係り受けアノテーション基準の比較

浅原 正幸 (国立国語研究所コーパス開発センター) *

Comparison of Syntactic Dependency Annotation Schemata

Masayuki Asahara (Center for Corpus Development, NINJAL)

1. はじめに

言語処理の分野でアノテーションデータに基づく統語解析の研究が盛んにおこなわれている。句構造もしくは係り受け構造が付与されたコーパスアノテーションに基づいて、さまざまな統語解析アルゴリズムと構造学習手法が提案されている一方、アノテーションの基準そのものに興味を持つ者は少ない。

英語において係り受け解析器の開発は、句構造がアノテーションされた Penn Treebank (Marcus et al. (1993)) を主辞規則 (Head percolation rules) などにより変換した係り受けアノテーションに基づいて行われている。主辞規則は係り受け解析アルゴリズムの計算量の観点から非交差制約 (projective) に基づいたもの (Magerman (1994), Collins (1999), Yamada and Matsumoto (2003)) が多く、Wh 疑問文・話題化 (topicalization)・分裂文 (cleft)・並列構造などの長距離係り受け関係については単純化されている。係り受け解析器の誤りの多くはこのような係り受け関係であるが、アノテーションの単純化による限界という指摘もあり、Johansson and Nugues (2007) は並列構造や従属節に対する係り受け関係の再定義を行い、分裂文や空所 (gapping) を Penn Treebank に付与されている二次辺 (secondary edge) や痕跡 (trace) の情報を用いて精緻化した。

日本語では文節係り受け構造が京都大学テキストコーパス、KNB コーパス (Kyoto-University and NTT Blog コーパス)、日本語話し言葉コーパス、現代日本語書き言葉均衡コーパスに付与されているが、ほとんどの係り受け解析器が京都大学テキストコーパスのアノテーションに基づいて構成されている。本稿では日本語で係り受け解析器が誤りやすい現象は各コーパスにおいてどのようなアノテーション基準に基づいて表現されているかを明らかにするために、係り受けアノテーション基準の比較を行う。対象は京都大学テキストコーパス基準 (以下KC; 黒橋ほか (2000))、日本語話し言葉コーパス基準 (以下CSJ; 内元ほか (2004))、現代日本語書き言葉均衡コーパス基準 (以下BCCWJ; 浅原 (2013)) の三つとする。KNB コーパスのアノテーション基準は京都大学テキストコーパス基準に準じているものとする。

2. 本稿における係り受け・並列構造の表現

本稿では図 1 にように係り受け・並列構造を表現する。

* masayu-a@ninjal.ac.jp

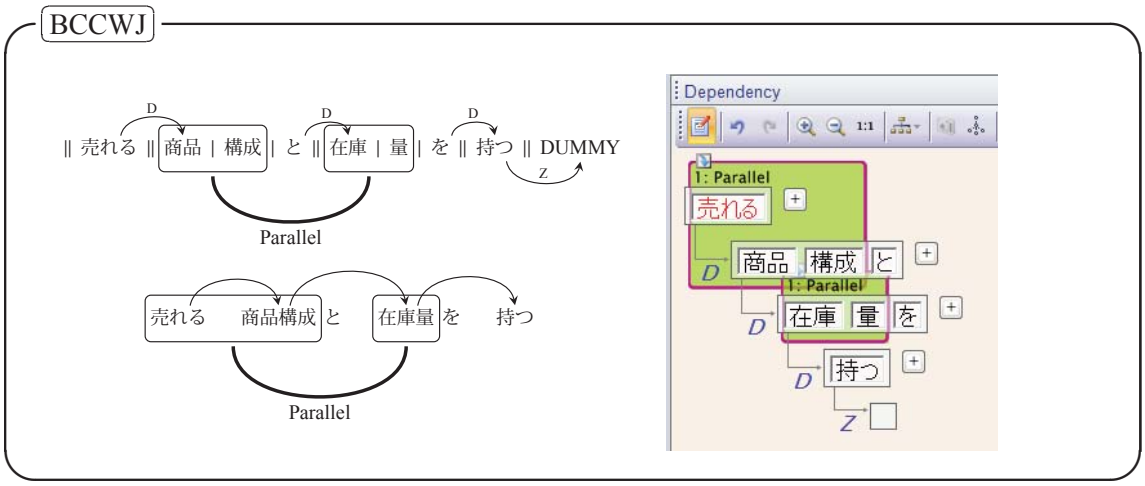


図 1　係り受け・並列構造アノテーションの表現方法

左上図中 ‖ が文節境界、| が短単位形態素境界、例文上のラベル “D” 付矢印が係り受けラベル “D” である係り受け関係を表す。例文下のラベル “Z” 付矢印が文末要素を表現する関係を表す。BCCWJ では並列構造などをセグメントとよばれる短単位形態素境界を最小単位とする範囲で複数切り出し、グループ化する。角丸四角と例文下のラベル “Parallel” 付曲線は並列構造範囲とその対応関係を表現する。他に、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” は係り先なしを表現するための要素である。アノテーションツール ChaKi (Matsumoto et al. (2005)) 上では右図のような形で表示される。

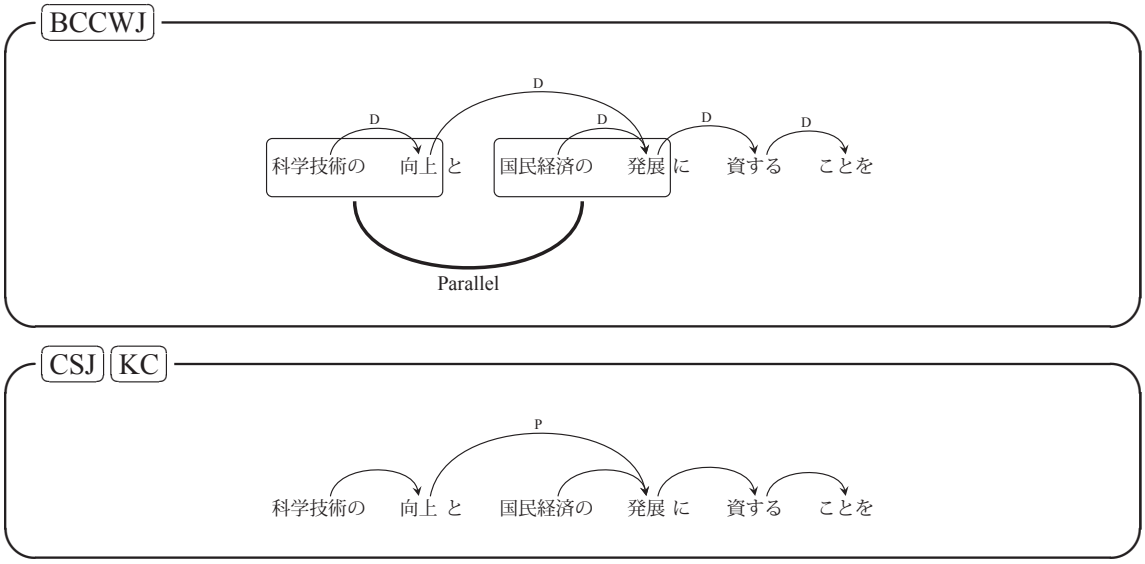
同じ文を、左下図のように略記することもある。文節境界記号と短単位形態素境界記号は範囲指定が不要な場合は省略し、文節境界の間に空白を入れて表現する。文末以外に係り先なしの関係がない場合には “DUMMY” を省略する。「通常の係り受け」は CSJ でラベルなし、KC, BCCWJ ではラベル “D” を用いるが、複数の基準の通常の係り受け関係を表現する際にはラベルなしとする。尚、CSJ においてラベル “D” は言いよどみを意味する。

3.　係り受け関係の比較

以下では三つの係り受けアノテーション基準で差異がある部分を対比的に示す。

3.1　並列構造

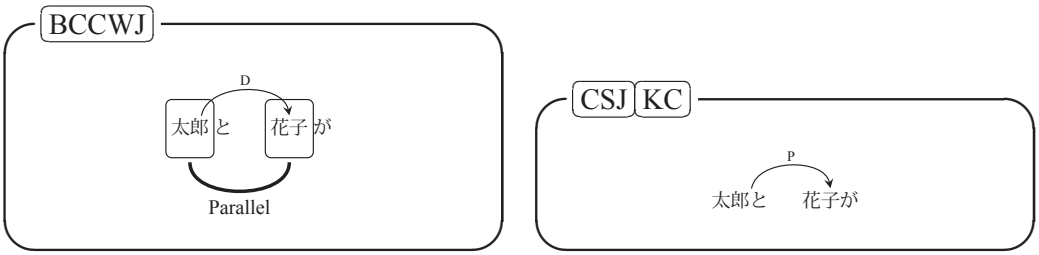
並列構造は日本語係り受け解析において頻出する扱いが難しい構造の一つである。BCCWJ のアノテーション基準の特色として、並列構造の範囲と対応する並列句を、係り受け木とは独立に範囲を付与する点がある。以下の例で、BCCWJ 基準では、係り受け関係ラベルを全て “D” としたうえで、「科学技術の向上」と「国民経済の発展」が対応する並列構造として、セグメント Parallel で切り出され、グループ化される。一方、CSJ、KC では、並列構造の構成句の最右要素動詞をラベル “P” でかける。



以下、様々な並列構造について示す。

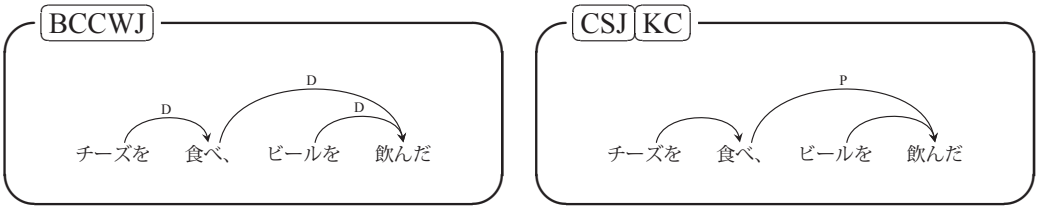
3.1.1 名詞句の並列

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。一方、CSJ KCにおいては、ラベル “P” によりアノテーションを行う。



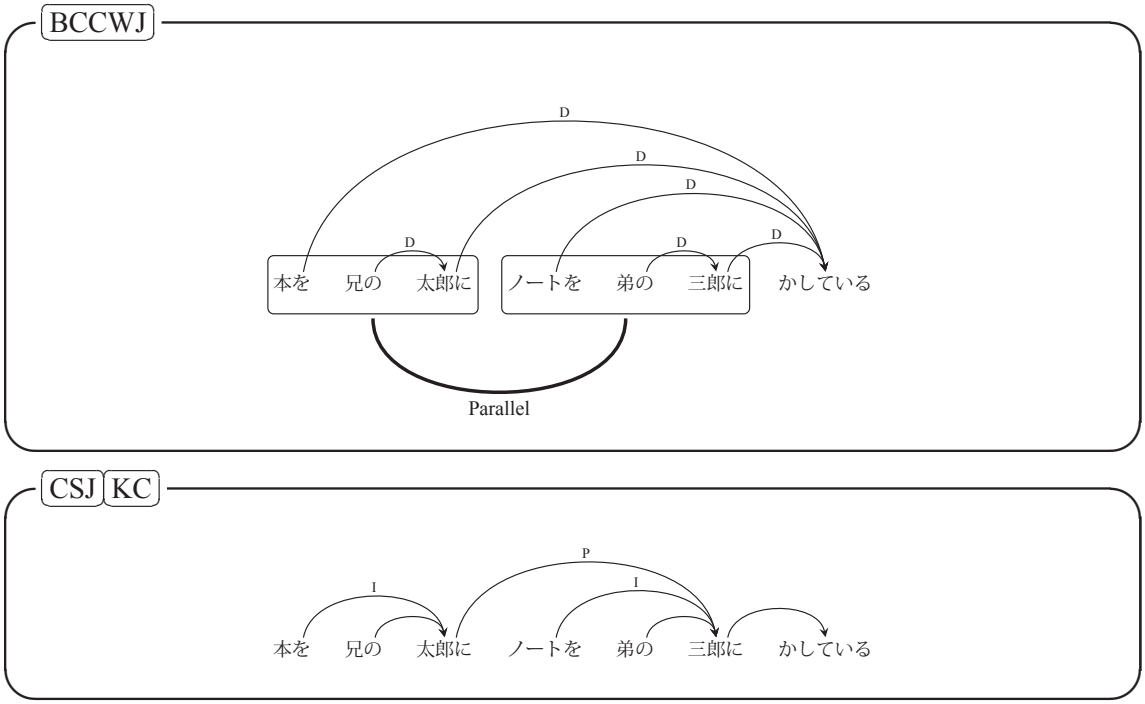
3.1.2 述語並列

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においては、全ての述語並列を並列とみなさず、通常の係り受けとして定義する。



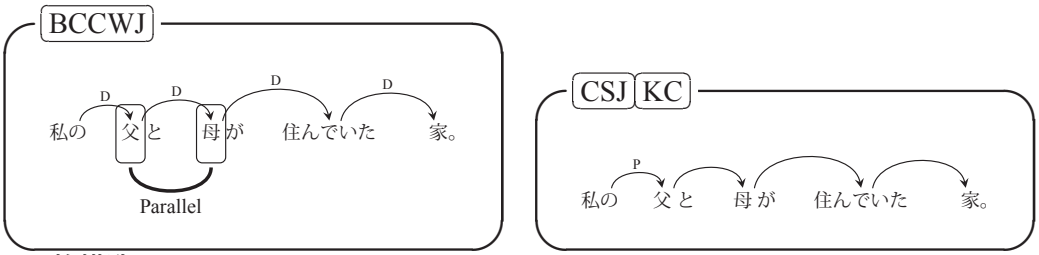
3.1.3 部分並列内の関係

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係けている。このようにラベルに交差の情報を持たせて、非交差条件を満たす木に変換する手法は pseudo projective と呼ばれる (Nivre and Nilsson (2005))。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係ける。



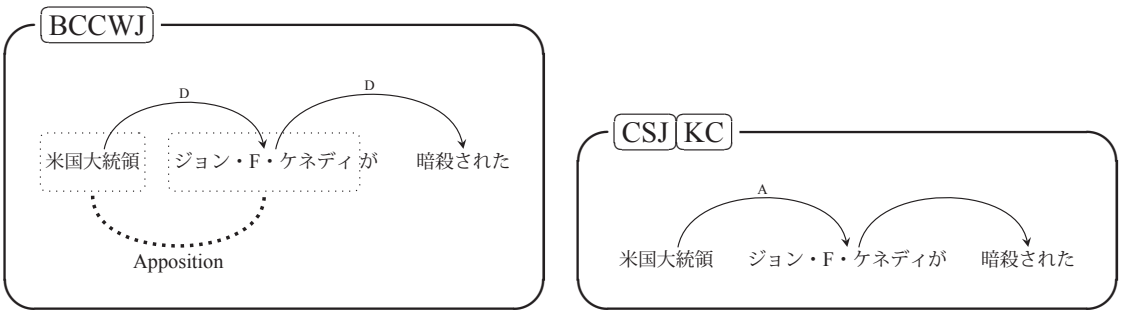
3.1.4　並列構造の複数の要素に左から係る場合

以下のように「オ（リックス）は」は「オーストリア」と「オーストラリア」の両方に係る場合には、BCCWJにおいては当該部分を並列構造範囲から外す。最左要素である「オーストリア」に係けることにより、両方に係っていることを表現する。



3.2　同格構造

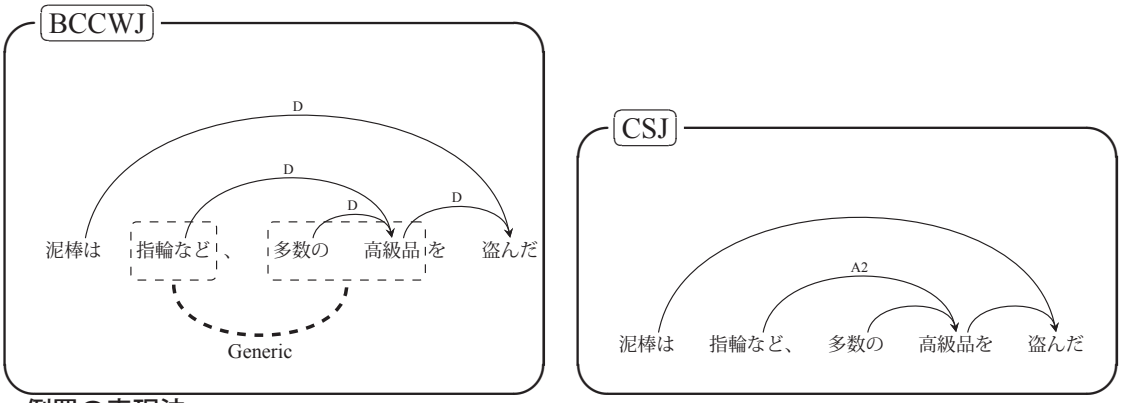
BCCWJにおいて、通常の同格関係は、対応する名詞句をセグメント Apposition で切り出し、グループ化する。係り受け関係は通常の係り受けと同じラベル “D” を付与する。一方、CSJ KCにおいては、ラベル “A” によりアノテーションを行う。



BCCWJとCSJは次に示す広義の同格を認定し、上に示した狭義の同格と区別するのに対し、KCは同格の意味を広めにとる傾向にある。

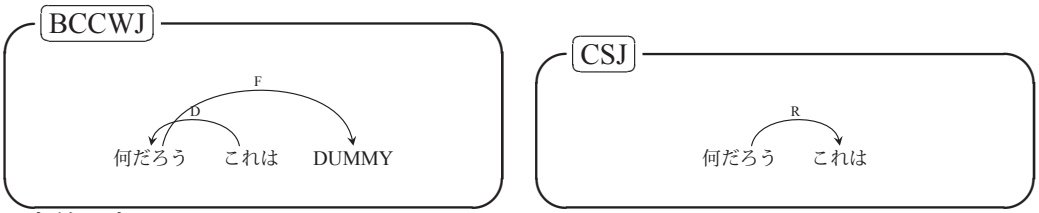
3.3 広義の同格

BCCWJとCSJは広義の同格として具体例と総称の同格関係、具体例と数詞の同格関係を狭義の同格と別のラベルで認定する。BCCWJでは、対応する名詞句をセグメント “Generic” 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。CSJでは、ラベル “A2” によりアノテーションを行う。KC においてはこの広義の同格を識別する方策は規定されていない。



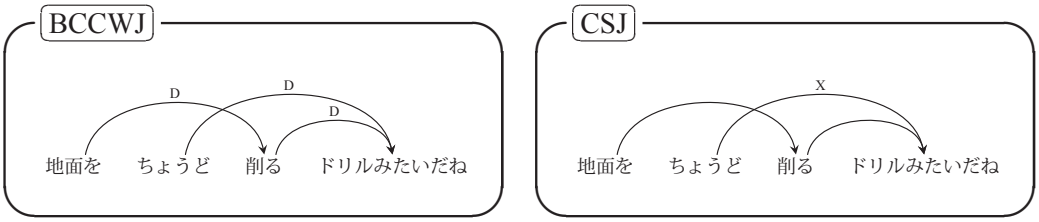
3.4 倒置の表現法

KCの基準においては、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” を用いて明示するが、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」は係り先なしの根ノードになるが、アノテーションツール上では末尾の DUMMY ノードに係けることにより表現する。



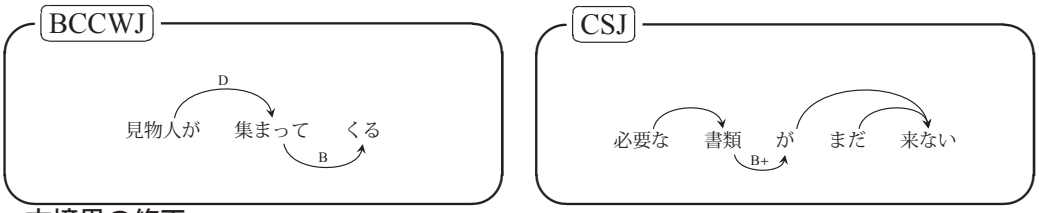
3.5 交差の表現

KCの基準においては、非交差制約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。ChaKi.NET の Dependency Panel 上では、交差があった場合には係り受け関係の色が自動的にオレンジに変更される。



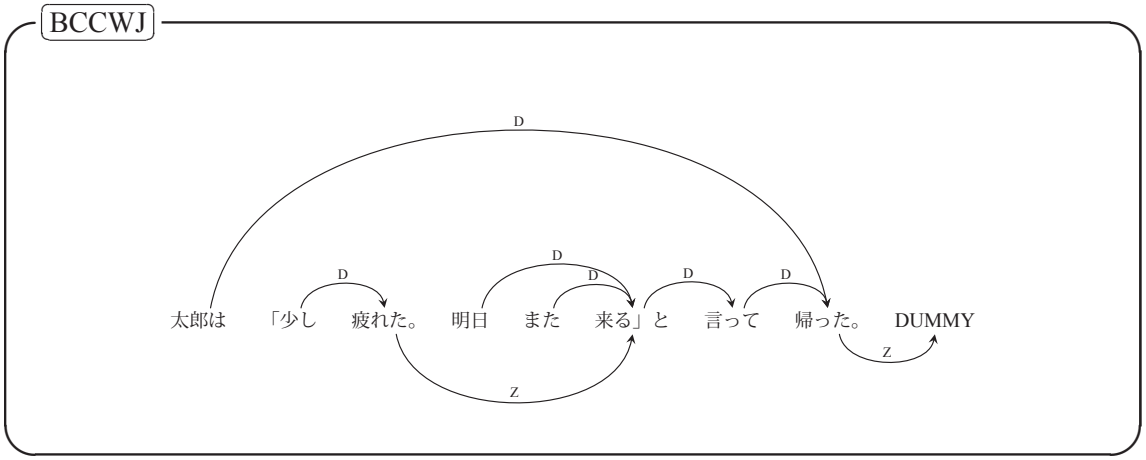
3.6 文節の連結

KCが文節係り受けを付与することを目的として文節単位を規定しているのに対し、BCCWJとCSJは形態論情報のみに基づいて文節単位を規定しており、係り受けを付与するためにそぐわない文節出現する。さらにCSJでは文節および節境界を元の音声ファイルのポーズによっても認定するために、文法的に不自然な単位が認定される場合がある。これに対応するために、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 “B” ラベル、CSJにおいて “B+” を用いて表現することがある。KCではこのような規定は存在しない。



3.7 文境界の修正

BCCWJ は文単位の定義として文の入れ子を許している。文書構造（レイアウト）に基づいて、一番外側の文について ⟨superSentence⟩ タグが付与されている。本来文の構造としては ⟨superSentence⟩ タグが付与されるべきものであって、文書構造中改行がある場合など ⟨superSentence⟩ タグが付与されていない場合、係り先のない文節が隣接文に出現する場合がありうる。このようなことのないように、BCCWJ 係り受けアノテーションにおいては、係り受けアノテーション向けに前処理で文書構造を考慮せずに、⟨superSentence⟩ 相当情報を追加で付与する。この際、文内に文境界相当の文節端が出現する場合がある。そのような場合には、BCCWJでは係り先なしとし、ラベル “Z” を付与する。一方、CSJは係り受けアノテーションを付与する単位として節を用いておりこのような問題は発生しない。また、KCではこのような規定は存在しない。



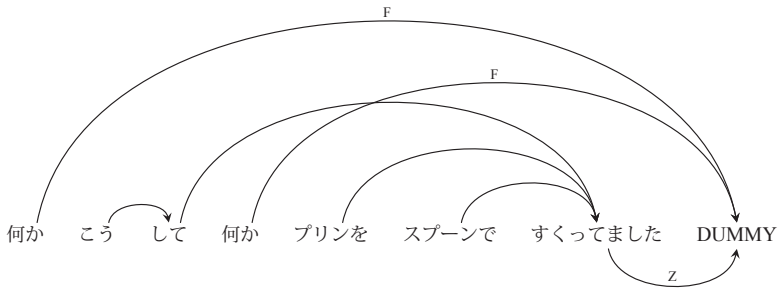
3.8 係り先なしの要素

KCでは係り先なしの文節要素を文末以外に認定していないのに対し、BCCWJとCSJでは係り先なしの文節要素を文末以外にも許している。特にCSJでは係り先なしの文節をラベルで細分化している。以下では、係り先なしの要素について比較する。

3.8.1 フィラー

CSJは、ラベル “F” を用い、フィラーの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、同様に、ラベル “F” を用い、フィラーの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。

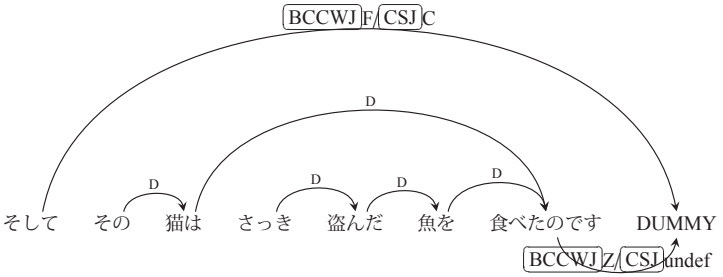
BCCWJ CSJ



3.8.2 接続詞

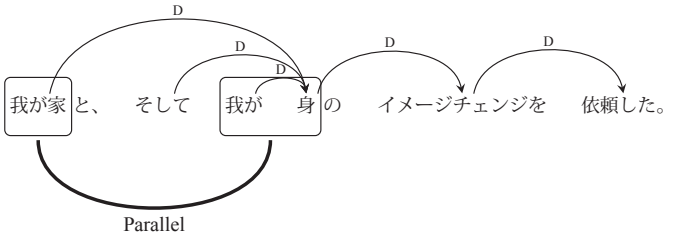
CSJは、ラベル “C” を用い、接続詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、文頭の接続詞で係り先判定が難しい際にラベル “F” を用い、接続詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。

BCCWJ/CSJ



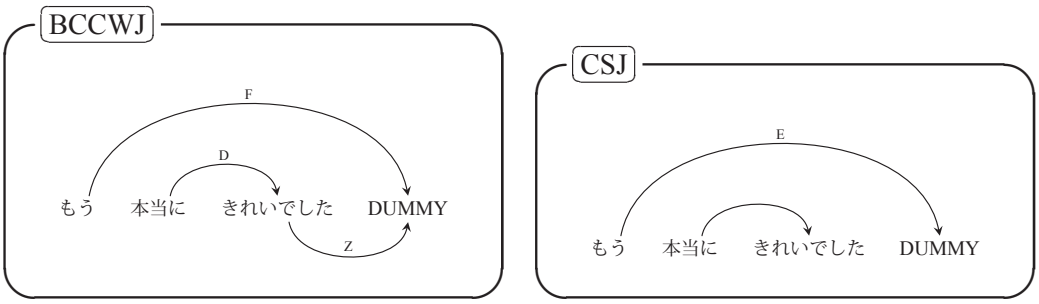
BCCWJにおいて、並列構造などを伴い、並列句の間に接続詞が出現する場合には、右隣接する並列句の最右文節に通常の係り受け関係 (ラベル “D”) として係ける。

BCCWJ



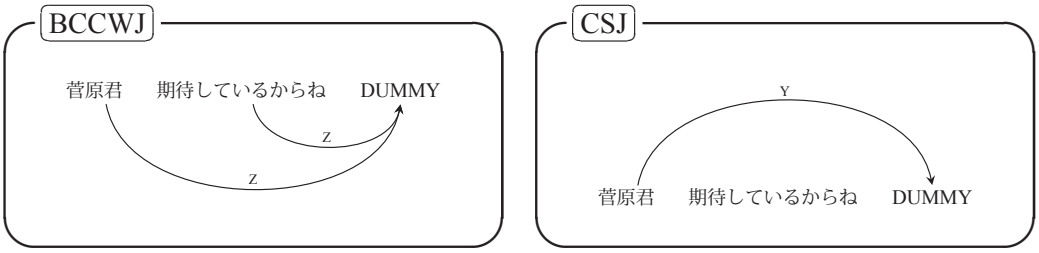
3.8.3 感動詞

CSJは、ラベル “E” を用い、感動詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、ラベル “F” を用い、感動詞の係り先は定義しない。DUMMY に係けることによって係り先なしを示す。



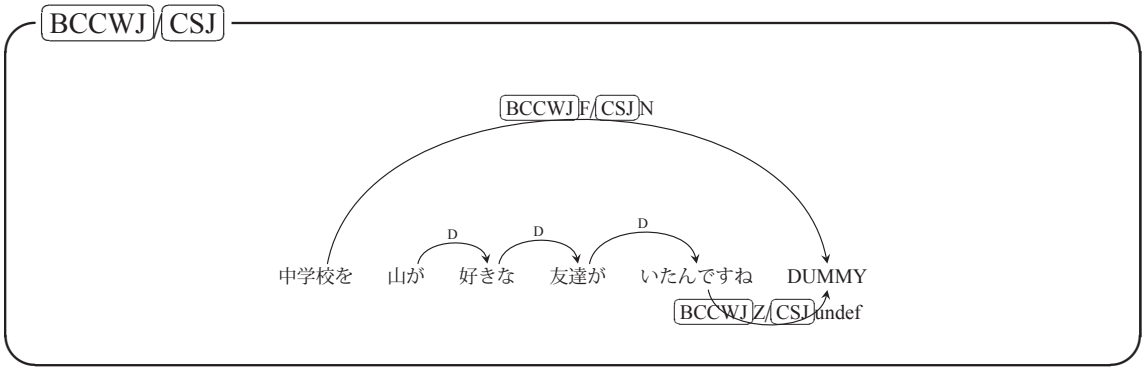
3.8.4 呼びかけ

CSJは、ラベル “Y” を用い、呼びかけの係り先は定義しない。DUMMY に係けることによって係り先なしを示す。BCCWJでは、ラベル “Z” を用い、呼びかけのあとに文境界相当の区切りを付与する。DUMMY に係けることによって係り先なしを示す。



3.8.5 係り先が消失している場合に付与するラベル

CSJは、ラベル “N” を用い、DUMMY に係けることによって係り先なしを示す。BCCWJは、ラベル “F” を用い、DUMMY に係けることによって係り先なしを示す。



3.9 格要素が複数の述語に係る場合

係り先を認定するのが難しい事例として、格要素が複数の述語に係る事例がある。並列する複数の述語の場合は等位接続とみなし係りうる遠いものに係ける。一方、複数の述語がそれぞれ従属節・主節に含まれている場合には、主題相当文節（「は」「も」）、主語相当文節（「が」）、それ以外の格要素（「に」「を」）など文節要素ごとに厳密に規定すべきである。BCCWJでは、このあたりの関係を南 (1974) の節分類などに基づき精緻化した。詳細については浅原 (2013)

を参照されたい。

3.10 その他

表 1 に各コーパスの係り受け関係ラベルの違いを示す。

表 1　係り受け関係ラベルの比較

係り受け関係のラベル	BCCWJ	(グループ | セグメント)	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
同格	D	(Apposition)	A	A
同格（総称、数詞）	D	(Generic)	A2	A
言いよどみ	D	(Disfluency)	D	未定義
倒置	D	-	R	未定義
文節境界に関するラベル	BCCWJ	-	CSJ	KC
後続文節と接続	B	-	B+	未定義
その他	BCCWJ	(セグメント)	CSJ	KC
フィラー	F	-	F	未定義
顔文字	F	-	未定義	未定義
接続詞	F or D		C	D
感動詞	F or D	-	E	D
呼びかけ	Z	-	Y	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
係り受け関係の交差	D	-	X	未定義 (A のみ)
英単語・ローマ字文・漢文	D	(Foreign)	未定義	未定義
古文	D	(Foreign)	K(|S1|E1)	未定義
文境界相当	Z	-	未定義	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 | を | に」)	未定義
	F	(Disfluency)	S:複数文節の言い直し (|S1|E1)	未定義

以下、言及していない基準間の違いについて簡単に述べる。

- 言い直し・言いよどみ
  CSJでは言いよどみをラベル “D” で付与する。また複数文節の言い直しについては “S:複数文節の言い直し” ラベルに開始タグ (S1) と終了タグ (E1) を付与し範囲指定する。BCCWJでは言いよどみ相当句に Disfluency セグメントを規定し、言い直した表現に通常の係り受け関係で係ける。
- 顔文字・非言語音
  BCCWJでは、格要素などにならない顔文字表現については、副詞的用法であっても、句読法的な用法であっても区別せずに、ラベル “F” とし、DUMMY ノードに係ける。CSJでは、非言語音は通常の係り受けとして扱う。BCCWJでは、顔文字と同様に扱う。
- 記号・補助記号・URL・空白
  BCCWJでは、係り先が判定しにくい、リスト項目マーカー相当の記号・補助記号については、ラベル “F” とし、DUMMY ノードに係ける。URL・空白も同様に扱う。
- 英単語・ローマ字文・漢文・古文
  CSJでは、古文相当を係り受けラベル “K” で扱う。古文が複数文節にわたる場合にはラベル “K” に開始タグ (S1) と終了タグ (E1) を付与し範囲指定する。BCCWJでは、係り受け木とは独立にセグメント “Foreign” として英単語・ローマ字文・漢文・古文の範

囲を指定する。係り受け関係は通常の係り受けとしてみなす。

- 格表示誤り
  CSJでは、発話者の格表示誤りと想定される文節について、ラベル “S” に “格表示誤り(「が | を | に」)” をつけて付与する。

4. おわりに

本稿では、日本語の係り受けアノテーション基準間の差異について概観した。より詳細な比較については浅原 (2013) を参照されたい。

謝辞

本研究は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

Collins, Michael J. (1999). “Head-driven statistical models for natural language.” Unpublished doctoral dissertation, University of Pennsylvania.

Johansson, Richard, and Pierre Nugues (2007). “Extended constituent-to-dependency conversion for english.” *Proc. of The 16th Nordic Conference of Computational Linguistics (NODALIDA-2007)*.

Magerman, David M. (1994). “Natural language parsing as statistical pattern recognition.” Unpublished doctoral dissertation, Stanford University.

Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). “Building a large annotated corpus of english: the penn treebank.” *Computational Linguistics*, 19:2, pp. 313–330.

Matsumoto, Yuji, Masayuki Asahara, Kou Kawabe, Yurika Takahashi, Yukio Tono, Akira Ohtani, and Toshio Morita (2005). “Chaki: An annotated corpora management and search system.” *Proc. of the Corpus Linguistics Conference Series (Corpus Linguistics 2005)*.

Nivre, Joakim, and Jens Nilsson (2005). “Pseudo-projective dependency parsing.” *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pp. 99–106. Ann Arbor, Michigan: Association for Computational Linguistics.

Yamada, Hiroyasu, and Yuji Matsumoto (2003). “Statistical dependency analysis with support vector machines.” *Proc. of 8th International Workshop of Parsing Technologies (IWPT-2003)*.

浅原正幸 (2013).「『現代日本語書き言葉コーパス』係り受け・並列構造アノテーション作業メモ (Version 0.6)」 Technical report, 国立国語研究所コーパス開発センター.

内元清貴・丸山岳彦・高梨克也・井佐原均 (2004).「『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)」 Technical report, 『日本語話し言葉コーパス』の解説文書.

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)」 Technical report, 京都大学.

南不二男 (1974).『現代日本語の構造』 大修館書店.