Text Mining Maniax

フリーソフトで始める日本語計量テキスト分析［増補版］

後藤和智
（後藤和智事務所 OffLine）

Text Mining Maniax

フリーソフトで始める 日本語計量テキスト分析 [増補版]

著：後藤和智（後藤和智事務所 OffLine）

初版発行：2016 年 8 月 14 日（コミックマーケット 90）
増補版発行：2021 年 12 月 31 日（コミックマーケット 99）

注意

1. 本 書 は、『Text Mining Maniax』 に、『Text Mining Maniax Advance』の第 2 章を第 5 章として増補したものです。
2. 本書を著作権法の定める私的使用の範囲外で公開などを行うことを禁じます。また、本書の使用により生じた問題についての責任は負いかねます。

はじめに

テキストマイニングはいいぞ。

というわけで57冊目の同人誌となります、後藤和智です。今回はいろいろな方面から期待されている(？)、テキストマイニングの解説書になります。テキストマイニングは一部では「計量テキスト分析」とも呼ばれていますね。2016年5月に参加した「第23回文学フリマ東京」(2016年5月1日、東京流通センター)では、少なくない一般参加者の方々(女性含む)から「テキストマイニングの解説書はまだですか」と言われてきました。2013年初旬に刊行された『統計学が最強の学問である』(西内啓：著、ダイヤモンド社、2013年)などで統計学が注目されて(とはいえ一時期よりは注目されなくなってしまっていますけど……)、その中で統計学の一分野であるテキストマイニングにも注目が集まっているように思えます。特に同イベントに参加された女性の参加者は、アニメのキャラクターソングを分析して、その特徴を分析したいということを熱心に語ってくれました。歌詞のテキストマイニングというのは私もそういう研究を読んだことがあり(もしかしたら本文中で触れるかもしれません)、興味を持っています。

簡単に説明すると、テキストマイニングとは、文章を対象としたデータマイニングの手法と言うことができます。とはいえそもそも「データマイニング」というものが、取り扱う範囲が多すぎてわかりづらい、という方もいるかもしれません(というか私もその一人です)。データマイニングは確かに魅力的なものではありますが、なんのために分析するのかという目的がはっきりしていないと、分析手法に逆に「使われてしまう」ことになりかねません。分析の仕方を知るためには、何よりデータの使い方を知っておくのが大切なのです。

本書では、主に「使用する文章の特徴を知る」という目的でのテキストマイニング、という目的に特化したテキストマイニングの解説を行おうと思います。その観点はこのような感じです。

・分析する文章のある部分(章など)に特徴的な単語はなんなのか。
・ある単語と同時に使われている単語(関係の強い単語)はなんなのか。
・単語を用いて文章をカテゴライズ、クラスタリングすることはできるのか。

私がテキストマイニングを始めたのは2013年3月に刊行した『統計学で解き明かす成人の日社説の変遷——平成日本若者論史5』(後藤和智事務所 OffLine、2013年(杜の奇跡20))ですが、複数の文章の解析によるクラスタリングと、クラスター分類に基づいた各クラスターの文章の特徴の確認というスタンスはほぼ一貫してとり続けています。最初に成人の日の社説を分析したときは、行ったのは単語の解析とクラスタリングだけでしたが、本書で採り上げるフリーのテキストマイニングソフト「KH Coder」を導入してからは、文章のクラスタリングはもとより、各クラスターにおける特徴の抽出や、コーディングを用いた分析、単語間の関係の

分析など、様々な分析ができるようになりました。ただ KH Coder にはできない機能もいくつかあり、それについては本書で紹介する RMeCab や TinyTextMiner を用いて分析することになるでしょう。

私見になりますがテキストマイニングの最大の醍醐味は異なる文章の間の比較にあります。私は 2014 年から若者論の著作のテキストマイニングを何回かやっているのですが、一冊だけの分析だと、たとえ違う章ごとの比較ができたとしても同じ著作の中のものなので分析が若干グダグダになりがちです。それに比べて複数の著作を比較する分析なら、著作ごとに論調の違いが分析することができて、分析する側としても楽しくなっています。

また、本書は、フリーソフトを用いて分析できることを解説しています。本書で採り上げる分析ソフトには「RMeCab」「KH Coder」「TinyTextMiner」がありますが、これらすべてがフリーソフトであり、さらにこれらを動かすためのソフトである R や RStuidio、MeCab もフリーソフトです。今まで私が『市民のための統計解析』シリーズや『R Maniax』シリーズ、そして東方 Project や艦隊これくしょんの数学系の講座系二次創作同人誌でフリーの統計ソフト「R」の使い方を解説してきたのは、フリーソフトを使うことによって統計学的な手法の普及を目的としたからです。その問題意識を受け継ぎ、本書でも使用するのはフリーソフトとします。

最後になりますが、本書が刊行される夏コミにおいて、本書の姉妹編である『艦娘たちの書斎――「艦これ」文学統計解析論序説』（後藤和智事務所 OffLine、2016 年。表紙は Azel 氏（シュレ猫 Online）が担当）を刊行します。これは「艦隊これくしょん～艦これ～」のノベライズ 4 シリーズ（「一航戦、出ます！」（鷹見一幸：著、電撃文庫）「陽炎、抜錨します！」（築地俊彦：著、ファミ通文庫）「鶴翼の絆」（内田弘樹：著、富士見ファンタジア文庫）「瑞の海、鳳の空」（むらさきゆきや：著、角川スニーカー文庫）。それぞれ 3 巻、7 巻、6 巻、3 巻の計 19 冊）のテキストマイニングを行った同人誌です。本書の執筆や教育用データの作成などと並行して行っているため、本書の「実践編」として読んでいただけると幸いです。

注

本書を作成した環境は次の通りです。本書は Windows ユーザー向けに作成しております。Macintosh、Linux などについては対応しておりませんのでご注意ください。

Windows 10 Professional 64bit
R 3.3.0
KH Coder 2.00f 及び 3.Alpha.07b
MeCab 0.996
Cabocha 0.53

目次

注：本書第 5 章には、2018 年刊行の『Text Mining Maniax Advance：R, KH Coder, Excel による計量テキスト分析の拡張』の第 2 章を収録しております。

第 1 章 ソフトのインストール

1.1 はじめに

まえがきでも述べた通り、本書はフリーソフトを用いて、日本語の文章を統計学的に解析する手法を解説する同人誌です。

私が『市民のための統計解析』シリーズや『R Maniax』シリーズ、そして『紅魔館の統計学なティータイム』などの統計学を取り扱った講座系二次創作同人誌でも触れているとおり、フリーの統計ソフト「R」の進化は凄まじいものがあります。R は（パッケージを使う必要がありますが）フリーソフトであるにもかかわらず商用ソフトには引けを取らない多彩な分析に対応し、ビッグデータの解析用に R をもとにした法人向けの商品が作られたこともありました。そして 2016 年には、商用版の R である「Revolution R Enterprise」の開発元を、なんとマイクロソフトが買収し、「Microsoft R Server」として供給するようになっているのです（http://www.atmarkit.co.jp/ait/articles/1601/08/news147.html）。

フリーソフトでも高度な統計解析ができるようになっています。そしてこのような状況は、統計学を市民の知として普及させるためには絶好の環境と言っていいでしょう。政治、経済、社会、医療、科学などのあらゆる分野において、物事を統計的に捉えるということは、よりよい社会の発展をもたらしたり、あるいは権力に対するチェック機能を強化したり、またあるいは現状をより性格に見通す指針として欠かせないものです。近年では歴史や文化、文学でさえも、統計学を用いて分析しようという機運が高まっております（例えば、エレツ・エイデン、ジャン＝バティースト・ミシェル：著、阪本芳久：訳『カルチャロミクス――文化をビッグデータで計測する』草思社、2016 年）。

私も、まえがきで述べた通り、弊サークルのメインフィールドである（はずの）若者論の研究に、段階的に統計学を導入しています。統計学を用いると、若者論者がどのような主張を行い、また他の論者とはどのような違いがあるのかということを知ることができて、質的、歴史的な研究や批判では行き詰まっていたものがより明確に見えるようになってきました。残念ながら若者論の研究というものそれ自体はあまり好意的には捉えられていないのですが、それでも統計学を用いることによって若者論研究の意義もより強く主張することができるようになったと考えております。

閑話休題、テキストマイニングには、2 つのプロセスがあります。第一に、文章を解析すること。これを形態素解析と言い、文章を「形態素」と呼ばれる要素に分けることを言います。形態素解析には幾つかの種類があり、例えば「僕は友達が少ない」という文を形態素に分けるとすると、次のような 3 パターンがあります。

文字：僕 / は / 友 / 達 / が / 少 / な / い
単語：僕 / は / 友達 / が / 少ない
品詞の種類：名詞 / 助詞 / 名詞 / 助詞 / 形容詞

本書で使うのは、文章を単語ごとに分ける解析です。この解析に使うソフトが「MeCab（めかぶ）」です。MeCab はフリーの日本語形態素解析ソフトで、京都大学と NTT コミュニケーション科学基礎研究所の共同開発で作られたものです。なぜ本書では MeCab を使用するのかというと、第一に無料で手に入れることができること、第二に本書で使うテキストマイニングツールである RMeCab、KH Coder、TinyTextMiner のすべてに対応していること、第三に辞書のカスタマイズが比較的容易であることが挙げられます。辞書のカスタマイズについては、専門的な語句や、あるいは表記の統一が必要な場合に行う必要があるからです。

また、テキストマイニングで用いる文章の解析には形態素解析以外にも構文（係り受け）の解析があります。これを行うためのソフトとして「CaboCha（かぼちゃ）」があります。係り受けの解析というのは、例えば「僕は友達が少ない」という文章には「僕（は）- 少ない」と「友達（が）- 少ない」という係り受けが存在することになります。これを用いた解析を行うこともできます（本書では触れませんが……）。

文章を形態素や構文に解析したら、次に行うのが分析です。分析には、単純な集計から、多変量解析まで様々な方法があります。手順としては、まず単語の頻度を見つつ、対応分析などで文章ごとに特徴的な語句を見たり、Jaccard 係数と呼ばれる係数を計算することによって単語の間のつながりを見たりという操作を指します。これについては次章以降で実例で紹介することにします。

テキストマイニングないし計量テキスト分析とは、こういった一連の流れを指します。文章を解析することにより、目で見ることができない「行間」を分析することができると筆者は捉えています。

本書では、無料で手に入れることができる文章を用いて分析を行います。ただし、無料で手に入ると言っても、元の文章の著作権は放棄されていないこともあるので、使用方法には注意する必要があります。

図 1.1 RMeCab で解析してみた例

1.2 下準備：R、MeCab をインストールする

本書で行うテキストマイニングの下準備として、テキストマイニングを行うためのツールをインストールしますが、その前にツールを動かすためのソフトをインストールします。ここでインストールするのは、フリーの統計ソフト「R」と、形態素解析ソフト「MeCab」です。

1.3.1 R

R はオープンソースで開発されているフリーソフトであり、パッケージを用いることで様々な分析を可能にします。本書で用いる RMeCab も、R のパッケージの一つです。まずは R 本体をインストールします。R の本体は、CRAN という R のサーバーから取得します。検索エンジンで「CRAN」と検索して、CRAN のサイトに移動します。

CRAN サイト：https://cran.r-project.org/

もしくは、CRAN のサイトから CRAN のミラーサイトに移動してそこからインストールするのもいいでしょう。ミラーサイトには、CRAN のサイトに左側にある「Mirrors」から各地のミラーサイトに飛ぶことができます。2016 年 7 月現在で、日本には統計数理研究所にミラーサイトがあります。

CRAN ミラーサイトの一例（日本・統計数理研究所）：https://cran.ism.ac.jp/

CRAN のサイトないしミラーサイトに飛んだら、トップページの「Download R for Windows」をクリックし（Macintosh や Linux 向けのダウンロード案内もありますが、本書では Windows で用いる場合を解説します）、さらに「base」をクリックして、一番上の「Download R X.X.X for Windows」をクリックしてインストーラーをダウンロードします。インストールするソフトはできるだけ最新のものを使いましょう。

インストーラーをダウンロードしたら、インストーラーを起動してソフトをインストールしますが、ここで一点注意（あるいは提案）しておきたいのは、「R などのフリーソフトは、Program Files フォルダにはインストールしない」ということです。というのも、これは特に MeCab で顕著なのですが、Program Files フォルダの中に入っているファイルは、管理者権限でないと中のファイルを編集できないという特徴があります。これは MeCab のように自分でファイルを変更する必要があったり（詳しくは第 4 章を参照）、あるいはソフト内のツールによらず直接ファイルをダウンロードする必要があったりするときは妨げになる可能性が高いからです。

本書で提案するのは、C ドライブ直下に「usr」フォルダ、さらに「local」フォルダを作ってそのフォルダにインストールするという方法です。この方法はかつて LaTex などのフリーソフトをインストールする際にこのフォルダに入れられていたという経緯があったというものがあります。このフォルダならファイルの編集なども容易であり、また位置もわかりやすいのでおすすめです（ただしそれは裏を返せばミスによるソフトの改変を防ぐことが難しいという

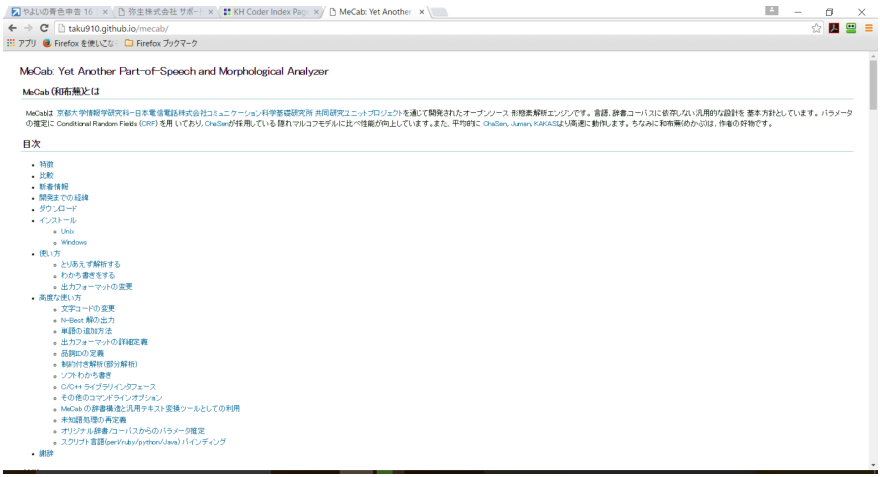

図 1.2 MeCab の公式サイト

ことでもあります。必要なファイル以外は編集しないよう心がける必要があります)。

統計ソフトとしてのRを使うなら、Rはテキストマイニング以外にも頻繁に使うと予想されるので、Windows8 以降のソフトを使っているなら画面下段のタスクバーにピン留めしておくことが望ましいです。Rを起動したら、Rのアイコンを右クリックして「タスクバーにピン留め」をクリックすれば、いつでもタスクバーからRを起動させることができます。

次に行うのは、パッケージのインストールです。Rを起動したら、メニューバーの「パッケージ」の「パッケージのインストール」をクリックします。CRANのミラーサイトの設定を求められたら、ミラーサイトを選択します。そしてどのパッケージを選択するかの画面になったら、下記の方法ですべてのパッケージを選択します。

1. 最上段のパッケージをクリックする。
2. スクロールバーを一番下に移動する。
3. Shift キーを押しながら、最下段のパッケージをクリックし、すべてのパッケージが選択された状態にする。
4. 「OK」をクリックする。

こうすることで、すべてのパッケージがインストールされます。ただし、全てのパッケージのサイズを合わせると数 GB あり、パッケージのインストールには時間がかかるので、この作業は仕事や学校、睡眠の間など、長い時間席を外すときに行うのが望ましいでしょう。

1.3.2 MeCab

次に MeCab をインストールします。MeCab は下記のサイトから取得することができます(検

索エンジンで「MeCab」と検索すると一番上に出てくるかと思います)。

MeCab 公式サイト：http://taku910.github.io/mecab/

MeCab は（2016 年 7 月 20 日現在）バージョン 0.998 で止まっているので、このバージョンをインストールします。メニューの「ダウンロード」をクリックし、「mecab-0.996.exe: ダウンロード」の「ダウンロード」をクリックすると、インストーラーをインストールすることができます。インストーラーでは、保存するフォルダを「c:\usr\local」にしてインストールしましょう。

1.3 RMeCab、KH Coder をインストールする

1.3.1 RMeCab

RMeCab は、石田基広によって開発されているパッケージで、MeCab を R 上で動かすことができるパッケージです。このパッケージは CRAN には登録されていないため、開発者のサイトから直接ダウンロードする必要があります。

とはいえ、ダウンロードするときは、開発者のサイトにアクセスする必要はなく、R に下記のコマンドを打ち込めばインストールすることができます。この作業を一回行えば、作業の度にインストールする必要はなくなります。インターネットが繋がっていない状態で使用することも可能です。

```
install.packages("RMeCab", repos = "http://rmecab.jp/R")
# ※注意：R は大文字と小文字を区別します。間違いないようにしましょう。
```

ただ、開発者のサイトには RMeCab の使い方が掲載されているので、ブックマークはしておいたほうがいいでしょう。

「R と Linux と ...」http://rmecab.jp/wiki/index.php?RMeCab

もう一つ行っておくと便利なのは、空の作業スペースを作っておくことです。空の作業スペースは、R を起動させて、なにも操作せずにメニューバーの「ファイル」から「作業スペースの保存」を選択して保存するというものです。これで空の作業をスペースを作った場合、この作業スペースを適当なファイルにコピーして開くと、そのフォルダが作業フォルダに設定されるという非常に便利な裏技を使うことができます。

RMeCab を起動させるには、R を起動させたのち、次のように入力します。

```
> library(RMeCab)
```

RMeCab が正しくインストールされている場合、このコマンドを打ち込むと RMeCab の機能を使うことができるようになります(これは R を起動するたびに入力する必要があります)。

図 1.3 RMeCab 作者のサイト

図 1.4 KH Coder の公式サイト

1.3.2 KH Coder

本書で頻繁に利用するテキストマイニングソフト「KH Coder」は、立命館大学の樋口耕一らによって開発されているソフトです。近年ではこのソフトを用いた研究も多く発表されており、2016 年 7 月現在で 1,200 件以上の研究で使用されています。こちらのソフトは現在も開発中であり、通常版であるバージョン 2 と、アルファ版であるバージョン 3 があります。

KH Coder 公式サイト：http://khc.sourceforge.net/

KH Coder は日本語や英語のテキストマイニングを行うソフトですが、バージョン 3 は、バージョン 2 からさらに分析結果の表示が拡張されていたり、バージョン 2 では対応していない韓国語や中国語、ロシア語などの分析に対応している一方で、環境によっては分析することができないなどの問題もあります。バージョン 3 でしかできない機能もいくつかありますが、バージョン 3 は正式版ではないので、バージョン 2 と 3 の両方をインストールする必要があるでしょう。

Windows であれば、KH Coder は本体のみインストールすれば大丈夫です（正確には、KH Coder を動かすための MySQL や R などのソフトは本体に同梱されている）。バージョン 2 はトップページの「KH Coder 安定版のダウンロード」から、バージョン 3 は「最新アルファ版のダウンロード」からダウンロードすることができます。

ただしサイト内でも警告されているとおり、ダウンロードサイトには紛らわしい広告がいくつもあります。ダウンロードサイトに着いたら、何もクリックせずに 5 秒ほど放置すれば、自動的にダウンロードが始まります。サイズはだいたい 400MB 程度であり、回線が遅いとダウンロードに失敗することもあるので注意しましょう。なお、バージョン 2 については「Vector」（http://www.vector.co.jp/）からインストールすることもできます。

ダウンロードしたファイルの中にある解凍ソフトを開くと、KH Coder がインストールされます。ただし初期状態では c ドライブ直下の「khcoder」（バージョン 3 の場合は「khcoder3」）フォルダに解凍されてしまうので、頻繁に使うフォルダの中に「khcoder」というフォルダを作って（フォルダ名は自由ですが）その中に解凍するようにしましょう。

KH Coder を起動するときは、「kh_coder.exe」というアプリケーションを直接開いて起動します。

1.4 そろえておくべき書籍など

1.4.1 Office もしくは OpenOffice.org など

ここでは、RMeCab ないし KH Coder を用いたテキストマイニングの際に必要になる、もしくは入れておいた方がいいソフトについて解説します。

まず必須なのが表計算ソフトです。表計算ソフトは、分析の結果を保存した csv ファイル（分析の結果は csv ファイルに保存した方が好ましいです）や Excel ファイル（KH Coder のみ）を閲覧するほか、一部の分析にも使いますし、中にはファイルの編集を行うときにも使います。そのため必須のソフトと言っていいでしょう。

推奨するのは（フリーソフトではありませんが）やはり Microsoft Office です。ファイルの編集についてはやはり Office が一番使いやすいと思います。Office は高いと思われるかもしれませんが、近年は月額 1,100 円程度、ないし年額 12,000 円程度で Office のソフトを使うことができる月額ないし年額課金制のサービス「Office 365」があり（solo と business がある）、これらのサービスでなら常に最新のバージョンを使用することができます。できるだけ高速のインターネット環境が必要ですが、導入コストも安いので是非とも導入しておくべきでしょう。

また Office には幾つかの類似・互換ソフトがありますが（Thinkfree Office、KINGSOFT Office など）、「OpenOffice.org（Apache OpenOffice）」や「Libre Office」はその中でも無料で使えるソフトであり、有償の互換ソフトを使うならこれらのフリーソフトを使うのが好ましいでしょう。なおこれらはあくまでも互換ソフトであり、操作性については本家 Office と異なることも多いので注意しましょう。

1.4.2 RStudio

RStudio は、オープンソースで開発されている、R の入力支援ソフトです。入力する部分（コンソール）と、現在使っている変数やグラフはもとより、ファイルの管理も一元的に行うことができるソフトです。

RStudio はいちどインストールしてしまえば、起動の度に最新の R を探してくれます（起動の度にファイルを選択する画面が現れることがありますが、キャンセルを押して無視してしまって結構です）。ただし注意すべきなのは、RStuido はコンソールの入力には日本語にも対応していますが、グラフの描画など日本語に対応していない部分もあるということです。特に KH Coder から出力した R のソースファイルを読み込んでグラフを確認したり、あるいはワードクラウドを描画したりすることができないということなので、この点は気をつけておく必要があるでしょう。

RStudio 公式サイト：https://www.rstudio.com/home/

1.4.3 その他のソフト

その他、分析に用いるテキストファイルの編集のために、Windows でデフォルトでインストールされているメモ帳やワードパッドよりも高度な検索・編集ができるソフトを入れておくとなおいいでしょう。

本書で推奨しているのが、シェアウェアのテキストエディタ「秀丸エディタ」です。秀丸エディタは正規表現による検索や置換が可能であるため、第 2 章で述べるような特定の語を含んだパターンの置換による編集が可能です。

1.4.4 書籍

最後に、本書の分析を行う上でそろえておくべき、あるいはそろえておいた方がいい書籍を案内します。

・樋口耕一『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、2014 年

KH Coder の開発者による解説書であり、KH Coder を用いる際には参考文献に同書（ないし同書に収録されている論文）を挙げることが推奨されています。同書には KH Coder の仕様や使い方はもとより、事例もいくつか書かれているので、KH Coder を使うなら是非とも手元に置いておくべきでしょう。また KH Coder を用いたテキストマイニングの解説書として、

石川慎一郎、前田忠彦、山崎誠（編）『言語研究のための統計入門』（くろしお出版、2010 年）などがあるようです。

・石田基広『R によるテキストマイニング入門』森北出版、2008 年

RMeCab の開発者による解説書です。公式サイトでは少ししか触れられていない RMeCab の各種機能について、詳細な形で触れられているので、RMeCab を用いるなら購入しておきましょう。ただし、RMeCab には、同書の出版後に追加、ないし修正された機能もあります。また姉妹編として『R で学ぶ日本語テキストマイニング』（ひつじ書房、2013 年）があります。

・その他、R のテキストマイニング関係の書籍

石田基広、金明哲『コーパスとテキストマイニング』（共立出版、2012 年）、石田基広ほか『R のパッケージおよびツールの作成と応用』（共立出版、2014 年）などが R のテキストマイニング関係の書籍として挙げられます。KH Coder に触れられているものもあります。

・テキストマイニングの実例が掲載されている書籍

近年では社会学関係の著作にもテキストマイニングが多く見られるようになりました。特に KH Coder については、『社会調査のための計量テキスト分析』が発売されてからは爆発的に広がっているように思えます。それらの中でも特におすすめなのが、高史明『レイシズムを解剖する――在日コリアンへの偏見とインターネット』（勁草書房、2015 年）です。これはネット上に見られるようなレイシズム（人種差別）について、KH Coder を用いて分析した結果を中心に研究されているもので、共起ネットワークの描画やコーディングを用いた分析など、実例として大変優れています。KH Coder を用いた分析の実例としては、中村隆志（編）『恋愛ドラマとケータイ』（青弓社、2014 年）に収録されている、谷本奈穂「ポピュラー音楽の歌詞における携帯電話の意味」などがあります。

また RMeCab や KH Coder を用いた分析ではありませんが、「エロマンガ統計」シリーズのサークル「でいひま」はテキストマイニングを用いた同人誌も幾つか刊行されています（牧田翠『一般人な俺と魔王な彼女のライトノベルが形態素的にこんなにエロいだなんて！？』でいひま、2012 年）。それらの同人誌は総集編である『エロマンガ統計 STARS』（でいひま、2014 年）に収録されております。

弊サークルの同人誌にもいくつかテキストマイニングを使ったものがあります。筆者としておすすめしておくのが、冒頭で採り上げた『統計学で解き明かす成人の日社説の変遷』の他、『都条例メディア規制の形成――平成日本若者論史 8』（後藤和智事務所 OffLine、2013 年）、『「劣化言説の時代」のメディアと論客：言論と論客の「再帰性」をめぐって――平成日本若者論史 Special』（後藤和智事務所 OffLine、2015 年）で、いずれも Kindle で配信されています。また、テキストマイニングの技術を用いた分析を含む同人誌として『東方キャラソート統計――多変量解析で見る「愛され方」の研究』（後藤和智事務所 OffLine、2016 年）があり、こちらはメロンブックス DL にて配信予定です。

14

・増補版に寄せて

本書の原書を刊行したのは 2016 年ですが、その後、私は様々な分野でテキストマイニングの同人誌を刊行しております。例えば、東方 Project 分野では、2017 年の「東方名華祭 / 幻想郷フォーラム」にて、『東方人気投票コメント統計：計量テキスト分析で見る「愛され方」の研究』を刊行し、翌年の名華祭では、口頭発表を行うと共に、いくつかのテーマに基づいた分析を行った『スカーレットの軌跡』も出しました。

また、2017 年からは、ゲーム「ステーションメモリーズ！」（駅メモ）のゲーム内の掲示板である「駅ノート」の分析を行い、路線や地域ごとの特徴を分析するということもやっております。これに関しましては弊サークルの「駅ノートの思い出」シリーズをご覧ください。

そのほか、近年はツイッター分析を通じて我が国の社会問題、得に性差別の問題についての分析を行っております。分析の解説書としては 2019 年刊行の『Twitter Analysis Maniax：twitteR、Excel VBA、KH Coder による最強（？）のツイッター分析』、また分析の実例については『ツイッターにおける女性差別についての考察』『続・ツイッターにおける女性差別についての考察』をご参照ください。

これらの同人誌は筆者の BOOTH（巻末参照）にて購入できるほか、評論と東方の同人誌については BOOK ☆ WALKER でも配信しております。

第 2 章 西田幾多郎『善の研究』を分析する

本章で習得すること…
・RMeCab の基本的な動作を覚える。
・RMeCab で単語数の分析と R-gram の出力が行えるようにする。
・KH Coder のデータの構造と基本的な動作を覚える。
・KH Coder で共起ネットワーク、対応分析、多次元尺度法、自己組織化マップの描画を行えるようにする。

2.1 はじめに

本書では、統計ソフトの使い方を、実例を通じて解説していきます。前章で述べたとおり、本書で分析する文章は、基本的に無料で手に入るものを使っております。ただし著作権などの関係もあり、画面に示すことができないこともあるのでご注意ください。

テキストマイニングの手慣らしのために、まずは比較的長い文章を解析してみましょう。無料で手に入る長めの文章を手に入れることができるサイトとして、「青空文庫」（http://www.aozora.gr.jp/）があります。知っている人は少なくないとは思いますが、このサイトは著作権が切れた文章を公開しているサイトで、このサイトの意義は大変大きいものだと思います。このサイトのために作業されている皆様には感謝の気持ちでいっぱいです。このサイトには、有名な文豪の小説や、評論、思想などといった分野に至るまで、様々な文章が掲載されています。

ただ、テキストマイニングに適しているのは比較的長文の、現代仮名遣いで書かれた評論文ですが（古典的仮名遣いについては若干分析が難しいかと思います）、青空文庫にはそのような条件を満たす文章があまりないのが現状です（あっても小説であったり、あるいは古典的仮名遣いであったり）。

今回分析するのは、我が国の代表的な哲学書である、西田幾多郎の『善の研究』（http://www.aozora.gr.jp/cards/000182/card946.html）です。こちらの文章は青空文庫に収録されている文章の中でも比較的長く、また現代仮名遣いと現代の漢字を使って書かれているので分析はしやすい部類に入ります。

この文章を解析するには、文章を分析に適した形に変換する必要があります。本章ではそこから解説していこうと思います。

2.2 KH Coder の形式にデータを修正する

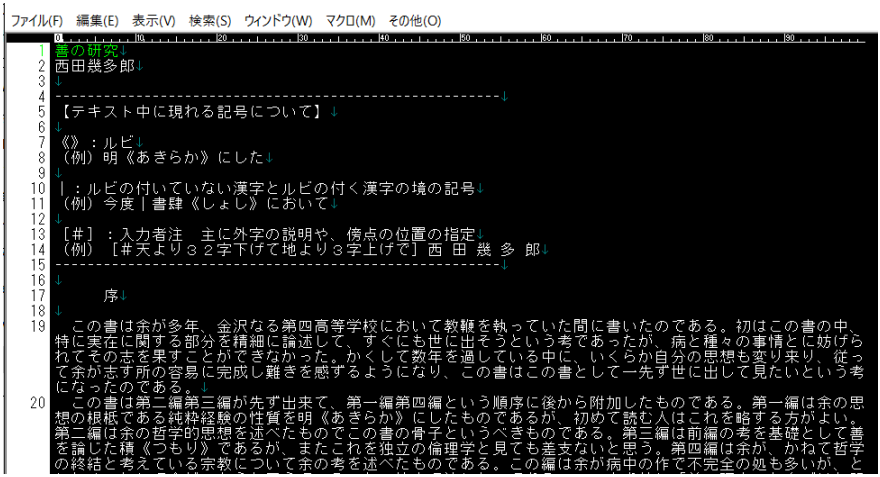

図 2.1 青空文庫の『善の研究』の元ファイル

青空文庫の文章は、基本的にルビ付きのテキストファイル、同じくルビ付きの html ファイル、そして電子書籍形式で公開されています。いずれもサイトからダウンロードは可能ですが、今回の分析ではテキストファイルを使います。ただし、分析に当たり、まずはルビを消す必要があります。

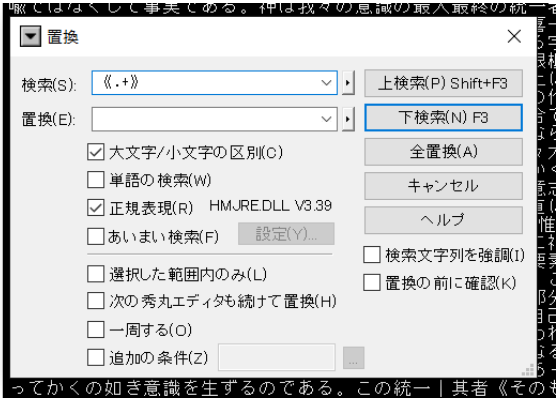

図 2.2 秀丸エディタによるルビの消去

青空文庫のテキストファイルでは、ルビがある部分は下記のような形式で表示されています。

> 経験するというのは事実｜其儘《そのまま》に知るの意である。全く自己の細工を棄てて、事実に従うて知るのである。純粋というのは、普通に経験といっている者もその実は何らかの思想を交えているから、毫《ごう》も思慮分別を加えない、真に経験其儘の状態をいうのである。たとえば、色を見、音を聞く刹那《せつな》、未だこれが外物の作用であるとか、我がこれを感じているとかいうような考のないのみならず、この色、この音は何であるという判断すら加わらない前をいうのである。（第１部「純粋経験」、第１章「純粋経験」より）

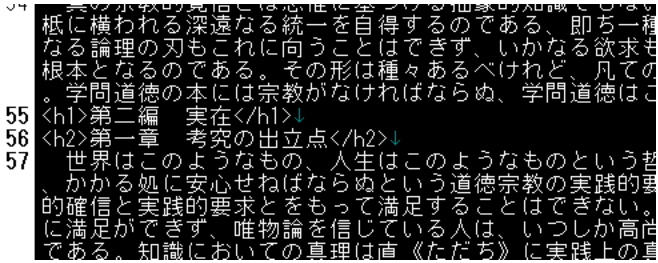

図 2.3 KH Coder 用のタグ付け

このように、ルビは「《》」（二重山括弧）の中に入った形で書かれています。これが残置されていると分析の邪魔になるので、まずはこれを除去する必要があります。この作業を行うには、正規表現による置換を行うことができるテキストファイルを使うのがいいでしょう。ここでは、秀丸エディタを用いてルビを除去する方法を解説します。

秀丸エディタでは、下記のように入力して、ルビを取り除くことができます（図 2-2）。

置換する文字列：《.+》
置換後の文字列：（何も入力しない）

秀丸エディタの正規表現では、「.」（ピリオド）は「任意の文字」、「+」は「1 回以上繰り返し」を指します。これを入力することで、ルビを一気に消すことができます。同様に、字下げ（［＃天より３２字下げて地より３字上げで］など）や改ページ（［＃改ページ］）示す記号なども消してしまいます。

次に、このテキストを KH Coder で読み込める形式に変換します。KH Coder で読み取るためのテキストファイルは、次のような構造にする必要があります。

```
<h1>大見出し 1</h1>
....
<h2>中見出し 1-1</h2>
....
<h2>中見出し 1-2</h2>
....
<h3>小見出し 1-2-1</h3>
....
<h1>大見出し 2
....
```

このような感じの階層構造にする必要があります（字下げなどをする必要はありません。なお最大 h5 までつけることができます）。本章では、各編を h1、各章を h2 とした構造にしてみます。この作業は、残念ながら手作業で行う必要があります……。各編に「<h1> 編タイトル </h1>」、各章に「<h2> 章タイトル </h2>」というふうにタグをつける作業をすると、このようになるはずです。

<h1> 序 </h1>
　この書は余が多年、金沢なる第四高等学校において教鞭を執っていた間に書いたのである。.....
<h1> 再版の序 </h1>
　この書を出版してから既に十年余の歳月を経たのであるが、この書を書いたのはそれよりもなお幾年の昔であった。京都に来てから読書と思索とに専なることを得て、余もいくらか余の思想を洗練し豊富にすることを得た。.....
<h1> 版を新にするに当って </h1>
　この書刷行を重ねること多く、文字も往々鮮明を欠くものがあるようになったので、今度｜書肆として置くの外はない。
　今日から見れば、この書の立場は意識の立場であり、心理主義的とも考えられるであろう。然の立場を介して絶対意志の立場に進み、更に「働くものから見るものへ」の後半において、ギリシャ哲学を介し、一転して「場所」の考に至った。.....
<h1> 第一編　純粋経験 </h1>
<h2> 第一編第一章　純粋経験 </h2>
　経験するというのは事実｜其儘に純粋の経験ではない。真の純粋経験は何らの意味もない、事実其儘の現在意識あるのみである。
　右にいったような意味において、如何なる精神現象が純粋経験の事実であるか。感覚や知覚がこれに属することは誰も異論はあるまい。.....
<h2> 第一編第二章　思惟 </h2>
　思惟というのは心理学から見れば、表象間の関係を定めこれを統一する作用である。その最も単一なる形は判断であって、即ち二つの表象の関係を定め、これを結合するのである。しかし我々は判断において二つの独立なる表象を結合するのではなく、かえって或一つの全き表象を分析するのである。.....
<h2> 第一編第三章　意志 </h2>
　余は今純粋経験の立脚地より意志の性質を論じ、知と意との関係を明て意志の目的という者も直接にこれを見れば、やはり意識内の事実である、我々はいつでも自己の状態を意志するのである、意志には内面的と外面的との区別はないのである。.....
<h2> 第一編第四章　知的直観 </h2>
.....
<h1> 第二編　実在 </h1>
<h2> 第二編第一章　考究の出立点 </h2>
　世界はこのようなもの、人生はこのようなものという哲学的世界観および人生観と、人間はかくせねばならぬ、かかる処に安心せねばならぬという道徳宗教の実践的要求とは密接の関係を持っている。.....
（以下略）