# 節境界に基づく独話文の係り受け解析とその評価

大野 誠寛† 松原 茂樹†‡§ 丸山 岳彦§ 柏岡 秀紀§ 田中 英輝♮ 稲垣 康善♮

†名古屋大学大学院情報科学研究科 ‡名古屋大学情報連携基盤センター

§ATR 音声言語コミュニケーション研究所 ♮NHK 放送技術研究所 ♮愛知県立大学情報科学部

ohno@inagaki.nuie.nagoya-u.ac.jp

## 1 はじめに

話し言葉は，一人の話者のみが話す「独話」と複数の話者が交替で話す「対話」に分類できる．これまでの話し言葉解析の研究は，対話文を対象としたものがほとんどであり，非文法性に対して頑健に対処する手法が提案されてきた (例えば，[6, 8])．しかしその一方で，独話文を対象とした研究はほとんどないのが現状である．

独話文は，対話文に比べ，一文の長さが長く文の構造が複雑であるといった特徴をもつ．そのような文に対して解析を実行すると，一般に，解析時間が長くなるうえ，高い解析精度の達成が難しくなる．高い性能を備えた独話文解析を実現するために，適切なユニットに文を分割し，簡単化することが効果的な方法である．

そこで本稿では，文分割に基づく独話文の係り受け解析手法を提案する．本手法では，節レベルと文レベルの二段階で係り受け解析を実行する．まず，節境界解析により文を節に分割し，各節に対して係り受け解析を行うことにより，節内の係り受け関係を同定する．次に，節境界をまたぐ係り受け関係を定め，文全体の係り受け構造を作り上げる．独話文係り受け解析実験の結果，本手法により解析精度が低下することなく解析時間を大幅に短縮できることを確認した．

## 2 節境界と係り受け構造

節とは，述語を中心としたまとまりであり，複文や重文の場合，文は複数の節から構成される．本研究で提案する手法では，「文は一つ以上の節の連接であり，各節を構成する文節は，節の最終の文節を除き，その節内の文節に係る」とみなす．例として，「警察法という法律には警察を管理するという曖昧な言葉でしか公安委員会の役割を示していません」の係り受け構造を図 1 に示す．この文は 3 つの節「警察法という法律には」,「警察を管理するという」,「曖昧な言葉でしか公安委員会の役割を示していません」から構成され[1]，各節が係り受け構造を形成し，それらが節の最終文節からの係り受け関係でつながっている．このような仮定を設ける理由は以下の通りである．

- 一文が長い独話文では，文を短く分割することにより係り受け関係の探索範囲が狭められ，解析時間を短縮できる．
- 節は述語を中心としたまとまりであるため，その内部で係り受けがまとまりやすい．そのため，上述の仮定に逸脱する係り受けは少なく，解析精度の低下への影響は小さい．実際，独話文に茶筌 [7], CaboCha[2] で自動的に付与した係り受け構造を分析した結果，約 90 %の節が上述の性質を満たすという結果が報告されている [1].

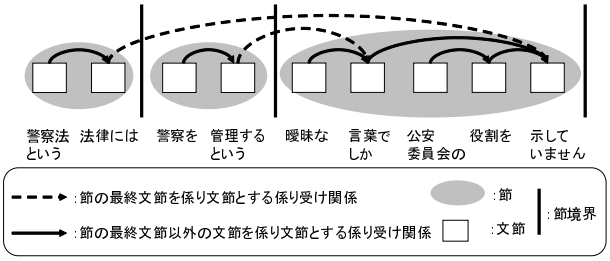


図 1: 節境界と係り受けの関係

なお，独話文の節への分割は，節境界解析により実現できる [5]．この節境界解析装置は，局所的な形態素解析列のみを手がかりとして，節境界の位置と種類を特定することができる．日本語形態素解析ツール茶筌 [7] で形態素解析した文を入力すると，入力文中に含まれるすべての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．挿入される節境界ラベルは /テ節/ や /並列節ケレドモ/ など，合計 144 種類である．

## 3 節境界に基づく係り受け解析手法

本手法では，前節で設けた仮定に基づき，形態素解析，節境界解析及び文節まとめ上げが施された文を入力とする．以下の手順で解析を実行する．

1. 一文中のすべての節に対して，各節ごとにその内部の係り受け構造を解析する．
2. 節の最終文節を係り文節とする係り受け構造を解析する．

なお，以下では，一文中の節列を $C_1, ..., C_m$，節 $C_i$ 中の文節列を $b^i_1, ..., b^i_n$，文節 $b^i_k$ を係り文節とす

[1]「主題ハ」は「述語を中心としたまとまり」という節の定義から逸脱しているが，統語的に大きな切れ目になると考えられるため，節境界と見なしている [5]．5.3.1 節参照．

る係り受け関係を $dep(b_k^i)$，一文の係り受け構造を $\{dep(b_1^1), ..., dep(b_{n_m-1}^m)\}$ と記す．

本手法では，まず，節 $C_i$ が入力されるごとに，節内部の係り受け構造 $\{dep(b_1^i), ..., dep(b_{n_i-1}^i)\}$ を求める．その後，文末まで，節の最終文節の係り受け構造 $\{dep(b_{n_1}^1), ..., dep(b_{n_{m-1}}^{m-1})\}$ を求める．なお，いずれの解析においても，係り受けは非交差性，後方修飾性，係り先の唯一性を満たすものとする．

### 3.1 節内部の係り受け解析

節内部の係り受け解析は，入力節 $C_i$ 中の文節列 $b_1^i, ..., b_{n_i}^i$ を $B_i$ とし，$P(S_i|B_i)$ を最大にする係り受け構造 $S_i(=\{dep(b_1^i), ..., dep(b_{n_i-1}^i)\})$ を求める．なお，節内部の係り受け解析では，節の最終文節 $b_{n_i}^i$ の受け文節は決定しない．

それぞれの係り受け関係は独立であると仮定すると，$P(S_i|B_i)$ は以下の式で計算できる．

$$P(S_i|B_i) = \prod_{k=1}^{n_i-1} P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i) \quad (1)$$

ここで，$P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ は，入力文節列 $B_i$ が与えられたときに，文節 $b_k^i$ が $b_l^i$ に係る確率を表す．最尤の係り受け構造は，式 (1) の確率を最大とする構造であるとして動的計画法を用いて計算する．

次に，$P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ の計算について述べる．まず，係り文節における自立語の原形を $h_k^i$，その品詞を $t_k^i$，係りの種類を $r_k^i$ とし，受け文節における自立語の原形を $h_l^i$，その品詞を $t_l^i$ とする．また，文節間距離を $d_{kl}^{ii}$ とする．ここで，係りの種類とは，係り文節が付属語を伴うときはその付属語の語彙，品詞，活用形であり，そうでない場合は一番最後の形態素の品詞，活用形である．

以上の属性を用いて，確率 $P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ を以下のように計算する．

$$\begin{aligned} & P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i) \quad (2) \\ \cong\ & P(b_k^i \overset{rel}{\rightarrow} b_l^i|h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii}) \\ =\ & \frac{F(b_k^i \overset{rel}{\rightarrow} b_l^i, h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii})}{F(h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii})} \end{aligned}$$

ただし，$F$ は共起頻度関数である．

### 3.2 節の最終文節の係り受け解析

節の最終文節の受け文節を同定する．一文の文節列を $B(=B_1, ..., B_m)$ とし，節の最終文節を係り文節とするような係り受け構造 $\{dep(b_{n_1}^1), ..., dep(b_{n_{m-1}}^{m-1})\}$ を $S_{last}$ とするとき，$P(S_{last}|B)$ を最大とする $S_{last}$ を求める．$P(S_{last}|B)$ は以下の式で計算できる．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B) \quad (3)$$

ここで，$P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B)$ は，一文の文節列 $B$ が与えられたときに，文節 $b_{n_i}^i$ が $b_l^j$ に係る確率を表す．最尤の係り受け構造は，式 (3) の確率を最大とする構造であるとして動的計画法を用いて計算する．本手法では，先に解析した節内部の係り受け構造を前提として決定する．すなわち，後方に位置するすべての文節を受け文節の候補として計算するのではなく，節内部の係り受け構造から非交差性を満たすものだけを受け文節の候補として計算する．例の場合，文節「管理するという」の受け文節は「曖昧な」,「言葉でしか」,「示していません」のいずれかであるとして計算する．

なお，$P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B)$ は，式 (2) と同様に計算する．

表 1: 実験で使用したデータ

| | あすを読む (テストデータ) | 京大コーパス (学習データ) |
|---|---|---|
| 文数 | 200 | 7,758 |
| 節数 | 951 | 27,060 |
| 文節数 | 2,430 | 72,393 |
| 形態素数 | 6,017 | 205,731 |

表 2: 実験結果（解析時間）

| | 本手法 | 従来手法 |
|---|---|---|
| 平均解析時間 (秒/文) | 0.012 | 0.055 |

注) 実装言語:LISP, 使用計算機:Pentium 4 2.4GHz, Linux

## 4 解析実験

独話文の係り受け解析における本手法の有効性を評価するため，解析実験を行った．

### 4.1 実験の概要

実験で使用したデータを表 1 に示す．NHK の解説番組「あすを読む」の書き起こしデータに形態素解析，節境界解析，文節まとめ上げを施した 200 文をテストデータとして用いた．正解の係り受けは人手で付与した．係り受け文法は京大コーパス [3] に準拠した．ここで，本手法では，節の最終文節を係り文節とする係り受け関係を除いて，係り受け関係は節境界をまたがないことを前提としているが，この前提を満たさない係り受け関係は，テストデータの正解中に 94 個存在した．これは，本手法による係り受け正解率 (文末を除く) が 95.8%(2,136/2,230) を超えることはないことを意味する．

一方,「あすを読む」に対する十分な量の係り受けデータは存在しないため，新聞記事ではあるが京大コーパス 7,758 文を学習データとして用いた．

なお，節に分割することなく文の係り受け構造を一度に求める手法（以下，従来手法）によっても係り受け解析を行い，本手法と比較した．

### 4.2 実験結果

両手法の解析時間を表 2 に示す．本手法の解析速度は従来手法に比べて，平均して約 5 倍向上した．文の長さと解析時間の関係を図 2 に示す．従来手法では文の長さが 12 文節を超えたあたりから，急激に解析時間が上昇

表 3: 実験結果（係り受け正解率）

| | 本手法 | 従来手法 |
|---|---|---|
| 節の内部 | 83.6%(1,236/1,479) | 82.6%(1,222/1,479) |
| 節の最終文節 | 58.1%( 436/ 751) | 55.0%( 413/ 751) |
| 合計 | 75.0%(1,672/2,230) | 73.3%(1,635/2,230) |

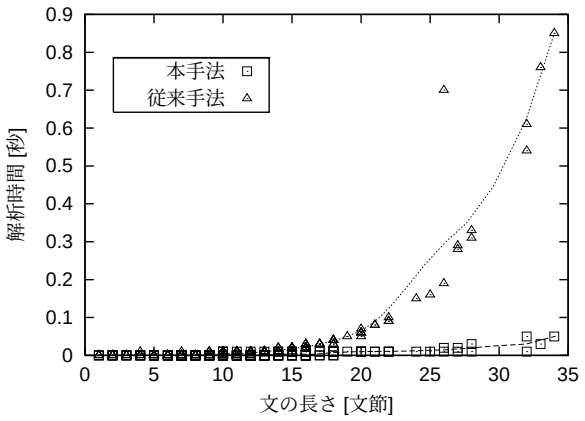

図 2: 文の長さと解析時間の関係

するのに対し，本手法の解析時間はあまり変化していない．実験で使用した 200 文の平均文節数は 12.2 であり，平均以上の長さをもつ独話文に対する本手法の効果を確認した．

次に，両手法の係り受け正解率を表 3 に示す．表 3 の第 1 行は，節の最終文節を除く節内の全ての文節に対する正解率を，第 2 行は，文末を除く全ての節の最終文節に対する正解率を示す．節の内部，最終文節とも，従来手法に劣らない解析精度を備えていることがわかる．

以上の結果から，本手法によって，解析精度を従来手法と同程度に維持したまま解析時間を短縮できることを確認した．

# 5 考察

本手法は，節内部の係り受け関係は節境界をまたがないとして係り受け解析を行っているため，正解データにおいて節内部にあり節境界をまたぐ係り受け関係は同定できない．しかし，実験では，節の内部，節の最終文節のいずれにおいても本手法の精度が従来手法に劣ることはなかった．

以下では，まず，本手法の解析精度に対する効果を，節の内部と節の最終文節の二つに分けて考察する．その後，本手法の問題点である，節内部の節境界をまたぐ係り受け関係の解析について検討する．

## 5.1 節内部の解析に対する考察

表 4 は，節の内部の係り受け関係の解析結果における両手法の正解，不正解の関係を示したものである．節の内部の係り受け関係 1,479 個のうち，両手法においてともに正しく解析された係り受け関係は 1,204 個であった．本手法で正解し，従来手法で不正解となったものは

表 4: 節の内部の係り受け解析結果の詳細

| | | 本手法 | | 合計 |
|---|---|---|---|---|
| | | 正解 | 不正解 | |
| 従来手法 | 正解 | 1,204 | 18 | 1,222 |
| | 不正解 | 32 | 225 | 257 |
| 合計 | | 1,236 | 243 | 1,479 |

表 5: 節の最終文節の係り受け解析結果の詳細

| | | 本手法 | | 合計 |
|---|---|---|---|---|
| | | 正解 | 不正解 | |
| 従来手法 | 正解 | 404 | 9 | 413 |
| | 不正解 | 32 | 306 | 338 |
| 合計 | | 436 | 315 | 751 |

32 個にのぼる．これは，本手法が受け文節の候補を節の内部に絞った効果を示している．

一方，従来手法のみで正しく解析できた係り受け関係は 18 個であった．このうち 14 個は節境界をまたぐ係り受け関係であり，本手法でそもそも同定できないものである．これは，節境界をまたぐ係り受け関係を除けば，従来手法で正しく解析される係り受け関係のほとんどを本手法によって正しく解析できることを意味する．

## 5.2 節の最終文節の解析に対する考察

表 3 が示すように，節の最終文節の係り受け正解率は，節の内部のものと比べて両手法ともかなり低い．これは，節の最終文節を係り文節とする係り受け関係の同定が難しいことを意味している．

表 5 に，節の最終文節 (文末を除く) の解析結果における両手法の正解，不正解の関係を示す．節の最終文節を係り文節とする係り受け関係 751 個のうち，両手法がともに正しく解析した係り受け関係は 404 個あった．本手法でのみ正解した係り受け関係は 32 個で，従来手法でのみ正しく解析できた係り受け関係 9 個を大幅に上回った．これは，本手法が，先に解析した節内部の係り受け解析結果を前提とすることにより，節の最終文節の受け文節となる候補を効果的に絞った結果であると考えられる．

## 5.3 節境界をまたぐ係り受け関係

表 6 に，節境界をまたぐ係り受け関係に対する両手法の正解率を示す．本手法は，このような係り受け関係は存在しないとして解析を行っているため，一つも正しく解析できない．一方，従来手法は，テストデータに存在する節境界をまたぐ係り受け関係 94 個のうち 14 個を正しく解析した．従来手法においてもその正解率は 14.9%にとどまっており，このような係り受け関係の同定はそもそも困難であることがわかる．

文献 [1] で議論されているように，解析精度の向上のために，節をまたぐ係り受け関係を考慮した処理が望まれる．本稿では，人手で正解を付与したテストデータ 200 文を用いてその解析可能性について検討する．図 3 に，節境界をまたぐ係り受け関係が存在した節の種類

表 6: 節境界をまたぐ係り受け関係の解析結果

| | 本手法 | 従来手法 |
|---|---|---|
| 正解率 | 0%( 0/94) | 14.9%(14/94) |

(節を区切る節境界のラベル名) とその割合を示す．“主題ハ” が最も多く，次いで，“連体節”，“テ節” の順であった．以下では，全体の 73.4%を占めるこの上位３つの節についてそれぞれ述べる．

#### 5.3.1 節の種類 “主題ハ”

節境界をまたぐ係り受け関係 94 個のうち，28 個は，節 “主題ハ” にその係り文節が存在した．節 “主題ハ” は「述語を中心としたまとまり」という節の定義に逸脱しているが，統語的に大きな切れ目になると考え [5]，本研究ではこれについても節としている．

このような節を調べてみると，節 “主題ハ” 内に述語が存在しないために，述語に係るような文節は節外に位置する述語に係る現象が多く見られた．この場合，述語に係る文節については節外に係り先があるとみなし，そのようなルールを作成し検出することが考えられる．

例）**七年前アラファト議長と当時のラビン首相は/主題ハ/歴史的な和解に踏み切りました**
「七年前」(副詞句) が「踏み切りました」(述語) に係るため，節境界 “主題ハ” をまたいでいる．

#### 5.3.2 節の種類 “連体節”

節境界をまたぐ係り受け関係のうち 27 個は，節 “連体節” にその係り文節が存在した．これらを調べてみると，節内部の文節がこの連体節が修飾する文節と並列関係や同格関係になっている現象が多く見られた．一般に，係り受け解析において，並列関係や同格関係の同定は難しいが，これらを検出する手法が報告されており (例えば，[4])，本手法への導入が考えられる．

例）**渡来系の人々と縄文人の系統をひく/連体節/人々の骨が出てきています．**
「(渡来系の) 人々と」と,「人々の」が並列し係り受け関係にあり，節境界 “連体節” をまたいでいる．

#### 5.3.3 節の種類 “テ節”

節境界をまたぐ係り受け関係のうち 14 個は，節 “テ節” にその係り文節が存在した．これらの中で多く見られたのは，文全体の係り受け構造としては，節境界をまたぐ係り受け関係になるが，節内部にも意味的には受けとなる文節が存在する現象である．この場合,「係り先は唯一である」という制約を緩めて柔軟に評価する，すなわち，節内部にある受け文節についても正解とすることが考えられる．

例）**小泉さんが二倍近くの差をつけて/テ節/圧勝した**
「小泉さんが」が文全体での係り受け構造としては「圧勝した」に係るが，同一節内の「つけて」の主語は「小泉さんが」であり，そのような係り受け関係が必ずしも誤りであるというわけではない．

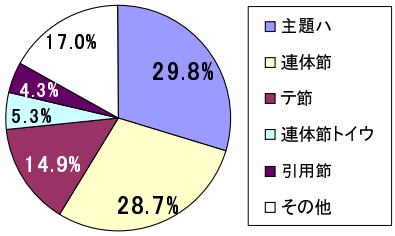


図 3: 係り受けがまたぐ節境界の種類とその割合

## 6 おわりに

本稿では，節境界に基づく独話文の係り受け解析手法を提案した．本手法の有効性を評価するために，独話文を用いて係り受け解析実験を行った．実験の結果，本手法の有効性を確認した．今後は，節境界をまたぐ係り受け関係を検出し，その係り先を同定する手法について検討したい．

**謝辞** 本研究は通信・放送機構の研究委託「大規模コーパスベース音声対話翻訳技術の研究開発」により実施したものである．

## 参考文献

[1] 柏岡秀紀，丸山岳彦，田中英輝: 節境界と係り受け解析，言語処理学会第 9 回年次大会論文集，pp.117-120 (2003).

[2] 工藤 拓，松本 裕治: チャンキングの段階適用による係り受け解析，情報処理学会論文誌，Vol.43，No.6，pp.1834-1842 (2002).

[3] 黒橋 禎夫，長尾 真: 京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会発表論文集，pp.115-118 (1997).

[4] 黒橋 禎夫，長尾 真: 長い日本語文における並列構造の推定，情報処理学会論文誌，Vol.33，No.8，pp.1022-1031 (1992).

[5] 丸山岳彦，柏岡秀紀，熊野正，田中英輝: 節境界自動検出ルールの作成と評価，言語処理学会第 9 回年次大会論文集，pp.517-520 (2003).

[6] Matsubara, S., Murase, T., Kawaguchi, N. and Inagaki, Y.: Stochastic Dependency Parsing of Spontaneous Japanese Spoken Language, *Proc. of 19th COLING*, Vol.1, pp.640-645 (2002).

[7] 松本 裕治，北内 啓，山下 達雄，平野 善隆，松田 寛，高岡 一馬，浅原 正幸: 形態素解析システム『茶筌』，version2.2.9，使用説明書 (2002).

[8] 大野誠寛，松原 茂樹，河口 信夫，稲垣 康善：日本語音声対話文の統計的係り受け解析とその評価，情報処理学会第 65 回全国大会講演論文集，Vol.2，pp.1-2 (2003).