『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[‡]
（国立国語研究所 コーパス開発センター）

Temporal Information Annotation on Balanced Corpus of Contemporary Written Japanese

Hikari Konishi, Masayuki Asahara and Kikuo Maekawa
(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ある時区間もしくは時点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が起きる日時が文書作成日時と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) (R. Grishman and B. Sundheim 1996) で、タグづけ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タグづけ基準 TimeML (J. Pustejovsky et al. 2003b) が検討され、TimeML に基づくタグつきコーパス TimeBank (J. Pustejovsky et al. 2003a) などが整備された。2007 年には、時間情報表現－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 におけるサブタスク TempEval (M. Verhagen et al. 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 におけるサブタスク TempEval-2 (M. 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、フランス語を含めた 6 言語が対象となった。

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。我々は TimeML に基づいた日本語に対する時間情報タグづけ基準を定義し、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。本稿ではタグづけ基準を示すとともに、タグづけしたコーパスの詳細について示す。

２．対象とする時間情報表現

まず以下の例文を見て欲しい。

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

[*] hkonishi at ninjal.ac.jp
[†] masayu-a at ninjal.ac.jp
[‡] kikuo at ninjal.ac.jp

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 3 つの分類に分けられる。①日付表現(“DATE”,相当)・②時刻表現(“TIME”相当)は「2008 年 4 月」「昨日」「今朝 9 時」といった、時点及び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(“DURATION”相当)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(“SET”相当)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。

曖昧性を解消しながら時間情報表現を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、方策が必要である。

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。

３．TimeML〈TIMEX3〉タグに基づいた時間情報タグづけ基準の概略

本節では日本語時間情報表現に対するタグづけ基準の概略を示す。タグづけ基準は、言語資源管理に関する国際標準 ISO/TC 37/SC 4[3]において 2009 年に採用された ISO 24617-1 (SemAF/Time)の基になっている TimeML (J. Pustejovsky et al. 2003b) ⟨TIMEX3⟩ タグの仕様に準拠している[4]。以下、日本語の例を用いながら⟨TIMEX3⟩のタグの仕様を説明する。細かな点で日本語に合うように変更しているがタグ名は⟨TIMEX3⟩をそのまま利用している。

３.１　タグづけ対象

タグづけ対象は日付表現(“DATE”)・時刻表現(“TIME”)・時間表現(“DURATION”)・頻度集合表現(“SET”)の 4 種類である。図 1 にタグづけ事例を示す。

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 1 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

---

[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 4 ケタの数字が、MM は月を表す 2 ケタの数字が、DD は日を表す 2 ケタの数字が、hh は 24 時間制で時刻を表す 2 ケタの数字が、mm は分を表す 2 ケタの数字が、ss は秒を表す 2 ケタの数字が入る。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 ― 日付及び時刻の表記」参照のこと。

[3] http://www.tc37sc4.org/

[4] 2003 年の TimeML と区別するために ISO 24617-1 の基準を ISO-TimeML と呼ぶ。

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をタグづけ対象である時間情報表現とする。

現在のタグづけ基準では⟨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連接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明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前六時ごろまで」は一つの時間表現として切り出さず、「午前十時ごろ」と「午前六時ごろ」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前六時まで」は現在のところ頻度集合表現として切り出していない。

⟨sentence type="quasi"⟩⟨TIMEX3 tid="t1" type="DATE" value="2003-10-20" valueFromSurface="2003-10-20"⟩二〇〇三年十月二十日⟨/TIMEX3⟩　⟨TIMEX3 tid="t2" type="DATE" value="2003-W43-1" valueFromSurface="XXXX-WXX-1"⟩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"⟩ 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 タグづけ例　（PB59_00001）

３.２　⟨TIMEX3⟩ の属性

⟨TIMEX3⟩ タグの属性のうち　@tid, @type, @value, @valueFromSurface, @temporalFunction , @freq, @quant, @mod を概説する。

@tid 属性は 1 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 4 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。属性にわりあてる値の詳細を３．３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。

@temporalFunction 属性は true, false のいずれかの値を持ち、@valueFromSurface が文脈情報により曖昧性解消可能か否かを表す。定時間情報が得られる不定時間情報表現は true の値を持ち、その他の時間情報表現は false の値を持つ。

@freq, @quant 属性は頻度集合表現に付与される頻度情報及び量化子情報である。属性にわりあてる値の詳細を３．４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」をタグづけするために@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３．５節に示す。

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、タグづけ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。

３.３　@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。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。

表 1 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 2 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

日付表現に対する値の事例を表 1 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 2 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策がとられていない。このため独自拡張として XXXX-XX-W3-3 のように YYYY-MM のあと月内の暦週の番号を WX で表記することを許す。実際のタグづけでは @valueFromSurface には月内の暦週の番号に基づく値を、@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。

表 3 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-XXTDN
		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-XXTMN

時刻表現に対する値の事例を表 3 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独自に導入した値である。厳密に TimeML の⟨TIMEX3⟩ 互換にする際にはどちらも「夜」と同じく TNI をわりあてる。

表 4 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	P1D
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 5 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

時間表現に対する値の事例を表 4 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 5 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。

３.４ 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

⟨TIMEX3 type="SET" value="P1W" freq="2X"⟩週に２回⟨/TIMEX3⟩

のようにタグづけする[6]。

@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。

⟨TIMEX3 type="SET" value="P1D" quant="EACH"⟩ 毎日⟨/TIMEX3⟩

⟨TIMEX3 type="SET" value="XXXX-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前六時まで」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タグづけにおいて今後対処していきたい。

---

[5] ISO-8601 では時間を表現するために Time interval 形式 と Duration 形式の２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。

３.５　モダリティ修飾子 @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子である。表 6 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

⟨TIMEX3 type="DATE" value="196X-XX-XX" mod="START"⟩ ６０年代初頭⟨/TIMEX3⟩

⟨TIMEX3 type="DATE" value="XXXX-10-XX" mod="MID"⟩ １０月半ば ⟨/TIMEX3⟩

⟨TIMEX3 type="DURATION" value="P40Y" mod="APPROX"⟩ 約４０年⟨/TIMEX3⟩

のようにタグづけする。

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

⟨TIMEX3 type="DATE" value="1998" mod="BEFORE"⟩１９９８年以前⟨/TIMEX3⟩

のようにタグづけする。

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

⟨TIMEX3 type="DURATION" value="PT10M" mod="EQUAL_OR_LESS"⟩１０分以内⟨/TIMEX3⟩

のようにタグづけする。

表 6 @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」

４．タグの分析

BCCWJ のコアデータ[7]の一部である、全ジャンル(OW(白書), PB(書籍), PN(新聞), OC(Yahoo! 知恵袋), PM(雑誌), OY(Yahoo! ブログ)) の部分集合“A”と比較的時間表現が多いジャンルである PN(新聞) の部分集合“B”について人手によりタグづけした。

表 7 にデータの概要を示す。表中「ファイル数」はタグづけしたファイルの数、ファイル数の下のカッコ内の数字「時間表現あり」は時間表現一つ以上含むファイルの数を表す。

まず、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時間表現

[7]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! ブログからなり、それぞれ約 5 万語単位で、タグづけすべき優先順位をもつ部分集合（A > B > C > D > E）が設定されている。

表 7 @type 属性毎の出現数と文脈による曖昧性解消可能性

ジャンル	ファイル数 （うち時間表現あり）	DATE	TIME	DURATION	SET	合計
		（文脈に曖昧性解消可能なものの数）				
OW	17 16	596 (414)	0 (0)	191 (0)	6 (0)	703
PB	25 25	209 (51)	28 (12)	105 (0)	14 (0)	356
PN	110 110	1323 (999)	193 (162)	553 (0)	41 (0)	2110
OC	518 250	341 (95)	70 (19)	184 (0)	37 (0)	632
PM	23 23	333 (108)	37 (2)	131 (0)	28 (1)	529
OY	257 198	632 (215)	161 (58)	117 (1)	22 (0)	932

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 の中で唯一、時間表現が一つも含まないサンプルは平成 16 年度の森林・林業白書であった。

ジャンル毎の文書作成日時を示すタグを除いた@type 毎のタグの出現数を“DATE”，“TIME”，“DURATION”，“SET” 表 7 右に示す。合計はジャンル毎の時間表現の合計を表す。カッコ内は、文脈より曖昧性解消が行うことができたものの数を表す。日付表現の曖昧性解消は、和暦から西暦への換算や、西暦 2 ケタ表記から西暦 4 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る。ジャンル間の差異において OW(白書) が時刻表現を一つも含まないという特色がわかった。

OW は和暦西暦換算の事例が多い一方、PN は文書作成日時がメタデータ中に明示的に含まれているため、曖昧性解消が文脈によって行える事例が多かった。他のジャンルは曖昧性がない表現が多いわけではなく、曖昧性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。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。時間表現で文脈により曖昧性解消される事例は、「３年くらい前」という相対的に何年に起きたかを日付表現に置き換えられるもの一例であった。集合表現で文脈により曖昧性解消される事例は、「１時間おき」という具体的に「１１月１４日の１時から６時の１時間おき」といったように時刻が特定できるもの一例であった。

ここで日付表現の曖昧性解消とは不定表現を完全に定表現に変換することだけではなく、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) も含まれる。さらに時刻表現の曖昧性解消の際、日付の情報が含まれていない場合には時刻レベルの補完のみにとどめられている場合[8]がある。実際に補完すべき時刻情報がより多くあり、時間情報表現の正規化が重要であることがわかる。

表 8 頻度集合表現の統計

	@freq=nil	@freq= n X	otherwise
@quant=nil	3	43	2
@quant=EACH	75	2	5
@quant=EVERY	18	0	0

頻度集合表現はタグづけしたテキスト中に 148 件出現した。@quant 属性、@freq 属性別

8 時刻レベルの補完のみにとどめられた時刻表現の正規化は、今後、時間的順序関係を表す〈TLINK〉タグを、文書作成日時を含めた日付表現への参照表現としてタグづけすることにより解決する。

の統計を表 8 に示す。(@quant="EACH", @freq=nil) に分類される表現（例「毎日」）が最も多く、次に(@quant=nil, @freq="nX") に分類される表現（例「１日３回」「週２度」）が多かった。その他複雑な表現として「１カ月あたり１時間」(@value="PT1H", @freq="P1M") といった表現があった。

表 9 @mod 属性の統計

@type	DATE	TIME	DURATION	SET
@mod=START	27	11	1	0
@mod=MID	5	0	2	0
@mod=END	72	0	5	1
@mod=APPROX	19	35	95	2
@mod=BEFORE	0	5	−	0
@mod=AFTER	0	6	−	0
@mod=ON OR BEFORE	7	0	−	0
@mod=ON OR AFTER	36	21	−	0
@mod=EQUAL OR LESS	−	−	16	0
@mod=EQUAL OR MORE	−	−	29	0
@mod=LESS THAN	−	−	13	0
@mod=MORE THAN	−	−	5	0

各時間情報表現に付与された @mod の統計を表 9 に示す。日付表現で多かった @mod の値は END であり、「年末」「月末」といった表現が例示される。それ以外の表現では「約」「ごろ」が付与された APPROX が多かった。日付表現では「五日以前」「４月以降」といった ON_OR_BEFORE, ON_OR_AFTER がある一方、BEFORE, AFTER は存在しなかった。時刻表現においては逆に「八時前」「５時すぎ」といった BEFORE, AFTER が出現する一方、ON_OR_BEFORE, ON_OR_AFTER はほとんど出現しなかった[9]。

５．関連研究

表 10 に関連研究を示す。

表 10 関連研究

英語に関する関連研究		
MUC-6 (R. Grishman and B. Sundheim 1996)	評価型会議	時間情報表現の切り出しのみ
(A. Setzer 2001)	タグづけ基準	時間情報表現の切り出しと正規化
TERN (DARPA TIDES 2004)	評価型会議	時間情報表現の切り出しと正規化
TimeML (J. Pustejovsky et al. 2003b)	タグづけ基準	事象間の時間的順序関係
TimeBank (J. 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	コーパス	TimeML 基準によるタグつきコーパス
(B. Boguraev and R. Kubota Ando 2005)	解析手法	時間情報表現- 事象間の時間的順序関係解析
(I. Mani 2006)	解析手法	二事象間の時間的順序関係解析
TempEval (M. Verhagen et al. 2007)	評価型会議	時間情報表現- 事象間/ 二事象間の時間的順序関係解析
TempEval-2 (M. Verhagen et al. 2010)	評価型会議	時間情報表現- 事象間/ 二事象間の時間的順序関係解析
日本語に関する関連研究		
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
関根らの拡張固有表現体系 (S. Sekine et al. 2002)	タグつけ基準	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

英語においては、評価型国際会議 MUC-6 (R. Grishman and B. Sundheim 1996) の１タスク

[9] 数少ない ON_OR_AFTER の時刻表現は 「夜来」と「昼以降」の２事例。

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。タグづけ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 のタグづけ基準 ⟨TIMEX⟩ に対し、Setzer は時間情報表現の正規化に関するタグづけ基準を提案している(A. 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。ISO-8601 形式を拡張した @value 属性などが設計され、時間表現の正規化が自動解析対象となっている。その後、Pustejovsky らによりタグづけ基準 TimeML (J. 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タグづけすることを目的に設計され、TimeBank (J. 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(B.Boguraev and R. Kubota Ando 2005; I. 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、扱いにくいものであった(B.Boguraev and R. Kubota Ando 2006) 。2007 年に開かれた SemEval 2007 の 1 タスク TempEval(M. 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 (M. 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、スペイン語に関しても同様の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) における 1 タスクとして、固有表現抽出タスクが設定された。IREX における時間情報では、日付・時刻表現を対象にし、相対的な表現が定義に含めている。また、関根らは拡張固有表現体系(S. Sekine et al. 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 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。

６　おわりに

本稿では作成している日本語時間情報タグつきテキストコーパスについて説明した。タグつきデータはタグの情報のみ github 上に公開する。BCCWJ 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－文書作成日時および解析対象表現－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づ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 における⟨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 における EVENT@type) 、テンス・アスペクト体系の整備(同 MAKEINSTANCE@tense, MAKEINSTANCE@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。現在は事象表現を動詞に限定し、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)を基にした階層的ラベルを設計し付与している。階層的ラベルの上位の情報を得ることにより TimeML で定義されている

EVENT@type の 8 分類に対応する設計になっている。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。今後 TimeML に準じた事象表現に対するタグづけを行い、最終目標である事象表現に対する時間情報付与の研究を進めていきたい。

謝　辞

本研究は、文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築 : 21 世紀の日本語研究の基盤整備」(平成 18～22 年度、領域代表者 : 前川喜久雄) 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。

文　献

B. Boguraev and R. Kubota Ando (2005). “TimeML-Compliant Text Analysis for Temporal Reasoning.” In Proc. of the 19th International Joint Conference on Artificial Intelligence (IJCAI-05) , pp. 997–1003.

B. Boguraev and R. Kubota Ando (2006). “Analysis of TimeBank as a Resource for TimeML parsing.” In Proc. of the 5th International Conference on Language Resources and Evaluation (LREC-06) .

DARPA TIDES (2004). The TERN evaluation plan; time expression recognition and normalization. Working papers, TERN Evaluation Workshop.

R. Grishman and B. Sundheim (1996). “Message Understanding Conference-6: a brief history.” In Proc. of the 16th International Conference on Computational Linguistics (COLING-96), pp. 466–471.

IREX 実行委員会(1999). IREX ワークショップ予稿集.

I. Mani (2006). “Machine Learning of Temporal Relations.” In Proc. of the 44th Annual Meeting of the Association for Computational Linguistics (ACL-2006), pp. 753–760.

J. Pustejovsky et al. (2003a). “The TIMEBANK Corpus.” In Proceedings of Corpus Linguistics 2003, pp. 647–656.

J. Pustejovsky et al. (2003b). “TimeML: Robust Specification of Event and Temp oral Expressions in Text.” In Proc. of the 5th International Workshop on Computational Semantics (IWCS-5) .

S. Sekine et al. (2002). “Extended Named Entity Hierarchy.” In the proc. of the Third International Conference on Language Resources Evaluation (LREC-02) .

A. Setzer (2001). Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study . Ph.D. thesis, University of Sheffield.

M. Verhagen et al. (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” In Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pp. 75–80.

M. Verhagen et al. (2010). “SemEval-2010 Task 13: TempEval-2.” In Proc. of the 5th International Workshop on Semantic Evaluations (SemEval-2010), pp. 57–62.

工藤真由美(1995)『アスペクト・テンス体系とテクスト ―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

中村ちどり(2001)『日本語の時間表現』、くろしお出版

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.