『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーション

浅原 正幸 †　　松本 裕治 ‡
国立国語研究所　コーパス開発センター †
奈良先端科学技術大学院大学　情報科学研究科 ‡
masayu-a -at- ninjal.ac.jp, matsu -at- is.naist.jp

1　はじめに

本発表では『現代日本語書き言葉均衡コーパス』(BCCWJ) に対する係り受け・並列構造アノテーションの現況について報告する。我々は 2006-2010 年の特定領域研究「日本語コーパス」ツール班に所属し、係り受け解析器や係り受けアノテーション支援環境 ChaKi.NET [1]を開発し、実際に係り受けアノテーションに従事してきた。係り受けアノテーション基準として、図 1 のように[2]係り受け構造と並列・同格構造を分離する基準を採用し、対象は BCCWJ のコアデータ相当部分 (約 130 万形態素) として作業を続けている。以下では、工程の概要と進捗について報告し、現在までの問題点と解決の指針について示す。

2　アノテーション作業の進捗

2.1　全行程の概略

以下、2013 年現在の工程の概略を示す。2011 年時点での工程の詳細については文献 [6] を参照されたい。

まず、上流工程として、コーパスのサンプリングとテキスト化・位取り記数法の変更・短単位形態論情報付与・文単位認定・長単位形態論情報付与がある。

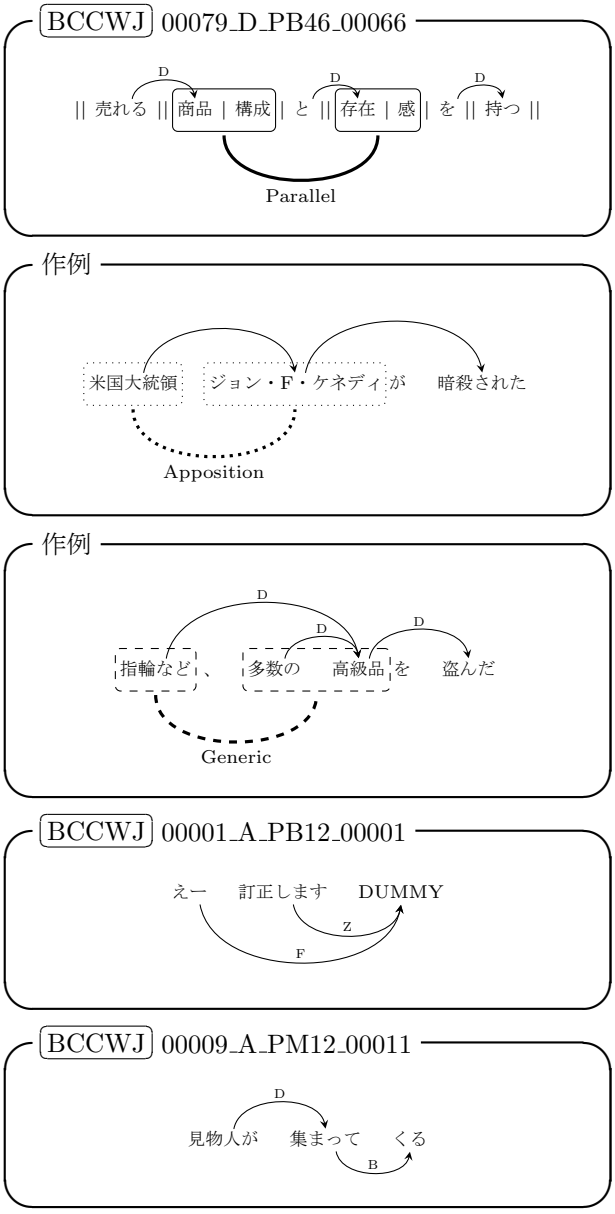

図 1: BCCWJ 係り受け・並列構造アノテーション例

[1] sourceforge.jp/projects/chaki/releases/

[2] 18 ケタの記号は、最初 5 ケタがアノテーション優先順位、7 ケタ目のアルファベットがアノテーション優先部分集合記号、9 ケタ目以降がサンプル ID を表す。また、図中 || が文節境界、| が短単位形態素境界、例文上のラベル “D” 付矢印が係り受け関係、例文下のラベル “F” 付矢印が係り先なしを表現する関係、例文下のラベル “Z” 付矢印が文末要素を表現する関係、例文下のラベル “B” 付矢印が文節単位の修正のための文節の連結を表現する関係、角丸四角と例文下のラベル “Parallel” 付曲線が並列構造範囲とその対応関係、点線角丸四角と例文下のラベル “Apposition” 付点線曲線が同格構造範囲とその対応関係、破線角丸四角と例文下のラベル “Generic” 付破線曲線が具体例-総称間同格構造範囲とその対応関係を示す。 “DUMMY” は文外に係ることを表現するための要素。文節境界と短単位形態素境界は範囲指定が不要な場合は省略する。

並列・同格範囲アノテーションは短単位形態論情報付与と文単位認定までを前提としており、係り受けアノテーションは長単位形態論情報により確定する文節境界付与を前提としている。

BCCWJ モニター公開版で短単位形態論情報が付与されたサンプルデータから、人手で並列・同格範囲アノテーションを行う。上流工程の文節境界付与が終わり次第、文節係り受けを係り受け解析器により付与し、並列・同格範囲アノテーションとの重ね合わせを行う。係り受け一次アノテーション作業においてはこの重ね合わせたデータを人手修正で行う。

係り受け一次アノテーション作業に用いたデータはモニター公開版を底本としている。我々の作業とは別に、上流工程において随時変更が行われるために、最新版との齟齬が発生する。係り受け一次アノテーション作業が完了し、正式版が般公開された時点で、底本の齟齬の解消作業を行う。

係り受け二次アノテーション作業は、前工程で基準が未定義である部分や、その後の議論で基準を見直した部分について修正作業を行う。また係り受け構造を表現するために問題となっている部分（国語研文節単位や自動付与された BCCWJ の文単位）の吸収するための作業を行う。

係り受け二次アノテーション作業終了後、英単語部分の英語としての係り受けアノテーション、文接続詞係り先を決定するために必要な節のスコープ認定、格表示誤りの情報付与を計画している。

以下各工程について解説する

2.2 並列・同格範囲アノテーション作業

アノテーション基準として、係り受け構造と並列・同格範囲構造を分離し、2008 年より並列・同格範囲アノテーション作業を開始した。当時のアノテーション基準については文献 [5] を参照されたい。

並列・同格構造は構成する要素を国語研短単位に切り出し、対応する要素をリンクづけすることにより行う。図 1 の一つ目の例が並列構造の例で、二つ目と三つ目の例が同格構造の例である。並列・同格構造は入れ子を許して付与する。本作業時は同格構造を広い意味で取り、二つ目の例（Apposition; 狭義の同格構造）と三つ目の例（Generic; 具体例-総称関係など）を同じラベルで付与していたが、係り受け二次アノテーション作業でこの二つの分化を行う。

作業は訓練された作業者一人により、全データに対して約二年半かけて実施された。アノテーションツールとして Microsoft Excel を利用した。形態論情報を帳票形式で出力したものに BIO/bio ラベル[3]を付与することによりアノテーションを行った。本データは係り受け一次/二次アノテーション作業時に他の作業者が修正を加えるために、作業者が一人とし、一致率などの分析は行っていない。

2.3 係り受け一次アノテーション作業

部分的に国語研長単位形態論情報が付与されはじめた 2009 年より係り受け一次アノテーション作業を開始した。当時のアノテーション基準については文献 [5] を参照されたい。

国語研長単位形態論情報から決まる文節単位に、京都大学テキストコーパス (以下 KC) 基準に基づく係り受けアノテーションを付与するために、CaboCha のモデルを再構成したもの[4]で自動解析を行った。

作業は訓練された作業者十五人により、全データに対して約二年かけて実施された。主に KC と BCCWJ での基準の差異や、山本らの解析器の誤りの類型化 [7] に注意して作業を行った。作業者採用時には、従前より長期雇用している作業者を除いて、作業前に三つのテスト[5]を行い、採用の基準とした。

作業内容は並列範囲修正、同格範囲修正、係り受け関係修正の三つの作業である。アノテーションツールとして ChaKi.NET[6] を利用した。

新規採用した五人について、基準の統制のために二種類の方法を試みた。一つ目は個別に同じファイルにアノテーションを行い一致度を評価する方法である。二つ目はペアプログラミング的な手法に基づいて二人で画面を共有しながらアノテーションを行う方法である。一つ目の方法で評価した一致度について表 1 に示す。極端に一致度が低い作業者が一人いたために作業者から排除した。二つ目の方法では作業者に負担がかかる一方、よりアノテーション修正個所が増える傾向が見られた。

また一次アノテーション終了後、並列構造により制約づけによる係り受けアノテーションの誤り検出を行

[3] “B” が並列構造の構成要素の先頭要素、“I” が並列構造の構成要素の先頭要素、“O” が並列構造の構成要素間の要素、“b” が同格構造の構成要素の先頭要素、“i” が同格構造の構成要素の先頭要素、“o” が同格構造の構成要素間の要素。

[4] KC を MeCab/UniDic で解析したうえで、KC の文節単位と係り受けアノテーションを重ね合わせて構成したデータにより訓練を行ったもの。

[5] テストは「副詞を三つ示せ」「連体詞を三つ示せ」「格助詞を知っている限り全て示せ」というものであった。格助詞を五つ以上示したものを正解としても、全問正解者はいなかったため、格助詞を五つ以上示したものを正解とし、二問以上正解の方を採用した。

[6] sourceforge.jp/projects/chaki/releases/

表 1: 係り受け一次アノテーションにおける作業者間の一致率：白書データ (%)

	並列範囲 (文単位)	同格範囲 (文単位)	係り受け関係 (文単位)	係り受け関係 (文節単位)
最大値	94.4	96.4	62.5	94.5
平均値	81.2	94.1	58.5	92.2
最小値	75.9	87.5	48.0	89.9

い、検出された部分について人手で修正した。詳しくは文献 [3] を参照されたい。

2.4 上流工程の変更にともなう修正作業

並列・同格構造アノテーション、係り受け一次アノテーションを行っている間でも、上流工程では、表層文字列、形態論情報や文境界情報を変更され、また、公開に際して個人情報などの伏字化作業が行われていた。このため、我々がアノテーションしているデータと 2011 年 12 月一般頒布 DVD 版とでアノテーションの齟齬が生じていた。

この問題に対処するために 2012 年 6-8 月に齟齬の吸収作業を行った。まず、文字列レベルの齟齬を diff により定量的に評価した。文字列レベルの齟齬を含めた、国語研短単位および文節単位の境界情報の多層にわたる齟齬を、本問題を解決するために特化した編集距離に基づく手法に基づいて定量的に評価した。各層で一対一、多対一対応する部分を自動で修正し、一対多、多対多対応する部分を人手で修正した。最後に文単位について、文断片相当 (`@type="quasi"`であるもの) を含む、文構造タグ ⟨sentence⟩ 相当の部分に写像し、係り受け関係が未定義になる部分について人手で修正した。現時点では、この修正作業を施したものを Web 上[7]に公開している。

2.5 係り受け二次アノテーション作業

これまでの作業で作成されたデータについて、様々な方から意見や助言などをいただいた。現在、既存の係り受けアノテーション基準の差異を定性的に分析 [4] しながら、係り受け二次アノテーション作業において、問題点の修正作業を行っている。この修正作業の基本方針として、研究者により意見が割れている部分については、どちらか一方の意見のみを採用するのではなく、係り受け木を双方が求める構造に写像できるような情報をできる限り残す方法をとる。以下に修正作業の概要について説明する。

[7] `github.com/masayu-a/BCCWJ-DepPara`

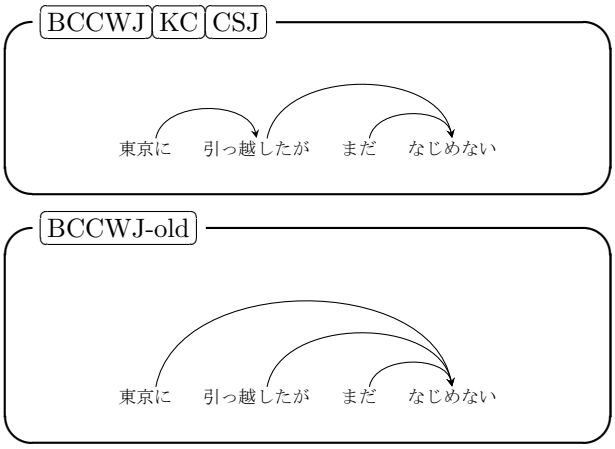

図 2: 複数の格要素と述語間の係り受け

文境界の再修正　現在公開されている一般頒布 DVD 版に収録されているデータの文境界情報は、サンプリング元のデータのレイアウト情報（紙面もしくは HTML タグ）に基づいて、半自動的に生成されたもので、斉一性がない。そこで係り受けアノテーションに問題が波及する部分について人手で修正を行う。具体的には文の入れ子を許した ⟨superSentence⟩ 相当の範囲で文を認定し、内部の文境界 (`@type="quasi"`でない ⟨sentence⟩ の最右要素) については、係り受け関係にラベル “Z” を付与することで文境界であったことを明示する。この文境界の認定基準については、文献 [1] を参照されたい。

文節境界の再修正　国語研文節単位 [2] は、形態論的な分析を目的として斉一性を持たせるべく設計されている。一方、可能性に基づく品詞体系により自立語-非自立語の分化を行わないまま単位設計を行っているために、係り受けを表現するためにそのまま利用するのが難しい。そこで、動詞-非自立可能を含む文節については、左隣接する要素と連結して語彙的な複合動詞をなす場合については、係り受け関係ラベル “B” 相当で一文節をなすことを表現する（図 1 の五つ目の例）。その他、レイアウト上の空白や記号により分断されている語彙的な複合語についても同様に連結作業を行う。我々の文節単位を利用する方はラベル “B” で連結される単位を一文節として、国語研文節単位を利用する方はそのまま利用する。

格要素と複数の述語の関係　一次アノテーション作業基準においては、ある要素が複数の要素に係りうる場合、並列構造に左から係る場合を除いて、最右要素に係ける基準を採用していた [5]。格要素と複数の述語の

関係においては適さない[8]ために、既存のアノテーション基準と同様に、主題や主語[9]以外の格要素[10] については最近接述語に係ける (図 2) ように変更した[11]。

係り先なしの要素　「日本語話し言葉コーパス」(CSJ) の基準においては、一文節をなす感動詞、記号、文接続詞などについて係り先なしの要素を認めている。BCCWJ においても図 1 の四つ目の例のように、係り先なしの要素を認め、ラベル “F” を付与する。厳密に KC 相当のアノテーションに写像する場合には DUMMY ノードに係けられている係り受け関係を文末要素に修正することにより復元する。

同格構造の分化　KC の基準においては広義の同格構造を一つのラベルで表現する一方CSJ の基準においては、同格構造を狭義の同格構造 (CSJでは A ラベル, BCCWJでは Apposition グループ) と具体例-総称間のような同格構造 (CSJでは A2 ラベル,BCCWJでは Generic グループ) とを分化させている。BCCWJ においても図 1 の二つ目と三つ目の例のように分化させることとした。

英単語/古文/漢文/ローマ字文・言い直し　CSJ の基準においては、古文などについてはラベル K を用い、言い直しについてラベル D もしくは S を用いて表現している。BCCWJの基準においては、前者を英単語/漢文/ローマ字文も含めて Foreign セグメントにより範囲指定し、後者を Disfluency セグメントに範囲指定することとした。

3　おわりに

本稿では『現代日本語書き言葉均衡コーパス』に対する係り受け・並列構造アノテーションの現況について報告した。今後、係り受け二次アノテーション作業において行わなかった、英単語などの範囲内のアノテーション、接続詞の係り先を厳密に規定するために必要な節のスコープ認定、CSJ にアノテーションされている格表示誤りなどのアノテーションを行っていく予定である。基準やデータに対する質問・コメントについては第一著者まで。

[8] ニ格、ヲ格相当は近接述語に係けるべきと様々な方からコメントをいただいた。言語処理的には述語項構造解析において文内ゼロ代名詞の認定基準が変更される。認知的には Miyamoto が日本語格要素と複数の述語の関係の人間の文処理傾向を心理言語実験により調査している。理論的には南の節分類毎の項の共有についての議論や X-bar における PRO の扱いなどがある。

[9] X-bar における specifier 相当。

[10] X-bar における complement, adjunct 相当。

[11] 一次アノテーション作業基準の情報は確保したうえで、述語項構造アノテーションを用いることで復元できるようにする

謝辞

本研究を行うに際し、アノテーションを実施された作業者の方、奈良先端大自然言語処理学研究室のメンバー、国語研コーパス開発センターのメンバー、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」のメンバー、筑波大学の Edson T. Miyamoto 氏、グーグルの工藤拓氏ほか、多くの方よりコメントをいただきました。

本研究は文科省科研費特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築 : 21 世紀の日本語研究の基盤整備」、国語研基幹型共同研究プロジェクト「コーパスアノテーションの基礎研究」および国語研「超大規模コーパス構築プロジェクト」によるものです。

参考文献

[1] 小西光, 小山田由紀, 浅原正幸, 柏野和佳子, 前川喜久雄. 『現代日本語書き言葉コーパス』の係り受け関係アノテーションのための文境界の再認定. 第 3 回コーパス日本語学ワークショップ, 2013.

[2] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 原裕. 『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（上下巻）, 2011.

[3] 岩立将和, 浅原正幸, 松本裕治. 並列構造アノテーションの制約を利用した係り受けアノテーション支援. 情報処理学会研究報告 2012-NL-205, pp. 1–7, 2012.

[4] 浅原正幸. 係り受けアノテーション基準の比較. 第 3 回コーパス日本語学ワークショップ, 2013.

[5] 浅原正幸, 岩立将和, 松本裕治. BCCWJ コアデータへの係り受け・並列構造アノテーション —進捗と課題—. 特定領域『日本語コーパス』平成 21 年度公開ワークショップ, 2010.

[6] 浅原正幸, 岩立将和, 松本裕治. BCCWJ コアデータへの係り受け・並列構造アノテーション. 『現代日本語書き言葉コーパス』完成記念講演会予稿集, pp. 71–76, 2011.

[7] 山本悠二, 増山繁. 日本語係り受け解析における誤りの類型化と文構造の曖昧性について. 言語処理学会第 15 回年次大会発表論文集, pp. 789–792, 2009.