BCCWJ-DepParaPAS:『現代日本語書き言葉均衡コーパス』係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化

浅原 正幸
人間文化研究機構 国立国語研究所
言語資源研究系・コーパス開発センター

大村 舞
奈良先端科学技術大学院大学
情報科学研究科

masayu-a@ninjal.ac.jp

1 はじめに

国立国語研究所の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」では、『現代日本語書き言葉均衡コーパス』(以下 “BCCWJ”) に対する様々なアノテーションに取り組んできた。本稿では、その中で進めてきた係り受け・並列構造アノテーション BCCWJ-DepPara [1] と BCCWJ-PAS [5] の重ね合わせと ChaKi.NET による可視化について示す。

2006 年より特定領域研究のプロジェクトとして BCCWJ の構築がはじまり、2007 年に最初のコアデータサンプル (OC,OW,PB,PN) が領域内公開された。当初の短単位形態論情報を元に並列構造・同格構造のアノテーションおよび、述語項構造・共参照アノテーションが並行して行われた。2008 年に文節の情報が利用できるようになり、さらに並行して文節係り受けアノテーションが進められた。全てのアノテーションにおいて原文文字列と文境界の情報を前提としており、並列構造・同格構造・述語項構造・共参照は短単位形態論情報を、文節係り受けアノテーションは文節境界を前提としている。文献 [2] の基準に基づき、並列構造・同格構造は Microsoft Excel を用いて、文節係り受けは『ChaKi.NET』[1] を用いてアノテーション作業を行った。2011 年 12 月の BCCWJ DVD 1.0 版の公開まで、前提となる上流工程の情報修正に応じて、構造化 diff を用いて修正を続けてきた。2012 年にコアデータのみ文節係り受けに適した文境界定義が行われた [3]。文節係り受けアノテーションは 2011 年以降も続けられたが、述語項構造・共参照アノテーションにおいてはアノテーション作業自体が中断された。2015 年に述語項構造アノテーション 1 次チェック済み (一部 2 次チェック済み) のデータをとりまとめ、述語項構造・共参照アノテーションの最終チェック作業を行った [5]。述語項構造アノテーション作業は NAIST Text Corpus の基準 [2] に基づき、アノテーションツール『Tagrin』[3] を用いて行った。先に述べた通り、文境界は独自のもの [3] を採用しており BCCWJ DVD 1.1 版のものへの統合は行っていない。

2 重ね合わせのデータ形式と可視化

2.1 拡張 CaboCha 形式の概要

拡張 CaboCha 形式 [4] は係り受け解析器 CaboCha の出力に対して、スタンドオフ形式でセグメント・リンク・同値類のアノテーションを`#!`ではじまる行で導入する形式である。表 1 に拡張 CaboCha 形式の概要について示す。

文内で閉じた要素のセグメント (`SEGMENT_S`) と文間の関係を指定する要素のセグメント (`SEGMENT`) の 2 種類に対して以下のようなオフセット値の計算を行う。

```
0 駒 1 と 2 盤 3 は 4 も 5 っ 6 て 7
------------------------------------
#! SEGMENT_S Parallel 0 1 "駒"
#! SEGMENT_S Parallel 2 3 "盤"
```

これらのセグメントは文内・文間それぞれ 0-origin の ID `<SegSNo>`・`<SegNo>`を持ち、以下のように同値類を定義する。リンクも同様に定義できる。

```
#! GROUP_S Parallel 0 1 ""
```

[1] https://osdn.jp/projects/chaki/

[2] https://sites.google.com/site/ryuiida/ntc-annotation-scheme/

[3] http://kagonma.org/tagrin/

表 1: 拡張 CaboCha 形式の概要

タグ	摘要
##	コメント記号
#! DOC <id>	文書開始タグ (ID の宣言)
#! DOCID\t<id>\t<Bibinfo>	文書単位の書誌情報
#! SEGMENT_S <TagName> <StartLPos> <EndLPos> "<Comments>"	文内で閉じたセグメント
#! SEGMENT <TagName> <StartGPos> <EndGPos> "<Comments>"	文間の関係を指定するセグメント
#! LINK_S <TagName> <FromSegSNo> <EndSegSNo> "<Comments>"	文内有向リンク
#! LINK <TagName> <FromSegNo> <EndSegNo> "<Comments>"	文間有向リンク
#! GROUP_S <TagName> <SegSNo> <SegSNo> ... "<Comments>"	文内同値類
#! GROUP <TagName> <SegNo> <SegNo> ... "<Comments>"	文間同値類

セグメント・リンク・同値類に対する属性情報を表現するために SEGMENT(_S)・LINK(_S)・GROUP(_S) に ATTR を後置させることができる。

```
#! SEGMENT_S <TagName> <StartLPos> <EndLPos> "<Comments>"
#! ATTR <Key1> "<Value1>"
#! ATTR <Key2> "<Value2>"
#! ATTR <Key3> "<Value3>"
```

さらに<TagName> に名前空間<ns> を前置し、要素名と属性名の名前衝突を回避することができる。

```
#! SEGMENT_S <ns>:<TagName> <StartLPos> <EndLPos> ".."
#! ATTR <ns>:<Key> "<Value>"
```

CaboCha 形式自体が係り受け解析器の出力であるため、文節係り受け・並列構造・同格構造には名前空間用のラベルを規定しない。述語項構造・共参照に対しては bccwj-pas を名前空間用のラベルとして用いる。

2.2 述語項構造の NAIST Text Corpus 形式

述語項構造は Tagrin というツールによりアノテーションされる。Tagrin の .tgr 形式を、NAIST Text Corpus 形式に変換したものを公開する。

図 1 に述語項構造の NAIST Text Corpus 形式の例を示す。係り受け解析結果のタブ区切りの最右列に述語項構造の情報を付与する。名詞句相当の「書店」に id が付与され、述語相当の「忘れ」に ga="exo1" o="3" o_dep="dep" type="pred"が付与されている。

表 2 に NAIST Text Corpus 形式の属性を示す。格要素はガ (ga)・ヲ (o)・ニ (ni)・ガ/ニ (ga/ni:助動詞に対する)・ハ (ha) からなる。文章内に格要素がある場合、対象の項名詞句に id を付与する。節照応 (ana_cla) や外界照応 (exo1,exo2,exog) の場合対応するラベルを付与する。格要素と述語の関係を ga_dep, o_dep, ni_dep, ha_dep にラベル {dep（直接係り受けあり）, zero（ゼロ代名詞）} を与える。述語のタイプとして type に {aux(助動詞), noun(名詞述語), pred(用言述語)} を与える。項名詞句の mention に対して id を uniq に与え、共参照名詞句の entity にたいして eq を uniq に与える。このうち ana_cla と aux は、今回新たに定義したラベルである。

2.3 述語項構造の拡張 CaboCha 形式

NAIST Text Corpus 形式では、形態素単位に属性として述語項構造の情報を与えていた。他のアノテーションと重ね合わせるために、述語項構造をスタンドオフ化した拡張 CaboCha 形式で表現する。図 2 に述語項構造の拡張 CaboCha 形式の例を示す。

共参照が文間の関係を規定することがあるため、述語項構造・共参照は文間の要素が定義できる SEGMENT・LINK・GROUP を用いて規定する。

SEGMENT は項を表す名詞句の主辞形態素 bccwj-pas:np と述語の表す要素 bccwj-pas:pred と述語にならない機能語相当表現 bccwj-pas:func の 3 つを定義する。

項名詞句と述語の関係は LINK に述語と項名詞句の SEGMENT の id を与えて表現するが、節照応・外界照応については、述語の属性 ATTR で表現する。共参照情報は GROUP により表現する。

2.4 可視化

拡張 CaboCha 形式の述語項構造データは、コンコーダンサ ChaKi.NET により GUI 上で可視化できる。図 3 に表示例を示す。ChaKi.NET の Dependency Panel (係り受け表示部) に係り受けに重ね合わせて、述語と項名詞句のセグメントと述語項構造リンクが表示される。セグメントもしくはリンクにマウスオーバーすることにより Attributes Panel (属性表示部) にアノテーションの各種情報が表示されるほか、Lexeme Panel (形態論情報表示部) に短単位形態論情報が表示

NAIST Text Corpus 形式：述語項構造

```
* 0 1D 1/2 0.000000
書店 名詞, 普通名詞, 一般,*,*,*, ショテン, 書店,*,*,*,*, 漢,*,*,*,* id="3"
名 名詞, 普通名詞, 助数詞可能,*,*,*, メイ, 名,*,*,*,*, 漢,*,*,*,* _
は 助詞, 係助詞,*,*,*,*, ハ, は,*,*,*,*, 和,*,*,*,* _
* 1 -1Z 0/2 0.000000
忘れ 動詞, 非自立可能,*,*, 下一段-ラ行, 連用形-一般, ワスレル, 忘れる,*,*, 忘れる,*, 和,*,*,*,*
ga="exo1" o="3" o_dep="dep" type="pred"
まし 助動詞,*,*,*, 助動詞-マス, 連用形-一般, マス, ます,*,*, ます,*, 和,*,*,*,* _
た 助動詞,*,*,*, 助動詞-タ, 終止形-一般, タ, た,*,*, た,*, 和,*,*,*,* _
。 補助記号, 句点,*,*,*,*,,。 ,*,*,*,*, 記号,*,*,*,* _
EOS
```

図 1: NAIST Text Corpus 形式：述語項構造

表 2: NAIST Text Corpus 形式の属性

属性	摘要	値
ga	ガ格	項名詞句の id or exo1, exo2, exog, ana_cla †
o	ヲ格	exo1: 外界一人称
ni	ニ格	exo2: 外界二人称
ga/ni	ガ/ニ格	exog: 外界その他
ha	ハ格	ana_cla: 節照応
ga_dep	ガと述語の関係	{dep, zero}
o_dep	ヲと述語の関係	dep: 直接係り受けあり
ni_dep	ニと述語の関係	zero: ゼロ代名詞
ha_dep	ハと述語の関係	
type	述語のタイプ	{aux(助動詞)†, noun(名詞述語), pred(用言述語)}
id	項名詞句の id	Integer
eq	共参照名詞句の id	Integer

† は今回新たに定義したタグ。

される。Dependency Panel 上で簡単な編集を行うこともできる。

3 おわりに

上にのべた形式のデータ 3 種 (NAIST Text Corpus 形式・拡張 CaboCha 形式・ChaKi.NET SQLite DB ファイル) を 2016 年 3 月に `https://bccwj-data.ninjal.ac.jp/mdl/` より、パッケージ BCCWJ-DepParaPAS として配布する予定である。

謝辞

本研究の一部は科研費萌芽「近代語コーパスに対する統語情報アノテーション基準策定」(15K12888) および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] 浅原正幸, 松本裕治. 『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. 言語処理学会第 19 回年次大会発表論文集, pp. 66–69, 2013.

[2] 浅原正幸. 係り受け関係アノテーション基準の比較. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 81–90, 2013.

[3] 小西光, 小山田由紀, 浅原正幸, 柏野和佳子, 前川喜久雄. 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 4 回コーパス日本語学ワークショップ予稿集, pp. 135–142, 2013.

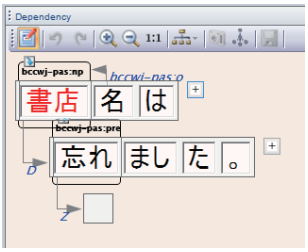
[4] 松吉俊, 浅原正幸, 飯田龍, 森田敏生. 拡張 CaboCha フォーマットの仕様拡張. 第 5 回コーパス日本語学ワークショップ予稿集, pp. 223–232, 2014.

[5] 植田禎子, 飯田龍, 浅原正幸, 松本裕治, 徳永健伸. 『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション. 第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214, 2015.

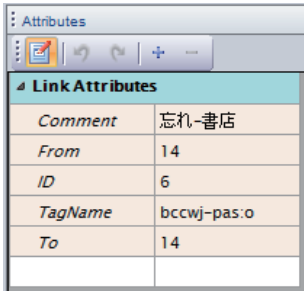
拡張 CaboCha 形式：述語項構造

```
...
* 0 1D 1/2 0.000000
書店 名詞,普通名詞,一般,*,*,*,ショテン,書店,*,*,*,*,漢,*,*,*,*
名 名詞,普通名詞,助数詞可能,*,*,*,メイ,名,*,*,*,*,漢,*,*,*,*
は 助詞,係助詞,*,*,*,*,ハ,は,*,*,*,*,和,*,*,*,*
* 1 -1Z 0/2 0.000000
忘れ 動詞,非自立可能,*,*,下一段-ラ行,連用形-一般,ワスレル,忘れる,*,*,忘れる,*,和,*,*,*,*
まし 助動詞,*,*,*,助動詞-マス,連用形-一般,マス,ます,*,*,ます,*,和,*,*,*,*
た 助動詞,*,*,*,助動詞-タ,終止形-一般,タ,た,*,*,た,*,和,*,*,*,*
。 補助記号,句点,*,*,*,*,,。,*,*,*,*,記号,*,*,*,*
EOS
...
#! SEGMENT bccwj-pas:np 97 99 "書店"
#! SEGMENT bccwj-pas:pred 101 103 "忘れ"
#! ATTR bccwj-pas:type "pred"
#! ATTR bccwj-pas:ga "exo1"
#! ATTR bccwj-pas:o_dep "dep"
...
#! LINK bccwj-pas:o 11 10 "忘れ-書店"
...
#! GROUP bccwj-pas:"よむ" 17 3 5
```

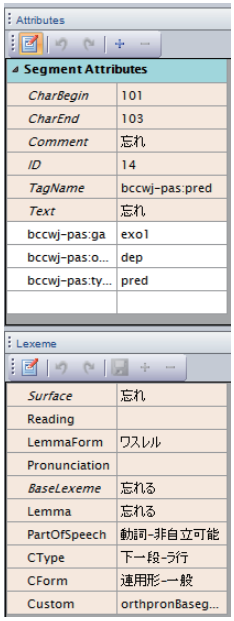
図 2: 拡張 CaboCha 形式：述語項構造



【係り受け表示】



【述語項構造リンク属性表示】



【述語セグメント属性表示】

【名詞句セグメント属性表示】

図 3: ChaKi.NET による可視化の例