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。

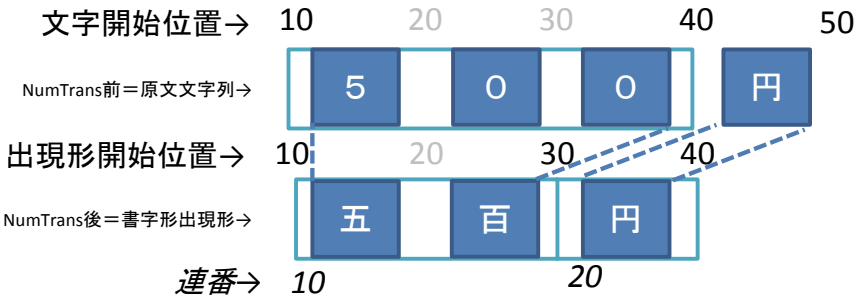

図 6-2: NumTrans されたテキストの文字位置と連番の対応

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.