Vol.2021-NL-251 No.8
Vol.2021-SLP-139 No.8
2021/12/1

定型表現に着目した情報の抽出・可視化による防災・安全のためのテレビアーカイブアナリティクス

片山 紀生[1,2,a)]　孟 洋[1,2,b)]　佐藤 真一[1,c)]

概要：インターネットや SNS の普及により情報入手の手段が多様化しており，テレビ報道の役割は相対的に小さくなっているが，防災・安全に関する情報源としての役割は今でも大きい．そのため，そのようなテレビの映像・音声を大量に集積したテレビアーカイブは，防災・安全に関する情報を発見・理解する手段として有用であると考えられる．本研究では，テレビアーカイブをそのような目的で解析するための手法のひとつとして，テレビ報道における定型表現に着目した重要語句等の抽出，解析，視覚化を行うことにより，テレビ報道に対する気付きや再認識を支援できることを示す．

キーワード：テレビアーカイブ，データアナリティクス, 防災, 安全, 定型表現

Television Archive Analytics for Disaster and Risk Prevention with Information Extraction and Visualization Focusing on Conventional Expressions

Abstract: Data analytics of news contents in television archive is useful for disaster and risk prevention since it helps to discover and understand various information on disasters and risks. Today we have various types of information acquisition methods and younger generation is more likely to use online media rather than television. However, television still plays an important role in disaster and risk prevention. This paper shows information extraction and visualization methods with focusing on conventional expressions that typically appears in TV news scripts.

Keywords: Television Archive, Data Analytics, Disaster Prevention, Risk Prevention, Conventional Expression

1. はじめに

インターネットや SNS の普及により情報入手の手段が多様化しており，テレビ報道の役割は相対的に小さくなっているが，防災・安全に関する情報源としての役割は今でも大きい．そのため，そのようなテレビの映像・音声を大量に集積したテレビアーカイブは，防災・安全に関する情報を発見・理解する手段として有用であると考えられる．本研究では，テレビアーカイブをそのような目的で解析するための手法のひとつとして，テレビ報道における定型表現に着目した重要語句等の抽出，解析，視覚化を行うことにより，テレビ報道に対する気付きや再認識を支援できることを示す．

[1] 国立情報学研究所
National Institute of Informatics
[2] 総合研究大学院大学
The Graduate University for Advanced Studies, SOKENDAI
[a)] katayama@nii.ac.jp
[b)] mo@nii.ac.jp
[c)] satoh@nii.ac.jp

2. 背景

2.1 テレビアーカイブの活用

近年の計算機技術の進歩により，大量の映像をデジタル記憶することが可能になり，テレビ放送を長期間に渡って蓄積したテレビアーカイブが作られるようになっている．国立情報学研究所でも，学術研究目的のテレビアーカイブである NII TV-RECS [1], [2] を構築しており，2009 年 8 月より東京地区の地上波 7 チャンネルの放送映像を連日 24 時間蓄積している．また，映像情報に加えてテキスト形式の情

ⓒ 2021 Information Processing Society of Japan

報として，字幕放送の文字字幕（クローズドキャプション）および番組情報も蓄積している．このようなテレビアーカイブを構築する動きは海外でも進められており，アメリカ合衆国の Vanderbilt Television News Archive，TV News Archive（Internet Archive），UCLA Library NewsScape，フランスの INA MEDIAPRO Television Collection，オランダの Beeld en Geluid（Sound and Vision）などがある．このようなテレビアーカイブには，様々な災害，事件，事故に関するニュースが蓄積されており，防災・安全意識を高めるためのリソースとして活用することが考えられる．

インターネットの普及により若者のテレビ離れが進んでいることが指摘されており，情報獲得手段としてのテレビ放送の役割は，相対的に小さくなっていると考えられる．しかし，テレビ放送は，同報性や広域性の点では現在でもメリットがあり視聴者も多い．視聴者が多いことで，衆人環視による評価の目も厳しくなるため，品質や信頼が保たれやすい面もある．実際，総務省の調査によると，テレビに対する信頼度は，新聞に次いで高いことが示されている [3]．また，日本新聞協会による調査においても，テレビは新聞に次いで信頼度の高いメディアとなっており，かつ，コロナ禍で接触頻度が最も増えたメディアとされている [4]．そのため，特に災害時等での情報伝達では，現在でも大きな役割を担っていると考えられる．

2.2 マルチメディアアナリティクス

今世紀に入り，インターネット，および，センサーや記憶装置などの電子デバイスの高性能化により，大量のデータを収集・蓄積できるようになり，いわゆるビッグデータの時代が到来している．そして，それに呼応して，2010 年頃より，ビッグデータを解析するための基盤として，Data Analystics や Data Science に注目が集まるようになっている [5]．また，マルチメディアデータを対象としたアナリティクスとして，Multimedia Data Analytics，あるいは，Multimedia Analytics にも関心が高まり，大量のマルチメディアデータを解析および可視化するための基盤として，方法論やプラットフォームの研究開発が進められている [6], [7], [8], [9], [10], [11], [12].

一方，テキストデータを対象としたビッグデータの解析では，早くから，地震や台風などの災害発生時の SNS データ（ツイートなど）を用いた技術や応用が注目されており，人々の振舞いを把握・解析する上で，位置情報付きのツイートなどが有効であることが示されている [13], [14], [15], [16].

そのような中，国立情報学研究所では，学術研究目的のテレビアーカイブである NII TV-RECS [1], [2] を用いたマルチメディアアナリティクスの研究に取り組んでおり，2011 年の東日本大震災の際には，「NII 研究用テレビジョン放送アーカイブを用いた東日本大震災の社会的影響の学術的分析」というテーマで共同研究を公募し，震災時のテレビ放送について，幅広い視点での解析が行われた [17], [18]．筆者らも，映像の構図による分類に基づく震災テレビ報道の分析を行った [19]．また，2019 年には西日本豪雨（平成 30 年 7 月豪雨）の際のテレビ報道の傾向解析 [20] を行うとともに，2020 年にはテレビアーカイブを防災・安全意識を維持・高めるための道具として応用する可能性について考察した [21]．2021 年には新型コロナ禍での報道を対象として関連ニュースショット検出の有用性を検証した [22].

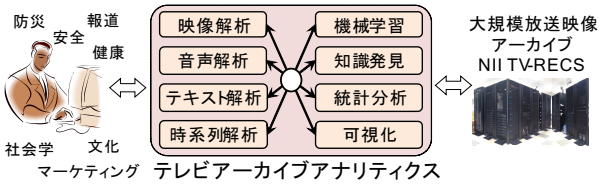


図 1 テレビアーカイブアナリティクス
Fig. 1 Television archive analytics.

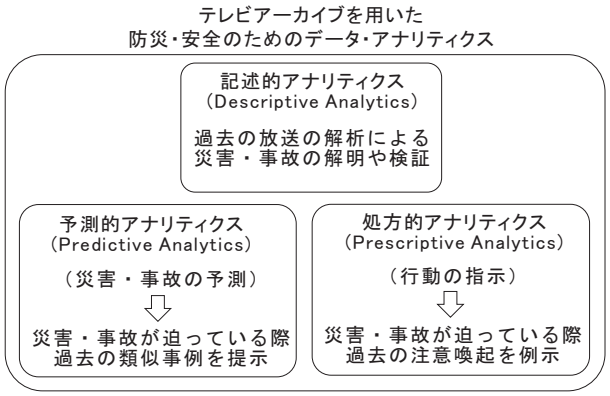


図 2 防災・安全のためのデータ・アナリティクス
Fig. 2 Data analytics for disaster and risk prevention.

3. テレビアーカイブを用いたアナリティクス

3.1 防災・安全のためのアナリティクス

データアナリティクスは，しばしば以下の 3 つの段階に分けて考えられる [11], [23].

- 記述的アナリティクス（Descriptive Analytics）
- 予測的アナリティクス（Predictive Analytics）
- 処方的アナリティクス（Prescriptive Analytics）

記述的アナリティクスは，過去に何が起きたかを解析するものであり，予測的アナリティクスは，将来に何が起きるかを予測するものであり，処方的アナリティクスは，何をすべきかを提案するものである．これを，テレビアーカイブを解析対象とし，かつ，防災・安全を目的としてどのようなことが可能か考えると以下のようになる．

まず，記述的アナリティクスは既に広く行われていることであり，大きな災害や事故が起きた場合にテレビアーカイブを解析することで何が起きたかを明らかにできる．実際，前節で述べたとおり，東日本大震災や西日本豪雨を対象とした解析が行われている．

ⓒ 2021 Information Processing Society of Japan

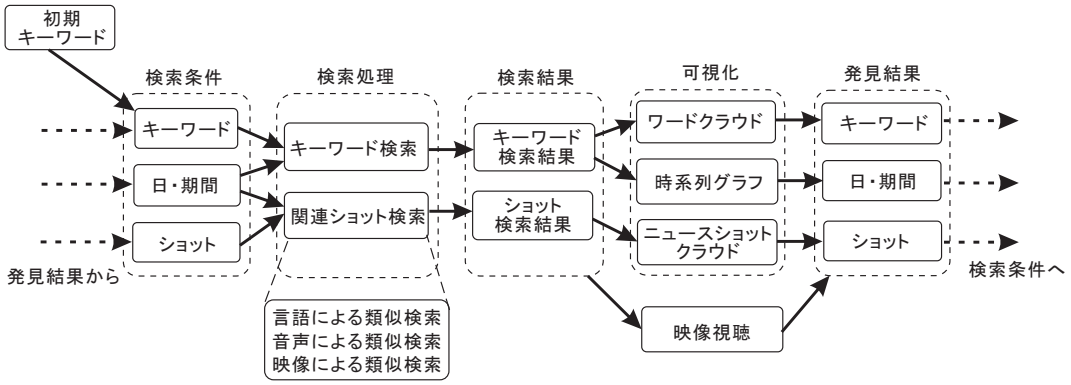


図 3 対話的プロセスの流れ
Fig. 3 Flow of interactive process.

次に，予測的アナリティクスは将来を予測することになるが，災害や事故をテレビアーカイブを用いて正確に予測することは用意ではない．しかし，何か災害や事故が迫っている際に，過去の類似事例を探索することで，起こるかもしれないことを見つけ出すことは可能だと考えられる．

そして，処方的アナリティクスは，データアナリティクスの３つの段階の中で，最も進んだ段階，すなわち，最も難しい段階と言われており，テレビアーカイブを用いるだけではできることが限られていると考えられる．しかし，予測的アナリティクスによって過去の類似事例が見つかった場合には，その際にどのような注意喚起が行われていたかを抽出すれば，どのような行動を取るべきかについて参考となる情報を提示できる可能性がある．

これらのことから，テレビアーカイブを用いた防災・安全のためのアナリティクスとしては，図２のように取り組むことが考えられる．予測的アナリティクス，および，処方的アナリティクスについては，完全な予測や完全な処方は達成できないと考えられるが，参考となる情報を提示することは可能であると考えられる．

3.2 対話的プロセス

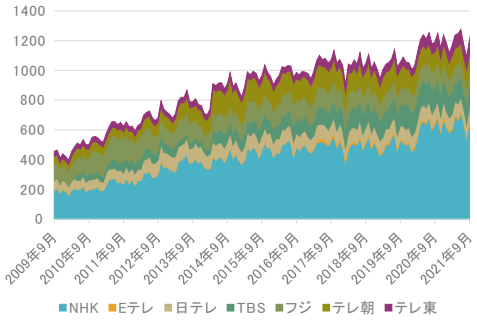
アナリティクスによる情報の探索は，視覚化を行いながらの対話的なプロセスになる．テレビアーカイブを対象に，情報を抽出・視覚化するツールを用いながら探索する過程は，図３のようになる．検索条件として，キーワード，ショット，期間を使いながら検索を行うとともに，検索結果を可視化したり，映像を視聴することによって，新たなキーワードやショットを発見しながら探索を進める．テレビアーカイブは一般に長期間に渡ってテレビ放送を蓄積しているため，対象とする期間を適切に絞り込むことも重要である．

4. 定型表現に着目した字幕の解析

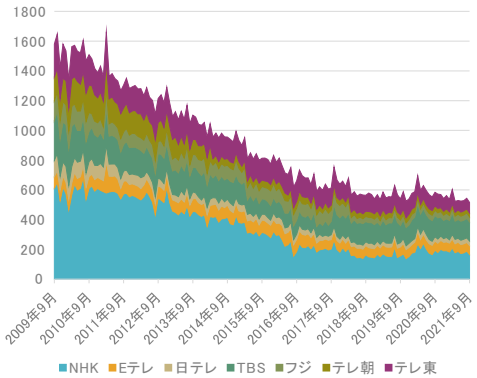
4.1 テレビの字幕放送

テレビアーカイブを解析する際，字幕放送として付与されている文字字幕（クローズドキャプション）は重要な解析対象である．音声をオフにした状態での視聴や聴覚障害者の視聴を支援するものであり，以下の制約があるものの，テレビが伝えている情報を把握するのに適している．

- 全ての番組に付与されている訳ではない．
  特に生中継を含むニュース番組の場合，音声認識および人手によるタイプ入力が必要になる．そのためかつては字幕が付与されているニュース番組は限られていたが，現在は図４に示したとおり，字幕が付与されているニュース番組の方が多くなっている．



(a) 字幕あり



(b) 字幕なし

図 4 字幕付きニュース番組の数
Fig. 4 Number of TV news programs with closed captions.

ⓒ 2021 Information Processing Society of Japan

- MeCab から出力される項目

  表層形　　品詞 1，品詞 2，品詞 3，品詞 4，活用型，活用形，原形，読み，発音

- MeCab からの出力例（入力：最新の台風情報にご注意ください）

  最新　　名詞，一般,*,*,*,*，最新，サイシン，サイシン
  の　　助詞，連体化,*,*,*,*，の，ノ，ノ
  台風　　名詞，一般,*,*,*,*，台風，タイフウ，タイフー
  情報　　名詞，一般,*,*,*,*，情報，ジョウホウ，ジョーホー
  に　　助詞，格助詞，一般,*,*,*，に，ニ，ニ
  ご　　接頭詞，名詞接続,*,*,*,*，ご，ゴ，ゴ
  注意　　名詞，サ変接続,*,*,*,*，注意，チュウイ，チューイ
  ください　　動詞，非自立,*,*，五段・ラ行特殊，命令，くださる，クダサイ，クダサイ
  EOS

- 形態素列（活用型，活用形，読み，発音は省く）※紙面の制約で折り返している

  <最新#名詞，一般,*,*，最新><の#助詞，連体化,*,*，の>
  <台風#名詞，一般,*,*，台風><情報#名詞，一般,*,*，情報><に#助詞，格助詞，一般,*，に>
  <ご#接頭詞，名詞接続,*,*，ご><注意#名詞，サ変接続,*,*，注意><ください#動詞，非自立,*,*，くださる>

図 5　MeCab の出力と正規表現照合の対象となる形態素列の例
Fig. 5　An example of MeCab output and morpheme sequence.

- 字幕が付与されていない区間がある．
  インタビューなど事前に録画された映像については，文字字幕ではなく，映像中にスーパーインポーズされた字幕（オープンキャプション）が付与されており，それらを解析するにはビデオ OCR が必要になる．本稿では，ビデオ OCR は今後の課題としている．
- 数十文字の短い行に分割されている．
  字幕は画面に表示することを目的としているため，文字数の少ない短い行に分割されている．そのため，句読点が正確に付けられている番組がある一方で，詩や散文のように句読点を付けずに，行末や空白で，文や節の区切りを表している場合がある．そのため，文単位で解析しようとすると，行から文への変換が必要になる．本稿では，ジャンルがニュースとなっている番組のみを対象としており，ほとんどの番組が句点によって文末を区切るスタイルが取られていた．一部に詩や散文のスタイルを取っているものが見られたが，それらについては行が表示される間隔の長さの評価，および，形態素解析結果に基づく文末である可能性の評価によって字幕を分割し，それらを文として扱った．

テレビ放送の字幕の解析はこれまでにも様々な試みがあり，例えば，ニュース番組のトピックを長期間に渡って追跡するトピックスレッディング [24] や，テレビ放送の字幕からの音声コーパスの構築 [25] が行われている．

4.2　形態素列に対する正規表現照合

本研究を進めるにあたり，定型表現の解析を容易にすることを目的として，形態素列に対する正規表現照合のツールを実装した．定型表現を解析する際には，形態素解析の結果として得られる表層形および素性を組み合わせたパターンによる照合が有用であるが，そのパターンを正規表現の形式で容易に記述することを可能にしている．ここで正規表現を用いているのは，以下の理由からである．

- プログラミング言語に実装されている高速な照合機構を利用可能である．
- 正規表現の記法に慣れていれば正規表現の記述力を最大限活用可能できる．

本研究では形態素解析には MeCab[26] を使用しており，出力される形態素の表層形と素性を，以下のように井桁文字 (#) でつなぎ，それらを山括弧 (<>) で形態素ごとに括った文字列を形態素列としている．

<表層形 1#素性 1><表層形 2#素性 2>...

このとき，表層形はすべて全角文字に変換しており，素性については，活用型，活用形，読み，および，発音は，本研究では使用しないため省略している．図 5 に MeCab の出力および正規表現照合に使用する形態素列の例を示す．

このような形態素列に対して，表層形と素性を組み合わせたパターン照合を行う場合，morfgrep[27], [28] のように独自の記法を用いる方法も考えられるが，本研究では上記の目的のために正規表現による照合を用いている．ただし，このような形態素列に対する正規表現は複雑なものになるため，ユーザがゼロから正規表現を記述するのは現実的ではない．そこで以下の略記法を導入し，ユーザは略記されたパターンを記述し，それを前処理において完全な正規表現に変換し，それを用いて照合処理を行うことにした．

- % を [^<>#] に変換し形態素内のワイルドカードとする．これを使えば形態素をまたがってマッチすることがなくなる．
- 表層形や素性の省略を可能にする．<○○○> および <○○○#> を <○○○#%> に，<#△△△> を <%#△△

△> に変換する．

- 素性の前方一致を可能にするため，閉じ山括弧の直前にカンマがある場合にはワイルドカードを加える．例えば <○○○#品詞 1,> は <○○○#品詞 1,%> に変換する．
- 山括弧 < > でくくられていない文字列は自動的に品詞を問わない形態素列に変換する．例えば，「最新の台風情報」は「<最新><の><台風><情報>」に変換され，さらに上述の略記法が適用される．なお，形態素への変換結果は一意ではなく前後の語によって変化し得るため，自動変換の結果がユーザの意図したものと一致しないことが起こり得るので形態素解析の結果が一意になりにくい文字列を使うときには注意が必要である．
- 山括弧 < > および井桁 # を，丸括弧 ( ) と同じ優先度の区切り文字と解釈し，<○○○#△△△> は (<(○○○)#(△△△)>) に変換する．これにより，<○○○>* と書いた場合に，繰り返しの範囲が閉じ山括弧 > １文字ではなく山括弧で括られた全体になる．また，<です|ます#助動詞,> と書いたときに，| の適用範囲が山括弧や井桁を越えなくなる．
- 頻繁に使うパターンを略記できるように，<:名詞句:> や <:動詞句:> などのマクロを用意している．
- 標準的な正規表現では，(○○○) がキャプチャリンググループ，(?:○○○) が非キャプチャリンググループとなっている．しかし，実際には，キャプチャしたいのはパターン中の少数箇所であることが多く，それ以外の全ての開き丸括弧を (?: と書かなければならなくなる．これを簡単にするために (??○○○) という記法を導入し，この丸括弧のみをキャプチャリンググループに変換し，それ以外の丸括弧はすべて非キャプチャリンググループに変換することにしている．これにより，キャプチャリンググループを使う際のパターンの長さを短くすることが可能になる．

この略記法を用いると，図 5 の形態素列に以下のパターン照合すると，それぞれ下記の結果が得られることになる（紙面の都合で照合結果については，素性を省略し表層形のみ示している）．

- ください
  → <ください>
- <#名詞, サ変接続,>
  → <注意>
- <#名詞,>+<#助詞,>
  → <台風><情報><に>
- <#接頭詞,>*<#名詞,>+<#名詞, 接尾,>*ください
  → <ご><注意><ください>

4.3 抽出・可視化ツール

アナリティクスは，対話的に使用するプラットフォームであり，本研究では情報の抽出・可視化のためのツールとして以下のものを実装している．

- 頻出語句の出現頻度の時系列グラフ表示
- 頻出定型句の出現頻度の時系列グラフ表示
- 頻出 KWIC の出現頻度の時系列グラフ表示
- 頻出語句・定型句のワードクラウド表示
- 頻出語句・定型句の共起ネットワーク表示

時系列グラフ表示は，出現頻度の時間的変化を把握するためのものであり，チャンネル別，番組ジャンル別，番組名別のプロットも可能になっている．出現頻度としては，出現した回数，出現した文の数，出現した番組の数でのプロットが可能になっている．他方，ワードクラウドは，多数の頻出語句や定型句を包括的かつ直観的に把握するのに適しており，共起ネットワークは，頻出語句や定型句の文中での共起の頻度を把握するのに適している．

なお，頻出 KWIC (keyword in context) は，語句とその前後の語句の三つ組を抽出するものであり，定型句の一種とも見なせるが，語の用法の把握などに特に有用であるため，専用のツールとして実装している．このツールでは，単純に直前・直後の語を抽出するだけでなく，上述の正規表現照合を用いて，助詞・助動詞を挟んだ直前・直後の名詞句などの条件で抜き出すことも可能になっている．

5. 定型表現に着目した情報抽出・可視化の活用

5.1 テレビ報道における定型表現

テレビ報道は，不特定多数を対象として情報が伝えられるものであり，わかりやすさが重要であるため，奇をてらった表現が使われることは少なく，一般的に知られている表現が使われることが多い．そのため，定型表現を情報を発見するためのパターンとして使うことによって，大量のテレビ報道の中から特定の情報を抽出可能になると期待できる．

そのためのツールとしては，定型表現を見つける段階では，KWIC による用例の抽出，および，対話的な試行による正規表現パターンの探索が重要になる．また，有用な定型表現が見つかった場合には，その表現に合致する定型句を検索し，時系列グラフ，ワードクラウド，共起ネットワーク等を用いて可視化し，出現の傾向を解析することが可能になる．

このような考え方のもと，本研究では定型表現に着目した情報抽出・可視化の活用事例として，テレビ報道における注意喚起の解析，および辞書整備のための新語の探索を行ったのでその結果について述べる．

5.2 テレビ報道における注意喚起の解析

2018 年に発生した西日本豪雨（平成 30 年 7 月豪雨）は被害が甚大だったことから，事前の注意喚起の重要性が改めて再認識された出来事となっている．特に，ハザードマッ

ⓒ 2021 Information Processing Society of Japan

Vol.2021-NL-251 No.8
Vol.2021-SLP-139 No.8
2021/12/1

KWIC	合計	7/5	7/6	7/7	7/8
情報 に [注意] して	17	1	7	9	0
十分 な [注意] が 必要	8	1	1	0	6
もご [注意] ください 。	6	2	1	2	1
も 十分 [注意] する よう	6	2	4	0	0
に ご [注意] ください 。	6	2	2	2	0
情報 に [注意] するとともに	6	0	3	3	0
など に [注意] するとともに	6	0	0	6	0

(a) 注意

KWIC	合計	7/5	7/6	7/7	7/8
警戒 が [必要 です] 。 $	245	51	74	74	46
注意 が [必要 です] 。 $	31	10	7	4	10
警戒 が [必要 です] ね 。	7	3	1	3	0
注意 が [必要 です] ね 。	4	3	1	0	0
警戒 は [必要 です] 。 $	3	0	0	3	0

(b) 必要です

KWIC	合計	7/5	7/6	7/7	7/8
して [ください] 。 $	454	47	172	165	70
続けて [ください] 。 $	66	1	16	31	18
ないで [ください] 。 $	59	5	27	18	9
を ご覧 [ください] 。 $	26	9	5	4	8
ご 注意 [ください] 。 $	24	7	7	7	3
取って [ください] 。 $	24	1	13	8	2

(c) ください

KWIC	合計	7/5	7/6	7/7	7/8
ように [してください] 。 $	171	16	58	69	28
警戒 を [してください] 。 $	53	7	16	26	4
を 確保 [してください] 。 $	48	2	16	12	18
に 警戒 [してください] 。 $	24	2	7	12	3
に 行動 [してください] 。 $	21	0	11	6	4
避難 を [してください] 。 $	14	1	6	4	3
で 確認 [してください] 。 $	11	1	7	1	2

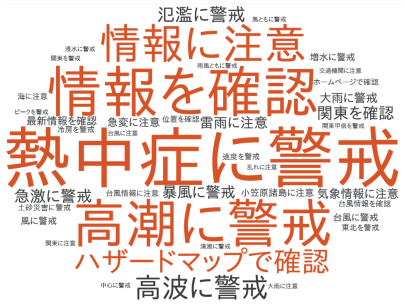
(d) してください

図 6 KWIC による定型表現の探索
Fig. 6 Searching conventional expressions with KWIC.

プの重要性が再認識される一方で，テレビ報道におけるハザードマップの取り上げ方を調査すると，西日本豪雨の発生後には大きく取り上げられたものの，発生前にはあまり取り上げられておらず，翌月に台風が近付いたときと比べても，発生前の取り上げ方が少ない傾向が見られた [20]．そこで，西日本豪雨で被害が集中したのが 2018 年 7 月 7〜8 日，気象庁の緊急会見が行われたのが 7 月 5 日だったことから，7 月 5〜8 日の間にどのような注意喚起が行われたのか調査した．

まず，頻出 KWIC の抽出ツールによって，「注意」という語の前後に出現する語句を調査した．図 6 (a) に，前後の 2 語を抽出した結果を示す．この結果から，「ご注意ください」，「注意してください」，「注意が必要です」という表現が多いことがわかる．そこで，同様に，「ください」や「が

(a) 西日本豪雨の直前（2018 年 7 月 5〜8 日）



(b) 台風 13 号の直前（2018 年 8 月 6〜8 日）

図 7 ニュース番組の字幕から定型表現によって抽出された注意喚起
Fig. 7 Warnings in closed captions of TV news programs.

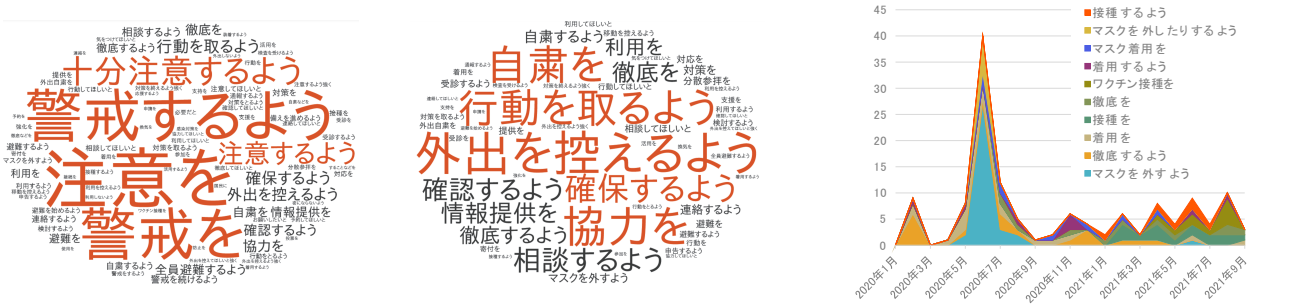
必要です」という表現にどのようなものがあるか調べるためにこれらの KWIC を調べたところ，同図の (b) および (c) の結果が得られた．また，(c) より「ください」は「してください」の形で頻出することがわかったため，「してください」の KWIC を抽出したところ同図の (d) の結果が得られ，注意喚起としては，「〜ください」，「〜が必要です」という表現が頻出し，中でも何か対象物への注意を促す表現としては，「〜に注意」，「〜に警戒」，「〜を確認」という表現がしばしば使われていることが明らかになった．

そこで，以下のパターンで注意喚起の対象を抽出することにした．

```
(??<:名詞句:>)<#助詞, 副助詞,>*
((??<に|を#助詞, 格助詞,>).*(??注意|警戒)|
 (??<で|を#助詞, 格助詞,>).*(??確認)).*
(ください|必要です)
```

このパターンを西日本豪雨直前の 2018 年 7 月 5 日から 2018 年 7 月 8 日のニュース番組の字幕に適用したところ，図 7 (a) の語句が抽出された．一方，翌月の台風 13 号直前の 2018 年 8 月 6 日から 2018 年 8 月 8 日のニュース番組の字幕に適用したところ，同図の (b) の語句が抽出された．これらの図から，西日本豪雨の際には水害および土砂災害への警戒が中心だったのに対して，台風 13 号の際には熱中症および高潮への警戒が中心だったことがわかる．また，ハザードマップについては，西日本豪雨の際には頻出していなかったのに対して，台風 13 号の際には上位 5

ⓒ 2021 Information Processing Society of Japan



(a) 「警戒」と「注意」を含む場合　(b) 「警戒」と「注意」を除いた場合　(c) 「マスク」または「ワクチン」を含む字幕の場合

図 8　2020 年 1 月〜2021 年 9 月に「呼びかけています」の形で行われた注意喚起
Fig. 8　Warnings in the form of reporting public appeals from Jan. 2020 to Sep. 2021.

件に入るほど頻出していたことがわかる．

次に，コロナ禍においても同様の注意喚起が行われたか調べたところ，上記の定型表現は気象災害でのみ使われており，新型コロナに関連するものは見られなかった．そこで，上述の KWIC による定型表現の探索によってさらに調べたところ，「〜呼びかけています」という表現が使われていることがわかった．そこで，以下のパターンを用いてどのような注意喚起が行われたか抽出することにした．

(??<:名詞句:>(?!.*<:名詞句:>).*) 呼びかけています

このパターンを 2020 年 1 月から 2021 年 9 月までのニュース番組の字幕に適用したところ，図 8 の結果が得られた．(a) は気象災害で現れやすい「警戒」と「注意」を含めた結果であり，(b) はそれらを除いた結果である．また，(c) はコロナ禍で特に中心的な話題となった「マスク」および「ワクチン」を含む字幕に限定した場合の結果である．(b) の「警戒」と「注意」を除いた結果を見ると，新型コロナに関して「呼びかけています」という形で様々な注意喚起が行われたことがわかる．気象災害の場合には，「してください」と放送局が主体となって注意喚起していたのに対して，新型コロナについては「呼びかけています」という形で他者の要請を伝達する形で注意喚起されている点が特徴的である．さらに (c) を見ると，2020 年の夏にマスクを外すように促されているが，これは熱中症対策としての注意喚起である．また，マスク着用に関する注意喚起が 2020 年 11 月頃に増加していることや，ワクチン接種に関する注意喚起が 2021 年に入って次第に増加している点も特徴的である．

5.3　辞書整備のための新語の探索

形態素解析を行う場合，辞書の整備は重要であり，ワードクラウドなど可視化の結果に与える影響が大きい．MeCab で使用可能な辞書としては，IPADIC，UniDic，NEologd が知られているが，どの語を 1 語と見なすかは用途によって変わってくることがあるとともに，新語への対応も必要

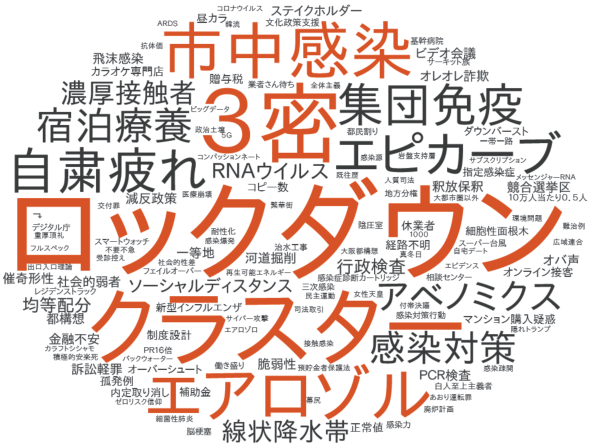


図 9　「いわゆる〜という」から抽出された語
Fig. 9　Word obtained from “what is called ...”.

になる．そこで，テレビ放送における新語を定型表現で抽出することを試みた．特にコロナ禍では，新型コロナウイルス自体が未知のウイルスであったために，様々な見慣れない語がテレビ放送に登場した．そこでそれらを探索することを目的とした．

上述の定型表現の探索と同様に，新語が現れやすい定型表現を探したところ，「いわゆる〜という」表現がしばしば使われることが明らかになった．そこで，以下のパターンを用いて，２語以上の名詞に分割されているもの，および，未知語となっているものを抽出した．未知語は原形が * になるため，以下のパターンでは <#名詞,%,*> という記述を使用している．

いわゆる (??<#名詞,>{2,}|<#名詞,%,*>) という

抽出した結果を図 9 に示す．なお，この結果を得るための形態素解析では IPADIC を使用している．図のとおり，様々な語が抽出されており，新語の候補を探索可能になっている．このパターンでは漏れている新語もあり得るため，今後さらに定型表現パターンを探索するとともに，逆に新語として登録する必要のない語も含まれ得るため，それらを選別する手法についても探究したいと考えている．

ⓒ 2021 Information Processing Society of Japan

6. まとめと今後の課題

本稿では，テレビアーカイブを用いた防災・安全のためのアナリティクスについて，定型表現に着目した情報抽出・可視化の有効性について検証するとともに，そのための情報抽出の手段として，形態素列に対する正規表現照合の略記法を提案した．そして，解析事例として西日本豪雨やコロナ禍でのテレビ報道に着目し，テレビ報道における注意喚起の状況を把握するのに有効であることを検証した．テレビアーカイブを用いた防災・安全のためのデータアナリティクスは，災害・事故の解明・予防にとって有用であると期待できる．今後は，映像中にスーパーインポーズされた字幕への対応を検討するとともに，対象番組をニュース番組のみならずワイドショーなどの情報番組に広げること，より複雑な内容解析を可能にすることなどを進めたいと考えている．

謝辞 本研究は，JST CREST（グラント番号 JPMJCR19F4）および JSPS 科研費（JP18K11386 ならびに JP21K11950）の支援を受けました．

参考文献

[1] 片山紀生，孟　洋，佐藤真一：映像インデクシング研究のための大規模放送映像アーカイブシステムの試作，情報処理学会研究報告 (DBS), 2002(41)，pp. 17–23 (2002).
[2] Katayama, N., Mo, H., Ide, I. and Satoh, S.: Mining Large-Scale Broadcast Video Archives towards Inter-Video Structuring, *Pacific Rim Conference on Multimedia (PCM2004) LNCS, vol.3332*, pp. 489–496 (2004).
[3] 総務省情報通信政策研究所：情報通信メディアの利用時間と情報行動に関する調査報告書 (2020).
[4] 日本新聞協会：新型コロナウイルスとメディア接触・信頼度調査 (2020).
[5] Cao, L.: Data Science: A Comprehensive Overview, *ACM Comput. Surv.*, Vol. 50, No. 3, pp. 43:1–43:42 (2017).
[6] Chinchor, N. A., Thomas, J. J., Wong, P. C., Christel, M. G. and Ribarsky, W.: Multimedia Analysis + Visual Analytics = Multimedia Analytics, *IEEE Computer Graphics and Applications*, Vol. 30, No. 5, pp. 52–60 (2010).
[7] Jónsson, B., Worring, M., Zahálka, J., Rudinac, S. and Amsaleg, L.: Ten Research Questions for Scalable Multimedia Analytics, *MultiMedia Modeling (MMM 2016), LNCS, vol.9517*, pp. 290–302 (2016).
[8] Wu, Y., Cao, N., Gotz, D., Tan, Y.-P. and Keim, D. A.: A Survey on Visual Analytics of Social Media Data, *IEEE Trans. on Multimedia*, Vol. 18, No. 11, pp. 2135–2148 (2016).
[9] Kurzhals, K., John, M., Heimerl, F., Kuznecov, P. and Weiskopf, D.: Visual Movie Analytics, *IEEE Trans. on Multimedia*, Vol. 18, No. 11, pp. 2149–2160 (2016).
[10] Renoust, B., Le, D.-D. and Satoh, S.: Visual Analytics of Political Networks From Face-Tracking of News Video, *IEEE Trans. on Multimedia*, Vol. 18, No. 11, pp. 2184–2195 (2016).
[11] Pouyanfar, S., Yang, Y., Chen, S.-C., Shyu, M.-L. and Iyengar, S. S.: Multimedia Big Data Analytics: A Survey, *ACM Comput. Surv.*, Vol. 51, No. 1, pp. 10:1–10:34 (2018).
[12] Zhang, W., Yao, T. and Zhu, S.: Deep Learning-Based Multimedia Analytics: A Review, *ACM Trans. on Multimedia Computing, Communications, and Applications*, Vol. 15, No. 1s, pp. 2:1–2:26 (2019).
[13] Imran, M., Castillo, C., Diaz, F. and Vieweg, S.: Processing Social Media Messages in Mass Emergency: A Survey, *ACM Comput. Surv.*, Vol. 47, No. 4, pp. 67:1–67:38 (2015).
[14] Li, T., Xie, N., Zeng, C., Zhou, W., Zheng, L., Jiang, Y., Yang, Y., Ha, H.-Y., Xue, W., Huang, Y., Chen, S.-C., Navlakha, J. and Iyengar, S.: Data-Driven Techniques in Disaster Information Management, *ACM Comput. Surv.*, Vol. 50, No. 1, pp. 1:1–1:45 (2015).
[15] Alfarrarjeh, A., Agrawal, S., Kim, S. H. and Shahabi, C.: Geo-Spatial Multimedia Sentiment Analysis in Disasters, *IEEE Intl. Conf. on Data Science and Advanced Analytics (DSAA 2017)*, pp. 193–202 (2017).
[16] Nazer, T. H., Xue, G., Ji, Y. and Liu, H.: Intelligent Disaster Response via Social Media Analysis A Survey, *ACM SIGKDD Explorations Newsletter*, Vol. 19, No. 1, pp. 46–59 (2017).
[17] 高野明彦，吉見俊哉，三浦伸也：311 情報学 ―メディアは何をどう伝えたか，岩波書店 (2012).
[18] 伊藤　守：テレビは原発事故をどう伝えたのか，平凡社 (2012).
[19] 片山紀生，孟　洋，佐藤真一：視覚的情報の役割に着目したニュースショット分類による震災テレビ報道の分析，情報処理学会研究報告, 2012-IFAT-106(6)，pp. 1–8 (2012).
[20] 片山紀生，孟　洋，佐藤真一：マルチメディアアナリティクスによる防災・災害テレビ報道の傾向解析，電子情報通信学会技術研究報告 (PRMU) 118(513)，pp. 109–112 (2019).
[21] 片山紀生，孟　洋，佐藤真一：防災・安全を目的とする記憶補完支援へのテレビアーカイブの応用可能性，電子情報通信学会技術研究報告 (PRMU) 119(481)，pp. 187–189 (2020).
[22] 片山紀生，孟　洋，佐藤真一：テレビアーカイブを用いたアナリティクスのための関連ニュースショット検出，情報処理学会研究報告 2021-IFAT-142(6)，pp. 1–6 (2021).
[23] Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. and Buyya, R.: Big Data computing and clouds: Trends and future directions, *Journal of Parallel and Distributed Computing*, Vol. 79-80, pp. 3–15 (2015).
[24] 井手一郎，木下智義，高橋友和，孟　洋，片山紀生，佐藤真一，村瀬　洋：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).
[25] 安藤慎太郎，藤原弘将：テレビ録画とその字幕を利用した大規模日本語音声コーパスの構築，情報処理学会研究報告 (SLP), 2020-SLP-134(8)，pp. 1–7 (2020).
[26] Kudo, T., Yamamoto, K. and Matsumoto, Y.: Applying Conditional Random Fields to Japanese Morphological Analysis, *Proc. 2004 Conf. on Empirical Methods in Natural Language Processing*, pp. 230–237 (2004).
[27] 中西恒夫，吉村賢治，乙武北斗，田辺利文，古庄裕貴，西浦洋一，浅野雅樹：形態素パターンマッチャ morfgrep とそのソフトウェア開発における応用，電子情報通信学会技術研究報告 (SS), SS2019-60，pp. 113–118 (2019).
[28] 中西恒夫，吉村賢治，乙武北斗，田辺利文，古庄裕貴，西浦洋一：morfawk: 形態素パターンマッチング／処理言語，電子情報通信学会技術研究報告 (SS), SS2020-4，pp. 1–6 (2020).

ⓒ 2021 Information Processing Society of Japan