『現代日本語書き言葉均衡コーパス』に対する時間情報表現アノテーションの再修正作業

浅原 正幸 (国立国語研究所) *
坂口 智洋 (京都大学)
渡邊 友香 (統計数理研究所)

Correction of Temporal Information Annotation on 'Balanced Corpus of Contemporary Written Japanese'

Masayuki Asahara (National Institute for Japanese Language and Linguistics)
Tomohiro Sakaguchi (Kyoto University)
Yuka Watanabe (The Institute of Statistical Mathmatics)

要旨

小西ほか (2013) は，『現代日本語書き言葉均衡コーパス』(Maekawa et al. (2014)) に対してジャンル横断的に時間情報表現の正規化情報を TimeML (Pustejovsky et al. (2003)) に準ずる規定に基づき付与した。その後，同規定に基づく時間情報正規化プログラム (坂口 (2015a), 坂口・黒橋 (2015b)) が開発された。今回，アノテーションデータとプログラム出力の齟齬の対照比較を行うことにより，再修正作業を行った。さらに，基準の曖昧な点を見直し，新たな属性を導入したので報告する。

1. はじめに

情報抽出において，事象表現の生起時刻（実時間軸上の時区間）や時間的順序関係を推定するために時間情報解析が行われている。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 et al. (2003)) が検討され，TimeML に基づくタグつきコーパス TimeBank (Pustejovsky et al. (2003)) 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 のサブタスク TempEval (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 のサブタスク TempEval-2 (Verhagen et al. (2010)) では，英語

* masayu-a@ninjal.ac.jp

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 年に開かれた 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。2013 年に小西ほか (2013) は，『現代日本語書き言葉均衡コーパス』(BCCWJ)(Maekawa et al. (2014)) の一部のジャンル横断的に時間情報表現の正規化情報を TimeML に準ずる規定に基づき付与した。その後，時間情報表現と事象表現との時間的順序関係として，TimeBank の TLINK 相当の情報を被験者実験的に付与し (保田ほか (2013))，BCCWJ-TimeBank(Asahara et al. (2014)) として公開された。

小西ほか (2013) の作業開始時は，原始的な時間情報表現解析器を開発していたものの，解析器の出力を確認しながらアノテーションの修正を行う MATTER サイクル (Pustejovsky and Stubbs (2012)) を行うには至らず，作業者 1 名と教示者 1 名によりディスプレイを共有しながらペアプログラミング的にアノテーションを行う変則 MAMA サイクルを行うにとどまっていた。

その後，同規定に基づく時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) が開発された。第 2 著者である解析器開発者よりアノテーションの誤りが報告され，修正を実施した。解析器開発者より都合 3 回の誤り報告を受けながら，アノテーションデータと解析器出力の齟齬の対照比を行うことにより，ゆるい MATTER サイクルにより再修正作業を行った。再修正作業において，解析器構築や実応用の観点から，基準の修正を行った。基準の修正に際して，新たな属性を導入したので報告する。

2. アノテーション開発サイクルと修正作業

2.1 アノテーション開発サイクル

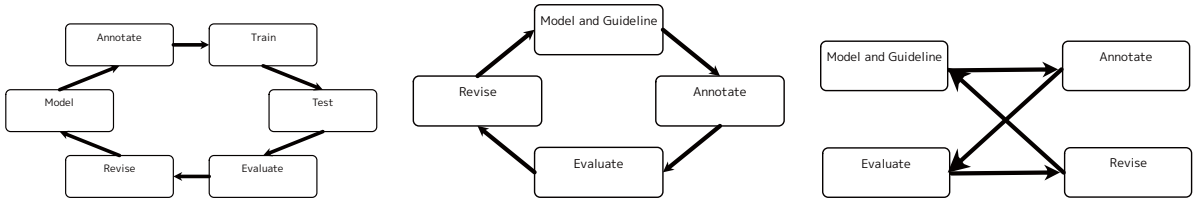
本節ではアノテーション開発サイクルについて述べる。アノテーション開発サイクルにおける各段階について Pustejovsky and Stubbs (2012) が次のように定義している(1)。

- Model: モデル化
  データを通した経験的な観察から生成される，理論的に説明可能な属性を与える構造的描写
- Annotate: アノテーション入力データの特定の構造的描写や性質をコード化する，特徴量集合を過程したアノテーション体系
- Train: 訓練
  対象となる特徴量集合がアノテーションされたコーパスを用いた解析器の訓練
- Test: 検証
  訓練データとは別に定義したテストデータ上での解析器の検証

(1) Pustejovsky and Stubbs (2012) pp.22-32 もしくは Pustejovsky (2006)

- Evaluate: 評価
  解析結果に対する標準化された評価
- Revise: 再検討
  アノテーションが機械学習アルゴリズム中で用いられる際により頑健で信頼できるものにするためにモデルとアノテーション基準を再検討する

これらの 6 つのステップを介したアノテーション開発サイクルを MATTER サイクルと呼ぶ。図 1 左に MATTER サイクルを示す。



MATTER サイクル　　MAMA サイクル　　変則 MAMA サイクル

図 1　アノテーション開発サイクル

しかしながら，アノテーション初期は適切な解析器が構成されることは少なく，MATTER サイクルの一部である MAMA サイクル (図 1 中) を用いる事が多い。「Evaluate: 評価」のステップにおいて，MATTER サイクルでは人手付与データ (GOLD) と解析器出力データ (SYS) との一致度をはかることが行われるが，MAMA サイクルではアノテーション作業者間の一致率 (Inter Annotator Agreement: IAA) をはかることが行われる。Passonneau and Carpenter (2014) はこの作業者間の一致率に対して確率モデルを導入することを提案している。

また，小西ほか (2013) は，BCCWJ-TimeBank の時間情報表現アノテーションにおいて，「Model(モデル化)」「Evaluate(評価)」を行う人と「Annotate(アノテーション)」「Revise(再検討)」を行う人と分業したうえで，同一の画面を見ながら作業を行うペアプログラミング的な方法を導入した。

保田ほか (2013) は，不定な時間的順序関係情報付与において，あらかじめ 3 人の作業者に基準を教示してから，その後相互に作業過程をフィードバックしないようにして，被験者実験のように作業者間でゆれる言語現象を明らかにするような方法を導入した。

本研究では，小西ほか (2013) のアノテーション作業が MAMA サイクルでとどまっていること，さらに同規定に基づく時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) が開発されたことから，解析器を用いた MATTER サイクルに基づく修正を行う。

尚，本稿では，上記に述べた MAMA サイクルや MATTER サイクルを介さない，人手によるパターン・規則のみに基づく解析器の出力をアノテーションとは呼ばない立場をとる。

2.2　解析器の概要

本節では，本研究で用いた時間情報正規化解析器 (坂口 (2015a), 坂口・黒橋 (2015b)) について簡単に示す。

解析器は時間情報表現を表す ⟨TIMEX3⟩ タグの 1. 認識， 2. TYPE 属性推定，3. VALUE 属

性推定の 3 段階からなる。時間情報表現の認識とは，⟨TIMEX3⟩ の開始タグと終了タグの挿入すべき位置を推定することである。TYPE 属性推定とは ⟨TIMEX3⟩ タグに付与された type 属性 4 種 { “DATE(日付表現)”, “TIME(時刻表現)”, “DURATION(時間表現)”, “SET(頻度集合表現)”} を推定することである。VALUE 属性推定とは時間情報表現が指し示す時刻・時間を正規化して機械可読形式に変換することである。⟨TIMEX3⟩ タグには，valueFromSurface 属性と呼ばれる表層文字列のみを用いて推定する正規化情報と，value 属性と呼ばれる前後文脈などの情報を用いて推定する正規化情報の 2 種類が定義されている。

坂口 (2015a) の時間情報正規化解析器は，「1. 認識」と「2.TYPE 属性推定」を系列ラベリング問題として同時に解き，「3.VALUE 属性推定」を正規化ルールに基づく valueFromSurface 属性推定と照応解析の手法を用いた value 属性推定の二段階の手法を用いて解いている。

「1. 認識」と「2.TYPE 属性推定」は，形態素解析器 JUMAN と係り受け解析器 KNP の出力から抽出した特徴量を用いた条件付確率場 (Lafferty et al. (2001)) による。特徴量として，見出し語・品詞・原形などの形態論情報，係り受け先の動詞，JUMAN 辞書に登録されている時相動詞か否か，記号か否か，数値の大きさを表すカテゴリ，200 近くの正規化ルールとの適合などを用いている。

「3.VALUE 属性推定」は最初に正規化ルールに基づく書き換え系により valueFromSurface 属性を復元する。valueFromSurface 属性が特定の時区間を示す定時間情報表現の場合には valueFromSurface 属性の情報がそのまま value 属性になる。曖昧性が残るような不定時間情報表現の場合に，曖昧性解消を行うことで value 属性を復元する。曖昧性解消は先に言及された定時間情報表現を参照し，復元する。参照すべき定時間表現の探索を，照応解析問題の一つである橋渡し参照問題として定式化し，SVM-Rank(Joachims (2003)) を用いたランキング学習を用いて解析する。

本研究の修正以前の解析器の性能は以下の通りである：

表 1　解析器の性能 (坂口・黒橋 (2015b))

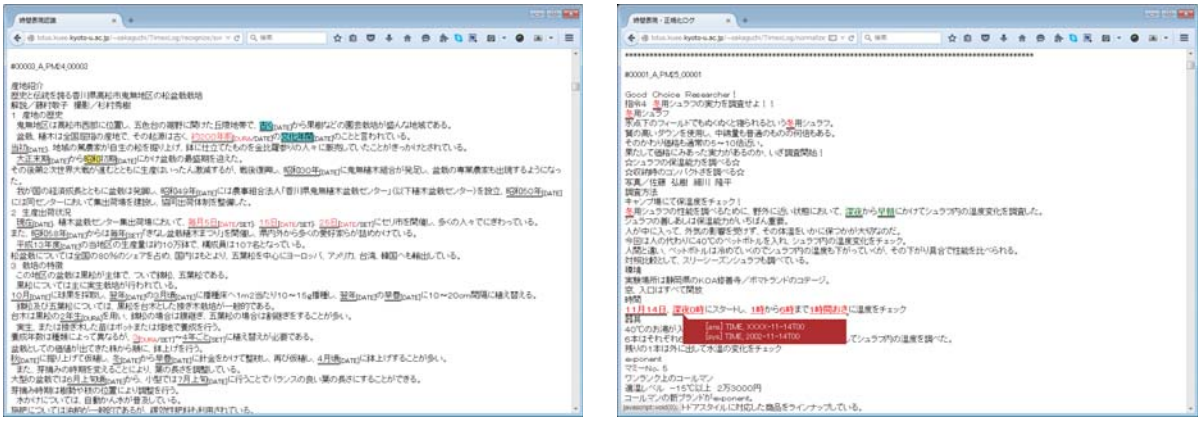
	Precision	Recall	F-value
「1. 認識」「2.TYPE 属性推定」	0.86	0.82	0.84
「3. VALUE 属性推定」	0.64	0.61	0.62

2.3　修正作業

修正作業は解析結果を図 2 のように可視化したものを，作業者に提示して行う。作業者は解析器の出力を見て，解析器が正しいか，既存のアノテーションが正しいかを確認しながら，oXygen XML Editor(図 3) 上で人手により修正を行う。

3.　基準の修正

3.1　括弧と時間情報表現範囲



認識結果　　正規化結果

図 2　修正作業に用いた解析結果の可視化

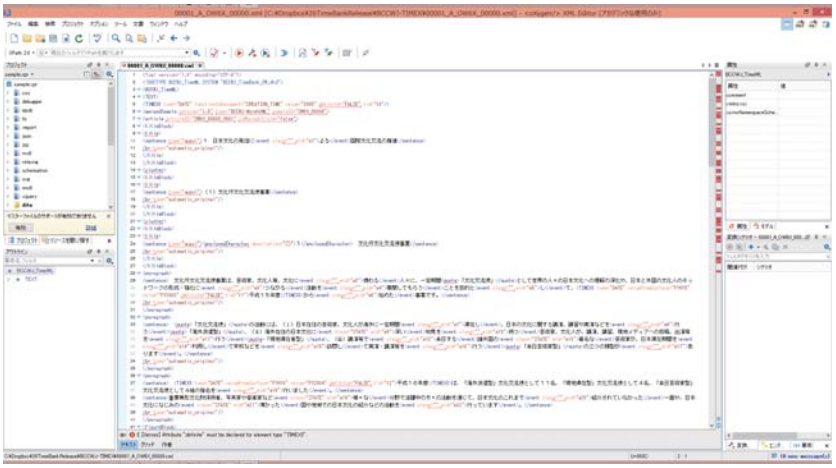


図 3　oXygen XML Editor

以前の基準 (浅原 (2012))[(2)]では「十四日午後一時半 (日本時間同七時半) 過ぎ」を「十四日午後一時半」と「 同七時半) 過ぎ」に分割して時間情報表現範囲を規定していたが，「日本時間」も時間表現の一部で value にも反映されていることから，まとめて「時間表現が連続する場合は括弧を含め，そうでない場合は括弧を含めない」という基準にする。

これによって西暦と和暦の併記の問題も同様に扱える。文書中に「1999(平成 11) 年」などの表現が出現した後は「2000(平成 12) 年 5 月」を「00(同 12) 年 5 月」などと省略することがあり，この場合，以前の基準では「00」と「(同 12) 年 5 月」に分割していたが，西暦と和暦を併記するときは両方が同じ年を指すため，まとめて「00(同 12) 年 5 月」にタグを付けることとする。

(2) 特に指定しない限り，「以前の基準」とは 2012 年 6 月 30 日現在の Version 0.10 のマニュアル相当のものとする。

3.2　type=DATE/TIME の value，valueFromSurface における DURATION の値の扱い

以前の基準では、type=DATE/TIME の value や valueFromSurface に DATE/TIME の形式の値　だけでなく DURATION の形式の値を取ることも許しており、type によって DURATION の形式を　異なる意味で用いているという問題、type=DATE/TIME の一部の時間情報表現では value や　valueFromSurface の値を DATE/TIME の形式と DURATION の形式の両方で書くことが可能なため　どちらで書くか明確に決まっていなかったという問題があった。

例えば、日付表現「３日目」「３日前」「３日後」や時刻表現「５時間前」「５時間後」は value に DURATION 形式を許している。

「３日目」
<TIMEX3 type="DATE" value="P3D">３日目</TIMEX3>
「３日前」
<TIMEX3 type="DATE" value="P3D">３日前</TIMEX3>
「５時間前」
<TIMEX3 type="TIME" value="PT5H">５時間前</TIMEX3>
「３日後」
<TIMEX3 type="DATE" value="P3D">３日後</TIMEX3>
「５時間後」
<TIMEX3 type="TIME" value="PT5H">５時間後</TIMEX3>
cf.)「３日間」
<TIMEX3 type="DURATION" value="P3D">３日間/TIMEX3>

上記の方針の場合，DURATION の値が DATE や TIME の valueFromSurface でも用いられ，DURATION の valueFromSurface と同じ記号で異なる意味を表すことになる。この混同を防ぐため，DATE や TIME の valueFromSurface 用に DURATION とは異なる，また，より表層的な情報を反映させることができる新たな記号 Q を定義する。

「３日目」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q3D">３日目</TIMEX3>
DATE のため，value の値は XXXX-XX-XX の形式となる。
「３日前」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q-3D">３日前</TIMEX3>
「前」という情報を残すために - (マイナス) を付与する。
「５時間前」
<TIMEX3 type="TIME" value="TXX" valueFromSurface="Q-T5H">５時間前</TIMEX3>
時間を表すために T を，また「前」という情報を残すために - を付与する。
「３日後」
<TIMEX3 type="DATE" value="XXXX-XX-XX" valueFromSurface="Q+3D">３日後</TIMEX3>
「後」という情報を残すために + (プラス) を付与する。
「５時間後」
<TIMEX3 type="TIME" value="TXX" value="PT5H">５時間後</TIMEX3>
時間を表すために T を，また「後」という情報を残すために + を付与する。

3.3 総称表現の識別

VALUE に含まれる X には 2 つの意味合いがある。1 つ目はある特定の時間を表すわけではない一般的な表現 (総称) で，2 つ目は特定の時間を表すがその文書での情報のみでは判別できない表現である。

例えば次の 2 つの例の「夏」は両方とも同じ VALUE となるが，X を用いる理由が異なると考えられる。

1. 特定の時間を表さない総称の例
   ・「京都の夏は暑い。」の「夏」が”XXXX-SU”
   (文書作成日時等に関わらず，XXXX となる)
2. 特定の時間を表すが，文書の情報だけでは判別できない例
   ・「その年の夏は暑かった。」の「夏」が”XXXX-SU”
   (もし「その年」が文脈から，例えば 2015 年と分かる場合は”2015-SU”となる)

これらの違いを機械で判断するのは難しい。解析器側で照応解析相当の処理を行っているが，その際に各時間情報表現が総称表現かあらかじめ情報として付与されていれば，参照すべきか否かを判別することができる。

そこで新たに属性 general を付与する。時間情報表現が総称表現である場合に general=TRUE を付与する。

```
<TIMEX3 type="DATE" value="XXXX-SU" valueFromSurface="XXXX-SU" general=TRUE>
夏</TIMEX3>
```

ここで総称表現として想定しているのは，具体的には次のようなものである。

- 「大津西小学校は，夏休み に家庭訪問を実施していた」
- 「でも，それだけだとなんなので，夏 野菜たっぷりサラダ」
- 「お問い合わせは 平日 10 時-18 時」
- 「五千万円を上限とする 年末 ローン残高の 1% を控除する」
- 「「レゴレゴ (0505)」と読めることなどから，5 月 5 日 は「レゴの日」」

4. 修正件数

表 2 に修正前件数と修正後件数を示す。約 6000 件のデータについて，修正がされなかった時間情報表現は 866 件であった。

表 2　修正前後の件数の変化

	DATE	TIME	DURATION	SET	ALL
修正前件数	4543	501	1211	151	6406
修正後件数	4755	513	1235	158	6661

表 3 に valueFromSurface に付与された Q+, Q- の件数を示す。

また，今回付与した general 属性が TRUE であった件数を表 4 に示す。

表 3　value 属性に付与された Q+, Q- の件数

	Q+	Q-	ALL
件数	13	42	55

表 4　総称表現の件数

	DATE	TIME	DURATION	SET	ALL
general=True	100	28	0	0	128

5. おわりに

本稿では，2015 年春に行った BCCWJ-TimeBank の時間情報表現アノテーションの再修正作業について報告した。

今後の課題として以下をあげる：

- TLINK 情報の再付与
  保田ほか (2013) は，時間的順序関係について 3 人の作業者に被験者実験的に行った。6688 個の関係については 3 人のアノテーションが一致したが，3011 個の関係が 2 人の一致にとどまり，545 個の関係は全く一致していない。これらの分布を評価するために，3 人の作業者で一致しなかった関係に対する情報付与を数千人規模で被験者実験的に行う。
- SLINK 情報のアノテーション
  SLINK は主節 (matrix clause)-従属節 (subordinate clause) 間の関係を規定するアノテーションである。英語以外のデータに対して付与されていることが少なく，多言語化における課題となっている。SLINK の関係ラベルは，’MODAL’, ’EVIDENTIAL’, ’NEG_EVIDENTIAL’, ’FACTIVE’, ’COUNTER_FACTIVE’, ’CONDITIONAL’ と事象の事実性に関するもので構成されている。英語においては FactBank (Saurí and Pustejovsky (2009))[(3)]として事実性解析用途に拡張している。これらのラベルを日本語に適合するために，鳥バンクの節分類を第 3 階層 (池原悟 (2007)) のレベルで付与を行い，その後，節分類の情報をもとに時制節性 (有田 (2007)) を付与することで SLINK 相当の情報を付与していきたいと考えている。

謝辞

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

有田節子 (2007).『日本語条件文と時制節性』　くろしお出版.

(3) FactBank 1.0 https://catalog.ldc.upenn.edu/LDC2009T23

Asahara, Masayuki, Sachi Kato, Hikari Konishi, Mizuho Imada, and Kikuo Maekawa (2014). “Bccwj-timebank temporal and event information annotation on japanese text.” *International Journal of Computational Linguistics and Chinese Language Processing*, 19:3, pp. 1–24.

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

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

池原悟 (2007).「意味類型パターン記述言語仕様書」　Technical report, 独立行政法人科学技術振興機構, 戦略的基礎研究事業, 高度メディア社会の生活情報技術.

IREX 実行委員会 (1999).『IREX ワークショップ予稿集』.

Joachims, T. (2003). “Optimizing search engines using clickthrough data.” *Proc. of the ACM Conference on Knowledge Discovery and Data Mining*.

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.

Lafferty, J. D., A. McCallum, and F. C. N. Pereira (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *Proc. of 18th International Conference on Machine Learning*, pp. 282–289.

Maekawa, Kikuo, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.

Passonneau, Rebecca J, and Bob Carpenter (2014). “The benefits of a model of annotation.” *Transactions of the Association for Computational Linguistics*, 2, pp. 311–326.

Pustejovsky, J. (2006). “Unifying linguistic annotations: A timeml case study.” *Proceedings of the Text, Speech, Dialogue Conference*.

Pustejovsky, J., and A. Stubbs (2012). *Natural Language Annotation*.: O’Reilly.

Pustejovsky, J., J. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, and G. Katz (2003). “TimeML: Robust Specification of Event and Temporal Expressions in Text.” *Proceedings of the 5th International Workshop on Computational Semantics (IWCS-5)*, pp. 337–353.

Pustejovsky, J., P. Hanks, R. Saurí, A. See, R. Gaizauskas, A. Setzer, B. Sundheim, L. Ferro, M. Lazo, I. Mani, and D. Radev (2003). “The TIMEBANK Corpus.” *Proceedings of Corpus Linguistics 2003*, pp. 647–656.

坂口智洋 (2015a).「時間表現の解釈に基づく言明の抽出と整理」　修士論文, 京都大学大学院情報学研究科.

坂口智洋・黒橋禎夫 (2015b).「多様な時間表現の解釈に基づく言明の抽出と整理」　情報処理学会第 77 回全国大会, pp. 2–201–2–202.

Saurí, Roser, and James Pustejovsky (2009). “Factbank: A corpus annotated with event factuality.”

Language Resource and Evaluation, 43:3, pp. 227–269.

Verhagen, M., R. Gaizauskas, F. Schilder, M. Hepple, G. Kats, and J. Pustejovsky (2007). “SemEval-2007 Task 15: TempEval Temporal Relation Identification.” *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp. 75–80.

Verhagen, M., R. Saurí, T. Caselli, and J. Pustejovsky (2010). “SemEval-2010 Task 13: TempEval-2.” *Proceedings of the 5th International Workshop on Semantic Evaluations (SemEval-2010)*, pp. 57–62.

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情報・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.

浅原正幸 (2012).『BCCWJ-Timebank 日本語時間表現タグづけ基準 version 0.10』.