文字コードとタグによる漢字字体の記述

高田智和　間淵洋子　西部みちる　北村雅則　山口昌也
独立行政法人 国立国語研究所

１．はじめに

国立国語研究所では，現在，2006 年～2010 年を開発期間として「現代日本語書き言葉均衡コーパス ("Balanced Corpus of Contemporary Written Japanese" 以下，"BCCWJ"と略す)」の構築を進めている。BCCWJ は，1976 年から 2005 年までの 30 年間に出版された，幅広い日本語の書き言葉を収録対象とし，総語数約 1 億語を目指す，大規模バランストコーパスである。収録テキストには，文字情報，文書構造情報，形態論情報など，豊富な研究用付加情報が与えられ，日本語学・自然言語処理・辞書編集・国語施策等，さまざまな分野での利用に資するものとして期待されている[1]。

BCCWJ の収録対象期間である過去 30 年間には，当用漢字表の廃止と常用漢字表の制定（1981 年）や JIS 漢字の制定（1978 年）と 5 回の改訂（1983 年，1990 年，1997 年，2000 年，2004 年）など，社会での漢字字体の使用に影響を与える出来事があった。

BCCWJ では，これらの出来事が出版資料の文字・表記に及ぼした影響を観測できるように，文字・表記の実態を詳細に記述することを目的として，JIS X 0213:2004 に依拠して漢字字体を記述すると共に，文字コードで表現できない文字(外字)については，XML 形式による独自のタグを設けて記述を試みている。これに加えて，今回は XML の拡張性を生かし，文字コードで再現できない漢字字体の差異を記述するための新たなタグを設計した[2]。

本稿では，既に作成がほぼ終了した，BCCWJ 収録予定の白書データ約 500 万語を資料として，文字コードとタグによる漢字字体の記述が，それぞれ，どのような事象について有効性をもつのかを調査･確認する。その上で，調査結果に，先に挙げた常用漢字表の制定や JIS 漢字の改訂の影響を見出せることを指摘する。

２．資料概要

まず，今回資料として用いたデータの概要を示す。

BCCWJ は，(1)「生産実態サブコーパス」(2)「流通実態サブコーパス」(3)「非母集団サブコーパス」の 3 つのサブコーパスからなる。(1)，(2)は，母集団からのランダムサンプリングによって，統計的な代表性を確保しつつサンプルを取得する。一方(3)は，母集団を設定せずにサンプルを取得する。(1)(2)からは十分な量が得られないものの，書き言葉の実態を把握する上で重要な資料として，公共性の高い文章を一定の基準で選択し，収録するものである。この「非母集団サブコーパス」に収録されるものの一つが，今回資料とした，中央省庁刊行の白書である。

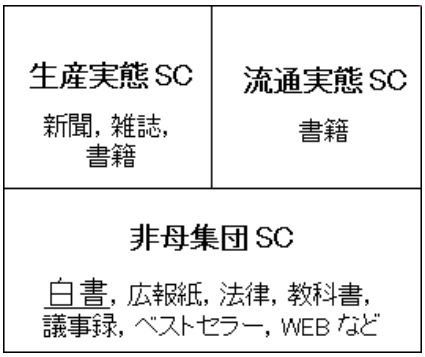


図 1：BCCWJ の構成

BCCWJ では，分析目的の違いを考慮し，長さを異にする以下の 2 種類のサンプルを用意している。

固定長サンプル：サンプル長を 1000 字に固定して取得するサンプル

可変長サンプル：記事，章などの論理的な構造を持つ文章の一まとまりを取得するサンプル

2 種のサンプルは，サンプル取得対象からランダムに選ばれる 1 文字（サンプル抽出基準点と呼ぶ）を基準として，同時に取得される。コーパス作成時は，この両者をいずれも包含する形でデータを構築しており，

[1] BCCWJ の設計については，山崎（2007）を，研究用付加情報と電子化形式については，間淵他（2006）を参照。
[2] 文字コードで再現できない漢字字体の差異を記述するための新たなタグは，漢字字体研究のための拡張タグであり，BCCWJ のタグの基本仕様には含まれていない。

公開時には，このデータから，それぞれのサンプルを抽出し提供する。

今回の調査には，固定長・可変長の各サンプルを抽出する前の，両サンプルが包含されたデータを用いる。本稿の以下の部分では，便宜的に，この作成時のデータを「サンプル」と呼ぶことにする。

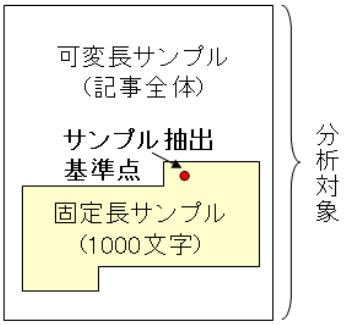


図 2：サンプルの形態

BCCWJ には，1500 サンプル，約 500 万語分の白書データが格納される。過去 30 年を 5 年刻みで 6 期に分け，各期に発行された白書から，データ量がほぼ均等になるよう，ランダムに 250 ずつサンプルを取得した。サンプリングの対象となった白書はさまざまで，環境白書，通商白書，科学技術白書，警察白書など，40 タイトル 1006 冊。コーパス収録対象となったのは，各期 120 冊前後，計 730 冊であった。今回調査対象とした文字に関するデータ量を，以下の表 1 に示す。

表 1：白書データの文字量

	文字数	1 サンプルあたりの文字数	対象冊数
第 1 期 (1976–1980 年)	約 150 万字	約 6000 文字	112
第 2 期 (1981–1985 年)	約 140 万字	約 5600 文字	113
第 3 期 (1986–1990 年)	約 140 万字	約 5600 文字	120
第 4 期 (1991–1995 年)	約 140 万字	約 5700 文字	121
第 5 期 (1996–2000 年)	約 130 万字	約 5300 文字	133
第 6 期 (2001–2005 年)	約 150 万字	約 5900 文字	131
合計	約 850 万字	約 5680 文字	730

3．白書データと JIS X 0213:2004

BCCWJ は，JIS X 0213:2004 を文字入力のための文字セットとしている。

情報交換用符号化文字集合として規定される JIS 規格は，1978 年に第 1・第 2 水準漢字と非漢字からなる約 6,800 文字が制定され（JIS X 0208），83 年，90 年，97 年の改訂を経て，2000 年には，第 3・第 4 水準漢字と非漢字約 4,000 字を拡張した（JIS X 0213）。2004 年には，国語審議会答申の「表外漢字字体表」で定める「印刷標準字体」を全面的に採用し，およそ 150 字種の字体を変更する改訂を行っている（「葛」は「葛」，「辻」は「辻」となる）。BCCWJ の文字処理で依拠する JIS 規格は，2004 年版の規格であり，第 1～第 4 水準漢字と非漢字からなる約 11,000 字の文字セットである。

JIS X 0213:2004 による白書データの符号化の状況を示すと，次の表 2 のようになる。

表 2：白書データの符号化

	異なり字数	延べ字数
第 1 水準漢字	2,407	3,965,586
第 2 水準漢字	231	1,065
第 3 水準漢字	15	56
第 4 水準漢字	1	1
X0208 非漢字	319	4,547,464
X0213 非漢字	107	12,287
外字	3	7
合計	3,083	8,514,179

JIS X 0213:2004 を用いることで，白書サンプルのほぼすべての文字が符号化される。また，第 3・第 4 水準漢字と X0213 非漢字の使用率，つまり，JIS 規格の拡張文字の使用の点では，漢字よりも非漢字の符号化に対して，JIS X 0213:2004 の有効性を見出すことができる。

4．文字コードによる漢字字体の記述

常用漢字の新旧字体など JIS X 0213:2004 に規定された異体字の組であれば，サンプルに出現した漢字字

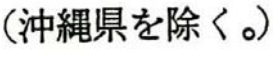
体をコーパスの中で再現することが可能である。

例えば,「縄―繩」の異体字の組は，1 面 38 区 76 点に「縄」，1 面 69 区 74 点に「繩」のように，JIS 漢字に両方の字体が登録されているので，資料に出現した字体に即して入力し分けることができる。以下に入力例を示す。

防災白書昭和 55 年版

→　沖縄気象台

防災白書昭和 57 年版



→　沖縄県を除く。

白書データにおいて,「縄―繩」の出現状況を期ごとに示すと次の表 3 のようになる。用例はいずれも「沖ナワ」である。

表 3：「沖ナワ」の字体

	1 期	2 期	3 期	4 期	5 期	6 期
縄	72	61	25	59	35	27
繩	19	0	0	0	0	0

第 1 期（1976-1980 年）では「縄」と「繩」の両方が出現するものの, 第 2 期 (1981-1985 年) 以降は「縄」のみが使われ,「繩」が出現しなくなる。第 2 期の始まりにあたる 1981 年は，常用漢字表が制定された年である。「縄」は常用漢字表に新たに追加された漢字であり，それ以前の当用漢字表には含まれていない。漢字使用の基準がないために，第 1 期の白書では康熙字典体（旧字体）の「繩」と簡略字体（新字体）の「縄」の双方が使われてゆれを生じ, 第 2 期以降の白書では, 常用漢字表で通用字体と示された「縄」に収束したものと考えられる。中央省庁の刊行物という白書の性格が使用される漢字字体の選択に反映された事例であると同時に，常用漢字表の影響が出版資料に現われた事例と見なされる。

５．タグによる漢字字体の記述

「縄―繩」のように文字コードで区別できる異体字の組であれば，JIS X 0213:2004 を運用することによって漢字字体を記述することができるが，「捗―  」のように文字コードで区別できない異体字の組であれば，漢字字体を記述するためには，文字コードによらない別の手段を講じる必要が生じる。

文字コードで区別できない異体字の組には，過去の JIS 規格の規格票例示字体の変更に関連する字種と重なるものがある。「捗―」もその一つである。1978 年の第 1 次規格では，康熙字典体の「」が例示され，1983 年の第 2 次規格（83JIS）では，常用漢字の新字体に準じた「拡張新字体」の「捗」に変更された。そして, 2004 年には,「印刷標準字体」を採用して「」に変更された。つまり，JIS 規格の例示字体は，1978 年に康熙字典体で始まり，1983 年から「拡張新字体」に変わり, 2004 年に再び康熙字典体に戻るという変遷を遂げている（図 3 参照）。

1978 年　→　1983 年　→　2004 年
康熙字典体　「拡張新字体」　康熙字典体

図 3：JIS 漢字の字体変更

文字規格の変更はコンピュータでの実装に反映され，コンピュータの実装字体（特に 83JIS の「拡張新字体」）が出版資料の文字・表記に影響を及ぼしていると言われている。しかし，このことを実証的に解明する調査はまだ行われておらず，近年のコンピュータ文字の影響を捉えることは，文字研究のテーマとして意義あるものと考えられる。

そこで，2004 年の JIS 規格改訂に際して,「拡張新字体」から康熙字典体（「印刷標準字体」）に変更されたおよそ 150 字種に限り，資料に出現した漢字字体が変更前後のどちらの字体であったのかをコーパスで区別して記述する試みとして, 研究用の拡張タグを設け, BCCWJ の白書データについて字体タグを施した。

jis2004：変更後の康熙字典体（「印刷標準字体」）
jis2000：変更前の「拡張新字体」

以下に字体タグの付与例を示す。

障害者白書平成 8 年版

→　進<jis2004>捗</jis2004>状況

障害者白書平成 15 年版

→　進<jis2000>捗</jis2000>状況

白書データにおいて字体タグを施した字種のうち，全期を通して出現する字種について，字体の分布を示すと右の表 4 のようになる。「葛—  」を除いて，第 6 期（2001-2005 年）には「拡張新字体」が用いられ，康熙字典体は出現しない。

表 4 にまとめた調査結果から，白書において俯瞰的には，康熙字典体と「拡張新字体」とでゆれがあった字種は「拡張新字体」に収束し，康熙字典体が主流であった字種は「拡張新字体」に移行している現象を指摘することができる。出版・印刷の実情に関する調査が今後の課題として残るものの，パソコンの普及と，それに伴う電子入稿の一般化から推測するに，第 6 期における「拡張新字体」への収束は，83JIS の影響が出版資料に現われた事例と解すべきであろう。

６．おわりに

BCCWJ は書籍部分の構築を進めており，書籍についても，文字コードと字体タグの併用によって，漢字字体の記述を行うことを予定している。

また，2004 年の JIS 規格の字体変更は，マイクロソフト社の次期 OS に反映され，ごく近い将来に社会での漢字字体の使用に影響が現われるものと予測される。「拡張新字体」から康熙字典体への移行が，白書や他の出版資料において行われる可能性があり，今後も継続して観察する必要があろう。

【参考文献】

山崎誠（2007）「現代日本語書き言葉均衡コーパス」の基本設計について．『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』．国立国語研究所.

間淵洋子他（2006）代表性を有する書き言葉コーパスの電子化フォーマットについて．『言語処理学会第 12 回年次大会発表論文集』．

表 4：字体変更の漢字

	1 期	2 期	3 期	4 期	5 期	6 期
茨	9	3	10	7	5	26
	7	6	19	32	7	0
葛	1	0	0	0	0	1
	3	3	5	3	2	2
揃	3	1	1	3	6	6
	18	0	3	6	2	0
遜	0	0	0	0	0	3
	2	2	1	1	1	0
捗	0	0	0	1	8	15
	12	18	13	15	13	0
賭	0	0	0	0	1	4
	10	5	4	5	8	0
灘	0	0	0	0	3	3
	8	1	5	4	1	0
逼	0	0	0	0	1	3
	3	5	3	4	1	0

《付記》

本研究は，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）による補助を得た。