国立国語研究所学術情報リポジトリ

『現代日本語書き言葉均衡コーパス』利用の手引 第1.1版

メタデータ	言語: Japanese 出版者: 国立国語研究所 公開日: 2021-04-23 キーワード: 現代日本語書き言葉均衡コーパス 作成者: 国立国語研究所　コーパス開発センター, National Institute for Japanese Language and Linguistics, Center for Corpus Development メールアドレス: 所属:
URL	https://doi.org/10.15084/00003228

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 International License.

『現代日本語書き言葉均衡コーパス』
利用の手引

第 1.1 版

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

2015 年 3 月

BCCWJ User's Manual

Version 1.1

© Center for Corpus Development

The National Institute for Japanese Language and Linguistics (NINJAL)

March 2015

目　次

第１章　『現代日本語書き言葉均衡コーパス』入門 1
1.1　はじめに 1
1.2　BCCWJ の特徴 1
1.2.1 均衡コーパス 1
1.2.2 形態論情報 3
1.2.3 その他のアノテーション 6
1.2.4 現代語 7
1.2.5 著作権処理 7
1.3　データの形式と内容 7
1.4　BCCWJ-DVD 版の意義 13
1.5　BCCWJ の参考文献 14
1.6　BCCWJ 構築の経緯 16
1.6.1 Version 1.0 の公開まで 16
1.6.2 Version 1.1 における修正 16
1.7　謝辞 17
付録：BCCWJ 開発メンバー 18
第２章　『現代日本語書き言葉均衡コーパス』の設計 20
2.1　はじめに 20
2.2　BCCWJ の設計 20
2.2.1 基本方針 20
2.2.2 基本概念の定義 21
2.2.3 BCCWJ の基本構成 21
2.2.4 BCCWJ の規模 22
2.2.5 各サブコーパスの特徴 22
2.2.6 コアデータ 23
2.3　サンプルの長さとタイプ 23
2.3.1 問題点 23
2.3.2 サンプルのタイプ 24
2.3.3 サンプルの重なり 24
2.3.4 レジスターとサンプルのタイプ 24
2.4　電子化 25
2.4.1 文字入力 25
2.4.2 タグの仕様 26

2.5　解析単位（短単位、長単位）……26
参考文献……27
第３章　サンプリング……28
3.1　BCCWJ 構築の基本理念……28
3.2　BCCWJ を構成する三つのサブコーパス……28
3.2.1 出版（生産実態）SC……29
3.2.2 図書館（流通実態）SC……29
3.2.3 特定目的 SC……29
3.3　BCCWJ を構成する 2 種類のサンプル……30
3.3.1 固定長（FIXED）サンプル……30
3.3.2 可変長（VARIABLE）サンプル……30
3.4　BCCWJ に収録するテキストの条件……30
3.5　BCCWJ-DVD 版に収録されているサンプルの一覧……31
3.6　サンプリング方法……31
3.6.1 出版 SC「書籍」……32
3.6.2 出版 SC「雑誌」……33
3.6.3 出版 SC「新聞」……34
3.6.4 図書館 SC「書籍」……35
3.6.5 特定目的 SC「白書」……36
3.6.6 特定目的 SC「教科書」……37
3.6.7 特定目的 SC「広報紙」……38
3.6.8 特定目的 SC「ベストセラー」……38
3.6.9 特定目的 SC「Yahoo!知恵袋」……39
3.6.10 特定目的 SC「Yahoo!ブログ」……40
3.6.11 特定目的 SC「韻文」……41
3.6.12 特定目的 SC「法律」……42
3.6.13 特定目的 SC「国会会議録」……43
参考文献……44
第４章　文書構造情報付き文字ベース XML（C-XML）……46
4.1　はじめに……46
4.2　文書構造タグセットの種類とサブコーパス・レジスターとの関係……46
4.3　可変長タグセット……47
4.4　固定長タグセット……51
4.5　Yahoo!知恵袋タグセット……51
4.6　その他のタグセット……51
4.7　文字入力仕様……52

4.7.1 基本方針 52
4.7.2 文字符号化方式と文字集合 53
4.7.3 包摂規準 53
4.7.4 外字 53
4.7.5 特殊表記 54
4.7.6 レイアウト 54
4.7.7 誤植 55
4.8 M-XML との相違点 55
参考文献 56
第 5 章 形態論情報 58
5.1 BCCWJ の言語単位 58
5.1.1 語彙調査の調査単位 58
5.1.2 BCCWJ の言語単位の設計方針 60
5.1.3 BCCWJ の言語単位 61
5.1.4 長単位・短単位の長所 62
5.1.5 形態素解析用辞書 UniDic について 64
5.2 長単位 66
5.2.1 文節認定規程 66
5.2.2 長単位認定規程 69
5.2.3 付加情報の概要 73
5.3 短単位 75
5.3.1 最小単位認定規程 75
5.3.2 短単位認定規程 78
5.3.3 付加情報の概要 83
5.4 CSJ からの変更点 87
5.5 終わりに 90
参考文献 91
付録 5-A: 複合辞（助詞相当句） 92
付録 5-B: 複合辞（助動詞相当句） 93
付録 5-C: 連語 94
付録 5-D: 接頭的要素 95
付録 5-E: 接尾的要素 96
第 6 章 形態論情報付きデータ（TSV） 100
6.1 形態論情報付きデータの概要 100
6.2 数字変換処理（NumTrans） 100
6.2.1 数字変換処理と 2 種類の本文 100

6.2.2 BCCWJ のバージョンと数字変換処理 102
6.2.3 数字変換処理と短単位・長単位の語数 102
6.3 総語数 103
6.4 TSV 形式データ 104
6.4.1 短単位 TSV のフィールド 104
6.4.2 長単位 TSV のフィールド 105
6.4.3 文字位置と連番 105
6.5 M-XML の形態論情報タグ 107
6.5.1 短単位タグ（SUW）の属性 108
6.5.2 長単位タグ（LUW）の属性 108
参考文献 109
第 7 章 書誌情報データベース 110
7.1 均衡コーパスにおける書誌情報の役割 110
7.2 書誌情報データベースの構成 110
7.3 「書誌情報データ」（Bibliography.txt） 111
7.3.1 「書誌情報データ」の概要 111
7.3.2 書誌 ID 113
7.3.3 タイトル 115
7.3.4 副題 115
7.3.5 巻号 115
7.3.6 責任表示 116
7.3.7 出版者 116
7.3.8 出版年 116
7.3.9 ISBN 117
7.3.10 判型 117
7.3.11 ページ数 117
7.3.12 ジャンル(1)～(4) 117
7.3.13 責任表示 ID 117
7.4 「サンプル情報データ」（Sample.txt） 118
7.4.1 「サンプル情報データ」の概要 118
7.4.2 サンプル ID 119
7.4.3 書誌 ID 124
7.4.4 サンプル抽出基準点ページ 124
7.4.5 サンプル抽出基準点座標 124
7.4.6 投稿日時 124
7.5 「人名録データ」（Directory.txt） 125

7.5.1 「人名録データ」の概要 125
7.5.2 人名 ID 125
7.5.3 人名 125
7.5.4 性別 125
7.5.5 生年代 125
7.6 記事情報データ（Article.txt） 126
7.6.1 「記事情報データ」の概要 126
7.6.2 サンプル ID 126
7.6.3 記事 ID 126
7.6.4 人名 ID 127
7.6.5 役割 127
7.6.6 初出情報 127
7.6.7 初刊情報 128
付録 7-A: 書誌情報データ「ジャンル」情報の詳細 129
付録 7-B: サンプル ID ベース書誌情報データの構成 145
第８章 文境界情報 146
8.1 はじめに 146
8.2 BCCWJ-DVD 版（Version 1.0） の文境界認定基準 146
8.2.1 文境界認定基準についての手がかり 146
8.2.2 BCCWJ-DVD 版（Version 1.0）における文境界認定基準の概要 147
8.3 BCCWJ-DVD 版（Version 1.1）における文境界認定基準 149
8.3.1 BCCWJ-DVD 版（Version 1.1）における文境界認定の作業方針 149
8.3.2 BCCWJ-DVD 版（Version 1.1）における文境界認定基準の詳細 150
8.3.2.1 基準の前提 150
8.3.2.2 処理 M(α)：修正率の高いパターン・認定基準 150
8.3.2.3 処理 M(β)：修正率の低いパターン・認定基準 154
8.3.3 BCCWJ-DVD 版（Version 1.1）における廃止事項 157
8.4 BCCWJ-DepPara における文境界認定 157
参考文献 158
第９章 形態論情報付き統合形式 XML（M-XML） 160
9.1 M-XML の概要 160
9.1.1 固定長と可変長の統合 160
9.1.2 異なる文書型定義の統合 161
9.2 要素の階層構造 161
9.3 C-XML と M-XML の相違点 162
9.3.1 数字変換（NumTrans タグ） 162

9.3.2 分数（fraction タグ） 163
9.3.3 ルビの処理 164
9.3.4 その他の追加されたタグ 165
参考文献 165
索 引 166

第 1 章　『現代日本語書き言葉均衡コーパス』入門

前川 喜久雄

1.1　はじめに

『現代日本語書き言葉均衡コーパス』（Balanced Corpus of Contemporary Written Japanese、以下 BCCWJ）は、国立国語研究所が中心となって開発した日本語に関する初めての大規模均衡コーパスである。2011 年 8 月以来、BCCWJ は 2 種類の検索インターフェースを用いて、オンライン公開されている。全文検索専用のインターフェースは『少納言』（http://www.kotonoha.gr.jp/shonagon/ ）、形態素解析済データ検索用のインターフェースは『中納言』（https://chunagon.ninjal.ac.jp/）と呼ばれている。

これにくわえて、2011 年 12 月にはデータ全体を DVD に記録して公開した。これを以下では BCCWJ-DVD 版（Version 1.0）と呼ぶ。BCCWJ-DVD 版（Version 1.0）はその後広く内外で利用されたが、公開後早い時期から文境界の認定に問題があることが指摘されていた。また数字を桁単位に形態素解析するために導入した NumTrans（第 6 章参照）の仕組みについても、かえってデータの使い勝手を阻害しているとの指摘があった。

今般、これらの問題を中心にその他若干の問題を解消した新データを公開し、これを BCCWJ-DVD 版（Version 1.1）と呼ぶことにする。本文書は BCCWJ-DVD 版のマニュアルである。Version1.1 を公開するにあたり、本文書にも必要な改訂をくわえたので、タイトルを「『現代日本語書き言葉均衡コーパス』利用の手引 第 1.1 版」に修正した。旧版（マニュアル第 1.0 版）と新版（同第 1.1 版）の主要な相違点は以下の 3 点である。

① 旧版では第 7 章を『中納言』の操作法にあてていたが、今回の改定に際して割愛した。『中納言』は毎年機能拡張を重ねて進化してきている。最新の操作法については『中納言』のオンラインマニュアルを参照していただきたい。
② 新版の第 8 章は新規に追加したもので、文境界の認定について BCCWJ-DVD 版（Version 1.0）から同（Version 1.1）への修正がどのように行われたかを説明している。
③ 旧版第 6 章では TSV データ（後述）と M-XML（Morphology-base XML）データ（後述）をまとめて解説したが、新版ではこれらを第 6 章（TSV）と第 9 章（M-XML）に分割した。

1.2　BCCWJ の特徴

1.2.1 均衡コーパス

BCCWJ は現代日本語の均衡コーパス（balanced corpus）である。現代日本語書き言葉

のできるだけ多くの変種をとりあげ、日本語の全体像を明らかにするための偏りのないサンプルを提供することを目標とした設計が施されている（第 2 章参照）。

BCCWJ は日本語に関する初の均衡コーパスであるが、その設計にあたっては、先行する諸外国の均衡コーパスを参考にしており、いくつかの点で先行コーパスに勝った設計がなされている。例えば、厳密な無作為抽出を可能なかぎり実施していること（第 3 章参照）、平均サンプル長を British National Corpus などに比べると短めに抑えることによって文献による語彙の偏りを低減していることなどである。

第 2 章および第 3 章で詳しく触れるが、BCCWJ は 3 個のサブコーパス、すなわち出版サブコーパス、図書館サブコーパス、特定目的サブコーパスから構成されている。

図 1-1 は、均衡コーパスが必要とされるひとつの事例を示している。この図は「食べ始める」「食べ続ける」のように用いられる補助動詞「～始める」「～続ける」が漢字を用いて表記される割合を BCCWJ のレジスター（register）（表 2-1 参照）ごとに示している。グラフ横軸に示されているレジスターについては 3.5 節以下参照。

最初に「～続ける」の結果を見ると、いずれのレジスターにおいても漢字表記率は 70%から 95%の水準にある。この場合、任意のレジスター、例えば新聞の分析によって得られた結論を他のレジスターに一般化することに大きな問題はない。

しかしながら「～始める」においては、レジスター間に顕著な差が存在している。そのため新聞データの分析から得られた結論は、雑誌・広報紙・教科書などのレジスターに対して一般化することができない。このような問題の存在は、均衡コーパスを分析することによって初めて知ることができるものである。

このようなレジスター間ないし語彙項目間の差は、あるいは何らかの一般的な要因に起因するものであり、従って予測可能であるかもしれない。しかし、そのような要因を発見するためにも均衡コーパスが必要とされるに違いない。

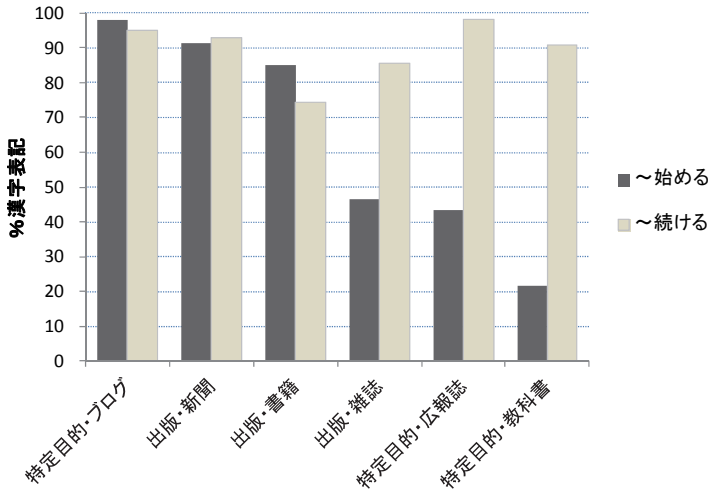


図 1-1: 補助動詞の漢字表記率のレジスター差（BCCWJ の解析結果）

もちろん BCCWJ にも種々の限界がある。例えば BCCWJ ではとりあげることのできなかった日本語書き言葉の重要なレジスターがある。その代表は漫画と広告である。これらのレジスターが現代日本語の動向（特にいわゆる新語の普及）に一定の影響を及ぼしていることは間違いない。しかし、画像情報への依存度が高いために他レジスターと同一の方法でのコーパス化が困難であること（この問題は雑誌サンプルの一部にも認められることがコーパスの構築過程で判明した)、および、著作権の処理に極度の困難が予想されることの二つの原因から、BCCWJ の収録対象とすることを断念した。

1.2.2 形態論情報

A. 短単位

BCCWJ にはアノテーションが施されている。最も重要なアノテーションは形態論情報、つまり文字列を語に分割して個々の語に品詞情報を付与した情報であろう。日本語のテキストは通常分かち書きされていないから、形態素解析されていないプレイン・テキストのデータから「国語」という文字列を単純に検索すると、目指す「国語」の他に「外国語」「韓国語」「中国語」「母国語」「自国語」等のごみが大量に生じてしまう。従来、日本語のコーパス言語学的分析では、正規表現を駆使して、プレイン・テキストから目指す文字列だけを得たり、後処理でごみを排除できることが研究者の基礎スキルとされてきたが、このような手法で常に目的を達することができるとは限らない。

正規表現を書くためにはあらかじめすべての表記上の可能性を把握しておかねばならないが、語によっては極端に異表記の多いものがある。例えば人名の「ヒロシ」には BCCWJ だけで 71 通りの表記がある。同じく、複合動詞の「ワキオコル」には終止形だけで 20 通り、活用形も含めると 324 通りの表記がある。これらの可能性をあらかじめ把握できる研究者は極めて稀であろう。

この問題は正しく形態素解析されたデータがあれば回避することができる。ただし日本語の形態論はいわゆる膠着語的な性格のために、「語」の規定に様々な問題が生じる。例えば「日本語」は全体でひとつの語とみてもよいが、「日本」と「語」の 2 語からなる複合語とみることもできる。

言うまでもないことだが、上記二つの解釈の間で言語分析上の絶対的な優劣を議論することには意味がない。重要なのは、どちらの解釈を採用するにしても、一旦ひとつの解釈を採用したら、その解釈の基礎となった言語学的観点を保持しながら、コーパス全体を分析できているかどうかである。

この点で従来の日本語形態素解析用辞書にはかなり深刻な問題が認められる。例えば ChaSen legacy の標準辞書として広く利用されている IPADIC では「国立国会図書館」は「国立＋国会図書館」の 2 語に分析されるが、「国立科学博物館」は「国立」「科学」「博物館」の 3 語に分析される。また「国立歴史民俗博物館」は「国立＋歴史民俗博物館」と 2

語に分析されるが、接尾辞「～学」を追加した「国立歴史民俗学博物館」は「国立＋歴史＋民俗＋学＋博物館」と 5 語に分析されてしまう。

言語学的な研究の基礎資料として用いるには、語認定におけるこのような一貫性の欠落は何としても回避したいところである。BCCWJ では上掲の例は以下のように分析される。（接）はその語の品詞が接尾辞であることを示しており、他はすべて名詞である。形態論的に一貫した言語単位が認定されていることがわかる。

国立国会図書館　　⇒　国立＋国会＋図書＋館（接）
国立科学博物館　　⇒　国立＋科学＋博物＋館（接）
国立歴史民俗博物館　⇒　国立＋歴史＋民俗＋博物＋館（接）
国立歴史民俗学博物館 ⇒　国立＋歴史＋民俗＋学（接）＋博物＋館（接）

BCCWJ が採用している形態論上のこの単位をわれわれは短単位（short unit word）と呼んでいる。短単位の認定基準については第 5 章参照。

B.　長単位（二重形態素解析）

短単位で一貫した言語単位を検索できるようになったのはよいが、短単位には副作用もある。上述の『中納言』を利用して、短単位の語彙素「ヒンディー語」を含む用例を検索すると 1 件もヒットしない。それならばと「中国語」を検索してみても同様である。これらの「語」は短単位としては「ヒンディー」と「語」、「中国」と「語」の 2 単位連鎖に分析されるからである。事実、これら 2 単位の共起関係を指定して検索すると、前者には 16 個、後者には 901 個の用例が見つかる。

しかし、これらの頻繁に利用される複合語を直接検索できないのは不便である。そこで BCCWJ には、主に複合語を把握する目的のために、長単位（long unit word）と呼ばれる単位に基づいた解析も施してある[1]。表 1-1 は同一のテキスト「公害紛争処理法における公害紛争処理の手続は」が短単位と長単位で、それぞれどのように解析されるかを比較したものである。

長単位の認定手順は第 5 章に詳しく説明されているのでここでは省略に従うが、結果として認定された長単位には以下のような特徴が認められる。

①複数の短単位から構成されている長単位には、「公害紛争処理法」のような実質語だけでなく、「における」のような機能語（いわゆる複合辞）がある。
②日本語のいわゆる膠着語的な性格を反映して「公害紛争処理」と「公害紛争処理法」がともに長単位として認定されている。BCCWJ を検索すると、さらに「公害紛争」「公害紛争処理制度」「公害紛争事件」「公害紛争処理機関」「公害紛争処理情報」等々

[1] 短単位と長単位による二重形態素解析は『日本語話し言葉コーパス』において最初に実施された。『日本語話し言葉コーパス』における短単位・長単位の定義と『現代日本語書き言葉均衡コーパス』における短単位・長単位の定義には、外来語の扱いなどに若干の相違があるが、大部分は一致している。

が長単位に認定されていることがわかる。

③長単位解析の結果は、短単位解析同様、解析対象テキストがもれなく長単位に分割されるという制約に従っている。そのため、いわゆる複合語（複合辞）だけが長単位に認定されるのではなく、短単位が単独で長単位に認定されることがある。表 1-1 の場合、最後の 3 行がこれに該当している。

表 1-1: 短単位と長単位の比較

短単位	短単位品詞	長単位	長単位品詞
公害	名詞-普通名詞-一般	公害紛争処理法	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
法	名詞-普通名詞-一般		
に	助詞-格助詞	における	助詞-格助詞
おけ	動詞-一般		
る	助動詞		
公害	名詞-普通名詞-一般	公害紛争処理	名詞-普通名詞-一般
紛争	名詞-普通名詞-サ変可能		
処理	名詞-普通名詞-サ変可能		
の	助詞-格助詞	の	助詞-格助詞
手続	名詞-普通名詞-サ変可能	手続	名詞-普通名詞-一般
は	助詞-係助詞	は	助詞-係助詞

短単位・長単位の認定基準を正確に理解するのは容易でないが、ユーザーは『中納言』の文字列検索機能を利用することで、検索したい文字列の単位構成についての知識を得ることができる。例えば「サーモンピンク色」が短単位としてどのように解析されるかを知りたければ、この文字列を文字列検索する際に、「結果表示単位」として「短単位」を指定すればよい[2]。検索結果の文字列には単位境界を示す縦線が挿入されて、以下のように表示される。

｜濃い｜サーモン｜ピンク｜色｜に｜なる｜。

また結果表示単位として「長単位」を指定した場合の表示は、

｜濃い｜サーモンピンク色｜に｜なる｜。

となるので、「サーモンピンク色」全体が 1 個の長単位として解析されていることがわかる。

C. 解析誤り

最後に、形態論情報について最も重要な情報に触れておく。形態論情報には解析誤りが含まれている。BCCWJ 全体の精度は 98%、コアデータ（第 2 章参照）に限れば 99%以上である。これは現在の形態素解析技術の最高水準を示す数字ではあるが、コアデータでも平均して 100 語に 1 語程度は誤りがあることになる。

解析誤りには、品詞を分類し間違えているもの、品詞は正解だが語彙素の細分類が誤っ

2 詳しくは『中納言』のオンラインマニュアル参照。

ているものなど、様々なタイプがある。もっとも深刻なのは、短単位境界そのものを分割し間違っている場合である。この場合、解析誤りが連続して出現することがあるので、注意が必要である。表 1-2 に解析誤りの例をいくつか示す。前文脈、後文脈中の縦線（｜）は短単位境界である。

表 1-2: 解析誤りの例

No	前文脈	キー	後文脈	語彙素読み	語彙素	品詞
(1)	｜ここ｜ん｜とこ｜、｜窮屈｜な｜こと｜ばかし｜で｜さ｜、｜	いやん	｜なっ｜ちゃう｜ったら｜ありゃ｜し｜ない｜...｜	イヤ	嫌	形状詞−一般
(2)	｜彼女｜は｜目｜を｜三角｜に｜し｜て｜部屋｜の｜中｜を｜	歩き	｜(まわっ)｜た｜。｜ルーク｜に｜この｜お｜礼｜は｜たっぷり｜し｜て｜あげる｜わ｜。	アルク	歩く	動詞−一般
(3)	｜奇妙｜な｜ほど｜	宮崎	｜(作品)｜に｜は｜家族｜、｜とりわけ｜親子｜関係｜の｜描写｜が｜避け｜られ｜て｜いる｜。	ミヤザキ	ミヤザキ	名詞−固有名詞−地名−一般

(1)は助動詞「に」の口語的な音便形を誤解析した例であり、短単位境界の認定誤りも生じている。(2)はいわゆる理論依存的な誤解析の例である。BCCWJ では「歩きまわる」全体が 1 個の短単位に分析されなければならないのだが（第 5 章参照）、このサンプルでは「まわる」が「歩く」から切り離されて 1 個の短単位に分析されている。(3)は短単位境界も語彙素の読みも正解だが、品詞分析で人名を地名に誤った例である。

誤解析の原因には様々なものがありうるが、BCCWJ の形態素解析では、コアデータを学習用コーパスとして解析器の機械学習を行っているので、学習用コーパスでカバーされていない語形の変異や品詞の細分類には対応が困難である。上例も学習用コーパスの限界による可能性が高い。

1.2.3 その他のアノテーション

形態論情報の他に、BCCWJ では文書構造と文字に関するアノテーションも提供されている（第 4 章参照）。これらは談話の研究や表記の研究に有益であると考えて施したアノテーションである。『中納言』では検索できないので、これらのアノテーションを利用するには BCCWJ-DVD 版が必要である。

また BCCWJ のサンプルには詳細な書誌情報が提供されている（第 7 章参照）。書誌情報はいわゆるメタ情報であり、言語の社会的側面の研究のために提供する情報である。書誌情報の一部は『中納言』の検索結果に表示されているが、『中納言』では書誌情報を検索条件に含めることはできない。書誌情報をキーとした検索を行うためには BCCWJ-DVD 版が必要である。

1.2.4 現代語

BCCWJ は現代語のコーパスであるが、ブラウンコーパスのように、或る特定の 1 年をきりとる形でデータを収集しているわけではない。一定の時間幅をもったサンプルが収録されており、その時間幅はサブコーパスないしレジスターによって変動している（表 3-1 参照）。

出版サブコーパスでは 2001 年から 2005 年までの 5 年間の幅であるが、図書館サブコーパスでは、これが 1986 年から 2005 年までの 20 年間に広がっている。特定目的サブコーパスに収められた種々のレジスター間にも相違があり、白書は 1976 年から 2005 年までの 30 年間をカバーしているのに対して、広報紙は 2008 年 1 年間だけである。すべてのレジスターが同一の時間幅をもっていることが望ましいのは言うまでもないが、実際にはデータの入手可能性が様々に異なることから、散らばりが生じている（第 2、3 章参照）。

1.2.5 著作権処理

コーパスの要件のひとつは、有償・無償を問わず、それが公開されていて誰でも利用できることである。そのためには、現代語コーパスの場合、著作権処理が必要になる。BCCWJ でもサンプルの性格に応じた著作権処理を実施した。

法律にはもともと著作権が存在しない。著作権が放棄されているテキスト（国会会議録と白書の一部）は、管理者にあたって著作権が放棄されていることを確認した。法人が著作権を有するテキスト（新聞記事、白書の大部分、雑誌記事の一部、広報紙等）は当該法人と交渉して許諾をもらった。

著作権の所属が明瞭でないテキスト（インターネット掲示板やブログ）の場合は、プロバイダ（Yahoo! Japan）の協力を得て、研究目的でデータを外部提供する可能性をネット上で告知した上で、告知の翌日以降に書き込まれたデータを提供してもらった。

個人の著作物のうち、権利者が日本文藝家協会等の著作権管理団体に所属しているものについては、管理団体の協力を得て、権利者に連絡をとることができた。しかし、例えば書籍の場合、このような方法で接触できる著者は全体の 2 割強であり、大部分のサンプルについては権利者の連絡先から調査を始める必要があった。

著作権データベース、各種紳士録、インターネット検索等で連絡先が判明することもあるが、それは例外的であり、多くの場合、連絡先を把握できない。その場合は、出版社に連絡をとって権利者への連絡を依頼するなどの方法で、多数の権利者と接触し、無償での利用を依頼した。

1.3 データの形式と内容

BCCWJ-DVD 版では、ユーザーの利便性に配慮して、サンプリングした言語データをさまざまな形式で提供している。Version 1.1 において提供されているデータは表 1-3 のとおりである。

NumTrans 非適用のデータは、第 6 章と第 9 章で説明するように Version 1.1 で新規追加されたデータである。最後の列（ディスク）に示したのは Version 1.1 を構成する 4 枚の DVD のうちどれにデータが保管されているかの情報である。さらに、この表には示していないが、書誌情報データとドキュメント類が Disc 1 に保存されている（1.5 節参照）。

表 1-3: BCCWJ-DVD 版（Version 1.1）に含まれる文書形式とデータの内容

文書形式	NumTrans	サンプル長	形態論情報	文書構造情報	ディスク
TSV	適用	統合	有	無†	Disc 2
	非適用	統合	有	無†	Disc 4
M-XML	適用	統合	有	有††	Disc 1
	非適用	統合	有	有††	Disc 3
C-XML	非適用	固定	無	有	Disc 1
	非適用	可変	無	有	Disc 1

† 文頭位置の情報（文頭ラベル）は提供されている（第 6 章参照）
†† C-XML（Character-base XML）の文書構造情報とは部分的に異なる（第 9 章参照）

(1) **TSV 形式と XML 形式**：データをタブ区切りテキストファイル（TSV）形式で提供しているか、タグ付き XML 文書として公開しているかの別である。TSV データは形態論情報を表形式で公開する目的に利用されており、短単位と長単位の情報は別のファイルに格納されている。XML 文書には 2 種類の別がある（下記(3) 参照）。

(2) **NumTrans 版と非 NumTrans 版**：「１９９９年」のように数字を含んだテキストを形態素解析するために事前に「千九百九十九年」のように形態素解析しやすい形にテキストを加工しているか（NumTrans 版）、していないか（非 NumTrans 版）の別である（第 9 章参照）。Version1.0 では NumTrans 版だけが公開されていたが、今回、非 NumTrans 版も追加公開する。NumTrans 版と非 NumTrans 版では、数字部分の短単位語数も形態論情報も異なることに注意が必要である。

(3) **C-XML 形式と M-XML 形式**：文書構造の情報だけを構造化したのが文書構造情報付き文字ベース XML（C-XML）である（第 4 章参照）。C-XML には後述する固定長（FIXED）サンプルと可変長（VARIABLE）サンプルの区別がある。形態論情報付き統合形式 XML（M-XML）は形態論情報を構造化したものであり、あわせて C-XML に含まれる文書構造情報の一部も構造化している（第 9 章参照）。

(4) **サンプル長**：BCCWJ のサンプルには固定長サンプル（1,000 字固定）と可変長サンプル（長さは様々。1 万字以内）がある。そしてレジスターによって、固定長と可変長の両サンプルを持つものと可変長サンプルだけのものとがある（第 2、3 章参照）。C-XML ではこれら両方のサンプルを別々に XML 化しているが（第 4 章）、一方、M-XML では、固定長と可変長を統合して重複部分を省いた統合形式サンプルに対して XML 化を

施している（第 9 章参照）。

(5) **コアデータと非コアデータ**：約 100 万短単位からなるコアデータに含まれるサンプル（コアサンプル）は、それ以外（非コアサンプル）に比べて形態論情報の解析精度が高い（第 5 章参照）。

(6) **書誌情報**：サンプルの書誌情報に関するメタデータを TSV 形式で提供している（第 7 章参照）。

(7) **文字符号化方式**:BCCWJ のすべての文書は文字符号化方式として UTF-8（BOM なし）を採用している。

図 1-2A-D に、BCCWJ-DVD 版（Version 1.1）の 4 枚のディスクのディレクトリ構成を示す。Disc 1（図 1-2A 参照）のルートディレクトリには 4 個のディレクトリがある。DOC ディレクトリ直下には、書誌情報データと著作権注釈情報データが格納されている。また DOC ディレクトリ下の MANUAL ディレクトリには、本文書、BCCWJ 構築時に蓄積したマニュアル類、さらに BCCWJ 公開後に出版された論文が格納されている。

書誌情報データについては第 7 章に詳しい説明がある。著作権注釈情報データは、権利者との交渉過程で、利用許諾に際して表示することを要請された注釈情報である。この情報は『中納言』でも当該サンプルがヒットした場合には表示される仕組みになっている。

C-XML には、文書構造タグ（第 4 章参照）を付したサンプルの XML データが、固定長（FIXED）と可変長（VARIABLE）に分かれて格納されている。

M-XML_NT （NumTrans）には、形態論情報付き統合形式 XML（M-XML）データ（第 9 章参照）が格納されている。この文書には固定長・可変長の区別はない。

C-XML 下の FIXED と VARIABLE および M-XML_NT の 3 ディレクトリの直下には、レジスターに対応するディレクトリがあり、各レジスターに属するサンプルが ZIP 圧縮されている（圧縮の方式については後述）。FIXED 直下のディレクトリは 書籍（PB）、雑誌（PM）、新聞（PN）、図書館 SC（LB）、白書（OW）の 5 個だけであるが、VARIABLE と M-XML_NT ディレクトリ直下には 13 個のディレクトリが存在する。CORE_NT ディレクトリについてはすぐ後で触れる。

Disc 2 は NumTrans 版の TSV データを格納している（図 1-2B 参照）。短単位（TSV_SUW_NT）、長単位（TSV_LUW_NT）の各ディレクトリ直下に、Disc 1 と同様に 13 のレジスターごとに圧縮されたデータが格納されている。

Disc 1 の CORE_NT ディレクトリには、BCCWJ コア（第 2 章参照）の対象となったサンプルの M-XML データ の NumTrans 版と TSV データの NumTrans 版（短単位と長単位）が格納されている。これはコアだけを処理したいユーザーの便宜を図ったものであり、このディレクトリのデータはすべて、Disc 1 の M-XML_NT、Disc 2 の TSV_SUW_NT、TSV_LUW_NT と重複して格納されている。

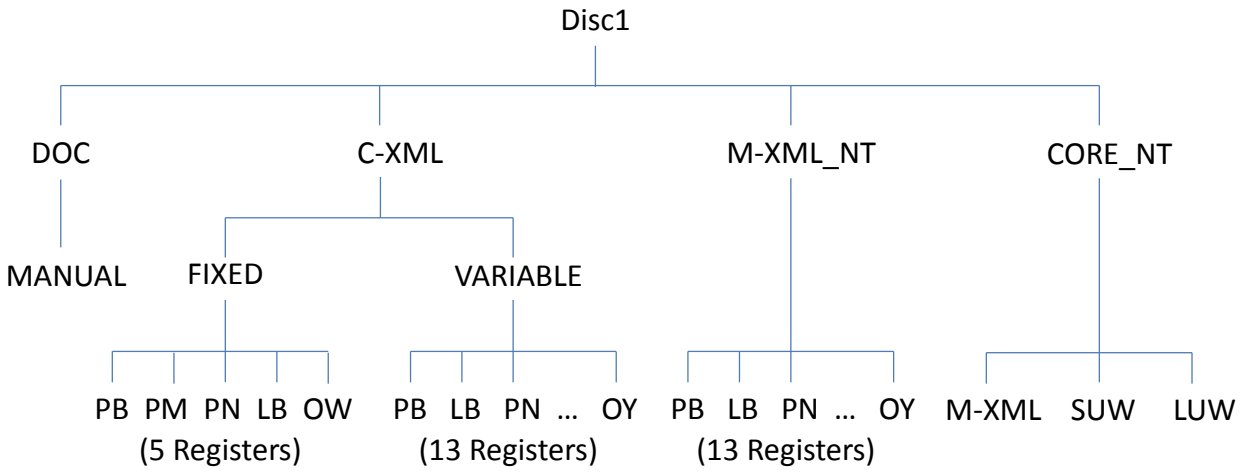


図 1-2A: BCCWJ-DVD 版（Version 1.1） Disc 1 のディレクトリ構成

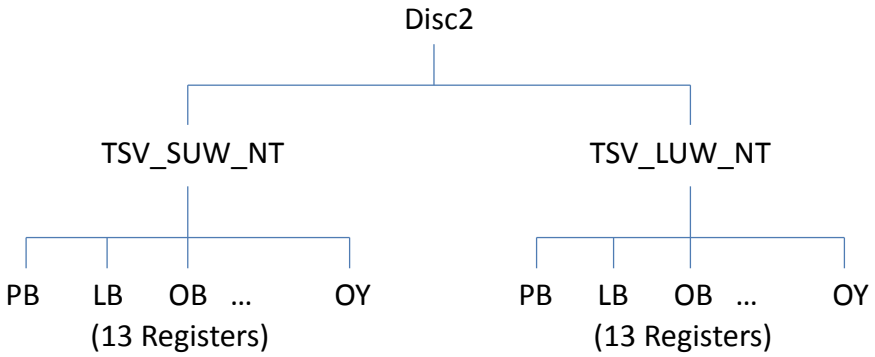


図 1-2B: BCCWJ-DVD （Version 1.1） Disc 2 のディレクトリ構成

Disc 3 と Disc 4 は、Version1.1 で新規に公開する非 NumTrans 版データを格納している。Disc 3（図 1-2C 参照）の M-XML_OT ディレクトリには M-XML の非 NumTrans 版が格納されており、CORE_OT ディレクトリには BCCWJ コアデータに含まれるサンプルの M-XML データと TSV データの非 NumTrans 版が格納されている[3]。前者は Disc 3 の M-XML_OT ディレクトリ内文書と、後者は後述する Disc 4 の TSV_SUW_OT、TSV_LUW_OT ディレクトリ内のデータと重複して格納されている。

最後に Disc 4 は、非 NumTrans 版の TSV データを保管している（図 1-2D 参照）。ディレクトリ構造は Disc 2 に準じている。

[3] ディレクトリ名に含まれる OT は original text の意味である。

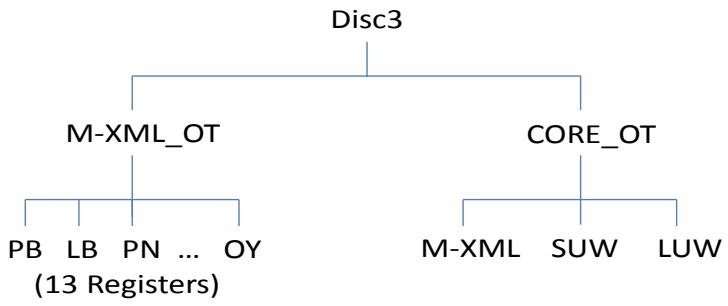


図 1-2C: BCCWJ-DVD 版（Version 1.1） Disc 3 のディレクトリ構成

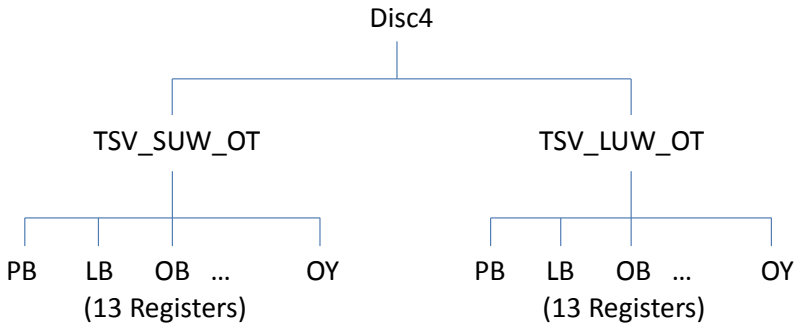


図 1-2D: BCCWJ-DVD 版（Version 1.1） Disc 4 のディレクトリ構成

これらのディスク中の圧縮ファイルを解凍すると、データサイズは数倍に増加するので、解凍時にはハードディスクに十分な残量を確保しておく必要がある。解凍前後でのデータサイズの変化を表 1-4A、B にまとめた。表 1-4A は XML 文書類の場合、表 1-4B は TSV データの場合をまとめており、表中の「前」「後」は「解凍前」「解凍後」の意味である。

PB（書籍）、LB（図書館 SC）、OC（Yahoo!知恵袋）、OY（Yahoo!ブログ）はファイル数、データ量が過大なので、圧縮に工夫を凝らしている。Disc 1 では、これらのディレクトリの圧縮ファイルを解凍すると複数のサブディレクトリに分けてファイルが格納される仕様になっている（表 1-4A、B でこれらのディレクトリの「後」はサブディレクトリ群を合計した値を示している）。

Disc 2、Disc 4 では、これらのディレクトリの圧縮ファイルを解凍すると TSV データが現れる。大部分のレジスターでは、そのレジスターの全データを含む 1 個のファイルが現れるだけであるが、LB と PB に関しては、TSV_SUW_NT、TSV_SUW_OT、TSV_LUW_NT、TSV_LUW_OT いずれも解凍後のデータサイズが 2GB を超えるので、ユーザーが利用している PC のファイルシステムが 2GB を超えるサイズのファイルに対応していない場合に配慮して、データを複数（5～20 個）のファイルに分割している。ユーザーはこれらのファイルを結合（concatenate）して当該レジスター用の TSV データを構成する必要がある。

表 1-4A: XML データのファイルサイズの解凍前後での変化（単位はメガバイト）

Register	C-XML				M-XML			
	FIXED		VARIABLE		NT		OT	
	前	後	前	後	前	後	前	後
PB*	20.5	59.2	63.6	243.0	1,157.6	9,597.0	1,153.3	9,583.4
PM	4.3	12.6	11.4	45.5	192.7	1517.3	191.6	1,516.8
PN	3.3	8.7	3.1	8.3	58.8	445.2	58.4	444.3
LB*	21.8	63.6	70.0	265.0	1,250.8	10,307.8	1,247.4	10,296.1
OB	--	--	9.3	37.1	155.4	1291.2	155.2	1,290.8
OW	2.9	8.0	8.9	35.4	181.3	1,513.7	178.8	1,503.9
OP	--	--	8.3	40.3	151.5	1,233.3	149.1	1,226.4
OL	--	--	1.4	7.8	34.2	322.0	34.2	322.0
OM	--	--	7.7	31.0	188.7	1,629.8	188.5	1,629.3
OT	--	--	2.3	9.2	37.4	317.4	37.1	316.5
OV	--	--	0.8	4.3	9.6	73.6	9.6	73.6
OC*	--	--	60.4	119.0	519.2	3,516.1	518.2	3,516.0
OY*	--	--	48.9	123.0	500.1	3,663.1	497.8	3,658.2
合計	52.8	152.1	296.2	968.9	4,437.0	35,427.6	4,419.1	35,377.2

*解凍後の値はサブディレクトないし複数ファイルにわけて格納されているデータの合計値

表 1-4B: TSV データのファイルサイズの解凍前後での変化（単位はメガバイト）

Register	NT				OT			
	SUW		LUW		SUW		LUW	
	前	後	前	後	前	後	前	後
PB*	864.4	4,823.8	617.6	3,572.9	842.1	4,827.4	591.6	3,568.2
PM	146.4	769.9	100.7	563.4	141.3	773.9	95.6	562.4
PN	44.0	229.1	29.5	161.2	43.0	229.5	28.4	160.9
LB*	930.7	5,112.0	672.7	3,862.3	911.4	5,113.2	647.6	3,858.5
OB	114.3	630.5	83.6	485.8	112.8	630.6	81.4	485.7
OW	139.2	820.4	91.4	535.1	132.0	820.9	85,2	532.1
OP	122.0	675.7	78.3	436.9	115.4	679.3	72.3	434.6
OL	24.4	173.6	16.1	114.7	24.3	173.6	16.1	114.7
OM	142.0	844.1	103.1	617.6	139.0	844.2	96.3	617.4
OT	27.9	160.1	20.2	118.9	26.9	160.2	19.2	118.7
OV	7.0	33.7	4.9	26.5	7.0	33.6	4.9	26.5
OC*	296.0	1,658.3	214.7	1,258.1	294.4	1,661.7	213.3	1,257.7
OY*	337.6	1,780.4	236.8	1,334.3	331.9	1,784.8	232.2	1,332.4
合計	3,195.9	17,711.7	2,269.7	13,087.8	3,121.6	17,732.9	2,183.9	1,3069.9

*解凍後の値はサブディレクトないし複数ファイルにわけて格納されているデータの合計値

1.4 BCCWJ-DVD 版の意義

『中納言』を利用できる環境にあるユーザーにとって、BCCWJ-DVD 版の存在意義はどこにあるだろうか。『中納言』は「語」（短単位ないし長単位）を単位としてコーパスを検索するツールである。語や語の連鎖を対象とした検索ならば、『中納言』でかなりのところまで用が足りる。

一方、『中納言』では検索できない情報もある。語の属性であっても現在の『中納言』では検索条件に指定できない属性が関与している場合（①,②）、「語」以外の単位が検索条件に関与している場合（③,④,⑤,⑥,⑦）、語ではなくサンプルの属性の検索（⑧,⑨）などは、『中納言』では実施不可能であるか、後処理を必要とする[4]。

① 特定の長さの語を検索する
② 和語だけを検索する
③ 文や段落の長さを測る
④ 文や段落の冒頭に生じやすい語を調査する
⑤ 個々のサンプルの語数を知る
⑥ サンプルごとに「ですます」体と「である」体の生起率を調べる
⑦ 常用漢字の出現頻度リストを作成する
⑧ 書き手の性別や年齢を検索条件に含めて語を検索する
⑨ 書き手の生年の分布を知る

BCCWJ-DVD 版を用いることによって、検索の可能性が大きくひらけてくる。ただしそれは検索に必要な情報を活用できるようになるという意味であって、万能の検索環境が提供されるという意味ではない。BCCWJ-DVD 版には検索ツール類は一切ふくまれていないので、ユーザーは自力で検索環境を構築する必要がある。本文書を読んで BCCWJ-DVD 版の購入を検討しているユーザーは、この点に特に留意していただきたい。

BCCWJ-DVD 版に適した検索環境は何かという問いあわせを受けることがある。ユーザーのスキルによって回答は異なってくるのだが、最も多くのユーザーに当てはまると考えられるのは、TSV 形式のデータはそのままの形でリレーショナルデータベース（RDB）にインポートできるので、MySQL、PostgreSQL、SQL Server などの RDB を利用して、SQL 言語で検索するのが便利ではないか、という回答であろう。

XML 文書を利用するためには、どうしてもある程度のプログラミングスキルが必要である。Ruby、Perl、Python 等のスクリプト言語でそれぞれの XML 処理用ライブラリを利用することが多いだろうが、XSLT のような XML 文書専用の言語もある。

---

[4] 後処理とは『中納言』の検索結果をダウンロードして、そこに含まれる情報を表計算ソフトやリレーショナルデータベース（RDB）などで集計する作業のことである。

1.5　BCCWJ の参考文献

BCCWJ は、構築途上で公開された数種類の「モニター版」も含めて、2011 年の公開以来、国内外の多くの研究者、研究機関によって利用されてきている。その結果、BCCWJ を参照・引用した研究文献も多数出版されている。本稿執筆の時点（2015 年 2 月）で確実に確認されているものだけで、内外 500 件以上の文献があり、国立国語研究所コーパス開発センターのホームページに文献リストが掲載されている[5]。

研究論文で BCCWJ を参照するにはどのような文献を引用すればよいかという問い合わせをもらうこともある。引用の目的によってどの文献が最適かは異なってくるが、以下にいくつか代表的な文献を紹介しておくことにする。

まず英文文献としては以下が代表的である。Disc 1 の MANUAL サブディレクトリにはこの論文の PDF が保管されている（LRE_2014.pdf）[6]。

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. "Balanced corpus of contemporary written Japanese". *Language Resources and Evaluation* 48 (2), pp.345-371 (DOI 10.1007/s10579-013-9261-0), 2014:06.

和文であれば、以下の書籍が代表的である。

山崎誠[編]『書き言葉コーパス ―設計と構築―』講座日本語コーパス 2, 朝倉書店, 2014 (ISBN978-4-254-51602-9 C3381).

この本は全 6 章と付録からなるが、そのうち以下の 5 章で BCCWJ の設計と構築に関する問題が多面的に論じられている。

第 1 章　コーパスの設計	[山崎誠・前川喜久雄]
第 2 章　サンプリング	[丸山岳彦・柏野和佳子]
第 3 章　文書構造の電子化	[山口昌也]
第 4 章　形態論情報	[小椋秀樹]
第 5 章　形態素解析	[小木曽智信]

本マニュアルを引用する場合は以下の書誌情報に準拠していただきたい。

---

[5] http://www.ninjal.ac.jp/corpus_center/bccwj/list.html　このリストは定期的にアップデートされる。

[6] この論文の扱いは Creative Commons Attribution 4.0 International (CC BY)に従う。

国立国語研究所コーパス開発センター「『現代日本語書き言葉均衡コーパス』利用の手引第 1.1 版」国立国語研究所, 2015.

Version 1.1 の Disc 1 の MANUAL サブディレクトリには、本マニュアルの他に BCCWJ の開発過程で蓄積された以下の作業用マニュアル類も保管されている[7]。BCCWJ の設計と構築の詳細情報はこれらの文献から得ることができる。

[1] 丸山岳彦・秋元祐哉「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法 －現代日本語書き言葉の文字数調査－」(JC-D-06-02.pdf)
[2] 丸山岳彦・秋元祐哉「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法(2) －コーパスの設計とサンプルの無作為抽出法－」(JC-D-07-01.pdf)
[3] 柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠「『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例」(JC-D-08-01.pdf)
[4] 丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子「『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用」(JC-D-01.pdf)
[5] 丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子「『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装」(JC-D-10-02.pdf)
[6] 高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也「JIS X 0213:2004 運用の検証」(JC-D-09-01.pdf)
[7] 西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」(JC-D-10-03.pdf)
[8] 山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる「『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」(JC-D-10-04.pdf)
[9] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕「『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)」(JC-D-10-05-01.pdf)
[10] 小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕「『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)」(JC-D-10-05-02.pdf)
[11] 小木曽智信・中村壮範「『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」(JC-U-10-01.pdf)

[7] さらに多くの BCCWJ 関連文書が国立国語研究所コーパス開発センターのホームページで公開されている。
http://www.ninjal.ac.jp/corpus_center/bccwj/doc.html

1.6　BCCWJ 構築の経緯

1.6.1 Version 1.0 の公開まで

BCCWJ の構築は、その構想段階にまで遡ると 2004 年に始まった。同年春に『日本語話し言葉コーパス』の公開を終えた後、国立国語研究所研究開発部門（当時）の有志が集まって、コーパス利用の可能性を探るなかで、現代日本語を対象とした書き言葉均衡コーパスの必要性に対する認識が共有され、後に BCCWJ となる均衡コーパスの概念設計が始まった。翌 2005 年には文科省科学研究費（基盤研究 C, 課題番号 17632002, 代表者:前川喜久雄）の補助を得て、100 万語規模のパイロット版コーパスの構築実験を実施した。

BCCWJ の本格的な構築作業は、国立国語研究所のコーパス整備計画 KOTONOHA 計画の一部として 2006 年 4 月に 5 年計画で始まり、2011 年 7 月末に終了した。その間、2007 年末から 2009 年秋にかけては、独立行政法人に関する行政改革の一環として、国立国語研究所が独立行政法人から大学共同利用機関法人へと移管される騒動があり、BCCWJ 開発チームにもその影響が及んだ。しかし開発メンバーの結束と努力によって、オンライン版も DVD 版も大幅に遅延することなく公開を果たすことができたのは幸いであった。本章冒頭で述べたように Version 1.0 の DVD 版を公開したのは 2011 年 12 月のことであった。

BCCWJ の開発資金には、国立国語研究所の運営費交付金にくわえて、文科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」（略称、特定領域研究「日本語コーパス」、領域代表者：前川喜久雄、2006-2010 年度）の補助を受けた。両資金の分担関係としては、書籍に関するデータ（サンプル ID が PB、LB、OB で始まるサンプル、第 3 章参照）の構築に特定領域研究の研究費をあて、それ以外を運営費交付金でまかなった。

1.6.2 Version 1.1 における修正

BCCWJ-DVD 版（Version 1.0）の公開後、ユーザーから寄せられた意見のうち、早急な対応を必要としたのが、文境界の認定基準の見直しであった。書き言葉において文末を認定し、文境界を設定することは、句読点などの記号類が用いられている以上、容易であると思われるかもしれない。しかし、実際に 1 億語相当のサンプルを処理してみると、文末が記号類で明示されていないサンプルが頻出することにくわえ、複雑長大な引用の存在、果ては文末であるのか否かを文法的には解決不能と思われるサンプルの存在まで、複雑多岐な問題に直面する。Version 1.0 は 5 年間という強い時間的制約の下で開発したため、文境界認定の基準が十分に練り上げられておらず、問題の複雑さに対処しきれていなかった。

文境界認定の異同は、文数・文長などの計量言語学的指標に影響するだけでなく、係り受け構造や述語項構造などの言語アノテーション作業にも深刻な影響を及ぼす。そこで 2013 年初夏には国立国語研究所コーパス開発センターで、文境界認定基準の再検討を開始した。

その後、約 1 年の検討期間を経て、2014 年春には文境界修正方針の成案を得たので、実

際の修正作業に着手した。今回の修正で文末認定に関するすべての問題が解決されたわけではないが、Version 1.0 に比較すれば大幅に問題が軽減されているものと信じる。

また Version 1.1 の公開を機に非 NumTrans 版データも公開することにした。NumTrans は先に述べたように数字を形態素解析しやすくするための前処理であるが、短単位と数字の対応をとるためにすべての数字を漢字表記に変換する。もちろん原文の表記情報が失われているわけではなく、XML 文書中にタグを付して保存されているのだが、『中納言』その他でユーザーの目にとまるのが漢字に変換された文字列であるため、原文を改変してしまっているとの誤解を生じる原因となった。また自然言語処理の研究者からも、処理の煩雑さを厭う声があがっていた。非 NumTrans 版の公開によって、これらの批判にも前向きに応えることができたと信じる。

文境界認定基準の再検討には浅原正幸・小木曽智信・山口昌也・山崎誠・丸山岳彦・中村壮範・小西光・田中弥生と筆者が、その後のデータ修正作業には、上記にくわえて立花幸子・加藤祥・今田水穂・間淵洋子が参加した。

1.7 謝辞

サンプルの利用許諾をいただいた延べ 1 万人を超える個人著作権者のみなさまに、心より感謝申しあげる。

また先に 1.2.5 節で述べたように、BCCWJ の著作権処理では、多くの法人、団体のご協力をいただいた。以下にその名称を記して感謝のしるしとしたい。

公益社団法人日本文藝家協会、社団法人日本推理作家協会、社団法人日本児童文学者協会、社団法人日本児童文芸家協会、社団法人日本ペンクラブの各団体には、文芸分野でのサンプルの著作権者への広報および依頼状発送業務にご協力いただいた。また鷹羽狩行、篠弘の両氏には韻文関係のサンプル選定についてご指導をいただいた。

社団法人教科書協会、一般社団法人教学図書協会には、教科書出版各社との連絡を仲介していただいた。

一般社団法人日本音楽著作権協会には、歌詞に関係するサンプルの利用を許諾していただいた。

㈱朝日新聞社、㈱読売新聞社、㈱産業経済新聞社、㈱毎日新聞社、㈱京都新聞社、㈱中日新聞社、㈱高知新聞社、㈱神戸新聞社、㈱西日本新聞社、㈱北海道新聞社、㈱新潟日報社、㈱河北新報、㈱琉球新報社、㈱中国新聞社、一般社団法人共同通信社、㈱時事通信社からは新聞記事サンプルの利用を許諾していただいた。

ヤフー株式会社からは、Yahoo!知恵袋および Yahoo!ブログのデータを提供していただき、著作権の一括処理にご尽力いただいた。

白書の著作権処理に関しては中央省庁における担当部署に、また広報紙の著作権に関しては地方自治体の担当部署に、それぞれご協力いただいた。

衆議院記録部、参議院記録部、国会図書館の関係者からは国会会議録の著作権処理方針

について種々ご教示をいただいた。

個人著作権者との交渉に際しては、権利者との連絡をとるための窓口として、出版社に接触することが多かった。そのなかで、㈱アカデミー出版、㈱ヴィレッジブックス、㈱オライリー・ジャパン、㈱オレンジページ、㈱学習研究社、㈱経済界、㈱光人社、㈱小学館、㈱新潮社、㈱誠文堂新光社、㈱世界文化社、㈱ナツメ社、㈱南江堂、㈱日本実業出版社、㈱ハーレクイン、㈱PHP 研究所、㈱文芸社、㈱マガジンハウス、㈱みすず書房の各社においては格別に好意的なご対応をいただいた。

書籍、雑誌、新聞類の原本の閲覧、および書誌情報データの入手においては、大阪市立中央図書館、国立国会図書館、埼玉県立浦和図書館、埼玉県立久喜図書館、埼玉県立熊谷図書館、自治大学校図書室、湘北短期大学図書館、立川市図書館、東京都立多摩図書館、東京都立中央図書館、東京都立日比谷図書館、日本図書館協会、八王子市図書館、一橋大学附属図書館、横浜市中央図書館に便宜を図っていただいた。

付録：BCCWJ 開発メンバー

秋元祐哉	阿左美厚子	稲益佐知子	内元清貴	大石有香
大島一	大矢内夢子	小川志乃	小木曽智信	小椋秀樹
小沼悦	柏野和佳子	神野博子	河内昭浩	北村雅則
小磯花絵	小澤俊介	小西光	小林正行	小松祐美
近藤明日子	佐野大樹	鈴木翼	相馬さつき	高田智和
竹内ゆかり	田中牧郎	田中弥生	伝康晴	中村壮範
西部みちる	長谷川愛	服部龍太郎	原裕	平本智弥
平山允子	冨士池優美	前川喜久雄	間淵洋子	丸山岳彦
宮内佐夜香	舞木右	森本祥子	山口昌也	山崎誠
山田篤	吉田谷幸宏	渡部涼子	浅原正幸†	今田水穂†
加藤祥†	立花幸子†			

†Version 1.1 から参加

第２章　『現代日本語書き言葉均衡コーパス』の設計

山崎 誠

2.1　はじめに

本章では、『現代日本語書き言葉均衡コーパス』（以下、BCCWJと省略）の設計の概要について説明する。

BCCWJは日本で初めての本格的な書き言葉均衡コーパスである。BCCWJは次のような点で日本語研究の質の向上に貢献する。従来、日本語研究においてコーパスとみなして利用されてきたデータはいくつかあったが、それらは新聞記事データ集や青空文庫などの単一の種類のテキストの集まりであり、書き言葉の一面を捉えているにすぎなかった。それに対して、BCCWJは書籍、新聞、雑誌、白書、ブログ等異なるレジスターのテキストの集まりであり、書き言葉の多様な実態を捉えることができるデータになっている。

また、従来の書き言葉データの多くはプレイン・テキストであり、その使い方は文字列検索が中心であったため正規表現を使っても限界があった。BCCWJは言語単位の情報（形態論情報）や書誌情報などの研究用のアノテーションが施されており、複雑な検索結果をもとに、より深い分析が可能である。

2.2　BCCWJ の設計

2.2.1 基本方針

BCCWJを構築するにあたっては、以下の四つの点を念頭に置いて設計した（前川 2008、山崎 2009）。

（1）現代日本語の縮図となるコーパス

従来、国立国語研究所が行ってきた語彙調査の手法を生かし、コーパスがその母集団の統計的な縮図になり、母集団に対し代表性（representativeness）を持つように設計する。これにより、母集団における言語的諸特性の分布が過不足なく表現できることになり、データの信頼性を高めることが出来る。

（2）汎用的な目的に供するコーパス

言語研究（語彙・文法・文字）以外にも、応用面として日本語教育や国語教育、国語政策、辞書編集、自然言語処理などの分野でも活用することを目的として、多様な日本語の姿を捉えることができるよう設計する。

（3）公開可能なコーパス

収録する著作物について利用許諾を得て公開する。公開形態は、オンラインでの簡易検索のほか、形態論情報を使って共起条件を詳しく指定できるオンライン詳細検索、DVDに

よる全文提供の3種類である。コーパスが学界の共有財産となることによって、研究の追試が可能になったり、日本語を母語としない研究者が研究を行いやすくなるなどのメリットがある。

（4）既存のコーパスとの調和

XMLによる文書構造の記述、2種類の言語単位（短単位、長単位）による形態論情報の付与により、『太陽コーパス』『日本語話し言葉コーパス』との整合性を保つ。

2.2.2 基本概念の定義

BCCWJは、現代日本語の書き言葉を収録するコーパスであるので、「現代」「日本語」「書き言葉」のそれぞれについて、以下のような基準を決めて資料選定にあたった。詳細な取り扱いについては、第3章「サンプリング」及び丸山他（2011a）を参照されたい。

【現代】

明治時代以降を現代とする。したがって、「源氏物語」などの江戸時代より以前の作品は対象外となる。ただし、古典の現代語訳は現代語として扱った。また、短歌、俳句などの韻文で使われる古語は現代語として扱った。

【日本語】

方言を含む日本語が対象である。英語、中国語などの外国語は対象外である。テキストによっては、日本語と外国語が混じっているものがある。そのような場合、段落単位で外国語かどうかの認定を行い、対象範囲を確定した。

【書き言葉】

文字で記録された言葉。インタビューの書き起こしなどを含む。

2.2.3 BCCWJ の基本構成

BCCWJは、出版（生産実態）サブコーパス、図書館（流通実態）サブコーパス、特定目的サブコーパス三つのサブコーパスから構成される（図2-1参照）。

<table>
<tr><td>出版（生産実態）サブコーパス
約3,500万語
書籍、雑誌、新聞
2001年～2005年</td><td>図書館（流通実態）サブコーパス
約3,000万語
書籍
1986年～2005年</td></tr>
<tr><td colspan="2">特定目的サブコーパス
約3,500万語
白書、教科書、広報紙、ベストセラー
Web掲示板、ブログ、韻文、法律、国会会議録
対象期間はさまざま</td></tr>
</table>

図2-1: BCCWJの構成

各サブコーパスは、さらにいくつかのレジスターから構成される。表2-1は各レジスターのサンプル数と短単位で数えた場合の延べ語数を示したものである。語数は品詞欄が空白・補助記号・記号のものは数えていない。また、固定長サンプルと可変長サンプルがあるレジスターについてはそれらを合わせて重複部分を差し引いた範囲を対象としている。

表2-1: 各レジスターのサンプル数と語数

サブコーパス	レジスター	サンプル（個）	NumTrans 版の語数（万）	非 NumTrans 版の語数（万）
出版サブコーパス	書籍(PB)	10,117	2,855	2,866
	雑誌(PM)	1,996	444	450
	新聞(PN)	1,473	137	138
図書館サブコーパス	書籍(LB)	10,551	3,038	3,044
特定目的サブコーパス	白書(OW)	1,500	488	494
	教科書(OT)	412	93	93
	広報紙(OP)	354	376	383
	ベストセラー(OB)	1,390	374	375
	Yahoo!知恵袋(OC)	91,445	1,026	1,030
	Yahoo!ブログ(OY)	52,680	1,019	1,028
	韻文(OV)	252	23	23
	法律(OL)	346	108	108
	国会会議録(OM)	159	510	510
合計		172,675	10,491	10,542

2.2.4 BCCWJ の規模

表2-1に示すように、BCCWJ全体の規模は短単位で数えて約１億語である。レジスター別では、LB（図書館書籍）が最も大きく約3,000万語、PB（出版書籍）もほぼ同じサイズであり、合わせると、BCCWJ全体の約6割は書籍で占められていることになる。それぞれのレジスターにおける延べ語数が異なるため、レジスター間で出現頻度を比較する場合は、それぞれの語数で割った出現率で比較しなければならない。

2.2.5 各サブコーパスの特徴

以下、各サブコーパスについて概括を述べるが、それぞれのサブコーパスに含まれるレジスターとその選定方法については、第3章「サンプリング」及び丸山他（2011a、b）を参照されたい。

A. 出版サブコーパス

書き言葉を生産する書き手の立場を重視したもので、売れ行きや知名度にかかわらず、出版された書き言葉であれば、どの書籍（雑誌、新聞）も同じ確率で選ばれるようにする。後述の流通実態を捉えたサブコーパスに比べると語彙やコロケーションなど言語的属性の

多様性が確保されることが期待される。

このサブコーパスには成人向けの書籍が一定の割合で含まれている。教育現場で使用する際には注意されたい。

B. 図書館サブコーパス

書き言葉が書き手と読み手との間で、社会的に流通している実態を図書館の所蔵から捉えたサブコーパスである。広い意味で社会の需要を反映している書き言葉とも言える。このサブコーパスは、極端に専門的な書籍や成人向け書籍が排除されることによって、より一般的な用語用字を調べるのに適していると期待される。また、資料年代にある程度の時間的な幅があり、短期間であるが通時的な観察が可能になる。

C. 特定目的サブコーパス

出版サブコーパス、図書館サブコーパスでは十分な分量が集まりにくい資料を中心に収録したサブコーパスである。例えば、政府の白書は上記二つのサブコーパスからでは分析に必要なだけの量が得られないため、白書のみを母集団としたデータからサンプリングを行い、サブコーパスに収録した。同様に、教科書・広報紙・ベストセラー・韻文・法律・国会会議録を収録した。また、ウェブの書き言葉（Yahoo!知恵袋、Yahoo!ブログ）も収録し、紙媒体の言語と比較できるようにした。

2.2.6 コアデータ

BCCWJに付与されている形態論情報などのアノテーションは、ほとんど自動付与であるが、BCCWJ全体の約100分の1の量に相当する約110万語については、人手により解析精度を高めている。この部分を「コアデータ」と呼んでいる。BCCWJ全体の解析精度が約98%であるのに対してコアデータの解析精度は99%以上である。コアデータを構成するレジスターは、出版書籍（PB）、雑誌（PM）、新聞（PN）、白書（OW）、Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）の六つである。

コアデータには、さまざまなアノテーションが施されており、順次、次のURLで公開される予定である。

http://www.ninjal.ac.jp/corpus_center/anno/

2.3 サンプルの長さとタイプ

2.3.1 問題点

コーパスに収録する1サンプルの長さをどのように決めるかはコーパスの設計にとって、コストにも影響する重要な問題である。1サンプルの長さが長くなれば収録するサンプルの数が少なくなり（著作権処理の負担減にも直結する）、労力も少なくて済むが、語彙的なかたよりが生じる。

また、1サンプルの長さについて、それが一定かどうかという、サンプルのタイプも重要な問題である。一定の長さのサンプルは計量的な分析に向いているが、多くの場合文が途

中で切れてしまうことになり、文脈を把握するような分析には向いていない。意味的なまとまりを重視するとサンプルの長さがまちまちになる。

BCCWJでは、サンプルの長さをとタイプについて、それぞれの長所を生かす以下のような設計を行った。

2.3.2 サンプルのタイプ

A. 固定長（FIXED）サンプル

固定長サンプルは、ひとつのサンプルの長さを1,000字とする（句読点などの補助記号は含めない）。固定長サンプルは、母集団からの抽出比率に基づいた統計的な処理、語彙表や漢字表の作成に適している。ちなみに、1サンプル1,000字は短単位で約590語であり、文庫本でいうと見開きより少し多いくらいの言語量である。

固定長サンプルのデータは、係り受けの関係が理解できるよう、サンプルの開始点を含む文の文頭からサンプルの終了点を含む文の文末までが収録されている。そのため、実際のひとつのサンプルの文字数は1,000字より多いが、サンプルの開始点と終了点がマークアップされており、その間がちょうど1,000文字となる。

B. 可変長（VARIABLE）サンプル

可変長サンプルは、文章のまとまりをもとに長さを決める。そのためひとつのサンプルの長さは一定ではない。多くの書籍では、節、章などのまとまりが1サンプルとなる。ただし、無制限に長いサンプルができるとそのサンプルの影響が強く出てしまうので、長さの上限を1万字としている。可変長サンプルは文章の論理構造を対象とした分析に適している。

2.3.3 サンプルの重なり

コーパス構築に当たって固定長サンプルと可変長のサンプルを別々に取得するのは作業コストがかかりすぎるため、BCCWJでは1回のサンプリングで当たった同一箇所から固定長と可変長の二つのサンプルを取得している。そのため、固定長サンプルと可変長サンプルとの間には包含関係を基本とする3種類のパターンが生じる。いちばん多いパターンは、固定長サンプルが可変長サンプルの中に完全に含まれる場合である。次に多いのが、固定長サンプルが可変長サンプルの末尾からはみ出す場合である。また、数は少ないが、固定長サンプルと可変長サンプルが重なり合わないパターンもある。

2.3.4 レジスターとサンプルのタイプ

表2-2にレジスターとサンプルのタイプの関係を示した。可変長サンプルは全てのレジスターにあるが、固定長サンプルは、出版サブコーパス全体、図書館サブコーパス全体と特定目的サブコーパスの白書だけに存在する。

表2-2: レジスターとサンプルのタイプ

サブコーパス	レジスター	サンプルのタイプ
出版サブコーパス	書籍(PB)	固定長、可変長
	雑誌(PM)	固定長、可変長
	新聞(PN)	固定長、可変長
図書館サブコーパス	書籍(LB)	固定長、可変長
特定目的サブコーパス	白書(OW)	固定長、可変長
	教科書(OT)	可変長
	広報紙(OP)	可変長
	ベストセラー(OB)	可変長
	Yahoo!知恵袋(OC)	可変長
	Yahoo!ブログ(OY)	可変長
	韻文(OV)	可変長
	法律(OL)	可変長
	国会会議録(OM)	可変長

2.4 電子化

2.4.1 文字入力

出版サブコーパスおよび図書館サブコーパスのように原文が紙媒体（原資料の媒体についての詳細は表4-1を参照）である場合には、電子化するための基準が必要である。文字入力については、以下の方針を立てた。

(1) JIS X 0213:2004 規格に基づき字形を詳細に区別する

この文字セットの採用により、ほとんどの文字を入力し分けることができる。詳細は、高田他（2009）を参照されたい。

(2) 記号・改行の意味による統制、統一的な表記

例えば、「コーパス」という語を表記する際の2文字目の中央位置横線は、通常「ー（長音符号）」が用いられるが、資料によっては「－（マイナス）」や「―（ダッシュ）」が用いられているものや、形状からはどの文字かを判別できない場合がある。また、「ー（マイナス）」を用いた「コーパス」という表記を、そのままコーパス本文に採用すると、語の検索や形態素解析を困難にする。そのため、原文における見え方ではなく、その意味によって入力し分ける。ダッシュ、ハイフン、長音、漢数字の「一」、丸記号、漢数字の「〇」、ローマ字の「０」などが対象となる。また、改行やスペースは、レイアウトではなく、論理的に意味をもつもののみを再現する。例えば、語や文を句切る空白、段落冒頭の1字字下げは入力するが、レイアウトのための空白は入力しない。

(3) 組み文字・半角文字を使わない

㈱、㌢のようないわゆる組み文字は「（株）」、「センチ」のようにすべて１字ずつ切

り離して入力する。また、半角文字は使用せずすべて全角で入力する。

文字入力の具体的な記述は西部他（2011）を参照されたい。

2.4.2 タグの仕様

BCCWJのタグの特徴は、形態論情報のタグだけでなく、『太陽コーパス』で行ったタグ付けの経験を生かし、文書構造が的確に再現されるようにしている点である。以下に主なタグの種別と特徴を挙げる。タグの詳細は、第4章および山口他（2011）を参照されたい。

(1) 文書構造情報

記事、見出し、段落、引用、文などのタグを付与し、文章を構造化・階層化して表現する。

(2) 文字情報

文字の読みに関するルビ、誤植などの校正注、文字集合に含まない文字や記号（外字）などの情報を付与する。

(3) 形態論情報

短単位、長単位についての形態論情報（語彙素、出現形、品詞、語彙素読み、語種など）を付与する。

(4) サンプリング情報

サンプリング時に決定するサンプル抽出基準点（乱数による縦横交叉点から決まる文字。7.4.5節参照）の情報を付与する。

2.5 解析単位（短単位、長単位）

BCCWJでは柔軟な検索・分析に対応するために「短単位」「長単位」という2種類の言語単位を用いている。短単位はコーパスからの用例収集に適した単位であり、長単位はBCCWJに格納したレジスターの言語的特徴の解明に適した単位である。

解析単位は、大量のデータをコンピュータで処理するのに向いているという性質が必須である。BCCWJの構築にあたってもその趣旨に則って、解析単位を揺れの少ない規則の集合として定義した。その詳細は、第5章および小椋他（2011a）を参照されたい。

BCCWJはすべてのサンプルが短単位と長単位の二つの単位で解析されている。解析精度は品詞も含めた見出し語の認定のレベルで98%以上である（レジスターによって解析精度に若干差がある）。

短単位、長単位は、元々は国立国語研究所の語彙調査で開発された調査単位であり『日本語話し言葉コーパス』の構築においても使用された。前者は最小単位（形態素）の一次結合までを最大とする言語単位であり、後者はほぼ文節に近い長さの言語単位である。例えば、「国立国語研究所は人間文化研究機構に移管される。」という文は、短単位で「/国立/国語/研究/所/は/人間/文化/研究/機構/に/移管/さ/れる/。/」と14単位に分割されるが、長単位では、「/国立国語研究所/は/人間文化研究機構/に/移管さ/れる/。/」と7単位になる。

参考文献

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011a）「『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）」国立国語研究所内部報告書 LR-CCG-10-05-01

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011b）「『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.

高田智和・小林正行・間淵洋子・大島一・西部みちる・山口昌也（2009）「JIS X 0213:2004 運用の検証」特定領域研究「日本語コーパス」平成21年度研究成果報告書JC-D-09-01.

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也（2011）「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」国立国語研究所内部報告書LR-CCG-10-04.

前川喜久雄（2008）「KOTONOHA『現代日本語書き言葉均衡コーパス』の開発」日本語の研究, 4 (1), 82-95.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011a）「『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用」国立国語研究所内部報告書LR-CCG-10-01.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011b）「『現代日本語書き言葉均衡コーパス』に含まれるサンプルおよび書誌情報の設計と実装」国立国語研究所内部報告書LR-CCG-10-02.

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる（2011）「『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」国立国語研究所内部報告書LR-CCG-10-04.

山崎誠（2009）「代表性を有する現代日本語書籍コーパスの構築」人工知能学会誌, 24 (5), 623-631.

第3章　サンプリング

丸山 岳彦　柏野 和佳子　田中 牧郎

3.1　BCCWJ 構築の基本理念

『現代日本語書き言葉均衡コーパス』（以下、BCCWJと略記する）を構築する上での基本理念は、次の4点にまとめられる（第2章参照）。

(1) 現代日本語の縮図となるコーパス
　これまで研究所が行ってきた語彙調査の手法を生かし、コーパスがその母集団の統計的な縮図になるよう設計する。それにより、母集団における言語的諸特性の分布が縮図において過不足なく再現でき、母集団における分布を高い精度で推測できるようになる。
(2) 汎用的な目的に供するコーパス
　言語研究（語彙・文法・文字）以外にも、応用面として、辞書編集や言語政策、日本語教育などでも使えることを意図し、多様な日本語の姿を捉えることができるよう設計する。また、言語変化に対応するためには、同じ設計のコーパスを繰り返し構築するなど定点観測的な工夫も必要である。
(3) 公開可能なコーパス
　収録する著作物の利用許諾を得て、公開を目指す。インターネット上からの簡易検索のほか、共起条件を指定できる検索ツールなどもあわせて提供する。
(4) 既存のコーパスとの調和
　解析単位の仕様を『日本語話し言葉コーパス』に合わせ、短単位、長単位の2種類の解析を行う。

これらの基本理念のうち(1)と(2)は、コーパスの設計、およびサンプリングに関わる問題である。また、(3)は著作権処理、(4)は形態論情報の付与に関わる理念である。サンプリングに関わる問題のうち、(1)については、レジスターごとに母集団を厳密に定義して、層別ランダムサンプリングを実施することにより実現した。(2)については、サンプリングの際、固定長サンプル・可変長サンプルという2種類のサンプルを取得することにより、統計的な研究から文章研究までに対応できるサンプル抽出を実現した。

以下では、BCCWJの設計、およびサンプリング作業の概要について解説する。

3.2　BCCWJ を構成する三つのサブコーパス

まず、BCCWJの構成を、図3-1に示す。

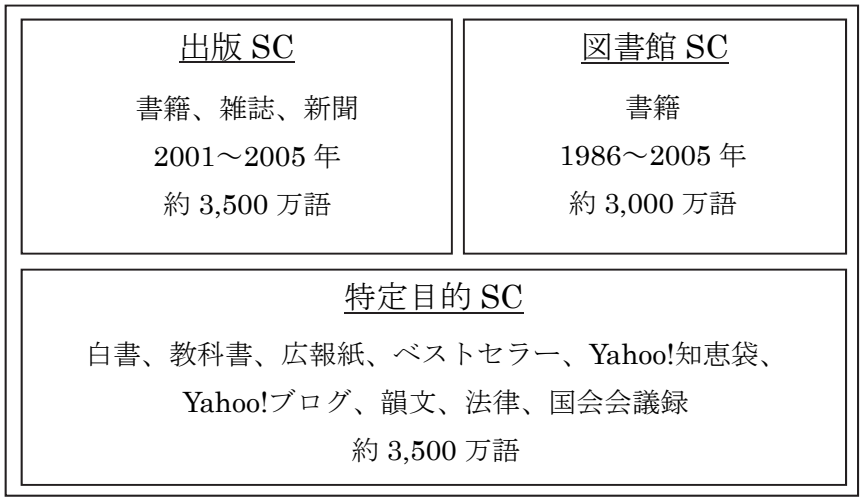


図3-1: BCCWJの構成

各サブコーパス（以下、SCと略記する）の概要を、以下に述べる。

3.2.1 出版（生産実態）SC

出版SCは、書き言葉の出版・生産という側面に着目するSCである。2001年から2005年の間に国内で出版されたすべての書籍・雑誌・新聞に含まれる文字の総体を母集団として、ランダムサンプリングによって得られる約3,500万語分のデータを収める。書き言葉が実際に出版された結果を、文字数という量的側面からできる限り忠実に反映することで、5年間における書き言葉の出版に関するありさまを捉えることを目的とする。

3.2.2 図書館（流通実態）SC

図書館SCは、書き言葉の流通・流布の実態という側面に着目するSCである。東京都内の公立図書館に所蔵されている書籍（ただし1986年から2005年の20年間に出版されたもの）を対象として、ランダムサンプリングによって得られる約3,000万語分のデータを収める。書き言葉（書籍）が世の中に流通している状態を公立図書館の所蔵状況によって近似的に把握し、世の中に広く行き渡っている書き言葉のありさまを捉えることを目的とする。

3.2.3 特定目的 SC

特定目的SCは、生産・流通という側面からは捉えきれない、あるいは、出版SC・図書館SCの母集団には入らないけれども、書き言葉の研究を遂行する上で必要と思われる種類の書き言葉を収めるSCである。白書、教科書、広報紙、ベストセラー、Yahoo!知恵袋、Yahoo!ブログ、韻文、法律、国会会議録を対象として、約3,500万語分のデータを収める。収録対象期間はレジスターによって異なる。

3.3 BCCWJ を構成する 2 種類のサンプル

三つのSCは、「固定長サンプル」「可変長サンプル」という2種類のサンプルによって構成する。

- 固定長サンプルの設計方針：
  統計的に厳密な言語調査に耐え得る設計にする。
- 可変長サンプルの設計方針：
  文体研究・テキスト研究に耐え得るよう、ある程度の文脈を確保した設計にする。

3.3.1 固定長（FIXED）サンプル

「固定長サンプル」は、母集団に含まれる全ての文字に対して等確率を与えた上で、ある1文字をランダムに指定し（この1文字を「サンプル抽出基準点」（7.4.5節参照）と呼ぶ）、その文字を始点として1,000文字目までの範囲を抽出するサンプルである。全ての文字に対して等確率を与えるために、母集団に含まれる文字の総数をあらかじめ推計しておく必要がある。母集団（＝推計された総文字数）からの抽出比が明確である点で、基本語彙表や漢字表の作成、語彙・文字調査など、統計的な言語研究に向く。また、母集団の層的かつ量的な構造を忠実に反映する点で、統計的な代表性を備えた均衡コーパスとしての性格を強く持つ。

3.3.2 可変長（VARIABLE）サンプル

「可変長サンプル」は、固定長サンプルと同様、母集団に含まれる全ての文字に対して等確率を与えた上で、ある1文字をランダムに指定し、その1文字を含む言語的な構造のまとまり（「章」や「節」など、ただし1万字を上限とする）を抽出するサンプルである。文章・談話としてのまとまりを重視したサンプルであるため、テキストの論理構造の把握や文脈の分析、文体の調査などに向く。

なお、可変長サンプルは、三つのSCの全てに対して提供される。一方、固定長サンプルは、統計的な言語調査を行う可能性の高いSC、すなわち、出版SC、図書館SC、および、特定目的SCの一部（白書）に対して提供される。

3.4 BCCWJ に収録するテキストの条件

BCCWJは現代日本語の書き言葉を収録したコーパスであるが、実際にサンプリング作業を実施するにあたり、「現代日本語書き言葉」をどのように定義すればよいか、という問題があった。そこで、「明治初年以降に」「日本語で」「書かれた」言葉を「現代日本語書き言葉」として定義し、これらの条件を満たすことをBCCWJに収録するテキストの条件とした。よって、江戸期以前に書かれた書き言葉は、基本的に（特定目的SC「教科書」レジスターの「国語」の一部を除いて）収録されていない。また、日本語の文章の中に外国

語が混在している場合は可能な限りそのまま収録しているが、例えばひとまとまりの英文が単独のパラグラフを構成している場合、その部分は収録対象から除外した。

3.5 BCCWJ-DVD 版に収録されているサンプルの一覧

BCCWJ-DVD版に収録されているサンプルの一覧を、表3-1に示す。なお、*が付与されているレジスターは、固定長サンプルと可変長サンプルの両方が、表3-1の「サンプル数」分それぞれ収録されている。*が付与されていないレジスターは、可変長サンプルのみが収録されている。

表3-1: 「BCCWJ-DVD版」に収録されているサンプルの一覧

SC	レジスター	対象期間	母集団	サンプル数
出版 SC（生産実態）	書籍 *	2001 年-2005 年	約 485 億文字	10,117
	雑誌 *	2001 年-2005 年	約 105 億文字	1,996
	新聞 *	2001 年-2005 年	約 64 億文字	1,473
図書館 SC（流通実態）	書籍 *	1986 年-2005 年	約 479 億文字	10,551
特定目的 SC	白書 *	1976 年-2005 年	1,006 冊	1,500
	教科書	2005 年-2007 年	145 冊	412
	広報紙	2008 年	100 自治体	354
	ベストセラー	1976 年-2005 年	951 冊	1,390
	Yahoo!知恵袋	2004 年-2005 年	約 312 万質問	91,445
	Yahoo!ブログ	2008 年-2009 年	約 346 万記事	52,680
	韻文	1980 年-2005 年	130 冊	252
	法律	1976 年-2005 年	718 法律	346
	国会会議録	1976 年-2005 年	32,925 会議	159

以下、3.6節では、BCCWJの構築において実施したサンプリング作業の方法について、各SCおよびレジスターごとに、概要を示す。なお、出版SC・図書館SCの設計の詳細については丸山・秋元（2007、2008）を、サンプリングの基準と実施手順の詳細については柏野他（2009）、丸山他（2011）を、それぞれ参照されたい。

3.6 サンプリング方法

以下では、BCCWJの構築において実施したサンプリング作業の方法について、各SC、およびレジスターごとに、その概要を示す。

3.6.1 出版 SC「書籍」

出版SC「書籍」は、2001年から2005年までの5年間に日本国内で出版されたすべての書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 国立国会図書館の書誌データ「J-BISC」を用いて、2001年から2005年までの5年間に出版された書籍を同定した。この際、漫画、写真集、電子資料、地図、学習試験図書、一般には流通しない官公庁刊行物、40ページ以下の書籍、ページ数の記録がない書籍などは除外した。その結果、5年間に出版された「書籍」は317,117冊、74,911,520ページという結果を得た。
- これらの書籍に印刷されている総文字数を推計した。「NDC（日本十進分類法）」および判型（本の高さ）の別にランダムに書籍を選び、そこからランダムに選んだページ内の文字数を実測した。合計227冊、1,135ページ分を実測した結果から1ページあたりの平均文字数を算出し、これを74,911,520ページに適用したところ、48,539,925,351文字という結果を得た。この総文字数を、出版SC「書籍」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計55層に層別した。
  - NDC（11層）： 国立国会図書館の蔵書目録「J-BISC」に書籍ごとに付与されているNDCの第1次区分（0～9）に、NDCが付与されていない「記録なし」を加えた、11分類。
  - **出版年（5層）**： 書籍の出版年である2001年から2005年までの、5分類。

サンプリング方法

- 母集団を55層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した10,117サンプルについて、NDCごとの内訳を、図3-2に示す。

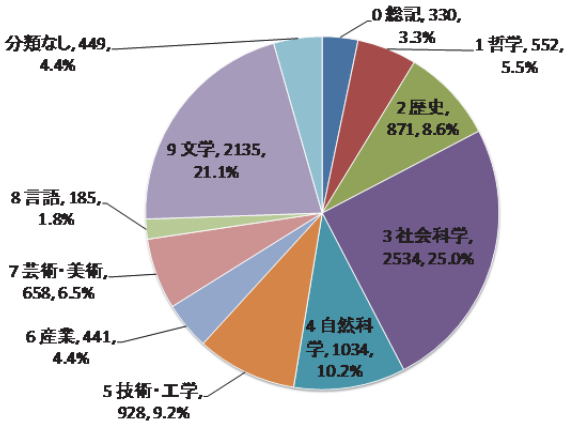


図3-2: サンプルの数と構成比率
（出版SC「書籍」、NDC別）

3.6.2 出版 SC「雑誌」

出版SC「雑誌」は、2001年から2005年までの5年間に日本国内で出版されたすべての雑誌を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 『雑誌新聞総かたろぐ』（メディア・リサーチ・センター発行）を用いて、2001年から2005年の間に社団法人日本雑誌協会に加盟していた出版社が出版した定期刊行物を同定した。この際、新聞・通信、コミック、要覧、非日本語による定期刊行物は除外した。その結果、5年間に出版された「雑誌」は、1,259タイトル、55,779冊、10,414,955ページという結果を得た。
- これらの雑誌に印刷されている総文字数を推計した。『雑誌新聞総かたろぐ』のジャンルおよび判型の別にランダムに雑誌を選び、そこからランダムに選んだページ内の文字数を実測した。合計53冊、265ページ分の実測した結果から1ページあたりの平均文字数を算出し、これを10,414,955ページに適用したところ、10,515,681,636文字という結果を得た。この総文字数を、出版SC「雑誌」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計30層に層別した。
  - **ジャンル（6層）**：『雑誌新聞総かたろぐ』で分類されているジャンル（1. 総合、2. 教育・学芸、3. 政治・経済・商業、4. 産業、5. 工業、6. 厚生・医療）による6分類。
  - **出版年（5層）**：雑誌の出版年である2001年から2005年までの5分類。

サンプリング方法

- 母集団を30層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された雑誌の指定されたページに含まれる文章を一定の手続きにより抽出した。なお、著作権処理の観点から、個人情報（一般人の氏名や住所、電話番号など）や出版社から要請のあった箇所に対して伏せ字処理を実施した。

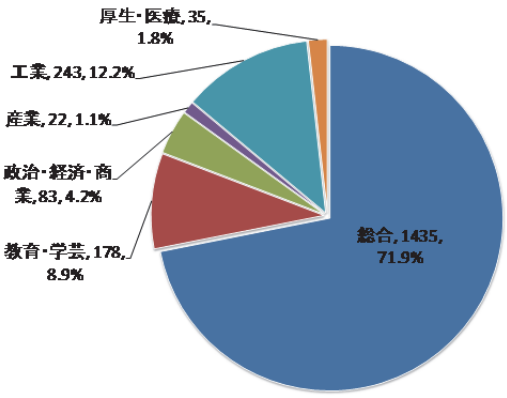


図3-3: サンプルの数と構成比率
（出版SC「雑誌」、ジャンル別）

- 収録した1,996サンプルについて、ジャンルごとの内訳を、図3-3に示す。

3.6.3 出版 SC「新聞」

出版SC「新聞」は、2001年から2005年までの5年間に日本国内で発行されたすべての新聞を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 『全国新聞ガイド』（社団法人日本新聞協会発行）を用いて、「全国紙」「ブロック紙」および各地の有力な地方紙をリスト化した。この結果、全国紙（朝日新聞、毎日新聞、読売新聞、日本経済新聞、産経新聞）、ブロック紙（北海道新聞、中日新聞、西日本新聞）、地方紙（河北新報、新潟日報、京都新聞、神戸新聞、中国新聞、高知新聞、愛媛新聞、琉球新報）を同定した。
- 上記の新聞に関するページ数や発行回数などを調査した結果、5年間に発行された「新聞」は、16タイトル、合計49,625冊、1,198,189ページという結果を得た。
- これらの新聞に印刷されている総文字数を推計した。全国紙4紙の朝夕刊を合計8冊を、曜日を考慮してランダムに選び、そこに含まれている211ページに印刷されている全文字数を実測した。この結果から1ページ当たりの平均文字数を面種ごとに算出し、1,198,189ページに適用したところ、6,416,070,114文字という結果を得た。この総文字数を、出版SC「新聞」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計80層に層別した。
  - **新聞タイトル（16層）**： 新聞タイトルによる16分類。
  - **発行年（5層）**： 新聞の発行年である2001年から2005年までの5分類。

サンプリング方法

- 母集団を80層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された新聞の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した1,473サンプルについて、新聞タイトルごとの内訳を、図3-4に示す。

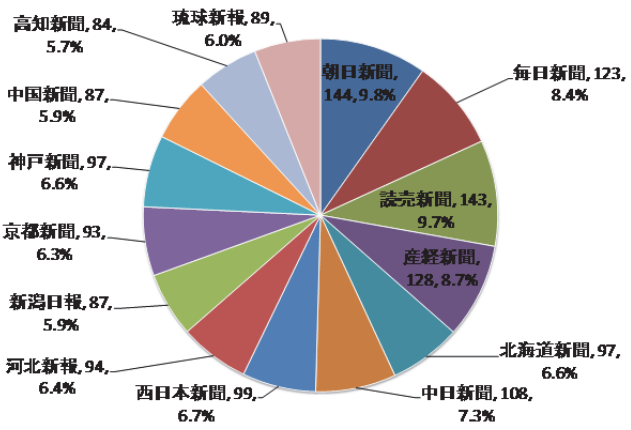


図3-4: サンプルの数と構成比率
（出版SC「新聞」、タイトル別）

3.6.4 図書館 SC「書籍」

図書館SC「書籍」は、1986年から2005年までの20年間に出版された書籍のうち、東京都内の公立図書館に所蔵されている書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 東京都立中央図書館作成の「ISBN総合目録」を用いて、東京都内の区市町村立図書館が所蔵する蔵書リストを作成した。
- 集計の結果、東京都内の13自治体以上で共通に所蔵されている335,721冊、85,363,019ページを対象とすると、推計総文字数が47,877,656,072文字となり、出版SC「書籍」の母集団とほぼ等しくなることが判明した。この総文字数を、図書館SC「書籍」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計220層に層別した。
  - NDC（11層）： 国立国会図書館の蔵書目録「J-BISC」に書籍ごとに付与されているNDCの第1次区分（0～9）に、NDCが付与されていない「記録なし」を加えた、11分類。
  - **出版年（20層）**： 書籍の出版年である1986年から2005年までの20分類。

サンプリング方法

- 母集団を220層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した10,551サンプルについて、NDCごとの内訳を、図3-5に示す。

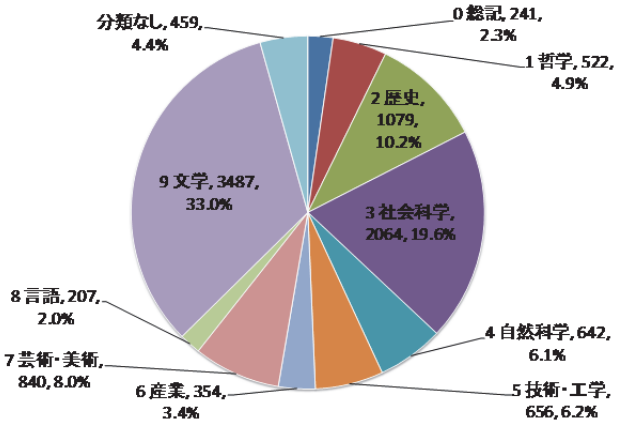


図3-5: サンプルの数と構成比率
（図書館SC「書籍」、NDC別）

3.6.5 特定目的 SC「白書」

特定目的SC「白書」は、1976年から2005年までの30年間に発行された政府系刊行物「白書」を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 2001年から2005年までに発行された白書のうち、『官報』に記載のあった白書タイトルを抽出した。これらについて、1976年以降、タイトルの変更や合併などの変遷を調査した。30年間にタイトルの変更や合併などがあったものは、まとめて扱った。例えば『土地白書』は、1989年以前は『国土利用白書』という別タイトルだったが、これは『土地白書（国土利用白書）』という1タイトルにまとめた。この結果、合計で40タイトル、1,006冊の白書が同定された。これらを特定目的SC「白書」の母集団として定義した。

層別方法

- 上記で定義した母集団を、以下の二つの基準により、合計54層に層別した。
  - **ジャンル（9層）**： 白書の内容に基づいて設定した、「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」という9分類。
  - **発行年（6層）**： 白書の発行年（1976年～2005年）の30年間を5年刻みにした、6分類。

    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年、
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年

サンプリング方法

- 第1期から第6期のそれぞれから250サンプルずつ、全体で1,500サンプル（約500万語）の取得を計画した。40タイトルごとに総ページ数を集計し、1,500サンプルに比例割当して、各期・各タイトルから取得するサンプル数を算出した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された白書の指定されたページに含まれる文章を一定の手続きにより抽出した。
- 収録した1,500サンプルについて、ジャンルごとの内訳を、図3-6に示す。

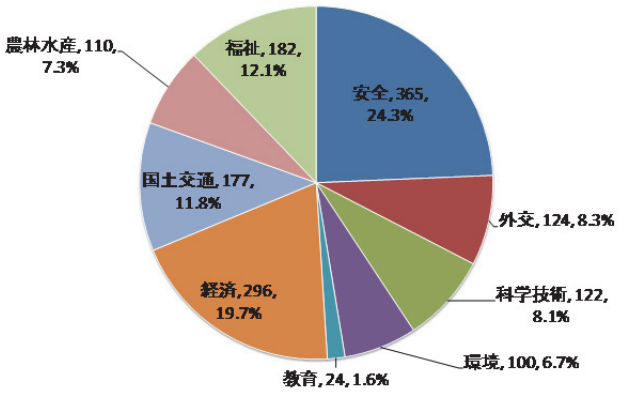


図3-6: サンプルの数と構成比率
（特定目的SC「白書」、ジャンル別）

3.6.6 特定目的 SC「教科書」

特定目的SC「教科書」は、小学校・中学校・高等学校で採用された各教科の教科書から、ランダムにサンプルを抽出したものである。

母集団の定義

- 小学校・中学校・高等学校の各学習指導要領（平成10～11年文部省告示、平成15年一部改正）に基づき、2005年度から2007年度に実際に使用された検定教科書を対象とした。ただし、専門に分化した高等学校の一部の科目（「農業」「商業」など）は除外した。
- 各校種・各学年・各教科から1種ずつの教科書を選出した。その際、できるだけ発行部数の多い教科書から順に選出した。この結果、145冊の教科書（推計総文字数7,859,456文字）が同定された。これらを、特定目的SC「教科書」の母集団として定義した。

層別方法

- 以下の二つの基準により、母集団を合計25層に層別した。
  - **教科（10層）**：「国語」「数学」「理科」「社会」「外国語」「技術家庭」「芸術」「保健体育」「情報」「生活」の10分類。
  - **校種（3層）**：「小学校」「中学校」「高等学校」の3分類。

サンプリング方法

- 母集団を25層に層別し、全体に対する各層の構成比率を取得サンプル数に比例割当した。
- 各層に含まれる全ページに対してランダムに優先順位を割り振った。優先順位の高い順に、指定された教科書の指定されたページに含まれる文章を一定の手続きにより抽出した（ただし、教科書であることを考慮し、書籍等の基準とは一部異なっているところがある）。
- 収録した412サンプルについて、教科ごとの内訳を、図3-7に示す。

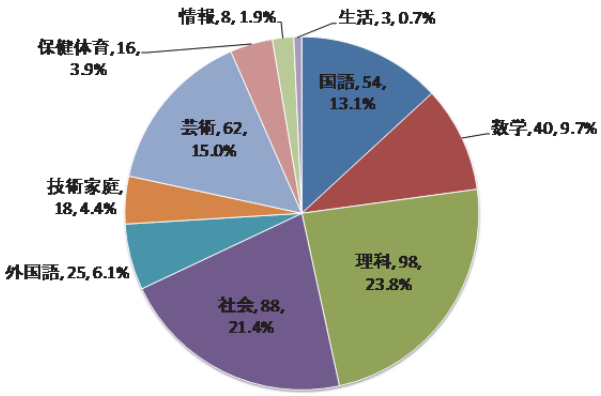


図3-7: サンプルの数と構成比率
（特定目的SC「教科書」、教科別）

3.6.7 特定目的 SC「広報紙」

特定目的SC「広報紙」は、日本の地方自治体において発行されている「広報紙」から、ランダムにサンプルを抽出したものである。

母集団の定義

- 全国各地から地域や人口構成比などを考慮して100の自治体（区市町村）を抽出し、その100自治体で2008年度に発行された広報紙を母集団として定義した。

層別方法

- 広報紙が発行している自治体の地域に応じて、母集団を合計8層に層別した。
  - **地域（8層）**： 北海道地方、東北地方、関東地方、中部地方、近畿地方、中国地方、四国地方、九州・沖縄地方

サンプリング方法

- 1自治体から6万字程度を取得することにした。入手した各自治体の広報紙からランダムに1冊（1号）を選び、そこに含まれる全文をサンプルとして取得した。
- また、著作権処理の観点から、外部著者による「寄稿」や、個人情報（一般人の氏名や住所、電話番号など）に相当する部分は伏せ字処理を実施した。
- 各自治体で6万字程度が取得できるまで、冊の取得を繰り返した結果、354サンプルを取得した。地域ごとの内訳を、図3-8に示す。

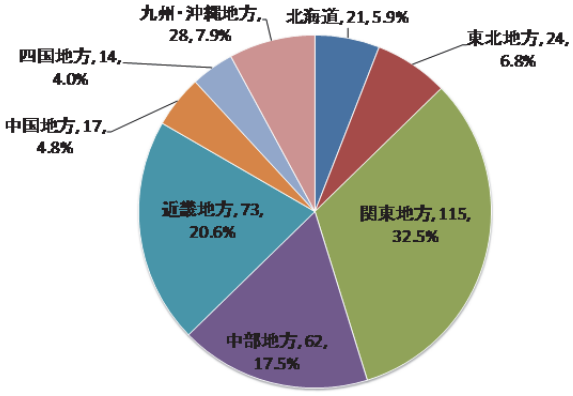


図3-8: サンプルの数と構成比率
（特定目的SC「広報紙」、地域別）

3.6.8 特定目的 SC「ベストセラー」

特定目的SC「ベストセラー」は、1976年から2005年までの30年間にベストセラーとなった書籍を対象として、ランダムにサンプルを抽出したものである。

母集団の定義

- 1976年から2005年までの30年間において、『出版年鑑』（出版ニュース社）および『出版指標年報』（全国出版協会出版科学研究所）のどちらかに、各年のベストセラーとして上位20位までに挙げられた書籍を調査した。その結果、951冊が同定された。これ

らを特定目的SC「ベストセラー」の母集団として定義した。

- なお、1971年に出版された本が1976年のベストセラーになったなど、出版年とベストセラーになった年との間にずれがあるものがある。

層別方法

- 「ベストセラー」という性格上、層別は実施しなかった。

サンプリング方法

- 1冊からランダムに2サンプルずつを取得することにした。
- 各冊に含まれる全ページに対して、ランダムに優先順位を割り振った。優先順位の高い順に、指定された書籍の指定されたページを開け、そこに印刷されている文章を一定の手続きにより抽出した。
- 951冊からは、合計1,902サンプルが取得できることになるが、作業上の理由（サンプリングできる箇所がない、当該の書籍が入手できないなど）により、すべてが取得できたわけではない。
- 収録した1,390サンプルについて、NDCごとの内訳を図3-9に示す。

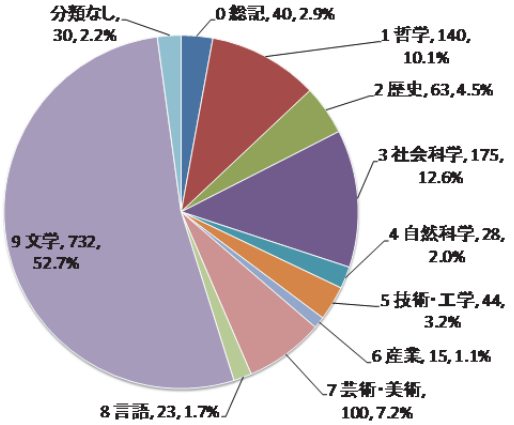


図3-9: サンプルの数と構成比率
（特定目的SC「ベストセラー」、NDC別）

3.6.9 特定目的 SC「Yahoo!知恵袋」

特定目的SC「Yahoo!知恵袋」は、Q&A形式のナレッジコミュニティサービス「Yahoo!知恵袋」の投稿データからランダムにサンプルを抽出したものである。

母集団の定義

- 「Yahoo!知恵袋」の元データには、2004年10月から2005年10月にかけて投稿された3,120,839の質問と、それに対する複数の回答が含まれていた。これらを、特定目的SC「Yahoo!知恵袋」の母集団として定義した。

層別方法

- 「Yahoo!知恵袋」の質問は、その質問内容に応じて、ある「カテゴリ」に分類されている。カテゴリは、15個の大カテゴリ・82個の中カテゴリ・279個の小カテゴリという

3階層に分かれている。このうち、小カテゴリによって、母集団を合計279の層に層別した。

サンプリング方法

- 母集団から、ひとつの質問とそれに対するひとつの回答の組を抽出して1サンプルとすることにした。複数の回答がある場合、「ベストアンサー」と呼ばれる回答を利用した。
- 全体で約1,000万語分のサンプルを取得することとし、1サンプルの平均長を試算して、対象データ全体から91,450サンプルを取得することを計画した。
- 279の各層に含まれる質問数を集計し、91,450サンプルに比例割当して、各小カテゴリから取得するサンプル数を算出した。この結果、取得対象となるのは14個の大カテゴリ、59個の中カテゴリ、130個の小カテゴリとなった。
- 各小カテゴリに含まれる質問から必要数をランダムに取得し、その質問に対する回答も同時に取得して、全体で91,445サンプルを取得した。大カテゴリごとの内訳を、図3-10に示す。

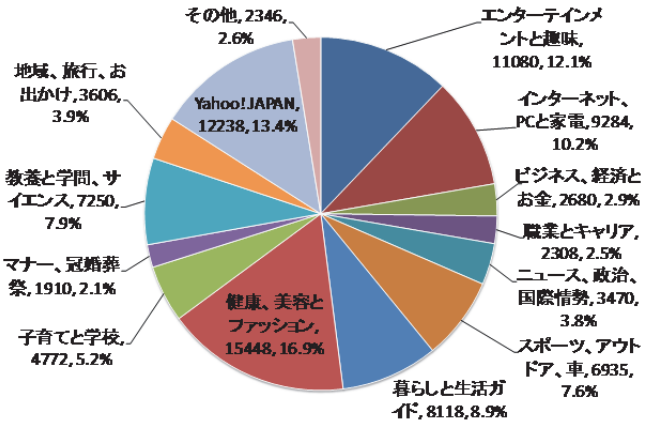


図3-10: サンプルの数と構成比率

(特定目的SC「Yahoo!知恵袋」、大カテゴリ別)

3.6.10 特定目的 SC「Yahoo!ブログ」

特定目的SC「Yahoo!ブログ」は、「Yahoo!ブログ」の記事データからランダムにサンプルを抽出したものである。

母集団の定義

- 「Yahoo!ブログ」の元データには、合計3,463,413の記事(ただし、以下の条件を満たすもの)が含まれていた。これらを、特定目的SC「Yahoo!ブログ」の母集団として定義した。
  1. 2008年4月26日から2009年4月25日までに投稿された記事。
  2. 抽出時点で1,000記事以上あるブログからの記事。
  3. 抽出時点で1か月以上掲載されており、かつ「公開」モードである記事。
  4. 転載(Yahoo!ブログ内のほかの記事の内容をコピーして、自分のブログに掲載す

ること）による記事は除外する。

5. ひとつの記事が全角20文字以下のものは除外する。

層別方法

- 「Yahoo!ブログ」の記事は、その内容に応じて、ある「カテゴリ」に分類される。カテゴリは、15個の大カテゴリ・54個の中カテゴリ・316個の小カテゴリという3階層に分かれているが、事前の層別には用いなかった。

サンプリング方法

- 全体で約1,000万語分のサンプルを取得することとした。サンプルは、記事タイトルやトラックバックを含まない、記事本文として記述されたテキストのみで構成するものとした。
- 対象データ全体を、投稿日時によって記事ごとに並び替え、等間隔サンプリングによって全体の1.8%を抽出した。
- ここから広告のみからなる記事などを除外した。結果、「ブログ」として、52,680サンプルを取得した。大カテゴリごとの内訳を、図3-11に示す。

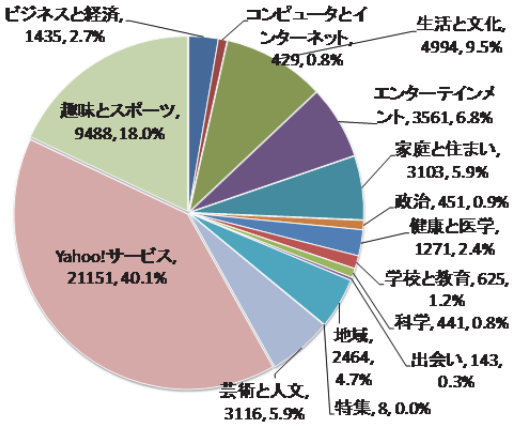


図3-11: サンプルの数と構成比率

（特定目的SC「Yahoo!ブログ」、大カテゴリ別）

3.6.11 特定目的 SC「韻文」

特定目的SC「韻文」は、短歌・俳句・詩の3種類について、代表的な作品からサンプルを抽出したものである。

母集団の定義

- 以下の作品を母集団として定義した。
  - 短歌: 『現代短歌全集』（筑摩書房、2002年刊） 第14巻～第17巻
  - 俳句: 『増補現代俳句大系』（角川書店、1980年～1982年刊） 第8巻～第15巻
  - 詩: 「現代詩文庫」シリーズ（思潮社、1986年～2005年刊） 118冊

層別方法

- 「短歌」「俳句」「詩」という3種類によって層別した。

サンプリング方法

- 短歌・俳句・詩からそれぞれ約5万語ずつを取得することとし、各歌集・句集・詩集からほぼ等量ずつのサンプルを抽出した。
- 収録した252サンプルについて、内訳を図3-12に示す。

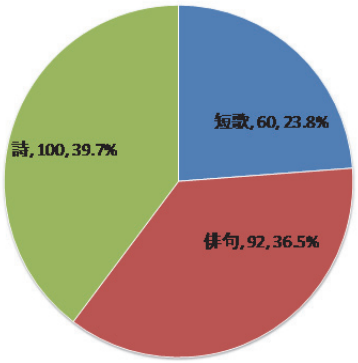


図3-12: サンプルの数と構成比率
（特定目的SC「韻文」）

3.6.12 特定目的 SC「法律」

特定目的SC「法律」は、1976年から2005年までの30年間に公布され、2009年時点でも施行されているすべての法律を対象として、そこからランダムにサンプルを抽出したものである。

母集団の定義

- Web上の「法令データ提供システム」（http://law.e-gov.go.jp/）から、1976年から2005年までの間に公布され、2009年9月の時点でも施行されている法律を検索したところ、718法律を得た。これらを特定目的SC「法律」の母集団として定義した。

層別方法

- 公布年により、母集団を合計6層に層別した。
  - **公布年（6層）**： 1976年から2005年までの30年間を5年刻みにした6分類。
    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年、
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年

サンプリング方法

- 第1期から第6期のそれぞれから30万文字ずつを取得し、約100万語分のサンプルを取得した。
- 各層に含まれる全法律に対して、それぞれ200箇所の文字を優先順位付きでランダムに選び、その文字を基準にして1万字を超えない一定範囲（条、節など）を取得した。その際、公布時以降に付け加えられた「附則」は取得の対象外とした。
- 収録した346サンプルについて、ジャンルごとの内訳を表3-2に示す。

表3-2: 取得したサンプルの数

（特定目的SC「法律」、ジャンル別）

憲法	2	国土開発	5	文化	2	航空	1
国会	3	土地	1	産業通則	18	貨物運送	3
行政組織	22	都市計画	7	農業	11	郵務	4
国家公務員	3	道路	1	林業	5	電気通信	13
行政手続	1	災害対策	6	水産業	3	労働	9
地方自治	4	建築・住宅	8	鉱業	2	環境保全	12
地方財政	1	財務通則	4	工業	10	厚生	17
司法	5	国税	18	商業	13	社会福祉	15
民事	36	専売・事業	4	金融・保険	40	防衛	1
刑事	7	国債	3	陸運	11	外事	6
警察	4	教育	3	海運	4	合計	346

3.6.13 特定目的 SC「国会会議録」

特定目的SC「国会会議録」は、1976年から2005年までの30年間における「国会会議録」からランダムにサンプルを抽出したものである。

母集団の定義

- Web上の「国会会議録検索システム」（http://kokkai.ndl.go.jp/）で公開されているデータのうち、第77回国会から第163回国会までに開かれた32,986会議の会議録データを特定目的SC「国会会議録」の母集団とした。
- このうち、「両院協議会」で開かれた61会議、発言部分の文字数が1,000文字以下の6,401会議、第77回国会のうち1975年に開催された33会議は除外した。

層別方法

- 以下の三つの基準により、母集団を合計48層に層別した。
  - **開催院（2層）**： 「衆議院」「参議院」による、2分類。
  - **開催時期（6層）**： 1976年から2005年までを5年刻みにした6分類。
    第1期：1976～1980年、第2期：1981～1985年、第3期：1986～1990年
    第4期：1991～1995年、第5期：1996～2000年、第6期：2001～2005年
  - **会議種別（4層）**： 「常任委員会」「特別委員会」「本会議」「その他」による4分類。

サンプリング方法

- 全体で約500万語分のサンプルを取得することを計画した。1サンプルは、1会議に含まれる発言部分のみで構成することにした。
- 48の各層に含まれる発言文字数を集計し、各層から取得するサンプル数を比例割当により算出した。各層に含まれる会議から必要数をランダムに取得し、全体で159サンプルを取得した。
- 収録した159サンプルについて、開催院・会議種別ごとのサンプル数と構成比率を図3-13に示す。

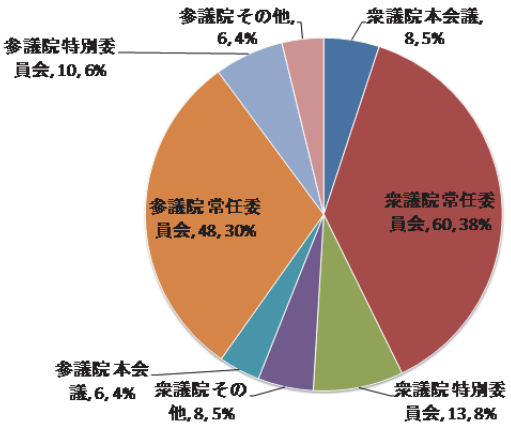


図3-13:サンプルの数と構成比率

（特定目的SC「国会会議録」、開催院・会議種別）

参考文献

柏野和佳子・丸山岳彦・稲益佐知子・田中弥生・秋元祐哉・佐野大樹・大矢内夢子・山崎誠　(2009)「『現代日本語書き言葉均衡コーパス』における収録テキストの抽出手順と事例」国立国語研究所内部報告書 LR-CCG-08-01.

丸山岳彦・秋元祐哉　(2007)「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法　－現代日本語書き言葉の文字数調査－」国立国語研究所内部報告書 LR-CCG-06-02.

丸山岳彦・秋元祐哉　(2008)「『現代日本語書き言葉均衡コーパス』におけるサンプル構成比の算出法(2)　－コーパスの設計とサンプルの無作為抽出法－」国立国語研究所内部報告書 LR-CCG-07-01.

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子　(2011)「『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用」国立国語研究所内部報告書 LR-CCG-10-01.

第4章　文書構造情報付き文字ベース XML（C-XML）

山口 昌也

4.1　はじめに

本章では、文書構造情報を付与した文字ベースの（形態論情報を含まない）XML 文書（Character-base XML、以下 C-XML と略記する）の仕様について、(1)文書構造タグ、(2)文字入力、(3)形態論情報付き XML 文書（M-XML：第 9 章参照）との相違点の三つに分けて説明する。なお、本章の内容の詳細については、山口他（2011）、西部他（2011）を参照されたい。

4.2　文書構造タグセットの種類とサブコーパス・レジスターとの関係

BCCWJ は複数のサブコーパス・レジスターから構成される。文書構造タグのセット（タグセット：TS）は、それぞれのサブコーパス・レジスターの特性に合わせて、表 4-1 のように規定される。個々のタグセットは、XML の文書型として定義される。なお、原資料が紙媒体のデータについては、sentence（後述、表 4-2 参照）など一部の要素を除き人手で付与しているが、電子媒体のデータについては、より多くの部分で自動付与を行うなど、個々に方法が異なる。同じタグであってもレジスターの種類によってタグの性質や付与の精度に差が生じることがあるため、注意が必要である。タグ付与方法の詳細については西部他（2011）を参照されたい。

タグセットは、次の 3 種類に大別される。表中で「可変長（一部修正）」とあるのは、可変長タグセットに部分的な変更を加えたタグセットであることを意味する。この後の節では、まず「可変長タグセット」「固定長タグセット」「Yahoo!知恵袋タグセット」について解説し、そのあとレジスターごとに個別の変更部分を説明する。

可変長タグセット（可変長 TS）：　可変長サンプル（ひとつのサンプルがひとつの「記事」に相当するサンプル）を記述するためのタグセット

固定長タグセット（固定長 TS）：　固定長サンプル（ひとつのサンプルに 1,000 文字を包含するサンプル）を記述するためのタグセット

Yahoo!知恵袋タグセット（Yahoo!知恵袋 TS）：　「Yahoo!知恵袋」レジスターのサンプルを記述するためのタグセット

表 4-1: サブコーパス・レジスターとタグセットとの関係

サブコーパス･レジスター	タグセット	原資料の媒体
出版サブコーパス（PB,PM,PN）	可変長 TS、固定長 TS	紙媒体
図書館サブコーパス（LB）	可変長 TS、固定長 TS	紙媒体
白書(OW)	可変長 TS、固定長 TS	紙媒体
教科書(OT)	可変長 TS（一部修正）	紙媒体
広報紙(OP)	可変長 TS	電子媒体
ベストセラー(OB)	可変長 TS	紙媒体
Yahoo!知恵袋(OC)	Yahoo!知恵袋 TS	電子媒体
Yahoo!ブログ(OY)	可変長 TS（一部修正）	電子媒体
韻文(OV)	可変長 TS（一部修正）	紙媒体
法律(OL)	可変長 TS	電子媒体
国会議事録(OM)	可変長 TS	電子媒体

4.3 可変長タグセット

可変長タグセットは、可変長サンプル（ひとつのサンプルがひとつの「記事」に相当するサンプル）を記述するためのタグセットである。タグの種類は、46 種類である。タグの一覧を表 4-2 に示す。また、紙媒体の原資料とタグづけ結果の例を図 4-1 に示す。

本タグセットによって付与される情報は、次の三つに大別される。

- サンプルに関するタグ：サンプルに関するタグには、sample と sampling がある。sample 要素は、ひとつのサンプルの範囲を表す。sampling タグは、サンプル抽出基準点などサンプリングに関する情報を表す。
- 文字・表記に関するタグ：この種のタグの役割は、(1)検索や計算機処理の利便性を高めること、(2)原資料に忠実に電子化テキストを記述することである。前者のタグの例として、correction タグ（誤植を修正した文字を表す）がある。

```
生活基<correction type="erratum" originalText="盟">盤</correction>に
伸びを示し<correction type="omission">て</correction>いる
整備を<correction type="excess" originalText="を" />図るべく
```

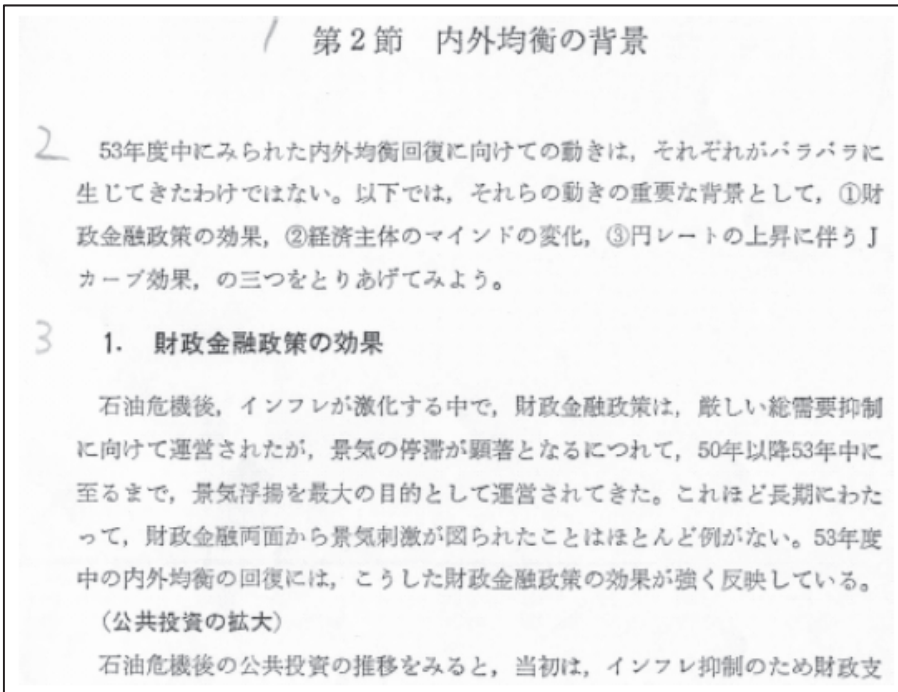
後者の例として、ruby タグ（ルビ付き文字を表す）、missingCharacter タグ（文字セット外字を表す）の例を次に示す。

```
<ruby rubyText="ご">語</ruby><ruby rubyText="い">彙</ruby>
<missingCharacter attribute="HanIdeograph" unicode="U+5AEB"
        daikanwa="M06673" description="女偏に莫">＝</missingCharacter>
```

- 文書構造に関するタグ：文書構造に関するタグは、見出し、概要、キャプション、注記など、文書中における論理的な役割が明確な文書要素に対して付与される。表 4-2 に示したとおり、この種のタグは、(a) 階層構造、(b) 図表、(c) 引用、(d)注記、(e)その他に分けられる。

  このうち、階層構造に関するタグについて、図 4-1 と対応づけて説明する。階層構造に関するタグは、article を最上位の階層として、cluster、paragraph、sentence といった言語的な階層構造を表現する。図 4-1 から、これらの要素に関係する部分を取り出すと次のようになる。なお、字下げは、下位の階層であることを示す。例えば、図 4-1 の article 要素直下の階層には、titleBlock 要素、paragraph 要素、cluster 要素があることがわかる。

```
article
  titleBlock 第２節　内外均衡の背景
  paragraph
  cluster
    titleBlock １．財政金融政策の効果
      cluster
        titleBlock （公共投資の拡大）
```



第２節　内外均衡の背景

53年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。以下では，それらの動きの重要な背景として，①財政金融政策の効果，②経済主体のマインドの変化，③円レートの上昇に伴うＪカーブ効果，の三つをとりあげてみよう。

1.　財政金融政策の効果

石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，50年以降53年中に至るまで，景気浮揚を最大の目的として運営されてきた。これほど長期にわたって，財政金融両面から景気刺激が図られたことはほとんど例がない。53年度中の内外均衡の回復には，こうした財政金融政策の効果が強く反映している。

（公共投資の拡大）

石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支

```
<?xml version="1.0" encoding="UTF-8" ?>
<?xml-stylesheet href="sc_check.xsl" type="text/xsl" ?>
<sample sampleID="OW1X_00000" version="20070208" type="variableLength">
<article articleID="OW1X_00000_V001" isWholeArticle="false">
<titleBlock><title><sentence type="quasi">第２節　内外均衡の背景
</sentence></title></titleBlock>
<paragraph>
<sentence>　５３年度中にみられた内外均衡回復に向けての動きは，それぞれがバラバラに生じてきたわけではない。</sentence><sentence>以下では，それらの動きの重要な背景として，...
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">１．財政金融政策の効果
</sentence></title></titleBlock>
<paragraph>
<sentence>　石油危機後，インフレが激化する中で，財政金融政策は，厳しい総需要抑制に向けて運営されたが，景気の停滞が顕著となるにつれて，５０年以降５３年中に至るまで，景気浮揚を最大の目的として運営されてきた。</sentence> ...
</paragraph>
<cluster>
<titleBlock><title><sentence type="quasi">（公共投資の拡大）</sentence></title></titleBlock>
<paragraph>
<sentence>　石油危機後の公共投資の推移をみると，当初は，インフレ抑制のため財政支出が抑制され，公共事業の伸びは低いものにとどまっていた。</sentence>
```

図 4-1: 原資料とその電子化テキストの例(『経済白書昭和 54 年版』から引用)

表 4-2: 可変長タグセット

	タグ名	内容
サンプル	sample	サンプリングによって 1 サンプルとされた文書要素
	sampling*	サンプル抽出基準点などサンプリングに関する情報
階層構造（文書構造）	article	同一著者による、同一テーマのひとまとまりの文書要素
	blockEnd	意味のまとまりや形式のまとまりを区切るためのマーカー
	cluster	titleBlock 要素が包括する文書要素全体
	titleBlock	title 要素とそれに付随する要素全体
	title	特定範囲の文書要素の内容を代表する記述
	orphanedTitle	不特定範囲の文書要素を代表する記述
	list	箇条書きなど、列挙された文書要素の集まり
	listItem	List 要素を構成する各並立要素
	paragraph	段落を表す文書要素
	sentence*	文に相当する文書要素
図表（文書構造）	figureBlock	図表・写真・絵などの要素と、それに付随する文書要素をまとめた要素
	figure	付随する文書要素のある図・表・写真・絵など
	caption	図表についてのタイトルや説明
	table	表
引用（文書構造）	quotation	当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし描写・書き起こし図表・写真・絵などの要素と、それに付随する文書要素をまとめた要素
	citation	当該 article 要素の本文において言及される、他文献からの引用要素
	source	引用文献についての情報(文献名、著者名、著者情報など)
	speech	発話の引用・書き起こし、心内発話の描写
	speaker	話者を明示的に表した文字列やマーク
	quote*	当該 article 要素とは異なる著作物からの引用や、発話・心内発話の引用・描写・書き起こし、「 」で表されるさまざまな表現
注記（文書構造）	noteBody	注記とその注記の範囲
	noteBodyInline*	傍注など行外に付随する形式で現れる注記
	noteMarker*	注番号や参考文献番号など、他の文書要素を参照する際の目印として機能する文字列
その他（文書構造）	abstract	article 要素、または cluster 要素の概要に相当する文書要素
	authorsData	著作者表示・署名にあたる要素
	contents	目次に相当する文書要素
	profile	著者や登場人物のプロフィールに相当する文書要素
	rejectedBlock	サンプル範囲内において、削除対象となったブロック要素の存在
	verse	詩、和歌、俳句、歌謡などの韻文
	verseLine	韻文における行
文字・表記*	ruby	ルビ付き文字
	correction	原文の誤植を訂正した文字
	missingCharacter	JIS X 0213:2004 で規定されている文字以外の文字（JIS 外字）
	enclosedCharacter	連続や参照などのラベルとして機能している囲み付きの文字
	cursive	変体仮名
	image	JIS X 0213:2004 が規定する諸記号に含まれていない記号類や絵文字
	superScript	数式や化学式などに用いる上付きの文字
	subScript	数式や化学式などに用いる下付きの文字
	fraction	帯分数の中の真分数部分
	delete	抹消線などによって削除された本文要素
	br	物理改行
	info	補助的な付与情報
	rejectedSpan	サンプル範囲内において、削除対象となったインライン要素の存在
	substitution	別の文字で代用入力されている JIS X 0213:2004 規定文字

※ 表中「*」付きの要素はインライン要素、それ以外の要素はブロック要素。

4.4 固定長タグセット

固定長タグセットは、固定長サンプル（ひとつのサンプルに 1,000 文字を包含するサンプル）を記述するためのタグセットである。可変長のタグセットとほぼ同じ仕様だが、固定長サンプルの収録範囲（文字数を基準に文を単位として限定される）に起因して、次の違いがある。

- 固定長タグセットのブロック要素は、当該要素の定義を満たす要素をすべて含むとは限らない。例えば、可変長タグセットにおける article 要素は「同一著者による、同一テーマのひとまとまりの文書要素」と定義され、記事や章などのまとまった文章範囲に相当するが、固定長の article 要素では、文章のまとまり全体を含まず、titleBlock 要素以外の本文が含まれない場合などもある。
- cluster 要素は認定しない。
- article 要素の isWholeArticle 属性は、IMPLIED （任意）である。

4.5 Yahoo!知恵袋タグセット

「Yahoo!知恵袋」レジスターのサンプルは、質問と回答の組という、一定の論理構造で構成される。しかし、可変長、固定長タグセットでは、この構造を十分記述することができないため、独立した文書型として定義した。タグの種類は、9 種類である。タグの一覧を表 4-3 に示す。また、サンプル例を図 4-2 に示す。

4.6 その他のタグセット

表 4-1 に示したとおり、レジスターの中には可変長タグセットを一部修正して記述しているものも含まれる。ここでは、可変長タグセットとの差異について説明する。

- Yahoo!ブログ
  - rejectedBlock タグの type 属性に ASCIIArt を追加した。これは、サンプル作成時に削除された、いわゆる「アスキーアート」を表す。
- 韻文
  - sample 要素の子要素に複数の article 要素を持つ。これは、「韻文」レジスターのサンプルには、1 サンプルに複数の作品が並列に含まれるためである。なお、可変長タグセットでは、sample 要素の子要素として、article 要素をひとつしか持たない。
- 教科書
  - 可変長タグセットに 5 種類のタグを追加するなど、「教科書」レジスター用に拡張している。詳細は、田中他（2011）「II 教科書コーパスの文字入力・タグ使用」を参照のこと。

表 4-3: 「Yahoo!知恵袋」レジスタータグセット

タグ名	内容
sample	質問本文と回答本文を対にしたもの
OCQuestion	質問本文を表す
OCAnswer	回答本文を表す
br	改行を表す
webLine	Web データに対して、自動で付与される、論理行相当の行を表す
sentence	文に相当するまとまりを表す
rejectedBlock	削除要素を表す
ncr	変換元データの数値文字参照を削除、または「=」に置換したことを表す
info	補助的な付与情報

```
<?xml version="1.0" encoding="UTF-8"?>
<sample sampleID="OC01_03216" type="chiebukuro" version="1.0">
<OCQuestion>
<webLine>
<sentence>ｗｉｎｄｏｗｓのＣＭで「税理士Ａの事件ファイル」という漫画をｗｅｂ上で公開しています、という男性が出ていますが、あのＣＭはフィクションですか？</sentence>
<sentence type="quasi">検索かけても出てきませんでした・・・・</sentence>
</webLine>
</OCQuestion>
<OCAnswer>
<webLine>
<sentence>税理士役も俳優さんらしいし、<br type="physicalLine_original" />完全なフィクションでしょう・・・。</sentence>
</webLine>
</OCAnswer>
</sample>
```

図 4-2: 「Yahoo!知恵袋」レジスターのサンプル例

4.7 文字入力仕様

本節では、BCCWJ に収録するデータを紙媒体（表 4-1）から作成する際の文字入力に関する仕様について述べる。なお、原資料が電子媒体のデータについては、データの性質上、この仕様に準拠しない点もある。詳細については、西部他（2011）の第 3 章を参照されたい。

4.7.1 基本方針

文字入力は、以下の基本方針に基づき行なった。

- 装飾、レイアウトなどの図形的情報を除いて文字を入力する（レイアウトの情報は、必要に応じて、タグで表現する）。

- 全ての文字種の入力に、いわゆる全角文字を用いる。
- 文字合成は行わない。
- 上記条件に抵触しない範囲で、原則として、原文を忠実に転記する。

4.7.2 文字符号化方式と文字集合

文字符号化方式は、以下に述べる文字集合を適切に符号化でき、テキストデータに対して施す形態素解析環境に適した方式として、UTF-8（BOM なし）を採用する。

文字集合は、JIS X 0213:2004 を用いる。ただし、次の文字については例外とし、それぞれ独自の方法で処理する。具体的な処理方法は、山口他（2011）を参照のこと。

- 入力対象外要素を構成する文字（例：ソフトハイフン、罫線素片）
- 装飾・デザインにかかわる文字（例：組み文字、分数、11 以上のローマ数字、囲み文字、上付き文字）
- 類似の非漢字
- 合成文字
- 入力が困難な文字（例：口偏に「七」の文字（「叱」面区点：1-47-52））

4.7.3 包摂規準

- 漢字
  - JIS X 0213 に準拠する。JIS X 0213:2000「6.6.3.1 漢字の字体の包摂規準の適用」（日本工業標準調査会 2000 参照）における包摂規準が適用される異体字については、これを区別しない。
- JIS X 0213 に定義されていない記号
  - JIS X 0213 に定義されていない記号であっても、原文の意味を損なわない場合、規格内の類似する記号に包摂してよいこととする。
- JIS X 0213 に定義されている記号
  - 字形の判別が困難な「長音記号」「負記号」「ダッシュ」「ハイフン」については、紙面上の形状ではなく、紙面上の意味によって入力し分けた。
  - その他の類似記号は独自に包摂規準を設けた。

4.7.4 外字

- 漢字、仮名、アルファベット
  - 漢字、仮名、アルファベットの JIS 外字は、当該の文字の代替として「**=**」（ゲタ）を入力すると共に、missingCharacter タグを用いて、タグ内部に属性として文字の情報を表す。

- 一般記号類
  - 入力対象外とする。ただし、語や文の構成要素になっているものについては、記号の代替として、image タグを挿入し、タグ内部に属性として記号の情報を表す。

4.7.5 特殊表記

- ルビ：ruby タグの rubyText 属性値として記述する。
- 上付き・下付き文字：それぞれ、superScript、subScript 要素として記述する。
- 囲み文字：囲みを無視して、囲まれている内部の文字を入力する。なお、連続・参照ラベルとして機能するもの（丸付き数字など）や、ある特定の語の略記号として機能するもの（「秘密」の意を表す丸付きの「秘」など）については、囲みの情報を、enclosedCharacter タグによって表す。
- 組み文字：組まれている文字をすべて 1 字ずつ切り離して入力する。
- 分数：「分子／分母」の形式に統一して入力する。ただし、帯分数の場合は、fraction 要素として記述する。
- 注記参照マーカー：「専門用語[2]」の上付きの「2」のような本文行から外れた位置にある注記参照用のマーカーは、noteMarker タグを付与する。
- 傍注：本文行の語や句の脇（行間など）に、注記が示されている「傍注」は、注記対象の語句の直後に、noteBodyInline タグを付与して示す。

4.7.6 レイアウト

- 空白
  - 入力対象となるもの：版面に現れる空白は、以下の場合に入力対象とする。その際、空白文字は常に 1 字分のみを入力する。
    - 段落冒頭の 1 字下げ
    - 語や文の区切り目を表すための空白
    - 「？」「！」などの後ろに挿入される空白
  - 入力対象とならないもの：上記以外の空白は、全てレイアウトによるものとみなし、無視する。例えば、以下のようなものをレイアウトとして入力対象としない。
    - 引用文、例文、項目等を本文行と区別するためのインデント
    - 中央揃え・右揃え・下揃え等の配置に伴うインデント
    - 文字幅を調整するためのスペース
- 改行

  改行は、版面の行の折り返しではなく、論理行（論理的に意味のある行。段落など意味のある切れ目で改行が施された行）で行う。具体的には、以下の要素の前後に改行を入れる。
  - 版面の行替えと一致する場合に改行するもの

- ✧ 段落
- ✧ 引用
- ✧ 韻文における行

➢ 版面の行替えと一致しない場合でも改行するもの

- ✧ タイトル
- ✧ 表の各セル

● リーダー・ダッシュ

リーダー・ダッシュが複数連続するものについては、すべて 1 字に置き換える。

4.7.7 誤植

原文に明らかな誤植がある場合は、これを訂正して入力する。ただし、原文の誤植を訂正した文字は、correction タグを用いて示し、原文の情報をタグ内部に originalText 属性として表す。以下に例を示す。

原文：

総トン数１００トン未満で長さ３０メートル<u>末</u>満の

タグづけ、および、修正：

総トン数１　０　０　トン未満で長さ３　０　メートル<correction type="erratum" originalText="末">未</correction>満の

なお、明らかな誤植とは、近似の字形の文字を誤って写植したもの（誤字）、前後の文字を逆に写植したもの（転倒）、脱字、衍字を指す。誤用や表記のゆれ、旧仮名遣い、仮名遣いの誤りなどは、これに含めない。詳細は、山口他（2011）を参照のこと。

4.8 M-XML との相違点

C-XML は、BCCWJ-DVD 版（Version 1.0）から変更されていない。特に、BCCWJ-DVD 版（Version 1.1）で加えられた文関連の修正が適用されていないため、次の点において、M-XML （第 9 章参照）と内容的に相違が生じている。利用する際は、注意されたい。

- C-XML では、文認定基準は Version 1.0 と同一であり、新規に追加された文認定基準が適用されていない。
- C-XML では、Version 1.0 と同様、文区切りを自動的に行っており、人手修正を行っていない。
- C-XML では、文認定の人手修正に伴い発見された文書構造タグの誤りが修正されていない（Version1.1 の M-XML では修正済み。8.3.1 節参照）。

参考文献

山口昌也・高田智和・北村雅則・間淵洋子・大島一・小林正行・西部みちる （2011） 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「『現代日本語書き言葉均衡コーパス』における電子化フォーマット ver.2.2」

西部みちる・大島一・間淵洋子・小林正行・田島孝治・高田智和・山口昌也 （2011） 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「『現代日本語書き言葉均衡コーパス』における電子化テキストの構築」

田中牧郎・相澤正夫・斎藤達哉・棚橋尚子・近藤明日子・河内昭浩・鈴木一史・平山允子（2011）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「言語政策に役立つ，コーパスを用いた語彙表・漢字表等の作成と活用」

日本工業標準調査会（2000）『7 ビット及び 8 ビットの 2 バイト情報交換用符号化拡張漢字集合 JIS X 0213:2000』日本規格協会.

第5章　形態論情報

小椋 秀樹　冨士池 優美

BCCWJ には、他の章でも解説されているように種々のアノテーションが施されている。これらのアノテーションを利用することで、コーパスを活用した現代日本語の研究が、今後、大きく進展することが期待される。

これら種々のアノテーションのうち形態論情報については、BCCWJ に格納したサンプルの言語的特徴の解明に適した「長単位」とコーパスからの用例収集に適した「短単位」の 2 種類の言語単位に解析し、それぞれの単位に見出し・品詞・語種等の情報を付与した。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[1]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（非コアデータ）は 98%以上である。

本章では、BCCWJ-DVD 版、オンライン版（中納言）で形態論情報を活用しようとする研究者の用に資するため、形態論情報、特に長単位・短単位の認定基準を中心に、その概要を紹介する[2]。

5.1　BCCWJ の言語単位

本節では、まず国立国語研究所がこれまでに行ってきた語彙調査における調査単位を概観し、続いて BCCWJ の言語単位の設計方針、採用した長短 2 種類の言語単位の長所、短単位の自動解析に使用した形態素解析用辞書 UniDic について述べる。

5.1.1 語彙調査の調査単位

国立国語研究所は、これまでに、マスメディアにおける書き言葉や話し言葉を中心に、合計 10 回の大規模な語彙調査を実施してきた。この語彙調査に当たっては、当然、語というものを規定することが必要となる。しかし、語の定義については研究者によって様々な立場があるため、語彙調査において語（調査単位）をどのように規定するかということは常に大きな問題となる。

国立国語研究所がこれまでに行った語彙調査では、調査単位の設計に当たって、語とは何かという本質的な議論の上に立って調査単位を設計するという立場は取っていない。それぞれの語彙調査の目的に応じて最もふさわしい単位を設計するという方針の下に、一貫して操作主義的な立場を取ってきた。そのため、表 5.1 に示すように、複数の調査単位が使われてきた[3]。

---

1 コアデータについては、第 2 章を参照。
2 BCCWJ の形態論情報の詳細については、小椋他（2011）を参照。
3 単位の概略と例については、林（1982:582-583）、中野（1998:171-172）を基にした。

表 5-1: 国立国語研究所の語彙調査における主な調査単位

	単位の名称	語　彙　調　査　名
長い単位の系列	α単位	現代の語彙調査・婦人雑誌の用語
	W単位	高校教科書の語彙調査、中学校教科書の語彙調査
	長い単位	雑誌用語の変遷、テレビ放送の語彙調査
短い単位の系列	β単位	現代の語彙調査・総合雑誌の用語、現代雑誌九十種の用語用字、雑誌２００万字言語調査
	M単位	高校教科書の語彙調査、中学校教科書の語彙調査

表 5-1 に挙げた各調査単位の概略と例とを次に示す。

【調査単位の概略】

(1) 長い単位の系列 ： 主として構文的な機能に着目して考えた単位。おおむね文節に相当する。

α 単 位　文節を基にした単位。「｜小学校｜卒業｜」「｜男児用｜外出着｜」のように長い語を分割する規定を設けている。

W 単 位　非活用語および活用語のうち終止・連体形、命令形、中止用法・修飾用法の連用形を 1 単位とする。また、それらに接続する付属語も 1 単位とする。

長い単位　文節に相当する単位。「テレビ放送の語彙調査」の長い単位は、複合辞を助詞・助動詞として扱っていること、人名・地名のほか書名・番組名・商品名なども固有名詞として扱っていることから、「雑誌用語の変遷」で採用した長い単位よりも長くなっている。

(2) 短い単位の系列 ： 主として言語の形態的な側面に着目して考えた単位。

β 単 位　原則として、現代語において意味を持つ最小の単位（最小単位）二つが、文節の範囲内で一次結合したものを 1 単位とする。

M 単 位　β 単位と同様に最小単位を基にした単位。漢語は、β 単位と同様に二つの最小単位が文節の範囲内で一次結合したものを 1 単位とするが、和語・外来語は 1 最小単位を 1 単位とする。

【調査単位の例】

(1) 長い単位の系列

α単位：　型　紙｜どおり　に｜裁断　し　て｜外出　着　を｜作り　まし　た｜

W単位：　型　紙　どおり｜に｜裁断　し　て｜外出　着｜を｜作り　まし　た｜

長い単位 （雑誌用語の変遷） ：

型　紙　どおり　に｜裁断　し　て｜外出　着　を｜作り　まし　た｜

長い単位 （テレビ放送の語彙調査） ：

型　紙　どおり　に｜裁断　し　て｜外出　着　を｜作り　まし　た｜

その｜問題について｜検討している｜

(2) 短い単位の系列

β 単位：　型　紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜

M単位：　型｜紙｜どおり｜に｜裁断｜し｜て｜外出｜着｜を｜作り｜まし｜た｜

調査単位の設計に当たって操作主義的な立場を取ってきたのは、「必要以上に学術的な議論に深入りし、実際上の作業がすすまないことをおそれたため」（国立国語研究所 1987:12）であり、「学者の数ほどもある「単語」の定義について、まず、意見を一致させてから、というのでは、見とおしがたたない。」（同:12）からである。

このような立場に対しては、当然のことながら「語というのは何なのか、調査のため便宜的に設けられた単位にすぎないのかという問題が残る。」（前田 1985:740）という批判がある。確かに、語というものを定義しようとする以上、語とは何かという本質的な議論を積み重ねていくことは重要なことである。しかし、国立国語研究所（1987:12）に、「原則的にただしい定義に達したとしても、それが現実の単位きり作業に役立たないならば、無意味である。語い調査というのは、現象の処理なのだから。」と述べられているように、語彙調査においては対象とする言語資料に現れた個々の事象を、的確に処理することも極めて重要である。このことから、これまでの語彙調査では、語とは何かという本質的な議論よりも、言語現象を的確に処理することを重視してきた。

このような立場で、各種の語彙調査を進めてきたことにより、「同じ資料の語彙調査を短単位と長単位との両方で行ってみてどのような違いが出てくるかを検討したことなどは、単位の区切り方を曖昧にしたまま「語彙調査」を行なうことに対する反省を促す」（前田 1985:740）など、日本語の計量的な研究を進める上で先駆的な役割を果たしてきたと言うことができる。国立国語研究所の語彙調査における調査単位の設計方針には批判もあるが、それにより現実の言語現象を的確に処理してきたことは、十分に意味があったと言える。

5.1.2 BCCWJ の言語単位の設計方針

BCCWJ の言語単位の設計に当たっては、語彙調査における調査単位の設計と同様の立場を取った。つまり、まず BCCWJ を日本語研究に利用するために、どのような言語単位が必要か整理し、その上で設計方針を立て、その方針に基づいて言語単位を設計したのである。

このような立場を取ったのは、語とは何かという本質的な議論の重要性はもちろん認めるところではあるが、コーパス構築という実務を考えた場合、BCCWJ に現れる言語現象を

的確に処理できる単位を設計することの方が、より重要であると考えたからである。このようにして大規模なコーパスを処理した結果をまとめておくことは、今後、言語単位論を進める上での基礎的な資料になると考えられる。

我々は、BCCWJ の言語単位の設計方針として、次の三つを掲げた。

方針 1：コーパスに基づく用例収集、各ジャンルの言語的特徴の解明に適した単位を設計する。

コーパスの日本語研究への活用としてまず考えられるのは、コーパスから用例を集めることである。そのため、BCCWJ を日本語研究で幅広く利用できるようにするには、用例収集に適した単位を設計する必要がある。

また BCCWJ は、新聞・雑誌・書籍といった複数の媒体を対象としたコーパスであり、内容も政治・経済・自然科学・文芸等と多岐にわたっている。このような BCCWJ の構成から、媒体別・ジャンル別の言語的な特徴を明らかにしていくことが重要な研究テーマになると考えられる。したがって、そのような分析に適した単位を設計することが必要になる。

方針 2：『日本語話し言葉コーパス』と互換性のある形態論情報を設計する。

国立国語研究所が既に構築したコーパスとして、現代の話し言葉を対象とした『日本語話し言葉コーパス』（Corpus of Spontaneous Japanese、以下 CSJ とする。）がある[4]。CSJ、BCCWJ は共に現代日本語を対象とした大規模コーパスであり、日本語研究の立場からは、両コーパスを活用した現代日本語の話し言葉・書き言葉の研究を進めていくことが重要なテーマとなる。このような研究を進めるためには、CSJ と BCCWJ とを統一的に扱うことのできる互換性を持った単位を設計する必要がある。

方針 3：国立国語研究所の語彙調査における知見を活用する。

国立国語研究所は、1949 年の『語彙調査 ―現代新聞用語の一例―』以来、合計 10 回の語彙調査を実施した。その中で、調査単位の設計や言語現象の処理に関して、様々な知見を蓄積している。そこで、BCCWJ の言語単位の設計や単位認定の際に、これら語彙調査の知見を活用していく。語彙調査の結果は、日本語研究でも様々に活用されており、言語単位の設計等に語彙調査の知見を活用していくことは、BCCWJ を使った日本語研究を進めていくためにも有用であると考えられる。

5.1.3 BCCWJ の言語単位

以上の方針の下、BCCWJ の言語単位について検討した結果、次のような結論を得た。

BCCWJ の言語単位には、方針 1 で挙げた、用例収集・各ジャンルの言語的特徴の解明

[4] CSJ の言語単位の概要については、小椋（2006）を参照。

という二つの利用目的に応じて、次に示す 2 種類を採用する。

① 用例収集を目的とした短単位

② 言語的特徴の解明を目的とした長単位

この短単位・長単位は、いずれも CSJ で採用した言語単位である。また短単位は国立国語研究所が行った現代雑誌九十種調査の β 単位を、長単位はテレビ放送の語彙調査の長い単位を基に設計したものである。このようにして、CSJ との互換性の保持と、国立国語研究所の持つ語彙調査の知見の活用とを図る。なお、長単位・短単位認定規程は、CSJ の規程をそのまま用いるのではなく、書き言葉用に修正・拡張を行っている。CSJ の規程からの変更点については、5.4 節で述べる。

5.1.4 長単位・短単位の長所

ここでは、長単位・短単位がコーパスの言語単位として、どのような長所を持つのかについて述べる。

(1) 長単位の長所

長単位の長所としては、次の 2 点が挙げられる。

長所 1：当該資料の性格を反映する特徴的な語を取り出しやすい。

一般に単位を短くすればするほど、取り出した単位はいわゆる基本的な語となる。短単位は基準が分かりやすくゆれが少ないため、用例収集を行う上では便利な単位であるが、合成語を構成要素に分割してしまうという問題点がある。これに対して長単位では、「国立国語研究所」「品詞比率」「分析する」のような合成語を 1 単位として認める。「を」「だ」のような付属語は単独で長単位とするのが原則であるが、「における」「ている」のような複合辞も付属語として 1 長単位としている。

コアデータを基に、どのような語と結合するかという点から、掲載媒体別の差異を見る。以下、「生活」という語を例に説明する。コアデータ約 100 万語中、「生活」は 508 例見られる。そのうち、「生活」単独で使われた例が 153 例、合成語の構成要素として使われた例が 355 例と、「生活」は合成語の構成要素として使われやすい傾向にあることが分かる。掲載媒体に注目して、「生活」を含む語を見てみよう。1 媒体のみに現れる語には、白書の「基礎的生活コスト」「国民生活選好度調査」「WHO国際生活機能分類」「労働者生活」、新聞の「生活面子育て相談室係」、web の「残業生活する」「入院生活する」「週末泥沼生活」等、資料の内容・性格を反映したものが見られる。

「労働者生活」を「労働」と「者」と「生活」とに、「残業生活する」を「残業」と「生活」と「する」とに分割するのではなく、全体でひとつとして扱う長い単位を使うことで、各ジャンルの特徴的な語を把握することができる。

長所 2：文脈に即した品詞が付与されている。

短単位には、「名詞-普通名詞-形状詞可能」のような曖昧性を持たせた品詞がある。これに対して、長単位では文脈に即して品詞を付与する方針を取り、「名詞-普通名詞-○○可能」といった品詞は設けず、その文脈での用法に基づき名詞・形状詞・副詞に判別する。「結果」を例にすると、短単位では一律に「名詞-普通名詞-副詞可能」という品詞が付与されるが、長単位では、「これらの結果に基づき」の場合は「名詞-普通名詞-一般」とし、「結果、様々な社会問題が発生し」の場合は「副詞」とする。

コアデータを基に、名詞・形状詞・副詞の判別を行った長単位データと判別を行わない短単位データとの品詞比率の差異を見ると、判別を行った場合、書籍の形状詞率、新聞の副詞率が判別を行わない場合よりも高くなっており、白書では判別による変化が少ないことがわかった（冨士池他 2011）。これらは、用法に基づき、名詞・形状詞・副詞の判別をした結果、媒体の特徴がより明確になったものと考えられる。このように文脈に即して品詞を判別した長い単位を使うことで、構文的な機能を見る際に、より精密な分析が可能になる。

以上の二つの長所から、長単位は構文的な機能に着目した、各媒体・各ジャンルの資料的な性格を反映する単位であり、言語的特徴を解明するという目的にかなうものと言える。

(2) 短単位の長所

短単位の長所としては、次の 2 点が挙げられる。

長所 1: 基準が分かりやすく、ゆれが少ない。

これは、短単位の基礎となる最小単位の認定に当たり、個人によって捉え方に幅のある要素を基準に持ち込んでいないことによる。

基準が分かりやすく、ゆれが少ないという短単位の長所は、作業効率の向上につながるだけでなく、コーパスの使いやすさにもつながる。基準が分かりやすければ、利用者が語を検索する際、どのように検索条件を指定すればよいか迷うことが少なくなる。また、ゆれの少なさ、つまりデータの精度の高さは、分析結果の確かさにもつながる。

長所 2: 取り出した単位が文脈から離れすぎない。

上で短単位はゆれが少ない単位であると述べたが、実は最もゆれが少ない単位は、短単位ではなく、その基礎となっている最小単位である。それにもかかわらず、最小単位を言語単位として採用しなかったのは、最小単位は文脈から離れすぎるため、日本語の研究に使いにくいからである。

例えば、短単位「気持ち」は「気」と「持ち」の二つの最小単位に分割することができる。もしこのような最小単位でコーパスが解析されていると、動詞「持つ」を検索した際

に、「荷物を持つ」などの「持つ」とともに、「気持ち」の「持ち」も検索結果として得られることになる。

しかし、動詞「持つ」の分析を行う際に、「気持ち」の「持ち」まで検索結果に含まれるのは望ましいとは言い難い。それは、実際の文脈の中では、動詞「持つ」として機能していないからである。したがって、コーパスから用例を収集し、分析することを考えた場合、正確に単位認定ができるとしても、最小単位のような単位では問題が多いということになる。

以上のように考えた場合、短単位は、基準の分かりやすさ、ゆれの少なさという条件を満たしつつ、用例を収集して分析を行うという利用目的にもかなう単位と言える。

5.1.5 形態素解析用辞書 UniDic について

BCCWJ は、1 億語からなる大規模なコーパスであるため、形態論情報の付与は自動解析システムにより行う。短単位解析には解析エンジン MeCab と形態素解析用辞書 UniDic[5] 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[6]を使う。ここでは、短単位解析で用いる形態素解析用辞書 UniDic についてその概要を紹介する。

短単位解析の解析用辞書に UniDic を採用したのは、UniDic の言語単位が CSJ の短単位とほぼ同じものであり、品詞等の情報についても BCCWJ と互換性を持つものであったことによる。

また、UniDic では、表記や語形の違いにかかわらず、同じ語であれば、同一の見出しを与えるという方針を取り、語を階層化した形で登録している。この階層構造の最上位を語彙素（国語辞典の見出しに相当）と呼んでおり、この語彙素の下に語形（語形の違いを区別する層）、更に語形の下に書字形（表記の違いを区別する層）という階層を設けている（図 5-1）。

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

図 5-1: UniDic の階層構造の例

このような階層構造で登録した辞書を用いて、コーパスを形態素解析することによって、例えば、ある語について、どのような語形の変異や表記のゆれが、どの程度あるのかという情報を容易に得られるなど、日本語研究の可能性が格段に広がることが期待される。こ

[5] UniDic については、伝他（2007）を参照。
[6] 長単位を自動構成する解析器については、Uchimoto 他(2007)を参照。

のこともUniDicを形態素解析用辞書として採用した理由である。

我々は、BCCWJの構築を始めた2006年の時点で既に伝康晴氏が中心になって構築していたUniDic（見出し語:約46,000語）を基に、国語辞典や国立国語研究所の語彙調査等を基に作成されたデータ、そしてBCCWJの解析結果からUniDicの未登録語を追加していく作業を継続して行った。この作業の過程で、単位の認定、品詞情報等についてすり合わせを行い、現在では、UniDicの解析結果と本書で述べるBCCWJの短単位、品詞情報等に違いがない状態となっている。

unidic-mecab 1.3.12の解析精度を、以下の表5-2に示した。

表5-2: UniDicの解析精度（レジスター別）

	白書	書籍（文学）	書籍（文学以外）	新聞	Web（Y!知恵袋）	話し言葉（CSJ）
単位境界	0.9992	0.9976	0.9975	0.9964	0.9947	0.9963
品詞	0.9957	0.9866	0.9896	0.9891	0.9841	0.9844
語彙素	0.9938	0.9857	0.9864	0.9864	0.9804	0.9795

表中の「単位境界」「品詞」「語彙素」の意味は以下のとおりである。

単位境界：単位境界が正解と一致するか否か。

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。

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。

通常、形態素解析辞書は、品詞レベルまでで精度評価を行っている。UniDicは、それよりも厳しい語彙素レベルについても精度評価を行い、書き言葉については、全てのジャンルで98%以上の精度を実現している。

UniDicは、BCCWJの構築期間中に複数回バージョンアップを行った。BCCWJのコアデータ・非コアデータの自動解析には、その時点における最新のバージョンを用いている。そのため、ジャンルや自動解析を行った時期によって使用したUniDicのバージョンが異なっている。

また、UniDicで自動解析を行った後、コアデータ・非コアデータとも人手修正を行った。コアデータは、データ全体に対して作業担当者を変えて3回にわたるデータチェックを行い、誤解析の発見と修正を行った。非コアデータは、コアデータに比べて規模が極めて大きいため、データ全体に対してチェックを行うことは不可能である。そこで、未知語に起因する誤解析と疑われる箇所を中心にチェックを行い、誤解析の発見と修正を行った。

以上、本節では、国立国語研究所の語彙調査における調査単位について概観した上で、BCCWJの言語単位について設計方針等、その概要を述べた。5.2節以降では、長単位と短単位の認定規程の概要を紹介するが、その際、以下の凡例に示した記号を用いて単位境界や単位のつなぎ目を示す。

《 凡 例 》

１．各規程に示した例は、コーパスに現れた例又は作例である。

２．文節・長単位・最小単位・短単位の境界を示すために次の記号を用いた。

文節の境界	‥‥‥‥‥ ┃	例：┃国立国語研究所の┃
長単位の境界（5.2節）	‥‥‥‥ ｜	例：｜国立国語研究所｜の｜
最小単位の境界	‥‥‥‥‥ ／	例：／国／立／国／語／研／究／所／
短単位の境界（5.3節）	‥‥‥ ｜	例：｜国立｜国語｜研究｜所｜
当該規定で着目している箇所	‥ ‖	例：┃国立国語研究所の‖

３．分割しないことを特に示す必要があるときには、次の記号を用いた。

文節・長単位のつなぎ目	‥‥‥ -	例：┃からかわれて-ばかり-いる┃
		｜大-丈夫｜です｜
当該規定で着目している箇所	‥ =	例：┃からかわれて=ばかり=いる┃
		｜大=丈夫｜です｜

４．着目している文節・長単位が分かりにくい場合は、当該箇所に下線を付した。

5.2 長単位

長単位は、言語の構文的な機能に着目して規定した言語単位である。長単位の認定は、文節の認定を行った上で、各文節の内部を規定に従って自立語部分と付属語部分とに分割していくという手順で行う。そのため、長単位の認定規程は、文節と長単位の二つの認定規程から成る。

以下、本節では文節認定規程・長単位認定規程のうち主要な規定を紹介する。また、長単位に付与する付加情報についても、その概要を述べる。

5.2.1 文節認定規程

文節の認定方法（区切り方）については、いわゆる学校文法によって広く知られているところである。ここでは、BCCWJ の文節認定規程のうち、学校文法における文節の認定方法と異なる点をはじめ、特に注意すべき事項について概要を示す。

学校文法と異なる点としては、同格・並列の扱いが挙げられる。学校文法では、同格の関係にある要素、並列の関係にある要素は、以下のように切り離される。

【例】

〔同格〕┃大江健三郎さんの┃長男‖光さん┃

〔並列〕┃公正‖妥当な┃実務慣行を┃集約した┃ものという┃意味で┃

しかし BCCWJ では、以下のとおり同格の関係にある要素、並列の関係にある要素を切り離さないこととした。

【例】

〔同格〕 ┃大江健三郎さんの┃長男=光さん┃
┃東海汽船の┃支店長=・=重久さんは、┃
┃機関誌=計量国語学-発行の┃年に┃
┃中国語日刊新聞=「=星島日報」┃

〔並列〕 ┃公正=妥当な┃実務慣行を┃集約した┃ものという┃意味で┃
┃麦=・=大豆=・=飼料作物の┃生産振興に┃資する┃水田の┃汎用化を┃
┃最も┃先進的な┃青森=・=岩手=・=秋田の┃北東北三県は、┃
┃東京の┃郊外の┃市=町=村と┃言うか┃

同格・並列の扱い以外で、特に注意すべき事項について、以下、その認定規定と例とを示す。

【句読点・空白に関する規定】

（1）句読点（句読点として用いられているカンマ・ピリオド・エクスクラメーションマーク・クエスチョンマーク、三点リーダー、並びにコロンを含む。）および空白の後ろで切る。

【例】 ┃不合格には、‖違いないでしょうが。‖
┃十五歳少女が┃最年少記録┃「エベレスト登頂」┃三浦さん最高齢記録┃
その┃日に…‖
┃第２部　‖森林┃及び┃林業に関して┃講じた┃施策┃
┃２　┃協力的自主国防推進：‖自主国防と┃米韓同盟が┃

ただし、文頭の空白の後ろでは切らない。

【例】 ┃　=それは、┃現実の┃世界情勢が┃

【付属語に関する規定】

（2）助詞・助動詞・接尾辞連続（言いよどみの助詞・助動詞・接尾辞も含む。）の後ろで切る。助詞・助動詞には付録 5-A・付録 5-B に挙げた複合辞を含む。

【例】 ┃地域活動への‖参加、┃地産地消といった‖小さな┃経済で‖充足感を‖
得る┃社会と‖なります。┃
┃ネットワークが‖形成さ=れ=にくい‖状況が‖生じており、┃
┃その┃目的が‖個人に‖絞られ‖過ぎている‖傾向が‖ある┃

複合辞の中に副助詞など（言いよどみの助詞・助動詞も含む。）が挿入された場合も、文節認定の上では全体でひとつの複合辞と見なす。

【例】 ┃お友達には┃からかわれて=ばかり=いる┃三枚目でもありました。┃

（2）－1　次に挙げる付属語の後ろでは切らない。

①付録 5-C に挙げた連語、1 短単位として認定された「…が～」「…の～」の中に現れる付属語

【例】

〔連　語〕｜サイドの｜ベルトが｜お気に=入りの｜ブーツは｜

〔…が～〕｜そこが｜万が=一｜倒産すると｜

〔…の～〕｜皮を｜よく｜亀の=子だわしで｜こすって｜洗い｜

②分割すると意味が不自然になるものの中に現れる付属語

【例】　｜しかたが=ない｜　　｜しようが=ない｜

【主語・主題に関する規定】

（3）次に示すような付属語を伴わない主語・主題の後ろでは、文節を切らない。

【例】　｜緑=あふれる｜風景の｜中に、｜

｜心=洗われる｜ような｜ステージに｜

｜気持ち=悪いから、｜

【敬語形式に関する規定】

（4）「お（ご）～する・できる・くださる・いただく・なさる・いたす・ねがう・もうしあげる・あそばす・になる」という形式の敬語表現は、全体を一続きとする。

【例】　｜ご理解と｜ご協力の｜ほど｜よろしく｜お=願い=申し上げます。｜

｜いかが｜お過ごしでしたか、｜お=聞か=せ=ください。｜

｜法廷にも｜全身ピンクずくめで｜お=出まし=になる。｜

【数を表す要素に関する規定】

（5）数を表す要素とその直前直後の要素とは切り離さない。

【例】　｜昭和十三年=八月=八日の｜荒木文部大臣の｜発言や｜

｜平均値=三．〇六と｜いうような｜値に｜なって｜

｜日米韓=三国の｜対応｜

｜パチスロの｜場合だったら｜一箱=三万ぐらいなんですけど｜

｜十年以上=前までは｜(F ま)｜規則合成って｜いう｜方式が｜

｜三十=～=五十代の｜主婦を｜対象に｜行った｜アンケートで、｜

ただし、直前の要素が数量の程度を表す場合は除く。

【例】　｜およそ‖十カ所で｜検問を｜受け、｜旅券を｜確かめられた。｜

｜笑うと｜同じ｜事を｜最低‖３回は｜言います。｜

5.2.2 長単位認定規程

長単位は、長単位認定規程の各規定に基づいて文節を分割する（または分割しない）ことによって得られた要素を 1 単位とする言語単位である。以下、長単位認定規程のうち、主要な規定と例を挙げる。

【句読点・空白・改行に関する規定】

句読点・空白・改行に関する規定は、他の全ての規定に優先して適用される。

（1）句読点（句読点として用いられているカンマ・ピリオド・エクスクラメーションマーク・クエスチョンマーク、三点リーダー、並びにコロンを含む。）および空白は1長単位とする。

【例】 ｜機動的｜に｜商業施設｜として｜活用する｜例｜など｜も｜ある‖。‖
｜米｜は｜湾岸戦争後‖、‖英‖、‖仏｜など｜と｜とも｜に｜
｜十五歳少女｜が｜最年少記録｜「｜エベレスト登頂｜」｜三浦さん最高齢記録｜その｜日｜に‖…‖
｜２‖　‖協力的自主国防推進‖：‖自主国防｜と｜米韓同盟｜が｜
｜　‖それ｜は｜、｜現実｜の｜世界情勢｜が｜

【記号に関する規定】

（2）記号は1長単位とする。

【例】 ｜「‖ー‖羨ましい｜な‖」‖
｜与野党逆転‖→‖海部政権誕生｜と｜の｜願望｜

記号のうち中点については、原則として切り出さない。

【例】 ｜平成｜３年度｜から｜コンピュータ=・=ネットワーク｜を｜利用し｜、｜
｜豪商=・=山崎屋｜の｜与五郎｜と｜遊女=・=吾妻｜と｜の｜恋｜を｜
｜麦=・=大豆=・=飼料作物｜の｜生産振興｜に｜資する｜水田｜の｜
｜D=・=N=・=A｜（｜ＡＢＣ｜＝｜深夜｜３=・=二十｜）｜

（2）－1　語と同じ働きをする記号・記号連続およびそれらを含む結合体は、全体で 1 長単位とする。

【例】 ｜Ａ‖が‖Ｂ‖に｜特定｜の｜法律行為｜を｜指図し｜た｜場合｜
｜南青山｜に｜ある｜敷地面積‖２，０００=㎡‖の｜土地｜は｜、｜
‖ＰＫＯ=地域訓練ワークショップ‖の｜開催｜や｜
｜一般会計｜の‖（=　=）=内‖は｜０３年度当初予算｜

【付属語に関する規定】

（3）付属語（付録 5-A・付録 5-B に挙げた複合辞を含む。）は 1 長単位とする。

【例】 ｜公害紛争処理法‖における‖公害紛争処理‖の‖手続‖は‖，｜原則‖として‖紛争当事者‖から‖の‖申請‖によって‖開始さ‖れる‖。｜
｜その｜目的‖が‖個人‖に‖絞ら‖れ‖過ぎ‖ている‖傾向‖が‖ある｜
｜「｜やむ‖を‖得‖ず‖型｜」‖の‖親同居未婚者｜

ただし、それを 1 長単位とすると、動詞的・形容詞的・形状詞的接尾辞および用言・助動詞の終止形・連体形以外に続く名詞的接尾辞が切り出されることになる場合の付属語は切り出さない。

【例】 ｜ネットワーク｜が｜形成さ=れ=にくい｜状況が｜生じ｜ており｜，｜
｜どんな｜使わ=れ=方｜を｜し｜た｜んだろう｜。｜

付属語を伴わない文節、および規定（3）によって付属語を切り出した後に残った形式（おおよそ文節の自立語部分に相当する形式）に以下の規定を適用する。それによって得られた各形式を 1 長単位とする。

（4）同格の関係にある体言連続、並列の関係にある体言連続は切り離さない。

【例】

〔同格〕｜大江健三郎さん｜の｜長男=光さん｜
｜機関誌=計量国語学-発行｜の｜年｜に｜
｜中国語日刊新聞=「=星島日報=」｜

〔並列〕｜公正=妥当｜な｜実務慣行｜を｜集約し｜た｜もの｜という｜
｜東京｜の｜郊外｜の｜市=町=村｜と｜言う｜か｜

（5）主語・主題の後ろで切る。

【例】 ｜緑‖あふれる｜風景｜の｜中に、｜
｜心‖洗われる｜よう｜な｜ステージ｜に｜
｜気持ち‖悪い｜から｜、｜

漢語形状詞の述部を持つ場合は、切らない。

【例】 ｜持続=可能｜な｜発展｜の｜ため｜
｜センス=抜群｜の｜クリエーター｜だっ｜た｜んだ｜な｜

（6）体言や副詞に形式的な意味の「する」「できる」「なさる」「いたす」が直接続く場合、体言・副詞と「する」「できる」「なさる」「いたす」とを切り離さない。

【例】 ｜まるで｜1つ｜の｜光点｜が｜往復運動=し｜ている｜よう｜に｜
｜私｜は｜この｜予選｜を｜1位｜で｜通過=できる｜と｜

｜目｜を｜きらきら=さ｜せ｜た｜

「こう」「そう」「ああ」「どう」に「する」「できる」「なさる」「いたす」が直接続く場合は、切り離す。

【例】 ｜こう‖し｜た｜サークル活動｜が｜盛ん｜に｜なる｜背景｜
｜そうこう‖し｜ている｜うち｜に｜

（7）体言＋用言という形式のうち、『岩波国語辞典』第 6 版、『日本国語大辞典』第 2 版のいずれか一方で見出し語（連語としての見出し語は除く。）になっているものは、体言と用言とを切り離さない。

【例】 ｜運転手｜は｜さり気=なく｜答え｜て｜アクセル｜を｜ゆるめ｜た｜。｜

（8）数を表す要素を含む自立語は、以下の規定に基づき長単位を認定する。

（8）－1　数を表す要素は、単位の変わり目の後ろで切る。

【例】 ｜平成｜１５年‖９月‖１５日‖午後｜７時‖３３分‖
｜１ｍ‖８０ｃｍ｜

（8）－2　数を表す要素の前で切る。

【例】 ｜平成‖１５年｜９月｜１５日｜午後‖７時｜３３分｜
｜南青山｜に｜ある｜敷地面積‖２，０００㎡｜の｜土地｜は｜、｜
｜おおむね‖十六歳以上‖二十歳未満｜の｜者｜を｜収容｜

（8）－3　数を表す要素とそれに続く体言・接辞とは切り離さない。

【例】 ｜残業時間｜が｜月｜８０時間=以上｜の｜者｜は｜
｜自家発電｜が｜すぐ｜に｜作動せ｜ず｜、｜約１時間=停電｜。｜
｜５年=連続=優勝｜を｜駒沢大｜が｜来年｜、｜達成できる｜か｜

ただし、数を表す要素とそれに続く体言・接辞の連続体の後ろに、これを受ける形式的な意味の「する」「できる」「なさる」「いたす」がある場合は、数を表す要素の後ろで切り離す。

【例】 ｜約１時間‖停電した｜。｜
｜箱根駅伝｜で｜４年連続‖優勝し｜た｜の｜は｜５校目｜。｜

（9）付録 5-C に挙げた連語、1 短単位として認定された「—が～」「—の～」およびそれらを含む結合体は、全体で 1 長単位とする。

【例】

〔連　語〕｜サイド｜の｜ベルト｜が｜お気に入り｜の｜ブーツ｜は｜
〔—が～〕｜そこ｜が｜万が=一｜倒産する｜と｜

〔―の～〕｜皮｜を｜よく｜亀の=子だわし｜で｜こすっ｜て｜洗い｜

以上の規定によって長単位を認定した例を次に示す。

｜平成｜４年度｜に｜創設さ｜れ｜た｜定期借地権制度｜は｜、｜借地契約｜の｜更新｜が｜なく｜、｜定め｜られ｜た｜契約期間｜で｜確定的｜に｜契約｜が｜終了する｜借地権制度｜である｜。｜貸し主｜（｜土地所有者｜）｜にとって｜は｜予定時期｜に｜土地｜の｜返還｜を｜受ける｜こと｜が｜保証さ｜れる｜とともに｜、｜一定期間｜の｜地代収入｜が｜安定的｜に｜得ら｜れ｜、｜また｜、｜借り主｜にとって｜は｜土地｜を｜取得する｜より｜も｜少ない｜負担｜で｜土地｜を｜利用できる｜こと｜から｜、｜双方｜にとって｜メリット｜が｜あり｜、｜借地｜の｜供給拡大｜による｜土地｜の｜有効利用｜を｜促進する｜もの｜として｜期待さ｜れ｜ている｜。｜定期借地権｜に｜は｜、｜一般定期借地権｜、｜建物譲渡特約付借地権｜、｜事業用借地権｜の｜３類型｜が｜ある｜（｜図表｜１｜-｜５｜-｜４｜）｜。｜

5.2.3 付加情報の概要

長単位認定規程によって認定された各単位に次に挙げる付加情報を付与する。

(1) 語彙素・語彙素読み

自立語の語彙素・語彙素読みは、同一語の活用変化・表記のゆれ（補助記号の有無を含む。）をグループ化するための情報である。

例えば、サ変動詞「構築する」の未然形「構築さ」、連用形「構築し」、終止形・連体形「構築する」には、いずれも同じ「コウチクスル【構築する】」という語彙素・語彙素読みが付与される。これによって、「構築さ」「構築し」「構築する」の各出現形が、ひとつの語（動詞「構築する」）の活用変化として扱われることになる。同様に、「打ち合わせ室」「打合せ室」という各出現形に対して「ウチアワセシツ【打ち合わせ室】」という同一の語彙素・語彙素読みが付与され、ひとつの語の表記のゆれとして扱われることになる。

出現形	語彙素読み	語彙素
構築さ	コウチクスル	構築する
構築し		
構築する		
打ち合わせ室	ウチアワセシツ	打ち合わせ室
打合せ室		

図 5-2: 長単位の語彙素・語彙素読みの例(1)

一方、表記以外のゆれ・音の転化・省略・融合等によって生じた異形態はグループ化しない。そのため、以下の図 5-3 に挙げた「あまり」と「あんまり」などについては、それぞれ異なる語彙素・語彙素読みが付与され、別語として扱われる。

出現形	語彙素読み	語彙素
あまり	アマリ	余り
余り		
あんまり	アンマリ	余り
ちょうふく 重複する	チョウフクスル	重複する
じゅうふく 重複する	ジュウフクスル	重複する
コンピューター	コンピューター	コンピューター
コンピュータ	コンピュータ	コンピュータ

図 5-3: 長単位の語彙素・語彙素読みの例(2)

付属語の語彙素は、同一語の活用変化・ゆれ・省略・融合等によって生じた異形態をグループ化するための情報である。

語彙素・語彙素読みは、原則としてコーパスに出現したすべての長単位に付与する。

出現形	語彙素読み	語彙素
からには	カラニハ	からには
からにゃ		
こととなる	コトトナル	こととなる
ことと成る		

図 5-4: 長単位の語彙素・語彙素読みの例(3)

(2) 品詞等の情報

各単位に品詞を付与する。活用する語には、活用型・活用形を付与する。

長単位の品詞・活用型・活用形については、表 5-3 から表 5-5 に一覧した。

表 5-3: 長単位品詞一覧

名詞-普通名詞-一般	接続詞	接尾辞-形状詞的
名詞-固有名詞-一般	感動詞-一般	接尾辞-動詞的
名詞-固有名詞-人名-一般	感動詞-フィラー	接尾辞-形容詞的
名詞-固有名詞-人名-姓	動詞-一般	記号-一般
名詞-固有名詞-人名-名	形容詞-一般	記号-文字
名詞-固有名詞-地名-一般	助動詞	補助記号-一般
名詞-固有名詞-地名-国	助詞-格助詞	補助記号-句点
名詞-数詞	助詞-副助詞	補助記号-読点
名詞-助動詞語幹	助詞-係助詞	補助記号-括弧開
代名詞	助詞-接続助詞	補助記号-括弧閉
形状詞-一般	助詞-終助詞	補助記号-ＡＡ-一般
形状詞-タリ	助詞-準体助詞	補助記号-ＡＡ-顔文字
形状詞-助動詞語幹	接頭辞	空白
連体詞	接尾辞-名詞的-一般	
副詞	接尾辞-名詞的-助数詞	

表 5-4: 長単位活用型一覧

五段-〇行	助動詞-ラシイ	文語助動詞-ザマス
上一段-〇行	助動詞-レル	文語助動詞-ザンス
下一段-〇行	無変化型	文語助動詞-ジ
カ行変格	文語四段-〇行	文語助動詞-ズ
サ行変格	文語上一段-〇行	文語助動詞-タリ-完了
形容詞	文語上二段-〇行	文語助動詞-タリ-断定
助動詞-ジャ	文語下一段-〇行	文語助動詞-ツ
助動詞-タ	文語下二段-〇行	文語助動詞-ナリ-断定
助動詞-ダ	文語カ行変格	文語助動詞-ナリ-伝聞
助動詞-タイ	文語サ行変格	文語助動詞-ヌ
助動詞-デス	文語サ行変格-ス	文語助動詞-ベシ
助動詞-ドス	文語ナ行変格	文語助動詞-マシ
助動詞-ナイ	文語ラ行変格	文語助動詞-マジ
助動詞-ナンダ	文語形容詞-ク	文語助動詞-ム
助動詞-ヌ	文語形容詞-シク	文語助動詞-ムズ
助動詞-ヘン	文語助動詞-キ	文語助動詞-メリ
助動詞-マイ	文語助動詞-ケム	文語助動詞-ラシ
助動詞-マス	文語助動詞-ケリ	文語助動詞-ラム
助動詞-ヤ	文語助動詞-コス	文語助動詞-リ
助動詞-ヤス	文語助動詞-ゴトシ	文語助動詞-ンス

表 5-5: 長単位活用形一覧

語幹-一般	連用形-融合	連体形-一般
語幹-サ	連用形-省略	連体形-〇音便
未然形-一般	連用形-ト	連体形-省略
未然形-サ	連用形-ニ	連体形-補助
未然形-セ	連用形-長音	仮定形-一般
未然形-撥音便	連用形-補助	仮定形-融合
未然形-補助	終止形-一般	已然形-一般
意志推量形	終止形-〇音便	已然形-補助
連用形-一般	終止形-融合	命令形
連用形-〇音便	終止形-補助	ク語法

5.3 短単位

短単位は、言語の形態的側面に着目して規定した言語単位である。短単位の認定に当たっては、まず現代語において意味を持つ最小の単位（最小単位）を規定する。その上で、最小単位を長単位の範囲内で短単位認定規程に基づいて結合させる（又は結合させない）ことにより、短単位を認定する。そのため、短単位の認定規程は、最小単位と短単位の二つの認定規程から成る。

以下、本節では、最小単位認定規程・短単位認定規程、および短単位に付与する付加情報について、その概要を述べる。

5.3.1 最小単位認定規程

最小単位は、現代語において意味を持つ最小の言語単位のことである。

最小単位は、和語・漢語・外来語・記号・数・人名・地名の種類ごとに、以下の規定によって認定する。

なお、以下に述べる最小単位は、短単位を認定するために必要な概念として規定するものであり、BCCWJ のデータに最小単位境界を示すことはしない。

（1）和語

和語の最小単位は、以下のように認定する。

【例】 ／母／親／　／青／白い／　／いい／加／減／な／
／本／箱／　／幾／人／　／オレンジ／色／
／わたし／で／も／できる／　／読み／終わり／まし／た／

（2）漢語

漢語（和製漢語を含む。）は、漢字 1 字で表されるものを 1 最小単位とする。

【例】 ／白／紙／　／安／価／　／含／有／量／　／数／百／

（3）外来語

外来語・外国語は原語で 1 単語になるものを 1 最小単位とする。

英語起源の外来語の最小単位の認定は『リーダーズ英和辞典』第 2 版（研究社）による。それ以外の言語を起源とする外来語については適宜判断する。

【例】 ／カラー／コピー／　／レーザー／プリンター／
／オレンジ／色／　／ビタミン／剤／

（4）記号

記号は 1 文字に当たるものを 1 最小単位とする。

【例】 ／表／Ａ／　／図／Ｂ／　／Ｕ／ターン／　／Ｖ／リーグ／
／甲／類／　／乙／種／

（4）－1　ローマ字を並べた略語は全体で 1 最小単位とする。ローマ字の間の中点・ピリオド等は 1 最小単位としない。

【例】 ／ＯＨＰ／　／ＯＳ／　／Ｄ・Ｎ・Ａ／　／Ｐｈ．Ｄ．／

（5）数

数字は 1 字に当たるものを 1 最小単位とする。

【例】 ／一／億／語／　／七／百／五／十／万／語／
／２／万／５／千／分／の／１／
／０／４／２／－／５／４／０／－／４／３／０／０／

（6）人名

人名は姓を１最小単位、名を１最小単位とする。

【例】　／星野／仙一／　　／マット／・／マートン／　　／林／威助／

通称・雅号・しこ名（その略称も含む。）等は、次のように最小単位を認定する。

【例】　／琴奨菊／　　／十返舎／一九／　　／古今亭／志ん生／

（7）地名

地名は、次の規定により最小単位を認定する。

（7）－１　行政区画を表す地名は「都・府・県・郡・市・区・町・村・字」を除いた部分をそれぞれ１最小単位とする。類概念を表す部分には、他の最小単位の認定規定を適用する。

【例】　／東京／都／北／区／西が丘／三／丁／目／九／番／十／四／号／

（7）－１－１　「北海道」は全体で１最小単位とする。

【例】　／北海道／夕張／郡／長沼／町／　　／明日／の／北海道／の／天気／

（7）－１－２　市区内の小区分の「～町」は「～町」を含めて１最小単位とする。

【例】　／大阪／府／豊中／市／待兼山町／　　／千代田／区／大手町／

（7）－２　外国の国名や行政区画名なども、日本のそれと同じに扱う。

【例】　／アメリカ／合／衆／国／　　／南アフリカ／共／和／国／
／中華／人／民／共／和／国／
／カリフォルニア／州／　　／広東／省／
／メキシコ／シティー／　　／ミズーリ／ステート／

以上の規定によって認定された最小単位を、短単位認定のために表 **5-6** のように分類する。

表 5-6: 最小単位の分類

分　　類	例
一般	和　語　：　山　川　白い　話す　言葉　…
	漢　語　：　社　会　用　研　究　所　…
	外来語　：　オレンジ　ボックス　アルゴリズム　…
付属要素	接頭的要素（付録5-D：接頭的要素に掲げたもの） ：　相　御(お)　各　御(ご)　…
	接尾的要素（付録5-E：接尾的要素に掲げたもの） ：　　致す　っぽい　性　的　…
記号	Ａ　Ｂ　ω　イ　ロ　ア　甲　乙　丙　ＮＨＫ　ＪＲ　…
数	一　二　十　百　千　…　幾　数　何
固有名	人　名　：　星野　仙一　ジェフ　ウィリアムス　橋　龍　…
	地　名　：　大阪　待兼山町　六甲　天六　…
助詞・助動詞	た　です　ます　か　から　て　も　…

5.3.2 短単位認定規程

短単位は、長単位の中で最小単位が以下の規定に基づいて結合した（又は結合しない（これは 0 回結合と考える。））結合体である。

短単位の認定に関する規定は、表 5-6 に示した種類ごとに適用すべき規定が定められている。以下、それを示す。

（1）一般

原則として、「一般」に分類した和語・漢語の最小単位二つの一次結合を 1 短単位とする。

【例】　｜母=親｜　　｜書き=言葉｜　　｜食べ=歩く｜　　｜音=声｜
｜無=口｜

「一般」に分類した外来語の最小単位のうち省略されたものは、和語・漢語の最小単位と同様に扱う。

【例】　｜パソ=コン｜　　｜オートマ=車｜　　｜塩=ビ｜

（1）－1　以下のものは、3 最小単位以上の結合であっても全体で 1 短単位とする。

①三つ以上の最小単位から成る組織の名称等の略称

【例】　｜統=数=研｜　　｜奈=文=研｜　　｜日=経=連｜

②切る位置が明確でないもの、あるいは切った場合とひとまとめにした場合とで意味にずれがあるもの

【例】　｜大統領｜　｜不可解｜　｜明後日｜　｜殺風景｜
｜輸出入｜　｜国内外｜　｜町村長｜　｜原水爆｜　｜市町村長｜
｜大袈裟｜　｜大雑把｜　｜大丈夫｜　｜一辺倒｜
｜十文字｜　｜二枚目｜　｜十八番｜

③「―が～」「―の～」の体言句

【例】
「―の～」　：　｜日=の=丸｜　｜床=の=間｜　｜竹=の=子｜
「―が～」　：　｜天=が=下｜　｜雁=が=音｜　｜剣=が=峰｜

（1）－2　以下に挙げるものは、1最小単位を1短単位とする。

①外来語・外国語の最小単位

【例】　｜オレンジ｜色｜　｜アウト｜オブ｜ドメイン｜

ただし、省略された外来語の最小単位との1次結合体は1短単位とする。

【例】　｜エア=コン｜　｜マス=コミ｜　｜デフレ=スパイラル｜

②最小単位が三つ以上並列した場合の、それぞれの最小単位

【例】　｜衣‖食‖住｜　｜松‖竹‖梅｜　｜都‖道‖府‖県｜

③名を表す部分と類概念を表す部分とが結合してできた固有名のうち、名を表す部分・類概念を表す部分が共に1最小単位である場合の、それぞれの最小単位

【例】　｜さくら‖屋｜　｜のぞみ‖号｜　｜くれない‖会｜

ただし、名を表す部分が1字の漢語である場合は、その1次結合体を1短単位とする。

【例】　｜阪=大｜　｜仏=教｜　｜李=朝｜　｜壮=族｜　｜礼=記｜

④感動詞

【例】　｜はい｜はい｜　｜おい｜おい｜　｜どれ｜どれ｜

⑤規定（1）、（1）－1、（1）－2の①から③によって得られた短単位に、前または後ろから結合した最小単位

【例】　｜内閣‖府‖　‖副‖大統領｜　｜橋本‖元‖首相｜
｜光｜ファイバー‖網‖　｜自衛‖隊‖　｜国立｜国語｜研究‖所‖

⑥単独で文節を構成する最小単位

【例】 ｜やっぱり｜これ｜も｜一｜つ｜の｜　｜オレンジ｜を｜食べる｜。｜
｜えーと｜、｜こちら｜の｜場合｜でし｜たら｜…｜…｜

（2）記号

記号は、1最小単位を1短単位とする。

【例】 ｜表｜A｜　｜図｜B｜　｜ＪＲ｜　｜ＮＴＴ｜　｜Ｌ．Ａ．｜

（3）数

数は、以下の規定によって単位認定する。

（3）－1　数は、ほかの最小単位と結合させない。

【例】 ｜四‖月｜の‖三十‖日｜ぐらい｜
｜私｜が‖一二‖年｜前｜まで｜住｜ん｜で｜い｜た｜

（3）－2　数の間どうしの結合については、一・十・百・千の桁ごとに1短単位とする。「万」「億」「兆」などの最小単位は、それだけで1短単位とする。小数部分は、1最小単位を1短単位とする。

【例】 ｜千‖九=百‖四=十‖二｜年｜十｜月｜二=十‖五｜日｜、｜
｜毎年｜何=十‖億‖円｜も｜の｜都民｜の｜税金｜を｜
｜都心｜から｜一｜時間｜半｜どころ｜か｜、｜三=、=四十｜分｜、｜
｜地形｜図｜２｜万‖５=千｜分｜の｜１｜
｜０｜４｜２｜－｜５｜４｜０｜－｜４｜３｜０｜０｜

（4）固有名

固有名（人名・地名）は、1最小単位を1短単位とする。

【例】

〔人　名〕　｜星野｜仙一｜　｜マット｜・｜マートン｜　｜林｜威助｜
｜琴奨菊｜　｜十返舎｜一九｜　｜お千代｜

〔国　名〕　｜アメリカ｜合衆｜国｜　｜ロシア｜共和｜国｜
｜南アフリカ｜共和｜国｜

〔行政区画名〕｜東京｜都｜立川｜市｜緑町｜十｜番｜二｜号｜
｜京都｜市｜上京｜区｜今出川｜通｜烏丸｜東入る｜

〔地域名〕　｜九州｜地方｜　｜四国｜地方｜　｜北海道｜地方｜
｜東海道｜　｜山陰道｜
｜東｜ヨーロッパ｜　｜南｜アメリカ｜

〔地形名〕　｜生駒｜山｜　　｜昭和｜新山｜　　｜サロマ｜湖｜
〔場所名〕　｜茨木｜市｜駅｜　　｜さいたま｜新｜都心｜駅｜
　　｜山陽｜本線｜　　｜大｜江戸｜線｜
　　｜東海道｜　　｜中山道｜
〔略　称〕　｜ちとから｜　　｜天六｜

（5）付属要素

付属要素は、1 最小単位を 1 短単位とする。

【例】　｜お‖母‖さん｜　　｜見‖にくい｜

（6）助詞・助動詞

助詞・助動詞は、1 最小単位を 1 短単位とする。

【例】　｜統一｜的‖な‖視点‖で‖切り‖ましょう‖
　　｜それ｜に｜つい｜て｜もっとも｜示唆｜に｜富む｜の｜は｜

（6）－1　1 短単位として認定された「―が～」「―の～」の中の助詞「が」「の」は、助詞・助動詞として扱わない。

【例】

「―の～」　：　｜日=の=丸｜　　｜床=の=間｜　　｜竹=の=子｜
「―が～」　：　｜君=が=代｜　　｜万=が=一｜

以上が短単位認定規程における主要な規定である。その他、短単位の認定に当たって注意すべき事項について規定を示す。

（7）可能動詞

可能動詞は、元になった五段活用動詞と同様に短単位を認定する。

【例】　｜読める｜　　｜行ける｜　　｜切り離せる｜　　｜話し合える｜

（7）－1　ら抜き言葉は語末の「れる」を切り出さない。

【例】　｜着=れる｜　　｜来=れる｜　　｜食べ=れる｜
　　｜見=れる｜　　｜透かし見=れる｜　　｜こじ開け=れる｜

（8）動詞「―（サ）ス」「―（サ）セル」

（8）－1　「―（サ）ス」という形の動詞は、語末「ス」「サス」を助動詞としない。

【例】　｜言わ=す｜　　｜書か=す｜　　｜食べ=さす｜　　｜受け=さす｜

（８）－２　五段・サ変動詞の未然形＋助動詞「セル」、五段・サ変以外の動詞の未然形＋助動詞「サセル」に分析可能なものは、語末「セル」「サセル」を助動詞とする。

【例】　｜書か‖せる｜　　｜食べ‖させる｜

ただし、以下に挙げるものは、語末の「（サ）セル」を分割しない。

①五段・サ変動詞の未然形＋助動詞「セル」、五段・サ変以外の動詞の未然形＋助動詞「サセル」と分析できないもの。

【例】　｜見=せる｜　　｜着=せる｜　　｜乗=せる｜　　｜寄=せる｜

②元の動詞が文語動詞であるもの、口語動詞であっても、現代語ではほとんど使われないもの。

【例】　｜くゆら=せる｜　　｜遅ら=せる｜　　｜そばだた=せる｜

③「―（サ）セル」という形の複合動詞（連用形が名詞化したものも含む。）。

【例】　｜言い聞か=せる｜　　｜言い聞か=せ｜続ける｜　　｜読み聞か=せ｜

④「―（サ）セル」という形の動詞（複合動詞は除く。）が複合語を構成している場合。

【例】　｜食わ=せ=物｜　　｜人騒が=せ｜　　｜人泣か=せ｜　　｜番狂わ=せ｜
｜役者｜泣か=せ｜

以上の規定によって短単位を認定した例を次に示す。

｜平成｜４｜年度｜に｜創設｜さ｜れ｜た｜定期｜借地｜権｜制度｜は｜、｜借地｜契約｜の｜更新｜が｜なく｜、｜定め｜られ｜た｜契約｜期間｜で｜確定｜的｜に｜契約｜が｜終了｜する｜借地｜権｜制度｜で｜ある｜。｜貸し主｜（｜土地｜所有｜者｜）｜に｜とっ｜て｜は｜予定｜時期｜に｜土地｜の｜返還｜を｜受ける｜こと｜が｜保証｜さ｜れる｜と｜とも｜に｜、｜一定｜期間｜の｜地代｜収入｜が｜安定｜的｜に｜得ら｜れ｜、｜また｜、｜借り主｜に｜とっ｜て｜は｜土地｜を｜取得｜する｜より｜も｜少ない｜負担｜で｜土地｜を｜利用｜できる｜こと｜から｜、｜双方｜に｜とっ｜て｜メリット｜が｜あり｜、｜借地｜の｜供給｜拡大｜に｜よる｜土地｜の｜有効｜利用｜を｜促進｜する｜もの｜と｜し｜て｜期待｜さ｜れ｜て｜いる｜。｜定期｜借地｜権｜に｜は｜、｜一般｜定期｜借地｜権｜、｜建物｜譲渡｜特約｜付｜借地｜権｜、｜事業｜用｜借地｜権｜の｜３｜類型｜が｜ある｜（｜図表｜１｜-｜５｜-｜４｜）｜。｜

5.3.3 付加情報の概要

短単位認定規程によって認定された各単位に、次に挙げる付加情報を付与する。

(1)　語彙素・語彙素読み

語彙素・語彙素読みは、同一語の活用変化・音の転化・ゆれ・省略・融合等によって生じた異形態や送り仮名の違い等の異表記をグループ化するための情報である。

例えば、動詞「取る」の未然形「取ら」、連用形「取り」、終止形・連体形「取る」には、いずれも同じ「トル【取る】」という語彙素・語彙素読みが付与される。これによって、「取ら」「取り」「取る」の各出現形がひとつの語（動詞「取る」）の活用変化として扱われることになる。同様に「打ち合わせ」「打合せ」という各出現形に対して「ウチアワセ【打ち合わせ】」という同一の語彙素・語彙素読みが付与され、ひとつの語の表記のゆれとして扱われることになる。

出現形	語彙素読み	語彙素
取ら	トル	取る
取り		
取る		
打ち合わせ	ウチアワセ	打ち合わせ
打合せ		

図 5-5: 短単位の語彙素・語彙素読みの例(1)

長単位の語彙素・語彙素読みでは、省略・融合等によって生じた異形態はグループ化しなかったが、短単位の語彙素・語彙素読みでは、以下の図 5-6 のように、それぞれ同じ語彙素・語彙素読みが付与され、同語として扱われる。

出現形	語彙素読み	語彙素
あまり	アマリ	余り
余り		
あんまり		
ちょうふく 重複	チョウフク	重複
じゅうふく 重複		
コンピューター	コンピューター	コンピューター
コンピュータ		

図 5-6: 短単位の語彙素・語彙素読みの例(2)

原則として、語彙素・語彙素読みは、コーパスに出現した全ての短単位に付与する。

(2) 品詞情報

各単位に品詞を付与する。活用する語には、活用型・活用形を付与する。

短単位の品詞・活用型・活用形については、表 5-7 から表 5-9 に一覧した。

表 5-7: 短単位品詞一覧

名詞-普通名詞-一般	連体詞	接尾辞-名詞的-サ変可能
名詞-普通名詞-サ変可能	副詞	接尾辞-名詞的-形状詞可能
名詞-普通名詞-形状詞可能	接続詞	接尾辞-名詞的-サ変形状詞可能
名詞-普通名詞-サ変形状詞可能	感動詞-一般	接尾辞-名詞的-副詞可能
名詞-普通名詞-副詞可能	感動詞-フィラー	接尾辞-名詞的-助数詞
名詞-普通名詞-助数詞可能	動詞-一般	接尾辞-形状詞的
名詞-固有名詞-一般	動詞-非自立可能	接尾辞-動詞的
名詞-固有名詞-人名-一般	形容詞-一般	接尾辞-形容詞的
名詞-固有名詞-人名-姓	形容詞-非自立可能	記号-一般
名詞-固有名詞-人名-名	助動詞	記号-文字
名詞-固有名詞-地名-一般	助詞-格助詞	補助記号-一般
名詞-固有名詞-地名-国	助詞-副助詞	補助記号-句点
名詞-数詞	助詞-係助詞	補助記号-読点
名詞-助動詞語幹	助詞-接続助詞	補助記号-括弧開
代名詞	助詞-終助詞	補助記号-括弧閉
形状詞-一般	助詞-準体助詞	補助記号-ＡＡ-一般
形状詞-タリ	接頭辞	補助記号-ＡＡ-顔文字
形状詞-助動詞語幹	接尾辞-名詞的-一般	空白

表 5-8: 短単位活用型一覧

五段-〇行	形容詞-一イ	文語ラ行変格
五段-カ行-一般	助動詞-ジャ	文語形容詞-ク-一般
五段-カ行-イク	助動詞-タ	文語形容詞-ク-多シ
五段-カ行-ユク	助動詞-タイ	文語形容詞-シク-シク
五段-マ行-一般	助動詞-ダ	文語形容詞-シク-ジク
五段-マ行-済ム	助動詞-デス	文語助動詞-キ
五段-ラ行-一般	助動詞-ドス	文語助動詞-ケム
五段-ラ行-アル	助動詞-ナイ	文語助動詞-ケリ
五段-ラ行-サル	助動詞-ナンダ	文語助動詞-コス
五段-ワア行-一般	助動詞-ヌ	文語助動詞-ゴトシ
五段-ワア行-〇ウ	助動詞-ヘン	文語助動詞-ザマス
上一段-〇行	助動詞-マイ	文語助動詞-ザンス
上一段-ラ行-一般	助動詞-マス	文語助動詞-ジ
上一段-ラ行-リル	助動詞-ヤ	文語助動詞-ズ
下一段-〇行	助動詞-ヤス	文語助動詞-タリ-完了
下一段-ア行-一般	助動詞-ラシイ	文語助動詞-タリ-断定
下一段-ア行-得ル	助動詞-レル	文語助動詞-ツ
下一段-サ行-一般	文語四段-〇行	文語助動詞-ナリ-伝聞
下一段-サ行-セル	文語四段-ハ行-一般	文語助動詞-ナリ-断定
下一段-ラ行-一般	文語四段-ハ行-〇ウ	文語助動詞-ヌ
下一段-ラ行-レル	文語四段-ハ行-イウ	文語助動詞-ベシ
下一段-ラ行-呉レル	文語上一段-〇行	文語助動詞-マシ
カ行変格	文語上二段-〇行	文語助動詞-マジ
サ行変格-スル	文語下一段-カ行	文語助動詞-ム
サ行変格-ズル	文語下二段-〇行	文語助動詞-ムズ
サ行変格-為ル	文語下二段-ハ行-一般	文語助動詞-メリ
形容詞-一般	文語下二段-ハ行-経	文語助動詞-ラシ
形容詞-無イ	文語カ行変格	文語助動詞-ラム
形容詞-良イ-イイ	文語サ行変格-ス	文語助動詞-リ
形容詞-良イ-ヨイ	文語サ行変格-ズ	文語助動詞-ンス
形容詞-〇イ	文語ナ行変格	無変化型

表 5-9: 短単位活用形一覧

語幹-一般	連用形-融合	連体形-一般
語幹-サ	連用形-省略	連体形-〇音便
未然形-一般	連用形-ト	連体形-省略
未然形-サ	連用形-ニ	連体形-補助
未然形-セ	連用形-長音	仮定形-一般
未然形-撥音便	連用形-補助	仮定形-融合
未然形-補助	終止形-一般	已然形-一般
意志推量形	終止形-〇音便	已然形-補助
連用形-一般	終止形-融合	命令形
連用形-〇音便	終止形-補助	ク語法

(3) 語種情報

語種とは、語をその出自によって分類したもののことである。原則として、コーパスに出現したすべての短単位に付与する。

BCCWJ で付与した語種は、次のとおりである。

①和語〔和〕

日本固有の語

【例】 暖かい　　言葉　　話す

②漢語〔漢〕

近世以前に中国から入った語

【例】 音楽　　国語　　報告

和製漢語も漢語とする。

【例】 大根　　返事

③外来語〔外〕

欧米系の諸言語から入った語

【例】 ゲーム　　コーパス　　データ

上記のほか、以下のものも外来語とする。

ａ．和製英語

【例】 アフレコ　　ナイター

ｂ．梵語等を中国で音訳した語に由来する語

【例】 阿羅漢　　盂蘭盆　　卒塔婆

c．アイヌ語から入った語

【例】 昆布　　鮭　　ラッコ

d．中国以外のアジア諸国語から入った語

【例】 キムチ　　パッチ

e．近代以降に中国から入った語

【例】 クーニャン　　シュウマイ　　メンツ

④混種語〔混〕

和語・漢語・外来語のうち異なる 2 種類以上の語種の語が二つ以上結合した語。漢語・外来語であったものの末尾が活用するようになった語

【例】 塩ビ　　トラブる　　本箱　　力む

⑤固有名〔固〕

人名・地名・商品名等。品詞が固有名詞となる語

【例】 大阪　　星野　　仙一　　ソニー

⑥記号〔記号〕

句読点・括弧などの補助記号や、箇条書きの項目名として使われた一字の片仮名などの記号。固有名以外のローマ字略語

【例】 ア　　イ　　A　　B　　OHP

(4) 用法

用法とは、「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」「名詞-普通名詞-サ変形状詞可能」の各語が、実際に当該文脈で名詞・形状詞・副詞のどの品詞で用いられているのか、また「名詞-普通名詞-助数詞可能」の語が名詞・助数詞のどちらの品詞で用いられているのかを示す情報である。

BCCWJ で付与した用法は、次のとおりである。

①名詞

「名詞-普通名詞-形状詞可能」「名詞-普通名詞-サ変形状詞可能」「名詞-普通名詞-副詞可能」の語が当該文脈で名詞として使われている場合に付与。

【例】 寛容，対話，協力を重んじる異文化間交流

ネットワーク担当の技術者が不足している

必要な場合には

②形状詞

「名詞-普通名詞-形状詞可能」「名詞-普通名詞-サ変形状詞可能」の語が当該文脈で形状詞として使われている場合に付与。

【例】 それらに必要な施設の整備
どの業種にも共通であるが

③副詞

「名詞-普通名詞-副詞可能」の語が当該文脈で副詞として使われている場合に付与。

【例】 笑福亭鶴笑氏が自ら考案した落語形式で
一時騒然とした雰囲気に包まれた

④助数詞

「名詞-普通名詞-助数詞可能」の語が当該文脈で助数詞として使われている場合に付与。

【例】 その約６割を落札している
前年と比べて１．８ポイント上昇している。

5.4 CSJ からの変更点

5.1 節で述べたように、BCCWJ では言語単位として CSJ と同じ長単位・短単位を採用した。しかし、長単位・短単位の認定規程は、CSJ の規程をそのまま用いるのではなく、修正等を行っている。また付加情報についても、CSJ とは異なるものとなっている。

そこで本節では、長単位・短単位および付加情報で、CSJ から変更した箇所のうち、主な箇所について述べることとする。

(1) 文節・長単位

文節・長単位の両方に関わる変更点としては、同格・並列の扱いがある。CSJ では学校文法と同様に、同格の関係にある要素、並列の関係にある要素を切り離していたが、BCCWJ では切り離さないこととした。

【例】

〔同格〕｜大江健三郎さんの｜長男=光さん｜

〔並列〕｜公正=妥当な｜実務慣行を｜集約した｜ものという｜意味で｜

長単位認定規程の主な変更点としては、次の 2 点が挙げられる。1 点目は、数量に関する規定である。CSJ では、以下に示すように、数量を表す要素は分割せず一続きとしていたが、長すぎるという指摘があった。

【例】 ｜１m=８０ｃｍ｜

そこで、BCCWJ では以下のように、単位の変わり目の後ろで分割することとした。

【例】　｜１ｍ‖８０ｃｍ｜

2 点目は、係り受けが関係する規定の簡素化である。CSJ では「体言連続の一部分が連体修飾語を受けている場合、その後ろで切る」「2 文節を受ける、若しくは 2 文節以上に係る接辞はその前後で切る」という規定があった。以下に例を示す。

【例】　｜項構造｜の｜曖昧性‖解消｜
｜円形劇場｜とか｜水路‖等｜

これらは、語と語との係り受けを厳密に考えようとしたところから作られたものである。しかし実際に単位分割をする際には、体言連続の一部分が連体修飾語を受けているかどうかの判定が難しいものがある。そのため、特に判定が難しい「体言＋以降、間（かん）、ごと、自体、達」という形式は、

【例】　｜住ん｜でる｜人=達｜

のように、体言と「達」などとを切り離さないという例外規定を設ける等、煩雑な規定となっていた。このことが単位認定のゆれにつながっていたため、BCCWJ では規定を簡素化することとした。具体的には、体言連続の一部分が連体修飾語を受けていても、体言連続を分割することなく、以下のように一続きとした。

【例】　｜項構造｜の｜曖昧性=解消｜
｜円形劇場｜とか｜水路=等｜

(2)　最小単位・短単位

CSJ の短単位や現代雑誌九十種調査の β 単位では、「一般」の外来語の最小単位も、和語・漢語と同様、二つの一次結合を 1 短単位としていた。例えば、「コールセンター」「オレンジ色」は共に 1 単位としていた。ただし、以下のような例外規定を設けた。

① 欧米語の冠詞・前置詞に当たるものは 1 最小単位を 1 短単位とする。

② β 単位では最小単位二つの一次結合が 7 拍を超える場合、短単位では同じく 10 拍を超える場合、結合させずに 1 最小単位を 1 短単位

外来語の最小単位二つの一次結合を 1 短単位とすることについては、CSJ の構築当初から和語・漢語に比べて長すぎるのではないかという指摘があった。このような指摘を踏まえ、上記②の拍数による例外規定を設けたが、10 拍を超える場合としたことに言語学的な意味があるわけではなく、そういう意味でこの例外規定にも問題があった。

そこで、BCCWJ では「一般」の外来語の最小単位は、原則として 1 最小単位を 1 短単位とし、和語・漢語の最小単位とは異なる扱いにした。

【例】　｜コール｜センター｜　　｜オレンジ｜色｜

(3)　付加情報

長単位・短単位とも品詞情報については、CSJ の品詞から大幅な改定を行った。普通名詞を例にして、CSJ の品詞体系と BCCWJ の品詞体系とを比較すると、表 5-10 のとおりである。

表 5-10: CSJ と BCCWJ との品詞の比較（名詞）

CSJ	BCCWJ（長単位）	BCCWJ（短単位）
名詞	名詞-普通名詞-一般	名詞-普通名詞-一般
		名詞-普通名詞-サ変可能
		名詞-普通名詞-形状詞可能
		名詞-普通名詞-サ変形状詞可能
		名詞-普通名詞-副詞可能
		名詞-普通名詞-助数詞可能
	名詞-助動詞語幹	名詞-助動詞語幹
名詞-固有名詞	名詞-固有名詞-一般	名詞-固有名詞-一般
	名詞-固有名詞-人名-一般	名詞-固有名詞-人名-一般
	名詞-固有名詞-人名-姓	名詞-固有名詞-人名-姓
	名詞-固有名詞-人名-名	名詞-固有名詞-人名-名
	名詞-固有名詞-地名-一般	名詞-固有名詞-地名-一般
	名詞-固有名詞-地名-国	名詞-固有名詞-地名-国
名詞-数詞	名詞-数詞	名詞-数詞

表 5-10 に示したとおり、BCCWJ では固有名詞を細分化するとともに、短単位において「名詞-普通名詞-サ変可能」「名詞-普通名詞-形状詞可能」「名詞-普通名詞-副詞可能」「名詞-普通名詞-助数詞可能」のように普通名詞を細分化した上で、「○○可能」という曖昧性を持たせた品詞を設けた。

BCCWJ の細分化した品詞体系は、形態素解析用辞書 UniDic の品詞体系に準拠したものである。BCCWJ は、1 億語から成る大規模なコーパスであるため、形態論情報の付与は自動解析システムにより行った。短単位解析には解析エンジン MeCab と解析用辞書 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を用いた。また、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。

このような自動解析システムの利用等の観点から、UniDic に準拠する形で CSJ から品詞体系を大幅に改定したのである。なお、品詞体系の改定に当たっては、UniDic の品詞体系をそのまま採用するのではなく、UniDic への未登録語の新規追加作業、コーパス修正作業の中で UniDic についても一部に改定を加える形で BCCWJ と UniDic とに共通な品詞体系を実現した。

なお、曖昧性を持たせた品詞は、短単位において、普通名詞以外にも、動詞、形容詞、名詞的接尾辞において設定している（表 5-7 参照）。一方、長単位では、実際の文脈において名詞として使われているのか、形状詞として使われているのかなどを判断し、それに基づいて品詞を付与したので、「名詞-普通名詞-○○可能」等の曖昧性を持たせた品詞は設けていない。

品詞情報のほか、CSJ では付与されなかった語種に関する情報も BCCWJ では付与した。

5.5　終わりに

以上、本章では、BCCWJ における長短 2 種類の言語単位の認定規程および付加情報について概略を述べた。

BCCWJ では、言語単位の設計に際し、①コーパスに基づく用例収集、各ジャンルの言語的特徴の解明に適した単位を設計する、②『日本語話し言葉コーパス』と互換性のある形態論情報を設計する、③国立国語研究所の語彙調査における知見を活用するという三つの方針を立てた。これら方針に沿って検討した結果、各ジャンルの言語的特徴の解明に適した長単位とコーパスに基づく用例収集短単位を採用した。

長単位・短単位は CSJ で採用した言語単位であるが、既に述べたように認定規程については、書き言葉用に修正・拡張を行った。また、長単位・短単位の解析に自動解析システムを活用したことなどから、品詞情報については、CSJ のものから大幅な改定を行った。

BCCWJ の長単位・短単位解析は、目標としていた解析精度 98%以上（コアデータについては 99%以上）を達成しており、高精度な解析を実現できたと言うことができる。

しかし、今後に残された課題もある。例えば、長単位の語彙素・語彙素読みの問題が挙げられる。語彙素・語彙素読みについては、表記や語形にかかわらず、同じ語であれば、同一の見出し（語彙素・語彙素読み）を付与するというのが基本的な方針であり、短単位については、そのとおりに設計されている。しかし長単位では、語形が異なる場合には、その語形に基づき異なる語彙素・語彙素読みを付与することとした。これにより、「あまり」と「あんまり」は、短単位では同一語と見なされるが、長単位では別語と見なされることなった。

これは、短単位解析結果を基に長単位を自動構成する際に、「語形」「書字形」の情報を利用したことによる。短単位解析結果を基に長単位の語彙素・語彙素読みを自動構成するに当たっては、小椋他（2011）に述べるように種々の問題があったが、その中で現時点での最善の手法として「語形」「書字形」の情報を利用することとした。しかしながら、今後改善が必要な点である。

参考文献

小椋秀樹（2006）「第 3 章 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』,133-186.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011a）「『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）」国立国語研究所内部報告書 LR-CCG-10-05-01

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011b）「『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.

国立国語研究所（1987）『国立国語研究所報告 89 雑誌用語の変遷』,秀英出版.

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』22,101-122,国書刊行会.

中野洋（1998）「言語の統計」『岩波講座言語の科学 9 言語情報処理』,149-199,岩波書店.

林大監修（1982）『角川小辞典 9 図説日本語』,角川書店.

冨士池優美・小西光・小椋秀樹・小木曽智信・小磯花絵（2011）「長単位に基づく媒体・カテゴリ間の品詞比率に関する分析」『特定領域研究「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』,273-280.

前田富祺（1985）『国語語彙史研究』,明治書院.

Uchimoto, K. and H. Isahara (2007). Morphological annotation of a large spontaneous speech corpus in Japanese, Proceedings of IJCAI, 1731-1737.

付録 5-A: 複合辞（助詞相当句）

語彙素読み	語彙素	品詞
カラシテ	からして	助詞-格助詞
カラスルト	からすると	助詞-格助詞
カラスレバ	からすれば	助詞-格助詞
サイニ	際に	助詞-格助詞
タメノ	ための	助詞-格助詞
トイウ	という	助詞-格助詞
トイッタ	といった	助詞-格助詞
トシテ	として	助詞-格助詞
ニアタッテ	にあたって	助詞-格助詞
ニアタリ	にあたり	助詞-格助詞
ニイタルマデ	に至るまで	助詞-格助詞
ニオイテ	において	助詞-格助詞
ニオケル	における	助詞-格助詞
ニカケテ	にかけて	助詞-格助詞
ニカンシテ	に関して	助詞-格助詞
ニカンスル	に関する	助詞-格助詞
ニサイシ	に際し	助詞-格助詞
ニサイシテ	に際して	助詞-格助詞
ニシテ	にして	助詞-格助詞
ニタイシ	に対し	助詞-格助詞
ニタイシテ	に対して	助詞-格助詞
ニタイスル	に対する	助詞-格助詞
ニツイテ	について	助詞-格助詞
ニツキ	につき	助詞-格助詞
ニトッテ	にとって	助詞-格助詞
ニヨッテ	によって	助詞-格助詞
ニヨリ	により	助詞-格助詞
ニヨル	による	助詞-格助詞
ニヨルト	によると	助詞-格助詞
ニヨレバ	によれば	助詞-格助詞
ニワタッテ	にわたって	助詞-格助詞
ニワタリ	にわたり	助詞-格助詞
ニワタル	にわたる	助詞-格助詞
ヲツウジテ	を通じて	助詞-格助詞
ヲハジメ	をはじめ	助詞-格助詞
ヲメグル	をめぐる	助詞-格助詞
ヲモッテ	をもって	助詞-格助詞
ダケデナク	だけでなく	助詞-副助詞

語彙素読み	語彙素	品詞
ニカギラズ	に限らず	助詞-副助詞
ノミナラズ	のみならず	助詞-副助詞
トイエドモ	といえども	助詞-係助詞
トイッテモ	といっても	助詞-係助詞
トキタラ	ときたら	助詞-係助詞
ニイタッテハ	に至っては	助詞-係助詞
ニシタッテ	にしたって	助詞-係助詞
ウエデ	上で	助詞-接続助詞
ウエニ	上に	助詞-接続助詞
ウエハ	上は	助詞-接続助詞
カトオモウト	かと思うと	助詞-接続助詞
カトオモッタラ	かと思ったら	助詞-接続助詞
ガハヤイカ	が早いか	助詞-接続助詞
カラトイッテ	からといって	助詞-接続助詞
カラニハ	からには	助詞-接続助詞
タトコロ	たところ	助詞-接続助詞
タトコロデ	たところで	助詞-接続助詞
タメニ	ために	助詞-接続助詞
トシタラ	としたら	助詞-接続助詞
トシテモ	としても	助詞-接続助詞
トスレバ	とすれば	助詞-接続助詞
トドウジニ	と同時に	助詞-接続助詞
トトモニ	とともに	助詞-接続助詞
トハイエ	とはいえ	助詞-接続助詞
ニカカワラズ	に関わらず	助詞-接続助詞
ニシタガイ	にしたがい	助詞-接続助詞
ニシタガッテ	にしたがって	助詞-接続助詞
ニシテハ	にしては	助詞-接続助詞
ニシテモ	にしても	助詞-接続助詞
ニシロ	にしろ	助詞-接続助詞
ニセヨ	にせよ	助詞-接続助詞
ニツレ	につれ	助詞-接続助詞
ニツレテ	につれて	助詞-接続助詞
ニモカカワラズ	にも関わらず	助詞-接続助詞
モノノ	ものの	助詞-接続助詞
ヤイナヤ	や否や	助詞-接続助詞
ワリニ	わりに	助詞-接続助詞

付録 5-B: 複合辞（助動詞相当句）

語彙素読み	語彙素	品詞
カモシレナイ	かもしれない	助動詞
カモシレマセン	かもしれません	助動詞
コトガアル	ことがある	助動詞
コトガデキル	ことができる	助動詞
コトトナル	こととなる	助動詞
コトニスル	ことにする	助動詞
コトニナル	ことになる	助動詞
コトハナイ	ことはない	助動詞
コトモアル	こともある	助動詞
コトモナイ	こともない	助動詞
ザルヲエナイ	ざるを得ない	助動詞
シカナイ	しかない	助動詞
ズニハイラレナイ	ずにはいられない	助動詞
タライイ	たらいい	助動詞
ツツアル	つつある	助動詞
テアル	てある	助動詞
デアル	である	助動詞
テイク	ていく	助動詞
テイタダク	ていただく	助動詞
テイル	ている	助動詞
テオク	ておく	助動詞
テオル	ておる	助動詞
テクダサル	てくださる	助動詞
テクル	てくる	助動詞
テクレル	てくれる	助動詞
テシカタガナイ	て仕方がない	助動詞
テシマウ	てしまう	助動詞
テショウガナイ	て仕様がない	助動詞

語彙素読み	語彙素	品詞
テタマラナイ	てたまらない	助動詞
デナイ	でない	助動詞
テナラナイ	てならない	助動詞
デハアリマセン	ではありません	助動詞
テハイケナイ	てはいけない	助動詞
デハナイ	ではない	助動詞
テハナラナイ	てはならない	助動詞
テホシイ	てほしい	助動詞
テミル	てみる	助動詞
デモアル	でもある	助動詞
テモイイ	てもいい	助動詞
テモラウ	てもらう	助動詞
テヤル	てやる	助動詞
ナイデハイラレナイ	ないではいられない	助動詞
ナクテハナラナイ	なくてはならない	助動詞
ナケレバナラナイ	なければならない	助動詞
ニスギナイ	に過ぎない	助動詞
ニチガイナイ	に違いない	助動詞
ニホカナラナイ	にほかならない	助動詞
ネバナラナイ	ねばならない	助動詞
ノダ	のだ	助動詞
ノデアル	のである	助動詞
ノデス	のです	助動詞
ノデハナイ	のではない	助動詞
バイイ	ばいい	助動詞
マデモナイ	までもない	助動詞
ワケニハイカナイ	わけにはいかない	助動詞

付録 5-C: 連語

語彙素読み	語彙素	品詞
イッケンラクチャク	一件落着	名詞
トントンビョウシ	とんとん拍子	名詞
ニクマレグチ	憎まれ口	名詞
ヒトイチバイ	人一倍	名詞
ミタメ	見た目	名詞
アマイモノギライ	甘い物嫌い	名詞・形状詞
オキニイリ	御気に入り	名詞・形状詞
クワズギライ	食わず嫌い	名詞・形状詞
タベズギライ	食べず嫌い	名詞・形状詞
マケズギライ	負けず嫌い	名詞・形状詞
イツモドオリ	何時も通り	名詞・副詞
イママデドオリ	今まで通り	名詞・副詞
カミヒトエ	紙一重	形状詞
コトバズクナ	言葉少な	形状詞
コトモナゲ	事も無気	形状詞
ワガモノガオ	我が物顔	形状詞
イイカゲン	良い加減	形状詞・副詞
シュタル	主たる	連体詞
アイカワラズ	相変わらず	副詞
アオアオト	青々と	副詞
イカニモ	如何にも	副詞
イツカ	何時か	副詞
イマヤ	今や	副詞
カナラズシモ	必ずしも	副詞
クログロト	黒々と	副詞
コツゼント	忽然と	副詞
シキリト	頻りと	副詞
シキリニ	頻りに	副詞
シゼント	自然と	副詞
ジツハ	実は	副詞
スクナクトモ	少なくとも	副詞
ソコハカト	其処は彼と	副詞
ダンジテ	断じて	副詞
ドウシテ	どうして	副詞
ドウシテモ	どうしても	副詞
ドウニカ	どうにか	副詞
ドウニモ	どうにも	副詞
ドウニモコウニモ	どうにもこうにも	副詞
ナニヨリ	何より	副詞
ナンダカ	何だか	副詞
ナンデモ	何でも	副詞
ナント	何と	副詞
ナントカ	何とか	副詞

語彙素読み	語彙素	品詞
フカブカト	深々と	副詞
ベツニ	別に	副詞
ホウフツト	彷彿と	副詞
ホソボソト	細々と	副詞
モウゼント	猛然と	副詞
シカシナガラ	然しながら	接続詞
シタガッテ	従って	接続詞
スルト	すると	接続詞
ソウシテ	そうして	接続詞
ソレカラ	其れから	接続詞
ソレデ	其れで	接続詞
ソレデハ	其れでは	接続詞
ソレデモ	其れでも	接続詞
ソレトモ	其れとも	接続詞
ダガ	だが	接続詞
ダカラ	だから	接続詞
ダケレド	だけれど	接続詞
ダケレドモ	だけれども	接続詞
ダッタラ	だったら	接続詞
ダッテ	だって	接続詞
チナミニ	因みに	接続詞
ッテイウカ	って言うか	接続詞
デスガ	ですが	接続詞
デスカラ	ですから	接続詞
デスケレド	ですけれど	接続詞
デスケレドモ	ですけれども	接続詞
デハ	では	接続詞
デモ	でも	接続詞
トコロガ	所が	接続詞
トコロデ	所で	接続詞
ナノデ	なので	接続詞
ナノニ	なのに	接続詞
ナラビニ	並びに	接続詞
マタハ	又は	接続詞
ユエニ	故に	接続詞
ヨウスルニ	要するに	接続詞
ヨッテ	因って	接続詞
オメデトウ	御めでとう	感動詞
ゴメンナサイ	御免為さい	感動詞
スマン	済まん	感動詞
スミマセン	済みません	感動詞
キモチワルガル	気持ち悪がる	動詞
カッコウヨイ	格好良い	形容詞

付録 5-D: 接頭的要素

語彙素読み	語彙素	品詞	注記
アイ	相	接頭辞	「相」と１最小単位との結合体が名詞である場合は除く。（相=乗り，相=討ち）
オ	御	接頭辞	次に挙げるものは，後の部分と併せて1最小単位とする。［お足，おいた，お家（芸・流），お薄，おかか，お鏡，おかき，お陰，おかず，お河童，おかま，おかみ，おから，おかわ，お冠，御形，おぐし，お好み（焼き），おこわ，お下げ（髪），お差し，おさつ，おざなり，おざぶ，おさん（どん），おしっこ，おしぼり，おしめ，おじや，おしゃぶり，お釈迦，お洒落，お節，お宅〈代名詞〉，お多福，お陀仏，お玉，おつむ，お手（上げ・の物），おでき，おでまし，お転婆，お伽（話），お腹，お成り，お握り，お主，お寝しょ，お萩，おはこ〈十八番の意〉，おはよう，お払い（箱），おひたし，お冷や，お袋，おふる，おまえ，おまけ，おませ，おまる，お巡り，お娘，おむすび，おむつ，お目見え，お漏らし，おやつ］
オン	御	接頭辞	次に挙げるものは，後の部分と併せて1最小単位とする。［御曹司，御大，御中，御身］
カク	各	接頭辞	漢語の1最小単位と結合したものは除く。（各=国，各=地）
コン	今	接頭辞	漢語の1最小単位と結合したものは除く。（今=回，今=度）
ゴ	御	接頭辞	次に挙げるものは，後の部分と併せて1最小単位とする。［御形，御供，御所，御新，御仁，御神火，御前，御饌，御託，御殿，御伝，御悩，御飯，御辺，御免，御覧，御料，御寮］
ショ	諸	接頭辞	漢語の1最小単位と結合したものは除く。（諸=国，諸=所）
ゼン	全	接頭辞	漢語の1最小単位と結合したものは除く。（全=国，全=社）
タイ	対	名詞-普通名詞-一般	漢語の1最小単位と結合したものは除く。（対=米，対=人）
ホン	本	接頭辞	「この」の意。漢語の1最小単位と結合したものは除く。（本件）
ミ	御	接頭辞	次に挙げるものは後の部分と併せて1最小単位とする。［御生，御門，御溝，御酒，御籤，御髪，御座，御食，御子，御輿，御言，御簾，御衣，御台，御霊，御堂，御息所，御幸，御代］
ホノ	仄	接頭辞	「ほのか」「ほのめく」「ほのぼの」「ほのめかす」は除く。

付録 5-E: 接尾的要素

語彙素読み	語彙素	品詞	注記
アガリ	上がり	接尾辞-名詞的-一般	前にその職業・身分だった者の意。
アグネル	あぐねる	動詞-非自立可能	
アソバス	遊ばす	動詞-非自立可能	
アタウ	能う	動詞-非自立可能	動作・状態の継続・進行を表す。
アタリ	当たり	接尾辞-名詞的-副詞可能	
アテ	宛	接尾辞-名詞的-一般	名あての意。「名宛(人)」の「宛」は除く。
アテ	宛	接尾辞-名詞的-一般	「〜に対して」の意。
アリ	有り	名詞-普通名詞-一般	「大有り」「神在」「徒有り」「訳有り」の「有り（在り）」は除く。
アル	有る	動詞-非自立可能	動作・状態の継続・進行を表す。
イタス	致す	動詞-非自立可能	
イル	居る	動詞-非自立可能	
ウエ	上	接尾辞-名詞的-一般	
エル	得る	動詞-非自立可能	「〜することができる」の意。
オエル	終える	動詞-非自立可能	
オオセル	果せる	動詞-非自立可能	「すっかり終える」の意。
オクレル	遅れる	動詞-非自立可能	
オル	居る	動詞-非自立可能	動作・状態の継続・進行を表す。
オワス	御座す	動詞-非自立可能	動作・状態の継続・進行を表す。
オワル	終わる	動詞-非自立可能	
カ	化	接尾辞-名詞的-一般	漢語の１最小単位と結合したものは除く。（特=化，液=化）
ガカル	がかる	接尾辞-動詞的	
カタ	方	接尾辞-名詞的-一般	「仕方」の「方」は除く。
ガタイ	難い	接尾辞-形容詞的	「有り難い」の「難い」は除く。
カタガタ	旁	接尾辞-名詞的-一般	
ガチ	勝ち	接尾辞-形状詞的	
ガテラ	がてら	接尾辞-名詞的-副詞可能	
カネル	兼ねる	接尾辞-動詞的	
ガル	がる	接尾辞-動詞的	助動詞「たがる」の「がる」は除く。
カワス	交わす	動詞-非自立可能	「互いに〜する」の意。
カン	間	接尾辞-名詞的-副詞可能	漢語の1最小単位と結合したものは除く。（空=間，車=間）
ギミ	君	接尾辞-名詞的-一般	
キル	切る	動詞-非自立可能	「すっかり〜し終える」の意。
クサイ	臭い	接尾辞-形容詞的	「〜めいた感じがする」という意。望ましくない意を強める用法。「かび臭い」「焦げ臭い」の「くさい」は除く。
クダサル	下さる	動詞-非自立可能	
グルミ	ぐるみ	接尾辞-名詞的-一般	
クン	君	接尾辞-名詞的-一般	「同君」の「君」は除く。
ゲ	気	接尾辞-形状詞的	
ケイ	系	接尾辞-名詞的-一般	漢語の1最小単位と結合したものは除く。（文=系，日=系）
ゴ	後	接尾辞-名詞的-一般	漢語の1最小単位と結合したものは除く。（戦=後，老=後）
ゴ	御	接尾辞-名詞的-一般	
コト	事	名詞-普通名詞-一般	
ゴト	ごと	接尾辞-名詞的-副詞可能	「〜も一緒に」の意。
ゴト	毎	接尾辞-名詞的-一般	そのもの一つ一つ，その時その時の意。
コナス	熟す	動詞-非自立可能	「うまく〜する」の意。
サ	さ	接尾辞-名詞的-一般	「そうだ」「過ぎる」が接続するときの「なさ」「良さ」の「さ」，ケシ型形容詞に付く「さ」，「憂さ」の「さ」は除く。
サス	さす	動詞-非自立可能	
サス	止す	動詞-非自立可能	
サマ	様	接尾辞-名詞的-一般	

語彙素読み	語彙素	品詞	注記
サン	さん	接尾辞-名詞的-一般	
ジ	時	接尾辞-名詞的-副詞可能	漢語の1最小単位と結合したものは除く。（戦=時）
シキ	式	接尾辞-名詞的-一般	形式・方法などの意。漢語の1最小単位と結合したものは除く。（洋=式，正=式）
シナ	しな	接尾辞-名詞的-副詞可能	
ジミル	染みる	接尾辞-動詞的	
ジュウ	中	接尾辞-名詞的-副詞可能	
ジョウ	上	接尾辞-名詞的-副詞可能	漢語の1最小単位と結合したものは除く。（機=上，車=上）
ジョウ	状	接尾辞-名詞的-一般	「～の形・有り様」の意。漢語の1最小単位と結合したものは除く。（液=状）
スギル	過ぎる	動詞-非自立可能	
ズク	尽く	接尾辞-名詞的-一般	
ズクメ	尽くめ	接尾辞-名詞的-一般	
スル	為る	動詞-非自立可能	漢語の1最小単位と結合したものは除く（対=する，信=ずる）。「～んずる」という形式は除く（甘ん=ずる，重ん=ずる）。
セイ	性	接尾辞-名詞的-一般	物事の性質・傾向の意。漢語の1最小単位と結合したものは除く。（特=性，急=性）
ソウ	そう	形状詞-助動詞語幹	様態の助動詞「そうだ」の語幹に当たるもの。
ソウ	そう	名詞-助動詞語幹	伝聞の助動詞「そうだ」の語幹に当たるもの。
ソコナウ	損なう	動詞-非自立可能	
ソコネル	損ねる	動詞-非自立可能	
ソビレル	そびれる	動詞-非自立可能	
ソンズル	損ずる	動詞-非自立可能	
タイ	対	名詞-普通名詞-一般	
ダス	出す	動詞-非自立可能	「～し始める」という意。
タチ	達	接尾辞-名詞的-一般	
タテマツル	奉る	動詞-非自立可能	
タマウ	給う	動詞-非自立可能	
ダラケ	だらけ	接尾辞-形状詞的	
チャン	ちゃん	接尾辞-名詞的-一般	
チュウ	中	接尾辞-名詞的-副詞可能	漢語の1最小単位と結合したものは除く。（空=中）
ツイデ	序で	名詞-普通名詞-一般	
ツキ	付き	接尾辞-名詞的-一般	「札付き」（知れわたっていること，悪い評判が世間に広まっている人の意）は除く。
ツクス	尽くす	動詞-非自立可能	「すっかり～する」という意。
ツケル	付ける	動詞-非自立可能	習慣の意。
ッコ	っこ	接尾辞-名詞的-一般	「～すること」の意。
ッコ	っこ	接尾辞-名詞的-一般	「～比べ」「互いに～する」という意。
ツヅク	続く	動詞-非自立可能	「引き続く」「打ち続く」等，動作継続の動詞に接続しないものは除く。
ツヅケル	続ける	動詞-非自立可能	「打っ続ける」等，動作継続の動詞に接続しないものは除く。
ヅライ	辛い	接尾辞-形容詞的	
テキ	的	接尾辞-形状詞的	漢語の1最小単位と結合したものは除く。（人=的，端=的）
デキル	出来る	動詞-非自立可能	
トウ	等	接尾辞-名詞的-一般	
ドウシ	同士	接尾辞-名詞的-一般	
トオス	通す	接尾辞-動詞的	
トオリ	通り	名詞-普通名詞-副詞可能	それと同じ状態であるという意。
ドノ	殿	接尾辞-名詞的-一般	
トモ	共	接尾辞-名詞的-副詞可能	全部の意。

語彙素読み	語彙素	品詞	注記
ドモ	共	接尾辞-名詞的-一般	
ナイ	内	接尾辞-名詞的-一般	漢語の1最小単位と結合したものは除く。（室=内，社=内）
ナガラ	乍ら	接尾辞-名詞的-一般	
ナサル	為さる	動詞-非自立可能	
ナシ	無し	名詞-普通名詞-一般	「有る無し」「形無し」「底無し」「台無し」「人で無し」「人無し」「幕無し」「間無し」「道無し」「文無し」「休み無し」の「無し」は除く。
ナミ	並み	接尾辞-名詞的-一般	その類と同じ，又は同じ程度であることを表す。
ナリ	形	接尾辞-名詞的-一般	そのもの相応である様の意。
ナリ	形	接尾辞-名詞的-一般	「～するまま」「～するに従う様」の意。
ナレル	慣れる	動詞-非自立可能	
ニクイ	難い	接尾辞-形容詞的	醜悪の意の「醜い」は除く。
ヌク	抜く	動詞-非自立可能	「終わりまでする」という意。
ハジメル	始める	動詞-非自立可能	
ハタス	果たす	動詞-非自立可能	「すっかり～し終える」の意。
ハテル	果てる	動詞-非自立可能	「すっかり～する」「～し終わる」という意。
ハナシ	放し	接尾辞-形状詞的	
バム	ばむ	接尾辞-動詞的	
ハン	版	名詞-普通名詞-一般	漢語の1最小単位と結合したものは除く。（新=版）
フウ	風	接尾辞-名詞的-一般	様子の意。漢語の1最小単位と結合したものは除く。（和=風，古=風）
ブリ	振り	接尾辞-名詞的-一般	それだけの時間が過ぎたという意を表す。
ブリ	振り	接尾辞-名詞的-一般	様子・状態の意。
ブル	振る	接尾辞-動詞的	「そのように振る舞う」という意。
ブン	分	名詞-普通名詞-一般	
ポイ	ぽい	接尾辞-形容詞的	形容詞語幹に接続する「ぽい」は除く。「いがらっぽい」の「ぽい」は除く。
ポッチ	ぽっち	接尾辞-名詞的-一般	
マエ	前	名詞-普通名詞-副詞可能	
マクル	捲る	動詞-非自立可能	
マス	坐す	動詞-非自立可能	
マワリ	周り	接尾辞-名詞的-一般	
ミタイ	みたい	形状詞-助動詞語幹	
ムキ	向き	接尾辞-名詞的-一般	
ムケ	向け	接尾辞-名詞的-一般	
メ	奴	接尾辞-名詞的-一般	ののしる語。
メ	奴	接尾辞-名詞的-一般	謙そんの意。
メ	目	接尾辞-名詞的-一般	順序を表す。
メク	めく	接尾辞-動詞的	擬態語的なものの「めく」は除く。（きら=めく，ざわ=めく）
モウス	申す	動詞-非自立可能	
ヤガル	やがる	接尾辞-動詞的	
ヤスイ	易い	接尾辞-形容詞的	
ヨイ	良い	形容詞-非自立可能	
ヨウ	様	形状詞-助動詞語幹	助動詞「ようだ」の語幹に当たるもの。
ヨウ	様	接尾辞-名詞的-一般	方法の意。
ヨウ	用	接尾辞-名詞的-一般	漢語の1最小単位と結合したものは除く。（学=用）
ラ	等	接尾辞-名詞的-一般	複数を表す。
ラ	等	接尾辞-名詞的-一般	事物をおおよそに指す。
ラシイ	らしい	接尾辞-形容詞的	助動詞「らしい」は除く。
リュウ	流	接尾辞-名詞的-一般	流派の意。
ルイ	類	接尾辞-名詞的-一般	漢語の1最小単位と結合したものは除く。（人=類）
ワスレル	忘れる	動詞-非自立可能	
ワタル	渡る	動詞-非自立可能	「辺り一面に～する」という意。
ワタル	渡る	動詞-非自立可能	「徹底的に～する」という意。

第 6 章　形態論情報付きデータ（TSV）

小木曽 智信

6.1　形態論情報付きデータの概要

本マニュアル第 1 章、第 2 章、第 5 章で述べたように、BCCWJ にはいわゆる形態素解析が施されており、コーパスの重要な特徴のひとつとなっている。形態素という表現は、自然言語処理と言語学とで異なる意味で用いられる傾向にあり、日本語の場合、特に誤解を招きやすいと考えられるので、我々は『日本語話し言葉コーパス』のときから、形態素情報と呼ばずに「形態論情報」という名称を用いてきている。BCCWJ には短単位と長単位による二重の形態論情報が付与されていることも既に述べたとおりである。

「形態論情報付きデータ」は BCCWJ の全サンプルのテキストに対して短単位・長単位の形態論情報（第 5 章参照）を付与したテキストデータである。形態論情報付きデータとして、表形式データ（TSV データ、タブ区切りテキスト）と形態論情報付き統合形式 XML データ（M-XML）の 2 形式を用意した。さらにそれぞれの形式について、後述する数字変換処理の有無による 2 種類のデータ（*_OT、*_NT）を用意した。したがって DVD には、つごう 4 種類の形態論情報付きデータが格納されている（データの格納場所は 1.3 節を参照）。

短単位・長単位の形態論情報は、TSV・M-XML の両形式とも同じ内容が付与されており、同一部分の短単位・長単位が異なって付与されていることはない。

短単位は、全体を UniDic によって解析した結果に対して部分的に人手による修正を施したものである。特定バージョンの UniDic で解析した結果そのままではないため、BCCWJ のテキストと UniDic を用いたとしても同一の内容を自動的に作成することはできない。長単位も同様である。長単位についても、長単位解析器 Comainu によって短単位を組み上げたのち、形態論情報データベース上での自動処理と人手による修正を経ているため、同一内容のデータを自動で作成することはできない。

すべての形態論情報は、冗長となることを恐れず、必要と考えられるすべての情報をテキストで保持している。短単位の形態論情報は、原則として UniDic の辞書見出しと対応づけることができるため ID のみで表現することも可能だが、あえてこの方法は採っていない。なお、TSV・M-XML の両形式とも、書誌情報は含んでいないので、必要な場合にはサンプル ID を元に別途取得する必要がある。

6.2　数字変換処理（NumTrans）

6.2.1 数字変換処理と 2 種類の本文

形態論情報付きデータは、BCCWJ の全てのテキストに対して形態素解析を行って情報を付与したものだが、形態論情報を付与するにあたって、本文をそのまま解析対象としたデ

ータ（M-XML_OT、TSV_OT）と、解析前に数字を解析しやすい表記に変換する処理（NumTrans）を行ったデータ（M-XML_NT、TSV_NT）の二通りを用意している。

NumTrans による変換とは、数字列を含む文章について、これを読みあげた場合の形態論情報を付与できるようにするために、形態素解析の前処理として数字列のテキストを解析しやすい表記に置き換えたものである。具体的には次の例のような処理である。なお、解析に影響を与えない一桁の数字は変換されない。

５００円　→　五百円
５０，０００円　→　五万円
２０１５年に公開した　→　二千十五年に公開した

元の本文「５００円」は「５」「０」「０」「円」（ゴ/レイ/レイ/エン）、「５０，０００円」は「５」「０」「,」「０」「０」「０」「円」（ゴ/レイ/ /レイ/レイ/レイ/エン）と解析されるが、NumTrans 後の本文「五百円」は「五百」「円」（ゴヒャク/エン）、「五万円」は「五」「万」「円」（ゴ/マン/エン）と短単位の規定どおりに解析される。また、「２０１５年」は「２」「０」「１」「５」「年」（ニ/レイ/イチ/ゴ/ネン）と解析されるが、NumTrans 後の本文「二千十五年」は「二千」「十」「五」「年」（ニセン/ジュウ/ゴ/ネン）と解析される。

分数が現れる箇所（fraction タグが付けられた箇所）では、次のように読み進める順にあわせて順序を入れ替える処理も NumTrans によって行なわれる。

２／４５　→　四十五 分 ２

これは、「２／４５」が「四十五（ヨンジュウゴ）分（ブン）ノ２（ニ）」と読み上げられるのに合わせた処理である。ただし、「／」は「分」（ブン）と変換されるが、通常なら読み添えられる「ノ」の部分は出力されない。

以上のように、元の本文が、数字列を個々の数字の連なりとして扱ったものとなるのに対し、NumTrans 後の本文は、当該部分を読み上げたものとしてそれを短単位に解析することになるため、当該部分の形態論情報は語数を含めて大きく異なるものとなる（表 6-1）。

この NumTrans 処理は、出現した文字列にもとづいて自動で行われているため、手作業で修正が施されたコアデータ以外のサンプルでは変換を誤っている可能性がある。

このような変換処理のため、NumTrans 処理が行われたデータ（M-XML_NT、TSV_NT）の表層文字列を組み上げたテキストは、文字ベースの C-XML（第 4 章参照）から抜き出したテキストとは一致しない。ただし、M-XML_NT、TSV_NT の両形式とも、C-XML と同じテキストを取り出すことができるように原文の情報が保持されている。形態論情報付きデータでは、元の文字列を「原文文字列（originalText）」、変換後の文字列（形態素解析の対象となった表層形）を「書字形出現形（orthToken）」と呼んで区別している。

表 6-1: NumTrans の有無と短単位

NumTrans	テキスト	発音形	語彙素読み	語彙素	品詞	語種
なし (*_OT)	5	ゴ	ゴ	五	名詞-数詞	漢
	0	レー	レイ	零	名詞-数詞	漢
	,			,	補助記号-読点	記号
	0	レー	レイ	零	名詞-数詞	漢
	0	レー	レイ	零	名詞-数詞	漢
	0	レー	レイ	零	名詞-数詞	漢
	円	エン	エン	円-助数詞	名詞-普通名詞-助数詞可能	漢
あり (*_NT)	五	ゴ	ゴ	五	名詞-数詞	漢
	万	マン	マン	万	名詞-数詞	漢
	円	エン	エン	円-助数詞	名詞-普通名詞-助数詞可能	漢

6.2.2 BCCWJ のバージョンと数字変換処理

M-XML_NT は、BCCWJ-DVD 版（Version 1.0）の M-XML に相当するものであり、TSV_NT は Version 1.0 の TSV に相当するものであるが、いずれも文境界の修正がなされアップデートされている（第 8 章参照）。一方、C-XML は Version 1.0 から変更されていない。まとめると表 6-2 のようになる。

表 6-2: BCCWJ Ver.1.0 データと Ver.1.1 データの関係

文書形式	NumTrans	Version 1.0	Version 1.1
TSV	適用	TSV	TSV_NT（更新）
	非適用	―	TSV_OT（新規）
M-XML	適用	M-XML	M-XML_NT（更新）
	非適用	―	M-XML_OT（新規）
C-XML	非適用	C-XML	C-XML（変更なし）

6.2.3 数字変換処理と短単位・長単位の語数

6.2.1 節で述べたとおり、NumTrans の有無によって短単位の語数は変化する。一方、長単位は NumTrans によって語数は変わらない。これは、NumTrans 後の短単位（NT）をベースに組み上げられた長単位（NT 長単位）のタグの範囲を変えないで、NumTrans 前の短単位（OT）を組み上げて長単位情報を付け直しているためである。すなわち、OT の長単位情報は NT の長単位境界を前提としてつけられている。この関係を以下に図示する。

OT テキスト：２０１５年に公開する

↓NumTrans

NT テキスト：二千十五年に公開する

NT 短単位	二千 ニセン	十 ジュウ	五 ゴ	年 ネン	に ニ	公開 コウカイ	する スル
NT 長単位	二千十五年 ニセンジュウゴネン				に ニ	公開する コウカイスル	

OT 短単位	2 ニ	0 レイ	1 イチ	5 ゴ	年 ネン	に ニ	公開 コウカイ	する スル
OT 長単位	２０１５年 ニレイイチゴネン					に ニ	公開する コウカイスル	

6.3　総語数

形態論情報付きデータの、レジスター別の短単位・長単位の数は表 6-3 のとおりである（TSV・M-XML 共通）。ここでは、コアを別立てし、空白・記号等は除外して計算している。

表 6-3: レジスターごとの短単位・長単位数

レジスター	サンプル数	短単位数 NT	短単位数 OT	長単位数 (OT・NT とも)
出版・新聞	1,133	1,061,729	1,067,236	773,395
出版・新聞コア	340	308,504	310,568	224,140
出版・雑誌	1,910	4,242,224	4,291,868	3,320,944
出版・雑誌コア	86	202,268	203,834	159,883
出版・書籍	10,034	28,348,233	28,450,702	22,688,156
出版・書籍コア	83	204,050	204,425	169,730
図書館・書籍	10,551	30,377,863	30,443,244	25,092,639
特定目的・白書	1,438	4,685,801	4,723,895	2,970,971
特定目的・白書コア	62	197,011	198,842	129,646
特定目的・ベストセラー	1,390	3,742,261	3,745,868	3,185,745
特定目的・知恵袋	90,507	10,162,945	10,208,917	8,534,253
特定目的・知恵袋コア	938	93,932	94,289	78,770
特定目的・ブログ	52,209	10,101,397	10,180,579	8,209,800
特定目的・ブログコア	471	92,746	93,367	75,242
特定目的・法律	346	1,079,146	1,079,156	706,313
特定目的・国会会議録	159	5,102,469	5,102,796	4,007,842
特定目的・広報紙	354	3,755,161	3,819,646	2,308,452
特定目的・教科書	412	928,447	933,356	746,170
特定目的・韻文	252	225,273	225,295	202,425
合計	172,675	104,911,460	105,377,883	83,584,516

6.4　TSV 形式データ

TSV 形式データは、上記の形態論情報をタブ区切りの表形式テキストデータにしたものであり、BCCWJ の Web 検索サービス『中納言』の元になっているデータである。短単位・長単位ごとに、別のテーブルとなっており、それぞれがレジスターごとに分割されている。テキストデータの文字符号化方式は UTF-8（BOM なし）である。

短単位･長単位 TSV はそれぞれ単独でも利用可能なように重複した情報を保持している。

6.4.1 短単位 TSV のフィールド

短単位 TSV のフィールド中身は表 6-4 のとおりである（左から順）。1 短単位が 1 レコード（行）となっている。文字開始/終了位置・連番・出現形開始/終了位置については 6.4.3 で解説する。

表 6-4: 短単位 TSV のフィールド

フィールド名	備考
レジスター	
サンプル ID	
文字開始位置	原文文字列のサンプル頭からのオフセット値（10 きざみ）
文字終了位置	
連番	サンプル内での長単位の並び順（10 きざみ）
出現形開始位置	書字形出現形のサンプル頭からのオフセット値（10 きざみ）
出現形終了位置	
固定長フラグ	0:固定長でない、1:固定長
可変長フラグ	0:可変長でない、1:可変長
文頭ラベル	M-XML の sentence タグ開始位置は「B」、それ以外は「I」
語彙表 ID	書字形出現形のレベルで語を識別する ID （桁数が大きいため bigint 型が必要）
語彙素 ID	UniDic の語彙素を識別する ID
語彙素	短単位情報
語彙素読み	
語彙素細分類	
語種	
品詞	
活用型	
活用形	
語形	
用法	
書字形	
書字形出現形	
原文文字列	
発音形出現形	

6.4.2 長単位 TSV のフィールド

長単位 TSV のフィールド中身は表 6-5 のとおりである（左から順）。1 長単位が 1 レコード（行）となっている。

表 6-5: 長単位 TSV のフィールド

フィールド名	備考
レジスター	
サンプル ID	
出現形開始位置	書字形出現形のサンプル頭からのオフセット値（10 きざみ）
出現形終了位置	
文節	B:文節、空文字:文節でない
短長相違フラグ	短単位と長単位の範囲が一致しているかどうか 0:短長一致、1:短長相違
固定長フラグ	0:固定長でない、1:固定長
可変長フラグ	0:可変長でない、1:可変長
語彙素	長単位情報
語彙素読み	
語種	
品詞	
活用型	
活用形	
語形	
書字形	
書字形出現形	
原文文字列	
発音形出現形	
連番	サンプル内での長単位の並び順（10 きざみ）
文字開始位置	原文文字列のサンプル頭からのオフセット値（10 きざみ）
文字終了位置	
文頭ラベル	B:文頭、I:文頭以外

6.4.3 文字位置と連番

TSV における「文字開始位置」「出現形開始位置」などのサンプル頭からのオフセット値は、図 6-1、表 6-6 のように 10 開始、10 きざみで文字間に割り振られている。「連番」は、短単位・長単位に対して 10 開始、10 きざみで振られている。

図 6-1: 文字位置と連番の対応

表 6-6: 形態素と文字位置・連番の対応

文字 開始位置	文字 終了位置	連番	出現形 開始位置	出現形 終了位置	書字形出 現形	原文文字 列
10	30	10	10	30	日本	
30	40	20	30	40	語	
40	50	30	40	50	の	

「文字開始位置」「出現形開始位置」の別は、6.2.1 節で述べた「原文文字列」「書字形出現形」に対応し、前者は NumTrans 前、後者は NumTrans 後のファイル先頭からの文字位置である。したがって「文字開始位置」と「出現形開始位置」は NumTrans 処理がなされたデータにおいてのみ違いがあり、NumTrans 処理がなされていない場合には一致する。終了位置についても同様である。

NumTrans 処理がなされたデータの「文字開始位置」「出現形開始位置」「連番」の対応は図 6-2 のようになる。

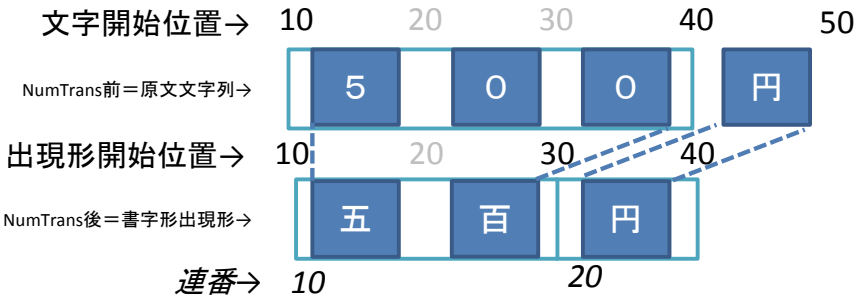


図 6-2: NumTrans されたテキストの文字位置と連番の対応

短単位情報中の「原文文字列」は、数字変換前の文字列であり、これも NumTrans 処理がなされたデータ（_NT）においてのみ当該箇所に出力される（表 6-7）。

表 6-7: NumTrans されたテキストの形態素と文字位置・連番の対応

文字 開始位置	文字 終了位置	連番	出現形 開始位置	出現形 終了位置	書字形出 現形	原文文字 列
10	40	10	10	30	五百	５００
40	50	20	30	40	円	

なお、NumTrans 後の文字列が複数の単位に分割される場合には、表 6-8 のように当該範囲内のすべてに同じ原文文字列が付与されている。

表 6-8: 数字変換箇所の原文文字列との対応例

文字 開始位置	文字 終了位置	連番	出現形 開始位置	出現形 終了位置	書字形出 現形	原文文字 列
10	50	10	10	30	二千	２０１５
10	50	20	30	40	十	２０１５
10	50	30	40	50	五	２０１５
50	60	40	50	60	年	

6.5 M-XML の形態論情報タグ

形態論情報付き統合形式 XML データ（M-XML）は、言語構造を一定程度反映させた XML フォーマットであり、形態論情報についても短単位・長単位の階層構造を維持したまま埋め込み、言語構造に関わる情報を扱いやすくしている。M-XML からこの部分だけを抜き出すと次のようになっている。

```
<LUW B="B" SL="v" l_lemma="公共工事請け負い金額" l_lForm="コウキョウコウジウケオイキンガク" l_wType="混" l_pos="名詞-普通名詞-一般" >
  <SUW lemma="公共" lForm="コウキョウ" wType="漢" pos="名詞-普通名詞-一般" pron="コーキョー">
    公共
  </SUW>
  <SUW lemma="工事" lForm="コウジ" wType="漢" pos="名詞-普通名詞-サ変可能" pron="コージ">
    工事
  </SUW>
  <SUW lemma="請け負い" lForm="ウケオイ" wType="和" pos="名詞-普通名詞-一般" pron="ウケオイ">
    請負
  </SUW>
  <SUW lemma="金額" lForm="キンガク" wType="漢" pos="名詞-普通名詞-一般" pron="キンガク">
    金額
  </SUW>
</LUW>
<LUW SL="v" l_lemma="の" l_lForm="ノ" l_wType="和" l_pos="助詞-格助詞" >
  <SUW lemma="の" lForm="ノ" wType="和" pos="助詞-格助詞" pron="ノ">
    の
```

```
    </SUW>
  </LUW>
  <LUW B="B" SL="v" l_lemma="動き" l_lForm="ウゴキ" l_wType="和" l_pos="名詞-普通名詞-一般" >
    <SUW lemma="動き" lForm="ウゴキ" wType="和" pos="名詞-普通名詞-一般" pron="ウゴキ">
      動き
    </SUW>
  </LUW>
```

長単位は LUW タグ、短単位は SUW タグで表現され、形態論情報はその属性値として与えられている。LUW 要素は、ひとつ以上の SUW 要素を子要素としてもつ。

6.5.1 短単位タグ（SUW）の属性

埋め込まれた短単位タグ（SUW）には表 6-9 の属性が付与されている。※印の属性は、出力する必要がない場合には、値だけでなく属性自体の出力を行っていない。

表 6-9: 短単位タグ（SUW）の属性

属性名	備考
start	原文文字列のサンプル頭からのオフセット値（10 きざみ）
end	
orderID	連番（TSV の連番と互換）
lemma	語彙素
lForm	語彙素読み
subLemma	語彙素細分類　※区別がある場合のみ出力
wType	語種
pos	品詞
cType	活用型　※活用語のみ出力
cForm	活用形　※活用語のみ出力
formBase	語形
usage	用法　※区別がある場合のみ出力
orthBase	書字形　※活用語のみ出力
originalText	原文文字列　※要素となるテキスト（＝書字形出現形）と異なる場合のみ出力
kanaToken	仮名形出現形　※語形と異なる場合のみ出力
pronToken	出現発音形

なお、TSV における書字形出現形は、SUW タグが囲んでいるテキストに相当する。

仮名形出現形は、テキストに対する読みがな（あるいは IME で入力する場合のカナ文字列）に相当するものである。

6.5.2 長単位タグ（LUW）の属性

埋め込まれた長単位タグ（LUW）には表 6-10 の属性が付与されている。※印の属性は、出力する必要がない場合には、値だけでなく属性自体の出力を行っていない。

また、TSV における「長短一致」など、M-XML の構造や、子要素となる短単位のタグから容易に取得可能な情報は属性としては付与していない。

表 6-10: 長単位タグ（LUW）の属性

属性名	備考
B	文・文節境界　　文節境界=B、文境界=S
SL	サンプル長　　固定長=f、可変長=v
l_lemma	語彙素
l_lForm	語彙素読み
l_wType	語種
l_pos	品詞
l_cType	活用型　※活用語のみ出力
l_cForm	活用形　※活用語のみ出力
l_formBase	語形
l_orthBase	書字形　※活用語のみ出力

参考文献

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),301-332.

小澤俊介・内元清貴・伝康晴（2014）「BCCWJに基づく長単位解析ツール Comainu」,『言語処理学会 第20回年次大会発表論文集』,582-585.

山田篤（2007）「数字列への読み付与―NumTrans と ChaOne―」,『特定領域「日本語コーパス」平成19年度全体会議予稿集』,85-90.

山田篤・小磯花絵（2008）『NumTrans マニュアル』, The UniDic Consortium.

第 7 章　書誌情報データベース

丸山 岳彦　中村 壮範

7.1　均衡コーパスにおける書誌情報の役割

一般に、均衡コーパスとは、さまざまなメディアやジャンルから抽出されたサンプルの集合体と見なすことができる。ある均衡コーパスがどのようなメディアやジャンルのサンプルを含むかは、そのコーパスがどのような設計に基づいているかに依存するが、どのような設計であっても、そこに含まれている各サンプルの出自が明示されていることが望ましい。均衡コーパスを検索した結果を分析したり解釈したりする際、その結果が幅広いメディアを通して一般的に観察される現象なのか、あるいは（例えば）「雑誌」に特有な現象なのか、といった違いを捉えるためには、各サンプルの出自を表す「書誌情報」が必要不可欠である。

BCCWJ の構築過程においては、サンプリングの作業と並行して、各サンプルの出自を示す「書誌情報データベース」を整備してきた。BCCWJ の利用者は、この書誌情報データベースを参照することにより、BCCWJ を構成するすべてのサンプルの出自と属性を知ることができる。厳密な手順で取得された大量のサンプルを、その書誌情報と関連づけて利用することにより、コーパスの分析結果が現代日本語書き言葉のどの位相に位置づけられるものであるかを明確にすることができるわけである。このような利点は、例えば Web をコーパスとして用いる方法論では得ることのできないものであり、均衡コーパスとしての BCCWJ が持つ意義を最大限に特徴づけるものであると言える。

7.2　書誌情報データベースの構成

BCCWJ-DVD 版で提供される書誌情報データベースは、以下のデータ群から構成される。

- 書誌情報データ（Bibliography.txt）： サンプルを取得した原本に関する情報。
- サンプル情報データ（Sample.txt）： サンプルのID や取得状況に関する情報。
- 人名録データ（Directory.txt）： サンプルの著者や著作権者などに関する情報。
- 記事情報データ（Article.txt）：記事に含まれる文章の著者および初出に関する情報。

以下、各データの構成について概略を示す。詳細は、以下の文献を参照。

丸山岳彦・山崎誠・柏野和佳子・佐野大樹・秋元祐哉・稲益佐知子・田中弥生・大矢内夢子（2011）『『現代日本語書き言葉均衡コーパス』におけるサンプリングの原理と運用』, 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-D-10-01）, 特定領域研究「日本語コーパス」データ班.

7.3 「書誌情報データ」（Bibliography.txt）

7.3.1 「書誌情報データ」の概要

書誌情報データ（Bibliography.txt）では、サンプルが抽出された出典元（原本）に関する書誌情報が、表 7-1 に示す 15 列によって表現されている。

表 7-1: 「書誌情報データ」の構成

1	**書誌** ID（Bib_ID）	サンプルを抽出した原本に対して付された ID
2	**タイトル**（Title）	原本のタイトル
3	**副題**（Subtitle）	原本の副題（サブタイトル）
4	**巻号**（Number）	原本の巻号
5	**責任表示**（Bib_author）	原本の責任表示（著者、編者、監修者など）
6	**出版者**（Publisher）	原本の出版者（出版社）
7	**出版年**（Year）	原本の出版年
8	ISBN（ISBN）	原本に付された ISBN（国際標準図書番号）
9	**判型**（Size）	原本のサイズ
10	**ページ数**（Pages）	原本のページ数
11	**ジャンル**(1)（Genre_1）	原本のジャンルに関する情報(1)
12	**ジャンル**(2)（Genre_2）	原本のジャンルに関する情報(2)
13	**ジャンル**(3)（Genre_3）	原本のジャンルに関する情報(3)
14	**ジャンル**(4)（Genre_4）	原本のジャンルに関する情報(4)
15	**責任表示** ID （Bib_author_ID）	原本の責任表示に対応する ID

書誌情報データの例を、表 7-2 に示す。実際には 15 列のタブ区切りテキストだが、ここでは折り返して表示している。「−」が表示されている列は、そのレジスターには情報が付与されないことを示す。

表 7-2: 「書誌情報データ」の例

Register	Bib_ID	Title	Subtitle	Number	Bib_author	Publisher
書籍	BK_20002488	龍臥亭事件	長編推理小説	上	島田荘司|著	光文社
雑誌	PM_00020404	ASAHI パソコン	-	2004 年 2 月 15 日号（通巻 353 号）	-	朝日新聞社
新聞	PN_01030302	朝日新聞	朝刊	2003/3/2	-	朝日新聞社
白書	WR_00000003	わが外交の近況	昭和 51 年版（上）	-	外務省	大蔵省印刷局
教科書	TB_01000009	国語　五上　銀河	-	-	宮地裕|ほか著	光村図書出版
広報紙	PR_14212017	広報あつぎ	-	2008 年 17 号	-	神奈川県厚木市
Yahoo! 知恵袋	YC_00297502	Yahoo!知恵袋	-	-	-	Yahoo!
Yahoo! ブログ	YB_00002691	Yahoo!ブログ	-	-	-	Yahoo!
韻文	VE_00010060	増補版現代短歌全集	紫木蓮まで・風舌	第 17 巻(昭和 55 年～昭和 63 年)	阿木津英|著	筑摩書房
法律	LA_S63HO108	消費税法	-	昭和六十三年十二月三十日法律第百八号	-	-
国会会議録	MD_02010001	国会会議録	-	第 154 回国会	-	-

（表 7-2 続き）

Register	Year	ISBN	Size	Pages	Genre_1	Genre_2	Genre_3	Genre_4	Bib_author_ID
書籍	1999	4334728898	16cm	577	9 文学	913	0193	-	00122924
雑誌	2003	-	A4 変型判	128	工業	電気機／電子	コンピュータ／情報処理	月２回刊	-
新聞	2003	-	ブランケット判	37	全国紙	-	-	-	-
白書	1976	-	-	-	外交	-	-	-	-
教科書	2006	-	-	-	国語	小	5	-	0045734
広報紙	2008	-	-	-	関東地方	神奈川県	-	-	-
Yahoo! 知恵袋	2005	-	-	-	子育てと学校	子育て, 出産	子育ての悩み	-	-
Yahoo! ブログ	2008	-	-	-	家庭と住まい	住まい	ガーデニング	-	-
韻文	2002	4480138374	23cm	500	短歌	-	-	-	00110019
法律	1988	-	-	-	23_国税	-	-	-	-
国会会議録	2002	-	-	-	衆議院	常任委員会	環境委員会	-	-

7.3.2 書誌 ID

書誌 ID（Bib_ID）列は、サンプルを取得した原本に対して一意に付された ID を表す。

「書籍」レジスターの書誌 ID

例：BK_20208020　→『うたかたの月』

1・2 桁目 「BK」「書籍（Book)」であることを表す。

3 桁目 「_」区切り記号。

4～11 桁目 原本に付された一意の ID。

「雑誌」レジスターの書誌 ID

例：PM_00030103　→『アサヒカメラ』、2001 年 3 号

1・2 桁目 「PM」「雑誌（Magazine)」であることを表す。

3 桁目 「_」区切り記号。

4～7 桁目 同一タイトルの雑誌に付された一意の ID。

※ 「0003」は『アサヒカメラ』に付与されたID。

8～9 桁目 発行年（2001 年から 2005 年の下 2 桁、01～05)。

10～11 桁目 その発行年における号数。

「新聞」レジスターの書誌 ID

例： PN_01010202　→ 朝日新聞・朝刊（0101)、2 月 2 日発行

1・2 桁目 「PN」「新聞（Newspaper)」であることを表す。

3 桁目 「_」区切り記号。

4～5 桁目 新聞タイトル・朝夕刊の別を表す ID。(7.A.3 を参照)

6～7 桁目 発行年（2001 年から 2005 年の下 2 桁、01～05)。

8～11 桁目 発行日（1 月 1 日から 12 月 31 日、0101～1231)。

「白書」レジスターの書誌 ID

例： WR_00000001　→『エネルギー白書』2004 年版

1・2 桁目 「WR」「白書」であることを表す。

3 桁目 「_」区切り記号。

4～11 桁目 原本に付された一意の ID。

「教科書」レジスターの書誌 ID

例： TB_01000001　→『こくご　一上　かざぐるま』

1・2 桁目 「TB」「教科書（TextBook)」であることを表す。

3 桁目 「_」区切り記号。

4 桁目 教科。

「0」= 国語　「3」= 社会　「6」= 芸術　「9」= 生活

「1」= 数学　「4」= 外国語　「7」= 保健体育

「2」= 理科　「5」= 技術家庭　「8」= 情報

5 桁目 学校。

「1」= 小学校 「2」= 中学校 「3」= 高校

6〜11 桁目 教科・学校ごとに分類された教科書の通し番号。

「広報紙」レジスターの書誌 ID

例： PR_14212017 → 『広報あつぎ』2008 年 17 号

1・2 桁目 「PR」「広報紙（Public Relations)」であることを表す。

3 桁目 「_」区切り記号。

4〜8 桁目 自治体に付された一意の ID。（7.A.6 を参照）

9〜11 桁目 その自治体における号数。

「Yahoo!知恵袋」レジスターの書誌 ID

例：YC_00297787 → 小カテゴリ「政治、社会問題」

1・2 桁目 「YC」「Yahoo!知恵袋（Yahoo! Chiebukuro)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 小カテゴリごとに付された一意の ID。（7.A.7 を参照）

「Yahoo!ブログ」レジスターの書誌 ID

例：YB_00000075 → 小カテゴリ「インテリア」

1・2 桁目 「YB」「Yahoo!ブログ（Yahoo! Blog)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 小カテゴリごとに付された一意の ID。（7.A.8 を参照）

「韻文」レジスターの書誌 ID

例：VE_89028672 → 『稲垣足穂詩集』

1・2 桁目 「VE」「韻文（Verse)」であることを表す。

3 桁目 「_」区切り記号。

4〜11 桁目 原本に付された一意の ID。

「法律」レジスターの書誌 ID

例：LA_S54HO004 →「民事執行法」（昭和五十四年三月三十日法律第四号）

1・2 桁目 「LA」「法律（Law)」であることを表す。

3 桁目 「_」区切り記号。

4〜6 桁目 法律の公布年（S54 → 昭和 54 年)。

7〜8 桁目 「法律（HO)」であることを表す。

4〜11 桁目 その年における法令番号。

「国会会議録」レジスターの書誌 ID

例：MD_79050005 → 国会会議録（1979 年第 91 回国会、参議院、常任委員会、外務委員会）

1・2 桁目 「MD」「国会会議録（Minutes of the Diet)」であることを表す。

3 桁目 「_」区切り記号。

4～5 桁目 会議の開催年。

6～7 桁目 開催院および会議種別。

「01」= 衆議院・常任委員会　「05」= 参議院・常任委員会

「02」= 衆議院・特別委員会　「06」= 参議院・特別委員会

「03」= 衆議院・本会議　「07」= 参議院・本会議

「04」= 衆議院・その他　「08」= 参議院・その他

8～11 桁目 会議種別ごとの会議に付された一意の ID。

7.3.3 タイトル

タイトル（Title）列は、原本のタイトルを表す。

例 「ファンの心をときめかせた世界の映画ベストセレクション」（書籍）

「塩狩峠; 道ありき」（書籍）

「週刊朝日」（雑誌）

「北海道新聞」（新聞）

「情報通信白書」（白書）

「こくご　一上　かざぐるま」（教科書）

「広報あげお」（広報紙）

「Yahoo!知恵袋」（Yahoo!知恵袋）

「Yahoo!ブログ」（Yahoo!ブログ）

「谷川俊太郎詩集」（韻文）

「民事保全法」（法律）

「国会会議録」（国会会議録）

7.3.4 副題

副題（Subtitle）列は、原本の副題・サブタイトルを表す。

例 「伝説の呼び屋・永島達司の生涯」（書籍）

「朝刊」（新聞）

「平成 4 年版」（白書）

「サラダ記念日」（韻文）

7.3.5 巻号

巻号（Number）列は、原本の巻号・巻次に関する情報を表す。

例 「第 6 巻」（書籍）

「3（神の星編）」（書籍）

「2002 年 4 月 15 日号（第 15 巻第 16 号、通巻 750 号）」（雑誌）

「サンデー毎日臨時増刊（第 80 巻第 49 号、通巻 4467 号）」（雑誌）

「2001/10/24」（新聞）
「2008 年 12 号」（広報紙）
「第 17 巻（55 年～昭和 63 年）」（韻文）
「平成元年六月二十八日法律第五十八号」（法律）
「第 154 回国会」（国会会議録）

7.3.6 責任表示

責任表示（Bib_author）列は、原本の責任表示（著者、編者、監修者など）の情報を表す。

例　「司馬遼太郎 | 著」（書籍）
「七田眞、七田厚 | 著」（書籍）
「高橋貞巳 | 監修；三菱総合研究所 | 著」（書籍）
「カフカ | 著；池内紀 | 訳」（書籍）
「ロナルド・A.　モース | 編著；日下公人 | 監修；時事通信社外信部 | ほか訳」（書籍）
「経済産業省；厚生労働省；文部科学省」（白書）
「宮地裕 | ほか著」（教科書）

7.3.7 出版者

出版者（Publisher）列は、原本の出版者（出版社）を表す。

例　「岩波書店」（書籍）
「日本図書刊行会；近代文芸社（発売）」（書籍）
「マガジンハウス」（雑誌）
「株式会社朝日新聞社」（新聞）
「大蔵省印刷局」（白書）
「光村図書出版株式会社」（教科書）
「北海道札幌市東区」（広報紙）
「Yahoo!」（Yahoo!知恵袋、Yahoo!ブログ）
「筑摩書房」（韻文）

7.3.8 出版年

出版年（Year）列は、4 桁の数字で、原本が出版された年を表す。

※　「Yahoo!知恵袋」と「Yahoo!ブログ」の場合、それぞれ「2005」「2008」の一通りとなる。質問や記事が実際に投稿された日時は、サンプル情報データ（Sample.txt）の「タイムスタンプ（Timestamp）」列を参照のこと。

※　「法律」の場合、法律が公布された年を表す。

※　「国会会議録」の場合、会議が開催された年を表す。

7.3.9 ISBN

ISBN（ISBN）列は、原本に付された ISBN（国際標準図書番号）を表す（10 桁）。

7.3.10 判型

判型（Size）列は、原本の大きさを表す。

7.3.11 ページ数

ページ数（Pages）列は、原本の総ページ数を表す。

7.3.12 ジャンル(1)～(4)

ジャンル(1)～(4)（Genre_1～Genre_4）列は、原本のジャンルに関連した情報を表す。レジスターごとに取るジャンル情報の種類を、表 7-3 に示す。

表 7-3: ジャンル情報の種類

レジスター	ジャンル(1)	ジャンル(2)	ジャンル(3)	ジャンル(4)
書籍	NDC(1 桁)＋分類名	NDC(3 桁)	C コード	
雑誌	大ジャンル名	中ジャンル名	小ジャンル名	刊行形態
新聞	配達エリア			
白書	ジャンル名			
教科書	教科名	学校種	学年	
広報紙	地域	都道府県名		
Yahoo!**知恵袋**	大カテゴリ名	中カテゴリ名	小カテゴリ名	
Yahoo!**ブログ**	大カテゴリ名	中カテゴリ名	小カテゴリ名	
韻文	韻文種別			
法律	ジャンル名			
国会会議録	開催院	会議種別	委員会名	

※ ジャンル情報の詳細については、付録 7-A を参照。

7.3.13 責任表示 ID

責任表示 ID（Bib_author_ID）列は、責任表示（Bib_author）列に記載されている人名・組織名などに対して付された ID を表す。記載されている ID は、人名録データ(Directory.txt)の「人名 ID （Directory_ID)」列に記載された ID に対応している。

例 「00685074」（書籍）
「00254659；00184422」（書籍）
「00113880；00166885；00124738」（教科書）
「00037561」（韻文）

7.4 「サンプル情報データ」（Sample.txt）

7.4.1 「サンプル情報データ」の概要

サンプル情報データ（Sample.txt）では、BCCWJ に収録された各サンプルの ID や抽出状況に関する情報が、表 7-4 に示す 5 列によって表現されている。

表 7-4: 「サンプル情報データ」の構成

1	**サンプル ID**（Sample_ID）サンプルに対して一意に付された ID
2	**書誌 ID**（Bib_ID）サンプルを抽出した原本に対して付された ID
3	**サンプル抽出基準点ページ**（Sampling_page）サンプル抽出基準点を取得したページ
4	**サンプル抽出基準点座標**（Sampling_point）サンプル抽出基準点を取得した交点
5	**投稿日時**（Timestamp）Yahoo!知恵袋の質問、Yahoo!ブログの記事の投稿日時

サンプル情報データの例を、表 7-5 に示す。

表 7-5: 「サンプル情報データ」の例

レジスター	Sample_ID	Bib_ID	Sampling_page	Sampling_point	Timestamp
出版・書籍	PB10_00047	BK_20205918	163	5D	-
雑誌	PM11_00053	PM_10550109	76	9F	-
新聞	PN1a_00013	PN_01010225	4	6C	-
図書館・書籍	LBa1_00004	BK_86049602	230	2H	-
白書	OW6X_00009	WR_00000066	285	4C	-
教科書	OT01_00008	TB_01000002	31	8A	-
広報紙	OP00_00001	PR_01103001	-	-	-
ベストセラー	OB0X_00001	BK_75079014	358	4D	-
Yahoo!知恵袋	OC01_00001	YC_00297514	-	-	2004/4/29 18:35
Yahoo!ブログ	OY01_00005	YB_00010571	-	-	2008/6/24 21:39
韻文	OV0X_00001	VE_00010001	-	-	-
法律	OL3X_00072	LA_H01HO058	-	-	-
国会会議録	OM11_00001	MD_80010001	-	-	-

7.4.2 サンプル ID

サンプル ID（Sample_ID）列は、各サンプルに対して一意に付された ID を表す。

- 出版サブコーパス「書籍」レジスターのサンプル ID

  例： PB10_00001

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「B」書籍（Book）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

  「2」＝ 2002 年　「4」＝ 2004 年

  4 桁目 「0～9,n」当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。

  「0」＝総記　　「4」＝自然科学　　「8」＝言語

  「1」＝哲学　　「5」＝技術・工学　「9」＝文学

  「2」＝歴史　　「6」＝産業　　「n」＝分類なし

  「3」＝社会科学　「7」＝芸術・美術

  5 桁目 「_」区切り記号。

  6～10 桁目 各出版年・各 NDC におけるサンプルの取得順位を表す。

- 出版サブコーパス「雑誌」レジスターのサンプル ID

  例： PM11_00002

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「M」雑誌（Magazine）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

  「2」＝ 2002 年　「4」＝ 2004 年

  4 桁目 「1～6」当該雑誌に付されたジャンルを表す。

  「1」＝総合　　「4」＝産業

  「2」＝教育・学芸　　「5」＝工業

  「3」＝政治・経済・商業　「6」＝厚生・医療

  5 桁目 「_」区切り記号。

  6～10 桁目 各雑誌タイトル・各出版年におけるサンプルの取得順位を表す。

- 出版サブコーパス「新聞」レジスターのサンプル ID

  例： PN1a_00001

  1 桁目 「P」出版サブコーパス（Publication）に所属することを表す。

  2 桁目 「N」新聞（Newspaper）のサンプルであることを表す。

  3 桁目 「1～5」出版年を表す。

  「1」＝ 2001 年　「3」＝ 2003 年　「5」＝ 2005 年

「2」＝ 2002 年　「4」＝ 2004 年

4 桁目 「a～o」新聞タイトルを表す。

「a」＝朝日新聞　「f」＝中日新聞　「k」＝神戸新聞
「b」＝毎日新聞　「g」＝西日本新聞　「l」＝中国新聞
「c」＝読売新聞　「h」＝河北新報　「m」＝高知新聞
「d」＝産経新聞　「i」＝新潟日報　「o」＝琉球新報
「e」＝北海道新聞　「j」＝京都新聞

5 桁目 「_」区切り記号。

6～10 桁目 各新聞タイトル・各出版年におけるサンプルの取得順位を表す。

- 図書館サブコーパス「書籍」レジスターのサンプル ID

例： LBa0_00002

1 桁目 「L」図書館サブコーパス（Library）に所属することを表す。

2 桁目 「B」書籍（Book）のサンプルであることを表す。

3 桁目 「a～t」出版年を表す。

「a」＝ 1986 年　「h」＝ 1993 年　「o」＝ 2000 年
「b」＝ 1987 年　「i」＝ 1994 年　「p」＝ 2001 年
「c」＝ 1988 年　「j」＝ 1995 年　「q」＝ 2002 年
「d」＝ 1989 年　「k」＝ 1996 年　「r」＝ 2003 年
「e」＝ 1990 年　「l」＝ 1997 年　「s」＝ 2004 年
「f」＝ 1991 年　「m」＝ 1998 年　「t」＝ 2005 年
「g」＝ 1992 年　「n」＝ 1999 年

4 桁目 「0～9,n」当該書籍に付された NDC（日本十進分類法）の第 1 次区分を表す。

「0」＝総記　「4」＝自然科学　「8」＝言語
「1」＝哲学　「5」＝技術・工学　「9」＝文学
「2」＝歴史　「6」＝産業　「n」＝分類なし
「3」＝社会科学　「7」＝芸術・美術

5 桁目 「_」区切り記号。

6～10 桁目 各出版年・各 NDC におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「白書」レジスターのサンプル ID

例：OW1X_00000

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「W」白書のサンプルであることを表す。

3 桁目 「1～6」出版時期を表す。

「1」＝第 1 期（1976～1980 年）　「4」＝第 4 期（1991～1995 年）
「2」＝第 2 期（1981～1985 年）　「5」＝第 5 期（1996～2000 年）
「3」＝第 3 期（1986～1990 年）　「6」＝第 6 期（2001～2005 年）

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 各出版時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「教科書」レジスターのサンプル ID

  例：OT01_00002

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「T」教科書（TextBook）のサンプルであることを表す。

  3 桁目 「0～9」教科を表す。

  「0」=国語 「3」=社会 「6」=芸術 「9」=生活

  「1」=数学 「4」=外国語 「7」=保健体育

  「2」=理科 「5」=技術家庭 「8」=情報

  4 桁目 「1～3」学校を表す。

  「1」= 小学校 「2」= 中学校 「3」= 高校

  5 桁目 「_」区切り記号。

  6～10 桁目 各教科・学校におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「広報紙」レジスターのサンプル ID

  例： OP00_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「P」広報紙（Public Relations）のサンプルであることを表す。

  3・4 桁目 「00～99」対象となった 100 自治体の通し番号を表す。

  5 桁目 「_」区切り記号。

  6～10 桁目 各自治体から取得したサンプルの取得順位を表す。

- 特定目的サブコーパス「ベストセラー」レジスターのサンプル ID

  例： OB0X_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「B」ベストセラー（Best-seller）のサンプルであることを表す。

  3 桁目 「0～6」出版時期を表す。

  「0」＝第 0 期（1975 年以前） 「4」＝第 4 期（1991～1995 年）

  「1」＝第 1 期（1976～1980 年）「5」＝第 5 期（1996～2000 年）

  「2」＝第 2 期（1981～1985 年）「6」＝第 6 期（2001～2005 年）

  「3」＝第 3 期（1986～1990 年）

  4 桁目 「X」ダミー記号。

  5 桁目 「_」区切り記号。

  6～10 桁目 各出版時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「Yahoo!知恵袋」レジスターのサンプル ID

  例： OC01_00001

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「C」Yahoo!知恵袋（Chiebukuro）のサンプルであることを表す。

  3・4 桁目 「01～15」質問が投稿された大カテゴリ ID を表す。

  「01」＝ 「エンターテインメントと趣味」

  「02」＝ 「インターネット、PC と家電」

  「03」＝ 「ビジネス、経済とお金」

  「04」＝ 「職業とキャリア」

  「05」＝ 「ニュース、政治、国際情勢」

  「06」＝ 「スポーツ、アウトドア、車」

  「08」＝ 「暮らしと生活ガイド」

  「09」＝ 「健康、美容とファッション」

  「10」＝ 「子育てと学校」

  「11」＝ 「マナー、冠婚葬祭」

  「12」＝ 「教養と学問、サイエンス」

  「13」＝ 「地域、旅行、お出かけ」

  「14」＝ 「Yahoo! JAPAN」

  「15」＝ 「その他」

  5 桁目 「_」区切り記号。

  6～10 桁目 各大カテゴリにおけるサンプルの取得順位を表す。

- 特定目的サブコーパス「Yahoo!ブログ」レジスターのサンプル ID

  例： OY01_00005

  1 桁目 「O」特定目的サブコーパスに所属することを表す。

  2 桁目 「Y」Yahoo!ブログ（Blog）のサンプルであることを表す。

  3・4 桁目 「01～15」記事が投稿された大カテゴリ ID を表す。

  「01」＝ 「ビジネスと経済」

  「02」＝ 「コンピュータとインターネット」

  「03」＝ 「生活と文化」

  「04」＝ 「エンターテインメント」

  「05」＝ 「家庭と住まい」

  「06」＝ 「政治」

  「07」＝ 「健康と医学」

  「08」＝ 「学校と教育」

  「09」＝ 「科学」

  「10」＝ 「出会い」

「11」＝「地域」
「12」＝「特集」
「13」＝「芸術と人文」
「14」＝「Yahoo!サービス」
「15」＝「趣味とスポーツ」

5 桁目 「_」区切り記号。

6～10 桁目 各大カテゴリにおけるサンプルの取得順位を表す。

- 特定目的サブコーパス「韻文」レジスターのサンプル ID

例：OV0X_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「V」韻文（Verse）のサンプルであることを表す。

3 桁目 「0～2」韻文の種類を表す。

「0」＝短歌 「1」＝俳句 「2」＝詩

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 サンプルの取得順位を表す。

- 特定目的サブコーパス「法律」レジスターのサンプル ID

例： OL1X_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「L」法律（Law）のサンプルであることを表す。

3 桁目 「1～6」法律の公布時期を表す。

「1」＝第 1 期（1976～1980 年） 「2」＝第 2 期（1981～1985 年）
「3」＝第 3 期（1986～1990 年） 「4」＝第 4 期（1991～1995 年）
「5」＝第 5 期（1996～2000 年） 「6」＝第 6 期（2001～2005 年）

4 桁目 「X」ダミー記号。

5 桁目 「_」区切り記号。

6～10 桁目 各公布時期におけるサンプルの取得順位を表す。

- 特定目的サブコーパス「国会会議録」レジスターのサンプル ID

例： OM11_00001

1 桁目 「O」特定目的サブコーパスに所属することを表す。

2 桁目 「M」国会会議録（Minutes of the Diet）のサンプルであることを表す。

3 桁目 「1～6」会議の開催時期を表す。

「1」＝第 1 期（1976～1980 年） 「4」＝第 4 期（1991～1995 年）
「2」＝第 2 期（1981～1985 年） 「5」＝第 5 期（1996～2000 年）
「3」＝第 3 期（1986～1990 年） 「6」＝第 6 期（2001～2005 年）

4 桁目 「1～8」会議の開催院・会議種別を表す。

「1」＝衆議院・常任委員会　「5」＝参議院・常任委員会
「2」＝衆議院・特別委員会　「6」＝参議院・特別委員会
「3」＝衆議院・本会議　「7」＝参議院・本会議
「4」＝衆議院・その他　「8」＝参議院・その他

5 桁目 「_」区切り記号。

6～10 桁目 各開催時期、開催院・会議種別におけるサンプルの取得順位を表す。

7.4.3 書誌 ID

書誌 ID（Bib_ID）列は、サンプルを取得した原本に対して一意に付された ID を表す。記載されている ID は、書誌情報データ（Bibliography.txt）の「書誌 ID（Bib_ID）」列に記載された ID に対応している（7.3.2 節参照）。

7.4.4 サンプル抽出基準点ページ

サンプル抽出基準点ページ（Sampling_page）列は、「サンプル抽出基準点」（7.4.5 参照）を含むページ番号を表す。

7.4.5 サンプル抽出基準点座標

サンプル抽出基準点座標（Sampling_point）列は、「サンプル抽出基準点」を同定する際、サンプル抽出基準点ページ内でランダムに指定されたある 1 点（交点）の座標を表す。

※ 横軸に0～9、縦軸にA～Jという目盛りを配置した10×10のマス目を準備し、それを印刷した透明なシートを実際のページに当て、ランダムに指定された交点（「3E」など）に最も近接している文字を「サンプル抽出基準点」として指定した。このサンプル抽出基準点をもとに、サンプルを取得した。

7.4.6 投稿日時

投稿日時（Timestamp）列は、「Yahoo!知恵袋」の質問、および「Yahoo!ブログ」の記事が投稿された日時を表す。

7.5 「人名録データ」（Directory.txt）

7.5.1 「人名録データ」の概要

人名録データ（Directory.txt）では、書誌データ（Bibliography.txt）の「責任表示（Bib_author)」列に記載されている人名や組織名（著者、編者、監修者など）や、各サンプルに含まれる記事を実際に執筆した著者名などの情報が、表 7-6 に示す 4 列によって表現されている。

表 7-6: 「人名録データ」の構成

1	**人名 ID**（Directory_ID）	人物や組織に対して一意に付された ID
2	**人名**（Name）	人物の氏名、または組織名
3	**性別**（Sex）	性別
4	**生年代**（BirthYear）	生年（10 年単位）

人名録データの例を、表 7-7 に示す。

表 7-7: 「人名録データ」の例

Directory_ID	Name	Sex	BirthYear
634	会田　雄次	男	1910
98948	アントニオ猪木	男	1940
153494	群　ようこ	女	1950
840303	厚生労働省労働基準局		
258003	講談社		
2502212	NHK「プロジェクト X」制作班		

7.5.2 人名 ID

人名 ID（Directory_ID）列は、人物の氏名または組織名に対して付された一意の ID を表す。

7.5.3 人名

人名（Name）列は、人物の氏名や組織名などを表す。

7.5.4 性別

性別（Sex）列は、人物の性別を表す。なお、組織の場合には記載していない。

7.5.5 生年代

生年代（BirthYear）列は、人物の生年を西暦の 10 年単位でまとめた年を表す。なお、性別・生年代については、原則として本人からの回答を記載しているが、国立国会図書館の典拠データなどから情報を補足しているものもある。また、組織の場合には記載していない。

7.6 記事情報データ（Article.txt）

7.6.1 「記事情報データ」の概要

記事情報データ（Article.txt）では、各サンプルに含まれる「記事」を対象として、「実著者」および「初出」に関する情報が、表 7-8 に示す 6 列によって表現されている。

表 7-8: 「記事情報データ」の構成

1	**サンプル** ID（Sample_ID）	各サンプルに対して一意に付された ID
2	**記事** ID（Article_ID）	各記事に対して一意に付された ID
3	**人名** ID（Directory_ID）	各記事を実際に執筆した著者に対して一意に付された ID
4	**役割**（Role）	著者の役割（実著者、原著者、翻訳者の別）
5	**初出情報**（First_appearance）	各記事の初出に関する情報
6	**初刊情報**（First_published）	各記事の初刊に関する情報

記事情報データの例を、表 7-9 に示す。

表 7-9: 「記事情報データ」の例

Sample_ID	Article_ID	Directory_ID	Role	First_ appearance	First_ published
LBa0_00002	LBa0_00002_V001	59986	実著者	1984	1986
LBq1_00026	LBq1_00026_F003	262756	実著者	2000-2001	2002
LBa1_00006	LBa1_00006_V001	459606	原著者	1986	
LBa1_00006	LBa1_00006_V001	108831	翻訳者	1986	
PB12_00059	PB12_00059_V001	189710	実著者	n.d.-n.d.	2001
PM11_00289	PM11_00289_F002	0	実著者	2001	
PN1a_00004	PN1a_00004_V003	256908	実著者	2001	

なお、記事情報データは、「書籍」「雑誌」「新聞」に対してのみ提供される。

7.6.2 サンプル ID

サンプル ID（Sample_ID）列は、サンプルに対して一意に付された ID を表す。記載されている ID は、サンプル情報データ（Sample.txt）の「サンプル ID（Sample_ID）」列に記載された ID に対応している。7.4.2 節を参照。

7.6.3 記事 ID

記事 ID（Article_ID）列は、サンプルに含まれる「記事」に対して一意に付された ID を表す。

例　PB15_00023_F001

LB29_00129_V001
PM11_00118_F002
PN1a_00013_F004

「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」を指す。ひとつのサンプル（可変長サンプル、固定長サンプルとも）は、ひとつの記事によって構成されている場合もあれば、複数の記事によって構成されている場合もある。記事 ID は、それが所属するサンプル ID の直後に「V001」「F002」などを続けて表される。「V001」は、そのサンプルに含まれる可変長（Variable_Length）サンプルの 1 番目の記事、「F002」は、そのサンプルに含まれる固定長（Fixed_Length）サンプルの 2 番目の記事であることを、それぞれ表す。

7.6.4 人名 ID

記事情報データにおける人名 ID（Directory_ID）列は、記事を実際に執筆した人物（著者）に対して付された ID を表す。記載されている ID は、人名録データ（Directory.txt）の「人名 ID（Directory_ID）」列に記載された ID に対応している。

各記事を実際に執筆した人物を「実著者」と判定し、その人名および人名 ID を記録した。翻訳書については、実著者に替えて、「原著者」と「翻訳者」の人名と人名 ID の組を記録した。実著者、原著者、翻訳者は、各サンプルの印刷紙面や、原本の目次、奥付に表示されている人名、著作権処理の過程で判明した実著者の情報などをもとに判定した。新聞については、実著者の記名が新聞記者と思われる場合は、その人名に替えて「朝日新聞社」などの新聞社名を記録した。なお、当該の文章を執筆した人名が確定できない場合は「実著者不明」として「0」という人名 ID を与えた。

7.6.5 役割

役割（Role）列は、「人名 ID（Directory_ID）」列に記載された ID に対応する人物の役割を表す。

例　「実著者」
　　「原著者」
　　「翻訳者」

7.6.6 初出情報

初出情報（First_appearance）列は、当該の記事に含まれる文章が、雑誌や新聞などで初めて発表・出版された年を表す。

7.6.7 初刊情報

初刊情報（First_published）列は、当該の記事に含まれる文章が、初めて書籍として刊行された年を表す。

なお、「初出情報」「初刊情報」は、次のような問題意識および方法によって、情報を取得した。ある書籍に含まれる文章は、その書籍の発行時において初めて世に発表されたものと、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、単行本が改版して出版される場合、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ている例もある。

そこで、取得した文章がそれ以前に出版されたことがあるかどうかについて、可能な限り調査した。原本の奥付や目次の周辺、後書きなどを確認し、初出・初刊に関する情報を取得した。同時に、『文芸雑誌小説初出総覧』（日外アソシエーツ）、『日本近代文学大事典』（講談社）も参照した。さらに、一部については、国立国会図書館（NDL-OPAC）を使って調査を行なった。情報が取得できた場合、その年を初出情報・初刊情報として記録した。なお、初出が確認できなかった場合は、初出情報は空欄とした。また、初刊が確認できなかった場合は、出版年を初刊情報として記録した。

なお、当該の書籍が一連のシリーズとして刊行された場合や、複数年にわたる連載記事として出版されていた場合は、「1965-1971」のように、年号をハイフンでつないで表示した。また、雑誌や新聞などに連載された原稿が書籍になった旨が明記してあるものの、その年が確定できない場合には、「n.d.（no date）」と記録した。

上記で述べた書籍の場合と同様に、雑誌・新聞についても初出情報を調査した。初出情報が確認できなかった場合は、出版年を初出の年とした。また、書籍として刊行された年を表す初刊情報は、雑誌・新聞には付与されていない。

付録 7-A: 書誌情報データ「ジャンル」情報の詳細

7.A.1 「書籍」のジャンル情報の詳細

- ジャンル(1)

「書籍」の「ジャンル(1)」列には、国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 1 次区分（類目）を表す数値と、その分類名が記載されている。

例　「0 総記」、「1 哲学」、「2 歴史」、「3 社会科学」、「4 自然科学」、
「5 技術・工学」、「6 産業」、「7 芸術・美術」、「8 言語」、「9 文学」、
「分類なし[1]」

- ジャンル(2)

「書籍」の「ジャンル(2)」列には、国立国会図書館で付与された「NDC（日本十進分類法）第 9 版」の第 3 次区分（要目）を表す数値が 3 桁で記載されている。詳細については、『日本十進分類法新訂 9 版』（日本図書館協会）などを参照。

- ジャンル(3)

「書籍」の「ジャンル(3)」列には、「C コード（図書分類コード)」が記載されている。「C コード」は日本図書コードの一部で、4 桁の数値で構成される。左から 1 桁目は「販売対象コード」で、対象読者を表す。2 桁目は「発行形態コード」で、発行形態を表す。3・4 桁目は「内容コード」で、書籍の内容を表す。詳細は、『ISBN コード／日本図書コード／書籍 JAN コード利用の手引き』（日本図書コード管理センター）などを参照。

※ 「C コード」の 1 桁目 「販売対象コード」の分類を、以下に示す。
「0」＝ 一般、「1」＝ 教養、「2」＝ 実用、「3」＝ 専門、「4」＝（欠番） 、
「5」＝ 婦人、「6」＝ 学参 I（小中） 、「7」＝ 学参 II（高校)、「8」＝ 児童、「9」＝ 雑誌扱い

※ 「C コード」の 2 桁目 「発行形態コード」の分類を、以下に示す。
「0」＝ 単行本、「1」＝ 文庫、「2」＝ 新書、「3」＝ 全集・双書、
「4」＝ ムック・その他、「5」＝ 事・辞典、「6」＝ 図鑑、「7」＝ 絵本、
「8」＝ 磁性媒体など、「9」＝ コミック

※ 「C コード」の 3・4 桁目 「内容コード」の分類を、表 7-10 に示す。

[1] 2005 年 10 月時点において国立国会図書館で NDC が付与されていなかった場合に相当する。

表 7-10: 「C コード」の 3・4 桁目 「内容コード」の分類

「00」= 総記
「01」= 百科事典
「02」= 年鑑・雑誌
「04」= 情報科学
「10」= 哲学
「11」= 心理（学）
「12」= 倫理（学）
「14」= 宗教
「15」= 仏教
「16」= キリスト教
「20」= 歴史総記
「21」= 日本歴史
「22」= 外国歴史
「23」= 伝記・系譜
「25」= 地理
「26」= 旅行
「30」= 社会科学総記
「31」= 政治（国防・軍事含む）
「32」= 法律
「33」= 経済、財政、統計
「34」= 経営
「36」= 社会
「37」= 教育
「39」= 民俗・風習
「40」= 自然科学総記
「41」= 数学
「42」= 物理学
「43」= 化学
「44」= 天文・地学
「45」= 生物学
「47」= 医学・歯学・薬学
「50」= 工学・工学総記
「51」= 土木
「52」= 建築
「53」= 機械
「54」= 電気
「55」= 電子通信
「56」= 海事・兵器
「57」= 採鉱・冶金
「58」= その他の工業
「60」= 産業総記
「61」= 農林業
「62」= 水産業
「63」= 商業
「65」= 交通・通信
「70」= 芸術総記
「71」= 絵画・彫刻
「72」= 写真・工芸
「73」= 音楽・舞踊
「74」= 演劇・映画
「75」= 体育・スポーツ
「76」= 諸芸・娯楽
「77」= 家事
「78」= 生活
「79」= コミックス・劇画
「80」= 語学総記
「81」= 日本語
「82」= 英（米）語
「84」= ドイツ語
「85」= フランス語
「87」= 各国語
「90」= 文学総記
「91」= 日本文学総記
「92」= 日本文学詩歌
「93」= 日本文学小説・物語
「95」= 日本文学評論・随筆・その他
「97」= 外国文学小説
「98」= 外国文学その他

7.A.2 「雑誌」のジャンル情報の詳細

- ジャンル(1)

「雑誌」の「ジャンル(1)」列には、表 7-11 に示す 6 種類の「大ジャンル」の情報が、雑誌タイトルごとに記載されている。

例　「1 総合」、「2 教育・学芸」、「3 政治・経済・商業」、「4 産業」、
「5 工業」、「6 厚生・医療」

- ジャンル(2)

「雑誌」の「ジャンル(2)」列には、表 7-11 に示す 27 種類の「中ジャンル」の情報が、雑誌タイトルごとに記載されている。

- ジャンル(3)

「雑誌」の「ジャンル(3)」列には、表 7-11 に示す 71 種類の「小ジャンル」の情報が、雑誌タイトルごとに記載されている。

- ジャンル(4)

「雑誌」の「ジャンル(4)」列には、雑誌タイトルの「刊行形態」が記載されている。

例　「月刊」「週刊」「月 2 回刊」「隔月刊」「隔週刊」「季刊」「年刊」
「年 2 回刊」「年 4 回刊」「年 5～6 回刊」「年 3 回刊」

なお、大ジャンル、中ジャンル、小ジャンル、および刊行形態の分類は、『雑誌新聞総かたろぐ』（メディア・リサーチ・センター）での記載に基づく。

表 7-11: 雑誌の大ジャンル・中ジャンル・小ジャンルの一覧

大ジャンル	中ジャンル	小ジャンル
総合	総記／マスコミ	総記
		マスコミ（新聞・放送）
		出版・読書・図書館
		出版情報・書評
	一般	一般週刊誌
		総合誌
		女性週刊誌
		婦人誌
		読み物
		東京都／タウン・地域誌
		関東地方／タウン・地域誌
		近畿地方／タウン・地域誌
	家庭／生活	生活情報
		ファッション
		料理・栄養
		住居・インテリア
		育児・家庭教育
	児童	少年
		少女
	娯楽／芸能	ヤング
		テレビ・ラジオ・芸能・映画
	レジャー／趣味	レジャー
		旅行・観光
		趣味の乗り物
		釣り・狩猟
		写真・カメラ
		家庭園芸
		ホビー・クラフト・日曜大工
		模型・無線・コンピュータゲーム
		音楽・オーディオ
		囲碁・将棋
		ペット

表 7-11: 雑誌の大ジャンル・中ジャンル・小ジャンルの一覧（続き）

大ジャンル	中ジャンル	小ジャンル
総合（続き）	スポーツ	スポーツ一般・陸上競技
		アウトドア・海／山
		球技
		ゴルフ
		武道・格闘技
教育・学芸	教育	教育技術
	学習／語学	小・中学生
		高校・大学生
	文学／芸術	文学文芸総合
		大衆文芸
		俳句
		短歌
		芸術・美術
	人文科学	宗教
	社会科学	歴史一般
	自然科学	自然科学一般
		地球宇宙科学
政治・経済・商業	政治／外交	国会行政
		海外情勢外交
	経済／経営	経営／経済
	金融／財政	金融財政
	商業／消費者	広告宣伝・ＰＲ
	国勢／民力	国勢／民力
		所得・物価・消費
産業	農林水産	農業経営
	食料／食品	醸造業
	運輸／通信	海事・海運・港湾
工業	工業一般	公害・環境保全
	建設／土木	建設一般
	機械	機械一般
		自動車・オートバイ・自転車
	電気機／電子	家電・弱電・照明
		エレクトロニクス
		コンピュータ／情報処理
		電波・電気通信
厚生・医療	厚生	福祉
	医学	医学総合
		家庭医学・健康

7.A.3 「新聞」のジャンル情報の詳細

- ジャンル(1)

「新聞」の「ジャンル(1)」列には、その新聞タイトルが配達される範囲の違いによって、以下の分類が記載されている。

例　　「全国紙」「ブロック紙」「地方紙」

なお、「新聞」の書誌 ID の 4～5 桁目で表される ID（01～31）は、新聞の母集団に含まれる 16 タイトル、および朝夕刊の別に対して独自に付与した ID である。各タイトルに対応づけられたジャンル（配達エリア）との対応関係を、表 7-12 に示す（著作権の都合で採録対象から外した 2 タイトルは表示していない)。

表 7-12: 新聞の書誌 ID（4～5 桁目）の内訳

ID	タイトル	朝夕刊	配達エリア	ID	タイトル	朝夕刊	配達エリア
01	朝日新聞	朝刊	全国紙	17	河北新報	朝刊	地方紙
02	朝日新聞	夕刊	全国紙	18	河北新報	夕刊	地方紙
03	毎日新聞	朝刊	全国紙	19	新潟日報	朝刊	地方紙
04	毎日新聞	夕刊	全国紙	20	新潟日報	夕刊	地方紙
05	読売新聞	朝刊	全国紙	21	京都新聞	朝刊	地方紙
06	読売新聞	夕刊	全国紙	22	京都新聞	夕刊	地方紙
09	産経新聞	朝刊	全国紙	23	神戸新聞	朝刊	地方紙
10	産経新聞	夕刊	全国紙	24	神戸新聞	夕刊	地方紙
11	北海道新聞	朝刊	ブロック紙	25	中国新聞	朝刊	地方紙
12	北海道新聞	夕刊	ブロック紙	26	中国新聞	夕刊	地方紙
13	中日新聞	朝刊	ブロック紙	27	高知新聞	朝刊	地方紙
14	中日新聞	夕刊	ブロック紙	28	高知新聞	夕刊	地方紙
15	西日本新聞	朝刊	ブロック紙	30	琉球新報	朝刊	地方紙
16	西日本新聞	夕刊	ブロック紙	31	琉球新報	夕刊	地方紙

7.A.4 「白書」のジャンル情報の詳細

「白書」の「ジャンル(1)」列には、白書のタイトルおよび内容によって分類した 9 種類のジャンル名（「安全」「外交」「科学技術」「環境」「教育」「経済」「国土交通」「農林水産」「福祉」）が記載されている。各ジャンルと白書のタイトルは、表 7-13 のように対応している。

表 7-13: 「白書」のジャンル情報

ジャンル	白書タイトル
安全	警察白書
	原子力安全白書
	原子力白書
	交通安全白書
	公害紛争処理白書
	消防白書
	犯罪白書
	防衛白書 / 日本の防衛
	防災白書
外交	外交青書 / わが外交の近況
	政府開発援助（ODA）白書 / 我が国の政府開発援助
科学技術	科学技術白書
	情報通信白書 / 通信白書
環境	環境白書
	循環型社会白書
教育	文部科学白書 / 我が国の文教施策
経済	エネルギー白書
	ものづくり白書 / 製造基盤白書
	経済財政白書 / 経済白書
	公益法人白書
	地方財政白書
	中小企業白書
	通商白書
	独占白書 / 独占禁止白書
	労働経済白書 / 労働白書
国土交通	観光白書
	国土交通白書 / 運輸白書 / 建設白書
	首都圏白書
	土地白書 / 国土利用白書
農林水産	食料・農業・農村白書 / 農業白書
	森林・林業白書 / 林業白書
	水産白書 / 漁業白書
福祉	厚生労働白書 / 厚生白書
	高齢社会白書
	国民生活白書
	少子化社会白書
	障害者白書
	人権教育・啓発白書
	青少年白書
	男女共同参画白書

※「防衛白書 / 日本の防衛」のように、「/」で区切られている白書タイトルは、1976 年から 2005 年までの間にタイトルの変更があったことを表す。

7.A.5 「教科書」のジャンル情報

- ジャンル(1)

「教科書」の「ジャンル(1)」列には、教科の別が記載されている。

例　「国語」「数学」「理科」「社会」「外国語」「技術家庭」「芸術」「保健体育」「情報」「生活」（ただし、「外国語」は中学校と高等学校のみ、「情報」は高等学校のみ、「生活」は 小学校のみとなる）

- ジャンル(2)

「教科書」の「ジャンル(2)」列には、学校の別が記載されている。

例　「小学校」「中学校」「高校」

- ジャンル(3)

「教科書」の「ジャンル(3)」列には、学年の別が記載されている。

例　「1」「2」「3」「4」「5」「6」「(空文字)」

※　「学年」の情報は、小学校・中学校の場合にのみ記載される。高校の場合は空文字になる。

7.A.6 「広報紙」のジャンル情報

- ジャンル(1)

「広報紙」の「ジャンル(1)」列には、当該の自治体の地域が記載されている。

例　「北海道地方」「東北地方」「関東地方」「中部地方」「近畿地方」「中国地方」「四国地方」「九州・沖縄地方」

- ジャンル(2)

「広報紙」の「ジャンル(2)」列には、当該の自治体の都道府県名が記載されている。

例　「北海道」「青森県」「秋田県」「沖縄県」など

なお、「広報紙」の書誌 ID の 4～8 桁目で表される ID は、「全国地方公共団体コード」の上 5 桁と一致しており、広報紙を発行している自治体に対応する。ID と自治体の対応関係を、表 7-14 に示す。

表 7-14: 「全国地方公共団体コード」と自治体名の対応

01103 北海道札幌市東区
01109 北海道札幌市手稲区
01213 北海道苫小牧市
01230 北海道登別市
01631 北海道十勝支庁音更町
02202 青森県弘前市
03208 岩手県遠野市
04101 宮城県仙台市青葉区
04361 宮城県亘理郡亘理町
05201 秋田県秋田市
06207 山形県上山市
07203 福島県郡山市
07447 福島県大沼郡会津美里町
08203 茨城県土浦市
08217 茨城県取手市
08235 茨城県つくばみらい市
09202 栃木県足利市
09213 栃木県那須塩原市
09361 栃木県下都賀郡壬生町
10201 群馬県前橋市
10205 群馬県太田市
10208 群馬県渋川市
11107 埼玉県さいたま市浦和区
11208 埼玉県所沢市
11219 埼玉県上尾市
11461 埼玉県北葛飾郡栗橋町
12104 千葉県千葉市若葉区
12203 千葉県市川市
12206 千葉県木更津市
12229 千葉県袖ケ浦市
13104 東京都新宿区
13108 東京都江東区
13112 東京都世田谷区

13116 東京都豊島区
13120 東京都練馬区
13203 東京都武蔵野市
13209 東京都町田市
13221 東京都清瀬市
14101 神奈川県横浜市鶴見区
14107 神奈川県横浜市磯子区
14114 神奈川県横浜市瀬谷区
14133 神奈川県川崎市中原区
14204 神奈川県鎌倉市
14208 神奈川県逗子市
14212 神奈川県厚木市
15106 新潟県新潟市南区
15210 新潟県十日町市
16202 富山県高岡市
17204 石川県輪島市
18201 福井県福井市
19201 山梨県甲府市
19208 山梨県南アルプス市
20203 長野県上田市
20385 長野県上伊那郡南箕輪村
21204 岐阜県多治見市
21217 岐阜県飛騨市
22103 静岡県静岡市清水区
22136 静岡県浜松市浜北区
22213 静岡県掛川市
22222 静岡県伊豆市
23113 愛知県名古屋市守山区
23211 愛知県豊田市
23302 愛知県愛知郡東郷町
24203 三重県伊勢市
24210 三重県亀山市
25206 滋賀県草津市
25209 滋賀県甲賀市

26108 京都府京都市右京区
26204 京都府宇治市
26407 京都府船井郡京丹波町
27109 大阪府大阪市天王寺区
27123 大阪府大阪市淀川区
27141 大阪府堺市堺区
27210 大阪府枚方市
27220 大阪府箕面市
28102 兵庫県神戸市灘区
28201 兵庫県姫路市
29204 奈良県天理市
29340 奈良県生駒市
30201 和歌山県和歌山市
30206 和歌山県田辺市
31202 鳥取県米子市
32201 島根県松江市
33461 岡山県小田郡矢掛町
34108 広島県広島市佐伯区
34205 広島県尾道市
35210 山口県光市
36341 徳島県名西郡石井町
37201 香川県高松市
38202 愛媛県今治市
39205 高知県土佐市
40203 福岡県久留米市
40305 福岡県筑紫郡那珂川町
41401 佐賀県西松浦郡有田町
42201 長崎県長崎市
43215 熊本県天草市
44202 大分県別府市
45206 宮崎県日向市
46218 鹿児島県霧島市
47209 沖縄県名護市

7.A.7 「Yahoo!知恵袋」のジャンル情報の詳細

「Yahoo!知恵袋」の「ジャンル(1)～(3)」列には、14 種類の「大カテゴリ」、59 種類の「中カテゴリ」、および 130 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を表 7-15 に示す。小カテゴリの違いは / で区切られている。

表 7-15: 「Yahoo!知恵袋」のジャンル情報

大カテゴリ	中カテゴリ	小カテゴリ
エンターテインメントと趣味	ゲーム	ゲーム / オンラインゲーム / トレーディングカード
	テレビ、ラジオ	テレビ、ラジオ / CM / ラジオ
	映画	映画
	音楽	音楽 / 楽器 / 邦楽 / 洋楽
	芸能人、タレント	芸能人、タレント / あの人は今 / 話題の人物
	占い、超常現象	占い、懸賞
	本、雑誌、コミック	本、雑誌、コミック / コミック / 雑誌
インターネット、PC と家電	インターネット	インターネット
	パソコン、周辺機器	パソコン、周辺機器
	家電、AV 機器	家電、AV 機器 / オーディオ
	携帯電話、モバイル	携帯電話、モバイル
ビジネス､経済とお金	家計、貯金	家計、貯金 / ローン / 家計、節約 / 貯金
	株と経済	株と経済 / 株式 / 経済、景気
	企業と経営	企業と経営 / 会計、経理、財務 / 会社情報、業界動向 / 企業法務、知的財産 / 起業
	保険、税金、年金	保険、税金、年金 / 税金 / 年金 / 保険
職業とキャリア	資格、習い事	資格、習い事 / 資格 / 専門学校、職業訓練
	就職、転職	就職、転職 / 就職活動 / 退職、入社手続き
	派遣、アルバイト、パート	派遣、アルバイト、パート / アルバイト、フリーター / パート / 派遣
	労働問題、働き方	労働問題、働き方 / 失業、リストラ / 労働条件、給与、残業 / 労働問題
ニュース、政治、国際情勢	ニュース、事件	ニュース、事件 / 事件、事故、流行 / 話題のことば
	政治、社会問題	政治、社会問題
スポーツ､アウトドア、車	アウトドア	アウトドア / キャンプ / 釣り
	スポーツ	スポーツ / オリンピック / サッカー / ダイビング、サーフィン / 格闘技、武術 / 野球
	バイク	バイク
	自動車	自動車 / 新車 / 中古車
暮らしと生活ガイド	ショッピング	ショッピング / これ、探してます
	ボランティア、環境問題	ボランティア、環境問題

	家事、住宅	家事、住宅 / 家事 / 不動産、引越し
	公共施設、役所	公共施設、役所 / 美術館、博物館、図書館 / 役所、手続き
	福祉、介護	福祉、介護
	法律、消費者問題	法律、消費者問題 / 消費者問題 / 法律相談
	料理、グルメ、レシピ	お酒、ドリンク / レシピ、調理法 / 飲食店、デパ地下 / 料理、食材 / 料理、グルメ、レシピ
健康、美容とファッション	コスメ、美容	コスメ、美容 / エステ、マッサージ / コスメ、化粧品
	ファッション	ファッション
	メンタルヘルス	カウンセリング、治療 / ストレス / 心の悩み、相談
	健康、病気、ダイエット	健康、病気、ダイエット / ダイエット / 病気、症状、ヘルスケア
	恋愛相談、人間関係の悩み	恋愛相談、人間関係の悩み
子育てと学校	子育て、出産	子育て、出産 / 子どもの病気とトラブル / 子育ての悩み / 妊娠、出産
	受験、進学	受験、進学
	小・中学校、高校	小・中学校、高校
	大学、留学	大学、留学 / 大学 / 留学
	幼児教育、幼稚園、保育園	幼児教育、幼稚園、保育園
マナー、冠婚葬祭	マナー	マナー / あいさつ、てがみ、文例
	冠婚葬祭	冠婚葬祭 / 結婚 / 葬儀
	祭りと年中行事	祭りと年中行事
教養と学問、サイエンス	一般教養	一般教養
	芸術、文学、歴史	芸術、文学、歴史
	言葉、語学	言葉、語学
	数学、サイエンス	数学、サイエンス
	天気、天文、宇宙	天気、天文、宇宙
	動物、植物、ペット	動物、植物、ペット
地域、旅行、お出かけ	海外	海外
	交通、地図	交通、地図
	国内	国内 / 花火大会
Yahoo!JAPAN	Yahoo!オークション	Yahoo!オークション
	Yahoo!サービス	Yahoo!サービス
	Yahoo!知恵袋	Yahoo!知恵袋
その他	アダルト	アダルト
	ギャンブル	ギャンブル

7.A.8 「Yahoo!ブログ」のジャンル情報の詳細

「Yahoo!ブログ」の「ジャンル(1)～(3)」列には、15 種類の「大カテゴリ」、54 種類の「中カテゴリ」、および 316 種類の「小カテゴリ」という 3 階層のカテゴリがそれぞれ記載されている。大カテゴリ・中カテゴリ・小カテゴリの一覧を表 7-16 に示す。小カテゴリの違いは / で区切られている。

表 7-16: 「Yahoo!ブログ」のジャンル情報

大カテゴリ	中カテゴリ	小カテゴリ
ビジネスと経済	金融と投資	通貨、為替 / 株式 / 保険 / 貯蓄、預金 / 銀行 / 不動産 / その他金融と投資
	雇用	就職 / 転職 / アルバイト / 人材派遣 / 失業、無職 / その他雇用
	ビジネス	会社経営 / 起業 / その他ビジネス
	職種	事務職 / 営業職 / 技術職 / 企画職 / 専門職 / 公務員 / その他職種
	経済	景気 / 国際経済 / その他経済
コンピュータとインターネット	インターネット	ホームページ / ネットサービス / その他インターネット
	コンピュータ	ソフトウェア / パソコン / 周辺機器 / Windows / Macintosh / その他コンピュータ / UNIX
生活と文化	祝日、記念日、年中行事	クリスマス / 正月 / 誕生日 / バレンタインデー / 花火 / ホワイトデー / 花見 / エイプリルフール / その他祝日、記念日、年中行事
	グルメ、ドリンク	レシピ / 飲食店 / 食べ物 / 飲み物 / 菓子、デザート
	環境問題	その他環境問題 / 省エネ / 自然保護 / リサイクル / ごみ問題 / 地球温暖化
	事件・事故	事件 / 事故 / 防犯
	災害	火災 / 地震 / 台風 / 火山活動 / その他災害
	文化活動	宗教 / ボランティア活動 / 祭りと伝統 / その他文化活動
	季節	冬 / 秋 / 夏 / 春
エンターテインメント	映画	俳優、女優 / その他映画 / 映画祭 / 映画レビュー / 映画監督
	テレビ	アナウンサー / コマーシャル / その他テレビ / ドラマ番組 / バラエティ番組
	音楽	その他音楽 / 音楽祭 / 洋楽 / 邦楽 / 音楽レビュー / ミュージシャン
	占い	心理テスト、性格診断 / タロット占い / 星占い / 血液型占い / 風水 / その他占い
	芸能人、タレント	男性 / 女性 / グループ
	超常現象	幽霊、心霊 / 都市伝説 / UFO / 超能力 / その他超常現象

	テーマパーク	ディズニーリゾート / ユニバーサル・スタジオ・ジャパン / 遊園地 / その他テーマパーク
家庭と住まい	住まい	ガーデニング / 修理とリフォーム / 住居 / インテリア
	ペット、動物	昆虫 / 観賞魚、水草 / 鳥 / ウサギ / ハムスター / 犬 / 猫 / その他ペット
	家庭電化製品	オーディオ / 季節家電 / 映像機器 / 調理器具 / その他家電
	家庭	家計 / 育児 / 家族 / 家庭環境
政治	政界と政治活動	政党､団体 / 選挙 / 政界 / 地方自治 / 軍事 / 国会 / その他政界と政治活動
	国際情勢	中東情勢 / アジア情勢 / アフリカ情勢 / アメリカ情勢 / ヨーロッパ情勢 / オセアニア情勢 / その他国際情勢
健康と医学	美容と健康	フィットネス / スキンケア / ボディケア / ネイルケア / ダイエット / その他美容と健康
	病気、症状	子どもの病気 / メンタルヘルス / 生活習慣病 / アレルギー / その他の病気 / 花粉症
学校と教育	学校	小学校 / 中学校 / 高校 / 専門学校 / 大学 / その他学校 / 受験
	教育	習いごと / 幼児教育 / 社会教育 / その他教育
科学	社会科学	人類学と考古学 / 経済学 / 心理学 / 政治学 / 法学 / その他社会学
	自然科学	化学 / 工学 / 物理学 / 天文学 / 気象学 / 生物学 / その他自然科学
出会い	恋愛	失恋 / 遠距離 / アドバイス / 片思い / 初恋 / その他恋愛
	結婚	離婚 / 結婚式 / 見合い / 再婚 / その他結婚 / 婚約、結納
地域	日本	北海道 / 青森県 / 岩手県 / 宮城県 / 秋田県 / 山形県 / 福島県 / 東京都 / 神奈川県 / 埼玉県 / 千葉県 / 茨城県 / 栃木県 / 群馬県 / 山梨県 / 新潟県 / 長野県 / 富山県 / 石川県 / 福井県 / 愛知県 / 岐阜県 / 静岡県 / 三重県 / 大阪府 / 兵庫県 / 京都府 / 滋賀県 / 奈良県 / 和歌山県 / 島根県 / 岡山県 / 広島県 / 山口県 / 徳島県 / 香川県 / 愛媛県 / 高知県 / 福岡県 / 佐賀県 / 長崎県 / 熊本県 / 大分県 / 宮崎県 / 鹿児島県 / 沖縄県
	世界の地方	アジア / アフリカ / オセアニア / 北アメリカ / 中東 / ヨーロッパ / ラテンアメリカ
特集	趣味とスポーツ	CLUBKEIBA
芸術と人文	芸術、アート	イラストレーション / 絵画 / 写真 / 工芸 / 書道 / その他芸術、アート

	文学	ノンフィクション、エッセイ / 小説 / 詩 / 俳句、川柳 / 短歌 / その他文学 / 伝記、自伝
	デザイン	ファッション / 工業デザイン / 建築デザイン / その他デザイン
	舞台、演劇	観劇 / 伝統芸能 / その他舞台、演劇
	人文科学	倫理学 / 哲学 / 歴史 / その他人文科学
Yahoo!サービス	Yahoo!ブログ	練習用
	Yahoo!オークション	出品 / 落札 / ウォッチリスト / Yahoo!オークションストア
	Yahoo!ゲーム	その他 Yahoo!ゲーム
	Yahoo!アバター	アバター作成
	Yahoo!スポーツ	ファンタジーサッカー
	Yahoo!ショッピング	Yahoo!ショッピングストア
趣味とスポーツ	スポーツ	野球 / サッカー / ゴルフ / テニス / 格闘技 / モータースポーツ / スキー / スノーボード / マリンスポーツ / その他スポーツ / 陸上競技 / バスケットボール / オリンピック / バレーボール / ラグビー / 卓球
	レジャー	旅行 / 釣り / 登山 / 散歩 / キャンプ / その他レジャー
	趣味	読書 / 漫画、コミック / アニメーション / ゲーム / おもちゃ / カラオケ / 携帯電話 / その他趣味
	乗り物	鉄道、列車 / 自動車 / オートバイ / その他乗り物 / 飛行機 / 自転車
	ギャンブル	パチンコ、パチスロ / 競馬 / 宝くじ / その他ギャンブル

7.A.9 「韻文」のジャンル情報の詳細

ジャンル(1)

「韻文」の「ジャンル(1)」列には、韻文の種別が記載されている。

例 「短歌」「俳句」「詩」

7.A.10 「法律」のジャンル情報の詳細

ジャンル(1)

「法律」の「ジャンル(1)」列には、データの取得元である「法令データ提供システム」で採用されている、法務省『日本現行法規』に基づく法律のジャンルが記載されている。一覧を表 7-17 に示す。

表 7-17: 「法律」のジャンル情報

「01 憲法」	「19 災害対策」	「35 金融・保険」
「02 国会」	「20 建築・住宅」	「37 陸運」
「03 行政組織」	「21 財務通則」	「38 海運」
「04 国家公務員」	「23 国税」	「39 航空」
「05 行政手続」	「24 専売・事業」	「40 貨物運送」
「07 地方自治」	「25 国債」	「42 郵務」
「08 地方財政」	「26 教育」	「43 電気通信」
「09 司法」	「27 文化」	「44 労働」
「10 民事」	「28 産業通則」	「45 環境保全」
「11 刑事」	「29 農業」	「46 厚生」
「12 警察」	「30 林業」	「47 社会福祉」
「14 国土開発」	「31 水産業」	「49 防衛」
「15 土地」	「32 鉱業」	「50 外事」
「16 都市計画」	「33 工業」	
「17 道路」	「34 商業」	

7.A.11 「国会会議録」のジャンル情報の詳細

ジャンル(1)

「国会会議録」の「ジャンル(1)」列には、開催院の別が記載されている。

例　「衆議院」「参議院」

ジャンル(2)～(3)

「国会会議録」の「ジャンル(2)～(3)」列には、4 種類の会議種別（「常任委員会」「特別委員会」「本会議」「その他」）と会議名称が、それぞれ記載されている。会議種別と会議名称は、表 7-18 のように対応している。

表 7-18: 「国会会議録」のジャンル情報

会議種別	会議名称
本会議	本会議
常任委員会	安全保障委員会、運輸委員会、科学技術委員会、外交防衛委員会、外務委員会、環境委員会、議院運営委員会、経済産業委員会、決算委員会、決算行政監視委員会、建設委員会、厚生委員会、厚生労働委員会、行政監視委員会、国土・環境委員会、国土交通委員会、財政・金融委員会、財政金融委員会、社会労働委員会、商工委員会、総務委員会、大蔵委員会、地方行政委員会、逓信委員会、内閣委員会、農林水産委員会、文教委員会、法務委員会、予算委員会
特別委員会	ロッキード問題に関する調査特別委員会、安全保障特別委員会、沖縄及び北方問題に関する特別委員会、科学技術振興対策特別委員会、個人情報の保護に関する特別委員会、交通安全対策特別委員会、公害対策及び環境保全特別委員会、国会等の移転に関する特別委員会、国旗及び国歌に関する特別委員会、国際平和協力等に関する特別委員会、災害対策特別委員会、世界貿易機関設立協定等に関する特別委員会、政治倫理の確立及び公職選挙法改正に関する特別委員会、青少年問題に関する特別委員会、物価等対策特別委員会、物価問題等に関する特別委員会
その他	議院運営委員会庶務小委員会、憲法調査会、国民生活・経済に関する調査会、国民生活・経済に関する調査特別委員会高齢化社会検討小委員会、産業・資源エネルギーに関する調査会、少子高齢社会に関する調査会、文教委員会入試問題に関する小委員会、予算委員会公聴会、予算委員会第三分科会、予算委員会第四分科会、予算委員会第五分科会、予算委員会第六分科会、予算委員会第八分科会

付録 7-B: サンプル ID ベース書誌情報データの構成

サンプル ID ベース書誌情報データ（Joined_info.txt）は、表 7-1（Bibliography.txt）、表 7-4（Sample_info.txt）、表 7-6（Directory.txt）、表 7-8（Article.txt）の情報を結合し、サンプル ID を単位として生成したものであり、『中納言』での書誌情報表示に用いられているデータである。ここに記録された「人名 ID」「人名」「生年代」「性別」の情報は、各サンプルを実際に執筆した人物に関する情報を表している。同一サンプルに対して複数のレコードが存在する場合は、重複する情報をスラッシュで区切って並べている。

表 7-19: サンプル ID ベース書誌情報データの構成

フィールド名称	結合元のファイル
サンプル ID	---
書誌 ID	Bibliography.txt
タイトル	Bibliography.txt
副題	Bibliography.txt
巻号	Bibliography.txt
責任表示	Bibliography.txt
出版者	Bibliography.txt
出版年	Bibliography.txt
ISBN	Bibliography.txt
サンプル抽出基準点ページ	Sample_info.txt
ジャンル 1	Bibliography.txt
ジャンル 2	Bibliography.txt
ジャンル 3	Bibliography.txt
ジャンル 4	Bibliography.txt
責任表示 ID	Bibliography.txt
人名 ID	Article.txt
人名	Directory.txt
生年代	Directory.txt
性別	Directory.txt
corpusName	サブコーパスの略称（新規追加）

第 8 章　文境界情報

浅原 正幸　小西 光　田中 弥生　間淵 洋子

8.1　はじめに

本章では『現代日本語書き言葉均衡コーパス』の文境界情報について説明する。文境界情報を規定する手がかりになるものとして、(1)文字情報を用いるもの、(2)形態論情報を用いるもの、(3)係り受け関係を用いるものなどが考えられる。BCCWJ-DVD 版（Version 1.0）は、文境界情報を含む文書構造タグの整備と形態論情報の整備とを並行して実施していたため、文字情報を手がかりとして用いた文境界認定作業にとどまっていた。また、工数の制約から知恵袋（OC）については論理行 [1] を表す<webLine>タグを付与するにとどめ、実質的な文境界修正作業を行っていなかった。その結果、BCCWJ-DVD 版 （Version 1.0）の文境界認定基準の妥当性については様々な指摘がなされた。

BCCWJ-DVD 版（Version 1.1）へのバージョンアップに際し、M-XML と TSV（第 6 章）に対して、形態論情報を手がかりとして用いた文境界基準を再策定することで、問題の解消を試みた。以下では、BCCWJ-DVD 版（Version 1.0）の問題点を示し、また、それに対しどのような文境界修正作業を行ったのか説明する。

本章の構成は以下のとおりである：8.2 節では BCCWJ-DVD 版 （Version 1.0）の文境界認定基準を示す。8.3 節では BCCWJ-DVD 版 （Version 1.1）の文境界認定作業について述べる。8.4 節では、コアデータに対するその他の文境界情報を紹介する。

8.2　BCCWJ-DVD 版（Version 1.0） の文境界認定基準

本節では BCCWJ-DVD 版 （Version 1.0）の文境界認定基準について述べる。はじめに文境界認定基準における手がかりについて概観する。

8.2.1 文境界認定基準についての手がかり

文境界認定においては、何らかの「手がかり」を用いて規則を人手で記述する必要がある。文境界認定作業をある程度自動化するためには何を「手がかり」に使うかが重要となる。以下では「手がかり」として、(1)文字情報を用いるもの、(2)形態論情報を用いるもの、(3)係り受け関係を用いるものの 3 種類について詳しく述べる。

(1) 文字情報に基づく認定とは、句点などに基づき文境界を認定する手法である。多くの形態素解析の前処理として、句点記号「。」「.」感嘆符「！」疑問符「？」などを手がかりとした文境界認定が行われている。少し高度な情報として、開き括弧や閉じ括弧

[1] 本節では紙面などの物理的制約によって指示される行を「物理行」「表示行」と呼ぶのに対して、改行コードやブロック要素などにより指示される行を「論理行」と呼ぶ。

を用いた規則を記述し、括弧の対応をとるという手法が存在する。

(2) 形態論情報に基づく認定とは、形態素解析により認定される品詞情報などを用いる手法である。句点のリストを第 5 章に示した短単位形態論情報（小椋他 2011）における品詞「記号-句点」などに汎化できるほか、開き括弧や閉じ括弧についても「記号-括弧開」「記号-括弧閉」と汎化して記述することができる。さらに、辞書に登録されている固有名詞や顔文字などに埋め込まれている記号などを文境界候補から除外することができる。その一方で、形態素解析誤りの影響をある程度見込んで処理する必要がある。

(3) 係り受け関係に基づく認定とは、文境界認定に係り受け関係のスパンを用いる手法である。括弧内の要素が文であるかどうかを認定するために括弧内の要素が連結係り受け木をなすかを判定したり、括弧の前後で係り受け関係があるかどうかで文要素の入れ子を認定したりする。

8.2.2 BCCWJ-DVD 版（Version 1.0）における文境界認定基準の概要

まず、BCCWJ-DVD 版 （Version 1.0）における文境界について述べる。BCCWJ-DVD 版 （Version 1.0）においては文字情報のみを含む C-XML（第 4 章）と形態論情報を含む M-XML （第 6、9 章）の 2 種類の XML 形式でデータが表現されている。文境界情報は XML 内の sentence 要素として表現されている。この 2 種類の形式において認定している文境界に差異がある。

C-XML における文境界認定：

C-XML（第 4 章）においては手がかりとして文字情報を用いた自動処理に基づく文境界認定が基本となっている。話し言葉や既存の書き言葉コーパスと異なり、元媒体のレイアウト情報に基づく文書構造情報（ブロック要素）が利用されている。以下 C-XML における文のスパンを表現する sentence 要素の認定規則について例（図 8-1）を示しながら解説する。自動認定においては句点記号「。」「.」感嘆符「！」疑問符「？」（以下文末記号）やブロック要素開始位置直前を文区切り位置とみなし、直前文の末尾を sentence 要素の始端とみなす処理（sentence タグ<sentence> </sentence> を付与）を行う（例 C-1）。文末記号によって認定される sentence 要素を正則な sentence 要素と呼ぶ。論理行頭からひとつ以上の sentence 要素の並びが存在し、かつ、行末に文末記号がない場合は sentence 要素とみなす（例 C-2）。論理行中にひとつも sentence 要素がなく文末記号もない場合、その論理行全体を sentence 要素とみなす（例 C-3）。これらの文末記号以外によって認定される sentence 要素は、特殊な文として属性 type="quasi"を付与する（例 C-2、C-3：以下 sentence@quasi 要素と略記）。文字情報として 9 種類の括弧の対応（括弧類A[2]）などを用いて、文認定時に sentence 要素の入れ子を許している。

括弧内にひとつも文末記号を含まない場合、括弧内に sentence 要素を認定しない（例

2 括弧類A：「補助記号−括弧開」「補助記号−括弧閉」のうち（ ）〔〕［ ］｛ ｝〈〉《》「」『』【】9 対

C-4)。括弧内にひとつ以上の文末記号が含まれる場合、括弧内に sentence 要素を認定する（例 C-5)。括弧内にひとつ以上の文末記号が含まれ、かつ、閉じ括弧直前に文末記号が出現しない場合、閉じ括弧直前までの部分を特殊な文とみなし、属性 type="quasi"を付与する（例 C-6)。

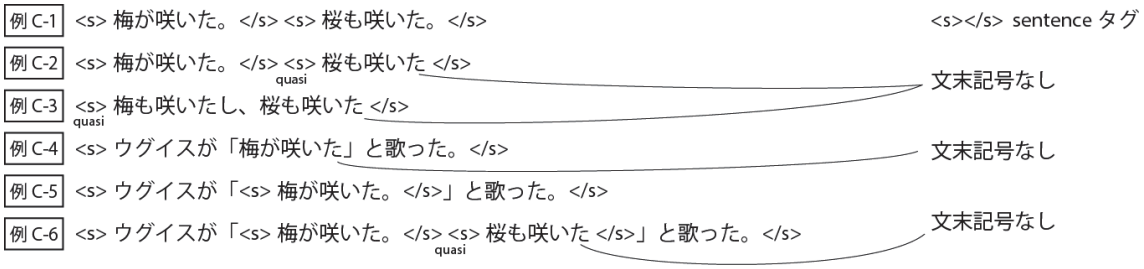


図 8-1:　C-XML における文境界認定

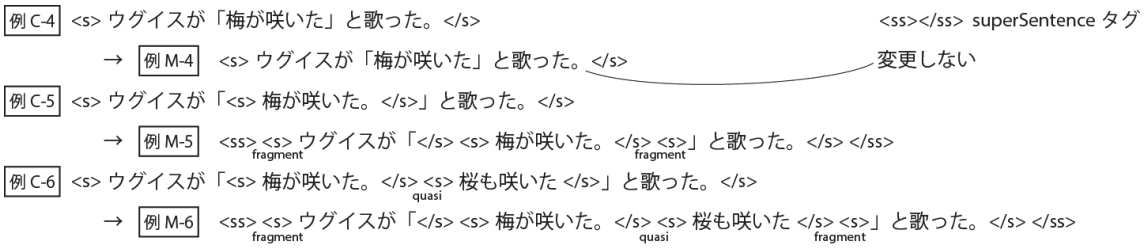


図 8-2:　C-XML から M-XML への変換

M-XML における文境界認定：

M-XML（第 6、9 章）においては、C-XML の文境界認定を基礎としつつ、C-XML とは異なる、より単純化した文境界認定を行う方針を採用した。C-XML の問題点として、sentence 要素がきわめて長くなる場合があること、形態素解析などの入力となる「文」が定めがたいこと、データを文番号で管理できないことの三つがあげられる。

M-XML では、C-XML において sentence 要素が入れ子になっている場合に、その最も内側（下位）にあるもののみを正則の sentence 要素とし、外側（上位）にある sentence は superSentence とする。その上で、superSentence の内側にありながら正則の sentence 要素の外側に位置する部分については、新たに sentence 要素と見なすとともに type="fragment"という属性（以下 sentence@fragment 要素と略記）を与えて、文断片であることを明示する。この際、括弧記号のみからなる文断片要素を作らないために、内側の sentence 要素に隣接する括弧記号を送り込む。最終的に superSentence と sentence の 2 階層からなる文境界情報が残される（図 8-2)。

例 C-4 においては sentence 要素に入れ子が発生していないため、C-XML と M-XML の sentence 要素は一致する（例 M-4)。

例 C-5 においては、括弧内の最内スパンの sentence 要素 "梅が咲いた。" を M-XML に

おける正則な sentence 要素と見なす（例 M-5）。例 C-5 における最外スパンは新たに superSentence 要素として認定する。正則 sentence 要素に含まれない最外スパンの連続文字列については、sentence@fragment 要素として認定する。ただし、正則 sentence 要素に隣接する括弧記号は sentence 要素に送り込む。

例 C-6 においては括弧内に正則な sentence 要素“梅が咲いた。”と sentence@quasi 要素“桜も咲いた”の二つが認定されている。例 C-6 における最外スパンを新たに superSentence 要素として認定する（例 M-6）。括弧内の 2 種類の sentence 要素（正則な sentence 要素と sentence@quasi 要素）を認定し、これに含まれない前後の連続文字列を sentence@fragment 要素として認定する。ただし、内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む。

しかし、例 M-5・M-6 における、「内側の sentence 要素に隣接する括弧記号は内側の sentence 要素に送り込む処理」が網羅的ではなかった。今回はこの問題を解決するために網羅的なパターンを記述し、再処理する。図 8-2 では、問題になるパターンを示した。

8.3 BCCWJ-DVD 版（Version 1.1）における文境界認定基準

8.3.1 BCCWJ-DVD 版（Version 1.1）における文境界認定の作業方針

以下に文境界認定の作業方針について述べる。BCCWJ-DVD 版（Version 1.0）の文字情報による自動処理と、BCCWJ-DepPara の係り受け関係の情報による人手修正との中間的な処理として、**形態論情報を用いた自動抽出結果の人手修正**をコアデータ・非コアデータ全体に対して実施する。

修正方法としては、まず C-XML における文字列レベルの情報を用いた文境界認定におけるバグ相当のものを自動抽出して人手修正し、次に M-XML に変換する際のバグ相当のものを、形態論情報を用いて自動抽出して、バッチ処理および人手修正を行う。基本的に最内スパンの正則な sentence 要素を認定するとともに、その作業に伴い発生する sentence@fragment 要素のような文が認定されることを許す。係り受け関係の整合性は検証しないが、括弧内の要素について最低限の確認作業（強調や補足の認定）を行う。詳細を以下に示す：

[処理 C] C-XML レベルで認定できる誤りの検出

BCCWJ-DVD 版 （Version 1.0）において、文字情報に基づく処理により 9 対の括弧（括弧類A）内に文末記号があるが文境界が設定されていない要素が約 6,000 箇所発見された。顔文字に埋め込まれた文末記号や括弧が対応していない事例について、全数人手で確認する。

[処理 M] M-XML レベルで認定できる誤り検出

処理 C が完了後、形態論情報を用いた誤り検出を行う。形態論情報を用いた誤り検出

においては、国立国語研究所に寄せられている様々な誤り報告事例や他のアノテーション作業時に問題となった事例をもとに、人手で形態論情報を用いたパターンを記述した。このパターンの認定においてはそのマッチする事例のうち修正率（真に修正すべき事例数／マッチする事例数）に基づいて 2 種類の処理を行う。

[M(α)] 修正率が高いパターン : マッチするほとんどの事例が真に修正すべき事例であるが、例外的に修正しなくてもよい事例が出現するパターン。これらについては、バッチ処理適用前に例外的な事例を排除するように人手で確認する。人手確認後バッチ処理で修正する（修正箇所自動抽出→人手例外確認→バッチ処理)。

[M(β)] 修正率が低いパターン : マッチする事例の一部のみを修正するパターン。全数確認は困難であるが、修正すべき事例が含まれうるパターンを先にバッチ処理で展開し、逐一人手で確認する（修正箇所自動抽出→人手修正処理)。

今回の修正は形態論情報を含む M-XML のみに対して実施し、C-XML については実施しない。この修正にともない、必要があれば形態論情報・文書構造タグも修正する。

8.3.2 BCCWJ-DVD 版（Version 1.1）における文境界認定基準の詳細

8.3.2.1 基準の前提

文境界認定基準の前提として、今回踏襲する BCCWJ-DVD 版 （Version 1.0）の文境界認定基準 3 点について示す。

1 点目は、現存する superSentence 要素を踏襲することを前提に sentence タグを付与することである。

2 点目は、助詞・助動詞から始まる、助詞・助動詞で終わる、助詞・助動詞のみの sentence 要素の発生を認めることである。

3 点目は、括弧内に文末記号が含まれない場合には sentence タグは付与しないことである（例 C-4、例 M-4 を踏襲)。

以下 8.3.2.2 節では括弧内に文末記号が含まれる場合に対してパターンを定義して行った修正作業について示す（処理 M(α))。8.3.2.3 節ではパターンに基づく機械処理で一括処理できない事例を中心に、人手で行う認定作業について示す（処理 M(β))。8.3.2.4 節では、今回廃止した BCCWJ-DVD 版 （Version 1.0）の属性とタグについて示す。以下、例文中、開始 sentence タグを <s>、終了 sentence タグを</s>と略記する。全角空白を□で表す。

8.3.2.2 処理 M(α) : 修正率の高いパターン・認定基準

以下修正率の高いパターンについて示す。これらは最初に修正箇所自動抽出を行い、次に人手で例外を確認し、最後にバッチ処理を行うという手続きで誤りが修正される。

1. 句点類Ｂ[3]のみ、もしくは、句点類Ｂの前に記号類Ｃ[4]があり、かつ、句点類Ｂと記号

[3] 句点類Ｂ :「補助記号−句点」。 ！ . ？ の 4 種。

類Cのみで構成されている sentence 要素は、前の sentence 要素の末尾に移動[5]

(1) PB26_00004

（9桁の英数字はサンプルID、1行が1 sentence 要素、横線上が修正前・横線下が修正後）

<s>でも、お客様が並んでしまったら、それより早めに放送してください」 </s>
<s>。</s>

<s>でも、お客様が並んでしまったら、それより早めに放送してください」。</s>

2. 【原則】〔括弧開〕[6]で終わっている sentence 要素は、次の sentence 要素の頭に〔括弧開〕を移動

(2) PN1b_00009

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ（</s> ←注目点
<s>異議なし！</s>
<s>)。</s>

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ</s>
<s>（異議なし！)。</s>

2-a.【例外処理】〔括弧開〕の前がすべて空白の場合も、それらすべてを次の sentence 要素の頭に移動

(3) OY14_12372

<s>□□□□□□『</s> ←注目点
<s>今度は□一緒にファーストで行きたいね□！！</s>
<s>□』</s>

<s>□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</s>

3. 【原則】〔括弧閉〕のみ、もしくは〔括弧閉〕で始まり、かつ、〔括弧閉〕と記号類D

---

[4] 記号類C：「補助記号−一般」（文境界を示す）― … ‒ ・ ～ 【 】〔 〕 - ･･ ♩ ♪ ♫ ♬ 《 》 — — ━ の20 種。

[5] 条件を規定する演算子は、打消の助動詞を否定とし、「かつ」を論理積とし、「もしくは」を論理和とした場合に、この順で優先順位が高い加法標準形で記述する。

[6] 今回は形態論情報により括弧として定義されている「補助記号−括弧開」「補助記号−括弧閉」の全12 種を用い、それぞれ〔括弧開〕・〔括弧閉〕と呼ぶ：‘’“”〈〉《》「」『』【】［ ］｛ ｝等。

[7]のみで構成された sentence 要素は、前の sentence 要素の末尾に移動

(4) PN1b_00009

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ（</s>
<s>異議なし！</s>
<s>）。</s>　←注目点

<s>それより「ブラボー砦の脱出」だ、「星のない男」だ</s>
<s>（異議なし！）。</s>

3-a.【例外処理】上記 3.を適用した結果、〔括弧閉〕（と記号類Dのまとまり）を移動した先の sentence 要素が〔括弧閉〕と記号類D・E[8]のみで構成されている場合は、それらを前の sentence 要素の末尾に移動

(5) PN2d_00008

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</s>
<s>？</s>
<s>）。</s>

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（</s>
<s>？）。</s>　←注目点：ここが記号のみ

<s>□真中に意中の人がいるか否かははっきりしないが、食べ物に反して男性の好みはうるさそう（？）。</s>

4. 【原則】〔括弧閉〕で始まり、かつ、〔括弧閉〕に任意の短単位が後続する sentence 要素は、前の sentence 要素の末尾に〔括弧閉〕のみを移動

(6) PN5f_00020

<s>（咽喉？</s>
<s>）…と其奴がね、異に蔑んだ笑い方をしたものです。</s>

---

[7] 記号類D：句点類B、記号類C、「空白」1 種、「補助記号-読点」2 種。

[8] 記号類E：「記号-一般」2,003 種、「記号-文字」255 種、「空白」1 種、「補助記号-AA-一般」78 種、「補助記号-AA-顔文字」2,405 種、「補助記号-一般」（文境界を示さない）444 種、「補助記号-括弧開」12 種、「補助記号-括弧閉」12 種。

<s>（咽喉？）</s>
<s>…と其奴がね、異に蔑んだ笑い方をしたものです。</s>

4-a【例外処理】〔括弧閉〕に記号類Ｆ[9]が続く場合は、記号類Ｆ以外の短単位が出現するまでの範囲を前の sentence 要素の末尾に移動

(7) OC06_00325 （この例では〔括弧閉〕と読点を移動）
<s> 峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して（</s>
<s>あおるつもりじゃないが。。</s>
<s>)、車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</s>

<s>峠や市街地でも、追い越し禁止道路で前を走る多少遅い車に接近して</s>
<s>（あおるつもりじゃないが。。)、</s>
<s>車が遠慮して道を譲ってくれた時、だいたい頭を下げて追い抜きます。</s>

4-b.【例外処理】空白で始まり、〔括弧閉〕と空白のみで sentence 要素を構成する場合は、それらすべてを前の sentence 要素の末尾に移動

(8) OY14_12372
<s>□□□□□□□『</s>
<s>今度は□一緒にファーストで行きたいね□！！ </s>
<s>□』</s> ←注目点

<s>□□□□□□□『今度は□一緒にファーストで行きたいね□！！□』</s>

4-c.【例外処理】上記 4-a.を適用した結果、「(？)」「(！)」の文字列を sentence 要素に含む場合には、前後の sentence 要素をひとまとまりにする（8.3.2.3 の “文境界認定を打ち消して文を結合する場合”の 1. を参照）

(9) PM41_00071
<s>この業界にしては珍しく（</s>
<s>？</s>
<s>)、可愛らしい女性編集長である。</s>

---

[9] 記号類Ｆ：記号類Ｃ、「補助記号−括弧閉」12 種。

<s>この業界にしては珍しく（？）、可愛らしい女性編集長である。</s>

5. 読点で始まっている場合は、前の sentence 要素の末尾に読点のみを移動

(10) PB45_00024
<s>「ブオノ・ヴェーロ？」</s>
<s>、美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</s>

<s>「ブオノ・ヴェーロ？」、</s>
<s>美味しいだろうと言ったオジサンはイタリア人で、ここに住む孫のためにナポリの店を引き払いやって来たのだという。</s>

8.3.2.3 処理 M(β) : 修正率の低いパターン・認定基準

以下の例は修正率が低いパターンで、手がかりにより候補を枚挙したうえで人手により修正すべきかどうかを判定する。大きく分けて「文境界を認定して分割する場合」と「文境界認定を打ち消して文を結合する場合」の 2 種類がある。これらは、最初に修正箇所自動抽出を行い、次に人手修正処理をすることで誤りを修正する。

文境界を認定して分割する場合（特に Web データ）

1. sentence 要素の中に顔文字を含み、かつ、その顔文字が文末表示だと考えられる場合は分割

(11) OC06_02963
<s>そーですよ＾＾一番左です＾＾</s>

<s>そーですよ＾＾</s>
<s>一番左です＾＾</s>

2. sentence 要素の中に（涙）等の（X）を含み、かつ、その（X）が文末表示だと考えられる場合は分割

(12) OY14_10161
<s>イプ『</s>
<s>違う！</s>

<s>作りすぎただけだっ（照）ナマモノだから今日中に食え』</s>

<s>イプ</s>
<s>『違う！</s>
<s>作りすぎただけだっ（照）</s>
<s>ナマモノだから今日中に食え』</s>

3. 【特殊事例】空白で文が区切られる場合等は分割

(13) OY14_12372
<s>□□□□□□□□『だね、ローマが一番だったよ□日曜なのでバチカンに行ってミサを聞いた</s>
<s>□□□□□□□□□ミケランジェロも見たよ』□うん、おいらはイタリアはしらない</s>

<s>□□□□□□□□『だね、ローマが一番だったよ□</s>
<s>日曜なのでバチカンに行ってミサを聞いた</s>
<s>□□□□□□□□□ミケランジェロも見たよ』□</s>
<s>うん、おいらはイタリアはしらない</s>

文境界認定を打ち消して文を結合する場合（特に雑誌・Web データ）

1. 係り受け関係を結べる要素が後続し、sentence 要素内に含めるべきと判断される「？」「！」は結合

(14) PM11_00263
<s>今が買い！</s>
<s>の中古ＭＦ一眼レフ</s>

<s>今が買い！の中古ＭＦ一眼レフ</s>

2. 補足を表す丸括弧（括弧内に句点を含まないものに限定）内に「？」「！」が含まれる、かつ、丸括弧内に含まれる要素が体言で終わる場合、結合

(15) OY01_00185
<s>この大会のチラシを、今夜（</s>
<s>昨夜？</s>

<s>）のハードルの練習中にわざわざ七夕ホールまで持ってきてくださったのです！</s>

<s>この大会のチラシを、今夜（昨夜？）のハードルの練習中にわざわざ七夕ホールまで持ってきてくださったのです！</s>

3. 【原則】係り受け関係を結べる要素が、原本レイアウト情報を反映した結果二つの sentence 要素に分割されていて、括弧内に文末記号が含まれない場合は結合

(16) PB1n_00024

<s>すると、</s> ←注目点：紙面上にて改行により sentence 要素が分割されている
<s>「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</s>

<s>すると、「溶岩流が危険だから、逃げるんです」という答えが返ってきたのである。</s>

3-a.【例外処理】括弧が強調やタイトル等の目的で用いられている場合

(17) OC01_03215

<s>ゆうべＰＭ９時から日本テレビ「</s>
<s>ものまねバトルオール新ネタ！</s>
<s>夏祭りＳＰ</s>
<s>」に出てましたよ。</s>

<s>ゆうべＰＭ９時から日本テレビ「ものまねバトルオール新ネタ！夏祭りＳＰ」に出てましたよ。</s>

4. 【特殊事例】〔括弧閉〕に丸括弧で注釈が後続する場合は結合しない

(18) PN4c_00011

<s>□だが、農業団体の韓国農業経営人中央連合会は、</s>
<s>「通貨危機で金利負担が膨らみ、農家は今も借金に苦しんでいる。</s>
<s>対策は成功していない」</s>
<s>（政策調整室）と批判的だ。</s>

8.3.3 BCCWJ-DVD 版（Version 1.1）における廃止事項

- BCCWJ-DVD 版（Version 1.0）に規定されていた以下の要素・属性を、BCCWJ-DVD 版（Version 1.1） の M-XML では廃止する。
  sentence タグの属性 type="quasi" : sentence タグの自動付与にあたり、文末記号以外によって認定される特殊な文であることを表すための属性である。今回、文末記号によらない新たな基準に基づき人手で文境界を認定したことで、文の属性（「quasi」は「擬似」の意）として不適切となるため廃止する。
- webLine 要素：「Yahoo!知恵袋」データに対する sentence タグの自動付与にあたり、「文を分断しない範囲で」データ上の物理行（改行記号により自動的に認定される行）を連結した上で認定した、論理行（意味的なまとまりを伴う行）相当のスパンを表す要素である。今回の文境界認定基準と、BCCWJ-DVD 版（Version 1.0）作成時に任意に「文を分断しない」と判断した行との間には矛盾が生じる場合もあるため、不要と判断し廃止する。

8.4 BCCWJ-DepPara における文境界認定

BCCWJ-DVD 版 （Version 1.0） （C-XML、 M-XML） や BCCWJ-DVD 版 （Version 1.1） （M-XML）とは異なる文境界認定基準として、係り受けアノテーションである BCCWJ-DepPara における文境界認定基準（小西 2013）がある。コアデータのみを利用する際には、BCCWJ-DepPara の文境界基準を利用することも考えられる。

係り受けアノテーション従事者は BCCWJ-DVD 版 （Version 1.0） における文境界の問題点として、基準の手がかりが文字列に基づく手法であるために係り受けを分断するような文境界が大量に発生すること、sentence@quasi 要素や sentence@fragment 要素においては要素内に係り先が存在せず、離れた別の sentence 要素に係り先を認定するような現象が起きること、全要素を xpointer などを用いないひとつの XML ファイルとして表現するために ad hoc な後処理がなされ文単位認定に無理が生じていること、実データを見ても必ずしも報告書どおりの処理がなされていないことの四つをあげている。

BCCWJ-DepPara は BCCWJ に対する係り受け・並列構造アノテーションである（浅原・松本 2013)。2012 年 10 月に BCCWJ-DVD 版（Version 1.0）を対象とした最初のバージョンが公開[10]されている。

基本方針として、元の文書構造タグを用いず、文の内容に即して “EOS” ラベルと“Z” ラベルの 2 種類の文境界を認定している（浅原 2013)。 “EOS” ラベルは、係り受け関係がつながる範囲で文を連結したもので、C-XML の最外スパンや M-XML の superSentence 要素に近い基準となっている。“Z” ラベルは、係り受け関係ラベルの一種で “EOS” ラベルで区切られる範囲内に出現する文末記号に対し付与される。“Z” ラベルは文末要素にしか付与されないが、“Z” ラベルを根とする係り受け木の最大スパンを確認することで、局所的な

[10] https://github.com/masayu-a/BCCWJ-DepPara

文の文頭要素が認定できるために、実質的に文の入れ子構造を認定している。括弧内の要素の扱いにおいては、コアデータに出現する括弧で括られた要素の機能を補足・発話・心内・引用・箇条書き・強調の 6 種類に分類し、要素の意味についても調査して、文認定を行っている。

参考文献

浅原正幸（2013）「係り受け関係アノテーション基準の比較」『第 4 回コーパス日本語学ワークショップ予稿集』,81-90.

浅原正幸・松本裕治（2013）「『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション」『言語処理学会第 19 回年次大会発表論文集』,66-69.

小西光・小山田由紀・浅原正幸・柏野和佳子・前川喜久雄（2013）「BCCWJ 係り受けアノテーション付与のための文境界再認定」『第 4 回コーパス日本語学ワークショップ予稿集』,135-142.

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011a）「『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）」国立国語研究所内部報告書 LR-CCG-10-05-01

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕（2011b）「『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.

第 9 章　形態論情報付き統合形式 XML（M-XML）

小木曽 智信　間淵 洋子　前川 喜久雄

9.1　M-XML の概要

形態論情報付き統合形式 XML（Morphology-base XML 以下、M-XML と略記する）は、文字ベースの XML（C-XML）フォーマットをもとにして、固定長・可変長サンプルを統合し、言語構造を一定程度反映させた XML フォーマットである。短単位・長単位の形態論情報を、階層構造を維持したまま埋め込み、言語構造に関わる情報を扱いやすくしている。XML ファイルの文字符号化方式は UTF-8（BOM なし）である。

第 6 章で述べたとおり、M-XML には、数字変換（NumTrans）処理を施した M-XML_NT と、数字変換を行っていない M-XML_OT の 2 種類の本文がある。それぞれのデータの格納場所については第 1 章を参照されたい。

短単位・長単位の形態論情報は、M-XML・TSV の両形式とも同じ内容が付与されており、同一部分の短単位・長単位が異なって解析されていることはない。

9.1.1 固定長と可変長の統合

C-XML では、固定長サンプルと可変長サンプルが別の XML 文書として構造化されている。しかし、2 種類のサンプルは同一の文書から採集されているため、多くの部分が重複している。こうしたデータに形態論情報を付与し整備する場合には、同一内容のテキストを 2 回処理する必要がないように、統合して扱うことができた方が望ましい。しかし、タグが交叉することになるため、別の構造を持つ二つの XML を単純に統合することはできない。そこで、統合形式では以下のような方法によって固定長と可変長を統合することとした。

そもそも、文書構造を意識して採集された可変長サンプルとは違い、均一な長さのサンプルを取得する目的で作られた固定長サンプルでは、文書構造を示すブロック要素タグは大きな意味を持たない。そこで、M-XML では、可変長サンプルの文書構造だけを保持し、固定長の範囲は形態論情報（長単位）タグに付与する属性で示すこととした。可変長部分から固定長部分がはみ出している場合には、はみ出した部分を単純なコンテナ（ ）で囲み、インライン要素だけを保持した。

M-XML は次のような属性を持つ mergedSample 要素をルートとして上記の要素をまとめ上げている。

```
<mergedSample sampleID="サンプル ID" type="BCCWJ-MorphXML" version="1.1">
```

なお、NumTrans 処理が行われた M-XML_NT のサンプルについては、次のように NumTrans 属性を付与して区別している。

```
<mergedSample sampleID="サンプル ID" type="BCCWJ-MorphXML" version="1.1"
NumTrans="true">
```

M-XML_NT のファイルであっても、対象となる数字列が存在せず、NumTrans 処理がなされていないものについてはこの属性は付与されていない。したがって、こうしたサンプルにおいては、M-XML_NT と M-XML_OT のファイルが完全に一致する。

9.1.2 異なる文書型定義の統合

C-XML は、サブコーパス・レジスターによって異なる文書型定義（DTD）が用いられている。Yahoo!知恵袋（OC）、Yahoo!ブログ（OY）、教科書（OT）、韻文（OV）は、おおよそ共通の構造を持ちながらも、一般の可変長サンプルとは異なるそれぞれ独自の文書型定義によっている。そのため、すべてのデータを統一的に処理しようとするとき問題となる場合がある。

そこで、M-XML では、タグセットを一部変更して、すべてのサブコーパス・レジスターについて共通の文書型定義で処理できるようにした。C-XML に比較してやや緩い制約での検証になるが、すべての XML ファイルは単一の XML スキーマで検証済みである。この統合に際してレジスター独自のタグを次のように一部変更している。

OC　:　`<OCQuestion>` → `<article articleID="`サンプル ID-Question`">`
　　`<OCAnswer>` → `<article articleID="`サンプル ID-Answer`">`
OC, OY :　`<br type='physicalLine_original' />` → `<webBr/>`
OT　:　`<root>` → `<squareRoot>`

9.2　要素の階層構造

BCCWJ における短単位・長単位・文節は、その定義から入れ子構造をなす。また、文節はこれが連なって文を構成するし、短単位は文字から構成されるから、BCCWJ の形態論情報は、結局次のような言語単位の階層構造の中に位置づけられることになる。

文章／文／文節／長単位／短単位／文字

文書構造タグや階層化された形態論情報を活用するためには、この階層構造・包含関係がそのまま XML フォーマットに反映されることが望ましい。この考え方に従い、M-XML では、次のような階層構造で形態論情報を付与した。

文書構造（ブロック）タグ／sentence（文）／LUW（長単位）／SUW（短単位）／文字

以下はそのサンプルとしてひとつの文（sentence 要素）を抜き出したものである（見やすさのため属性を省略した。形態論情報タグの詳細は第 6 章を参照のこと）。

```
<sentence>
  <LUW><SUW>公共</SUW><SUW>工事</SUW><SUW>請負</SUW><SUW>金額</SUW></LUW>
  <LUW><SUW>の</SUW></LUW>
  <LUW><SUW>動き</SUW></LUW>
（略）
</sentence>
```

以上の形態論情報の階層に C-XML の諸要素を当てはめるならば、図 9-1 のような階層構造が考えられる（網掛けはすべてのテキストに必須の要素）。このとき C-XML における諸要素がこの階層と齟齬を来すことが問題となるが、M-XML では、次節以降に示すように C-XML のタグに修正を加えることで対処している。

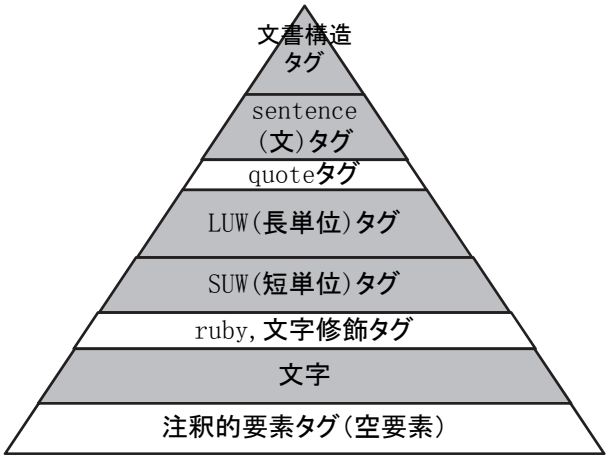


図 9-1: 形態論情報付き統合形式 XML フォーマットの階層構造

9.3 C-XML と M-XML の相違点

これまでに見てきたとおり、C-XML と M-XML の大きな相違は次の 4 点である。

- 固定長と可変長が統合されていること（9.1.1 節）
- レジスターごとに異なる文書型定義が統一されていること（9.1.2 節）
- 短単位（SUW）・長単位（LUW）の形態論情報が付与されていること（第 6 章）
- 文（sentence）タグが修正されていること（第 8 章）

ただし、これ以外にも M-XML ではいくつかの修正がなされている。以下、上記以外の C-XML と M-XML の相違点や追加されたタグについて述べる。

9.3.1 数字変換（NumTrans タグ）

第 6 章で述べたとおり、M-XML_NT においては、数字変換（NumTrans）処理がなされている。この処理が行われた箇所には、次のように NumTrans タグが付けられ、変換前の本文はこのタグの originalText 属性に保存されている。

OT テキスト：１９８６

NT テキスト：<NumTrans originalText="１９８６">千九百八十六</NumTrans>

実際には形態論情報が付与されているから、M-XML の当該箇所は以下のようになる（見やすさのため形態論情報の一部を省略した）。

- M-XML_OT：

```
<LUW l_lemma="一八九六" l_lForm="イチハチキュウロク">
  <SUW lemma="一" lForm="イチ">１</SUW>
  <SUW lemma="八" lForm="ハチ">８</SUW>
  <SUW lemma="九" lForm="キュウ">９</SUW>
  <SUW lemma="六" lForm="ロク">６</SUW>
</LUW>
```

- M-XML_NT：

```
<LUW l_lemma="千八百九十六" l_lForm="センハッピャクキュウジュウロク">
  <NumTrans originalText="１９８６">
    <SUW lemma="千" lForm="セン">千</SUW>
    <SUW lemma="八百" lForm="ハッピャク">八百</SUW>
    <SUW lemma="九十" lForm="キュウジュウ">九十</SUW>
    <SUW lemma="六" lForm="ロク">六</SUW>
  </NumTrans>
</LUW>
```

9.3.2 分数（fraction タグ）

C-XML では帯分数にのみ fraction タグが付与されているが、M-XML では帯分数以外の分数にも fraction タグが追加されている。

```
<fraction>１／１０</fraction>
```

M-XML_NT では、さらに NumTrans 処理によって分子（numerator）・括線（vinculum）・分母（denominator）が次のようにタグ付けされ、分子と分母の順が入れ替えられている。2 桁以上の数字の変換も合わせて行われている。

```
<fraction>
 <denominator><NumTrans originalText="１０">十</NumTrans></denominator>
  <vinculum><NumTrans originalText="／">分</NumTrans></vinculum>
  <numerator>１</numerator>
</fraction>
```

実際には形態論情報が付与されているから、M-XML の当該箇所は以下のようになる（見やすさのため形態論情報の一部を省略した）。

- M-XML_OT :

```
<fraction>
  <SUW lemma="一" lForm="イチ" pos="名詞-数詞">１</SUW>
  <SUW lemma="／" lForm="" pos="補助記号-一般">／</SUW>
  <SUW lemma="一" lForm="イチ" pos="名詞-数詞">１</SUW>
  <SUW lemma="零" lForm="レイ" pos="名詞-数詞">０</SUW>
</fraction>
```

- M-XML_NT :

```
<fraction>
  <denominator>
    <NumTrans originalText="１０"><SUW lemma="十" lForm="ジュウ">十</SUW></NumTrans>
  </denominator>
  <vinculum>
    <NumTrans originalText="／"><SUW lemma="分" lForm="ブン">分</SUW></NumTrans>
  </vinculum>
  <numerator>
    <SUW lemma="一" lForm="イチ">１</SUW>
  </numerator>
</fraction>
```

9.3.3 ルビの処理

形態論情報を付与する際、ルビのタグが形態論情報と齟齬を来す場合があるため、次のように対処した。

BCCWJ では、ふりがなは原則として 1 文字ごとに付与しているが、熟字訓や臨時的な読みでは複数文字を ruby タグで囲んでいる。たとえば次のような例がある。

1) 語彙(い)　　(短単位よりも短いルビ)
2) 時雨(しぐれ)　　(短単位と一致するルビ)
3) 喜望峰(ケープタウン)　　(短単位よりも長いルビ)
4) 新しい芸術(アール・ヌーヴォー)　　(長単位よりも長いルビ)

これらは C-XML では次のようにタグ付けされている。

1a) `<SUW>語<ruby rubyText="い">彙</ruby></SUW>`
2a) `<SUW><ruby rubyText="しぐれ">時雨</ruby></SUW>` もしくは
`<ruby rubyText="しぐれ"><SUW>時雨</SUW></ruby>`
3a) `<ruby rubyText="ケープタウン"><SUW>喜望</SUW><SUW>峰</SUW></ruby>`
4a) `<ruby rubyText="アール・ヌーヴォー"><SUW>新しい</SUW><SUW>芸術</SUW></ruby>`

M-XML では、1a) 2a)のように、短単位タグの内側に ruby タグを置くことができる場合

にはそのままとした。一方、3a) 4a)のように短単位を越えるルビについては、先頭の短単位を ruby タグで囲み、そのタグの属性値として本来のルビ範囲のテキストを保持することとした。これにより、元の状態に戻せるようにするとともに、複数単位に渡る特殊なルビを容易に取り出すことを可能にしている。

3a') <SUW><ruby rubyText="ケープタウン" rubyBase="喜望峰">喜望</ruby></SUW> <SUW>峰</SUW>

4a') <SUW><ruby rubyText="アール・ヌーヴォー" rubyBase="新しい芸術">新しい</ruby></SUW><SUW>芸術</SUW>

9.3.4 その他の追加されたタグ

改ページ位置を示す参考情報が空要素の info タグに残されている。

以上のように、できる限り互換性を保持するように努めているものの、各種の変更を加えているため、M-XML に付与されたタグと C-XML のタグとの間に完全な互換性はない。

参考文献

小木曽智信・間淵洋子・前川喜久雄（2011）「『現代日本語書き言葉均衡コーパス』における形態論情報付きXMLフォーマット」『言語処理学会第17回年次大会講演論文集』,352-355.

山田篤・小磯花絵 （2008） 『NumTrans マニュアル』, The UniDic Consortium.

索 引

B

BCCWJ 1
BCCWJ 構築の経緯 16
BCCWJ に収録するテキストの条件 30
BCCWJ の開発メンバー 18
BCCWJ の基本構成 21
BCCWJ の規模 22
BCCWJ の参考文献 14
BCCWJ の設計 20
BCCWJ-DepPara 157
BCCWJ-DVD 版 7
BCCWJ-DVD 版（Version 1.0） 1, 16, 55, 146
BCCWJ-DVD 版（Version 1.0）における文境界認定基準 147
BCCWJ-DVD 版（Version 1.1） 1, 8, 16, 55, 146
BCCWJ-DVD 版（Version 1.1）における文境界認定基準 149
BCCWJ-DVD版に収録されているサンプルの一覧 31
BCCWJ-DVD 版の意義 13
BCCWJ の規模 22

C

ChaSen legacy 3
Comainu 100
CSJ 61, 62, 64, 87
CSJ からの変更点 87
C-XML 8, 9, 101, 102
C-XML と M-XML の相違点 162
C-XML 形式 8
C コード 129

D

Disc 1 9
Disc 2 9
Disc 3 10
Disc 4 10
DTD 161

F

FIXED 9

I

IPADIC 3
ISBN
ISBN（書誌情報データ） 111, 117
ISBN 総合目録 35

K

KOTONOHA 計画 16

L

LUW 108, 162

M

MeCab 64
M-XML 10, 55, 100, 146, 160
M-XML_NT 9, 101, 160, 161, 163

M-XML_OT.....10, 101, 160, 161, 163
M-XML データ 1
M-XML における文境界認定..... 148
M-XML 形式..... 8

N

NumTrans.....1, 100, 101, 160, 162
NumTrans 版 8

R

RDB..... 13

S

sentence タグ 147, 162
SUW 108

T

TSV
TSV_LUW_NT..... 9
TSV_LUW_OT..... 10
TSV_NT 101
TSV_OT 101
TSV_SUW_NT..... 9
TSV_SUW_OT..... 10
TSV データ 1, 100
TSV 形式 8
TSV 形式データ..... 104

U

UniDic 64, 100
UTF-8..... 9, 104

V

VARIABLE 9
Version 1.1 における修正.....16

W

webLine タグ 52, 157
W 単位.....59

X

XML 形式.....8

Y

Yahoo!知恵袋タグセット.....46, 51

あ

アイヌ語
アイヌ語（語種情報） 86
圧縮ファイル.....11
アノテーション 3, 6, 157
α 単位.....59

い

一次結合.....78
一般
一般（最小単位認定規定） 78
一般（短単位認定規定） 78
韻文.....142

か

改行.....54
外字.....53

解析誤り 5
解析精度 65
階層構造 161
解凍 11
外来語
　外来語（語種情報） 85
　外来語（最小単位認定規定） 76
　外来語（短単位認定規定） 78
係り受けアノテーション 157
学年（教科書：ジャンル） 136
学校（教科書：ジャンル） 136
括弧類A（文境界修正規則） 147
活用型
　活用型（短単位） 84
　活用型（長単位） 74
活用形
　活用形（短単位） 84
　活用形（長単位） 74
可能動詞
　可能動詞（短単位認定規定） 81
可変長 24
　可変長サンプル 30, 160
　可変長タグセット 46, 47
漢語
　漢語（語種情報） 85
　漢語（最小単位認定規定） 76
　漢語（短単位認定規定） 78
巻号（書誌情報データ） 111, 115

き

記号
　記号（語種情報） 86
　記号（最小単位認定規定） 76, 78
　記号（短単位認定規定） 80
　記号（長単位認定規定） 69
　記号類C（文境界修正規則） 151
　記号類D（文境界修正規則） 152
　記号類E（文境界修正規則） 152
　記号類F（文境界修正規則） 153
記号類 ..... → 記号
記事 ID（記事情報データ） 126
記事情報データ 110, 126
教科（教科書：ジャンル） 136
均衡コーパス 1

く

空白 67
　空白（長単位認定規則） 69
句点類B（文境界修正規則） 150
句読点
　句読点（長単位認定規則） 69
　句読点（文節境界規則） 67

け

敬語表現 68
形状詞
　形状詞（用法） 87
形態素解析 3
形態論情報 3, 100
　形態論情報（文境界認定基準） 146
形態論情報付き統合形式 XML 160
形態論情報付き統合形式 XML データ ..100, 107
言語単位 60, 61
検索ツール 13
現代語 7
原文文字列 101, 106

こ

コアデータ 9, 23, 65, 157

語彙素 4, 64, 73
語彙素（短単位） 83
語彙素読み 73
語彙素読み（短単位） 83
誤解析 6
語形 64
語種 85
誤植 55
語数 22, 103
国会会議録 144
固定長 24
固定長サンプル 30, 160
固定長タグセット 46, 51
固定長と可変長の統合 160
異なる文書型定義の統合 161
固有名
固有名（語種情報） 86
固有名（最小単位認定規定） 78
固有名（短単位認定規定） 80
混種語
混種語（語種情報） 86

さ

最小単位 75
最小単位（CSJ からの変更点） 88
最小単位認定規定 75
サブコーパス 2, 29
サンプリング 28
サンプリング方法 31
サンプリング方法（韻文） 42
サンプリング方法（教科書） 37
サンプリング方法（広報紙） 38
サンプリング方法（国会会議録） 44
サンプリング方法（雑誌） 33
サンプリング方法（書籍） 32, 35
サンプリング方法（新聞） 34
サンプリング方法（白書） 36
サンプリング方法（ベストセラー） 39
サンプリング方法（法律） 42
サンプリング方法（Yahoo!知恵袋） 40
サンプリング方法（Yahoo!ブログ） 41
サンプル ID 119
サンプル ID ベース書誌情報データ 145
サンプル ID（記事情報データ） 126
サンプル情報データ 110, 118
サンプル数 22
サンプル抽出基準点座標（サンプル情報データ） 124
サンプル抽出基準点ページ（サンプル情報データ） 124
サンプル長 8
サンプルに関するタグ（文書構造タグ） 47

し

自治体（広報紙：ジャンル） 136
ジャンル
ジャンル(1)（書誌情報データ） 111, 117
ジャンル(2)（書誌情報データ） 111, 117
ジャンル(3)（書誌情報データ） 111, 117
ジャンル(4)（書誌情報データ） 111, 117
ジャンル（書誌情報データ「ジャンル」情報の詳細） 129
主語
主語（長単位認定規定） 70
主語（文節認定規定） 68
主題
主題（長単位認定規定） 70
主題（文節認定規定） 68
出現形開始位置 105
出版サブコーパス

出版 SC「雑誌」 33
出版 SC「書籍」 32
出版 SC「新聞」 34
出版サブコーパス 2, 22
出版社（書誌情報データ） 111, 116
出版年（書誌情報データ） 111, 116
小カテゴリ
小カテゴリ（Yahoo!知恵袋：ジャンル） . 138
小カテゴリ（Yahoo!ブログ：ジャンル） . 140
小ジャンル（書誌情報データ） 131
初刊情報（記事情報データ） 128
助詞・助動詞
助詞・助動詞（最小単位認定規定） 78
助詞・助動詞（短単位認定規定） 81
書誌 ID 113, 124
書誌 ID（書誌情報データ） 111
書字形 64
書字形出現形 101, 106
書誌情報 6, 9, 100, 110
書誌情報データ 9, 110
書誌情報データ「ジャンル」情報の詳細 . 129
書誌情報データベース 110
初出情報（記事情報データ） 127
助数詞
助数詞（用法） 87
人名
人名（最小単位認定規定） 77
人名（人名録データ） 125
人名 ID 125
人名 ID（記事情報データ） 127
人名録データ 110, 125

す

数
数（最小単位認定規定） 76, 78
数（短単位認定規定） 80
数を表す要素（長単位認定規定） 71
数を表す要素（文節認定規定） 68
数字変換処理 100, 101

せ

正規表現 3
生年代
生年代（人名録データ） 125
性別
性別（人名録データ） 125
責任表示（書誌情報データ） 111, 116
責任表示 ID（書誌情報データ） 111, 117
接頭的要素 95
接尾的要素 96
全国地方公共団体コード（広報紙：ジャンル）
..... 137

そ

層別方法
層別方法（韻文） 41
層別方法（教科書） 37
層別方法（広報紙） 38
層別方法（国会会議録） 43
層別方法（雑誌） 33
層別方法（書籍） 32, 35
層別方法（新聞） 34
層別方法（白書） 36
層別方法（ベストセラー） 39
層別方法（法律） 42
層別方法（Yahoo!知恵袋） 39
層別方法（Yahoo!ブログ） 41
その他のタグセット（文書構造タグ） 51

た

大カテゴリ
　大カテゴリ（Yahoo!知恵袋：ジャンル）. 138
　大カテゴリ（Yahoo!ブログ：ジャンル）. 140
大ジャンル（書誌情報データ）..... 131
タイトル（書誌情報データ）.....111, 115
代表性..... 20
短単位.....3, 62, 63, 75, 100, 160
　短単位（CSJ からの変更点）..... 88
　短単位 TSV のフィールド..... 104
　短単位タグの属性..... 108
　短単位認定規定.....78
　短単位の長所..... 63

ち

地名（最小単位認定規定）..... 77
中カテゴリ
　中カテゴリ（Yahoo!知恵袋：ジャンル）. 138
　中カテゴリ（Yahoo!ブログ：ジャンル）. 140
中ジャンル（書誌情報データ）..... 131
中納言..... 1, 9
調査単位..... 58
長単位.....4, 62, 63, 66, 69, 100, 160
　長単位（CSJ からの変更点）..... 87
　長単位 TSV のフィールド..... 105
　長単位タグの属性..... 108
　長単位の長所..... 62
著作権処理..... 7
著作権注釈情報データ..... 9

て

ディレクトリ構成..... 9
電子化..... 25

と

同格
　同格（長単位認定規定）.....70
　同格（文節認定規定）.....66
投稿日時（サンプル情報データ）.....124
特殊表記.....54
特定目的サブコーパス.....2, 23
　特定目的 SC「韻文」.....41
　特定目的 SC「教科書」.....37
　特定目的 SC「広報紙」.....38
　特定目的 SC「国会会議録」.....43
　特定目的 SC「白書」.....36
　特定目的 SC「ベストセラー」.....38
　特定目的 SC「法律」.....42
　特定目的 SC「Yahoo!知恵袋」.....39
　特定目的 SC「Yahoo!ブログ」.....40
特定領域研究「日本語コーパス」.....16
図書館サブコーパス.....2, 23
　図書館 SC「書籍」.....35
図書分類コード.....129

な

長い単位.....59, 62

に

日本語話し言葉コーパス.....16, 61

は

判型（書誌情報データ）..... 111, 117

ひ

非 NumTrans 版.....8, 10, 17

非コアデータ 9, 65
品詞
　品詞（CSJ からの変更点） 88
　品詞（短単位） 84
　品詞（長単位） 74

ふ

複合語 4
複合辞 67, 70
　複合辞（助詞相当句） 92
　複合辞（助動詞相当句） 93
副詞
　副詞（用法） 87
副題（書誌情報データ） 111, 115
付属語
　付属語（長単位認定規定） 70
　付属語（文節認定規定） 67
　付属要素（最小単位認定規定） 78
　付属要素（短単位認定規定） 81
プレイン・テキスト 3
文境界認定基準 146
文書型定義 161
文書構造タグ 9, 46
文書構造に関するタグ（文書構造タグ） 48
文節 66
　文節（CSJ からの変更点） 87
　文節認定規定 66
文タグ ..... → sentence タグ

へ

並列
　並列（長単位認定規定） 70
　並列（文節認定規定） 66
ページ数（書誌情報データ） 111, 117
β 単位 59, 62

ほ

包摂規準 53
法律 142
母集団の定義
　母集団の定義（韻文） 41
　母集団の定義（教科書） 37
　母集団の定義（広報紙） 38
　母集団の定義（国会会議録） 43
　母集団の定義（雑誌） 33
　母集団の定義（書籍） 32, 35
　母集団の定義（新聞） 34
　母集団の定義（白書） 36
　母集団の定義（ベストセラー） 38
　母集団の定義（法律） 42
　母集団の定義（Yahoo!知恵袋） 39
　母集団の定義（Yahoo!ブログ） 40

め

名詞
　名詞（用法） 86

も

文字・表記に関するタグ（文書構造タグ） 47
文字開始位置 105
文字集合 53
文字入力 25, 46, 52
文字符号化方式 9, 53, 104, 160

や

役割（記事情報データ） 127

よ

用法..... 63, 86

れ

レイアウト..... 54
レジスター..... 2, 7, 9, 22
連語..... 94
連番..... 105

ろ

論理行..... 54, 146, 157

わ

和語
　和語（語種情報）..... 85
　和語（最小単位認定規定）..... 76
　和語（短単位認定規定）..... 78
和製英語
　和製英語（語種情報）..... 85