節境界自動検出ルールの作成と評価

丸山 岳彦　　柏岡 秀紀　　熊野 正　　田中 英輝

ATR 音声言語コミュニケーション研究所

1 はじめに

近年，独話を対象とした自発音声コーパスの構築が進んでいる．講演やニュース，学会発表など，一人の話者が話し続ける独話は，対話よりも1文の長さが長くなったり文の構造が複雑化したりするという特徴を持つことが知られている [4]．さらに，自発的な発話になるほど，明示的な文末がはっきりしなくなり，文の境界を見つけることが困難になる [6]．

従来の自然言語処理技術では，「文」を基本的な処理単位とする場合が大半であった．しかし，1文が長く，文末が確定しにくいという性格を持つ独話を対象とする場合，文という処理単位は必ずしも適切ではない．文が長くなることによって構文解析の曖昧性が爆発するという問題や，文末がはっきりしないためにどこまで入力を待てばよいのか分からないという問題が発生するからである．さらに，独話を機械翻訳する場合，発話に追従して翻訳を出力する同時通訳としての運用が望ましいが [4]，文を処理単位とすると，同時通訳としての追従性に欠けるという問題がある．このため，発話の中から文よりも短い単位を随時検出し，各種の処理を漸進的に行なっていく手法が求められる．

我々は，文よりも短い処理単位として，述語を中心としたまとまりである「節」を提案してきた [2, 4]．節は，統語的・意味的にまとまった単位であり，翻訳や要約などの処理に有用であると考えられる．文を節に分割すると，(1) のような結果が得られる．

(1) その台湾の総統でありますけれども /
一応共産党との内戦に破れ /
国民党政権が台湾に移って以来 /
長い間その翼賛組織ともいうべき /
国民大会の場で選出されておりました。

我々は，形態素解析の結果のみを用いて，構文解析を行なうことなく，日本語の節境界を検出するルールを作成した．本稿では，我々が提案する節境界検出ルールの概要と性能について述べる．また，複数のコーパスに対して節分割を行なった結果を分析する．さらに，ルールの応用可能性についても述べる．

2 節境界検出ルール

「述語を中心としたまとまり」である節は，主節と従属節に分けることができる．従属節は，さらに次の4種類に分類することができる [3]．

補足節: 格助詞や引用形式を伴って，述語を補う．
太郎は，花子を町で見かけたことを思い出した．

副詞節: 述語を修飾したり，文全体を修飾する．
父はいつも新聞を読みながら朝食を食べる．

連体節: 名詞を修飾する．
太郎が撮った写真

並列節: 主節に対して対等に並ぶ関係で結びつく．
太郎は音楽が好きで，花子は映画が好きだ．

節境界を自動的に検出する手法としてまず考えられるのは，構文解析器を用いて文を解析した結果から，節境界に相当する位置を特定するという方法である．しかし，構文解析器は一般に入力として「文」を要求するものであり，文末が入力されて構文解析が済むまで，節境界の検出を始めることは難しい．この制約は，同時通訳のように入力を漸進的に処理していく必要がある場合，望ましくない．漸進的な処理を行なうためには，発話の入力中であっても，局所的な情報のみから節境界の位置を検出できることが望ましい．

日本語の節境界は，ある局所的な範囲内での形態素の連接を見ることによって，かなり正確に取り出すことができる．そこで我々は，形態素の局所的な連接関係のみを手がかりとして，構文解析を行なうことなく節境界を検出するルールを人手により作成した．

この節境界検出ルールは，節境界の位置を発見するための形態素列パタンと，節境界の種類を表す節境界ラベルの組から構成されている．我々はこのルールを Perl の正規表現置換の形式で実装した．入力を形態素解析し，各形態素を

出現形 _ 品詞 _ 活用形 _ 活用型

という4つ組に変換した列に対してこのルールを適用すると，節境界の位置にその節の種類を表す節境界ラベルが挿入される．ルールの例を図1に示す．

図 1: 節境界検出ルールの例

```
1. s/(が _ 助詞- 接続助詞 __) /$1 \/ 並列節ガ \/ /g;
2. s/((連用タ接続 | 連用形) たら _ 助動詞 _ 特殊・タ _ 仮定形) /$1 \/ 条件節タラ \/ /g;
3. s/((基本形 | 助詞- 副助詞／並立助詞／終助詞 __| 命令 ([^_]*)| 助詞- 終助詞 __| 感動詞 __) (と | って)_ 助詞- 格助詞-(引用 | 一般)__) /$1 \/ 引用節 \/ /g;
```

我々は，日本語形態素解析ツール「茶筌 Ver.2.2.9 [7]」の体系を用いて，361個のルールを作成した．全てのルールは，1～3個の連接する形態素から構成されるパタンを持つ．入力には読点が含まれていないことを想定し，パタンに読点は含めていない．

検出される節境界は，文献[3]に記述されている従属節の形態（補足節，副詞節，連体節，並列節）を増補・改編して作成したもので，合計144種類である．これには，統語的に大きな切れ目になると考えられる主題「は」，談話標識，感動詞を検出するパタンも含まれている．以下では，これらを含めて「節境界」と考えることにする．

このルールにより検出される節境界の分類と，挿入される節境界ラベルの種類を，以下に示す[1]．

補足節: 補足節，引用節，間接疑問節

副詞節: 連用節，形容詞連用節，理由節（カラ，ノデ），条件節（カギリ，ケッカ，タラ，ト，トコロ，ナラ，バ，モノノ），譲歩節（テモ，ノニ，命令形），時間節（アト，イマ，イライ，トキ，マエ），目的節，タメ節，ダケ節，ツツ節，テ節，ナガラ節，ナド節，ホカ節，ホド節，マデ節，ママ節，ヨウ節，ヨリ節

連体節: 連体節，連体節（カギリノ，タメノ，テノ，トイウ，ナドノ，ホドノ，マデノ，ヨウナ），形容詞連体節，形容動詞連体節

並列節: 並列節（ガ，ケレドモ，シ，タリ，デ，トカ）

その他: 感動詞，間投句，主題ハ，従属文，体言止，談話標識，文末

検出結果の例を図2に示す．節境界ラベルは，/境界名/という形で挿入されている．

[1] 実際には，各節境界がさらに細かく分類されている．例えば，「タメ節」の下位には「タメニ節」「タメニハ節」という節境界が設定してある．全てを合計すると144種類となる．

図 2: 節境界検出結果の例

自主避難が呼びかけられている /連体節/ 壮瞥町の滝之町地区では /主題ハ/ ほとんどの商店が休業する中 /連用節その他/ 営業している /連体節/ コンビニエンスストアに食料品などを買い求める /連体節/ 人が集中しています。 /文末/ 壮瞥町役場の近くで営業を続けている /連体節/ コンビニエンスストアには /主題ハ/ 車で十分あまりかかる /連体節/ 避難所からも住民が訪れ /連用節/ 食料品や生活用品などを買い求めています。 /文末/ この店では /主題ハ/ 有珠山の火山活動が活発になってから /テカラ節/ ミネラルウォーターや下着類を多めに仕入れたという /連体節トイウ/ ことです。 /文末/ しかし /談話標識/ 道路の通行止めで仕入の車が来れなくなったため /タメ節/ 弁当や生鮮食品が品切れになったほか /ホカ節/ 乾電池や生活用品なども在庫がほとんどなくなりました。 /文末/

3 実験

2節で示した節境界検出ルールの性能を評価するために，性質の異なる複数のコーパスに対してルールを適用し，結果を分析した．

3.1 対象コーパス

用意したコーパスは，以下の5種類である．

1. **『あすを読む』（ASU）**
   NHKの解説番組「あすを読む」を書き起こした独話コーパス．327番組分を収録している．
2. **NHKニュース原稿（NHK）**
   NHKニュースの原稿コーパス．1995年3月から2000年4月までの原稿を収録している．
3. **日本経済新聞（日経）**
   日本経済新聞・日経産業新聞・日経流通新聞・日経金融新聞の新聞記事データベース．1995年1月から2000年12月までの記事を収録している．
4. **バイリンガル旅行会話コーパス（SLDB）**
   ATRで作成された，旅行会話を題材とする模擬会話コーパス．618会話を収録している．
5. **旅行会話基本表現集（BTEC）**
   海外旅行で用いられる典型的な表現を収集したコーパス．

これら5種類のコーパスの規模を，表1に示す[2]．1文の長さは，**NHK**が突出して長く，**ASU**と日経がそれに続き，**SLDB**と**BTEC**は極端に短いことが分かる．

[2] 文節の認定には，日本語係り受け解析器「CaboCha/南瓜 Ver.0.21[8]」を用いた．

表 1: 各コーパスの規模

	形態素数	文数	形態素 / 文	文節 / 文
ASU	577.9 K	19.8 K	29.14	11.24
NHK	75.6 M	1.6 M	47.46	17.07
日経	499.5 M	18.6 M	26.83	8.92
SLDB	255.2 K	21.8 K	11.72	3.90
BTEC	1.4 M	174.2 K	7.87	2.71

3.2 結果

節境界検出ルールを各コーパスに対して適用し，節境界の検出を行なった．検出された節の数，1 文に含まれる平均節数，各節に含まれる平均形態素数と平均文節数を，表 2 に示す．

表 2: 節境界の検出結果

	節数	節 / 文	形態素 / 節	文節 / 節
ASU	93.0 K	4.69	6.21	2.40
NHK	10.2 M	6.41	7.40	2.66
日経	62.4 M	3.35	8.00	2.66
SLDB	45.2 K	2.08	5.64	1.87
BTEC	262.4 K	1.51	5.23	1.80

表2 から，一つの節の長さ（形態素数，文節数）は，コーパス間でほとんど差がないことが分かる．

3.3 評価

節境界検出ルールの性能を測定するため，各コーパスから 500 文を選択し，人手で節境界の検出と判定を行ない，正解データを作成した．検出ルールの結果と正解データとを照合し，適合率と再現率を求めた．結果を表 3 に示す．

表 3: 適合率と再現率

	適合率	再現率
ASU	97.49%	97.07%
NHK	97.02%	96.71%
日経	98.00%	97.24%
SLDB	98.91%	99.69%
BTEC	99.11%	98.01%

全てのコーパスにおいて，適合率と再現率ともに非常に高い精度で節境界が検出されていることが分かる．

3.4 問題点

以下では，節境界検出ルールの検出結果が誤っている場合について，分析を行なう．

3.4.1 形態素解析の誤りに起因する問題

まず，形態素解析の誤りが原因となって，誤った節境界が検出されているケースが多く見受けられた．例えば，次のようなものである．

(2) 借金があり /連用節/ その返済にあてるために /タメニ節/ 強盗には /主題ハ/ いった。/文末/ **(NHK)**

「はいった」は実際には動詞であるが，形態素解析の段階で「は」が「助詞 - 係助詞」と解析されたために，/主題ハ/ という節境界が検出されている．形態素解析が正しく行なわれることによって，検出ルールの性能はさらに向上すると考えられる．

3.4.2 検出が困難な節境界

節境界の位置を形態素列パタンのみから発見するという我々の手法では，検出することが困難な種類の節境界が見受けられた．以下では，形容詞述語，および名詞述語の問題について示す．

形容詞の連用形は，連用修飾要素として機能する場合と，述語として機能する場合がある．(3) は前者，(4) は後者の例である．

(3) 対等な関係で契約を結ぶためには /タメニハ節/ まだ課題が多く残されています。/文末/ **(ASU)**

(4) 今の借地借家法では /主題ハ/ 家主の立場があまりにも弱く賃貸住宅を貸そうという /連体節トイウ/ 意欲がわかないという /連体節トイウ/ ... **(ASU)**

(3) の形容詞「多く」は連用修飾要素であるが，(4) の形容詞「弱く」は述語として機能しており，その直後は節境界「連用節」として検出されるべき点である．しかし，局所的な形態素の連接関係のみから後者だけを選別することは困難である．なお，人手で作成した正解データ 500 文の中で，形容詞が述語として連用節を構成している場合は，**ASU** に 4 例，**NHK** に 3 例，**日経**に 2 例あった．

また，名詞は助動詞「だ」や補助動詞「する」を後続させることで述語になることができるが，助動詞・補助動詞を伴わない名詞が単独で述語位置に現れる場合がある．ここでは「名詞述語節」と呼ぶ．名詞述語節の境界を || で示すと，以下のようになる．

(5) 観光バスとトラックが正面衝突し /連用節/ トラックの運転手が死亡 || 旅行客五人が軽いケガをしました。/文末/ (NHK)

(6) コンチネンタル式が十ドル || 英国式が十二ドルとなっております。/文末/ (SLDB)

このような名詞述語節の境界の前後は，表層上は名詞の連続でしかなく，形態素の連接のみから節境界として検出することは困難である．なお，正解データ500文中に，名詞述語節は **ASU** に10例，**NHK** に30例，日経に12例，**SLDB** に10例あった．

4 応用

以下では，節境界検出ルールをどのように応用することが可能かについて述べる．

4.1 同時通訳のための文分割処理

節境界検出ルールによって出力された節境界を，文分割処理に応用することができる．節は，統語的にも意味的にもある程度まとまった単位であり，統語解析，翻訳，要約などの処理を効率的に進める上でも有用である[3]．特に1文が長くなる傾向を持つ独話を同時通訳する場面においては，入力される発話を漸進的に細かい処理単位に分割する過程が必要であるが，局所的な情報だけで決定可能であるという点で，節境界は有力な分割点候補として考えられる．

また，全ての節境界を分割点とするのではなく，発話の大きな切れ目になりやすいと考えられる節境界 — 例えば，ケレドモ節，ガ節，マシテ節など — のみを分割点として採用するという方略も考えられる [4]．統語的・意味的なまとまりを保持しつつ，分割結果の単位長を比較的自由に調整できるという点でも，節境界での分割は有用である．

4.2 句点のないコーパスを対象とした文認定処理

現在，『日本語話し言葉コーパス (CSJ: The Corpus of Spontaneous Japanese)[1]』という大規模な日本語自発音声コーパスの構築が進められている．このコーパスに納められている書き起こしテキストには句点が含まれておらず，通常の意味での「文」という単位が存在しない．このコーパスを対象として統語解析，翻訳，要約などの処理を行なうためには，前もって何らかの処理単位を検出しておく必要がある．このような場合には，局所的な形態素の連接のみから節境界を検出できる節境界検出ルールが有用である．句点の含まれないコーパスからも，節境界を検出し，それらを処理単位とすることができるからである．

CSJで採用されている形態素体系は，「茶筌」の形態素体系とは異なっているため，本稿で提示した節境界検出ルールを適用するためには，節境界を検出するパタンの仕様を変更する必要がある．そこで我々は，節境界検出ルールをCSJの形態素の仕様に改編し，CSJから節境界を検出するルールを作成して，発話の自動分割を行なった．この詳細については，文献 [6] を参照されたい．

[3] 節境界の検出が係り受け解析の効率化に及ぼす影響については，柏岡ほか [2] を参照．

5 まとめ

局所的な形態素列から日本語の節境界を検出するルールを作成し，その性能と問題点，および応用可能性について述べた．今後は，節境界の応用可能性とその妥当性について，より具体的に検討していきたい．

謝辞: 本研究は通信・放送機構の研究委託により実施したものである．

参考文献

[1] Maekawa, K., H. Koiso, S. Furui, H. Isahara Spontaneous Speech Corpus of Japanese. *Proceedings of LREC2000*, 947–952, 2000.

[2] 柏岡秀紀・丸山岳彦・田中英輝 2003. 節境界と係り受け解析. 『言語処理学会第9回年次大会発表論文集』, 言語処理学会.

[3] 益岡隆志・田窪行則 1992. 『基礎日本語文法 — 改訂版 —』. くろしお出版.

[4] 丸山岳彦・熊野正・柏岡秀紀 2001. 日本語における独話の特徴と文分割. 『言語処理学会第7回年次大会発表論文集』, 429–432.

[5] 南不二男 1974. 『現代日本語の構造』. 大修館書店.

[6] 高梨克也・丸山岳彦・内元清貴・井佐原均 2003. 話し言葉の文境界 —CSJコーパスにおける文境界の定義と半自動認定 —. 『言語処理学会第9回年次大会発表論文集』, 言語処理学会.

[7] 日本語形態素解析システム ChaSen 「茶筌」 (奈良先端科学技術大学院大学 松本研究室)

[8] 日本語係り受け解析器 CaboCha 「南瓜」 (奈良先端科学技術大学院大学 松本研究室)