# 漸進的構文解析における長距離依存関係の同定

加藤 芳秀 †　　松原 茂樹 ‡

† 名古屋大学情報基盤センター
‡ 名古屋大学大学院情報科学研究科

## 1　はじめに

長距離依存関係は，ゼロ代名詞の照応関係や構成素の移動といった言語現象を表現する．長距離依存関係の同定は，意味解析や応用システムを実現する上で，重要なタスクである．

本稿では，漸進的構文解析において長距離依存関係を同定する手法を提案する．漸進的構文解析は，自然言語文を単語の出現順序に従って解析し，文の入力途中の段階で，その構文構造を捉える枠組みであり，同時通訳システムやリアルタイム字幕生成システムなどの実時間音声言語処理システムの実現に必要不可欠な要素技術の一つである．漸進的構文解析において長距離依存関係を同定できれば，実時間音声言語処理システムの性能向上に繋がると期待できる．

## 2　長距離依存関係とその同定

本節では，まず長距離依存関係について簡単に説明する．次に，これまでに提案された長距離依存関係同定手法を概観する．

### 2.1　長距離依存関係

長距離依存関係の扱い方は言語理論により異なるが，基本的な考え方に大きな違いはないので，ここでは，Penn Treebank [9] におけるアノテーションに基づき説明する．長距離依存関係は，構文木において，**空要素** (empty element) と**フィラー** (filler) のペアとして表現される．長距離依存関係は，空要素の位置に，対応するフィラーがあるものとして解釈すべきことを示している．図 1 に Penn Treebank における構文木の例を示す．この構文木には，いくつかの長距離依存関係が含まれている．`-NONE-`は空要素であることを示すラベルである．`*T*`や`*`は空要素の種類を示しており，`*T*`は wh 移動の痕跡を，`*`は不定詞の省略された主語を表している．空要素に対して番号が与えられているとき，対応するフィラーが構文木中に存在し，それに対しても同一の番号が与えられる．例えば，`*T*-1` に対しては，`WHNP-1` が対応するフィラーであることを意味する．図中の点線の矢印は，これらの対応関係を示している．

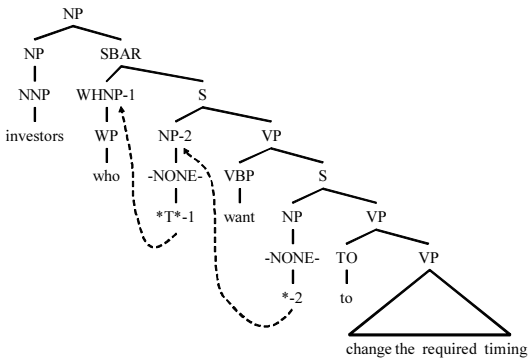


図 1: Penn Treebank の構文木

Penn Treebank には，様々な種類の長距離依存関係が存在するが，Levy らは，これらを以下の 3 つに分類している [8].

**unindexed empty element** 対応するフィラーが存在しない空要素で，関係代名詞や補文標識などの省略を表現する．

**dislocation** フィラーが，空要素の位置から移動したと考えられる長距離依存関係で，wh 疑問文，関係代名詞化，主題化などにおける要素の移動を表現する．Penn Treebank においては，空要素の種類が`*ICH*`，`*RNR*`，`*T*`であるものがこれに該当する．

**control** 不定詞の主語の省略や，受動化などにより出現する長距離依存関係で，省略された主語と実際の主語の対応関係などを表現する．Penn Treebank においては，空要素の種類が`*`であるものがこれに該当する．

Levy らは，長距離依存関係の種類に応じて，異なる戦略を用いてそれを同定する手法を提案しているが，本稿で提案する手法においても，長距離依存関係の種類に応じて，その同定方法は異なる．

### 2.2　長距離依存関係の同定

Penn Treebank に基づく構文解析は数多く提案されているが，そのほとんどは，長距離依存関係を取り除いた構文木を解析結果として返す．以下では，このような構文木を **CF 木**と呼ぶ．図 1 の構文木に対する CF 木を図 2 に示す．

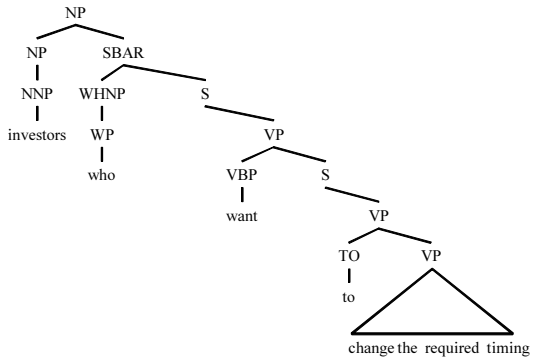


図 2: CF 木の例

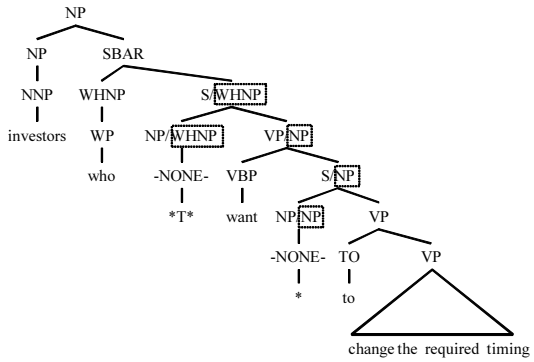


図 3: スラッシュ素性を付与した構文木

これに対して，長距離依存関係を同定する手法がこれまでにいくつか提案されている．これらの手法は，その処理戦略の違いによって，次の 2 種類に分類することができる．

- CF 木から長距離依存関係を復元する手法
- 構文解析に長距離依存関係の同定を統合した手法

前者の手法は，入力として CF 木を受け取り，長距離依存関係を含んだ構文木を出力する手法である．構文解析の後処理として長距離依存関係を同定する．Johnson は，パターンマッチングに基づき CF 木から長距離依存関係を復元する手法を提案している [6]．この手法では，長距離依存関係を復元するためのパターンを構文木コーパスから抽出する．Levy[8] らは，分類器に基づき，CF 木から長距離依存関係を復元する手法を提案している．CF 木中の各ノードが空要素を持つかなどを分類器により識別し，長距離依存関係を同定する．Campbell は，言語学的な原理をルールとして表現し，それに基づき長距離依存関係を復元する手法を提案している [2]．CF 木をルートからトップダウンに辿りながら，長距離依存関係を復元する．

一方，後者の手法は，構文解析と同時的に長距離依存関係を同定する．Dienes ら [4, 5]，Schmid[10]，Cai ら [1] は，スラッシュ素性に相当するアノテーションを構文木に付与する方法を提案している．この手法では，構文木コーパスにおいて，空要素からフィラーの兄弟までのパス上のノードに，図 3 のようにアノテーションを付与する．構文解析の解析結果は，スラッシュ素性を含んだ構文木となり，空要素からスラッシュ素性の与えられたノードをたどり，その兄弟の中からフィラーを選択し，長距離依存関係を同定する．

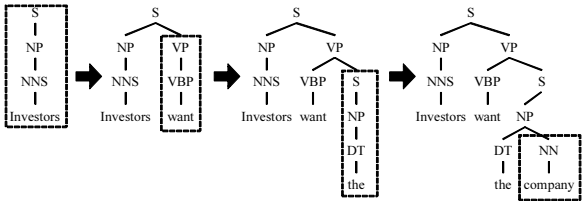


図 4: 漸進的構文解析の解析過程

# 3 漸進的構文解析における長距離依存関係の同定

本節では，漸進的構文解析の解析過程において，長距離依存関係を同定する手法を提案する．まず，漸進的構文解析について説明する．続いて，従来の長距離依存関係同定手法を漸進的構文解析に適用したときに生じる問題について論じ，その問題を解決するための方法を提案する．

## 3.1 漸進的構文解析

漸進的構文解析は，文を先頭から読み込み，読み込んだ文の断片に対してそれを覆う部分構文木を生成する枠組みである．Collins ら [3] や加藤ら [7] が提案している漸進的構文解析では，allowable chain と呼ばれる要素を部分構文木に付加することにより，解析が進行する．allowable chain は，終端記号と非終端記号からなる系列で，最後の要素が終端記号で，それ以外は非終端記号である．構文木中のあるノードから左端の子を順にたどるときに得られるラベルの系列に相当する．文の断片

Investors want the company to . . .

に対する解析過程を図 4 に示す．この例が示すように，漸進的構文解析では，文の断片に対してそれを覆う部分構文木を生成できる．

## 3.2 従来の手法の問題点

漸進的構文解析における長距離依存関係の同定を考えた場合，構文解析の後処理といったアプローチは，適切とはいいがたい．第一に，文の入力途中の段階で長距離依存関係を同定できない．さらに，部分構文木に対して手法を適用するとしても，文の断片に対する部分構文木においては，空要素，あるいはフィラーの片方しか具現化されてない場合があり，そのような場合の処理方法は明らかではない．

一方，構文解析と長距離依存関係同定を統合するアプローチにおいても，空要素からスラッシュ素性の付与されたノードをたどってフィラーを同定するとき，フィラーが構文木中に存在することを前提としているが，そのような前提は，漸進的構文解析が生成する部分構文木においては成り立たない．

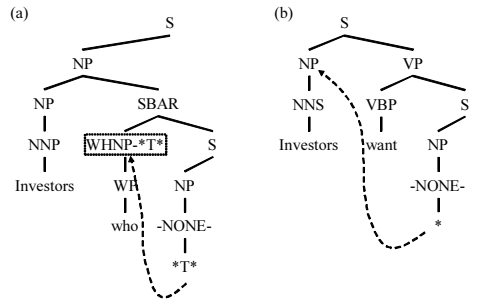


図 5: 対応するフィラーの同定

## 3.3 ルールに基づく長距離依存関係の同定

前節で述べた問題を回避するために，本節では，長距離依存関係を同定する別のアプローチを提案する．

空要素の同定については，スラッシュ素性に基づく従来の手法と同様である．すなわち，各単語間に空要素の存在を仮定して解析を進める．フィラーについては，フィラーに対するアノテーションを導入する．これにより，部分構文木中にフィラーが具現化されているか否かを明示的に示すことができる．空要素とフィラーの対応関係は，ルールに基づき同定する．提案する手法は，空要素とフィラーの両者が部分構文木上に具現化された時点で長距離依存関係を同定できるため，漸進的な解析処理に適している．

### 3.3.1 フィラーに対するアノテーション

本手法では，dislocation のフィラーに対して，フィラーであることを示すために，空要素の種類をアノテーションとして付与する（図 5(a) 参照）．アノテーションが付与されたノードを見つけることにより，解析過程で容易にフィラーを同定できる．一般に，dislocation のフィラーが出現する場所には，フィラー以外の構成素は出現しない．したがって，フィラーと一意に定めてしまっても問題ない．一方，control のフィラーには，アノテーションを与えない．control のフィラーは，図 1 のように主語の名詞句であることが多いが，これは，対応する空要素が存在するか否かに応じて，フィラーである場合もあればそうでない場合もある．

### 3.3.2 空要素とフィラーの対応関係の同定

本節では，長距離依存関係を同定するルールを提案する．提案するルールは，長距離依存関係を構成する要素を同定する順序に特徴がある．従来の手法では，長距離依存関係の種類に応じて，空要素とフィラーを同定する順序が異なる場合があるものの，それらの要素の文中での出現位置には依存しない．一方，本稿で提案する手法では，空要素とフィラーの出現順序により同定の方法が異なる．

空要素が後に出現する長距離依存関係は，次のルールにより同定される．

**対応するフィラーの同定（dislocation）** 種類が *ICH*，*T*のいずれかである空要素 $E$ が部分構文木中に具現化されたとき，次の条件をすべて満たすノード $F$ が部分構文木中に存在すれば，そのうちで $E$ に最も近いノードを $E$ に対応するフィラーとする（図 5(a) 参照）．

- $F$ は，フィラーのアノテーションを持つ．
- $F$ の種類と統語範疇は，$E$ と同一である[1]．
- $F$ は，対応した空要素を持たない．

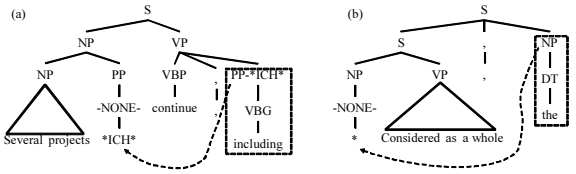


図 6: 対応する空要素の同定

**対応するフィラーの同定（control）** 種類が*である空要素 $E$ が部分構文木中に具現化されたとき，次の条件をすべて満たすノード $F$ が部分構文木中に存在すれば，そのうちで，$E$ に最も近いノードを $E$ に対応するフィラーとする（図 5(b) 参照）．

- $F$ は名詞句である．
- $E$ が主語でないならば，$F$ は主語である．
- $F$ は，$E$ を c-統御する．

これらのルールにより対応するフィラーが見つからなかった場合，対応するフィラーは空要素の後に出現するとみなされる．

フィラーが後に出現する長距離依存関係は，次のルールにより同定される．

**対応する空要素の同定 (dislocation)** フィラーのアノテーションを持つノード $F$ が部分構文木中に具現化されたとき，次の条件を満たす空要素 $E$ が部分構文木中に存在すれば，そのうちで $F$ に最も近いものを $F$ に対応する空要素とする（図 6(a) 参照）．

- $E$ の種類と統語範疇は，$F$ と同一である．
- $E$ は，対応するフィラーを持たない．

**対応する空要素の同定 (control)** 主語である名詞句 $F$ が部分構文木中に具現化されたとき，次の条件を満たす空要素 $E$ が部分構文木中に存在すれば，そのすべてを $F$ に対応する空要素とする（図 6(b) 参照）．

- $E$ の種類は*である．
- $E$ は対応するフィラーを持たない．
- $E$ は，$F$ に c-統御される．

[1] wh 句は WH を除去した上で判定する．例えば，wh 名詞句 WHNP と名詞句 NP は同一の統語範疇とする．

表 1: 長距離依存関係同定の精度と再現率

| スラッシュ素性 | 精度 (%) | 再現率 (%) | F 値 (%) |
|---|---|---|---|
| なし | 72.7 | 65.4 | 68.9 |
| あり | 75.6 | 71.0 | 73.2 |

表 2: CF 木のラベル精度とラベル再現率

| スラッシュ素性 | 精度 (%) | 再現率 (%) | F 値 (%) |
|---|---|---|---|
| なし | 87.3 | 86.1 | 86.7 |
| あり | 87.4 | 86.3 | 86.8 |

## 4 評価実験

提案手法の有効性を確認するために，長距離依存関係同定の精度と再現率を，Johnson[6] が提案した評価法に基づき評価した．Johnson の評価法では，長距離依存関係を「空要素の種類」,「空要素の統語範疇」,「空要素の位置」,「フィラーの範疇」「フィラーの位置」の組として表現し，精度，再現率を評価する．位置は，文中における位置であり，空要素であれば，文中のどの単語の間に位置するか，フィラーであれば，どの単語列に対する構成素かで位置を定める．

漸進的構文解析は，加藤ら [7] が提案した確率的な漸進的構文解析に，長距離依存関係の同定を統合した．文法等は，Penn Treebank の WSJ コーパスのセクション 2-21 から学習している．セクション 23 の 2416 文をテストデータとして評価した．

提案した長距離依存関係同定ルールを評価するために，漸進的構文解析において正解構文木を生成する解析について，長距離依存関係を同定できるかどうかを評価した．その結果，精度は 93.3%，再現率は 92.0%であった[2]．提案したルールはシンプルであるが，高い精度・再現率であり，構文解析に成功すれば有効に機能するルールであると考えられる．

次に，長距離依存関係を統合した漸進的構文解析の性能を評価する実験を行った．長距離依存関係同定の精度・再現率を表 1 に，解析結果の CF 木に関するラベル精度・ラベル再現率を表 2 に示す． スラッシュ素性を付与した構文木コーパスから学習したモデルと，付与していないコーパスから学習したモデルを使用した．スラッシュ素性の有無に関して，構文解析のラベル精度・ラベル再現率に大きな差は見られなかった．一方，長距離依存関係の同定の精度と再現率は，スラッシュ素性を付与したコーパスから学習したモデルのほうが，高い数値を示した．このことは，長距離依存関係を考慮したモデルでなければ，長距離依存関係を正しく復元できないことを示唆する．

## 5 おわりに

本稿では，漸進的構文解析において長距離依存関係を同定する手法を提案した．漸進的構文解析により生成される部分構文木から，ルールに基づき長距離依存関係を同定できることを示した．構文解析実験において，スラッシュ素性を構文木に付与することにより，高い精度で長距離依存関係を同定できるという結果が得られたが，今回提案したルールでは，スラッシュ素性を直接的には利用していない．今後の課題として，スラッシュ素性を考慮したルールの検討などが上げられる．

[2] 長距離依存関係の同定を評価するとき，フィラーを持たない空要素も含めて評価することが多いが，ここでは除いて評価している．正解構文木を生成する解析であるため，フィラーを持たない空要素は，必ず正解となるからである．

**謝辞**

本研究は一部，科研費基盤研究 (B)(No. 22300051) により実施した．

## 参考文献

[1] S. Cai, D. Chiang, and Y. Goldberg. Language-independent parsing with empty elements. *Proc. ACL-HLT-2011*, pp. 212–216, 2011.

[2] R. Campbell. Using linguistic principles to recover empty categories. *Proc. ACL-2004*, pp. 645–652, 2004.

[3] M. Collins and B. Roark. Incremental parsing with the perceptron algorithm. *Proc. ACL-2004*, pp. 111–118, 2004.

[4] P. Dienes and A. Dubey. Deep syntactic processing by combining shallow methods. *Proc. ACL-2003*, pp. 431–438, 2003.

[5] P. Dienes and A. Dubey. Antecedent recovery: Experiments with a trace tagger. *Proc. EMNLP-2003*, pp. 33–40, 2003.

[6] M. Johnson. A simple pattern-matching algorithm for recovering empty nodes and their antecedents. *Proc. ACL-2002*, pp. 136–143, 2002.

[7] Y. Kato and S. Matsubara. Incremental parsing with adjoining operation. *IEICE Transactions on Information and Systems*, E92-D(12):2306–2312, 2009.

[8] R. Levy and C. Manning. Deep dependencies from context-free statistical parsers: Correcting the surface dependency approximation. *Proc. ACL-2004*, pp. 327–334, 2004.

[9] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):310–330, 1993.

[10] H. Schmid. Trace prediction and recovery with unlexicalized PCFGs and slash features. *Proc. COLING-ACL-2006*, pp. 177–184, 2006.