社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2008－ＦＩ－ 89 (11)
2008－ＮＬ－183 (11)
2008／1／22

事態抽出のための事実性解析

原一夫　　乾健太郎
奈良先端科学技術大学院大学 情報科学研究科
〒 630-0192 奈良県生駒市高山町 8916-5
{kazuo-h,inui}@is.naist.jp

本研究では事態の事実性解析に焦点をあてる。テキストに記述されるすべての事態を対象にした情報抽出を行う上で、事実性解析は述語項構造解析とともに必須の技術といえるが、これまでのところ、十分には研究されていない。本稿では、事実性解析の応用例としてブログ記事からの経験抽出を想定し、そのための事実性解析を時間情報、極性、話者態度の観点から抽象化する枠組みを提案する。また、ブログ記事を対象とする予備実験についても報告する。

Factuality Analysis for Event Extraction

Kazuo Hara　　Kentaro Inui
Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma Nara 630-0192 Japan
{kazuo-h,inui}@is.naist.jp

In this paper we concentrate on the analysis of the events in terms of factuality. We consider that the identification of factuality is an inevitable step towards practical information extraction after completing the predicate argument structures, however, it has not been studied well up to now. Then, as a case study, we shall propose a framework for examining the factuality of personal experiences in terms of tense, aspect, and mood, and we also give experimental results using Web log texts.

1 はじめに

自然言語テキストからの情報抽出研究は、これまでは固有表現や関係の抽出を主な目的として進められてきた。たとえば、その先駆け的な研究プログラムである MUC (Message Understanding Conference) や ACE (Automatic Content Extraction) では、あらかじめタスクを設定し、抽出する関係を限定した上で (例えば、テロ事件のタスクでは事件の日付、場所、武器、実行犯、テロ標的を抽出するなど)、固有表現同定、照応解析、抽出パターンの自動獲得等の技術を向上させてきた。

これに対して、抽出する関係や事態の種類を限定せず、テキストに記述されるすべてのイベント／事態を抽出することを目的とする研究も行われるようになってきている。一般に、事態は格要素 (項) をともなう述語あるいは事態性名詞で表現されるため、正しく述語項構造解析できることが事態抽出のために必須となる。述語項構造解析については、英語では PropBank[5] や NomBank[4] などの資源を用いた研究が盛んであり、日本語でも例えば Kawahara ら [2] や Iida ら [1] の研究が成果を上げている。

その一方で、事態の事実性を同定する研究はあまり進んでいない。事態は、既に事実として起こったこと (テロ事件や企業合併結果など) や真実であると証明されたことだけでなく、まだ計画や仮説でしかない不確実なものも含む。事態の種類を限定しない情報抽出を行うとき、これらを混在させたままの解析結果を提示することは利用者にとってあまり意味がなく、抽出した事態の事実性をテキストから読み取り、解析結果に付与することが望まれる。

以上の背景のもと、本研究はあらゆる種類の事態に対する事実性解析を最終的な目的とするが、本稿ではまず手始めとして、ブログ記事から個人の経験を抽出するタスク (我々はこれを**経験マイニング**と呼んでいる) をアプリケーションとして想定し、経験抽出を目的とする事実性解析について論じる。

経験マイニングでは、商品やサービスなど、様々な事物 (トピック) に関する個人の経験を広く Web

文書集合から抽出し、述語項構造に基づく表現形式に構造化するとともに、トピック、著者、事態の評価極性、事実性情報等のきめ細かい情報で索引付けする。これによって、Webユーザが他者の経験情報を活用し、互いの経験から学び合う新しい情報サービスを構築することができると考えている。経験の分類において、事態の評価極性および事実性の情報はとくに重要である。例えば、次の例のように、ネガティブな事態が事実として起こっていれば、著者が「トラブル」を経験したことになるし、ポジティブな事態が成り立っていないことが述べられていれば、それは著者の「不満」の表出であると解釈できる。

(1) ランプがつかない (negative) ときがある (事実)

(2) もう少し 中低速のトルクがあれ (positive) ばいいのに (反事実)

このように、評価極性と事実性の情報を組み合わせれば、例えば個々の事態を「利用成功」「トラブル」「トラブルの未然回避」「トラブルの解消」「満足」「不満」「安心」「心配」といったクラスに分類することができ、これまで主として評価極性（ポジティブ／ネガティブ）だけで分類してきた意見分類を、それを包含する経験分類に一般化することができる。

2 事態の事実性解析

本稿では、言語で書かれた人間のあらゆる経験（事態）を解析の対象とする。事態は、コアとなる述語（事態表現と呼ぶ）および、それを修飾する時間情報（テンス・アスペクト）、極性（成立／不成立）、話者態度（モダリティ）から構成されると考えられる。たとえば、

(3) 商品Aは店舗Bで3割 値引きし てました

では、事態表現は状態変化を表わす「値引きする」であり、機能表現「て・ます」によって状態変化の結果残存状態というアスペクトが加えられ、「た」によって過去にその事態が成立していたという宣言（テンスとモダリティ）が付加される。別の例、

(4) 明日から飲料水Aを 飲み 始めるつもりです

では、行為を表わす「飲む」という事態表現が、「始める」による行為の開始のアスペクトと「つもり」による意志のモダリティが加えられている。

さらに、事態の記述には、次の3通りのタイプがあると考えられる。

事実認識 世界（個人の内外）の認識、すなわち事実性に関する記述

意図 認識した世界を前提として発生する、意志、関心、欲求、願望などの意図や、仮定についての記述

事実認識と意図 以上をどちらも含む記述

(3) の例は「値引きしていた」という事実認識のみの記述であるが、(4) は「今は飲んでいない」という事実とともに、これを前提として「明日から飲む」という意図も述べている[1]。

したがって、言語で書かれた事態を理解するためには、

- 事態表現の種類（行為／状態／状態変化）
- 事実認識が含まれている場合は確信度
- 意図が含まれている場合はその内容

について詳しく調べる必要がある。ただし、たとえば、次の例

(5) きのうは、もし安ければ、飲料水Aを 買う つもりだった

では、「買うつもりだった」意図と「買わなかった」事実認識の両方がおそらく言及されているが、「もし安ければ」の記述がなかったとしたら、「買わなかった」事実について言及されているというのは多少無理がある。

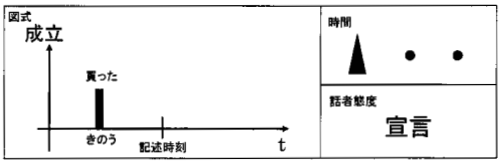
本研究ではできるだけ無理のない解釈ができる範囲で事実性に焦点を当て、まずは図1、図2の中の図式に示すような仕方で事態を把握することから始める。縦軸は事態の成立・不成立（上が成立、原点が言及なし、下が不成立）、横軸は時間（記述時刻をアンカーとしてもつ）である。テキストから読み取れる事実認識あるいは意図は、事態表現が瞬間的な行為や状態変化である場合は黒塗りの長方形（棒状）で表し、状態（結果残存状態を含む）である場合は網掛けの長方形で表わす。

この図式を基にして、事態をさらに、時間情報（テンス・アスペクト）、話者態度（ムード）の観点から抽象化する。以下、これらについて説明するが、時間情報についてのやや詳しい仕様は付録にまとめた。

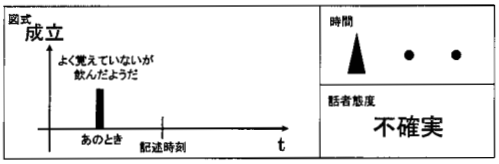
2.1 時間情報と極性

時間情報と極性（事態の成立／不成立）は、過去、現在（＝記述時刻）、未来のスロットからなる3つ組で表す。それぞれのスロットには、{▲、■、□、↑、↓、×、・} のいずれかのラベルが入る。単発的な瞬間的事態（状態変化や行為など）の成立を▲、瞬間的事態の反復的継続の成立を■、状態等の継続的事態の成立を□、反復的事態または継続的事態の開始を↑、終了を↓、事態の否定（不成立）を×、言及なしを・で表す。ただし、図の (j), (k), (l), (n) の例のように、記述時刻と関連がないことがある。その場合は、便宜的に過去と未来のスロットを言及なし（・）とし、現在のスロットのみに記号を入れる。なお、本研究では開始、継続、終了などのアスペクト表現（～し始める、～にハマる、～するのをあきらめる、～するのをやめる、等）は事態とみな

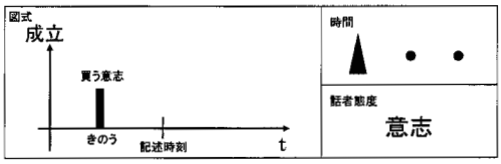
[1] 「明日から飲む」という意図から推論できる事実として、「今は飲んでいない」が成り立つと考えることもできる。



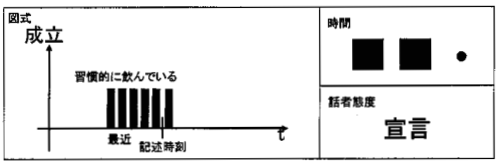
(a) きのう、飲料水Aを 購入して 飲んだ



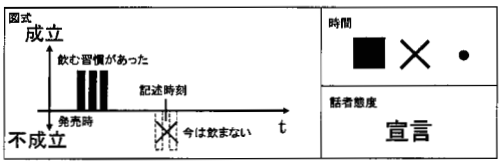
(b) そういえば、あのとき飲料水Aを 飲んだかもしれない



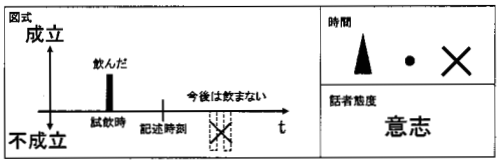
(c) きのうは、飲料水Aを 買うために スーパーへ行ってきた



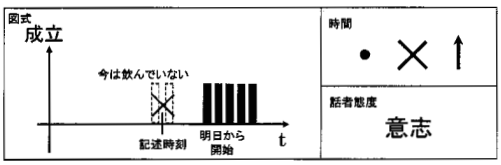
(d) 最近、飲料水Aを 飲むのに ハマっている



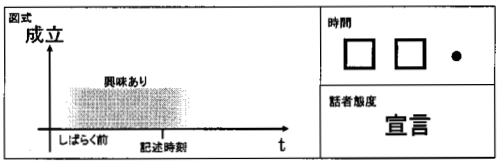
(e) 発売当初は、飲料水Aを 飲んでいた



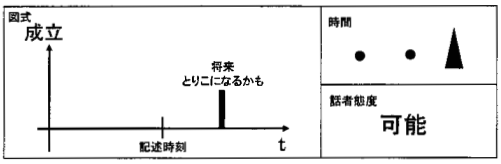
(f) もう二度と飲料水Aは 飲まないぞ



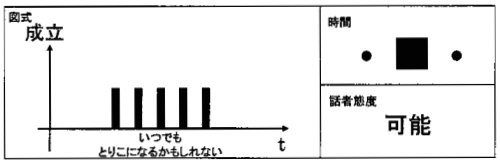
(g) 明日から飲料水Aを 飲み始めるつもりです



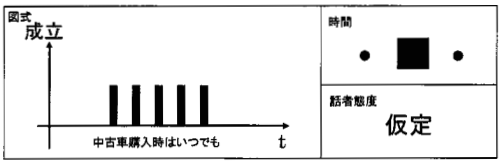
(h) しばらく前から、飲料水Aが 気になってるんですが



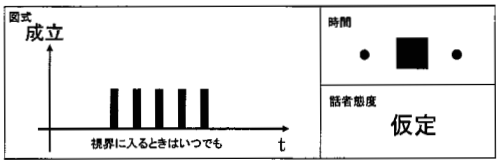
(i) これは、(自分にとって) クセになる 味かも



(j) これは、(一般的に) クセになる 味かも

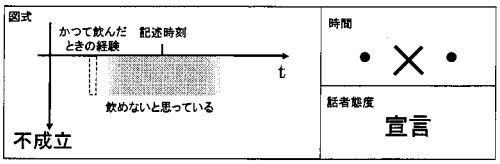


(k) (一般的に) 中古車を 買う ときは、注意が必要だ

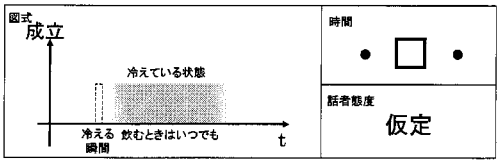


(l) 自動車Bが 走ってると、つい見ちゃう

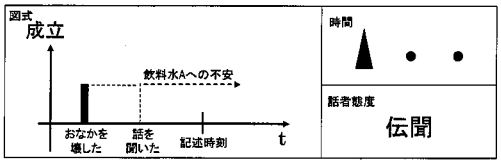
図 1: 事態を時間情報、極性、話者態度で抽象化する試み



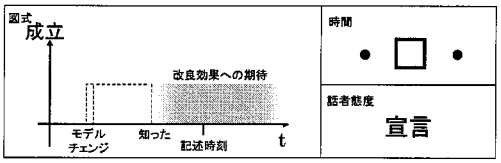
(m) 飲料水Aを まるごと1本飲む なんて不可能だ･･･



(n) 冷えた 飲料水Aを飲みたい



(o) 飲料水Aを飲んで おなかを壊した 人の話を聞いた



(p) 自動車Bは、エンジンにも 改良が加えられている

図 2: 事態を時間情報、極性、話者態度で抽象化する試み（続き）

さず、主となる事態を補助するための表現と位置づける。

2.2 話者態度

話者態度は、極性以外の心的状態を抽象する。事態成立の真偽を問うもの、すなわち、事態の事実性に関連するもの；

宣言（事実）、確信、不確実、伝聞、保留、疑問、反実仮想、反語

と、それ以外；

仮定、予定、可能、願望、意志、質問、推奨、当為

をタグとして用意する。一つの事態に対して、伝聞・予定などのように複数のタグが付与されることもある。従来、話者態度を表す表現としては、「よう」「まい」「らしい」のような機能語や「～するべきだ」「～するところだった」のような複合辞が主な分析対象だったが、本研究では、「～と思う」（宣言）、「～の感がある」（不確実）、「～という話を聞いた」（伝聞）、「～というのは信用しかねる」（疑問）など、内容語を含む表現もそれが話者態度を表すなら広く話者態度表現として解釈する。したがって、これを解析するには、従来の機能表現辞書（例えば松吉らの辞書 [7]）に加え、話者態度を表しうる表現を広く識別できる資源あるいはモデルの開発が必要である。

3 事実性解析のためのタグ付きコーパスの作成

以上のような事態の抽象化を、計算機で自動化して行うために、本研究では、まずテキストに出現する事態に対して各情報（時間情報、話者態度）を人手で付与してタグ付きコーパスを作成し、それを訓練データとして用いて解析モデルを学習する方法をとる。そのための予備的な試みとして、次の手順でタグ付けコーパスを作成した。

まず、ドメインを選び（ここでは、価格や買い換え頻度など、やや性質の異なる2つのドメインとして、飲料水と自動車を選択）、対象商品名を決定した後（飲料水 A、自動車 B）、商品名をキーとして blog をクローリングする。得られた blog を茶筌 [3] を用いて形態素解析および文節区切りを施す。タグを付与する対象は、動詞および動詞化しているサ変名詞（名詞-サ変接続の直後に動詞-自立がある場合）に限定する[2]。さらに、対象とする商品と直接関係のない表現（e.g. 飲料水 A を買って 帰りました）、アスペクト表現（e.g. ～し始める）、モダリティ表現（e.g. ～と思う、～の話を聞いた）、機能語相当表現（e.g. ～ことができる、～ことにする、～といっても、～てもらう、～といえば、～ことがある、～てみる、等）を人手によって同定し、タグ付け対象から外す。こうして、タグ付け対象を特定した上で、各情報（事態タイプ、時間情報、話者態度）のタグを前後一文を見て読み取れる範囲で付与した。時間情報タグについては、言及なしかどうか定かでない場合は、？付きのタグを用いた（□?、■?、等）[3]。

タグ付けは一人の作業者により行った。同一データに対し、2週間の間隔を空けて2度タグ付けをしてもらい、判定者内一致（intra-rater agreement）による κ 統計量は 0.8 以上を得た。

4 実験

前節のコーパスを使用して、時間情報と話者態度について予測実験を行った。学習モデルは、過去、現在、未来のラベル系列からなる時間情報に対し

[2]人間の経験は、形容詞や形容動詞によっても表現されるが（たとえば評価表現）、経験マイニングのための事実性解析を行う本稿では、意見マイニングでは扱わない行為に関連する事実性について重点的に調べたいため、こうした限定を行った。

[3]たとえば、図 1 の (h) で、「しばらく前から」という記述がない場合は（□?, □, ・）とする。

ては Hidden Markov SVM を、話者態度については multi class SVM を使用した [6] 。素性は、予測対象の文節、その前後の文節、文全体を区別した上で、品詞と原型を組み合わせたものを用いた。

leave-one-out による実験結果（正解率）を表 1 に示す。時間情報の欄には、過去、現在、未来が揃って正解した場合の正解率を記載した。また、正解率の良し悪しを評価するためのベースライン指標として、各タグの頻出ラベルの割合を表 2 に示した。

5 おわりに

本稿では、経験を時間情報、極性、話者態度の観点から抽象化する枠組みを提案した。さらに、ブログ記事からの経験抽出を応用例として想定し、実際にコーパスを作成し、予備的な実験を行った。

我々の目的は、例えば時相論理のような過度に複雑な意味表現を導入することなく、経験抽出/分類のような事態抽出の応用に広く有益で現実的な事実性解析の枠組みを設計し、それを実現する解析モデルを開発することである。本稿ではその最初の試みを報告したが、課題も多い。とくに、2 節で述べた枠組みでは表現できない情報が種々残っており、時間情報、極性、話者態度それぞれの精緻化が必要である。たとえば、

(6) 妹夫婦は 買った ばかりの新車で来ていた

では、買うという行為に注目すると、(▲, ・, ・) となるが、買ったばかりという状態と考えると、(・, □, ・) になる。また、次の例のような部分否定は現在の枠組みでは適確に表現できない。

(7) すべての自動車 B にその 故障が起こる ようではないみたいだけど

次に解析モデルについてであるが、今回の実験では個々の事態の事実性解析を独立した問題として扱った。しかし、次のような例からも容易に想像できるように、一連の文脈のなかでいくつかの事態が言及されている場合、それらの事実性の間には、例えば主節が過去であれば従属節も過去になりやすいといった依存関係があると考えられる。

(8) 飲料水 A は、体脂肪を 分解する のに 効果がある

(9) 飲料水 A を 飲みすぎて おなかを 壊した 人の話を聞いて

今後は、事態間の事実性の依存関係も含めて学習できるようなモデルを設計し、実験を進める予定である。

最後に、関連研究についてであるが、事態を抽象化する試みとしては、TimeML プロジェクト[4]がある。TimeML では、各事態の時間に関する曖昧性（たとえば事態間の時間的順序）を機械で解消するためのタグの仕様およびタグ付きコーパスを開発している。しかし、事実性については SLINK の一部で扱おうとはしているものの、

(10) 発売当初は飲料水 A を何度か 飲んでいた のですが

の例で、「現在は飲んでいない」という記述から読み取れる事実に対応できるようなタグは用意されていない。逆にいえば、本研究はこのような事実性解析へのチャレンジを行う。

[4] http://timeml.org/site/index.html

謝辞

本研究は、文科省科研費特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」の公募研究「経験マイニング：Web 文書からの個人の経験の抽出と分類」（19024057，代表：乾健太郎）、およびニフティ株式会社から支援を受けた。記して深く感謝する。

参考文献

[1] Iida, R., Inui, K. and Matsumoto, Y.: Exploiting Syntactic Patterns as Clues in Zero-Anaphora Resolution, *Proc. of COLING-ACL*, pp. 625–632 (2006).

[2] Kawahara, D. and Kurohashi, S.: A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis, *Proc. of HLT-NAACL*, pp. 176–183 (2006).

[3] Matsumoto, Y.: Morphological Analysis System ChaSen: Easy-to-Use Practical Freeware for Natural Language Processing, *Journal of Information Processing Society of Japan*, Vol. 41, No. 11, pp. 1208–1214 (20001115).

[4] Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B. and Grishman, R.: The NomBank Project: An InterimReport, *Proceedings of the HLT-NAACL Workshop on Frontiers in Corpus Annotation* (2004).

[5] Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles, *Computational Linguistics*, Vol. 31, No. 1, pp. 71–106 (2005).

[6] Tsochantaridis, I., Joachims, T., Hofmann, T. and Altun, Y.: Large Margin Methods for Structured and Interdependent Output Variables, *Journal of Machine Learning Research (JMLR)*, Vol. 6, pp. 1453 – 1484 (2005).

[7] 松吉俊, 佐藤理史: 体系的機能表現辞書に基づく日本語機能表現の言い換え, 言語処理学会第 13 回年次大会発表論文集, pp. 899–902 (2007).

A 付録 1：時間（テンス）

ここでは事態の成立時間を、過去、現在＝著者の記述時刻、未来によって識別する。

たとえば、

(11) 照明が落としてある（・, □, ・）

を用いると（コンテクストから、極性判定結果＝ポジティブな出来事、が得られているとする）、著者にとってポジティブな出来事が、現在事実として成立していることを表している。しかし、

(12) かつては照明が落としてあった（□, ×, ・）

になると、ポジティブな出来事が過去には成立していたが、現在は成立していないことになる。これに対して

(13) 昔から照明が落としてあった（□, □, ・）

表 1: 作成したタグ付きコーパスを用いた実験結果

ドメイン（データ数）	過去	現在	未来	時間情報話者態度
飲料水（591）	65.8%	63.6%	88.3%	51.1% 76.5%
自動車（484）	67.1%	58.5%	81.4%	41.9% 77.9%

表 2: 各タグの頻出ラベルの割合

ドメイン（データ数）	過去	現在	未来	時間情報
飲料水（591）	37.7%	45.7%	88.8%	37.6%
自動車（484）	46.1%	53.3%	82.0%	39.3%

では、ポジティブな出来事が過去から現在まで成立していることを表している。

これら 3 つの表現は、時刻を問わなければ、著者にとってポジティブな出来事が（どこかの時点で）事実として成立していたという意味で共通点をもつ。一方、時刻を現在に限定すると、(12) のみ、ポジティブな出来事の成立が否定されている。また、未来の願望、予定について言及する次の例

(14) 照明を落としてくれたらいいのに（・, ×, □）

(15) 明日から照明を落とすと聞いている（・, ×, □）

では、現在は成立していないことも暗に述べている。なお、未来を言及する事態はまだ事実として成立しておらず、予定、願望、仮定、意志、可能などの話者態度により分類することができる。

B 付録 2：時間（アスペクト）

まず、事態を行為と状態に分ける。行為は、動作主が意識してコントロールできる事態のことであり、それ以外のすべての事態を状態と呼ぶ。状態にはそのサブカテゴリとして、状態変化がある。

行為の例：「飲む」「乗る」「買う」

状態の例：「美味しい」「暗い」「飲める」

状態変化の例：「太る」「壊れる」「暗くなる」

本研究では、これらを、瞬間的事態（▲と略記する；以下同様）、反復的事態（■）、継続的事態（□）に再分類する。状態は、本来継続的なものであるため、継続的事態に分類する。行為と状態変化は、瞬間的事態、反復的事態のいずれかに分類する。しかし、行為と状態変化は、その結果残存状態に焦点がある場合は、もはや瞬間的か反復的かを問えるものではなくなり、継続的事態に分類する。

記号表現を用い、アスペクトについて、以下に具体例をあげて説明する。

瞬間的事態（▲）とは、単発的な行為や状態変化のことである。なお、瞬間的事態は、著者の記述時刻を含むことはないと定義し、現在のスロットはいつも（・）としておく。

単発的な行為：

(16) 飲料水 A をさっき飲んだ（▲, ・, ・）

(17) これから照明を落とす（・, ・, ▲）

単発的な状態変化：

(18) 飲料水 A を飲めるようになりたい（・, ・, ▲）

(19) いま照明が暗くなった（▲, ・, ・）

反復的事態（■）とは、習慣的あるいは不特定多数による多発的な行為や状態変化の繰り返しのことである。

習慣的な行為の繰り返し：

(20) 飲料水 A をよく飲んでいる（■, ■, ・）

(21) （いつもこの店では）照明を落としている（■, ■,・）

多発的な行為の繰り返し：

(22) 飲料水 A は最近よく飲まれているようだ（■, ■,・）

(23) （多くの店で）照明を落としていると聞いている（・, ■, ・）

習慣的な状態変化の繰り返し：

(24) （いつもこの店では）照明が暗くなっている（■, ■, ・）

多発的な状態変化の繰り返し：

(25) （多くの店で）照明が暗くなっている（・, ■, ・）

継続的事態（□）とは、状態、または、行為や状態変化の結果残存の状態のことである。

状態：

(26) 飲料水 A は美味しかった（□, ・, ・）

(27) いま飲料水 A を飲んでいる（・, □, ・）

(28) 照明が暗い（・□, ・）

行為の結果残存状態：

(29) 2 週間前から飲料水 A が発売された（□, □, ・）

(30) きょうは照明を落としています（・, □, ・）

状態変化の結果残存状態：

(31) 冷えた飲料水 A を飲みたい（・, □, ・）

(32) 照明が暗くなっているようだ（・, □, ・）

さらに、事態の開始（↑）、事態の終了（↓）、事態の否定（×）によって、より細かい情報を記述できる。

(33) その頃から毎日飲むようになった（↑, ■, ・）

(34) 飲むのを止めている（↓, ×, ・）

(35) 飲料水 A を飲まなかった（×, ・, ・）

(36) 明日から飲み始めよう（・, ×, ↑）

(37) もう飲まないぞ（・, ■, ↓）

(38) まもなく照明を落とし始める（・, ×, ↑）

なお、↑は×■または×□と、↓は■×または□×と等価である。