平成 26 年度　卒業論文
2015 年 2 月 3 日

ニュースデータから決める
2014 年の流行語

法政大学　理工学部　経営システム工学科
経営数理工学研究室
11X4009　井出健太
指導教員　五島洋行　教授

学科名	経営システム工	学籍番号	11X4009
申請者 指　名		井出　健太	
指導教員 氏　名		五島　洋行	

論文要旨

論文題目	ニュースデータから決める 2014 年の流行語

流行語とはその年の世相を言葉で表す指標の一つである. 一般的に㈱ユーキャンらによって決まる新語・流行語大賞の年間大賞を受賞した単語がその年の流行語となる. 年間大賞が発表される年末には, 多くのメディアや人々の間で今年の流行語予想が話題になる. これらように流行語は人々の間で広く関心のある話題である. しかし, 流行語, 流行語予想に関する研究, 論文は少ない. これより, 流行語を題材とした研究は流行, 流行語を明らかにしていくという視点で有用性があると考えられる.

本研究の目的は, 2014 年の流行語をニュースデータより機械的に決定する事である. これは, ㈱ユーキャンら決める年間大賞では最も流行した言葉を知ることが出来ないという不満, ニュースデータより機械的に流行語を決められるのではないかと言う疑問に答えるものである.

本研究では使用したニュースデータには, 更新頻度が高く, 文字数の制限が無い事, 多くの配信サイトがある事が挙げられる. これは, 流行語の定義より, 流行語の予想に適したデータだと考えられる.

研究の結果, 2014 年の流行語は「理化学研究所」に決まった. 流行語, 流行語候補リストに選ばれた単語は考察の結果, 2014 年の世相を表す話題に関連する単語を決めることが出来たと考えられる. これより, 本研究は 2014 年の流行語, 世相を表す言葉を選ぶことに成功し, ニュースデータよりは流行語を決めるのに適したデータである事, 本研究で使用した分析方法は流行語を決められることが分かった.

目次

第 1 章はじめに

本論文では, ニュースデータから決める 2014 年の流行語を提案する. これによって, 従来の㈱ユーキャンと㈱自由国民社によって決められていた新語・流行語大賞の受賞単語によらず, 2014 年に流行った言葉を知ることが出来る.

本章では流行語, 流行語予想に関する背景, 本研究の目的を説明する.

1.1. 研究背景

流行語とはその年の世相を言葉で表す指標の一つである. 一般的に㈱自由国民社と㈱ユーキャンによって年末に新語・流行語大賞として決める言葉がその年の流行語となる. その年の流行語が決まると私達の身の回りで話題にあがることは多い. そのため, 毎年年末の流行語大賞が決められる時期には, TV やラジオなど様々なメディアで今年の新語・流行語大賞に選ばれる言葉が予測されている. そして, その予想はメディア以外にも, 多くの人々の間でも話題になる. これらように, 流行語は人々の間で広く関心のある話題である.

しかし, 流行語, 流行語予想に関する研究, 論文は[1]によると少ないことがわかる. 文献[1]によると, これは流行語という言葉の定義の難しさによるものだと考えられている. その難しさの例として, 流行語は新語の一種であるのか否かがある. 仮に, 新語の一種とする場合, 2013 年の新語・流行語大賞の年間大賞を受賞した言葉[2]である「今でしょ」,「おもてなし」は流行語ではなくなる. 他の定義の難しさを挙げると, 流行語を文字としてとらえるのか, 文としてとらえるのかなどがある.

これらから, 流行語を題材とした研究は「流行語」または「流行」を明らかにしていくという視点で有用性があると考えられる.

そもそも流行語とは, デジタル大辞泉によると「ある時期, 多くの人々の間で盛んに使われる語や言い回し. はやりことば.」という意味である. 1984 年から㈱自由国民社と㈱ユーキャンにより新語・流行語大賞として決められている. ただし, [1]によると 1984 年よりも昔 1868 年明治元年から, 流行語のように認知されている言葉の存在を確認することができる. 毎年の年末には新語・流行語大賞に選ばれた単語が発表される. 加えて, 流行語の審査の過程でトップ 10 に入った単語や, 流行語候補となった言葉 50 個も発表される. ただし, 今のようにトップテンと年間大賞を選ぶようになったのは第 11 回からであり, それまでは, 新語部門, 流行語部門, 表現部門をつくり, それぞれの賞を受賞する単語を選出していた. また, 新語・流行語大賞の年間大賞ときくと受賞する言葉は一つのみと考えられるが, 実際には複数の言葉が受賞する年もある. この結

果，複数の言葉が年間大賞に選ばれた年は，最も流行した言葉を知る事は出来ない．例として 2013 年に新語・流行語大賞の年間大賞を受賞した言葉は四つあった．

そして，近年はネットワークの発達，IT 技術の急速な発達のおかげで多様なで大量の情報がデータとして保存され，ネットワークを通じ，大量に収集できるようになっている．その例として，Web ニュースがある．元来，新聞や TV のニュース番組で世の中の出来事など，ニュースは報道されてきた．しかし，現代においてはスマートフォンやタブレット端末の普及などによって，いつでも，どこでも，どんな端末からでもインターネットにアクセスできるようになった．その結果，日経電子版のように，日経新聞が電子化され，紙媒体である新聞を持たずに新聞を読むことが可能になった．ちなみに，このニュースの電子化の動きは日経新聞に限らず，朝日新聞や毎日新聞などの新聞社，NHK や Yahoo!など新聞社以外の企業も自社のホームページで Web ニュースを配信している．また，電子化されたニュースの特徴として更新頻度が高いこと，文字数の制限が無いこと，多くの配信サイトがあることが挙げられる．つまり，世の中の事件や出来事を起きた瞬間に報道でき，新聞のよりも具体的な報道が可能となった．そして，多くの配信サイトがあることから，複数のサイトが配信する同一の事件の報道を比較する事で，事件を様々な目線より俯瞰することが出来る．これらの特徴より，このニュースデータは，「ある時期，多くの人々の間で盛んに使われる語」と定義される流行語の予測に適したデータだと考えられる．上記のとおり，Web ニュースの報道には時差が少ない為，配信日時より言葉が使われた日時，期間を得ることが出来る．そして Web ニュースには文字数の制限が無いこと，多くの配信サイトがある事から，複数の配信サイトが配信するニュースを使用する事で，大量の具体的なテキストデータを得ることが出来，結果的にサイトごとの偏りのない流行語を決めることが出来ると考えられる．これらの事より，流行語をニュースデータより決めることが出来るのではないかと考えられる．

1.2. 研究目的

本研究の目的は，2014 年の流行語をニュースデータより機械的に決定することである．本研究を始めた動機は下記の二つである．

- ㈱ユーキャンと㈱自由国民社が決める新語・流行語大賞の年間大賞では最も流行した言葉を知ることが出来ないという不満
- 流行語を勘などにより感覚的に決めるのではなく，ニュースデータを使用することで，機械的に流行語を決められるのではないかと言う疑問

この二つの不満と疑問に答えるために，本研究はニュースデータより 2014 年に流行した言葉を決めるものである．

第 2 章　先行研究

第 2 章では，本研究に関わる流行語，バースト解析についての先行研究を紹介し，先行研究の特徴，課題点を挙げる．

2.1　流行語に関する研究

流行語予想に関する研究には[3]などがある．

文献[3]では新聞社が発行する新聞から，流行語を発見している．[3]では，新聞データを形態素解析した際に，未知語に分類される単語に，文書中の単語を重みづける評価尺度の一種である TF-IDF 法を用いて流行語を発見している．

つまり，[3]では流行語を新語の一種とみなしている．そして，新聞データから新語を抽出し，TF-IDF 法を用いて新語を評価し，上位に選ばれた単語を流行語と決めている．

文献[3]の課題点として，新語の流行語のみしか決めることが出来ない点，流行語になった新語が実際に流行語として認知されているのか評価していない点が挙げられる．その根拠として，㈱自由国民社と㈱ユーキャンが 2013 年に発表した新語・流行語大賞の年間大賞を受賞した単語[2]では，「今でしょ」，「倍返し」など，流行語になる以前から存在した単語が流行語に選ばれている．

2.2　バースト度に関する研究

バースト度に関する研究には[4]，[5]，[6]などがある．

バースト度とは, 時系列データにおいてある事象の頻度の急激な増減をした箇所を検出する評価尺度である．バースト度の検出アルゴリズムとして Kleinberg のバースト検知アルゴリズム[4]がある．そして，[4]のアルゴリズムを用いた研究には，[5]，[6]があり，ともに Twitter やニュースデータなどの時系列データからバースト度を求めることで，バーストキーワードや話題語など，ホットトピックの推定を行っている．[5]ではニュースデータに対しバースト度が高いキーワードをもとに, ニュース記事をクラスタリングすることで分類分けをしている．そして，分類分けされたニュース記事の集合を読むことで, キーワードが決められた背景を簡易に把握できることを提案している．[6]では, Twitter に書き込まれるツイートに対してバースト度とピアソン相関係数を用いる事で，ツイートから 24 時間周期と一週間周期で特徴的な話題語が存在する事を示して

いる．以上の[5]，[6]の研究結果からわかるとおり，バースト度を用いる事でツイート，ニュースの中で話題になっている言葉を抽出することが出来ると考えられる．

第 3 章　関連知識 － ニュースデータ －

本研究を理解するために必要な情報として，本研究で使用するニュースデータについて説明する．

本研究で用いるニュースデータとは，インターネット上で新聞社や Web ポータルサイトが配信する Web ニュースのことである．

本研究では 2010 年から 2014 年の 5 年間の中で配信された Web ニュース 1,638,646 件を使用している．ただし，ニュースデータ収集の際に，収集している Web ニュースサイトが配信方法を変える事があり，配信方法の変化に対応するまでの期間，ニュースデータを取得することが出来ない．そのため，今回使用するニュースデータには欠損がある．

本研究で分析する際に使用するニュースデータの配信サイト，ニュースデータの型を表 1，表 2 に記す．

本研究で使用したニュースデータの配信サイトは朝日新聞 DIGITAL，毎日新聞，Yomiuri ONLINE，NHK ニュース，Yahoo!ニュース，goo ニュースである．

本研究で使用したニュースデータの型は，ニュースサイト名，日付，タイトル，本文である．

表 1.ニュースサイト一覧

朝日新聞 DIGITAL	毎日新聞	Yomiuri ONLINE	NHK ニュース	Yahoo!ニュース	goo ニュース

表 2.ニュースデータの型

ニュースサイト名	日付	タイトル	本文
NHK ニュース	2014-09-29 13:45:22	両陛下 １２月に広島被災地訪問へ	天皇皇后両陛下は，土砂災害で大きな被害を受けた広島市をことし１２月に訪れ，被災した人たちを見舞われる見通しになりました．先月，広島市で起きた豪雨による土砂災害では７４人が亡くなり，宮内庁は，ことし１１月下旬を視野に両陛下の被災地訪問を検討していましたが，その後の現地の状況や今後の日程なども考慮した結果，両陛下は，１２月に広島市を訪問される見通しになりました．具体的には，１２月３日からの１泊２日の日程で検討が進められているということで，両陛下は，避難生活を送る被災者を見舞うほか，広島市の平和公園にも足を運び，原爆慰霊碑の前に花を供えて，犠牲者の霊を慰められる見込みいうことです.

.

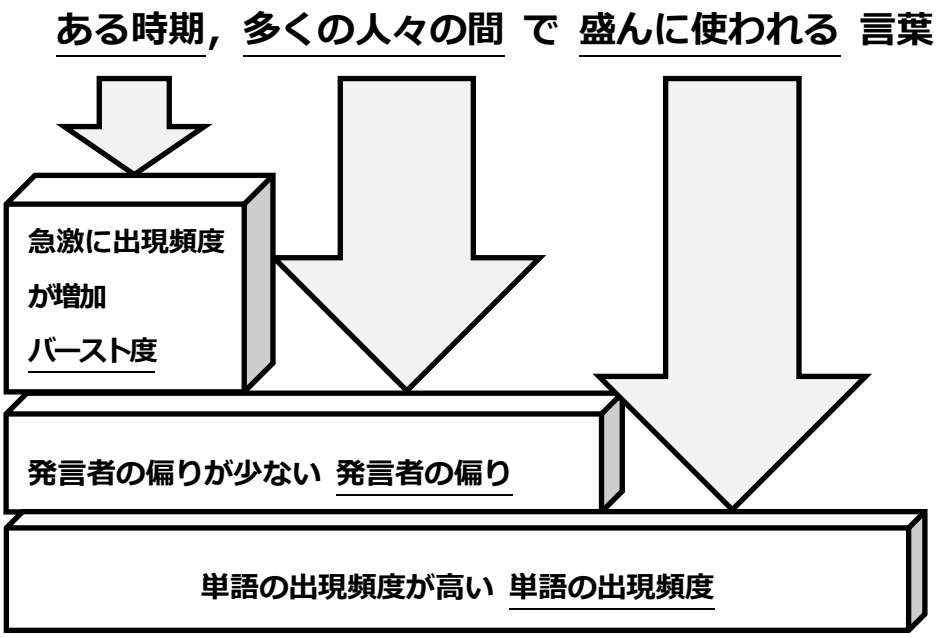

図 1.流行語決定に用いる評価尺度

第 4 章　分析方法

本研究では，ニュースデータから流行語を決める為に四つの評価尺度を用いる．使用した四つの評価尺度のうち三つは，流行語の意味である「ある時期，多くの人々の間で盛んに使われる語」から決めている．流行語の意味から決めた評価尺度は図 1 より，単語の出現頻度，発言者の偏り，バースト度が適していることがわかる．これらに加え，単語の品詞を流行語決定の評価尺度として用いた．

本章では，本研究で使用する四つの評価尺度，単語の出現頻度，単語の品詞，発言者の偏り，単語のバースト度の説明を行う．

4.1　単語の品詞

単語の品詞とは，文の中での意味によって分類する単語の区分けである．本研究で行った年間大賞を受賞した単語，ニュースデータ本文が含む単語の品詞分けには MeCab を用い, MeCab で単語を品詞分類する際に使用する辞書は, IPA 品詞体系を用いている．

本研究では，流行語候補リストに入れる単語の条件として単語の品詞を使用した．主に流行語候補リストの条件として単語の品詞を用いている．これにより，流行語にはなりにくい品詞を持つ単語を流行語候補リストから取り除けると考えられる．

流行語候補リストに入れる条件として，品詞が名詞の単語，品詞の種類が名詞の中でナイ形容詞語幹，引用文字列，数，接続詞的，動詞非自立的，特殊を持たない単語とした．

上記の条件の決定には，過去に新語・流行語大賞の年間大賞を受賞した単語の品詞を用いた．㈱自由国民社，㈱ユーキャンにより 1984 年から 2014 年に発表された流行語に形態素解析を行い品詞分類した結果を表 3 に載せる．表 3 より，年間大賞を受賞した単語の中で名詞を含む単語が 52%と過半数を超えており，品詞が名詞の単語は年間大賞を受賞しやすいと考えられる．これから，本研究で決める流行語は，品詞が名詞の単語のみを対象にする．

次に，新語・流行語大賞の年間大賞を受賞した単語で品詞が名詞の単語を，品詞の種類で分類を行った．表 4 がその結果である．表 4 より年間大賞受賞単語で，品詞が名詞の単語では品詞の種類が一般の単語が過半数を占め，品詞の種類でナイ形容詞語幹，引用文字列，数，接続詞的，動詞非自立的，特殊を持つ単語は年間大賞を受賞してない事がわかる，これより，品詞の種類が一般を持つ単語が名詞の中でも年間大賞を受賞しやすいとわかる．しかし，品詞の種類を一般の単語のみを対象にして流行語を決めるのは，対象とする単語の範囲を狭めると考える．品詞の種類における条件は年間大賞を受賞した単語が持つ品詞の種類のみを対象にする．つまり，品詞の種類でナイ形容詞語幹，引用文字列，数，接続詞的，動詞非自立的，特殊を持つ単語は流行語を決める際には用いない．

表 3. 流行語大賞の年間大賞を受賞した単語の品詞 頻度

品詞	名詞	助詞	助動詞	記号	動詞	接続詞	形容詞	副詞	感動詞	フィラー	/
頻度	85	20	19	12	10	7	4	2	1	1	1
割合	52%	12%	12%	7%	6%	4%	2%	1%	1%	1%	1%

表 4. 新語・流行語大賞の年間大賞受賞単語で使われた名詞の品詞の種類の頻度

品詞の種類	一般	固有名詞	サ変接続	接尾	非自立	副詞可能	形容動詞語幹	代名詞
頻度	43	16	12	7	3	2	1	1
割合	51%	19%	14%	8%	4%	2%	1%	1%

4.2 単語の出現頻度

単語の出現頻度とは，Web ニュースの本文に出現した単語の頻度のことである．単語の出現頻度からは，出現した日付近にその単語にかかわる事柄が話題になった事がわかる．本研究では流行語候補リストの作成，発言者の偏りの判定，単語のバースト度の検

出で用いられる．

通年頻出単語リストに入れる単語の出現頻度の条件は，その単語が毎日出現している事とする．また，本研究で用いるニュースデータは，5 年分 6 サイト 365 日分ある．これらから，通年頻出単語リストに入れる単語は 5 年間の出現頻度の合計が 10950 以上の単語とする．この値は以下の数式(1)より導いた．

$$5 \times 6 \times 365 = 10950 \tag{1}$$

この結果，表 5 からわかるとおり，「日」，「こと」，「若手」などよく見かけられる単語を除くことができる．

表 5.通年頻出単語リストに含まれる単語．上位下位五個

単語	出現頻度	単語	出現頻度
日	2,702,768	若手	10,954
こと	1,638,921	約束	10,956
人	1,337,886	致死	10,957
年	947,218	先行	10,958
市	859,940	黒田	10,958

4.3 発言者の偏り

発言者の偏りとは，各ニュースサイトにおける単語の出現頻度の偏りである．この評価尺度を使用する事で，各発言者の発言から，キーワードの発言者の偏りを確認することが出来る．本研究では，ニュースサイトを発言者とみなし，また各ニュースサイトが配信するニュース本文を発言者の発言とみなす．

発言者の偏りと似た評価尺度を用いた研究で[7]がある．文献[7]では四つの指標を用い，新たに辞書に載せるべき単語を決めている．四つの指標とは単語の出現頻度，出現頻度による順位，単語の使用者数，使用者数の順位である．ここで発言者の偏りと似た指標として単語の使用者数が使われている．

発言者の偏りの作成には，六つの Web ニュースサイトが配信するニュース本文に含まれる単語，一年間の出現頻度を用いる．そしてキーワードが含まれる 6 サイトごとの出現頻度の標準偏差が発言者の偏りとなる．

発言者の偏りのイメージを図 2 にあらわした. 図 2 の吹き出しに書かれている数字はキーワードの発言回数である．この説明では，仮に上段のキーワードはじぇじぇじぇ，中段は今でしょ，下段ではおもてなしと置く．

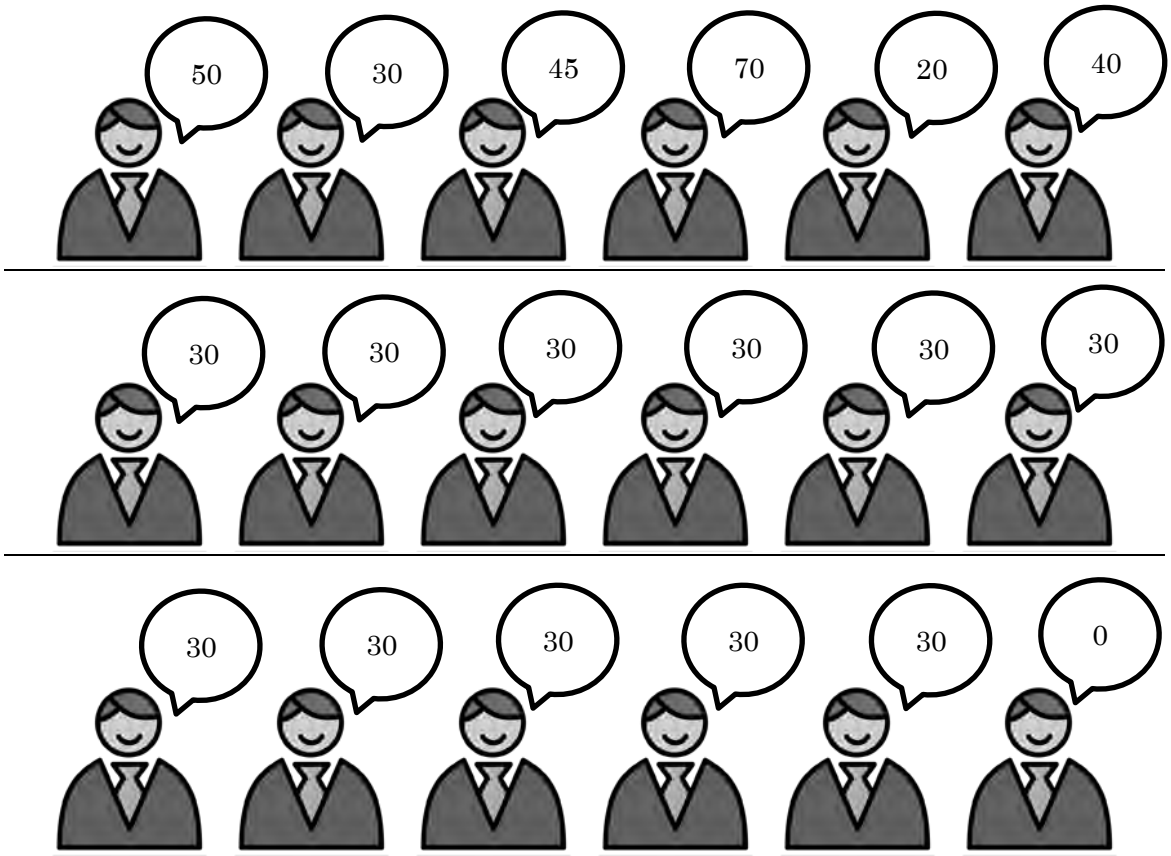

図 2.発言者の偏りイメージ

図 2 の上段，6 人それぞれの発言回数は 50,30,45,70,20,40 である．この値からじぇじぇじぇの使用者の偏りは 45.75 となる．中段での発言回数は，6 人とも 30 回である．この場合，発言者の偏りは 0 となる．下段では 5 人の発言回数が 30 回だが，一人だけ 0 回である．この場合，発言者の偏りは 66.66 となるが，1 人が発言してないので，この場合はこの単語を流行語候補リストから取り除く．三つの発言者の偏りはを順位付けすると，今でしょ，じぇじぇじぇとなる．

4.4 単語のバースト度

単語のバースト度とは，ある期間で単語の出現頻度の急激な増減を捉える評価尺度である．本研究ではバースト度の検出には Jon Kleinberg が提案したバースト検出アルゴリズム[4]を用いる．Kleinberg のバースト検出アルゴリズムとは，時系列データを用いることで，単位時間内におけるイベントの急激な増加を検出する方法の一つである．この手法を用いる事で，Web ニュースの本文内で，単位時間内に急激に増加した単語を検出できる．

Kleinberg のバースト解析には二つのバースト検出アルゴリズム，連続型と離散型とがある．本研究では離散型のバースト検出アルゴリズムを用いた．Web ニュース本文におけるキーワードのバースト度はコスト関数(2)によって求めることができる．そして，離散型のバースト検出アルゴリズムでは Web ニュース内の単語の出現確率(3)は二項分布に従うと仮定しており， コスト関数では i が 0 を取る値をバースト状態，1 を取る値を非バースト状態とおいている．これより，コスト関数が正の値を取る場合キーワードがバースト状態になり，負の値を取る場合は非バースト状態となる．本研究では一日ごとのバースト度を検出している．また，本研究ではバースト度の評価にはバースト状態のバースト度を使用している．つまり，正の値のバースト度のみを使用する．

本研究で行う検出期間$t_1,\cdots\cdots,t_n$において，一日ごとの Web ニュース集合$WN_1,\cdots\cdots,WN_n$におけるバースト度の検出方法を考える．検出期間における全ニュース件数を$D=\sum_{t=1}^{n} d_t$とおき，キーワードを含むニュース件数をR$=\sum_{t=1}^{n} r_t$とおく．次に検出期間内の，キーワードを含むニュース件数の期待値を$P_0 = R/D$とおき，P_1はP_0にパラメータsをかけた値$P_1 = P_0 s$ (4)とする．本研究ではsを 2 とする．

$$\sum_{t=t_1}^{t_2}(\sigma(0,r_t,d_t)-\sigma(1,r_t,d_t)) \quad (2)$$

$$\sigma(i,r_t,d_t)=\log(\, d_t C r_t P_0^{r_t}\,(i-P_i)^{d_t-r_t}) \quad (3)$$

$$P_1 = P_0 s \quad (4)$$

離散型バースト検出アルゴリズムには，以下の特徴がある．

1.単語によるバースト度の比較が可能.

本研究では流行語候補リストに含まれている単語どうしを比べ，流行語を決める．これを可能にするには，単語によるバースト度との比較が出来る離散型が適切と考えられる．

2.一部分だけサンプリングしたデータを用いてバースト検知を行える．

第 3 章の関連知識でも説明したとおり，本研究で扱うニュースデータには一部欠損がある．つまり全データを扱う事は出来ない．これより，全データを取得できなくともバースト検知を行える離散型が適切と考えられる．

これらより，流行語の決定とニュースデータを使用する本研究では，離散型のバースト検出アルゴリズムが適していると考えられる．

第 5 章　分析手順

本章では，ニュースデータから流行語を決める手順を記す．ニュースデータから流行語を決めるおおまかな手順を図 3 にあらわした．また，各リストの単語数は年間大賞の選定方法にならい，流行語候補リストを作成するための単語数を 50，流行語候補リストに含まれる単語の数は 10 とした．また，流行語の単位は，意味が分かる最小の単位である単語とした．新語・流行語大賞の選出を図 4 にあらわした．

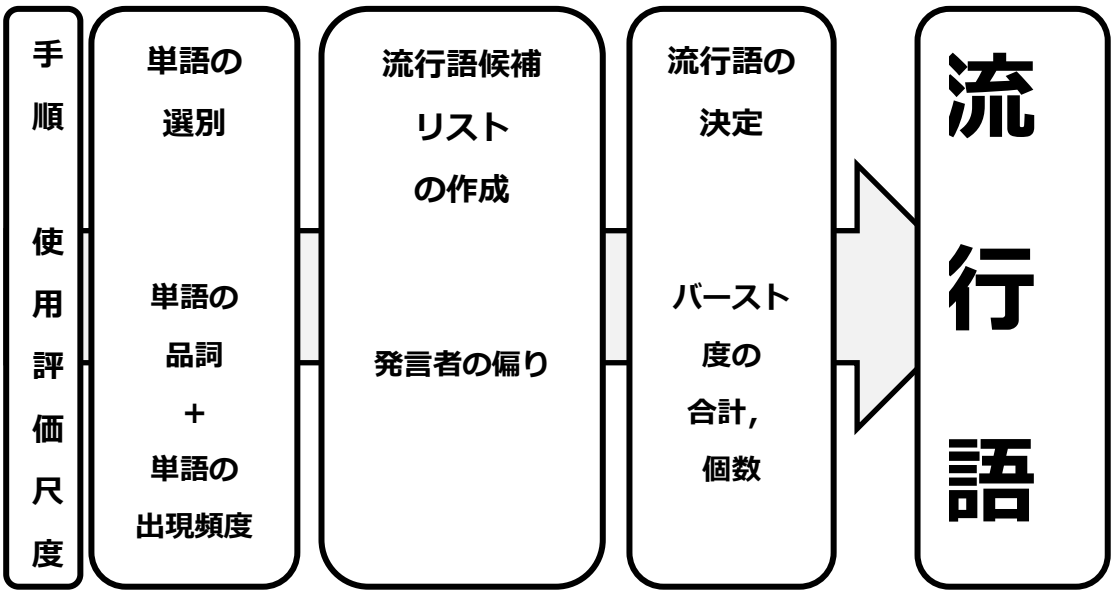

図 3.本研究での流行語決定のフローチャート

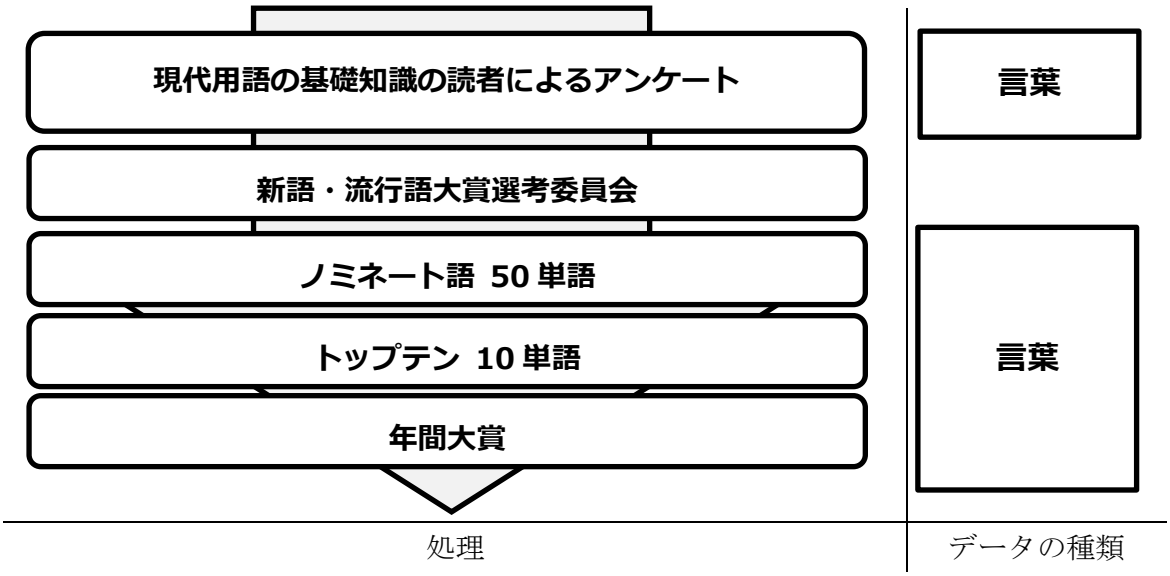

図 4.新語・流行語大賞の選出フローチャート

5.1 流行語候補リストの作成

まず，流行語候補を決める為に，ニュースデータの本文から流行語候補リストを作成する．流行語候補リストとは，バースト度を使用して流行語を決める為の単語リストの事である．流行語候補リストを作るにあたって使ったニュースデータは 2010 年から 2014 年間，5 年分の Web ニュースの本文である．Web ニュースのタイトルではなく本文を使用した理由は，単語数の多さと内容の詳しさによって社会の流行を表しやすいと考えられるからである．

流行語候補リストを作るには，単語の品詞，出現頻度，発言者の偏り，これら三つの評価尺度を用いる．

流行語候補リストを作るために，まず通年頻出単語リストを作る．通年頻出単語リストとは，「人」，「国」，「県」など出現頻度が一年中高い，普段の生活でよく見かけられ流行しているとは考えられない単語のリストである．これを作る事で出現頻度が高いが流行しているとは言えない単語を流行語候補リストに入れることを防ぐことが出来る．通年頻出単語リストの作成手順を図 5 にあらわした．

通年頻出単語リストの作成には 2010 年から 2014 年の 5 年間で配信された本文とタイトルで，「流行語」という単語を含まない本文を用いた．これは本研究では 5 年間 12 ヶ月のニュースデータ使用する為，ニュースデータには新語・流行語大賞が発表される前後には流行語を予想するニュース，発表された流行語に関わるニュースなどが含まれる．これらのニュースを使用して流行語を決めては，新語・流行語大賞を受賞した単語の影響を受けた流行語が出来ると考えられるためである．この本文に形態素解析を行い，品詞が名詞の単語，名詞の種類で「ナイ形容詞語幹」，「引用文字列」，「数」，「接続詞的」，「動詞非自立的」，「特殊」の単語は除いた．これはこれまで㈱ユーキャン，㈱自由国民社らが発表した流行語部門・金賞，年間大賞の単語を形態素解析し，単語単位で品詞分類した結果，品詞が名詞の単語が上位を占め，上記の品詞の種類を持つ単語が無かった為である．つまり，品詞が「名詞」の単語が流行語になりやすいが，名詞のうち，上記の種類を持つ単語は流行語になりにくいと考えられる．

次に，品詞分類して得られた単語を，出現頻度を基準に通年頻出単語リストに入れる．これにより「人」，「国」，「県」など出現頻度が一年中高く，流行しているとは言えない単語を除くことができる．切り捨てる基準となる出現頻度は「一年中毎日，ニュースに含まれている」という基準とし，10950 とするこれは本研究で使用するニュースデータが 5 年間，6 サイト分であり，日数で表すと数式(1)より，10950 となるからである．出現頻度がこの値以上の単語を，通年頻出単語リストに入れる．

次に流行語候補リストを作成する．流行語候補リストの作成手順を図 6 にあらわした．まず日付が 2014 年内のニュースで，タイトルと本文に流行語が含まれないニュースの

本文を取得する．その本文に形態素解析を行い，文章を単語に分ける．そして単語の品詞が名詞であり，名詞の種類が「ナイ形容詞語幹」,「引用文字列」,「数」,「接続詞的」,「動詞非自立的」,「特殊」ではない単語を取得する．次に，出現頻度降順に単語を並び替え，上位 50 単語未満を除く．残った 50 単語の発言者の偏りを作成し，発言者の偏り降順で並び替え，上位 10 単語未満を除く．その結果，残った単語 10 個が流行語候補リストとなる．

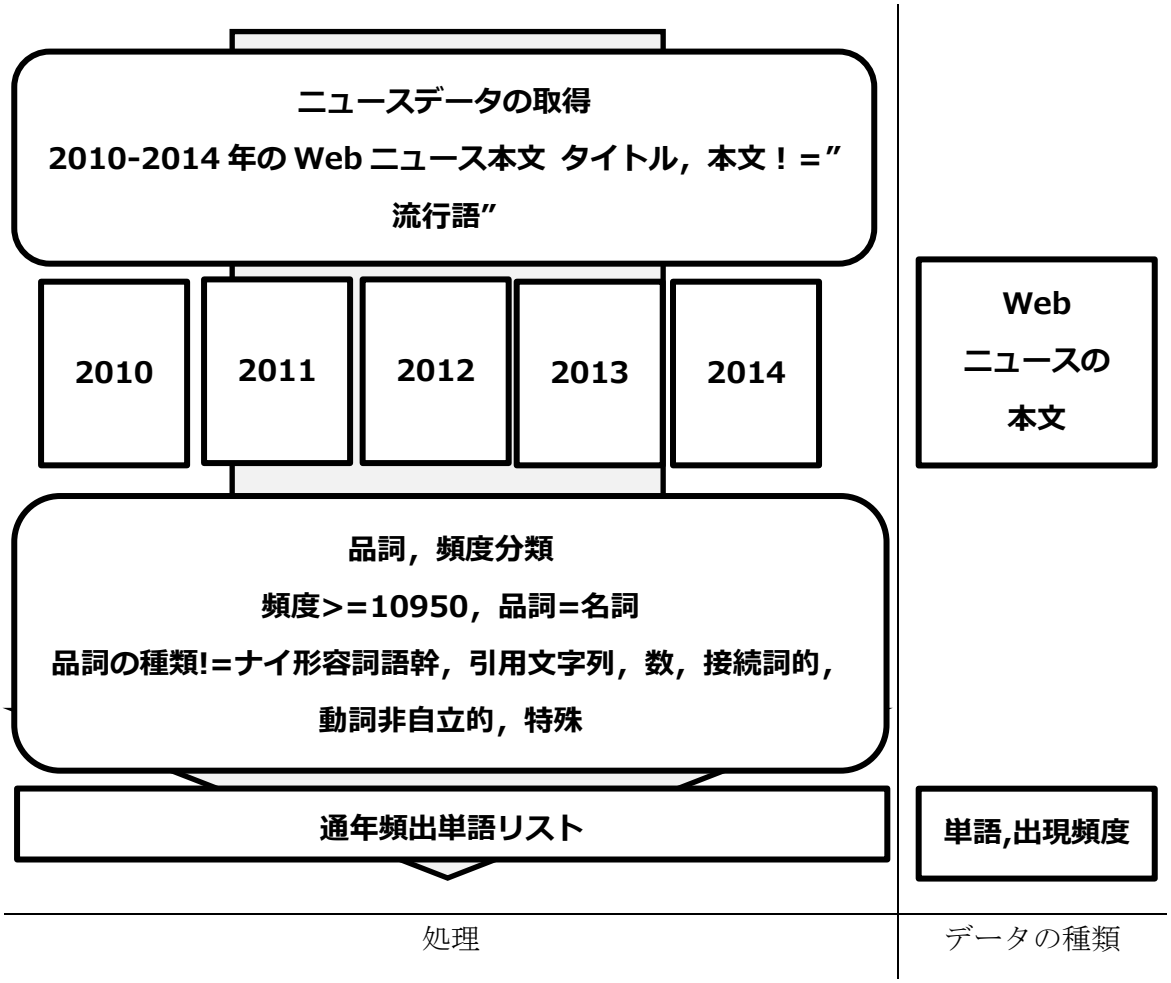

図 5.通年頻出単語リストの作成フローチャート

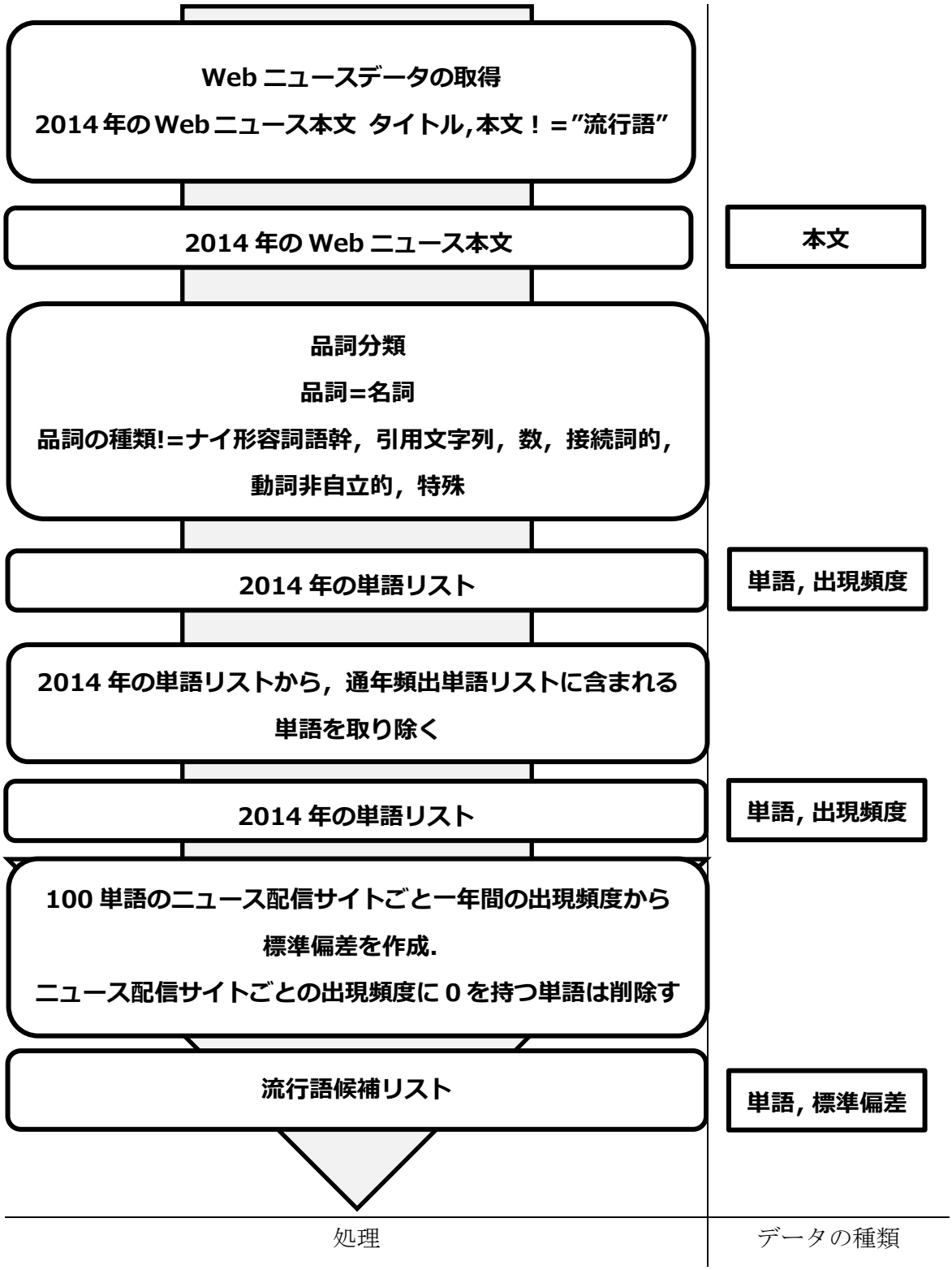

図 6.流行語候補リストの作成フローチャート

5.2 単語の評価尺度の作成

5.1 章で作成した流行語候補リストに含まれる単語の 365 日分，一日ごとのバースト度を作成する．バースト度は正の値を取る場合はその単語はバースト状態であり，負の値を取る場合は非バースト状態となる．本研究では，流行語候補リストがバースト状態であるときのバースト度のみを使用する．

5.3 単語の評価

本研究では，流行語候補リストに含む単語からバースト度を用いて流行語を決める．流行語を決定する手順を図 7 にあらわした．バースト度の評価方法は，単語ごと一年間のバースト状態のバースト度の合計と，バースト状態であるバースト度の出現個数を用いる．つまり，正の値をとるバースト度の出現個数が一年間で 80%以上の単語を流行語候補リストから取り除く．次に，流行語候補リストのうち，バースト度の合計が最も大きい単語を流行語と決める．

バースト状態のバースト度の合計値に関しては，[5]，[6]の研究でバースト度を評価する際に，バースト度の最大値を使用していることから正の値を使用している．また，本研究では一年間の流行語を決める為，一年間 365 日分のバースト度を使用して，単語を評価するべきと考えられるため，一年分のバースト度の合計を用いる．

次に，バースト度の個数に関しては流行語の意味より必要と考えられる．流行語の意味は「ある時期，多くの人々の間で盛んに使われる語や言い回し．」である．本研究では一年間のニュースデータを用いる為，仮にある単語の正の値のバースト度の出現個数が 100%，365 個である場合，一年間 365 日の一部では無く全部になってしまい，流行語の「ある時期」という意味とは一致していないと考えられる．

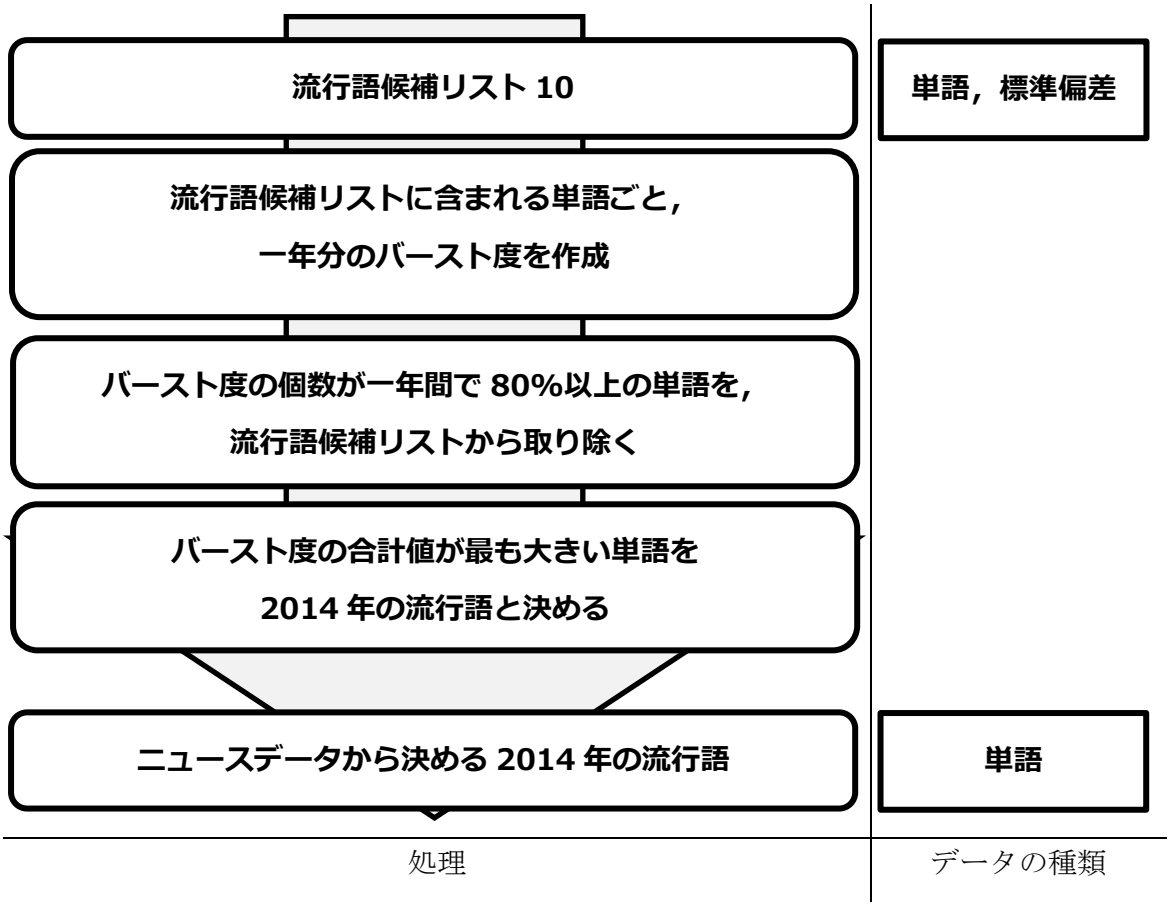
流行語候補リスト 10
単語，標準偏差
流行語候補リストに含まれる単語ごと，
一年分のバースト度を作成
バースト度の個数が一年間で 80%以上の単語を，
流行語候補リストから取り除く
バースト度の合計値が最も大きい単語を
2014 年の流行語と決める
ニュースデータから決める 2014 年の流行語
単語
処理
データの種類

図 7.流行語の決定

第 6 章　分析結果 - 2014 年の流行語 -

研究の結果，2014 年の流行語は「理化学研究所」に決定した．また，流行語候補リストに挙がった単語は表 6 のとおりである．流行語に決めた単語は太字にし，表を色付けしている. また, 表 7 は流行語候補リストの前身である 2014 年の単語リストである.

表 6. 2014 年の流行語候補リスト

順位	単語	順位	単語
1	**理化学研究所**	6	慰安
2	エボラ出血熱	7	舛添
3	ＳＴＡＰ	8	ガザ
4	クリミア	9	ドラッグ
5	小保	10	覚醒剤

表 7. 2014 年の単語リスト

単語				
ＳＴＡＰ	都知事	占拠	ＳＰ	甲信
小保	河野	空爆	パラリンピック	投球
理化学研究所	=	石破	リーダー	婦
ドラッグ	錦織	停戦	国境	特集
舛添	浅田	山頂	火山	竜
クリミア	安佐南	マレーシア	積雪	毎日新聞
覚醒剤	アルゼンチン	孤立	全て	露
ガザ	解釈	沈没	日本テレビ	将
慰安	御嶽山	全米	鶴	健
エボラ出血熱	画像	銀メダル	–	

第 7 章　検証

本章では研究の結果決めた流行語，流行語候補リストを新語・流行語大賞受賞単語らとの比較をおこない，新語・流行語大賞と決めた流行語が一致，含まれるか判定する．次に，流行語候補リストに含まれる単語が実際にどのようにして Web ニュースで取り上げられたのか説明する．

7.1　新語・流行語大賞受賞単語との比較

まず，2014 年に新語・流行語大賞を受賞した単語，ノミネート語は[2]より表 8，表 9 のとおりである．年間大賞を受賞した単語は太字にし，表を色付けしている．表 8 を見てわかるとおり，本研究で決めた流行語は新語・流行語大賞の年間大賞を受賞した言葉には一致しなかった．しかし，流行語候補リストに選ばれた STAP はノミネート語に選ばれた「STAP 細胞はあります」という言葉に含まれ，エボラ出血熱はノミネート単語の「エボラ出血熱」と一致した．これより，本研究で決めた流行語，流行語候補リストは新語・流行語大賞と一部共通する流行をとらえているが，全体的には異なる流行をとらえていると考えられる．

7.2　ニュース本文での流行語の取り上げられ方

本章では，本研究で決めた流行語，候補リストらが，実際に流行していたのか，Web ニュースの本文を参照する事で検証する．検証した結果，流行語候補リストに含まれる単語に関連する出来事をまとめたものが表 10 である．関連する出来事に記入した話題は，その年の流行語を含むニュース本文において，話題の割合が多かったものを記入した．

流行語候補リストに含まれる「理化学研究所」，「STAP」，「小保」は STAP 細胞論文ねつ造を表していると考えられる．そして，流行語候補リストに含まれる単語に関連する出来事は全体的にスポーツやエンタメではなく，社会や政治に関する内容が占めている事がわかる．また，国内の出来事が流行語候補リストの多くを占めたが，「クリミア」，「ガザ」と「エボラ出血熱」は海外の出来事である．これより，国内のニュースでも，国外の出来事を表す単語が流行する事がわかる．

表 8. 2014 年の間に新語・流行語大賞を受賞した単語一覧

ダメよ～ダメダメ	**集団的自衛権**
ありのままで	レジェンド
壁ドン	カープ女子
ごきげんよう	危険ドラッグ
マタハラ	妖怪ウォッチ

表 9.2014 年，新語・流行語大賞のノミネート語

STAP 細胞はあります	リベンジポルノ	壁ドン
まさ土	絶景	壊憲記念日
デング熱	ゆづ	雨傘革命
2025 年問題	塩対応	塩レモン
アイス・バケツ・チャレンジ	こじらせ女子	エボラ出血熱
マタハラ	号泣会見	輝く女性
レリゴー	集団的自衛権	バックビルディング
ごきげんよう	積極的平和主義	トリクルダウン
J 婚	カープ女子	ダメよ～ダメダメ
タモロス	ハーフハーフ	危険ドラッグ
家事ハラ	マイルドヤンキー	女装子
ありのままで	JK ビジネス	セクハラやじ
こぴっと	レジェンド	限定容認
リトル本田	妖怪ウォッチ	勝てない相手はもういない
ゴーストライター	マウンティング（女子)	ワンオペ
消滅可能性都市	ミドリムシ	イスラム国
昼顔	ビットコイン	

表 10.流行語候補リストの Web ニュースでの取り上げられ方

単語	話題
理化学研究所	STAP細胞論文ねつ造
エボラ出血熱	西アフリカで流行した感染病
ＳＴＡＰ	STAP細胞論文ねつ造
クリミア	クリミア半島の帰属をめぐってロシアとウクライナの間に生じた政治危機
小保	STAP細胞論文ねつ造
慰安	日韓の間にあるとされている従軍安婦問題，朝日新聞による従軍慰安婦問題の誤報道
舛添	新東京都知事の名字
ガザ	イスラエルとイスラム原理主義組織ハマスの戦闘，イスラエルによるガザ地区への軍事行動
ドラッグ	危険ドラッグの広がり
覚醒剤	歌手ASKA氏が覚せい剤取締法違の罪に問われた件、危険ドラッグの広がり、覚醒剤の密輸

第 8 章　おわりに

本研究は㈱ユーキャン，㈱自由国民社により新語・流行語大賞として決められる流行語を，ニュースデータを使用する事で機械的に流行語を決める事を目的とし分析を行ってきた．

先行研究では，新聞を使用し新語を対象に流行語を発見している．先行研究の課題点として，新語のみしか流行語を決めることができない点，決めた流行語が実際に流行しているのか検証していない点が挙げられる．したがって，本研究では新語以外の単語も流行語の対象とし，決めた流行語を新語・流行語大賞の年間大賞を受賞した単語と比較することで，実際に流行していたのか検証を行う．

研究の結果，本研究で決めた流行語，流行語候補リストは新語・流行語大賞とは一致はしなかった．しかし，検証の結果より 2014 年の世相を表す話題に関連する単語を決めることが出来ていると考えられる．加えて，流行語候補リストに含まれる単語に関連する出来事より，ニュースデータはその年に起きた出来事を広く報道していることが分かる．これより，ニュースデータはその年の世相を表す単語，流行語を予想するのに適しているデータであることがわかる．そして，ニュースデータを使用する前提をおくと，本研究で使用した分析方法は世相を表す言葉，流行語を探し出せることがわかった．

参考文献

[1] 吉田光浩，”「流行語」研究の諸問題 (上) 附・六種資料対照近現代流行語年表”，大妻女子大学紀要 文系，pp.145--168(1999)

[2] 自由国民社：”ユーキャン新語・流行語大賞”，http://singo.jiyu.co.jp/

[3] 山川侑吾，馬青：”新聞データからの「流行語」自動発見 ―「コンピュータ流行語大賞」を目指して―”，言語処理学会年次大会発表論文集，言語処理学会第 11 回全国大会(2005)

[4] Jon Kleinberg，”Bursty and Hierarchical Structure in Streams“，the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining，pp.5--18 (2002)

[5] 高橋佑介，宇津呂武仁，吉岡真治，“ニュースにおけるバーストキーワードの話題への集約”, 情報処理学会研究報告, The 3rd Forum on Data Engineering and Information Management (2011)

[6] 佐々木謙太朗，田村一樹，吉川大弘，古橋武，“Twitter における話題語の抽出と周期に基づく分類”，言語処理学会 第 19 回年次大会，pp.806--809(2013)

[7] 荒巻英治，増川佐知子，宮部真衣，森田瑞樹，保田祥：”頻出語ではなく使用者が多い語が自然な日本語である”， 言語処理学会 第 19 回年次大会 発表論文集，pp.544--547（2013）