『日本語話し言葉コーパス』における節単位認定

(Version 1.2)

高梨 克也　　内元 清貴　　(情報通信研究機構)
丸山 岳彦　(国立国語研究所 / ATR 音声言語コミュニケーション研究所*)

【目次】

* 2003 年度まで所属.

１．背景と目的

節単位境界認定作業では，形態素（短単位・長単位）（pos.pdf）と文節（bunsetsu.pdf）の情報を利用し，係り受け構造付与（dependency.pdf），重要文抽出（summarydata.pdf），談話境界付与（discourse.pdf）のための共通単位である「節単位 ClauseUnit」を認定する．対象講演はコア中の独話177 講演（対話・再朗読を除く）とコア以外のテストセット22 講演を含む199 講演である（talk_data.csv 参照）．

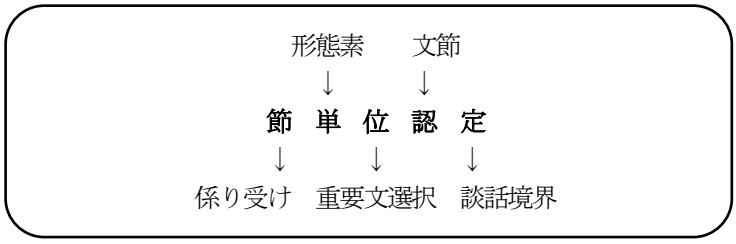

図１：節単位認定の位置づけ

従来，書き言葉に対する係り受け構造，談話境界などの情報付与や重要文選択などにおいては，その単位として「文」が用いられてきた．しかし，自発的な話し言葉を対象とする場合，文は必ずしも自明な単位ではない．書き言葉では書き手自身が句点によって区切りを確定するのに対して，話し言葉である CSJ には句点情報がなく，また文法的に明確な文末形式が頻繁に用いられるとは限らないため，いわゆる文末表現のみを境界認定の基準にしたのでは極端に長い単位が生じてしまう場合がある．他方，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（CSJ では転記基本単位（transcription.pdf）にほぼ相当する）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，自動要約などにとって意味のある単位であるとは限らない．

そこで我々は，発話を分割する位置として，「節」の境界に注目することにした．「節」とは，述語を中心としたまとまりであり，統語的にも意味的にもある程度完結した単位である．そのため，ポーズで区切られた文の断片に比べて，係り受け構造付与，重要文選択，談話境界付与のための共通単位の認定という目的により合致する．そこで，発話中からさまざまな種類の節境界を検出し，その統語的・意味的特性を考慮した上で，分割位置を特定することを考えた．

また，対象となるデータは独話であり，一人の発話者が一定時間継続して話し続けているため，対象データはいわば「一本の長い紐」のようなものであるといえる．加えて，一口に「節」といっても，形式上日本語にはさまざまな種類の節があるため，これらをひとつずつ独立の単位として切り離してしまうわけにはいかない．従って，本作業は，文の構成要素から内在的に単位の範囲を認定するのではなく，当該の節単位の自立性を連続する前後の部分との間の相対的関係から決定することを特徴としている．

さらに，自発的な話し言葉である CSJ では，言い直し，言い換え，言いやめなどの要因により文の範囲が確定しにくい場合や語や文の断片だけで発話が構成される場合がある．こうした箇所では自動検出された節境界以外の箇所で単位を切断することが必要になる場合もあるため，節単位の認定においては人手による確認と修正が必要となる．

以上のような状況を鑑みて，我々は，

(1) 形態素情報をもとにして，節の境界を自動的に検出する．
(2) ある種の節境界を発話分割位置として，デフォルトの発話分割結果（デフォルト単位）を自動で作成する．
(3) 発話の流れや前後の文脈を考慮し，さらに言いやめや体言止めなど節境界とは関わりのない発話の切れ目に対処するため，(2) の結果に対して人手修正を加える．

という手順によって発話分割を行なうことにした．こうして認定されたされたひとつひとつの単位をここでは「節単位」と呼ぶ．従って，認定された「節単位」は(2)で自動認定されたデフォルト単位のままのものか，あるいは(3)によって人手修正されたもののいずれかになる．節単位は「文」に代わる，話しことばにおける基本単位である．

本解説文書では，発話を分割して節単位を特定するための手続きについて述べる．以下の各節では，自動処理による「節境界ラベル」の挿入と「デフォルト単位」の認定（第２節），「デフォルト単位」に対する人手修正の必要性と作業基準（第３節），これらの二段階の作業によって認定された節単位に関する情報が XML ファイルにおいてどのように格納されているか（第４節）について，順に解説していく．

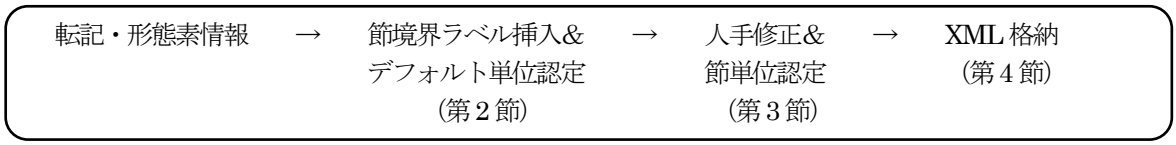

図2：節単位認定作業の流れ

2．節境界ラベルの挿入とデフォルト境界の認定

2.1 CBAP-csj による節境界ラベルの挿入

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，述語句の形態的情報（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，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．従来，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，節境界を機械的に検出するための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，節境界を自動的に検出するためのルールを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．

そこで我々は，日本語節境界検出プログラム CBAP（丸山他,2004）を利用して CSJ に含まれる節境界を検出し，発話分割処理の手がかりを得ることにした．CBAP は，日本語形態素解析器「茶筌[1]」によって形態素解析された日本語テキストを入力とし，ある形態素の前後 1～3 語を読み込んで，そこが節境界であるか否かを判定するプログラムである．節境界と判定された位置には，節境界の種類に応じた「節境界ラベル」が挿入される．我々は，CBAP が CSJ 仕様の形態素列を読み込んで動作するように改編し，新たに CSJ 仕様の節境界検出プログラム“CBAP-csj”を開発した[2]．

CBAP-csj は，CSJ に含まれる短単位および長単位の情報（pos.pdf ）を入力として，発話分割位置の候補になり得る 49 種類の節境界を検出することができる．発話分割位置の候補にはなり得ない節境界，例えば連体節，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．

検出されるすべての節境界ラベルは，節境界直後の切れ目の大きさという観点から「絶対境界」「強境界」「弱境界」という 3 つのレベルに区分されている．絶対境界は，形式上明示的な文末表現で，「思います」「できました」などの表現が相当する．強境界は，発話の大きな切れ目として考えられる従属節で，「けれども」「が」などの表現が相当する．弱境界は，通常は発話の切れ目になることはないと考えられる従属節で，「なら」「て」などの表現が相当する．さらに，厳密な意味では節境界ではないが，「そして」「ですけれども」のような接続詞についても，検出されるようになっている．3 つのレベルの節境界ラベルは，節境界名を囲む記号の種類によって表し分けられる．絶対境界は[文末]，強境界は/並列節ガ/，弱境界は<テ節>，接続詞は<接続詞>のようにそれぞれ表示される．

CBAP-csj が自動的に検出・挿入する「節境界ラベル」の一覧を表 1 に示す．また，CBAP-csj によって節境界ラベルが挿入された結果の例を次に示す．

ただし<接続詞> 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/ 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] そういうことで<並列節デ> 結構年のわりに歩いてるんですが/並列節ガ/ . . .

---

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

[2] 日本語節境界検出プログラム CBAP-csj は，研究目的に限り，無償で使用することができます．なお，使用者は，株式会社国際電気通信基礎技術研究所との間で「ソフトウェア使用許諾に関する覚書」を締結していただきます．ご希望の方は，ご所属・お名前・ご連絡先と「CBAP-csj 配布希望」との旨を明記した上で，次の連絡先までお申し込みください．折り返し，必要書類をお送りいたします．619-0288 京都府「けいはんな学研都市」光台 2-2-2 ATR 音声言語コミュニケーション研究所 企画担当 袋谷丈夫（takeo.fukuroya@atr.jp）

[絶対境界]　※デフォルト境界

文末，文末候補，と文末

/強境界/　※デフォルト境界

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>

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，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

<接続詞>

接続詞，接続詞L，接続詞C，接続詞CL

メタルール

弱境界に接続詞が後接した場合や，「テ節」「ヨウニ節」などの弱境界に「です」「ます」が前接した場合には，レベルを弱境界から強境界に変更する．

（従って，同じ節末表現が前後の形態素との関係で異なるレベルのラベルとして挿入される．）

表 1：CBAP-csj が検出・挿入する「節境界ラベル」の一覧 [3]

2.2 発話分割処理と「デフォルト単位」の確定

我々は，CBAP-csj によって検出された節境界のうち，絶対境界および強境界としてマークされた節境界を発話の分割位置として採用することにした．絶対境界は明示的な文末表現であるため，発話の切れ目となる可能性が極めて高い．また，「けれども」「が」などの強境界は，統語的にも意味的にも大きな切れ目となることが南不二男(1974,1993)などの研究で明らかにされている．このような文法論的な知見に基づいて抽出された単位は，一定以上のポーズで区切られた単位よりも統語的・意味的に有用な単位として，係り受け付与，談話構造分析，重要文抽出などの処理に利用することができる．そこで，絶対境界・強境界を「デフォルト境界」とし，これらのデフォルト境界で発話を分割した単位を「デフォルト単位」として確定した．先に示した例は，以下の 5 つのデフォルト単位に分割されることになる．

---

1. ただし<接続詞> こういう生活をしてますと/条件節ト/
2. 摂取カロリーが大変多くなりまして/テ節/
3. 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/
4. 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]
5. そういうことで<並列節デ> 結構年のわりに歩いてるんですが/並列節ガ/ . . .

以上のように CBAP-csj によって検出・確定されたデフォルト単位を，今回の節単位認定の作業における一次データとした．ただし，絶対境界・強境界で分割されたデフォルト単位が発話の切れ目として常に最良であるとは限らない．例えば次の例は，「タイトル」「夢の国ディズニーワールド」「私は旅行が大好きで...」という 3 つの単位に分割されるべきであるが，前者 2 つは名詞だけで発話が終わっているため，CBAP-csj では分割位置として検出することができない．

---

[3] 接続詞は厳密には節境界を表すものではないが，メタルールの適用に関わるものであるため，他の節境界ラベルと同様に挿入される．<接続詞>は「で」のように短単位のみから接続詞になるもの，<接続詞 L> は「それ+で」のように長単位のみから接続詞になるもの，<接続詞 C> は「けれど_接続詞」＋「も_助詞-副助詞」のように短単位の連接から接続詞になるもの，<接続詞 CL> は「ですけれど_接続詞 L」＋「も_助詞-副助詞」のように長単位と短単位の連接から接続詞になるものである．

タイトル夢の国ディズニーワールド私は旅行が大好きで<並列節デ>今までもあちこち行きましたけれども/並列節ケレドモ/

次の例は，「はいそうです 月曜日です」という部分が引用節の内部に収まる構造を取るが，局所的な形態素情報のみから節境界を検出する CBAP-csj では，そのような大局的な構造を解析できない．それゆえ，「そうです」までの部分がひとつのデフォルト単位となるが，これは適切ではない．

それから<接続詞>このはい<感動詞>そうです[文末]
月曜日ですって<引用節>相手の質問を繰り返したりすることが多々ありますが/並列節ガ/

次の例は，「千九百九十六年の七月です」「フリーソフトで出てるんですけど」が発話の途中に挿入された構造を取るが，この場合もまた，局所的な形態素情報のみから節境界を検出する CBAP-csj では解析できず，不適切なデフォルト単位が生成されている．

私が大学の二年の時に千九百九十六年の七月です[文末]
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]

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

以上のような問題から，我々は，CBAP-csj によって自動的に生成されたデフォルト単位を人手でチェックし，必要に応じて修正を施すことにした．次節では，我々が行なった人手修正操作について，現象ごとに作業基準を示す．なお，上記 199 講演以外のデータにも CBAP-csj で自動認定された節末ラベルの情報が同様に付与されるが，デフォルト単位の認定と人手による節単位認定は行っておらず，各種節末にラベルが挿入されるだけとなる．

3．人手修正作業基準

3.1 人手修正操作

前節で述べたCBAP-csjによる「デフォルト単位」の認定は局所的な形態素列のみを参照して境界を判定するものであるため，「体言止」などの特殊な節境界は発見できず，また，言い誤り・言いさしなどの自発的な話し言葉に特有の現象や談話構造との関係に不都合が生じる箇所には対処できないため，音声情報を参照しつつ，下記の基準に従ってデフォルト単位を人手修正した．人手修正の操作は次の三種類である．以下に例を示す[4]．

- 二つ以上のデフォルト単位の末尾を + でつなぐ．
- デフォルト単位の途中を – で切る．
- デフォルト単位内の要素を ( )，{ }，<< >> で囲む．

表２：人手修正操作記号

[4] 以下，修正後の例では，1 行がひとつの節単位に対応する．<P>は転記基本単位境界（transpription.pdf）であり，その箇所に原則 200 ミリ秒以上のポーズが存在することを示す．人手修正作業の際にもどのように発話が行なわれているかを間接的に知るための参考情報として利用した．また，例中では義務的コメント（3.2 節）を「;○○」の形式で当該節単位の末尾に生起順に記入してある．「%○○」の部分は本マニュアルでの説明の便宜のために付与したものであり，コーパスにおいて付与されている情報ではない．なお，見易さを考慮し，フィラーF や語断片 D などのタグが付与された要素（transcription.pdf）は削除して示している．

――＜修正前＞――
タイトル<P>夢の国ディズニーワールド<P>私は旅行が大好きで<並列節デ>今までもあちこち行きましたけれども<P>/並列節ケレドモ/
その中で一番楽しかった旅行をこれからお話しいたします<P>[文末]
――

↓

――＜修正後＞――
タイトル<P>－**;体言止**
夢の国ディズニーワールド<P>－**;体言止**
私は旅行が大好きで<並列節デ>今までもあちこち行きましたけれども<P>/並列節ケレドモ/ **+** その中で一番楽しかった旅行をこれからお話しいたします<P>[文末] **;主題の共有**
――

前節で述べたように，[絶対境界]と/強境界/で分割された単位を「デフォルト単位」と呼ぶ．人手修正作業では，このデフォルトの分割結果に対して，必要に応じて「デフォルト単位をつなぐ」あるいは「デフォルト単位を切る」という作業を行う．作業者は，以降に示す一連の基準に基づいて，デフォルトの分割結果を見ながら「つなぐ」あるいは「切る」ための明確な理由がないかどうかを判断し，必要な場合には操作を行った．

実際の人手修正作業では，各講演について，一次作業（3 名）→比較統合作業（1 名）→最終判定（1 名）と手順で作業を行った．その際，一次作業者３名の作業結果に不一致がある場合には，比較統合，最終判定とも必ずしも多数決で機械的な判断をするのではなく，適切だと思われる作業結果を採用してよく，また，最終判定では一次作業，比較統合の結果とは異なる操作を行ってもよいこととした．また，一次作業，比較統合作業，最終判定のいずれにおいても音声を参照した．

3.2 操作の種類と義務的コメント

人手修正作業においては，表２の三種類の操作を行うだけでなく，下記で定義されている操作の種類のうちのどれを適用したかを「義務的コメント」として記録した．

――
最後はプレジャーアイランド<P>－**;体言止**
ここはディスコの島でして<P>/テ節/ **+** もう<P>夜から十時からオープンの島です<P>[文末] **;主題の共有 % 「ここは」が「ディスコの島でして」と「オープンの島です」の両方に係る．**
――

操作の種類としては以下のものを定めた．原則として，義務的コメントごとに行うべき操作が一意に定められている[5]．本節の以下の部分では，これらの義務的コメント別に適用基準と操作の方法を解説する．

[5] 「タグミスーその他」のみ例外となる．

1．係り受けに関係するもの：

+	主題の共有，主題の飛び越し，格要素の飛び越し

2．日本語の文法構造に関係するもの：

−	体言止
{ : }	引用節構造，連体節構造

3．自発的な話し言葉の特徴に関係するもの：

( )+	挿入節，挿入文，フィラー文
<< >>−	倒置
−	言いさし，と文末

4．節間関係や談話のレベルに関わるもの：

−	話題導入表現，直後がまとめ表現，話題の転換点， 大きい切れ目ー係り先なし，大きい切れ目ーその他

5．形態素・節境界ラベル等の問題に対処するもの：

+	連体形，間投助詞，言い直しマーカー，述語の言い直し，例文など， 格助詞相当表現，非文末，強→弱，タグミスーその他
−	文末候補，タグミスーその他

表３：人手修正操作（義務的コメント）一覧

3.3 コメント別操作方法

3.3.1 係り受けに関係するもの

係り受けとは文節を単位として節単位の範囲内で文節間の統語的関係を付与するものである（dependency.pdf）．述語に係る要素に関しては，大きく分けて，1.主題，2.格要素，3.副詞要素，4.述語，という区分が可能だが，本作業ではこれらのうち，**述語から述語への係り受け（いわゆる「連用修飾」）はデフォルト境界を人手でつなぐ理由とはしない**ことに注意してほしい（3.3.4 も参照）．

【主題の共有 + 】[6]

「は」や「も」でマークされる主題文節は強境界を越える広いスコープを持つ場合があるため，先行する節の中にある主題文節が強境界を挟む二つの節のいずれともに係っている（主題が共有されている）と判定された場合には，強境界の直後が「+」で結合される．

私は旅行が大好きで<並列節デ>今までもあちこち行きましたけれども<P>/並列節ケレドモ/ + その中で一番楽しかった旅行をこれからお話しいたします<P>[文末]**;主題の共有**

それに対して<テ節>従来の手法は<P>やはり逆転現象が見られまして/テ節/ + ここが有意なくらいに差が逆転しています<P>[文末]**;主題の共有**

【主題の飛び越し + 】

強境界を挟む二つの節の間で，先行する節の中にある主題文節が，その節の中には係らず，後続する節にのみ係っていると判定された場合，強境界の直後が「+」で結合される．

で<接続詞>ここは毎晩十二時になりますと<P>/条件節ト/ + カウントダウンショーが行なわれます<P>[文末]**;主題の飛び越し**

---

[6] 以下【OO】は義務的コメント名を表しており，記号やブラケットは当該義務的コメントが付与される場合の操作を表す．

【格要素の飛び越し ＋ 】

強境界を挟む二つの節の間で，先行する節の中にある格要素（格助詞でマークされる名詞句及び副詞）がその節の中には係らず，後続する節にのみ係っていると判定された場合，強境界の直後が「+」で結合される.

太郎が 12 時になりますと/条件節ト/**+ 来るはずです**[文末] **;主題の飛び越し**

＊注意：

・「主題の共有」「主題の飛び越し」「格要素の飛び越し」によってデフォルト境界の直後が結合されるのはこの境界が強境界の場合のみであり，絶対境界をまたいではならない.
・「格要素の飛び越し」は認めるが，「格要素の共有」は認めない.
・接続詞やフィラーなど，係り受け関係付与の対象外となる要素については，共有や飛び越しの認定はしない.
・「主題の飛び越し」「格要素の飛び越し」は「挿入節」と判断を迷う場合もある．挿入節との区別については 3.3.3 節参照.
・「は」でマークされた主題が意味的には複数の節単位に係ると思われる（つまり文主題ではなく談話主題である）場合には，この主題の直後で単位を分割してもよい（3.3.4 節の「話題導入表現」参照）.

3.3.2 日本語の文法構造に関係するもの

CBAP-csj は日本語の局所的な形態素列のみから節境界を検出するため，以下のような場合にはデフォルト境界認定に問題が生じる.

・講演のタイトルなど，名詞句だけで独立の節が構成される場合，当該部分は前後の節から統語的に独立しているが末尾に述語句を持たないため，CBAP-csjでは境界を発見できない.
・引用部分の途中に絶対境界・強境界が現れた場合，CBAP-csj では主節の述語句よりも統語的に先に生起する埋め込み節の述語句が誤ってデフォルト境界として認定されてしまう.

なお，3.3.3 節の諸現象とは異なり，これらの現象は話し言葉に特有のものではなく，書き言葉を対象とした場合にも同様に問題となるものである.

【体言止 － 】

・名詞一語のみで発話が終わっている場合や「NP です」の「です」のような述語が省略されている場合，デフォルト単位の冒頭に「はい」のようなつなぎの感動詞がある場合には，その位置で人手でデフォルト単位を分割した.
・名詞句が不自然につながっている場合や，係り先がない場合，体言止になっているケースが多い.
・接続詞や副詞が前接する場合もある.

まずエプコットセンター<P>－**;体言止**
ここは各国のパビリオンがあって<P><テ節>規模が大きい万博っていう感じです<P>[文末]

それから<接続詞>話者は男性一名 －**;体言止**
これは私でございます<P>[文末]

はい<P><感動詞>－**;体言止**
タイトルは不運続きで<P><並列節デ>忍耐の<P>一年だったっていう<トイウ節>ことで話したいと<引用節>思います<P>[文末]

● 引用節構造・連体節構造

【引用節構造｛ ： ｝】

引用節「～と」「～って」やトイウ節「～という」,「～など」「～なんて」「～みたいな」などの引用構造の内部にデフォルト境界が含まれていた場合，デフォルト境界直後を：で結合すると共に埋め込み節を｛ ｝で囲むことによって，望ましくないデフォルト境界の直後を結合している.

それから<接続詞>この｛ はい<感動詞>そうです[文末]：月曜日です ｝って<引用節>相手の質問を繰り返したりすることが多々ありますが/並列節ガ/;引用節構造

ただし，操作の対象となるのは発言の明確な引用などの場合だけでなく，形式上埋め込み節となっており，その途中に望ましくないデフォルト境界が含まれている場合全てを対象とする.

で<接続詞>｛ 疑いでは第二ホルマントが相対的に高い<P>[文末]：で<P><接続詞>感心では第二ホルマントが相対的に低い<P> ｝という<トイウ節>ような<P>関係がこれは現在までに音響的な特徴を観察しました四名の話者に<P>一貫して<テ節>観察されました<P>[文末];引用節構造

【連体節構造｛ ： ｝】

連体修飾節の内部に絶対境界，強境界が含まれる場合，引用節構造の場合と同様の操作を行う.

｛ 逆転はしていませんけれども/並列節ケレドモ/：差が詰まっている ｝ことが見て<P><テ節>取れると<引用節>思います<P>[文末];連体節構造

で<接続詞>そこで<接続詞>｛ 二十ミリずつ<P>伸ばした/文末/：あるいは縮めた ｝場合にそれぞれの数値がどうなるかを検討します<P>[文末];連体節構造

＊注意：

・｛ ｝の範囲は文脈に基づいて判定する．その際，｝は文節の途中に生起することが不可避だが，｛は極力文節の途中には入れないようにする.
・引用節構造・連体修飾構造の内部にデフォルト境界が含まれていない場合には，たとえ内容上引用であることが明確な場合にも｛ ｝で囲む操作もコメントも不要である.
・この操作によってきわめて長い節単位が生じてしまう場合には，引用マーカーの直前や直後で分割してもよい（3.3.4節の「話題導入表現」「直後がまとめ表現」参照）.

【引用節構造ー内部切断｛ ー: ｝】

・引用節に相当する部分の内部について，例えば「体言止」などを認定する必要がある場合には，これらの直後を一旦人手で切断し，さらに「人手認定した文末」をデフォルト境界に準ずるものとみなし，「:」も挿入する.

で<接続詞>このように<P>分かりにくさを記述する為の情報として<P><テ節>｛ 分かる為にどういう情報が必要かという<トイウ節>観点<P> ー: それから<接続詞 L>問題の所在がどこにあるか<P>[文末候補]：それから<P><接続詞 L>分かりにくさの程度が<P>どういうものがあるか ｝と<引用節>いったような<P>幾つか観点が<P>必要になってくるという<トイウ節>ことが分かってきました<P>[文末];引用節構造ー内部切断;引用節構造

【連体節構造－内部切断｛ －: ｝】

・引用節構造－内部切断の場合に準ずる.

＊注意：

・原則として，他のデフォルト境界に関して「引用節構造」「連体節構造」の操作が必要になることによって｛｝が生じた場合に，さらにその引用節の範囲内に切断すべき箇所が生じている場合のみを操作の対象とする.

・従って，｛ ｝の範囲内に「:」でつながれるデフォルト境界が存在しない場合に，「内部切断」だけを行うことは任意である.

【引用節構造－末尾境界｛ }+ 】

・デフォルト境界が引用節の末尾にのみあり，そのことによりデフォルト単位に係り受けなどに関する問題が生じている場合，引用節部分を {} で囲むが，｛｝内に「:」や「+」を用いるのではなく，} の直後に「+」が付与される.

それで<接続詞 L>｛ もうこの<P>(A エー;A)さんとは<P>終わったんだな[文末候補]｝+ っつって<P>/引用節 /;引用節構造－末尾境界

3.3.3 自発的な話し言葉の特徴に関係するもの

CSJは自発的な独話のコーパスである．自発的な独話においては「何を始めに言い，何を次に言うか」という線状化問題が特に大きなものとなり，前もって形成されていた発話プランが発話途中で変更される場合がある．こうして生じた挿入節や倒置，発話中止などのさまざまな非流暢現象が望ましくないデフォルト単位を引き起こしている場合には下記のような人手修正を行った.

● 挿入

【挿入節（)+ 】

・発話の途中で発話プランがすることによって，途中まで開始された節の途中に強境界を伴う別の節が「但し書き」的に挿入されている場合には，デフォルト単位のままでは係り受け関係などに問題が生じる場合に限って，これらを「挿入節」と認定し，分割されている強境界の直後を「+」で結合し，挿入部分を（）で囲む.

・挿入節と認定してよい強境界の種類は，/並列節ガ/，/並列節ケ(レ)ド(モ)/，/条件節ト/，/ヨウニ節/のみとする.

色んなパターンを<P>（ ここに書いてある数字は頻度ですが/並列節ガ/ ）+ たくさん集めてみました<P>[文末] ;挿入節

ホテルの<P>部屋の中も早速（ 夜着いたんですけども/並列節ケドモ/ ）+ チェックしました<P>[文末] ;挿入節

＊注意：

・デフォルト単位が以下のようなものであれば，係り受け関係はおかしくないので，「ここに書いてある数字は頻度ですが」は挿入節とは認定しない.

ここに書いてある数字は頻度ですが/並列節ガ/
色んなパターンをたくさん集めてみました<P>[文末]　→ そのまま

・弱境界の従属節が挿入節のように見えることがあるが，挿入節の対象表現ではなく，強境界をまたぐものでもないため，挿入節とは認めない.

もう最初の一日目は先程言ったように夜着いたので<P><理由節ノデ>とにかく<P>部屋の中でのミッキー探しだけで<P>三時間ぐらいはたってしまって<P><テ節>なかなか寝れませんでした<P>[文末]　→ そのまま

＊注意：主題や格要素の飛び越し（3.3.1）との区別

・挿入節の場合には，挿入節の直前の要素が直後で繰り返されている，音声上高さや速さの変化がある，意味内容的にメタ表現的機能を果たしている場合やスライドなどの外部情報を参照していることがある，などのいくつかの特徴が観察されることが多い．ただし，決定的な基準を発見するのは困難であるため，音声を含むさまざまな情報を総合的に参照して判断している．

これは<P>正直言って<P><テ節>学部<P>（私工学部だったんですけど/並列節ケド/）+ そちらの勉強は<P>殆ど<P>しておりませんで<P>/テ節/ + 学部の成績で選ばれたら<条件節タラ>多分落ちていたと<引用節>思うんですけれども<P>/並列節ケレドモ/ **;挿入節 ;主題の共有 %「学部」→「そちらの」**

【挿入文ー判定詞（)+ 】

・挿入文とは，挿入節の場合と同様，ある節の途中で別の節が開始されているが，問題となるデフォルト境界が絶対境界の場合である．

・デフォルト単位のままでは係り受け関係がおかしくなる場合や通常の文の内側に別の文が注釈のような形で挿入されていると判定されできる場合に限り，「挿入文」と認定し，挿入節の場合と同様の方法で操作を行う．

・挿入されている部分の末尾が「です」のような判定詞であり，「名詞＋です」のようなきわめて短い節である場合には「挿入文ー判定詞」とする．

友人のエム君<P>で<接続詞>彼はですね<P>中森明菜（当時の<P>松田聖子さんと<P>双肩<P>結構双璧をなす<P>アイドルですね<P>[文末候補]）+ で<接続詞>こちらの<P>追っ掛けをやっておりました<P>[文末]；**挿入文ー判定詞 %挿入文の末尾が判定詞「です（ね）」.**

【挿入文ーその他（)+ 】

・複数の挿入文を含む部分が「の三種類」などの表現でまとめ上げられる場合である．

まず音源とマイクロホンとの距離が<P>音響モデルの学習時と<P>その<P>音声の認識時で同じ場合<P>（これをマッチドエイチエムエムと呼びます<P>[文末]）+ そして<接続詞>最大尤度基準に基づき<P><連用節>空間音響特性依存エイチエムエムを<P>選択して<テ節>認識した場合<P>（これをライクリーフットセレクションと呼びます<P>[文末]）+ そして<P><接続詞>まず音響の特性を学習していないエイチエムエムで認識した場合<P>アイピーエーエイチエムエム<P>の三種類の認識率で比較します<P>[文末] **;挿入文ーその他 %「の三種類」によってそれまでの部分がまとめられる.**

・位置的に挿入文が文節の途中や連体修飾関係にある文節間に生起する場合もある．

これはまずさまざまな空間音響特性を学習したエイチエムエム<P>（これを空間音響特性依存エイチエムエムと呼びます<P>[文末]）+ を複数用意します<P>[文末] **;挿入文ーその他 %「エイチエムエムを」が一文節.**

陸上<P>の（トラックって言うんですか[文末候補]）+ 四百メートルトラックも<P>あります<P>[文末] **;挿入文ーその他 %「陸上の」→「四百メートルトラック」.**

＊注意：

・一度発話された要素が，後方で同じ形または別の形で言い直される場合には，デフォルト単位のままで問題はないことが多く，その場合操作は不要である．

カレーの場合これをシーとします[文末]
うどんの場合これをユーとします[文末]
親子丼の場合これをオーとします[文末]
お寿司の場合これは回転寿司も含みます[文末]
これをエスとします[文末]
これら四つの場合メニューの人気度を比較しています[文末]　　→ そのまま

【フィラー文（)+ 】

・フィラーに相当する文形式が発話の途中に出現する場合には，挿入文と同様の操作を行う．
・「何て言うんですかね」「何て言うんでしょう」など定型的表現となる．

ちょうど中学校三年生で<並列節デ>こう<P>（何て言うんですかね[文末候補]）+ 人に会って<P><テ節>例えば土を掘って<テ節>鳥を埋めてるようなところを見つかると<条件節ト>ちょうど恥ずかしいという<トイウ節>ような年頃だったと<引用節>思うので<P><理由節ノデ>…;フィラー文

＊注意：

・意味的にはフィラー文とも見なせるが，デフォルト境界を含まない場合には操作は不要である．

で<接続詞>自分よりできる子が何か自分よりこう何と言うか下の役割みたいなものに決められてしまって/テ節/　→ そのまま

・きわめて少数だが，フィラー文の中で倒置が生じていることにより，例外的に（+）という構造が生じる場合がある．

● 倒置

話し言葉では，文節がその係り先となる述語句の直後に置かれる「倒置」が多く発生する．

倒置分類

倒置要素 RRR の前/後がデフォルト境界（| |）になっているか否かに応じて，次の 4 種類の区別が可能である．

	直前	直後	表記		操作
A	デフォルト境界	デフォルト境界でない	XXX | | RRR　YYY	→	「倒置ー つなぎ切り」
B	デフォルト境界	デフォルト境界	XXX | | RRR | | YYY	→	操作不要
C	デフォルト境界でない	デフォルト境界でない	XXX　RRR　YYY	→	「倒置ー後切り」
D	デフォルト境界でない	デフォルト境界	XXX　RRR　| | YYY	→	操作不要

＊注意：

・すべての場合において，倒置要素の直後は最終的に切れているというのが原則である．
・従って，デフォルト境界になっていない箇所を人手で切断する操作が必要な場合（A，C）のみを修正の対象とし，最初からデフォルト境界となっている箇所をそのまま修正しない場合（B，D）には操作は不要である．
・倒置要素の直前がデフォルト境界で，直後もデフォルト境界である B の場合には，デフォルト状態では倒置要素だけが独立した節単位になっており，この部分から直前の節単位への係り受けは「述語間の係り受け」となるため，人手でつなぐことはせず（3.3.4 参照），操作，コメントとも不要となる．
・ある要素が倒置か否かの判断は，音声情報を特に参照して行う．

【倒置ーつなぎ切り +<< >>- 】

・倒置要素の直前がデフォルト境界で，直後がデフォルト境界でない場合である.
・倒置要素の直前を結合し(+)，倒置された要素を << >> で囲み，倒置要素の直後を切断する(-).

なかなか言えないんですね[文末]**+<<** やっぱり<P>取ったでしょって<P>**>>-　;倒置ーつなぎ切り**
で<接続詞>結局ー か月ぐらい凄く悩んだんですけど<P>/並列節ケド/

【倒置ー後切り << >>- 】

・倒置要素の直前がデフォルト境界でなく，直後もデフォルト境界でない場合である.
・倒置要素の直前はつながっているのでそのままとし，倒置された要素を << >> で囲み，倒置要素の直後を切断する(-).

イラン人と目が合っても<P><テモ節>騒いだり指を指したり<P>してはいけない<P>[文末]
それでもし腹の立つことがあったら<P><条件節タラ>笑いながら怒れと<引用節>**<<**日本語で<P>**>>-;倒置ー後切り**
そういう風な先生の指示がありまして<P>/テ節/

＊注意：

・節単位認定作業では上記の基準に合致した倒置部分のみを認定するため，本作業で倒置と認定されなかった箇所が係り受け作業で倒置だと認定される場合もある（dependency.pdf）.

● 言いさし

・挿入や発話プランの変更により，発話が途中で言いやめられたり，文の構造が発話途中から変わってしまうことで，係り先のない要素ができる場合がある.
・特にデフォルト単位の冒頭に多く発生する.
・ただし，接続詞やフィラーなどもともと係り受け関係が付与されない要素については，極端に長い休止がある場合などを除き，わざわざ言いさしとして認定する必要はない.

【言いさしー係り先なし - 】

・ある要素の係り先がなく，かつ後方に言い直しの表現がない場合，「言いさし」と認定し，その位置でデフォルト単位を切断する.

今回の実験で<P>-**;言いさしー係り先なし**
この<P>あの<P>次ページの表に示しましたのは実験条件と実験結果の一覧です[文末]

【言いさしー言い直しあり - 】

・言いさしの要素は，後になって言い直されることが多い.
・言いさし要素と言い直し要素の間にデフォルト境界が介在する場合，「言いさし」要素の直後でデフォルト単位を分割し，独立させる.
・同一表現でなく，指示詞による場合も言い直しに含める.

ここの目玉は<P>シャークリーフと言いまして/テ節/
シュノーケルで<P>-**;言いさしー言い直しあり**
海の中とそっくりに作られた大きいプールがありまして<P>/テ節/
そこにシュノーケルで入るんですけれども/並列節ケレドモ/

＊注意：

・言いさしと言い直しの間に絶対境界，強境界が入らない場合はそのままでよい．

正解率今回の<P>提案手法であります音素で見た時の正解率がこのようになっております<P>[文末]
→ そのまま

【と文末 - 】

・引用節（「～と」「～って」）の形で発話が終わっていると思われる場合がある．
・通常[と文末] という節境界ラベルがついていることが多いが，このラベルがなくても，「～と」などで発話が切れていると判断された場合は，その位置でデフォルト単位を切断する．

これもデュレーションの結果ですが/並列節ガ/
全体が一様に伸縮する訳ではなくて<テ節>例えば最後が大きくまた<接続詞>頭がそれに次いで<テ節>大きく真ん中はあまり動かないと<引用節>-;と文末
これはモーラ単位で示しておりますが/並列節ガ/
そういう結果もこれもかなりコンシステントに観察される変化であります[文末]

3.3.4 節間関係や談話のレベルに関わるもの

以下のような理由に基づき，意味内容的にはこの部分全体がつながっているようにも感じられる場合にも，「述語から述語への係り受け」（連用修飾）はデフォルトの境界の直後を人手で文をつなぐ根拠とはしていない．

・「述語から述語への係り受け」が感じられるかどうかは，それぞれの述語の「意味的な」特徴によって左右されるものであるため，必ずしも純粋に統語的な関係だとは考えられない（例えば「本を」→「読む」という係り受けと比較してみれば明らか）．
・書き言葉の一文内における「述語から述語への係り受け」は著者が意味的なつながりを判断し推敲した結果決められたものであるため，ある程度の意味的なつながりが保障されているのに対して，話し言葉である CSJ ではこうしたことは当てはまらない．
・そもそも自動で「デフォルト単位」を認定しているのはこうした述語から述語への係り受けについての判断が作業者間で判断が一致しないことが多く，作業量も膨大なものとなってしまうためである．

それで<P>顔は雛の時に買って<テ節>二匹一緒に買ってきたので<P><理由節ノデ>そのせいか<P>どうか知らないですが/並列節ガ/
顔は<P>お互いに似てるので<理由節ノデ>見分けが付かなかったものですから<P><理由節カラ>顔の周りの<P>赤の濃さで見分けを付けていました[文末]

ここで，意味的には「買ってきた（ので）」→「（見分けが）付かなかった」のようなつながりがあると感じられるかもしれないが，こうした「述語から述語への係り受け」をデフォルト境界の直後を人手で結合する根拠としてはならない．

しかし，「述語から述語への係り受け」を理由にデフォルト境界の直後を人手で結合することが認められないため，

1．「AAA BBB || CCC」で，形式上は BBB の直後がデフォルト境界だが，内容的には BBB と CCC がまとまる（並列など）と感じられる場合
2．「AAA || BBB CCC」で，形式上は AAA の直後がデフォルト境界だが，内容的には AAA と BBB がまとまる（並列など）と感じられる場合

のそれぞれにおいて，全体を一つの節単位として結合することはできない（AAA，BBB，CCC は節，「||」はデフォルト境界を表す）．

そこで，主にこうした理由により不都合が感じられる場合について，逆に，1では「AAA の直後を人手で切る」，2では「BBB の直後を人手で切る」ことを可能にするための操作を定義した．なお，本 3.3.4 節で解説する各操作については，音声情報によって話者の意図が明確になる場合も多いため，音声情報に特に注意して作業を行った．

【話題導入表現 − 】[7]

・談話レベルの構造がデフォルト単位と交差してしまう場合がある.
・例えば,「AAA BBB || CCC」の構造において,談話構造の観点から見たときには節AAAがBBBとCCCにわたって続く話題を挿入しており,内容的にBBBとCCCがまとまる(並列など)と感じられる場合には,節AAAを「話題導入表現」と見なし,AAAの直後を人手で切断してよい.

それが分かったのはなぜかと<引用節>言うと<条件節ト>−**;話題導入表現**
コザクラインコっていうのは全体的には背中が緑色<P>で<P><並列節デ>青色がちょっと混ざったような感じなんですけれども/並列節ケレドモ/

でも<接続詞C>その写真の撮り方はやはり<P>日本とは違っていて<P><テ節>−**;話題導入表現**
日本のディズニーランドに行くと<P><条件節ト>みんな我を我をで<P>ミッキーを囲んじゃって<テ節>一体誰<P>と何か知らない人と私は写真を撮ってる<P>っていう<トイウ節>状態ですけど<P>/並列節ケド/
アメリカは<P>ちゃんと順番を待っていて<P><テ節>私がミッキーの側に行くと<条件節ト>周りの人は寄ってきません<P>[文末]

・「話題導入表現」としては,次のような「疑問詞+発言動詞(メタ表現)」のパターンが典型的である.

どういったことで始まったかって<引用節>言うと<P><条件節ト>−**;話題導入表現**
何が有名<P>あるかと<引用節>言うと<条件節ト>−**;話題導入表現**
それはどういうものかと<引用節>言うと<条件節ト>−**;話題導入表現**

・しかし,「話題導入表現」であるかどうかを内容から判断しなければならない場合や一部の「〜は」の箇所のように弱境界でもない箇所を「話題導入表現」として切断する場合もある.

その写真の撮り方はやはり<P>日本とは違っていて<P><テ節>−**;話題導入表現**
印象的だったのは<P>−**;話題導入表現**

【直後がまとめ表現 − 】

・「AAA || BBB CCC」で,形式上はAAAの直後がデフォルト境界だが,内容的にはAAAとBBBがまとまる(並列など)と感じられる時,CCCが複数文にわたって続いた話題をまとめる表現である場合がある.典型的には,「話題導入表現」と対になって用いられることが多い.

話題導入表現 − 語りの本体 − まとめ表現

・特に過去の出来事についての描写から現在の視点からの回顧に移行する箇所では,「直後がまとめ表現」となりやすい.

で<P><接続詞>挙げ句の果てはクリスマスの歌を<P>ドライバーさんと交えて<P><テ節>歌を歌って<P><テ節>ドライバーさん運転大丈夫っていうぐらい<P>もう大騒ぎで<P><並列節デ>−**;直後がまとめ表現**
あんな騒ぎながらバス乗ったのは初めてでした<P>[文末]

何か色々と楽しい催し物を<P>企画してくれて<テ節>そいでそれに<P>放課後は参加したりと<P><引用節>−**;直後がまとめ表現**
毎日が本当に学生に戻って<テ節>とても忙しくて<テ節>楽しかったです<P>[文末]

---

[7] 談話境界認定(discourse.pdf)においても談話セグメントの開始地点の特定が行われており,その際節単位認定における「話題導入表現」などのコメント情報も参照されているが,両作業では観点が異なるため,義務的な対応があるわけではない.

やっぱり<P>中絶が反対だと<引用節>言う<P>人達は<P>キリスト教で<並列節デ>そういう<P>人間の<P>せっかくの生命を<P>殺してはいけないという<P><トイウ節>意見が凄く多かったので<理由節ノデ>－**;直後がまとめ表現**
日本にはない宗教の違いを物凄く実感しました<P>[文末]

＊注意：

・「まとめ表現」の認定を行いたい場合にも，文節の途中で節単位を切断してはならない.

【話題の転換点 － 】

・デフォルト境界ではないが，話題が明らかに転換されている場合がある.
・特に，当該位置の直後に「後」のような接続詞の使用が使用されている場合が多い.

やっぱり<P>お財布とかに<P>結構気を付ける<P>ようになったっていう<トイウ節>のと<P> － **;話題の転換点**
後<P>何かこう<P>それまでは人をあんまり疑う<P>こととかはなかったのに<P><ノニ節>この人って<P>ってやっぱり一緒に仕事してても<テモ節>そういうことがある訳だから<P><理由節カラ>やっぱり考えるようになっちゃった<P>[文末]

【大きい切れ目ー係り先なし － 】

・末尾がデフォルト境界でないある節が同じ単位内に意味的に明らかに係り先を持たない場合，当該部分を切ってもよい.
・特に，「AAA BBB || CCC」において BBB が挿入節的なものである場合，節 AAA から CCC への係り受けは「述語間の係り受け」であり，この挿入節相当部分を飛び越して人手でデフォルト境界をつなぐことはできないため，逆に AAA の直後を切る場合などがある.

それで<P>顔は雛の時に買って<テ節>二匹一緒に買ってきたので<P><理由節ノデ>－**;大きい切れ目ー係り先なし**
そのせいか<P>どうか知らないですが/並列節ガ/ **%挿入節的**
顔は<P>お互いに似てるので<理由節ノデ>見分けが付かなかったものですから<P><理由節カラ>顔の周りの<P>赤の濃さで見分けを付けていました[文末]

・また，「AAA || BBB CCC」において，BBB が倒置された節であると感じられる場合，一方で意味的に BBB→CCC の係り受けはなく，他方で BBB→AAA の係り受けは述語間の係り受けであるため倒置を認定することはできないため，BBB の直後を切る必要がある.

で<接続詞>青山の昔のお話私が住んでた頃の昔のお話にこう戻していきたいと<引用節>思うんですが/並列節ガ/
現在の青山はあんまりよく知らないものですから<P><理由節カラ> － **;大きい切れ目ー係り先なし %倒置部分の可能性あり.**
昔青山<P>小学校も青山だったものですから<P><理由節カラ>あの辺りでも<P>生まれ育ち<連用節>かなり遊んだという<トイウ節>ことで<並列節デ>遊んだ場所ということで<P><並列節デ>青山の中の色々幾つか挙げていきたいと<引用節>思いますが<P>/並列節ガ/

・ただし，「大きい切れ目ー係り先なし」はこれらのパターンに限らない.

【大きい切れ目ーその他 － 】

・以上のいずれの下位分類にも該当しない場合にも，例外的だが明確な理由がある場合にはデフォルト単位を切断してもよい.

言語モデルですと/条件節ト/
例えばパープレキシティーという指標があったり<タリ節> － **;大きい切れ目ーその他 %ここで切らないと明確な並列関係が失われるため.**
デコーダーですと/条件節ト/
サーチエラー率<P>というものがあったりしまして<P>/テ節/

3.3.5 形態素・節境界ラベル等の問題に対処するもの

CBAP-csj によるデフォルト境界認定は形態素情報に基づいて機械的に判定されるものであるため，CBAP-csj に登録されていない表現が生起した場合や，あるいは元になる形態素情報についての判断が節境界認定作業における判断と相違している場合には，デフォルト単位を人手で修正する場合がある.

【連体形 ＋ 】

形態素情報において「連体形」となるべきところが誤って「終止形」となっているとデフォルトで [文末] となってしまうため，当該の望ましくないデフォルト境界の直後を人手で結合する必要がある[8].

一つは今言いました[文末] **+**コミュニティー内の情報の個人化をしたら<条件節タラ>いいんじゃないかという<トイウ節>論点です<P>[文末] **;連体形**

【間投助詞 ＋ 】

・文末に付く終助詞「ね」と同様のものが文末以外の文節末についたものを「間投助詞」と呼ぶ.
・文末でない位置に出てくる「ですね」の多くは間投助詞「ね」の丁寧形であり，同様に間投助詞であるとみなしてよい.
・デフォルト単位認定においては，こうした間投助詞の「ですね」が誤って文末であると判定されている箇所があるため，デフォルト境界の直後を人手で結合する必要がある.
・典型的な形式は「名詞＋格助詞＋ですね」「副詞＋ですね」「名詞＋ですね＋格助詞[9]」である.
・「ですね」の他，「ですか」や「ね」の省略された（聞き取れない）「です」の場合もある.

後程ですね[文末候補] **+** 聞いたところで<P>ございますと<P>/条件節ト/ **;間投助詞**

できる限りですね[文末候補] **+** 自分の<P>ペース<P>で勝利に結び付けようと<引用節>思いまして<P>/テ節/ **;間投助詞**

【言い直しマーカー ＋ 】

・「と言いますか」のような言い直しのメタ表現の直後が不適切なデフォルト境界となっている場合には，人手でつなぐ必要がある.

---

[8] 終止形と連体形の区別については，特に連体修飾述語と被修飾名詞の間に複数の形態素が介在している場合には，人手でも判断が揺れる場合がある．また，節単位認定作業対象となった講演のうちの一部については形態素情報が自動付与されているものもあり，これらのファイルでは終止形と連体形の区別に関する誤りが多くなっている.

[9] 「名詞＋ですね＋格助詞」は主に新情報の提示や共有知識の確認に関わるものであると思われ，統語的にも通常の間投助詞の場合とは異なるが，今回の作業では同様の扱いとした.

そうして<テ節>お葬式も<P>祖母の<P>遺言と言いますか[文末候補] + 言い残していったことでお葬式はしないで<P><テ節>それを<P>大学病院に<P>献体して<P><テ節>自分の体をこれからの医療に役立てて<テ節>ほしいという<トイウ節>ことで<P><並列節デ>～ ;言い直しマーカー

それで音素タスクの方は言ってみれば<条件節レバ>調音運動の目標と言いますか[文末候補] + ターゲットを表わすようなもので<P><並列節デ>連続性条件の方は調音器官の動的な振る舞いを<P>規定するものであると<引用節>考えられます<P>[文末] ;言い直しマーカー

＊注意：挿入文，フィラー文との区別の基準

・挿入文の場合，例えば「○○です」のような表現のうちの「○○」が（ ）で括られてしまったとしても，主節には統語的な問題は生じない.

・フィラー文の場合にはこの部分全体が「フィラー的」であるのに対して，「言い直しマーカー」の場合には，たとえその直前の要素が「言いさし（言い誤り）」であるとしても，この言い直しマーカー自体が言い直しであるわけではないため，この部分を（ ）に入れてしまうと，その前後の統語的なつながりが不明になる.

【述語の言い直し + 】

・述語の言い直しがある場合，言い直される述語の直後が誤って[文末]と判断されてしまう場合がある.

・こうした場合には，二つの述語が同一節単位内になるよう，望ましくないデフォルト境界の直後を人手でつながなければならない.

・述語の活用形が変更されるような単純な場合だけでなく，同義の別の語に言い換えられる場合や最初はなかった格要素等が追加されて言い直される場合もある.

みんなで揃って<テ節>大学へ行く<P>[文末] + 行きました<P>[文末] ;述語の言い直し

後は話がちょっと外れます[文末] + 飛びますが<P>/並列節ガ/ ;述語の言い直し

どうもありがとうって言われた時は<P>良かった[文末] + タイミングが良かったような<P>～ ;述語の言い直し

・ただし，次例のように，言い直しではなく強調などの意図による繰り返しだと思われる場合には，操作は不要である.

で<P><接続詞>都心から一時間以内のベッドタウンなんで<P><理由節ノデ>めちゃくちゃ通勤電車とかは込んでます<P>[文末]
込んでます<P>[文末]　　→ そのまま

【例文など + 】

・特に言語系の学会発表などでは，例文や著書からの引用部分など，書き言葉ならかぎ括弧に入るであろう表現が頻出するが，これらの部分の途中か最後に終止形があると[文末]と自動判定されてしまう場合があるため[10]，人手でつなぐ必要がある.

[10] 転記においてMタグ（transcription.pdf）が付与されている場合には，この情報を利用して [文末] の付与を回避している.

浮かぶ[文末] + 選ぶ[文末] + 共に咄本では<P>ま行表記の方が<P>多い訳なんですけれども<P>/並列節ケレドモ/ **;例文など ;例文など**

【格助詞相当表現 + 】

- 表１の「メタルール」のように，弱境界<テ節> の直前が丁寧形の場合や直後が接続詞の場合には，これが強境界の /テ節/ に変更される.
- しかし，「に対しまして」「に関しまして」「によりまして」「にあたりまして」「につきまして」「をもちまして」のような表現は，形式上は動詞だが，実際には他のより実質的な述語に係る格要素的な役割を果たしている（英語の前置詞と機能的に類似）場合が多いため，これらの表現は「格助詞相当表現」と見なし，強境界でなく弱境界とした.
- しかし，CBAP-csj では「格助詞相当表現」として扱われていない表現が機能的に格助詞相当表現だと思われる場合や，あるいは次例の場合のように，格助詞相当表現だと思われる「において」の直後に接続詞が生起するような場合には，メタルールの適用によってこの格助詞相当表現の直後がデフォルト境界となってしまうため，この箇所を人手でつなぐ必要がある.

で<接続詞>日本語のピッチアクセントは単語認知において/テ節/ + しかしながら<接続詞 L>英語に比べては<P><テハ節>語義活性化に大きく関与してるという<トイウ節>報告がされています[文末] **;格助詞相当表現**

【非文末 + 】

- 特に慣用表現の場合や転記の際に一部の音声が聞き取れなかった場合などに，「終止形」だが望ましくない箇所にデフォルト境界が挿入されてしまうことがあるため，こうした箇所は人手でつなぐ必要がある.

うちの犬がそういう音が嫌いで<並列節デ>大騒ぎして<テ節>植木は倒すわ<P>[文末候補] + 網戸は破くわで<文末候補>大変です<P>[文末] **;非文末 %「A だわ B だわ」という定型表現.**

で<接続詞>涙が出る[文末] + 出ない<P>っていう<トイウ節>ことで悲しい<P>とか<トカ節>そういうことを判断するってのも何かとは思いますけれども<P>/並列節ケレドモ/ **;非文末**

で<P><接続詞>更に<接続詞>原因を遡ると<P><条件節ト>学習データー<P>に付けられてる無音ラベル<P><ベル>そのものに問題がある<P>[文末] + ではないかという<トイウ節>ことで<P><並列節デ> - **;非文末** ;大きい切れ目－直後がまとめ表現 **%「の」が聞き取れなかったものと思われる.**

【強→弱 + 】

- 格助詞相当表現の場合を除き，節末表現「ますと」「ですと」「まして」「でして」は強境界となる.
- その中で，現在付与されている強境界のタグが実際には弱境界の方が望ましいのではないかと感じられるが，かといって「格助詞相当表現」でもないという場合があり，こうした箇所は人手でつないでもよい.
- 例えば，これらの表現の直後がデフォルト境界のままだとその前後に極端に短い節単位が生じてしまう場合などが対象となる.
- ただし，これを無制限に認めてしまうと，「境界の強弱の変更は認めない」という大原則が崩れ，節末ラベルの強弱を自動で付与したことの意味がほとんどなくなってしまうため，明らかに問題があると感じられる箇所についてのみ操作を行なう.

御覧いただきますと/条件節ト/ + 分かりますように<P>/ヨウニ節/ **;強→弱**

私が千歳烏山に住んでるっていう<トイウ節>ことを<P>言いますと<P>/条件節ト/ + 羨ましがられる<P>[文末] **;強→弱**

【文末候補 － 】

- 「のかな」のように，本来は文末となるべきだが，現行版の CBAP-csj では対応できていない表現や，明らかに文末的に使用されていると感じられる場合の「みたいな」「ような」などの直後を人手で切る必要がある場合がある．
- 特に，「みたいな」や「ような」については，常に文末的に用いられるとは限らないため，自動で文末に認定してしまうことができず，その都度適宜人手で判断・操作するしかない．

屋根がないと<P><条件節ト>雨が降った時にもしかしたら<条件節タラ>寒いのかな<P><文末候補> － **;文末候補 %「のかな」.**

その日取れた野菜とか<P>を<P>こう<P>山積みにして<P>/テ節/
で<接続詞>お金は<P>缶の中に入れてくださいみたいな<P>－**;文末候補 %「みたいな」.**
だから<P><理由節カラ>入れてる人もいれば<P><条件節レバ>入れてない人もいるから<P><理由節カラ>～

【タグミスーその他 + 】【タグミスーその他 － 】

- 上記のどの項目にも当てはまらないが，転記や形態素情報，CBAP-csj によって挿入された節ラベルなどに誤りや不都合があると感じられる場合には，デフォルト境界の直後を人手でつなぐ操作もデフォルト境界以外の箇所を人手で切る操作も許容する．原因が多様であり，一般化が難しいため，「タグミスーその他」として一括した．
- 結合の場合の典型的な例は並列に関わるものである．例えば，強境界である /並列節シ/ よりも後方の弱境界 <理由節ノデ> などの方が大きい切れ目であると感じられる場合のような，特定の強弱境界の組み合わせによる不都合や，弱境界の直後に並列の接続詞が生起することによってメタルールが適用されてしまう場合などである．

でも<接続詞 L>最近は真面目に働いてると<引用節>本人も言っていましたし/並列節シ/ + 周りの人も言ってたので<P><理由節ノデ>ちょっと<P>短期間ならいいだろうと<引用節>思って<テ節>その子も紹介しました<P>[文末] **;タグミスーその他 %「並列節シ」より「理由節ノデ」の方が大きい切れ目.**

で<接続詞>そうすると<条件節ト>それを最も徹底させ/連用節/ + かつ<接続詞>恐らくは意識的に行なっていたのが定家であるという<トイウ節>解釈が可能かと<引用節>思います<P>[文末] **;タグミスーその他 %連用節＋「かつ」.**

- 切断の場合の典型例は学会講演における対話例の引用やデモの提示箇所などである．

ここでは<P>虫よけ対策のスプレーについての<連体節テノ>話をしています<P>[文末]
(F あ)この前ね<P>(F あのー)あたしのいつも行き付けのねお花ねお花屋さんでね<P> － **;タグミスーその他 %予稿を参照し，対話例の話者交替を基準に切った.**
おお<P><感動詞>－;体言止
(F あのー)レモンの香り<P>－**;タグミスーその他 %予稿を参照し，対話例の話者交替を基準に切った.**
ああ<P><感動詞>－;体言止
レモンの香りのするスプレー売ってて<テ節>－**;タグミスーその他 %予稿を参照し，対話例の話者交替を基準に切った.**
ああ<P><感動詞>－;体言止
それは虫よけになりますって<引用節>言ったから<理由節カラ>何か － **;タグミスーその他 %予稿を参照し，**

対話例の話者交替を基準に切った.
うん<P><感動詞> －;体言止
おんなじようなものなんじゃない[文末] +<< 柑橘類のにおいって<P> >> －;倒置ー つなぎ切り ;倒置ー つなぎ切り

---

4. XML での格納方法と表示

・原則として談話>談話セグメント>節単位>文節>長単位>短単位という言語単位間の階層性が成り立つため，節単位は一つ以上の文節から構成されると定義できる.
・認定された節単位の末尾の短単位は必ずデフォルト境界の節境界ラベルか人手切断記号「-」を持ち，また原則として文節末になる[11].
・節単位途中のブラケットは文節の途中に入る場合がある．特に引用節構造，連体節構造の 〕は文節途中に挿入される場合も多い．当該文節は係り受け上は引用節構造，連体節構造内の文節ともその外部の文節とも同時に係り受け関係を持ちうる（dependency.pdf）.
・XML ファイルにおいては，特に節単位と文節の間の階層関係は直接的には表現されておらず，節単位認定に関する情報は全て SUW 要素（短単位に相当）の属性として格納されている（表４）[12].
・これらの情報のうち，節単位の範囲を表す「節単位 ID」は当該節単位の最初の SUW 要素の属性として記述されており，他の情報は特定の点を表すものであるため，当該の SUW 要素自体に付与されている.

ClauseUnitID (節単位 ID)	:	0 から始まる整数
ClauseBoundaryLabel (節境界ラベル)	:	表 1 参照
CU_OperationSign (節単位操作記号)	:	+, -, :, -: の 4 種類
CU_PreBraket (節単位前ブラケット)	:	(, {, << の 3 種類
CU_PostBraket (節単位後ブラケット)	:	), }, >> の 3 種類
CU_ObligateComment (節単位義務的コメント)	:	表 3 参照

表４：節単位関連情報の XML ファイルでの格納

・XML ファイルから節境界関連情報をテキストファイルなどの形式で取り出す場合，以下の方法が見やすいと考えられる.
  ・ClauseUnitID によって当該節単位の範囲を決める.
  ・範囲内の SUW 要素について，付与された情報を「CU_PreBracket, OrthographicTranscription, ClauseBoundary Label, CU_PostBracket, CU_OperationSign」の順に取り出し，一行として並べる.
  ・当該行の冒頭に ClauseUnitID を付与する.
  ・当該行の末尾に CU_ObligateComment を生起順に並べる．その際，各 CU_ObligateComment の冒頭に「;」などの特定の記号を付与するとよい.

・なお，節単位，係り受け，重要文抽出，談話のアノテーション結果を表示するためのビューワーを volume 18 に収めてある．このビューワーは XML 文書を表示用に変換した cuxml（/volume18/CU/ に所収）を InternetExplorer などのブラウザで表示できるようにしたものである．cuxml の仕様については/volume18/DOC/cuxml.pdf を，またビューワーの使用法については/volume18/DOC/cuxml_viewer.pdf を，それぞれ参照していただきたい.

---

[11] 文節認定基準(bunsetsu.pdf)に基づいて文節の後半に (?) や <FV>, <笑> などのタグ要素が含まれる文節について，これらの要素の直前にデフォルト境界の自動節ラベルが挿入されている場合には例外となる．こうした場合にのみ，デフォルト境界の節境界ラベルを持つ短単位が人手結合記号「+」「:」を持たず，かつ節単位末でない位置に生じることになる.
[12] 格納方法の詳細については，xml.pdfも参照のこと.

【参考文献】

丸山岳彦・柏岡秀紀・熊野正・田中英輝(2004)「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
南不二男(1974) 現代日本語の構造. 大修館書店.
南不二男(1993) 現代日本語文法の輪郭. 大修館書店.