社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

節の始端検出に基づく独話文の係り受け解析

大野　誠寛† 　松原　茂樹††,††† 　柏岡　秀紀†††,†††† 　稲垣　康善†††††

† 名古屋大学大学院国際開発研究科，名古屋市
†† 名古屋大学情報連携基盤センター，名古屋市
††† 情報通信研究機構音声言語グループ，京都府
†††† ATR 音声言語コミュニケーション研究所，京都府
††††† 愛知工業大学経営情報科学部，豊田市

あらまし　1 文の長さが長いという特徴をもつ独話文の高性能な係り受け解析を実現するため，節に分割し，節レベルと文レベルの 2 段階で係り受け解析を実行する枠組みが提案されており，その有効性が確認されている．しかし，上述の枠組みにおいて，節そのものに文を分割することはできないため，解析の処理単位として節に相当する単位を近似的に利用することになる．これまでの研究では，節の終端境界で挟まれた単位（以下，節境界単位）が解析の処理単位として利用されているが，その内部で係り受けが閉じない場合があることが問題となっていた．本論文では，節レベルと文レベルの 2 段階で解析を実行する枠組みに基づいて，節境界単位を拡張した完全に係り受けが閉じた単位を解析の処理単位とする係り受け解析手法を提案する．本手法では，ポーズや節境界タイプを考慮して，機械学習により節境界単位で閉じない係り受けの係り文節を検出し，この直後で節境界単位を再分割することにより，係り受けが閉じた単位を同定する．この単位を解析の処理単位として利用することにより，解析精度が改善することを確認した．
キーワード　節境界解析，構文解析，係り受け構造，音声言語，コーパス

Dependency Parsing of Spoken Monologue Based on Clause-Start Identification

Tomohiro OHNO†, Shigeki MATSUBARA††,†††,

Hideki KASHIOKA†††,††††, and Yasuyoshi INAGAKI†††††

† Graduate School of International Development, Nagoya University, Nagoya-shi
†† Information Technology Center, Nagoya University, Nagoya-shi
††† NICT Spoken Language Communication Group, Kyoto-fu
†††† ATR Spoken Language Communication Research Laboratories, Kyoto-fu
††††† Faculty of Management And Information Science, Aichi Institute of Technology, Toyota-shi

Abstract　A dependency parsing method based on segmenting a sentence into clauses has been proposed and confirmed to be effective. In this method, dependency parsing is executed in two stages: at the clause level and the sentence level. However, since a sentence can not be segmented into complete clauses, in the past research, a unit sandwiched between two clause-end boundaries (**clause boundary unit**) is adapted as an approximate unit of the complete clause. There has been a problem that the dependency structure of the clause boundary unit is not necessarily closed. This paper proposes a method for dependency parsing based on segmenting a sentence into units which corresponds to clauses and whose dependency structure is completely closed. Our method identifies a unit whose dependency structure is closed by redividing a clause boundary unit at modifier bunsetsus of dependency relations over clause-end boundaries. As the results of the experiment, we confirmed the improvement of the dependency accuracy by utilizing the unit as a parsing unit.
Key words　clause boundary analysis, parsing, dependency structure, spoken language, corpus

1. まえがき

独話データへの効率的なアクセスやその効果的な再利用を実現するために，単に独話を蓄積するだけでなく，その構造情報とともに蓄積することが望ましく，その要素技術の一つとして，独話文の高性能な係り受け解析が必要となる．独話には，極端に長い文が存在するため，従来の文単位での係り受け解析手法を適用すると，解析時間の増加や解析精度の低下が問題となる．これに対し，節が文よりも短く，係り受けが閉じた単位であることに着目し，節を文に代わる解析の処理単位として利用することが考えられ，節レベルと文レベルの 2 段階で係り受け解析を行う枠組みが提案されている [1]．この手法では，まず，節境界解析により，文を節に分割し，各節ごとに係り受け解析を行う．その後，各節の最終文節の係り先を定めることにより，文全体の係り受け構造を作り上げる．独話文を用いた解析実験により，その有効性が確認されている．

しかし，ある節に別の節が埋め込まれている場合など，構文解析の前処理として節の範囲を同定することは容易ではなく，たとえ節を同定できたとしても，語順を入れ替えることなく文を節に分割することはできない．そのため，係り受け解析の前に，節に相当し，かつ，係り受けが閉じた単位に文を分割できるかどうかが重要となる．上述した手法では，形態素列のパタンのみから検出可能な「節の終端境界」によって挟まれた単位（**節境界単位**）を節の近似として利用している．節境界単位は，多くの場合，節と一致するものの，埋め込み節が存在する場合には，節境界単位で閉じない係り受けが生じることになり，このような係り受けは上述した手法では解析できなかった．

本論文では，節レベルと文レベルの 2 段階で解析を実行する枠組み [1] に基づいて，節境界単位を拡張した完全に係り受けが閉じた単位を解析の処理単位とする係り受け解析手法を提案する．本手法では，まず，節境界単位で閉じていない係り受けの係り文節を検出し，この文節の直後で節境界単位を再分割した単位（**節断片**）を解析の処理単位として利用する．節断片は，節に相当し，完全に係り受けが閉じた単位となる．節断片を同定するためには，節境界単位で閉じていない係り受けの係り文節を係り受け解析の前に検出する必要がある．節境界単位で閉じていない係り受けは埋め込み節が存在するときに生じ，この係り文節を検出することは埋め込み節の始端境界を検出することに相当する．

日本語文を節に分割する手法のほとんどは，人手で作成したルールにより節の終端境界を検出し，分割するものであった．これは，日本語の場合，述語句が形態的に発達しており，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くことにより，節の終端境界がかなり正確に同定できることによる [2]．一方，日本語の節の始端境界は，英語の間接代名詞ように，節の先頭に特定の単語が現れるわけではないため，その検出は容易ではない [3]．一部，係り受け解析の結果を利用して節の始端境界を検出することは試みられているものの [3]，係り受け解析の前処理として節の始端境界の検出を試みた研究はほとんどない．一方，英語では，節の終端境界，始端境界のいずれもその検出手法が研究されており，人手で作成したルールに基づく手法 [4], [5] だけでなく，機械学習による手法 [6]〜[8] も提案されている．特に最近では，ACL2001 の併設ワークショップ CoNLL2001 で取り上げられる [9] など，機械学習によって節の終端境界，始端境界を検出する研究も行われており，高い成果が得られている．

本手法では，ポーズや節境界タイプを考慮した機械学習に基づいて節境界単位で閉じない係り受けの係り文節を検出し，この直後で節境界単位を再分割することにより，節断片を同定する．節断片は係り受けが閉じた単位となるため，全ての係り受けが解析対象となる．独話データを用いた実験の結果，本手法により，節境界単位で閉じていない係り受けが同定できるようになり，解析精度が改善することを確認した．

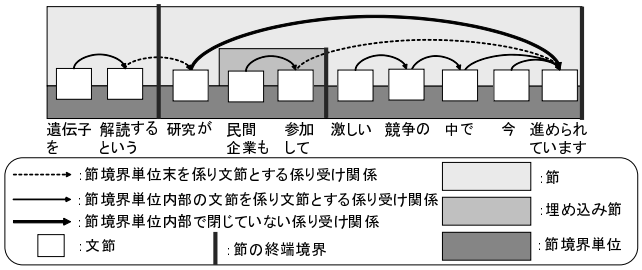

図 1　節と節境界単位

2. 独話文の解析単位

本研究では，文より短い節に相当する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 節と節境界単位

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．これまで我々は，節境界解析 [2] を用いることにより，節への分割を近似的に実現してきた．節境界解析では，局所的な形態素列のみを手がかりとして，節の終端境界と種類を特定することができる．この解析により検出される節の終端境界により挟まれた単位を**節境界単位**(注1)とよび，これを新たな解析単位として考えてきた．節境界単位は，ほとんどの場合，節と一致する．

しかし，図 1 に示すように，埋め込み節がある場合，節境界単位は節と一致しない．この例では，節「遺伝子を解読するという」と「民間企業も参加して」の終端境界に挟まれた「研究が民間企業も参加して」が節境界単位となるが，実際は「民間企業も参加して」で節を形成しており，節と節境界単位が一致しない．

2.2 節　断　片

本手法では，節境界単位を拡張し，節の終端境界だけでなく，節境界単位で閉じていない係り受けの係り文節の直後でも文を分割し，これらの境界によって挟まれた単位を**節断片**と呼び，これを新たな解析単位として採用する．なお，節断片はその内部で係り受けが必ず閉じる．また，節境界単位で閉じていない係り受けの係り文節の直後に節の始端境界が存在すると考えることができる(注2)．本手法では，「独話文は一つ以上の節断片の連接であり，各節断片を構成する文節は，節境界単位の最終文節を除き，その節境界単位の内部の文節に係る」とみなして，係り受け解析を実行する．

例として，独話文「遺伝子を解読するという研究が民間企業

(注1)：節境界単位の終端境界に付与されたラベル名をその節境界単位の種類とする．

(注2)：厳密には，節境界単位で閉じていない係り受けの係り文節の直後に節の始端境界があるとは限らない．実際，節境界単位で閉じていない係り受けの係り文節が 2 つ連続して存在した場合，その係り先が同じ節内の文節であれば，この 2 つの文節の境界は節の始端境界にはならない．

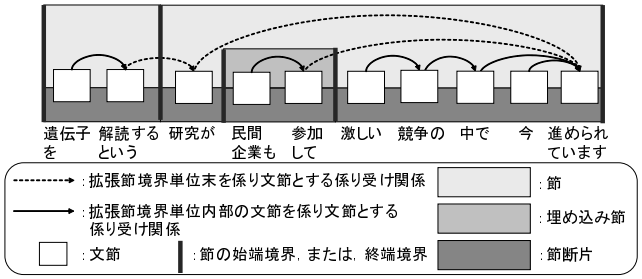

図 2　節と節断片

表 1　分析データ（「あすを読む」）の基礎統計

項目	数値
番組数	6
文数	315
節数	1,612
文節数	4,017
形態素数	9,973
節境界単位で閉じていない係り受け数	140

も参加して激しい競争の中で今進められています」の係り受け構造を図 2 に示す．この文は 4 つの節断片「遺伝子を解読するという」,「研究が」,「民間企業も参加して」,「激しい競争の中で今進められています」から構成され，各節断片が係り受け構造を形成し，それらが節断片の最終文節からの係り受け関係でつながっている．

3.　節境界単位で閉じていない係り受けの特徴

文を節断片に分割するためには，節の終端境界と節境界単位で閉じていない係り受けを検出する必要がある．このうち，節の終端境界については，節境界解析プログラム CBAP [2] によって高い精度で検出することができる．一方，節境界単位で閉じていない係り受けは，ある節が別の節に埋め込まれる場合に生じ，その検出にはある節が埋め込まれていることを示す構文的な情報が必要となると通常は考えられるため，構文解析の前に検出するための特徴はまだ明らかにされていない．本研究では，係り受け解析の前に節境界単位で閉じていない係り受けを検出する必要があるため，この特徴について分析した．

3.1　分析データ

分析には，NHK の解説番組「あすを読む」6 番組分の書き起こしデータ(注3)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．ここで，形態素解析には ChaSen [10] を，文節まとめ上げ，係り受け解析には CaboCha [11] を，節境界解析には CBAP [2] を用いた．なお，形態素解析は ChaSen [10] の IPA 品詞体系 [12] に，文節まとめ上げは CSJ 作成基準 [13] に，節境界解析は丸山らの基準 [2] に，係り受け文法は京大コーパスの作成基準 [14] にそれぞれ準拠して人手により修正している．ただし，話し言葉特有の現象については，新たに作成基準を設けた．具体的には，話し言葉特有の言い回し表現（「こっから」，「という」など）については，新たな辞書項目を設けて，形態素ごとに品詞を定めた．また，文節まとめ上げでは，形式名詞の前で一律に文節を区切る仕様とした．分析データの基礎統計を表 1 に示す．以下では，節境界単位で閉じていない係り受け 140 個について詳しく分析する．

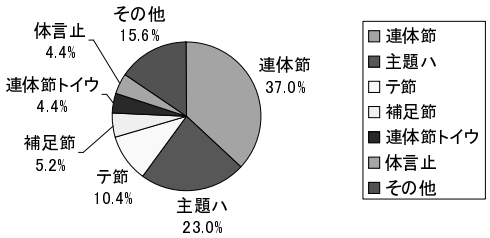

図 3　節境界単位で閉じていない係り受けの係り文節が存在した節境界単位の種類とその割合

3.2　ポーズ情報に基づく分析

2.2 で述べたように，節境界単位で閉じていない係り受けは，係り受け関係が埋め込み節をまたぐことにより生じるため，その係り受け距離が長くなる．したがって，埋め込み節の直前の文節の直後には，この文節の係り先が遠く離れていることを示唆するため，ポーズが入りやすいと考えられる．

そこで，200ms 以上のポーズの分布を分析した(注4)．全体の 76.8% (780/1,015) が節もしくは文の最終文節の直後に存在した．ここで問題となるのは，節内部の文節の中で，節境界単位で閉じていない係り受けの係り文節を検出することであるので，残りの 23.2% (235/1,015) のポーズが存在する節内部の文節について調査した．その結果，直後にポーズが存在する節内部の文節のうち，26.8% (63/235) が節の終端境界をまたぐ係り受けの係り文節であった．一方，節境界単位で閉じていない係り受けの係り文節 140 個のうち，45.0% (63/140) は，直後にポーズが存在することにもなるため，ポーズを考慮することにより，節境界単位で閉じていない係り受けの係り文節を高い再現率で検出できる可能性があることがわかった．

3.3　節境界単位の種類に基づく分析

節境界単位で閉じていない係り受けの係り文節が属している節境界単位の種類に着目した．これは，どのような節が埋め込み節になりうるかということを見ることになる．図 3 に，節境界単位で閉じていない係り受けの係り文節が存在した節境界単位の種類（節の終端境界のラベル名）とその割合を示す．“連体節” が最も多く，次いで，“主題ハ”，“テ節” の順であった．以下では，全体の 70.6%を占めるこの上位 3 つの節境界単位について，その特徴をそれぞれ述べる．

3.3.1　節境界単位の種類 “連体節”

節境界をまたぐ係り受け関係のうち 52 個は，節境界単位 “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，以下の 3 つの現象に大きく分類できることがわかった．

（1）「節境界単位内部の文節が，節境界単位内の述語に係らず，直後の述語に係る現象」が 25 個見られた．

例）　税の公平という見地から痛みも伴う/連体節/
　　税の構造改革に踏み込む/連体節/
　　段階を向かえようとしていると …

「見地から」が,「伴う」に係らず，直後の述語「踏み込む」に係っており，節境界単位 “連体節” で閉じていない．

（2）「節境界単位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 7 個見られた．

例）　大陸を統治する/連体節/
　　国と台湾を統治する/連体節/
　　国が存在している

(注3)：ATR と NHK の共同研究において使用している．

(注4)：分析データには 200ms 以上のポーズ情報が存在している．

「(大陸を統治する) 国と」と,「(台湾を統治する) 国が」が並列関係としての係り受け関係にあり，節境界単位 “連体節” で閉じていない.

(3)「節境界単位内部の文節がこの連体節が修飾する文節を修飾している現象」が 7 個見られた.

例) 警察不服審査審査庁のような独立した/連体節/
機関を作っては …

「警察不服審査審査庁のような」が,「独立した」が修飾する「機関を」を修飾しており，節境界単位 “連体節” で閉じていない.

以上の分析結果から，節境界単位 “連体節” で閉じていない係り受けは，直後の述語，もしくは，直後の名詞に係りやすいことがわかった.

3.3.2 節境界単位の種類 “主題ハ”

節境界をまたぐ係り受け関係のうち 31 個は，節境界単位 “主題ハ” にその係り文節が存在した．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 [2]，本研究ではこれについても節境界単位とした．このような節境界単位を調べてみると,「節境界単位 “主題ハ” 内に述語が存在しないために，直後の述語に係るような文節は節境界単位外に位置する述語に係る現象」が全体の 54.8% (17/31) を占めた.

例) キリシタン文化の流入にマカオは/主題ハ/
深く関わってきました

「流入に」が「関わってきました」(述語) に係っており，節境界単位 “主題ハ” で閉じていない.

3.3.3 節境界単位の種類 “テ節”

節境界をまたぐ係り受け関係のうち 16 個は，節境界単位 “テ節” にその係り文節が存在した．これらの中で,「節境界単位内部の文節が，節境界単位内の述語に係らず，直後の述語に係る現象」が多く見られ，10 個存在した.

例) イギリスが中国に迫って/テ節/
割譲させた/連体節/
植民地であります

文全体の係り受け構造を考えた場合,「イギリスが」は,「迫って」を飛び越えて,「割譲させた」に係るため，節境界単位 “テ節” で係り受けが閉じていない.

4. 節断片への分割

節断片を係り受け解析の処理単位として利用するためには，係り受け解析の前処理として文を節断片に分割する必要がある．本節では，節断片の同定手法について述べる．本手法では，形態素解析及び文節まとめ上げが施された独話文を入力とし，以下の手順により，文中の全ての節の終端境界と節境界単位で閉じていない係り受けの係り文節を検出し，節断片を同定する.

(1) 節境界単位の同定

節境界解析ツール CBAP [2] を用いて入力文に対して節の終端境界を検出し(注5)，節境界単位を同定する.

(2) 節境界単位の分割

節境界単位で閉じていない係り受けの係り文節を検出し，この文節の直後で節境界単位を再度分割することにより，節断片を同定する.

以下では，節境界単位で閉じていない係り受けの係り文節の検出について詳述する.

(注5)：CBAP では，統語的に大きな切れ目になると考えられる「主題ハ」や「談話標識」など,「述語を中心としたまとまり」という節の定義に逸脱した境界も一部検出する [2]．本研究ではこれらも節の終端境界として扱う.

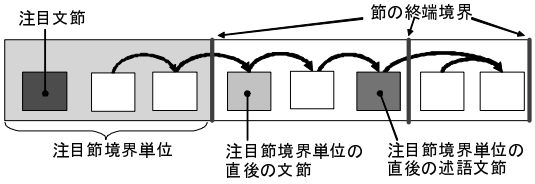

図 4 節境界単位で閉じていない係り受けの係り文節の検出において着目した言語単位

4.1 最大エントロピー法による節境界単位で閉じていない係り受けの検出

節境界単位で閉じていない係り受けの検出アルゴリズムでは，1 文の文節列を入力とし，節境界単位の最終文節でない文節に対して，節境界単位で閉じていない係り受けの係り文節であるか否かの判定を先頭の文節から順に繰り返すことにより，1 文中の節境界単位で閉じていない係り受けの係り文節を検出する.

ある文節が節境界単位で閉じていない係り受けの係り文節であるか否かの判定は，最大エントロピー法に基づくモデルにより，ある文脈において，その文節が節境界単位で閉じていない係り受けの係り文節である確率を推定することにより行う．この確率値が 0.5 以上であれば，この文節は節境界単位で閉じていない係り受けの係り文節であると判定する.

4.1.1 検出モデルに利用した素性

ここでは，節境界単位で閉じていない係り受けの係り文節を検出するための最大エントロピー法に基づくモデルにおいて利用した素性について説明する．本研究では，**3.** の分析結果に基づき，1）現在，節境界単位で閉じていない係り受けの係り文節であるか否かの判定を行っている文節（注目文節），2）注目文節が属している節境界単位（注目節境界単位），3）注目節境界単位の直後の文節，4）注目節境界単位の後から現れる述語のうち最も近い述語を含む文節（注目節境界単位の直後の述語文節），に着目した．図 4 に 1）～4）の言語単位の位置関係を示す．本手法では，これら 4 つの言語単位の以下に示す各素性を利用した.

1）注目文節
- 主辞の基本形，品詞（大分類，細分類）
- 語形の出現形，品詞（大分類，細分類）
- 助詞 1 の出現形，品詞細分類
- 助詞 2 の出現形，品詞細分類
- 直後にポーズがあるか否か

2）注目節境界単位
- ラベル名
- 注目文節以降に同一格の文節が存在するか否か

3）注目節境界単位の直後の文節
- 注目文節と同一主辞の基本形を持つか否か
（注目節境界単位が “連体節” の場合のみ）

4）注目節境界単位の直後の述語文節
- 注目節境界単位内のどの文節よりも
注目文節との係り受け確率(注6)が高いか否か
（注目節境界単位が “主題ハ”or“連体節”or“テ節”，
かつ，注目文節が述語に係りえる文節の場合のみ）

の各情報を素性として利用する.

ここで，**主辞**は，各文節内で，品詞の大分類が記号，助詞，名詞-接尾となるものを除き，最も文末に近い形態素を，**語形**は，各文節内で，記号を除き最も文末に近い形態素をそれぞれ

(注6)：**5.2** の係り受け確率 $P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B)$ と同様に計算する.

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

意味し，各文節内で，一番文末に近い助詞を**助詞１**，その次に文末に近い助詞を**助詞２**として表記している．また，５）で条件としてあげた，ある文節が述語に係りえる文節か否かは，文献 [12], [15] を参考にして，文節の最終形態素の品詞により決定した．表 2 の品詞(注7)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判定する．

5. 節境界に基づく係り受け解析

本手法では，形態素解析，文節まとめ上げ，及び節断片同定が施された文を入力とする．また，この手法では，係り受けの後方修飾性，係り先の唯一性，非交差性の 3 つの性質を絶対的制約とする．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**　1 文中のすべての節断片に対して，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**　1 文中のすべての節断片に対して，その最終文節の係り先を解析する．

なお，以下では，1 独話を構成する節断片列を $C_1 \cdots C_m$，節断片 C_i を構成する文節列を $b^i_1 \cdots b^i_{n_i}$，文節 b^i_k を係り文節とする係り受け関係を $dep(b^i_k)$，1 独話の係り受け構造を $\{dep(b^1_1), \cdots, dep(b^m_{n_m-1})\}$ と記す．

5.1 節レベルの係り受け解析

節レベルの係り受け解析では，節断片 C_i 中の文節列 $b^i_1 \cdots b^i_{n_i}$ 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 $S_i(= \{dep(b^i_1), \cdots, dep(b^i_{n_i-1})\})$ を求める．ここでは，節断片の最終文節 $b^i_{n_i}(1 \leq i \leq m)$ の受け文節は決定しない．

係り受け関係は互いに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．

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

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．$P(b^i_k \overset{rel}{\rightarrow} b^i_l|B_i)$ は，内元ら [16] の係り受け確率モデルを用いて最大エントロピー法により推定した．用いた素性は，内元らの手法 [16] とほぼ同様であるが，話し言葉を対象としているため，読点や括弧の素性は取り除いている．また，節レベルの解析では，内元らの手法で利用されている句点の情報を節末か否かという情報に置き換えた．

5.2 文レベルの係り受け解析

節断片の最終文節の受け文節を同定する．1 文の文節列を $B(=B_1 \cdots B_m)$ とし，節断片の最終文節を係り文節とするような係り受け構造 $\{dep(b^1_{n_1}), \cdots, dep(b^{m-1}_{n_{m-1}})\}$ 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．

$$P(S_{last}|B) \;=\; \prod_{i=1}^{m-1} P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B) \qquad (2)$$

ここで，$P(b^i_{n_i} \overset{rel}{\rightarrow} b^j_l|B)$ は，1 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 に係る確率を表し，**5.1** と同様に最大エントロピー法を用いて計算する．文レベルの解析では，節レベルの解析で利用した素性に，文末か否かの素性を付け加えた素性を利用した．最尤の係り受け構造は，式 (2) 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．

(注7)：この他の品詞として，感動詞や接続詞などがあるが，これらは CBAP により別の節境界単位になるので，ここには含めていない．

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１	学習データ２
文数	500	5,532	2,274
節数	2,237	26,318	10,852
文節数	5,298	65,762	27,027
形態素数	13,342	165,173	7,183

学習データ１：係り受け解析時の学習データ
学習データ２：節境界単位で閉じていない係り受けの係り文節検出時の学習データ

表 4　3 手法の実験結果（平均解析時間:ミリ秒/文）

	節断片 係り受け解析	節境界単位 係り受け解析	文単位 係り受け解析
解析時間	92.56	87.66	209.7

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

6. 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．

6.1 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 [1]．なお，節境界単位で閉じていない係り受け関係は，テストデータの正解中に 152 個存在した．一方，係り受け解析時の学習データには，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．このうち，時間情報が付与されている 2,274 文を節境界単位で閉じていない係り受けの係り文節の検出時の学習データとして利用した．

6.2 実験の概要

本手法の有効性を評価するために，上述したデータを用いて以下の 2 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比較した(注8)．

- **節断片に基づく係り受け解析手法**

3 節，4 節でそれぞれ述べた，節断片への分割，係り受け解析を順に行う．

- **節境界単位に基づく係り受け解析手法**

上述の手法のうち，3 節で述べた節断片への再分割は行わず，節境界単位を解析単位として，係り受け解析を行う．

(注8)：参考までに，節境界単位への分割を行わず，文を解析単位として，文全体の係り受け構造を一度に求める手法（以下，文単位の係り受け解析手法）の解析結果についても記載する．

表 5　3 手法の実験結果（係り受け正解率）

	節断片 係り受け解析	節境界単位 係り受け解析	文単位 係り受け解析
節内部	91.4% (2,798/3,061)	90.2% (2,762/3,061)	90.1% (2,759/3,061)
節末文節	75.8% (1,317/1,737)	75.8% (1,317/1,737)	75.4% (1,309/1,737)
全体	85.8% (4,115/4,798)	85.0% (4,079/4,798)	84.8% (4,068/4,798)

表 6　節境界単位で閉じていない係り受けに対する 3 手法の実験結果

	節断片 係り受け解析	節境界単位 係り受け解析	文単位 係り受け解析
再現率	28.9% (44/152)	1.3% (2/152)	40.8% (62/152)
適合率	53.7% (44/ 82)	11.8% (2/ 17)	40.3% (62/153)

表 7　節境界単位で閉じていない係り受けの係り文節の検出結果

再現率	48.7% (74/152)
適合率	58.7% (74/126)

なお，学習のための最大エントロピー法のツールとしては，文献 [17] のものを利用し，オプション等はデフォルトのまま使用した．

6.3 実験結果

各手法の解析時間を表 4 に示す．節断片と節境界単位の両解析手法の解析時間にはほとんど差がなかった．係り受け解析の前処理として，節境界単位で閉じていない係り受けの係り文節を検出することにより解析全体に与える時間的な影響はほとんどないことがわかった．

次に，各手法の係り受け正解率を表 5 に示す．表 5 の第 1 行は，節末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末文節に対する正解率を示す．節断片の解析手法は，節境界単位の解析手法と比べ，正解率がわずかに増加した．節境界単位で閉じていない係り受けは 152 個しかなく，節断片を解析の処理単位として利用することにより，これらが全て正解できたとしても，3.2%(152/4,798) しか増加しないことを考えると，本手法が，そもそも解析対象でない係り受けが存在するという従来手法の弱点を克服しつつ，正解率を 0.8%向上させたことは，本手法の効果を示している．

節境界単位で閉じていない係り受けに対する係り受け解析結果を表 6 に示す．節断片を解析の処理単位とすることによって，節境界単位で閉じていない係り受けを解析できるようになっていることがわかる．ここで，節境界単位で閉じていない係り受けの係り文節の検出結果を表 7 に示す．節境界単位で閉じていない係り受けの係り文節の検出における再現率・適合率はそれほど高くなく，節断片をより正確に同定することが望まれる．しかし，節断片の同定の精度がこの程度であっても，上述したように，節断片の解析手法は，節境界単位の解析手法と比べ，係り受け正解率が改善した．節断片の同定が多少不正確であっても，係り受け解析の段階でそのミスが吸収されるためだと考えられる．

以上の結果から，本手法によって，節境界単位の解析手法の解析速度を同程度に維持しつつ，解析精度を改善できることを確認した．

7. おわりに

本論文では，節境界単位に基づく係り受け解析を拡張し，これまでは解析できなかった節境界単位で閉じていない係り受け関係も解析可能な手法を提案した．解析実験の結果，本手法によって，節境界単位に基づく係り受け解析手法と同程度の解析時間を維持しつつ，解析精度を改善できることを確認した．今後は，節境界単位で閉じていない係り受けの係り文節をより正確に検出するため，節境界単位 “連体節” や “テ節” で閉じていない係り受けの係り文節を検出する手法について検討したい．

謝辞　本研究は，科学研究費補助金（特別研究員奨励費）「大規模音声言語コーパスを用いた独話データの構造化とその応用に関する研究」（課題番号 18・6433）により実施したものである．

文　献

[1] T. Ohno, S. Matsubara, H. Kashioka, T. Maruyama, and Y. Inagaki, “Dependency parsing of japanese spoken monologue based on clause boundaries,” Proc. of the Joint 21st COLING and 44th ACL, pp.169-176, 2006.

[2] 丸山岳彦，柏岡秀紀，熊野正，田中英輝，“日本語節境界プログラム CBAP の開発とその評価,” 自然言語処理，vol.11，no.3，pp.517-520，2004.

[3] 浜辺良二，内元清貴，河原達也，井佐原均，“話し言葉における引用節・挿入節の自動認定結果を利用した係り受け解析,” 言語処理学会第 12 回年次大会発表論文集，pp.133-136，2006.

[4] H.V. Papageorgiou, “Clause recognition in the framework of alignment,” in Recent Advances in Natural Language Processing, eds. R. MitKov, and N. Nicolov, John Benjamins, Amsterdam/Philadelphia, 1997.

[5] V.J. Leffa, “Clause processing in complex sentences,” Proc. of the 1st LREC, pp.937-943, 1998.

[6] X. Carreras, and L. Màrquez, “Boosting trees for clause splitting,” Proc. of CoNLL2001, pp.73–75, 2001.

[7] A. Molina, and F. Pla, “Clause detection using hmm,” Proc. of CoNLL2001, pp.70–72, 2001.

[8] E.F. Tjong Kim Sang, “Memory-based clause identification,” Proc. of CoNLL2001, pp.67–69, Toulouse, France, 2001.

[9] E.F. Tjong Kim Sang, and H. Déjean, “Introduction to the conll-2001 shared task: Clause identification,” Proc. of CoNLL2001, pp.53–57, 2001.

[10] 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』 and version2.2.9 and 使用説明書,，2002.

[11] 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.

[12] 浅原正幸，松本裕治，IPADIC ユーザーズマニュアル，version2.5.1,，奈良先端科学技術大学院大学，2002.

[13] 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.

[14] S. Kurohashi, and M. Nagao, “Building a Japanese parsed corpus while improving the parsing system,” Proc. of the 1st LREC, pp.719-724, 1998.

[15] 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.

[16] 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.

[17] L. Zhang, “Maximum entropy modeling toolkit for python and c++,” `http://homepages.inf.ed.ac.uk/s0450736/maxent\_toolkit.html`, 2007, [Online; accessed 6-September-2007].