報告番号	※甲　　第　　号

主　論　文　の　要　旨

論文題目　音声言語コーパスを用いた日本語話し言葉の構文解析に関する研究

氏　名　大野 誠寛

論　文　内　容　の　要　旨

近年，音声処理，ならびに，自然言語処理の技術の発展を背景に，音声対話や音声翻訳，音声要約，会話マイニングなど音声言語処理システムに関する研究が盛んに行われている．しかし，現状の音声言語処理システムの多くは，あらかじめ定められた単語や言い回しなどキーワードを処理する方式に留まっている．より高度な処理を実現するための次なるステップとして，音声言語の構文情報の活用が検討されつつあり，音声言語に対する高い性能を備えた構文解析器の開発が望まれている．

一方，構文解析の研究は，これまで，主に書き言葉である新聞記事を対象として行われてきた．言語はその出現形態によって書き言葉と話し言葉に分類でき，さらに話し言葉は，複数の話者が交替で話す「対話」と一人の話者のみが話す「独話」に分類できる．これらはそれぞれ性質が異なるため，書き言葉を対象とした従来の構文解析手法を単に話し言葉に適用しただけでは，様々な問題が生じる．

本論文では，上述した音声言語処理システムの高度な話し言葉処理の実現に必要となる話し言葉の高性能な構文解析器を開発することを目的とする．この開発には，音響情報処理，言語情報処理，視覚情報処理などの複数の分野が互いに関連しており，各分野における技術的向上が不可欠であるが，本研究では，そのうち，自然言語処理技術による解決が主に求められる課題に焦点を絞り，解決することを試みる．すなわち，本研究の目標は，話し言葉の高性能な構文解析を実現するための要素技術として，

1）頑健な構文解析手法

話し言葉（特に，自由対話）では，書き言葉にはない倒置やフィラー，言い淀み，言い直し，言い誤りなど，書き言葉の文法に逸脱する言語現象が頻出する．これら非文法的言語現象を含む文に対して，頑健な解析を実現する．

2）効率的な構文解析手法

解説や講演など一人の話者のみにより話される独話では，書き言葉と比べ，文の切れ目に対する意識が低下するため，極端に長い文が頻出する．一般に，文が長くなればなるほどその解析時間は指数関数的に増加するため，解析効率が低下することになる．このような話し言葉に頻出する極端に長い文に対して，従来の構文解析手法と同程度以上の解析精度を備えた高速な解析を実現する．

3）漸進的な構文解析手法

一般に，書き言葉では，情報が文字列として一度に提示されるのに対して，話し言葉では，情報が音素列として時間軸上で逐次提示される．このため，話し言葉では，入力に追従して情報を処理する応用システムが考えられ（例えば，同時通訳），その要素技術となる構文解析においては，話し言葉の入力に対して漸進的な処理を行うことが求められる．特に，独話では，文が長くなる傾向にあり，文全体の入力を待って解析を開始すると著しく同時性が損なわれることになるため，漸進的な解析が望まれる．そこで，従来の構文解析手法と同程度の解析精度を維持しつつ，話し手の話速に追従できる程度の漸進性を備えた解析を実現する．

を開発することである．本研究では，音声言語コーパスに基づく統計的な手法を用いることにより，これら3つの構文解析手法を開発する．なお，本研究では日本語を構文解析の対象とする．

本論文は全6章から構成される．第1章は本論文の序論であり，話し言葉の構文解析に関する課題及び研究動向を示すとともに，本論文の位置づけとアプローチを述べたものである．

第2章では，話し言葉の構文的特徴を明らかにするための分析データ及び統計的構文解析手法の学習データとして利用することを目的に構築した，対話と独話の2つの構文構造付き音声言語コーパスについて述べる．対話の構文構造付き音声言語コーパスはCIAIR車内音声対話コーパスに対して，独話の構文構造付き音声言語コーパスはNHKの解説番組「あすを読む」の書き起こしコーパスに対して，それぞれ構文構造を付与することにより構築する．両コーパスとも話し言葉に特有な言語現象に対しては新たな付与基準を設けている．また，構文構造付き音声独話コーパスには，節境界情報や複数の係り先を付与しているという特徴がある．構築した対話と独話の構文構造付きコーパスは，それぞれ，85,870形態素，192,495形態素規模を備えている．

第3章では，大規模音声言語コーパスを用いた話し言葉の頑健な係り受け解析手法を提案する．本章の研究では，フィラーや言い淀み，倒置などの非文法的言語現

象が頻出する対話文を構文解析の対象とした．実際に，第2章で構築したCIAIR構文構造付き音声対話コーパスを分析した結果，従来の係り受け解析手法では，係り受けの非交差性，後方修飾性，係り先の唯一性の3つの制約が用いられてきたが，対話音声では，後方修飾性を満たさない倒置現象や係り先の唯一性を満たさない文節などを含む発話が頻出することが分かった．そこで本手法では，後方修飾性の制約及び係り先の唯一性に関する制約は統計情報を反映させつつ緩和する．また，本手法では，構築した構文構造付き音声対話コーパスから各文節間の係り受け確率を統計的に獲得し，それを用いて係り受け構造の尤度を計算する．これにより，非文法的な特徴をもつ発話の解析が可能になる．CIAIR構文構造付き音声対話コーパスに対して係り受け解析実験を行い，その結果，本手法により，自然発話文に対しても，書き言葉を対象とした従来の係り受け解析手法と同等の高い精度で係り受けを抽出できることを確認した．特に，係り先を持たない文節と倒置，発話単位をまたぐ係り受けの解析に対する本手法の頑健性を明らかにした．

第4章では，文の分割に基づく話し言葉の効率的な係り受け解析手法を提案する．本手法では，文の分割単位として節を採用し，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．節は，構文的かつ意味的にまとまった単位であり，文に代わる解析単位として利用できると考えられるためである．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．これにより，話し言葉に出現する極端に長い文に対する効率的な解析の実現が期待できる．極端に長い文が頻出する独話データとして，第2章で構築した「あすを読む」構文構造付き音声独話コーパスを用いた係り受け解析実験を行い，その結果，本手法により，従来の係り受け解析手法と比べ，解析精度を改善しつつ解析時間を約1/5に短縮できることがわかった．

第5章では，話者による音声入力に従って順次，解析を行う漸進的係り受け解析手法を提案する．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，すで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，後続するいくつかの文節との係り受けの尤度を考慮した動的なタイミングで行う．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．「あすを読む」構文構造付き音声独話コーパスを用いた係り受け解析実験を行い，その結果，本手法により，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，解析の漸進性の向上が可能となることを確認した．

最後に，第6章において本論文を総括し，今後の研究課題ならびに将来の展望について示す．