特集論文　2009 年度 全国大会 近未来チャレンジ

時系列情報の値と変化に関する言語表現コーパスの構築

－動向情報の情報編纂に向けて－

Constructing a Linguistic Expression Corpus on the Values and Behaviors of Time Series Information

– Toward Compilation of Trend Information –

加藤 恒昭　東京大学
Tsuneaki Kato　The University of Tokyo
kato@boz.c.u-tokyo.ac.jp

松下 光範　関西大学
Mitsunori Matsushita　Kansai University
mat@res.kutc.kansai-u.ac.jp

神門 典子　国立情報学研究所
Noriko Kando　National Institute of Informatics
kando@nii.ac.jp

keywords: information compilation, corpus, information extraction, coding

Summary

Statistical information such as gasoline prices and approval ratings of political parties, that has time as one of its domains and numerical values as its range is called time series information. You can obtain the trend of a topic you are interested in by extracting time series information from documents on the topic and compiling it. You, in turn, use the trend for the understanding of and information access from those documents. In order to support mainly to design and implement such a information compilation system, we have constructed a corpus, by extracting linguistic expressions describing the values and the behaviors of time series information from documents, in this case newspaper articles, and coding those contents into a formal framework. This paper explains the coding schema used for this corpus construction, and shows its appropriateness. This paper also includes findings obtained through the corpus construction, such as some classification of syntactic patterns and the vocabulary for describing the behaviors of time series information.

1. は じ め に

ガソリン小売価格や政党支持率等，時間を独立変数に含み，数値を従属変数とするような情報を時系列情報と呼ぶ．本稿では，テキスト中に現れる時系列情報の値やその変化に関する言語表現を分析し，それらによって表されている情報を一定の形式で抽出したコーパスの構築と，そこから得られた知見について報告する．このコーパスの特徴は,「レギュラーガソリンの店頭小売価格（全国平均・１リットル当たり）は、１９９２年末の１２５円をピークに下がり始め、９７年１１月には１００円台を割り込んだ (981114074)」という表現[*1]から，ガソリンの小売価格が 1992 年末に 125 円であったというある時点の値の情報だけでなく，1992 年末から 97 年 11 月までの時間においてそれが下落を続けていた，97 年 11 月には（それまで 100 円以上であった価格が）100 円以下になっていたという変化の情報をあわせて抽出していることにある．変化に関するものを含めて，時系列情報についての広範な言語表現を収集・分析することによって，その要約やそこからの情報抽出をこれまでにない形で行う可能性を提供するものである．

本稿では，まず，近未来チャレンジ「情報編纂の基盤技術」の一環として我々が研究を進めている動向情報の情報編纂について述べ，その中で時系列情報の値やその変化を利用することの有効性を示すことを通じて，本稿で述べる時系列情報の値やその変化に関する言語表現コーパスの重要性を主張する（2 章）．つづいて，コーパス

*1　コーパス作成の対象としては 1998 年から 2001 年までの毎日新聞を用いた．例文の末尾に付与している 9 桁の数字は毎日新聞記事データベースでの記事番号（左 6 桁は掲載日付を表している）を示す．

作成とコーディング（言語表現からの一定の形式での情報抽出）の仕様について説明し，コーディングの実例を示すと共に，作成されたコーパスの規模等を報告する（3 章）．その後，このコーパスの作成を通じて明らかになった日本語新聞記事テキストでの時系列情報の値やその変化に関する言語表現の特徴を論じる．特に，これらの言語表現の統語パタンの分類，そこで用いられる語彙について議論する（4 章 ）．更に，今回のコーディングの問題点，コーディングを困難にしている要因について考察し，今後の改良の方向性を示す（5 章）．最後に全体をまとめる（6 章）．

2. 動機と位置づけ

2·1 情報編纂と動向情報

近未来チャレンジ「情報編纂の基盤技術」では，著者らが情報編纂と名づけた技術の確立を目指している．情報編纂とは，多量雑多な情報を知的に編纂し，その理解を容易にすると共にそれへのアクセスを支援するための技術で，そのコンセプトには２つの力点がある．第一は，様々なモード／メディア／ジャンルを横断して，現実世界の雑多な情報を入力として扱い，その出力においても視覚情報等の非言語情報と言語情報を協調させたマルチメディアプレゼンテーションを活用していくというモード横断性である．この横断性を通じて，既存の言語処理技術と情報可視化技術の建設的な融合を足掛かりとした新しい技術分野の創成を目指している．第二は，情報の理解・概観の支援と情報へのアクセスの支援とを縫い目なく統合することによる情報活用支援の連続性である．要約と情報アクセスインタフェースとして区別されてきたもの，要約の中でも，報知的，指示的と区別されてきたものを統一的に扱っていくような総合的な技術を目標としている [加藤 06].

このような情報編纂のひとつの対象として，動向情報があげられる．ここでいう動向情報とは,「今年に入って原油とガソリンの価格はどう動いているのだろう」「06 年からゲーム機業界はどんな感じになったのか」「去年の台風はひどかったのか」等で示される利用者の関心に対する最初の回答となるもので，利用者の関心に応える情報の全体像を概観させるという役割を持つ．ここに示されたような関心には，その中核となる時系列情報，例えば，原油価格やガソリン小売価格，ゲーム機の総売上や各社のシェアや，もしくは出来事，例えば，台風の発生や上陸を持つので，それらを表現する言語情報，非言語情報（この場合は主に数値情報）に関連した要約と可視化が重要となる．このように，非言語情報と言語情報の処理とそれらを協調させた提示が重要であること，利用者の関心に対する最初の回答ということで，情報の概観とその後の情報アクセスの起点となることから，動向情報の情報編纂（動向情報編纂）は情報編纂の格好の素材である[*2]．動向情報編纂の枠組みを検討しそこに必要な要素技術を確立していくことにより，情報編纂の参照モデルを明らかにすることができる．そして，それらの要素技術はそのまま情報編纂の基盤技術として位置づけられる．もちろん，情報編纂は更に広がりを持った技術である（例えば [西山 08, 井出 08]）が，動向情報編纂がその核のひとつであることは間違いない．

著者らは，このような動向情報の要約と可視化を主たる研究テーマとするワークショップ，MuST（動向情報の要約と可視化に関するワークショップ）をオーガナイザとして運営し [加藤 09]，同時に参加者としても関わることで，動向情報編纂の研究を続けている．MuST は，参加者が協調的かつ競争的に研究を進めていくことを目的に，共通の素材を用いて，緩い意味で共通の課題に取り組むことによる議論と研究の活性化，研究コミュニティの形成，ツールやコーパス類の蓄積を図る非評価型ワークショップとして開始した．そしてその後，多くの参加者が取り組んでいた問題を評価課題として具体化し，評価型ワークショップの側面を合わせ持つようになっている．本稿で論じる時系列情報の値と変化に関する言語表現コーパスは，そこでの検討，そして MuST の他の参加者の研究から明らかになってきた動向情報編纂の枠組みを視野に入れ，その実現の検討で用いられることを目的として作成された研究資源である．2·2 節で，時系列情報を中核とする動向情報編纂の枠組みを提案する．そして 2·3 節で，そのような枠組みを実現するための研究資源として，本稿の主題である時系列情報の値と変化に関する言語表現コーパスを位置づける．

2·2 動向情報編纂と時系列情報

時系列情報である原油価格について述べている以下の文章を考えてみる．

> また、原油価格（ドバイ原油）も、昨年１０月ごろ１バレル＝約２０ドルをつけたのをピークに下落が続き、今年１月下旬には同約１２ドル５０セントまで落ち込んだ。その後、イラク情勢の緊迫化で一時上昇したものの、現在はまた１２ドル前後で低迷、「原油はだぶつき気味」（同庁）だ。(980214080)

ここに含まれる情報は，少なくとも以下の３つに分類することができ，それぞれが動向情報を構成する素材となる．

- 統計量名，時点，値の３つ組で表現できるようないわゆる統計情報（**時点数値情報**）
- 「ピーク」「一時上昇した」「低迷」のような統計量の変化を述べる情報（**変化情報**）

*2 実際の経緯はむしろ逆で，動向情報の要約と可視化を行う技術をより一般化する形で情報編纂のコンセプトが提案された [加藤 06].

- 「イラク情勢の緊迫化で」「原油はだぶつき気味」等の状況変化の理由や状況に対する評価（**状況情報**）

時点数値情報は，白書等の非言語情報から容易に獲得できるものではあるが，それをテキストから抽出するのは，それが新聞記事等で言及されたことに意味があるためである．上例の「昨年１０月ごろのピーク」からもわかるように，現れるのは統計量の動向を要約する節目節目の情報である．また，一定の間隔で調査や発表が行われる一部の統計量（内閣支持率等）は別にして，その時点の統計量が発表されたことにも意味がある*3．従って，このような情報の抽出とその結果の可視化によって，扱っている文書集合が持つ関心に従った動向情報の提示が可能となる．

時点数値情報の抽出はいわゆる情報抽出 [Sarawagi 07] の枠組みによって行われるが，以下に示すような難しさと同時に利用可能な特徴を有しており，それらに着目した研究が進められている．

- 統計量名の正式な名称は長めの複合語（「レギュラーガソリンの全国平均小売価格」）で構成されることが多いが，文書中ではそれがそのまま現れるとは限らず，異なる表現（「平均価格」と「価格の平均」）や省略が頻繁に見られるため，その同定が困難である [Mori 08, Nanba 08].
- 同一記事内で複数の統計量に言及されることが多いため，ある数値がどの統計量の値であるかの同定が困難である．
- 時間表現は文脈を受けて省略されることが多く，広い範囲の解析が必要となる．加えて，発表日等，雑音となる日付も混在し，同定が難しい [Nanba 08].
- 一定期間にわたってある統計量の一連の値を抽出するということから，とりうる値の範囲や値の連続性等を仮定するヒューリスティクスを用いて，精度の向上を図ることができる [Yoshida 08, Inoue 08].
- 以前の値との差等，相対的な値の表現も含まれており，それらに着目することで，再現率の向上を図ることができる [Uenishi 08].

変化情報は，以下のような興味深い特徴を持つ．

- 「つけた」「下落が続き」「落ち込んだ」という用言によって表現されることが多く，言語表現の分節の単位に対応している．
- 「今年１月下旬には同約１２ドル５０セントまで落ち込んだ」に見られるように時点数値情報は変化の終点や起点であり，変化情報が中心で時点数値情報はそのパラメータと考えられる．
- 時点数値情報に較べ，「ピーク」「低迷」等，ある時間幅や特徴的な時点の変化に関しての要約になっている*4．

*3　ただし，これはその時点での判断で，後に振り返った時にその重要性が維持されている保証はない．またその情報は記事掲載日から復元できないこともない．

*4　時点数値情報も「１０ドル台」「約２０ドル」のような概数は要約的な性格を持つ．

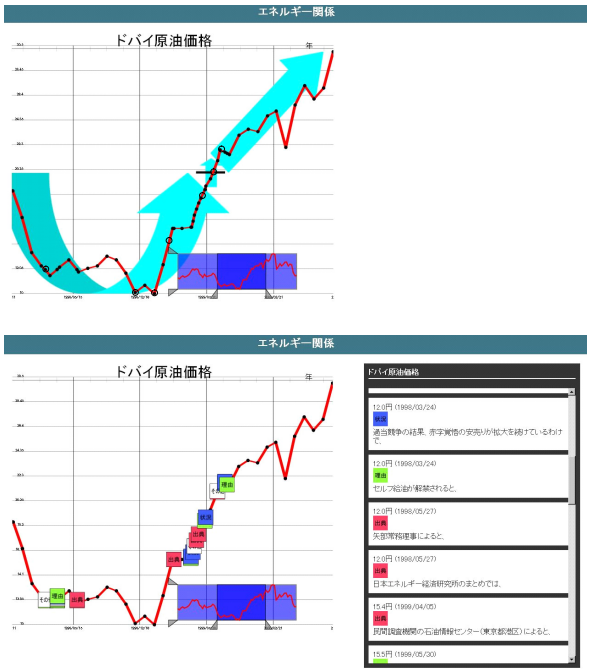


図 1　動向情報の可視化

- 「ピーク」「上昇した」のように客観的で類型化できるものに加えて，「落ち込んだ」「低迷」のように含意や付随的な意味を持つ用言が用いられることがあり，それによって主観的な評価が上乗せされる．

状況情報は変化情報によって示される状況に関連する言語的な記述となっている．これは主観情報抽出 [Pang 08] と同様の技術によって抽出が可能と考えられる．

これら３種類の情報を区別し，それらをテキストから抽出し，白書等から得られる統計量の数値データ（統計データ）と合わせて，時間軸に重ねて配置することで組織化し，この統計量に関する動向を可視化して概観を許すと共に，この対話的なプレゼンテーションを詳細な情報，例えば，記事全文へアクセスする際の入り口として用いるという枠組みを提案する．

この枠組みでの表示例を示す*5．図 1 上では．時点数値情報と変化情報が重ねられている．変化情報はその種類を表現する矢印アイコンによって示されている．ここには示されていないが，例えば，上向き矢印に「急騰」等と注釈づけることにより，変化情報に含まれる主観的な評価を上乗せすることも可能である [加藤 07].

図 1 下では，状況情報が，理由，状況，出典等の種類で色分けされたアイコンでグラフと関係づけられている [松下 09]．ここでは，グラフ中のアイコンと右側に並んだスニペット（状況情報として抽出された部分）が関連づけられており，グラフ中で着目した状況情報のアイコンを指定することでその内容を知ることができ，関心を

*5　上例の文章で表現されたものとは異なる時期が表示されている

持ったスニペットを指定することで，それがグラフ中のどの時点についてのものかを明らかにできる．更にそれらと，それを抽出した文，それが含まれる記事の見出し，記事全文とを関係づけ，指定された場合にポップアップさせること等ができる．図 1 上においても，グラフ中の矢印アイコンや時点数値情報の点に同様の関連づけが行える．記事の掲載時点ではなく，記事の内容が言及している時点に関連づけられる点が重要である．

図 1 でグラフの右下にあるのは，パンやズームのためのパネルで，これにより注目点を移動させたり，細部に注目したりすることができる．これらの画面を相互に行き来することも可能である．このような視覚的な関心の移動と，上で説明したアイコンの指定による文書情報への関心の移動を通じて，Shneiderman の情報可視化のマントラ “Overview first, zoom and filter, then details on demand”[Shneiderman 09] が実現されることになる．

2·3 動向情報編纂のための研究資源

2·2 節で述べた動向情報編纂を実現するためには，利用者の関心に基づいて収集されたテキストから，時点数値情報，変化情報，状況情報を抽出することが重要となる．これらの情報が動向情報の要約と可視化の素材として利用される．この抽出の検討のために構築されたのが，***MuST* データセット**と本稿で報告する**時系列情報の値と変化に関する言語表現コーパス**[*6] である．

近年，テキスト中の特定の情報を註釈づける試みがなされている．これらの註釈づけは言語情報の分析を支援すると同時に，そのような情報を抽出するシステムを設計する場合に参照されたり機械学習に基づく手法での学習データとして用いられる．時間情報の註釈のための TimeML[Pustejovsky 05]，場所情報の註釈のための SpatialML[Mani 08]，事象の因果関係の註釈づけの仕様 [Bethard 08] 等が提案されている．また，評価ワークショップの正解データとしてテキストに註釈づけが行われる．例えば，内容抽出に関する評価ワークショップ ACE では，人物や組織等の固有表現とその間の関係を註釈している [NIST 08]．意見抽出を対象とした NTCIR MOAT では，ある文が意見かどうかの判断や，意見である場合にそのテーマや持ち主等が註釈されたコーパスを提供している [Seki 08]．このような正解データも同種の課題に関するシステム設計で参照されるものとなる．註釈づけというより言語資源としての側面が強いが，PropBank は動詞の項構造に関する定義とそれに基づいた註釈を行ったコーパスを提供し，語彙情報を提供すると共に意味解析（格役割付与）設計のための資源となっている [Palmer 05]．

MuST データセットと時系列情報の値と変化に関する言語表現コーパスも，MuST 全般が関わる動向情報に関する様々な処理の設計で参照されたり，機械学習に基づく手法での学習データとして使われることを目的としている．前述のように MuST はその初期において評価型ワークショップではなかったため，このような資源を共通の研究素材として構築することで，その活性化を図っていた．また，時系列情報の値と変化に関する言語表現コーパスについては，2·2 節で述べた動向情報編纂の検討を含めて，時系列情報の値と変化に関する言語表現一般がどのような構成要素を持ち，それらがどう関連しあって全体としての意味を構成しているかについての考察に用いることもその目的であった．

MuST データセットは，1998 年から 2001 年の毎日新聞記事を知識源と考え，そこから「ガソリン価格」「自動車生産」「通信機器」等，31 のトピックについて，関連する時系列情報（統計量）や出来事（各トピックについて 3 つ前後，計 90 種類）に関する情報を含んだ新聞記事 702 記事を収集し，注釈づけを行ったものである．このうちで時系列情報を中心とするものは，27 トピック，606 記事となっている．

トピック毎に収集した新聞記事集合は情報収集によって得られた言語情報に相当する．各記事への注釈（以下，***MuST* 注釈**と呼ぶ）は，要約における重要文の抽出と，情報抽出における固有表現抽出と時間表現解析を含む参照表現解析の結果に相当する．MuST 注釈の例を図 2 に示す．時系列情報（統計量）や出来事に言及している文が unit 要素として抽出され，そこに含まれる統計量名（name 要素），時点（date 要素），時系列情報の値（val 要素）等が注釈づけられている．統計量名は「レギュラーガソリンの全国平均店頭価格」が「ガソリン価格は全国平均でレギュラー１リットル当たり９２円」のようにテキスト中に分散して表現されることが多いため，その位置づけを明示する part 属性を備えている．相対的な時間表現にはその絶対値が注釈される（abs 属性として yyyymmdd で表現）．これらは時点数値情報の 3 つ組を構成する要素となる．加えて，時系列情報の値そのものではない，値の差分，割合，順位は，rel 要素として val 要素とは別に注釈される．この区別は，時点数値情報の抽出において雑音となる情報を分離する指針を設計する参考となることに加えて，変化の表現にも関係するものである．rel 要素は type 属性によって，差分（diff），割合（prop），順位（ord）等へと分類されている．時間の幅も時点である date 要素とは区別され，dur 要素とされる．時系列情報の値や変化に直接言及していない部分は，del 要素とされるが，これが状況情報として抽出すべき部分に対応しており，type 属性によって，出典，原因（rsn），評価，状況（sit）等へと分類されている．

特に，統計量名への註釈，rel 要素の val 要素からの分離とその分類，del 要素の存在とその分類の 3 点によって，MuST 注釈は単なる固有表現の注釈と区別され，動向情報編纂の枠組みの実現を検討するのに適したものとなっている．MuST データセットの全容や具体的な仕様

*6 他の文献（[加藤 09] 等）ではより簡単に「変化表現コーパス」と呼んでいる．

<unit stat="ドバイ原油価格">また、<name part="head">原油価格（ドバイ原油）</name>も、<date gra="月" abs="199710">昨年１０月ごろ</date><rft id="980214080_1"><name part="foot">１バレル＝</name></rft><val>約２０ドル</val>をつけたのを<rel type="ord">ピーク</rel>に下落が続き、<date gra="旬" abs="19980121">今年１月下旬</date>には<pro ref="１バレル＝" id="980214080_1" >同</pro><val>約１２ドル５０セント</val> まで落ち込んだ</unit>。<unit stat="ドバイ原油価格"><ins type="name"><name>ドバイ原油価格</name>は</ins><date gra="旬" abs="19980121">その</date>後、<del type="rsn">イラク情勢の緊迫化で</del>一時上昇したものの、<date gra="不明" abs="19980214">現在</date>はまた<val>１２ドル前後</val>で低迷、<del type="sit">「原油はだぶつき気味」（同庁）だ</del></unit>。

図 2 MuST 注釈の例

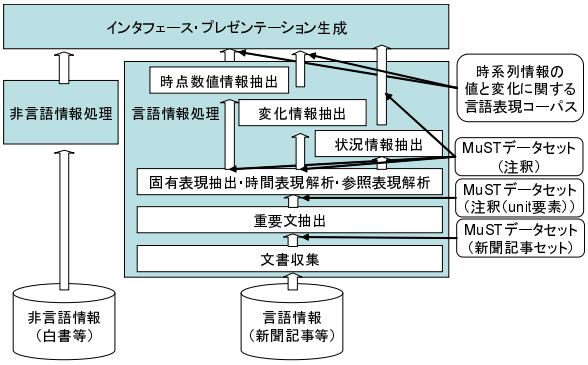


図 3 動向情報編纂の枠組と研究資源

は MuST HP[*7] と文献 [Kato 07] に詳しい.

時系列情報の値と変化に関する言語表現コーパスは，MuST コーパスにおいて unit 要素として抽出された部分を対象に，時点数値情報と変化情報の抽出結果となるであろうものと，それが抽出された言語表現を整理したものである．このコーパスの分析を通じて，どのような変化情報がありうるのか，それをどのような形式的表現で捉えられるのか，それらがどのような言語表現に対応しているのかを明らかにできる．

これら２つの研究資源と 2·2 節で述べた枠組みとの関係を図 3 に示す．多くの構成要素の出力に相当する情報をこれらの資源が提供しており，多くのサブシステムの設計や評価，正解データの作成で利用できることがわかる．特定のサブシステムに研究を集中するために，それ以前の処理結果としてこれらを利用することもできる．例えば，MuST では，動向情報の要約と可視化に関する要素技術について客観的定量的評価を行うものとして，評価課題を設計したが，時点数値情報を文書から抽出することに相当する T2N 課題の正解データはこのコーパスから作成された [Kato 08].

3. 時系列情報に関する言語表現コーパス

本章では時系列情報の値と変化に関する言語表現コーパスの構築における値や変化の捉え方と，それに基づいたコーディングの基準を述べる．幾つかの細かい規定について述べた後，構築したコーパスからコーディングの実例を示すと共に，その規模等の情報を述べる．

3·1 変化の捉え方

2·2 節で議論したように，ひとつの変化情報は，多くの場合，テキストの節 (clause) に対応し，時点と値の対である時点数値情報は変化の終点や起点として表現される．その意味で，変化の情報が中心的であり，そのパラメータとして値があると捉えることが自然である．

節の中心となる用言によって，表現される変化やその項（パラメータ）となる値との関係が規定される．例えば「$10 であった」という表現は，言及されている時点（やや不正確かもしれないが，参照時点と呼ぶこととする）の統計量の値が$10 に等しいという情報であり，値の変化には言及していない．一方,「$10 になった」はその時点の値が$10 に等しいことに加えて，その直前に何からの変化があったことを表すし，逆に「$10 とかわらない」はそこに変化がなかったことを表している．その時点の値との関係は同値関係だけでなく,「$10 を上回っている」の場合は参照時点の値が$10 より高いことを示している．

項となる値との関係は言及されている時点（参照時点）のものに限らない.「$10 を上回った」の場合，上昇方向の変化があり，値が$10 に等しい時点が参照時点より過去にあったことを意味している．そして，含意として参照時点の値は$10 を上回っていることが示される.「$10 に迫っている」の場合は，予測される未来の時点との同値関係が表され，それに向けての方向が明示されない変化があることを示してる．

このような捉え方に基づいて，値と変化についての記述を，参照時点，項となる値，参照時点との時間関係，項となる値との大小関係，変化の方向（変化の有無を含む）で，表現する．典型的な表現を図 4 に示す．ここで丸印は変化がないのではなく，変化に関する言及がないことを示す．これらは必ずしも直交するものではない．項となる値との関係が同値関係でないのは，参照時点について述べられている場合だけであるし，その場合，変化へ

*7 http://must.c.u-tokyo.ac.jp

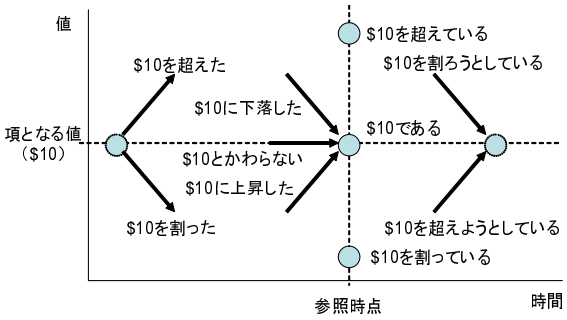


図 4 値と変化の状況とその表現

の言及はできない*8．参照時点以外の言及の場合は変化への言及がないことは許されない（図では破線の丸印で表現している）．

複合的な用言を含んだ節，もしくは統語的には単純であるが複雑な意味を持つ用言によって，変化の始まりや継続等，変化の変化が表現されるとする．「上昇し始めた」は上昇方向の変化が参照時点で始まったことを示すし，「上昇に転じた」は参照時点以前に下降方向の変化があったことを示す．用言ではないが，「ピークとして」という表現も，同じように変化の変化に関する言及であると考える．また，「上昇」に対する「急騰」，「とかわらない」に対する「低迷」のような用言については，類型化できる情報に加えて，その変化や値についての含意，付随的な意味，執筆者による主観的な評価の情報が上乗せされていると考える．

3·2 コーディング

コーディングのより具体的な仕様を以下に示す．ここで述べる仕様はコーパス作成開始前に固定されたものではなく，作成中に現れた事例について検討し改良していった結果である．その一方で，すべての問題が解決されているわけでもない．その意味では，開発途上の仕様であるが，一定の基準に達しており，作成された言語資源の実際の仕様となっているものである．解決されていない問題については主に 5 章で述べる．

コーディングは MuST コーパスで unit 要素とされた表現を対象としている．一つの unit 要素から 1 つ以上の情報（情報単位）が抽出され，コーディングされる．抽出された情報単位毎に，まず，抽出元の情報として以下の要素が記述される．

トピック その情報が属するトピック（トピックは記事単位に設定されている）．

テキスト 抽出対象となった（その情報が記述されている）記事の部分（unit 要素）の平文．

注釈付テキスト 上記テキストの MuST 注釈を施した版．

通番 ファイル内通し番号．

記事番号 抽出対象となった記事の毎日新聞記事データベースでの記事番号．

unit 番号 抽出対象となった unit 要素が当該記事中で何番めのものか．

どの統計量に関するどの時点の情報かについて，以下の要素が記述される．

統計量（name） どの統計量に関する情報か．統計量名はテキスト中の表現ではなく，MuST 注釈で name 属性の値として用いられている正規化された名称を用いる．

パラメータ（par） メーカー別シェアのように時間以外のパラメータを持つ統計量の場合のパラメータ（シェアの場合はメーカー名）．この情報は MuST 注釈で par 要素として注釈されているので，<par>タグ付の形で記述する．

予測（pros） 見込みや予測として述べられる情報であるか．MuST 注釈において，unit 要素毎に同様の情報が付与されているが，ここでは情報単位の付与となり，より細かい情報となる．例えば「今年は昨年の$10 を越えるだろう」は全体として予測であるが，そこから抽出される「昨年は$10 であった」という情報は予測とはしない．

時点（time） その情報で言及している参照時点．一般に date 要素となるので，<date>タグ付の形で記述する．従って，相対的な時間表現であっても，その絶対値が abs 属性の値として付与されている．

値と変化に関する情報は，値の状況に関する情報（type0），変化に関する情報（type1），変化の変化に関する情報（type2）とそれぞれのパラメータで表現する*9．

type0 の情報が存在することは値に関する言及があることを示し，それが，参照時点との関係，項となっている値との関係，項となっている値という 3 つの要素で表現される．

参照時点との関係 (t-rel) “during”, “before”, “after” のいずれか．それぞれ，値との関係が参照時点のものであること，参照時点より過去，あるいは未来のものであることを示す．

項となっている値との関係 (v-rel) “equal”, “larger”, “smaller” のいずれか．それぞれ，その値との同値関係，それを上回っているという関係，それを下回っているという関係を示す．参照時点との関係が “before”, “after” の場合は，“equal” のみが許される．なお作業負担軽減のため，“equal” の場合は入力を省略して空欄のままとすることを許している．

項となる値（val） 一般には val 要素となる．その場合，タグを含めないで記述する．

*8 「$10 以上となった」は，項となる値が「$10 以上」であると捉え，値との同値関係とする．「$10 を超えた」とほぼ同じ内容が異なる情報にコーディングされることは好ましくはないが，項については「$10 台」「約$10」等様々なものがあることと，コーディングの判断は原則として用言部分に基づいて行いたいということから，このような方針をとった．

*9 type1，type2 等の名前は，変化を 1 次微分量，変化の変化を 2 次微分量と考えたことに由来する．

type1 は変化に関する情報で，変化の種別と，変化があってその程度が具体的に表現されている場合に，その程度が，いつからもしくはどれだけの期間で，どれだけ（差や割合）変化したかという要素で表現される．

種別（c） “none”, “some”, “upward”, “downward” のいずれか．それぞれ，変化がないこと，方向が明らかでない変化，上昇方向の変化，下降（下落）方向の変化があることを示す．

比較時点/変化期間（c-t/dur） いつの時点（比較時点）と較べての，もしくはどれだけの期間（変化期間）での変化か．一般に時点の場合は date 要素，期間の場合は dur 要素となるので，それぞれのタグを含めた形で記述することで区別する．

差分（diff） 変化によって生じた差もしくは変化の割合．一般に rel 要素となり，<rel>タグを含めた形で記述する．差であるか割合であるか等は rel 要素の type 属性で区別される．

type2 は変化の変化に関する情報で，その種別と，その間隔や期間を表す要素で表現される．

種別（cc） “continue”, “start”, “reverse”, “speed_up”, “speed_down”, “convex” のいずれか．それぞれが示すものは以下の通り．

“continue” 同じ変化が連続している．

“start” 現在の変化が直前のものと異なる（これまではなかった変化が，もしくは，これまでとは別の変化が始まった）．

“reverse” 現在の変化が直前の変化と逆方向のものである．

“speed_up” 変化の度合い（増加や減少の幅）が大きくなった．

“speed_down” 変化の度合いが小さく（鈍く）なった．

“convex” 現在とは逆方向の変化が以前にあり（値の推移としては凸もしくは凹で），過去に同じ値となった時点があった．

開始時点/継続期間（s-t/dur） 同じ変化が継続している（“continue”）であれば，それがいつの時点（開始時点）からもしくはどのくらいの期間（継続期間），継続しているのか．以前と異なる変化が生じた（“start”, “reverse” 等）のであれば，それが以前に生じたのはいつの時点か，もしくはどのくらいの期間ぶりに生じたのか．過去に同じ値であった（“convex”）のであれば，その過去はいつの時点あるいはどのくらいの期間ぶりなのか．一般に時点の場合は date 要素，期間の場合は dur 要素となるので，それぞれのタグを含めた形で記述することで区別する．

言語表現との関係や，その他の非定型な情報を以下の要素として記述する．

主たる表現 その情報の表現の中心となる用言.「ピーク」等の名詞も許す．Copula（「だ」等），体言止め，名詞的に表現されるために適当なものがない場合は，その旨の表示として “-” を用いる．

付随情報有無 執筆者による主観的な評価の情報が上乗せされているか.「低迷」「急騰」等の用言の利用に加えて,「わずかに上昇した」における「わずかに」等の副詞の存在もそのような情報と考える．

備考 典型的な情報の形式でない場合や作業中に気づいた問題がある場合に利用する．

§1 情報抽出の単位

ひとつの情報単位は，前出の要素の並びであり，言及のない場合，対応する要素は空欄となる．抽出は節を単位とし，ある節の中で記述されている情報をできるだけまとめて抽出する．例えば,「先月より$1 高い$10 になった」という文に含まれる「$10 になった」「先月から$1 高くなった」という type0, type1 の情報をまとめてひとつの情報単位とする．もちろん，同じ統計量，同じ参照時点の情報しかまとめることはできないので，それが異なると別の情報になる．例えば,「先月の$9 から$1 高い$10 になった」では,「先月$9 であった」という情報は参照時点が異なるため，この文に含まれる他の情報とは別の情報単位として抽出される．また，ある情報に同じ type の要素はひとつしか含まれないので,「$11 と$10 を超えた」「98 年 4 月からの 3 カ月で$2 上昇した」のような場合,「$11 になった」と「$10 を超えた」,「98 年 4 月から$2 上昇した」と「3 カ月で$2 上昇した」は，それぞれ異なる情報単位として記述される．この点については，5·1 節で再考する．

§2 明示されていない値の扱い

コーディングすべき情報がテキスト中で明示されておらず，文脈や常識から理解される場合がある．この場合，以下の基準に従って適切な値を入力する，もしくは空欄のまま残すようにしている．

- 記事の抽出部分に参照時点が明示されていない場合，時点（time）は，以下のように処理する．その情報が予測（pros=1）でなく，記事先頭で述べられるような記事内基準時点というべき時点が参照時点と思われる場合，もしくはそのような基準時点が示されていない場合で記事発行日が参照時点と判断される場合には，空欄のままとする．そうではなく，明示されていない過去の情報である（参照時点が不定の過去である）と推測される場合は，“<date gra="" abs="0">過去</date>” と記入する． 情報が予測である場合は，指定されていない未来と推測される場合に空欄のままとする．
- type1 の情報の比較時点/変化期間（type1:c-t/dur）と差分（type1:diff）は，両方に値がないか両方に値があるかのどちらかとする．つまり，どちらか一方の情報しか明示されていない場合は，以下に基づいて他方の値を判断し入力するようにする．比較時点/変化期間が示されていない場合（例えば単に「$1 上昇した」とある場合）は，予測でない場合，参照時点

として示された時点の粒度（date 要素に gra 属性として記述されている）で，参照時点のひとつ前の時点，つまり「今月」であれば先月,「今年」であれば去年と推測できるのであれば，そのことを “-” で示す．予測の場合は記事内基準時点と判断されるときに同様に “-” で示す．差分が示されていない場合（例えば単に「先月より上昇した」とある場合），正の値であること以上の情報がなければ，そのことを “-” で示す．単に正の値ではなく，文脈や常識から特定の値であることが推論される場合，その推論される値を注釈付の形式で記入し，属性として inf=”1”を加える．

§3 変化とその変化の扱い

type2 の情報は type1 で示された変化すべて（type1:c の部分だけでなく type1:diff 等を含めた全体）の変化や継続を表現する．従って，

type1:c=”upward”, type1:c-t/dur=”1 年”, type1:diff=”$1”, type2:cc=”start”, type2:s-t/dur=”3 年”

が示している[*10]のは,「３年ぶりに１年で$１の上昇があった」「１年$1 の上昇は３年ぶり」であって,「３年ぶりに上昇に転じ，今年は$1 上昇した」ではない．こちらは,「３年ぶりに上昇に転じた」「今年は$1 上昇した」が異なる情報単位となり，それぞれ以下のように表現される．

type1:c=”upward”, type2:cc=”start”, type2:s-t/dur=”3 年”
type1:c=”upward”, type1:c-t/dur=”1 年”, type1:diff=”$1”

従って，以下の例は一つの情報単位として抽出される．

携帯電話の加入台数はここ３年は毎年１０００万台以上増え続けている (990422027)
レギュラーガソリンの全国平均店頭価格のひと月で２円の上昇は９２年９月以来 (990624031)

一方，以下の例では，情報単位をふたつに分ける必要がある．

ＰＨＳは同［前月比］８万５０００台減の６５６万８０００台で、９カ月連続の減少となった (980708071)
「支持する」は８月の前回電話調査を８ポイント下回る３２％で、…、初めて２カ月連続の減少となった (990907005)

最初の例では,「８万５０００台減少した」ことと「９カ月連続で減少が続いた」ことが異なる情報単位となる．これらの例からわかるように，言語表現における分節と情報単位の分節とが適切に対応づけられた仕様となっている．

3·3 実施例と実施規模

時系列情報の値と変化に関する言語表現コーパスのコーディングは，9 トピック，27 統計量に関する 219 記事，893unit 要素に対して実施された．抽出された情報単位は 3·2·1 節に示した方針で数えて，約 1,800 である[*11]．トピックの一覧と，それぞれの記事数，統計量の一部を表 1 に示す．幾つかの典型的な表現とそのコーディングを図 5 に示す．

4. 値と変化に関する言語表現

本章では，時系列情報の値と変化に関する言語表現コーパスの作成を通じて明らかになった日本語新聞記事テキストでの時系列情報の値やその変化に関する言語表現の特徴を論じる．まず，時系列情報の値やその変化に関して，どのような言語表現（統語的なパタン）が用いられ，それによってどのような情報が表現されているかを眺める．その後，そこで用いられる語彙について議論する．

具体的な数字はそれぞれの節で述べるが，言語表現についてはパタン化が比較的容易な定型的なものが多く，語彙も分野に依存しないものが多い．評価，判断を示す付随表現についても大きく分類できることがわかった．これらは，時系列情報の値や変化に関する情報を分野依存性が少ない方式を用いて高い精度で抽出できることを示唆している．また，そこで表現されている情報についても，情報抽出や可視化で重要となる局所的で特定的なものについては，ここで提案している枠組みにほぼ収まることが明らかとなった．このことは，この枠組みがそのような情報抽出のテンプレートとして充分機能することを示している．

4·1 値と変化に関する表現の分類

2·1 節で述べたように，変化に関する情報の単位は，言語表現の分節の単位に対応していることが期待され，値の状況や変化を節の主な用言で表現するような，図 5 にあげたものを典型例とする表現が多いと考えられる．実際のコーパスを検証しても，中心となるのはこのような文型であることがわかる（unit 要素単位で 62%）．これに加え，このような節が「と報告された」「ということが明らかになった」や「と見られる」「という予測だ」のような報告や予想であることを示す表現に埋め込まれたり，その節に「と，〜であった」のような評価の節が繋がることも多い（21%）．節の用言によって述べられる情報に加えて，その一部に，例えば比較対の形で「〜年の〜に較べて」等の異なる参照時点の情報が付随することもあり（全 unit 要素の 9%），そこからも情報を抽出することができる．

これらの言語表現（83%）についてはパタン化も比較的容易で，その内容もコーディング可能であるので，情報抽出の対象として有望である．一方で以下に述べるような，その整理が難しい言語表現も散見される．

*10 本例を含めて，以後，要素の値に付与されているはずの MuST 注釈は問題のない範囲で省略する．

*11 5 章で述べる分析はこのうち 5 トピック 99 記事を対象にした．示されている割合等はその範囲での数字である．

表 1 コーパス作成対象トピックおよび統計量

トピック	記事数	統計量
ガソリン	32	レギュラーガソリンの全国平均店頭価格，ドバイ原油価格
パソコン	20	国内出荷台数，国内出荷額，メーカー毎のシェア
通信機器	45	固定電話，ＰＨＳ，携帯電話の加入台数
内閣支持率	33	内閣支持率，政党支持率
自動車生産	16	主要メーカーの国内，海外，総生産台数，主要メーカーの輸出台数
鉱工業	17	鉱工業生産指数，鉱工業出荷指数，鉱工業在庫指数
i モード	20	i モード，EZweb，J-スカイ加入者数
デジカメ	8	デジカメ，フィルムカメラの出荷台数，同出荷金額
センター試験	18	受験者数，倍率，参加私立大学数

(1) ＰＨＳ（…）は３カ月連続で減少し、昨年末で６９９万台と７００万台を割り込んだ（980110065）
(1)-1 time=”記事基準時点”, type1:c=”downward”, type2:cc=”continue”, type2:s-t/dur=”３カ月”
(1)-2 time=”昨年末”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”６９９万台”
(1)-3 time=”昨年末”, type0:t-rel=”before”, type0:v-rel=”equal”, type0:val=”７００万台”

(2) 簡易型携帯電話（…）は同月末時点で６８２万台と頭打ちになってはいるものの（980512242）
(2)-1 time=”同月末”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”６８２万台”, type1:c=”none”, type2:cc=”start”

(3) 原油価格（ドバイ原油）も、昨年１０月ごろ１バレル＝約２０ドルをつけたのをピークに下落が続き、今年１月下旬には同約１２ドル５０セントまで落ち込んだ (980214080)
(3)-1 time=”昨年１０月ごろ”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”約２０ドル”, type1:c=”downward”, type2:cc=”reverse”
(3)-2 time=”過去”, type1:c=”downward”, type2:cc=”continue”, type2:s-t/dur=”昨年１０月ごろ”
(3)-3 time=”今年１月下旬”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”約１２ドル５０セント”, type1:c=”downward”

(4) １９９８年度第２四半期（…）の…国内出荷台数は１７９万７０００台で、対前年同期比９％増と１年３カ月（５四半期）ぶりに前年実績を上回った (981103066)
(4)-1 time=”１９９８年度第２四半期”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”１７９万７０００台”
(4)-2 time=”１９９８年度第２四半期”, type1:c=”upward”, type1:c-t/dur=”前年同期”, type1:diff=”９％”
(4)-3 time=”１９９８年度第２四半期”, type1:c=”upward”, type1:c-t/dur=”前年同期”, type1:diff=”-”, type2:cc=”start”, type2:s-t/dur=”１年３カ月（５四半期）”

(5) 小渕内閣の内閣支持率は昨年１２月の支持率１６％を底に上昇に転じた (990518006)
(5)-1 time=”昨年１２月”, type0:t-rel=”during”, type0:v-rel=”equal”, type0:val=”１６％”, type1:c=”upward”, type2:cc=”reverse”

図 5 時系列情報の値と変化に関する表現とそのコーディング

「〜の変化（伸び）は〜」のように統計量の変化を主題にしている文がある（2%）．これらの文で表現される情報は，ほとんど上記の表現に置き換えることができるが，「固定電話の加入者の伸びは鈍化した (980421070)」のような上昇や下降の速度が大きくなった，小さくなったという情報はこの形でしか表現することができない．また，このような文の場合，「携帯電話の昨年１年間の増加数が１０００万台を超え (980110065)」のように type2:diff の値となるべきもの（「１０００万台以上」）が用言を含むことがあり，用いられている表現の抽出という形では情報が記述できないことがある．また，「〜したのは〜ぶり」「〜したのは、〜のためと思われる」のようにある具体的な変化を主題として，その属性を述べたり，それを論評する場合がある（4%）．

その他は整理が更に難しく，一定期間の変化の形状を俯瞰的に述べる，例えば「小渕内閣の内閣支持率は…「反転上昇型」(990130010)」「ドバイ原油価格における過去の急激な上昇は、第１次石油ショック（７３〜７４年）時の３ドル→１２ドル、湾岸危機（９０〜９１年）の１７ドル→３７ドルなど (990817066)」のような場合や，様々な内容の文の中で値の状況や変化に触れられる場合等がある（9%）．なお，新聞記事中に表形式で情報が含まれる場合もある（3%）．ただし，これは政党支持率等，特定の統計量に限られる．

4·2 表現されている情報

現在の枠組みで表現できる変化に関する情報は図 5 に例を示した通りで，type0, type1, type2 の情報が様々に組み合わされた表現があることがわかる．71%の unit 要素単位について提案したコーディングで情報が記述できて

いる*12．この数字はパタン化できる言語表現が80%を超えることとあわせて，この枠組みでの情報抽出が有望であることを示している．

それ以外のものは非定型的な記述となるような，あるいは表現できない情報を含んでいる．前者には，値以外との比較や順位の表現が含まれ，これらを含めていく枠組みを検討する必要がある．後者は，大局的な変化や個別的特定的でない変化の記述であり，いわゆる情報抽出の枠組みよりも要約の枠組みで利用されるような，2·2節の分類でいえば状況情報に近いものであると考えられる．以下，これらについて述べる．

値以外のもの，つまり過去の時点や，他の事物の統計計量の値，異なる統計量の値との比較の表現は難しい．なお，時点との比較は変化の記述と捉えることもできるので，大抵の場合大きな問題とならないが．それでも幾つかの場合に不都合を生じることがある（この点は5·3節で論じる）．これらの比較については，val要素ではない非定型的な値が示されている考えて，type0:valに注釈を含んだそれらを記述することで対応している．例えば「自民党より支持率が低い」であれば，type0:val="<par>自民党</par>"と表現する．ただ，このような比較の表現では,「民主党の政党支持率も自民党より１０ポイント以上も少なく (990201038)」等，その差への言及があることがあり，これは現在の枠組みでは扱えない．そもそも，このような比較の表現が，現状のコーディングで扱っているものと同じレベルの情報なのか，それとも個々の状況の記述の上に被せられる上位の情報であるのかを検討する必要がある．数は少ないが,「最安値から８０%以上上昇し (990717070)」のようにtype1:c-t/durの要素が非定型となる場合もある．

企業シェアや支持率のような統計量では，順位について頻繁に言及される．これについても非定型的な値と考えることができるが，こちらの場合，企業シェアの順位という，1位，2位という値をとる統計量と考えることもできる．また,「携帯電話の加入者２０００年には３人に１台 (980422135)」のような統計量の値そのものではないものを値かのように扱う記述も稀に現れるが，これも同様に捉えられる．

「昨年８月末に初めて７００万台を超え (980110065)」の「初めて」というような大局的な位置づけが記述できない．現在のコーディングでは原則として局所的な変化を記述する枠組みとなっている．例えば,「毎月史上最低価格を更新し（990624031)」の場合，価格が毎月下落し続けたということは記述できるが，その価格が「史上最低」であることは記述できない.「１９９７年２月以来の高値を更新した (990824079)」という場合の表現は更に難しい．加えて，前述した一定期間の変化の形状を俯瞰的に述べるものや,「各社とも前年度に比べ増加が目立つ (980422080)」というような複数の変化を集約したようなものもコーディングの対象にできない．

個別的特定的でない情報，例えば「ポケベルは一時、１０００万台を超えたが (980620233)」「ピーク時は１０００万台を超えたものの (980414084)」のように時点が明示されておらず，ある状況が生じたことだけを述べ，いつ生じたかを述べていないものは記述できない.「加入台数が…、１０００万台を超えるまでには４年、２０００万台突破は１年だった (980422135)」のように推論を通じてある程度まで時期が読み取れるものについても，その情報を記述することは困難である．

4·3 変化の表現に用いられる語彙

§1 主たる表現となる語彙

変化を表現する文において，特定の統計量に依存するような（分野依存の）用言が用いられることは少ない．もちろん,「高騰」等の表現は，金銭を値とするような統計量だけで使われ,「支持率が高騰する」と表現されることはないが，それでも意味分類としては「上がる」に含まれる．73%がそのような値や変化を記述する一般的な用言 (「だ」「となった」等を含む) である．そして，その周辺に「落とす」「伸ばす」「後退する」「つける（９０円をつける)」「(まで）行く」「確保する」「持ち直す」「回復する」「戻す」「突破する」「衰える」「頭打ちになる」「維持する」「膨らむ」「とどまる」（これについては，後述）等，一般には数値の記述というより程度や調子の変化に用いられる用言等が比喩的慣用的に用いられるような表現が広がる（21%）．表現としては「好調だった」「躍進する」「快進撃を続ける」のような評価に相当するものが変化の表現を埋め込む形で続く場合もある．

少ない例として，統計量に依存するものとして,「(加入者を）獲得する」「奪われる」「(シェアを）浸食する」のように，統計量の性質・意味付けを反映した用言が使われる場合がある．更に,「ソニーが前期比２・２倍の２２万台を出荷し (991104064)」「民主は、…５月の調査より最高２６ポイントも支持を伸ばした (980804005)」のように，用言の部分を含めて統計量（この例では出荷台数と政党支持率）を表す場合もある．この最後の場合以外では，統計量名の記述は主題の部分に含まれていることになる．つまり「ガソリンは$10で売られている」という表現ではなく,「ガソリンの価格は$10である」という表現となっているわけで，この事実は統計量名抽出や変化情報抽出の指針となる．いずれにせよ，これら2つの場合はここに挙げた例でほぼ網羅できる程，稀（3%）である．

このように分野に依存した語彙が少ないことも時系列情報の値や変化に関する情報抽出を有利にしている．

§2 付随表現

type0, type1, type2の情報として類型化できる値や変化の記述に，それへの評価，価値判断が上乗せされて述

*12 同じ文に問題なく表現できる情報要素も含まれているので，情報単位では71%より多くの情報が抽出できている．

べられている場合がある．21%の情報単位がそのような情報を含んでいる．その方法は３種類に分類できる．まず，用言そのもの（用言性の名詞を含む）にそのような判断が含まれている場合（50%），第２に副詞あるいは用言性の名詞を修飾する形容詞の利用（24%），第３に項となる値への修飾や助詞の利用である（17%）．

判断を含んだ用言としては,「急上昇」「急落」「急増」「急伸」のように変化の速度に言及しているもの,「大幅増」「微増」「横ばい」のように変化の量に関するもの,「高騰」「低迷」のように値そのものに関するものがある．また,「達成」「突破」「盛り返す」や「確保」「持ち直す」にもある種の評価が読み取れる[*13]．評価ではないが,「倍増」「半減」という語は，変化だけでなくその程度を表現する用言である.「横ばい」は,「内閣支持率は「支持しない」は２９％（前回比３ポイント増）…で横ばいだった (990615023)」という表現の中に現れる．数値的には多少の変化があるわけで，そこに注目して type1:c=”upward” としてコーディングした場合,「横ばい」という表現はそれが非常に小さくて変化として意味を持たないという判断を加えていることになる．もうひとつ興味深い振る舞いをするのは,「とどまった」という語で,「前年度並みにとどまった (980422080)」のように変化がないことを示すだけでなく,「同５・５%減と１ケタ減にとどまった (981028080)」のように変化が小さいことや,「〜の３８%に対し、社会党政党支持率は２４%にとどまった (980811029)」のように値が小さいことの判断を付加している．

副詞あるいは用言性名詞への形容詞としては,「激しく」「急激に」「爆発的に」「大幅に」「わずかに」(「〜な」の形も現れる）に加えて,「高い（伸び)」「穏やかな（下落傾向)」「かろうじて」等が使われる.「〜の勢いで」「予想以上に」「順調に」にも評価が読み取れる．

値への修飾や助詞の利用は，時間幅に関して「わずか〜で」「〜足らずで」，値について「〜しか（〜ない)」「〜も」「わずか（〜しかない，〜にすぎない)」がある.「〜の大台」という表現にも価値判断が加わっている.「超安値」「高支持」のように対象に依存した語彙もある．

この他に，状況情報との境界線が明確でない表現がある.「高（低）水準」「上昇（減少）基調」「急成長」「躍進」「好転」「失速」は，値の大小や変化と何らかの関係があり，それらの情報を示しつつ，評価を上乗せしているとも感じられるが,「快進撃」「(絶）好調」「堅調」「苦戦」のようなものについては，type0, type1, type2 で表現される類型的な情報とどのように対応しているかが必ずしも明確でない．

*13　ここにはあげていないが，順位の表現では,「返り咲く」「転落する」「躍り出る」等，評価を上乗せした用言が多用される．

5. コーディング再考

本章では，作成したコーパスの仕様，特にコーディングについて改善が必要と思われる点を述べる．既に指摘した問題として，4·2 節で議論した，現在の仕様では表現できない変化の記述がある．情報抽出の仕様という観点では，非定型的な事象として扱わないでおくという判断もありうるが，ここでのコーディングを時系列情報の値や変化に関する情報を形式的に記述する枠組みと考えた場合は大きな問題である．ただ，ここでそれを再度論じることはせず，コーパス作成の指針という観点からの問題に限る．つまり，コーディングは可能であるが，どうコーディングすべきかの判断が難しい（5·1 節,5·2 節），あるいはコーディングの結果得られる情報が直感にあわない（5·3 節–5·5 節）等の問題を指摘する．前者の問題については，明確な作業基準作りが必要となる．後者についてもそのような基準によって救われる面はあるが，自然言語が本来的に持つ表現もしくは現象の捉え方の多様性を反映した問題でもあり．本質的な解決は難しいように思われる．

5·1 情報単位の認定

現在のコーディングは節毎に一つの情報単位としてまとめることを薦めているが，この節という単位は必ずしも自明ではない．例えば，以下のようなヴァリエーションが考えられるが，どこまでをひとつの情報単位としてまとめるかは難しい．この点についての指針が必要である．

前月比 100 万台増［の | と | で］700 万台となった
700 万台［と | で］前月比 100 万台増だった
700 万台［と | で］前月に比べ 100 万台増加した
700 万台となり，前月比 100 万台増だった
前月に較べて 100 万台増加し，700 万台となった

また逆に，一つの情報単位において，あるタイプの情報は一つしか含められないために，表現的にも内容的にもひとつの状況や変化と考える方が自然なものを複数に分割せざるをえない場合がある（3·2·1 節を参照）．これは不自然であると同時に，情報の重複した投入が必要となり，望ましくない．より明確で自然な情報単位の認定と記述方法が求められる．

5·2 抽出すべき情報の認定

「93 年に減少が始まった」「93 年から減少している」「93 年から減少が続いている」という表現すべてから，会話の含意を含めて以下の２つの情報を読み取ることができる．

- 93 年に減少が始まった
  time=93 年, type1:c=”donward”, type2:cc=”start”
- 93 年から現在まで減少が続いている

time=記事内基準時点, type1:c=”donward”,
type2:cc= ”continue”, type2:s-t/dur=93 年

どの表現の場合にこれらの情報のどちらあるいは両方を抽出するかの判断が必要である．このような会話の含意より強い推論の帰結として「３年ぶりに$10 となった」という表現には「３年前に$10 であった」ことも情報に含まれるが，これを抽出するかも問題となる．同様に「ピークの 93 年には$10 に達した」は,「93 年から変化が反転して下降が始まった」（type1:c=”downward”, type2:cc=”reverse”）と表現すれば情報的には充分であるが，表現に合わせると「93 年に$10 まで上昇があった」（type1:c=”upward”）ことも抽出したくなる．このような抽出すべき情報の選定基準を明確にする必要がある．

情報単位での抽出の判断より細かい問題として，例えば「倍以上に高騰し、…２１～２２ドルになった (990924069)」という表現において,「～になった」の部分から変化があったという情報（type1:c=”some”）だけを記述するか,「高騰」と合わせて，上昇方向の変化があった（type1:c=”upward”）とするかの判断がある．

このように局所的な表現には現れていないが，文脈や一般常識から読み取れる情報は幾つかある．例えば,「サービス開始当初は…爆発的な伸びをみせ、…加入台数は、２年間で７００万台を超えた (980422135)」は０台からの増加であることが明らかである．加えて,「１カ月で$9 から$10 になった」「昨年の 120%になった」という表現から$1 や 20%の差分を読み取って，表現されていない値を記述するかも問題となる．この場合，コーディングの仕様が表現を直接反映した形で情報を記述することを妨げているため，情報を落とすか，表現から乖離して，直接は述べられていない内容で抽出結果を補完するかという判断が求められる*14．

5·3 事象の反復や継続の扱い

type2 は，type1 で表現される変化に関する情報を修飾するような（主にその変化を示す）情報である．ところが，この中で type2:cc=”convex”だけは，type0 で表現される値の状況に関する情報への修飾となっている．例えば,「ＮＹ市場で原油続騰、９７年末以来の水準に (990714086)」は明らかに値の状況に言及している．type2:cc=”convex” を付与される表現は常にこのような言及であるので，記述上の矛盾を招くわけではないが，整合性という点で望ましくない．5·1 節で述べた情報の単位という面でも，他の type2 の情報は type1 の情報と分離できないが，type2:cc=”convex”の場合は type0 の情報と分離できないという違いを生じる．

これに関連して,「～ぶりに」という表現に対応する type2 の情報が変化（type1）に関する場合と，値の状況（type0）に関する場合とがあることが問題となる.「３年ぶりに上昇した」は過去３年間上昇しておらず，３年前に上昇したこと，つまり「上昇が３年ぶりに生じた」ことを表現しているが,「３年ぶりに$10 まで上昇した」は.「$10 への上昇が３年ぶりに生じた」のではなく,「３年ぶりに$10 という値になった」，より厳密に言うと「になった」のではなく「であった」ことを表現している．前者は３年ぶりに上昇が始まったということで，type2:cc=”start”となり，後者は３年ぶりに$10 であるということで type2:cc=”convex” を割り当てることになる．そもそも「～ぶりに」という表現が曖昧性を持つので*15，それに応じて情報が異なってくることは当然とはいえ，この２つの対立は直感にあうものではない．

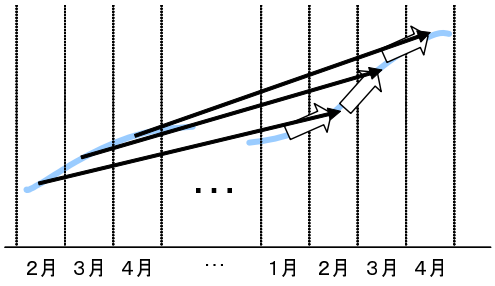


図 6 変化の継続と状況の繰り返し

更に,「3 年ぶりに$10 を割り込んだ」という場合,「$10 を割り込んだ」は，$10 以下になるということで，type0:t-rel=”before”として表現されており，$10 以下であることの表現（type0:t-rel=”during”, type0:v-rel=”smaller”）とは異なるので，この表現は２つの情報単位として「$10 を割り込んだ」「3 年ぶりに$10 以下である」とを分けて記述する必要がある．ここにも表現と情報単位の乖離が生じる．

「～ぶりに」というような同じ状況の反復だけでなく，同じ状況が継続して生じているという表現も直感に反する記述となる場合がある.「３カ月連続で前月を上回っている」という表現は，上昇（上回る）という変化が３カ月継続していることを示しており，図式的には図 6 の白抜き矢印で表されるような状況を表現している．一方，生産量や販売量の比較でよく用いられる「３カ月連続で前年同月を上回っている」は，同じ図の黒矢印で表されるが，ある種の変化が継続しているというよりもある種の状況（事態）が繰り返し生じているという印象を受ける．現在のコーディングでは，この２つは，もちろん type1 の記述は異なるが，type2 はまったく同じ記述となる．変化の継続と状況の繰り返しを同じ記述としてよいかには疑問が残る．

*14 これについては，3·2·2 節で述べた，明示されていない値の扱いに従い，補完した記述を行う仕様としている．

*15 “John opened the window again.” に現れるような文を修飾する “again” に，その出来事が過去にも起こったという解釈とそれによって生じた事態が過去にも起こっていたという解釈の２つがあるというよく知られた問題（例えば [Jäger 99]）と同じである．

5·4 変化と値の状況の曖昧性

5·3 節末尾の問題とも若干関係するが，本コーディングは値の状況よりも変化を基本的なものとして扱っている．例えば,「$10 を下回った」という情報を，$10 以下であるという値の状況ではなく，下落しており$10 であった過去の時点があったと表現する．同様に「昨年を下回った」も「昨年から下落している」として表現する．しかし，同じことを意味している表現で,「前年割れした」「同［前年比］３・１ポイント低い８２・７％となった (980122071)」等，変化とは捉えにくいものがある．また「９７年度末の固定電話の加入台数６０３８万台は、ピークとなった９６年度（６１４６万台）だけでなく、９５年度（６０７７万台）をも下回っている (980421070)」という表現も大局的に見れば 95 年から減少してはいるが，減少が続いていたわけではないので，この「下回っている」を減少と記述するのは適切と思えない*16．

更に，このコーディングでも「$10 を上回っている」は値の状況として表現するが,「昨年を上回っている」は「昨年を上回った」と同じ扱いをするしかないという統一性を欠いた状況となっている．また,「$10 を上回った 95 年に較べ」という表現もされるが，これはむしろ「$10 を上回っていた 95 年」なのではないかと推察され，ここから「95 年に$10 を上回った」という情報を抽出することにはためらいがある．このような表現と抽出される情報の乖離は作業の正確性と効率性を落とす原因として懸念される．

5·5 時間に基づく記述の扱い

時間の情報については，幅のある時間ではなく，時点に基づいた記述となっている．もちろん時点として取られる時間の粒度（幅）は様々である．これは統計量そのものでも同じで，例えば 98 年の出荷台数は，98 年という時間幅の間に出荷された台数であるが，統計量としては 98 年の値として一つの点となる．同じ扱いが変化についてもなされていて,「98 年の出荷台数は 97 年から 300 万台増加した」という表現は，97 年という時点から 98 年という時点に渡って上昇という変化があったことを示している．大抵の場合，このような時間の捉え方で問題はないが，対象としたテキストの中には稀にこれに収まらない表現が現れる．例えば.「９９年度に２３０万台減少する (990227098)」における「９９年度」は時点ではなく時間幅で，時点に基づいた記述では「この後，９９年度末までに２３０万台減少する」となる．このような表現をどう扱うかも，表現との整合性と抽出される情報の正確性をバランスして検討する必要がある．また，そうして規定されたコーディングの仕様が作業者にとって理解し，遵守しやすいものであることも重要である．

*16 現在ではこれを避けるために type0:val に特別な値を入れている（4·2 節を参照）．

6. お わ り に

動向情報編纂研究の資源となることを目標として構築した時系列情報の値と変化に関する言語表現コーパスについて報告した．テキスト群に含まれる時系列情報の値と変化は，動向情報の要約と可視化の素材を提供し，そこへの情報アクセスの起点ともなる．その点でこれらをテキスト中から抽出することは動向情報編纂の要となるものである．

本稿では値の変化に基づいたコーディングを提案し，それに基づいたコーパス構築を報告した．構築されたコーパスを検討することで，今回対象とした日本語新聞記事テキストでの時系列情報の値と変化に関する表現の特徴をまとめた．その結果，大部分の情報がパタン化しやすい表現を持っていること，そこで用いられている語彙も分野依存性が少ないこと等，情報抽出に有利な特徴を持つことが明らかとなった．また，多量のテキストに対してコーディングを行ったことによって明らかとなったコーディングの問題点を考察した．こちらについてあげられた問題は多いが，その原因は設計されたコーディングの稚拙さではなく，言語表現の多様さである．多様な表現から必要充分な情報を意味的な整合性を持って抽出・記述できる枠組であると同時に，作業者の負荷やその間の揺れを少なくするような対象となる表現を素直に反映した枠組であることは難しい．とはいえ，これらの問題点を整理できたことは，本コーパス構築の成果であると考える．

以上のような点を明らかにできたことが現時点での本コーパス構築の意義であるが，それに加えて，構築されたコーパスが多くの研究における資源として活用されることで更に大きな意義を持つことを期待している．

謝　　辞

MuST の運営とそれに関連する研究は，NTT と東京大学との産学連携共同研究，ならびに国立情報学研究所の NTT と東京大学との公募型共同研究によって支援されています．ご支援をここに感謝いたします．MuST の活動全般と本稿の執筆は MuST に参加頂いた皆様によって可能となったものです．あらためて感謝いたします．

◇ 参 考 文 献 ◇

[Bethard 08] Bethard, S., Corvey, W., et al.: Building a Corpus of Temporal-Causal Structure, in *Procs of LREC2008*, pp. 909-915, 2008.

[井出 08] 井手 一郎・木下 智義 他「大量ニュース映像を対象とした時系列意味構造に基づく情報編纂手法の提案」 人工知能学会論文誌, Vol. 23, No. 5, pp.282-292, 2008.

[Inoue 08] Inoue, T., Yamamoto, T., et al.: Extraction of Chronological Statistics Using Domain Specific Knowledge, in [NII 08], pp. 494-501, 2008.

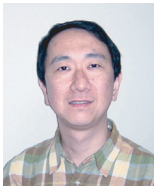
[Jäger 99] Jäger, G. and Blutner, R.: Against Lexical Decomposition in Syntax, in *Procs of the Fifteenth Annual Conference of the Israel*

Association for Theoretical Linguistics, pp.113-137, 1999.
[加藤 06]　加藤 恒昭・松下 光範「情報編纂（Information compilation）の基盤技術」第 20 回人工知能学会全国大会, 1D3-2, 2006.
[加藤 07]　加藤 恒昭・松下 光範「時系列情報の抽出と可視化に基づく情報アクセスためのマルチモーダルインタフェース –情報編纂の基盤技術に向けて–」人工知能学会論文誌, Vol. 22, No. 5, pp. 553 - 562, 2007.
[Kato 07]　Kato, T., Matsushita, M. and Kando, N.: Expansion of Multimodal Summarization for Trend Information –Report on the First and Second Cycles of the MuST Workshop –. in *Procs of the 6th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 235–242, 2007.
[Kato 08]　Kato, T. and Matsushita, M.: Overview of MuST at the NTCIR-7 Workshop: Challenges to Multi-modal Summarization for Trend Information, in [NII 08], pp. 475-488, 2008.
[加藤 09]　加藤 恒昭・松下 光範・神門 典子「動向情報の要約と可視化とその展開 – MuST（動向情報の要約と可視化に関するワークショップ）活動報告–」情報処理学会自然言語処理研究会, 2009-NL-190-12, 2009.
[Mani 08]　Mani, J., Hitzeman, J., et al.: Spatial ML: Annotation Scheme, Corpora, and Tools, in *Procs. of LREC2008*, pp. 410-415, 2008.
[松下 09]　松下 光範・加藤 恒昭「情報編纂研究促進のための試み」人工知能学会論文誌, Vol.24, No.2, pp.272-283, 2009.
[Mori 08]　Mori, T. and Miyazaki, R.: A Simple Baseline Method for NTCIR-7 MuST T2N Task - Yokohama National University at NTCIR-7 MuST T2N –, in [NII 08], pp. 502-508, 2008.
[Nanba 08]　Nanba, H.: Extraction of Trend Information from Newspaper Articles: Hiroshima City University at NTCIR-7 MuST, in [NII 08], pp. 489-493, 2008.
[NII 08]　National Institute of Informatics: *Procs of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, 2008.
[西山 08]　西山 莉紗・竹内 広宜 他「新技術が持つ特長に注目した技術調査支援ツール」, 人工知能学会論文誌, Vol. 24, No. 6, pp.541-548, 2009.
[NIST 08]　http://www.itl.nist.gov/iad/mig/tests/ace/
[Palmer 05]　Palmer, M., Gildea, D. and Kingsbury, P.: The Proposition Bank: A Corpus Annotated with Semantic Roles, in *Computational Linguistics*, Vol. 31, No. 1, pp. 71-106, 2005.
[Pang 08]　Pang, B. and Lee, L.: Opinion Mining and Sentiment Analysis, in *Foundations and Trends in Information Retrieval*, Vol. 2, Nos. 1-2, pp. 1-135, 2008.
[Pustejovsky 05]　Pastejovsky, J., Sauri, I.B., et al.: The Specification Language TimeML, in Mani, I., Pustejovsky, J. and Gaizauskas, R. (eds.), *The Language of Time: A Reader*, pp. 545-557, Oxford University Press, 2005.
[Sarawagi 07]　Sarawagi, S.: Information Extraction, in *Foundations and Trends in Databases*, Vol. 1, No. 3, pp. 261-378, 2007.
[Seki 08]　Seki, Y., Evans, D. K., et al.: Overview of Multilingual Opinion Analysis Task at NTCIR-7, in [NII 08], pp.185-203, 2008.
[Shneiderman 09]　Shuneiderman, B., and Plaisant, C.: *Designing the User Interface (fifth edition)*. Addison-Wesley, 2009.
[Uenishi 08]　Uenishi,Y., Masui, F., et al.: Trend Information Extraction Based on Relative Expression Participated on MuST T2N Subtask, in [NII 08], pp. 509-514, 2008.
[Yoshida 08]　Yoshida, M., Sugiura, T., et al.: TDU Systems for MuST: Attribute Name Extraction, Text-Based Stock Price Analysis, and Automatic Graph Generation, in [NII 08], pp. 520-527, 2008.

〔担当委員：阿部 明典〕

2010 年 2 月 2 日　受理

著　者　紹　介



加藤　恒昭（正会員）

1959 年生まれ．1981 年東京工業大学工学部電気電子工学科卒業，1983 年東京工業大学大学院総合理工学研究科電子システム専攻修士課程修了．同年，日本電信電話公社（現 NTT）に入社．2000 年東京大学大学院総合文化研究科言語情報科学専攻准教授，2010 年同教授，現在に至る．質問応答，マルチモーダル情報アクセス，語彙意味論に関する研究に従事．博士（工学）．電子情報通信学会，情報処理学会，言語処理学会，ACL 各会員．

松下　光範（正会員）

1969 年生まれ．1993 年大阪大学工学部精密工学科卒業．1995 年大阪大学大学院基礎工学研究科物理系専攻制御工学分野博士前期課程修了．同年 4 月，日本電信電話 (株) 入社．2008 年関西大学総合情報学部准教授、2010 年同教授，現在に至る．インタラクションデザイン，自然言語理解に関する研究に従事．博士 (工学)．情報処理学会，日本ファジィ学会，日本バーチャルリアリティ学会，ACM 各会員．

神門　典子（正会員）

1994 年慶應義塾大学文学研究科博士課程修了．博士（図書館・情報学）．同年学術情報センター助手．1995 年米国シラキウス大学客員研究員，1996～1997 年デンマーク王立図書館情報大学客員研究員．1998 年学術情報センター助教授．2004 年より国立情報学研究所教授，現在に至る．テキスト構造を用いた検索と情報活用支援，情報アクセスシステムの評価等の研究に従事．ACM-SIGIR，ASIS&T，言語処理学会，情報処理学会，日本図書館情報学会各会員．