コラム
column

# データにもとづいた考え方とは(2)
## ―ことばのデータ分析―

東北学院大学　情報学部データサイエンス学科
教授　鈴木　努 氏

前回の記事では、身近なデータを使って論理的に考える方法として事例比較の方法を紹介しました。今回は、また別の観点から、私たちの身近なデータである「ことば」を取り上げたいと思います。現在、言語とデータサイエンスに関して最も注目されている話題はChatGPTのような対話型のAIや、そこで使われている大規模言語モデルの技術でしょう。これらの話題については岡野原（2023）などの丁寧な解説書がありますので詳しくはそちらを参照してほしいのですが、これらの技術を可能にしたのは、インターネット上などにある大量のテキストデータと計算機の性能の大幅な向上だといわれています。それによって、少し前までは難しいと考えられていた、人間のように言語を扱うことができるシステムが登場したのです。

ことばのデータを分析するというと、上で述べたような大量のテキストデータが必要と思うかもしれません。そのような大量のデータを集めたり分析したりすることは個人には無理です。逆に個人で集められる程度のテキストデータであれば、データ分析の方法を用いるまでもなく、実際に読んでみれば済むとも思われるでしょう。しかし、私たちの身の回りにあるような、それほど大量でないテキストデータでも、データ分析を通して見ることで、新たな視点や気づきを得る助けになります。

私たちのコミュニケーションや仕事の多くの場面でデジタル化が進んだことで、テキストデータは身近に多く存在するようになりました。SNSやニュースサイトの記事を集めれば立派なテキストデータになります（ただし、使用に際しては著作権法や利用規約に違反しないように注意する必要があります）。また、ビジネスにおいても、日々の業務の報告書など社内のさまざまな文書はデジタル化されているでしょうし、ウェブやメールで集められた顧客アンケートなどもデジタル化されています。手書きの文書でもスマートフォンで写真にとりOCR（光学的文字認識）により簡単にテキストデータ化できるアプリやサービスがあります。

アンケートなどのテキストデータを分析してくれるサービスもありますが、基本的な分析なら自分で行うのもそれほど難しくありません。注目するのは、ことばの「頻度」と「共起」です。頻度とは文字通り、ある言葉がどれくらい多く使われているかです。よく使われることばは、それだけ重要だったり、注目されたりしていることを表します。報告書であれば、業務において問題になっていることは何か、顧客アンケートであれば、消費者が重視していることは何かを知る手掛かりになります。

ことばの頻度を知るには、文章に登場する単語を数える必要がありますが、ここで問題があります。それは文章をひとつひとつの単語に切り分けなければならないことです。日本語の文章の場合、英語と違って単語と単語の間にスペースがありません。人間は自然に単語の区切りを認識して読むことができますが、機械的にカウントするにはまず単語に切り分ける必要があるのです。そこで用いられるのが形態素解析という方法です。

形態素解析とは、文章をより小さい単位に切り分けて調べる方法です。このとき、例えば文字1字ずつに切り分けると意味をもったまとまりになりません。そこで、意味のわかるなるべく小さいまとまり、例えば単語に切り分けていきます。ただし、単語の中にはさらに分割しても意味が分かるもの（「中／小／企業」など）がありますので、単語より細かい単位に分かれることもあります。これを形態素といいます。さらにそれらの品詞や活用形などの文法的な特徴を調べます。それによって活用によって違う語形になっている場合でも同じ単語であると判断したり、品詞の異なる同音異義語（名詞の「たとえ」と副詞の「たとえ」など）を区別したりすることができます。

テキストデータを形態素解析するための無料のソフトウェアはインターネットでいくつか公開されていますが、もっと簡単に使うことができるのが、文章を入力またはテキストデータをアップロードして形態素解析を実行してくれるウェブサイトです。ここでは国立国語研究所が公開している「Web茶まめ」（https://chamame.ninjal.ac.jp/）を使ってみましょう。この連載の第1回の記事のテキストデータを形態素解析した結果の一部を示したのが表1です。

表1　形態素解析の結果の一部

| 書字形（＝表層形） | 語彙素 | 語彙素読み | 品　詞 | 活用型 | 活用形 |
|---|---|---|---|---|---|
| データ | データ-data | データ | 名詞-普通名詞-一般 | | |
| に | に | ニ | 助詞-格助詞 | | |
| もとづい | 基づく | モトヅク | 動詞-一般 | 五段-カ行 | 連用形-イ音便 |
| た | た | タ | 助動詞 | 助動詞-タ | 連体形-一般 |
| 考え | 考える | カンガエル | 動詞-一般 | 下一段-ア行 | 連用形-一般 |
| 方 | 方 | カタ | 接尾辞-名詞的-一般 | | |

形態素解析の結果はExcelファイルとして出力できるので、Excelの集計機能を使って、使われている頻度の高いことばを名詞と動詞に分けて示したのが表2です。本文ではひらがなで書かれていたものが漢字になっているなどの表記法の違いはありますが、使われていることばを眺めてみると、前回の記事のテーマがなんとなく分かるのではないでしょうか。もちろん、数字などそれだけでは意味が分からないものも含まれていますので、実際には単純に頻度を見るだけではなく、内容に応じた取捨選択が必要です。

表２　使用頻度の高い名詞と動詞

| 名　詞 | 頻　度 | 動　詞 | 頻　度 |
|---|---|---|---|
| 商談 | 29 | 為る | 64 |
| ネクタイ -necktie | 24 | 居る | 21 |
| 事 | 24 | 有る | 17 |
| 一 | 20 | 出来る | 14 |
| データ -data | 19 | 言う | 13 |
| 二 | 15 | 因る | 7 |
| 成立 | 13 | 成る | 6 |
| 要因 | 13 | 考える | 5 |
| 事例 | 11 | 生かす | 5 |
| 条件 | 10 | 付ける | 5 |
| 成功 | 10 | 用いる | 5 |

このようにして文章中から頻度の高いことばを抽出することで、ニュースやアンケート、報告書などの中から、重要なことば、注目の集まっていることばを見つけ出すことができます。それらの多くは、ざっと読んだだけでも重要性の高さに気づくものでしょうが、その中に意外なことばが含まれていたら、それは今まで気づかなかったけれども実は重要なことばかもしれません。

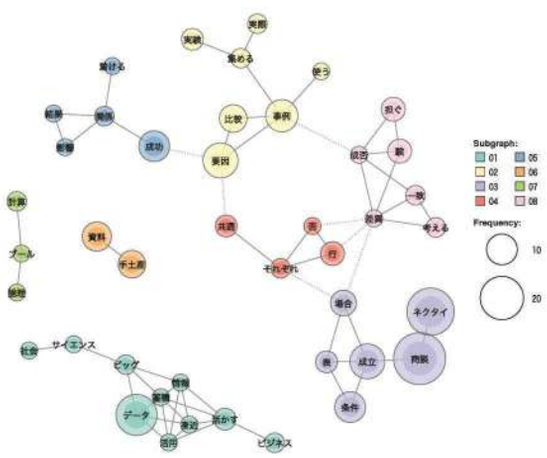
しかし、頻度だけでは分からないこともあります。例えば、顧客アンケートで「価格」がよく用いられていることが分かったとしても、それが「高い」と言われているのか「安い」と言われているのかは、いっしょに使われていることばを見なければ分かりません。つまり「共起」することばどうしの関係を知る必要があるということです。

ことばの共起関係を抽出して可視化する方法は少しややこしくなりますので、ここではテキスト分析の専用ソフトウェアであるKH Coder（3.Beta.07c）を使った例を示します。KH Coderについては開発者の樋口耕一先生（立命館大学）による解説書があります（樋口 2020）。

前回の記事の中でいっしょに使われることの多かった単語の共起ネットワークを示したのが図１です。頻度の高いことばは大きな丸で描かれているのが分かります。また、それだけでなく「商談」が「成立」する「条件」とか、「事例」の「要因」を「比較」するといったようにことばをつなげて解釈することができます。このように頻度の高いことばに着目するだけでなく、それらのことばどうしがどのように結びついているかを可視化できるのが共起ネットワークの利点です。また逆に、データでは結びついていないことばどうしをつなぐことで、新しい発想やアイデアが生まれるかもしれません。

データ分析というと数字で表された統計データの分析がすぐ思いつくでしょうが、私たちが日常で使っていることばも、データとして分析することで学術研究やビジネスに活かすことができるのです。

図１　共起ネットワークの例





**参考文献**

●岡野原大輔，2023，『大規模言語モデルは新たな知能か　ChatGPTが変えた世界』岩波書店.

●樋口耕一，2020，『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　第２版』ナカニシヤ出版.



〈プロフィール〉

福島県いわき市出身。東京都立大学で博士（社会学）取得後、東京工業大学、情報システム研究機構（国立情報学研究所）での博士研究員を経て、2013年より東北学院大学教養学部人間科学科教員。2023年、情報学部データサイエンス学科の開設とともに同学科教員として、主に社会調査やデータ分析に関する授業を担当。専門は人間関係などの「つながり」の構造を数理的に解析する社会ネットワーク分析。関心の中心は、人間・社会・文化・情報といった分野横断的な教育、研究。著書に『Rで学ぶデータサイエンス8 ネットワーク分析』（共立出版）、共著書に『テキスト計量の最前線―データ時代の社会知を拓く』（ひつじ書房）。