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 —

高梨 克也[†] 丸山 岳彦[‡*] 内元 清貴[†] 井佐原 均[†]

[†] 独立行政法人 通信総合研究所 [‡] 独立行政法人国立国語研究所

[*] ATR 音声言語コミュニケーション研究所

1 はじめに

『日本語話し言葉コーパス The Corpus of Spontaneous Japanese (CSJ)』[2] は,『話し言葉工学』プロジェクトで作成されている,自発的な話し言葉の大規模コーパスである.このコーパスは,主に,「学会講演 (APS: academic presentation speech)」と「模擬講演 (SPS: simulated public speech)」という二種の独話から構成されており,今後の話し言葉研究のための有益な資源として用いられることが期待されている.

一般に,統語解析,機械翻訳,テキスト要約などの処理において,基本的な処理単位となるのは「文」である.しかし,これらは書き言葉を対象とした場合で,文の境界に句点が置かれていることが前提となる.一方,話し言葉にはそもそも句点が含まれておらず,CSJにも句点が書き起こされていないため,基本的な処理単位として文を利用することはできない.このようなCSJを対象として,自動要約,係り受け解析,談話構造分析[5]などを行なうためには,話し言葉の処理に有用な「文」相当の処理単位をあらかじめ定義し,検出しておくことが必要である.

本発表では,CSJから「文」境界を半自動的に検出するための手法について説明する.初めに,「文」境界の候補として,節境界を自動的に検出する手法について述べる.次いで,この自動検出結果を所定の基準に従って人手修正する方針について述べる.この人手修正は,体言止や倒置など,話し言葉に特有の現象に対して適用される.最後に,パイロット的に人手修正を行なった結果と,今後の作業に用いる支援ツールについて述べる.

2 話し言葉における「文」の境界

さまざまな言語処理や言語学的な分析に用いられる基本的な単位は,ほとんどの場合,「文」である.しかし,自発的な話し言葉を分析・処理の対象とする場合,文は必ずしも最適な単位であるとは言えない.話し言葉には,そもそも,文の終端をマークする句点が存在しないからである.さらに,自発的な発話の中から明確な文の境界を確定すること自体が難しいからである.自発的な発話では,常に完全な文の形が発話されるとは限らない.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かったり,時には語や文の断片だけで発話が構成されることもある.書き起こしの中に句点を含んでいないCSJを対象として,自動要約,係り受け解析,談話構造分析などを行なうためには,文に代わる有用な単位(話し言葉における「文」)を定義し,発話の中からそのような境界をあらかじめ検出しておく必要がある.

我々は,いわゆる文に代わる単位として,「節」を採用する.述語を中心として意味的なまとまりを成す節は,述語部分の形態的特徴に着目することによって,文よりも容易にその境界を特定することができる.そこで我々は,CSJの書き起こしテキストを節に分割することにより,これらを文に代わる単位として利用することにした.

3 節境界の自動検出

日本語では,述語の活用形や接続助詞などの形態素情報を利用することにより,非常に局所的な情報だけで,節の境界を自動的に検出することが可能である[3].そこで我々は,丸山ら[3]が作成した節境界自動検出ツールをCSJ仕様に改編し,CSJから節境界を自動的に検出するルールを作成した.

この節境界自動検出ルールは,ある形態素の前後1～3語を読み込んで,節境界の種類を判別し,その種類に応じたラベルをテキスト中に挿入するというものである.形態素情報は,CSJに付与されている形態素が「出現形_品詞_活用型_活用形」という四組に整形されて表現されている[6].ルールの実態はperlで実装されており,形態素列を読み込んで,あらかじめ人手で用意した節パタンに該当する部分を発見したらその直後にラベルを挿入するという,パタンマッチを用いた正規表現として記述されている.規則の例を以下に示す.

1. `s/(けど_接続助詞__) /$1 \/並列節ケド\/ /g;`
2. `s/(*_(動詞|助動詞|形容詞)_*_(連用形|連用形-促音便)(たり|だり)_助詞-副助詞__) /$1 <タリ節> \n/g;`

1. は，接続助詞「けど」の直後に /並列節ケド/ というラベルを挿入する規則である．2. は，「連用形」「連用形-促音便」という活用形を取っている動詞，助動詞，形容詞に，副助詞「たり」「だり」が後続した場合，その直後に <タリ節> というラベルを挿入するものである．現在，142 個の規則が準備されており，発話の分割点の候補として考えられる 33 種類の節境界を検出することが可能である．この検出ツールの出力例を，(1) に示す．

(1) 本を読んでいると <条件節ト> 想像力を働かせることができるという <トイウ節> ことです /文末/ よく自分が自分が読んだ本が映画になったり <タリ節> テレビのドラマになることがあるかと <引用節> 思うのですが /並列節ガ/ その時大体自分が想像していたものよりも...

節境界のラベルには，「絶対境界（[***]）」「強境界（/***/）」「弱境界（<***>）」という三段階のレベルを設けてある．これら三つのレベルは，節直後の切れ目の大きさという観点から区別される．絶対境界は，通常の意味での「文末」の表現に相当する．強境界は，いわゆる文末ではないけれども，発話の大きな切れ目として考えられる節境界である．接続助詞の「が」「けれども」などで導かれる従属節がこれに相当する．弱境界は，節境界ではあるけれども，通常は発話の切れ目になることはないと考えられる節境界である．「引用節」「条件節」などがこれに相当する．

南不二男 [4] は，節境界の形態の違いから従属節を複数のクラスに分類し，それらを統語的 / 意味的な自立性の度合いと関連づけた．ここで設定した三つのレベルは，南の形態的な分類を経験的な知見に基づいて修正したものである．このようなレベルを分化させておくことにより，節の種類ごとに異なる文法的な振る舞い — 主題や格要素の共有，モダリティ要素のスコープの違いなど — をあらかじめ予測し，分類しておくことができる．

この節境界検出ツールを CSJ の 338 講演（804,983 形態素）に適用し，節境界を自動的に検出した．結果を表 1に示す．

表 1: CSJ に現れた節境界

節境界	頻度	比率
絶対境界	21,693	(25.00%)
強境界	14,350	(16.53%)
弱境界	50,770	(58.48%)
合計	86,813	(100 %)

我々は，自動検出される節境界のうち，絶対境界および強境界を，発話の分割点候補として採用することにした．これら二つの節境界は，いずれも発話の大きな切れ目として考えられる境界であり，統語的 / 意味的なまとまりを備えている点で，さまざまな処理や分析にとって有用な単位の境界であると考えられるからである．また，弱境界のラベルはテキスト中に残し，後に行なわれる人手修正の候補点として用いることにした．

以上で述べた自動検出処理によって，形態素付与された CSJ から，短く分割された単位を自動的に取り出すことができる．ただし，後述する「体言止」などの特殊な節境界や，言い誤り・言い差しなど発話の過程で発生する問題については，パタンマッチによる規則では扱えないため，自動的に検出することができない．そこで，自動分割結果を人手でチェックし，あらかじめ定義した修正方針に従って，問題点を修正する作業が必要となる．次節では，我々が定義した問題点と，その修正方針について述べる．

4 人手修正

上述のように，節境界自動検出ルールは，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」や「倒置」など，話し言葉に特有な現象を適切に処理することができない．統語的にも意味的にも適切な処理単位を認定するためには，音声情報を参照しつつ自動分割結果を人手修正する必要がある．我々は，次の三種類の操作からなる修正作業を定義した．

- 二つ以上のデフォルト単位を「+」でつなぐ．
- デフォルト単位を「-」で切る．
- 要素を (***)，{***}，《***》で囲む．

作業者は所定の基準に従って自動分割結果を修正する．以下では，このうち「体言止」「主題の共有」「引用節」「挿入節」「倒置」を取り上げ，現象の定義と操作について説明する．

4.1 体言止

話し言葉では，名詞句だけで独立の節が構成される場合がある．典型的なものは講演のタイトルなどであり，こうした部分は前後の節から統語的に独立している．こうした「体言止」の要素は，末尾に述語句を持たず，ルールでは検出できないため，後続部分から切り離す必要がある．

AAA BBB*
→ AAA - # BBB* [1]

操作: もし AAA が BBB とは統語的に独立した体言止要素と判定されるならば，AAA を BBB から分離しなければならない．

(2) タイトル - ; 体言止 [2]
夢の 国 ディズニーワールド - ; 体言止
私は旅行が大好きで <並列節デ>...

[1] * はデフォルト境界 ([絶対境界] または /強境界/), # は修正後の境界位置，-, +, ( ), { }, 《 》は適用された操作を，それぞれ示す．
[2] ; の直後に操作の種類を記録する．

4.2 主題の共有

日本語では，主題要素は「は」や「も」でマークされる．主題要素は強境界を越える広いスコープを持つ傾向があるが，こうした場合には，主題を共有する二つ以上のデフォルト単位を一つの単位に結合する必要がある．

XXX は AAA / 強境界 / BBB *
→ XXX は AAA / 強境界 / + BBB *

操作: もし主題要素「XXX は（も）」が AAA だけでなく BBB にも係るならば（「XXX は（も）」が AAA と BBB の述語句に共有されているならば），AAA を BBB と結合しなければならない．

(3) 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/ ; 主題の共有

4.3 引用節

デフォルト単位が引用節として他の節に埋め込まれている場合がある．節境界検出ルールは発話の全体を解析しているわけではないため，主節の末尾に先行する埋め込み節の末尾がデフォルト境界として認定されてしまう．このような場合，埋め込み節を { } で囲み，デフォルト境界を : で置き換える．

AAA BBB* CCC DDD*
→ AAA { BBB * : CCC } DDD*

操作: AAA が DDD に係り，その間に BBB と CCC が埋め込まれている場合，BBB と CCC を { } で囲み，埋め込み節内のデフォルト境界位置を「:」でつながなければならない．

(4) もう 何度も これは テレビで 見て <テ節> { いいな /文末候補/ : いつか 行きたいな } と <引用節> もう ずっと 思っていた ところで <並列節デ> ; 引用節構造

4.4 自発的な話し言葉に特有の諸現象

自発的な話し言葉の産出の際，音韻的，語彙的，統語的（語順）問題によって，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．特に自発的な長い独話の場合ほど，話し手に課せられる線状化問題（linearization problem : 何を始めに言い，何を次に言うか）[1] は大きいものとなり，挿入節や倒置，言いさし（発話中止）などさまざまな非流暢現象が引き起こされる．こうした現象により，望ましくないデフォルト単位が生み出されてしまう場合には，人手修正が必要になる．

挿入節

自発的な話し言葉においては，話し手が発話の最中に発話プランを変更することにより，ある節の途中に別の節が挿入されることがある．これらの部分は補足，注釈，言い換えなどの機能を果たす．

AAA BBB / 強境界 / CCC*
→ AAA (BBB / 強境界 /) + CCC*

操作: BBB が挿入節であると判断され，かつ AAA から CCC への係り受け関係がある場合，BBB を ( ) で囲み，BBB の直後のデフォルト境界を，CCC と結合しなければならない．

(5) 色んな パターンを (' 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 [文末] ; 挿入節

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」が発生する場合がある．これらは補足や後からの思いつきのような，発話産出上の理由によるものである．

AAA BBB* CCC DDD*
→ AAA BBB * + 《CCC 》 - # DDD*

操作: 文節 CCC が BBB 内の述語句に係る倒置要素であると判断された場合，BBB の直後のデフォルト境界を CCC の直後に移動し，さらに CCC を 《 》 で囲むとともに DDD から切り離さなければならない．

(6) 家に いる 時間に 必ず 掛かってくるんですね [文末] + 《一日に 三四回 》 - ; 倒置
で 何か 今 よく テレビで ストーカーの 話とか やってるけれども /並列節ケレドモ/...

言いさし — 発話中止

話し手が節の途中で事前の発話プランを変更することにより，後続部分を発話することをやめる場合がある．このように破棄された部分は前後の単位から取り残されるため，係り受け関係の付与などにとって有用な単位とはならない．

AAA BBB*
→ AAA - # BBB*

操作: 要素 AAA が発話中止によって生じた節の断片であり，有意味な単位を構成しないと判断された場合には，AAA を BBB から切り離さなければならない．

(7) 今回の 実験で - ; 言いさし
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]

5 議論

5.1 人手修正結果

前節で述べた基準に基づき，独話 43 講演（学会講演 15 講演，模擬講演 28 講演）について人手修正を行い，我々の定義する「文」単位を抽出した．自動分割結果に対して，各講演とも二人以上の作業者が作業を行い，作業結果を比較・統合した．結果を表 2に示す．

表 2: 人手修正結果

	デフォルト	修正後	人手修正率	全操作数	+	−	( )	{ }	《 》
学会講演	102.1	94.4	16.9%	17.2	10.8	4.8	3.5	1.7	0.1
模擬講演	89.6	84.5	24.4%	21.9	12.4	8.5	4.4	1.5	0.5
平均	94.0	88.0	21.6%	20.3	11.9	7.2	4.1	1.6	0.4

表 2の「デフォルト」「修正後」はそれぞれ自動分割結果と人手修正結果における 1 講演あたりの平均単位数を示す．「人手修正率」はデフォルト単位数に対する全操作数[3]の割合である．「全操作数」は，人手操作の総数である．各操作の平均回数も併せて示す．

注目すべき点は，学会講演 (16.9%) と模擬講演 (24.4%) という人手修正率の違いである．これは両講演の自発性の違いに基づくものであると考えられる．学会講演は多くの場合事前原稿に基づいて話されることが多いのに対して，模擬講演はあまり形式ばらずに話された個人的な内容についての語りである．このため，話し言葉に特有な非流暢性によって，模擬講演の方がより多くの人手修正が必要になっているものと思われる．この傾向は，模擬講演にはより多くの音韻的・形態的な非流暢要素が含まれるという指摘 [2] ともほぼ合致する．

5.2 人手修正支援ツール

対象 43 講演の人手修正率の平均は 21.6% であった．この結果は，第 4 節で述べた各現象が考慮に値すべきものであると考えるならば，人手修正は決して単なる例外的な作業とはいえないということを示すものである．人手修正作業を支援するツールを導入することによって，こうした人手作業では不可避な操作誤りを減少させるとともに，作業者への負荷を軽減することが必要となる．

そこで我々は，人手修正のためのアノテーションツールを用意した（図 1）．このツールは記号の使用法及び記号と操作の種類を示すコメントとの間の対応関係を限定するものである．また，このツールによって，複数作業者による作業結果の比較やデータ管理が効率的に行なえる．

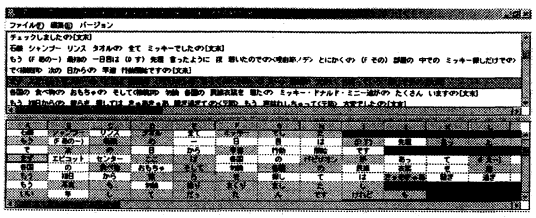

図 1: 人手修正ツール

[3]( ) の直後には「+」が伴うなどの記号間の規則的な共起を考慮した．

6 まとめ

自発的な話し言葉における有用な処理単位を特定することは，不可欠だが困難な課題である．本発表では，『日本語話し言葉コーパス CSJ』において「文」に相当する単位を半自動的に認定する方法を提案した．本プロジェクトでは，CSJ のうちの約 50 万語（約 50 時間）分についてこうした「文」認定作業を行なう予定である．「文」認定が施された CSJ コーパスは自動要約，係り受け解析，談話構造分析などのさまざまな話し言葉処理技術の発展に貢献するものである．

参考文献

[1] Levelt, W.J.M. (1989) *Speaking: From Intention to Articulation.* The MIT Press.

[2] Maekawa, K. (forthcoming) Corpus of Spontaneous Japanese: its design and evaluation. In *Proceedings of The ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR 2003).*

[3] 丸山岳彦・柏岡秀紀・熊野正・田中英輝. (2003) “節境界自動検出ルールの作成と評価.” 『言語処理学会 第 9 回年次大会 発表論文集』. 言語処理学会.

[4] 南不二男. (1974) 『現代日本語の構造』. 大修館書店.

[5] 森本郁代・高梨克也・竹内和広・小磯花絵・井佐原均. (2003) “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』. 言語処理学会.

[6] 内元清貴・野畑周・山田篤・関根聡・井佐原均. (2003) “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』. 言語処理学会.