【第59回全国大会公開講演】

スタイロメトリーから連想する計量的表現研究

金　明哲

1. まえがき

本稿では、大会の発表内容に沿って、理解しやすいように整理し、「スタイロメトリーとは」「テキストアナリティクスの流れ」「データの形式と分析の例」「データ集計および分析のツール」の順で要点をまとめる。

2. スタイロメトリーとは

ポーランドの哲学者のヴィンツェンティ・ルトスワフスキ（Wincenty Lutosławski）は、プラトン（Platon）の作品の年代について数量的に分析した論文“Principes de stylométrie appliqués à la chronologie des œuvres de Platon”を発表した（Lutosławski, 1898）。ルトスワフスキがこの論で用いたのがstylométrieという手法である。Stylométrieに対応する英語表記はstylometry（スタイロメトリー）であり、近年ではstylometricまたはstylometricsの語も用いられている。日本語に訳すと、計量文体、または文体計量となるが、「文体」という言葉から書き言葉のみを対象とする分野であるかのようなイメージが先行するのは好ましくない。スタイロメトリーは、書き言葉のスタイルの分析から始まったが、英語のstyleは文体だけではなく、周知のとおりもっと広い意味を持っている。

スタイロメトリーの発端は、著名の作品の作者が問題となる著作物を計量的に分析することであった（Lutosławski, 1898; Mendenhall, 1887）。Mendenhall（1887）は、ディケンズ（Dickens C.　1812～1870）、サッカレー（Thackeray W.M.　1811～1863）、ミル（Mill J.S.　1806～1873）の文章に使われた単語の長さを調べ、それが作家によって異なり、作家の特徴になることを示した。

単語の長さは、単語を構成する要素（アルファベット、音素など）の数によって表現される。メンデンホールは、単語の長さに関するデータに基づいて描いた図1のような折れ線グラフを用いて分析を行った（Mendenhall,1901）。図1の横軸はアルファベットの数であり、縦軸は単語の使用率である。図1から、シェイクスピアは4文字の単語を最も多く使用し、ベーコンは3文字の単語を最も多く使用していることが分かる。この結果を根拠としてメンデンホールは、「シェイクスピアという人物は実在せず、単にベーコンが圧政に抗議するために、シェイクスピアという偽名で一連の風刺劇を書いた」という説を否定した。

1975年になってこの問題の再考を促す研究が発表された（Williams, 1975）。ウィリアムズは、16世紀のイングランドの詩人シドニー（Philip Sidney, 1554-1586）の著作を調べ、同一人物の著作であっても散文（prose）と韻文（verse）では、最も多く使われている単語の長さの値が異なる場合があることを示した。そして、シェイクスピアの文章とベーコンの文章とでは最も多く使われている単語の長さが異なることは、著者が別人である可能性を否定するものではないが、メンデンホールが分析に用いたのは、シェイクスピアの散文とベーコンの韻文であったため、文章の形式の差による違いである可能性もありうると指摘している。

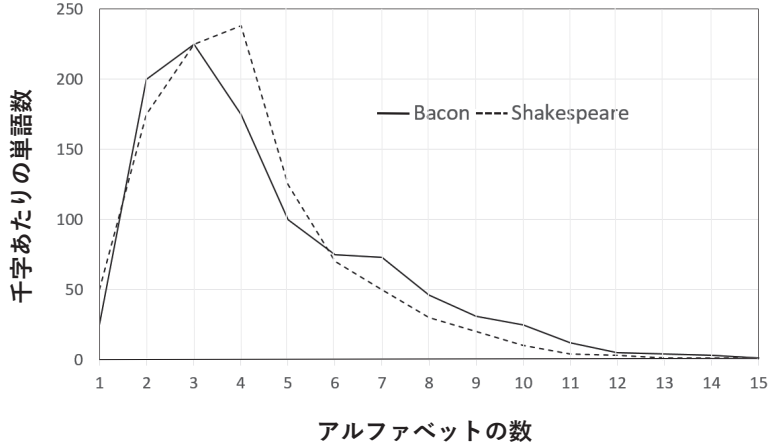


図1　単語の長さの分布（Mendenhall（1901）のデータを用いて作成し直した）

日本語においても単語の長さの分布に書き手の特徴が現れることが報告された（金 1995, 1996）。日本語の場合は、何を単位として長さを計算するかをまず決める必要がある。例えば「日本語」という単語について、表記の文字で数えれば「日本語」という3文字なので長さは3、仮名であれば「にほんご」となり長さが4、ローマ字表記「nihongo」に置換えた場合の長さは7となる。

メンデンホールの研究に啓発され、文の長さ、段落の長さ、単語の使用率、品詞の使用率、文節のパターン、構成要素のn-gramなどが文体分析に用いるようになった。1960年代からはコンピュータを用いた多変量データ解析の手法を用いるようになり、1990年代では、テキスト型データ（文章）を研究対象としてデータマイニングを行うことをテキストマイニング（text mining）と呼ぶことになった。テキストマイニングの用語を用いる前からテキストの計量分析を行った方から見ればテキストマイニングはスタイロメトリーを一般化したことであり、研究手法には何ら斬新さも感じられない。

スタイロメトリーは、音楽や美術・絵画や舞踊などを対象とした分野でも用いられている。近年、プログラムのソースコードに関するCode stylometry研究もある。

Code stylometryは、プログラムのソースコードからプログラミングのスタイルの特徴分析をしたり、ソフト作者を特定したりすることを目的としている。

3. テキストアナリティクスの流れ

昨今、コンピュータを用いてテキスト分析を行う分野をテキストマイニング、またはテキストアナリティクス（text analytics）と呼ぶことになっている。テキストについて、どのような目的で、いかなるアプローチで分析するかは問題によって異なるが、電子化されたテキストアナリティクスの流れを図2に示す。

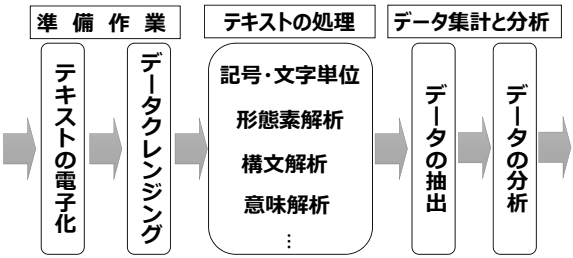


図2　テキストアナリティクスの流れ

3.1　テキストの電子化

電子メールなどを含むインターネット上のテキスト、CDRやDVDなどに記録されたテキストはすでに電子化されている。紙媒体の書籍などの内容について分析を行うためには、電子化する必要がある。書籍物になっているテキストを電子化する方法としては、光学読み取り機器（OCR）を用いてコンピュータに画像として読み込み、さらに文字認識ソフトを用いて画像データを文字データに変換する方法がある。近年スマートフォンの普及と機能の向上により、スマートフォンのOCRアプリを用いて、簡単に紙媒体の文章をテキストデータに変換できるようになった。また、音声言語をテキストに置き換えることも可能である。このような技術は日進月歩であり、ますます便利になっている。

3.2　データクレンジング

インターネットでダウンロードしたテキスト、音声言語をテキストに変換したもの、OCRで読み取ったテキストの中には必要がない記号や誤った文字列などが含まれている場合がある。テキストの中の必要ではない記号・文字列（ゴミ）を取り除いたり、間違った文字列を訂正したりすることは不可欠である。このような作業をデータクレンジングと呼ぶ。データクレンジングの作業は、単純ではあるが非常に重要である。

3.3 テキストの加工

データクレンジングの作業が終わったら、記号・文字単位でデータを集計することが可能である。例えば、文が何文字で構成されたか、文章のサイズはどれぐらいであるか、読点などの記号がどの文字の前後に打たれているかなどに関する集計は十分可能である。しかし、単語を単位として分析を行うためには、文を単語単位に切り分けないとデータの集計が不可能である。文を単語単位に切り分け、品詞の情報などを加える作業を形態素解析と呼ぶ。形態素解析を行うソフト（形態素解析器）としてはMeCabやJUMANなどがある。また、文節を単位として分析する場合は、文を文節単位に切り分けたり、文節の係り受け関係に関する情報を付け加えたりすることが必要である。文節の係り受けを解析するソフトとしてCabochaやKNPなどがある。

4. データの形式と分析の例

具体的な例を用いて説明することにする。川端康成の『雨の日』の冒頭の1文「今年は冬が暖かだったが、春になってから寒い。」をMeCabで形態素解析した結果を整形したものを次に示す。

今年＜名詞＞は＜助詞＞冬＜名詞＞が＜助詞＞暖か＜名詞＞だ＜助動詞＞た＜助動詞＞が＜助詞＞、＜記号＞春＜名詞＞に＜助詞＞なる＜動詞＞て＜助詞＞から＜助詞＞寒い＜形容詞＞。＜記号＞

4.1 データの形式

川端康成と三島由紀夫のそれぞれ10作品を上記のように処理し、各作品に使用されている助詞「の」「は」「に」「て」「を」「が」「と」「も」「で」「から」「か」「へ」「という」「ながら」「や」「ば」「ので」「まで」「ほど」「だけ」「かも」「など」「より」とその他の助詞をまとめた「OTHERS」を集計したデータの画面コピーを図3に示す。このようなデータの行を個体、列の項目を変数と呼ぶ。

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
	の	は	に	て	を	が	と	も	で	から	か	へ	という	ながら	や	ば	ので	まで	ほど	だけ	かも	など	より	OTHERS
K_たまゆ	20.46	13.6	12.82	9.69	9.45	8.3	5.9	3.97	2.71	2.53	3.01	0.54	0.66	0.54	0.3	0.3	0.48	0.3	0.42	0.24	0.6	0.24	0.3	2.65
K_みづうみ	17.72	13.79	12.02	12.32	11.14	8.23	5.78	3.71	3.58	2.09	2	0.81	0.51	0.7	0.37	0.46	0.29	0.44	0.38	0.28	0.42	0.28	0.33	2.32
K_再婚者	16.86	12.84	12.93	10.91	8.99	8.73	6.93	5.01	3.23	1.83	2.18	0.91	0.91	0.67	0.56	0.47	0.43	0.58	0.42	0.39	0.69	0.35	0.37	2.81
K_小春日	19.53	15.17	10.7	12.69	9.08	8.21	6.72	3.48	3.98	1.74	0.87	0.62	0.62	0.87	0.12	0.87	0.5	0.12	0.37	0.5	0.25	0.75	0.37	1.87
K_少年	21.52	12.1	12.8	9.54	8.06	7.2	6.76	5.09	4.3	2.43	1.3	0.57	1.58	0.51	0.6	0.47	0.35	0.63	0.19	0.25	0.19	0.44	0.41	2.72
K_岩に菊	24.7	10.16	11.99	9.24	6.68	8.23	6.77	7.5	3.02	1.92	1.1	0.27	0.64	0.55	1.46	0.27	0.18	0.64	0.37	0.18	0.27	1.19	0.18	2.47
K_故郷	20.76	12.16	11.35	12.5	9.06	9.63	5.05	5.39	3.1	2.87	1.83	1.15	0.69	0.8	0.11	0.34	0.46	0	0.23	0.34	0.23	0	0.11	1.83
K_横町	14.21	13.67	12.06	14.75	10.01	8.49	9.2	3.75	3.66	2.14	1.34	0.27	0.36	1.16	0.18	0.36	0.27	0.27	0.27	0.27	0.18	0.27	0.36	2.5
K_自然	16.89	15.84	11.92	9.95	7.69	9.5	9.65	4.68	3.92	2.56	1.36	0.15	0.3	1.06	0.9	0	0.3	0.15	0.15	0	0.45	0.15	0.3	2.11
K_雨の日	17.21	14.16	11.22	13.29	8.39	9.91	7.3	4.47	3.38	2.51	1.85	0.76	0.65	0.54	0.22	0.22	0.54	0	0.33	0.44	0.54	0	0.33	1.74
M_風関間	16.19	11.6	11.78	13.43	11.82	10.17	6.2	2.99	3.84	1.83	1.12	1.34	1.16	0.36	0.54	0.49	1.25	0.54	0.09	0.4	0.18	0.13	0.13	2.45
M_家庭最	17.27	12.52	12.17	11.56	13.56	9.42	5.76	1.96	4.84	1.57	1.09	0.83	1.35	0.48	0.26	0.44	0.78	0.52	0.57	0.31	0.09	0.04	0.04	2.57
M_携帯用	15.98	14.11	13.01	11.53	13.64	8.95	6.51	2.92	3.88	1.39	0.96	1.58	0.77	0.38	0.29	0.57	0.33	0.43	0.29	0.48	0.05	0.05	0.05	1.87
M_月	20.42	13.12	12.51	12.13	12.57	9.28	4.72	2.69	3.62	1.81	0.93	1.37	0.11	0.22	0.55	0.44	0.27	0.11	0.05	0.49	0	0	0.05	2.52
M_果実	19.43	11.71	14.19	10.48	14.38	9.9	4.19	3.71	3.33	2	0.86	1.33	0.86	0.29	0.48	0.48	0.57	0.19	0.38	0.19	0	0	0.19	0.86
M_海と夕	21.9	13.47	16.19	11.97	10.75	7.89	3.81	2.72	2.45	2.04	0.95	2.31	0.14	0	0.82	0.14	0.14	0.27	0.14	0	0.14	0.14	0.27	1.36
M_詩を書	19.44	13.99	12.07	9.78	11.21	9.1	5.94	4.09	1.98	1.73	0.93	0.93	1.55	0.37	0.93	0.74	0.12	0.12	0.19	0.19	0.12	0.31	0.37	3.78
M_近世姑	19.83	12.28	12.52	12.4	10.56	9.52	4.73	2.76	3.93	1.53	1.1	1.17	0.55	0.61	0.86	0.68	0.8	0.55	0.31	0.25	0	0.25	0.06	2.76
M_遠慮会	21.98	14.25	13.05	10.36	12.54	8.64	4.58	2.18	3.15	1.72	0.92	1.43	0.57	0.52	0.63	0.52	0.29	0.23	0.17	0.17	0	0.11	0	2
M_鳥獣	17.9	12.29	13.05	11.14	10.95	9.71	5.33	3.9	3.14	1.62	0.29	2.48	0.48	0.29	0.67	0.76	0.86	0.29	0.38	0.48	0.1	0.38	0.38	3.14

図3　川端康成と三島由紀夫の20作品における主な助詞の集計表の画面コピー

テキストアナリティクスは、図3 のようなデータについて統計分析、または機械学習を行い、その結果に基づいて考察・分析を行う。目的によって分析の手法が異なる。説明の便宜上、ここでは両者の助詞の使用特徴を分析する例を次に示す（金: 1997, 2002, 2003）。

4.2 統計分析の例

4.2.1 助詞ごとの平均プロット

図3のデータを用いて作成した助詞ごとの平均プロットを図4に示す。平均プロットは、変数ごとの平均の差異を分析する統計グラフである。図4から助詞「の」は、両者には平均の差が見られないが、助詞「を」「に」「も」「へ」には大きな差があることが分かる。

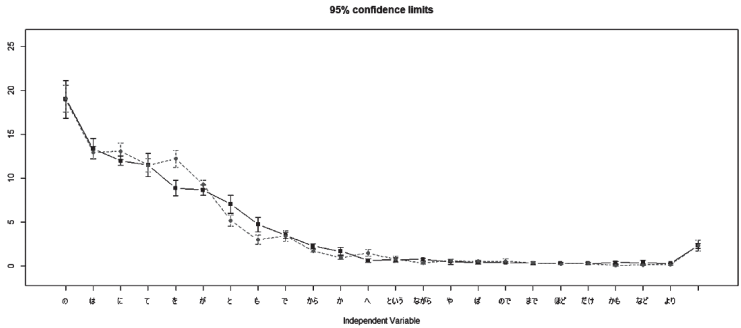


図4　川端康成と三島由紀夫の作品における助詞データの平均プロット

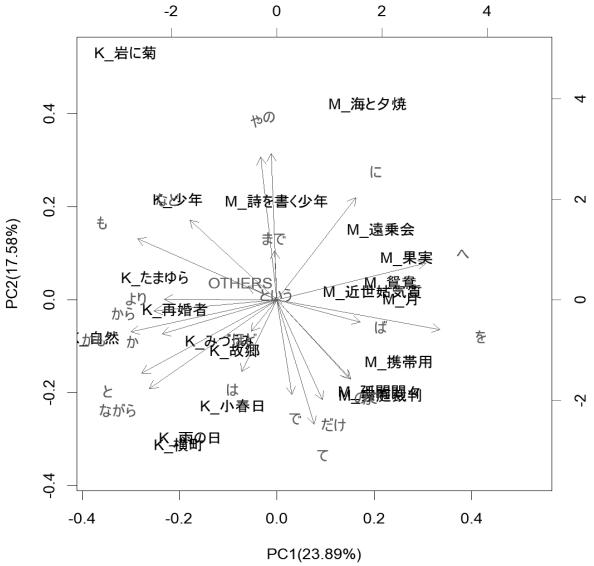
4.3 特徴分析

図3のデータを用いた主成分分析の結果を図5(a)、対応分析の結果を図5(b)に示す。

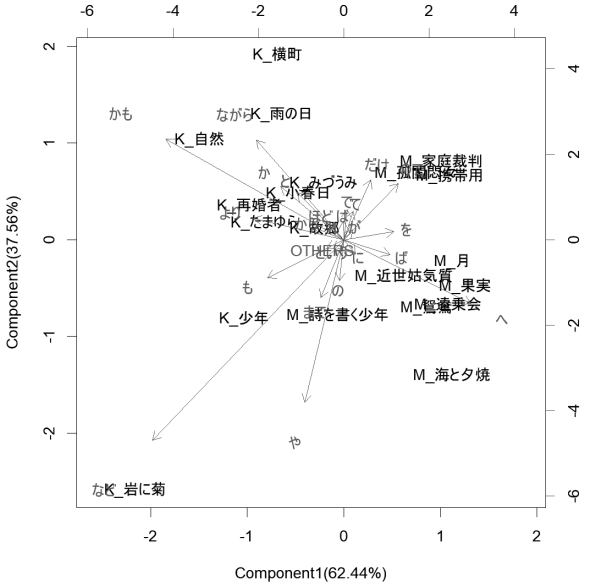
図5は24次元（変数）のデータを2次元に縮約したものである。図5（a）は主成分分析、図5（b）は対応分析の第1、2成分の散布図である。情報の損失はあるが、データの全貌を視覚的に概観できる。両図の右側に三島由紀の作品、左側に川端康成の作品が分かれて配置されていることが分かる。また右方向には「を」「へ」の矢印が長く伸びている。これは、三島由紀夫はこれらの助詞を好んで用いていることを意味する。このような方法で作品や作者の特徴を分析することができる。

4.4 分類分析

研究の目的によっては、テキストを分類したいときがある。図3に示したデータを用いて階層的クラスター分析を行い、作成した樹形図を図6に示す。樹形図は大きく二つのブランチに分かれていることが確認できる。左は三島由紀夫の作品のブランチ、右は川端康成の作品のブランチになっている。



(a)



(b)

図5　川端康成と三島由紀夫の作品における助詞データの次元縮約後のバイプロット

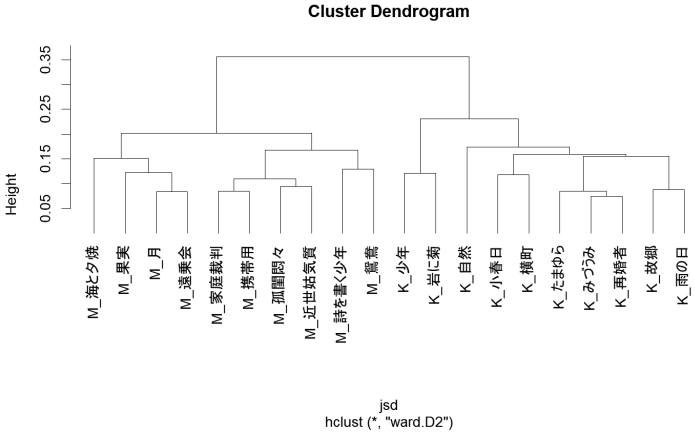


図6　川端康成と三島由紀夫の作品における助詞データのクラスター分析の樹形図

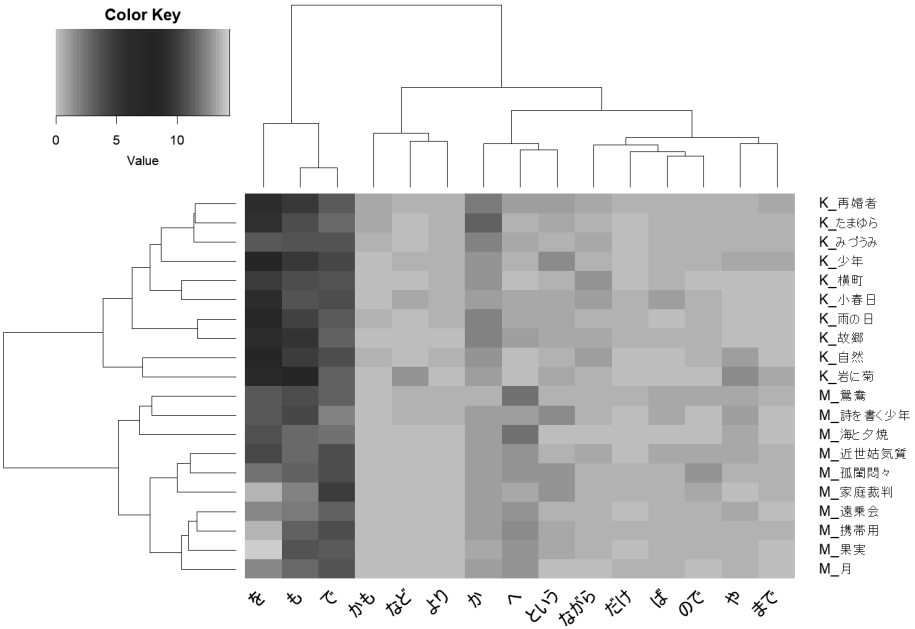


図7　川端康成と三島由紀夫の作品における助詞データのバイクラスタ・ヒートマップ

クラスタリングは、図3のデータ形式の行（個体）単位、または列（変数）単位で行うことが可能である。図6は行単位で行ったものである。行単位と列単位のクラスタリングを同時に行い、値の大小を色で示すバイクラスタ・ヒートマップを作成する方法がある。図6で用いたデータから情報が少ないいくつかの助詞を取り除いて作成したバイクラスタ・ヒートマップを図7に示す。図7の個体のクラスター樹形図は、図6と同様に20作品が著者ごとに2つのブランチに分かれている。これは、用いた助詞の使用率は書き手によって異なることを意味する。ヒートマップの色からどの変数が個体のクラスターに与えている影響の度合いを考察することができる。つまり、書き手の特徴を分析することができる。例えば「を」の色から三島由紀夫は「を」を比較的に多く使用し、「かも」「など」「より」の使用率は川端康成より相対的に低いことが読み取れる。これは平均プロットや主成分分析の結果と一致する。

このようなデータ分析方法は、数多く提案されており、研究の目的に合わせて適切な方法を選択して用いることが必要である。

5. データ集計および分析のツール

作成したコーパスから図3のようなデータ表を作成するためには、何らかのツールがあると便利である。プログラミング言語を用いることが最も望ましい。近年、広く用いられているのはPythonとR言語である。本稿の読者の殆どは、これらの言語を自由自在に使える状況ではないと推察している。そこで、これらのプログラミング言語を用いて詳細のコードを書くことなく、メニュー操作でデータの集計とデータ解析が可能なツールMTMineR（エム・ティ・マイナー）を紹介する。

MTMineR（Multi Lingual Text Miner with R）は20数年前から、私個人用のプログラムをベースにし、研究と教育用として開発・拡張し続けているソフトである。MTMineRは、複数のテキストを同時に処理することを前提としており、日本語、中国語、韓国語、英語、ドイツ語、フランス語、イタリア語などを扱うことができる。

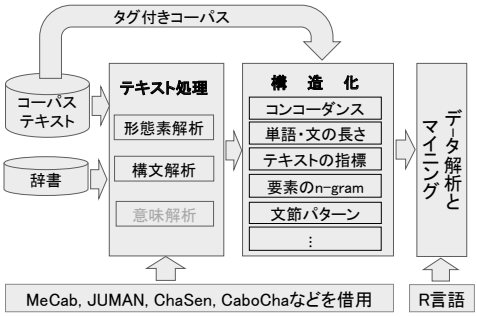


図8　MTMineRの構成イメージ

MTMineRは、Java言語で書かれ、Windowsで検証が行われている。Java言語がインストールされているマシンであれば、ダウンロードして直接使用できる。ツールは大きく三つの部分に分かれる。(1) 既存の形態素解析器などを借用してテキストを一括処理する。(2) ユーザーの指定に基づいてデータを集計する。(3) 集計したデータを統計的解析・機械学習する。その構成イメージを図8に示す。

用いるテキストコーパスは、テキスト型 (*.txt) である。構成要素や出現頻度を集計する際、テキストはタグを付けていない平テキスト (Plain Text) とタグ付きのテキスト (Tagged Text) に分けられる。タグは、決まったルールに従えば、自由に付けることができる。

ツールの操作は、GUI画面上でメニュー操作やオプションの指定などで簡単に分析を進めることができる。MTMineRは、現時点では次のサイトからダウンロードできる。

https://mjin.doshisha.ac.jp/MTMineR/mt.html

上記のサイトには、MTMineRのマニュアルがある。MTMineRはzip形式で圧縮されている。これを解凍し、ファイルMTMineR_WithToolsをダブルクリックすると図9(a)の画面が開かれる。図9(b)は分析したいテキストを取り組んだ画面である。

MTMineRの終了は、メニューの「File」⇒「Exit」をクリックするか、画面の右上の×印をクリックする。

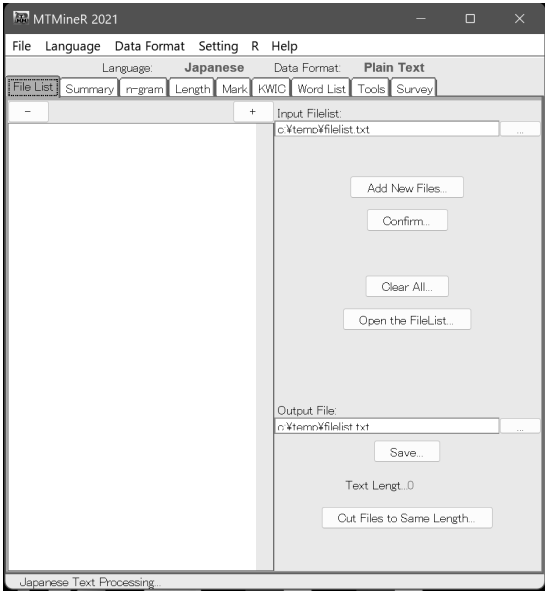
5.1 言語種類と形態素解析

形態素解析は、日本語はMeCab、JUMAN、ChaSen、中国語はJieba、韓国語はHanDic、英語、ドイツ語、フランス語、ドイツ語、イタリア語などはTreeTaggerを借用する。日本語の構文(係り受け)解析は、CaboChaとJuman|KNPを用いる。日本語以外の言語を形態素解析に適する文字コードはUTF-8である。文字コードは、MTMineRのCode Conversionで一括変換できる。

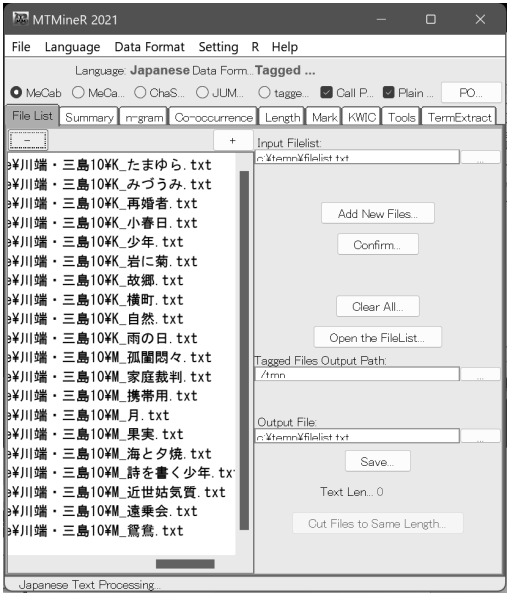
テキストコーパスを読み込むには、ボタン[Add New File]を押して、準備したテキストファイルを読み込む。同時に複数のファイルをリストに繰り返し加えることができる。ファイルの読み込みの指定が終わったらボタン[Conform]を押し、テキストの加工や構成要素の集計に進む。図9 (b) が形態素解析を行うためファイルを読み込んだ画面である。

5.2 集計と分析機能

テキストを構成する要素の集計では、単語と文などの長さ、文字列・形態素・形態素のタグ・文節およびそれぞれのn-gram(n=1〜6) と共起、文節の係り受け、文節のパターンなどについて図3のような表形式でまとめる。集計結果は保存して、使い慣れたツール、またはデータ解析ソフトRと連結して分析することができる。



(a) 起動された画面



(b) 形態素解析を行う画面

図9　MTMineRのGUI画面

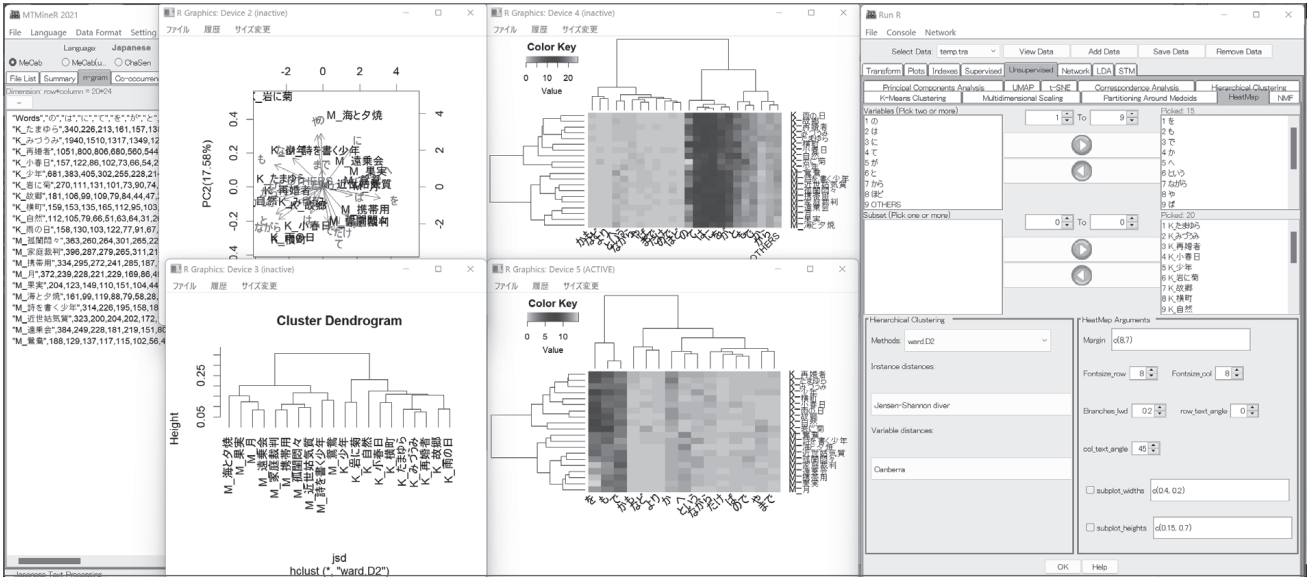


図10　MTMineRの分析画面コピー

分析は実装されている方法の中からメニュー操作で方法を選択し、オプションを指定し、実行ボタンを押すだけで行うことが可能である。本稿に示した例の処理の画面コピーを図10に示す。図10の右側のものがデータ分析を行うためのGUI画面である。データ分析の方法はタブごとに分かれている。

また、メニューに実装していない分析機能に関しては、ツールに実装されているRのコンソール画面上で分析用のRスクリプトを作成し、実行することができる。

MTMineRに関してより詳しいことは、金（2021, p.289-309）、Web上のマニュアルを参照されたい。

6. あとがき

本稿では、100年前から行われているスタイロメトリーの例を用いてテキストアナリティクスのプロセスを説明した。スタイロメトリーの研究は個人やグループ間の表現特徴を抽出して計量的に分析を行っている。したがって、計量的表現研究を考えるとスタイロメトリーが連想される。

参考文献

Lutosławski, W.（1898）. Principes de stylométrie appliqués à la chronologie des œuvres de Platon, Revue des Études Grecques, 11（41）, 61-81.

Mendenhall T. C.（1887）. The Characteristics Curves of Composition. Science, Ⅸ, 237-249.

Mendenhall T. C.（1901）. A mechanical Solution of a Literary problem, Popular Science Monthly, 60, 97-105.

Williams, C. B.（1975）. Mendenhall's Studies of Word-length Distribution in the Works of Shakespeare and Becon, Biometrika, 62, 207-211.

金 明哲（1995）. 動詞の長さの分布に基づいた文章の分類と和語および合成語の比率, 自然言語処理, 2（1）, 57-75.

金 明哲（1996）. 動詞の長さの分布と文章の書き手, 社会情報, 5（2）, 13-22.

金 明哲（1997）. 助詞の分布に基づいた日記の書き手の認識, 計量国語学, 20（8）, 357-367.

金 明哲（2002）. 助詞のn-gramモデルに基づいた書き手の識別, 計量国語学, 23（5）, 225-240.

金 明哲（2003）. 自己組織化マップと助詞分布を用いた書き手の同定及びその特徴分析, 計量国語学, 23（8）, 369-386.

金 明哲（2021）. テキストアナリティクスの基礎と実践, 岩波書店.

（同志社大学）