Internal Use Only（非公開）

TR－SLT－0063

# 日本語節境界検出プログラム CBAP の開発と評価
# Japanese Clause Boundary Annotation Program
## --- The Development and Evaluation ---

丸山 岳彦　　柏岡 秀紀
MARUYAMA, Takehiko　　KASHIOKA, Hideki
熊野 正
KUMANO, Tadashi

２００４年３月２３日

## 概要

従来の文分割研究において，文の分割点として利用されてきたのは，「節」の境界である．しかしながら，実際に文の分割点として用いられる節境界はごく一部の種類のものに限られており，文に含まれる節境界を網羅的に検出する手法は考えられてこなかった．我々は，日本語の文に含まれる節境界の位置を網羅的に検出し，その種類を特定するプログラム ``CBAP (Clause BoundaryAnnotation Program)" を開発した．CBAP は，形態素解析の結果を入力とし，局所的な形態素の連接を対象としたパタンマッチによって，147 種類の節境界を検出する．CBAP を性質の異なる 5 種のコーパスに適用したところ，いずれのコーパスでも 97¥% 以上の検出性能が確認された．この検出結果を利用することにより，言語学的に意味のある文の分割点を特定することができ，従来の手法よりも柔軟に文分割を行なうことができる．また，1〜3 形態素という非常に局所的な範囲のみから節境界を検出できるため，発話に追従して処理を進めていく漸進的構文解析や同時通訳システム，また，句点を含まない音声コーパスを対象とした発話分割処理などに有用である．本稿では，CBAP による節境界の検出手法を示し，節境界を用いて文分割・発話分割処理を行なった事例をもとに，節境界検出の有用性を述べる．

（株）国際電気通信基礎技術研究所
音声言語コミュニケーション研究所
〒619－0288「けいはんな学研都市」光台二丁目 2 番地 2 TEL：0774－95－1301

Advanced Telecommunication Research Institute International
Spoken Language Translation Research Laboratories
2-2-2 Hikaridai "Keihanna Science City" 619-0288,Japan
Telephone:+81-774-95-1301
Fax　　:+81-774-95-1308

# 目次

# 表一覧

# 1 はじめに

従来の言語処理技術において，その処理単位として利用されてきたものは，基本的に「文」であった．文が処理単位として用いられてきた理由としては，文が言語活動の基本的な単位として考えられること，書きことばでは句点で区切られているためにその境界が明確であること，統語的・意味的に完結していると考えられること，などが挙げられる．

しかしながら，文は場合によって処理単位としては長すぎることがある．例えば，テレビニュースや講演などの話しことばでは1文の長さが顕著に長くなる傾向がある [1, 2]．明示的な文末表現が現れるまでが非常に長い (1) のような1文を，構文解析や機械翻訳，重要文抽出の処理単位として採用することは，妥当ではない．

(1) 身長の二乗掛ける二十二が標準体重ということになっとりまして私の標準体重は六十四キロなんですがそれから見ると約七キロぐらいは減量が必要ということで運動をする方がいいことになってまして本当は食事量を減らすということなんでしょうけどなかなかそれは難しいので私は専ら運動の方で健康を維持しようということに努めとります。

そこで従来，長い文の処理を行なうための前処理として，長文を短く分割する研究が行なわれてきた．2 節で詳しく見るように，これらの研究のほとんどは，特定の「節」の境界を分割候補点としてリストアップしておいた上で，ある条件を満たした場合にそれらを分割位置として採用する，というものである．「述語を中心としたまとまり [3]」と定義される節は，統語的も意味的にもある程度のまとまりを備えていることから，局所的に構文解析を行なったり [4]，部分的に翻訳結果を出力したり [5]，あるいは重要文抽出の抽出単位に利用したりすることができると考えられる．しかしながら，従来の文分割研究で利用されている節境界は特定の種類のものに限られており，節境界を網羅的に利用するという手法は考えられてこなかった．

文中から節境界の位置と種類を網羅的に検出することは，形態素解析や文節解析と同様，言語学的に意味のある単位を検出する技術の1つであると言える．ところが，節境界の種類を体系的に分類し，それらを自動的に検出する手法について論じた研究は，管見の限り見当たらない．一方，日本語の記述的な文法研究においては，日本語の節境界が形態的・機能的な特徴から体系的に分類されている．本稿では，日本語の記述的文法研究の知見をもとに，日本語の節境界の位置を網羅的に検出し，その種類を特定するための手法を提案する．

節境界を検出する手法としてまず考えられるのは，既存の構文解析器を用いて文を解析し，その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．ところが，入力文が長くなればなるほどそもそもの解析に失敗してしまうという根本的な問題がある．節の終端に述語句が配置される構造を持つ日本語の場合，述語句部分の形態的情報（述語の活用形や接続助詞の種類など）をパタンとして記述しておけば，構文解析を行なわなくても，単純なパタンマッチによって節境界の位置と種類とをかなり正確に検出することができる．我々は，形態素の局所的な連接を対象としたパタンマッチによって，節境界の位置を網羅的に検出し，その種類を特定するプログラム “CBAP (Clause Boundaries Annotation Program: 節境界検出プログラム)” を開発した．このプログラムの利点としては，日本語の節境界を形態素の連接パタンとして記述してあるため，文長・文体の違いやニュース・旅行会話などドメインの違いに関わらず高い精度で節境界を検出できること，節境界を局所的な範囲のみから決定的に検出できるため，発話に追従して処理を進めていく漸進的構文解析や同時通訳システムなどに向くこと，日本語の記述的文法研究の知見をもとに人手でルールを作成してあるため，言語学的に妥当な情報を大量の学習用データを用いずに検出できることなどが挙げられる．

以下，論文構成について述べる．まず2節では，これまで行なわれてきた節境界を利用した研究について概観した後で，日本語の述語句を形態素列のパタンとして捉えることにより，節境界の位置およびその文法的な特性を正確に把握できることについて述べる．3節では，日本語の記述的文法研究から見た節境界の分類を提示した上で，本稿で

扱う節境界を定義し，さらに我々が開発した“CBAP”による節境界の検出手法を示す．4節では，CBAPによる節境界の検出性能を評価した結果について示し，ベースライン手法との比較，さらに節境界を検出するルールの使用状況について示す．5節では，節境界が誤って検出された例について分析を行なう．6節では，CBAPの出力を利用して文を分割することにより，異種コーパスに含まれる文長のばらつきが均質化されること，節境界を選択的に利用して文分割を行なうことにより，従来の手法よりも柔軟に，かつ言語学的にも妥当な文分割が行なえることを示す．さらに，CBAPを『日本語話し言葉コーパス[6]』の発話分割処理に適用した事例をもとに，節境界検出の有用性を示す．

## 2 先行研究

### 2.1 文分割研究における節境界の検出手法

従来，特にニュース原稿などを対象とした処理において常に問題となってきたのは，長文の処理であった．1文が長くなればなるほど構文解析の曖昧性が増大し，解析誤りや翻訳誤りの原因となるため，その前処理として長文を短く分割する手法が検討されてきた．

例えば，武石・林 (1992) は，長文の推敲機能を実現する目的から，従属節の接続表現の違いに基づいて「接続構造の解析」を行ない，接続構造の違いによって文の分割位置を決定するという文分割手法を提案している [7]．また，金・江原 (1993)，木村他 (1989) らは，長文によって構文解析や機械翻訳の精度が低下するという問題を避けるために，日英機械翻訳の際の前編集として文分割を行なっている [8, 9]．金らは，ニュース文に現れる接続法を「連用中止法」「引用法」「連体節法」などに分類し，あらかじめ人手で用意した分割パタンとのパタンマッチを最長一致法で行なうことで，文の分割位置を決定している．さらに，福島他 (1999) は，自動要約・重要文抽出の観点から，文分割が自動要約に与える影響の調査を行なっている．福島らは文分割の条件として19の規則を用意し，文が一定の長さを超えた場合に文分割を行なった上で，このような文分割が自動要約における情報の欠落や文字数圧縮に効果があることを示している [1]．福島らの規則によると，(2) の1文は3文に分割される（例文は福島ら）．

(2) 千葉市に本店がある京葉銀行の成田西支店の女子行員が，他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まり，京葉銀行ではきょう，この女子行員を懲戒解雇するとともに，千葉県警察本部に被害を届け出ました．

1. 京葉銀行の成田西支店の女子行員が，他人名義のカードローンを悪用しておよそ三億円を着服していた疑いが強まりました．
2. そして，京葉銀行ではきょう，この女子行員を懲戒解雇しました．
3. それとともに，千葉県警察本部に被害を届け出ました．

これらの研究に共通しているのは，文分割位置の候補として文中に含まれる「節」の境界が利用されているという点である．統語的に大きな切れ目になる可能性のある特定の従属節（「連用中止形」，「〜が」，「〜ので」など）をあらかじめリストアップし，それらを検出するためのルールを人手で作成しておいた上で，ある条件を満たした場合に当該の節境界を文の分割位置として採用する，という手法である．

### 2.2 機械学習に基づく節境界検出

上記に見たように，日本語の節境界を自動的に検出する研究は人手で作成したルールによるものが大半であった．一方，英語を対象とした節境界の検出に関する研究では，人手で用意したルールによる検出だけでなく，機械学習による検出手法も見られる．人手で作成したルールによって英語の節境界の検出を行なっているものとしては，音声合成の韻律単位として節境界を利用した Ejerhed (1988) [10]，漸進的構文解析のために節境界を利用した Abney (1990) [11]，パラレルテキストのアライメントに節境界を利用した Papageorgiou (1997) [12]，機械翻訳に節境界を利用した Leffa (1998) [13] などが挙げられる．これらの研究では高い精度で節境界が検出されているが，人手でルールを作成するのにコストがかかるという問題がある．

一方，近年では，機械学習によって英語の節境界を検出するという研究も見られる．例えば，ACL2001 の併接ワークショップとして行なわれた CoNLL2001 では，shared task として “Clause Identification” が採用されており [14][1]，

[1] http://cnts.uia.ac.be/conll2001/clauses/

PennTreeBank-2 に含まれるデータを対象として，「節の始点の検出」「節の終点の検出」そしてその両者から導かれる「節全体の同定」という 3 つのタスクが挙げられている．節の同定には，各語に付与された品詞タグ [15] とチャンクタグ [16]，そして PennTreeBank に含まれる clause タグが利用される．

このタスクに対して 6 システムが参加し，AdaBoost アルゴリズム，隠れマルコフモデル，用例ベースモデル，コネクショニストモデルなどによるアプローチが提出された．6 システム中もっとも成績が良かったのは AdaBoost アルゴリズムによる Carreras (2001) [17] のシステムで，3 つのタスクの F 値は，「節の始点の検出」が 91.72，「節の終点の検出」が 89.22，「節全体の同定」が 78.63 であった．彼らは節の始点および終点を検出するために，節境界候補点前後の文脈情報，1 文全体のチャンクタグ，コンマやピリオドなど大局的な情報を素性として利用しており，ベースライン手法，および他の 5 システムよりもよい結果を得ている．

### 2.3 記述的文法研究から見た節境界の分類

さて，日本語では述語句が形態的に発達しており，表層の形態素列から豊富な文法情報を獲得することができる．また，英語の場合とは異なり，日本語の述語句は節の終端に配置されるため，述語の活用形や接続助詞の種類などをパタンとして記述しておけば，1 文全体を構文解析したり，正解データを大量に用意して機械学習を行なったりすることなく，節の終端境界の位置およびその文法的な特性をかなり正確に把握することができる．さらに，日本語の記述的文法研究では節境界が形態的および機能的な特徴から詳細に分類されており，節境界を検出するパタンを作成する際の指針として利用できる．以上のような状況を考え合わせると，節境界の検出ルールを人手で作成する条件として，日本語は比較的有利な立場にあると言ってよい．

そもそも，2.1節で見たような長文の分割処理に関して，ある種の節境界が統語的に大きな切れ目として文の分割位置になり得るということは，日本語の記述的文法研究において早い時期から指摘されてきた．従属節の種類とその従属度との対応関係を論じた三上 (1953,1959,1963) [18, 19, 20] や南 (1974,1993) [21, 22] などは，その代表である．例えば三上 (1953 : p.182) は，「代表的な活用形の平均的陳述度を分数値で表し」たもの，すなわち従属節の種類と文中の切れ目の大きさとの関係として，表 1 に示すような対応を示している．

表 1: 従属節の種類と切れ目の大きさの対応

| | | |
|---|---|---|
| 中止連用形 | 何々シ（テ） | 1/4 |
| 連体形 | 何々シタ | 1/2 |
| 仮定形 | 何々スレバ | 3/4 |
| 終止形 | 何々シタ，セヨ | 1 |

さらに，「連用補語を食止めるか否か」「連体として収まるか否か」「普通体を丁寧体に変更することがふさわしいか否か」という統語的テストによって，従属節（三上は「活用形」と呼ぶ）の各種類を「単式」「軟式」「硬式」という 3 つのクラスに割り当てている．これらのクラスは「句切りの力」の違いに対応づけられており，後に述べる南 (1974,1993) による従属節の分類に引き継がれていくことになる．

また，三上 (1963 : pp.73–74) では，長文内に現れる「終止法（「が」「けれども」などに前接する述語の活用形）」と文分割の関係について，次のように述べている．三上の記述は，無論，言語の機械処理を念頭に置いたものではないが，今日の言語処理研究の観点から見ても，大きな示唆を含んでいると思われる．

> 文の途中の終止法は，次の二つの特徴を持っている．
>
> 1) 文をそこで切ることができる．

2) 全文を丁寧調にするとき，文末に次いで，そこが丁寧化する．

（中略）(1) は終止法の定義ともいうべき重要な性質である．文中の終止法をそのまま言い切りにしてピリオド（句点）を打ち，次をソレニ，ダカラ，シカシなどの接続詞を使わなくてすむこともある．そんな場合には，文を切らないことが悪文を生みやすい．

幸い，全員〔四人〕乗ることができたのです ga，すわる座席がないので通路にゴザをひいて，
すわって東京までいったのです ga，車中の人々は，どこかの団体の人たちだったのです ga，
ケンカをする，大声でさわぐ，人を茶かす，また，話の内容と言えば最低で...

次に，南 (1974,1993) [21, 22] は，従属節の内部に現れ得る要素の範囲，および従属節相互の包含可能性が従属節の種類によって異なるという統語的な事実に注目し，従属節が備える「文らしさ」という観点から，従属節を「A 類」「B 類」「C 類」という 3 つのクラスに分類した．

**A 類:** つつ，て（様態），ながら（継続），連用形反復...

**B 類:** て（理由，原因），と，ながら（逆接），ので，連用形...

**C 類:** が，から，けれど，し，て，連用形...

A 類から C 類に進むほど従属節内部に現れ得る要素の範囲は広くなり，それに応じて主節への従属度が低く（主節からの独立度が高く）なる．つまり，C 類に近づくほど従属節に「文らしさ」が備わっていくということであり，その節境界を文の分割位置として利用できる可能性が高くなることを示している．

南による分類・記述は，従属節の形態ごとに従属度が対応づけられている点で，工学的にも利用しやすいものとなっている．例えば先述の武石・林 (1992) による接続構造の解析は，南の従属節の分類がもとになっている．また，白井他 (1995) [23] は，南の分類を改編して述語句表現を細分類することにより，従属節（用言）間の係り受け解析の精度を大幅に向上させている．

先に見た三上による観察・記述は，長文の分割処理だけでなく，話しことばを書きことばに変換するための言い換え処理や，推敲支援などの言語処理技術に対しても有用であると考えられる．また，南による従属節の形態と従属度の対応付けは，文分割や係り受け解析の精度向上に有効な知見を提供している．これらの点において，文中に含まれる節の境界を網羅的に検出し，その種類を特定しておくことは，形態素解析や文節解析と同様，言語処理の様々な側面において有用な技術であると言える．

## 3 節境界検出プログラム “CBAP” の開発

我々は，形態素解析された日本語テキストを入力として，文中の節境界を網羅的に検出するプログラム “CBAP (Clause Boundary Annotation Program: 節境界検出プログラム)” を開発した．以下では，本稿における節境界の定義，およびCBAPによる節境界の検出手法について述べる．

### 3.1 本稿での節境界の定義

節とは「述語を中心としたまとまり [3]」と定義される文の構成要素である．本稿での節の定義もこれに従う．節は，文末に配置される「主節」とそれ以外の「従属節」とに分けることができる．従属節の形態および機能に着目して節の種類を体系的・網羅的に記述した文法研究には，寺村 (1981)，北條 (1989)，益岡・田窪 (1992) [24, 25, 3] などがある．益岡・田窪 (1992) では，形態的および機能的な観点から，従属節（益岡らは「接続節」と呼ぶ）が次の4種類に分類されている．

**副詞節:** 述語または文全体を修飾する．（健はテレビを見ながら夕食を食べる）

**補足節:** 格助詞や引用形式を伴い，述語の補足語になる．（健は凛を見たことを思い出した）

**連体節:** 名詞を修飾する．（健が撮った写真）

**並列節:** 主節に対して対等に並ぶ関係で結びつく．（健は音楽が好きで凛は映画が好きだ）

我々は，上記4種の従属節を大分類とした上で，表2に示す10種類のクラスを小分類として設けた．これらのクラスは，益岡・田窪 (1992) による形態的・機能的な分類を参考にして作成したものである．表2の括弧内の数字は，各クラスに登録された節境界の種類の数を示す．

表 2: 節境界の大分類と小分類

| 大分類 | 小分類 |
|---|---|
| 副詞節 (102) | 条件・譲歩節 (23)，原因・理由節 (8)，時間節 (21)，様態節 (12)，副詞節その他 (38) |
| 補足節 (10) | 補足節 (2)，引用節 (5)，間接疑問節 (3) |
| 連体節 (15) | 連体節 (15) |
| 並列節 (12) | 並列節 (12) |

これら10種の小分類に対して，益岡・田窪 (1992) に記述されている節境界の表現をすべて登録した．例えば，「～すれば」という節境界の表現は，「条件・譲歩節」のクラスに「条件節レバ」という節境界名で登録した．それぞれの節境界名は，その形態的特徴，および節間で結ばれる関係的意味という文法的な特性の違いを考慮して，「理由節カラ」「連体節ヨウナ」「並列節ケレドモ」のように表現した．

さらに，益岡・田窪 (1992) には記述されていない節境界の表現をカバーするため，複数のコーパスを観察し，登録すべき節境界の種類を内省により決定・追加しながら，調整を行なった（4.1 節参照）．その結果，登録した節境界の合計は139種類となった．全10種類のクラスに登録した139種類の節境界のリストを，稿末の Appendix. に示す．

### 3.2 CBAP による節境界検出手法

以下では，CBAP による節境界の検出手法について述べる．CBAP 本体は，「節境界検出ルール」の集合から構成されている．節境界検出ルールは，節境界の位置を検出・特定するための「形態素列パタン」と，節境界の種類をラベリングするための「節境界ラベル」の組から構成される．ルールの実装は Perl で行なった．形態素解析した日本語入力文に対して CBAP を適用すると，パタンマッチによって節境界の位置が特定され，その種類を表す節境界ラベルが挿入される．形態素解析には，日本語形態素解析システム「茶筌[2]」を用いた．形態素列パタンは「茶筌」で用いられている日本語辞書 IPADIC の品詞体系に準じて記述されている[3]．「茶筌」の解析結果から「出現形」「品詞」「活用形」「活用型」という 4 つの情報を取り出し，1 つの形態素を「出現形 _ 品詞 _ 活用形 _ 活用型 」という形式で表現した．

節境界検出ルールの例を，図 1 に示す．例えばルール 1. は，出現形が「けれども」かつ品詞が「助詞 - 接続助詞」という 1 形態素から成る形態素列パタンを発見したら，その直後に <並列節ケレドモ> という節境界ラベルを挿入する，というものである．

```
1. s/(けれども _ 助詞 - 接続助詞 __) /$1 <並列節ケレドモ> /g;

2. s/(基本形 よう _ 名詞 - 非自立 - 助動詞語幹 __ な _ 助動詞 _ 特殊・ダ _ 体言接続 /$1 <連体節ヨウナ> /g;

3. s/((基本形 | 助詞 - 副助詞／並立助詞／終助詞 __| 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;
```

図 1: 節境界検出ルールの例

我々は，332 個の節境界検出ルールを人手により作成した．全てのルールは，1 ～ 3 個の形態素から構成される形態素列パタンを含んでいる．文末の検出には句点を利用しているため，入力文中には句点が含まれていることを前提としている[4]．検出対象となる節境界は先に示した 139 種類であるが，これに加えて，次の要素を検出するためのパタンも準備した．

表 3: CBAP の検出対象（その他）

| 大分類 | 小分類 |
|---|---|
| その他 (8) | 文末，主題ハ，談話標識，体言止，間投句，感動詞，従属文，従属文その他 |

このうち，「文末」以外の要素は厳密には節境界ではないが，統語的に大きな切れ目になると考えられるため，検出対象に含めた．全てのラベルを合計すると，147 種類になる．

冒頭に挙げた例 (1) に対して CBAP を適用すると，文中に含まれる節境界の位置と種類が検出され，節境界ラベルが挿入された (3) のような出力を得る．

(3) 身長の二乗掛ける <連体節> 二十二が標準体重ということになっとりまして <テ節> 私の標準体重は <主題ハ> 六十四キロなんですが <並列節ガ> それから見ると <条件節ト> 約七キロぐらいは <主題ハ> 減量が必要という

[2] http://chasen.aist-nara.ac.jp/

[3] ただし，CBAP による節境界の検出手法そのものは，特定の品詞体系に依存するものではない．6.3 節参照．

[4] 句点の含まれていないテキストに対する処理については 6.3 節を参照．

ことで <並列節デ> 運動をする <連体節> 方がいいことになってまして <テ節> 本当は食事量を減らすという <連体節トイウ> ことなんでしょうけど <並列節ケレドモ> なかなかそれは <主題ハ> 難しいので <理由節ノデ> 私は <主題ハ> 専ら運動の方で健康を維持しようという <連体節トイウ> ことに努めとります。 <文末>

# 4 評価

性質の異なる5種類のコーパスに対してCBAPを適用し，節境界の検出を行なった．その結果を人手で作成した正解データと比較することによって，CBAPによる節境界検出の性能を評価した．

## 4.1 対象コーパス

対象として用いたのは，以下の5種類のコーパスである．各コーパスの概要と例文を示す．

「あすを読む」（**ASU**）：NHKで放映されている10分間のニュース解説番組「あすを読む」の発話を書き起こしたコーパス．327番組分を収録している．

(4) 今晩は。ニュースでもお伝えしていますが冷えきっていました日本と北朝鮮朝鮮民主主義人民共和国との間の対話が動き始めました。今回の合意の背景と今後の課題を中心に整理したいと思います。

「**NHKニュース原稿**」（**NHK**）：NHKで放送される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，31,7205記事を収録している．

(5) 三十一日のニューヨーク外国為替市場は、ロンドン市場で円高が進んだ流れを引き継いで円を買う勢いが強まり、円は一時一ドル百二円五銭まで円高ドル安が進みました。円が一ドル百二円台で取引されるのは、今年初め以来ほぼ三ヶ月ぶりのことです。

「**日本経済新聞**」（**Nikkei**）：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．

(6) 科学技術庁無機材質研究所は水に溶けるフィルム状の複合材料を開発した。粘土を寒天やでんぷんと混ぜ合わせ、乾燥して作る。材料は半透明でセロハン紙並みの強度を持つ。水に入れると溶解し、リサイクルも可能という。

「**バイリンガル旅行会話コーパス**」（**SLDB**）：旅行会話を題材とする模擬対話を収録し，その発話を書き起こしたコーパス．618対話を収録している．[26]

(7) — ありがとうございます、ニューヨークシティホテルでございます。
— もしもし、わたし田中弘子といいますが、そちらのホテルの予約をしたいのですが。

「**旅行会話基本表現集**」（**BTEC**）：海外旅行の場面で用いられる典型的な対話表現を収集したコーパス．[27]

(8) — 旅行の目的は何ですか。
— 観光です。一か月です。

なお，3.1節で述べたように，節境界検出ルールを作成する際に検出対象としてまず登録したのは，益岡・田窪(1992)に記述されていた節境界の表現であった．さらに，益岡・田窪(1992)には記述されていない節境界の表現をカバーするため，検出対象とすべき節境界の増補を内省により行なった．さらに**ASU**を中心に5つのコーパスの観察を行ない，必要に応じて検出対象の追加を行なった．全体的な調整を行ない，新たに追加すべき節境界がないと判断した時点で，検出対象とする節境界の種類を確定した．

### 4.2 評価

CBAP による節境界の検出性能を評価するため，人手で節境界の位置と種類を認定した正解データを作成した．評価用データは，5 つのコーパスから，節境界検出ルールを作成する際に参照していない 500 文，合計 2,500 文を選択した．CBAP の検出結果と正解データとを照合し，節境界の位置が正しく検出されているか，および節境界ラベルが正しく付与されているかという 2 点に関して評価し，精度と再現率を求めた．また，両者の総合的な指標として，F 値を求めた．F 値は次の式で，$\beta$ =1 として計算した．結果を表 4 に示す．

$$F = \frac{(\beta + 1)PR}{\beta P + R}$$

表 4: CBAP による精度，再現率，F 値

| | 精度 | 再現率 | F 値 |
|---|---|---|---|
| **ASU** | 97.28% (2185/2246) | 97.50% (2185/2241) | 97.39 |
| **NHK** | 97.50% (2417/2479) | 97.15% (2417/2488) | 97.32 |
| **Nikkei** | 97.93% (1800/1838) | 97.24% (1800/1851) | 97.59 |
| **SLDB** | 97.13% (947/975) | 98.75% (947/959) | 97.93 |
| **BTEC** | 98.17% (699/712) | 99.57% (699/702) | 98.87 |

いずれのコーパスでも 97% 以上という高い結果が得られている．また，各コーパスの結果に顕著な差は見られない．つまり，CBAP は，話しことばと書きことばの違い，解説・ニュース・新聞記事・旅行会話などのドメインの違い，文長や文体などの違いによらず，常に安定した性能で節境界を検出することができていると言える．この結果は，日本語は述語部分が形態的に発達しており，局所的な形態素列のみから節の境界と種類を正確に把握できる，という本稿での見方を裏付けるものと考えられる．

### 4.3 ベースライン手法との比較

次に，評価用のベースラインとして，5 つのコーパスに対して文節解析を行ない，動詞または助動詞を含む文節の末尾を節境界の位置と見なしたデータを作成した[5]．ただし，正解データには「主題ハ」「談話標識」「感動詞」などの境界も含まれているため，動詞・助動詞を含む文節の終端のみを節境界と見なしたデータとは単純に比較することができない．そこで，表 3 に示した「その他」の境界うち，文末以外の境界を正解データから削除した上で，ベースラインとの比較を行なった．また，ベースライン手法では節境界の種類を特定することができないため，節境界の位置についてのみ評価を行なった．ベースライン手法による精度と再現率，F 値を表 5 に示す．
正解データおよび検出対象が異なっているため，表 4 に示した CBAP の検出性能と単純に比較することはできないものの，少なくともベースライン手法では表 4 を上回る結果は 1 件も見られなかった．さらに CBAP は，節境界の位置の検出だけでなく，その種類の特定まで行なっている．この点において，CBAP はベースライン手法よりも優れた検出性能・検出能力を有していると言える．

### 4.4 節境界検出ルールの使用状況

最後に，節境界検出ルールの使用状況について見ておく．3.2 節で述べたように，CBAP は人手で作成された 332 個の節境界検出ルールから構成されている．これらのルールの使用状況を確かめるために，2,500 文の評価用データ

[5] 文節の認定には，日本語係り受け解析器「南瓜 Ver.0.21」を用いた．
http://cl.aist-nara.ac.jp/~taku-ku/software/cabocha/

表 5: ベースライン手法による精度，再現率，F 値

| | 精度 | 再現率 | F 値 |
|---|---|---|---|
| **ASU** | 80.83% (1383/1711) | 83.82% (1383/1650) | 82.30 |
| **NHK** | 81.72% (1614/1975) | 82.05% (1614/1967) | 81.88 |
| **Nikkei** | 67.05% (871/1299) | 61.77% (871/1410) | 64.30 |
| **SLDB** | 90.55% (613/677) | 88.58% (613/692) | 89.55 |
| **BTEC** | 94.90% (502/529) | 89.96% (502/558) | 92.36 |

に含まれる節境界を検出する際，332 個のルールのうちどれだけが使われたかを調べた．さらに，処理対象となる文数が増加するにしたがって，検出に使われるルールの数がどのように増加するかを調べた．結果を図 2 に示す．

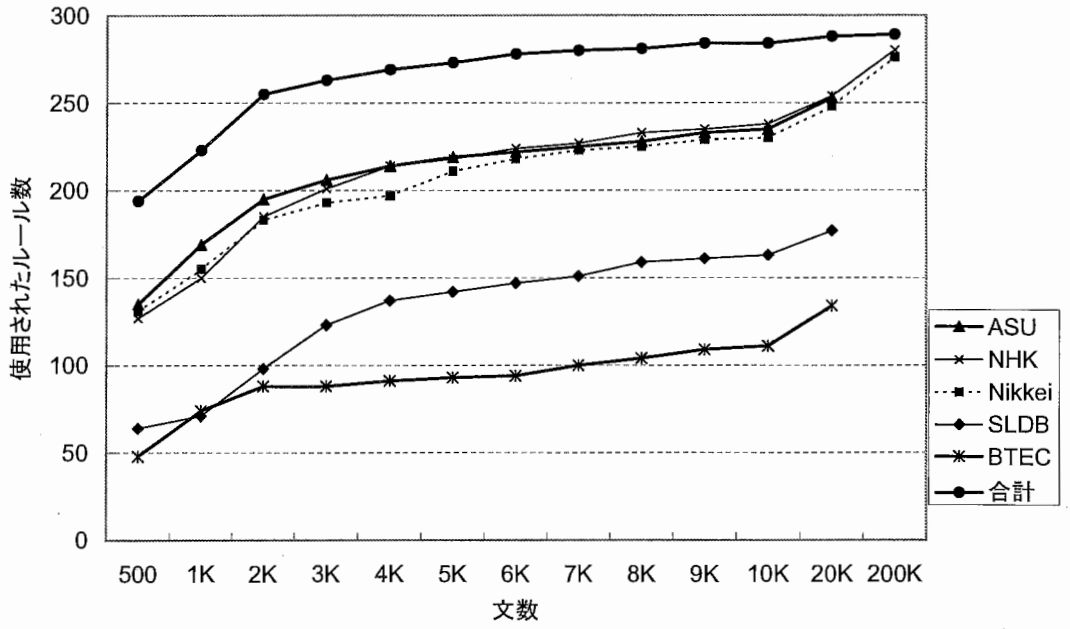


図 2: データサイズの増加と使用された節境界検出ルールの数

各コーパスから 500 文を選択した評価用データを対象とした場合，実際に使われたルールの数は，**ASU**，**NHK**，**Nikkei** で約 130 個，**SLDB** と **BTEC** では約 50 個であった．また，2,500 文の評価用データ全体で使われたルールの数は 194 個であった．さらに，処理対象となる文数が増加するにつれて使われるルールの数は増え，各コーパスそれぞれ 2 万文を処理対象とした場合には，**ASU**，**NHK**，**Nikkei** で約 250 個，**SLDB** で 177 個，**BTEC** で 134 個のルールが使われていた．計 10 万文全体の処理に使われたルール数は，合計 288 個であった．

この結果から，処理対象となる文数が増えるほど多様な節境界の表現が出現し，それに応じたルールが使われていると見ることができる．節境界検出ルールは部分的に内省によって作られているため，出現頻度の低い節境界にも対応することが可能になっている．出現頻度の低い節境界に対しては，文法論的な知見を取り入れて人手で作成したルールによる対処が効果的であると言える．

## 5 検出誤りの分析

以下では，CBAP が節境界を誤って検出した場合について，その原因を分析する．まず，評価用データ 2,500 文全体における検出誤り箇所の内訳を，表 6 に示す．

表 6: 検出誤り箇所の内訳

| 頻度 | 検出誤りの内容 |
|---|---|
| 133 (41.2%) | 形態素解析の誤りに起因する検出誤り |
| 77 (23.8%) | そもそも検出することが困難な節境界 |
| 57 (17.7%) | 節境界検出ルール自体の問題 |
| 46 (14.2%) | 複合動詞に関する検出誤り |
| 10 (3.1%) | その他 |
| 323 (100%) | 合計 |

以下では，これらの原因ごとに例を挙げながら分析を行なう．

### 5.1 形態素解析の誤りに起因する検出誤り

節境界が誤って検出された例のうち，最も多かったのは形態素解析の誤りに起因するもので，2,500 文中 133 例あった．これは，全検出誤りの 41.2% に相当する．この 133 例について，形態素解析誤りの内訳と，それが節境界の検出に及ぼす影響を，表 7 に示す．

表 7: 節境界の検出誤りに関わる形態素解析の誤り

| 頻度 | 形態素解析誤りの内容 | 節境界検出への影響 |
|---|---|---|
| 28 (21.1%) | 助動詞デを格助詞デと誤解析 | <並列節デ> 未検出 |
| 25 (18.8%) | 格助詞デを助動詞デと誤解析 | <並列節デ> 誤検出 |
| 15 (11.3%) | 単語境界が誤って解析され，本来は存在しない述語句が発生 | 本来は存在しない節境界を誤検出 |
| 12 (9.0%) | 格助詞トを接続助詞トと誤解析 | <引用節> を <条件節ト> と誤検出 |
| 11 (8.3%) | 名詞を動詞と誤解析 | <連用節> 誤検出 （「お隣り」など） |
| 8 (6.0%) | 連体詞を動詞と誤解析 | <連体節> 誤検出 （「ある」など） |
| 8 (6.0%) | 接続助詞トを格助詞トと誤解析 | <条件節ト> を <引用節> と誤検出 |
| 26 (19.5%) | その他 （未知語など） | — |
| 133 (100%) | 合計 | |

以下，形態素解析誤りが原因となって節境界が誤検出された例を示す．

(9) 貿易収支は <主題ハ> 史上最悪の水準 で しかも外国から大量に入ってきた <連体節> 資金でこれを賄う <連体節> 形となっています。 <文末> **(ASU)**

(10) 電気 かみ <連用節>そり のコンセントは <主題ハ> どこですか。 <文末> **(BTEC)**

(11) 実戦力を地元経済の活性化に結び付けようと <条件節ト> 「インキュベーター」構想を描く。<文末> (**Nikkei**)

(9) の「水準で」の「で」は実際には助動詞であるが，格助詞と誤解析されたために，本来は挿入されるべき <並列節デ> が挿入されていない．(10) の「かみそり」は実際には 1 語の名詞であるが，単語境界が誤って解析され，本来は存在しない「かみ」という動詞が発生したために，本来は存在しないはずの <連用節> という節境界が誤って検出されている．(11) の「結び付けようと」の「と」は実際には引用を表す格助詞であるが，接続助詞と誤解析されたために，<引用節> となるべきラベルが <条件節ト> になっている．

### 5.2 節境界検出ルール自体の問題

節境界検出ルール自体の問題によって誤った節境界が検出されていた場合が，57 例あった．このうち 11 例は，節境界として検出すべき表現が節境界検出ルールに登録されていなかったために検出できなかった場合であった．例えば，以下のようなものである．

(12) 社会的な影響が極めて大きい上 噴火しなかった場合 <条件節バアイ> 情報を出した <連体節> 責任は <主題ハ> どうなるのかという <連体節トイウ> 指摘も... (**NHK**)

(13) よくそんなに切り刻むわねと <引用節> 言いたくなるくらい 土地が細分化されている。<文末> (**Nikkei**)

(12) の「大きい上」，(13) の「言いたくなるくらい」の直後は，本来は <連用節その他> として検出されるべき節境界である．しかし，これらの表現を検出するためのパタンが節境界検出ルールに登録されていなかったため，検出することができなかった．

その他の場合は，節境界検出ルールの記述に曖昧性が含まれていたことに起因するものが大半であった．これらはいずれも，検出ルールの追加・修正によって正しく検出されるようになる．

### 5.3 複合動詞に関する検出誤り

複合動詞が節境界の誤検出に関与する場合が 46 例あった．例えば，以下のようなものである．

(14) 新たにできた <連体節> 火口から噴煙が吹き <連用節> 上がった 際に <時間節その他> 起きた <連体節-形式名詞> ものと考えられると <引用節> 話しています。<文末> (**NHK**)

(15) 情報をどう使いこなすか <間接疑問節> こうした問題がますます大きな課題として私たちに突きつけられて <テ節> 来るでしょう。<文末> (**ASU**)

(14) の「吹き」と「上がる」，(15) の「突きつけられて」と「来る」がそれぞれ自立的な動詞句として形態素解析されたため，途中に <連用節> <テ節> という節境界ラベルが挿入されてしまっている．これらはそれぞれ，「吹き上がる」という 1 語の複合動詞，「〜てくる」というアスペクト表現を伴った 1 つの述語句と見なすべきであり，<連用節> <テ節> という節境界ラベルは正しくない．しかしながら，文脈によっては，まったく同じ表現に対して<連用節> <テ節> が挿入されるべき場合もあり得るため，単純に形態素解析の誤りと考えることは妥当ではない．このような複合動詞の認定は，CBAP による節境界検出だけでなく，構文解析においても問題となるところである．

### 5.4 そもそも検出することが困難な節境界

局所的な形態素の連接パタンのみを手がかりとして節境界を検出するというCBAPの手法では，そもそも検出することが困難な節境界が77例あった．このような節境界を検出するためには，文全体を構文解析して構文要素間の依存関係を明らかにする必要があり，CBAPによる検出能力を超える問題となる．77例の内訳を，表8に示す．以下では，それぞれの場合について詳しく見ていく．

表 8: 検出が困難な節境界の内訳

| 頻度 | 検出誤りの内容 |
|---|---|
| 28 (36.4%) | 述語格名詞句 |
| 21 (27.3%) | 形容詞連用節 |
| 21 (27.3%) | 形容詞連体節 |
| 4 (5.2%) | 形容動詞連体節 |
| 3 (3.9%) | 部分並列 |
| 77 (100%) | 合計 |

#### (5.4.1) 述語格名詞句・部分並列

名詞句が述語になる場合，基本的には助動詞「だ」や補助動詞「する」を後接させる必要がある．しかし，場合によって助動詞・補助動詞を伴わない名詞句が単独で述語として機能することがある．このような名詞句を，ここでは「述語格名詞句」と呼んでおく．述語格名詞句の境界を || で示すと，以下のようになる．

(16) 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)**

(17) 点訳結果は <主題ハ> ＷＳでデータベース化 || 視覚障害者は <主題ハ> 主に点字図書館でプリントしてもらい <連用節> 利用する。 <文末> **(Nikkei)**

上記の例で，名詞句「死亡」「データベース化」は単独で述語として機能しており，その直後は節境界として検出されることが望ましい．しかし，局所的な形態素の連接パタンのみを手がかりに節境界を検出するCBAPの手法では，このような境界を検出することは非常に困難である．また，いわゆる「部分並列」構造も，CBAPの手法で検出することは非常に困難である[6]．部分並列の境界を || で示す．

(18) 壮瞥町では <主題ハ> 五か所の避難所に百十九人が || 伊達市でも八か所の避難所に千二百七十人が避難しています。<文末> **(NHK)**

(19) 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)**

[6] それゆえ，CBAPが検出する節境界の定義に「述語格名詞句」「部分並列」は含まれていない．

### (5.4.2) 形容詞連用節・形容詞連体節・形容動詞連体節

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た連用節を形成する述語として機能する場合とがある．(20) は前者，(21) は後者の例である．

(20) 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)**

(21) 銀行などの金融業が百三人で最も多く次いで製造業の七十四人保険業の四十一人などとなっていて <テ節> ... **(NHK)**

(20) の「多く」は単独で述語句「残されています」を修飾する連用修飾要素であるが，(21) の「多く」は「金融業が」などを補足語として従えた連用節を構成する述語として機能しており，その直後は節境界 <形容詞連用節>として検出されるべき箇所である．

また，形容詞の基本形は，単独で連体修飾要素になる場合と，１つ以上の補足語を従えた連体節を構成する述語として機能する場合とがある．(22) は前者，(23) は後者の例である．

(22) 中国で作った <連体節> 安い衣料品が大変な売れ行きです。 <文末> **(ASU)**

(23) 地価が安い山梨県の郡部に工場があるから <理由節カラ> まったく問題ない。 <文末> **(Nikkei)**

(22) の「安い」は単独で名詞句「衣料品」を修飾する連体修飾要素であるが，(23) の「安い」は補足語「地価が」を従えた連体節を構成する述語として機能しており，その直後は節境界 <形容詞連体節> として検出されるべき箇所である．

同様に，「形容動詞語幹＋な」という表現は，単独で連体修飾要素になる場合と，１つ以上の補足語を従えた連体節を構成する述語として機能する場合とがある．

(24) ＢＳデジタル放送は <主題ハ> ハイビジョン方式の美しい <形容詞連体節> 映像が楽しめる <連体節> 「デジタルハイビジョン放送」と天気予報など自分の住んでいる <連体節> 地域の情報を必要なときにいつでも引き出せる <連体節> 「データ放送」が二つの柱となっています。 <文末> **(NHK)**

(25) 特に <談話標識> 被災地では <主題ハ> 水没した <連体節> 地域に取り残されている <連体節> 人たちを救助したり <並列節タリ> 食糧を運ぶために <タメニ節> 最も必要なヘリコプターが足りないのが <補足節> 現状です。 <文末> **(NHK)**

(24) の「必要な」は単独で名詞句「とき」を修飾する連体修飾要素であるが，(25) の「必要な」は「水没した」から「最も」までの部分を従えて連体節を構成する述語として機能しており，その直後は節境界 <形容動詞連体節>として検出されるべき箇所である．

ある形容詞・形容動詞が述語として機能しているか否かは，局所的なパタンマッチを用いたCBAPの手法では判断することができない．このような節境界を検出するためには，より広い範囲を構文解析し，大局的に見てそこが節境界であるか否かを判定する必要がある．

以上，節境界が誤って検出された場合について分析した結果を示した．そもそも検出することが困難な節境界や，大局的な構造を見なければ決定できない節境界など，CBAP による検出手法では本質的に対処できない問題があることを論じた．

## 6 節境界を利用した文分割

本節では，CBAP によって検出された節境界ラベルを文分割に利用する事例について述べる．まず，節境界の位置で文を分割することによって，異なるコーパスに含まれる文長のばらつきが均質化されることを示す．また，2.3 節で示した「従属節の従属度」の概念に基づいて節境界を選択的に利用することにより，分割長を柔軟に操作できることを示す．さらに，CBAP による節境界の検出手法を，句点の含まれない音声コーパスの発話分割処理に適用した事例を示し，その有用性について述べる．

### 6.1 文長のばらつきの均質化

前節で示した5種類のコーパスを対象に，CBAP が検出した節境界の位置で文を分割することにより，異種コーパス間に見られた文長のばらつきが均質化されることを示す．まず，各コーパスに含まれる形態素数，文節数，文数，1文あたりの形態素数と文節数を表9 に示す．

表 9: 各コーパスの規模

| | 形態素数 | 文節数 | 文数 | 形態素 / 文 | 文節 / 文 |
|---|---|---|---|---|---|
| **ASU** | 577,862 | 222,930 | 19,828 | 29.14 | 11.24 |
| **NHK** | 75,624,654 | 27,196,464 | 1,593,442 | 47.46 | 17.07 |
| **Nikkei** | 499,482,879 | 166,082,917 | 18,619,709 | 26.83 | 8.92 |
| **SLDB** | 255,163 | 84,811 | 21,769 | 11.72 | 3.90 |
| **BTEC** | 1,371,643 | 471,563 | 174,242 | 7.87 | 2.71 |

1文あたりに含まれる形態素数および文節数を比較すると，従来からの指摘通り，ニュース文（**NHK**）の文長が特に長いことが分かる [1]．これに対して，情報交換を主な目的とする旅行会話（**SLDB**，**BTEC**）では，文長は比較的短くなっている．**ASU** と **Nikkei** はその中間に位置している．各コーパスにおける文長の分布を図 3 に示す．

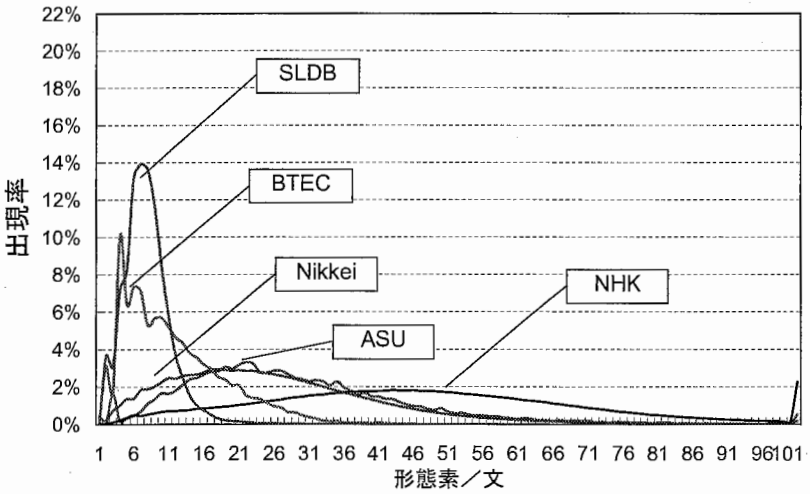


図 3: 文長の分布

次に，各コーパスに対して CBAP を適用し，節境界の検出を行なった．検出された節境界の位置で文を分割し，分割された各単位を「節単位」とした．検出された節境界の数，1文あたりに含まれる節境界の平均数，1節単位に含

まれる形態素数と文節数の平均を，表 10 に示す．

表 10: 節境界の検出結果

| | 節境界数 | 節境界 / 文 | 形態素 / 節単位 | 文節 / 節単位 |
|---|---|---|---|---|
| **ASU** | 92,539 | 4.67 | 6.24 | 2.40 |
| **NHK** | 9,569,742 | 6.01 | 7.90 | 2.84 |
| **Nikkei** | 60,606,934 | 3.25 | 8.24 | 2.74 |
| **SLDB** | 44,558 | 2.05 | 5.73 | 1.90 |
| **BTEC** | 261,926 | 1.50 | 5.24 | 1.80 |

1 文に含まれる節境界の数は文長に比例して **NHK** が最も多く，**ASU** と **Nikkei** がそれに続いている．一方 **SLDB** と **BTEC** では 1 文に含まれる節境界の数が 1.5 から 2 であることから，節境界が <文末> のみである文，すなわち単文が多く占めていることが分かる．

ここで注意すべきは，1 つの節単位に含まれる形態素数および文節数に，コーパス間でほとんど差が見られないという点である．文長の分布には各コーパス間でかなりのばらつきが見られたが，節単位の平均長にはコーパス間で顕著な差は見られない．各コーパスにおける節単位の長さの分布を，図 4 に示す．文を節単位に分割することによって，図 3 で見られた文長のばらつきがほぼ均等になり，全体が均質化されていることが分かる．

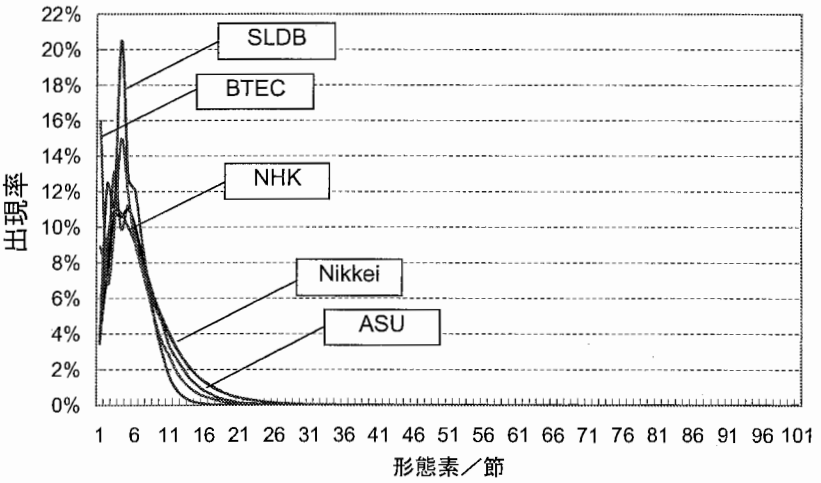


図 4: 節単位長の分布

節境界の位置で分割された節単位を，構文解析，機械翻訳，重要文抽出などの処理単位として利用することにより，文長のばらつきを含む複数のコーパスを統一的な視点から扱うことが可能になる．また，処理の対象を局所的な範囲に限定することにより，各処理の局所化・効率化に寄与することができる．例えば，柏岡 (2003) は，**ASU**，**NHK**，**Nikkei** の 3 種のコーパスを対象に，節単位の内部で係り受け構造が完結するかどうかを調査している [4]．その結果，**ASU** の 91.2%，**NHK** の 87.7%，**Nikkei** の 89.0% の節単位は，その内部で係り受け構造が完結していることが確認されている．この結果は，節単位が漸進的な係り受け解析にとって有効な単位であることを示している．

また，検出された節境界ラベルの分布を調べることにより，各コーパスに含まれる節境界の種類と出現率を比較し，そのコーパスに含まれる文の性質を大まかに推察することができる．5 つのコーパスに現れた従属節の種類（大分類）および<文末> の出現率を図 5 に示す．

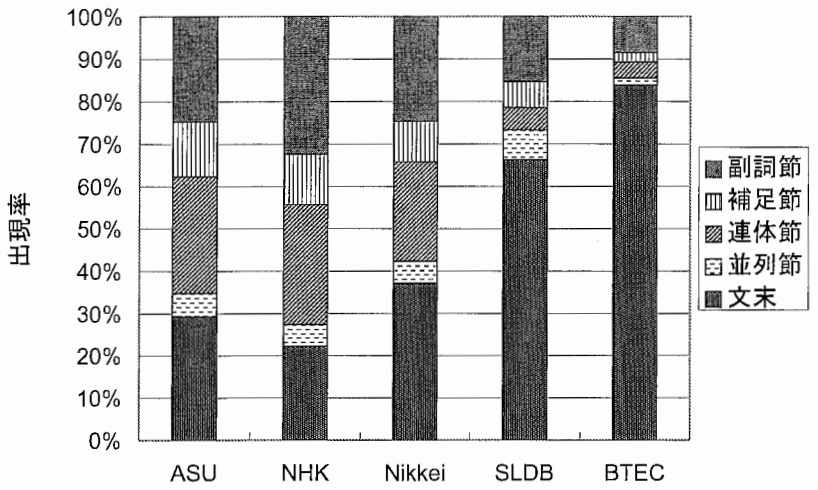


図 5: 各コーパスにおける節境界の分布

**ASU**, **NHK**, **Nikkei** では <文末> がそれぞれ 20–40% の割合でしか出現していないのに対して, **SLDB**, **BTEC** での <文末> の割合は 60–80% と非常に高く, 逆に従属節は少数しか現れていない. 特に **BTEC** では 85% 近くを <文末> が占めており, 従属節を含まない単文が多く現れていることを示している.

### 6.2 節境界を選択的に利用した文分割

以上では, 節境界の位置で文を分割することにより, 異種コーパス間の文長のばらつきが均質化されることを見た. しかしながら, 節境界で分割された節単位の長さは平均 1.8 ～ 2.8 文節という長さでしかなく, 処理単位としてはかえって問題になることがある. 例えば, 節境界で文が分割されることによって係り受け関係が付与されるはずの 2 要素が分断されてしまったり, 分割された単位が短すぎるために翻訳結果の品質が低下したりするという問題である. そこで以下では, 「従属節の従属度」の概念を利用し, 検出された節境界ラベルを選択的に利用することにより, 統語的・意味的な完結性を備えた分割結果を得る手法について述べる.

2.3 節で述べたように, 南 (1974,1993) は, 従属節の形式に応じて 3 段階の従属度を割り当てた. この記述に基づいて, 検出された節境界のうち従属度の低いもの (独立度の高いもの) から順に文の分割位置として採用すると, 分割されたそれぞれの単位は, 統語的に完結し, 意味的にも充足している可能性が高い. ここでは, 検出された節境界のうち, 南の言う C 類の従属節, すなわち, <並列節ガ> <並列節ケレドモ> <並列節シ> <理由節カラ> <連用節> <並列節デ> <テ節>[7] という節境界でのみ文を分割する処理を考える. すると, (26) ( = (1),(3) ) の文は // の位置で分割されることになる. | は分割点として採用されなかった節境界の位置を表す.

(26) 身長の二乗掛ける | 二十二が標準体重ということになっとりまして //
私の標準体重は | 六十四キロなんですが //
それから見ると | 約七キロぐらいは | 減量が必要ということで //
運動をする | 方がいいことになってまして //
本当は食事量を減らすという | ことなんでしょうけど //
なかなかそれは | 難しいので | 私は | 専ら運動の方で健康を維持しようという | ことに努めとります。 //

従属度の低い節境界のみを分割点とした場合の各単位は, すべての節境界を分割点とした場合に比べて, 統語的・

[7] ただし, テ節に関しては「です」「ます」が「て」に前節する場合のみとする.

意味的な完結性と充足性をより適切に備えていると思われる．ここではC類の節境界のみを分割点としたが，すべての節境界ラベルについて分割点となる可能性の大小を考慮し，必要に応じて従属度の低いものから順に分割点とすることにより，従来の手法よりも柔軟に，かつ言語学的にも意味のある基準によって，文分割を実現することができる．

ただし，このような手法で分割された単位が，全ての場合において，統語的・意味的に完結しているとは限らない．比較的少数ではあるものの，上記の節境界で文を分割することによって本来係り受け関係を持つはずの2要素が分断され，その単位内では係り受け関係が付与できなくなる場合がある．

(27) これに対して | 小渕総理大臣は | 「政権合意の実現に取り組む | 意欲は | 十分あるが //
自由党が度々離脱問題を起こして | 連立政権の中の信頼関係が崩れている」と | 述べました。 **(NHK)**

(28) これは | 職場に公正で //
透明なルールを確保するという | ことであります。 **(ASU)**

(27)の「小渕総理大臣は」は文末の「述べました」に係る要素であるが，引用節内にある節境界 <並列節ガ> で文が分割されてしまったため，分割された単位内での係り先が消失してしまっている．また，(28)は連体節内の節境界 <並列節デ> で文が分割されてしまったため，文頭の「これは」の係り先が消失してしまっている．CBAPが局所的な範囲から節境界を検出するものであり，文の大局的な構造を解析するものではない以上，上記のような問題が生じることは不可避的である．ただし，表層的な情報（括弧記号など）を用いることによって，このような過分割の問題を部分的に回避できる可能性はある[28]．

### 6.3 発話分割処理への応用

従来，句点の含まれない音声コーパスを対象に何らかの処理を行なう場合，一定以上のポーズで区切られた区間を発話の分割点として利用することが多かった．しかし，ポーズで区切られたそれぞれの単位が，構文解析や翻訳処理，重要文抽出などにとって意味のある単位であるとは限らない．むしろ，発話中から節境界を検出して分割点として利用する方が，ポーズで区切られた文の断片に比べて，統語的・意味的な妥当性を備えた処理単位を抽出できる可能性が高い．以下では，CBAPを利用して『日本語話し言葉コーパス』の発話分割処理を行なった事例を示し，その有用性を示す．

『日本語話し言葉コーパスCSJ（Corpus of Spontaneous Japanese; 以下CSJと記す）』は，文科省科学技術振興調整費開放的融合研究制度研究課題「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」プロジェクト（1999–2003）によって構築が進められている，大規模な自発音声コーパスである[6]．学会発表や模擬講演など，独話を中心とした日本語の自発的な発話（650時間，700万語）が書き起こされ，形態素情報，イントネーションラベル，係り受け構造，談話構造など，さまざまな研究用情報が付与されている．発話の書き起こしにはポーズ情報が含まれているものの，文の境界を示す句点は含まれていない．このような句点の含まれないコーパスに対して係り受け構造の付与[29]，談話構造分析[30]，重要文抽出[31]などを行なおうとする場合，まず必要になるのは，発話を分割して処理単位を特定することである．

我々は，CSJを対象として，発話中に含まれる節境界をCBAPによって検出し，各処理で用いる基本単位を確定することにした[32, 33]．ただし，CSJに付与されている形態素の仕様は，「茶筌」のものとは異なっている．そこで，「茶筌」仕様で開発されたCBAPをCSJの形態素仕様に改編し，CSJ仕様の節境界検出プログラム “CBAP-csj” を新たに開発した．

開発したCBAP-csjは，292個の節境界検出ルールによって49種類の節境界を検出する．茶筌版のCBAPに比べて検出される節境界の種類が少なくなっているのは，発話の分割処理に特化するために，検出する節境界を限定した

ことによる．CBAP-csj では，従属度が低いために発話の分割位置とは認められない節境界（連体節，ナガラ節，ツツ節，など）はそもそも検出されないようになっている．

さらに，南 (1974,1993) による従属節の形態と従属度の対応付けを参考にして，節境界直後の切れ目の大きさという観点から，節境界に「絶対境界」「強境界」「弱境界」という 3 つのレベルを設けた．絶対境界とは，通常の意味での文末表現に相当する境界である．強境界とは，いわゆる文末表現ではないが，発話の大きな切れ目と考えられる従属節の境界である．「並列節ガ」や「並列節ケレドモ」などがこれに相当する．弱境界とは，通常は発話の切れ目になることはないが，稀に発話の大きな切れ目になり得る従属節の境界である．「理由節ノデ」「条件節」などがこれに相当する．3 つのレベルは，すべての節境界ラベルに対して個別に割り当ててある[8]．CBAP-csj で検出される節境界ラベルの一覧を，図 6 に示す．[ ] は絶対境界，/ /は強境界，< >は弱境界を表す．

[絶対境界] （いわゆる文末表現）

文末，文末候補，と文末

/強境界/ （発話の大きな切れ目になる節境界）

並列節ガ，並列節ケド，並列節ケドモ，並列節ケレド，並列節ケレドモ，並列節シ

<弱境界> （稀に発話の大きな切れ目になる節境界）

タリ節，タリ節 - 助詞，テカラ節，テカラ節 - 助詞，テハ節，テモ節，テ節，テ節 - 助詞，トイウ節，トカ節，トカ節 - 助詞，ノニ節，フィラー文，ヨウニ節，引用節，引用節 - 助詞，引用節トノ，感動詞，間接疑問節，間接疑問節 - 助詞，条件節タラ，条件節タラバ，条件節ト，条件節ナラ，条件節ナラバ，条件節レバ，接続詞，接続詞 C，接続詞 CL，接続詞 L，接続詞 M，並列節ダノ，並列節デ，並列節ナリ，理由節カラ，理由節カラ - 助詞，理由節カラニハ，理由節ノデ，連体節テノ，連用節

図 6: 節境界ラベルの一覧 (CBAP-csj)

CBAP-csj を 396 講演分のデータ（1,026,969 形態素）に適用し，節境界の検出を行なった．検出された節境界の数を表 11 に示す．検出された節境界のうち上位 20 位までを，表 12 に示す．さらに，CBAP-csj によって節境界ラベルが挿入された結果の例を (29) に示す．(F　) で囲まれた部分は，フィラーを現す．

表 11: CSJ （396 講演）に出現した節境界

| | 頻度 | |
|---|---|---|
| 絶対境界 | 25,664 | (24.3%) |
| 強境界 | 16,948 | (16.0%) |
| 弱境界 | 63,159 | (59.7%) |
| 合計 | 105,771 | (100%) |

(29) 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/

[8] 「テ節」「ヨウニ節」などの弱境界に「です」「ます」が前接した場合，または弱境界に接続詞が後接した場合は，レベルを弱境界から強境界に引き上げるという規則が設けてある．

表 12: CSJ（396 講演）に出現した節境界の内訳（上位 20 位）

| 出現数 | | 節境界名 | 出現数 | | 節境界名 |
|---|---|---|---|---|---|
| 21,656 | (20.5%) | [文末] | 2,941 | ( 2.8%) | /並列節ケレドモ/ |
| 13,706 | (13.0%) | <テ節> | 2,528 | ( 2.4%) | <理由節ノデ> |
| 11,646 | (11.0%) | <接続詞> | 2,228 | ( 2.1%) | <条件節ト> |
| 6,838 | ( 6.5%) | <引用節> | 1,983 | ( 1.9%) | <連用節> |
| 6,725 | ( 6.4%) | <トイウ節> | 1,879 | ( 1.8%) | /並列節ケド/ |
| 4,304 | ( 4.1%) | <接続詞 L> | 1,351 | ( 1.3%) | <トカ節> |
| 4,145 | ( 3.9%) | /並列節ガ/ | 1,240 | ( 1.2%) | /条件節ト/ |
| 3,615 | ( 3.4%) | /テ節/ | 1,206 | ( 1.1%) | /並列節ケドモ/ |
| 3,456 | ( 3.3%) | [文末候補] | 1,065 | ( 1.0%) | <条件節レバ> |
| 3,090 | ( 2.9%) | <並列節デ> | 997 | ( 0.9%) | <理由節カラ> |

(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/ ...

我々は，検出された節境界のうち，従属節の従属度の概念および経験的知見に基づいて，絶対境界と強境界を発話の分割位置として採用することにした．これにより，一定以上のポーズで区切られた範囲よりも，統語的・意味的により有用な単位を抽出することができ，係り受け付与，談話構造分析，重要文抽出などの処理単位の確定に大きく寄与することができる．

ただし，自発的な発話に特有な問題によって，CBAP-csj による発話分割の結果が適切でなくなる場合がある．例えば，発話しながら逡巡しているような場合，明示的な文末表現や強境界/並列節ケド/が，発話の途中に現れることがある．

(30) 私が大学の二年の時に (F えーっと) 千九百九十六年の七月です [文末] (F えー) 学科の仲間と一緒にキャンプに行ったことについて話します[文末]

(31) 我々が開発してある (F あ) (F ま) フリーソフトで出てるんですけど /並列節ケド/ (F ま) 形態素解析エンジンがありまして (F ま)...

下線を引いた節境界は分割点として採用されるため，デフォルトの状態では，発話が統語的・意味的に充足しないうちに分割されてしまうことになる．さらに，「体言止」や「一語文」などの特殊な節境界や，言い差し・言い誤り・言いやめ・倒置・文のねじれなど，自発的な発話に特有な諸現象については，パタンマッチを用いた CBAP-csj の手法によって適切に発話分割を行なうことは非常に困難である．そこで，CBAP-csj の出力を人手でチェックし，一定の基準に沿って修正する作業を行なった [32, 33]．この人手修正の結果を発話分割の最終的な結果として，係り受け構造の付与，談話構造分析，重要文抽出などの基本処理単位として採用した．

以上，句点の含まれない音声コーパスを対象とした発話分割処理に CBAP を適用した事例について示し，その有用性を示した．

## 7 おわりに

日本語の記述的文法研究の知見に基づいて，日本語の節境界の位置を網羅的に検出し，その種類を特定するプログラム CBAP を開発した．日本語の節境界は，述語句の活用形や接続助詞などの形態的特徴により，詳細に分類することができる．節境界の位置と種類を特定するための節境界検出ルールを人手で記述した結果，CBAP は 97% 以上という高い精度で 147 種類の節境界を検出できることが確認できた．また，局所的な形態素列のみを手がかりにしているため，対象となるデータのドメインの違いや文体・文長の違いなどにはほとんど影響されることがない．

CBAP で検出された節境界の位置で文を分割することにより，異なるコーパスの間で見られる文長のばらつきが均質化されることを示した．また，「従属節の従属度」の概念に基づいて，分割点とする節境界を選択的に採用することにより，統語的に完結し，意味的にも充足した分割単位が得られることを述べた．さらに，句点の含まれない音声コーパス CSJ を対象とした発話分割処理に応用した事例について示し，その有用性を示した．

本稿で示した節境界の検出手法は，さらに次のような処理に利用することができると考えられる．

- 構文解析を節境界で区切られた範囲（節単位）に限定して行なうことにより，解析処理の局所化・高速化に寄与することができる．
- 多言語パラレルコーパスを構築する際，節単位ごとに対訳を付与することにより，文単位の対訳付与よりも細かいアライメントを行なうことができる．[5]
- 機械翻訳において，節単位ごとにアライメントされた対訳ペアを用例として採用することにより，部分的・同時的な訳出を行なう単位として利用できる．
- 重要文抽出において，節単位を抽出単位とすることにより，過剰な部分の抽出や情報の欠落を回避し，より効果的な抽出を行なうことができる．
- 節境界を局所的・決定的に検出できる特徴を利用して，漸進的な言語処理技術（漸進的構文解析，同時通訳システムなど）における処理単位の抽出に用いる．
- コーパス中に現れる節境界の種類と頻度の比較，節境界の連接関係の統計調査など，コーパスの特徴分析に用いる．

# Appendix.

CBAP が検出する 139 種の節境界のリストを，以下に示す．なお，冒頭に *が付されている節境界は，「-助詞」が後続することがあるものである．例えば，「*引用節」は，「～すると<引用節>」「～するとは<引用節-助詞>」の両方に対応する．全てを合わせると 139 種類となる．

1. 副詞節 (102)

   (a) 条件・譲歩節: (23) <*条件節カギリ> <条件節ケッカ> <条件節タラ> <条件節タラ-引用> <条件節タラバ> <条件節ト> <条件節ト-引用> <条件節トコロ> <*条件節トコロデ> <条件節ナラ> <条件節ナラ-引用> <条件節ナラバ> <条件節ナラバ-引用> <条件節バ> <条件節バ-引用> <*条件節バアイ> <条件節モノノ> <譲歩節ノニ> <譲歩節命令形> <テモ節>

   (b) 原因・理由節: (8) <理由節カラ> <理由節カラ引用> <理由節ノデ> <*タメ節> <*タメニ節> <タメニハ節>

   (c) 時間節: (21) <*時間節その他> <*時間節アト> <*時間節アトデ> <*時間節アトニ> <時間節アトノ> <*時間節イマ> <時間節イライ> <時間節トキ> <時間節トキニ> <時間節トキニハ> <時間節トキノ> <*時間節マエ> <*時間節マエニ> <時間節マエノ>

   (d) 様態節: (12) <*ママ節> <ママ節-引用> <ママ節-補足語> <*ママデ節> <*ナガラ節> <ナガラモ節> <ナガラ節-引用> <*ツツ節>

   (e) 副詞節その他: (38) <連用節> <連用節その他> <目的節> <*ダケ節> <ダケ節-引用> <ダケ節-補足語> <*ダケニ節> <テ節> <*テカラ節> <テハ節> <*ナド節> <ナド節-補足語> <*ホカ節> <*ホカニ節> <*ホド節> <*ホドニ節> <マデ節> <マデ節-引用> <マデニ節> <マデニハ節> <マデハ節> <マデモ節> <*ヨウ節> <*ヨウニ節> <ヨウニ節-引用> <*ヨリ節> <形容詞連用節>

2. 補足節 (10)

   (a) 補足節: (2) <補足節> <補足節-並立>

   (b) 引用節: (5) <*引用節> <引用節-補足語> <引用節トノ> <連続引用>

   (c) 間接疑問節: (3) <*間接疑問節> <間接疑問節-補足語>

3. 連体節 (15)

   (a) 連体節: (15) <連体節> <連体節-形式名詞> <連体節カギリノ> <連体節タメノ> <連体節テノ> <連体節トイウ> <連体節ナドノ> <連体節バアイノ> <連体節ホドノ> <連体節マデノ> <連体節ヨウナ> <連体補足節> <形容詞連体節-形式名詞> <形容詞連体節> <形容動詞連体節>

4. 並列節 (12)

   (a) 並列節: (12) <*並列節ガ> <*並列節ケレドモ> <*並列節シ> <*並列節タリ> <並列節デ> <並列節デハ> <*並列節トカ>

## 参考文献


[1] 福島, 江原, 白井：“短文分割の自動要約への効果”, 自然言語処理, **6**, 6, pp. 131–147 (1999).

[2] 丸山, 熊野, 柏岡：“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, pp. 429–432 (2001).

[3] 益岡, 田窪：“基礎日本語文法 — 改訂版 —”, くろしお出版 (1992).

[4] 柏岡, 丸山, 田中：“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, pp. 117–120 (2003).

[5] H. Kashioka, T. Maruyama and H. Tanaka: “Building a parallel corpus for monologue with clause alignment”, Proceedings of the Ninth Machine Translation Summit, pp. 216–223 (2003).

[6] K. Maekawa: “Corpus of spontaneous japanese: its design and evaluation”, Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003), pp. 7–12 (2003).

[7] 武石, 林：“接続構造解析に基づく日本語複文の分割”, 情報処理学会論文誌, **33**, 5, pp. 652–663 (1992).

[8] 金, 江原：“日英機械翻訳のための日本語長文自動短文分割と主語の補完”, 情報処理学会論文誌, **35**, 6, pp. 1018–1028 (1993).

[9] 木村, 野村, 平川：“日英機械翻訳前編集における日本語文分割処理について”, 自然言語処理研究会, 情報処理学会 (1989).

[10] E. I. Ejerhed: “Finding clauses in unrestricted text by finitary and stochastic methods”, Proceedings of the Second Conference on Applied Natural Language Processing, pp. 219–227 (1988).

[11] S. Abney: “Rapid incremental parsing with repair”, Proceedings of the 6th New OED Conference: Electronic Text Research, pp. 1–9 (1990).

[12] H. V. Papageorgiou: “Clause recognition in the framework of alignment”, Recent Advances in Natural Language Processing (Eds. by R. Mitkov and N. Nicolov), John Benjamins, Amsterdam/Philadelphia (1997).

[13] V. J. Leffa: “Clause processing in complex sentences”, Proceedings of the First International Conference on Language Resources and Evaluation, pp. 937–943 (1998).

[14] E. F. Tjong Kim Sang and H. Déjean: “Introduction to the conll-2001 shared task: Clause identification”, Proceedings of CoNLL-2001 (Eds. by W. Daelemans and R. Zajac), Toulouse, France, pp. 53–57 (2001).

[15] E. Brill: “Some advances in rule-based part of speech tagging”, Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), pp. 722–727 (1994).

[16] E. F. Tjong Kim Sang: “Text chunking by system combination”, Proceedings of CoNLL-2000 and LLL-2000, pp. 151–153 (2000).

[17] X. Carreras and L. Màrquez: “Boosting trees for clause splitting”, Proceedings of CoNLL-2001 (Eds. by W. Daelemans and R. Zajac), Toulouse, France, pp. 73–75 (2001).

[18] 三上：“現代語法序説 — シンタクスの試み —”, 刀江書店 (1953).

[19] 三上：“続・現代語法序説 — 主語廃止論 —”, 刀江書店 (1959).

[20] 三上：“日本語の構文”, くろしお出版 (1963).

[21] 南：“現代日本語の構造”, 大修館書店 (1974).

[22] 南：“現代日本語文法の輪郭”, 大修館書店 (1993).

[23] 白井, 池原, 横尾, 木村：“階層的認識構造に着目した日本語従属節間の係り受け解析の方法とその精度”, 情報処理学会論文誌, **36**, 10, pp. 2353–2361 (1995).

[24] 寺村：“日本語教育指導参考書 5, 日本語の文法（下）”, 国立国語研究所 (1981).

[25] 北條：“日本語教育指導参考書 15, 談話の研究と教育 II”, 国立国語研究所 (1989).

[26] T. Morimoto, N. Uratani, T. Takezawa, O. Furuse, Y.Sobashima, H. Iida, A. Nakamura, Y. Sagisaka, N. Higuchi and Y. Yamazaki: “A speech and language database for speech translation research”, Proceedings of International Conference on Spoken Language Processing, pp. 1791–1794 (1994).

[27] T. Takezawa, E. Sumita, F. Sugaya, H. Yamamoto and S. Yamamoto: “Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world”, Proceedings of LREC 2002 3rd International Conference on Language Resources and Evaluation, pp. 147–152 (2002).

[28] 張, 丸山, 柏岡, 浦谷, 江原：“ニュースの同時通訳ための短文分割手法について”, 情報処理学会 第 61 回 全国大会講演論文集, 第 2 巻, pp. 163–164 (2000).

[29] 内元, 野畑, 山田, 関根, 井佐原：“日本語話し言葉コーパスの形態素解析”, 言語処理学会 第 9 回年次大会 発表論文集, pp. 113–116 (2003).

[30] 森本, 高梨, 竹内, 小磯, 井佐原：“話し言葉コーパスへの談話構造タグ付与”, 言語処理学会 第 9 回年次大会 発表論文集, pp. 695–698 (2003).

[31] 野畑, 関根, 内元, 井佐原：“話し言葉コーパスにおける文の切り分けと重要文抽出”, 第 2 回 話し言葉の科学と工学ワークショップ, pp. 93–100 (2002).

[32] 高梨, 丸山, 内元, 井佐原：“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, pp. 521–524 (2003).

[33] K. Takanashi, T. Maruyama, K. Uhimoto and H. Isahara: “Identification of “sentence” in spontaneous japanese –detection and modification of clause boundaries –”, Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003), pp. 183–186 (2003).