論　文

節境界に基づく独話の漸進的係り受け解析

大野　誠寛[†a)]　松原　茂樹[††,†††]　柏岡　秀紀[†††]　加藤　直人[†††*]
稲垣　康善[††††]

Incremental Dependency Parsing of Japanese Spoken Monologue Based on Clause Boundaries

Tomohiro OHNO[†a)], Shigeki MATSUBARA[††,†††], Hideki KASHIOKA[†††], Naoto KATO[†††*], and Yasuyoshi INAGAKI[††††]

あらまし　同時通訳や字幕生成のように，独話を同時的に処理する音声言語処理システムでは，音声入力に従って順次，解析を実行する漸進的解析技術が必要である．本論文では，節を解析単位とする独話の漸進的係り受け解析手法を提案する．本手法では，節境界解析に基づき，話者による音声入力と同時的に節を同定する．節が入力されるたびにその節の内部の係り受け構造を作成し，既に入力された節との係り受け関係を動的に決定する．独話文全体が入力される前の段階で係り受け関係を出力することが可能であり，同時的な音声理解のための言語解析技術として利用できる．独話データを用いた解析実験により，本手法が，文を解析単位とした係り受け解析と同程度の解析性能を備えていることを確認した．

キーワード　音声言語，係り受け構造，文分割，コーパス，構文解析

1. まえがき

音声同時通訳や自動字幕生成のように，音声を入力と同時に処理するような音声言語処理システムでは，話者による音声入力に従って順次，解析することが求められる．特に，独話を対象とする場合，一般に，文が長くなる傾向にあり，従来の言語処理技術と同様に文単位での逐次的な解析を行うと，入力に対する出力の同時性が損なわれることになるため，漸進的な解析が必須となる．実際，これまでにも漸進的な構文解析に関する研究がいくつか行われており（例えば，[1]〜[4]），そこでは，解析処理の単位，すなわち，どのような言語単位ごとに処理を実行し，結果を出力するのかが問題となる．構文解析の漸進性と正確さの双方を満たすために，文より短く，かつ，構文的なまとまりを備えた言語単位を解析処理の単位として採用することが望ましい．

そこで本論文では，節を解析単位とする独話の漸進的係り受け解析手法を提案する．節は単文に相当し，構文的にも意味的にもまとまった言語単位であるとともに，文が長くなりがちな複文や重文は，複数の節に分割できることから，節は，漸進的な解析処理の単位として適している．特に，長文の構文的あいまい性を軽減することを目的に，節に着目した研究がいくつか行われ，構文解析の精度 [5]〜[7] や機械翻訳等の自然言語処理応用システムの性能 [8]〜[10] の向上が報告されており，節ごとの処理による解析の正確さへの効果が期待できる．

本手法では，独話音声に対して，節が入力されるたびにその節の内部の係り受け構造を作り上げるとともに，既に入力されている節の係り先を決定することを試みる．節の係り先となる文節の決定は，後続するい

† 名古屋大学大学院情報科学研究科，名古屋市
Graduate School of Information Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464–8601 Japan
†† 名古屋大学情報連携基盤センター，名古屋市
Information Technology Center, Nagoya University, Furo-cho, Chikusa-ku, Nagoya-shi, 464–8601 Japan
††† ATR 音声言語コミュニケーション研究所，京都府
ATR Spoken Language Translation Research Laboratories, 2–2–2 Hikaridai, Seika-cho, Soraku-gun, Kyoto-fu, 619–0288 Japan
†††† 愛知県立大学情報科学部，愛知県
Faculty of Information Science and Technology, Aichi Prefectural University, 1522–3 Ibaragabasama, Kumabari, Nagakute-cho, Aichi-gun, Aichi-ken, 480–1198 Japan
* 現在，NHK 放送技術研究所
a) E-mail: ohno@el.itc.nagoya-u.ac.jp

くつかの文節との係り受けのゆう度を考慮した動的なタイミングで実行する．これにより，独話の入力途中の段階で構造情報を随時出力する漸進的な解析が可能となる．

更に，本手法では，文境界が付与されていない独話データ全体に対してその係り受け構造を解析する．これは，独話には明示的な文末標識がなく，あらかじめ文単位に区切ることは容易ではないという独話の特徴に対応している．独話データを用いた解析実験の結果，本手法により，文を解析単位とした係り受け解析手法と同等の解析精度を維持しつつ，係り受け解析の漸進性を実現できることを確認した．

本論文の構成は以下のとおりである．次章で独話の解析単位について述べ，**3.** で節境界に基づく漸進的係り受け解析手法を示す．**4.** で漸進的係り受け解析アルゴリズムについて説明し，**5.** で解析実験について述べる．**6.** で提案手法について考察し，**7.** で本論文のまとめと今後の課題について述べる．

2. 独話の解析単位

本研究では，解析の処理単位として節を採用し，節が入力されるたびにその時点までの係り受け構造を可能な限り決定する漸進的な独話係り受け解析システムを実現する．本章では，節を独話の漸進的係り受け解析における処理単位とすることについて検討を与える．

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．更に，節は，構文的，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と考えられる．

節を単位とした漸進的係り受け解析とは，節が入力されるたびに，それまでの入力に対する解析結果を出力することを意味し，そのためには，独話の入力と同時的に節への分割を実行できる必要がある．しかし，複文において従属節が主節に埋め込まれる場合など，構文的な解析の前処理として節を漸進的に検出することは必ずしも容易ではない．

そこで，本研究では，節境界解析 [11] を用いることにより，節への漸進的な分割を近似的に実現する．節境界解析では，局所的な形態素列のみを手掛りとして，節の終端境界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と呼び，これを新たな解析単位として考える．節境界単位は，埋込み節がある場合には，節と一致しないものの，それ以外の場合であれば，節と完全に一致する．

2.2 節境界単位の分析

漸進的係り受け解析の漸進性と正確さの双方を実現するために，その処理単位が，節同様，文より短く，かつ，構文的にまとまっていることが望まれる．そこで，本節では，節境界単位がこのような性質を持ち併せているかについて調査した．具体的には，文と比較した節境界単位の長さ，また，係り受けが節境界単位でどの程度閉じているのかを，実際の独話データを用いて分析した．

分析には，NHK の解説番組「あすを読む」の書き起こしデータ(注1)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．ここで，自動解析器として，形態素解析には ChaSen [12] を，文節まとめ上げ，係り受け解析には CaboCha [13] を，節境界解析には CBAP [11] を用いた．なお，形態素解析は ChaSen [12] の IPA 品詞体系 [14] に，文節まとめ上げは CSJ 作成基準 [15] に，節境界解析は丸山らの基準 [11] に，係り受け文法は京大コーパスの作成基準 [16] 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．

200 文の基礎統計を表 1 に示す．まず，節境界単位と文の長さに着目して分析した．文の平均文節長は 12.2 であるのに対して，節境界単位の平均文節長は 2.6 であった．これは，文ごとに解析する場合と比べ，節境界単位ごとに解析することによって大幅に解析の漸進性が改善されることを示唆している．次に，節境界単位と係り受け構造の関係について述べる．総文節

表 1 「あすを読む」200 文の基礎統計
Table 1 200 sentences in “Asu-Wo-Yomu.”

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

(注1)：ATR と NHK の共同研究において使用している．

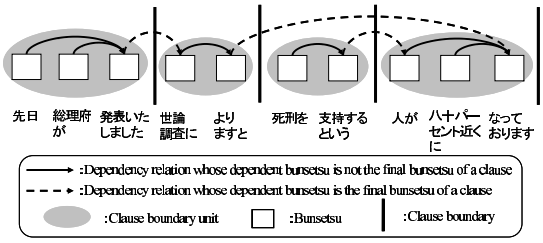

図 1 節境界と係り受け構造の関係
Fig. 1 Relation between clause boundary and dependency structure.

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) の係り受け関係が節境界単位で閉じていることを意味しており，節と同様にある程度構文的にまとまった単位であることを示している．以上の分析結果から，節境界単位が漸進的係り受け解析の処理単位として利用可能であることを確認した．

2.3 漸進的係り受け解析の処理単位としての節境界単位

前節の分析結果に基づき，本研究では，「独話は一つ以上の節境界単位の連接であり，各節境界単位を構成する文節は，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「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．この文は四つの節境界単位「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，各節境界単位が係り受け構造を形成し，それらが節境界単位の最終文節からの係り受け関係でつながっている [17]．

3. 節境界に基づく漸進的係り受け解析

本章では，節境界単位に基づく漸進的係り受け解析について述べる．本手法では，音声入力に対して節境界を随時判定し，節境界単位が同定されると，その時点までの入力に対して係り受け解析を実行する．節境界の判定は節境界解析 [11] により実行する．係り受け解析は，入力された節境界単位の内部の係り受け構造を解析するとともに，既に入力された節境界単位の最終文節の係り先を可能であれば決定する．

本手法では，形態素解析，文節まとめ上げ，及び節境界解析が施された 1 独話を入力とする．ここで，入力データには，文境界が付与されていないことに注意されたい．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，非交差性の三つの性質を絶対的制約とする．解析の手順は以下のとおりである．なお，具体的なアルゴリズムは **4.** で述べる．

（1） 節レベルの係り受け解析
1 独話中のすべての節境界単位に対して，その内部の係り受け構造を解析する．

（2） 独話レベルの係り受け解析
1 独話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．

以下の **3.1**，**3.2** では，各レベルにおける係り受け構造を求める計算方法について述べる．本研究では，係り受け構造を求める際，統計的な手法を用いた．これまでに，統計的係り受け解析手法の学習モデルとして，決定木 [18] や最大エントロピー法 [19]，SVM [13] を用いた手法などが提案されているが，本研究では，節境界を用いることによる解析性能の向上を目的とするため，学習モデルとしては素朴なモデルを用いることとし，具体的には，文献 [20] で提案された共起確率に基づく統計的係り受け解析手法を採用した．

なお，以下では，1 独話を構成する節境界単位列を $C_1 \cdots C_m$，節境界単位 C_i を構成する文節列を $b_1^i \cdots b_{n_i}^i$，文節 b_k^i を係り文節とする係り受け関係を $dep(b_k^i)$，1 独話の係り受け構造を $\{dep(b_1^1), \cdots, dep(b_{n_{m-1}}^m)\}$ と記す．

3.1 節レベルの係り受け解析

節レベルの係り受け解析では，節境界単位 C_i 中の文節列 $b_1^i \cdots b_{n_i}^i$ 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 $S_i(= \{dep(b_1^i), \cdots, dep(b_{n_i-1}^i)\})$ を求める．ここでは，節境界単位の最終文節 $b_{n_i}^i (1 \leq i \leq m)$ 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．

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

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．最ゆうの係り受け構造は，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．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場合は文節末の形態素の語彙，品詞，活用形であり，そうでない場合は文節末の形態素の品詞，活用形である [21]．なお，これらの属性は，従来の係り受け解析手法 [13], [19]〜[21] で用いられてきたものと同様である．

更に，本手法では，受け文節が節境界単位の最終文節であるか否かを示す属性 e_l^i を導入する．文単位で係り受け解析を行う従来手法 [13], [19], [20] では，受け文節が文末であるか否かを示す属性がよく用いられているが，本手法では，文の概念はなく，それに相当する単位として節境界単位を考えているためである．

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．

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

ただし，F は共起頻度関数である．

なお，本手法では，式 (2) により $P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ を計算するときに起こるデータスパースネスの問題を解決するために，藤尾ら [20] のスムージング手法を用いている．すなわち，式 (2) 中の $F(h_k^i, h_l^i, t_k^i, t_l^i, r_k^i, d_{kl}^{ii})$ が 0 である場合は，次式 (3) を用いて $P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i)$ を計算する．

$$\begin{aligned} &P(b_k^i \overset{rel}{\rightarrow} b_l^i|B_i) \\ &\quad \cong P(b_k^i \overset{rel}{\rightarrow} b_l^i|t_k^i, t_l^i, r_k^i, e_l^i, d_{kl}^{ii}) \\ &\quad = \frac{F(b_k^i \overset{rel}{\rightarrow} b_l^i, t_k^i, t_l^i, r_k^i, e_l^i, d_{kl}^{ii})}{F(t_k^i, t_l^i, r_k^i, e_l^i, d_{kl}^{ii})} \end{aligned} \quad (3)$$

3.2 独話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．1 独話の文節列を $B(=B_1 \cdots B_m)$ 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 $\{dep(b_{n_1}^1), \cdots, dep(b_{n_{m-1}}^{m-1})\}$ 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．

$$P(S_{last}|B) \ = \ \prod_{i=1}^{m-1} P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B) \quad (4)$$

ここで，$P(b_{n_i}^i \overset{rel}{\rightarrow} b_l^j|B)$ は，1 独話の文節列 B が与えられたときに，C_i の最終文節 $b_{n_i}^i$ が b_l^j に係る確率を表し，**3.1** と同様に式 (2)，(3) を用いて計算する(注2)．最ゆうの係り受け構造は，式 (4) 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．

ただし，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．

4. 漸進的係り受け解析アルゴリズム

上述した二つのレベルの解析のうち，節境界単位内部の係り受け解析は，**3.1** で述べた方法により解析すればよい．それに対して節境界単位の最終文節に対する係り受け解析は，その受け文節がいつ入力されるかはあらかじめ明らかであるわけではないため，それを決定するタイミングが問題となる．本研究では，節境界単位の最終文節が入力されてからその受け文節が入力されるまでが格段に長くなることはまれであると考え，ある程度解析が進んだ時点でその受け文節を決定することとした．具体的には，節境界単位が入力されるたびにその時点での最ゆうの係り受け構造を **3.2** で述べた方法により解析し，ある最終文節の受け文節が変化しなかった回数（不変回数）があるしきい値に達したとき，その文節をこの最終文節の受け文節として決定する．以下では，このしきい値を不変化しきい値と呼ぶ．

本章の以下では，節境界単位の最終文節に対する係り受け解析について説明する．

4.1 独話レベルの漸進的解析アルゴリズム

係り受け解析の流れを図 2 に示す．解析では，節境界単位 C_i が入力されるごとに，既に入力された節境界単位 $C_1, \cdots, C_{i-1}$ の各最終文節 $b_{n_1}^1, \cdots, b_{n_{i-1}}^{i-1}$ に対する係り受け構造 $D = \{(dep(b_{n_j}^j), k) | \ 1 \leq j \leq i-1\}$

(注2)：**2.2** の 200 文を分析した結果，節境界単位の最終文節も，その 70.6% (522/751) が別の節境界単位の最終文節に係ることが分かった．このため，独話レベルの解析においても属性 e_l^i を用いる．

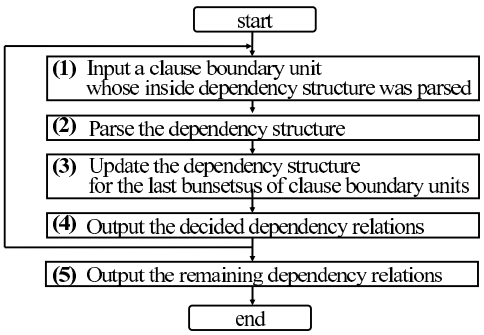

図 2 漸進的係り受け解析の流れ
Fig. 2 Flow of incremental dependency parsing.

を更新することにより実行する．ここで k は $dep(b^j_{n_j})$ の不変回数を示す．以下に係り受け解析アルゴリズムを示す．なお，不変化しきい値を λ とする．

（1） 内部の係り受け構造が決定された節境界単位 C_i を入力する．

（2） 節境界単位の最終文節のうち，係り先が未決定な文節に対して，それを係り文節とする係り受け関係を **3. 2** で説明した方法により求める．

（3）（2）で生成された係り受け関係 $dep(b^j_{n_j})$ に基づき，最終文節に対する係り受け関係 D を更新する．$dep(b^j_{n_j})$ が同一の場合は不変回数を $k+1$ とし，異なる場合は 1 とする．

（4） $k=\lambda$ を満たす係り受け関係 $(dep(b^j_{n_j}),k) \in D$ に対して，文節 $b^j_{n_j}$ の係り先が決定したとして $dep(b^j_{n_j})$ を出力する．

（5） すべての節境界単位が入力された時点で，$k<\lambda$ の $(dep(b^j_{n_j}),k) \in D$ に対して，その係り受け関係 $dep(b^j_{n_j})$ を出力する．

なお，本手法では，文末は係り先がないとして解析する．そのため，節境界単位末の解析では係り先なしを候補に含める．具体的には，式 (4) において，係り先のない文節はそれ自身に係る（すなわち，$b^i_{n_i}=b^j_l$）とし，係り先なしとなる確率も計算する．

4.2 解 析 例

図 3 に，独話「正当な事由がない限り契約期間が切れたといっても明渡しを請求できない点にあるといわれています」の節境界単位末の文節の係り先を解析する様子を示す．(a)〜(f) の六つの過程から構成され，それぞれ上部に係り受け構造を，下部に節境界単位の最終文節の係り受け関係を示す．$(dep(b^j_{n_j}),k) \in D$ の $dep(b^j_{n_j})$ が係り文節 (dependent bunsetsu) 及び受け文節 (head bunsetsu) に，k が不変回数 (continuation) に，それぞれ相当する．なお，ここでは不変化しきい値が 3 であるとして説明する．

(a) は，最初の節境界単位 I が入力された状態を，(b) は，節境界単位 II が入力され，係り受け構造 {dep(限り)} が解析された状態を示す．dep(限り) は上部の点線矢印に相当し，「限り」の係り先が「切れた」であり，不変回数は 1 であることが下部に記録される．同様にして，(c)，(d) は，それぞれ節境界単位 III，IV が入力されたときの最ゆうの係り受け構造 {dep(限り), dep(切れた)}，{dep(限り), dep(切れた), dep(いっても)} が解析された状態を示す．

(e) は，節境界単位 V が新たに入力され，最ゆうの構造 {dep(限り), dep(切れた), dep(いっても), dep(請求できない)} が求まった状態を示している．このとき，係り受け関係 dep(切れた) の不変回数が不変化しきい値として設定した 3 に達したため，この関係を決定し出力する．

(f) は，節境界単位 VI が新たに入力され，最ゆうの係り受け構造 {dep(限り), dep(いっても), dep(請求できない), dep(あると)} が求まった状態を示す．(e) と同様に不変回数が不変化しきい値に達している係り受け関係 dep(限り), dep(いっても) を決定し出力する．

5. 解 析 実 験

独話の漸進的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

実験には，NHK の解説番組「あすを読む」（番組当りの長さは約 10 分）の書き起こしデータを使用した(注3)．使用したデータの概要を表 2 に示す．テストデータとして，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 7 番組（470 文）を用いた．節境界，及び，係り受けの正解は人手で作成した．一方，学習データとしては，形態素，文節まとめ上げ，節境界，係り受けに関する情報が与えられた 95 番組（5,532 文）を用いた．なお，これらのアノテーションは **2. 2** 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 145 個存在した．これは，本手法の係り受け正解率（番組末を除く）が 97.1% (4,902/5,047) を超えることはないことを意味する．

(注3)：**2. 2** の分析で使用した 200 文とは異なるものを用いた．

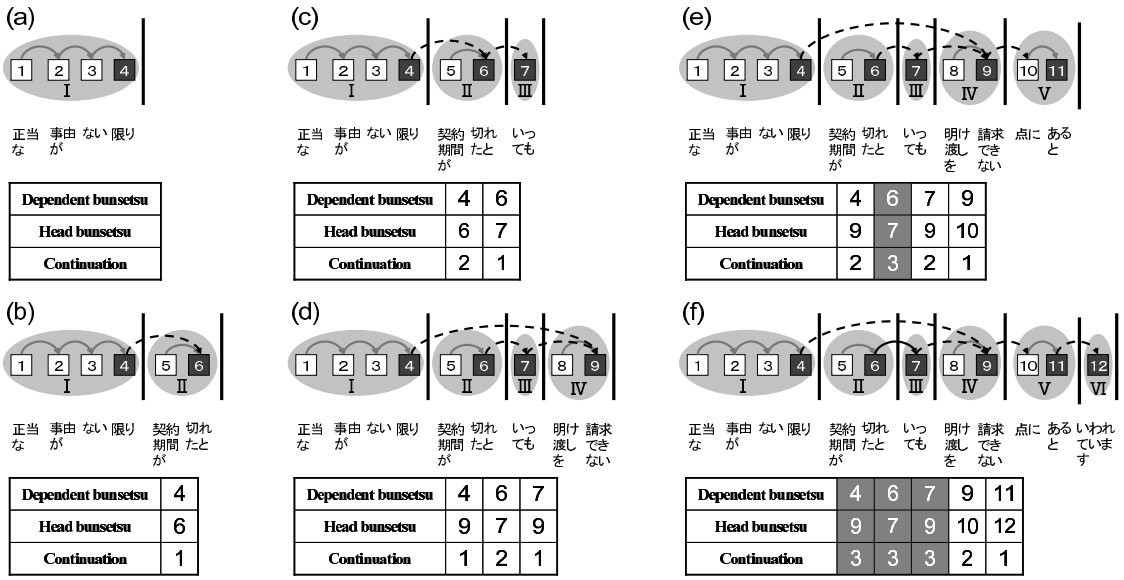

図 3　漸進的係り受け解析の例（不変化しきい値 3 の場合）
Fig. 3　Example of incremental dependency parsing. (in case that continuation threshold is 3)

表 2　実験で使用したデータ（「あすを読む」）
Table 2　Experimental data set. (“Asu-Wo-Yomu”)

	テストデータ	学習データ
番組数	7	95
文数	470	5,532
節境界単位数	2,140	26,318
文節数	5,054	65,821
形態素数	12,753	165,129

5.2　実験の概要

上述したデータを用いて本手法により解析を行った．**4.1** で説明した不変化しきい値を 1 から 12 まで変化させて計 12 回実験し，解析精度，解析時間，解析の漸進性を評価した．

解析精度は係り受け正解率を，解析時間は文節が入力されるごとに実行される解析に要した時間の合計を，それぞれ求めた．また，解析の漸進性を評価するために，遅延時間という尺度を定義し，これを測定した．日本語の係り受けにおける後方修飾性のために，係り受け関係はその受け文節が入力されるまで決定できない．そこで，遅延時間を，受け文節が入力されてからどの程度遅れてその係り受け関係が決定されたかを示す指標として導入することとし，以下のとおり定義する．

番組 p の文節列 $b_1 \cdots b_{n_p}$ を解析するとき，ある文節 b_x が入力された時点で解析システムが決定し出力した係り受け関係の集合を $O_x(p)$ とし，係り受け関係 $\sigma \in O_x(p)$ の受け文節を $f(\sigma)$ とする．また，ある文節 b_x が入力されたときの時間を $g(b_x)$ とする(注4)．このとき，係り受け関係の決定に至る遅延時間の平均

$$Delay = \frac{\sum_{p}\sum_{x=1}^{n_p}\sum_{\sigma \in O_x(p)} \{g(b_x) - g(f(\sigma))\}}{\sum_{p}\sum_{x=1}^{n_p} |O_x(p)|} \quad (5)$$

を計算し，これを漸進性の評価値とする．なお，出力された係り受け関係 σ の受け文節 $f(\sigma)$ は，文節 b_x が入力された時点で既に入力されており，$g(b_x) - g(f(\sigma))$ が負の値になることはない．

比較のため，節境界解析を行わず，文ごとに 1 文の係り受け構造を求める従来手法でも係り受け解析を行った．1 文の文節列が与えられたときの文節間の係り受け確率は，式 (2)，(3) の属性 e_l^i を，受け文節 b_l^i が文末であるか否かを示す属性 s_l^i に変更した式を用いて **3.1** と同様に計算する．実験で使用した書き起こしデータには，句点が付与されており，文ごとに解析可能である．なお，本手法により係り受け解析を行う

(注4)：「あすを読む」の書き起こしデータでは，200 ms 以上のポーズで区切られた発話ごとに発話時間を付与している．文節の発話時間をモーラ数から近似的に算出し，これを文節の入力時間とした．

表 3　不変化しきい値ごとの係り受け正解率
Table 3　Dependency accuracy for each continuation threshold.

不変化しきい値	節境界単位末	全体
1	57.6% (1,228/2,133)	74.9% (3,778/5,047)
2	**60.8%** (1,296/2,133)	**76.2%** (3,847/5,047)
3	**60.8%** (1,296/2,133)	**76.2%** (3,847/5,047)
4	60.4% (1,289/2,133)	76.1% (3,840/5,047)
5	59.8% (1,276/2,133)	75.8% (3,827/5,047)
6	59.4% (1,268/2,133)	75.7% (3,819/5,047)
7	58.8% (1,254/2,133)	75.4% (3,805/5,047)
8	58.6% (1,251/2,133)	75.4% (3,803/5,047)
9	58.7% (1,253/2,133)	75.4% (3,805/5,047)
10	58.4% (1,245/2,133)	75.2% (3,797/5,047)
11	57.6% (1,229/2,133)	74.9% (3,780/5,047)
12	57.9% (1,235/2,133)	75.0% (3,786/5,047)

表 4　CBAP の節境界解析結果
Table 4　Experimental result of clause boundary analysis.

再現率	97.6% (2,088/2,140)
適合率	99.1% (2,088/2,106)

ときは，句点を取り除き 1 番組分の発話を連結した．

これらの解析システムを GNU Common LISP で実装し，CPU が Pentium4 2.40 GHz，メモリが 2 GByte の Linux PC 上で実験した．

5.3 実験結果

本手法の不変化しきい値ごとの係り受け正解率を表 3 に示す．表 3 の第 1 列は，番組末を除くすべての節境界単位末に対する正解率を，第 2 列は，番組末を除くすべての文節に対する正解率を示す．不変化しきい値が 2 及び 3 のときに，節境界単位末に対する正解率が最も高く，全体の正解率は 76.2%となった．なお，節境界単位末を除く節境界単位内に対する解析の正解率は 87.5% (2,551/2,914) であった(注5)．表 4 に，CBAP の節境界解析の精度を示す．これは，CBAP による節境界の検出性能を適合率，再現率により評価した結果である．適合率，再現率ともに高く，後に行われる解析への影響は小さい．一方，従来手法の係り受け正解率は，470 の文末文節を除くすべての文節に対して評価し，76.1% (3,490/4,584) であった．このことから，本手法は，文境界情報を利用していないにもかかわらず，従来手法と同程度の解析精度で，係り受け構造を同定できることが分かる．

本手法の不変化しきい値 (continuation threshold) と 1 番組当りの解析時間 (average parsing time) の関係を図 4 に示す．不変化しきい値を大きくするに従って解析時間が増加している．解析時間が最も短かったのは，不変化しきい値が 3 のときで，全 7 番組で 125.3 秒，1 番組当り 17.8 秒だった．不変化しきい値が 2 のときも，解析時間は 1 番組当り 18.3 秒であり，ほとんど差がなかった．なお，この解析時間には，CBAP による節境界解析の時間も含まれている．節境界解析の解析時間は 1 番組当り 3 秒程度である．一方，従来手法の解析時間は 1 番組当り 6.4 秒であった．本手法は，従来手法と違い，文境界解析を行っていない独話全体の文節列を解析の対象にしているにもかかわらず，従来手法の 3 倍程度の時間で 1 番組を解析できていることが分かる．

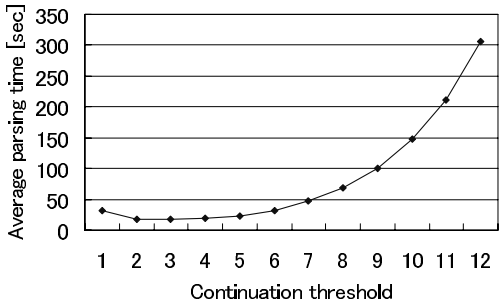

図 4　不変化しきい値と 1 番組当りの解析時間の関係
Fig. 4　Relation of continuation threshold and average parsing time.

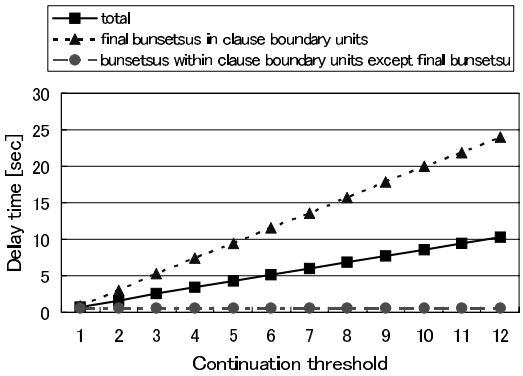

図 5　不変化しきい値と平均遅延時間の関係
Fig. 5　Relation of continuation threshold and average delay time.

図 5 に，本手法の不変化しきい値と係り受け関係決定の平均遅延時間 (average delay time) の関係を示す．この図では，係り文節が節境界単位の最終文節で

(注5)：このうち，節境界単位末から二つ目の文節に対しては，正解率が 97.7% (1,287/1,317) であった．本手法による節境界単位内部の解析では，節境界単位末から二つ目の文節は必ず節境界単位末の文節にかけることになるが，実際，ほとんどの節境界単位で係り受けは閉じているので，高い解析精度が得られたと考えられる．

ある場合と節境界単位内の文節である場合に係り受け関係を分類し，それぞれの平均遅延時間も示している．最も正解率が高かった，不変化しきい値が 2 や 3 のときの平均遅延時間は，それぞれ 1.6 秒，2.5 秒であった．図から，遅延時間のほとんどが，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解析遅延によるものであることが分かる．実際，節境界単位内の文節を係り文節とする係り受け関係の平均解析時間は，不変化しきい値の値に関係なく 0.5 秒程度である．一方，従来手法の平均遅延時間(注6)は，3.2 秒であった．このことから，本手法は，従来手法とくらべて出力タイミングの漸進性を大幅に改善できることが分かる．

以上の結果から，本実験においては不変化しきい値が 2 若しくは 3 のとき，最も高い性能を示しており，文単位を入力とする従来手法と比較して，同程度の解析精度を達成し，なおかつ，解析の漸進性を大幅に改善していることを確認した．

6. 考　　察

6.1 節境界をまたぐ係り受け関係

本手法は，節境界単位内部の係り受け関係は節境界をまたがないとして解析を行っているため，このような係り受け関係を同定することができない．本節では，節境界をまたぐ係り受け関係の解析について考察する．

表 5 に，節境界をまたぐ係り受け関係に対する本手法と従来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，本来，一つも正しく解析できない．実験結果では，本手法により節境界をまたぐ係り受け関係を一つ同定できているが，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，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 145 個のうち 32 個を正しく解析した．従来手法においてもその再現率は 22.1%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．

解析精度の向上のためには，節境界をまたぐ係り受け関係を考慮した処理が望まれる．そこで，以下では，実験に使用したテストデータ 470 文に存在した節境界をまたぐ係り受け関係 145 個を用いて，その解析可能性について検討する．図 6 に，節境界をまたぐ係り受け関係の係り文節を含んでいる節境界単位の種類(注7)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 71.0%を占めるこの上位三つの節境界単位についてそれぞれ述べる．

表 5　節境界をまたぐ係り受け関係に対する実験結果
Table 5　Parsing accuracy for dependency relations over clause boundaries.

	本手法	従来手法
再現率	0.7% (1/145)	22.1% (32/145)
適合率	25.0% (1/ 4)	37.2% (32/ 86)

6.1.1 節境界単位の種類 “主題ハ”

節境界をまたぐ係り受け関係 145 個のうち 41 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 [11]，本研究ではこれについても節境界単位としている．

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，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」が多く見られた．この場合，述語に係る文節については節境界単位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）　次々と賃貸マンションアパートは/主題ハ/
　　建ってまいります…

「次々と」が「建ってまいります」（述語）に係るため，節境界 “主題ハ” をまたいでいる．

6.1.2 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 36 個は，節境

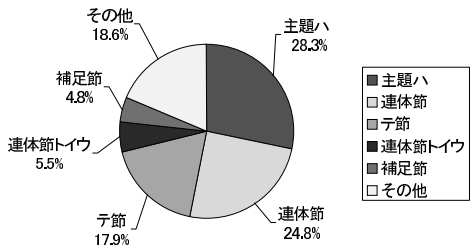

図 6　節境界をまたぐ係り受け関係の係り文節を含む節境界単位の種類とその割合
Fig. 6　Types of clause boundary units which contain the dependent bunsetsus of dependency relations over clause boundaries.

(注6)：1 文ごとに解析されるので，1 文中のどの係り受け関係も文末の入力時点で決定されるとする．
(注7)：節境界解析器 CBAP は，局所的な形態素列のみから，節の終端境界と種類を特定し，テ節や並列節ケレドモなど 144 種の節ラベルを付与することができる [11]．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している．

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，「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, [22], [23]），本手法への導入が考えられる．

例） 貸す/連体節/
側と借りる/連体節/
側の関係を根本から覆すという/連体節トイウ/
ものです…

「（貸す）側と」と，「（借りる）側の」が並列関係としての係り受け関係にあり，節境界 “連体節” をまたいでいる．

6.1.3 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 26 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，節境界単位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，節境界単位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） 検察側が死刑を求めて/テ節/
上告しておりました…

「検察側が」が文全体での係り受け構造としては「上告しておりました」に係るが，同一節境界単位内の「求めて」の主語は「検察側が」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．

6.2 文末検出性能

本手法では，文境界が付与されていない独話データを一度に解析する際，本来文末であるとされる文節は係り先がないとして解析を実行している．すなわち，係り先なしと解析された文節を文末であるとみなすことができる．このような観点から，本手法の文末検出性能を評価した．表 6 に文末検出の適合率，再現率，F 値を示す．不変化しきい値 3 のときに最も高い F 値を示した．独話の文境界検出手法はこれまでにもいくつか提案されている [24], [25]．本手法は，これらと比べ，精度において劣っているものの，漸進的係り受け解析と同時的に文末を検出できるという特徴がある．また，解析精度についても，従来研究 [24] において，ポーズ長などの音響情報を利用する効果が報告されており，今後，これらを利用することにより，精度向上が期待できる．

表 6 文末検出の適合率・再現率・F 値
Table 6 Precision, recall and f-measure for sentence boundary detection.

不変化しきい値	適合率	再現率	F 値
1	54.6% (337/617)	**72.1%** (334/463)	62.1
2	69.8% (312/447)	67.4% (312/463)	68.6
3	74.6% (296/397)	63.9% (296/463)	**68.8**
4	75.7% (278/367)	60.0% (278/463)	66.9
5	76.8% (271/353)	58.5% (271/463)	66.4
6	76.9% (260/338)	56.2% (260/463)	64.9
7	78.5% (252/321)	54.4% (252/463)	64.3
8	78.7% (247/314)	53.3% (247/463)	63.6
9	**81.1%** (249/307)	53.8% (249/463)	64.7
10	80.3% (245/305)	52.9% (245/463)	63.8
11	79.9% (238/298)	51.4% (238/463)	62.6
12	79.8% (233/292)	50.3% (233/463)	61.7

7. むすび

本論文では，節境界単位での漸進的な独話係り受け解析手法を提案した．本手法は，文境界が特定されていない独話発話全体に対して，漸進的に係り受け関係を同定する．文末は係り先がないとして解析する．本手法の有効性を評価するために，独話コーパスを用いて係り受け解析実験を行った．実験の結果，本手法が，文単位を入力とする従来の係り受け解析手法と同程度の解析精度と解析時間を備えつつ，解析の漸進性を向上させることができることを確認した．

今後は，1）節境界をまたぐ係り受け関係を解析可能にする，2）不変化しきい値を節境界単位の種類や単語によって動的に変化させる，3）節境界情報やポーズ情報を考慮した確率モデルを検討する，4）より精緻な学習モデルを採用する，などにより，漸進的係り受け解析の更なる性能向上を図る予定である．

謝辞 日ごろから御指導頂いている名古屋大学大学院情報科学研究科教授の坂部俊樹先生に感謝致します．独話文係り受けコーパスの作成に御協力頂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皆様に感謝致します．査読者の方々には有益な助言，コメントを頂きました．記して感謝致します．本研究は，一部，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」，並びに，総務省戦略的情報通信研究開発推進制度の研究委託「講演など独話データの知的構造化に関する研究開発」，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] J. Nivre, “Incrementality in deterministic dependency parsing,” Proc. ACL Workshop Incremental Parsing: Bringing Engineering and Cognition Together, pp.50–57, 2004.

[2] 加藤芳秀，松原茂樹，外山勝彦，稲垣康善，“確率文脈自由文法に基づく漸進的構文解析，” 電学論（C），vol.122-C, no.1, pp.2109–2119, 2002.

[3] 加藤芳秀，松原茂樹，外山勝彦，稲垣康善，“主辞情報付き文脈自由文法に基づく漸進的な依存構造解析，” 信学論（D-II），vol.J86-D-II, no.1, pp.86–97, Jan. 2003.

[4] B. Roark, “Probabilistic top-down parsing and language modeling,” Computational Linguistics, vol.27, no.2, pp.249–276, 2001.

[5] M. Kim and J. Lee, “Syntactic analysis of long sentences based on s-clauses,” Proc. 1st International Joint Conference on Natural Language Processing, pp.420–427, 2004.

[6] 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.

[7] 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，” 情処学論，vol.36, no.10, pp.2353–2361, 1995.

[8] V.J. Leffa, “Clause processing in complex sentences,” Proc. 1st International Conference on Language Resources and Evaluation, pp.937–943, 1998.

[9] 金　淵培，江原暉将，“日英機械翻訳のための日本語長文自動短文分割と主語の補完，” 情処学論，vol.35, no.6, pp.1018–1028, 1994.

[10] 武石英二，林　良彦，“接続構造解析に基づく日本語複文の分割，” 情処学論，vol.33, no.5, pp.652–663, 1992.

[11] 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，pp.517–520, 2004.

[12] 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.

[13] 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，” 情処学論，vol.43, no.6, pp.1834–1842, 2002.

[14] 浅原正幸，松本裕治，IPADIC ユーザーズマニュアル version2.5.1，2002.

[15] 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.

[16] 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.

[17] 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.

[18] 春野雅彦，白井　諭，大山芳史，“決定木を用いた日本語係り受け解析，” 情処学論，vol.39, no.12, pp.3177–3186, 1998.

[19] 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，” 情処学論，vol.40, no.9, pp.3397–3407, 1999.

[20] 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，” 情処学論，vol.40, no.12, pp.4201–4211, 1999.

[21] T. Ohno, S. Matsubara, N. Kawaguchi, and Y. Inagaki, “Robust dependency parsing of spontaneous Japanese spoken language,” IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.545–552, March 2005.

[22] R. Agarwal and L. Boggess, “A simple but useful approach to conjunct indentification,” Proc. 30th Annual Meeting of the Association for Computational Linguistics, pp.15–21, 1992.

[23] 黒橋禎夫，長尾　真，“長い日本語文における並列構造の推定，” 情処学論，vol.33, no.8, pp.1022–1031, 1992.

[24] 下岡和也，内元清貴，河原達也，井佐原均，“話し言葉の係り受け解析と文境界推定の相互作用による高精度化，” 自然言語処理，vol.11, no.1, pp.119–126, 2005.

[25] 田島幸恵，難波英嗣，奥村　学，“形態素解析器を利用した講演書き起こしの文境界検出について，” 2003 FIT 情報科学技術フォーラム講演論文集，pp.155–156, 2003.

（平成 18 年 4 月 3 日受付，8 月 31 日再受付）

大野　誠寛　（学生員）

2003 名大・工・情報卒．2005 同大大学院情報科学研究科博士前期課程了．現在，同博士後期課程在学中．2006 より日本学術振興会特別研究員．自然言語処理，音声言語処理の研究に従事．情報処理学会，言語処理学会各会員．

松原　茂樹　（正員）

1993 名工大・工・電気情報卒．1998 名大大学院博士課程了．博士（工学）．同年，同大助手を経て，2002 名古屋大学情報連携基盤センター助教授．この間，日本学術振興会特別研究員，ATR 音声言語コミュニケーション研究所客員研究員，NICT 知識創成コミュニケーション研究センター客員研究員を兼任．自然言語処理，情報検索，ディジタル図書館の研究に従事．情報処理学会，人工知能学会，言語処理学会，IEEE，ACM 各会員．

柏岡　秀紀

1993 大阪大学大学院基礎工学研究科博士後期課程了．博士（工学）．同年 ATR 音声翻訳通信研究所入社．1998 同研究所主任研究員（現 ATR 音声言語コミュニケーション研究所）．1999 奈良先端科学技術大学院大学情報科学研究科客員助教授（兼任）．2006 情報通信研究機構専攻研究員（兼任）．2006 ATR 音声言語コミュニケーション研究所音声言語処理研究室室長．主に自然言語処理，機械翻訳，音声言語処理の研究に従事．

加藤　直人　（正員）

1986 早大・理工・電気卒．1988 同大大学院修士課程了．同年日本放送協会（NHK）に入局．NHK 放送技術研究所に勤務．この間，ATR 音声翻訳通信研究所，ATR 音声言語コミュニケーション研究所に出向．博士（情報科学）．機械翻訳，自動要約など自然言語処理の研究に従事．情報処理学会，言語処理学会各会員．

稲垣　康善　（名誉員：フェロー）

1962 名大・工・電子卒．1967 同大大学院博士課程了．同大助教授，三重大学教授を経て，1981 名古屋大学工学部教授．2003 名古屋大学名誉教授，愛知県立大学情報科学部教授．工博．コンピュテーションとコミュニケーションの理論，オートマトン言語理論，ソフトウェア基礎論，自然言語処理に関する研究に従事．本会情報・システムソサイエティ会長，副会長等を歴任，功績賞受賞．情報処理学会名誉会員，日本ソフトウェア科学会，人工知能学会，言語処理学会，IEEE，ACM，EATCS 各会員．