『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

小西 光[†]・浅原 正幸[†]・前川喜久雄[†]

時間情報表現は、テキスト中に記述される事象の生起時刻を推定するための重要な手がかりである。時間情報表現を含む数値表現の抽出は、固有表現抽出の部分問題として解かれてきた。英語においては、評価型国際会議が開かれ、時間情報表現のテキストからの切り出しだけでなく、曖昧性解消・正規化のための様々な手法が提案されている。さらに、時間情報と事象とを関連づけるアノテーション（タグづけ）基準 TimeML の定義や新聞記事にアノテーションを行ったコーパス TimeBank の整備が進んでいる。一方、日本語においては時間情報処理に必要なアノテーション基準の定義及びコーパスの整備が進んでいない。本稿では、TimeML の時間情報表現を表す〈TIMEX3〉タグに基づいた時間情報のアノテーション基準を日本語向けに再定義し、『現代日本語書き言葉均衡コーパス』（BCCWJ）コアデータの一部にアノテーションを行った。問題点を検討し、今後事象の生起時刻を推定するために必要な課題を考察する。

キーワード：時間情報処理，コーパスアノテーション

Temporal Information Annotation on ‘Balanced Corpus of Contemporary Written Japanese’

HIKARI KONISHI[†], MASAYUKI ASAHARA[†] and KIKUO MAEKAWA[†]

Temporal information is important for grounding event expressions on a timeline. Temporal expression extraction has been performed as numerical representation extraction, which is a subtask of named entity extraction. For English texts, evaluation workshops were held in which temporal expressions were extracted and normalized. An annotation schema, *TimeML*, was designed to annotate events and temporal expressions, and several annotated corpora of newswire texts were developed. However, a schema for temporal information and normalization of Japanese texts has not been designed. This paper proposes an annotation schema, which is based on *TimeML*, for Japanese temporal information. We annotate the temporal information in parts of the ‘Balanced Corpus of Contemporary Written Japanese’. We identify several problems in the annotation and discuss the steps to be taken to ground Japanese event expressions on a timeline.

Key Words: *Temporal Information Processing, Corpus Annotation*

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

1 はじめに

情報検索や情報抽出において、テキスト中に示される事象を実時間軸上の時点もしくは時区間に関連づけることが求められている。Web 配信されるテキスト情報に関しては、文書作成日時 (Document Creation Time: DCT) が得られる場合、テキスト情報と文書作成日時とを関連づけることができる。しかしながら、文書作成日時が得られない場合や、文書に記述されている事象発生日時と文書作成日時が乖離する場合には他の方策が必要である。テキスト中に記述されている時間情報解析の精緻化が求められている。

時間表現抽出は、固有表現抽出の部分問題である数値表現抽出のタスクとして研究されてきた。英語においては、評価型国際会議 MUC-6 (the sixth in a series of Message Understanding Conference)(Grishman and Sundheim 1996) で、アノテーション済み共有データセットが整備され、そのデータを基に各種の系列ラベリングに基づく時間表現の切り出し手法が開発されてきた。TERN (Time Expression Recognition and Normalization)(DARPA TIDES 2004) では、時間情報の曖昧性解消・正規化がタスクとして追加され、様々な時間表現解析器が開発された。さらに、時間情報表現と事象表現とを関連づけるアノテーション基準 TimeML (Pustejovsky, Hanks, Saurí, See, Gaizauskas, Setzer, Radev, Sundheim, Day, Ferro, and Lazo 2003b) が検討され、TimeML に基づくタグつきコーパス TimeBank (Pustejovsky, Castaño, Ingria, Saurí, Gaizauskas, Setzer, and Katz 2003a) などが整備された。2007 年には、時間情報表現一事象表現間及び 2 事象表現間の時間的順序関係を推定する評価型ワークショップ SemEval-2007 のサブタスク TempEval (Verhagen, Gaizauskas, Schilder, Hepple, Katz, and Pustejovsky 2007) が開かれ、種々の時間的順序関係推定器が開発された。後継のワークショップ SemEval-2010 のサブタスク TempEval-2 (Verhagen, Saurí, Caselli, and Pustejovsky 2010) では、英語だけでなく、イタリア語、スペイン語、中国語、韓国語を含めた 5 言語が対象となった。2013 年に開かれる SemEval-2013 のサブタスク TempEval-3 では、データを大規模化した英語、スペイン語が対象となっている。

一方、日本語においては IREX (Information Retrieval and Extraction Exercise) ワークショップ (IREX 実行委員会 1999) の固有表現抽出タスクの部分問題として時間情報表現抽出が定義されているのみで、時間情報の曖昧性解消・正規化に関するデータが構築されていなかった。

そこで、我々は TimeML に基づいた日本語に対する時間情報アノテーション基準を定義し、時間情報の曖昧性解消・正規化を目的とした時間情報タグつきコーパスを構築した。他言語のコーパスが新聞記事のみを対象としているのに対し、本研究では均衡コーパスである『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese; 以下 “BCCWJ”) を対象としており、新聞記事だけでなく、一般書籍・雑誌・ブログなどに出現する多様な時間情報表現を対象としている。本稿ではアノテーション基準を示すとともに、アノテーションしたコーパスの詳細について示す。

以下, 2 節では時間情報表現についての背景について概観する. 3 節では, 対象とする時間情報表現について詳しく述べる. 4 節では策定した日本語時間情報表現に対するアノテーション基準を示す. 5 節でアノテーションにおける日本語特有の問題について説明する. 6 節でアノテーション作業環境を示す. 7 節で実際にアノテーションしたコーパスの分析を行う. 最後にまとめと今後の課題を示す.

2 背景

2.1 時間情報表現に関する関連研究

テキスト中の時間情報表現を分析する研究は日本語以外の言語で進んでおり, 時間表現の文字列の切り出しや正規化のみならず, 時間表現と事象表現の関連づけなどが行われている. 表 1 に英語もしくは日本語を対象とした時間情報表現に関連する研究を示す. 以下, まず英語の時間情報表現に関する代表的な研究を俯瞰する. 次に日本語の数少ない時間情報表現に関する研究を示す.

英語においては, 評価型国際会議 MUC-6 (Grishman and Sundheim 1996) の 1 タスクである

表 1 関連研究

英語の時間情報表現に関する関連研究		
MUC-6 (Grishman and Sundheim 1996)	評価型会議	時間情報表現の切り出しのみ
(Setzer 2001)	基準	時間情報表現の切り出しと正規化
TERN (TIMEX2) タグ (DARPA TIDES 2004)	評価型会議	時間情報表現の切り出しと正規化
TimeML (TIMEX3) タグ (Pustejovsky et al. 2003b)	基準	時間情報表現の切り出しと正規化
TimeML	基準	事象間の時間的順序関係
TimeBank (Pustejovsky et al. 2003a)	コーパス	TimeML 基準によるタグつきコーパス
Aquaint TimeML Corpus (Boguraev and Kubota Ando 2005)	コーパス	TimeML 基準によるタグつきコーパス
(Mani 2006)	解析手法	時間情報表現一事象間の時間的順序関係解析
TempEval (Verhagen et al. 2007)	解析手法	二事象間の時間的順序関係解析
TempEval-2 (Verhagen et al. 2010)	評価型会議	時間情報表現一事象間／二事象間の時間的順序関係解析
TempEval-3 (2013 年実施)	評価型会議	時間情報表現一事象間／二事象間の時間的順序関係解析
日本語の時間情報表現に関する関連研究		
IREX (IREX 実行委員会 1999)	評価型会議	時間情報表現の切り出しのみ
拡張固有表現体系 (Sekine et al. 2002)	タグつけ基準	時間情報表現の切り出しのみ
(橋本, 中村 2010)	コーパス	時間情報表現の切り出しのみ
本論文	コーパス	時間情報表現の切り出しと正規化

固有表現抽出の中に時間情報表現の抽出が含まれていた。MUC-6 で定義されている時間情報表現タグ (TIMEX) は日付表現 (@type="DATE") と時刻表現 (type="TIME") からなる。アノテーション対象は絶対的な日付・時刻を表す表現にのみ限定され、"last year" などといった相対的な日付・時刻表現は含まれていない。この MUC-6 のアノテーション基準 (TIMEX) に対し、Setzer は時間情報表現の正規化に関するアノテーション基準を提案している (Setzer 2001)。評価型国際会議 TERN (DARPA TIDES 2004) では、時間情報表現検出に特化したタスクを設定している。TERN で定義された時間表現情報タグ (TIMEX2) は、相対的な日付・時刻表現、時間表現や頻度集合表現が検出対象として追加されている。時間表現の正規化情報を記述する ISO-8601 形式を拡張した @value 属性などが設計され、こちらも自動解析対象となっている。その後、Pustejovsky らによりアノテーション基準 TimeML (Pustejovsky et al. 2003b) が提案されている。その中では、TERN で用いられている (TIMEX2) を拡張した (TIMEX3) が提案され、さらに時間情報表現と事象表現の時間的順序関係を関連づけるための情報が付加される。これらの情報は人手でアノテーションすることを目的に設計され、TimeBank (Pustejovsky et al. 2003a) や Aquaint TimeML Corpus などの人手によるタグつきコーパスの整備が行われた。これらのコーパスに基づく時間情報表現の自動解析 (Boguraev and Kubota Ando 2005; Mani 2006) が試みられたが、タグの情報に不整合があったり、付与されている時間的順序関係ラベルに偏りがあったりなど扱いにくいものであった (Boguraev and Kubota Ando 2006)。2007 年に開かれた SemEval 2007 の 1 タスク TempEval (Verhagen et al. 2007) では、時間的順序関係のラベルを簡略化し、人手で見直したデータによる時間的順序関係同定のタスクが行われた。このタスクでは、時間表現に対して正規化された @value 属性などが付与されており、事象表現の時間的順序関係同定に利用してよい。TempEval-2 (Verhagen et al. 2010) では、英語だけでなく、イタリア語、スペイン語、中国語、韓国語に関しても同様なデータを利用したタスクが設定された。

日本語においては、IREX (IREX 実行委員会 1999) の 1 タスクとして、固有表現抽出タスクが設定された。IREX の時間情報では、日付・時刻表現を対象にし、相対的な表現が定義に含められている。また、関根らは拡張固有表現体系 (Sekine, Sudo, and Nobata 2002) を提案し、辞書／オントロジやコーパスの作成などを行っており、BCCWJ にも同じ体系の拡張固有表現タグが付与されている (橋本, 中村 2010)。日本語においては、表現の分類の体系化が進んでいるが、正規化のための研究は他言語と比べて遅れをとっている。

2.2 コーパスアノテーションの標準化

本研究はコーパスアノテーションにおける標準化活動の観点から重要である。

国際標準化機構 (International Organization for Standardization: ISO) の標準化技術委員会 (Technical Committee) TC 37 は "Terminology and other language and content resources" と題し、言語資源に関するさまざまな標準化を提案している。そのなかに分科会 (Structure of

the committee) が四つ設定されているが, TC 37/SC 4 が言語資源管理 (Language resource management; LRM) に関する国際規格の規定を行っている. TC 37/SC 4 はさらに以下に示すような作業部会を六つ設定しており, さまざまな形式・出自の一次言語データに対するアノテーションや XML に代表される汎用マークアップ言語に基づくアノテーションの表現形式についての仕様記述言語を設計している.

- TC 37/SC 4/WG 1 Basic descriptors and mechanisms for language resources
言語資源に関する情報を記述するための作業部会
- TC 37/SC 4/WG 2 Semantic annotation
アノテーションと表現方法を議論する作業部会
- TC 37/SC 4/WG 3 Multilingual information representation
多言語対訳テキストに特化した作業部会
- TC 37/SC 4/WG 4 Lexical resources
言語資源そのものに関する作業部会
- TC 37/SC 4/WG 5 Workflow of language resource management
言語資源管理の作業手順を議論する作業部会
- TC 37/SC 4/WG 6 Linguistic annotation
言語情報アノテーションを議論する作業部会

TimeML 開発者は, 作業部会 TC 37/SC 4/WG 2 と連携を取りながら, Semantic Annotation Framework(SemAF)-Time (ISO-24617-1: 2012) を 2012 年に正式に制定した.

日本の言語資源整備は, 実データを作成せずに標準化活動を行うものと実データを作成するが標準化活動を見捨てて行うものとに二極化している. 一方, 国際標準の中には, 実世界上の特定のモノもしくはコトに関係づけられる言語横断的な標準化が有効なアノテーションと, 言語の表現形態・表現機能のような言語横断的な標準化がそぐわないアノテーションとが混在している. 後者の作業の失敗が顕在化しており, アノテーションの標準化作業を低く評価する傾向がある.

我々のグループは 2006 年より TimeML 開発者から TimeML 関連の情報を得ながら時間情報表現アノテーションと事象表現アノテーションに取り組んできた. 標準化に適した時間情報表現アノテーションと, 標準化に適さない事象表現アノテーションを切り分けたうえで, 前者について ISO-TimeML に準拠する日本語版 (TIMEX3) アノテーション基準を検討し策定した. この部分が本研究の内容に相当する. 一方, 後者についてはモダリティが豊かな日本語の事象表現を国際標準に合わせてアノテーションすることが困難であり, 別の方策でアノテーションすることを検討中である.

日本において, 国際標準に準拠している他のアノテーションデータとして, 策定中の SemAF-NE Named Entities (ISO-24617-3 制定中) に準拠している BCCWJ に対する拡張固有表現体系

アノテーション (橋本, 中村 2010) がある.

3 対象とする時間情報表現

まず以下の例文を見て欲しい.

例文:

彼は **2008 年 4 月** から週に **3 回** ジョギングを **1 時間** 行ってきたが, 昨日ケガをして走れなくなり, **今朝 9 時** に病院に行った.

本稿の研究対象である時間情報表現¹は時間軸上の時点もしくは時区間を表現するテキスト中の文字列とする. 時間情報表現は以下の四つの分類に分けられる. ①日付表現・②時刻表現は「2008 年 4 月」「昨日」「今朝 9 時」といった, 時点もしくは時区間の時間軸上の位置を定義することを目的として用いられる表現である. ③時間表現は「1 時間」といった, 時間軸上の位置に焦点をあてずに時区間幅を定義することを目的として用いられる表現である. ④頻度集合表現は「週に 3 回」といった, 時間軸上複数の時区間を定義することを目的として用いられる表現である.

これら時間情報表現の曖昧性を解消しながら時間軸上の特定の区間に写像することを正規化と呼ぶ. 日付・時刻表現において, 表層の情報だけで正規化ができる表現と, 文脈の情報を用いなければ正規化ができない表現がある. 前者を定時間情報表現 (fully-specified temporal expression) と呼び, 後者を不定時間情報表現 (underspecified temporal expression) と呼ぶ. 上の例では「2008 年 4 月」が定時間情報表現であり, 「昨日」「今朝 9 時」が不定時間情報表現である. 時間情報表現の正規化には計算機で扱う日付や時刻を扱うための国際標準 ISO-8601 形式²への変換が一般的である. しかしながら, 自然言語では表現できるが, ISO-8601 形式では直接表現できない時間情報表現がある. 例えば, 時間表現や頻度集合表現は時間軸上不定な場合が多く ISO-8601 形式だけでは表現できず, TimeML の `<TIMEX3>` においてさまざまな拡張形式が提案されている.

想定する時間情報表現解析では手がかりとして, テキストが書かれた日付・時刻を表す文書作成日時を用いることを仮定している. 例えば, 文書作成日時が 2008 年 9 月 1 日であれば, 「昨日」は 2008 年 8 月 31 日 (ISO-8601 形式では “2008-08-31”) を表し, 「今朝 9 時」は 2008 年 9 月 1 日午前 9 時 (同 “2008-09-01T09:00”) を表す.

¹ 「時間情報表現」は①「日付表現」(“DATE”)・②「時刻表現」(“TIME”)・③「時間表現」(“DURATION”)・④「頻度集合表現」(“SET”)の 4 種類の表現を包含するものを指す.

² 日付や時刻を YYYY-MM-DDThh:mm:ss などといった数値と記号列で表記する標準. YYYY は年を表す四ケタの数字が, MM は月を表す二ケタの数字が, DD は日を表す二ケタの数字が, hh は 24 時間制で時刻を表す二ケタの数字が, mm は分を表す二ケタの数字が, ss は秒を表す二ケタの数字が入る. T は日付と時刻を分割する記号. 様々な略記方法が提案され, 例えば「2008 年 4 月」は “2008-04” と表記する. 詳細については ISO-8601 に対応する日本工業規格 JISX0301「情報交換のためのデータ要素及び交換形式一日付及び時刻の表記」参照のこと.

4 TimeML (TIMEX3) タグに基づいた日本語時間情報アノテーション基準

本節では日本語時間情報表現に対するアノテーション基準の概略を示す。アノテーション基準は、言語資源管理に関する国際標準 ISO/TC 37/SC 4³ において、2009 年に採用された ISO-24617-1 の元になっている TimeML (Pustejovsky et al. 2003b) (TIMEX3) タグの仕様に準拠⁴している。以下、(TIMEX3) のタグの日本語適応について説明する。細かな点で日本語に合うように変更しており、合わない部分については次節で説明する。

4.1 アノテーション対象

アノテーション対象は日付表現・時刻表現・時間表現・頻度集合表現の 4 種類である。図 1 にアノテーション事例を示す。

日付表現は「一九二九年二月」「前日」のような日暦に焦点をあてた表現である。時刻表現は「午前十時ごろ」「午後六時ごろ」「昼」「九日昼」のような一日のうちのある時点に焦点をあてた表現である。日付表現と時刻表現の区別は時間軸上の粒度の区別でしかない。便宜上不定の現在を表す「今」という表現を時刻表現に分類する。時間表現は「その間」のような時間軸上の両端に焦点をあてておらず、期間を表すことに焦点をあてている表現である。頻度集合表現は「毎日」のような複数の日付・時刻・時間に焦点をあてた表現である。この分類は、解析の方便のために導入したものである。時間軸上一つもしくは複数の時点・時区間を表現するもの

(出典) PB59_00001 (優先順位 00003)

```
<sentence type="quasi"><TIMEX3 @tid="t1" @type="DATE" value="2003-10-20"
@valueFromSurface="2003-10-20" @definite="true">二〇〇三年十月二十日 </TIMEX3>
<TIMEX3 @tid="t2" @type="DATE" @value="2003-10-W3-1" @valueFromSurface="XXXX-
WXX-1" @definite="true">月曜日 </TIMEX3></sentence> <br @type="automatic.original"
/> <sentence @type="quasi"><TIMEX3 @tid="t3" @type="TIME" @value="2003-10-
20T17:30:XX" @valueFromSurface="XXXX-XX-XXT17:30:XX" @definite="true">午後五時
三十分 </TIMEX3></sentence> <br @type="automatic.original" /> <blockEnd /> <paragraph>
<sentence> ステイシーはだらけた姿勢でモニターの前に陣取り、白黒の画像に見入っていた。
</sentence> <sentence> 彼女は伸びをし、腕時計に目をやった。 </sentence> <sentence><TIMEX3
tid="t4" type="DURATION" value="PT2H30M" valueFromSurface="PT2H30M">二時間半
</TIMEX3>で収穫ゼロ。 </sentence>
```

図 1 アノテーション例

³ <http://www.tc37sc4.org/>

⁴ 2003 年の TimeML と区別するために ISO-24617-1 の基準は ISO-TimeML と呼ばれている。国際標準として、英語だけでなく他言語の時間情報表現をアノテーションするために仕様が拡張されている。本研究の (TIMEX3) タグは ISO-TimeML にも準拠している。

をアノテーション対象である時間情報表現とする。

現在のアノテーション基準では `<TIMEX3>` タグの入れ子を許さない。日付・時刻表現の線形結合はこれを一つの日付・時刻表現として切り出す。例えば「九日昼」のように日付表現と時刻表現が接続する場合には一つの時刻表現として切り出す。時間を表す際に、開始時点と終了時点を示している場合には、開始時点と終了時点とを別々の日付・時刻表現として切り出す。例えば「午前十時ごろから午後六時ごろまで」は一つの時間表現として切り出さず、「午前十時ごろ」と「午後六時ごろ」の二つの時刻表現として切り出す。事象が起こる期間を表すために、今後、関連する事象表現に対し、この二つの時刻表現への参照関係を付与する予定である。頻度集合表現は、文字列上できるだけ短い単位を切り出す。例えば「毎日」を頻度集合表現として切り出すが、「毎日午前十時ごろから午後六時まで」は現在のところ頻度集合表現として切り出していない。

4.2 `<TIMEX3>` の属性

`<TIMEX3>` タグの属性のうち `@tid`, `@type`, `@value`, `@valueFromSurface`, `@freq`, `@quant`, `@mod` を概説する。また、作業・分析用に導入した `@definite` について説明する。

`@tid` 属性は一文書中の各時間情報表現に付与される識別子である。各時間情報表現を一意に同定するために用い、今後同一指示、参照、事象表現との時間的順序を表す際に用いる。

`@type` 属性は `DATE`, `TIME`, `DURATION`, `SET` の四つの値を持つ。それぞれ日付表現・時刻表現・時間表現・頻度集合表現を意味する。

`@value` 及び `@valueFromSurface` 属性は時間情報表現が含意する日付・時刻・時間の値を表す。値として ISO-8601 形式を自然言語表現向けに拡張したものをを用いる。このうち `@value` は文脈情報を用いて正規化を行った値を付与し、`@valueFromSurface` 属性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。

ここで `@value` と `@valueFromSurface` 属性の違いについて例を用いて説明する。「2013年4月」という表現は、文脈を用いなくても表層の文字列から時間軸上に一意に曖昧性解消ができる。「4月」という表現に対して、文脈情報からそれが2013年の「4月」であるとわかる場合には以下のように記述する（ここで属性にわりあてる値の詳細については4.3節に示す）。定時間情報表現は `@value` と `@valueFromSurface` の値は同じになるが、不定時間情報表現は同じになるとは限らない。

@value と @valueFromSurface

```

<TIMEX3 @type="DATE" @value="2013-04" @valueFromSurface="2013-04"> 2 0 1
3 年 4 月 </TIMEX3>

<TIMEX3 @type="DATE" @value="2013" @valueFromSurface="2013">          2  0
1 3 年 </TIMEX3> の 予 定 で す が <TIMEX3 @type="DATE" @value="2013-04"
@valueFromSurface="XXXX-04"> 4 月 </TIMEX3> は…

```

@freq, @quant 属性は頻度集合表現に付与される頻度情報及び量子化子情報である。属性にわりあてる値の詳細を 4.4 節に示す。

@mod 属性は時間情報表現のモダリティを表す。例えば「2000 年以前」をアノテーションするために @mod 属性に ON_OR_BEFORE という値をわりあてることにより「以前」というモダリティを表現する。属性にわりあてる値の詳細を 4.5 節に示す。

@definite 属性は“true”, “false”のいずれかの値を持ち、@value 属性が、文脈情報により定時間情報が得られる時間情報表現は“true”の値を持ち、その他の時間情報表現は“false”の値を持つ。言い換えると、日付・時刻表現が時間軸上の特定時区間に写像できる場合と時間・頻度集合表現の時間幅が特定できる場合に“true”の値を持ち、そうでない場合に“false”の値を持つ。例えば、以下の例はともに @definite が“false”の例である。「10 日」という表現は、文脈から「4 月」ということがわかるが、何年かまではわからないために定時間情報が得られない。なお、@definite 属性は作業・分析の便宜上導入したもので、元の ISO-TimeML の <TIMEX3> には規定されていない。

@definite

```

<TIMEX3 @type="DATE" @value="XXXX-04" @valueFromSurface="XXXX-04"
@definite="false"> 4 月 </TIMEX3> の 予 定 で す が <TIMEX3 @type="DATE"
@value="XXXX-04-10" @valueFromSurface="XXXX-XX-10" @definite="false">
1 0 日 </TIMEX3> は…

```

作成したコーパスに対し上記属性を付与した。他の属性として、記事配信日時など特別な意味を表す時間情報表現に付与する @functionInDocument, 同一指示を表す @anchorTimeID, 時間表現の開始位置と終了位置を表す @beginPoint, @endPoint, アノテーション時の問題点を自由記述する @comment がある。これらの情報は作業者が気づいた範囲で付与を行ったが完全ではない。

4.3 @value 及び @valueFromSurface

各表現に付与する @value 及び @valueFromSurface は ISO-8601 形式を基として、自然言語が表す時間情報向けに拡張したものである。@value は文脈情報を用いて正規化を行った値を付

与し、@valueFromSurface 属性は文脈情報を用いずに文字列の表層表現のみから判定できる値を付与する。ISO-8601 の標準表記では、日付・時刻表現を XXXX-XX-XXTXX:XX:XX の形で表す。

日付表現に対する値の事例を表 2 に示す。自然言語向けの拡張により、ISO-8601 では表現できない季節・四半期・年度などが表現できるようになっている。曜日表現に対する値の事例を表 3 に示す。曜日表現が表す WXX の数値部分は年内の暦週の番号を表す。日本語でよく用いられる「第 3 水曜」のような月内の暦週の番号を表す方がとられていない。これについては 5.2 節で詳説する。

時刻表現に対する値の事例を表 4 に示す。自然言語向けの拡張により、「朝」「昼」「夜」などが表現できるようになっている。* が付与されている「未明」と「深夜」は日本語新聞記事に頻出したために @valueFromSurface の値を独自に導入した。@value にはどちらも「夜」と同じく TNI をわりあてる。詳しくは 5.3 節で説明する。

表 2 日付表現に対する @value

単位	記号	日付表現例	@value
年月日	XXXX-XX-XX	1980年7月7日	1980-07-07
曜日	XXXX-WXX-X	水曜日	XXXX-WXX-3
季節	XXXX-{SP,SU,FA,WI}	冬	XXXX-WI
四半期	XXXX-QX	第一四半期	XXXX-Q1
年度	FYXXXX	1998年度	FY1998
世紀	XXXX	11世紀	10XX
紀元前	BCXXXX	紀元前202年	BC0202
		4000年前	KA4
		2億年前	MA200

表 3 曜日表現に対する @value

曜日表現例	@value
月曜日	XXXX-WXX-1
火曜日	XXXX-WXX-2
水曜日	XXXX-WXX-3
木曜日	XXXX-WXX-4
金曜日	XXXX-WXX-5
土曜日	XXXX-WXX-6
日曜日	XXXX-WXX-7
週末	XXXX-WXX-WE

表 4 時刻表現に対する @value

単位	記号	時刻表現例	@value
時刻	XXXX-XX-XXTXX:XX:XX	2006年8月8日午前8時45分30秒	2006-08-08T08:45:30
時刻 (略記)	TXX:XX:XX	午前8時45分30秒	T08:45:30
その他	XXXX-XX-XXTXX	未明*	XXXX-XX-XXTNI (DN)
		朝	XXXX-XX-XXTMO
		昼	XXXX-XX-XXTMI
		日中	XXXX-XX-XXTDT
		午後	XXXX-XX-XXTAF
		夕方	XXXX-XX-XXTEV
		夜	XXXX-XX-XXTNI
		深夜*	XXXX-XX-XXTNI (MN)

時間表現に対する値の事例を表 5 に示す. 基本的に ISO-8601 の時間表現⁵と同じであり, 接頭辞として P を付与し, その後に数値とともにそれぞれ年, 月, 日, 時間, 分, 秒, 週を表す Y, M, D, H, M, S, W を接尾辞として付与する. 月 (M) と分 (M) を区別するために日と時間の境界に T を付与する.

「今」「近年」「今後」など不定な表現に対する値の事例を表 6 に示す. これらは全て自然言語向けに導入した値である.

頻度集合表現は上記 @value 属性を流用しながら次節に示す @freq, @quant 属性を組み合わせることによって表現される.

4.4 頻度集合表現に対する @freq 及び @quant 属性

頻度集合表現は @value, @freq, @quant 属性を組み合わせることにより複雑な時間情報を表現する.

頻度情報を表すためには, 期間を表す @value 属性とともに, @freq 属性に nX をわりあてることにより, 焦点をあてている期間中に事象が n 回起こることを示す. 例えば「週に 2 回」を表現する際には以下のようにアノテーションする⁶.

「週に 2 回」

`<TIMEX3 type="SET" value="P1W" freq="2X"> 週に 2 回 </TIMEX3>`

@quant 属性には「毎日」「毎週」「毎 10 月」といった表現に EACH をわりあて, 「10 日おき」「3 日毎」といった表現に EVERY をわりあてる. この際 @value 属性には期間を表す値だけでなく, 日付・時刻を表す値が入ることがある. 以下に例を示す.

表 5 時間表現に対する @value

単位	記号	時間表現例	@value
年	PnY	3 年間	P3Y
月	PnM	2 ヶ月	P2M
日	PnD	5 日	P5D
時間	PTnH	3 時間	PT3H
分	PTnM	30 分	PT30M
秒	PTnS	9 秒 80	PT9.80S
週	PnW	1 週間	P1W

表 6 不定な表現に対する @value

時間表現例	@value
「今」「現在」	PRESENT_REF
「近年」「以前」	PAST_REF
「今後」「将来」	FUTURE_REF

⁵ ISO-8601 では時間を表現するために Time interval 形式 と Duration 形式の二つがあるが, ここでは Duration 形式を用いる.

⁶ 説明に不要な属性は省略して表示. 以下同様.

「毎日」「毎10月」「10日おき」

```
<TIMEX3 type="SET" value="P1D" quant="EACH"> 毎日 </TIMEX3>
<TIMEX3 type="SET" value="XXX-10" quant="EACH"> 毎10月 </TIMEX3>
<TIMEX3 type="SET" value="P10D" quant="EVERY"> 10日おき </TIMEX3>
```

頻度集合表現は、できるだけ文字列上小さな単位で切り出しているため、現在のところ上記定義で意味論的表示に曖昧性が生じていない。例えば「毎日午前十時ごろから午後六時まで」のような表現の場合、表現全体の単位で切り出すとすると、@value, @freq, @quant 属性のみで曖昧性なく意味論的表示に落とすことは困難である。これは、時間情報表現間の時間的順序関係のアノテーションにおいて今後対処していきたい。

4.5 モダリティ修飾 @mod 属性

時間情報表現は接尾表現をともない、様々なモダリティを表現する。@mod 属性は時刻・時間表現に対するモダリティ修飾情報である。表7に取りうる値の一覧と例を示す。

日付・時刻・時間表現に共通して用いられる @mod 属性として START, MID, END, APPROX がある。例えば、「60年代初頭」「10月半ば」「約40年」は以下のようにアノテーションする。

「60年代初頭」「10月半ば」「約40年」

```
<TIMEX3 type="DATE" value="196X" mod="START"> 60年代初頭 </TIMEX3>
<TIMEX3 type="DATE" value="XXX-10-XX" mod="MID"> 10月半ば </TIMEX3>
<TIMEX3 type="DURATION" value="P40Y" mod="APPROX"> 約40年 </TIMEX3>
```

表7 @mod 属性に対する値

値	定義	例
@mod=START	日付時刻表現の初期	「初め」「初頭」
@mod=MID	日付時刻表現の中期	「半ば」「中ごろ」
@mod=END	日付時刻表現の後期	「末」「暮れ」
@mod=APPROX	近似表現	「ごろ」
@mod=BEFORE	日付時刻表現より前	「前」
@mod=AFTER	日付時刻表現より後	「過ぎ」
@mod=ON_OR_BEFORE	日付時刻表現以前	「以前」
@mod=ON_OR_AFTER	日付時刻表現以後	「以降」「以来」
@mod=EQUAL_OR_LESS	時間表現の範囲以下	「以内」
@mod=EQUAL_OR_MORE	時間表現の範囲以上	「以上」
@mod=LESS_THAN	時間表現の範囲未満	「未満」「近く」
@mod=MORE_THAN	時間表現の範囲超過	「余り」「過ぎ」

日付・時刻表現に対する @mod 属性として, BEFORE, AFTER, ON_OR_BEFORE, ON_OR_AFTER がある. 例えば「1998年以前」は以下のようにアノテーションする.

「1998年以前」
`<TIMEX3 type="DATE" value="1998" mod="ON_OR_BEFORE"> 1998年以前`
`</TIMEX3>`

時間表現に対する @mod 属性として, EQUAL_OR_LESS, EQUAL_OR_MORE, LESS_THAN, MORE_THAN がある. 例えば「10分以内」は以下のようにアノテーションする.

「10分以内」
`<TIMEX3 type="DURATION" value="PT10M" mod="EQUAL_OR_LESS"> 10分以内`
`</TIMEX3>`

5 <TIMEX3> 日本語適応上の問題点

本節では <TIMEX3> の日本語適応について問題となった事例について個別に紹介する. 基本的には @valueFromSurface に日本語特有の表示を用い, @value に TimeML に適合した表示を用いる. 必要に応じて韓国語のデータ K-TimeML (Im, You, Jang, Nam, and Shin 2009) に準拠した Korean Timebank 1.0/2.0 でどのような扱いを行っているのかを付記する.

5.1 暦

日本語に <TIMEX3> を適応するうえで一番大きな問題として, 和暦の問題がある. 日本語では元号法に基づいた「昭和」「平成」などの年表記がテキスト中出现する. これについては @valueFromSurface に「明治」「大正」「昭和」「平成」を表現する “M”, “T”, “S”, “H” の接頭子を二ケタの数字表現に付与することで記述し, @value に西暦に換算したものを記述することとする. 上記四つ以外の元号については @value 相当に西暦を記述するのみで対処する. 元号以外の時代名 (例:「江戸時代」) については, 時間情報表現として切り出しを行うのみで, @value 相当は空白とした.

和暦の扱い
`<TIMEX3 type="DATE" @value="1985" @valueFromSurface="S60"> 昭和 60 年`
`</TIMEX3>`

旧暦に対しては, @valueFromSurface において日付表現の表示の末尾に “Q” を付与し旧暦を表し, 明示的に新暦の暦日が記述されている場合にのみ @value 相当に記述する.

皇紀に対しては, @valueFromSurface において表示の最初に "JY" を付与し皇紀を表し, @value 相当に西暦を記述する.

韓国語 (韓国・北朝鮮), 中国語 (中国・台湾) など同様の問題が起きうるが, 公開されている文書を見る限り @value 相当に西暦を記述することで対処している.

5.2 月次の週番号の扱い

欧米地域では年単位の週番号を利用する傾向がある一方, 東アジア地域では「第三木曜」など月単位の週番号を利用する傾向がある. このため, 韓国より <TIMEX3> に対して月単位の週番号を記述可能にする拡張が提案されている. 具体的には以下のように @valueFromSurface の日相当部分に週番号を記述する. @value にはカレンダーを参照することにより ISO-8601 の標準表記 XXXX-XX-XX 形式の値をわりあてた.

月次の週番号の扱い

```
<TIMEX3 type="DATE" @value="2013-01-17"
@valueFromSurface="XXXX-XX-W3-4"> 第三木曜 </TIMEX3>
```

5.3 その他日本語特有の表現

以下では日本語適応において問題となった雑多な事例について紹介する. 基本的に @valueFromSurface において日本語に限定した形式で正規化を行い, @value 相当は正規表現などを用いて正規化を行う.

- 旬 (月の細分類) 上旬, 中旬, 下旬

@valueFromSurface においては "1J" (上旬), "2J" (中旬), "3J" (下旬) など日本語に限定した形式を導入し, @value には "XXXX-XX-(0[123456789]|10)" (上旬) のように正規表現を用いて正規化を行った.

- 未明・午前・真夜中

@valueFromSurface においては新聞記事などで頻出する表現に対応するため "DW" (未明; @value は "NI" (夜) 相当)・"FN" (午前; @value は "MO" (朝) 相当)・"MN" (真夜中; @value は "NI" (夜) 相当) を導入した.

- 学期

@valueFromSurface においては学期を表現する記号として "G" を導入した. @value は何学期制かが不明なので空白である事例が多い.

- 国民の祝日

国民の祝日は, 国立天文台が官報で公表する「春分日」「秋分日」やハッピーマンデー制度が適用される四つの祝日も含めて @value 相当にその年の正しい日付を記述する.

- 休日
時間情報表現として切り出すが, 具体的にいつなのかわからない場合が多く, @value は空白とした.
- お盆
@value 相当に “SU” (夏) を記述する.
- 十干十二支
時間情報表現として切り出すが, 具体的にいつなのかわからない場合が多く, @value は空白とした.

6 作業環境と作業対象

本節では作業環境と作業対象である BCCWJ について詳しく説明する.

6.1 作業環境

アノテーション作業には XML Editor oXygen⁷ を利用した. DTD や XML Schema を記述することにより時間表現の切り出し部分や属性にわりあてる値などを統制することができる. 時

図 2 XML Editor oXygen によるアノテーション

⁷ <http://www.oxygenxml.com/>

図 3 ペアプログラミングによるアノテーション作業

間表現の切り出しはマウスもしくはキーボードで対象となる文字列を選択したうえで Ctrl-e とタイプし、タグを選択することで、XML タグ（閉じタグも含む）が挿入される。この状態で画面右の属性項目を記述することにより、XML ファイルを編集することができる。

言語資源に対するアノテーションにおいて、ある一定の基準を守ったうえで複数の作業者の主観を尊重してそれぞれの作業者間の判断の揺れを許す場合⁸、基準を厳格化し作業者間の判断の揺れを許さない場合の2通りの統制手法がある。本論文の時間情報表現の特定の時間軸上への写像作業は後者の統制手法を取るために図3のようにペアプログラミングのような手法を取った。1台のPCに、キーボード二つ・マウス二つ・ディスプレイ二つを接続し、ディスプレイはミラーリングを行う。一つのPCを共有したうえで、作業者がアノテーション作業を行い、作業監督者がアノテーション仕様の改訂を行う。

アノテーション作業は1人の作業者と1人の作業監督者により行った。1回目のアノテーション時の作業時間は7時間15分週2日勤務で約2か月だった。1回目のアノテーション直後（1回目見直し）と6か月後（2回目見直し）に見直しを行った。1回目の見直し作業はアノテーション結果を帳票形式で出力したうえで、属性に関する見直しを行った。2回目の見直し作業は再度XML Editor上で行った。アノテーションデータは1回目の見直し作業が終わった時点で公開しており、利用者から誤りの指摘があった場合に修正を行っている。

このアノテーション作業の準備段階の〈TIMEX3〉の切り出しの一致率は95%以上であり、切り出しの不一致の多くが単純な見落としてであった。属性についての一致率は、一つの時刻・時間に複数の意味論的表示を許す @value, @valueFromSurface⁹を除いて90%以上である。

6.2 作業対象

作業対象である BCCWJ (国立国語研究所コーパス開発センター 2011) について概説する。

⁸ 例えば例文「柔道家が先輩に勝った仲間を紹介した」に対して一般的な統語的な制約基準を教示したうえで、本質的に2通りの解釈が残る場合に片方の解釈に矯正せずに作業者間の選択選考性の差異の判断を委ねるような場合。

⁹ 例えば、「30分」を表現する @value として “PT30M” と “PT0.5H” の両方が許されている。

表 8 作業対象データ

レジスタ	ファイル数	うち時間表現あり	文数	うち時間表現あり	長単位形態素数	短単位形態素数
白書 OW (A)	17	16 (94%)	1439	405 (28%)	40690	58336
書籍 PB (A)	25	25 (100%)	2568	289 (11%)	50257	57929
新聞 PN (A,B)	110	110 (100%)	5582	1562 (28%)	88733	116834
知恵袋 OC (A)	518	250 (48%)	3479	488 (14%)	51240	60086
雑誌 PM (A)	23	23 (100%)	3066	413 (13%)	49715	59372
ブログ OY (A)	257	198 (77%)	3986	765 (19%)	53333	63459

BCCWJ のコアデータは、OW: 白書, PB: 書籍, PN: 新聞, OC: Yahoo! 知恵袋, PM: 雑誌, OY: Yahoo! ブログの六つのレジスタからなり、それぞれ約 5 万語単位で、アノテーションすべき優先順位に基づいた部分集合が規定されている¹⁰。上記全レジスタの部分集合“A”と比較的時間表現が多いレジスタである PN(新聞)の部分集合“B”をアノテーション対象とした。

表 8 にデータの概要を示す。表中「ファイル数」はアノテーションしたファイルの数、ファイル数の「うち時間表現あり」は時間表現を一つ以上含むファイルの数を表す。

OC, OY などのユーザー生成コンテンツはサンプリングの長さにもよるが 1 ファイルに時間情報表現が一つも含まれないものがある一方、OW, PB, PN, PM などのユーザー生成コンテンツ外のほとんどは時間表現が必ず一つ以上含まれている。OW の中で時間表現を一つも含まない 1 ファイルは「平成 16 年度森林・林業白書」であった。文単位では、OW, PN が時間情報表現を含む割合が最も多く、PB が時間情報表現を含む割合が最も少なかった。

7 タグの分析

本節ではアノテーションした情報について、時間情報表現の正規化の観点から分析を行う。表 9 に文書作成日時を示すタグを除いた@type ごとのタグの出現数を曖昧性解消の観点から二つの視点で四つに分割して示す。一つ目の視点は @definite が “true” か “false” かである。“true” の場合、時間軸上に時区間が特定可 (“DURATION” と “SET” は時間幅が特定可)であることを意味し、“false” の場合、時間軸上に時区間が特定不可であることを意味する。二つ目の視点は @value と @valueFromSurface の値が一致する (“=” で表記) か、一致しない (“≠” で表記) かである。一致する場合人手による文脈を用いた正規化作業が行われていないことを意味し、一致しない場合人手による文脈を用いた正規化作業が行われたことを意味する。

まず 日付表現 (“DATE”) について、時区間特定可 (@definite が “true”) であるものの多くが、人手による曖昧性解消が行われている (@value ≠ @valueFromSurface) ことがわかる。こ

¹⁰ BCCWJ においては、均衡性を保つためにコアデータにアノテーションする際の優先順位 ($A > B > C > D > E$) が定義されており、その優先順位に基づいて部分集合が定義されている。

表 9 @type 属性ごとの出現数と文脈による曖昧性解消可能性

@definite		true (特定可)			false (特定不可)		
@value と @valueFromSurface		all	=	≠	all	=	≠
OW	DATE	533 (86%)	108 (17%)	425 (69%)	87 (14%)	85 (14%)	2 (0%)
OW	TIME	0	0	0	0	0	0
OW	DURATION	100 (96%)	100 (96%)	0	4 (4%)	4 (4%)	0
OW	SET	5 (83%)	5 (83%)	0	1 (17%)	1 (17%)	0
PB	DATE	85 (38%)	36 (16%)	49 (22%)	139 (62%)	128 (57%)	11 (5%)
PB	TIME	6 (18%)	1 (3%)	5 (15%)	28 (82%)	20 (59%)	8 (24%)
PB	DURATION	96 (89%)	96 (89%)	0	12 (11%)	12 (11%)	0
PB	SET	13 (93%)	13 (93%)	0	1 (7%)	1 (7%)	0
PN	DATE	1145 (80%)	136 (10%)	1009 (71%)	279 (20%)	247 (17%)	32 (2%)
PN	TIME	161 (81%)	0	161 (81%)	37 (19%)	34 (17%)	3 (2%)
PN	DURATION	537 (93%)	537 (93%)	0	39 (7%)	39 (7%)	0
PN	SET	39 (89%)	37 (84%)	2 (5%)	5 (11%)	5 (11%)	0
OC	DATE	82 (23%)	16 (5%)	66 (19%)	271 (77%)	243 (69%)	28 (8%)
OC	TIME	0	0	0	70 (100%)	49 (70%)	21 (30%)
OC	DURATION	169 (91%)	169 (91%)	0	17 (9%)	17 (9%)	0
OC	SET	33 (89%)	33 (89%)	0	4 (11%)	4 (11%)	0
PM	DATE	146 (41%)	35 (10%)	111 (31%)	211 (59%)	174 (49%)	37 (10%)
PM	TIME	0	0	0	39 (100%)	35 (90%)	4 (10%)
PM	DURATION	116 (88%)	116 (88%)	0	16 (12%)	16 (12%)	0
PM	SET	21 (72%)	21 (72%)	0	8 (28%)	7 (24%)	1 (3%)
OY	DATE	223 (33%)	50 (7%)	173 (26%)	451 (67%)	398 (59%)	53 (8%)
OY	TIME	21 (13%)	0	21 (13%)	141 (87%)	101 (62%)	40 (25%)
OY	DURATION	111 (91%)	110 (90%)	1 (1%)	11 (9%)	11 (9%)	0
OY	SET	20 (83%)	20 (83%)	0	4 (17%)	4 (17%)	0

(括弧内はレジスタ・@type ごとに集計した割合)

のことから本アノテーションの目的とする時間表現の正規化作業の重要性がうかがえる。レジスタにもよるが、OW 白書や PN 新聞など、出版年・発行年月日が明らかであるものほど曖昧性解消がよく行われる傾向がある。日付表現の曖昧性解消は、和暦から西暦への換算や、西暦二ケタ表記から西暦四ケタ表記への換算、さらに年が省略されている表現の文脈や文書作成日時に基づく年の補完によるものがあり、白書の多くの事例がこの暦の換算作業であった。他のレジスタは曖昧性がない表現が多いわけではなく、文書作成日時など曖昧性を解消するに足る情報がデータ中に含まれていないことにより、具体的な時間軸上の区間を指し示すことができない事例が多かった。部分的に情報を補完すること（例えば「3日」という表現に対し9月であることまでがわかるが、何年であるかまではわからないので、@value="XXXX-09-03"をわりあてること）も含まれており、時区間特定不可（@definite が “false”）であるが人手による曖昧性解消が行われている（@value≠ @valueFromSurface）ものがこれにあてはまる。

次に時刻表現（“TIME”）については、PB の 1 件を除いて時区間特定可であるものの殆どが

人手による曖昧性解消が行われている。時刻表現の曖昧性解消は、日付が省略されている場合の日付の補完のほか、午前と午後の曖昧性解消が含まれる。日付表現と同様に出版年月日がわかる新聞記事の曖昧性解消がよく行われている一方、OC Yahoo! 知恵袋や PM 雑誌は、店の営業時間など時間軸上の特定の時刻を表現しないものが多かった。

時間表現 (“DURATION”) と頻度集合表現については、時間軸上の時区間を特定することを目的とせず、時間幅が特定できれば `@definite` が “true” になると定義している。実際に時間軸上の時区間に写像する際には、日付・時刻表現や事象表現との時間的順序関係 (TimeML の `TLINK`) を定義することが必要になる。

8 おわりに

本稿では BCCWJ に対する日本語時間情報アノテーションについて説明した。アノテーションは各国で進められている国際標準 ISO-TimeML に定義された `TIMEX3` タグに準拠している。他言語においては対象を新聞記事に限定しているのに対し、本研究は 6 種類のレジスタを対象にアノテーションを行い、レジスタ横断的な分析を行うとともに、日本語適応における問題点を調査した。作成したアノテーションデータはスタンドオフ形式で公開する。BCCWJ の DVD を入手することでタグつきテキストコーパスが復元できる。

以下、今後の展望を示す。

今回作成したテキストコーパスをベンチマークとして正規化を行う日本語時間表現解析器の開発を現在進めている。作成中の解析器では、まず、表層文字列からわかる値をラティス上に展開し、セミマルコフモデルを用いて曖昧性解消を行う。解析対象表現一文書作成日時および解析対象表現一隣接時間情報表現の時間的順序関係を今回作成したタグつきコーパスを用いて機械学習器を用いて推定することにより、不定時間情報表現に対する情報補完を行う。

今後、TimeML で行われている事象表現と時間表現間の時間的順序関係 (TimeML 中の `TLINK`) 付与を進めていきたい。そのためには、対象となる事象表現の策定、事象表現に対する分類 (TimeML 中の `EVENT@type`)、テンス・アスペクト体系の整備 (同 `MAKEINSTANCE@tense`, `@aspect`)、節間の関係定義 (同 `SLINK`) など解決すべき問題は山積している。事象表現に対する分類として工藤らの動詞分類 (工藤 1995, 2004) を基にしたうえで、`EVENT@type` を付与している。テンス・アスペクト体系については中村らのテンス・アスペクトの解釈 (中村 2001) を参考にしてラベルを設計する予定である。日本語のアスペクト表現は多様であるために `MAKEINSTANCE@aspect` については、元の TimeML から拡張する予定である。今後 TimeML に準じた事象表現に対するアノテーションを行い、最終目標である事象表現に対する時間情報付与の研究へと進んでいきたい。

謝 辞

本研究は、国立国語研究所基幹型共同研究「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「超大規模コーパス構築プロジェクト」による補助を得ています。

参考文献

- Boguraev, B. and Kubota Ando, R. (2005). “TimeML-Compliant Text Analysis for Temporal Reasoning.” In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05)*, pp. 997–1003.
- Boguraev, B. and Kubota Ando, R. (2006). “Analysis of TimeBank as a Resource for TimeML parsing.” In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-06)*, pp. 71–76.
- DARPA TIDES (2004). *The TERN evaluation plan; time expression recognition and normalization*. Working papers, TERN Evaluation Workshop.
- Grishman, R. and Sundheim, B. (1996). “Message Understanding Conference-6: a brief history.” In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pp. 466–471.
- 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—白書, 書籍, Yahoo! 知恵袋コアデータ. 言語処理学会第 16 回年次大会発表論文集, pp. 916–919.
- Im, S., You, H., Jang, H., Nam, S., and Shin, H. (2009). “KTimeML: specification of temporal and event expressions in Korean text.” In *Proceeding of ALR7 Proceedings of the 7th Workshop on Asian Language Resources*, pp. 115–122.
- IREX 実行委員会 (1999). IREX ワークショップ予稿集.
- 国立国語研究所コーパス開発センター (2011). 『現代日本語書き言葉均衡コーパス』利用の手引 (第 1.0 版).
- 工藤真由美 (1995). アスペクト・テンス体系とテキスト—現代日本語の時間の表現—. ひつじ書房.
- 工藤真由美 (2004). 日本語のアスペクト・テンス・ムード体系—標準語研究を超えて—. ひつじ書房.
- Mani, I. (2006). “Machine Learning of Temporal Relations.” In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006)*, pp. 753–760.
- 中村ちどり (2001). 日本語の時間表現. くろしお出版.
- Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., and Katz,

- G. (2003a). “The TIMEBANK Corpus.” In *Proceedings of Corpus Linguistics 2003*, pp. 647–656.
- Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., and Lazo, M. (2003b). “TimeML: Robust Specification of Event and Temporal Expressions in Text.” In *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, pp. 337–353.
- Sekine, S., Sudo, K., and Nobata, C. (2002). “Extended Named Entity Hierarchy.” In *The Third International Conference on Language Resources Evaluation (LREC-02)*, pp. 1818–1824.
- Setzer, A. (2001). *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield.
- Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., and Pustejovsky, J. (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 75–80.
- Verhagen, M., Saurí, R., Caselli, T., and Pustejovsky, J. (2010). “SemEval-2010 Task 13: TempEval-2.” In *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pp. 57–62.

略歴

小西 光：2005 年上智大学文学部卒業。2007 年上智大学文学研究科博士前期課程修了。2008 年より国立国語研究所コーパス開発センタープロジェクト奨励研究員。現在に至る。『日本語書き言葉均衡コーパス』『日本語話し言葉コーパス』『日本語超大規模コーパス』の整備に携わる。

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特任准教授。現在に至る。博士（工学）。自然言語処理・コーパス言語学の研究に従事。

前川喜久雄：1956 年生。1984 年上智大学大学院外国語学研究科博士後期課程（言語学）中途退学。国立国語研究所教授。言語資源系長。コーパス開発センター長。博士（学術）。専門は音声学ならびに言語資源学。

(2012 年 10 月 24 日 受付)

(2012 年 12 月 28 日 再受付)

(2013 年 2 月 10 日 採録)