絵入源氏物語の統計的な テキスト解析 web アプリケーションの設計

土山 玄（同志社大学 研究開発推進機構）

『源氏物語』は平安時代に成立した古典文学作品であり、また計量的な手法がアプローチが可能な研究課題を有していると考えられる。そこで、本研究では国立情報学研究所よりオープンデータとして公開されている『絵入源氏物語』のテキストデータを用い、統計解析を目的とした web アプリケーションを作成した。一般に、テキストデータの統計解析にはプログラミング言語である R が用いられる。しかし、そのためにはある程度 R に習熟する必要があり、テキストデータの統計解析の実行は敷居が高いと言える。それゆえ、本研究では統計学の初学者でも容易に統計解析が可能な web アプリケーションを設計し、これについて報告する。

Web Application Development for Statistical Analysis of Text Data in “The Illustrated Tale of Genji”

Gen Tsuchiyama (Organization for Research Initiative and Development, Doshisha University)

“The Tale of Genji” is considered to be a classic that was written around the peak of the Heian period. It has some related research topics for which statistical analyses are thought to be an effective approach. The National Institute of Informatics has now released a variety of related text data, including the work itself, into the public domain. Text mining commonly uses R, which is a well-known programming language, but users must be acquainted with R to use it. In this respect, it is difficult for scholars of the work to analyze its text data if they are not familiar with R. Therefore, we developed a web application to statistically analyze the text data of the tale. In this study, we report on the development of a web application for users who are beginners in statistics.

1．はじめに

近年、オープンデータの活用の重要性は様々な学術領域において認識されており、人文学領域においても多様なオープンデータが公開され始めている。そもそもオープンデータとはあらゆる制限から自由であり、誰もが再利用及び再配布が可能なデータを意味し、人文学領域のオープンデータには画像データや文学作品のテキストデータなどが含まれる。

現在は国立情報学研究所より「国文研古典籍データセット（第 0.1 版）」[1]が公開されており、この「国文研古典籍データセット（第 0.1 版）」は国文学研究資料館が所蔵する約 30 万点の古典籍を画像化したデータベースの構築を目指す「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」において作成されたデータセットである。このデータセットには膨大なテキストデータが含まれることから、文学研究に資するオープンデータであり、それ故、人文学領域のオープンデータを取り巻く環境も整ってきたと考えられる。

文学作品のテキストデータは文章の計量分析と親和性が高いことから、従来から広く学術的に利用されている。このような研究では主に著作権の消滅した、あるいは著者が許可した近現代の小説や随筆などの文学的文章のテキストデータが用いられる。しかし、その一方で、古典文のテキストデータは作成するために国語国文学の専門的知識を必要とするということ、及び校訂本文の取扱いに注意を払う必要が大きいことから、古典文学作品などの統計的な分析に適したテキストデータは容易に入手することが難しかった。このような背景から、古典文の計量的な研究は近現代の文学的な文章を対象とした研究に比べると、十分に展開しているとは言い難いと思われる。

また、文章の計量分析における統計解析では R と称されるプログラミング言語が一般的に用いられる。しかし、テキストデータを対象に、簡単な統計解析を行う場合においてもある程度 R に習熟する必要があり、統計学の初学者にとってはテキストデータの統計解析は容易ではない。

[1] http://www.nii.ac.jp/dsc/idr/nijl/nijl_list.html

そこで、本研究では「国文研古典籍データセット（第 0.1 版）」において公開されているオープンデータの 1 つである『源氏物語』のテキストデータを用いて、統計的な処理を行うことを目的とした web アプリケーションを開発した。本稿において報告する web アプリケーションは、主に『源氏物語』における単語の出現率に着目し、各巻における単語の出現傾向の可視化、またこの出現率を特徴量とした多変量解析を行えることを目的とした。

2．『源氏物語』のデータセット

上掲の「国文研古典籍データセット（第 0.1 版）」に含まれる『源氏物語』のテキストデータは江戸時代に出版された刊本である『絵入源氏物語』を底本としたデータセットである。『絵入源氏物語』とは江戸時代に広く流布した『源氏物語』の版本の 1 つである。『源氏物語』は、平安時代に著され、各時代を通じて読み継がれてきた古典作品であり、平安時代の著名な女流作家である紫式部が表した『紫式部日記』の記述から、紫式部の手により執筆されたと考えられる現存最古の 54 巻にわたる長編物語の 1 つである。また、『源氏物語』の自筆原稿は散逸しており、主に書写による写本によって受け継がれてきた。従って、『源氏物語』に触れられるのは写本を入手できる一部の人に限られていた。そのような中、『絵入源氏物語』は「絵入」とあるように本文だけではなく絵が挿入され、多くの人の目に触れるために作成された版本である。なお、『絵入源氏物語』は何度か刊行されているが、本研究で用いるテキストデータは承応三年に刊行された版本を底本とするテキストである。

また、この『絵入源氏物語』のテキストデータには活用事例があり、国文学研究資料館より古典選集本文データベースの 1 つとして『絵入源氏物語データベース』[2]が公開されている。この『絵入源氏物語データベース』は所謂 KWIC 索引であり、ユーザが文字列を入力することで、該当文字列が含まれる前後の文章を表示するものである。

3．web アプリケーションについて

3.1　web アプリケーションの概要

本研究において報告する web アプリケーションについて以下に概観する。すでに述べたように、この web アプリケーションは文字ではなく単語を単位とした計量的な分析を行うことを想定しており、品詞別に集計された各巻の単語の出現率を可視化することを目的の 1 つとしている。従って、本研究ではオープンデータとして公開されているテキストデータについてまず形態素解析を行った。用いた形態素解析ツールは国立国語研究所から公開されている「Web 茶まめ」[3]である。

しかし、『絵入り源氏物語』のテキストデータには「Web 茶まめ」によって形態素解析することができない文字列が混入していることから、テキストデータのクリーニングの必要がある。これについては後述する。

また、先にふれたように、本研究において報告する web アプリケーションの統計的な処理については R を用いた。R は統計解析向けのプログラミング言語であり、多様な統計処理を容易に行うことが可能である。また、R には多くのパッケージが用意されており、これを R にインストールすることで機能を拡張することができる。

そのような R のパッケージの 1 つに Shiny がある。これは解析結果を埋め込んだインタラクティブな web アプリケーションの作成を支援するパッケージであり、Shiny を用いてアプリケーションを公開することも可能である。本研究においてもこの Shiny を用い、これによって『絵入源氏物語』における語の出現傾向を可視化する web アプリケーションを開発した。

先にふれたように、本研究において開発した web アプリケーションの目的とするところは単語の出現傾向の可視化である。そこで、この web アプリケーションにおいては現状においていくつかの可視化する機能を実装した。1 つは任意の単語の各巻における出現率の可視化である。もう 1 つの機能は品詞別に集計された単語の出現率を特徴量とした主成分分析の分析結果の可視化である。これらに加えて、形態素解析を行う際に、各単語には品詞に関する情報が付与されることから、本研究ではこれを用い、各巻の品詞の比率を可視化する機能も追加している。

このように、本研究による web アプリケーションは基礎的な統計処理によって得られる対語の出現率と、応用的な統計解析によって得られる単語の出現傾向を可視化することが可能である。

3.2　データの加工

先にふれたように、本研究において用いたテキストデータには形態素解析を行う上で不適な文字列がある。これは踊り字に係わる記号であり、古典文における踊り字には「一の字点」、「同の字点」、「くの字点」の 3 種類に大別される。「一の字点」及び「同の字点」はそれぞれ「ゝ」と「々」という記号で表記され、これらは踊り字の直前の 1 文字を繰り返すことを意味する。『絵入源氏物語』のテキストデータにおいても、これらの記号が用いられ、また「Web 茶まめ」においても「こゝろ」や「日々」といった単語の形態素解析に大きな問題は生じない。しかし、「くの字点」は図 1 に示されるような踊り字であり、「くの字点」の

[2] http://base1.nijl.ac.jp/infolib/meta_pub/g0001501genji

[3] http://chamame.ninjal.ac.jp/

直前の 2 文字以上の文字列を繰り返すことを意味する。

さう〳〵しく
御かた〴〵
こゝろ〴〵

図 1　くの字点の一例

図 1 において示したこれらの文字列は「国文研古典籍データセット（第 0.1 版）」における『絵入源氏物語』のテキストデータにおいては「さうさう＋＋しく」、「御かた＋＋$」、「こゝろ＋＋$」というように「＋＋」という記号をあてて表記されている。なお、ここで「$」は繰り返される文字列に濁音になることを含意している。これらの記号は「Web 茶まめ」において形態素解析できないことから、「＋＋」の部分を適宜修正する必要がある。

また、「御かた＋＋$」は「御かたがた」、「こゝろ＋＋$」は「こころごころ」となり、これはすなわち、本研究において用いたテキストデータにおいて、くの字点を意味する「＋＋」は、「こゝろ＋＋$」が「こころごころ」となるように「＋＋」の指示する直前の文字列が必ずしも 2 文字であることを意味する訳ではない。3 文字以上の文字列を指示する場合も多く認められる。この「＋＋」の加工については自動で修正するための参照資料がないため、「＋＋」が指示する文字列については文脈から判断するしかない。そこで本研究では原文にあたり、手作業でくの字点についての加工を行った。このようなくの字点は『絵入源氏物語』に 3241 箇所あり、これをすべて手作業によって形態素解析が可能な形式にテキストデータの加工を行った。

このような処理を通じて加工されたテキストデータに対し、「Web 茶まめ」を用い形態素解析を行った。形態素解析を行う際には中古和文の辞書を用いた。

3.3　web アプリケーションの機能

本稿において述べる web アプリケーションには以下の 3 つの機能が含まれる。

(1) 各巻における品詞の比率の可視化
(2) 任意の単語の各巻における出現率の可視化
(3) 単語の出現率について主成分分析の実行

以下において、これらの機能について順に概観する。

まず、各巻における品詞の比率の可視化についてである。本研究の目的とすることは単語の出現率及び出現傾向の可視化であるが、形態素解析の際に各単語の品詞情報が付与されることから、この情報を用い、品詞の比率の可視化を試みた。これは棒グラフによって示される。出力結果は図 2 の通りである。

本研究による web アプリケーションはデフォルトで、品詞の比率を計算されるタブが開かれており、図 2 における「巻の選択」をクリックすることで、目的となる巻の比率を可視化することができる。また、「比較する巻を追加」にチェックを入れることで先に選択した巻における品詞の比率と比較対象となる他の巻を選択でき、グラフに比較対象となる巻の棒が追加される。図 2 においては、第 1 巻「桐壺」と第 2 巻「帚木」の比較結果が示されている。この棒グラフから、「桐壺」は名詞や動詞などの比率が「帚木」より高く、助詞や助動詞の比率が「帚木」より低いことが分かる。

次に、任意の単語の各巻における出現率の可視化についてである。web アプリケーション上部にある「単語の出現率」というタブをクリックすることで、単語の出現率を計算することが可能となる。単語の出現率は入力された単語の品詞の延べ

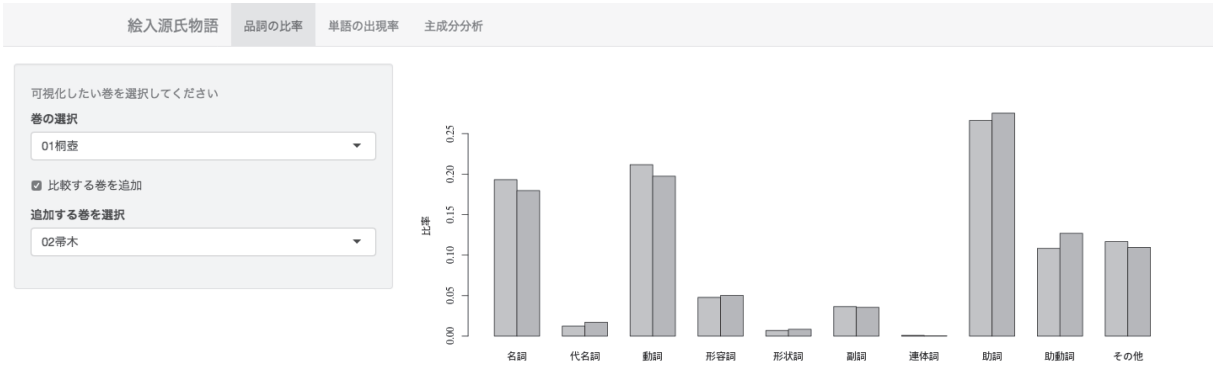


図 2　品詞の比率の出力結果

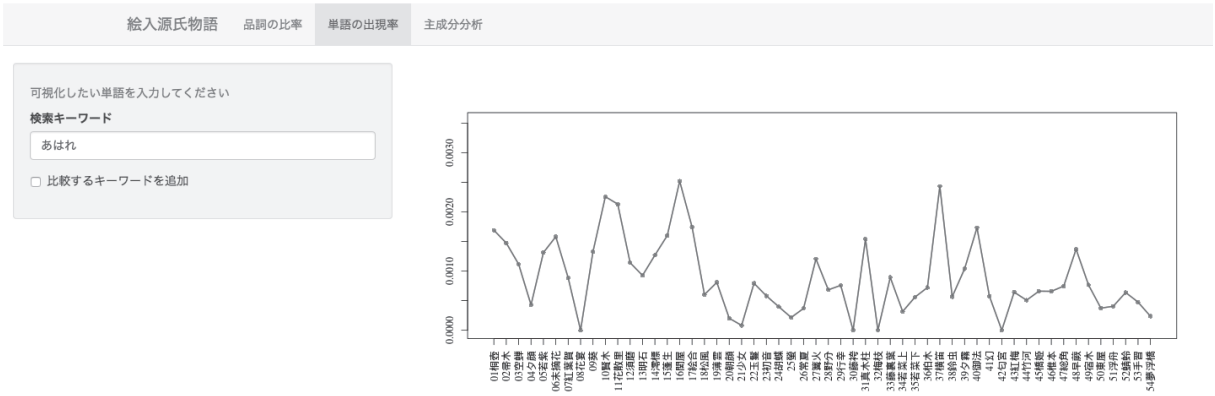


図 3　1 単語の出現率の出力結果

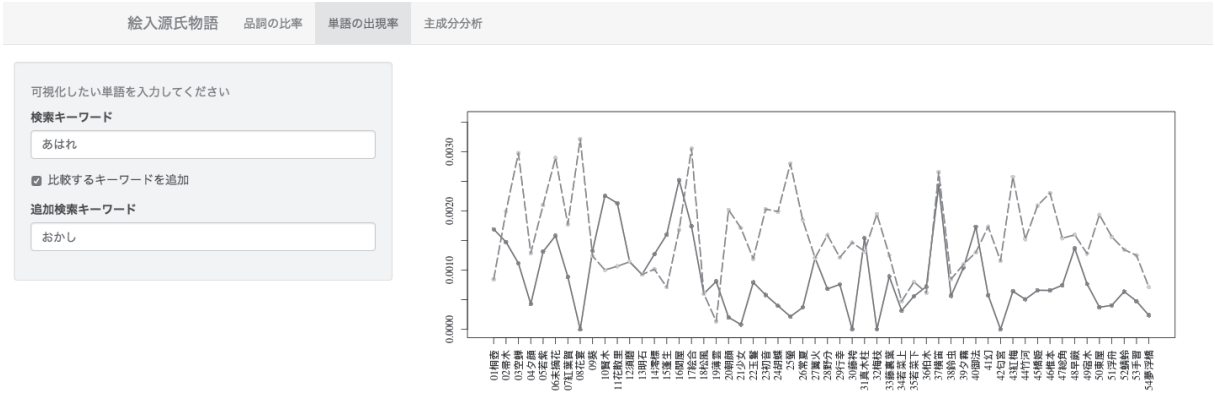


図 4　2 単語の出現率の出力結果

語数に対する割合である。計算された結果は図 3 において示すように折れ線グラフとして図示される。

図 3 における「検索キーワード」に任意の単語の基本形を入力することで、その単語の出現率が計算される。なお、図 3 では「あはれ」を検索し、その出現率が図示されている。品詞の比率の可視化と同様に、単語の出現率の可視化においても、他の単語の出現率との比較が可能である。図 3 における「比較するキーワードを追加」にチェックを入れることで、図 4 に示すように「追加検索キーワード」という入力フィールドが現れ、任意の単語を追加検索することができ、追加された単語は破線の折れ線でその出現率が図示される。図 4 では「あはれ」と「おかし」という 2 単語の各巻の出現率が可視化されている。なお、『絵入源氏物語』のテキストデータでは「をかし」ではなく「おかし」が用いられていることから、図 4 においても「おかし」が検索されている。

このように、単語の出現率の可視化については、分析者が自由に任意の単語を入力することが可能であり、本研究における web アプリケーションは分析者の興味のある単語の量的な特徴についての直感的な理解を支援できると考えられる。

最後に、単語の出現率に対する主成分分析についてである。主成分分析とは多次元データに対する次元縮約の手法であり、元のデータの変数より新たに合成変数を求めることで情報の縮約を行う統計手法である。本研究における変数とは単語の出現率のことである。分析に用いる単語数が増えるとデータは高次元になり、各巻の量的な関係性の理解は困難になる。主成分分析はこのような問題の解決に適した手法である。なお、本研究では相関行列を用いた主成分分析を行っている。

2 つの単語の出現傾向を考察するのであれば、先に述べた単語の出現率の可視化においても可能であり、例えば図 4 において示した出力結果から、「あはれ」と「おかし」という 2 単語の出現傾向が類似している巻として第 37 巻「横笛」が指摘でき、その反対に第 8 巻「花宴」では上掲の 2 単語の出現傾向は類似しているとは言えない。しかし、3 単語以上については図 4 のような折れ線グラフでは解釈が困難になることから、文章の

計量分析では主成分分析などの手法が用いられる。

本研究の web アプリケーションによって出力される分析結果は、図 5 に示すように主成分分析によって求められた第 1 主成分と第 2 主成分の主成分得点の散布図として出力される。散布図の横軸は第 1 主成分、縦軸は第 2 主成分である。また、各主成分に含まれている情報量は寄与率によって評価され、第 1 主成分及び第 2 主成分の寄与率は散布図に併記した。

図 5 においては助動詞の出現頻度上位 20 語を用いた主成分分析の分析結果である。図 5 にあるように、「品詞の選択」において名詞や動詞などの品詞から分析を行う品詞を選択し、「分析に用いる単語数 (出現頻度順)」の入力フィールドに単語数を入力することで、主成分分析が実行される。なお、分析に用いる単語数は 1 単語刻みで変更することが可能である。

次いで、主成分分析によって得られた散布図の軸の解釈には因子負荷量が用いられることが一般的であるが、これは散布図の下に出力され、本研究の web アプリケーションでは第 1 主成分及び第 2 主成分の因子負荷量はそれぞれ上位 5 語と下位 5 語が示される。これに基づいて、図 5 に示されている分析結果を解釈すると、図 5 の散布図において横軸の正の領域にある巻は「なり」、「ず」、「む」、「ぬ」、「べし」の出現率が高く、負の領域にある諸巻は「まし」、「けむ」、「まじ」、「さす」、「めり」の出現率が高いと言える。他方、縦軸においては正の領域にある諸巻は「けむ」、「き」、「す」、「さす」、「けり」の出現率が高く、負の領域にある諸巻は「まじ」、「り」、「めり」、「る」、「まし」の出現率が高いと言える。

また、この分析結果において、第 34 巻「若菜上」、第 35 巻「若菜下」、第 39 巻「夕霧」、第 47 巻「総角」、第 49 巻「宿木」、第 50 巻「東屋」、第 51 巻「浮舟」、第 52 巻「蜻蛉」、第 53 巻「手習」の 9 巻と他の諸巻との間に、助動詞の出現傾向の相違が認められると考えられる。特に「蜻蛉」と「手習」の 2 巻は孤立して位置していることから、助動詞の出現頻度上位 20 語についての分析においては類似した傾向を有していると考えられる。

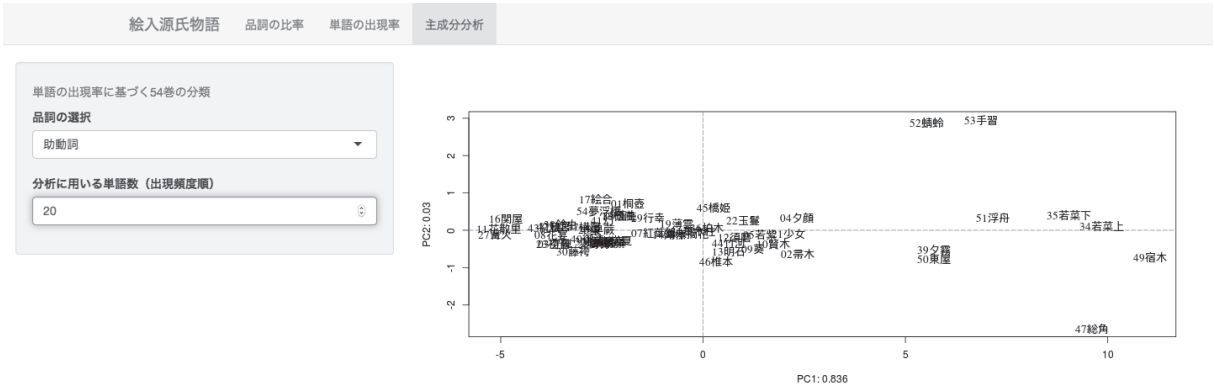
このように、主成分分析を行うことで、単語の出現率を可視化した折れ線グラフからは解釈が困難である『絵入源氏物語』の各巻の量的な関係性を概観することが可能となる。

3.4 web アプリケーションの公開

この web アプリケーションは現在 web 上で公開されており、その URL は https://gtcym.shinyapps.io/genjiapp/である。本稿において概観した機能はすべて実行可能である。

4. まとめ

本研究において開発した web アプリケーションの機能については以上に概観した通りである。本研究において目的としたところは、統計学の初学者にとっても容易に統計処理が行えることである。単語の頻度や出現率は、従来から様々な研究に用いられており、これを統計的に可視化することは単語の出現傾向の直感的な理解を促進できると考えられる。また、これに加えて、本研究における web アプリケーションでは主成分分析を行うことで、単語の出現傾向から傾向の類似し



	PC1因子負荷量上位5語	PC1因子負荷量下位5語	PC2因子負荷量上位5語	PC2因子負荷量下位5語
1	なり 0.242	まし 0.185	けむ 0.531	まじ -0.321
2	ず 0.241	けむ 0.199	き 0.295	り -0.303
3	む 0.238	まじ 0.212	す 0.266	めり -0.297
4	ぬ 0.237	さす 0.215	さす 0.194	る -0.268
5	べし 0.236	めり 0.217	けり 0.176	まし -0.209

図 5　主成分分析の出力結果

た巻のグルーピングも可能である。

このように、この web アプリケーションは『絵入源氏物語』を対象とした研究を行うユーザに、統計手法の知識がなくとも『絵入源氏物語』の量的特徴あるいは量的な傾向を示すことで、新たな研究視点や資料を提供できると考えられる。

今後の展望として、主成分分析に加えて他の多変量解析の実装を考えている。想定している分析手法としては階層的クラスター分析や多重対応分析である。階層的クラスター分析は主成分分析と異なり、どのような要因によって『絵入源氏物語』の諸巻がグルーピングされるのか明示的ではないという問題点が予想されるが、クラスター分析を用いることで、統計学の初学者にとって主成分分析より直感的に単語の出現傾向に基づき『絵入源氏物語』の諸巻のグルーピングすることが可能になると考えられる。多重対応分析についても同様の意図から実装を考えている手法である。主成分分析は因子負荷量の解釈が難しいと思われるが、多重対応分析の場合、分析によって得られる出力結果についてはより容易に解釈できると考えられる。

また、より直感的に操作できる GUI を改良する予定であり、これに加え、各分析を行う際により細かいオプションも追加することで、統計手法についてある程度の知識を有しているユーザまで利用対象を拡大することを考えている。

参考文献

1) 橋本雄太. (2015). 人文学資料オープンデータの可能性と現状. 情報の科学と技術, 65(12), 525-530.

2) Open Knowledge Foundation. (2012). オープンデータとは何か? – Open data handbook. <http://opendatahandbook.org/guide/ja/what-is-open-data/> (参照 2016-10-04)

3) 村上征勝. (2002). 文化を計る−文化計量学序説. 朝倉書店.

4) 金明哲. (2009). 文章の執筆時期の推定―芥川龍之介の作品を例として―. 行動計量学, 36(2), 89-103.

5) 計量国語学会. (2009). 計量国語学事典. 朝倉書店.

6) 金明哲,張信鵬. (2013).テキスマイニングツール MTMineR のコンセプトと機能. 日本行動計量学会大会発表論文抄録集, 41, 360-363.

7) 樋口耕一. (2015). フリーソフトウェア「KH Coder」による計量テキスト分析: 手軽なマウス操作による分析からプラグイン作成まで. 研究報告人文科学とコンピュータ (CH), 2015(9), 1-2.

8) 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.

9) 土山玄. (2016). 絵入源氏物語のテキストデータに対する統計解析 web アプリケーションの設計. 研究報告人文科学とコンピュータ (CH), 2016(1), 1-4.