節境界に基づく独話文係り受け解析の効率化

大野 誠寛[1]　松原 茂樹[2,4]　丸山 岳彦[3]
柏岡 秀紀[4]　田中 英輝[5]　稲垣 康善[6]

[1]名古屋大学大学院情報科学研究科　[2]名古屋大学情報連携基盤センター
[3]国立国語研究所　[4]ATR 音声言語コミュニケーション研究所
[5]NHK 放送技術研究所　[6]愛知県立大学情報科学部
E-mail: ohno@el.itc.nagoya-u.ac.jp

概要

独話文は，対話文に比べ，１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．高い性能を備えた独話文解析を実現するためには，適切な単位に文を分割し，簡単化することが効果的な方法である．そこで本論文では，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により節境界解析に基づく本手法の有効性を確認した．

キーワード　係り受け解析, 独話, 節境界解析

Efficient Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries

Tomohiro Ohno[1]　Shigeki Matsubara[2,4]　Takehiko Maruyama[3]
Hideki Kashioka[4]　Hideki Tanaka[5]　Yasuyoshi Inagaki[6]

[1]Graduate School of Information Science, Nagoya University
[2]Information Technology Center, Nagoya University
[3]The National Institute for Japanese Language
[4]ATR Spoken Language Translation Research Laboratories
[5]NHK Science & Technical Research Laboratories
[6]Faculty of Information Science and Technology, Aichi Prefectural University
E-mail: ohno@el.itc.nagoya-u.ac.jp

Abstract

Generally speaking, spoken monologue sentences are longer and more complex than spoken dialogue sentences. To achieve high parsing performance for spoken monologue, it could prove effective to simplify the structure by dividing a sentence into suitable language units. This paper proposes a method for dependency parsing of Japanese monologue based on sentence segmentation. In this method, dependency parsing is executed in two stages: a clause level and a sentence level. An experiment using a spoken monologue corpus shows the method to be effective for efficient dependency parsing of Japanese monologue sentences.

key words　dependency parsing, monologue, clause boundary

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，[10, 13])．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

な単位に文を分割し，簡単化することが効果的な方法である．

そこで本論文では，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．

次節で独話文の解析単位について述べ，3 節，及び，4 節で節境界解析手法，係り受け解析手法をそれぞれ示す．5 節で本手法に対する解析実験について説明し，6 節で実験結果に基づいて考察する．

2　独話文の解析単位

本研究では，文よりも短い単位を解析単位とすることにより，解析を効率化する．一文が長い独話文では，係り受け関係の探索範囲が狭められ，解析時間を短縮することができる．

2.1　節と係り受け

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．さらに，節は，統語的，意味的にまとまった単位であるため，文に代わる解析単位として利用できる．

そこで本研究では,「文は一つ以上の節の連接であり，各節を構成する文節は，節の最終文節を除き，その節の内部の文節に係る」とみなし，それに基づく係り受け解析手法を提案する．

例として，独話文「先日総理府が発表いたしました世論調査によりますと死刑を支持するという人が八十パーセント近くになっております」の係り受け構造を図 1 に示す．この文は 4 つの節「先日総理府が発表いたしました」,「世論調査によりますと」,「死刑を支持するという」,「人が八十パーセント近くになっております」から構成され，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 [4].

2.2　節境界単位

節を文に代わる解析単位とするためには，係り受け解析の前処理として独話文を節に分割する必要がある．しかし，節には，主節の中に埋め込まれた従属節も存在するため，本来，文を節に一次元的に分割することは困難である．

ただし，節への分割は，節境界解析により近似的に実現できる [9]．すなわち，節境界解析では，節の終端位置を検出することにより節に相当する単位を検出する．節境界の検出では，局所的な形態素列のみを手がかりとして，節の終端位置と種類を特定し，144 種の節ラベルを付与する．

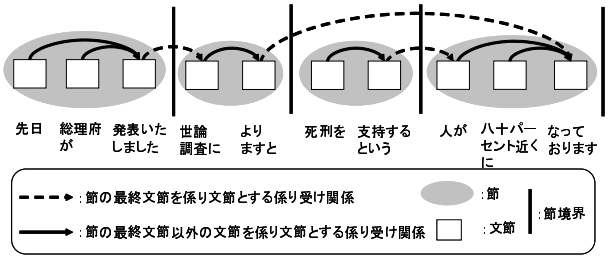

図 1: 節境界と係り受けの関係

表 1: 「あすを読む」200 文の基礎統計

項目	数値
文数	200
節境界単位数	951
文節数	2,430
形態素数	6,017
節境界をまたぐ係り受け数	94

本研究では，節境界解析により検出された節境界ではさまれた単位を**節境界単位**と呼び [3]，これを新たな解析単位と考える．なお，節境界単位の終端位置に付与されたラベル名をその節境界単位の種類とする．

2.3　節境界単位と係り受け構造の関係

節境界単位と係り受け構造の関係を明らかにするために独話文コーパスを用いて分析した．分析には，NHK の解説番組「あすを読む」の書き起こしデータ 200 文に対して形態素解析，文節まとめ上げ，節境界解析，係り受け解析を自動的に行い，人手で修正したものを用いた．なお，形態素解析は ChaSen[11] の IPA 品詞体系 [1] に，文節まとめ上げは CSJ 作成基準 [7] に，節境界解析は丸山らの基準 [9] に，係り受け文法は京大コーパスの作成基準 [6] にそれぞれ準拠している．

200 文の基礎統計を表 1 に示す．総文節数 2,430 文節のうち，節境界単位の最終文節 (951 文節) を除いた 1,479 文節の中で，94 文節のみが節境界単位の外に位置する文節に係っていた．これは，全体の 93.6%(1,385/1,479) の係り受けが節境界単位で閉じていることを意味しており，本研究で設定した仮定がある程度妥当なものであることを確認した．

一方でこれは，節境界単位内部の係り受けのうち，6.4%が節境界単位で閉じていないことも同時に意味しており，必ずしも無視できる現象ではない．そこで，この 94 個の係り受けについてさらに詳しく分析

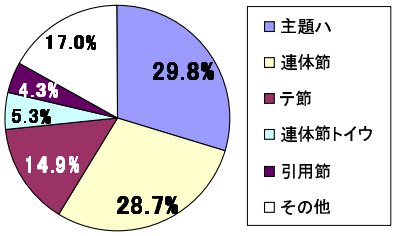

図 2: 係り受けがまたぐ節境界の種類とその割合

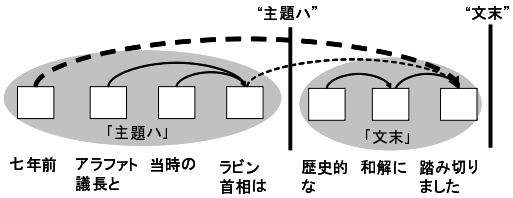

図 3: 節境界 “主題ハ” をまたぐ係り受け関係の例

した.

図 2 に，節境界をまたぐ係り受け関係が存在した節境界単位の種類（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．これらの多くは，節境界単位が純粋な意味で節ではないために生じている．以下では，28 個と最も多く存在した節境界単位「主題ハ」において閉じていない係り受けに着目して論じる．

節境界単位「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 [9]，本研究ではこれについても節境界として検出している．

節境界単位「主題ハ」では，述語が存在しないため，その中に位置する述語に係る文節が直後の節境界単位内の述語に係る現象が多く見られた．図 3 に，独話文「七年前アラファト議長と当時のラビン首相は歴史的な和解に踏み切りました」の節境界と係り受け構造を示す.「七年前」が「踏み切りました」に係り，節境界 “主題ハ” をまたいでいる．このような場合，例えば，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．

3　節境界単位への分割

節境界解析により独話文を節境界単位に分割する．節境界解析では，形態素解析及び文節まとめ上げが施された文を入力とする．解析は以下の手順で実行する．

1. 節境界解析ツールによる解析
   節境界解析ツールを用いて入力文に対して節境界を付与する．

2. 節境界単位の修正
   節境界をまたいで係り先をもつ文節を検出し，係り受けが閉じるように節境界単位を修正する．

3.1　節境界解析

節境界解析ツール CBAP[9] を用いて，入力文の文節列に節境界を付与し，一文中の節境界単位をすべて同定する．ただし，CBAP は入力を形態素列とし，形態素列パタンのみから形態素の切れ目に節境界を付与するため，文節の切れ目でない箇所に節境界が付与される場合がある．この場合，節境界によって文節が分割されることを防ぐため，文節内部で検出された節境界は無視し，節境界と文節境界が一致した場合のみ，それを節境界と認めた．

3.2　節境界単位の修正

本節では，節境界をまたぐ係り受け関係も解析可能にするため，そのような文節を検出し，係り受けが閉じるように節境界単位を修正する手法について説明する．ただし，本論文では，2.3 節での分析に基づき，節境界単位「主題ハ」で閉じていない係り受けに着目する．それ以外の種類については今後さらなる検討が必要である．

まず，節境界単位「主題ハ」で閉じていない係り受け関係を検出する．次に，検出した係り文節をこれより前に位置する文節列とともに直後の節境界単位に移動し，節境界単位を修正する．

3.2.1　節境界をまたぐ係り受け関係の検出

節境界単位「主題ハ」で閉じていない係り受け関係の大半は，内部に述語となるような文節が存在しないために，述語に係るような文節が単位外に位置する述語に係ることにより生じる．したがって，これらを検出するには，節境界単位「主題ハ」の中で述語に係るような文節を検出すればよい．

そこで，文節が述語に係るか否かを判定するために，文献 [1, 8] を参考にして，述語に係る文節の最終形態素の品詞を定めた．その一覧を表 2 に示す[1]．表 2 の品詞と，文節の最終形態素の品詞が一致するとき，その文節は述語に係る文節であるとして判定する．

3.2.2　移動手順

検出した，節境界をまたぐ係り受けの係り文節より左側の文節列を，節境界単位「主題ハ」から取り除き，直後の節境界単位の先頭に連結する．独話文

[1] この他の品詞として，感動詞や接続詞などが考えられるが，これらは CBAP により別の節境界単位になるので，ここには入れていない．

表 2: 述語に係る文節の最終形態素の品詞

品詞	品詞細分類
助詞	格助詞-一般
	格助詞-引用
	格助詞-連語
	係助詞
	副助詞
	副詞化
名詞	副詞可能
	非自立-副詞可能
	非自立-助動詞語幹
	接尾-副詞可能
	接尾-助数詞
副詞	一般
	助詞類接続

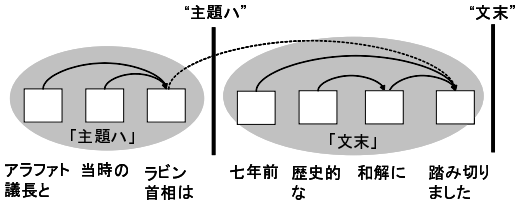

図 4: 節境界単位を修正した係り受け構造

を構成する節境界単位列 $C_1, ..., C_m$ において，節境界単位 C_x が「主題ハ」であり，文節列 $b_1, ..., b_n$ で構成されているとする．ここで，b_y が述語に係る文節のうちの，最右の出現であるとすると，$b_1, ..., b_y$ を C_x から取り除き，これを次の節境界単位 C_{x+1} の先頭に連結する．

3.2.3 修正の例

節境界単位を修正する様子を図 3 の文を例に説明する．節境界単位「主題ハ」を構成する文節のうち，最終文節「ラビン首相は」を除く文節について，述語に係るか否かを判定する．この場合，文節「七年前」が述語に係る最右の文節としてが検出される．次に，この文節より左側の文節列「七年前」を，節境界単位「主題ハ」から取り除き，直後の節境界単位「文末」の先頭に連結する．この節境界単位を修正した結果を図 4 に示す.

4 節境界に基づく係り受け解析

本手法では，2.1 節で設けた仮定に基づき，形態素解析，文節まとめ上げ，及び節境界解析が施された文を入力とする．解析の手順は以下の通りである．

1. 節レベルの係り受け解析
   一文中のすべての節境界単位に対して，各節境界単位ごとにその内部の係り受け構造を解析する．
2. 文レベルの係り受け解析
   一文中のすべての節境界単位に対して，その最終文節の係り先を解析する．

なお，以下では，一文を構成する節境界単位列を $C_1, ..., C_m$，節境界単位 C_i を構成する文節列を $b^i_1, ..., b^i_{n_i}$，文節 b^i_k を係り文節とする係り受け関係を $dep(b^i_k)$，一文の係り受け構造を $\{dep(b^1_1), ..., dep(b^m_{n_m-1})\}$ と記す.

本手法では，まず，あらゆる節境界単位 C_i に対して，節境界単位内の係り受け構造 $\{dep(b^i_1), ..., dep(b^i_{n_i-1})\}$ を求める．その次に，節境界単位の最終文節の係り受け構造 $\{dep(b^1_{n_1}), ..., dep(b^{m-1}_{n_{m-1}})\}$ を求める．なお，いずれの解析においても，係り受けの非交差性，後方修飾性，係り先の唯一性を満たすものとする．

4.1 節レベルの係り受け解析

節レベルの係り受け解析は，節境界単位 C_i 中の文節列 $b^i_1, ..., b^i_{n_i}$ を B_i とするとき，$P(S_i|B_i)$ を最大にする係り受け構造 $S_i(= \{dep(b^i_1), ..., dep(b^i_{n_i-1})\})$ を求める．なお，節境界単位の係り受け解析では，節境界単位の最終文節 $b^i_{n_i}$ 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．

$$P(S_i|B_i) = \prod_{k=1}^{n_i-1} P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i) \quad (1)$$

ここで，$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ は，入力文節列 B_i が与えられたときに，文節 b^i_k が b^i_l 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，$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ 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^i_k，その品詞を t^i_k，係りの種類を r^i_k とし，受け文節における自立語の原形を h^i_l，その品詞を t^i_l とする．また，受け文節が節境界単位の最終文節であるか否かを e^i_l とし，文節間距離を d^{ii}_{kl} 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．これらの属性は，従来の係り受け解析手法 [2, 5, 15] で用いられてきたものと同様である．以上の属性を用いて，確率 $P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ を以下のように計算する．

$$\begin{aligned} & P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i) \quad (2) \\ \cong\ & P(b^i_k \overset{rel}{\rightarrow} b^i_l|h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl}) \\ =\ & \frac{F(b^i_k \overset{rel}{\rightarrow} b^i_l, h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl})}{F(h^i_k, h^i_l, t^i_k, t^i_l, r^i_k, e^i_l, d^{ii}_{kl})} \end{aligned}$$

ただし，F は共起頻度関数である．

表 3: 実験で使用したデータ (あすを読む)

	テストデータ	学習データ
文数	500	5,532
節数	2,237	26,318
文節数	5,298	65,762
形態素数	13,342	165,173

4.2 文レベルの係り受け解析

節境界単位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，節境界単位の最終文節を係り文節とするような係り受け構造 $\{dep(b^1_{n_1}), ..., dep(b^{m-1}_{n_{m-1}})\}$ 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．

$$P(S_{last}|B) = \prod_{i=1}^{m-1} P(b^i_{n_i} \stackrel{rel}{\rightarrow} b^j_l|B) \qquad (3)$$

ここで，$P(b^i_{n_i} \stackrel{rel}{\rightarrow} b^j_l|B)$ は，一文の文節列 B が与えられたときに，C_i の最終文節 $b^i_{n_i}$ が b^j_l 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，先に解析した節境界単位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，節境界単位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．図 1 の場合，文節「支持するという」の受け文節は「人が」または「なっております」のいずれかであるとして計算する．

なお，$P(b^i_{n_i} \stackrel{rel}{\rightarrow} b^j_l|B)$ は，式 (2) と同様に計算する．

5 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．

5.1 実験に使用したデータ

実験で使用したデータを表 3 に示す．テストデータとして，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，文節まとめ上げを施した 500 文を用いた．正解の節境界，及び，係り受けは人手で付与した．なお，これらのアノテーションは 2.3 節で述べた基準に準拠している．また，節境界をまたぐ係り受け関係は，テストデータの正解中に 152 個存在した．これは，節境界に基づく係り受け解析手法が節境界 “主題ハ” をまたぐ係り受け関係に対処しないとき，係り受け正解率 (文末を除く) が 96.8%(2,521/3,061) を超えることはないことを意味する．

表 6: 節境界解析ツール (CBAP) の実験結果

再現率	95.7% (2,140/2,237)
適合率	96.9% (2,140/2,209)

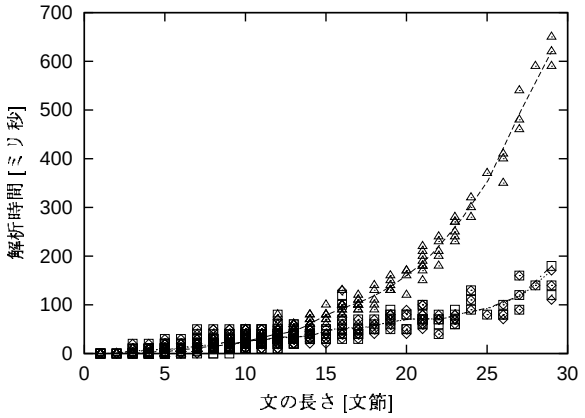

図 5: 文の長さと解析時間の関係

一方，学習データには，形態素解析，文節まとめ上げ，節境界解析，係り受け解析が施された「あすを読む」の書き起こし 5,532 文を用いた．

5.2 実験の概要

本手法の有効性を比較評価するために，上述したデータを用いて以下の 3 つの手法で解析を行い，それぞれの解析時間と解析精度を求めた．

- **節境界に基づく係り受け解析手法（節境界単位の修正あり）**
  3 節，4 節でそれぞれ述べた，節境界解析，係り受け解析を順に行う．
- **節境界に基づく係り受け解析手法（節境界単位の修正なし）**
  上述の手法のうち，3.2 節で述べた節境界単位の修正は行わない．
- **文単位の係り受け解析手法**
  上述の手法のうち，節境界解析を行わず，文全体の係り受け構造を一度に求める．

5.3 実験結果

各手法の解析時間を表 4 に示す．節境界単位の解析手法の解析速度は文単位の解析手法に比べて，平均して約 3 倍向上した[2]．文の長さと解析時間の関係を図 5 に示す．文単位の係り受け解析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

[2]節境界解析ツール CBAP の 1 文あたりの平均解析時間は 1.2 ミリ秒だった．したがって係り受け解析の前処理として CBAP を用いて節境界解析をする時間的な負担は無視できるほど小さい．

表 4: 3 手法の実験結果（解析時間）

	節境界単位係り受け解析（節境界単位の修正あり）	節境界単位係り受け解析（節境界単位の修正なし）	文単位係り受け解析
平均解析時間 (ミリ秒/文)	31.8	31.6	91.1

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

表 5: 3 手法の実験結果（係り受け正解率）

	節境界単位係り受け解析（節境界単位の修正あり）	節境界単位係り受け解析（節境界単位の修正なし）	文単位係り受け解析
節境界単位の内部	87.9% (2,690/3,061)	87.5% (2,677/3,061)	85.4% (2,615/3,061)
節境界単位の最終文節	63.6% (1,104/1,737)	64.1% (1,114/1,737)	60.3% (1,047/1,737)
全体	79.1% (3,794/4,798)	79.0% (3,791/4,798)	76.3% (3,662/4,798)

が上昇するのに対し，節境界に基づく両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 6032 文の平均文節数は 11.8 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．

各手法の係り受け正解率を表 5 に示す．表 5 の第 1 行は，節境界単位末を除く節境界単位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境界単位末に対する正解率を示す．節境界単位の内部，最終文節とも，文単位係り受け解析手法に劣らない解析精度を節境界単位の両係り受け解析手法が備えていることがわかる．なお，節境界単位の両係り受け解析で用いた CBAP の解析結果を表 6 に示す．適合率，再現率ともに高く，後に行われる解析に悪影響を与えることはほとんどなかったと考えられる．また，節境界単位の係り受け解析手法のうち，節境界単位を修正する手法が修正しない手法の正解率をわずかに上回った．これは，節境界単位を修正した効果を示している．

以上の結果から，節境界に基づく係り受け解析手法によって，解析精度を従来の文単位係り受け解析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．

6 考察

本節では，まず，節境界単位で係り受け解析を行う効果について，文単位の係り受け解析手法と節境界単位の係り受け解析（単位修正なし）手法との実験結果の比較により分析する．次に，できるだけ係り受けが閉じるように節境界単位を修正する効果について，単位修正をしない手法と修正する手法の比較をもとに考察する．

6.1 節境界単位への分割による効果

実験では，節境界単位の内部，節境界単位の最終文節のいずれにおいても節境界に基づく解析手法の精度が文単位解析手法に劣ることはなかった．以下では，節境界単位の内部と節境界単位の最終文節の二つに分けて考察する．

表 7: 節境界単位解析（修正なし）と 文単位解析の比較

（節境界単位の内部の係り受け解析結果）

		節境界単位解析（修正なし）		合計
		正解	不正解	
文単位解析	正解	2,553	62	2,615
	不正解	124	322	446
合計		2,677	384	3,061

6.1.1 節境界単位の内部の解析に関する考察

表 7 は，節境界単位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，不正解の関係を示す．節境界単位内の係り受け関係 3,061 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 2,553 個であった．節境界単位の解析手法で正解し，文単位の解析手法で不正解となったものは 124 個にのぼる．これは，節境界単位の解析手法が受け文節の候補を節境界単位内に絞った効果を示している．

一方，文単位の解析手法のみで正しく解析できた係り受け関係は 62 個であった．このうち 32 個は節境界をまたぐ係り受け関係であり，節境界単位の解析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，文単位の解析手法で正しく解析される係り受け関係のほとんどを節境界単位の解析手法によって正しく解析できることを意味する．

6.1.2 節境界単位の最終文節の解析に関する考察

表 5 が示すように，節境界単位の最終文節（文末を除く）の係り受け正解率は，節境界単位内解析のものと比べて両手法ともかなり低い．これは，節境界単位の最終文節を係り文節とする係り受け関係の

表 9: 節境界をまたぐ係り受けに対する 3 手法の実験結果

	節境界単位係り受け解析（節境界単位の修正あり）	節境界単位係り受け解析（節境界単位の修正なし）	文単位係り受け解析
再現率	17.1% (26/152)	1.3% (2/152)	21.1% (32/152)
適合率	51.0% (26/ 51)	40.0% (2/ 5)	37.2% (32/ 86)

表 8: 節境界単位解析（修正なし）と 文単位解析の比較

（節境界単位の最終文節の係り受け解析結果）

		節境界単位解析（修正なし）		合計
		正解	不正解	
文単位解析	正解	943	104	1,047
	不正解	171	519	690
合計		1,114	623	1,737

表 10: 節境界単位解析（修正あり）と（修正なし）の比較

（節境界単位の内部の係り受け解析結果）

		節境界単位解析（修正あり）		合計
		正解	不正解	
節境界単位解析（修正なし）	正解	2,665	12	2,677
	不正解	25	359	384
合計		2,690	371	3,061

表 11: 節境界単位解析（修正あり）と（修正なし）の比較

（節境界単位の最終文節の係り受け解析結果）

		節境界単位解析（修正あり）		合計
		正解	不正解	
節境界単位解析（修正なし）	正解	1,098	16	1,114
	不正解	6	617	623
合計		1,104	633	1,737

同定が難しいことを意味している．このような係り受け関係の解析に関連して，一文中に複数の従属節が存在する場合にそれらの節間の係り受け関係を解析する研究が行われている [12, 14].

表 8 に，節境界単位の最終文節 (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．節境界単位の最終文節を係り文節とする係り受け関係 1,737 個のうち，両手法でともに正しく解析された係り受け関係は 943 個あった．節境界単位の解析手法でのみ正解した係り受け関係は 171 個で，文単位の解析手法でのみ正しく解析できた係り受け関係 104 個を上回った．これは，節境界単位の解析手法が，先に解析した節境界単位内の係り受け解析結果を前提とすることにより，節境界単位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．

6.2 節境界単位の修正による効果

節境界単位の係り受け解析手法のうち，できるだけ係り受けが閉じるように節境界単位を修正する処理を追加した手法と修正しない手法の実験結果を比較し，節境界単位の修正効果を考察する．表 5 が示すように，節境界単位を修正する手法が，修正しない手法と比較し，節境界単位内解析では高い正解率を得たが，節境界単位末解析では逆の結果となった．以下では，両手法の節境界単位の内部と最終文節の解析結果を分けて考察する．

6.2.1 節境界単位の内部の解析に関する考察

表 10 に，節境界単位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，不正解の関係を示す．節境界単位内の係り受け関係 3,061 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 2,665 個であった．節境界単位の修正あり手法で正解し，修正なし手法で不正解となったものは 25 個で，修正なし手法のみで正解した 12 個を上回った．これは，節境界単位の修正あり手法が，できるだけ係り受けが閉じるように節境界単位を修正した結果，節境界をまたぐ係り受けを一部解析できるようになったためである．

表 9 に，節境界をまたぐ係り受け関係に対する 3 つの手法の実験結果をそれぞれ示す．単位修正をしない手法は，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，そもそも一つも正しく解析できない．実験結果では，2 個の節境界をまたぐ係り受け関係を同定しているが，これは，節境界解析の段階で誤った節境界が付与されたために同定できたものである．

一方，単位修正あり手法は，節境界単位を修正することによって，節境界をまたぐ係り受け関係のうち，17%を正しく解析できた．この 17%は，文単位解析手法の再現率 21%と比べ下回っているものの，単位修正なし手法と比べると大幅に改善されている．また，単位修正あり手法の適合率は，文単位解析と比べ上回っており，この手法が比較的悪影響なく，効果的に節境界をまたぐ係り受け関係を同定できていることがわかる．

6.2.2 節境界単位の最終文節の解析に関する考察

表 11 に，節境界単位の最終文節 (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．節境界単位の最終文節を係り文節とする係り受け関係 1,737 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 1,098 個あった．単位修正あり手法でのみ正解した係り受け関係は 6 個で，単位修正なし手法でのみ正しく解析できた係り受け関係 16 個を下回った．これは，節境界単位を修正する手法の場合，節境界をまたぐ係り受けの係り文節を検出し直後の節境界単位に連結するため，節境界単位末の解析時に係り先の候補が増え，解析のあいまい性が高まるためだと考えられる．したがって，節境界単位「主題ハ」を移動した文節列は，その節境界単位末の係り先の候補として除くことが考えられる．

7 おわりに

本論文では，節境界に基づく独話文の係り受け解析の効率化手法を提案した．さらに，節境界 “主題ハ” をまたぐ係り受け関係を同定するために，節境界単位を修正する手法を提案した．これらの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，節境界単位の係り受け解析手法が，従来の文単位係り受け解析手法と比べて，解析精度を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．また，節境界単位を修正することにより，節境界 “主題ハ” をまたぐ係り受け関係を同定できるようになり，解析精度が向上した．

今後は，節境界 “主題ハ” 以外の節境界をまたぐ係り受け関係を検出し同定する手法について検討したい．

謝辞 独話文係り受けコーパスの作成に御協力いただいた名古屋大学大学院国際言語文化研究科の大学院生のみなさまに感謝致します．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] 浅原 正幸，松本 裕治: IPADIC ユーザーズマニュアル, version2.5.1 (2002).

[2] 藤尾 正和，松本 裕治: 語の共起確率に基づく係り受け解析とその評価，情報処理学会論文誌, Vol.40, No.12, pp.4201-4211 (1999).

[3] 柏岡秀紀，丸山岳彦: 節境界単位による翻訳 - 連体節について -，言語処理学会第 10 回年次大会論文集，pp.460-463 (2004).

[4] 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).

[5] 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).

[6] 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).

[7] 前川 喜久雄, 籠宮 隆之, 小磯 花絵, 小椋 秀樹, 菊池 英明: 日本語話し言葉コーパスの設計, 音声研究, Vol.4, No.2, pp.51-61 (2000).

[8] 益岡 隆志，田窪 行則: 基礎日本語文法 - 改訂版 -，くろしお出版 (1992).

[9] 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).

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

[11] 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).

[12] 宇津呂 武仁，西岡山 滋之，藤尾 正和，松本 裕治 : コーパスからの日本語従属節係り受け選好情報の抽出およびその評価，自然言語処理，Vol.6, No.7，pp.29-60 (1999).

[13] 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).

[14] 白井諭, 池原悟, 横尾昭男, 木村淳子 : 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, Vol.36, No.10, pp.2353-2361 (1995).

[15] 内元 清貴，関根 聡，井佐原 均: 最大エントロピー法に基づくモデルを用いた日本語係り受け解析，情報処理学会論文誌，Vol.40, No.9, pp.3397-3407 (1999).