拡張 CaboCha フォーマットの仕様拡張

松吉 俊 (山梨大学大学院医学工学総合研究部)
浅原 正幸 (国立国語研究所コーパス開発センター)
飯田 龍 (東京工業大学大学院情報理工学研究科)
森田 敏生 (総和技研)

RFC: Requirements in Extended CaboCha Formats

Suguru Matsuyoshi (Interdisciplinary Grad. School of Med. and Eng., Univ. of Yamanashi)
Masayuki Asahara (Center for Corpus Development, NINJAL)
Ryu Iida (Dept. of Computer Science, TITECH)
Toshio Morita (Sowa Research Co., Ltd.)

1. はじめに

各研究機関において「現代日本語書き言葉均衡コーパス」(BCCWJ) に対する様々なレベルの言語情報のアノテーションが進められている。形態論情報が人手により修正されているコアデータを中心に、共通のサンプルに複数のアノテーションが施され、これらを統合して分析・モデル化することが可能になりつつある。実際に分析・モデル化を行うためにはこれらを統合的に記述できる形式が必要になる。コーパス管理ツール『ChaKi.NET』の入出力で利用されている拡張 CaboCha フォーマット[*1]は意味的に異なるアノテーションを共通の形式で表現することを目的とし、形態論情報と係り受け情報に対する追加のアノテーションを表現するために、部分文字列を表現する SEGMENT、SEGMENT 間の有向関係を表現する LINK、SEGMENT 間の同値類を表現する GROUP を規定している。しかし、実際に重ね合わせ作業を進めてみると、アノテーション要素に属性情報を付与できない、複数のアノテーション間で名前の衝突が起きるなどの問題が発生した。そこで本研究ではこれらを解決する仕様拡張について示し、実際のアノテーションの形式表現について紹介する。

2. 現在の拡張 CaboCha フォーマットの仕様と問題点

拡張 CaboCha フォーマットは、形態論情報・文節・文節係り受けを表現する日本語係り受け解析器 CaboCha の出力フォーマット (`-f1` オプション) の出力に対して、**部分文字列の範囲を表現するセグメント**と**セグメント間の有向関係を表すリンク**と**セグメントの同値類（推移律を持つ無向リンク）を表すグループ**を表現できるようにしたものである。前提として拡張 CaboCha フォーマットはアノテーションとして形式のみを規定し、どのような情報を格納するかは自由に規定してよい。表 1 に従来の拡張 CaboCha フォーマットの概要を示す。

以下、各要素について解説する。`##` は大域的な付加情報を表現するコメント行を示す。

[*1] `http://sourceforge.jp/projects/chaki/wiki/`拡張 CaboCha フォーマットへのエクスポート を参照。

表 1　従来の拡張 CaboCha フォーマットの概要

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> <SegS> ... "<Comments>"	文内同値類
#! GROUP <TagName> <SegNo> <SegNo> ... "<Comments>"	文間同値類

DOC は文書単位に <ID> 相当の自然数 (1-origin) を割り当てる。DOCID は文書単位の書誌情報を格納する。書誌情報を表す<Bibinfo>は XML フラグメントで与えることとする。先に述べた 3 種類のアノテーションは識別子として <TagName> を持つ SEGMENT(_S)、LINK(_S)、GROUP(_S) により表現する。セグメント間の関係はセグメント番号に対するリファレンスにより表現し、セグメント番号は以下に述べる範囲で上から順に 0-origin で付与する。_S をつける場合は、関連するアノテーションが全て文内で閉じている場合に利用し、後で示すオフセット値 (<StartLPOS>: セグメント開始位置、<EndLPos>：セグメント終了位置) やリファレンスセグメント番号 (<FromSegSNo>：リンク元、<ToSegSNo>：リンク先、<SegSNo>：同値類）も文内で閉じたものを割り当てる。_S をつけない場合は、<DOC>行で分割される範囲で閉じたオフセット値 (<StartGPOS>: セグメント開始位置、<EndGPos>：セグメント終了位置) やリファレンスセグメント番号 (<FromSegNo>：リンク元、<ToSegNo>：リンク先、<SegNo>：同値類) を割り当てる。いずれのオフセット値も 0-origin の文字位置で表現する。以下にオフセット値の計算例を示す。

オフセット値の計算

```
0 駒 1 と 2 盤 3 は 4 も 5 っ 6 て 7 い 8 ま 9 せ 10 ん 11 。12
-----------------------------------------------------------------
#! SEGMENT_S Parallel 0 1 "駒"
#! SEGMENT_S Parallel 2 3 "盤"
```

<Comments> には局所的な付加情報を表現するコメントを記述する。本稿では便宜的に SEGMENT(_S) の<Comments>に表層文字列を記述する。

以上の現在の拡張 CaboCha フォーマットに対して今回扱いたい問題が二つある。

一つ目は属性情報である。XML などで表現されるアノテーションは XML タグに属性情報が表現できる。しかしながら現状の拡張 CaboCha 形式は属性情報を表現する手立てがない。

二つ目は名前空間である。形態論情報・文節・文節係り受けが付与されたうえに多様なアノテーションを重ね合わせる際にタグ名が衝突する事態が起きうる。その際に衝突を回避する手立てがない。

次節ではこの二つの問題を解決するための仕様拡張案について示す。

3. 仕様拡張案：属性情報と名前空間

前節で示した問題点に対処するために、新たに次の二つの機構を導入する。

一つ目は SEGMENT(_S)、LINK(_S)、GROUP(_S) に対する属性情報を表現するための ATTR の導入である。代表して以下に SEGMENT_S に対する属性付与方法を示す。

属性付与

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

ATTR は、その上に存在する一番近い要素に対する属性と解釈される。1 つの要素に対して付与できる属性の数に制限はない。拡張 CaboCha フォーマット形式のファイルを可視化するツールは、要素の属性を綺麗な形で表示できることが望ましい。

二つ目は <TagName> に対して名前空間<ns> を付与し名前衝突を回避する機構の導入である。次の記法により、要素名と属性名に対して、名前空間を記述する。

名前空間定義

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

名前空間を使用する目的は情報提供者を明示し、複数のアノテーションを重ね合わせた時に要素名が衝突するのを防ぐためである。例えば、pred(Pred) という名前の要素が二つ以上のアノテーションで異なる意味で用いられている場合、bccwj-pth:pred と bccwj-pas:Pred と表記することにより、これらを容易に区別することができる。

4. 仕様拡張案の使用例

表 2 に各機関で進められている BCCWJ に対するアノテーションと表現方法を示す。

表 2　BCCWJ に対するアノテーションと表現方法

アノテーション	開発者	名前空間	セグメント	リンク	グループ
並列・同格構造	奈良先端大・国語研	-	○	×	○
時間情報	国語研	bccwj-timebank	○	○	×
拡張固有表現	東工大	bccwj-ene	○	×	×
語義	東工大	bccwj-wsd	○	×	×
項構造シソーラス	岡山大	bccwj-pth	○	○	×
れる・られる	国語研	bccwj-auxv	○	×	×
日本語フレームネット	慶應大	bccwj-jfn	○	×	○
拡張モダリティ	奈良先端大・東北大	bccwj-eme	○	×	×
否定の焦点	山梨大	bccwj-wsb	○	○	×
述語項構造	奈良先端大	bccwj-pas	○	○	○

以下では様々な機関で実施されている BCCWJ に対するアノテーションを例に仕様拡張案の使用例を示す。セグメントとリンクとグループに属性情報を付与することで全てのアノテー

ションが表現可能である。また、10 種類のアノテーションの間で名前の衝突は頻繁に起こっており、pred(Pred)、event(Event)、type、class などが複数のアノテーションで利用されていた。新たに名前空間の機構を与えることにより衝突を回避することが可能になる。尚、いくつかのアノテーションについては、元データに含まれていた最終更新日時や表示時の色指定などの属性情報など、言語情報に直接関係しないものを省略して示す。

4.1 アノテーション事例

■並列構造・同格構造　並列構造・同格構造は奈良先端大・国語研において係り受けとともにアノテーションが進められている (浅原・松本 (2013))。並列構造・同格構造は文内に閉じており、同値類として三つ以上のセグメントに対しても規定できるため SEGMENT_S、GROUP_S により表現する。係り受けと一体となっているため名前空間は定義しない。

例: OC01_00001（サンプル名を表す; 以下同様）

```
* 0 1D 0/0 0
駒 名詞,普通名詞,一般,*,*,*,コマ,駒,駒
と 助詞,格助詞,*,*,*,*,ト,と,と
* 1 2D 0/0 0
盤 名詞,普通名詞,一般,*,*,*,バン,盤,盤
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 2 -1D 0/0 0
持っ 動詞,一般,*,*,五段-タ行,連用形-促音便,モツ,持つ,持っ
て 助詞,接続助詞,*,*,*,*,テ,て,て
い 動詞,非自立可能,*,*,上一段-ア行,連用形-一般,イル,居る,い
ませ 助動詞,*,*,*,助動詞-マス,未然形-一般,マス,ます,ませ
ん 助動詞,*,*,*,助動詞-ヌ,終止形-撥音便,ズ,ず,ん
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S Parallel 0 1 "駒"
#! SEGMENT_S Parallel 2 3 "盤"
#! GROUP_S Parallel 0 1 ""
```

■時間情報　時間情報のアノテーションは国語研でアノテーションが進められた (小西ほか (2013), 保田ほか (2013))。時間情報表現 (TIMEX3)・事象表現 (EVENT) を SEGMENT(_S) で切り出し、時間的順序関係 (TLINK) を LINK(_S) で表現する。名前空間として bccwj-timebank を用いる。

例: PN4b_00001 (bccwj-timebank)

```
* 0 2D 0/0 0
　空白,*,*,*,*,*,,　,
「 補助記号,括弧開,*,*,*,*,*,,「,「
十 名詞,数詞,*,*,*,*,ジュウ,十,十
一 名詞,数詞,*,*,*,*,イチ,一,一
月 名詞,普通名詞,助数詞可能,*,*,*,ガツ,月,月
に 助詞,格助詞,*,*,*,*,ニ,に,に
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 1 2D 0/0 0
ジョージ 名詞,固有名詞,人名,一般,*,*,ジョージ,ジョージ,ジョージ
・ 補助記号,一般,*,*,*,*,,・,・
ブッシュ 名詞,固有名詞,人名,一般,*,*,ブッシュ,ブッシュ,ブッシュ
を 助詞,格助詞,*,*,*,*,ヲ,を,を
* 2 -1D 0/0 0
倒す 動詞,一般,*,*,五段-サ行,終止形-一般,タオス,倒す,倒す
ぞ 助詞,終助詞,*,*,*,*,ゾ,ぞ,ぞ
」 補助記号,括弧閉,*,*,*,*,,」,」
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-timebank:TIMEX3 2 5 "１１月"
#! ATTR bccwj-timebank:tid "t4"
#! ATTR bccwj-timebank:type "DATE"
#! ATTR bccwj-timebank:definite "FALSE"
#! ATTR bccwj-timebank:valueFromSurface "XXXX-11"
#! ATTR bccwj-timebank:value "2004-11"
#! SEGMENT_S bccwj-timebank:EVENT 17 19 "倒す"
#! ATTR bccwj-timebank:class "OCCURRENCE"
#! ATTR bccwj-timebank:eid "e8"
#! ATTR bccwj-timebank:modality "UNDEF"
#! ATTR bccwj-timebank:tense "UNDEF"
#! ATTR bccwj-timebank:eventID "e8"
#! ATTR bccwj-timebank:aspect "UNDEF"
#! ATTR bccwj-timebank:cardinality "UNDEF"
#! ATTR bccwj-timebank:polarity "UNDEF"
#! ATTR bccwj-timebank:signalID "UNDEF"
#! ATTR bccwj-timebank:eiid "ei8"
#! ATTR bccwj-timebank:nf_morph "UNDEF"
#! LINK_S bccwj-timebank:TLINK 0 1 ""
#! ATTR bccwj-timebank:task "T2E"
#! ATTR bccwj-timebank:relatedToEventInstance "ei8"
#! ATTR bccwj-timebank:relTypeA "contains"
#! ATTR bccwj-timebank:relTypeB "contains"
#! ATTR bccwj-timebank:relTypeC "contains"
#! ATTR bccwj-timebank:timeID "t4"
EOS
```

■拡張固有表現　拡張固有表現は東工大においてアノテーションされた (橋本・中村 (2010))。固有表現の階層構造を属性値に level1-4 というキーを与え表現する。名前空間として bccwj-ene を用いる。

例: OC11_00001 (bccwj-ene)

```
* 0 3D 0/0 0
ラーメン 名詞,普通名詞,一般,*,*,*,ラーメン,ラーメン,ラーメン
花月 名詞,普通名詞,一般,*,*,*,カゲツ,花月,花月
の 助詞,格助詞,*,*,*,*,ノ,の,の
* 1 3D 0/0 0
ホッピー 名詞,固有名詞,一般,*,*,*,ホッピー,ホッピー,ホッピー
の 助詞,格助詞,*,*,*,*,ノ,の,の
* 2 3D 0/0 0
百 名詞,数詞,*,*,*,*,ヒャク,百,百
円 名詞,普通名詞,助数詞可能,*,*,*,エン,円,円
で 助詞,格助詞,*,*,*,*,デ,で,で
* 3 5D 0/0 0
焼酎 名詞,普通名詞,一般,*,*,*,ショウチュウ,焼酎,焼酎
追加 名詞,普通名詞,サ変可能,*,*,*,ツイカ,追加,追加
サービス 名詞,普通名詞,サ変可能,*,*,*,サービス,サービス,サービス
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 4 5D 0/0 0
何 名詞,数詞,*,*,*,*,ナン,何,何
回 名詞,普通名詞,助数詞可能,*,*,*,カイ,回,回
まで 助詞,副助詞,*,*,*,*,マデ,まで,まで
* 5 -1D 0/0 0
頼める 動詞,一般,*,*,下一段-マ行,連体形-一般,タノム,頼む,頼める
の 助詞,準体助詞,*,*,*,*,ノ,の,の
でしょう 助動詞,*,*,*,助動詞-デス,意志推量形,デス,です,でしょう
か 助詞,終助詞,*,*,*,*,カ,か,か
？ 補助記号,句点,*,*,*,*,,？,？
#! SEGMENT_S bccwj-ene:Name 0 6 "ラーメン花月"
#! ATTR bccwj-ene:level1 "名前"
#! ATTR bccwj-ene:level2 "施設名"
#! ATTR bccwj-ene:level3 "ＧＯＥ"
#! ATTR bccwj-ene:level4 "ＧＯＥ_その他"
#! ATTR bccwj-ene:ENE "GOE_Other"
#! SEGMENT_S bccwj-ene:Name 7 11 "ホッピー"
#! ATTR bccwj-ene:level1 "名前"
#! ATTR bccwj-ene:level2 "製品名"
#! ATTR bccwj-ene:level3 "食べ物名"
#! ATTR bccwj-ene:level4 "食べ物名_その他"
#! ATTR bccwj-ene:ENE "Food_Other"
#! SEGMENT_S bccwj-ene:Numex 12 14 "１００円"
#! ATTR bccwj-ene:level1 "数値表現"
#! ATTR bccwj-ene:level2 "金額表現"
#! ATTR bccwj-ene:ENE "Money"
```

■語義　語義は東工大においてアノテーションされた (奥村ほか (2013))。SENSEVAL-2 Japanese WSD タスクとして語義曖昧性解消のベンチマークデータとして利用された (奥村・白井 (2009))。セグメントとして文字列を切り出し、属性として語義を表す ID を付与する。ID は岩波国語辞典の語釈文に対応しており、これを属性値 explanation として展開する。名前空間として bccwj-wsd を用いる。

例: PN1c_00001 (bccwj-wsd)

```
* 0 1D 0/0 0
キャスター 名詞,普通名詞,一般,*,*,*,キャスター,キャスター,キャスター
　空白,*,*,*,*,*,,　,
* 1 -1D 0/0 0
蓮舫 名詞,固有名詞,人名,一般,*,*,レンホウ,レンホウ,蓮舫
さん 接尾辞,名詞的,一般,*,*,*,サン,さん,さん
#! SEGMENT_S bccwj-wsd:wsd 0 5 "キャスター"
#! ATTR bccwj-wsd:sense "11335-0-0-1-0"
#! ATTR bccwj-wsd:explanation "&lt;1&gt; ●ニュースキャスター。"
EOS
```

■項構造シソーラス　項構造シソーラスの情報は岡山大学でアノテーションされている (竹内・上野 (2013))。述語と項をセグメントとして切り出し、述語項関係をリンクで表現する。名前空間として bccwj-pth を用いる。本アノテーションは複数人によりアノテーションされており、アノテータを識別するための userid が属性値に含まれている。名前空間としてはアノテーションを識別するが、複数のアノテータを識別する機構がなく属性値を利用する。

例: PB12_00001 (bccwj-pth)

```
* 0 2D 0/0 0
単に 副詞,*,*,*,*,*,タンニ,単に,単に
* 1 2D 0/0 0
コーヒー 名詞,普通名詞,一般,*,*,*,コーヒー,コーヒー,コーヒー
と 助詞,格助詞,*,*,*,*,ト,と,と
* 2 4D 0/0 0
間違え 動詞,非自立可能,*,*,下一段-ア行,連用形-一般,マチガエル,間違える,間違え
```

```
て 助詞,接続助詞,*,*,*,*,テ,て,て
* 3 4D 0/0 0
紅茶 名詞,普通名詞,一般,*,*,*,コウチャ,紅茶,紅茶
を 助詞,格助詞,*,*,*,*,ヲ,を,を
* 4 7D 0/0 0
飲ん 動詞,一般,*,*,五段-マ行,連用形-撥音便,ノム,飲む,飲ん
だ 助動詞,*,*,*,助動詞-タ,連体形-一般,タ,た,だ
だけ 助詞,副助詞,*,*,*,*,ダケ,だけ,だけ
で 助動詞,*,*,*,助動詞-ダ,連用形-一般,ダ,だ,で
、 補助記号,読点,*,*,*,*,,、,、
* 5 6D 0/0 0
衝撃 名詞,普通名詞,一般,*,*,*,ショウゲキ,衝撃,衝撃
的 接尾辞,形状詞的,*,*,*,*,テキ,的,的
な 助動詞,*,*,*,助動詞-ダ,連体形-一般,ダ,だ,な
* 6 7D 0/0 0
味覚 名詞,普通名詞,一般,*,*,*,ミカク,味覚,味覚
体験 名詞,普通名詞,サ変可能,*,*,*,タイケン,体験,体験
に 助詞,格助詞,*,*,*,*,ニ,に,に
* 7 -1D 0/0 0
なっ 動詞,非自立可能,*,*,五段-ラ行,連用形-促音便,ナル,成る,なっ
て 助詞,接続助詞,*,*,*,*,テ,て,て
しまう 動詞,非自立可能,*,*,五段-ワア行,終止形-一般,シマウ,仕舞う,しまう
と 助詞,格助詞,*,*,*,*,ト,と,と
は 助詞,係助詞,*,*,*,*,ハ,は,は
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-pth:semrole 0 2 "単に"
#! SEGMENT_S bccwj-pth:semrole 2 11 "コーヒーと間違えて"
#! SEGMENT_S bccwj-pth:pred 7 9 "間違え"
#! SEGMENT_S bccwj-pth:semrole 11 14 "紅茶を"
#! SEGMENT_S bccwj-pth:pred 14 16 "飲ん"
#! LINK_S bccwj-pth:semrole6 4 0 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 0 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
#! LINK_S bccwj-pth:semrole6 4 1 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 1 ""
#! ATTR bccwj-pth:role "副詞相当"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
#! LINK_S bccwj-pth:semrole6 4 3 ""
#! ATTR bccwj-pth:role "対象"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "6"
#! LINK_S bccwj-pth:semrole9 4 3 ""
#! ATTR bccwj-pth:role "対象"
#! ATTR bccwj-pth:flag "0"
#! ATTR bccwj-pth:userid "9"
EOS
```

■れる・られるの用法　国語研により助動詞れる・られるの用法をアノテーションした(小山田ほか (2012))。対象となる助動詞をセグメントで切り出し、属性情報として用法分類のラベルを付与する。名前空間として bccwj-auxv を用いる。

例: PN2f_00003 (bccwj-auxv)

```
* 0 2D 0/0 0
俺 代名詞,*,*,*,*,*,オレ,俺,俺
は 助詞,係助詞,*,*,*,*,ハ,は,は
* 1 2D 0/0 0
もう 副詞,*,*,*,*,*,モウ,もう,もう
* 2 3D 0/0 0
やら 動詞,非自立可能,*,*,五段-ラ行,未然形-一般,ヤル,遣る,やら
れ 助動詞,*,*,*,助動詞-レル,連用形-一般,レル,れる,れ
た 助動詞,*,*,*,助動詞-タ,終止形-一般,タ,た,た
と 助詞,格助詞,*,*,*,*,ト,と,と
* 3 -1D 0/0 0
思っ 動詞,一般,*,*,五段-ワア行,連用形-促音便,オモウ,思う,思っ
た 助動詞,*,*,*,助動詞-タ,終止形-一般,タ,た,た
ね 助詞,終助詞,*,*,*,*,ネ,ね,ね
。 補助記号,句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-auxv:auxv 6 7 "れ"
#! ATTR bccwj-auxv:label "受身"
#! ATTR bccwj-auxv:direct "直接"
#! ATTR bccwj-auxv:sentiment "有情"
#! ATTR bccwj-auxv:agentivity "-"
#! ATTR bccwj-auxv:projection "-"
#! ATTR bccwj-auxv:existence "-"
#! ATTR bccwj-auxv:inducement "-"
EOS
```

■日本語フレームネット　日本語フレームネットは慶應義塾大学でアノテーションされている (小原 (2013))。本稿で説明するアノテーションの中で最も複雑な構造を持っている。フレームに参加する表現をセグメントで切り出し、その代表表現の属性情報 FE-feID にフレーム辞書へのリファレンスを含める。セグメント間の関係はフレームに参加するセグメント全てについてグループを定義して表現する。名前空間として bccwj-jfn を用いる。項構造シソーラスと同様に複数人によりアノテーションされており、属性情報にアノ

テータの情報が含まれている場合がある。

例: PB42_00003 (bccwj-jfn)

```
* 0 1D 0/0 0
マザー 名詞, 普通名詞, 一般,*,*,*, マザー, マザー, マザー
テレサ 名詞, 固有名詞, 人名, 一般,*,*, テレサ, テレサ, テレサ
の 助詞, 格助詞,*,*,*,*, ノ, の, の
* 1 -1D 0/0 0
施設 名詞, 普通名詞, サ変可能,*,*,*, シセツ, 施設, 施設
にて 助詞, 格助詞,*,*,*,*, ニテ, にて, にて
#! SEGMENT_S bccwj-jfn:SEG 0 7 "マザーテレサの"
#! ATTR bccwj-jfn:Lemma-name "N"
#! ATTR bccwj-jfn:FE-feID "1598"
#! ATTR bccwj-jfn:FE-name "Name"
#! SEGMENT_S bccwj-jfn:SEG 7 9 "施設"
#! ATTR bccwj-jfn:Lemma-name "N"
#! ATTR bccwj-jfn:Target-name "Target"
#! ATTR bccwj-jfn:GF-name "Target"
#! ATTR bccwj-jfn:PT-name "Target"
#! ATTR bccwj-jfn:Postpos-name "Target"
#! GROUP_S bccwj-jfn:Frame 0 1 "0"
EOS
```

■拡張モダリティ　拡張モダリティは奈良先端大・東北大でアノテーションされている (松吉ほか (2011))。対象となる事象表現をセグメントで切り出し、属性情報として用法分類のラベルなどを付与する。名前空間として bccwj-eme を用いる。

例: PB49_00005 (bccwj-eme)

```
* 0 2D 0/0 0
　空白,*,*,*,*,*,,　,
平七郎 名詞, 固有名詞, 人名, 名,*,*, ヘイシチロウ, ヘイシチロウ, 平七郎
は 助詞, 係助詞,*,*,*,*, ハ, は, は
* 1 2D 0/0 0
花川戸 名詞, 固有名詞, 地名, 一般,*,*, ハナカワド, ハナカワド, 花川戸
河岸 名詞, 普通名詞, 一般,*,*,*, カシ, 河岸, 河岸
へ 助詞, 格助詞,*,*,*,*, へ, へ, へ
と 助詞, 格助詞,*,*,*,*, ト, と, と
* 2 -1D 0/0 0
急い 動詞, 一般,*,*, 五段-ガ行, 連用形-イ音便, イソグ, 急ぐ, 急い
で 助詞, 接続助詞,*,*,*,*, テ, て, で
い 動詞, 非自立可能,*,*, 上一段-ア行, 連用形-一般, イル, 居る, い
た 助動詞,*,*,*, 助動詞-タ, 終止形-一般, タ, た, た
。 補助記号, 句点,*,*,*,*,,。,。
#! SEGMENT_S bccwj-eme:Event 12 14 "急い"
#! ATTR bccwj-eme:morphIDs "80"
#! ATTR bccwj-eme:pseudo ""
#! ATTR bccwj-eme:source "wr:筆者"
#! ATTR bccwj-eme:time "非未来"
#! ATTR bccwj-eme:conditional "0"
#! ATTR bccwj-eme:pmtype "叙述"
#! ATTR bccwj-eme:actuality "成立"
#! ATTR bccwj-eme:evaluation "0"
#! ATTR bccwj-eme:focus ""
#! ATTR bccwj-eme:degree ""
#! SEGMENT_S bccwj-eme:Event 15 16 "い"
#! ATTR bccwj-eme:morphIDs "100"
#! ATTR bccwj-eme:pseudo ""
#! ATTR bccwj-eme:source "wr:筆者"
#! ATTR bccwj-eme:time "非未来"
#! ATTR bccwj-eme:conditional "0"
#! ATTR bccwj-eme:pmtype "叙述"
#! ATTR bccwj-eme:actuality "成立"
#! ATTR bccwj-eme:evaluation "0"
#! ATTR bccwj-eme:focus ""
#! ATTR bccwj-eme:degree ""
EOS
```

■否定の焦点　否定の焦点は山梨大でアノテーションされている (松吉ほか (2013))。対象となる否定辞・焦点・手がかり句をセグメントで切り出し、属性情報として用法分類のラベルなどを付与し、否定辞-焦点間, 焦点-手がかり句間にリンクを付与する。名前空間として bccwj-wsb を用いる。

例: PN2f_00003 (bccwj-wsb)

```
* 0 1D 0/0 0
　空白,*,*,*,*,*,,　,
カ 名詞, 普通名詞, 一般,*,*,*, チカラ, カ,*, チカラ
を 助詞, 格助詞,*,*,*,*, ヲ, を,*, オ
* 1 2D 0/0 0
出し 動詞, 非自立可能,*,*, 五段-サ行, 連用形-一般, ダス, 出す,*, ダシ
切っ 動詞, 非自立可能,*,*, 五段-ラ行, 連用形-促音便, キル, 切る,*, キッ
て 助詞, 接続助詞,*,*,*,*, テ, て,*, テ
* 2 3D 0/0 0
敗れ 動詞, 一般,*,*, 下一段-ラ行, 連用形-一般, ヤブレル, 敗れる,*, ヤブレ
た 助動詞,*,*,*, 助動詞-タ, 連体形-一般, タ, た,*, タ
* 3 -1Z 0/0 0
わけ 名詞, 普通名詞, 一般,*,*,*, ワケ, 訳,*, ワケ
で 助詞, 格助詞,*,*,*,*, デ, で,*, デ
は 助詞, 係助詞,*,*,*,*, ハ, は,*, ワ
ない 形容詞, 非自立可能,*,*, 形容詞, 終止形-一般, ナイ, 無い,*, ナイ
。 補助記号, 句点,*,*,*,*,,。,*,
#! SEGMENT_S bccwj-wsb:Negation 15 17 "ない"
#! ATTR bccwj-wsb:morphID "4450"
#! ATTR bccwj-wsb:POS "否定複合辞"
#! SEGMENT_S bccwj-wsb:Focus 5 7 "切っ"
#! ATTR bccwj-wsb:morphID "4380"
#! ATTR bccwj-wsb:argTypes "テ節; 動詞句"
```

```
#! ATTR bccwj-wsb:class "付加-連用修飾"
#! ATTR bccwj-wsb:description "様態の副詞句が使用されている"
#! SEGMENT_S bccwj-wsb:Clue 11 17 "わけではない"
#! ATTR bccwj-wsb:morphIDs "4420.4430.4440.4450"
#! LINK_S bccwj-wsb:FocusOfNegation 0 1 ""
#! LINK_S bccwj-wsb:Clue 1 2 ""
EOS
```

■述語項構造・照応関係　述語項構造は奈良先端大でアノテーションされている (小町・飯田 (2011))。対象となる述語と項をセグメントで切り出し、述語-項間にリンクを付与する。また照応関係を同一の実体を指し示す同値類として捉えてグループを付与する。名前空間として bccwj-pas を用いる。

例: PB40_00003 (bccwj-pas)

```
* 0 1D 0/0 0
客先 名詞，普通名詞，一般,*,*,*，キャクサキ，客先，客先
の 助詞，格助詞,*,*,*,*,*，ノ，の，の
* 1 2D 0/0 0
安全 名詞，普通名詞，形状詞可能,*,*,*，アンゼン，安全，安全
基準 名詞，普通名詞，一般,*,*,*，キジュン，基準，基準
が 助詞，格助詞,*,*,*,*,*，ガ，が，が
* 2 -1D 0/0 0
厳しく 形容詞，一般,*,*，形容詞，連用形-一般，キビシイ，厳しい，厳しく
て 助詞，接続助詞,*,*,*,*,*，テ，て，て
* 3 4D 0/0 0
守れ 動詞，一般,*,*，下一段-ラ行，連用形-一般，マモル，守る，守れ
そう 形状詞，助動詞語幹,*,*,*,*,*，ソウ，そう，そう
に 助動詞,*,*,*,*，助動詞-ダ，連用形-ニ，ダ，だ，に
* 4 -1D 0/0 0
ない 形容詞，非自立可能,*,*，形容詞，終止形-一般，ナイ，無い，ない
EOS
    :
#! SEGMENT bccwj-pas:Np 0 2 "客先"
#! SEGMENT bccwj-pas:Np 5 7 "基準"
#! SEGMENT bccwj-pas:Pred 8 11 "厳しく"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT bccwj-pas:Pred 12 14 "守れ"
#! ATTR bccwj-pas:type "述語"
    :
#! LINK bccwj-pas:Ga 2 1 ""
#! LINK bccwj-pas:O 3 1 ""
    :
#! GROUP bccwj-pas:Coref 0 36 37 ""
```

外界照応のような参照対象が文書中に出現しないようなリンクは、セグメントに対する属性表現 Exophora_Ga により表す。

```
#! SEGMENT bccwj-pas:Pred 392 395 ""
#! ATTR bccwj-pas:type "述語"
#! ATTR bccwj-pas:Exophora_Ga "その他"
```

4.2　重ね合わせ事例

アノテーションを重ね合わせた事例を稿末の付録Ａ．に示す。時間情報 (bccwj-timebank) と語義 (bccwj-wsd) と述語項構造 (bccwj-pas) と日本語フレームネット (bccwj-jfn) を重ね合わせたものである。また、ChaKi.NET による可視化事例についても稿末に示す。セグメント相当にマウスカーソルを合わせることで属性情報を見ることが可能である。

5.　おわりに

本稿では拡張 CaboCha フォーマットに対して、属性情報と名前空間を表現する機構を新たに追加することを提案し、BCCWJ に対して付与されている十種類のアノテーションについてどのように表現できるか例を示した。言語学・言語処理双方の利用にとって、このような表現が有効か否かについて広く意見を求めたい。

今後、アノテーション開発者・配布者・利用者の意見を伺いながら、仕様の最終案を策定し、各アノテーションの重ね合わせを進めたい。また、コーパス管理ツール ChaKi.NET (Matsumoto et al. (2005)) に重ね合わせたアノテーション情報に対する検索・問い合わせを行う機構を実装し、利用者系の整備も進めていきたいと考える。

謝辞

本研究の一部は科研費若手研究 (B)「否定焦点コーパス構築と焦点自動解析に関する研究」（課題番号：25870278、代表：松吉俊）、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

浅原正幸・松本裕治 (2013).「『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.

橋本泰一・中村俊一 (2010).「拡張固有表現タグ付きコーパスの構築」　言語処理学会第 16 回年次大会発表論文集.

小町守・飯田龍 (2011).「BCCWJ に対する述語項構造と照応関係のアノテーション」　『現代日本語書き言葉均衡コーパス』完成記念講演会.

小西光・浅原正幸・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパス』に対する時間情報アノテーション」　自然言語処理, 20:2, pp. 201–222.

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)*.

松吉俊・佐尾ちとせ・乾健太郎・松本裕治 (2011).「拡張モダリティタグ付与コーパスの設計と構築」　言語処理学会第 17 回年次大会発表論文集.

松吉俊・大槻諒・福本文代 (2013).「日本語における否定の焦点アノテーション」　第 3 回コーパス日本語学ワークショップ予稿集.

小原京子 (2013).「日本語フレームネット：文意理解のためのコーパスアノテーション」　言語処理学会第 19 回年次大会発表論文集.

奥村学・白井清昭 (2009).「BCCWJ を用いた新しい語義曖昧性解消タスク」　言語処理学会第 15 回年次大会発表論文集.

奥村学・白井清昭・古宮嘉那子・横野光 (2013).「BCCWJ 上の語義タグ付コーパス」　言語処理学会第 19 回年次大会（ライトニングトーク）.

小山田由紀・柏野和佳子・前川喜久雄 (2012).「助動詞レル・ラレルへの意味アノテーション作業経過報告」　第 2 回コーパス日本語学ワークショップ予稿集.

竹内孔一・上野真幸 (2013).「日本語コーパスに対する動詞項構造シソーラスの概念と意味役割のアノテーション」　言語処理学会第 19 回年次大会発表論文集.

保田祥・小西光・浅原正幸・今田水穂・前川喜久雄 (2013).「『現代日本語書き言葉均衡コーパス』に対する時間表現・事象表現間の時間的順序関係アノテーション」　自然言語処理, 20:5, pp. 657–682.

関連 URL

- 「現代日本語書き言葉均衡コーパス」Web ページ：`http://www.ninjal.ac.jp/corpus_center/bccwj/`

- 「CaboCha」 Web ページ：https://code.google.com/p/cabocha/
- 「ChaKi.NET」 Web ページ：http://sourceforge.jp/projects/chaki/releases/
- 「BCCWJ-DepPara」Web ページ：https://github.com/masayu-a/BCCWJ-DepPara
- 「BCCWJ-DepPara2」Web ページ：https://github.com/masayu-a/BCCWJ-DepPara2
- 「BCCWJ-TimeBank」Web ページ：https://github.com/masayu-a/BCCWJ-Timebank
- 「拡張固有表現タグ付きコーパス」Web ページ：http://www.gsk.or.jp/catalog/gsk2013-b/
- “SENSEVAL-2 Japanese WSD task” Web ページ：http://www.lr.pi.titech.ac.jp/wsd.html
- 「岡山大学動詞項構造シソーラス」 Web ページ：http://cl.cs.okayama-u.ac.jp/rsc/data/index.html
- 「BCCWJ-AUXV」 Web ページ： https://github.com/masayu-a/BCCWJ-AUXV
- 「日本語フレームネット」Web ページ：http://jfn.st.hc.keio.ac.jp/ja/
- 「拡張モダリティタグ付与コーパス」Web ページ：http://cl.cs.yamanashi.ac.jp/nldata/modality/
- 「否定の焦点アノテーション」Web ページ：http://cl.cs.yamanashi.ac.jp/nldata/negation/
- 「BCCWJ:述語項構造と照応関係のアノテーション」Web ページ：http://cl.naist.jp/nldata/bccwj/pas/

付録A. 重ね合わせの例

例: PB22_00002

```
* 0 1D 0/0 0
若く 形容詞，一般,*,*，形容詞，連用形-一般，ワカイ，若い，若く
* 1 2D 0/0 0
し 動詞，非自立可能,*,*，サ行変格，連用形-一般，スル，為る，し
て 助詞，接続助詞,*,*,*,*，テ，て，て
* 2 3D 0/0 0
散っ 動詞，一般,*,*，五段-ラ行，連用形-促音便，チル，散る，散っ
た 助動詞,*,*,*，助動詞-タ，連体形-一般，タ，た，た
* 3 -1D 0/0 0
夢 名詞，普通名詞，一般,*,*,*，ユメ，夢，夢
#! SEGMENT_S bccwj-timebank:event 0 2 "若く"
#! ATTR bccwj-timebank:eid "e1"
#! ATTR bccwj-timebank:class "STATE"
#! SEGMENT_S bccwj-timebank:event 2 3 "し"
#! ATTR bccwj-timebank:eid "e2"
#! ATTR bccwj-timebank:class ""
#! SEGMENT_S bccwj-timebank:event 4 7 "散った"
#! ATTR bccwj-timebank:eid "e3"
#! ATTR bccwj-timebank:class ""
#! SEGMENT_S bccwj-wsd:wsd 4 6 "散っ"
#! ATTR bccwj-wsd:sense "33687-0-0-4-0"
#! ATTR bccwj-wsd:explanation "&lt;4&gt; 比ゆ的に、人がいさぎよく死ぬ。多く、戦死にいう。「花と―」"
#! SEGMENT_S bccwj-wsd:wsd 7 8 "夢"
#! ATTR bccwj-wsd:sense "52859-0-0-3-0"
#! ATTR bccwj-wsd:explanation "&lt;3&gt;（現在のところ実現してはいないが）将来は実現させたい願い・理想。「彼の―は大きい」「―を描く」"
#! SEGMENT_S bccwj-pas:Pred 2 3 "し"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT_S bccwj-pas:Pred 4 6 "散っ"
#! ATTR bccwj-pas:type "述語"
#! SEGMENT_S bccwj-pas:Np 7 8 "夢"
#! SEGMENT_S bccwj-jfn:JFN 4 6 "散っ"
#! ATTR bccwj-jfn:Lemma-name "V"
#! ATTR bccwj-jfn:Target-name "Target"
#! ATTR bccwj-jfn:GF-name "Target"
#! ATTR bccwj-jfn:PT-name "Target"
#! SEGMENT_S bccwj-jfn:JFN 7 8 "夢"
#! ATTR bccwj-jfn:Lemma-name "V"
#! ATTR bccwj-jfn:FE-feID "6803"
#! ATTR bccwj-jfn:FE-name "Entity"
#! LINK_S bccwj-pas:Ga 6 7 ""
#! GROUP_S bccwj-jfn:JFN 8 9 "0"
EOS
```