BCCWJ 係り受け関係アノテーション付与のための文境界再認定

小西 光（国立国語研究所コーパス開発センター）†
小山田 由紀（国立国語研究所コーパス開発センター）
浅原 正幸（国立国語研究所コーパス開発センター）
柏野 和佳子（国立国語研究所言語資源研究系）
前川 喜久雄（国立国語研究所言語資源研究系／コーパス開発センター）

Revision of Sentence Boundaries in BCCWJ for Syntactic Dependency Structure Annotation

Hikari Konishi (Center for Corpus Development, NINJAL)
Yuki Oyamada (Center for Corpus Development, NINJAL)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Wakako Kashino (Dept.Corpus Studies, NINJAL)
Kikuo Maekawa (Dept.Corpus Studies/Center for Corpus Development, NINJAL)

１．はじめに

『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）では，文を機械的に自動認定し，コアデータのみ人手による文境界の修正を行っている．このコアデータの文境界情報を元に係り受け関係アノテーションを付与しようとすると，「係り先のない文」が出現する．これは，文書を電子化した際に認定したレイアウトに基づく階層構造[1]の影響であったり，係り受け関係にあると判断されるものが自動文認定の際に一文とされなかったりしたことに由来する．そこで，BCCWJ のコアデータに関して「係り受け関係アノテーション付与」を目的とした「文」の再認定を行うこととした．

２．自動文認定—sentence 要素—

sentence 要素は，「文に相当するまとまりを表す要素」として機械的に自動認定されている．BCCWJ の電子化フォーマットでは，自動認定は以下のように行われる．

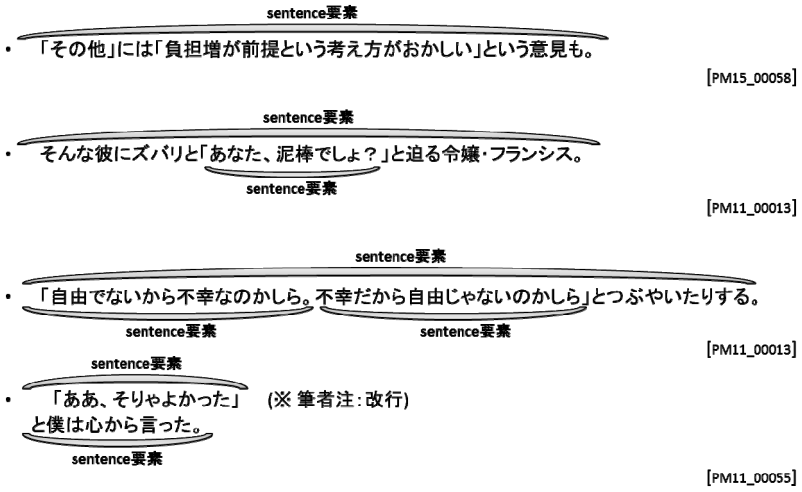

図 1 sentence 要素の自動認定

† hkonishi@ninjal.ac.jp

1 山口ほか(2011)によると上位から article> cluster/titleBlock>paragraph>sentence という階層構造を持つ

自動認定により sentence 要素冒頭と判断された箇所には XML 形式で<sentence>タグを，末尾と判断された箇所には</sentence>タグを挿入する．

現在公開されている BCCWJ（DVD 版）の電子化フォーマットのうち，XML で構成されるものとして文字ベース XML(C-XML)と形態論情報付きの統合形式 XML（M-XML）の二種類がある．C-XML は sentence 要素の入れ子構造を認めるが， M-XML では sentence 要素の入れ子構造を認めず，C-XML で入れ子構造の外側の sentence 要素を<superSentence>としている[2]（図２・上）．また C-XML では，<sentence>タグに属性“quasi”（文区切り文字以外の基準により自動付与された sentence 要素）と“verse”（韻文内の sentence 要素）を付与しており，M-XML では，その二属性に加えて入れ子構造外側の sentence 要素に対して“fragment”という属性を新たに導入している．

M-XML

```
<superSentence>
  <sentence type="fragment">　声明は、同大統領の法案署名へ歓迎と感謝を表明し</sentence>
   <quote>
     <sentence>「米国の支持は、台湾の（ＷＨＯ参加への）努力が既に友邦の理解を得たことを意味する。</sentence>
      <sentence type="quasi">今後、全力を挙げて国際社会の全面的な賛同を得られるよう努力する」</sentence>
   </quote>
  <sentence type="fragment">と述べている。</sentence>
</superSentence>
<br type="automatic_original"/>
```

C-XML

```
<paragraph>
  <sentence>　声明は、同大統領の法案署名へ歓迎と感謝を表明し
    <quote>
       「<sentence>米国の支持は、台湾の（ＷＨＯ参加への）努力が既に友邦の理解を得たことを意味する。</sentence>
        <sentence type="quasi">今後、全力を挙げて国際社会の全面的な賛同を得られるよう努力する</sentence>
       」
    </quote>
       と述べている。
  </sentence>
<br type="automatic_original" />
</paragraph>
```

図２ M-XML(上)と C-XML(下)の比較（PN4g_00001）

２.１　sentence 要素の現状と問題点

係り受け関係アノテーション付与を目的とした場合，BCCWJ の sentence 要素における問題点は，大きく以下の二つに分けられる．

① 文境界と判断されるべき箇所に<sentence>タグが付与されていない．
② 文境界と判断されるべきでない箇所に<sentence>タグが付与されている．

まず①について，前述のとおり<sentence>タグはほぼ自動付与である．そのため，現段階で本来複数の文とされるべき発話や引用・補足などが認定基準が原因でひとつの sentence 要素となっており，分割されていない場合がある．例えば「<sentence>「受験勉強に明け暮れて、東大に入って、官僚になってもちっとも幸福じゃない」最近では、そんなセリフを大人も子供も口にします。</sentence>」(PB33_00032)のように自動認定の基準から外れており，機能的には括弧・引用符と同様に用いられているのだが sentence 要素とされていないような場合である．ただし，これらは人手修正のチェック漏れということも考えられる．

2 小木曽ほか(2011)「上位の文は superSentence として文書構造タグの一種とした。下位の sentence はそのまま残し、superSentence の一部分を新たに sentence で囲み type=”fragment”とした」(p.39)

次に②であるが，これには二つの問題がからんでいる．一つは資料原本のレイアウト情報を元に認定した文書階層構造により発生している問題であり，もう一つは自動及び人手修正による sentence 要素の認定基準と，係り受け関係アノテーションを目的とした「文」の認定基準とが異なるという問題である．

一つ目の問題は，例えば原本が図３のように会話文に入る直前の地の文で改行（<br type=”automatic_original”/>）されるようなレイアウトだと，sentence 要素より上位の文書構造である<paragraph>タグ[3]や<quotation>タグ[4]に阻まれ，文が続いているにもかかわらず一つの文としては認定されていない．

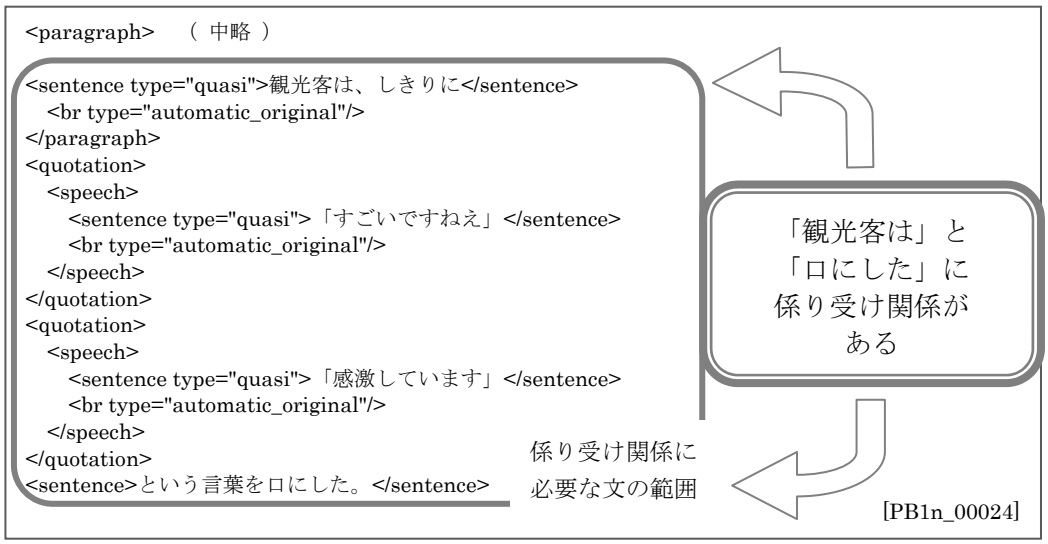

図３ 階層構造により複数の sentence 要素となった例(M-XML)

図３では「観光客は、しきりに」という sentence 要素と「という言葉を口にした。」という sentence 要素が認定されており，「観光客は」の本来の係り先「（口に）した」は「文」を越えて存在することになる．文を越えた係り受け関係は付与できないため，「観光客は」の係り先は不明となり，正しい係り受け関係アノテーションの付与ができないこととなる．

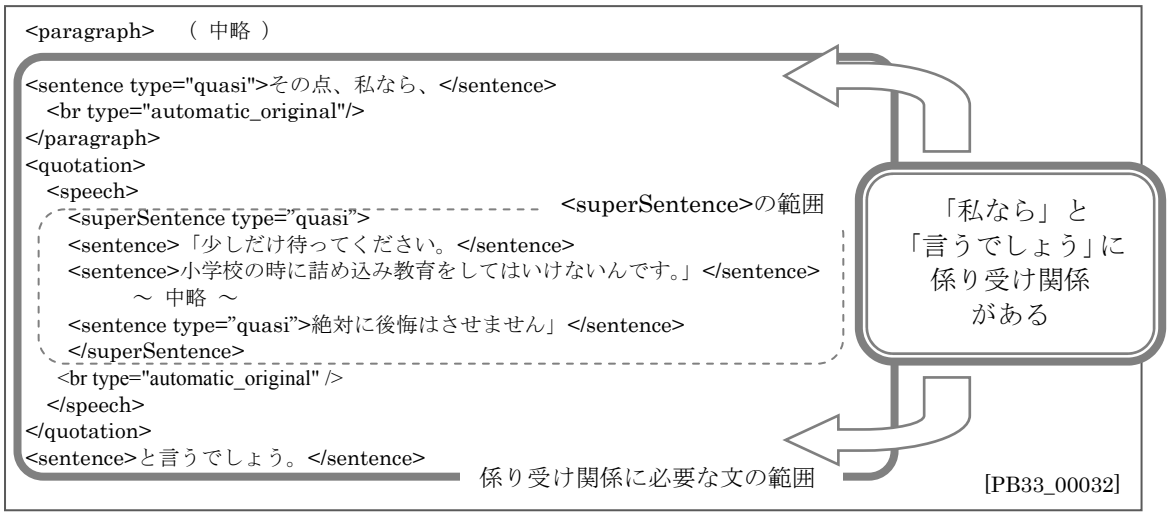

図４ <superSentence>タグの範囲（M-XML）

[3] 段落を表す文書構造要素．原則として，一字下げで始まる．sentence 要素よりも上位階層となり，sentence 要素が上位階層の要素をまたぐことはない．

[4] 山口ほか(2011)「当該記事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・書き起こしを表す」

二つ目の問題は，今回作業対象としたコアデータについては人手による修正が行われているものの，それらは「係り受け関係を付与する」という基準で作業されていないため，係り受け関係アノテーション付与を目的とした文認定が再度必要となる．M-XML に付与された<superSentence>タグを利用した文の認定も可能だが，図４のような過不足のある範囲となっている場合もあるため，自動的に抽出することは難しい．

またこれとは別に，C-XML から M-XML に変換する際に図５のような問題も生じている．

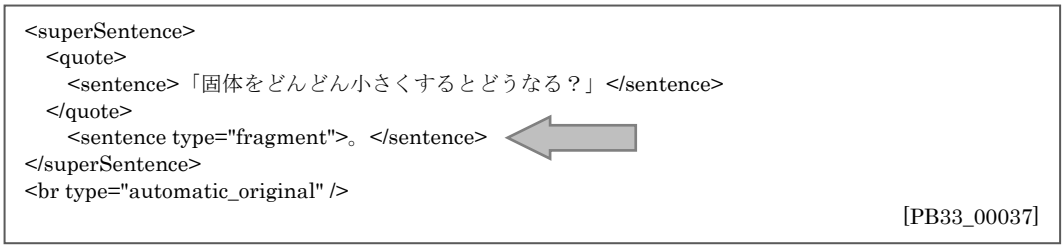

図５　句点「。」のみで 1 文と認定されている例（M-XML）

以上のことからも，係り受け関係アノテーションを付与する場合，現状の「文」境界の認定では問題があるため，今回は係り受け関係アノテーション付与に影響の大きい②の「文境界と判断されるべきでない箇所に<sentence>タグが付与されている」sentence 要素についてのみ文境界の再認定作業を行った．なお浅原(2013)によると，①は係り受け関係アノテーション作業時に「文境界」を表現する係り受け関係ラベル（“Z”ラベル）を導入している．

３．文境界再認定作業

３.１　作業対象

BCCWJ のコアデータ全 60,374 文を対象とする．XML データの<sentence>タグや<superSentence>タグを修正するのではなく，XML データの sentence 要素を参考にした係り受け関係アノテーション用の文を別途認定する．

３.２　認定基準

まず前提として以下の二点を示す．

- 係り受け関係アノテーション付与を目的としたもっとも長い単位としての「文」の認定を行う
- 現在 XML 等に付与されている改行やタグ情報（<sentence>タグ・<paragraph>タグ等）には縛られない

３.２.１　一文と認定するもの

現状の sentence 要素では係り受け関係アノテーション付与に問題があり，以下の三点のいずれかを満たすものを「文」と再認定する．

① 括弧や引用符などの括り記号で括られた発話や引用・補足部分を挟んだり，引用の助詞「と」で受けたりして係り受け関係を結べる要素が前・中・後に接続する
② 箇条書き（改行を伴う）を内包する要素が前・中・後に接続する（主にウェブ媒体）
③ 本来一文であるべきものが，書き手による意図的な改行で分割されている（主にウェブ媒体）

３.２に記したとおり係り受け関係がもっとも長い単位としての文境界越えないことを基準とするが，例えば「「掛け給え」
と部長は言った。」や「手でひたいをおさえて、

「なにをいっているんだ、わたしは？」」のように括られた要素に対して後ろのみ，前のみに接続する場合がある．この場合は「「掛け給え」と部長は言った。」「手でひたいをおさえて、「なにをいっているんだ、わたしは？」」という文を認定した．接続詞のみの場合や助詞「と」だけで括られた要素を受ける場合も同様に処理する．

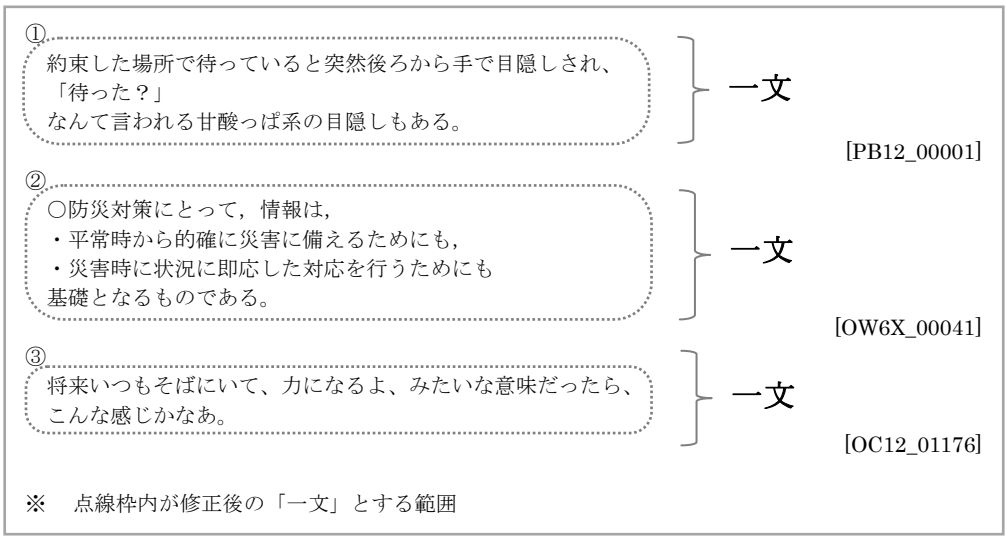

図６　一文と認定するもの

３.２.２　一文と認定しないもの

以下の場合は，現状のままひとつの文にまとめ上げることはしない．

① 倒置部分が改行されている
② 改行を伴って文がねじれている
③ 接続助詞ではなく接続詞「と」「っと」と判断されるものが文頭にくる
④ 前後の sentence 要素と括弧や引用符などで括られた要素がそれぞれ独立して係り受け関係にない

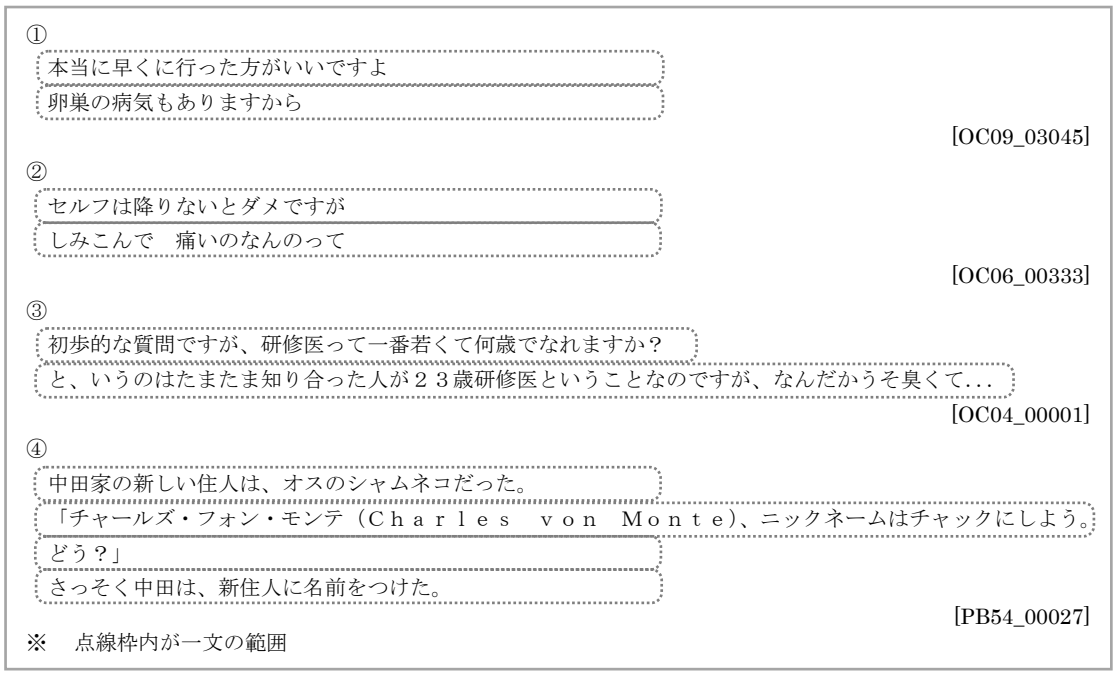

図７　一文と認定しないもの

３.３　作業手順

M-XML を元に sentence 要素, <superSentence>タグの範囲情報, <sentence>/<superSentence>タグの type 情報（fragment，quasi，verse），sentence 要素の冒頭と末尾の品詞情報を抽出し，それらを参考にして作業を行った．

例えば，sentence 要素が助詞や括弧閉，読点・カンマで始まっている場合は，一つ前（もしくはそれ以前から）の sentence 要素を受けていると考えられる．また sentence 要素が読点やカンマ，助詞，括弧開で終わっている場合は，「文」の途中で分割されており，係り先となるはずの文要素がそれ以降に後続していると考えられる．このように手がかりを見つけ次第，その前後の sentence 要素を確認して「文」の範囲を認定した．

４．作業結果

４.１　再認定結果

表１に作業結果をまとめた．

コアデータ全文 60,374 文に対する修正箇所の割合は，4,585 文と約 7.6%である．この 4,585 文を係り受け関係アノテーション用の「文」に修正すると 1,385 文となる．これは修正前の約三文が一つの文にまとまるという割合になる（まとめ上げ「文」数）．

修正箇所全体の約 66%は，「私は「～。」と言った。」のように括弧や引用符等で括られた要素を前と後ろに挟んで係り受け関係を結べるもの（「前後型」とする）である．また約 25%は，「“～。”と言った。」のように括られた要素を後ろのみで受けるもの（「後型」とする）である（表２[5]）．

レジスター別では，Yahoo!ブログ（OY）と新聞（PN）の修正する割合が高い．各レジスターのまとめあげ「文」数を見てみると，それぞれの特徴の一端を示している．

表１　文境界再認定の結果

レジスター	<sentence>タグ数＝「文」数	修正箇所「文」数（チェック率）＝A	修正後「文」数（修正対象のみ）＝B	まとめ上げ「文」数（A/B）
OW（白書）	6,067	385(6.3%)	113	3.41
PB（出版書籍）	10,095	592(5.9%)	166	3.57
PN（新聞）	17,136	**1,530(8.9%)**	415	**3.69**
OC（Y!知恵袋）	6,435	514(8.0%)	161	3.19
OY（Y!ブログ）	7,651	**915(12.0%)**	332	**2.76**
PM（雑誌）	12,990	649(5.0%)	198	3.28
合計	**60,374**	**4,585(7.6%)**	**1,385**	**3.31**

括弧内が複数文

文途中での改行頻度が高い

接続詞・接続助詞で改行

新聞は，約 3.69 文が一つの文にまとめ上げられているのに対し，Yahoo!ブログは約 2.76 文が一つの文にまとめ上げられている．これは，新聞が括り記号内に文区切り文字で区切られる sentence 要素が複数含まれ，再認定作業後の「文」が長文化するのに対し，Yahoo!ブログは，図８のようにブログ執筆者によって接続詞や接続助詞の後ろなど文の途中で改行されている例が修正箇所全体の約三分の一（100 例）と多くを占め，それら

表２　修正箇所の構造（単位：文）

レジスター	前後型	後型	合計
OW	74	10	84
PB	85	47	132
PN	**254**	**75**	329
OC	83	36	119
OY	80	31	111
PM	96	**54**	150
合計	672	253	**925**

[5] 前後型・後型以外にも「前中後型」「前中型」「中後型」「前型」があるが，それらは数が少ないためここでは省いた．

文の断片を結びつけるための単純な再認定である場合が多い.

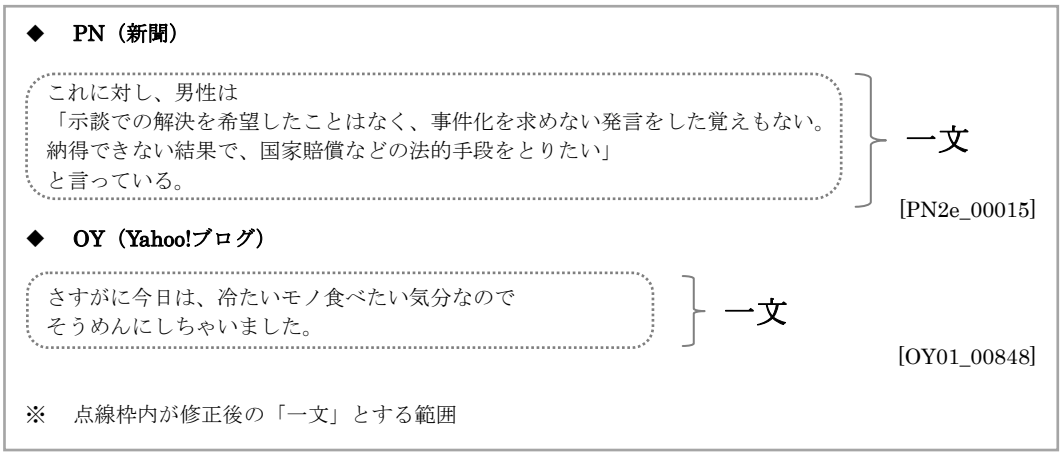

図８　新聞と Yahoo!ブログの比較

４.２　括弧・引用符等の機能別分類

４.１の作業結果をもとに括弧・引用符等の機能別に以下の分類を試みたので，レジスター別の特徴を示す.

- 補足 ： 語や文を補う目的で用いる（主に（ ））
  補足部分がなくても文が成立する
- 発話 ： 「と言う」等で受ける発話
- 心内 ： 「と思う」等で受ける心内語
- 引用 ： 上記以外のもの
- 箇条書き ： 行頭の中点等記号および改行によって複数の項目を列挙したもの
- 強調 ： 主に括弧を用いて他の文字列よりも強調するために用いる
  （書籍名やタイトル等も含む）

表３ 機能別分類（単位：文）

レジスター	補足	発話	心内	引用	箇条書き	強調	合計
OW	40	4	0	23	15	2	84
PB	1	107	4	26	7	1	146
PN	0	307	4	42	0	5	358
OC	12	26	4	62	23	5	132
OY	16	32	24	46	2	8	128
PM	0	108	9	46	3	6	168
合計	69	585	45	240	50	27	1016

表３を見ると，レジスターごとに特徴が表れている.

白書(OW)は，丸括弧による補足と箇条書きが多用されている.

新聞(PN)は，括られた要素の前後で係り受け関係を結べるような発話が多用される．これは，文頭に発話者の情報や状況が来て，続いて引用部分を挟み，引用の「と」等でそれを

受けて係り受け関係を結ぶというある種の「文型」が決まっていると考えられる[6]．また書籍(PB)や雑誌(PM)でも発話が多用されている．

Yahoo!知恵袋(OC)は，Q&A 形式の特徴（答える際に文献からの引用や列挙を用いる）が引用や箇条書きの多用に表れている．

Yahoo!ブログ(OY)は，他のレジスターより心内語が多用され，ブログ執筆者の心情を表わす傾向をとらえている．

５．まとめ

係り受け関係アノテーション付与を目的とした文境界の再認定作業について報告を行った．修正を必要とする 4,585 文（7.6%）のみではあるが，各媒体の特徴の一部が明らかになった．またランダムサンプリングではないデータではあるが，文を単位とした括弧・引用符等の機能別でのアノテーションもレジスター分析に有効な指標を設定するための予備調査に位置づけることができるだろう．

今回の報告により文を自動で認定する困難さが具体的なものとなり，また文分析の可能性の一端を示すことができた．今後はより精度の高い自動文認定解析の確立を待ちつつ，係り受け関係アノテーション付与の研究に着目していきたい．

謝　辞

本研究を行うにあたり，助言いただきました丸山岳彦氏に感謝いたします．また本研究は，国立国語研究所基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国立国語研究所「超大規模コーパス構築プロジェクト」による補助を得ています。

文　献

小木曽智信、間淵洋子、前川喜久雄 (2011)「階層的形態論情報を考慮した『現代日本語書き言葉均衡コーパス』の公開用 XML フォーマット」『現代日本語書き言葉均衡コーパス』完成記念講演会 予稿集，pp.35-42，JC-G-11-01

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『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

浅原正幸 (2013)「係り受けアノテーション基準の比較」本予稿集

6 会話×前後型で 375 例あった．これが全体に占める割合は 36.9%である．