APIR Discussion Paper Series No.47
2020/11

テキストデータを利用した新しい景況感指標の開発と応用(上)
― 入門編：基礎的概念と分析手法の解説 ―

生田祐介　木下祐輔　松林洋一

大阪産業大学経営学部商学科
一般財団法人アジア太平研究所（APIR）
神戸大学大学院経済学研究科

APIR 一般財団法人 アジア太平洋研究所
ASIA PACIFIC INSTITUTE OF RESEARCH

テキストデータを利用した新しい景況感指標の開発と応用(上)
― 入門編：基礎的概念と分析手法の解説 ―

生田　祐介*，木下　祐輔†，松林　洋一‡

【要旨】

本稿では，テキストデータを用いて数量的考察を行う際の基礎的概念を紹介し，同データを用いて経済動向を分析する際の基本的な手続きについて，入門レベルの解説を行っています。テキストデータとは，新聞や書籍に書かれるような文字列で表されている情報を数値化したものであり，経済のみならず様々な分野において新たなデータソースとして注目を集めています。テキストデータの収集，整理，解析は総称して「テキストマイニング」と呼ばれていますが，その内容は高度に専門的なため，初心者には理解しにくい面もあります。そこで本稿では，まずテキストデータの特徴，テキストマイニングの基本的手続きをできるだけ平易かつ丁寧に解説していきます。続いてテキストマイニングを用いることは，経済動向を理解する上で，有用な手法であることを説明していきます。

一国の経済動向を把握するにあたり，これまでは数量的なマクロデータが用いられてきました。しかし，こうしたデータは，家計や企業といったミクロ経済主体の段階まで情報を分解することができず，収集から公開まで時間を要するという課題があります。そこで，情報量が豊富であり，かつ高頻度で発生するテキストデータを用いることが，マクロ経済のより精緻な情勢判断と予測において有効であると考えます。そこで本稿では内閣府の「景気ウォッチャー調査」のテキストデータを用いた簡単な分析結果を示し，テキストデータを用いた経済分析の一例を紹介します。併せて経済分析においてテキストマイニングを用いることの難しさについても検討します。

* 大阪産業大学経営学部商学科, ikuta@dis.osaka-sandai.ac.jp（照会先）
† 一般財団法人アジア太平洋研究所（APIR）, kinoshita-y@apir.or.jp
‡ 神戸大学大学院経済学研究科, myoichi@econ.kobe-u.ac.jp

本稿は，APIR における研究プロジェクト「テキストデータを利用した新しい景況感指標と開発と応用」の一環であり，その成果の一部を初心者向けに平易に解説したものです（本稿と応用編である下巻の２分冊から構成されており，いずれも入門的な解説を意図しています）。

本稿の作成に当たっては，宮原秀夫 APIR 所長，猪木武徳先生，本多佑三先生，稲田義久先生，青山秀明先生，池田裕一先生，岩野宏 APIR 代表理事，中山明 APIR 総括調査役から有益なコメントをいただきました。ここに記して感謝申し上げます。なお本稿における誤謬はすべて筆者の責任です。

はじめに

昨今，世界経済を取り巻く不確実性は日に日に増しており，足元の経済動向を正確に把握することが極めて重要となっています。これまで，一国の経済の良し悪しを把握するためには，GDP や物価を始めとするマクロデータを用いることが一般的でした。しかし，数量的に表現されるマクロデータには 2 つの課題があります。第一に，個々の経済主体の活動を調査して，それらを集計してデータとして公表するまでには，ある程度の期間を必要とするため，高頻度で発生しないということです。第二に，数字の背後にいる個々の経済主体が何を考えているかというミクロの要因が捨象されてしまっているということです。

他方で，情報技術の急速な進展により，国内外の経済活動において生成される大規模なデータ（ビッグデータ）が様々な形で利用可能になり始めています。筆者の知る限り，ビッグデータとは何であるかを示す共通の定義は見当たりませんが，平成 24 年版の情報通信白書（総務省，2012）では，多量性，多種性，リアルタイム性の 3 点を兼ね備えたデータであるとされています[4]。こうしたビッグデータを用いることで，個人の購入履歴や移動履歴をはじめとする詳細な行動パターンを把握でき，それをもって事業に役立つ知見を導出することが求められています。きわめて豊富な情報を内包しているビッグデータの活用は，マクロ経済動向に関しても，より精緻な情勢判断と予測のために大いに役立つと考えられます。

ビッグデータには，質的にも量的にも様々な種類がありますが，本稿ではマクロ経済の動向を迅速かつ詳細に把握するという目的のため，「テキストデータ」に注目します。テキストデータとは，数字ではなく，新聞や書籍に書かれるような文字列で表された自然言語データのことです。テキストデータに注目する理由は，人々が経済の変化にさらされているとき，それと関連する「言葉」が各種メディアや調査（サーベイ）において，多く登場すると考えるからです。数量データは，家計や企業が何らかの行動をとった結果であり，そうした行動の兆しとなる心情や意志は言葉で表されます。したがって，経済活動の足元と先行きをリアルタイムかつ低コストで捉える手法を開発するにあたり，テキストデータを利用することに意義があると考えています。

本稿である上巻は，別途下巻と併せて，テキストデータの経済分析への利用可能性について解説する内容となっています。この上巻では，テキストデータに馴染みのない読者に向けて，テキストデータがあらゆるデータの中でどのような位置にあるかということから始め，単語の頻度を利用した単純な経済予測方法まで含んでいます。下巻では，単語の順序を考慮した複雑な解析をすることで，まるで人間のように景況感を推定するという研究を報告し

---

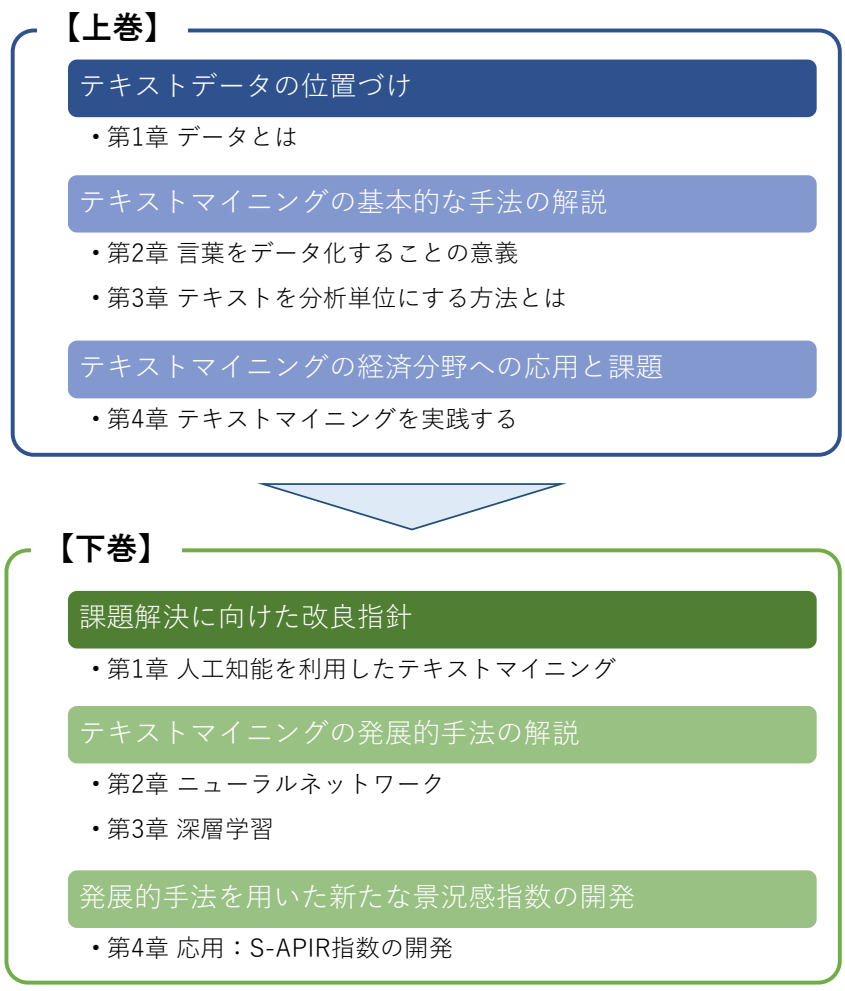
[4] 同白書によると，ビッグデータは，それ自体を導出する観点から特徴付けることができると述べています。その特徴は 3 点あり，「高解像（事象を構成する個々の要素に分解し，把握・対応することを可能とするデータ）」，「高頻度（リアルタイムデータ等，取得・生成頻度の時間的な解像度が高いデータ）」，「多様性（各種センサーからのデータ等、非構造なものも含む多種多様なデータ）」です。これらの特徴を満たすために，結果的に「多量（ビッグ）」のデータが必要となります。

他方で，2000 年台半ばにゲノム解析や天文学の分野において，大量の情報があふれ出てきたことを経験したため，ビッグデータという言葉が誕生したとも言われています（照井 2016）。

ます。具体的に，人工知能の一種である深層学習（ニューラルネットワークという人間の脳神経回路を模したモデルを構築し，コンピュータに機械学習させること）を，テキストデータの解析に用います。上巻では単語の頻度解析のみ扱うため比較的単純な理論が登場しますが，下巻では単語の順序解析を扱うため，より高度な理論が登場するのだと思ってください。

本稿の構成は，以下の通りです。まず，第１章では，テキストデータの位置づけを知るために，一般的にデータとはどういうものであるのか説明します。第２章では，「言葉をデータとして扱う」とはどういうことなのか解説します。第３章では，文字列を単語に区切る「形態素解析」という方法について学びます。最後の第４章では，内閣府の「景気ウォッチャー調査」のテキストデータを用いて，簡単な分析を行った結果を紹介します。景気動向を表す代表的な経済統計と我々が作成した指標との間には差異が生じました。その理由についても，経済用語の複雑性という視点から解説するとともに，改良の方向性を述べます。

図表 全体構成



第１章 データとは

本稿全体を通して読者に伝えたいことは，テキストデータを経済分析にどのように利用していくか，ということです。そのために，テキストデータとはどういうモノなのか知る必要があります。そして，テキストデータが，量的にも質的にも様々な種類のデータがある中でどのような位置付けにあるのか，なぜテキストデータを使って分析できるのか，という仕組みを知ることが必要です。こうした理由から，第１章では，テキストデータについて理解する事前の知識として，一般的にデータにはどういう種類があり，それぞれがどのような特徴を持っているのか知ることから始めます。本章を読むことで，世の中に存在する「データ」について，頭の中で大まかに整理することができるでしょう。

本章は，５節から構成されています。第１節では，データの分類について概要を説明します。その中でも定性的データについて第２節で，そして定量的データについては第３節で，それぞれの特徴を説明します。第４節では，データが持つ４つの尺度水準について説明します。最後の第５節では，データにおける加工の有無という観点から，経済分析で用いられる定量的データをマクロデータとミクロデータに分けて説明します。また，そうした定量的データが持つ課題を克服するためのテキストデータの利用可能性にも触れます。

1-1. データの分類

経済活動を把握するためには，データの利活用が何より重要であることは言うまでもありません。しかし，データにはどのような種類があり，それぞれがどのような関係であるかを理解している人は少ないのではないでしょうか。データとは，記録であり，そこから物事の全体を推論する基礎となる情報であるため，統計資料とも呼ばれます。日本統計学会によると，データには，以下の三つの性質があります[5]。

> 1）ある調査や実験・観察の結果を表す情報の表現（数値や分類項目など），またはそれらから作成されたものである。
> 2）得られた結果には，何らかの意味で不確実性がともなう。
> 3）それらの結果の背景には，現実的または仮想的な集団・メカニズムを想定することができる。

データはこうした性質を持つことから，データを単に眺めるだけでなく，データの発生元となるグループに対して，その全貌を推論し，データを発生させる仕組みについて理解を深めることも重要です。そうした目的のために使われる道具が統計学です。

[5] 日本統計学会「改訂版　日本統計学会公式認定 統計検定３級対応 データの分析」(2020 年)，第１章 １節データの種類，２項

図表 1-1：尺度水準とデータ例

	尺度水準	例
定性的データ	名義	・ 都道府県コード(1=北海道, 2=青森, 3=岩手県,…,47=沖縄県) ・ 人名（第 x 代目の首相） ・ 学籍番号（番号と名前） ・ 血液型（1=A,2=O,3=B,4=AB） ・ 戸籍上の性別（0=男，1＝女）
	順序	・ 景気判断(1=良い, 2=やや良い, 3=どちらでもない,4=やや悪い, 5=悪い) ・ 成績の順位（S,A,B,C,F） ・ 学歴（大学院，大学，専門学校，高校，中学） ・ デザインの好み（Mac, Windows),（赤，白，黒，青),（濃味，薄味）
定量的データ	間隔	・ 日付（和暦：令和元年,2 年),（西暦：2019 年,2020 年） ・ 知能指数（IQ:120,100,90), ・ 摂氏温度（100℃, 37.5℃, 0℃，マイナス 40℃）
	比率	・ 絶対温度（摂氏マイナス 273℃＝絶対 0℃） ・ 時間の経過（10 秒間, 30 分間, 1 時間） ・ 金額（100 円, 100 ドル） ・ 身長（175.3cm, 163cm）

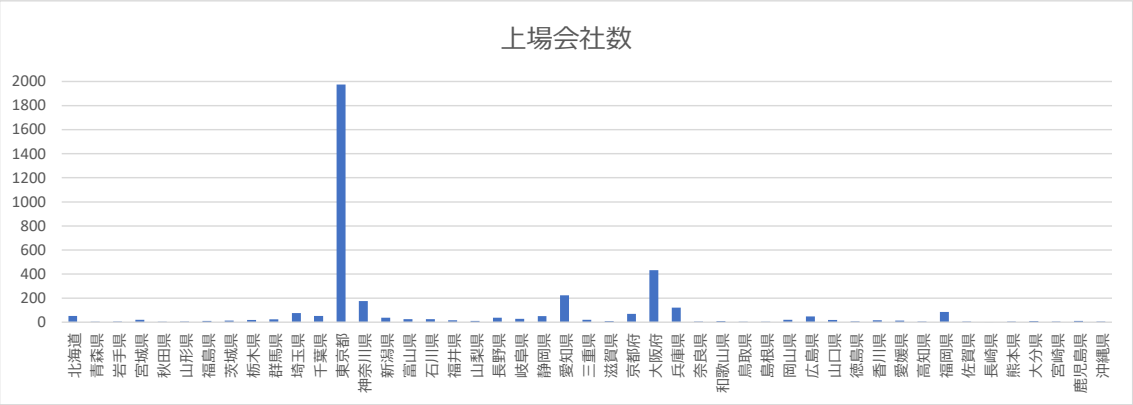
(出所)筆者作成

名義尺度と順序尺度は定性的データ，間隔尺度と比率尺度は定量的データに含まれる。テキストデータは名義尺度と順序尺度を持つ定性的データであるが，間隔尺度と比率尺度を持つ定量的データに変換することで，統計解析が可能となる。

一般的にデータは，①演算の可否の観点と，②尺度水準の観点から分類でき，これらの観点は互いに関係しています。演算可能なデータを定量的データとよび，数値の大小に意味があります。演算不可能なデータを定性的データとよび，カテゴリーを区別することに意味があります。尺度水準には，名義，順序，間隔，比例の 4 つがあります。そして，定性的データは名義尺度と順序尺度を持ち，定量的データは間隔尺度と比例尺度を持ちます。**図表 1-1** は，これらの分類について例を挙げて整理したものです。尺度水準の詳細については，後で詳しく説明します。

さて，テキストデータは，このようにデータを分類したとき，どこに位置づけることができるのでしょうか。先取りして説明しましょう。テキストデータは，本来，名義尺度を持つ定性的データとして位置づけられます。しかし，自然言語処理や機械学習の分野における研究の発展にともない，近年，テキストを構成する単語を，まるで言葉の地図上における住所のように，数値で表現できるようになっています。これは，「単語を演算する」ことが可能であることを意味します。これにより，テキストデータを比率尺度をもった定量的データとして扱うことが可能となり，統計解析ができるデータとして利用できるようになりました。

図表 1-2：上場会社数



（出所）「上場企業サーチ.com」のデータを元に筆者作成(https://xn--vckya7nx51ik9ay55a3l3a.com/analyses/number_of_companies)

これにより，分析の幅が大きく広がりました。これらのことを意識しながら，以降の内容を読んでいただくと，テキストデータに対する理解がしやすくなると思います。

1-2. 定性的データ

定性的データは，「カテゴリーの集合の中から 1 つのカテゴリーを選ぶようなデータ」のことです。例えば，本社所在地，自動車の色，景気判断などです。これらは，カテゴリーを区別するためのデータであり，計算するためではありません。ただし，区別を容易に行うために数値を使うことがあります。例えば，上場企業の本社所在地を都道府県で部類する場合，「1.北海道, 2.青森県,…,47,沖縄県」というように，1 から 47 の数値を割り振ることで，カテゴリーを区別することができます[6]。そして，各カテゴリーに該当する企業数を度数として集計することができます。

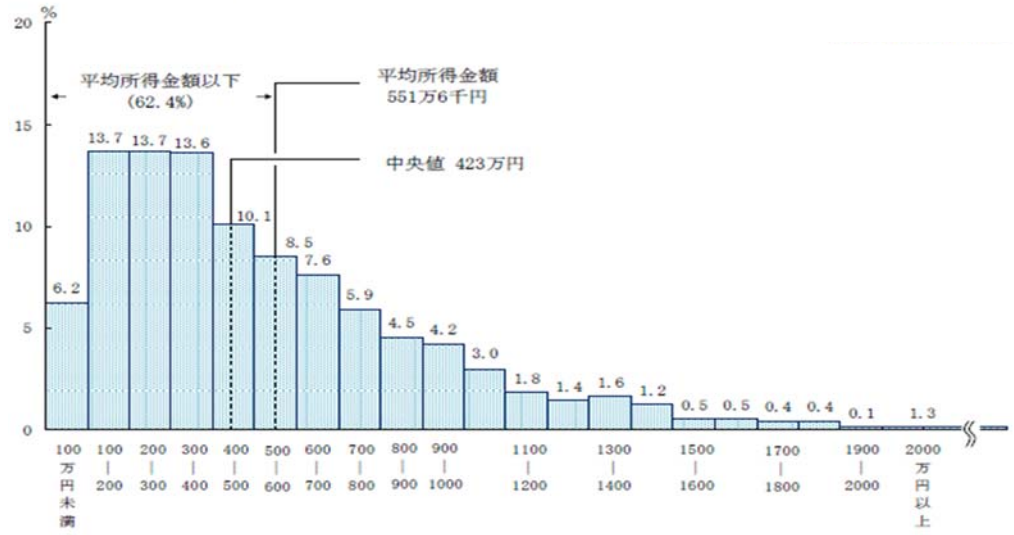
図表 1-2 は，横軸にカテゴリー，縦軸に度数をとることで，47 都道府県別の上場会社数（度数）の分布を表したものです。棒が 1 番目に高いのは「東京」で，2 番目に高いのは「大阪」です。このように，度数の分布（ちらばり）を棒グラフで表すことで，どこにデータが集中しているのか視覚的に分かります。

1-3. 定量的データ

定量的データは，「数値で記録された演算可能なデータ」のことであり，その数値の性質によって離散変数と連続変数に分けることができます。**変数**とは，「観察対象がとりうる状

[6] この数値は，総務省が設定した「都道府県コード」に基づきます。

図表 1-3：所得金額階級別世帯数の相対度数分布（単位%）



(出所)厚生労働省「国民生活基礎調査」2018（平成 30）年より引用

態を数値で表したもの」です。例えば，価格のデータが，“高い”あるいは“低い”でなく，“60 円”や“63 円”といった様々な値をとる場合，それは変数であるといえます。**離散変数**は，「世帯人数，購入回数，特許取得件数のように，小数点以下で表すことに意味がなく，区切りのある値（整数値）で表された変数」のことです。変数に切れ目があるため，定性的データと同じように，度数の分布は棒グラフで表すことができます。**連続変数**は，「体重，所得金額，金利のように，小数点以下の値（実数値）まで，途切れることなく細かく測ることができる変数」のことです[7]。変数に切れ目がないため，通常棒グラフ同士の隙間を空けないヒストグラムで分布を表します。

図表 1-3 は，厚生労働省「国民生活基礎調査」から引用した，所得金額ごとの世帯数の分布です。横軸に所得金額の各階級を，縦軸に相対度数をとっています。**階級**とは，「変数（所得金額）の取りうる値が多いため，変数の範囲をグループ分けしたもの」です。**相対度数**とは，「全体の度数に占める各階級の度数」を表します。したがって各階級の相対度数を足し合わせると 1 になります（単位がパーセントの場合は 100）。

1-4. データの尺度

定性的データは，計測方法の観点から，名義尺度と順序尺度に分けることができます。そして，定量的データは，計測方法の観点から，間隔尺度と比率尺度に分けることができます。

[7] 桁数が大きい、または、とりうる値の多い数値であるほど、連続データとして考えることができる。例えば、所得金額は 3.45 万円と表すことができる。

(1) 名義尺度

名義尺度とは，「カテゴリー同士を単に区別するための数値」であるため，演算することができません。先ほど説明したように，本社所在地が「1.北海道」，「2.青森県」，または「3.岩手県」であることの間に順序はありません[8]。つまり，名義尺度では平均に意味はありません。これら 3 つの所在地を表す数値の算術平均をとり，$\frac{1+2+3}{3}=2$ となるため，平均は「2.青森県」であるとは言えないのです。そのため，分布の中心を表す代表値は，平均値ではなく**最頻値（モード）**といい，「度数が最も多いデータ」が使われます。例えば，上場会社数の分布の中で，東京で 1975 件と最も多く観察されれば，「13.東京」が最頻値となります。

また，名義尺度では，平均値が意味をなさないため，ちらばりの尺度の一つである**偏差**を求めることにも意味はありません。偏差は「各観測値から平均値を引いたもの」であり，

$$\text{偏差}=\text{観測値}-\text{平均値}$$

と定義されます。名義尺度では偏差を実質的に計算できないので，統計解析で用いられる分散や標準偏差も求めることができないと考えたらよいでしょう。

(2) 順序尺度

順序尺度は，景気判断のように，「カテゴリー間の大小関係を表すことができる数値」です。景気判断おける「1.良い」は「2.やや良い」よりも順位が高いことが分かります。しかし，これらは大小関係があっても演算できないため，定量的データとは異なるのです。大小関係の幅については，身長や通貨のように単位がはっきりしていないこともあり，一般的にカテゴリー間で均等であると見なされます。例えば，「1.良い」，「2.やや良い」「3.どちらともいえない」，「4.やや悪い」，「5.悪い」という 5 つのカテゴリーには，4 つの間隔があります。このとき，「1.良い」と「2.やや良い」の間隔は，「2.やや良い」と「3.どちらともいえない」の間隔と等しいと考えます。

順序尺度においても，分布の代表値を求める際に注意が必要です。いま，**図表 1-4** のように 5 段階の景気判断について 10 人からデータが得られたとします。各判断を区別する数値を使い算術平均をとると，$\frac{(1\times2)+(2\times2)+(3\times1)+(4\times3)+(5\times2)}{10}=3.1$ です。この結果をどう解釈したらよいでしょうか。これは，「3.どちらともいえない」に近似できるかもしれません。しかし，翌月に集計したデータから同様に平均値を求めた結果，2.5 ならば，どう判断できるでしょうか。それは「2.やや良い」と「3.どちらともいえない」の中間と考えるべきなのかどうか，判断が難しいところです。

[8] 名義尺度を持ち，性質の有無を 0 か 1 の二値で表す変数をダミー変数と言います。

図表 1-4：景気判断の度数分布表

判断	1.良い	2.やや良い	3.どちらともいえない	4.やや悪い	5.悪い
度数	2	2	1	3	2

(出所)筆者作成

こうした景気判断の例のように，順序尺度では，回答のカテゴリーは主観的かつ心理的なものです。また，好みの味の順位やマラソンの到着順位のように，数値の大小が意味を持つにもかかわらず，その間隔に厳密な単位がありません。このため，たとえ平均値を出したとしても，その解釈には十分な注意が必要です。

そのため，順序尺度では分布の中心を表す代表値として，平均値の代わりに最頻値あるいは**中央値（メディアン）**を使う方が良いかもしれません。中央値とは，「データを順位の低い順から並べたときに，真ん中に位置する値」のことです[9]。この場合のデータ数は 10 のため，中央値は，下から 5 番目に位置する「3.どちらともいえない」と，6 番目に位置する「4.やや悪い」の間となり，$\frac{3+4}{2} = 3.5$ となります。

(3) 間隔尺度

間隔尺度は，暦年や温度のように，「評価の基準点を持たず（0 に意味がない），間隔だけが意味を持ち，加減を計算することができる数値」のことです[10]。例えば，暦年について説明すると，グレゴリウス暦で 2000 年から 2020 年の間には 20 年分の期間があるだけで，「2000 年から 1 %の年が経過した」とは言いません。なぜなら時間には，その大きさを測るための基準点がないからです。また，間隔尺度では，異なる基準で相互に数値を変換可能です。例えば，グレゴリウス暦で 2020 年から 2018 年を引いたものが，和暦の令和 2 年です。このことより，数字の 0 には，何か状態を表す意味はないということが分かります。

次に，定性的データの順序尺度であっても，そこに評価値を与えることで，実際には定量的データである間隔尺度として扱われることが多いということを説明します。これを理解するための良い例は，大学生の成績です。大学生の成績を，順序が高い順に「S,A,B,C,F」の 5 段階で評価するならば，これは順序尺度です。しかし，「S」という評価を，試験の得点が「100 点から 90 点」（1 点刻み）の区間の学生に与えるならば，評価のもとになる得点は間隔尺度です。

[9] データ数 n が奇数の場合、$\frac{n+1}{2}$ 番目のデータを中央値とする。データ数 n が偶数の場合，$\frac{n}{2}$ 番目のデータと $\frac{n+1}{2}$ 番目のデータの平均値を中央値とする。

[10] 摂氏温度は 1 気圧の下で，水の凝固点を 0，水の沸点を 100 と見なして，その間を 100 等分したものであるため間隔尺度である。また，時間（日付）は間隔尺度であるが，時間の経過（例えば，30 分間の 2 倍は 60 分間）は比例尺度である。

図表 1-5：評点付き景気判断の度数分布表

判断 （評点）	1.良い (2)	2.やや良い (1)	3.どちらともいえない (0)	4.やや悪い (-1)	5.悪い (-2)
度数	2	2	1	3	2

(出所)筆者作成

もう一つ例を挙げると，**図表 1-5** は，先ほどの景気判断に，大小関係が分かる評点を割り当てた分布表です。平均を計算すると，$\frac{(2\times2)+(1\times2)+(0\times1)+(-1\times3)+(-2\times2)}{10}=-0.1$ となります。先ほどの**図表 1-4** の順序尺度の下では，平均値が 3.1 でした。どちらの平均値も，それらが意味することは，「3.どちらともいえない」から，わずかに下回った評価であるということです。つまり，各平均値によって示されるデータの中心は近いように見えます。しかし，**図表 1-5** においては加減の演算ができる間隔尺度のため，平均値として意味が高まるのです。このため，間隔尺度では，分布の中心を表す代表値として，最頻値，中央値，そして平均値が使われています。

(4) 比率尺度

比率尺度は，価格や体重のように，「評価の基準点を持ち（0 に意味がある），間隔だけなく比率にも意味があり，加減乗除の計算ができる数値」のことです[11]。例えば，価格について説明すると，60 円から 63 円へ 3 円高まることを，「60 円から 5%上昇した」と言います。なぜなら価格自体，その大きさを測ることができるからです。比率尺度は，物理学における基本量（fundamental scale）と組立量（derived scale）に分けることができます。経済学の文脈では，基本量を“価格”とすると，“需要関数”という変換機能を通じて“需要量”という組立量を求めることができます[12]。そして，価格が 0 であることは，財の取引に貨幣を必要としないという観点から，その財は無価値であることを意味します。

比率尺度を用いた分布の代表値として，最頻値，中央値，算術平均のほか，幾何平均が使われます。幾何平均は，変化率の平均値を求めるときに使います。**図表 1-6** は，ある商品価格の変化率（前年同月比）のデータです。6 カ月間での平均値は，算術平均でなく，幾何平均を用いて $\sqrt[6]{1.021\times1.034\times1.058\times1.106\times1.128\times1.098}=\sqrt[6]{1.530}=1.073$ となるので，価格変化率の平均値は 7.3%となります。

[11] 絶対温度は，0 度が物理的な意味を持つため比例尺度であると考えられます。つまり，摂氏温度-273℃の下では原子や分子の熱運動が止まることから，それを絶対温度における 0 度としているのです。
[12] 需要関数を用いることで，需要量変化率の価格変化率に対する比率も計算できます。この比率は，“需要の価格弾力性”といい，組立尺度です。

図表 1-6：価格変化率

月次	4 月	5 月	6 月	7 月	8 月	9 月
価格変化率	2.1%	3.4%	5.8%	10.6%	12.8%	9.8%

(出所)筆者作成

比率尺度は，加減乗除の演算ができるため，統計解析において必要な多くの情報を与えてくれます。まずは，データのちらばり具合を知るために用いられる，**分散**を求めることができます。分散は，「各観測値の偏差を二乗した値の平均を求めたもの」です。すなわち，

$$分散 = \frac{\sum \left(観測値 - 平均値 \right)^2}{観測数}$$

です。式の中の $\sum$（ギリシア文字でシグマ)は足し算するという意味です。分散は，元のデータを二乗しているため，例えば，データが“服の丈（cm）”であるのに，分散を求めると，“服の面積（cm^2）”へと単位が変換されてしまいます。測定単位を元に戻すために，**標準偏差**とよばれる「分散の平方根」をとります。すなわち，

$$標準偏差 = \sqrt{分散}$$

です。

二つのグループ間で平均値が異なっている場合，各グループの標準偏差を用いることで，データのばらつきを正しく比較することができます。例えば，去年は大企業 1 社当たりの年間の営業利益の平均が 26.6 億円，標準偏差が 7.5 億円であったのに，今年になると，平均が 117.5 億円，標準偏差が 23.8 億円だとします。去年と比べて今年の方が，企業間での営業利益の格差は広がっているのか知りたい場合，**変動係数**という指標が役立ちます。変動係数は，

$$変動係数 = \frac{標準偏差}{平均値}$$

と定義されます。変動係数は，「平均値に対してデータがどれだけばらついているか」を，

グループ間で相対的に比較することができます。先ほどのデータを用いると，去年の変動係数は 0.28，今年の変動係数は 0.20 であり，今年の方が企業間の利益格差が小さくなっていることが分かるのです。

(5) 尺度間の関係

図表 1-1 において、尺度水準は、上の段階へ戻るほど日常の多くの場面で成立しますが，下の段階へ進むほど日常において満たす条件が厳しくなります。青森産リンゴと長野産リンゴというデータに対しては，1 つの尺度（名義）でしか違いを測ることができません。しかし，青森産リンゴ 60 円と長野産リンゴ 63 円というデータであれば，4 つ全ての尺度（比率，間隔，順序，名義）で評価できます。したがって，尺度間での情報量の大小関係は

「 比率 ＞ 間隔 ＞ 順序 ＞ 名義 」

となります。

ところで，知りたい情報は同じであっても，質問の仕方や回答方法を変えることによって，その情報を評価する方法が変わってしまいます。例えば，「リンゴの価格」という情報を知りたい場合，「先月は 60 円で，今月は 63 円」と回答してもらう場合と，「先月よりも今月の値段は高い」と回答してもらう場合とでは，比率尺度から順序尺度へと変わってしまいます。順序尺度と比べて比率尺度のデータの方が，情報量は多くなるため精密な分析ができますが，データを提供する側の負担は大きくなってしまうというジレンマがあります。

図表 1-7 は，これまで説明したデータの尺度をまとめたものです。各行の内容は，上から順に累積しています。すなわち，名義尺度が，最も少ない条件で成立し，有する情報量も最小です。これに対して，比率尺度は，成立するために多くの条件が必要ですが，最も多くの情報量を備えています。そして，データの種類と尺度の関係を見ると，定性的データは名義尺度と順序尺度，定量的データは間隔尺度と比率尺度という特徴を有します[13]。

[13] こうしたデータの尺度は，経済学における消費者行動の理論と関係しています。例えば，経済学の多くの分野で想定されている序数的効用は，消費から得られる効用以前の選好順序を重視しています。このため，序数的効用を前提とした消費者行動の推定では，順序尺度を有するデータ（例えば，色の好み）を使い，順序ロジットモデルで回帰分析することが適当であると言えるでしょう。

図表 1-7：尺度によるデータの分類

	尺度水準	尺度の特徴	意味ある比較（可能な変換）	分布の指標	例
定性的データ	名義	$A = B,$ $A \neq B.$	同一性判定 （一対一変換）	最頻値	都道府県コード 人名　国名 血液型　性別
	順序	$A > B,$ $A = B,$ $A < B.$	大小判定 （単調増加変換）	中央値	景気判断 成績の順位　学歴 デザインの好み
定量的データ	間隔	$(A - B) +$ $(B - C) =$ $A - C.$	加減演算 （アフィン変換）	平均値	時間 知能指数 摂氏温度
	比率	$A = kB,$ $B = lC,$ $\rightarrow A = klC.$	乗除演算 （原点から線形変換）	分散 標準偏差 変動係数	価格 絶対温度 身長　体重

（出所）筆者作成

1-5. 経済分析における定量的データとテキストデータの利用可能性

経済分析で多く使われてきたデータは，経済活動の結果が数字で表された「定量的データ」です[14]。定量的データは，マクロデータとミクロデータという名称で分けられることがあります。**マクロデータ**は，「個々の企業や消費者などが一定期間に行った経済活動の結果を一つのデータに集計したもの」であるため，**集計データ**とも言われます。例えば，全国のスーパーや百貨店の総販売額，国内の工場で生産された電子部品の総個数，国内で生産された電子部品のうち国外へ輸出された総額などを指します。**ミクロデータ**は，マクロデータとは対照的に，「集計される前の企業や消費者といった個別の経済主体についてのデータ」です。したがって，ミクロデータは個別の経済主体を調査し記録したデータであることから，**個別調査票（個票）データ**とも言われます。

マクロデータとミクロデータには，どのような特徴があるのでしょうか。マクロデータは，全体の経済動向を把握するための有益な情報源であることから，政策当局，金融機関，民間企業等で一般的に使われています。しかし，ある時点で発生した経済活動がマクロデータとして公開されるまでには，集計作業が必要なこともあり一定の時間を要します。つまり，マクロデータは認知の「即時性（リアルタイム性）」に欠けるという欠点があります[15]。

[14] 定量的データは、その数値の発生した構造が明確に定義できて、発生した数値に意味を与えることができるため、より一般的に「構造化データ」と言います。

[15]「鉱工業生産指数」の場合，10 月のデータは速報値であっても 11 月に公開されるため，1 ヵ月程度のタイムラグが発生してしまいます。

このことは，マクロ経済の“今”を知り，速やかに情勢判断を下して政策や企業経営の意思決定に反映する際の足かせとなっていると考えられます。一方で，ミクロデータは集計作業が必要でないこともあり，個々の経済的な行動が発生してから，データとして公開されるまでにそれほど多くの時間はかかりません[16]。したがって，ミクロデータはマクロデータに欠けているリアルタイム性を補完していると言えます。

ところで，ミクロデータは，個々の経済主体がとった行動の結果を数字として記録しているものですが，その行動の背景には，人が持つ“思いつき”，“ためらい”，“執着心”，“惰性”等の心理が少なからず作用しているでしょう。そうした心理面の情報を定性的データとして得ることができれば，経済を動かす源流である景況感を推定でき，それを以て経済の情勢判断に役立てることができそうです[17]。しかし，ミクロデータは行動した結果を数字で表すのみであり，心理面の情報まで持ち合わせていません。

こうした定量的データの限界を踏まえると，より深い経済分析を行うためには，まだ数量で表されていないデータを使うことも視野に入れることが必要でしょう。中でも，人々の経済に対する態度や心理について表された情報は，とても魅力的なものです。そのような魅力的な情報は，主に人々の発言内容に含まれます。つまり，人々が経済の変化にさらされているとき，それと関連する「言葉」が各種メディアや調査（サーベイ）において，数多く登場します。数量データは，家計や企業が何らかの行動をとった結果であり，そうした行動の兆しとなる心情や意志は言葉で表されます。近年，これらの情報を文字で記録された「テキストデータ」として手に入れて，分析ができるようになりました。

ここまでの内容を振り返り，テキストデータを尺度水準のどこに位置付けることができるのか改めて整理しましょう。テキストデータは，まず，名義尺度を有します（都道府県コードにおける，「1.北海道」と「2.青森」のように，カテゴリーを区別できる）。また，順序尺度も有します（景気判断における，「1.良い」「2.やや良い」のように，順序関係を区別できる）。このため，テキストデータは，定性的データに分類できることが分かります。

他方で，自然言語処理と機械学習における研究の発展により，テキストデータは，間隔尺度と比率尺度を有する定量的データとしても扱えるようになっています。これにより，テキストデータに対して統計的に解析できる範囲が広がりつつあります。この背景には，**分散表現**といい，「各単語を多次元の数値ベクトルで表現できる」ようになったことが大きく貢献しています（坪井ほか，2017）。これにより，単語を多次元空間における一つの点と捉え，点と点の距離を定義することができます。すると，例えば，

「 王様 - 男性 + 女性 = 女王 」

[16] 近年経済分野でも分析が進められている「POS（販売時点情報管理）データ」は，店頭のレジで商品が消費者に販売されたときに記録されるデータのため，リアルタイム性の高いデータであるといえます。
[17] 景況感（business sentiment）とは，経済の今，あるいは比較的直近を対象とした先行きに対する家計や企業の感情の程度（楽観的あるいは悲観的）を意味します。

というように，単語と単語を，類似度や関連性といった観点から演算できるようになるのです[18]。ただし，テキストデータは，元々名義尺度であるため，統計解析の結果については分析内容に応じて注意が必要でしょう。

本稿の残りの章では，経済分析を念頭において，テキストデータの利用方法を解説します。テキストデータは文字で記録された定性的データであるため，そのままではコンピュータで統計的に処理できません。したがって，次の第２章では，テキストデータはどのように数量化され，コンピュータで処理できる状態になるか解説します。

第２章 「言葉をデータ化する」ことの意義

第１章では，われわれが目にし，分析に使うデータには，どのようなものがあるのか概観しました。通常はデータというと，１，２，３・・・と数えることができる定量的なデータを思い浮かべます。しかし，データの種類はそれだけではありません。毎朝目を通す新聞に書かれている，誰がどこで何をやったかという数値で表されていない定性的な情報も，一工夫加えることでデータとして取り扱うことができるようになります。本章では，人が意思疎通を図る際に使う“言葉”に注目し，言葉をデータとして扱うとはどういうことなのか説明します。

2-1. テキストの定義

本節では，われわれの関心対象である「テキスト」がどういうものか定義することから始めます。「テキスト」という言葉を聞くと，何を想像するでしょうか。「教科書」を頭に思い浮かべる方が多いかもしれません。テキストとは何か，その意味を国語辞典で調べてみましょう。「テキスト」という言葉は本来英語なので，“text”を著名な英英辞典である Cambridge Dictionary で調べると，“the written words in a book, magazine, etc., not the pictures”という結果が表示されました[19]。この記述から，“text”は，図表以外の文字全般のことだと言えるでしょう[20]。同様に，日本語の「テキスト」をデジタル大辞泉で調べてみます。すると，「1.書物の本文，2.教材とする書物，3.コンピュータで扱う文字列や文章」という結果が表示されました。日本語の「テキスト」には，英語の“text”が持つ意味に加え，「教科書」の意味も含まれているようです[21]。特に注目していただきたいのが３つ目の「コンピュータで扱う文

[18] **埋め込み表現**というベクトルの次元数を縮退させる技術の発展により，単語間の演算が容易になってきています（例えば，変数を 100 次元から２次元へ）。

[19] 他に“a sentence or piece of writing from the Bible that a priest or minister reads aloud in church and talks about”というように，キリスト教の聖書に書かれた語句という意味もあります。

[20] 「教科書」を英語では“text”ではなく“textbook”といいます。

[21] 石田・金（2012）によると，テキストとは，「文字列で記述した文章・文書，文字列で記述された遺伝情報，情報処理分野のアクセス情報を記号列で記述したログ情報，音楽の音符を記号列で記述したものなど」を指すとされています。

字列や文章」です。これは後で説明するように，テキストマイニングを行うために必要なテキストデータのことを表していると考えられます。これで「テキスト」という言葉の意味を知ることができました。

次に，「テキスト」の種類について見てみたいと思います。全体像をつかむために，もう少しだけ我慢して議論にお付き合いください。ここでは，テキストを①種類と②保存状態という二つの観点で分類します。まず，①種類の観点ですが，テキストを「自然言語」と，「非自然言語」という名称で分けます。ここで「自然言語（Natural Language）」とは，日本語，英語，中国語等の文字で表され，人と意思疎通を図るために日常的に使われる言語のことを指します。なお，自然言語は音声でも表現されますが，音声は文字化することで，テキストとして扱うことが可能です。「非自然言語」とは，自然言語以外の手段で情報を伝えるもので，例えば，数学的表現，プログラミング言語，マークアップ言語等が該当します[22]。なお，こうした分類は学術的に厳密性を欠いているかもしれませんが，本稿の趣旨に専念した解説を行うため，これ以上立ち入って考えることはせず，一般向けに分かりやすさを優先して言語を分類することにします。今後，われわれは，人間が日常の社会生活で意思疎通を図る際に使う“言葉”に注目し，そうした言葉が文字化されたものを「テキスト」と言うことにします。

また，②保存状態の観点からは，テキストを印刷されたものと，コンピュータで扱うことができるものに分けることができます。印刷されたテキストとは，書籍，雑誌，新聞紙等の紙媒体に掲載されたテキストです。また，コンピュータで扱うことができるテキストとは，コンピュータで処理可能なファイル形式で保存されたテキストのことです。例えば，Word のファイル形式（ファイル名.docx）や Excel の CSV ファイル形式（ファイル名.csv）などがそれに該当します[23]。まとめると，テキストは種類別と保存状態別の組み合わせにより，合計 4 グループに分類できることがわかります。**図表 2-1** は，こうしたグループ分けを示したものです。

ここまでの整理から，本稿が関心対象とするテキストは，自然言語，特に日本語で表されており，かつコンピュータ処理可能なテキスト，つまり，**図表 2-1** のグループ C に該当する日本語テキストです。以後，コンピュータ処理可能な自然言語を，「テキストデータ」と呼びます。本稿では，日本語のテキストデータを解析することで，そこから得た結果を経済分析へ応用する考え方を解説します。テキストデータの解析は「テキストマイニング」と呼ばれます。次節では，テキストマイニングとは何か，それを支える自然言語処理について解説します。

[22] 本稿では，自然言語と対峙する言語を，便宜的に非自然言語と呼んでいますが，正確には，形式言語と人工言語があります。形式言語は文法や意味が形式的に与えられ，数学的に記号で表現できる言語です。人工言語はコンピュータプログラミング言語のことです。

[23] Word，Excel は米国 Microsoft Corporation の米国およびその他の国における登録商標または商標です。

図表 2-1：テキストのグループ

テキストの分類 （保存状態別／種類別）	**自然言語** ● 日本語 ● 英語 ● 中国語等	**非自然言語** ● 数学的表現 ● プログラミング言語 (Java, Python, C) ● マークアップ言語等 (HTML, XML, LaTeX)
印刷されたテキスト ● 書籍 ● 雑誌	A	B
コンピュータ処理可能なテキスト ● テキストファイル ● CSV ファイル	C	D

（出所）筆者作成

テキストは保存状態と種類の観点から，4 つのグループ(A,B,C,D)に分類することができる。例えば，グループ A は，日本語の文章が紙面に印刷されたものであり，日本語の新聞紙が含まれる。他方，テキストファイルで電子化された新聞記事はグループ C に位置付けられる。

2-2. テキストマイニングと自然言語処理

(1) テキストマイニングとは

テキストマイニングとは，「数量データではなく，文字列で記述したコンピュータ処理可能な自然言語（テキストデータ）から，情報を析出すること」です[24]。数量データは，その数値の発生した構造が明確に定義でき，発生した数値に意味を与えることができるため，一般的に**構造化データ**といいます。これに対して，テキストデータは**非構造化データ**に該当します。テキストマイニングは，自然言語処理と，統計学や機械学習といったデータ・マイニング手法の二段階で構成されています。データ・マイニング手法は構造化データを扱うことができるのですが，テキストデータは非構造化データであるため，そのままの形では数量的に分析できません。そこで，次で述べる自然言語処理を用いることによって，テキストデータを数量的に扱うことができる構造化データへと変換し，それを統計的な手法で解析するのです。この一連の作業をテキストマイニングといいます。

[24] テキストマイニングを直訳すると「テキストを採掘すること」です。これは，鉱山からダイアモンドを採掘するがごとく，大量のテキストの中から情報価値の高い言葉を見つけることを意味します。

図表 2-2：自然言語処理の流れ

（出所）筆者作成

自然言語処理の第一ステップは，一般的な文章であるテキストを単語に分解することである。この作業により，単語の種類の分析や頻度を計算するといった数量的な分析が可能となる。

---

(2) 自然言語処理

テキストマイニングの第一段階である自然言語処理とは，テキストを数量的に扱うことができる形へ加工する作業です。ここで，多くの方は，テキストという数値ではない文字列をどうやって数量化するのかと疑問を持たれたかもしれません。これに対する回答を簡単に述べると，一つのテキストを単語に分解して，単語の種類がいくつあるのか，各単語がいくつあるのかというように，数えることのできる単位を作ることです。そうすれば，テキストを数量的に特徴付けることができます。

図表 2-2 を見てください。テキストに限らず，ある全体像の特徴を数量的に把握するには，全体を構成する各要素に分解することから始める必要があります。したがって，自然言語処理を行う際にも，テキストを何らかの要素に分解することから始まります。こうした考え方により，テキストを構造化データへと加工するのです。

さて，テキストを構成する要素は，先ほど単語であると述べましたが，正確にいうと形態素と呼ばれるものです。**形態素**とは，言語学的に重要な概念であり，「それ以上分割しては言語として意味を成さない最小単位のこと」をいいます。形態素は，厳密には単語とは異なる概念です。しかし，本稿では理解のしやすさを優先するため，形態素に該当するものを全て単語と呼称することにします。

自然言語処理は，自然言語で書かれたテキストを，コンピュータを用いて単語に分解して品詞を判別するという形態素解析を基本とします[25]。形態素解析については第３章で詳しく説明しますが，英文とは異なり，単語間が明確に区切られていない日本語の文章を，適当な単語に区切る方法のことです。実際の形態素解析の作業は，コンピュータがどのように文字列を単語として認識しているかで結果が大きく異なります。コンピュータに意味が通る適

[25] 自然言語処理には，形態素解析の結果を利用して，文中の単語と単語の係り受け関係（修飾・被修飾関係）を明らかにする「構文解析」，「意味解析」，そして「文脈解析」があります。

切な単語に区切ってもらうためには，区切り方を認識できる学習環境を整える必要があります。適切な学習環境が無ければ，コンピュータはテキスト内の大量の文字列を，どのように区切るのか判断することはできません。そこで，次節では，自然言語処理のためにコンピュータが単語を学習するための環境である，辞書とコーパスについて解説します。

2-3. 辞書とコーパス

本節では，自然言語処理を行うコンピュータがテキストを読み込んだときに，文字列の中から単語を認識するために必要な環境である，**辞書**と**コーパス**について説明します。われわれが書物を読んでいると，文中に知らない単語が登場することがあります。そこで，単語の意味を知るために辞書を参照するでしょう。実は自然言語処理を行うコンピュータにとっても同様で，知らない単語に出会うたびにコンピュータ用の辞書を引きながら学んでいきます。人間とコンピュータとでは，辞書を使って単語を学習するという点は似ていますが，学習のために使う辞書の在処に違いがあります。人間は，辞書で学習した単語を次々と自らの脳に蓄積させることによって，脳内の辞書を充実させていきます。そして，知らない単語は外部の辞書から，知っている単語は頭の内部の辞書から，それぞれ引き出して，その単語の読み方，品詞，活用形，意味等を理解しています。これに対して，コンピュータは，常に外部にあるコンピュータ用の辞書とつながることで，単語を判別し，文字列中のどこで適当な単語として区切って良いかを判断しているのです。

それでは，自然言語処理を行うコンピュータは，どのような辞書で単語を理解しているのでしょうか。辞書には，大きく分けて**単語辞書**と**シソーラス**の 2 種類があります。これらの辞書の内容とその使い方は，われわれが日常的に使う「広辞苑」や「デジタル大辞泉」といった人間用の辞書とは若干異なります。例えば，単語辞書で単語を検索すると，読み，品詞，活用形等の文法に関する情報は表示されますが，人間用の辞書と異なり，その単語の意味と例文はありません。単語辞書は，文字列の中から単語を認識するに足る情報を提供し，主に形態素解析を行うために使われます。もう一つの辞書であるシソーラスには，ある単語と似た意味の単語が収録されています。シソーラスの構造を描画的に表現すると，単語の概念が，より抽象的な層から，より具体的な層へと階層別に分けられており，また，各階層内に類似した意味を持つ複数の単語が存在しています。このため，シソーラスは，単語同士の類似度を計算するために使われます。ちなみに，最大規模の日本語シソーラスは「日本語語彙大系」です。

次に，**コーパス**とは，単語が実際にどのように使われているかを一覧できる，電子化されたデータベースのことです。コーパスは，英語で“corpus”と書き，その意味は「文書の集積，資料の総体」であることから，大量の言語資料に書かれている様々な文を格納しています。ある単語をコーパスで検索すると，その単語がどのような話し言葉や書き言葉で出現しているかが分かるよう，単語の前後の文と共に表示されます。つまり，コーパスのデータを利用することで，文法的に正しい単語の使われ方と，より好ましい単語の使われ方を，統計的

に明らかにできるのです。ちなみに，最大規模の日本語コーパスは，「現代日本語書き言葉均衡コーパス（BCCWJ)」です[26]。

さて，コンピュータが言語を学習するために，辞書とコーパスはどのように使われるのでしょうか。このことを理解するために，子どもが言語（日本語）を学習する過程を考えてみましょう。コンピュータが言語を習得するためには，第一段階として，語彙力を高めることが必要です。このために，辞書を利用します。ただし，単語を多く知っているからといって，意味が正確に伝わるように文を書いたり，書き手の意図する通りに文を読んだりできるとは限りません。

一つ目の例として，「すももももももものうち。」という文字列を考えてみましょう。どこで単語を区切ればよいでしょうか。候補は多そうです。候補を絞り込むためには，文法的に正しく，単語として意味のあるように文字列を区切るという規則を課すことが必要です。この規則を満たすような文字列の区切り方は「すもも/も/もも/も/もも/の/うち。」となります。

二つ目の例として，「東京都へ行く。」という文字列も考えてみましょう。単語の区切り方は以下に示すように，二つの候補があります。いずれの区切り方であっても，文法的にも単語としても意味は正確に通じます。ただし，一般的に多くの人が実際に好んで使う単語の区切り方は，「東京都/へ/行く/。」だと思われます。「東/京都/へ/行く/。」の場合，「東」「京都」とはどこのことを指すかわかりません。以上の二つの例から分かることは，意味が通じる文というのは，文法的に正しく配置された単語の列であり，同時に，単語は多くの話者（読み手）の好みも反映しながら解釈されているということです。

次に，第二段階として，文法的に正しく，かつ，常識的に意味が通じるような単語の選び方を学ぶ必要があります。専門用語で言い換えると，**制約**と**選好**を満たすように文字列を単語として区切る方法を学ぶということです。制約とは，「必ず守らなければならない強い規則のことであり，ここでは文法を指します。文法が正しくなければ意味が通じないことは先の例からも明らかです。選好とは，「他の選択肢よりも優先して守るという弱い規則」のことであり，ここでは多数の話者の好みと考えてください。

制約と選好について，具体例で考えてみましょう。子どもは周りの大人たちの会話を聴いたり，本に書かれている文章を読んだりしながら，文の読み書きの技能を習得しています。これを，コンピュータが言語を学ぶことに当てはめると，言葉がどう話され，書かれたかという大量の実例データを参照することで，正しく言葉を使う際に必要となる制約と選好を理解するということです。コーパスとは，こうした大量の実例データを収録している電子上の保管庫のことであり，そのデータを統計的に解析することで，制約と選好を導き出しています。このように，コンピュータは，品詞の並び順を定める文法を，単語辞書とコーパスを参照することで学習しているのです。

[26] 大学共同利用機関法人国立国語研究所（https://tokuteicorpus.jp）

第３章 テキストを分析単位にする方法とは

第２章では，テキストマイニングが対象とするテキストとは何か，そして，自然言語処理を行うために欠かせないコンピュータ上の言語学習環境である辞書とコーパスについて説明しました。本章では，そうした言語学習環境を素材にして，文字列を単語に区切る「形態素解析」という方法について学びます。

3-1. 形態素解析とは

辞書とコーパスを利用して自然言語を学習することにより，コンピュータは単語についての知識（単語の読み方，品詞，原形）を獲得できるようになります。そうすることで，コンピュータは文中から単語を探すこともできるようになります。文中から単語を探すということは，言い換えると，文字列を単語に区切るということです。これを**形態素解析**と言い，「文字の塊であったテキストを単語に分解して分析可能な状態する作業」のことであり，自然言語処理の第一段階にあたります。

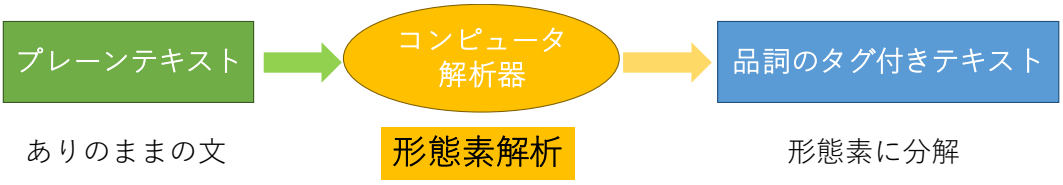
文字列を単語に区切る作業は，意外と単純ではありません。特に，日本語で書かれた文は，英語と異なり，単語と単語の区切りが明確でなく，候補も複数あり得ます。例えば，「ここではきものをぬいでください」という文字列を意味が通じるようにするには，どのように単語を区切れば良いでしょうか。これに対する答えは，「ここで/は/きもの/を/ぬいでください」と「ここで/はきもの/を/ぬいでください」の二つが考えられます。どちらの区切り方も文法的に正しいのですが，前者は「着物」を，後者は「履物」をそれぞれ単語として捉えているため，文の意味は異なります。このように，文字列をどこで区切るかという判断は，たとえ人の目を通しても一度考えなくてはなりません。

文字列を区切る作業を考えるにあたり，まずは言語学上の概念を定義します。先ほど第２章の 2-2 で述べたように、**形態素**とは，「文法上意味を持つ最小単位の文字列」のことを指します。これに対して，これまでに使ってきた**単語（語）**とは，「文法上一つの意味のまとまりを持つ最小単位」であり，一つ以上の形態素から構成されたものです。形態素は，正確には単語と異なりますが，日本語においては形態素と単語の違いはほとんどないため，やはり本章でも分かりやすさを優先し，形態素のことも単語と書くことにします[27]。

形態素という概念に基づいて，文中から単語を探し出し，単語の品詞と活用形は何であるのか解析することを形態素解析と言います。われわれが本や新聞で見るありのままの文章をプレーンテキストといいます。形態素解析を行うことで，プレーンテキストは単語に区切って書き下され，品詞のタグ付きテキストになります（**図表 3-1**）。こうした文字列の表記を**分かち書き**と言います。

[27] 形態素という概念は欧米語の言語研究において導入されたという経緯があるため，日本語の言語研究とは相性は良くないと言われています（土屋 2015）

図表 3-1：形態素解析の流れ



（出所）筆者作成

形態素とは文章を構成する要素で，意味を持つ最小の単位のことを指す。形態素解析を行うことで，本や新聞で見るプレーンテキストは単語に分割され，品詞のタグ付きテキストになる。これにより，数量的な分析が可能となる。

より具体的に，形態素解析は，**①単語分割**，**②品詞付与**，**③原形復元**，という三つの処理から構成されています。単語分割は，「文を単語に分割すること」です。品詞付与は，「分割された単語の品詞を特定すること」です。そして，原形復元は，「文中で活用変化された単語を原形に戻すこと」です。これら三つの処理は，第 2 章で述べた単語辞書と，本章第 2 節で述べる連接可能性行列を利用したルールに基づいて行われます。単語辞書には，単語の読み，品詞，活用形が登録されているため，コンピュータが単語辞書を用いて単語分割を行うと，品詞付与と原形復元も同時に行うことができるのです。

3-2. 自然言語の構造と品詞

形態素解析の処理の中心は，文を単語に区切ることです。ところで，形態素解析を始めるにあたり，そもそも文とは何かを明らかにしておきましょう。本節では始めに，文と形態素の関係を含めて，自然言語の構造を学びます。形態素解析では品詞の情報が重要な意味を持ちます。なぜなら，形態素解析では，品詞の並び順に従って文から単語を探し出しているからです。このため，どのような種類の品詞があるか学ぶことが必要です。最後に，品詞の並び順を定める文法について説明しますが，これは第 3 節で述べる形態素解析を実践するためのアルゴリズムと関係しています。

形態素とは，先述した通り「意味を持つ最小の言語単位であり，一つ以上の文字で構成」されています。日本語では，形態素は単語とほぼ同じ機能を持ちます。そして，**文節**とは，「一つ以上の単語から構成されて，意味と発音の観点から不自然でない程度の最小の言語単位」のことをいいます。以上より，**文**とは，「一つ以上の文節から構成されて，まとまった内容を持つ，形の上で完結した言語単位のこと」をいいます。このように，自然言語は各部品が重なった構造的なものといえます。

図表 3-2 で示す通り，日本語の品詞は 10 種類あります。中でも特に，動詞，形容詞，形容動詞，助動詞については，本稿が関心対象とする経済の情勢判断を行う上で重要な品詞

図表 3-2：日本語の品詞

品詞名	例	品詞名	例
1.動詞	買う，減る，上がる	6.連体詞	いろんな製品，大きな家
2.形容詞	大きい，良い，高い	7.接続詞	それで，だが，または
3.形容動詞	満足だ，積極的だ，急だ	8.感動詞	あら，ええ，こんにちは
4.名詞	消費税，大阪，輸出	9.助動詞	高くない，低そうだ
5.副詞	とても，やや，すぐに	10.助詞	関西が，今から，将来は

（出所）筆者作成

日本語には 10 種類の品詞があり，**図表 3-2** では，各品詞とそれらに対応した例をあげている。中でも，連体詞，助動詞，そして助詞については，単独で成立しないため，下線部が品詞例となる。

であるため，第 4 章で詳細に検討します。コンピュータは，品詞の並び順を定める文法を，単語辞書とコーパスを参照することで学習します。そうした学習において，品詞の**連接可能性行列**という，「これまでの学習を元に作成された品詞間の接続のしやすさが行列で表されたデータベース」が用いられます。品詞の連接可能性行列については次節で解説します。

形態素解析を行うにあたっては，品詞の並び順を定める文法が重要ですが，文法体系は数多く存在します。そのため，「正しい文法」は存在しません。正しい文法のことを**正書法**といいますが，自然言語には正書法がないことが知られています（土屋，2015）。様々な文法があることの証拠に，ある品詞が省略されて話されたり，書かれたりすることがあります。例えば，「百貨店は前年より上昇」という文は，「百貨店の売上高は前年より上昇した」から，「売上高」という名詞を省略したものと考えられます。

また，言葉は時代に応じて変化するため，新しい単語は，まだ単語辞書に登録されていない可能性もあります。そうした場合，その単語の品詞も分からないため，文法的にその単語をどのように処理すれば良いかも判断できません。例えば，「スマホで撮った動画がインスタ映えしたからバズった」という文を考えてみましょう。下線の「スマホ」，「インスタ映え」，「バズった」は，新しい商品やサービスの登場に伴い使われつつある新しい単語ですが，既存の単語辞書で十分に対処できるのでしょうか。これらが単語辞書に未登録の場合，「スマホ」と「インスタ映え」は名詞とし，「バズった」は「バズる」を原形とする動詞として，単語辞書に追加することが望ましいでしょう。

3-3. 形態素解析のアルゴリズム

本節では，実際に形態素解析がどのように行われるのか説明します。**図表 3-3** は，ある入

図表 3-3：形態素解析の例

【入力テキスト】（日本経済新聞 2019 年 3 月 6 日朝刊より）

「中国が景気減速をにらんだ大規模な景気対策を打ち出した。」

【出力結果 1】（形態素原形を出力）

名詞 / 助詞 / 名詞 / 名詞 / 助詞 / 動詞 / 助動詞 / 接頭詞 / 名詞
“中国” “が” “景気” “減速” “を” “にらむ” “だ” “大” "規模"
/ 助動詞 / 名詞 / 名詞 / 助詞 / 動詞 / 助動詞/ 記号
“だ” “景気” “対策” “を” “打ち出す” “だ” “。”

【出力結果 2】（形態素表層形を出力）

名詞 / 助詞 / 名詞 / 名詞 / 助詞 / 動詞 / 助動詞 / 接頭詞 / 名詞
“中国” “が” “景気” “減速” “を” “にらん” “だ” “大” "規模"
/ 助動詞 / 名詞 / 名詞 / 助詞 / 動詞 / 助動詞/ 記号
“な” “景気” “対策” “を” “打ち出し” “た” “。”

(出所)日本経済新聞 2019 年 3 月 6 日朝刊より筆者作成

形態素解析では 2 つの出力結果を表示させることが可能である。出力結果 1 は，文を単語に分割する際に，単語を原形で表示した結果である。出力結果 2 は，文を単語に分割する際に，単語を原形に戻さずに入力テキストの活用形のまま表示した結果である。

力テキストに対して形態素解析を行い，その出力結果を表したものです（解析には，後述する MeCab というソフトウェアを利用しています)。出力結果 1 を見ると，文が単語に分割されるとともに，品詞が特定され，活用変化した単語が原型に戻されています。これが，単語分割，品詞付与，原形復元という，形態素解析の 3 つの処理になります。また，出力結果 2 は，原形復元を行わずに表示したものです。人の目には簡単に映る形態素解析の結果ですが，実際にどのような仕組みで実行されているのでしょうか。

形態素解析は二段階で構成されています。第一段階は，単語辞書を参照することで，文字列の始まりから終わりまで一文字ずつ対象範囲をずらしながら，単語を探索します。具体的に，入力テキスト中の各文字の位置から始まる単語を，**図表 3-4** のような単語辞書を参照することで探し出します。単語辞書は五十音順で単語を掲載しているため，ここでは例文中

図表 3-4：単語辞書の例

見出し語	読み	品詞	活用型	活用形	基本形
打ち	うち	動詞	タ行五段	連用形	打つ
うち	うち	名詞			うち
打ち出し	うちだし	名詞			打ち出し
打ち出し	うちだし	動詞	サ行五段	連用形	打ち出す
打ち出した	うちだした	動詞	サ行五段	過去形	打ち出す
が	が	助詞			が
規模	きぼ	名詞			規模
景気	けいき	名詞			景気
減速	げんそく	名詞			減速
大	だい	接頭辞			大
大規模	だいきぼ	名詞			大規模
大規模な	だいきぼな	形容動詞		連体形	大規模だ
対策	たいさく	名詞			対策
出し	だし	名詞			出し
出し	だし	動詞	サ行五段	連用形	出す
出した	だした	動詞	サ行五段	過去形	出す
た	た	助動詞	特殊型	連体形	た（だ）
中国	ちゅうごく	名詞			中国
にらんだ	にらんだ	動詞	マ行五段	連用形	にらむ
を	を	助詞			を

（出所）筆者作成

日本語の単語辞書には，膨大な数の単語が 50 音順に格納されている。例えば，「中国が景気減速をにらんだ大規模な景気対策を打ち出した。」という文から単語を探し出そうと単語辞書を参照するとする。その結果，形態素解析のソフトウェアは，単語辞書の中から当該の入力文に含まれていそうな単語の候補を示す。その際候補となる単語は，上記のように表される。

に含まれる単語の候補のみを五十音順で表示しています。

次に，形態素解析の第二段階では，品詞の**連接可能性行列**を参照することによって，二つの単語が品詞の観点から連接する可能性が高いかどうか判断して単語を探索します。連接可能性行列とは，品詞間の接続のしやすさを行列で表したデータベースです。文中で二つの単語がそれぞれどの品詞であれば，連続して現れやすいかを表しています。

図表 3-5 は，任意の二つの単語（単語 A と単語 B）についての，連接可能性行列の例で

図表 3-5：連接可能性行列の例

単語 B / 単語 A	文末	名詞	動詞	助詞	接続詞
文頭	0	1	1	0	1
名詞	1	1	1	1	0
動詞	1	1	1	0	1
助詞	1	1	1	1	1
接続詞	0	1	1	0	1

（出所）奥村(2010)を参考に筆者作成

単語 A と単語 B は，文中に順番に続けて現れる任意の二つの単語である。単語 A と単語 B は，文頭か文末であるか，または何かの品詞であったりするため，様々な組み合わせが考えられる。組み合わせの中でも，文法的に連接しやすいものと，連接しにくいものがある。この表では，連接可能性を単純に考え，連接する組み合わせを 1，連接しない組み合わせを 0 と表している。

す。表中の「文頭」と「文末」とは，文の始まりと終わりの位置を示しています。先の例文に文頭と文末を明示すると，「(文頭)中国・・・(省略)・・・打ち出した。(文末)」と表記できます。人間とは異なり，コンピュータは教えられなければ自然言語で書かれた文の開始位置と終了位置を知らないため，処理の上ではこうした表記を施すのが一般的です[28]。

ここでは単純に，品詞が連接する場合と連接しない場合を考え，「1」は品詞が連接することを，「0」は品詞が連接しないことを表しています。具体的に，単語 A が文頭で単語 B が文末という連接は成立不可能です。そのような連接は，そもそも文が存在しないことを意味するからです。しかし，単語 A が文頭で単語 B が接続詞という連接，また，単語 A が助詞で単語 B が文末という連接は，それぞれ成立可能です。これは，「(文頭) しかし・・・である (文末)」という文が成立することから分かります。

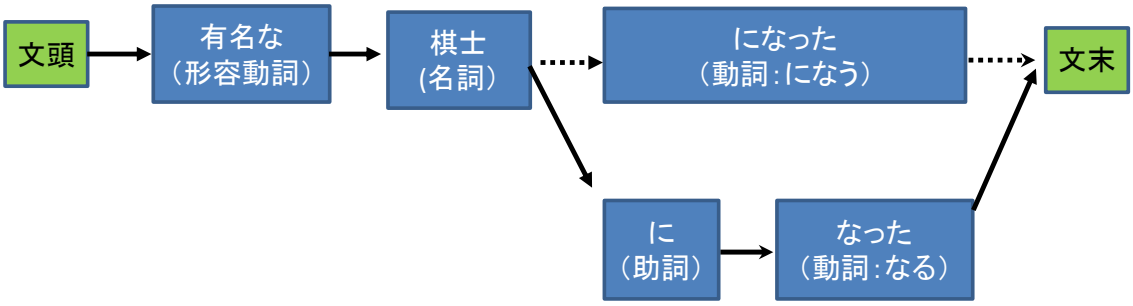
形態素解析では単語辞書と連接可能性行列を参照することが分かりました。次に，形態素解析はどのように実行されるのかを，例文を用いて説明します。**図表 3-6** は，「有名な棋士になった」という文に対する形態素解析の結果を，ラティス構造で表したものです[29]。これをみると，単語の区切り方には二つの経路の可能性があることがわかります。どちらの経路が実際の解析結果となるかは，図にコストの概念を取り入れた**コスト最小法**で導きます。これは，完成文という目的地までの移動手段を表していると解釈してください。

---

[28] 実際のコンピュータ処理では，文頭が BOS (Begin Of Sentence)と，文末が EOS(End of Sentence)と表記されます。

[29] ラティス構造とは構造を構成する各要素を束状（ラティス）の図で表したものです。テキストデータにおけるラティス構造とは，文を形成する単語との単語の切り分けパターンを，まるで一本の束のようにまとめたものとなります。

図表 3-6：形態素解析の実行



(出所)筆者作成

形態素解析は，図のように単語と単語を線で結んだラティス構造で表すことがある。例文としてあげた「有名な棋士になった」を形態素解析すると，二つの経路が解析結果の候補となる。しかし，このままでは，どちらの経路を解析結果として採用すべきか判断できない。そこで，単語コストと連接コストを考慮する必要がある。コストを考慮した解析は，次の**図表 3-7** を参照。

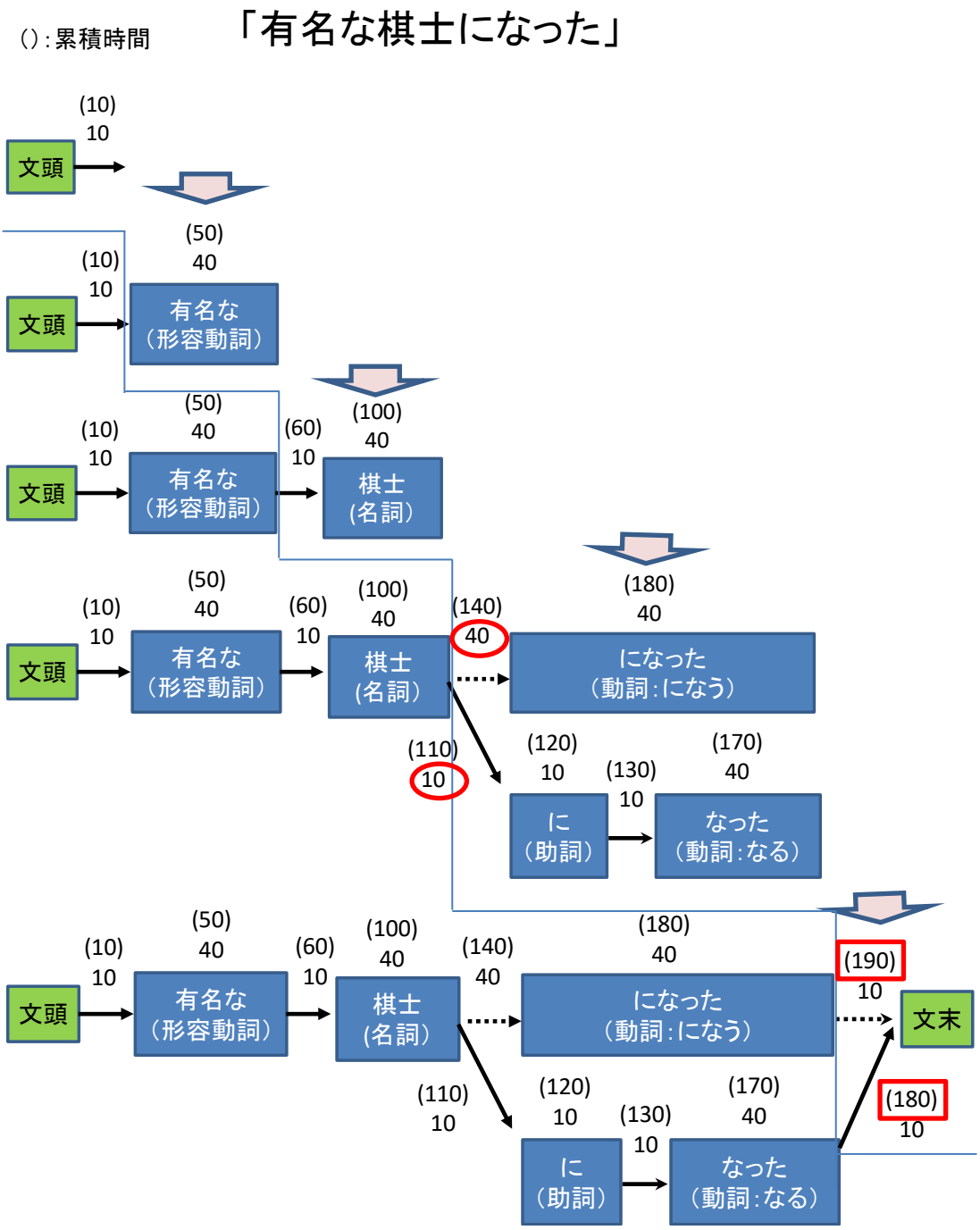
コスト最小法では，単語と単語間の連接にそれぞれコストを与えて，解析結果候補となる各径路の総コストを出し，その中で総コストが最小となるような経路を最適解として導きます。ここでのコストとは，単語を探し出して文を単語に切り分けるために要する「時間」のようなもので，単語コストと連接コストの２つに分けることができます。**単語コスト**とは，単語自体に与えられるコストであり，単語（品詞）の出現頻度が小さいほど，大きいコストが与えられます。**連接コスト**とは，単語(品詞)間の連接可能性に与えられるコストであり，連接しにくい(前後がつながりにくい)品詞同士であるほど，大きいコストが与えられます。

図表３-７は，コストを含めて形態素解析の実行を表したものです。図の矢印と四角形の上にある数字は，それぞれ連接コストと単語コストであり，それらは先述の通り，単語を探して，文を単語に切り分けるために要する時間と解釈できます。括弧内の数字は，文頭から各単語までの切り分けに要した部分コストであり，累積時間と解釈できます。解析が進み文末に至ると，部分コストは解析経路の総コストとなります。

図表３-７を見ながら，形態素解析の流れを追っていきましょう。文を入力すると，まずは文頭を認識して，「有名な（形容動詞)」と，続けて「棋士（名詞)」という単語に切り分けられます。ここまでの経路は一つですが，次に，「になった（動詞：になう)」と「に（助詞)」という２つに分かれます。第一経路の「になった（動詞：になう)」を選んだ場合，その直後が文末です。一方で，第二経路の「に（助詞)」を選ぶと，その直後には「なった（動詞：なる)」が現れて，その次に文末となります。２つの経路のうちどちらを選ぶかは，各径路の総コストを比較して小さい方で決まります。

形態素解析の結果，総コストの小さい第二経路が選ばれます。なぜなら，名詞の後は，動詞よりも助詞が連接する可能性が高いためです。これは，「になった(動詞：になう)」への

図表 3-7：連接コストと単語コストを考慮した形態素解析の実行



（出所）筆者作成

単語コストと連接コストの合計コスト（累積時間）を最小化するような経路が形態素解析の結果として選択される。単語辞書と連接可能性行列を参照しながら処理を進めると，実線の経路の合計コストは 180，点線の経路の合計コストは 190 となり，実線の経路が解析結果として示される。

接続コスト 40 よりも，「に(助詞)」への接続コスト 10 の方が小さい数字として表されています。最終的に，実線の経路の合計コストは 180，点線の経路の合計コストは 190 となり，実線の経路が解析結果として示されます。つまり，総コストが小さい経路とは，統計的に発生する可能性の高い品詞の接続系列であり，自然な文を構成するような単語の切り分け方であるといえるのです。

こうしたコスト最小法による形態素解析の最適解を求める方法では，原理的に全ての解析経路について文頭から文末までの総コストを計算して比較します。しかし，そうした正攻法を用いると膨大な労力（処理能力）を要します。そこで，一気に各経路を展開して総コストを計算する代わりに，文頭から文末にかけて経路を少しずつ展開して行き，部分コストを最小化するような経路を選ぶという処理を行います。こうした処理の名前を，動的計画法の一種である**ビタビアルゴリズム**といいます。

実際にコンピュータで形態素解析を行うときは，統計的な解析手法を実装しているソフトウェアを使います。そうしたソフトウェアは「形態素解析器」と呼ばれます。日本語に対応した代表的な無償の形態素解析器は三つあり，それぞれ JUMAN（ジュマン），Chasen（チャセン），MeCab（メカブ）といいます。この中で，**図表 3-7** で行った形態素解析では，MeCab（メカブ）が使用されています[30]。

さて，以上でテキストデータ分析のための基本的な道具が揃いました。いよいよ，次章では，実際のデータを用いて分析を行ってみたいと思います。

[30] 本章では，MeCab の公式ウェブサイト（taku910.github.io/mecab/） に基づいて，JUMAN および Chasen と比較した MeCab の優位性を述べることにします。なお，MeCab は工藤拓氏によって開発されました。
解析器として MeCab が使われる理由として，MeCab は次の三つの特徴を持つため他に比べて優れていると言われています。第一に，「条件付確率場(CRF)」という確率モデルを採用してコスト最小法を実行できることです。第二に，品詞情報を利用して，解析および推定を行うことが可能です。第三に，単語辞書で定義されていない未知語の品詞推定を，高速で行うことができます。他の形態素解析器である JUMAN と CHASEN は，「隠れマルコフモデル」という確率モデルを採用してコスト最小法を行っていますが，条件付確率場を使う方が，コスト最小法を効率的に実行できることが分かっています。以上の理由から，MeCab が日本語の形態素解析で多く利用されています。

第 4 章　テキストマイニングを実践する

本章ではテキストマイニングの実践編として，内閣府の「景気ウォッチャー調査」のテキストデータを用いて，簡単な分析を行った結果を紹介します。第 1 節では使用するデータについて説明し，第 2 節では代表的な分析方法である頻度分析を，第 3 節ではセンチメント分析について説明し，第 4 節では分析結果を示します。

結論を述べると，われわれが作成した「テキスト版センチメント指標」は，同種の政府統計と大きくずれた動きをしていました。なぜずれてしまったのか。その理由についても，経済用語の複雑性という視点から解説するとともに，改良の方向性を述べたいと思います。

4-1. 使用するデータの説明

昨今では，COVID-19（新型コロナウィルス感染症）、米中貿易摩擦の高進、英国の EU 離脱や中国経済の減速など，世界経済を取り巻く不確実性は日に日に増しており，足元の経済動向の正確な把握が重要となっています。はじめにで見たように，一国の経済の良し悪しを把握するためには，GDP や物価を始めとする集計データ（マクロデータ）を用いることが一般的でした。しかし，集計データには，収集から公表までにある程度の期間を必要とし速報性に欠けること，数字の背後にいる個々の経済主体の事情といったミクロの要因が捨象されてしまっているという課題があります。

他方で，情報技術の急速な進展により，国内外の経済活動において生成される大規模なデータ（ビッグデータ）が様々な形で利用可能になり始めています。個人の購入履歴や行動パターンを始め，きわめて豊富な情報を内包しているビッグデータの活用は，マクロ経済動向に関する，より精緻な情勢判断と予測に大いに役立つと考えられます。こうした問題意識から，本章では，前章までの分析道具を使って，ビッグデータの一つであるテキストデータを分析してみたいと思います。

日本では消費や景況感に関する統計が多く公表されています（**図表 4-1**）。今回，われわれが分析に利用するのは，内閣府の「景気ウォッチャー調査」のテキストデータです。「景気ウォッチャー調査」は，足下の「景況感（Business sentiment）」を示す経済指標の一つです[31]。中でも「景気ウォッチャー調査」は，地域の景気に関連の深い動きを観察できる立場にある人々の協力を得て，地域ごとの景気動向を的確かつ迅速に把握し，景気動向判断の基礎資料とすることを目的に 2000 年から毎月実施されています。また，調査時期は毎月 25 日から月末までで，翌月中旬には結果が公表されるため，速報性も高い経済指標です[32]。そのため，多くのエコノミストが景気判断の参考にしています。

---

[31] 通常，景気循環（business cycle）は，ある一定期間（1 カ月，1 四半期（3 カ月間）など）を対象期間とする経済動向を意味しており，例えば GDP 成長率の推移などを用いて表されます。これに対し，景況感は，比較的直近を対象とした経済の先行きに対する家計や企業の見通しを意味するものといえます。

[32] 調査対象は北海道から沖縄までの 11 地域別に，家計動向，企業動向，雇用などの経済活動を敏感に観察できる業種の適当な職種の中から約 2,000 人を選び，毎月調査が行われています。調査形式はインタビュー形式で行われ，電話・WEB・電子メールの 3 通りの方法で回収されています。

図表 4-1：国内における消費や景況感に関する主な統計

統計の名称 （公表機関）	概要	公表 頻度	メリット	デメリット
景気ウォッチャー調査：現状判断DI・先行き判断DI(内閣府)	･景気に敏感な職業の人に対するサーベイ調査 ･足下と3カ月先	月次	･調査月の翌月に公表され，速報性が高い	･アンケートのため，イベントや海外情勢の影響を受けやすい ･特定の業種が対象で，網羅性がない
消費者態度指数(内閣府)	･世帯を対象としたサーベイ調査	月次	･調査月の翌月に公表され，速報性が高い	･アンケートのため，イベントや海外情勢の影響を受けやすい
日銀短観： 業況判断 DI (日本銀行)	･経済動向に対する企業の判断を問うサーベイ調査	四半期	･業種別･規模別に企業のマインドを把握できる	･アンケートのため，企業に関するイベント，政策動向の影響を受けやすい
地域別支出総合指数(RDEI)： (内閣府)	･地域の支出の動向を迅速かつ総合的に把握する目的で試算	四半期	･消費，住宅投資，設備投資，公共投資の４項目を都道府県別に把握	･輸出動向が含まれておらず，総合的な経済状況の把握に課題
県民経済計算： (内閣府)	･都道府県別の付加価値総生産額	年次	･都道府県別の総合的な経済状況を時系列で比較可能	･GDP 公表から約２年遅れで発表されるため，速報性に欠ける

(出所)各種資料から筆者作成

政府が公表する消費や景況感に関する統計（集計データ）は数多く存在する。これらの統計は，足元の景況感を知る上で非常に貴重な情報を与えてくれるものの，公表までにある程度の期間を必要とすることに加え，速報性に欠けるなどの課題があることに注意が必要である。

「景気ウォッチャー調査」には，景気の現況に対する認識を尋ねる「現状判断 DI」と，現在から２～３カ月後の先行きに対する認識を尋ねる「先行き判断 DI」の２つの指数があります。いずれも「①良くなっている，②やや良くなっている，③変わらない，④やや悪くなっている，⑤悪くなっている」の５つの選択肢から回答者は適切と考える内容を選択します（**図表４-２**の上図の【質問２】と【質問５】，その結果である下図の点線部分）。そして，回答者の数だけ集められた「景気認識」は指数化して集計され，景気の現状（または先行き）を判断するために用いられます。

図表 4-2：内閣府「景気ウォッチャー調査」調査票と回答例

【調査票】

【質問２】
景気が上向きか下向きか、どちらの方向に向かっているかの質問です。今月のあなたの身の回りの景気は、３か月前と比べて良くなっていると思いますか、悪くなっていると思いますか。
次の５つの中から、答えたい番号のプッシュ・ボタンを押し、最後に ＃ を押してください。
【電話方式】
次の５つの中から、答えたい番号を選択し、右の回答欄に該当の番号を打ち込んでください。
【電子メール方式】
次の５つの中から、答えたい番号を選択し、下の該当の番号のボタンをクリックしてください。【Ｗｅｂ方式】
（各分野共通）
①良くなっている　②やや良くなっている　③変わらない　④やや悪くなっている　⑤悪くなっている

【質問３】
質問２のご回答の理由として、どのような点に特に着目しましたか。
次の６（５）つの中から、最も適当と思われる番号のプッシュ・ボタンを押し、最後に ＃ を押してください。【電話方式】
次の６（５）つの中から、最も適当と思われる番号を選択し、右の回答欄に該当の番号を打ち込んでください。【電子メール方式】
次の６（５）つの中から、最も適当と思われる番号を選択し、下の該当の番号のボタンをクリックしてください。【Ｗｅｂ方式】
（家計動向関連の方の場合）
①来客数の動き　②販売量の動き　③単価の動き　④お客様の様子　⑤競争相手の様子　⑥それ以外
（企業動向関連の方の場合）
①受注量や販売量の動き　②受注価格や販売価格の動き　③取引先の様子　④競争相手の様子　⑤それ以外
（雇用関連の方の場合）
①求人数の動き　②求職者数の動き　③採用者数の動き　④雇用形態の様子　⑤周辺企業の様子　⑥それ以外

【質問４－２】　質問３において次の番号を選んだ方への質問です。
（家計動向関連）④～⑥を選択された方。
（企業動向関連）③～⑤を選択された方。
（雇用関連）④～⑥を選択された方。
今のご回答について、具体的な状況を教えてください。
ピーという発信音が鳴ったら、30 秒以内でお話しください。回答が終了したら、 ＃ を押してください。【電話方式】
今のご回答について、具体的な状況を 300 字以内で教えてください。
【電子メール方式】
質問３のご回答について、具体的な状況を 300 字以内で教えてください。
【Ｗｅｂ方式】
（各分野共通）
自由回答

【質問５】
将来の景気についての質問です。今後２～３か月先のあなたの身の回りの景気は、今月より良くなると思いますか、悪くなると思いますか。
次の５つの中から、答えたい番号のプッシュ・ボタンを押し、最後に ＃ を押してください。
【電話方式】
次の５つの中から、答えたい番号を選択し、右の回答欄に該当の番号を打ち込んでください。
【電子メール方式】
次の５つの中から、答えたい番号を選択し、下の該当の番号のボタンをクリックしてください。【Ｗｅｂ方式】
（各分野共通）
①良くなる　②やや良くなる　③変わらない　④やや悪くなる　⑤悪くなる

【質問６】
質問５で、そのように回答した理由を教えてください。
ピーという発信音が鳴ったら、30 秒以内で自由にお話しください。回答が終了したら、 ＃ を押してください。【電話方式】
質問５で、そのように回答した理由を 300 字以内で教えてください。
【電子メール方式、Ｗｅｂ方式】
（各分野共通）
自由回答

【回答結果の抜粋(2018 年 7 月調査)】

７．近畿（地域別調査機関：りそな総合研究所株式会社）

（－：回答が存在しない、＊：主だった回答等が存在しない）

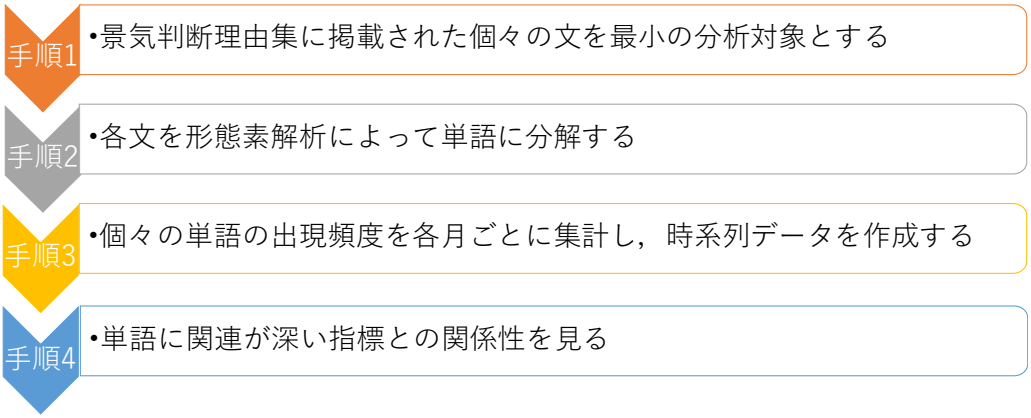
分野	景気の現状判断	業種・職種	判断の理由	追加説明及び具体的状況の説明
家計動向関連（近畿）	◎	百貨店（営業担当）	お客様の様子	・高級ブランド婦人服のクリアランスでの売上が好調であり、ブランド宝飾品も例年以上に引き合いがあるなど、売上は更に上向いている。
	◎	百貨店（販売推進担当）	来客数の動き	・来客数は７か月連続で前年実績を上回っており、固定客化や、地元沿線での顧客の囲い込みが更に進んでいる。また、消耗品の売行きが継続的に好調であり、来店頻度も増加している。
	◎	家電量販店（人事担当）	販売量の動き	・気温の上昇に伴い、エアコンを中心とした季節商材の販売が伸びている。また、高付加価値商品の販売に注力することで利益も確保されている。
	○	一般小売店［事務用品］（経営者）	お客様の様子	・新年度に入って、官公庁では新たな予算の策定時期を迎えており、参考見積を含めて、案件が多くなっている。
	○	百貨店（売場主任）	お客様の様子	・今月の売上は目標、前年の水準共に上回る見込みである。ただし、インバウンドの動きやバーゲン商品は好調であるものの、売れるアイテムは限定的である。客は不要不急の商品の購入には消極的であるが、季節商材などの実需商品や、値段が高くても客にとって価値のある商品は、購入する傾向が強い。
	○	百貨店（企画担当）	お客様の様子	・富裕層による外商売上は、前年から微増である。一方、ボリューム層である自社カード顧客の売上は、クリアランスセールを６月に前倒しした影響もあり、若干の前年割れとなっている。特徴的なのはインバウンド売上であり、今月は前年比で78%増えている。来客数、客単価共に前年を大幅に上回っており、全体の売上を押し上げている。
	○	百貨店（営業担当）	単価の動き	・富裕層の購買意欲が回復しており、宝飾品、絵画、時計などの高額商品の動きも活発化しつつある。また、化粧品や高級ブランドの雑貨を中心とした、インバウンドの購買意欲も高い。

集計データはこの情報を使用　　以降の分析で利用するテキストデータ

(出所)内閣府「景気ウォッチャー調査」調査票に筆者加筆

「景気ウォッチャー調査」調査票では，景気判断の認識に関する選択肢と合わせて，なぜそのように判断したかという理由をセットで尋ねている。この判断理由の自由記述には，様々な情報が含まれており，経済の現状と先行きを知るために重要な情報を提供するものと考えられる。

図表 4-3：頻度分析の手順



(出所)筆者作成

頻度分析では，始めに個々の文を単語レベルまでバラバラに分解し，個々の単語の出現頻度を集計することから始める。本稿では，出現頻度が多い時期に何が起こったのか，単語に関連が深いと思われる他の指標と並べて関係性（主に相関関係）を確認するという手続きで分析を行っている。

しかし，集計後のデータを見るだけでは調査を十分に活用しているとは言えません。なぜなら，集計する過程で，回答者が肌で感じた「なぜそのように判断したか」という背景情報が捨象されてしまっているからです。実は，「景気ウォッチャー調査」のホームページでは，景気判断の認識に関する選択肢と共に，なぜそのような判断をしたかという理由について，自由記述形式で書かれた「景気判断理由集」が，コンピュータ処理可能なテキストファイルの形式で公開されています（**図表 4-2** の下図の実線部分）。この「テキストデータ」を分析することで，回答者が日々の業務をこなす中で感じ取った微妙な景況感の変化を析出できる可能性があります。こうした試みは，街角の景気の変化を敏感に察知する立場にある人の「生の声」を取り入れるということですから，景気の先行きや転換点を読み解くうえで，非常に価値があると考えられます。

以下では，この「景気判断理由集」を用いてテキスト分析を実践します。具体的には，基本的な分析方法である**頻度分析**について説明した後，発展的な手法である**センチメント分析**の結果を報告します。

4-2. 頻度分析－文中に登場する言葉の頻度を計る－

(1) 分析手順

第 2 節では，**頻度分析**という，「文章の中における各々の言葉の出現頻度に着目した分析手法」について説明します。この分析の手順は**図表 4-3** に示した通りです。以下，それぞれの手順について解説していきましょう。

〇手順１：景気判断理由集に掲載された個々の文を最小の分析対象とする

足下の景気動向を把握することが目的のため，内閣府「景気ウォッチャー調査」の「現状判断 DI」の「景気判断理由」に掲載されたテキストデータを利用します。分析対象期間は，2011 年 1 月から 2017 年 5 月までの 77 カ月間で，分析対象は近畿地域（滋賀県，京都府，大阪府，兵庫県，奈良県，和歌山県の 2 府 4 県）における総計 1 万 3,321 件のコメントです。

〇手順２：各文を形態素解析によって単語に分解する

次に，「景気判断理由」のテキストデータから単語を取り出すために**形態素解析**を行います。これにより，文章を単語に分割し，それぞれの単語に品詞を付与することで，分析可能なデータに変換することができます。ここでは，「KH Coder」というフリーソフトウェアを用いて形態素解析を行いました[33]。

〇手順３：個々の単語の出現頻度を各月ごとに集計し，時系列データを作成する

先ほどの手順 2 で形態素解析によって分解されたテキストデータは，名詞・動詞・形容詞・形容動詞といった品詞別に分けて出力されます（**品詞付与**）。ここでは，上位 30 語について，個々の単語の出現頻度を機械的にカウントしました。

図表 4-4 は 2017 年 5 月における近畿地域の「景気判断理由」を用いて，品詞別に上位 10 位まで集計したものです。調査時期が 5 月のため，名詞の列を見ると，大型連休であるゴールデンウィークに言及したコメントが多く，形容詞を見ても，「良い」が 23 回，「多い」が 11 回と概ね好調であることを示す語彙が上位を占めていることがわかります[34]。登場する単語の多さに注目することで，大まかな景気の良し悪しを推し量ることも可能でしょう。

また，各単語の出現頻度を見るだけでなく，例えば，名寄せを行い，概念を一つにまとめることで，分析者の関心があるテーマについて分析することも可能です。例えば**図表 4-5** では，「猫」とそれに類すると思われる単語（ひらがなの「ねこ」，カタカナの「ネコ」、猫の名前に多い「タマ」，猫の代表的な品種である「ヒマラヤン」）の出現回数を一つにまとめています。次節では，景況感と経済政策との関係を探るという趣旨から，財政政策や金融政策に関係が深いと考えられる単語を名寄せし，分析を行っています。

---

[33] 本章で利用した「KH Coder」は立命館大学産業社会学部の樋口耕一先生によるテキスト型（文章型）データを統計的に分析するためのフリーソフトウェアです（http://khcoder.net/，最終閲覧-2019 年 4 月 18 日）。初心者でもわかりやすい解説書として樋口（2014）があります。なお，「KH Coder」では形態素解析器として，茶筅（ChaSen）が使われています。

[34] 形態素解析を行う際，不自然な言葉が出現することがあります。例えば，KH Coder では，「ゴールデンウィーク」という単語は「ゴールデン」と「ウィーク」の 2 つの異なる単語として認識されました。そのため，「ゴールデンウィーク」という一つの単語として登録し，事前にコンピュータに認識させることで強制抽出するという作業を行っています。

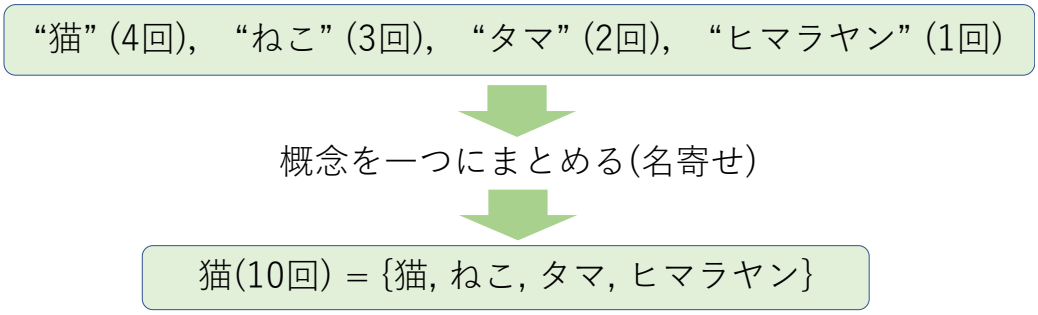
図表 4-4：品詞別抽出語上位 10 位(2017 年 5 月：近畿地域)

順位	名詞	出現回数	順位	動詞	出現回数	順位	形容詞	出現回数	順位	形容動詞	出現回数
1	売上	43	1	増える	28	1	良い	23	1	好調	26
2	単価	28	2	続く	26	2	多い	11	2	堅調	7
3	動き	25	3	変わる	17	3	悪い	10	3	安定	4
4	ｺﾞｰﾙﾃﾞﾝｳｨｰｸ	22	4	減る	14	4	厳しい	9	4	不振	4
5	来客	21	5	伸びる	14	5	少ない	8	5	顕著	3
6	傾向	18	6	感じる	12	6	大きい	8	6	高額	3
7	状況	16	7	上回る	11	7	高い	5	7	高級	3
8	景気	14	8	思う	7	8	長い	4	8	順調	3
9	衣料	13	9	出る	6	9	固い	3	9	大幅	3
10	企業	13	10	上がる	6	10	低い	2	10	同様	3

(出所)筆者作成

5 月は大型連休があったことから，名詞では「ゴールデンウィーク」や「来客」が，動詞や形容詞では「増える」「良い」といった好調さを感じさせる単語が多く出現していることがわかる。

図表 4-5：単語の出現頻度の集計イメージ

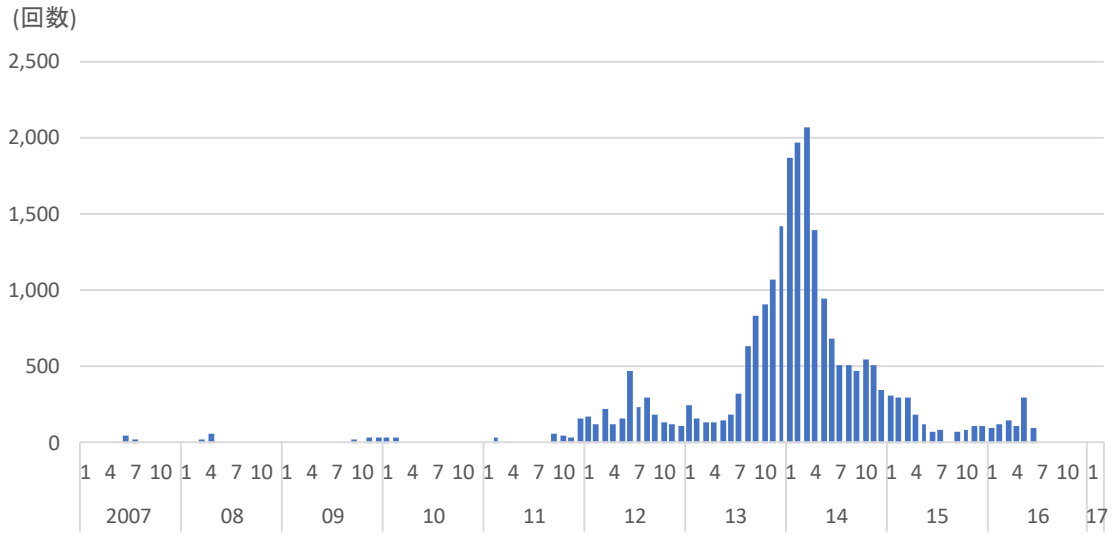


(出所)筆者作成

あるテキストを形態素解析したところ，「猫」とそれに関連すると思われる単語（ひらがなの「ねこ」，カタカナの「ネコ」、猫の名前に多い「タマ」，猫の代表的な品種である「ヒマラヤン」）が登場していた。ここでは，これを「猫」という概念にまとめ，合計で 10 回登場したとして解釈している。

手順 1 から手順 3 の作業を，毎月のテキストデータに対して行うことで，2011 年 1 月から 2017 年 5 月までに登場した単語について，時系列データを作成することができます。

図表 4-6：「景気判断理由集」に登場した財政政策関連語句の出現回数



(出所)筆者作成

上は財政政策関連語句の出現回数を示したものである。回数が多くなっている時期を丁寧に見ると，政権の消費税に関連する決定に対して，即座に反応している（言及される頻度が高い）ように見受けられる。

〇手順４：単語に関連が深い指標との関係性を見る

最後に，分析したい単語と関係が深い経済・社会指標を並べ，両者の関係性について分析を行います。実は，この作業が一連のプロセスの中で最も重要です。なぜなら，単なる”数字の羅列”でしかなかった単語の頻度に，経済的・社会的な意味を持たせることができるからです。どのような経済・社会指標を参照するか，分析者のセンスが試されます。

(2) 分析結果

以下では，景気ウォッチャー調査の「景気判断理由集」を用いて頻度分析を行った結果を３つ紹介します[35]。

始めに，**図表 4-6** を見てください。これは，2007 年 1 月から 17 年 1 月にかけて，財政政策関連語句の登場回数を時系列で並べたものです。ここでは，「増税」「減税」「税率」「年金」「予算」という５つの単語を財政政策関連語句として名寄せし，登場回数を棒グラフで表しました。

---

[35] 本稿では，簡単な分析結果を示すにとどめています。より高度な分析について学びたい方は，下巻の応用編をご覧ください。

図表 4-6 について，少し丁寧に推移を追ってみることにしましょう。始めに目につくのは，2012 年 1 月以降，登場回数が増え始めていることです。特に 12 年 6 月が大きな山になっているようです。これは当時の野田内閣で消費税率を 14 年に 8 %へ，そして 15 年に 10%へと引き上げる「社会保障と税の一体改革」関連法案に関する三党合意が行われた時期と一致します。

これにより，増税が実施されるという期待が高まりました。登場回数の増加はこうした背景によるものと考えられます。その後，登場回数は減少していきますが，13 年 7 月以降再び急増しています。これは安倍内閣が集中点検会合を開催し，消費税率の 8 %への引き上げの是非の検討を行った時期と重なります。その後，14 年 4 月の登場回数が最も多くなっていますが，これは言うまでもなく 5 %から 8 %へと消費増税が実施された月です。ここから言えることは，調査に答えた街角の景気ウォッチャーたちは，財政政策（特に増税に関する動き）に関して非常に大きな関心を持っている（反応している）ということです。

また，決定の翌月には登場回数が増加していることから，即座に反応しているように見受けられます。通常，増税前に消費が盛り上がる駆け込み需要は約半年前頃から生じ，引き上げ後は反動減が生じるとされます。実際に**図表 4-6** を見ても，約半年前から財政政策関連語句の登場回数が増加し，14 年 4 月をピークとして，その後徐々に減少し，約 1 年後の 2015 年 5 月頃にはほぼ言及がなくなりました。詳細な分析が必要ではありますが，駆け込み需要と反動減の影響が落ち着く（織り込まれる）には，2 年程度の期間を要するということかもしれません[36]。このように，登場回数だけ見ても，非常に興味深いことがわかります。

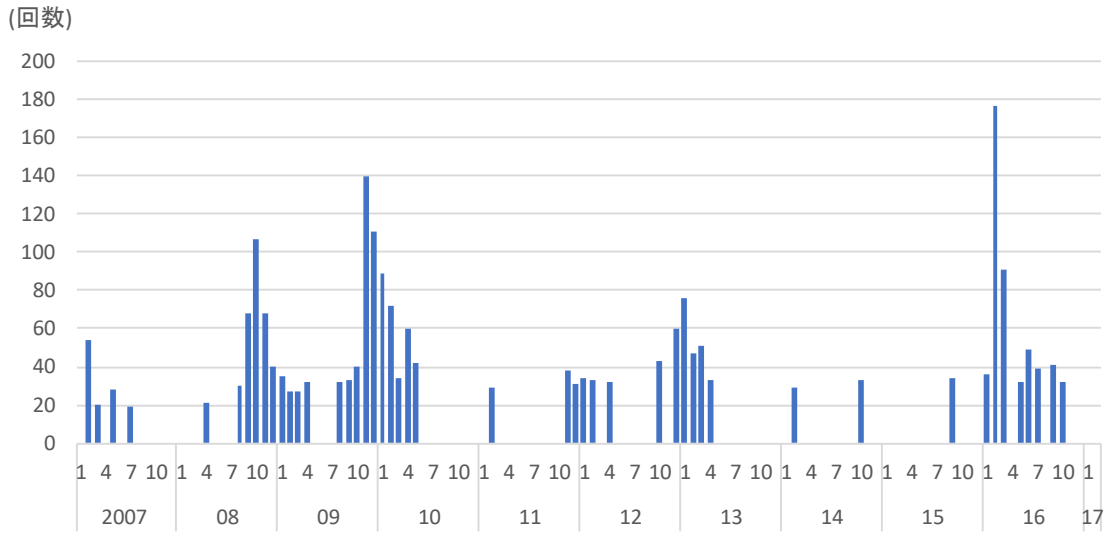
次に，金融政策についてはどうでしょう。**図表 4-7** は同じ期間について金融政策関連語句（「デフレ」「金融」「金利」）の登場回数を見たものです。同じく時系列で見ると，世界経済の動きや日銀の金融政策の決定に対して反応しているように見受けられます。例えば，2008 年の後半以降登場回数が急増しているのは，2008 年 10 月にリーマンショックが起きたことで世界的なデフレ懸念が広がったことに伴い，金融緩和の必要性が議論されたことによるものでしょう。2009 年の 12 月はドバイショックにより世界的に株式相場が急落，ユーロ安や円高が生じるなど，日銀に対して金融緩和の要望が高まった時期です。2011 年 11 月は資産買入増額による金融緩和強化，12 年後半から 13 年前半にかけてはアベノミクス第 1 の矢への期待があったものと思われます。金融政策関連語句の登場回数が最も多くなっているのは 2016 年 2 月ですが，これは前月のマイナス金利の導入が大きく影響したものと思われます。

財政政策・金融政策といったマクロ経済政策の動きに対して，経済主体は敏感に反応しているように思われます。しかし，中味を見ると，財政政策関連語句と比べて，金融政策関連語句の登場回数は圧倒的に少ないようです。この背景には，金融政策の効果が，マイナス金利を除くと直接人々の生活に影響するものが少ないといった理由があるのかもしれません。

---

[36] その後 2016 年 5 月に再び登場回数が増加していますが，これは同月に開催された伊勢志摩サミットで安倍首相が 8 %から 10%への消費税再増税の先送りを表明したことを受けたものと思われます。

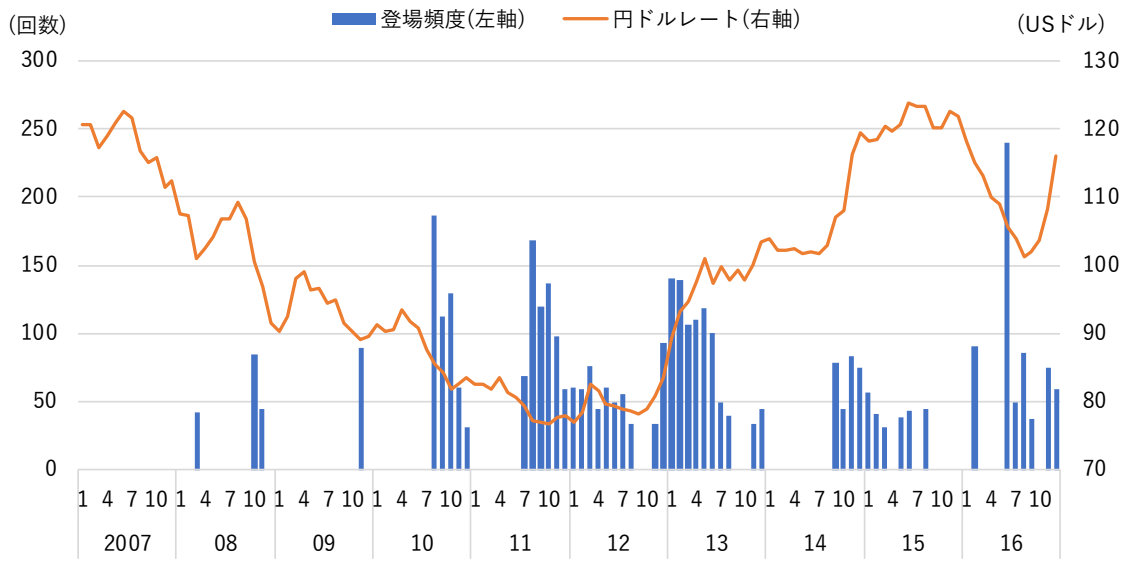
図表 4-7：「景気判断理由集」に登場した金融政策関連語句の出現回数



(出所)筆者作成

上は金融政策関連語句の出現回数を示している。主要な政策変更に対しては反応が見られるものの，財政政策関連語句の推移と比べると，登場回数自体が圧倒的に少ないことがわかる。

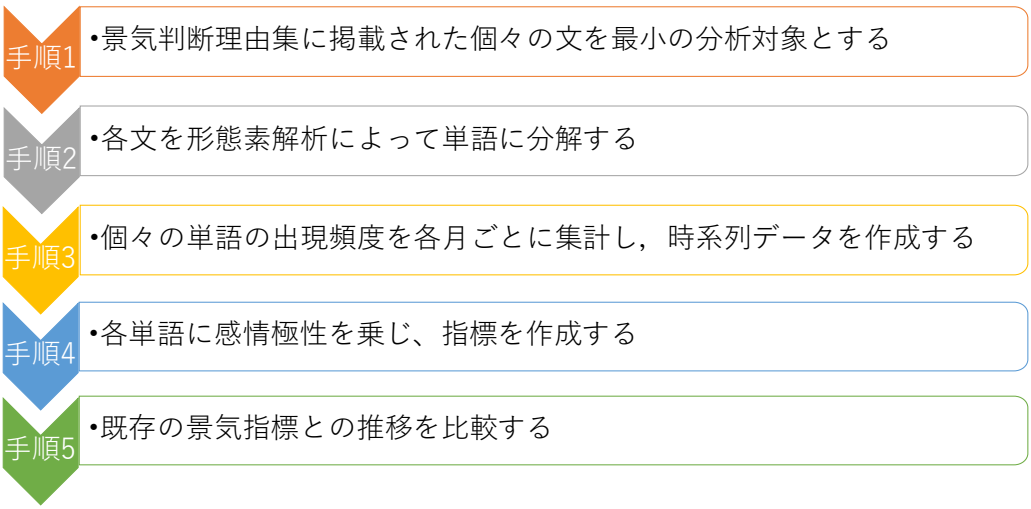
図表 4-8：為替関連語句の出現回数と円ドルレートの推移



(出所)内閣府「景気ウォッチャー調査」，IMF 資料より筆者作成

上の図は為替関連語句の頻度と円ドルレートの推移を重ねたものである。これを見ると，おおむね円ドルレートが 100 円を下回る時期には出現回数が増加していることがわかる。

図表 4-9：センチメント分析の手順



（出所）筆者作成

センチメント分析の手順について，手順 1 ～手順 3 は頻度分析と同様だが，抽出された各単語に感情極性表の点数を乗じてセンチメント指標を作成する点が大きく異なる。

最後に，**図表 4 - 8** は，為替関連語句（「円高」「円安」「為替」）の登場回数と円ドルレートの推移を示したものです。この図からは，円ドルレートが円高局面では 1 ドル 100 円を下回る場合に出現頻度が急激に増加していることがわかります。また，2012 年から 13 年にかけては為替が円高から円安へと大きく変化しましたが，ここでも登場回数が急増しています。特に輸出企業や外国との取引を行う商社などが敏感に反応した可能性があります。このように，出現頻度を円ドルレートという経済変数と結びつけると，人々がどの程度まで円高・円安が進めば反応（言及）するかという「閾値」のようなものが見えてくる可能性もあります。

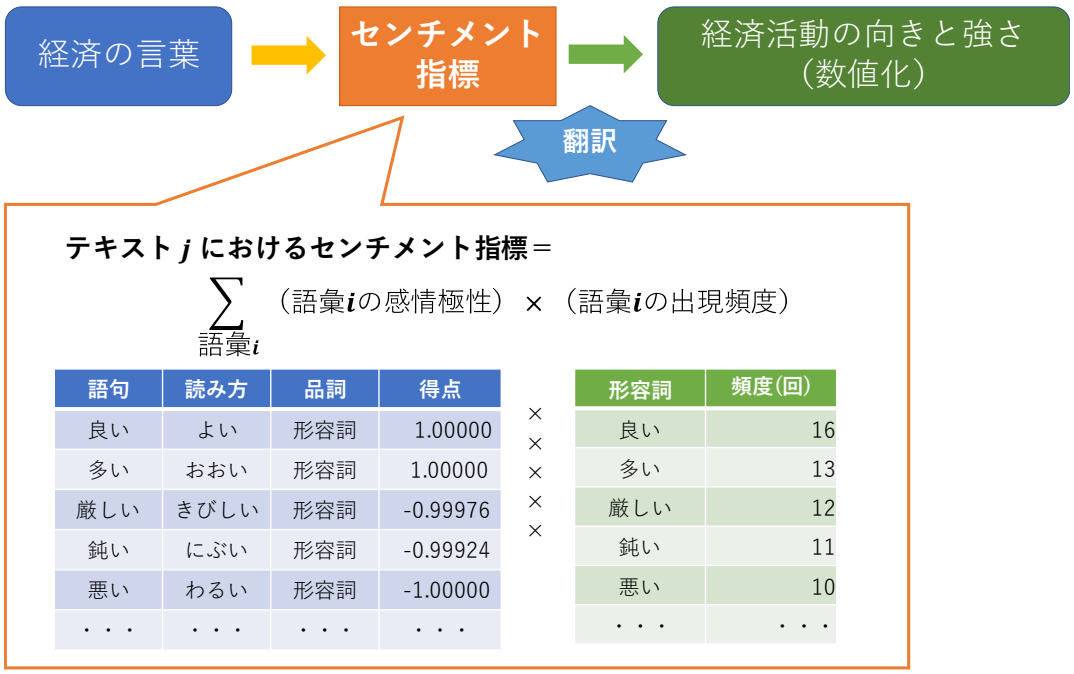
4-3. センチメント分析－言葉に隠された「感情」を読み解く

(1) 分析手順

本節では，**センチメント分析（感情分析）**と呼ばれる手法を用いて分析を行います。これは，「各単語の頻度に感情極性という係数を乗じることで，センチメント指標という数値を算出し，その指標の動きを分析する手法」です[37]。センチメント分析の具体的な手順を**図表 4-9** に示しています。手順 1 ～手順 3 までは頻度分析と共通ですので割愛し，手順 4 ～ 5 について説明します。

[37] 形態素解析については，第 3 章を参照してください。

図表 4-10：センチメント指標作成のイメージ



語句	読み方	品詞	得点
良い	よい	形容詞	1.00000
多い	おおい	形容詞	1.00000
厳しい	きびしい	形容詞	-0.99976
鈍い	にぶい	形容詞	-0.99924
悪い	わるい	形容詞	-1.00000
・・・	・・・	・・・	・・・

×
×
×
×
×

形容詞	頻度(回)
良い	16
多い	13
厳しい	12
鈍い	11
悪い	10
・・・	・・・

(注)出現頻度は 2017 年 3 月の景気ウォッチャー調査を分析したものである
(出所)筆者作成

上の図はセンチメント指標作成のイメージを示している。各文について，形態素解析を行った後，抽出された単語の感情極性に出現回数を乗じることで指標作成を行っている。

○手順 4 ： 各単語から「センチメント指標」を作成する

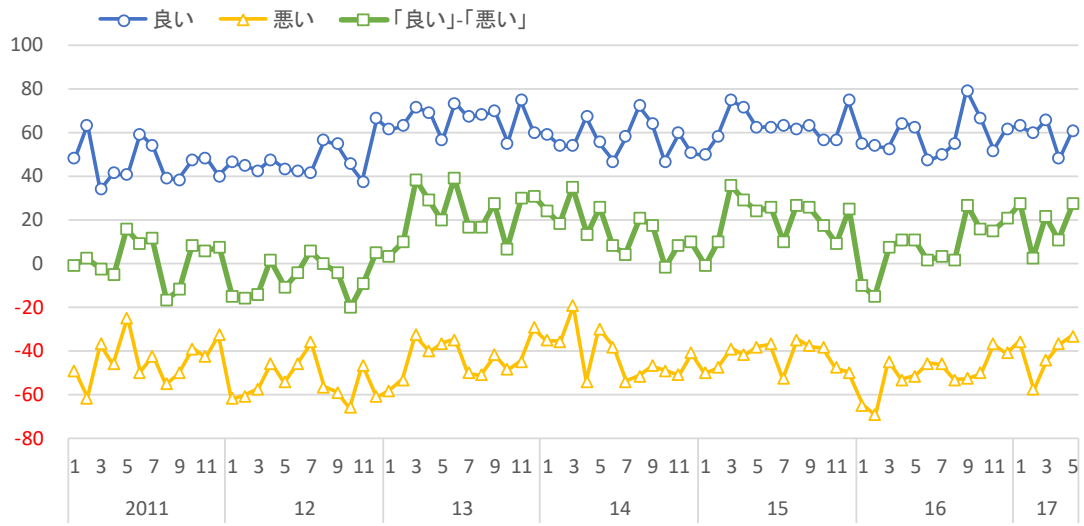
「テキスト版センチメント指標」作成のイメージを**図表 4-10** に示しています。今回，われわれは景況感の「改善」「悪化」に関心があるため，抽出された単語の中でも，「良い」「悪い」「厳しい」といった「形容詞」に着目しました。

ここでは，岡崎・敦賀（2015）と同様，各単語が持つポジティブ／ネガティブな印象を数値化した高村他（2006）の「感情極性対応表」における得点を用いて，個々の単語の出現頻度を加重和することで指標を作成しています[38]。

---

[38] 高村氏の「単語感情極性対応表（http://www.lr.pi.titech.ac.jp/~takamura/pndic_ja.html，最終閲覧 2019 年 4 月 19 日）」は，もともと景況感分析に対応した表ではないため，「景気判断理由集」にある単語の中には，該当する得点が存在しない単語もあります。これらの単語については，個々の語彙の内容を確認し，-1(悪い)か 1(良い)の得点を機械的に与えました（例えば，**図表 4-4** の単語の中では，「多い」が該当します)。

図表 4-11：「テキスト版センチメント指標」の推移



(出所)筆者作成

図表 4-9 の手順に従って作成したテキスト版センチメント指標の推移を示している。「良い」はポジティブな感情を示す語彙について集計したもので，「悪い」はネガティブな感情を示す語彙を集計したものある。「良い」から「悪い」を引いたもの（緑の折れ線グラフ）を以下では確認していく。

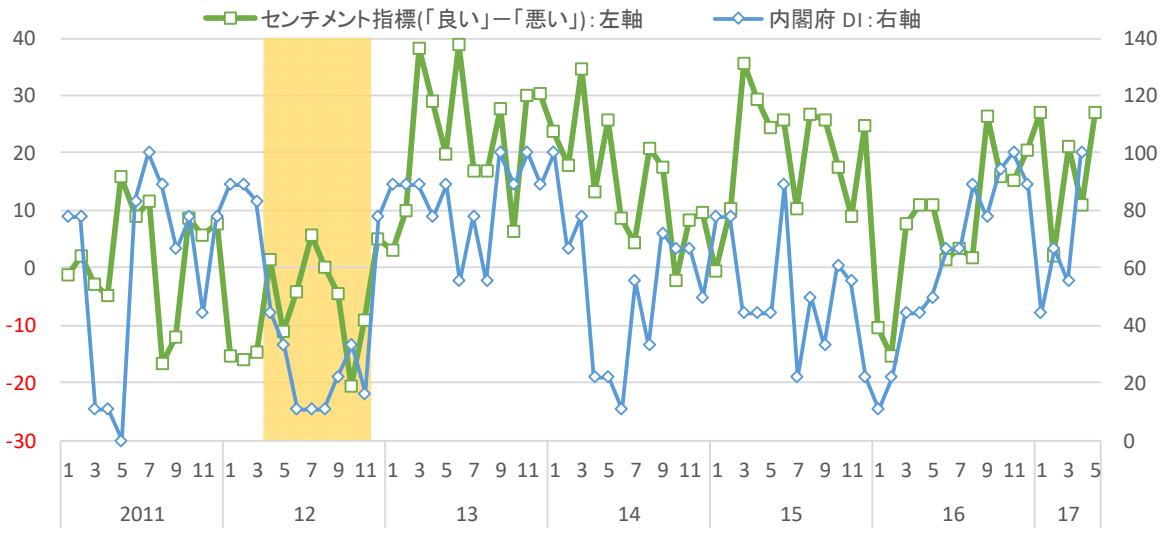
〇手順 5 ：既存の景気指標との推移を比較する

最後に，作成した「テキスト版センチメント指標」を既存の景況感を示す指標と比較し，両者の相関係数を確認することで，「テキスト版センチメント指標」がどの程度既存の景況感指標と整合的か確認します。ここでは，センチメント指標中でも代表的な経済指標である内閣府の「Diffusion Index（DI），全国値」，そして同じく内閣府の「消費者態度指数（近畿）」の値と比較してみます。

(3)「テキスト版センチメント指標」の動き

作成したセンチメント指標の推移を見たものが**図表 4-11** です。果たしてわれわれが作成した「テキスト版センチメント指標」は，景況感を示す既存指標と整合的な動きをしているのでしょうか。**図表 4-12**，**4-13** を見ると，推計期間全体（2011 年 1 月～2017 年 5 月）では，ある程度既存指標と似た動きを示しているように見えます。しかし，詳細に見ると，逆の動きをしている期間も多く見られます。全期間において相関係数を計算したところ，内閣府の「Diffusion Index（DI），全国値」とは 0.1849，「消費者態度指数（近畿）」とは 0.2187 となり，相関関係は大きくない，つまり既存の景況感指標の動きを捉えることができていない結果となってしまいました。

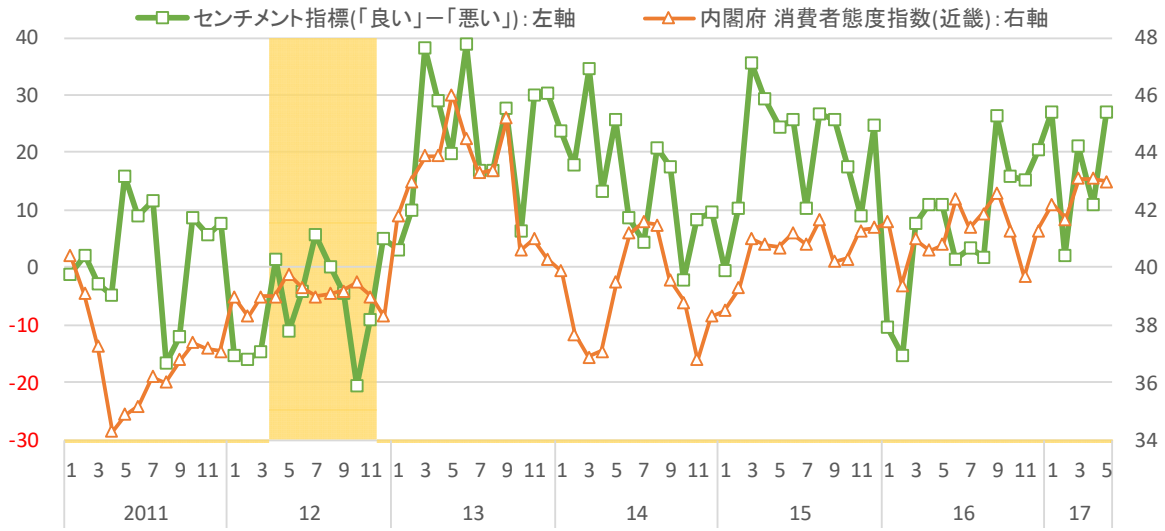
図表 4-12：「テキスト版センチメント指標」と景気動向指数との比較



(出所)内閣府「景気ウォッチャー調査,「景気動向指数」より筆者作成

テキスト版センチメント指標と内閣府の「景気動向指数（全国)」を並べたものである。同じ動きをしている期間も見られる一方で，逆の動きをしている期間も目立つ。両者の相関係数は 0.1849 であった。

図表 4-13：「テキスト版センチメント指標」と消費者態度指数（近畿）との比較



(出所)内閣府「景気ウォッチャー調査,「消費動向調査」より筆者作成

テキスト版センチメント指標と内閣府「消費者態度指数（近畿地域，原数値)」を並べたものである。2013 年 11 月～14 年 6 月のように逆の動きをしている期間が目立つ。両者の相関係数は 0.2187 だった。

図表 4-14：景気ウォッチャー調査のコメントにおける「高い」の使われ方

コメント１	イートインコーナーの利用率が相変わらず高い。今月から更に座席数を増やしたところ，それでも常に満席状態が続き，食品と飲料の売上増につながっている。
コメント２	予約の入込の動きが速く，価格の高い低いにかかわらず，動きが良い。
コメント３	客の様子では，単価が高い商品よりも安い商品を選ぶことが多く，以前からの節約志向は変わらない。
コメント４	単価の高い化粧品や一般食品，日用雑貨などは，良い物で安い商品が単品で購入される。
コメント５	商品単価の動きや，商品購入までの時間をみると，購買意欲が高いとは思えない。

(出所)内閣府「景気ウォッチャー調査」(2017 年 3 月調査）コメントから筆者作成

一例として,「高い」という単語がどのように使われているかを抜粋した。コメント 1 はプラスの景況感の得点を与えることができる一方で，コメント 5 は「高い」という言葉を「思えない」という言葉で打ち消しており，文章全体でみると，ネガティブな印象を受ける。よって，個々の単語ではなく，文章全体でプラスかマイナスか評価することが重要であるといえる。

4-4. なぜ既存の景況感指標と「ずれた」のか

残念ながら「テキスト版センチメント指標」は既存の景況感指標の動きとずれていました。このような「ずれ」はなぜ生じたのでしょうか。その理由としては，**①経済用語の特殊性と複雑性**，**②基にした「感情特性辞書」の限界**の 2 点が考えられます。

順番に説明していきましょう。まず，**①経済用語の特殊性と複雑性**ですが，景気ウォッチャー調査の「景気判断理由」に掲載されたコメントを詳細に確認すると，「高い」という言葉一つをとっても，様々な形で使われていることがわかります（**図表 4-14**)。例えば，コメント 1 は「利用率が相変わらず高い」ということですから，ポジティブな感情であり，プラスの景況感の得点を与えることができるでしょう。一方でコメント 5 を見ると，「購買意欲が高いとは思えない」となっています。「高い」という言葉が「思えない」という打ち消しの言葉によって否定されています。コメント 5 を一つの文章としてみると，ネガティブな印象を受けますので，プラスの得点を与えることは妥当ではありません。

もうおわかりでしょう。今回われわれが行ったセンチメント分析では，個々の単語に注目し，その単語が持つ方向性を確認し，プラスかマイナスの得点を与えていました。しかし，正確にコメントの「イイタイコト」を読み取るためには，個々の単語ではなく，文章全体で評価する必要があるということです。個々の単語に着目するか，文脈全体で評価するか，この違いが既存の景況感指標との大きなずれを生じさせていたと考えられます。

次に，**②基にした「感情特性辞書」の限界**については，今回用いた「単語感情極性対応表」

がもともと小説の表現など人文系のテキストデータを主たる分析対象として検討されたものであり，社会科学系のテキストデータへの適用については考えられていなかったためと考えられます。そのため，「良い」「高い」「低い」など，景況感を示すと思われる単語については，適切な得点が与えられていないものが散見されます。今回われわれは個々の単語の意味だけを見て得点を与えましたが，この方法が妥当かどうかについても検討が必要でしょう。場合によっては社会科学系のテキストデータを扱うことが可能な「感情極性辞書」を作成する必要もあるのではないでしょうか。これらの要因をより一般的な形で整理したものが**図表 4 -15** です。経済でよくみられる文章について，**①真逆の意味**，**②文末否定**，**③真逆の意味＋文末否定**，**④経済主体ごとに異なる印象**という 4 つに類型化しています。普段何気なく使っている用語の複雑さがわかっていただけるのではないでしょうか。

4-5. まとめと今後の展望

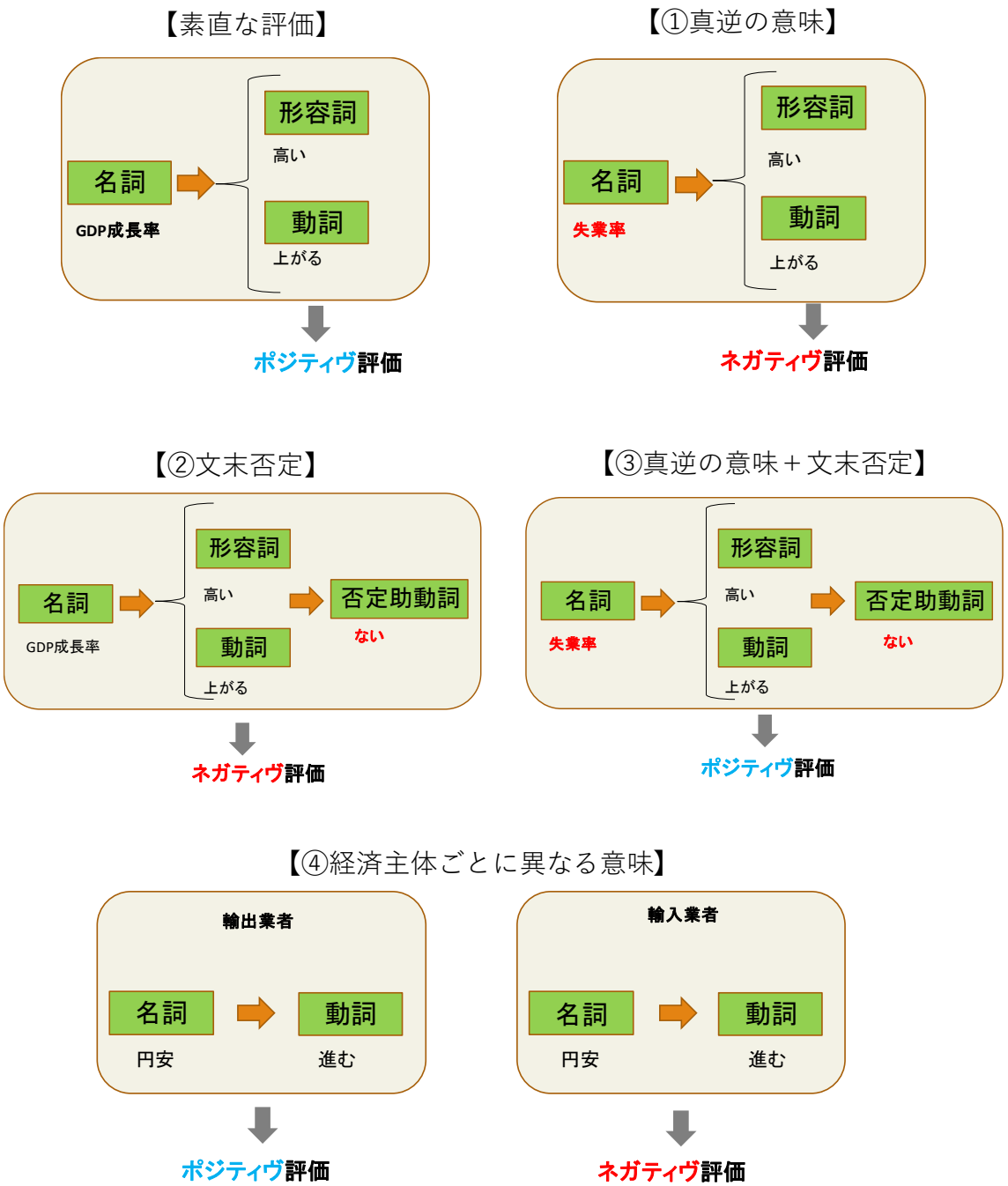
本章では，内閣府の「景気ウォッチャー調査」のテキストデータを用いて，「テキスト版センチメント指標」を作成し，既存の景気指標と比較を試みました。その結果，作成したセンチメント指標と既存の景気指標との間には大きな乖離が生じており，大幅な改良が必要であることがわかりました。

経済用語は複雑です。一つの経済現象でも，観察者の数だけ判断の数が存在します。今回われわれは「景気判断理由」に掲載されたコメントから，コメントの裏に隠された景況感を読み解こうとしましたが，良い結果は得られませんでした。繰り返しになりますが，コメントの「イイタイコト」を正確に読み取るためには，やはり個々の単語では不十分であり，文章全体でその意味を評価する必要があるといえるのではないでしょうか。

今後の展開としては，やはり経済用語の複雑性に配慮した分析手法の開発と適用が望まれるところです。その手掛かりは「景気ウォッチャー調査」の中で既に示されています。本章第 1 節で「景気ウォッチャー調査」の最大の特徴は，景気判断の方向性と共に，その理由を示した「景気判断理由集」が使えることだと述べました。判断理由を集めたテキストデータという「質」のデータだけでは，回答者が景況感の改善のつもりで答えているのか，悪化のつもりで回答しているのかわかりません。

ここで**図表 4-2** の下図をもう一度見てください。今回テキストデータとして用いた「追加説明および具体的状況の説明」の 3 列左に「景気の現状判断」という列があり，◎や○の記号が書かれています。全部で 5 つ（◎：良，○：やや良，□：不変，▲：やや悪，×：悪）の種類がありますが，これは回答者本人の景況感への評価を示しています。つまり，「景気判断理由集」の各コメントに 5 種類の景気判断の方向性という「量」のデータを一対一で対応させることで，機械学習で重要な「教師役」として使えるということです。文章全体をポジティブなのかネガティブなのか評価し，機械学習を通じてコンピュータ自身が学び，判断を精緻化していくと同時に景況感指標を改善していく。これにより，既存の統計との大きなずれを解消できる可能性があります。

図表 4-15：経済用語の複雑さの類型化



(出所)筆者作成

経済用語の複雑さについて，実際のコメントを基に，①真逆の意味，②文末否定，③真逆の意味＋文末否定，④経済主体ごとに異なる印象という 4 つに類型化した。個々の単語ではなく，どのような文脈で使われているか，「文章単位」で読み解くことが必要といえよう。

本稿の続きにあたる下巻（生田ほか 2020）では，先述した「テキスト版センチメント指標」の課題を克服するために，深層学習をテキストデータの解析に応用することで，景況感を推定する方法について報告します。そうした推定作業を理解するためには，本稿で学んだテキストデータの解析方法に加えて，新たに深層学習の知識が必要になります。具体的には，まずは深層学習を機能させるためのモデルである，単純なニューラルネットワークから解説を始めます。その中で，テキストデータがどのようにニューラルネットワークで処理されるのか学びます。そして，ニューラルネットワークを複雑にすることで，人間が経済関連のテキストを読んで景気を判断するような状態へとモデルを近づけます。このように，下巻では，深層学習について段階的に理解できるよう工夫して，最後に，テキストデータを利用した景況感の推定結果を述べます。

参考文献

生田祐介・関和広・松林洋一(2020，近刊)。

石田基広(2008)『R によるテキストマイニング入門』森北出版。

石田基広・金明哲編著（2012),『コーパスとテキストマイニング』共立出版。

石田基広(2012)「R で学ぶデータ・プログラミング入門 -R Studio を活用する-」共立出版。

伊藤新(2016)「政府の政策に関する不確実性と経済活動」RIETI Discussion Paper Series 16-J-016。

和泉潔・後藤卓・松井藤五郎(2009)「テキスト情報による金融市場の逐次外挿予測」『人工知能学会ファイナンスにおける人工知能応用研究会』SIG-FIN-03-02 pp6-14。

今井耕介(2018)「社会科学のためのデータ分析入門（上・下)」岩波書店。

大高一樹・菅和聖(2018)「機械学習による景気分析－「景気ウォッチャー」調査のテキストマイニング」，日本銀行ワーキングペーパーシリーズ No.18-J-8。

奥村学(2010)『自然言語処理の基礎』コロナ社。

奥村学(2010)『言語処理のための機械学習入門』コロナ社。

岡崎陽介・敦賀智裕（2015)「ビッグデータを用いた経済・物価分析について―研究事例のサーベイと景気ウォッチャー調査のテキスト分析の試み―」BOJ Reports & Research Papers。

工藤拓(2018)『形態素解析の理論と実装』近代科学社。

黒橋禎夫・柴田知秀(2016)『自然言語処理概論』サイエンス社。

小林雄一郎(2017)『R によるやさしいテキストマイニング』オーム社。

小林雄一郎(2017)『R によるやさしいテキストマイニング[機械学習編]』オーム社。

五島圭一・高橋大志（2017)「株式価格情報を用いた金融極性辞書の作成」『自然言語処理』Vol.24，No.4 pp547-577。

五島圭一・高橋大志・山田哲也（2019）「自然言語処理による景況感ニュース指数の構築とボラティリティ予測への応用」IMES Discussion Paper Series No.2019-J-3。
総務省(2012)『平成２４年版　情報通信白書』。
塩野剛士(2018)「人工知能とテキストデータを活用した数量分析」IMES Discussion Paper Series No.2018-J-9。
白井ゆかり(2014)「日本銀行の金融緩和とコミュニケーション政策～サーベイ調査に基づくレビュー～」，コロンビア大学における講演の邦訳（於，米国ニューヨーク市，2 月 27 日）日本銀行 2014 年 2 月 28 日。
高村大也・乾孝司・奥村学（2006)「スピンモデルによる単語の感情極性抽出」『情報処理学会論文誌』Vol.47 No.2 pp627-637。
高村大也(2010)『言語処理のための機械学習入門』コロナ社。
土屋誠司(2015)『はじめての自然言語処理』森北出版株式会社。
坪井祐太・海野裕也・鈴木潤（2017)『深層学習による自然言語処理』講談社。
照井伸彦(2018)「ビッグデータ統計解析入門－経済学部/経営学部で学ばない統計学－」日本評論社。
東京大学教養学部統計学教室編(1991)「統計学入門」東京大学出版会。
東京大学教養学部統計学教室編(1994)「人文・社会科学の統計学」東京大学出版会。
西山慶彦・新谷元嗣・川口大司・奥井亮(2019)『計量経済学』有斐閣。
日本統計学会編(2020)「改訂版　日本統計学会公式認定　統計検定３級対応　データの分析」東京図書。
日本統計学会編(2020)「改訂版　日本統計学会公式認定　統計検定４級対応　データの活用」東京図書。
樋口耕一(2014)『社会調査のための計量テキスト分析－内容分析の継承と発展を目指して』ナカニシヤ出版。
松本裕治・奥村学(2017)『コーパスと自然言語処理』朝倉書店。
三菱トラスト投資工学研究所（2018)『実践金融データサイエンス』日本経済新聞出版社。
山内長承(2018)「Python による統計分析入門」オーム社。
山本裕樹・松尾豊(2016)「景気ウォッチャー調査の深層学習を用いた金融レポートの指数化」The 30th Annual Conference of the Japanese Society for Artificial Intelligence.
Baker, Scott R., Nicholas Bloom, Steven J. Davis, (2016) “Measuring Economic Policy Uncertainty,” *The Quarterly Journal of Economics*.
Catalinac A, and K. Watanabe (2018) 「日本語の量的テキスト分析」未公刊。
Church, K, Gale, W., Hanks, P., Hindle, D. (1991) “Parsing, Word Associations and Typical Predicate-Argument Relations, ” in Tomita, M.(ed.) Current Issues in Parsing Technology, Kluwer Academic Publishers, Boston, MA.
Elif C. Arbatli, Steven J. Davis, A. Ito, N. Miake, and I.Saito, (2017) “Policy Uncertainty in

Japan," IMF Working Paper No.17/128.

Gentzkow, M., B. T. Kelly, and M. Taddy (2019) "Text as Data," *Journal of Economic Literature*.

Ishijima, H., (2014) "Quantifying Sentiment for the Japanese Economy as Predictors of Stock Prices," Columbia Business School CJEB Working Paper 338.

Kamihigashi, T., K. Seki, and M. Shibamoto, (2017) "Measuring Social Change Using Text Data: A Simple Distributional Approach," RIEB Discussion Paper Series DP2017-16.

MeCab : Yet Another Part-of-Speech and Morphological Analyzer, Available at taku910.github.io/mecab/

Shapiro, A. Hale, M. Sudhof, and D. Wilson, (2017) "Measuring news sentiment," Federal Reserve Bank of San Francisco Working Paper 2017-01, Federal Reserve Bank of San Francisco.

Stevens, S. S, (1946) "On the Theory of Scales of Measurement," *Science*, Vol. 103, No. 2684. pp. 677-680.