『現代日本語書き言葉均衡コーパス』に対する 文節係り受け・並列構造アノテーション

浅原 正幸[†]・松本 裕治^{††}

本稿では『現代日本語書き言葉均衡コーパス』のコアデータに対する文節係り受け・並列構造情報のアノテーションについて述べる。統語構造のアノテーションに対して、文節係り受け情報と並列・同格構造を分離してアノテーションする方法を提案する。さらに節境界を越える係り受け関係について、節の分類に基づきスコープを決めることでよりアノテーションの精緻化を行う。実作業の工程上の問題などにも言及しながら、アノテーション基準を概説する。また、アノテーションデータの基礎統計量について示す。

キーワード：コーパス、アノテーション、係り受け、並列、同格

Bunsetsu-based Dependency Relation and Coordinate Structure Annotation on ‘Balanced Corpus of Contemporary Written Japanese’

MASAYUKI ASAHARA[†] and YUJI MATSUMOTO^{††}

This article presents syntactic annotation for ‘Balanced Corpus of Contemporary Written Japanese’. We propose a syntactic annotation schema wherein the *bunsetsu* dependency and coordinate structure are separated. In addition, we propose an annotation standard to determine the attachments beyond clause boundaries. Our annotation schema and standard have issues associated the hierarchical annotation processes. Furthermore, we present the basic statistics of the annotation data.

Key Words: *Corpus, Annotation, Dependency Relation, Coordination, Apposition*

1 はじめに

文節係り受け解析は情報抽出・機械翻訳などの言語処理の実応用の前処理として用いられている。文節係り受け解析器の構成手法として、規則に基づく手法とともに、アノテーションを正解ラベルとしたコーパスに基づく機械学習に基づく手法が数多く提案されている (Uchimoto, Sekine,

[†] 人間文化研究機構国立国語研究所, National Institute for Japanese Language and linguistics, Japan

^{††} 奈良先端科学技術大学院大学情報科学研究科, Nara Institute of Science and Technology, Graduate School of Information Science, Japan

and Isahara 1999; Kudo and Matsumoto 2002; Sassano 2004; Iwatate, Asahara, and Matsumoto 2008; Yoshinaga and Kitsuregawa 2010, 2014). 文節係り受け情報は、新聞記事(黒橋, 居倉, 坂口 2000)・話し言葉(内元, 丸山, 高梨, 井佐原 2003)・ブログ(橋本, 黒橋, 河原, 新里, 永田 2011)などにアノテーションされているが、使用域(register)横断的にアノテーションされたデータは存在しない。

我々は『現代日本語書き言葉均衡コーパス』(以下 BCCWJ)に対する文節係り受け・並列構造アノテーションを整備した。対象は BCCWJ のコアデータで新聞・書籍・雑誌・白書・ウェブデータ(Yahoo!知恵袋・Yahoo!ブログ)の6種類からの使用域からなる。これらに対して係り受け・並列構造を付与したものを BCCWJ-DepPara として公開した。

本稿では、アノテーション作業における既存の基準上と工程上の問題について議論し、どのように問題を解決したかについて解説する。既存の基準上の問題については、主に二つの問題を扱う。一つ目は、並列構造・同格構造の問題である。係り受け構造と並列構造は親和性が悪い。本研究では、アノテーションの抽象化としてセグメントとそのグループ(同値類)を新たに定義し、係り受け構造と独立して並列構造と同格構造を付与する基準を示し、アノテーションを行った。二つ目は、節間の関係である。我々は、節境界を越える係り受け関係に対する判断基準を示し、アノテーションを行った。工程上の問題においては、文節係り受けアノテーションのために必要な先行工程との関係について述べ、作業順と基準により解決を行ったことを示す。本論文の貢献は以下のとおりである。

- 使用域横断的に 130 万語規模のコーパスにアノテーションを行い、アノテーションデータを公開した。
- 係り受けと並列・同格構造の分離したアノテーション基準を策定した。
- 節境界を越える係り受け関係に対する判断基準を明示した。
- 実アノテーション問題における工程上の問題を示した。

2 節では『現代日本語書き言葉均衡コーパス』の概要について述べる。3 節ではアノテーション作業で扱った問題について紹介する。4 節では先行研究である京都大学テキストコーパスのアノテーション基準(黒橋 他 2000)や日本語話し言葉コーパス(内元 他 2003)のアノテーション基準と対比しながら基準を示す。5 節では基準の各論について示す。6 節ではまとめと今後の課題について述べる。

また、以下では二文節間に係り受け関係を付与することを便宜上「かける」と表現する。

2 『現代日本語書き言葉均衡コーパス』

本節では『現代日本語書き言葉均衡コーパス』の概要(国立国語研究所コーパス開発センター 2015)について述べる。

『現代日本語書き言葉均衡コーパス』は、約1億語からなる現代日本語の均衡コーパスである。このうち約130万語からなるコアデータは、新聞(PN)・書籍(PB)・雑誌(PM)・白書(OW)・Yahoo! 知恵袋(OC)・Yahoo! ブログ(OY)の6種類の使用域からなる。この使用域を表すアルファベット2文字は、サンプルの識別子であるサンプルIDの接頭辞として用いられている。2006年に整備がはじまり、2008年以降、毎年サンプル版が内部公開された。2011年にversion 1.0が一般公開され、2015年に修正版version 1.1が公開された。

すべてのサンプルについて、自動解析による短単位(Short Unit Word: SUW)(小椋, 小磯, 富士池, 宮内, 小西, 原 2011a)・長単位(Long Unit Word: LUW)(小椋, 小磯, 富士池, 宮内, 小西, 原 2011b)二つの形態論情報が提供されている。コアデータについては、人手で短単位・長単位形態論情報が全数チェックされ、長単位に基づく文節境界が付与されている。2008年～2010年のサンプル版はコアデータの量が増えるとともに、形態論情報の誤りの修正がなされた。

2011年に公開されたBCCWJ version 1.0の文境界は不完全であることが報告されている(田野村 2014)。これを受けて、非コアデータ全体に対して特定の誤りパターンに基づく文境界修正(小西, 中村, 田中, 間淵, 浅原, 立花, 加藤, 今田, 山口, 前川, 小木曾, 山崎, 丸山 2015)が行われた。このデータはBCCWJ version 1.1として、2015年に公開された。なお、version 1.0からversion 1.1への更新時には、文境界にかかわる形態論情報のみが修正された。

コアデータに対しては、さまざまな研究機関がアノテーションを行った¹。できるだけアノテーションが重なるようにするために、アノテーションの優先順位が研究者間で共有されている²。成果物はBCCWJのDVD版所有者のみが利用可能なダウンロードサーバ³で公開されており、本研究の成果物も含まれている。

3 アノテーション作業で扱った問題

本節では、アノテーション作業で扱った問題について示す。

アノテーション基準の策定には、扱う言語表現をどのように記号化するかを決める必要がある。BCCWJに対する文節係り受け・並列構造アノテーションでは、既存の係り受けアノテーション基準で表現できなかった構造を表現することを目標とし、基準の再設計を行った。具体的には並列・同格構造を文節係り受けと分離してアノテーションを行い、新聞記事を対象にしていたときには問題にならなかった、倒置や係り先のない文節の認定を行った。

アノテーションの実作業では、上流工程としてサンプリング・電子化・著作権処理・文境界認定・短単位形態論情報付与・長単位形態論情報付与・文節認定などがあった。これらの作業

¹ https://pj.ninjal.ac.jp/corpus_center/anno/

² <https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER/>

³ <https://bccwj-data.ninjal.ac.jp/mdl/>

と並行して行ったために、工程上に発生する問題を解決しながら行う必要があった。工程上の問題は、上流工程が完了するまで作業が開始できないという作業順の問題と、上流工程で付与されたアノテーションが文節係り受け・並列構造を付与するのに適さないという基準上の問題の二つがあった。

3.1 基準で扱う言語学的な問題

アノテーション作業および基準において、重点的に取り扱った言語学的な問題について示す。一つ目は並列・同格構造である。並列・同格構造は文節係り受け関係との親和性が悪く、特に部分並列 (non-constituent conjuncts) の場合に表現に無理が生じる。以下の例文では、1 文節目から 3 文節目までと 4 文節目から 6 文節目までの二つの構成要素による並列構造をなすが、この並列構造を係り受け関係により自然に表現することは難しい。

部分並列 (non-constituent conjuncts) の例

本を 兄の 太郎に ノートを 弟の 三郎に かしている

本稿ではこの例も含めて、係り受け構造・並列構造・同格構造をどのように表現するか、過去の研究と対照しながら議論する。

我々の基準では、文節係り受けとは別に並列や同格構造の構成要素の範囲を形態素単位（国語研短単位）に同定し、構成要素間に無向リンク（同値関係）をはることでグループ（同値類）としてアノテーションを行う。並列や同格のアノテーションを分離したうで、文節係り受け関係は本来係かるべき要素に係かるようにする。

この結果、非交差条件を満たさない構造 (non-projectivity) を許す。さらに並列構造の構成要素二つ以上のヲ格を含む構成要素から一つの述語に係かることを許すため、二重ヲ格制約を満たさない。

テ形や連用中止の連続などの述語並列については、並列構造を認定しない方針をとる。これは、日本語の節間の並列関係を従属関係と識別するのが困難であるために作業から除外する。そのかわり、節に対する格要素の係り受け関係の曖昧性については、南の節分類 (南 1974) を参考に、(野田 1985) の議論を参考にしながら厳密に行う。

Web 上のテキストには倒置であったり、係り先のない文節が含まれたりする。これらについては基準において主辞後置制約 (strictly head final) を緩和し右から左に係ることを許すほか、係り先のない文節にラベルを付与したうで連結根付き木 (single rooted) 制約も緩和してアノテーションを行う⁴。

現実のテキストには外国語や顔文字など、日本語の係り受け関係が認定できないものも存在

⁴ 正確には単一の文外の要素 DUMMY を設定し、連結根付き木として表現する。

する。これらについては、係り受けアノテーションを放棄するセグメントを導入して対処する。

3.2 作業順で扱う工程上の問題

係り受けの上流工程として、サンプリング・電子化・著作権処理・数値表現正規化・文境界認定・短単位形態論情報付与・長単位形態論情報付与がある。

サンプリングはコーパスに収録するテキストを選択する工程である。電子化は紙媒体のテキストを機械可読な形式に入力する工程である。著作権処理はコーパスに採録して公開してよいかを権利者に確認する工程である。数値表現正規化は、アラビア数字や漢数字など多様な数値表現を位取り記数法を用いて正規化する工程 (NumTrans) である。文境界認定は日本語文として適切な文単位を認定する工程である。BCCWJ において形態論情報付与は、二重形態素解析と呼ばれ、短単位・長単位の二つの単位を付与する。短単位形態論情報付与 (小椋他 2011a) は、形態素解析器 MeCab⁵ と辞書 UniDic⁶ による解析を行い、人手で修正を行う工程である。長単位形態論情報付与 (小椋他 2011b) は、長単位解析器 Comainu (小澤, 内元, 伝 2014)⁷ により文節境界と長単位形態論情報を推定したうえで人手で修正を行う工程である。

2008 年の時点で、新聞・書籍・白書・Yahoo!知恵袋のデータに短単位形態論情報が付与されたデータが内部公開された。文節係り受け付与には長単位形態論情報が必要なために、並列構造・同格構造を先行して付与した。その後、雑誌・Yahoo!ブログのデータが得られ次第、並列構造・同格構造を付与した。この作業は基本的に 1 人の作業者により全データに対してアノテーションを行った。

2009 年の時点で、長単位形態論情報が得られ、文節係り受けアノテーションを開始した。文節係り受けアノテーションは既に付与されている並列構造・同格構造アノテーションと、文節係り受け解析器 CaboCha⁸ による自動解析結果を重ね合わせたものを修正することで行った。作業者は十数名により並行で行った。この時点でのデータを BCCWJ-DepPara version 1 とした。

2011 年 12 月に BCCWJ DVD version 1.0 が公開された。BCCWJ-DepPara version 1 と BCCWJ DVD version 1.0 とで、係り受けアノテーション作業と形態論情報修正作業が並行して実施されたために、二つのデータ間で形態論情報に齟齬が発生した。この齟齬を動的計画法を用いて検出し、一意に構造が決められる 1 対多対応もしくは多対 1 対応の齟齬については自動的に修正した。一意に構造が決められない多対多の構造については、人手で確認することで解消を行い、BCCWJ-DepPara version 2 を公開した。

その後、BCCWJ DVD version 1.0 は文境界に問題があることが報告された (田野村 2014)。

⁵ <https://taku910.github.io/mecab/>

⁶ <https://unidic.ninjal.ac.jp/>

⁷ <https://ja.osdn.net/projects/comainu/>

⁸ <https://taku910.github.io/cabocha/>

係り受けアノテーションにおいては根が文末にあるため、文末の定義が重要である。このため、DVD 版とは異なる独自の文境界基準を規定したうえで、文境界の修正と節境界を越える係り受け関係の修正に着手した。文境界修正については基準で扱う工程上の問題として次小節で扱う。

2015 年に BCCWJ DVD version 1.1 が公開され、これには数値表現正規化前のデータ (OT) も含めることになった。表層文字列を数値表現正規化後のデータ (NT) ではなく数値表現正規化前のデータ (OT) に変更するという作業を行い、それに基づいて発生した齟齬の吸収作業を行った。この BCCWJ DVD version 1.1 に対応する、節境界を超える係り受け関係を修正したものを BCCWJ-DepPara version 3 として公開した。並行して非コアデータを含むコーパス全体の文境界修正を行ったために、BCCWJ DVD version 1.1 の文境界 (小西 他 2015) と BCCWJ-DepPara version 3 の文境界 (小西, 小山田, 浅原, 柏野, 前川 2013) との間に文境界の齟齬が発生している。前者の文境界は文境界の右端を認定したもので文の入れ子の最小スパンを文として規定しているが、後者の文境界は文の入れ子を許した最大スパンを文として規定しており、基準にはカギカッコ内の要素の意味に基づいた判断を行っている。このため、節間の関係が複雑になり、今回改めて節境界を超える係り受け関係を整理した。なお、BCCWJ-DepPara version 3 における文境界は、BCCWJ コアデータについて付与されているが、BCCWJ 非コアデータ全体について付与することは作業工数が膨大になるために断念した。

3.3 基準で扱う工程上の問題

工程上、他の基準のデータをもとに作業を行うが、上流工程で付与された情報が必ずしも文節係り受けを付与するために適した基準でない場合がある。この基準の齟齬を解決するため、アノテーションラベルに齟齬の情報を持たせる方策をとる。

一つは文節境界である。国語研文節境界は形態論情報に基づき定義する作業が上流工程で行われているが、係り受け構造を厳密に判定しながら決められたものではない。文節係り受け構造をアノテーションする際、文節が細かく分かれているがためにアノテーションしがたい構造については、「文節をつなぐ」というラベルを定義し、それを付与することにより解決する。

もう一つは文境界である。文境界は係り受け木の根を決める重要な概念であるが、BCCWJ においては version 1.0 の時点では文境界認定を実質的に放棄しており、DVD が頒布された時点で、BCCWJ の整備とは独立に修正を始めた (小西 他 2013)。その際に節境界を作業者が内省により分類しながら、節を越える文節のスコープを再定義した。なお、本研究の文境界修正 (小西 他 2013) と DVD 版の文境界 (小西 他 2015) は独立のものである。

4 既存のアノテーション基準との違いの概要

本節では、係り受け関係ラベルを比較することで先行研究の基準との違いを概観する。『現代日本語書き言葉均衡コーパス』(BCCWJ)の基準については[BCCWJ]と記し、『日本語話し言葉コーパス』(CSJ)の基準については[CSJ]と記し、『京都大学テキストコーパス』(KC)の基準については[KC]と記す。

各基準の違いをまとめたものを表1に示す。

表1 係り受けアノテーション基準の比較

係り受け関係のラベル	[BCCWJ]	(グループ)	[CSJ]	[KC]
通常の係り受け	D	—	ラベルなし	D
並列	D	(Parallel)	P	P
部分並列	D	(Parallel)	I	I
狭義の同格	D	(Apposition)	A	A
広義の同格 (総称)	D	(Generic)	A2	A
倒置	D	—	R	未定義
係り受け関係の交差	D	—	X	未定義 (Aのみ)
倒置により係り先が不定になるもの	F	—	未定義	未定義
境界修正に関するラベル	[BCCWJ]	—	[CSJ]	[KC]
文境界相当	Z	—	未定義	未定義
後続文節と接続	B	—	B+	未定義
その他	[BCCWJ]	(セグメント)	[CSJ]	[KC]
接続詞	F or D	—	C	D
感動詞	F or D	—	E	D
言い直し・言いよどみ	F or D	(Disfluency)	D	未定義
	F or D	(Disfluency)	S:複数文節の言い直し	未定義
	F or D	(Disfluency)	S:複数文節の言い直し:S1	未定義
	F or D	(Disfluency)	S:複数文節の言い直し:E1	未定義
フィラー	F	—	F	未定義
顔文字	F	—	未定義	未定義
非言語音	F	—	ラベルなし	未定義
係り先のない文節	F	—	N	未定義
記号・補助記号	F	—	未定義	未定義
URL・空白	F	—	未定義	未定義
外国語・ローマ字文・漢文	D	(Foreign)	未定義	未定義
古文	D	(Foreign)	K	未定義
古文	D	(Foreign)	K:S1	未定義
古文	D	(Foreign)	K:E1	未定義
呼びかけ	Z	—	Y	未定義
コメント	[BCCWJ]	(セグメント)	[CSJ]	[KC]
	未定義	—	S:格表示誤り「が」	未定義
	未定義	—	S:格表示誤り「を」	未定義
	未定義	—	S:格表示誤り「に」	未定義

まず、本論文の主貢献である並列構造関連の違いについて簡単に説明する。中俣は「並列関係とは元来範列的關係であったものを言語の線条性の要請に従って統合的關係に変換したものである。そのために非現実性を帯びる」と述べている(中俣 2015)。いわゆる部分並列(non-constituent conjuncts)のような統語木の部分木を満たさないものが範列的關係であった場合に、並列構造を木構造として表現する際に無理が生じる。**[KC]**においては、並列構造に対してラベル P、同格構造についてラベル A を準備し、係り受け関係にラベルを付与することにより表現する。部分並列についてはラベル I を付与した上で、係り受け関係が交差しないよう並列構造の構成要素の最右要素に係り先を決める。**[CSJ]**においては、**[KC]**の基準をもとに同格構造を狭義の同格(ラベル A)と総称や数値表現などの広義の同格(ラベル A2)を区別し、係り受け関係にラベルを付与することにより表現する。**[BCCWJ]**においては、並列構造・同格構造のような範列的關係を、係り受け関係で表現することを放棄する。並列構造・同格構造はその構成要素の範囲(セグメント)とその同値類(グループ)を定義することで表現する。セグメント・グループに対するラベルとして Parallel(並列構造)・Apposition(狭義の同格)・Generic(総称など)を定義する。なお、以下の例では Parallel を実線角丸四角で、Apposition を点線角丸四角で、Generic を破線角丸四角で囲み、それぞれの隣接構成素をラベル矢印のない太いスプライン曲線(Parallel は実線・Apposition は点線・Generic は破線)で結ぶことにより表現する。

非交差制約(projectivity)や主辞後置制約(strictly head final)を違反する場合のラベルについて示す。**[KC]**においては、対象が新聞記事であるため、この二つの制約は常に満たすという立場でアノテーションを行っており、特別なラベルを設定していない。**[CSJ]**においては、交差する関係にはラベル X を付与し、倒置などで右から左に係る関係には左から右に係る関係をラベル R として付与する。**[BCCWJ]**においては、交差する関係および倒置などで右から左に係る関係を許すが、それらは正しい係り先に係る関係を導入することにより表現し、ラベルとしては通常の係り受け関係と同じラベル D を用いる。

次に、上流工程のアノテーションを修正するラベルについて示す。**[KC]**においては、すべてのアノテーションが一機関で作成されているために工程間で発生する齟齬を吸収するという機構は基準に組み込まれていない。**[CSJ]**においては、ポーズが入ったがために書き起こしレベルで文節境界が入ったが、統語的には文節境界が入るべきではないという場合がある。その場合には、後続文節と接続するためのラベル B+ を定義する。**[BCCWJ]**においては、文節境界アノテーションと係り受けアノテーションが別機関で作成されており、これらの間の文節境界の齟齬を吸収するためのラベル B を定義する。また、**[BCCWJ]**には文の入れ子を許した文境界単位を認定するため、文中に文境界が出現する場合がある。その場合にはラベル Z を定義する。

係り受けアノテーションを放棄するラベルについて示す。

[KC]においては、対象が新聞記事であるため、基本的には係り受けアノテーションを放棄するという方策はとられていない。

一方, [CSJ]においては, 係り受け関係が不定なものとして, 接続詞 (ラベル C)・感動詞 (ラベル E)・言い直し (ラベル D⁹ (or S1:複数文節の言い直し))・フィラー (ラベル F)・係り先のない文節 (ラベル N)・古典 (ラベル K (or K-S1, K-E1: 複数文節の場合))・呼びかけ (ラベル Y) などにラベルを定義している.

[BCCWJ]においては, 接続詞・感動詞に対して, 係り受け関係が認定できるものはラベル D を付与し, 認定できないものはラベル F を付与し係り先を不定とする (根にかける). 言い直しに対して, 言い直す前のものから言い直した後のものへの係り受けをラベル D で付与するが, 言い直した後のものが存在しない場合にはラベル F を付与し, 根にかける. また, 言い直す前の範囲に対してセグメント Disfluency を規定する. フィラー・顔文字・非言語音・係り先のない文節・記号・補助記号¹⁰・URL・空白など係り先が不定になるものに対してラベル F を付与し, 根にかける. 外国語・ローマ字文・漢文・古文などの係り受けを放棄する範囲については Foreign のセグメントを定義し, セグメント内の係り受け関係は基本的には右連接要素にかける. 呼びかけや文末要素であって係り受けが決まらないものに対しては, ラベル Z を定義するとともに, 根にかける¹¹. なお, 以下の例では Disfluency や Foreign を破線角丸四角で囲むことにより表現する.

その他, [CSJ]特有のラベルとして, 話し手の文法的誤りを示すラベル S:格表示誤りがある. [BCCWJ]においては, 文法的誤りを示すラベルは認定していない.

5 各論

以下では, 本論文における係り受け・並列構造アノテーション基準を既存のアノテーション基準と比較しながら用例を用いて示す.

[BCCWJ]の右に付与されている XXXXX_Y_ZZZZ_ZZZZZ は用例が出現したファイルを示す. XXXXX が作業の優先順位¹², Y が優先順位を表す部分集合名¹³, ZZZZ_ZZZZZ が DVD に収録されている元データのファイル名である. 明示されていないものは, 各コーパスのマニュアルの用例などに基づいた作例である.

⁹ ここでラベル D とあるのは [CSJ]におけるものであり, [BCCWJ]における通常の係り受けを含むラベル D と異なり, 通常の係り受けを含まない.

¹⁰ アイテマイズの記号など, 単独で文節をなすもの.

¹¹ 文末相当について, 引用などで係り先が明らかな場合には正しい係り先にかける.

¹² BCCWJ コアデータのアノテーションにおいては, 研究者間でアノテーションを付与する優先順位を共有し, 優先順位が高いサンプルにできるだけアノテーションが重なるようにしている.

¹³ 先に述べた優先順位は使用域 (register) 毎に 2-5 の部分集合 (A,B,C,D,E) に分割され, 各部分集合内では代表性を有する比率でサンプリングされている.

5.1 係り受け関係

5.1.1 係り受け関係の表現法

以下では係り受け関係を次のような図で示す.

| は国語研短単位形態素境界 (**BCCWJ****CSJ**) もしくは JUMAN 品詞体系形態素境界 (**KC**) を表す. 係り受け関係の説明に形態素境界が問題にならない場合には省略する.

|| は国語研文節境界 (**BCCWJ****CSJ**) もしくは京都大学テキストコーパス文節境界 (**KC**) を表す. 係り受け関係の説明に文節境界が問題にならない場合には省略する.

係り受け関係は文節列上に矢印で表現する. 倒置を除いて格要素 → 述語, 修飾語 → 被修飾語などの方向に付与する.

5.1.2 倒置

KCの基準においては, 主辞後置 (strictly head final) の原則から常に左から右に係る. **BCCWJ****CSJ**の基準においては, 右から左に係ることを許す.

CSJでは右から左に係ることをラベル R を用いて示すが, **BCCWJ**においては特に明示しない.

BCCWJにおいて, 最初の「何だろう」は係り先なしになるが, アノテーションツール上では末尾の文外の根ノード (以下 ROOT と呼ぶ) にかけることにより表現する. 倒置箇所は通常の係り受けとしてラベル D を用いる. 倒置などにより係り先が不定になる箇所には ROOT にかけてラベル F を用いる.

5.1.3 交差

【KC】の基準においては、非交差制約 (projectivity) の原則から係り受け関係が同格表現以外においては交差することを許さない。

【BCCWJ】【CSJ】の基準においては、係り受け関係が交差することを許す。【CSJ】では係り受け関係が交差することをラベル X を用いて明示するが、【BCCWJ】においては特に明示しない。

【BCCWJ】

【CSJ】

5.1.4 係り先のない文節

【KC】の基準においては、最右文節以外に係り先のない文節を許さない。【BCCWJ】【CSJ】の基準においては、最右文節以外でも係り先のない文節を許す。

【BCCWJ】においては係り先のない文節を ROOT (CaboCha 出力における文節 ID -1) に係るように表現する。係り先のない文節の分類については、後述する係り受け関係ラベル分類に基づいて記述する。

尚、【CSJ】においては係り先のない文節を ラベル N により表現する。

以下の例では「中学校を」が ROOT に係る。文末の「いたんですね」も係り先がなく ROOT に係る。【BCCWJ】において、文末以外の係り先がないノードはラベル F、文末はラベル Z として区別する。

BCCWJ

CSJ

5.1.5 文節境界修正

文節境界を修正する記述を係り受け関係ラベル **BCCWJ** においてラベル B, **CSJ** においてラベル B+ を用いて表現する。係り受けアノテーションにそぐわない文節単位のみを、上のような係り受け関係を用いて表現する。このような例は、**BCCWJ** においては複合動詞に、**CSJ** においては発話時にポーズが入った場合に出現する。

BCCWJ

CSJ

5.2 並列構造・同格構造とアノテーションを放棄する範囲

本小節では[BCCWJ]における並列構造・同格構造とアノテーションを放棄する範囲について示す。これらは[KC][CSJ]においては定義されていないが、係り受け関係のラベルにより定義されているものである。

5.2.1 セグメントとグループの導入

[BCCWJ]において、係り受け関係とは別に、文に含まれる並列構造・同格構造をアノテーションする。並列構造もしくは同格構造の構成要素に対しては、国語研短単位形態素境界に基づいてセグメントと呼ばれる範囲を付与し、複数のセグメントをグループ化することによりアノテーションを行う。

5.2.2 並列構造の表現法

並列構造のセグメントとグループを Parallel と呼ぶ。並列構造を以下のように表現する。並列構造の構成要素は文節単位ではなく形態素単位でセグメントにより範囲指定する。図中、実線の角丸四角を用いて Parallel を示す。

[BCCWJ] 00079_D_PB46_00066

並列構造は形態素単位・文節単位・複数文節単位・節単位など様々な単位に規定できる。基本的には入れ子を許して全ての単位に対して付与する。

BCCWJ 00006_A_OW6X_00009

名詞句については，対応する名詞句をセグメント Parallel で切り出し，グループ化する．係り受け関係は通常の係り受けと同じラベル D を付与する．

BCCWJ

尚，CSJ KCにおいては，ラベル P を付与する．

CSJ KC

BCCWJにおいては，形態素単位・文節単位・複数文節単位・節単位など様々な単位に規定できる並列構造を，基本的には入れ子を許して全てのレベルに対してアノテーションする．しかしながら述語並列については並列かどうかの判定が困難なため，全て並列とみなさず，通常の係り受けとして定義する．

CSJ KCでは一部の述語並列について，並列構造を認定しラベル P を付与しているが，BCCWJでは通常の係り受けとしてラベル D を付与する．

BCCWJ

CSJ KC

次に部分並列 (non-constituent coordination) について示す. [CSJ][KC]では以下のような構造について, 非交差制約を順守するためにラベル I を付与し, 真の係り先でないものにかかる. [BCCWJ]においては, 範囲を規定したうえで, 通常の係り受け関係として真の係り先にかかる.

BCCWJ

KC CSJ

[BCCWJ]において, 並列句の間に接続詞が出現する場合には, 右隣接する並列句の最右文節にかかる.

BCCWJ 00144_C_PN2c.00010

次に、並列構造の複数の要素に左から係る場合について示す。以下のように「赤い」が「シャツと」と「ジャケットを」の両方に係り受け関係がある。その場合には、最左要素である「シャツと」にかける。並列構造の外側から、並列構造の要素にかける場合には両方が係り先であることを意味する。これは、項を共有するなど、部分文字列を共有するような係り受け関係を対象とする。

BCCWJ

以下のように「赤い」が「パプリカと」に係り、「黒オリーブが」に係らない場合、並列構造の範囲を「赤いパプリカ」と「黒オリーブ」に定義して、「赤い」の係り先を限定する。

BCCWJ

BCCWJにおいて、3つ以上の並列構造の場合、隣接する並列句にかける。3つのセグメントによって並列句のグループ化を行うことによりこれを表現する。並列句の間に接続詞が出現する場合には、右隣接する並列句の最右文節にかける。以下の事例では、「風合い」「風格」「高級感あふれる質感」が並列構造をなす。

BCCWJ 00033_B.PB35_00013

5.2.3 狭義の同格構造

BCCWJにおいては、同格構造についても形態素単位のセグメントとグループで表現する。同格構造のセグメントとグループを Apposition と呼ぶ。一方、KC CSJにおいては、同格構造を文節単位の係り受け関係としてラベル A を用いて表現する。

BCCWJ

KC CSJ

5.2.4 具体例と総称の同格関係・具体例と数詞の同格関係

BCCWJにおいては、Parallel と Apposition の他に Generic という「セグメント」と「グループ」を規定する。Generic は具体例と総称の同格関係や具体例と数詞の同格関係に対して規定する。

BCCWJ

CSJ

KC

5.2.5 アノテーションの放棄

[BCCWJ]においては、Foreign, Disfluency などのセグメントを規定する。Foreign は外国語・ローマ字・漢文・古典などを表し、Disfluency は言い直し・言いよどみを表す。基本的にはセグメント内の係り受けアノテーションを放棄し、範囲内のすべての要素は右隣接要素にかける。

BCCWJ 00004_A_PB22.00002

言い直した後の要素が存在する場合、言い直しのセグメント外の係り先をその要素に認定する。

BCCWJ 00329_A_OC03.00651

5.3 節境界を越える係り受け関係

BCCWJでは節境界を越える係り受け関係について、次に示す基準を明確にして修正を行った。従属節が埋め込まれている場合、埋め込まれている節境界より前の要素について係り受け関係が判定しにくい場合が多い。具体的には、述語項関係が認められるが、節のスコープに入る要素の制限により係り受け関係が認められない場合である。項の省略により主節と従属節（もしくは複数の従属節）で一つの項を共有する場合に、作業者が述語項関係と係り受け関係を混同し、誤った節間の係り受け関係を認定してしまう。そこで、表2に示す南の節分類に基づいて節間の係り受け関係の再認定を行った。南のA類・B類・C類を、その構成要素（表2上部分）と述語的部分の要素（表2左下部分）により判別し、述語的部分以外の節内要素が当該節のスコープに入る(+)か入らない(-)かを定める。具体的には、提示の「は」は、A類・B類のスコープに入らないが、C類のスコープに入る。一方、主語の「が」はA類のスコープに入らないが、B類・C類のスコープに入る。

次の「つつ」は付帯状況を表す節末表現でA類である。以下の二例において、「彼女が」および「彼女は」は、「飲みつつ」と「食べた」の双方の主語になるという述語項構造を持つ。南の分類においては、目的語の「を」はA類のスコープに入るが、主語の「が」や提示の「は」はA類のスコープに入らない。ゆえに、「彼女が」および「彼女は」の係り先は「食べた」になる。

表 2 南 (1974) の従属節分類と内部に現れる要素 (p128-129)

従属節の種類 構成要素	A 類	B 類	C 類	述語の部分以外の節内要素		述語の部分の要素	
				名詞＋格助詞（ガ以外） 状態副詞 程度副詞 A 類従属節 主語（～ガなど） 時の修飾語 場所の修飾語 ジツニ、トニカク、ヤハリの類 評価的意味の修飾語 B 類従属節 提示のことば（～ハなど） オソラク、タブン、マサカの類 C 類従属節	用言 使役形 受身形 受給の形 尊敬の形 丁寧の形 打消の形 過去形 用言＋形式名詞 意志形 推量形		
従属節の種類 構成要素	‡～連用形（形容詞・形容動詞）	(+) (+) (+)	+	名詞＋格助詞（ガ以外） 状態副詞 程度副詞 A 類従属節 主語（～ガなど） 時の修飾語 場所の修飾語 ジツニ、トニカク、ヤハリの類 評価的意味の修飾語 B 類従属節 提示のことば（～ハなど） オソラク、タブン、マサカの類 C 類従属節	用言 使役形 受身形 受給の形 尊敬の形 丁寧の形 打消の形 過去形 用言＋形式名詞 意志形 推量形		
	‡連用形反復	+ + + +	+ + + +				
	†～テ（継起句・付帯状況）	+ + + +	+ + + + + +				
	～ツツ（付帯状況）	+ + + +	+ + + + + +				
	～ナガラ（継続・付帯状況）	+ + + +	+ + + + + +				
	～ナイデ	+ + + + + + + + + +	+ + + + + + (+)				
	～ズ（ズニ）	+ + + + + + + + + +	+ + + + + + (+)				
	‡～連用形（継起節）	+ + + + + + + + + +	+ + + + + +				
	†～テ（継起節）	+ + + + + + + + + +	+ + + + + + (+)				
	～テモ（仮定）	+ + + + + + + + + +	+ + + + + + + +				
	～ナラ（仮定）	+ + + + + + + + + +	+ + + + + + + + + +				
	～タラ（仮定）	+ + + + + + + + + +	+ + + + + + + +				
～バ（仮定）	+ + + + + + + + + +	+ + + + + + + +					
～ノニ（理由：焦点）	+ + + + + + + + + +	+ + + + + + + + + +					
～ノア（理由：焦点）	+ + + + + + + + + +	+ + + + + + + + + +					
～ナガラ（逆接）	+ + + + + + + + + +	+ + + + + + + + +					
～ト（継起節・仮定節）	+ + + + + + + + + +	+ + + + + + + +					
†～テ	+ + + + + + + + + +	+ + + + + + (+)					

+はある要素がその節の中に存在可能であることを示す
 -はある要素がその節の中に存在不可能であることを示す
 () 内のものは、そう認めることに問題があるものを示す
 * の形は述語的部分の要素に対して言語テストを行い4つに分類すること
 * 連用形は述語的部分の要素に対して言語テストを行い4つに分類すること

「が」とツツ節（付帯状況：A類）

「は」とツツ節（付帯状況：A類）

次の「ないで」は逆接を表す節末表現でB類である。以下の二例において、「彼女が」および「彼女は」は、「飲まないで」と「食べた」の双方の主語になるという述語項構造を持つ。南の分類においては、主語の「が」や目的語の「を」はB類のスコープに入るが、提示の「は」はB類のスコープに入らない。ゆえに、「彼女が」の係り先は「飲まないで」になるが、「彼女は」の係り先は「食べた」になる。

「が」とナイド節（逆接：B類）

「は」とナイド節（逆接：B類）

次の「から」は理由を表す節末表現でC類である。以下の二例において、「彼女が」および「彼女は」は、「飲まないから」と「食べた」の双方の主語になるという述語項構造を持つ。南

の分類においては、提示の「は」・主語の「が」・目的語の「を」全てがC類のスコープに入る。ゆえに、「彼女が」および「彼女は」の係り先は「飲まないから」になる。

「が」とカラ節（理由：C類）

「は」とカラ節（理由：C類）

この判定は一人の作業者により、BCCWJ-DepPara version 2 から version 3 への変更の際に全数見直しを行った。

6 BCCWJ-DepPara の基礎統計

本節では、BCCWJ-DepPara の基礎統計量を示す。

表 3 は、各レジスタの短単位数 (SUW)、長単位数 (LUW)、文節数、係り受けラベル数 ‘D’・

表 3 BCCWJ-DepPara の基礎統計（形態素数とラベル数）

レジスタ	サンプル数	SUW 数	LUW 数	文節数	‘D’	‘B’	‘F’	‘Z’	‘EOS’
PN (新聞)	340	308,504	224,140	116,955	96,892 (82.8%)	1,652 (1.4%)	2,017 (1.7%)	16,394 (14.0%)	16,042
PB (書籍)	83	204,050	169,730	84,733	72,340 (85.3%)	1,091 (1.2%)	1,425 (1.7%)	9,877 (11.7%)	9,678
PM (雑誌)	86	202,268	159,883	83,077	67,618 (81.4%)	1,187 (1.4%)	1,629 (2.0%)	12,643 (15.2%)	12,542
OW (白書)	62	197,011	129,646	68,449	59,320 (86.6%)	359 (0.5%)	2,927 (4.3%)	5,843 (8.5%)	5,825
OC (知恵袋)	938	93,932	78,770	36,740	29,753 (81.0%)	323 (0.9%)	428 (1.2%)	6,236 (17.0%)	6,110
OY (ブログ)	471	92,746	75,242	38,576	29,650 (78.9%)	337 (0.9%)	1,501 (3.9%)	7,088 (18.4%)	7,059

割合 (%) は、ラベル数 {‘D’, ‘B’, ‘F’, and ‘Z’} / 総文節数。

表 4 BCCWJ-DepPara の基礎統計 (並列構造と同格構造)

レジスタ	Parallel			Apposition			General		
	seg	grp	seg/grp	seg	grp	seg/grp	seg	grp	seg/grp
PN (新聞)	8,446	3,844	2.19	3,440	1,713	2.01	1,026	513	2.00
PB (書籍)	4,640	2,060	2.25	704	352	2.00	304	152	2.00
PM (雑誌)	5,513	2,454	2.24	1,313	651	2.02	280	140	2.00
OW (白書)	10,709	4,613	2.32	1,326	662	2.00	656	328	2.00
OC (知恵袋)	1,586	715	2.21	292	146	2.00	62	31	2.00
OY (ブログ)	1,603	682	2.35	262	131	2.00	58	29	2.00
合計	32,497	14,368	2.26	7,337	3,655	2.01	2,386	1,193	2.00

seg はセグメント数, grp はグループの数, seg/grp はグループあたりのセグメント数.

‘B’・‘F’・‘Z’, 文末数 (‘EOS’) である. ラベル ‘F’ は白書 (OW) と Yahoo! ブログ (OY) とで他のレジスタよりも, 割合が大きい傾向にある. 白書は箇条書きの行頭記号が多く出現する一方, ブログは顔文字などが多く出現し, ラベル ‘F’ が多い傾向にある. BCCWJ-DepPara では, BCCWJ の元データとは異なる文境界を認定しており (小西 他 2013), 文の入れ子を許しているため, 入れ子のすべての文末を表すラベル ‘Z’ が, アノテーション単位としての文末 ‘EOS’ よりも多い傾向にある.

表 4 は, 並列構造・同格構造の統計である. seg はセグメントの数 (並列構造の部分構成要素数), grp はグループの数 (並列構造全体の数), seg/grp は 1 グループあたりのセグメントの数 (並列構造を構成する平均部分構成要素数) を表す. 白書 (OW) は並列構造が多い傾向にある. 並列構造 (Parallel) は 2 つ以上の構成要素を許し, 平均部分構成要素数 (seg/grp) は 2.19–2.35 であった. 同格構造 (Apposition, General) は, 基本的に構成要素対に基づく構造である. しかしながら, 新聞 (PN) と書籍 (PB) に, 2 回以上言い換える例外的な表現があった.

7 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について, その基準とともに概説した. 基準においては, 係り受け構造と並列構造を分けてアノテーションを行うことを提案し, 並列構造は, 構成要素範囲 (セグメント) と構成要素集合 (グループ) により表現する方法を提案した.

節間の関係については, 南の節分類 (南 1974) に基づき, 係り先を決めることを提案した. 3 次チェック時に節を超える係り先を再確認することで, 整合性を保たせた. 節間のラベルについては, 現在『鳥バンク』 (池原 2007) の節間の関係 (池原 2009) を付与しつつある (松本, 浅原, 有田 2017; 松本 2017, 2018; Matsumoto, Asahara, and Arita 2018). BCCWJ-DepPara データは, 『現

代日本語書き言葉均衡コーパス』DVD版を購入した方に <http://bccwj-data.ninjal.ac.jp/mdl/> より公開している.

本提案手法に適合する係り受け解析モデルとして, (岩立 2012) は並列構造と係り受け構造を双対分解 (dual decomposition) して, 推定する手法を提案している.

謝 辞

本研究の一部は, 国立国語研究所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」, および国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものです. 本研究は JSPS 科研費 特定領域研究 18061005, 基盤研究 (A) 23240020, 基盤研究 (B) 25284083, 挑戦的萌芽研究 15K12888, 基盤研究 (A) 17H00917 の助成を受けたものです.

参考文献

- 橋本力, 黒橋禎夫, 河原大輔, 新里圭司, 永田昌明 (2011). 構文・照応・評価情報つきブログコーパスの構築. 自然言語処理, **18** (2), pp. 175–201.
- 池原悟 (2007). 鳥バンク (Tori-Bank). <http://unicorn.ike.tottori-u.ac.jp/toribank>.
- 池原悟 (2009). 非線形言語モデルによる自然言語処理. 岩波書店.
- 岩立将和 (2012). *Development of Pairwise Comparison-based Japanese Dependency Parsers and Application to Corpus Annotation*. Ph.D. thesis, Nara Institute of Science and Technology, Japan.
- Iwatate, M., Asahara, M., and Matsumoto, Y. (2008). “Japanese Dependency Parsing Using a Tournament Model.” In *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 361–368.
- 国立国語研究所コーパス開発センター (2015). 『現代日本語書き言葉均衡コーパス』利用の手引き第 1.1 版. 大学共同利用機関法人人間文化研究機構.
- 小西光, 中村壮範, 田中弥生, 間淵洋子, 浅原正幸, 立花幸子, 加藤祥, 今田水穂, 山口昌也, 前川喜久雄, 小木曾智信, 山崎誠, 丸山岳彦 (2015). 『現代日本語書き言葉均衡コーパス』の文境界修正. 国立国語研究所論集, **9**, pp. 81–100.
- 小西光, 小山田由紀, 浅原正幸, 柏野和佳子, 前川喜久雄 (2013). BCCWJ 係り受け関係アノテーション付与のための文境界再認定. 第 3 回コーパス日本語学ワークショップ発表論文集, pp. 135–142.
- Kudo, T. and Matsumoto, Y. (2002). “Japanese Dependency Analysis using Cascaded Chunking.”

- In *Proceedings of the 6th Conference on Natural language learning*, pp. 63–69.
- 黒橋禎夫, 居倉由衣子, 坂口昌子 (2000). 形態素・構文タグ付きコーパス作成の作業基準 (Version 1.8). テクニカル・レポート, 京都大学.
- 松本理美, 浅原正幸, 有田節子 (2017). 『現代日本語書き言葉均衡コーパス』に対する節の意味分類情報アノテーションー基準の策定、仕様書の作成の必要性についてー. 言語資源活用ワークショップ 2016, pp. 336–346.
- 松本理美 (2017). 従属節の意味分類基準策定についてー鳥バンク基準互換再構築の検討ー. 言語資源活用ワークショップ 2017, pp. 39–50.
- 松本理美 (2018). 日本語研究のための日本語従属節意味分類基準の試案ー「鳥バンク」節間意味分類体系の再構築ー. 言語処理学会第 24 回年次大会, pp. 769–772.
- Matsumoto, S., Asahara, M., and Arita, S. (2018). “Japanese Clause Classification Annotation on the ‘Balanced Corpus of Contemporary Written Japanese’.” In Shirai, K. (Ed.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Paris, France. European Language Resources Association (ELRA).
- 南不二男 (1974). 現代日本語の構造. 大修館書店.
- 中俣尚己 (2015). 日本語並列表現の体系. ひつじ書房.
- 野田尚史 (1985). はとが. 日本語文法セルフ・マスターシリーズ (1). くろしお出版.
- 小椋秀樹, 小磯花絵, 富士池優美, 宮内佐夜香, 小西光, 原裕 (2011a). 『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版 (下). 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書, 国立国語研究所.
- 小椋秀樹, 小磯花絵, 富士池優美, 宮内佐夜香, 小西光, 原裕 (2011b). 『現代日本語書き言葉均衡コーパス』形態論情報規定集第 4 版 (上). 特定領域研究「日本語コーパス」平成 22 年度研究成果報告書, 国立国語研究所.
- 小澤俊介, 内元清貴, 伝康晴 (2014). 長単位解析器の異なる品詞体系への適用. 自然言語処理, **21** (2), pp. 279–401.
- Sassano, M. (2004). “Linear-Time Dependency Analysis for Japanese.” In *Proceedings of the 20th International Conference on Computational Linguistics*, pp. 8–14.
- 田野村忠温 (2014). BCCWJ の資料特性. 田野村忠温 (編), コーパスと日本語学, 講座日本語コーパス 6. 朝倉書店, 東京.
- 内元清貴, 丸山岳彦, 高梨克也, 井佐原均 (2003). 『日本語話し言葉コーパス』における係り受け構造付与 (Version 1.0). テクニカル・レポート, 『日本語話し言葉コーパス』の解説文書.
- Uchimoto, K., Sekine, S., and Isahara, H. (1999). “Japanese Dependency Structure Analysis Based on Maximum Entropy Models.” In *EACL '99: Proceedings of the 9th Conference on European Chapter of the Association for Computational Linguistics*, pp. 196–203.

- Yoshinaga, N. and Kitsuregawa, M. (2010). “Kernel Slicing: Scalable Online Training with Conjunctive Features.” In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 1245–1253.
- Yoshinaga, N. and Kitsuregawa, M. (2014). “A Self-adaptive Classifier for Efficient Text-stream Processing.” In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, pp. 1091–1102.

略歴

浅原 正幸：2003 年奈良先端科学技術大学院大学情報科学研究科博士後期課程修了。2004 年より同大学助教。2012 年より国立国語研究所コーパス開発センター准教授。現在に至る。博士（工学）。自然言語処理・認知言語学の研究に従事。

松本 裕治：1977 年京都大学工学部情報工学科卒。1979 年同大学大学院工学研究科修士課程情報工学専攻修了。同年電子技術総合研究所入所。1984～85 年英国インペリアルカレッジ客員研究員。1985～87 年（財）新世代コンピュータ技術開発機構に出向。京都大学助教授を経て、1993 年より奈良先端科学技術大学院大学教授。現在に至る。2016 年 9 月より理研 AIP 知識獲得チーム PI 兼務。工学博士。専門は自然言語処理。情報処理学会、人工知能学会、AAAI、ACL、ACM 各会員。情報処理学会フェロー。ACL Fellow。

(2017 年 12 月 26 日 受付)

(2018 年 3 月 26 日 再受付)

(2018 年 4 月 17 日 採録)