
東方
一番
きっかけ
これら
期待
流れる
出
東方人気投票
コメント分析で学ぶ
計量テキスト分析
著・後藤和智事務所offline　表紙イラスト・ラパメリ

東方人気投票コメント分析で学ぶ
計量テキスト分析

月刊テキストマイニングレポート Vol.25
2019年10月6日号

著：後藤和智（後藤和智事務所 Offline）
表紙イラスト：ラパメリ（ラパパメリカ）
発行：2019年10月6日
（第6回博麗神社秋季例大祭）

注意

1. 本書は、同人サークル「上海アリス幻樂団」の作品「東方Project」の二次創作作品です。本書は東方Projectの二次創作ガイドラインに従って製作されているものであり、また著者と原作作者及び作者のサークルとは一切関係がありません。そのほか、登場人物の口調などが原作と異なる場合があります。
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

霧雨魔理沙（以下、魔理沙）：このたびは「後藤和智事務所 OffLine」82 冊目の同人誌を手に取ってくれてありがとう。本書は、改めて東方 Project の二次創作で、本書の版元サークルが頻繁に活用している「計量テキスト分析」、テキストマイニングについて紹介するというものだ。ここまで、2017 年に刊行した『東方人気投票コメント統計』、2018 年の『スカーレットの軌跡』と、「東方 Project 人気投票」のコメントを計量テキスト分析を使って分析した同人誌を出してきたが、評論分野での計量テキスト分析の解説書は複数冊あるが、東方の同人誌を読む人向けの計量テキスト分析の解説書がなかったため、改めて製作しようと思った次第だ。

アリス・マーガトロイド（以下、アリス）：正直このサークルの同人誌って、学術分野ではそろそろネタ切れが迫ってきたのよね。あと、いろいろと都合がつかないこともあって、かなり新刊 RTA 状態で書いてしまっているのよねぇ。

古明地こいし（以下、こいし）：あれ？　この本って、82 冊目なの？　81 冊目じゃなくて？

パチュリー・ノーレッジ（以下、パチュリー）：それについては私が説明しよう。元々 81 冊目は、2019 年 3 月に、岩手県にある JR 山田線の宮古〜釜石間が復旧し、三陸鉄道に移管された記念の乗車レポート「三陸鉄道　ゆくぞ 35 周年」（仮題）と、2019 年 2 月から 3 月にかけて多くの部分が開通し、全線開通を間近に控えた三陸自動車道のレポート本「V やねん！三陸自動車道　全線開通目前号」（仮題）のリバーシブル本『岩手・宮城　東日本大震災沿岸被災地の現在』になる予定だったんだ。それが、4 月 6 日に三鉄に乗ろうとしたところ、強風でダイヤが大幅に乱れて、大船渡から予約していたバスに乗れなくなったことがわかり断念、三陸自動車道単独で書こうとしたけれど、やっぱり岩手、宮城の現状を伝えるなら三鉄と一緒にしなければならないと思い断念したという経緯がある。

古明地さとり（以下、さとり）：あのとき三鉄に乗れていれば、2019 年 4 月 21 日の駅メモ（「ステーションメモリーズ！」モバイルファクトリー、2014 年〜）のオンリーイベント「フットバーしま〜す！ 8」で、文学フリマ新刊の先行頒布扱いとして頒布する予定だったのに、それが潰れてしまったのですものね。まあそういう経緯があったとはいえ、番号順通りに出せないのは、このサークルの「月刊テキストマイニングレポート」もそうなのですが。

魔理沙：というわけで今回は、2019 年まで「東方 wiki」の管理人の主催で行われていた「東方 Project 人気投票」（第 10 回まで「東方シリーズ人気投票」）のコメントの分析を通じて、このサークルが行っている計量テキスト分析、テキストマイニングについての簡単な解説を行うことにしたい。なお、この人気投票は、2020 年に予定されている回からは他の有志に引き継がれることが既に決まっている。

アリス：何回か東方や評論、駅メモなどの同人誌で解説してきたと思うけど、このサークルがテキストマイニングに取りかかるようになったのは 2013 年のことね。元々、懇意にさせてもらってるサークル「でいひま」の牧田翠さんが、「エロマンガ統計」シリーズの一つとして『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！？』を 2012 年の冬コミで出していて、それに影響を受けたのよね。そこから、

2013 年の初頭に、同人誌『統計学で解き明かす成人の日社説の変遷』や、津田大介さんが主催するメールマガジンで簡単なテキストマイニングを使った論考を書いていたんだけど、本格的に始めたのは 2014 年かな。

パチュリー：2014 年に本格的にテキストマイニングに取りかかるようになったのは、本書で紹介するフリーソフト「KH Coder」に触れたのがきっかけだね。どういう経緯かは本書の著者は忘れてしまったのだけど、ツイッターか何かを通じて KH Coder に触れ、『「ヤンキー」論の奇妙な位相』やなどで KH Coder を使い、それ以降の評論本はほとんどがテキストマイニングになってしまったくらいだ。2015 年以降の評論本でテキストマイニング要素がないのは、福島県の「酪王カフェオレ」を扱った『酪 Fan』くらいじゃない？

こいし：そもそも KH Coder ってなんなの？　この本の筆者さんが、「このソフトを使えば確実にテキストマイニング沼にはまる」ってツイッターか何かで言ってたような気がするんだけど。

魔理沙：KH Coder というのは、立命館大学の樋口耕一氏らが、1999 年から開発を進め、2000 年から公開した無料のテキストマイニングソフトだ。ちなみに本書の著者が KH Coder を知った 2014 年は、それまで公式サイトの PDF や樋口氏による学術論文でしか示されていなかった KH Coder の使い方が、書籍としてみられるようになった年でもある。『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めざして』（ナカニシヤ出版、2014 年）だな。恐らくこれをきっかけに、学術書や論文でも KH Coder を使ったものが増え、さらに元 nifty で、現在はイッツコミュニケーションが運営している「デイリーポータル Z」、ITMedia の「ねとらぼ」、そして文藝春秋の「文春オンライン」など、ネットで KH Coder を用いた記事を書くライターも出てくるようになった。本書の著者も講談社の「現代ビジネス」やサイゾーの「wezzy」で KH Coder を用いた記事を書いている。KH Coder の登場が、文章を分析して、書くという行為のハードルを大幅に引き下げたというのは論を俟たないだろうな。

さとり：ライターの大山顕氏は、「マンションポエム」と呼ばれる分譲マンションの宣伝文句を、それまでは普通に読み込んでいたものを、最近は KH Coder を使って量的に分析するようになっているわね。ネット上の KH Coder を使った記事の多くは、そのような宣伝コピーだったり、あるいはアニメやアイドルの楽曲の歌詞だったりすることが多く、本書の著者のように社会問題の解明に応用するのは少ないけれど、まあそれは問題意識の違いであって、手法は同じだからいろいろと学べることも多いわ。

魔理沙：本書の著者なんか、最近は書籍をまるごとスキャンして OCR にかけて、それを 100 冊も 200 冊も分析するという狂気の沙汰のような同人誌を作っているもんな。分析にかかる時間の長さが喜び、とか言っていやがる。何を考えているんだか。

パチュリー：KH Coder のツイッターアカウント（@khcoder）は、KH Coder を使った学術研究やネット上の記事を盛んに紹介したりリツイートしたりしているから、KH Coder を使いたいなら是非ともフォローしておくといい。また、J-STAGE では、樋口氏による「計量テキスト分析および KH Coder の利用状況と展望」（『社会学評論』第 87 巻 3 号、2017 年 https://www.jstage.jst.go.jp/article/jsr/68/3/68_334/_article/-char/ja）

3

という、KH Coder での主に社会学における使われ方を概観した論文が無料で公開されているから、呼んでみるといいよ。

魔理沙：というわけで本書は、多くの人をテキストマイニング沼に招致すべく書かれたものだ。だいたいの手法については『Text Mining Maniax』で書いているが、ここでは KH Coder に特化した説明をすることとしたい。

目　　次

第1章 分析の下準備

1.1 KH Coder と MeCab のインストール

こいし：で、分析しようにも、KH Coder がないと始まらないよね。

魔理沙:そうだな。というわけで、まずは KH Coder のインストールから始めたい。とはいえ、基本は「公式サイトからダウンロードして、解凍する」だけだ。まずは、次の公式サイトにアクセスする。

http://khcoder.net/

魔理沙：ここの「KH Coder 3（最新版）ダウンロード」をクリックして、その先の Windows 版パッケージのリンクを押せば、そのままダウンロードされる。ダウンロードした exe ファイルをそのまま実行して、「Unzip」というボタンを押して解凍すればいい。高度なカスタマイズのために、保存する場所は、C ドライブ直下の「c:\khcoder3」から動かさない方が身のためだ。KH Coder をインストールしたら、MeCab もインストールしておきたい。

さとり：ちなみに「MeCab」は、文章を単語に分ける形態素解析と呼ばれる方法を行うため

図 1-1 KH Coder の公式サイト

のソフトのひとつで、同じく形態素解析のソフトである「茶筌」と共に KH Coder に既にインストールされているわ。

こいし：だったらわざわざインストールする必要なんてないじゃん。

さとり：改めて MeCab をインストールする理由は、KH Coder にインストールされている MeCab は、Neologd という新語や流行語に対応した辞書を使うためのものにして、別にインストールした MeCab は、東方の人気投票などといった、固有名詞などの一般に使われない専門用語が多いものの分析をするという、要するに「二刀流」的な使い方をするためね。

パチュリー：もっとも、これについては、本書の著者が KH Coder に既にインストールされている MeCab をカスタマイズする方法を知らないことでもあるんだけどね。KH Coder の中の MeCab は Unicode で、Windows10 の標準の文字コードは Shift-JIS だからというのもあるのかもしれないけど。

アリス：Neologd の導入については、本書の著者が『Twitter Analysis Maniax——twitteR, Excel, KH Coder による最強（？）のツイッター分析』（後藤和智事務所 OffLine、2019 年 / コミックマーケット 98）で解説しているので、ここでは省略します。BOOK ☆ WALKER で電子版も配信しているので、ご参照ください。

魔理沙：MeCab は次の公式サイトからインストールする。

https://taku910.github.io/mecab/

魔理沙：2019 年 9 月 20 日現在で、最も新しいバージョンは、2013 年 2 月 18 日に公開された 0.996 だ。「ダウンロード」のところから「mecab-0.996.exe」というファイルをダウンロードして、解凍する。解凍する場所は、カスタマイズのしやすさを考えて、C ドライブ直下に「usr」というフォルダをつくり、その下に「local」、さらにその下に「mecab」というフォルダを作って、そこに解凍するのが好ましいぜ。デフォルトだと Program Files、または Program Files（x86）というフォルダに解凍されるが、Program Files のフォルダは管理者権限がないとカスタマイズできないからな。そして、「mecab」というフォルダに「bin」や「dic」などのフォルダができれば成功だ。

1.2 辞書をカスタマイズする

パチュリー：改めて述べるけど、KH Coder に MeCab が既に入っているのに、改めて MeCab を別にインストールする必要があるのは、辞書をカスタマイズするための MeCab を別に用意しておいた方が分析にはちょうどいいというものがある。

こいし：辞書のカスタマイズって、それって解析……だっけ？　で出てくる単語を操作できるっていうこと？

アリス：そうよ。例えば医療系の学術論文と、東方の人気投票のコメントと、「艦隊これくしょん」のノベライズでは、出てくる用語が違うじゃない。だから、必要に応じてカスタマイ

6

ズする必要があるのよ。

魔理沙：辞書のカスタマイズは、Windows のコマンドプロンプトで行う。ここまでの説明通り行っていれば、MeCab の実行ファイル mecab.exe は「c:\usr\local\mecab\bin」というフォルダに入っているはずだから、次のように入力する。

```
cd c:¥usr¥local¥mecab¥bin
```

魔理沙：MeCab 辞書の設定ファイルである ipadic は、「c:\usr\local\mecab\dic\ipadic」の中に入っている。そしてこの中には様々な辞書ファイルが入っているから、中の辞書ファイルを参考にして自分で辞書を作り、それを「c:\usr\local\mecab\bin」に CSV ファイルとして保存する。CSV ファイルというのが重要だ。仮に「dictionary.csv」、出力する辞書を「dictionary.dic」としようか。そして、コマンドプロンプトに次のように入力する。

```
mecab-dict-index -d c:¥usr¥local¥mecab¥dic¥ipadic -u dictionary.dic dictionary.csv
```

魔理沙：こうすれば、bin フォルダの中に「dictionary.dic」という辞書ファイルが生成される。あとは、ipadic のフォルダにある、拡張子なしの設定ファイル「ipadic」に、次の行を書き加えれば、カスタマイズした辞書が使えるようになるぜ。

```
userdic=c:\usr\local\mecab\bin\dictionary.dic
```

パチュリー：高度な内容になるから本書では説明しないけど、KH Coder の中の MeCab をそのままにしておく必要があるのは、カスタマイズした MeCab とは別に、MeCab の強化版である mecab-ipadic-NEologd という、ネット上を中心とした新語に対応した辞書を使えるようにして、それと使い分けるという目的がある。必要に応じて、自分でカスタマイズした辞書と、mecab-ipadic-NEologd を適用した辞書を使い分けるという、いわば二刀流にした方が、様々な分析に対応できると思うんだ。

魔理沙：インストールした MeCab を使うには、「プロジェクト」から「設定」を選び、「mecab.exe のパス」のところに「c:\usr\local\mecab\bin\mecab.exe」と入力して、「unicode 辞書」のチェックは外す。逆に、KH Coder に既に入っている MeCab を使う場合は、チェックは入れておく。

1.3 MeCab 辞書のカスタマイズの詳細について

魔理沙：MeCab に登録するための辞書の構造を知りたいなら、MeCab が入っているフォルダの中にある dic フォルダのさらに中にある ipadic に入っている CSV ファイルを適当に開いてみるといい。図に示したのは、形容詞の場合の辞書だ。

表1-1「可愛い」の表記統一

かわいい	43	43	1000	形容詞	自立	*	*	形容詞・イ段	基本形	可愛い	カワイイ	カワイイ
かわい	45	45	1000	形容詞	自立	*	*	形容詞・イ段	文語基本形	可愛い	カワイ	カワイ
かわいから	47	47	1000	形容詞	自立	*	*	形容詞・イ段	未然ヌ接続	可愛い	カワイカラ	カワイカラ
かわいかろ	46	46	1000	形容詞	自立	*	*	形容詞・イ段	未然ウ接続	可愛い	カワイカロ	カワイカロ
かわいかっ	50	50	1000	形容詞	自立	*	*	形容詞・イ段	連用タ接続	可愛い	カワイカッ	カワイカッ
かわいく	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイク	カワイク
かわいくっ	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイクッ	カワイクッ
かわいゅう	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュウ	カワイュー
かわいゅぅ	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュゥ	カワイュー
かわいき	44	44	1000	形容詞	自立	*	*	形容詞・イ段	体言接続	可愛い	カワイキ	カワイキ
かわいけれ	40	40	1000	形容詞	自立	*	*	形容詞・イ段	仮定形	可愛い	カワイケレ	カワイケレ
かわいかれ	48	48	1000	形容詞	自立	*	*	形容詞・イ段	命令ｅ	可愛い	カワイカレ	カワイカレ
かわいけりゃ	41	41	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約１	可愛い	カワイケリャ	カワイケリャ
かわいきゃ	42	42	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約２	可愛い	カワイキャ	カワイキャ
かわい	39	39	1000	形容詞	自立	*	*	形容詞・イ段	ガル接続	可愛い	カワイ	カワイ
可愛い	43	43	1000	形容詞	自立	*	*	形容詞・イ段	基本形	可愛い	カワイイ	カワイイ
可愛	45	45	1000	形容詞	自立	*	*	形容詞・イ段	文語基本形	可愛い	カワイ	カワイ
可愛から	47	47	1000	形容詞	自立	*	*	形容詞・イ段	未然ヌ接続	可愛い	カワイカラ	カワイカラ
可愛かろ	46	46	1000	形容詞	自立	*	*	形容詞・イ段	未然ウ接続	可愛い	カワイカロ	カワイカロ
可愛かっ	50	50	1000	形容詞	自立	*	*	形容詞・イ段	連用タ接続	可愛い	カワイカッ	カワイカッ
可愛く	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイク	カワイク
可愛くっ	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイクッ	カワイクッ
可愛ゅう	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュウ	カワイュー
可愛ゅぅ	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュゥ	カワイュー
可愛き	44	44	1000	形容詞	自立	*	*	形容詞・イ段	体言接続	可愛い	カワイキ	カワイキ
可愛けれ	40	40	1000	形容詞	自立	*	*	形容詞・イ段	仮定形	可愛い	カワイケレ	カワイケレ
可愛かれ	48	48	1000	形容詞	自立	*	*	形容詞・イ段	命令ｅ	可愛い	カワイカレ	カワイカレ
可愛けりゃ	41	41	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約１	可愛い	カワイケリャ	カワイケリャ
可愛きゃ	42	42	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約２	可愛い	カワイキャ	カワイキャ
可愛	39	39	1000	形容詞	自立	*	*	形容詞・イ段	ガル接続	可愛い	カワイ	カワイ
かわゆい	43	43	1000	形容詞	自立	*	*	形容詞・イ段	基本形	可愛い	カワイイ	カワイイ
かわゆ	45	45	1000	形容詞	自立	*	*	形容詞・イ段	文語基本形	可愛い	カワイ	カワイ
かわゆから	47	47	1000	形容詞	自立	*	*	形容詞・イ段	未然ヌ接続	可愛い	カワイカラ	カワイカラ
かわゆかろ	46	46	1000	形容詞	自立	*	*	形容詞・イ段	未然ウ接続	可愛い	カワイカロ	カワイカロ
かわゆかっ	50	50	1000	形容詞	自立	*	*	形容詞・イ段	連用タ接続	可愛い	カワイカッ	カワイカッ
かわゆく	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイク	カワイク
かわゆくっ	51	51	1000	形容詞	自立	*	*	形容詞・イ段	連用テ接続	可愛い	カワイクッ	カワイクッ
かわゆゅう	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュウ	カワイュー
かわゆゅぅ	49	49	1000	形容詞	自立	*	*	形容詞・イ段	連用ゴザイ接続	可愛い	カワイュゥ	カワイュー
かわゆき	44	44	1000	形容詞	自立	*	*	形容詞・イ段	体言接続	可愛い	カワイキ	カワイキ
かわゆけれ	40	40	1000	形容詞	自立	*	*	形容詞・イ段	仮定形	可愛い	カワイケレ	カワイケレ
かわゆかれ	48	48	1000	形容詞	自立	*	*	形容詞・イ段	命令ｅ	可愛い	カワイカレ	カワイカレ
かわゆけりゃ	41	41	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約１	可愛い	カワイケリャ	カワイケリャ
かわゆきゃ	42	42	1000	形容詞	自立	*	*	形容詞・イ段	仮定縮約２	可愛い	カワイキャ	カワイキャ
かわゆ	39	39	1000	形容詞	自立	*	*	形容詞・イ段	ガル接続	可愛い	カワイ	カワイ
かわゆす	43	43	1000	形容詞	自立	*	*	形容詞・イ段	基本形	可愛い	カワユス	カワユス
カワユス	43	43	1000	形容詞	自立	*	*	形容詞・イ段	基本形	可愛い	カワユス	カワユス

こいし：これって、形容詞とか動詞みたいな活用がある単語を登録するときには、活用形を全部登録しなきゃいけないの！？　めんどくさいよー！

魔理沙：まあそう言いたくなる気持ちもわからんでもないが、そうしなきゃいけないんだよ。ただ、こういう構造になっていることにより、異なる表記のものをまとめることができるんだ。例えば、東方の人気投票のコメントにおいて頻出するのは「かわいい」という言葉だが、表記としては「かわいい」のほか「可愛い」「カワイイ」「カワユス」などというものが挙げられるよな。その場合は、それぞれの活用形を登録し、K 列に「可愛い」と書け

8