節境界認定に関する諸問題

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 言語資源研究系)

Issues of Clause-Boundary Detection

Satoshi Sato　(Graduate School of Engineering, Nagoya University)
Takehiko Maruyama　(National Institute for Japanese Language and Linguistics)

要旨

本稿では、文中の節境界を認定するために必要な処理について議論する。我々は、既存のCBAP と異なり、まず、文節境界を認定したのち、節境界を認定する方法を採用する。この方法の採用により、節境界認定問題を、(1) 文節境界 (および文節) をどのように定義するか、(2) 文節にどのような属性を認定するのか、(3) どこを節境界と定義するか、の 3 つの部分問題に分割できる。これらの問題に対する現時点での方針を述べ、BCCWJ のコアデータの一部への節境界付与の見通しについて述べる。

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**という単位がある。これまでの日本語の言語処理において、節という単位はそれほど重要視されてこなかった。しかしながら、我々は、センター試験の国語問題を解くシステムを開発する過程で、長い文をいくつかの部分に分割する必要性に遭遇し、そのための基礎となる節境界の認定が不可欠であるという認識に至った [2, 3]。

節境界検出プログラムには、丸山らの CBAP [4] がすでに存在する。しかしながら、CBAP は、特定の形態素解析システム (ChaSen/IPAdic) に依存していること、および、形態素解析結果の文字列を書き換える方式で実装されているため、保守性・拡張性に難がある。このため、CBAP を改良するのではなく、完全に新しいシステムを作成する方針を採用し、節境界認定システム Rainbow を試作した [2]。

Rainbow の特徴は、(1) 文節境界の認定、(2) 文節属性の認定、(3) 節境界の認定、という 3 つの段階を踏んで、節境界を認定する点にある。このような段階を踏むことにより、節境界認定問題を、3 つの部分問題に分割して解く。もちろん、これらの部分問題は相互に関連しているが、それぞれにおいて解くべき問題はかなり明確であり、保守性・拡張性の高いシステムを作成することができる。

我々は、現在、「現代日本語書き言葉均衡コーパス (BCCWJ)」に節境界を付与するために、Rainbow の新しいバージョン (Rainbow3) を実装している。これに合わせて、上記の (1)(2)(3) に対し、より明確な基準を定めようとしている。本稿では、その内容について報告する。

なお、Rainbow3 の内部は、原則として、益岡・田窪文法 [1] に準拠した文法体系を採用している。

2　3ステップによる節境界認定

節境界認定とは、より正確には、**節の末尾の境界**を認定することを意味する。たとえば、以下の文では、-C-が節境界となる。

(1) 太郎が荷物を軽々と運んだので-C-花子は驚いた-C-

この境界を認定するために、まず、文中の文節境界を認定する。文節境界は-B-で表す。

(2) -S-太郎が-B-荷物を-B-軽々と-B-運んだので-B-花子は-B-驚いた-S-

文節境界で区切られたものを文節と認定する。なお、文の先頭と末尾には、それぞれ文境界 (-S-) があるものとみなす。

次に、それぞれの文節の属性を認定する。ここで最も重要な属性は、その文節が文中で述語として働いているかどうかを表す属性である。この例では、「運んだので」と「驚いた」の2つの文節が述語として働いていると認定する。

(3) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-B-花子は-B-**驚いた**-S-

最後に、この結果に基づいて、節境界 (-C-) を認定する。なお、文末の文境界も節境界であるが、これは、そのまま-S-と表記する。

(4) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-C-花子は-B-**驚いた**-S-

3　境界と句読点

前節に示したように、ここでの解析は、単位を中心とした解析ではなく、**境界を中心とした解析**である [5, 6]。多くの場合、境界は文字と文字の間に存在するが、句読点や括弧などの補助記号は、それ自身が境界を表すとみなす。つまり、これらは実体を持った境界である。句点は「文境界を表す記号」であり、読点は「文中の比較的大きな境界を表す記号」である。実体を持った境界は、以下のように角括弧付きで表す。

(5) -S-太郎が-B-荷物を-B-軽々と-B-**運んだので**-C[、]-花子は-B-**驚いた**-S[。]-

4　文節境界認定

文節 [7] は、文を構成する単位の中で、おそらく最も合意が取りやすい (個人差が少ない) 単位であろう。以下のように文を文節に区切ることは、多くの人々にとって自然である。

(6) -S-太郎が-B-荷物を-B-軽々と-B-運んだので-B-花子は-B-驚いた-S-

しかしながら、文節境界の認定にも、合意が取りにくい (すなわち、明確な定義が必要な) 場合もある。

Rainbow3 では、**助詞を中心とした**文節境界認定を採用する。すなわち、複合語や派生語を含む**長い単位の語**を助詞と助詞以外 (W) に分け、次のような境界を認定する。これが、長い単位の語の間の境界のすべてである。

助詞-j-助詞
助詞-B-W
W-A-助詞
W-B-W

この結果、文節は、次のいずれかの形式をとることになる。

-B-W-B-
-B-W-A-助詞の列-B-

ここで、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、Rainbow3の文節モデルの根幹をなす大原則である。このような文節モデルにおいては、助詞の集合を定義すれば、文節境界がほぼ定まる。(より正確には、どんな単位を長い単位の語と認定するか—すなわち、長い単位の語の中に存在しうる全ての境界—を明確に定義する必要がある。)

我々は、文節境界のみが、節境界となりうる境界と考える。

5　節境界と節末形式

日本語の文や節では、述語が他の要素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。

[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。

5.1　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。

(拡大述語文節)
[[述語文節]-B-[助動詞文節]]-C-　　例：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。

5.2　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。

(7) a. 太郎が荷物を軽々と-B-運んだ-B-**こと**-A-**が**-C-花子を驚かせた。
　　b. 太郎が荷物を軽々と-B-運んだ-B-**とき**-C[、]-...
　　c. 太郎が荷物を軽々と-B-運ぶ-B-**あいだ**-A-**に**-C[、]-...

5.3　節境界の決定

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。

I. 述語文節-C-
II. 述語文節-B-節末機能文節-C-

すでに述べたように、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。すなわち、若干の例外を除き、述語文節が連体修飾不可能な形式であれば、I 型となる。述語文節が連体修飾可能な形式の場合、後続文節が節末機能文節であれば II 型、そうでなければ I 型となる。

以上の議論から明らかなように、我々のモデルでは、次の情報があれば、節境界を定めることができる。

1. 文節境界 (文節境界だけが、節境界となる可能性がある)
2. 文節が述語文節か否か
3. 文節が助動詞文節か否か (拡大述語文節の判定に必要)
4. 文節が連体修飾可能か否か (述語文節と助動詞文節のみ)
5. 文節が節末機能文節か否か

これらのうち、2–5 の 4 つは、**文節の属性**と考えることができる。つまり、先に述べたように、

Step1 文節境界を認定し、
Step2 その結果認定される各文節の 4 つの属性の値を決定すれば、
Step3 それらの情報のみを用いて、節境界を認定する

ことができる[1]。

残された問題は、Step1 と Step2 の詳細化である。これらのうち、以下では、述語文節の認定と、節末機能文節の認定について議論する。

6　述語文節の認定

節は述語を中心としたまとまりである。つまり、述語があって、初めて節が構成される。そのため、述語を認定することが、節境界を認定する前に必要である。この段階では、文は文節に分割されているので、述語として働いている文節 (述語文節) を認定する問題となる。

述語文節の認定のために、文節の主要部 W の品詞を定義する。Rainbow3 では、次の 10 品詞を採用する。これは、益岡・田窪文法 [1] の品詞体系を踏襲している (指示詞は設けない)。

名詞 (代名詞、数詞を含む)、動詞、形容詞 (ナ形容詞とイ形容詞)、副詞、
連体詞、接続詞、感動詞、助動詞、判定詞、助詞

このうち、助詞を除く 9 品詞が文節の主要部となるが、文節が述語文節となりうるのは、原則として、主要部が

動詞、形容詞、判定詞、助動詞

[1] 述語文節の後ろに、節末機能文節が連続して接続する III 型を考える必要があるかどうかは、現時点では保留とする。ただし、III 型を導入しても、3 ステップ節境界認定法は堅持できる。

の 4 品詞の場合である。助動詞は、ほとんどの場合、動詞、形容詞、判定詞に後続するが、名詞に直接接続する場合があるので、便宜上、述語文節を構成するとみなす。同様に、「か」「かしら」などの一部の終助詞は、名詞に直接接続する場合があるので、「学生かしら」のような文節は、述語文節と認める必要がある。

述語文節の認定の難しさは、文節内の情報のみからでは、その文節が述語文節かどうか判定できない場合があることにある。これには、節の定義の問題も関連する。

問題となるのは、主に、形容詞の連用形と連体形である。イ形容詞[2]は、単体で述語を副詞的に修飾している場合 (連用修飾) は述語として働いていない (節を構成しない) 一方で、補足語 (格要素) を支配する場合は述語として働いている (節を構成する) とみなすのが一般的である。ただし、これらを区別せず、両者とも節とみなす立場もある。

(8) a. 花が-B-美しく-B-咲いた-S[。]-　(連用修飾語)
　　b. 花が-B-美しく-C-香りも-B-いい-S[。]-　(並列節)

形容詞の連体形も同様である。

(9) a. きれいな-B-女性-B-　(連体修飾語)
　　b. 声が-B-きれいな-C-女性-B-　(連体節)

以上の例からわかるように、これらの区別は、当該文節の範囲内では決定できないのは明らかである。

文節の属性を計算するという立場から考えれば、ここで行うべきことは、文節内の情報に基づいて、その文節が述語文節となりうる可能性を持つか否かの判定である。この段階では、可能性を持つものには、すべて「述語性」という属性を付与する。

その先の処理には、いくつかの選択肢がある。

1. 一般的な述語および節の定義に従うように、述語文節を認定するルール (ヒューリスティック) を実装する。
2. 境界ラベルあるいは節ラベルを工夫することにより、格要素を伴わない場合には節とならないという情報を明示する。
3. 格要素を伴わない場合でも節と認める。

最終的な決定は保留しているが、現時点では、2 番目の選択肢を中心に検討している。

7　節末機能文節の認定

節末機能文節が関与する節は、主に、副詞節と補足節である。これらの節では、連体節との区別が問題となる。

7.1　副詞節と連体節

副詞節か連体節かの判定は、後続文節を節末機能文節と認定するか否かに帰着させる。

(10) a. わたしが 16 だった-B-**とき**-C[、]-彼女はまだ 7 つでした。 (副詞節)
　　b. わたしが 16 だった-C-年、彼女はまだ 7 つでした。 (連体節)

[2] ナ形容詞の並列節はテ形をとるので、ナ形容詞の連用形は連用修飾語とみなしてよい。

節末機能文節が「に」や「で」を伴う場合、「に」や「で」を節に含める。これは、「に」や「で」を助詞とみなすことに相当する。以下の最後の例に示すように、「で」は助詞ではなく、判定詞「だ」のテ形の場合もあるが、この場合の「で」も、特別な助詞 (判定詞由来の助詞) として助詞扱いとする。

(11) a. 採決が終わった-B-**後**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
b. 採決が終わった-B-**後に**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
c. 採決が終わった-B-**後で**-C[、]-大勢の人が反対意見を言い始めた。 (副詞節)
d. 大勢の人が反対意見を言い始めたのは、採決が終わった-B-**後で**-C[、]-それが問題を引き起こした。　(並列節)

これに対して、副詞節の形式に判定詞「だ」の基本形が後続する場合は、判定詞の前を節境界とする。(判定詞は助詞ではないので、文節の主要部となる。)

(12) 大勢の人が反対意見を言い始めたのは、採決が終わった-B-**後**-C-だ。 (副詞節)

「〜せいで」は副詞節を作れるが、「〜せい」は作れないので、次のような扱いとする。

(13) a. 電車が止まった-B-**せいで**-C[、]-会議に行けなかった。 (副詞節)
b. 会議に行けなかったのは、電車が止まった-C-せいだ。 (連体節)
c. 会議に行けなかった。電車が止まった-B-**せいで**-C-だ。 (副詞節)

なお、「電車が止まったせいで (も) ある」の扱いは、現時点では保留である。いずれにしても、判定詞が後続する場合はすっきりしないことは免れない。

7.2　補足節と連体節

補足節か連体節かの判定も、後続文節を節末機能文節と認定するか否かに帰着させる。補足節を作る節末機能文節の主要部は、「の」「こと」「ところ」に限られるため、どのような助詞を伴うかが焦点となる。

(14) a. 花子は太郎がその店に入る-B-**ところを**-C-見かけた。 (補足節)
b. 太郎はその店に入る-C-ところで、花子はそれを見かけた。 (連体節)
c. 太郎は、店の勝手口に入る-C-ところで、花子と会った。 (連体節)
d. 太郎はその店に入る-C-ところだ。 (連体節)
e. あの勝手口が、太郎がお店に入った-C-ところだ。 (連体節)
f. 結婚する-B-**ことに**-C[、]-母が反対した。 (補足節)
g. 母が反対したのは、結婚する-B-**ことに**-C-だ。 (補足節)
h. 君にあげた-B-**のは**-C[、]-この指輪だ。 (補足節)
g. この指輪は、君にあげた-の-だ。 (見分けがつかない)

原則として、格助詞あるいは係助詞「は」「も」を伴う場合は補足節とみなし[3]、それ以外の場合は連体節とみなす。

[3] 理想的には、述語と格関係にある場合は補足節とみなすべきであるが、高い精度で機械的にそれが判定できるかどうかは不明である。

7.3 「という」「ような」

「という」と「ような」は、後続が形式名詞であっても、そこで区切る。ただし、これらの後ろが「の」の場合は、そこで切らない。

(15) a. 彼が書いたらしいという-C-ことが、 (連体節)
b. 彼が書いたような-C-ことは、 (連体節)
c. 論文を書くというのが-C-望ましい。 (補足節)
d. 似たようなのが-C-他にも 2 個以上ある。 (補足節)

「という」に接続助詞が後続する場合は、接続助詞までを節とする。

(16) a. 彼は書いたというし-C[、]- (連用節)
b. 彼は書いたというが-C[、]- (連用節)
c. 彼が書くらしいというので-C[、]-期待が高まった。 (副詞節)

8 BCCWJ コアデータへの節境界付与

Rainbow3 が前提としている文法体系は、益岡・田窪文法 [1] を文節文法に焼き直したものである。一方、BCCWJ の解析済データは、これとは異なる文法体系を前提としている。そのため、BCCWJ の解析済データに節境界を付与するためには、その不整合を吸収する必要がある。

現時点では、まず、長単位 (LUW) の TSV データを入力とし、それに節境界を付与することを先行させている。これは、LUW の TSV データには、文節区切りの情報が含まれていること、長単位認識のための解析が不要であることの 2 つの理由による。

LUW の TSV データに対して節境界を認定するための前処理 (不整合の吸収) は、おおよそ次の 2 種類に分類できる。

1. 品詞の付け替え
2. 単位の調整 (LUW と Rainbow3 の長い単位の語の不整合を調整)

現時点では、BCCWJ 体系から Rainbow 体系への変換は必要最小限に止めているが、その中で最も煩雑なのは、BCCWJ の助動詞の変換である。BCCWJ の助動詞の多くは、Rainbow 体系では、活用語尾、接尾辞、動詞 (複合動詞後件) 扱いとなる。

BCCWJ の文節と Rainbow3 の文節は、大体の場合は一致する。一致しないのは、主に、判定詞、助動詞、形式名詞が関わる場合と、複合辞が関わる場合である。これ以外に、テ形複合動詞の扱いが一部異なる。たとえば、「引返してゆく」を BCCWJ は 2 文節とみなすが、Rainbow3 では 1 文節とみなす。

残された問題は、節境界ラベルの設計である。応用の立場からは、意味的な節境界ラベルが望ましい。一方、認定処理の立場からは、ほぼ一意に決定できる境界ラベルが望ましい。現時点では、形式的な境界ラベル集合 (たとえば、「とき節」) と意味的な境界ラベル集合 (たとえば、「副詞節-時間」) の両方を設計し、それらの両方を付与する (意味的境界ラベルは可能な候補を付与する) 方針を立てている。

図 1 にサンプル PB12_00001 の冒頭部分の節境界認定例を示す。この図では、節境界を認定した部分で改行している。最右欄に示す節境界ラベルは仮のものである。

-S[ ]-パソコン-A-の-B-画面-A-や-B-本-A-など-j-に-B-集中し-A-ながら-C[、]-　ながら節
-C[、]-自分-A-の-B-入れ-k-た-C-　動詞-タ形連体形
-C-飲み物-A-に-B-手-A-を-B-伸ばし-C[、]-　動詞-連用形
-C[、]-飲み物-A-に-j-は-B-まったく-B-目-A-を-B-遣ら-n-ない-B-まま-C-　ママ節
-C-飲む-A[、]-という-B-の-A-は-C-　という-ノ節=は
-C-だれ-A-で-w-も-B-やる-B-こと-C-　コト節
-C-だろう-S[。]-　判定詞-意志推量形
-S-自分-A-で-B-入れ-k-た-B-のだ-A-から-C[、]-　から節
-C[、]-それ-A-が-B-なん-B-な-B-の-A-か-j-は-C-　ノ節=か-は
-C-見-n-なく-k-て-A-も-C-　イ形容詞-テ形=も
-C-わかる-S[。]-　動詞-終止形
-S-だから-J-たいがい-B[、]-なん-A-の-B-問題-A-も-B-ない-S[。]-　イ形容詞-終止形
-S[ ]-ところが-J[、]-ごくごく-B-稀-k-に-B[、]-変-k-な-B-こと-A-が-C-　コト節=が
-C-起こる-S[。]-　動詞-終止形
-S-たとえば-B[、]-紅茶-A-を-B-入れ-k-た-A-の-w-に-C[、]-　のに節
-C[、]-どう-w-いう-B-わけ-A-か-C[、]-　ワケ節=か
-C[、]-コーヒー-A-を-B-入れ-k-た-A-と-C-　と節
-C-勘違いし-t-てしまう-S[。]-　動詞-終止形
-S[ ]-手-A-を-B-カップ-A-に-B-伸ばす-S[。]-　動詞-終止形
-S-頭-A-は-B-コーヒー-A-を-B-入れ-k-た-A-と-C-　と節
-C-思い込ん-t-でいる-A-から-C[、]-　から節
-C[、]-口-A-は-B-すっかり-B-コーヒー-A-を-B-受け容れる-C-　動詞-連体形
-C-態勢-A-に-B-なっ-t-ている-S[。]-　動詞-終止形

図 1: 節境界認定例 (PB12_00001 の冒頭部分)

-j-を除く小文字の境界記号は、語内の境界を表す。-J-は接続詞の末尾の境界を表す。

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, 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.
[5] 佐藤理史. 境界認定の提案: (1) コンセプトと実現法. 情報処理学会自然言語研究会, NL–164, pp. 25–32, 2004.
[6] 佐藤理史. 境界認定の提案: (2) 背景と思想. 情報処理学会自然言語研究会, NL–164, pp. 33–44, 2004.
[7] 橋本進吉. 国文法体系論. 岩波書店, 1959.