BCCWJに対する節境界ラベルのアノテーション

丸山 岳彦

国立国語研究所 言語資源研究系

1　はじめに

本稿では、日本語コーパスに対する言語学的情報のアノテーションとして、「節境界ラベル」の付与を取り上げる。『現代日本語書き言葉均衡コーパス』(BCCWJ)の一部に対して節境界解析を実施し、自動的に節境界ラベルを付与した実験の結果について報告する。

節の終端境界を自動的に検出し、その形態的・文法的特性を表すラベル(「節境界ラベル」)を付与する処理を、節境界解析と呼ぶことにする。2004年に公開された『日本語話し言葉コーパス』(CSJ)では、アノテーション作業の一環として節境界ラベルの自動付与が実施された[1, 2]。これを引き継ぐ形で、現在、BCCWJの短単位データベースに対して節境界ラベルを自動的に付与する実験を行なっている。そこで本稿では、節境界解析の設計と付与結果を示し、解析誤りの事例を検討する。さらに、付与された節境界ラベルをもとに、書き言葉と話し言葉の比較・対照を行なう。

2　節境界解析とは何か

文節(または句)より大きく、文よりも小さな統語的単位として、「節(clause)」の存在が挙げられる。あるテキスト中に現れる節の終端境界について、その位置と形態・統語的なタイプを特定する処理を、ここでは「節境界解析」と呼ぶことにする[1]。

一例として、新聞記事テキストに現れた節境界の位置と種類を示すと、図1のようになる。「|」は節境界の位置(その直後で原文を改行している)、太字は当該の節境界の形態・統語的なタイプを表わす。

県警日光署によると|**連用節(条件節ト)**
男性が１人で歩いていたところ|**連用節(時間節)**
横からいきなりクマが飛びかってきたという。|**文末**
クマは男性を襲った後|**連用節(時間節)**
そのまま逃げ|**連用節(連用形接続)**
捕まっていない。|**文末**
男性は通りかかった|**連体節(内の関係)**
夫婦に助けを求め|**連用節(連用形接続)**
女性が１１９番通報した。|**文末**

図1: 節境界の位置と形態・統語的なタイプ

節境界ラベルのアノテーションは、言語学的な調査・研究(特に文法研究や文体・変異研究)に利用したり、言語処理研究のための基礎データとして利用したりするなど、幅広い応用分野を想定することができる。

言語学的な調査・研究について言えば、あるテキストに現れる従属節の種類と数を集計したり、類似した接続形式の出現傾向を文体ごとに調査したりする際に、節境界解析の結果が利用できる。例えば、接続助詞「が」と「け(れ)ど(も)」という5つの接続形式や、「れば」「たら」「なら」「と」という4つの条件形式が、文体ごとに異なる出現傾向を示すか否かを検討する際、節境界ラベルは分析の基礎データとなる(これらの具体的な分析については、5節で示す)。

言語処理への応用について言えば、例えば、長文の発話分割処理に節境界ラベルが利用できる。CSJの構築過程においては、形態素解析された転記テキストに対して節境界解析を実施し、付与された節境界ラベルを基準として発話分割処理が行なわれた[2]。また、Ohno et.al. (2007)は、1文が長くなりがちな独話を節境界で分割し、局所的な統語解析を実施することにより、解析時間が大幅に短縮されることを示した[4]。

このように、節境界ラベルのアノテーションは、言語学・言語処理の研究のさまざまな側面にとって有用な情報を提供するものと言える。

3　CSJで実施された節境界解析

以下では、CSJで実施された節境界解析の概要を示す。CSJに収録されている自発音声の大半は、基本的にデスマス体の丁寧な発話スタイルを取る独話である。ただし、句点類は書き起こされていない。このようなデータに対して係り受け構造、談話構造などの情報付与を考える際、共通した統語的単位が必要となり、「節単位」という統語的単位が設計された[2]。その際、CBAP-csjと呼ぶルールセットを用いた節境界解析が実施され、アノテーションされた節境界ラベルが発話分割位置の候補として用いられた。CSJに付与された節境界ラベルの例を、図2に示す。太字の部分が、付与された節境界ラベルである。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

ただし <**接続詞**> (F えー) こういう生活をしてますと **/条件節 ト/** 摂取カロリーが大変多くなりまして **/テ節/** 本来だったら <**条件節タラ**> 私の年では一日千五百キロカロリー取れば <**条件節レバ**> 十分なんですが **/並列節ガ/** (F まー) 大体平均すると <**条件節 ト**> 二千五百カロリーぐらいは取ってるんじゃないかなと <**引用節**> そういう気がいたします **[文末]** (F ま) そういうことで <**並列節デ**> (F えー) 結構年のわりに歩いてるんですが **/並列節ガ/** . . .

図 2: CSJ に付与された節境界ラベルの例

```
IF EXISTS ( SELECT * FROM PB
 WHERE sample_ID = @sample_ID AND ID = @ID AND OT = ’けど’ AND pos = ’助詞-接続助詞’ )
AND EXISTS ( SELECT * FROM PB
 WHERE sample_ID = @sample_ID AND ID = @ID +1 AND NOT ( lemma = ’も’ AND pos = ’助詞-係助詞’ ))
BEGIN SET @RetVal = ’並列節_ケド’ END
```

図 3: 節境界解析のための規則の例

CSJ の節境界ラベルは、形態素解析の結果を読み込んで、人手で用意した規則とのパタンマッチによって自動的に付与される。292 組の検出規則によって、49 種類の節境界ラベルを付与する。実装は Perl で行なった。

CSJ に付与された節境界ラベルは、その直後の統語的な切れ目の大きさという点から、3 つのクラスに分類してある。[　] で囲まれたラベルは絶対境界と呼ばれ、文末表現に相当する。直後は完全な統語的な切れ目となる。/　/ は強境界と呼ばれ、南 (1974) の分類 [3] における「C 類の従属句」にほぼ相当する、直後の統語的な切れ目が大きい節境界である。通常、発話の分割位置となる。<　> は弱境界と呼ばれ、南 (1974) の「B 類の従属句」にほぼ相当する、直後の統語的な切れ目が小さい節境界である。一定の条件を満たした場合に、発話の分割位置になり得る。

今回、話し言葉コーパスである CSJ に対して付与された節境界ラベルを、その種類を拡張する形で、書き言葉コーパスである BCCWJ に対しても付与することにした。次節では、BCCWJ に対する節境界ラベルの自動付与実験と、その結果について述べる。

4　BCCWJ に対する節境界解析

4.1　データ

BCCWJ は、現代日本語のさまざまな書き言葉を収録した、1 億語超の均衡コーパスである。このうち、形態素解析の結果を人手で修正した 100 万語分の「コアデータ」がサブセットとして設定され、集中的にアノテーションを実施する対象として扱われている。

今回、節境界解析の試行として、比較的整った文体のテキストが含まれる書籍・雑誌・新聞のコアデータを対象とした。対象データの語数（短単位数）は、書籍が 234,400 語、雑誌が 239,440 語、新聞が 360,526 語となっている。

4.2　節境界ラベルの付与規則

形態素解析用辞書 UniDic と形態素解析器 MeCab による形態素解析（UniDic-mecab 1.3.12）の結果を格納した短単位データベースに対して、節境界解析を実施する。ここで用いる短単位データベースは、SQL Server 上に展開されている。このうち、「短単位 ID、文頭ラベル、書字形出現形、語彙素、品詞、活用形、サンプル ID」という 7 組の情報を用いる。例を図 4 に示す。

ID	BND	OT	lemma	pos	cForm	sample_ID
265	B	則ち	即ち	接続詞		PB10_00047
266	I	彼	彼	代名詞		PB10_00047
267	I	が	が	助詞-格助詞		PB10_00047
268	I	まっ先	真っ先	名詞-普通名詞-一般		PB10_00047
269	I	に	に	助詞-格助詞		PB10_00047
270	I	引返し	引き返す	動詞-一般	連用形-一般	PB10_00047
271	I	て	て	助詞-接続助詞		PB10_00047
272	I	ゆく	行く	動詞-非自立可能	連体形-一般	PB10_00047

図 4: SQL Server に格納された短単位データベース

UniDic に基づく形態素解析の結果に対して、節境界解析を実施するための規則を、テーブル値関数として準備した。これを CBAP-UniDic と呼ぶ。CBAP-UniDic に記述された規則の一例を、図 3 に示す。

図 3 に引用したのは、書字形出現形が「けど」でかつ品詞が「助詞-接続助詞」である短単位に、語彙素が「も」でかつ品詞が「助詞-係助詞」である短単位が後接しない場合に、前者の短単位「けど」のレコードに「並列節_ケド」という節境界ラベルを付与するための規則である。4 行目の NOT を削除すれば、「並列節_ケドモ」というラベルを付与する別の規則になる。

現時点における開発版では、人手で記述された 163 組の規則によって、80 種類の節境界ラベルを付与するようになっている。

4.3　結果

対象データに対して、節境界解析を実施した。結果を表 1 に示す。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

表 1: 節境界解析の結果

サンプル数		総短単位数	節境界数	節境界率
書籍	83	234,400	22,601	9.64%
雑誌	86	239,440	22,975	9.60%
新聞	340	360,526	28,759	7.98%

節境界の総体的な出現率という点で言えば、書籍と雑誌はほぼ同様の結果であった。一方、新聞では出現率が低かった。次に、短単位データベースに対して節境界ラベルを付与し、テキストに整形した結果を図 5 に示す。太字の部分が付与された節境界ラベルである。

雛人形の起源は形代・人形にあり **連用節_動詞述語** 、古くは草や紙でそれを作って **連用節_テ** 身をぬぐい **連用節_動詞述語** 、息を吹き込んでは **連用節_テハ** 病気や穢れをそれに移して **連用節_テ** 、川や海に流し送ったのだと **引用節_ト** いう。**文末_一般**

図 5: 節境界ラベルの付与結果の例

100 万語あたりに出現した節境界ラベルの数について、上位 10 位までをメディアごとに表 2 に示す。

表 2: 節境界ラベルの 100 万語あたりの出現数

書籍		雑誌	
文末_一般	36,770	文末_一般	39,338
連用節_テ	11,280	連用節_テ	9,130
連用節_動詞述語	6,706	連用節_動詞述語	7,179
引用節_ト	4,480	文末_非文末境界	7,092
文末_引用内	4,330	文末_引用内	4,903
並列節_ガ	3,430	引用節_ト	3,141
条件節_レバ	2,491	並列節_ガ	2,865
条件節_ト	2,304	条件節_ト	2,017
連体節_トイウ	1,950	並列節_デ	1,662
並列節_デ	1,843	条件節_レバ	1,616

新聞	
文末_一般	32,439
連用節_動詞述語	8,876
文末_非文末境界	8,074
連用節_テ	6,646
文末_引用内	5,131
引用節_ト	2,688
並列節_ガ	2,552
条件節_ト	1,304
並列節_デ	1,187
条件節_レバ	1,004

表 2 からは、3 種類のメディアの間で、節境界ラベルの分布が異なっている様子を見て取ることができる。**文末_一般**は、雑誌で最も多く、新聞が最も少ない。総体的に見て、新聞では 1 文が長くなる傾向にあるのだろう。

文末_非文末境界は、句点で終わらない文末相当の位置である。雑誌・新聞のみに出現しているのは、記事タイトルやリード文などにこのラベルが付与されているからであろう。**連用節_テ**は、書籍・雑誌に多く、新聞に少ない。逆に**連用節_動詞述語**（動詞の連用形による接続）は、新聞に多く、書籍・雑誌に少ない。これは、書籍・雑誌と新聞との間で、継起や並列などを表す接続形式の選好性が異なることを示すものである。

4.4 解析誤りの分析

節境界解析の結果のうち、書籍の約 1 万語分を人手でチェックし、解析誤りを収集していくつかのタイプに分類した。以下、3 種類の誤解析の例を挙げておく。

規則の不備による誤り 準備した規則が十分でなかったため、検出できない節境界があった。これらは、適切な規則を準備することで改善することが可能である。

(1) 刀を抜いて負けた以上、命はない
(2) 死を高める代わりに、気味の悪い恐怖にしてしまう
(3) 訊いてまいりますゆえ、三田村さまは釜山の町中でも
(4) ハコを取り替えたって、人の生活が簡単に変わるわけは

節境界ラベルの過剰な付与 節境界ラベルが付与されたもののの、付与の対象外とすべき事例があった。

(5) 彼らにとっては **連用節_テハ** かけがえのない
(6) 暮し方はいつだって **引用節_ッテ** 変化しているし
(7) どういう時に **連用節-時間節_トキニ** 不安を感じますか。

これらの問題については、事前に複合辞を認定しておくことや、付与対象データとして短単位ではなく長単位を用いることによって、回避できる可能性がある。

局所的な検出規則では対応できない例 ある語の前後を参照するだけでは対応できない例が存在する。

(8) はるかに野太くそして元八郎の記憶にある声だ。
(9) 太郎が右側に、僕は左側に寝た。
(10) オーガスチンが電報を手に、私を起こしに来た

(8) は、形容詞が連用節の述語を形成する例である。形容詞の連用形は、単独で連用修飾要素として機能する場合と、1 つ以上の補足語を従えて連用節を形成する述語として機能する場合とがある。これらは形態的に区別することができないため、連用修飾要素か連用節かを決めるためには、構文解析の結果を参照しなければならない。(9) は、述語を共有する並列構造の例である。「右側に」の直後はある種の（述語を持たない）節境界の一種であると考えられるが、このような境界を局所的に検出する規則は書くことができない。さらに (10) は、やはり述語を持たない、イディオマティックな節境界を含む例である。このような境界もまた、局所的な検出規則を書くことはできない。これらのような例については、現状の検出規則を拡張したとしても、適切に対処することは難しい [1]。

5 節境界ラベルを使った分析

最後に、BCCWJ と CSJ に付与された節境界ラベルを用いて、書き言葉と話し言葉を比較・対照する形で、並列節と条件節の分析を行なう。

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.

5.1 並列節の分析

接続助詞「が」「け（れ）ど（も）」によって導かれる並列節は、相互に入れ替えが可能な類義表現である。これらが、書籍・雑誌・新聞という BCCWJ の書き言葉、および CSJ の「学会講演」「模擬講演」という話し言葉の中でどのように分布するかについて、付与された節境界ラベルを集計した結果を図 6 に示す。

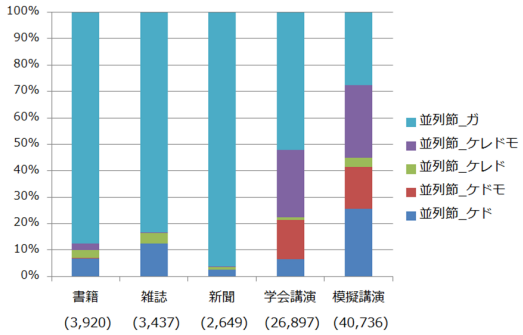

図 6: 並列節ガ・ケ（レ）ド（モ）の分布

「が」と「け（れ）ど（も）」の分布は、書き言葉と話し言葉の間で大きく異なっていることが分かる。書き言葉では「が」の比率が 90%前後を占めるのに対して、話し言葉では 30〜50%に留まっていることから、書き言葉では「が」に強い選好傾向があると言える。

このうち、「け（れ）ど（も）」の分布のみを抜き出した結果を、図 7 に示す。

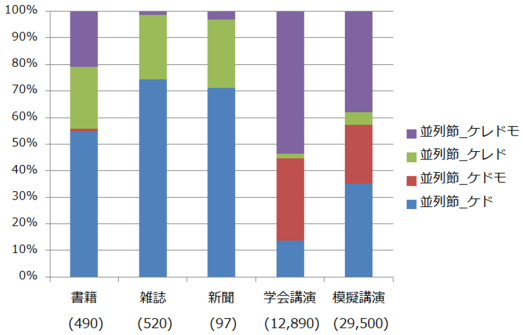

図 7: 並列節ケ（レ）ド（モ）の分布

「けれども」と「けど」という形式では、「けど」の方が口語的な印象を受けるが、図 7 を見ると、「けど」の比率はむしろ書き言葉の方が圧倒的に高い。原文を確認したところ、これらの例の大半は、小説等の会話部分に現れたものであった。一方、CSJ で「けれども」の比率が高くなっているのは、基本的にデスマス体の丁寧な発話スタイルで話されているためと解釈できる。

また、書き言葉では「けれど」が多く、話し言葉では「けども」が多い、という対照的な結果を見て取ることができる。理由は明らかではないが、書き言葉と話し言葉の間で選好性に明確な違いがあることになる。

5.2 条件節の分析

日本語の条件節には、「れば」「たら」「なら」「と」という 4 つの接続形式がある。付与された節境界ラベルの集計結果を、図 8 に示す。

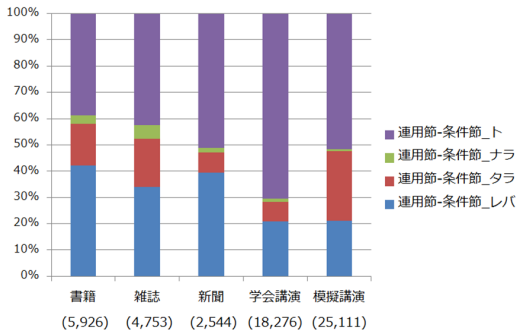

図 8: 条件節の分布

4 つの条件形式の中では、比較的「たら」が口語的であると考えられるが、これは「新聞」「学会講演」という硬いテキスト・口調の中で「たら」が少ない比率になっていることと符合する。一方、硬い書き言葉の新聞では「れば」が、硬い話し言葉の学会講演では「と」が、それぞれ選好されるようである。

6 まとめ

本稿では、BCCWJ に対する節境界ラベルのアノテーション実験について報告した。また、付与された節境界ラベルを用いて、書き言葉と話し言葉の文法形式の比較・対照を行なった。同じ設計に基づくラベルをアノテーションすることによって、異なるメディアに属するテキストの統一的な分析が可能なことを示した。

現時点での開発版では、規則の不備によりまだ検出できない節境界が存在する。また、CSJ で実装した節境界のクラス分類についても未実装の状態である。今後、規則を拡張することによって、これらの問題に対処していく予定である。

参考文献

[1] 丸山岳彦, 柏岡秀紀, 熊野正, 田中英輝. 日本語節境界検出プログラム CBAP の開発と評価. 自然言語処理, Vol. 11, No. 3, pp. 39–68, 2004.

[2] 丸山岳彦, 高梨克也, 内元清貴. 第 5 章 節単位情報. 日本語話し言葉コーパスの構築法, 国立国語研究所報告書 124, pp. 255–322. 国立国語研究所, 2006.

[3] 南不二男. 現代日本語の構造. 大修館書店, 1974.

[4] Tomohiro Ohno, Shigeki Matsubara, Hideki Kashioka, Takehiko Maruyama, Hideki Tanaka, and Yasuyoshi Inagaki. Dependency parsing of Japanese monologue using clause boundaries. *Language Resources and Evaluation*, Vol. 40, No. 3-4, pp. 263–279, 2007.

Copyright(C) 2013 The Association for Natural Language Processing.
All Rights Reserved.