard (2004) によると、コーパスに付与されるメタデータは、図 2 のように分類できる。以下、本稿では、Burnard (2004) の分類に基づいて BCCWJ に付与されている「書誌情報データ」をメタデータとして整理し、それらを用いることで、どのように BCCWJ を利用または評価することができるかについて述べる。

† maruyama@ninjal.ac.jp

1. Corpus identification： コーパスの識別情報。コーパスの名称、作成者、配布元など。
2. Corpus derivation：サンプルの出自に関する情報。
   (a) Bibliographic description：書誌情報
   (b) Extent：サンプルサイズ
   (c) Languages：使用言語、言語コード
   (d) Classification：層別情報、カテゴリ
3. Corpus encoding：コーパスの編纂に関する情報
   (a) Project Goals：作成の目的
   (b) Sampling and extent：サンプリングの方法と結果
   (c) Editorial practice：編集方法（修正、追加、包摂など）
   (d) Markup scheme：マークアップの方法
   (e) Reference scheme：コーパスを参照するための情報（文番号、単語番号など）
   (f) Classification scheme：カテゴリの分類方法

図 2: コーパスに付与されるメタデータの分類例（Burnard, 2004）

3 BCCWJ の書誌情報データベース

BCCWJ は、「出版サブコーパス」「図書館サブコーパス」「特定目的サブコーパス」という 3 つのサブコーパスから構成される、約 1 億語の均衡コーパスである[1]。各サブコーパスには、「書籍」「雑誌」「新聞」「白書」「教科書」「広報紙」「Yahoo!知恵袋」「Yahoo!ブログ」「韻文」「法律」「国会会議録」から無作為抽出したテキストが、合計 172,675 サンプル収録されている。アノテーション情報として、2 種類の形態論情報（短単位、長単位）、および文書構造タグが付与され、XML 文書として構造化されている。これに加えて、メタデータとして、サンプルを取得した原本の書誌情報やジャンル情報、著者情報、サンプリングの状況などを記録した「書誌情報データベース」が同梱されている。サンプルごとに一意に付与された ID（「サンプル ID」）で関連付けることによって、コーパス本体と書誌情報データベースを結合することができるようになっている。

書誌情報データベースは、次の 4 つのデータから構成される。

書誌情報データ： サンプルを取得した原本やジャンルに関する情報。

サンプル情報データ： サンプルの ID や取得状況に関する情報。

記事情報データ： 記事に含まれる文章の初出および著者に関する情報。

人名録データ： サンプルの著者や著作権者などの人名録。

書誌情報データには、サンプルの原本に関する書誌情報、およびジャンルやカテゴリーに関する情報が記録されている。サンプル情報データには、サンプルごとに一意に付与された ID の他に、出版物（書籍、雑誌、新聞、白書、教科書）のサンプリングを実施した結果（サンプリングの基準となったページ数とページ内の座標）が記録されている。記事情報データは、書籍・雑誌・新聞のサンプルに含まれる「記事[2]」に関する初出情報、および実際に記事を執筆した著者に関する情報が記録されている。人名録データは、書誌情報や記事情報に現れる人名 ID に対応する人名を記録したデータである。4 つのテーブルを関連付けた結果を、図 3 に示す。

[1] 以下では、BCCWJ の全データを記録した「BCCWJ-DVD 版」を前提に話を進める。
[2] 「記事」とは、「同一著者によって、同一のテーマについてまとまりをもって書かれた文章の範囲」のことを指す。

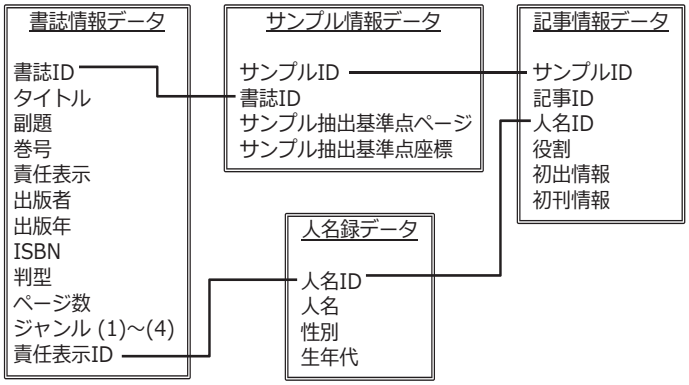


図 3: 書誌情報データベースの構成と関連付け

これらの書誌情報データベースによって提供される情報は、図 2 に示したメタデータのうち、2-(a) “Bibliographic description”、2-(d) “Classification”、および 3-(b) “Sampling and extent”（の一部）に相当する。 1 “Corpus identification” や、2-(b) “Extent”、2-(c) “Language”、および 3 “Corpus Encoding” については、「BCCWJ-DVD 版」に付属するマニュアル (国立国語研究所, 2011) にその中身が詳述してある。将来的には、これらの書誌情報データを、ダブリン・コアに基づくメタデータ記述の基本要素にどのように配置するかという問題を検討することも考えられるが、現時点ではその問題について触れる余裕がないので、ここでは擱く[3]。

BCCWJ を検索して用例を収集したりその出現数を集計したりするとき、分析者がまず考慮することは、どのような性質を持つテキストにどれだけの用例が出現したかという点であろう。多様なテキストの集合体である大規模コーパスを利用する場合、ジャンル、執筆年、執筆者など、収録されているテキストを切り分けるための分類指標ができるだけ多く付与されていることが望ましい。どのような出自や属性を持つデータ集合にどのような言語的特徴が認められるのか、という問いに的確に答えるためには、多角的な観点から詳細に記述されたメタデータの存在が欠かせない。

次節以降では、書誌情報データベースに記録された情報を用いて、BCCWJ をどのように利用・評価することができるかについて論じる。次の 4 節では、書誌情報データベースを用いて検索結果を分類する例として、モダリティ形式の分析例を示す。5 節では、書誌情報データベースに記録された「初出情報」をもとに、BCCWJ に収録された書籍のサンプルを評価する例を示す。

4 書誌情報データベースを用いた BCCWJ の検索と集計 —モダリティ形式の分析—

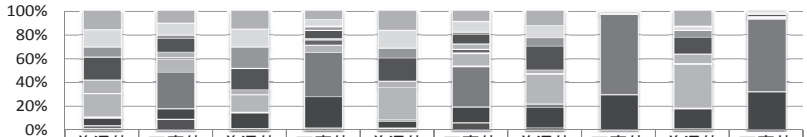
書誌情報データベースを用いて BCCWJ を検索・集計する例として、メディア[4]やジャンルの違いによって文末のモダリティ形式がどのような出現傾向を示すのかについて見てみよう。具体的には、ダロウ・ヨウダ・カモシレナイ・ラシイ・ミタイダという形式を取り上げる。これらは、事態に対する話し手の認識的なとらえ方を表す「認識のモダリティ」として扱われるが (日本語記述文法研究会, 2003) 、実際にそれらが現れやすい（または現れにくい）条件や使用場面についての記述はない。

そこで、「中納言[5]」を用いて、BCCWJ 全体を対象に検索を実施した。上に示した 5 種類のモダリティ形式と、それらを丁寧体にしたデショウ・ヨウデス・カモシレマセン・ラシイデス・ミタイデスという合計 10 通りの形式について、直後に句点が後接する事例を、短単位検索によって検索した。検索条件の例を (1) に挙げる。

[3] 千葉他 (2006) は、「青空文庫」で公開されている書誌情報をメタデータ化した研究である。
[4] サンプルを取得した媒体の種類（書籍、雑誌、白書、法律など）のことを、ここではメディアと呼ぶことにする。
[5] 中納言 RC2 (Released at 2011-11-10)、2012 年 1 月に検索を実施。

(1) (出現書字形 = "だろう" AND 品詞 LIKE "助動詞%") AND 後方共起: 出現書字形 = "。"
ON 1 WORDS FROM キー IN (subcorpusName="生産・書籍" AND core="true")
OR (subcorpusName="生産・書籍" AND core="false") WITH OPTIONS unit="1"
AND tglWords="20" AND tglKugiri="|" AND tglFixVariable="2"

5 種類・10 通りのモダリティ形式について出現数を集計し、各メディアごとに 100 万語あたりの出現数を求めた。結果を図 4 に示す。



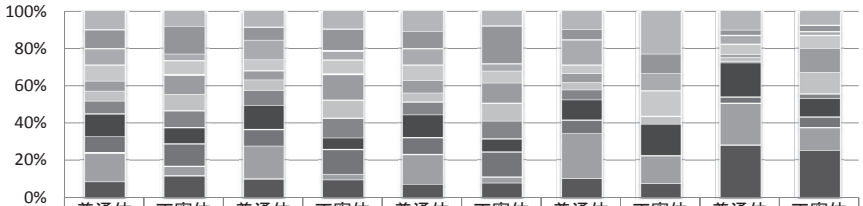
	普通体 (2,268)	丁寧体 (2,192)	普通体 (462)	丁寧体 (722)	普通体 (745)	丁寧体 (445)	普通体 (629)	丁寧体 (97)	普通体 (28)	丁寧体 (131)
	だろう。 (4,460)		ようだ。 (1,184)		かもしれない。 (1,190)		らしい。 (726)		みたいだ。 (159)	
出版書籍 (952)	359	227	70	50	122	41	78	1	4	1
雑誌 (875)	325	220	70	45	110	38	63	2	1	2
新聞 (461)	186	57	80	24	59	7	45	0	1	1
図書館書籍 (1,142)	450	245	86	47	149	36	123	1	4	2
白書 (9)	5	1	1	0	1	0	0	0	0	0
教科書 (455)	247	100	18	9	35	19	23	0	2	1
広報紙 (82)	2	31	1	31	1	15	1	0	0	0
ベストセラー (1,224)	444	244	68	46	202	48	160	0	10	2
Yahoo!知恵袋 (1,285)	16	675	1	268	8	154	18	65	0	80
Yahoo!ブログ (873)	151	187	63	187	43	59	106	29	5	42
韻文 (88)	56	8	4	4	4	0	12	0	0	0
法律 (0)	0	0	0	0	0	0	0	0	0	0
国会会議録 (272)	26	197	0	10	10	27	1	0	0	0

図 4: 100 万語あたりのモダリティ形式の出現数（メディア別）

5 種類のモダリティ形式の出現数は、メディアの違いに関わらず、ダロウが圧倒的に多い。各形式の普通体と丁寧体の違いを見てみると、広報紙、Yahoo!知恵袋、Yahoo!ブログ、国会会議録において、丁寧体が多く用いられていることが分かる。不特定多数の読み手にメッセージを発信したり、特定の聞き手に話しかけたりするスタイルの文体の中では、丁寧体が優先されるためと解釈できる。ラシイとミタイダの丁寧体は、Yahoo!知恵袋と Yahoo!ブログでのみ特に出現数が多いことから、一般人が書く文章に特徴的に現れる形式であると考えられる。法律や白書ではモダリティ形式の出現数が極端に少ないが、これは、法規範の羅列である法律や、国内外の情勢を客観的に記述する白書において、話し手の主観的な認識を表すモダリティ形式が現れにくいためであると解釈できる。

次に書籍（出版、図書館、ベストセラー）について、書誌情報データの「ジャンル (1)」列に記載されている「NDC（日本十進分類法）」によって、各モダリティ形式の分布がどのように異なるかを見てみよう。NDC とは書籍をその主題・内容に基づいて分類したコードであり、BCCWJ では国立国会図書館における NDC の分類に準拠している。NDC の第 1 次区分（10 種）ごとに、100 万語あたりに出現する各モダリティ形式の出現数を求めた。結果を、図 5 に示す。

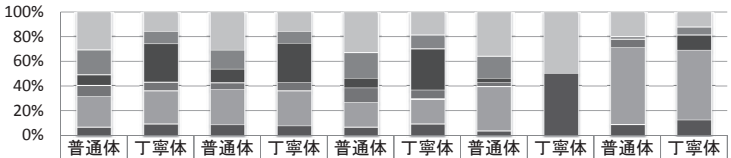
図 5 で普通体と丁寧体を比較すると、特に「文学」において、普通体の割合が丁寧体に比べて顕著に高い。これは、普通体で書かれた小説が「文学」の中に多く含まれるためであると考えられる。一方、「哲学」「自然科学」「工業」などの硬い内容を持つと思われる NDC で丁寧体の割合が高いという（一見すると意外な）結果が出ているが、これらの中には口語的な読み物のサンプルも一定数含まれており（『あなたを変える 3 つのレッスン（哲学）』『ぼくがすすめるがん治療（自然科学）』『電子レンジで朝ごはん（工業）』など）、丁寧体で書かれているこれらのサンプルにモダリティ形式が頻出すること（かつ、モダリティ形式が出現しにくい専門的な内容を持つサンプルは普通体で書かれていること）が原因であると考えられる。



	普通体 (3,271)	丁寧体 (2,696)	普通体 (787)	丁寧体 (594)	普通体 (1,105)	丁寧体 (488)	普通体 (700)	丁寧体 (5)	普通体 (54)	丁寧体 (15)
	だろう。 (5,967)		ようだ。 (1,382)		かもしれない。 (1,593)		らしい。 (705)		みたいだ。 (69)	
総記 (1,100)	396	271	82	68	143	47	82	1	7	1
哲学 (1,299)	392	468	66	83	122	117	49	1	2	1
歴史 (831)	337	115	95	34	114	21	111	1	3	0
社会科学 (857)	318	243	56	53	105	37	39	1	4	1
自然科学 (863)	215	321	42	96	88	63	34	0	1	2
工業 (766)	203	284	54	70	62	55	35	0	1	2
産業 (911)	273	294	76	76	92	55	44	0	0	0
芸術 (1,220)	469	279	121	42	161	41	92	1	12	2
言語 (1,149)	342	378	83	92	117	75	58	0	2	1
文学 (1,375)	589	162	161	21	206	19	200	1	14	2
分類なし (1,103)	328	368	93	67	94	46	84	0	18	4

図 5: 100 万語あたりのモダリティ形式の出現数（NDC 別）

最後に、同じく書籍について、書誌情報データの「ジャンル (3)」列に記載されている「C コード」を利用する。C コードは日本図書コードの一部で、「販売対象コード」、「発行形態コード」、「内容コード」から構成されている。ここでは、「販売対象」による分類（「一般（広く一般が対象）」「教養（知識階層が対象）」「実用（実務家が対象）」「専門（専門家学究者層が対象）」「児童（中学生以下の児童・生徒が対象）」）によって、各モダリティ形式の分布がどのように異なるかを見てみよう。販売対象ごとに、100 万語あたりに出現する各モダリティ形式の出現数を求めた。結果を、図 6 に示す。



	普通体 (1,627)	丁寧体 (1,567)	普通体 (404)	丁寧体 (319)	普通体 (515)	丁寧体 (226)	普通体 (368)	丁寧体 (2)	普通体 (45)	丁寧体 (15)
	だろう。 (3,194)		ようだ。 (723)		かもしれない。 (741)		らしい。 (370)		みたいだ。 (60)	
一般 (1,283)	499	248	126	51	170	43	133	1	9	2
教養 (783)	333	156	63	31	108	25	65	0	1	1
実用 (911)	143	493	45	101	39	75	13	0	0	2
専門 (386)	142	108	22	22	60	17	12	0	3	0
児童 (1,336)	403	412	115	90	104	45	131	0	28	9
その他 (389)	106	149	35	24	34	21	13	1	4	2

図 6: 100 万語あたりのモダリティ形式の出現数（C コード「販売対象」別）

普通体と丁寧体を比較すると、「実用」で丁寧体の割合が極めて高いことが分かる。「実用」の中には、『すぐ役立つ家庭の電気百科』『よくわかる会社更生法改正』のようなハウツー本が多く、読者に語りかける文体のテキストの中で、丁寧体のモダリティ形式が多用されている実態がうかがえる。また、ミタイダの出現数が、全カテゴリ中「児童」で際立って多い。ミタイダには、複数の辞書に「「ようだ」の口語的表現（新明解国語辞典 第六版）」「「ようだ」よりもくだけた言い方（例解新国語辞典 第八版）」のような記述があることから、そのような文体的特徴によって、文体のやわらかい児童書に多く出現しているという結果を解釈することができる。

以上で見たような、ある言語形式の出現傾向について、メディア、ジャンル、文体的な特徴などの

観点から定量的に分析するという方法論は、書誌情報データベースがあって初めて可能になることである。このような分析結果は、例えば辞書の編纂や日本語教育の現場において特に需要が高いと思われるが、従来の記述文法書の中では、このような視点からの記述は皆無であった。多種多様なテキストを収録した BCCWJ と、そのメタデータとしての書誌情報データベースを組み合わせて利用することによって、さまざまな位相における文法現象の分布について、定量的に記述することができる。今後は、どのようなメタデータを用いるとどのような言語形式の出現傾向を記述することができるか、その方法論と実践、そしてそれを可能にするメタデータの検討が求められると思われる。

5　書誌情報データベースを用いた BCCWJ の評価 ―初出情報の利用―

次に、本節では、書誌情報データベースの「記事情報データ」に記録された初出情報を利用して、書籍のサンプルを評価することについて述べる。

ある書籍に含まれる文章は、出版時において初めて世に発表されるものと、そうでないものとに分かれる。このうち前者は、一般的には「書き下ろし」と呼ばれる。一方、後者には、雑誌や新聞に連載されていた小説が単行本として出版される場合や、単行本が文庫として出版される場合などがある。中には、100 年以上前に出版された本が 2005 年に文庫として出版される例もある。

BCCWJ の書籍（出版、図書館、ベストセラー）には合計 22,058 サンプルが含まれているが、中には上記のような理由によって、古い時代に執筆されたテキストも収録されている（例えば、2005 年に出版された夏目漱石『吾輩は猫である』など）。無論、出版の実態を反映した結果であるので、これらの作品が収録されていることはサンプリングの結果としては正しい。しかしながら、『吾輩は猫である』のテキストに対して、「2005 年」という出版年だけでなく、初出に関する情報が付与されていることが、検索結果を利用する上では望ましい。また、あるコーパス（の部分集合）にどのようなテキストが収録されているかを評価する上でも、初出情報は重要である。

書誌情報データベースの「記事情報データ」は、このような問題意識によって作成されたメタデータである。各サンプルに含まれる「記事」を単位として、その文章がそれ以前に出版された経緯の有無を、原本の奥付や目録類などを参照して可能な限り調査した。そして、当該の記事が雑誌や新聞などで初めて発表・出版された年が判明した場合は「初出情報」として、当該の記事が初めて書籍として刊行された年が判明した場合は「初刊情報」として、それぞれ記録した。なお、初刊が確認できなかった場合や、書き下ろしであることが判明した場合は、出版年を初刊情報として記録した。

調査の結果、全 26,915 記事の 25.1%にあたる 6,755 記事から、初出・初刊に関する情報を取得することができた。初出・初刊のうち古い方の年（これを初出年と呼ぶことにする）が、出版年からどのくらい開いているかを集計した結果を、図 7 に示す。

初出年から 3 年程度は、新聞や雑誌での連載が単行本化されたり、単行本が文庫化されたりする場合が圧倒的に多い。初出年から 15 年で 100 記事以下に減少するが、個人全集の出版や名著の文庫化などにより、古い時代に書かれた文章が再度出版されるケースがロングテールで続いている。NDC 別に集計すると、初出年と出版年に開きがある書籍の数は、圧倒的に「文学」が多かった。

なお、BCCWJ に収録された書籍のサンプルのうち、初出年と出版年の開きが最大だったのは、福澤諭吉の『学問のすすめ』であった（1872 年初刊、2002 年出版、130 年の開き）。

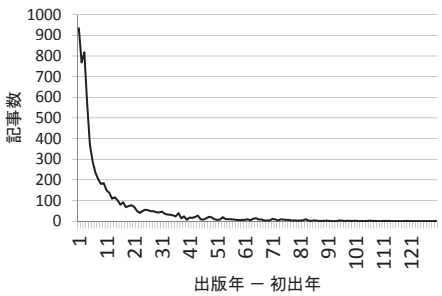


図 7: 書籍における出版年・初出年の開き

以下では、上記の初出情報を利用して、収録されているサンプルを語彙の面から特徴付けることを試みる。図 7 に示した初出年と出版年に開きのある記事のうち、50 年以上の開きがある記事は 281 記事あり、全記事の約 1%を占める。句読点等を含む語数は、合計で 890,314 語である。これに対して、2005 年に出版された書籍で、初出情報のない記事のうち、出版サブコーパスと図書館サブコーパスから 250 記事ずつを無作為に取得した。語数は、前者が 801,976 語、後者が 833,395 語である。

これで、50 年以上の開きがある記事のセットと、2005 年の書き下ろしと見なせる記事のセットを比較することができる。ここでは、2 つのセットのうちどちらかにしか出現しない語を、そのセットの「特徴語」と呼ぶことにする。2 つのセットそれぞれの特徴語（出現書字形）を品詞別に抽出したところ、表 1 のような結果を得た。上段は 50 年以上の開きがあるセット、下段は 2005 年の書き下ろしのセットから得た特徴語の例である。いずれも、出現数の多いものから順に挙げてある。

表 1: 2 つのセットにおける特徴語の例

名詞	まゝ、小吉、兵卒、細君、心持、氣、称呼、君子、攘夷、騾馬、等親、蠅、コサック、黒猫、空地、書生、燈籠、為め、牡丹、山猫、歌壇、爲、聲、義姉、世尊、高直、先き
	保育、ラッキー、メディア、補強、給付、データ、関数、福祉、再生、合金、データ、テレビ、細胞、原価、学習、コミュニケーション、介護、ベジタリアン、フード、暗号
形容詞	いゝ、なかつ、なかる、珍らしい、少い、危く、宜、蒼い、物凄い、新らしい、ゆる、旨く、少、よかつ、柔かい、まるい、よろしゅう、危い、巧い、明かるく、おそろしく
	すごく、美味しい、幼、きつい、やわらかく、温かく、ものすごく、やばい、上手く、厳し、きつい、ややこしく、長かっ、ややこしい、ひくい、嬉しかっ、しかたない
副詞	かう、忽ち、何う、先づ、たとい、悉く、已に、恰も、一ぱい、曾て、いろ、暫らく、少しく、兎に角、至極、すつかり、迚も、終に、頻り、唯、ちやんと、稍、矢張り、嘗て
	じっくり、そー、どー、ササ、わりと、良く、仲良く、ずばり、グスン、やっぱ、適宜、ずーっと、のろのろ、ちょくちょく、ぎくしゃく、ひくひく、とっとと、
動詞	思つ、行つ、言つ、考へ、言ふ、しまつ、云い、來、いつ、云わ、しまひ、起っ、依、をり、起る、向、帰つ、貰、於け、呉れ、御座い、出で、切つ、当る、あらう、言ひ
	超える、変える、分ける、抜き出し、捉える、受け入れる、取り付ける、取り組む、たらす、いかれ、贖い、注ぐ、摂る、図り、振りかぶり、斬り下ろす、欠かせ、囲う、ずれ
接続詞	或ひは、すなはち、只、然し、然も、然して、ち、尤も、もつとも、ぢゃ、乃ち、偖、んて、若くは、併し、して
	じゃぁ、ふんで、んじゃ

（上段：50 年以上の開きがあるセット、下段：2005 年書き下ろしのセット）

記事に対して付与された初出情報を利用することによって、書籍サンプルの中に見られる語彙の時代的な特徴を、表 1 のような形で把握することができる。BCCWJ 全体から見れば少数ではあるが、現代において出版されている書籍に含まれる語彙にどのような時代性が見られるかを、初出情報を参照することによって知ることができるわけである。

初出情報というメタデータの存在は、あるコーパス（の部分集合）に含まれるテキストがどのような性質を備えているかを評価する上で、重要な役割を果たすと思われる。図 7 で見たように、書籍とは部分的に再生産を繰り返す媒体であり、100 年以上前に出版された文章が現在でも出版され続けている。時間的に幅の広い日本語が混在している状態が書籍というメディアの性質であり、そこから無作為抽出した書籍サンプルの中身を把握する上で、初出情報の存在は欠かせないと言える。

6　おわりに

大規模コーパスの利用におけるメタデータの役割について、BCCWJ の書誌情報データベースを例に論じた。具体的には、メディア・ジャンル別に見られるモダリティ形式の分布、および初出情報を利用したテキストの評価と特徴語の抽出、という 2 点について示した。

今後のコーパス日本語学が取り得る方向性として、(1) 既存のコーパスを用いた言語学的分析、(2) 既存のコーパスに対するアノテーション情報・メタデータの付与、(3) 新規コーパスの設計・開発 という 3 つが考えられる。このうち (1) については、従来英語を中心に研究が進んできたコーパス言語学の方法論を現代日本語の諸側面に適用できるという点で、大きな進展が期待される。(2) については、これまでにも自然言語処理の分野においてさまざまなアノテーションが活発に行なわれてきているが、人文系の研究者にとっては使いにくい状況にあった。これに対して、国立国語研究所の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」が 2010 年度から開始されており、共通のデータ集合（BCCWJ のコアデータ）に対して集中的にアノテーションを実施する計画が進行中である。言語学的な分析とアノテーション情報の付与を関連させながら並行して進めることで、より効果的な検索方法や分類の基準、言語の使用実態に関する新たな知見の発見が期待できる。

この流れの中で、既存のコーパスに対する新たなメタデータの設計と付与もまた、今後の課題の一つとなるだろう。個々のサンプルに対する統計値のセットを設計して自動的に値を付与したり、テキストの難易度 (佐藤・柏野, 2012) や硬軟 (柏野他, 2012)、スタイルなどに関する情報を人手で付与したりすることが考えられる。そのようにして構築されたメタデータ群を利用して分析した結果は、メタデータを参照できないデータ（典型的には Web をクロールして得たメタデータのないデータ）を分析した結果よりも、はるかに信頼性と妥当性の高いものになると思われる。

謝辞：本研究は、国立国語研究所共同研究プロジェクト（基幹型）「コーパス日本語学の創成」、「コーパスアノテーションの基礎研究」によるものである。

参考文献

Burnard, L. (2004). Metadata for corpus work. In Wynne, M. (Ed.), *Developing Linguistic Corpora: a Guide to Good Practice*, pp. 30–46. Oxford: Oxbow Books.

千葉庄寿, 夷石寿賀子, 陳君慧 (2006).「「青空文庫」を言語コーパスとして使おう—メタデータ構築による歴史的・社会言語学的研究への応用の試み—」.『言語処理学会第 12 回年次大会発表論文集』, 915–918.

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 (2012).「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—」.本予稿集所収.

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』.（BCCWJ-DVD 版に収録）.

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.

佐藤理史, 柏野和佳子 (2012).「テキストの難易度に対する人間の判断と機械の判断」.本予稿集所収.