漸進的係り受け解析の出力構造

ー人間の文解析過程のアノテーションー

大野 誠寛　　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
{ohno，matubara}@nagoya-u.jp

1　はじめに

同時通訳や音声対話，字幕生成などの音声言語システムにおいて高度な言語処理を実現するためには，構文情報の利用が不可欠となる．これらの音声言語システムでは音声入力に追従した出力が求められるため，構文情報を利用するためには，音声入力と同時的に処理を進める解析技術が必要となる．このような要請に応えるべく，これまでにも，解析処理を漸進的に進めていく係り受け解析技術が開発されてきた [1, 2, 3, 4, 5, 6]．しかし，これらの自動解析処理は必ずしも十分な性能を達成していない．

一方，人間の言語理解過程において漸進性があることは認知科学や心理言語学の分野において広く認識されている．実際，このことを示唆する実験結果の報告 [7] や，人間の言語処理過程をモデル化する研究 [8, 9] が行われている．しかし，ガーデンパス文などの特殊な文だけでなく，広く一般の文を対象として，その係り受け構造を漸進的に解析するという観点から，人間の言語処理過程を大規模に分析した研究はない．人間による漸進的な係り受け解析結果を定量的に分析し，その能力や振舞いを明らかにすることができれば，漸進的係り受け解析器の性能を向上させるための知見が得られる可能性がある．

そこで本稿では，漸進的な係り受け解析器の性能改善を目的に，人間による言語処理過程を表出した大規模データを構築する．本データの特徴は，新聞記事中の 2,502 文を対象として，文節が文頭から順に 1 つ提示されるたびに，人間が係り受け構造を解析し，その解析結果をタグ付けしている点にある．また，本稿では，構築したデータを用いて簡単な分析を実施し，人間の漸進的係り受け解析能力の一端を明らかにする．

2　漸進的係り受け解析の出力構造

人間の言語処理過程を表出する方法としては，どのような時点でどのような情報をアノテーションするのかによって，様々な方法が考えられる．本研究では，漸進的係り受け解析器の評価や性能改善に向けて，構築したデータを活用することを想定しているため，解析器の出力を模する形で，人間の言語処理過程を表出することとする．そこで本節では，本研究が想定する，音声言語システムのための漸進的係り受け解析器について述べ，その出力構造を解説する．

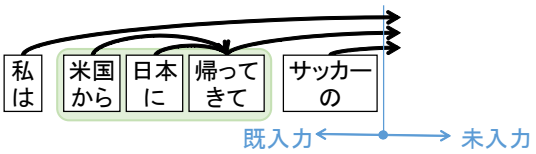

図 1: 既入力のどの文節とも依存関係にないことを明示した係り受け構造 [5]

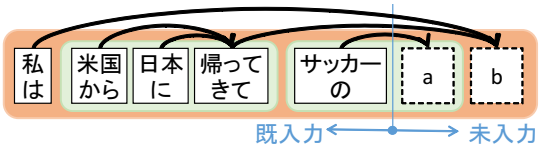

図 2: 未入力文節との依存関係を明示した係り受け構造 [6]

漸進的な係り受け解析は，文の入力途中の段階で文節間の依存関係を同定するものであり，入力音声に追従した出力が求められる音声言語システムにとって有用である．しかし，解析器が提示する結果にどのような情報が含まれるべきかについては，これまでほとんど検討されていない．また，これまでの漸進的な係り受け解析 [1, 2, 3, 4] では，一般的な係り受け解析と同様，係り（係り元）と受け（係り先）の組を同定することとして問題を捉え，その同定処理の実現方法に焦点を当てている．そこでは，入力が進むごとに解析の途中結果を更新し，係りと受けの組を同定するたびに，その組を解析結果として出力することを想定している．そのため，係りと受けのどちらか一方だけでなく双方が入力されるまで，その係り受け関係を出力できないという問題がある

この問題に対し，著者らはこれまでに，音声言語システムによって必要とされる係り受け解析の新たな出力方式を提案している [5, 6]．文献 [5] では，文の入力途中において，係り先が既に入力されている文節については係り先文節との依存関係を，また，係り先が入力されていない文節については既入力のどの文節とも依存関係にないことを，係り受け構造が明示することを提案している．その係り受け構造の例を図 1 に示す．この例は，文「私は米国から日本に帰ってきてサッカーのワールドカップを見ました」のうち，文節「サッカーの」まで入力された段階で出力する係り受け構造を示している．係り先が未だ入力されていない

という情報を早い段階で提示できれば，係り先が入力される前に，入力済み文節列内の構文的なまとまり[1]の成否を捉えることができる．例えば，図 1 からは，文節列「米国から日本に帰ってきて」が構文的にまとまっていて，文節「私は」や「サッカーの」とは同じまとまりを形成しないことがわかり，上位層のアプリケーションがその情報を利用できることになる．なお実際に，任意の文節列が構文的にまとまっているか否かという情報は，同時通訳における訳出タイミング [10] や，字幕生成での読みやすい改行位置 [11, 12] などの決定において，重要な手掛かりとして利用されている．

さらに文献 [6] では，文献 [5] の係り受け構造を拡張し，係り先が未入力の文節が複数ある場合は，それらの係り先が同一か否かを，係り受け構造が明示することを提案している．係り先が未入力である文節が複数存在したとき，それぞれの文節は別々の未入力文節に係ることもあれば，同一であることもあり，それらを同定できれば，構文的なまとまりをより詳細に捉えることが可能となる．図 2 に，図 1 と同様の入力において出力される文献 [6] の係り受け構造を示す．文節「私は」と「帰ってきて」の係り先は同一（未入力文節 b）であり，「サッカー」の係り先（未入力文節 a）とは異なることを明示している．このような係り受け構造を同定することができれば，「サッカーの」から未入力文節 a までの文節列は，構文的なまとまりを構成し，別の構文的なまとまり「米国から日本に帰ってきて」と共に，「私は」から未入力文節 b までの文節列からなるまとまりの中に埋め込まれていることが分かる

本研究では，上述した係り受け構造を解析するだけでなく，未入力の係り先文節に対して，その具体的な文字列も予測し出力する係り受け解析器を開発することを予定している．この解析器では，例えば，図 2 の場合，未入力文節 a が「ワールドカップを」，未入力部文節 b が「見ました」であることもそれぞれ予測し出力することになる．

本研究におけるデータ構築では，文節が入力されるごとに，図 2 の形式の係り受け構造の解析と，その構造の中の未入力係り先文節の文字列の予測とを作業者が行い，その結果をタグ付けすることとする．

3 人間による漸進的係り受け解析のアノテーション

本節では，人間による漸進的な係り受け解析結果をタグ付けした大規模データの構築について述べる．

本データは，既存コーパスに含まれる文の文節列が文頭から順に 1 文節ずつ提示されるごとに，それまでに提示された文節列に対して，図 2 の形式の係り受け構造と，その構造内の未入力係り先文節の文字列とを，作業者がタグ付けすることにより構築する．以下では，アノテーションの対象とする既存コーパスについて述べ，その後，1 文に対するアノテーション手順を詳述し，全データに対するアノテーション作業の実施について述べる．

[1]ある文節列が構文的にまとまっているとは，その文節列内で係り受け構造が閉じている，すなわち，その文節列外の文節に係る文節が，その文節列末の文節以外には存在しないことを意味する．

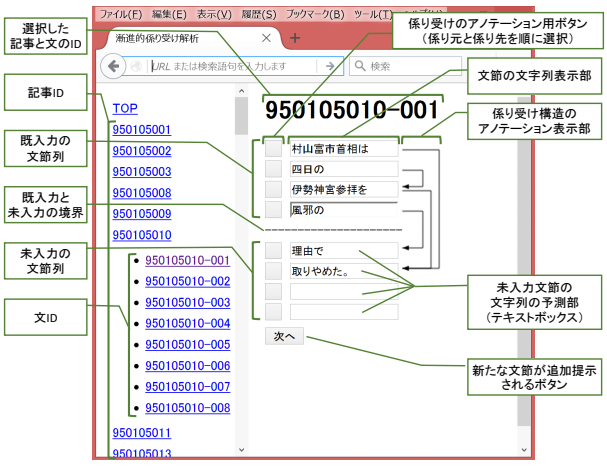

図 3: Web インタフェース画面

3.1 アノテーションの対象データ

漸進的解析器の改善に構築したデータを活用することを念頭におくと，発話されるごとに徐々に情報が提示される音声言語を収集したコーパスをアノテーションの対象データとして利用することが考えられる．しかし，話し言葉には，非文法的な言語現象が頻出するなどの原因で，たとえ発話全体が提示された状態であっても，その係り受け構造を同定することがそもそも難しいという問題が存在する．本研究では，人間の言語処理過程の表出に問題の焦点を絞るため，読みやすく整形されていると考えられる新聞記事を利用することとした．

また，本データ構築では作業者が今後入力される文字列の予測も行う．その予測精度は，対象となる文の内容を理解するための知識や文体への馴染みといった個人差に大きく影響を受ける可能性がある．一方，新聞記事は，一般常識といえる内容が，多くの人にとって慣れ親しんでいる文体で一貫性をもって記述されており，これらの個人差を比較的排除できると考えられるため，本データのアノテーション対象として適していると判断した．

具体的には，京都大学テキストコーパス Version 4.0[13] に含まれる毎日新聞記事の 2,502 文（1 月 5 日及び 1 月 9 日の一般記事）をアノテーションの対象データとした．このコーパスには，形態素情報や係り受け構造，格関係，照応・省略関係，共参照の情報が付与されており，これらの情報はアノテーション結果の分析に利用することができる．

3.2 1 文に対するアノテーション手順

作業者は，1 文が文頭から順に 1 文節ずつ提示されるという状況において，新たな文節が提示されるごとに，各時点で提示されている文節列に対して，図 2 の形式の係り受け構造の解析と，その構造内の未入力係り先文節の文字列の入力予測を行い，その結果をアノテーションする．

具体的なアノテーション作業は，図 3 の Web インタフェースを用いて，以下の手順で実施する．

1. アノテーションの対象文の選択
   Web インタフェースの左フレーム上で，新聞記事 ID を上から順に選択すると，その記事に含まれる文 ID の一覧が展開され表示されるので，上から順に文 ID を選択する．人間は文脈知識を利用して次の入力を予測していると一般に考えられるため，作業者は，記事ごとに，その記事内の文を順番にアノテーションすることとした．
2. 選択した文に対するアノテーション
   上記で 1 文を選択すると，Web インタフェースの右フレームに，その文が文頭から順に 1 文節ずつ提示されるので，新たな文節が提示されるごとに，それまでに提示された文節列に対して，以下の 2 つの作業を時間制限を設けず実施する．
   (a) 図 2 の形式の係り受け構造を付与する．具体的には，まず，係り元文節の「係り受けのアノテーション用ボタン」をクリックし，その後，その係り先と考えられる文節のボタンをクリックすることにより，その係り元文節から係り先文節への矢印が「係り受けのアノテーション表示部」に表示される．この操作を「既入力の文節列」内の全文節に対して繰り返す．その際,「未入力の文節列」内の文節に係ることも許す．図 3 では，文節「村山富市首相は」と「伊勢神宮参拝を」が同一の未入力文節に係り，それとは異なる未入力文節に文節「風邪の」が係るとしてアノテーションされている．
   (b) 上記で付与した係り受け構造において，係り先文節として未入力文節がある場合，それら未入力文節の文字列を可能な限り予測し，その文字列を該当のテキストボックスに入力する．図 3 では，2 つの未入力の係り先文節に対して，それぞれ「理由で」と「取りやめた。」の文字列が予測されている．

   上記 2 つの作業が終わると,「次へ」ボタンをクリックする．これまでに提示された文節列に加えて新たな文節が提示されるので，再度，上述の作業 (a) と (b) を実施する．これを文末文節が提示されるまで繰り返す．
3. アノテーション結果と正解の確認
   上記で 1 文に対するアノテーション作業が完了した後に，そのアノテーション結果に対する得点（京大テキストコーパスの情報を正解とみなしたときの一致度）と正解の係り受け構造が表示されるので，作業者は自分のアノテーション結果と正解とを比較し，係り受けのタグ付け基準や文体などの確認を行う．なお，得点表示は，作業者のモチベーションを保つことを目的に実施した．

3.3 アノテーション作業の実施

作業者 1 名が約 5 か月間かけて，前節で述べた 1 文に対するアノテーション作業を繰り返すことにより，対象データ全 2,502 文（総文節数 25,336 文節，1 文あたりの平均文節数 10.13）に対するアノテーションを実施した．なお，作業者には事前に，京大テキストコーパスのタグ付け基準を熟読すること，ならびに，文字列の入力予測については思いつく範囲でできる限り実施することを指示した．

4 人間の漸進的言語処理能力の分析

構築したデータに対する分析に基づいて，人間の漸進的言語処理に関する能力の一端を明らかにする．以下では，係り受け構造の解析と文字列の入力予測の二つに分けて分析する．

4.1 人間の係り受け解析能力

構築したデータを用いて，作業者の係り受け解析能力を以下の 3 つの観点で評価した．

- **文単位解析**: 文末文節が提示された後に 1 文全体に対して作業者が付与した係り受け構造のみを対象として，正解（京大テキストコーパス上）の係り受け構造とどの程度一致しているかを評価した．これは，文全体の文節列が一度に入力され，その係り受け構造を出力するという一般的な係り受け解析を人間が施した結果の精度に相当する．なお，Web インタフェースにおいて，新たに提示された文節が文末文節であるか否かは，作業者に分かるようになっている．
- **文節単位解析 A**: 作業者が付与した係り受け構造を図 1 に示す形式の係り受け構造とみなして評価した．すなわち，係り先が未入力である文節については，その係り先が他の文節の係り先と同一か否かという情報は無視し，係り先が未入力であるか否かが正解と一致しているか否かを評価する．係り先が既入力である場合は，その係り先が正解の係り先文節と一致しているか否かを評価する．なお，新たな文節が提示されるごとに付与された係り受け構造を毎回評価するため，例えば，文頭文節を係り元とする係り受け関係は，1 文の文節数だけ複数回評価されることになる．
- **文節単位解析 B**: 作業者が付与した係り受け構造を図 2 に示す形式の係り受け構造とみなして評価した．作業者が付与した係り受け構造は，係り先が未入力である文節について，その係り先が他の文節と同一か否かという情報を持っているものの，その未入力文節の文節位置を具体的に決めるわけではないため，正解と一致するかを単純には判定できない．そのため，アノテーション結果と正解を比較し，一致する係り受け関係の数が最も多くなるように，正解とアノテーション結果の係り先文節を対応付け，一致した係り受け関係の数を評価した．なお，文節単位解析 A と同様に，新たな文節が提示されるごとに付与された係り受け構造を毎回評価した．

各評価結果を表 1 に示す．2 列目は文単位解析と文節単位解析 A，文節単位解析 B の係り受け正解率を示しており，それぞれ文献 [14]，[5]，[6] の定義を用いて測定した．ただし，文単位解析の係り受け正解率は，文末文節を係り元とする係り受け関係も含めて計測した．3 列目は文正解率を示しており，各観点において，

表 1: 係り受け解析の正解率

解析名	係り受け正解率	文正解率
文単位	94.8% (24,027/25,336)	63.9% (1,598/2,502)
文節単位 A	94.4% (170,262/180,301)	54.9% (1,373/2,502)
文節単位 B	91.8% (165,501/180,301)	31.3% (784/2,502)

表 2: 文字列の入力予測精度

	完全一致のみ	部分一致も含む
再現率	5.4% (1,972/36,760)	12.4% (4,998/36,760)
適合率	13.6% (1,972/15,862)	31.5% (4,998/15,862)

1 文に対するの全てのアノテーション結果が正解と完全に一致している文の割合である．

実験結果から，図 2 の形式の係り受け構造を文節が入力されるごとに付与する文節単位解析 B は，他の 2 つの解析と比べ，最も難しいタスクであることが分かる．このことは想像に難くなく，他の解析と比べ文節単位解析 B では最も多くの情報を付与する必要があるためだと考えられる．次に，各観点での評価結果の差分に着目すると，文単位解析と文節単位解析 A の差より，文節単位解析 A と文節単位解析 B の差が大きいことがわかる．これは，人間にとって，係り先が未入力文節であることを単に示すこと（文節単位解析 A）はそれほど難しいタスクではないが，その未入力文節との間の係り受け関係を示さなければならなくなる（文節単位解析 B）と，途端に難しいタスクとなることを示唆している．

4.2 人間の文字列の入力予測能力

本節では，作業者が未入力文節の文字列をどの程度正しく予測できたかを評価した．表 2 に，作業者が予測した文字列の再現率と適合率を示す．再現率は，正解の係り受け構造において未入力の係り先となる文節 36,760 個のうち，作業者が文字列を正しく予測した文節の割合を，適合率は，作業者が何らかの文字列を予測した未入力係り先文節 15,862 個のうち，正しく文字列を予測した文節の割合を，それぞれ意味する．なお，2 列目，3 列目はそれぞれ，完全に文字列一致した場合のみを正しく予測したと判定する場合と，部分的に文字列が一致した場合[2] も正しく予測したと判定する場合の値である．この結果から，未入力の係り先文節の文字列を予測することは，人間にとって難しいタスクであるものの，その一方で，1 割程度の未入力文節についてはその文字列（の一部）を予測できており，全く予測できないわけではないことを示唆している．

5 おわりに

本稿では，人間の言語処理過程における漸進性を表出した大規模データを構築した．具体的には，漸進的係り受け解析器と同一の出力方式により，作業者が文の漸進的解析を行い，その解析結果をタグ付けする形で，新聞記事中の 2,502 文に対するアノテーションを実施した．また，本稿では，構築したデータの簡単な分析を実施し，人間の漸進的係り受け解析能力の一端を明らかにした．

今後は，構築したデータをさらに詳細に分析し．人間の漸進的係り受け解析能力を明らかにするとともに，そこで得られた知見に基づき漸進的係り受け解析器の改善を進める予定である．

謝辞 本研究は一部，科研費若手研究 (B)（No. 25730134）及び科研費基盤研究 (B)（No. 26280082）により実施した．

[2] 形態素解析を施し，形態素の出現形が一致する組が 1 個でも見つかれば部分一致とみなした．

参考文献

[1] R. Johansson and P. Nugues. Incremental dependency parsing using online learning. In *Proc. EMNLP-CoNLL2007*, pp. 1134–1138, 2007.

[2] J. Nivre. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, Vol. 34, No. 4, pp. 513–553, 2008.

[3] 加藤, 松原, 外山, 稲垣. 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, Vol. J86-DII, No. 1, pp. 86–97, 2003.

[4] 大野, 松原, 柏岡, 加藤, 稲垣. 節境界に基づく独話の漸進的係り受け解析. 信学論, Vol. J90-D, No. 2, pp. 556–566, 2007.

[5] 大野, 松原. 文節間の依存・非依存を同定する漸進的係り受け解析. 信学論, Vol. J98-D, No. 4, pp. 709–718, 2016.

[6] 村田, 大野, 松原. 未入力文節との構文的関係を考慮する漸進的な係り受け解析. 言語処理学会第 20 回年次大会発表論文集, pp. 193–196, 2014.

[7] G. T.M. Altmann and M. J. Steedman. Interaction with context during human sentence processing. *Cognition*, Vol. 30, pp. 191–238, 1988.

[8] E. P. Stabler. The finite connectivity of linguistic structure. In C. Clifton, L. Frazier, and K. Rayner, editors, *Perspectives on Sentence Processing*, pp. 303–336. Lawrence Erlbaum, 1994.

[9] P. Sturt and M. W. Crocker. Monotonic syntactic proceesing: A cross-linguistic study of attachement and reanalysis. *Language and Cognitive Processes*, Vol. 11, No. 5, pp. 449–494, 1996.

[10] 笠, 松原, 稲垣. 英日同時翻訳のための依存構造に基づく訳文生成手法. 信学論, Vol. J92-D, No. 6, pp. 921–933, 2009.

[11] 村田, 大野, 松原. 読みやすい字幕生成のための講演テキストへの改行挿入. 信学論, Vol. J92-D, No. 9, pp. 1621–1631, 2009.

[12] 大野, 村田, 松原. 講演のリアルタイム字幕生成のための逐次的な改行挿入. 電学論, Vol. 133-C, No. 2, pp. 418–426, 2013.

[13] 黒橋, 長尾. 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.

[14] 内元, 関根, 井佐原. 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.