現代日本語書き言葉均衡コーパスに対する節境界付与

佐藤理史
名古屋大学大学院工学研究科

丸山岳彦
国立国語研究所

夏目和子
名古屋大学大学院工学研究科

1 はじめに

日本語の文を構成する単位の一つに、「複文を構成するところの、述語を中心とした各まとまり[1]」と定義される**節**がある。この単位は、言語学において、非常に重要な単位であり、言語処理においても、特に長い文を処理する際には、認定すべき有用な単位と考えられる[2, 3]。

我々は、現代日本語書き言葉均衡コーパスのコアデータに対して節末境界(以下、節境界と略記する)を付与することを構想し、その実現に取り組んでいる[4, 5]。本稿では、現時点での節境界の設計と認定規則の詳細について報告する。

2 節境界の付与の方針

現代日本語書き言葉均衡コーパス(BCCWJ)への節境界の付与に際し、以下のような方針を立てた。

1. コアデータの長単位TSVデータに対して節境界を付与する。コアデータは6つのレジスタから構成されるが、書籍(PB)を先行させ、順次、他のレジスタに拡大する。
2. 強い区切りの節境界の付与を先行させ、順次、付与する節境界の種類を拡大する。
3. プログラムで節末境界を付与し、必要に応じて人手で修正する。このプログラムを新たに作成する。

我々は、節境界の認定規則を書き下し、それを駆動して節境界認定器を構成するという方法をとる。この選択は、「大半の節境界の認定は比較的単純な規則で実現できる」という直感と経験[6, 2]に基づく。もう一つの理由は、節境界に対する厳格かつ網羅的な定義が存在するわけではなく、「設計しなければならない節境界」が存在するからである。

たとえば、次のような例文を考えよう。なお、本論文では節境界の箇所を-C-と表記する。

(1) 彼女が到着したとき-C-雷が鳴った。
(2) 雷が鳴ったのは-C-彼女が到着したとき-?-だ。

例文1の「彼女が到着したとき」を時間節と認定し、例文2の「雷が鳴ったのは」を補足節と認定するのはゆるぎない。しかしながら、例文2の「彼女が到着したとき」を節と認定するかどうかは定かではない。我々は、これを時間節と認定する立場をとるが、「彼女が到着したときだ」を主節と認定する立場もあろう。一般に、判定詞が絡む場合の節の設計は、かなり悩ましい。

規則で節境界を認定するメリットは、我々が定めたモデル(規則)を実際のデータに対して即座に適用でき、その結果を観察できることにある。そして、その観察に基づき速やかにモデルを修正できることにある。本研究(BCCWJに対する節境界付与)の主眼は、節境界認定器を作ることにあるのではなく、標準的な節境界を定めることにある。

3 Rainbow4の節境界モデル

本節では、節境界付与に用いるシステムRainbow4が採用している節境界モデルについて説明する。

3.1 文節モデル

Rainbow4では、節境界の認定に先立ち、まず、文節境界を認定する。その理由は、文節境界のみが節境界となりうると考えるからである。

Rainbow4では、**助詞を中心とした**文節境界認定を採用する。具体的には、文節は次のいずれかの形式をとるものとする。

-B-W-B-
-B-W-A-助詞の列-B-

ここでは、-B-は文節境界を、Wは複合語や派生語を含む**長い単位の語**[1](助詞以外)を表す。文節のWの部分を**主要部**、助詞の列を**機能部**と呼び、主要部と機能部の境界をA境界(-A-)と呼ぶ。「助詞は文節の機能部のみに現れ、文節の機能部は助詞のみで構成される」ことと、「文節の主要部は、長い単位の語1語で構成される」ことの2点が、Rainbow4の文節モデルの根幹をなす大原則である。

[1] BCCWJの長単位よりも長い。

ほとんどの場合、Rainbow4の文節境界は、BCCWJのそれと一致する。一致しないのは次のような場合であり、‘-’を文節境界から主要部内境界に書き換える。

1. 一部のテ形複合動詞。例：書いて-いただく
2. 形容詞+ない。例：美しく-ない、確実で-ない
3. 指示詞。例：こう-いう、この-ような

3.2 節境界と節末形式

日本語の文や節では、述語が他の要素(項、修飾句など)より後ろに配置されるという制約がある。このため、述語を含む文節(以下、**述語文節**と呼ぶ)の末尾の境界が、節境界の第一候補となる。

(3) [述語文節]-C-

我々は、これをオーバーライトする場合を規定することにより、節境界の位置を定める。

拡大述語文節 述語文節の直後に、助動詞を主要部とする文節(助動詞文節)が後続する場合、助動詞文節を含めて、拡大述語文節と考える。まれに、助動詞文節が連続することがあるが、この場合も、連続する助動詞文節を含め、拡大述語文節とみなす。

(4) [[述語文節]-B-[助動詞文節]-C-] (拡大述語文節)

(5) 書く-B-らしいが-C-

以下の説明では、特に断らないかぎり、述語文節は拡大述語文節を含むものとする。

節末機能文節 述語文節の直後に、ある特定の文節が後続するとき、これを節に含め、後続文節の末尾境界を節境界と認定する。このような認定を行なう特定の文節を節末機能文節と名付ける。節末機能文節が後続する場合、述語文節は、原則として連体修飾の形式をとる。(若干の例外がある。)

節末形式 節末機能文節が後続する場合にかぎり、「節境界は述語文節の末尾境界」という原則をオーバーライトする。以上により、節末形式は、次のいずれかとなる。

(5) Type I. [述語文節]-C-
Type II. [述語文節]-B-[節末機能文節]-C-

3.3 節境界認定規則

節境界認定規則を上記の形式で区分して、それぞれ表に示す。

I-f 述語文節の機能部による認定 — 表1
I-m 述語文節による認定規則 — 表2
II-t 節末機能文節による認定規則 — 表3
II-pt 述語文節と節末機能文節の両者を考慮した認定規則 — 表4

これらの規則群は、ここに示した順番の後ろの方が優先度が高い[2]。記述形式は、以下の通りである。

1. **節ラベル**とは、節境界に付与する種類情報である。直後の数字は、節境界の強さを表すもので、現在、1(最も強い境界)と2(連体修飾節)の区分がある。‘[ ]’付きの行は、２つの規則を１行にまとめたもので、たとえば、「条件節トコロデ[-助詞]」は、「条件節トコロデ」と「条件節トコロデ-助詞」の２つの規則を表現している。その条件の違いは、‘[ ]’の部分で表現されている。
2. 文節の主要部は、以下のような形式で記述されている。

   品詞、活用形、品詞-活用形、リテラル、リテラル-活用形、P(=任意の述語)、T(=述語の連体修飾形式)、A(=任意)

3. 文節の機能部は、助詞の列であり、それぞれの助詞は、以下のような形式で記述されている。

   助詞区分(格、副、係、接続、終、判デ)、リテラル、リテラル-助詞区分、any(=任意の助詞)、any?(=0個か1個の助詞)、any*(=任意の助詞列)、any+(=長さ1以上の助詞列)

4. 「判デ助詞」は、Rainbow4の文法で認める、判定詞由来の助詞「で」である。(BCCWJでは判定詞として認定されている。)
5. 節認定規則II-tの述語文節に対する条件は、すべて、主要部は「T」、機能部は「なし」、である。
6. 節認定規則II-ptの先頭の2つの規則は特別な規則で、他のII-ptまたはII-tの規則が適用されて節末機能文節が認定された後に、節ラベルを書き換えるだけに用いられる規則である。
7. 「抑制ホカノ」は、節認定を抑制する規則である。

3.4 節境界認定のアルゴリズム

Rainbow4は、以下の手順で節境界を認定する。

1. 与えられた文(BCCWJの文を単位とする)の文節境界を認定する。
2. 文の先頭から文節境界を調べ、その直前が述語文節であった場合は、以下を行う。

[2]規則適用の制御の詳細は割愛するが、原則として、適用できる規則のうち条件が最も厳しい規則が適用されるように運用している。

表 1: 節境界認定規則I-f

節ラベル		述語文節 主要部	機能部
引用節[-助詞]	1	P	終 と-格 [any+]
引用節-補足語	1	P	終 と-格 格
引用節トノ	1	P	終 と-格 の
トカ節[-助詞]	1	P	終 と-格 か [any+]
引用節-助詞	1	P	終 と-格 か と
連用節その他	1	P	うえで
連用節その他[]	1	P	うえで-接続 [も]
間接疑問節[-助詞]	1	P	か [any]
引用節[-助詞]	1	P	か と [(も|の)]
トイウ節	1	P	か という
並列節ガ[-助詞]	1	P	が-接続 [終]
間接疑問節カドウカ[-助詞]	1	P	かどうか [any]
引用節[-助詞]	1	P	かどうか と [(も|の)]
トイウ節	1	P	かどうか という
引用節	1	P	から て
理由節カラ[-助詞]	1	P	から-接続 [と]
テ節	1	P	から とて
テ節	1	P	からといって-接続
連用節その他[]	1	P	くらい-副 [に]
並列節ケド	1	P	けど
並列節ケドモ[-助詞]	1	P	けど も [終]
並列節ケレド	1	P	けれど
並列節ケレドモ[-助詞]	1	P	けれど も [終]
連用節その他	1	P	際に-格
並列節シ[-助詞]	1	P	し [終]
ダケ節[-助詞]	1	P	だけ [any]
連体節ダケノ	1	P	だけ の
ダケ節-助詞	1	P	だけ と any
ダケニ節[-助詞]	1	P	だけ に [any]
タメニ節[-助詞]	1	P	ために [any]
連体節タメノ	1	P	ための
ツツ節[-助詞]	1	P	つつ [any]
引用節ッテ	1	P	って
条件節ト	1	P	と-接続
引用節[-助詞]	1	P	と-格 [any+]
引用節	1	P	と-格 any と
引用節-補足語	1	P	と-格 格
引用節トノ	1	P	と-格 の
トカ節[-助詞]	1	P	と-格 か [any+]
引用節	1	P	と-格 か any と
並列節トカ[-助詞]	1	P	と か [(で|も)]
並列節トカ-助詞	1	P	と か と か
トイウ節	2	P	という
条件節タラ	1	P	(ときたら|としたら)
連用節ドコロカ	1	P	どころ か
条件節バ	1	P	とすれば
連用節その他	1	P	と同時に-接続
連用節その他	1	P	とはいえ
ナガラ節[-助詞]	1	P	ながら [any]
ナガラ節-助詞	1	P	ながら と any
ナガラモ節	1	P	ながら も
ナド節[-助詞]	1	P	など [any]
トイウ節	1	P	など という
ナド節-補足語	1	P	など 格
ナド節-助詞[]	1	P	など と [any]
連体節ナドノ	1	P	など の
条件節ナラ[-助詞]	1	P	なら [と]
条件節ナラバ[-助詞]	1	P	ならば [と]
引用節ナンテ	1	P	なんて
連用節ニハ	1	P	に は
連用節ニモ	1	P	に も
譲歩節ニセヨ	1	P	にせよ-接続
理由節ノデ	1	P	ので
ノニ節	1	P	のに
ホド節[-助詞]	1	P	ほど [any]
ホドニ節[-助詞]	1	P	ほど に [any]
連体節ホドノ	1	P	ほど の
マデ節[-助詞]	1	P	まで [と]
マデニ節	1	P	までに
マデニハ節	1	P	までに は
マデハ節	1	P	までは
マデモ節	1	P	までも
連体節マデノ	1	P	までの
条件節モノノ	1	P	ものの
並列節ヤラ	1	P	やら
ヨリ節[-助詞]	1	P	より [any]
連用節その他	1	P	割に-接続

表 2: 節境界認定規則I-m

節ラベル		述語文節 主要部	機能部
連用中止節	1	A	で-判デ
連用中止節	1	A	副助詞 で-判デ
連用節	1	終止形	に
連用節ッテ	1	判定詞-終止形	って
引用節	1	動詞-命令形	て
連体節	2	動詞-連体形	
連体節	2	判定詞-連体形	
連体節	2	甲助動詞-連体形	
連用節	1	連用形	
連用節シニ[-助詞]	1	連用形	に [(は|も)]
ズニ節[-助詞]	1	ぬ/ズ-連用形	に [は]
イ形容詞連用節	1	イ形容詞-連用形	
条件節バ[-助詞]	1	基本条件形	[と]
条件節タラ[-助詞]	1	タ系条件形	[と]
条件節タラバ	1	タ系条件形	ば
テ節	1	テ形	
テカラ節[-助詞]	1	テ形	から [any]
テカラノ節	1	テ形	から の
テノ節	1	テ形	の
テハ節	1	テ形	は
テモ節	1	テ形	も
連用節	1	ナ形容詞-テ形	
並列節デハ	1	ナ形容詞-テ形	は
並列節タリ[-助詞]	1	タリ形	[(で|は|も|とか)]
連用節バカリ[-助詞]	1	(動詞-連体形|動詞-テ形)	ばかり [に]
ヨウ節[-助詞]	1	よう	[any]
ヨウ節-助詞	1	よう	で は
ヨウニ節[-助詞]	1	よう	に [any]
連体節ヨウナ	2	ようだ-連体形	
テモ節	1	動詞	にしても
テモ節	1	(動詞|判定詞)	(としても|といっても)
テモ節	1	動詞	(ほど|だけ) で も

(a) 拡大述語文節を認定する。その末尾の文節をPとし、その直後の文節をTとする。

(b) P-B-Tに対して、節境界認定規則II-ptまたはII-tが適用できたならば、Tを節末機能文節と認定し、Tの直後の境界を節境界とする。

(c) それ以外の場合は、Pの直後の境界を節境界とする。節境界認定規則I-mまたはI-fが適用できた場合は、節ラベルが定まる。それ以外の場合は、「未定義」とする。

(d) 節境界が文末境界と一致する場合の節ラベルには、「文末_」というprefixを付与する。

3. 文末境界に節ラベルが付与されなかった場合には、「文末-その他」を付与する。

4 現状と問題点

2015年末の時点で、付与の第1ラウンドが完了している。第1ラウンドでは、レジスタ書籍(PB)に対する節境界付与結果を調査し、規則の調整を行なった。現時点では、まだ、いくつかの問題を抱えており、それらは、節や文法の設計上の問題、現方式では解決できない問題、処理上の問題に大きく区分される。以下では、これらについて簡単に述べる。

節や文法の設計上の最大の問題は、判定詞の扱いで

表 3: 節境界認定規則II-t

節ラベル		節末機能文節 主要部	機能部
条件節タラ	1	A	(ときたら|としたら)
条件節バ	1	A	とすれば
時間節その他[-助詞]	1	あいだ	[any]
時間節アト[-助詞]	1	(あと|のち)	[any+]
時間節アトデ[-助詞]	1	(あと|のち)	で [any]
時間節アトニ[-助詞]	1	(あと|のち)	に [any]
時間節アトノ	1	(あと|のち)	の
連用節その他[]	1	(あまり|以上|一方|うえ|なか)	[(に|で)]
時間節イマ[-助詞]	1	いま	[any+]
時間節その他[-助詞]	1	(うち|おり|際|瞬間|たび)	[(に|は)]
時間節その他-助詞	1	(うち|おり|際|瞬間|たび)	に any
時間節その他[-助詞]	1	ころ	[any+]
条件節カギリ[-助詞]	1	かぎり	[any]
連体節カギリノ	1	かぎり	の
連用節その他	1	かわり	に
条件節ケッカ	1	結果	
タメ節[-助詞]	1	ため	[any]
条件節ナラ	1	ため	なら
連体節タメノ	1	ため	の
タメニ節[-助詞]	1	ため	に [any]
タメニハ節	1	ため	に は
時間節トキ[-助詞]	1	とき	[any]
時間節トキデ	1	とき	で
時間節トキニ	1	とき	に
時間節トキニハ	1	とき	に は
時間節トキノ	1	とき	の
時間節その他[-助詞]	1	(とたん|以前|瞬間)	[any+]
連用節その他	1	(半面|反面)	
ホカ節[-助詞]	1	ほか	[any]
抑制ホカノ	*	ほか	の
ホカニ節[-助詞]	1	ほか	に [any]
条件節バアイ[-助詞]	1	場合	[(副|に|は)]
連体節バアイノ	1	場合	の
時間節マエ[-助詞]	1	まえ	[any]
時間節マエニ[-助詞]	1	まえ	に [any]
時間節マエノ	1	まえ	の
ママ節[-助詞]	1	まま	[any]
ママデ節[-助詞]	1	まま	で [any]
補足節	1	N	(格|係|副)
補足節	1	N	(格|係|副|で) any
補足節	1	N	X
テ節	1	N	(に対して|によって)
条件節ナラ[-助詞]	1	N	なら [any]
条件節ト	1	N	によると-格
条件節タラ	1	こと	ときたら
間接疑問節[-助詞]	1	の	か [any]
引用節[-助詞]	1	の	か と [(も|の)]
引用節トイウ	1	の	か という

N = (こと|の|ん|もの|もん|ところ)
X = (と|たり|だの|だり|とか|なり|や|やら)

ある。「判定詞はどのように扱ってもすっきりしないところが残る」というのが我々の実感であり、「ある方針に従って決めるしかない」と考えている。節境界からの要請により、Rainbow4では、判定詞を独立とした文節と認めることとしたが、これはRainbow4が実装する文法における、当初設計からの最大の変更である。

現方式は、局所的な形式に基づき節境界を認定するので、局所的には定まらない場合、および、意味的にしか定まらない場合は、正しく節境界(節ラベル)を認定できない。前者の例としては、形容詞連体節[3]、後者の

[3]現時点の実装では、形容詞の連体形(–イ、–ナ)および連用形(–ク、–ニ)は述語として認定しないため、形容詞連体節は認定されない。

表 4: 節境界認定規則II-pt

節ラベル		述語文節 主要部	機能部	節末機能文節 主要部	機能部
時間節イライ					
	1	テ形		以来	any*
条件節トコロデ[-助詞]					
	1	タ形		ところ で [は]	
補足節	1	P	副	N	(格|係|副|で) any?
補足節	1	P	という	N	(格|係|副|で) any?
補足節	1	P	any という	N	(格|係|副|で) any?

N = (こと|の|ん|もの|もん|ところ)

例としては、以下のような「ところ」を含む節がある。

(6) 書いたところで-C-
(条件節トコロデ、連用節、補足節のいずれか)

処理上の問題として残っているものの 1 つは、終助詞の扱いである。終助詞は、広く、文節の末尾に現われうる。また、引用節の中の引用文の最後にも現われうる。これらの現象をすべて規則で記述するのは煩雑なため、システム側からのなんらかのサポートが必要である。

もう 1 つは、実体を持つ境界の扱いである。

(7) 「目をとじて-A[」]-は-C[、]-京都御所でのスナップ。

現在の実装では、境界として認定する記号(句読点や鉤括弧)は、透過的(あってもなくても同一)に扱われているため、上記の例文は、誤ってテハ節と認定される。日本語では、記号の使用に一貫性がないため、完全な解決は不可能であるが、なんらかの対処が必要であろう。

謝辞 本研究では、『現代日本語書き言葉均衡コーパス』を利用した。本研究は、JSPS 科学研究費基盤研究 (B)「文章の読解と産出のための言語処理技術」(課題番号 15H02748) の助成を受けている。

参考文献

[1] 益岡隆志, 田窪行則. 基礎日本語文法—改訂版—. くろしお出版, 1992.
[2] 加納隼人, 佐藤理史. 日本語節境界検出プログラムRainbowの作成と評価. 第13回情報科学技術フォーラム(FIT2014), E-005, 第2分冊, pp. 215–216, 2014.
[3] 加納隼人, 佐藤理史, 松崎拓也. 節境界検出を用いたセンター試験『国語』評論傍線部問題ソルバー. 情報処理学会自然言語研究会, NL–220–8, 2015.
[4] 丸山岳彦. BCCWJに対する節境界ラベルのアノテーション. 言語処理学会第19回年次大会発表論文集, pp. 154–157, 2013.
[5] 佐藤理史, 丸山岳彦. 節境界認定に関する諸問題. 第8回コーパス日本語学ワークショップ予稿集, pp. 225–232, 2015.
[6] 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.