

MTMineR

マニュアル

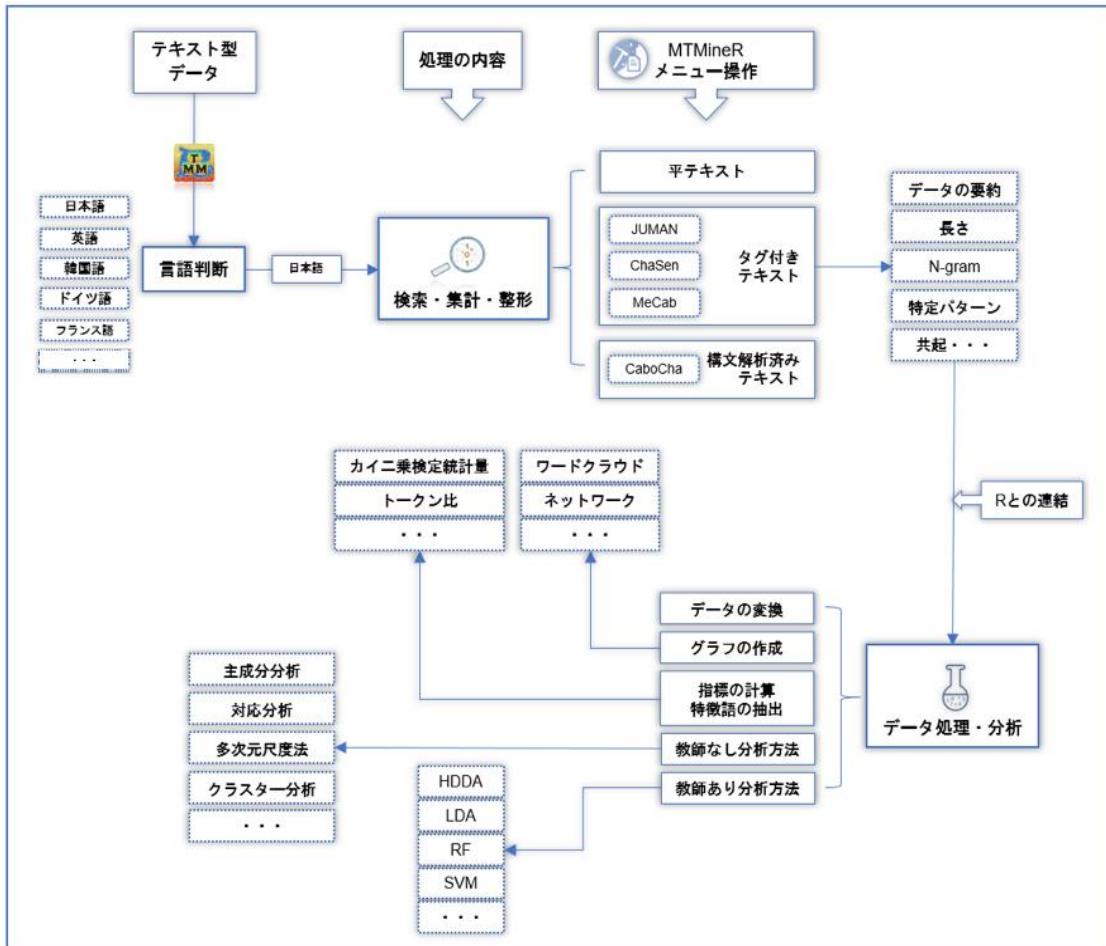
データサイエンス研究室

2018.06

MTMineR の概要

MTMineR(Multilingual Text Miner with R ; エム・ティ・マイナー)とは、テキスト型データを構造化して集計し、R を用いて統計的に分析するソフトウェアである。MLTP を高機能化したバージョンである。文学作品・アンケートの自由記述・新聞記事などさまざまなテキストを処理し、データを集計することができる。テキストの統計的解析を勉強する方々のため、無償で本ツールを公開する。ただし、著作権を放棄することではない。MTMineR では日本語、中国語、韓国語、英語、ドイツ語とフランス語等のデータを扱うことができる。データの構造化では平テキストやタグ付きのテキストから頻度を集計し表形式(行列形式)のデータを出力する。データの統計分析は、各自が使い慣れている統計ツールを用いることができるが、MTMineR は集計したデータを直接メニュー操作でデータ解析ソフト R で分析することもできる。

MTMineR の概略図を次に示している(日本語を例とする)



MTMineR の機能

1. データの検索・集計・整形のサポート機能

○ 平テキスト

平テキストの要約(ファイルのサイズ, 文字数, 文の数, 漢字数, 平仮名数, カタカナ数, ローマ字の数, 数値の数, 全角文字の数, 半角文字の数), 文字・記号の n-gram($n=1, \dots, 6$), 単語・文などの長さの分布, 指定した要素の前後のパターンの集計, KWIC 検索やテキストの前処理などの機能がある.

○ タグ付きテキスト

形態素解析済みのテキストにおけるデータの構造化：データの要約(ファイルのサイズ, 延べ語数, 異なり語数, 片仮名単語数, ローマ字単語数など), タグの n-gram, タグ付きの要素の n-gram, タグ単位の要素の長さの分布（文の長さ, タグ区切りの要素の長さなど）, 指定した要素の前後のパターンなどの集計, タグ単位の KWIC 検索, タグ付きデータの一括整形処理など機能がある.

○ 構文済みテキスト

構文解析済みのテキストにおけるデータの構造化：文節を単位とした長さの分布(文節単位の文の長さ, 文節の長さ), 文節の n-gram, 条件付きの文節の n-gram, 文節の共起(係り受け関係を無視した共起, 係り受け先を考慮した共起), 文節のパターンなどのデータを集計する機能がある.

2. データの処理と分析機能

構造化したデータセットは csv 形式とタブ区切り形式で保存し, データ解析・データマイングツールを用いて分析を行うことができる. 本システムではメニュー操作でデータ処理や解析を行う GUI 環境を備えている. メニュー操作による主な機能は, データの処理・加工, データの視覚化, 指標の計算と特徴語の抽出, 教師なしの分析方法, 教師ありの分析方法に分けられる.

○ データの変換

分析方法によって集計した度数データをテキストの長さに依存しない相対頻度に変換して用いることが必要である。MTMineR では行あるいは列の合計に基づいた比率の変換、データの標準化、行列の転置機能が設けられている。

○ グラフ作成

データの視覚化ワードクラウド、折れ線グラフ、Zipf の法則とグラフ、ネットワークグラフを作成する。

○ 指標の計算と特徴語の抽出

カイ二乗検定統計量、尤度比検定統計量、カルスカル・ワリス検定統計量、ランダムフォレストの正解率、ジニ分散指標などを用いた特徴語抽出機能やローカル TF-IDF、グローバル TF-IDF など 10 種類の重み計算、トークン比や Yull の K 特性値など数種類の語彙の豊富さ指標の計算機能が実装されている。

○ 教師なしの分析方法

主成分分析、対応分析、多次元尺度法、k-平均法、階層的クラスタリング法、トピックモデルなどの教師なしの分析の環境が備えている。クラスター分析、多次元尺度法などに用いる距離としてユークリッド距離、SKLD 距離、対称的カイ二乗距離など八種類の距離が用意されている。

○ 教師ありの分析方法

線形判別、K-近傍法、決定木、ランダムフォレスト、SVM、HDDA が実装されている。

多言語形態素解析

MTMineR は主に**日本語、中国語、韓国語、英語、ドイツ語とフランス語などのテキスト**解析のために作成した。本システムでは、日本語の形態素解析は JUMAN、ChaSen、MeCab、構文解析は CaboCha、中国語の形態素解析は NLPIR、英語、ドイツ語とフランス語などの形態素解析は TreeTagger を借用する。

データの構造化では平テキストやタグ付きのテキストから頻度を集計し表形式(行列形式)

のデータを出力する。データ統計分析は、集計結果を一旦保存し、各自が使い慣れている統計ツールを用いて分析することもできる。また、データ解析のフリーソフト R をインストールし、若干の設定を行うと直接 MTMineR からメニュー操作で用意されているデータ解析の方法を用いることができる。

平テキストは、我々が書いた一般的な文章形式を指す。

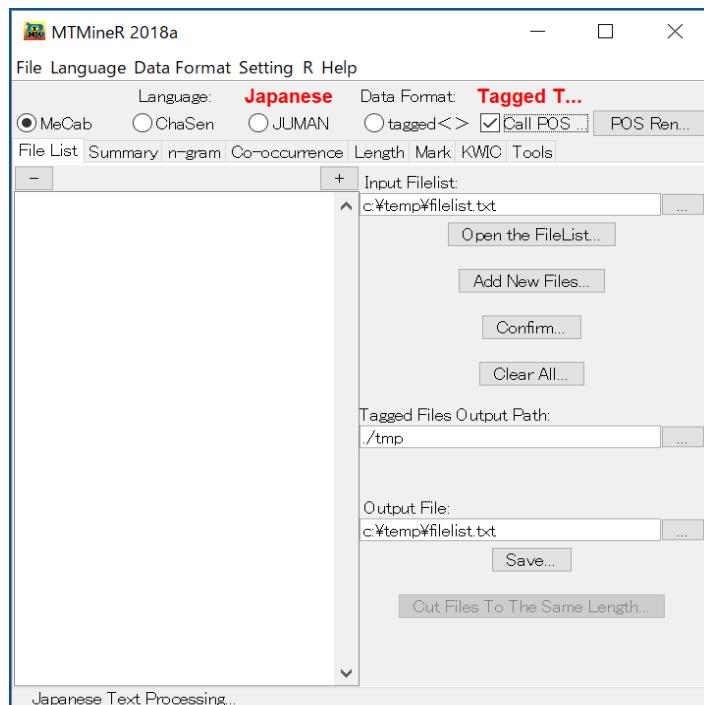
タグ付きは、平テキストを自由に切り分け、その部分の性質を <> の中に自由にタグをつけたテキストを指す。

ここで、MTMineR を用いて、各言語に対して形態素解析の操作を説明する。各形態素解析器を使う前に環境設定が必要であるため、[MTMineR 事前準備マニュアル](#)に参考してください。

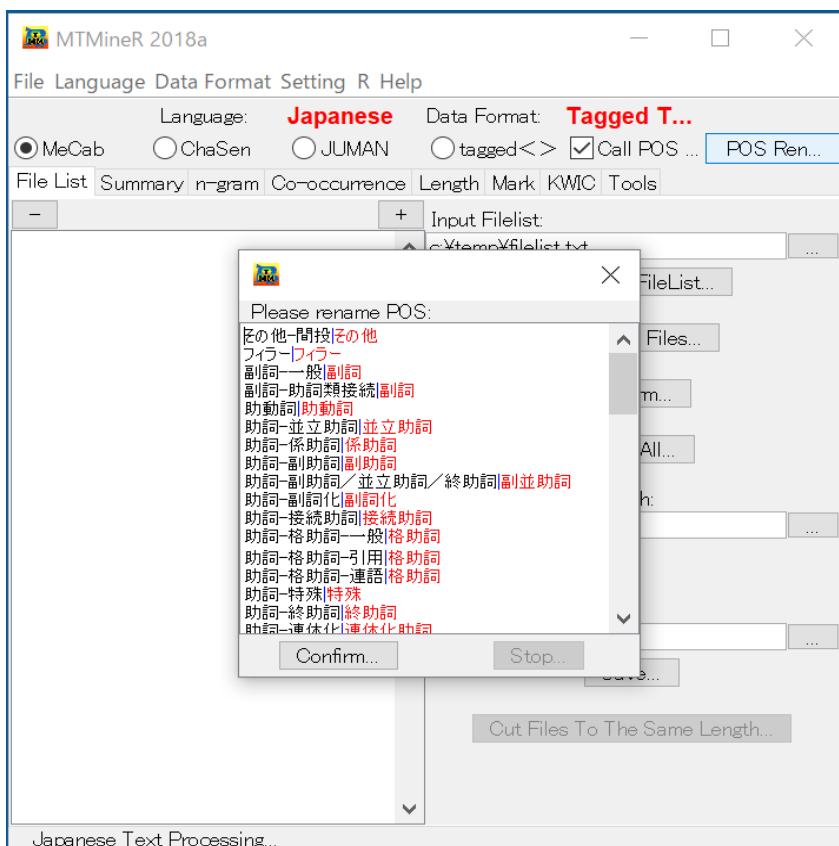
日本語

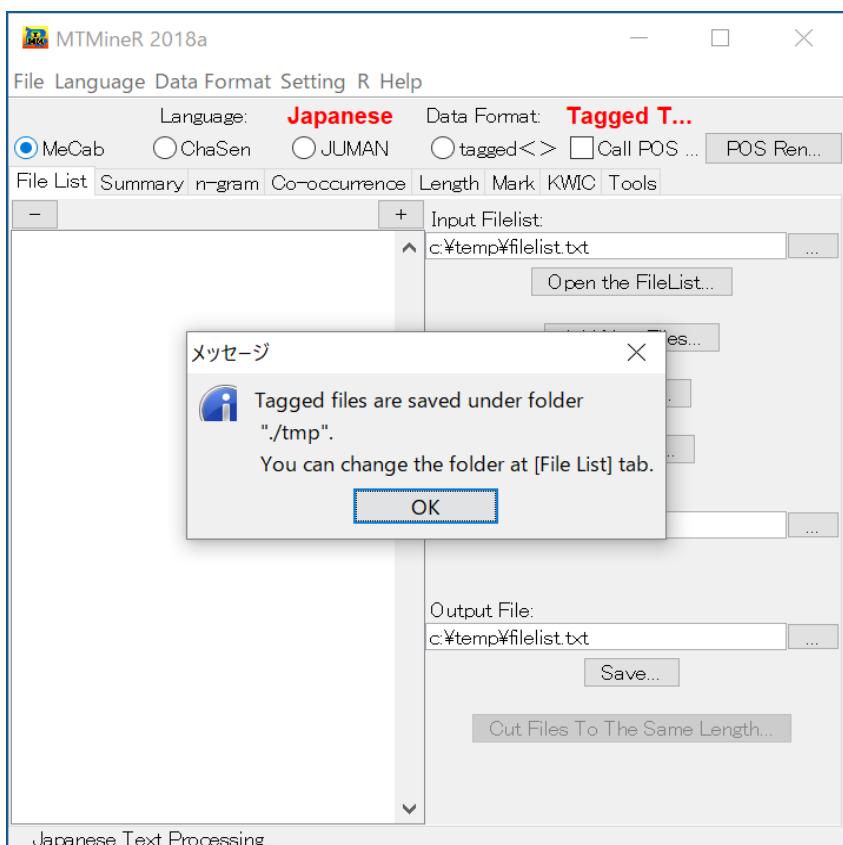
MeCab、ChaSen、JUMAN がインストールされ、パスが通されている環境では、平テキストを読み込み、MTMineR でメニュー操作により形態素解析を行い、タグを付け集計を行うことが可能である。

まず、メニューバーに[language]を選択し、[Data Format]に[Tagged Text]を選択する。



MeCab、ChaSen、JUMANによって形態素解析を行ったテキストを読み込み処理するときには、図に示す画面の上部の三種類の形態素解析器の名前にラジオボタンを選択する。ただし、形態素解析結果の中の品詞は階層化されている。たとえば、助詞「の」の第1層は助詞で、第2層は連帯化助詞になっている。そのまま用いてもよいが、MTMineRでは、各自が自由にタグを命名するステップを置いている。形態素解析器を選択し、さらに[POSRenaming]ボタンを押し、品詞の命名を行う。ボタン[POSRenaming]を押すと品詞を命名する窓が開かれる。黒字は形態素解析器の結果であり、青色縦棒の右の赤文字は自由に書き換えられる形態素の属性である。属性の命名が終わったら確認ボタン[Conform]を押す。これで、日本語形態素解析が終った。形態素解析結果を[temp]ファイルに保存している。

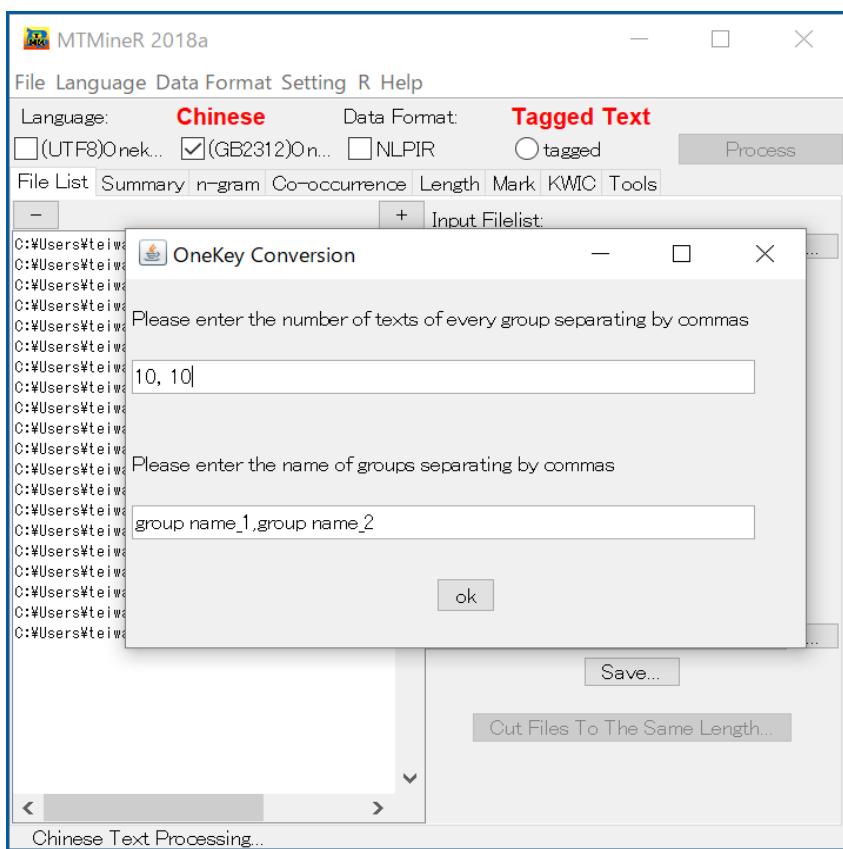




中国語

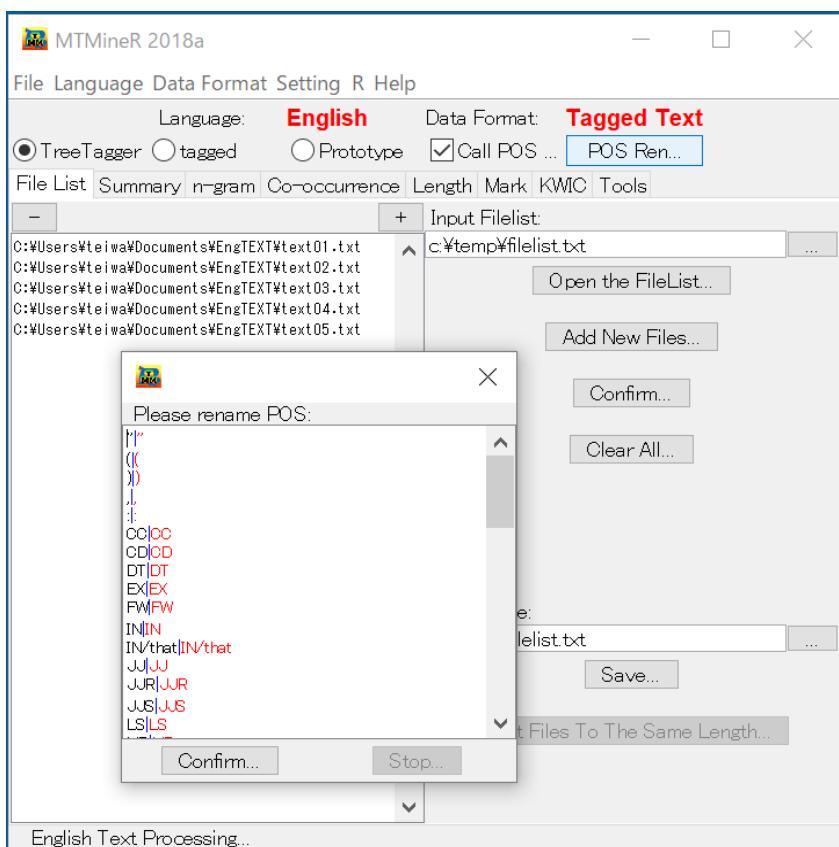
中国語の形態素解析は NLPIR を使っている。パスを通す必要がない。しかし、テキストは UTF-8 で保存する必要がある。また、ファイル名についてアルファベットしか認識できない。ファイル名はアルファベット以外の文字/記号が入っている場合は、[(UTF8)Onekey Convert]と[(GB2312)Onekey Convert]の機能でファイル名とテキストの保存コードを一括で変更できる。

たとえば、元テキストは GB2312 で保存されている場合は、[(GB2312)Onekey Convert]を選択すると、下に示しているようなフレームが出て来る。それぞれのグループのテキスト数とグループ名を設定し、[ok]を押す。生成したファイルを[temp]の中に確認できる。



英語、ドイツ語とフランス語

英語、ドイツ語とフランス語などの形態素解析は共に TreeTagger を借用する。操作は同じであるので、ここで、英語を例として説明する。[treeTagger]、[Call POS Tagger]を選択してから、[POS Renaming]を押し、日本語と同じくタグの名前を変更できる。[Confirm]を押すと、形態素解析を完了する。



TreeTaggerの形態素解析結果を下の図に示している。第1列はテキストの中に用いた単語、第2列はタグ、第3列は単語の原型である。[Prototype]を選択しないまま形態素解析する結果は、[テキストの中に用いた単語/タグ]になっている。一方、[Prototype]を選択すると[単語の原型/タグ]という結果になる。

test_tag - メモ帳		
ファイル(F)	編集(E)	書式(O)
133	CD	@card@
years	NNS	year
ago	RB	ago
,	,	,
Joseph	NP	Joseph
Hardy	NP	Hardy
Neesima	NP	Neesima
broke	VBD	break
new	JJ	new
ground	NN	ground
in	IN	in
Japanese	JJ	Japanese
education	NN	education
and	CC	and

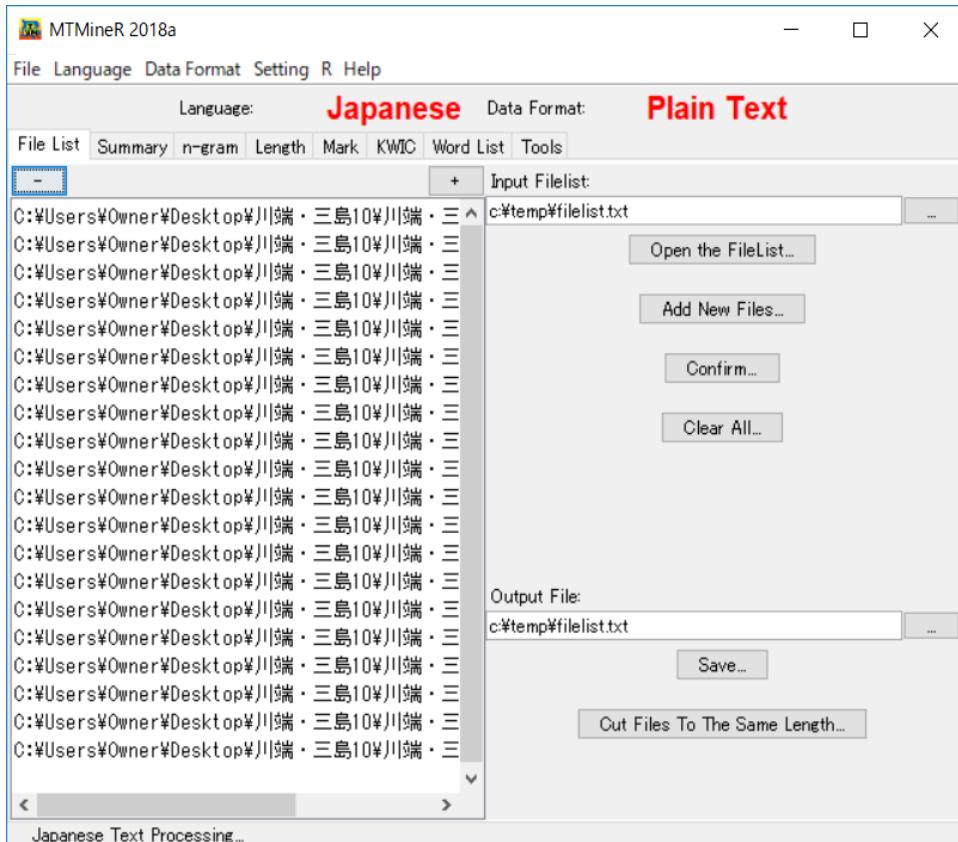
データ収集

平テキスト

平テキストの画面には、メニューの下に8つのタブが用意されている。

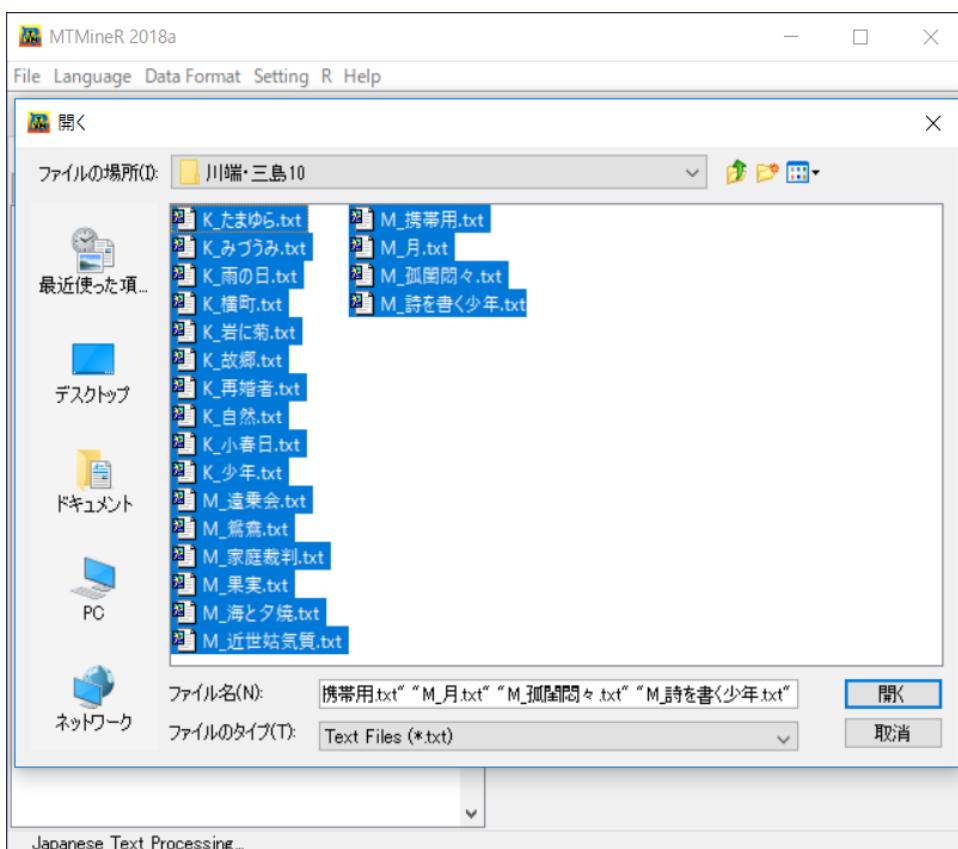
- File List
 - Summary
 - n-gram
 - Length
 - Mark
 - KWIC
 - Wordlist
 - Tools

1. File List (データの読み込み)



タブ File List の右側にボタン「Open the File List」「Add New Files」「Confirm」「Clear All」「Save」「Cut Files To The Same Length」を設けている。

一般的には、ボタン「Add New Files」を用いてファイルの読み込みを行う。ボタン「Add New Files」を押すとファイルが置かれている場所を指定する画面が開かれる。下図のようにドライブ、フォルダ、サブフォルダ順に選択し続け、読み取りたいテキストを選択し、画面上の「開く」ボタンを押すと選択されたテキストが MTMineR のタブ「File List」の左の窓にリストアップされる。



ほかのフォルダ中のファイルを追加したいときには、上記の操作を繰り返す。

「Save」ボタンを用いて、リストアップしたリストを保存しておくことができる。これにより、後日にリストアップしたファイルを利用する時、再びリストアップしなくて済む。ファイルリストの保存は、まず「output File:」の下の窓に保存する場所とファイル名前を指定し、次に「Save」ボタンを押す。ファイルの場所の指定は窓の右側のボタンを用いることが可能である。保存したファイルリストを読み込んで用いる際には、ボタン「Open the file list」を用いて保存しておいたファイルリストを指定する。

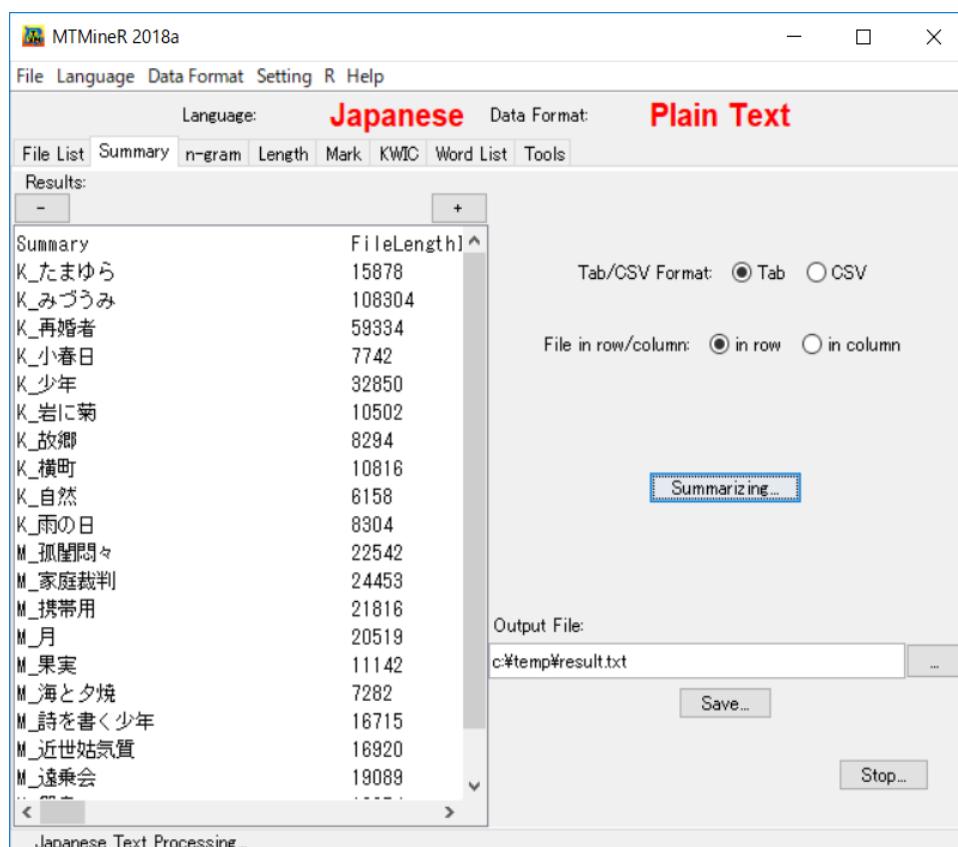
リストアップしたファイルについて処理を行う前に確認ボタン「Confirm」を押す。ファイルリスト画面のクリアはボタン「Clear All」を用いる。

ボタン「Cut Files To The Same Length」は、ファイルと同じの長さに前から切り取るためのボタンである。これを押すと、ファイルの長さ及び出力場所を指定する画面が開かれる。デフォルトではファイルの長さが 200(全角文字数)になっている。テキストの長さと出力場所を指定し、画面上の「OK」ボタンを押すと同じ長さに切り取ったテキストが保存され、同時にそのファイルが MTMineR にリストアップされる。

2. Summary(データの要約)

タブ Summary は、読み込んだテキストについてバイト数 (FileLengthInBytes) , 文字・記号数 (FileLengthInChars) , 文の数 (SentencesNum) , 文字の数 (CharNum) , 漢字の数 (KanjiNum) , 平仮名の数 (HiraganaNum) , 片仮名の数 (KatakanaNum) , ローマ字の数 (RomajiNum) , 数字の数 (NumberNum) , 全角記号の数 (ZenkakuKigoNum) , 半角記号の数 (HankakuKigoNum) を集計する。

ボタン「Summarizing」を押すと集計結果が左側の窓に返される。



データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」は個体(テキスト)を行に、「File in column」は個体を列に表示する。集計したデータを保存する時、保存の場所とファイル名前を指定し、ボタン「Save」を押すと保存される。

3. n-gram

タブ n-gram では、文字単位の n-gram のデータを集計する。

画面の右側の「Ngram Type」下の窓で n を指定する。中には Unigram(n=1), Bigram(n=2), Trigram(n=3), Fourgram(n=4), Fivegram(n=5), Sixgram(n=6) という 6 つの選択肢がある。

集計結果のサイズは cutoff 値（閾値）を用いてコントロールできる。例えば、Cutoff 値を 100 すると全対象テキストにおいて合計の頻度が 100 未満の項目はすべて、1 つの項目 “OTHERS” にまとめる。集計したデータは総度数が大きいものから降順にソートされている。ボタン「Processing」を押すと、集計結果が左側の窓に返される。結果の保存は、Output File の窓にフォルダを指定し、ファイルの名を付け、ボタン「Saving…」を押す。

4. Length(長さの分布)

タブ Length では、文の長さ (Sentence Length) と段落の長さ(Paragraph Length)、リズムの長さ(Rhythm Length)を集計する。文については句点、感嘆符、疑問符を文の終わりと判断する。これらの記号を用いず、改行を文の終わりとしている場合は画面上の「Use line break to split sentence」にチェックをいればよい。リズムはコロン、セミコロン、読点、句点、感嘆符、疑問符をリズムの区切りとする。

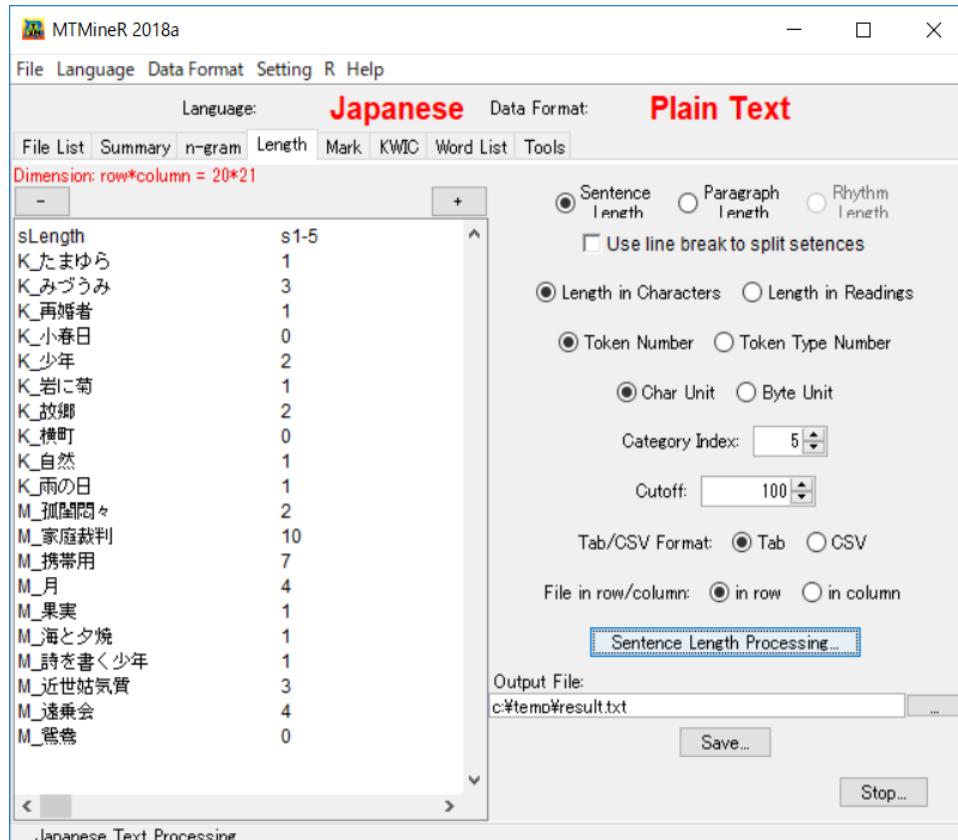
平テキストの場合は、長さを集計する際、一般的には文字単位として「Length in Character」を集計する。Mecab がインストールされている環境では、漢字を読み方に置き換え、読み方による長さ「Length in Reading」を集計することも可能である。

形態素解析済みの場合、一般的には延べ語数「Token Number」を単位として集計するが、異なり語数「Token Type Number」で集計ことも可能である。文字を単位とする時は「Char Unit」を選択し、バイトを単位とする時は「Byte Unit」を選択する。

長さの分布のデータを集計する際、1 文字ごとに一つの変数（項目）にするとデータのサイズ大きくなるので、いくつも文字を 1 つの項目にまとめて集計すると便利である。これは画面上の「Category Index」で自由に指定できる。例えば、文字を単位とした場合、Category

Index が 5 であると 1 文字から 5 文字を 1 項目、6 文字から 10 文字を 1 項目のように集計する。

画面上の「Cutoff」(閾値) を用いて集計サイズをコントロールすることができる。デフォルトは 100 になっている。Cutoff 値が 100 の場合は、100 文字以下の文はすべて一つの項目にまとめて集計する。



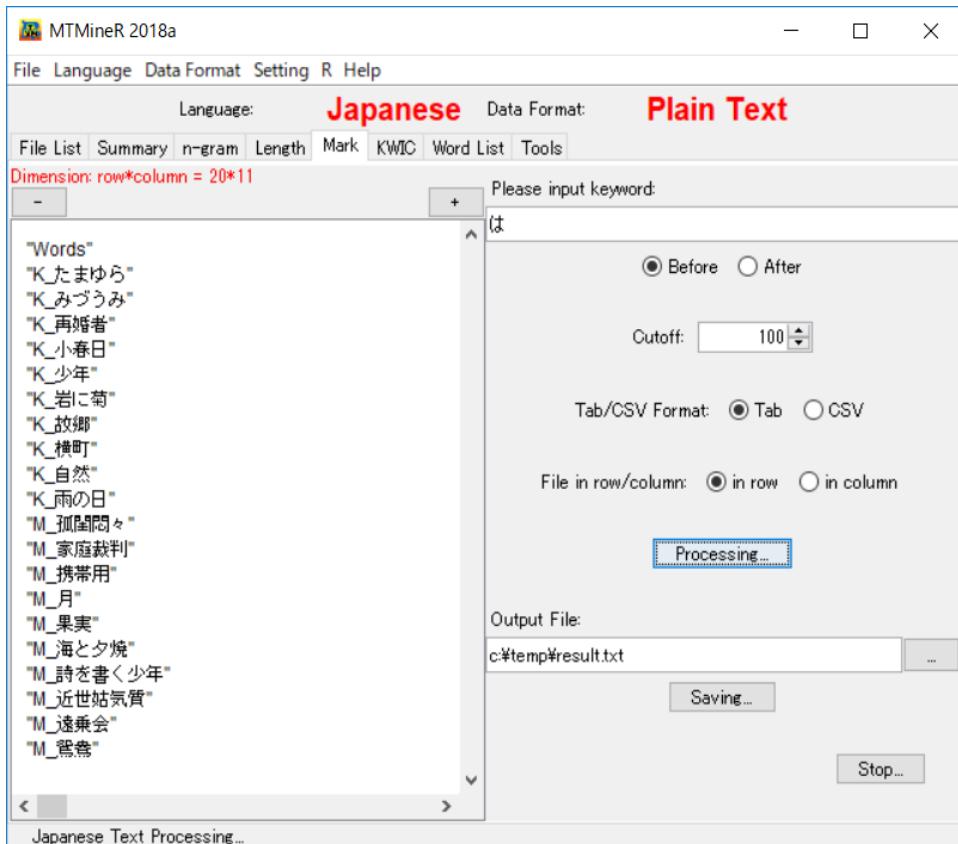
結果の保存は、「Output File」の窓にフォルダを指定し、ファイルの名を付け、ボタン「Save」を押す。

5. Mark (指定文字・記号の前後)

タブ Mark では、ある文字・記号の前後の文字を切り取ったデータを集計する。「Please input keyword」下の窓に指定の文字或は記号を入力して、当該文字・記号がどの文字の前に付いているかを集計する時、「After」にチェックを入れる。逆に当該文字・記号がどの文字の後に付いているかを集計する時、「Before」にチェックを入れる。タブ「Mark」で集計したデータは、文字単位の Bigram の一部分である。

結果の保存は、Output File の窓にフォルダを指定し、ファイルの名を付け、ボタン

「Saving…」を押す。



6. KWIC(クウィック検索)

タブ「KWIC」(Keyword in context)では、指定したキーワードについてすべてのテキストから、その前後の文脈を一定の長さで切り取って返す。右側の画面上の「Please input keyword」下の窓に検索したいキーワードを入力し、画面上の「No. Left」と「No. Right」を用いて前後切り取る長さを自由に指定し、ボタン「Process」を押すと、結果が左側に返される。返された結果は自由にソートすることができる。切り取った部分の前後を基準としたソートは、左側の画面上の「Left」或は「Right」の部分をクリックすると降順、昇順に入れ替わる。

The screenshot shows the MTMineR 2018a application window. At the top, the menu bar includes File, Language, Data Format, Setting, R, and Help. The Language is set to Japanese and the Data Format is Plain Text. Below the menu is a toolbar with tabs: File List, Summary, n-gram, Length, Mark, KWIC, Word List, and Tools. A message 'Obtained 6188 hits.' is displayed. To the right is a processing panel with the following controls:

- Use regular expression
- Please input keyword: は
- No. Left: 10, No. Right: 10
- Process... button
- Sentence button
- Output File: c:\temp\result.txt
- Save... button
- Stop... button

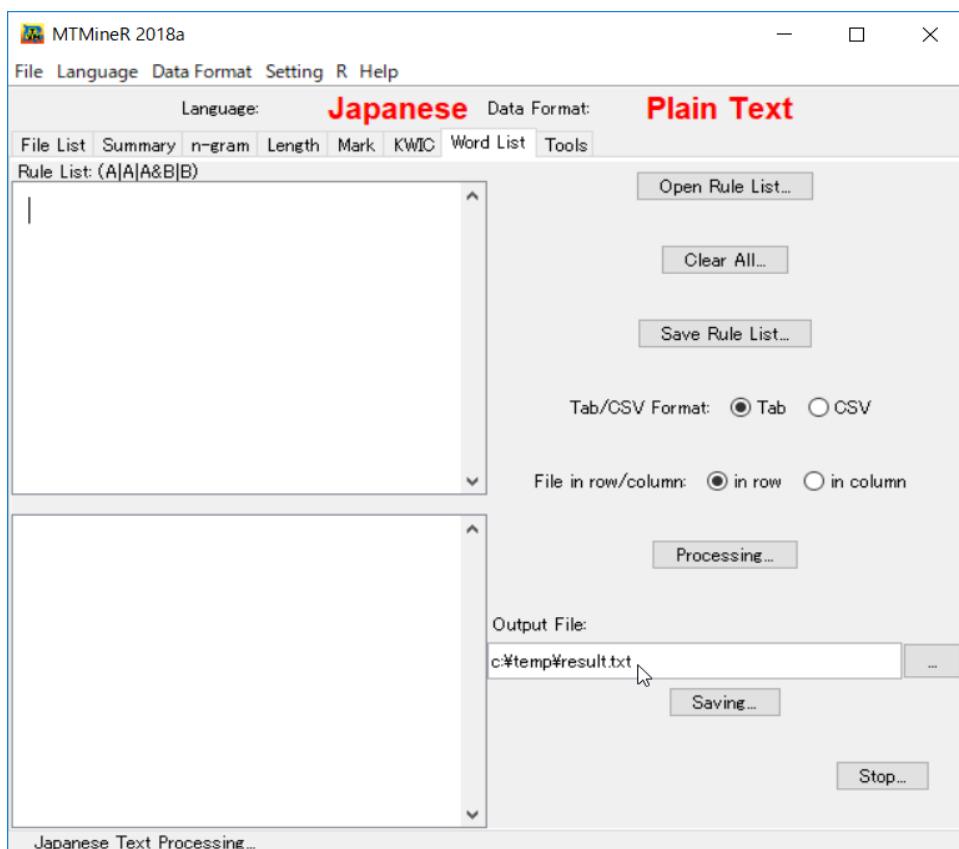
A note at the bottom left says 'Select a row to display the full text.'

返された結果の一行をクリックするとそれが含まれているテキストが左下側の空白欄に返される。また、キーワードは正規表現 (regular expression) で指定することが可能である。[Use regular expression]にチェックを入れると正規表現による KWIC 検索ができる。結果の保存は、「Output File」の窓にフォルダを指定し、ファイルの名を付け、ボタン「Save」を押す。

7. WordList

タブ「Word List」では、各自が作成したワードリストに指定している語句をテキストごとに集計する。ワードリストは直接画面の左側の「Rule List」の窓に直接記述できる。また、文章エディターで作成したファイルを、ボタン「Open Rule List」から読み込み用いることもできる。

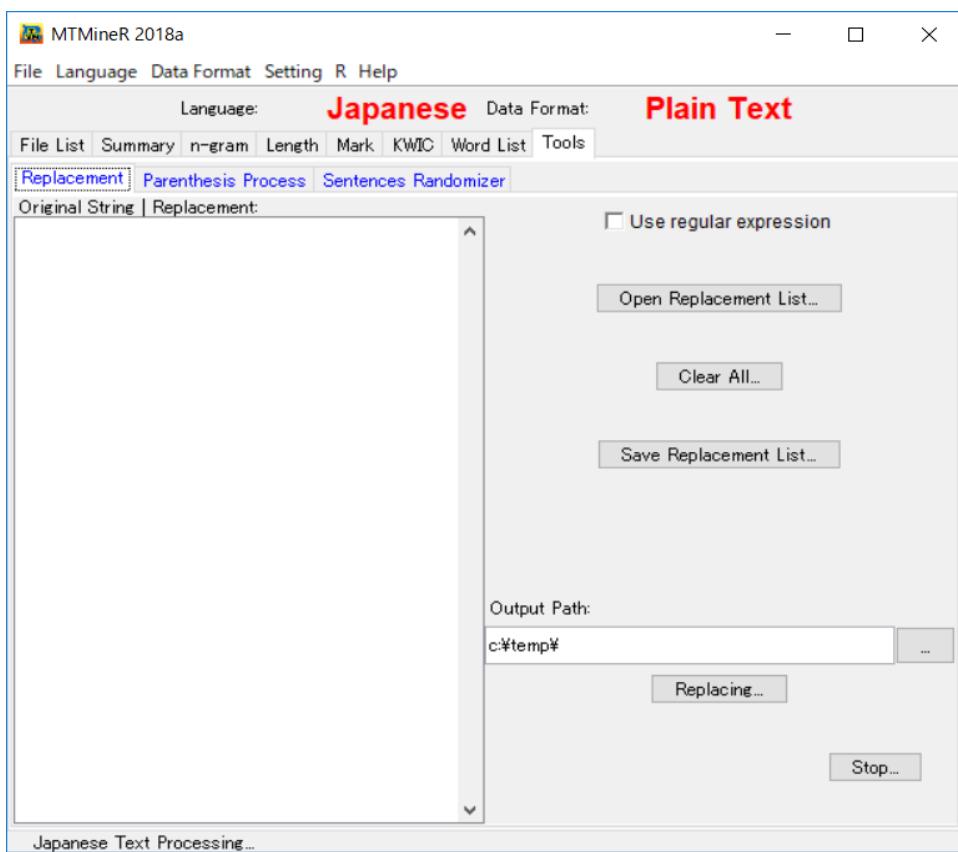
項目の記述は、1 行を一つの項目とする。また、記述には論理演算を用いることができる。かつ (and) 演算は半角の &、また (or) 演算は半角の 縦棒 | を用いる。



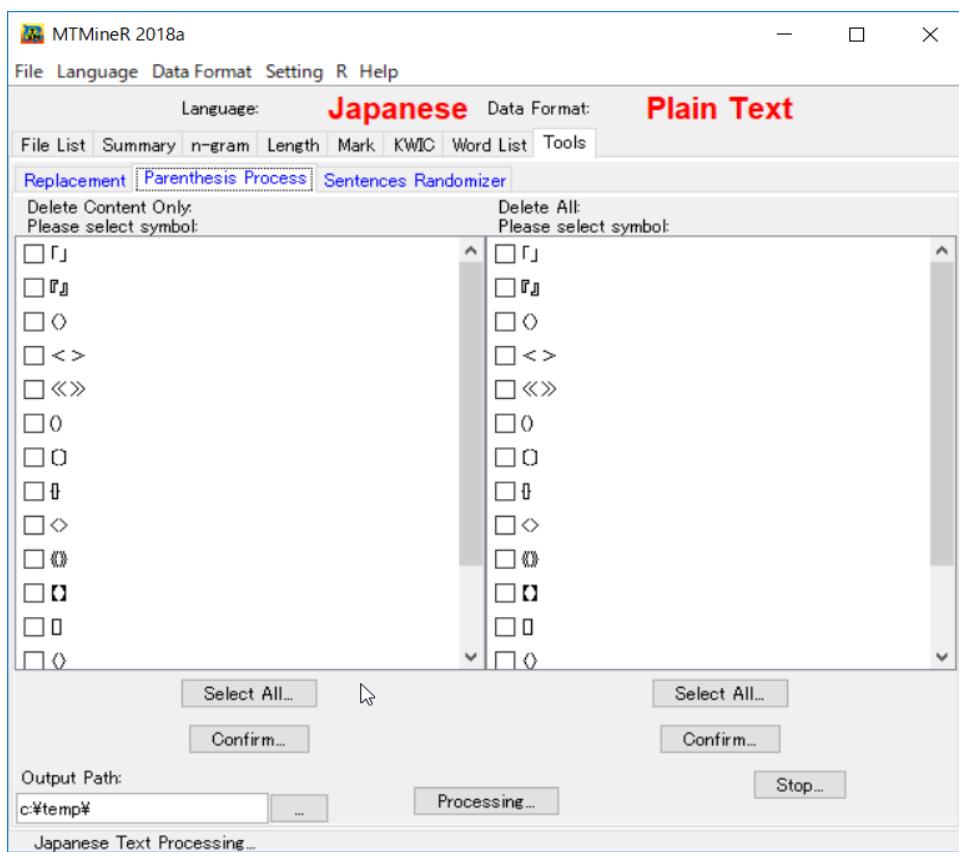
結果の保存は、「Output File」の窓にフォルダを指定し、ファイルの名を付け、ボタン「Save」を押す。

8. Tools (テキストの整形のための小ツール)

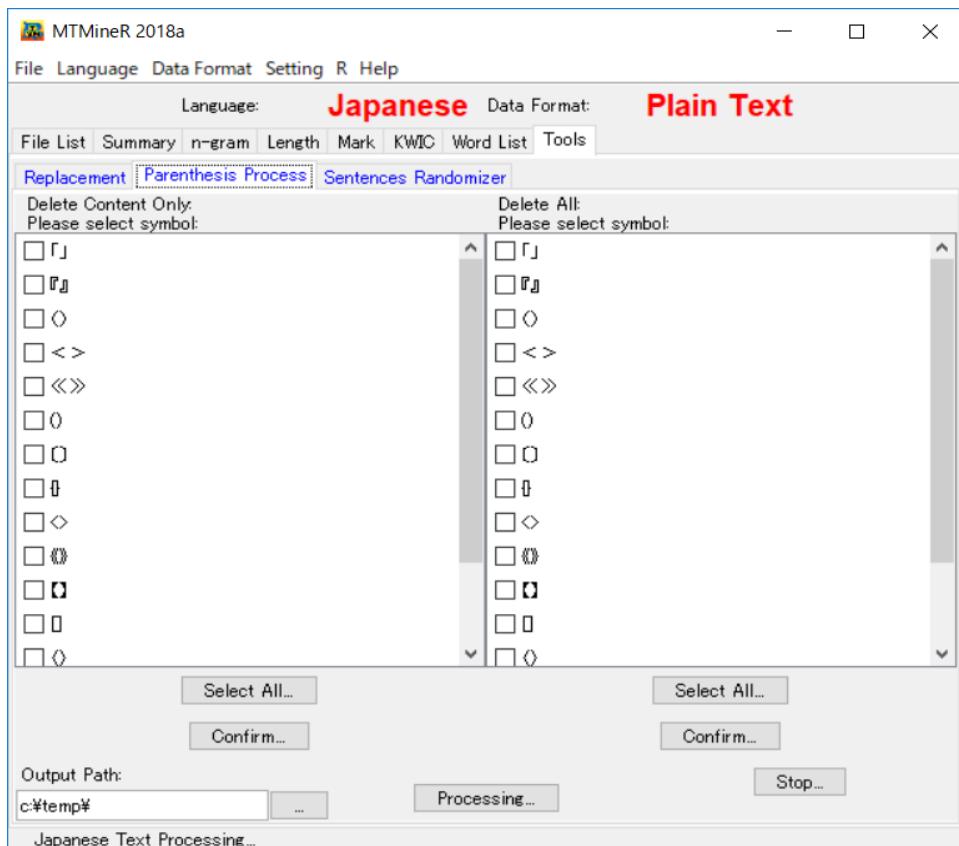
タブ Tools には、「Replacement」「Parenthesis Process」「Sentences Randomizer」という三つのサブタブがある。前の二つはテキストの整形や洗浄に必要な機能である。サブタブ「Replacement」では、テキストの中の記号・文字列を置き換える。記述は 1 行を一項目にする。また、置き換え前と置き換え後の文字列は半角の縦棒 | で切り分ける。また、正規表現を用いて記述することもできる。



サブタブ「Parenthesis Normalizer」では、さまざまな括弧の中のものを削除する機能である。括弧「」の中の会話文を削除したいときには、画面の左部分の「」をチェックし、その下の確認ボタン「Conform」押す。括弧「」の中身だけ削除したいときには、出力場所を指定した上でボタン「Normalizing」を押す。括弧「」の記号も削除したいときには右側の「」の前にチェックを入れる。



サブタブ「Sentences Randomizer」ではテキストから、ランダムに文を取り出す。取り出す文の数は画面の右側の「Num of Sentences」の窓で自由に指定できる。また、取り出すファイルの数は「Num of Files」の窓で選択できる。画面の右側の「Ngram Type」下の窓で n を選択できる。ボタン「Processing」を押すと、結果は左側の窓 Results に返される。抽出したデータを保存する時、保存場所とファイル名前を「Save」ボタン上の窓に指定し、ボタン「Save」を押すと保存される。



タグ付きテキスト

タグ付きのデータは大きく 3 種類：

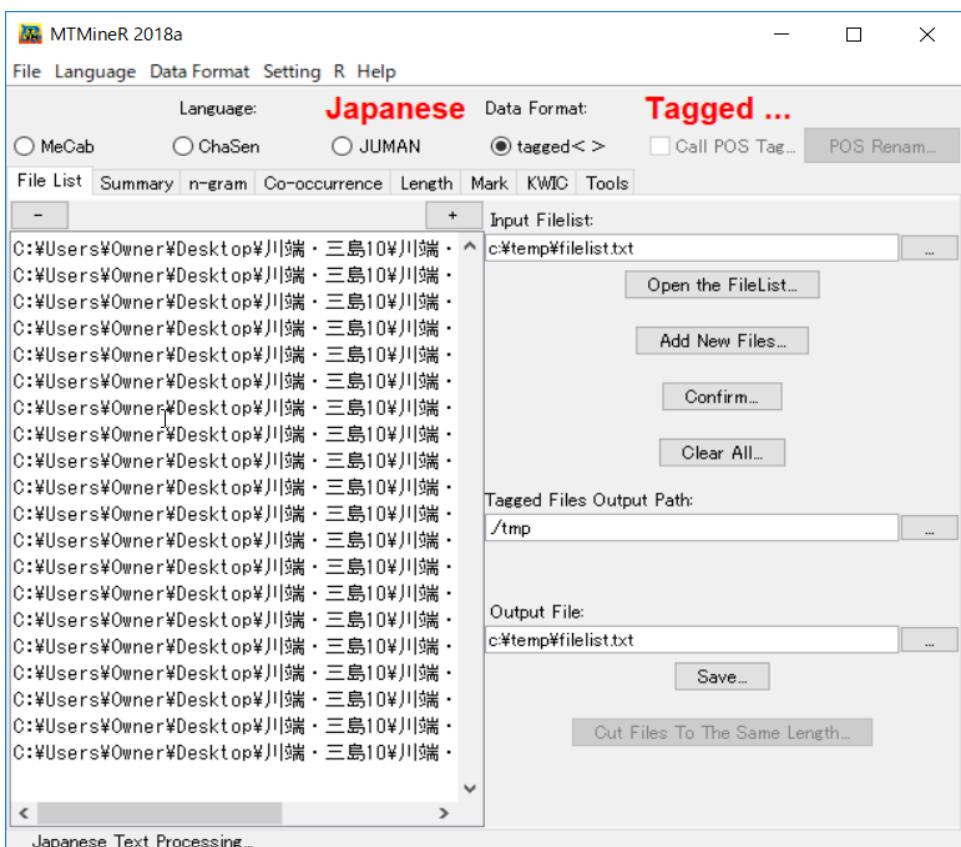
1. 自由に作成したタグ付きテキスト
 2. 形態素解析器により形態素の属性が付けられたテキスト
 3. Cabocha により文節に切り分けたテキストに分けられる

(1) (2) の処理は、メニューの「Data Format」から Tagged Text を選択する。デフォルトには自由に作成したタグ付きテキストの処理環境である。自由にタグをつける際のタグは全角記号 <> 中に記入する。

MeCab、ChaSen、JUMAN がインストールされ、パスが通されている環境では、平テキストを読み込み MTMineR でメニュー操作により形態素解析を行い、タグを付け集計を行うことが可能である。

下図から分かるように、これらのタグ付きテキストについても 8 つのタブが用意されている。

- [File List](#)
- [Summary](#)
- [n-gram](#)
- [Co-occurrence](#)
- [Length](#)
- [Mark](#)
- [KWIC](#)
- [Tools](#)

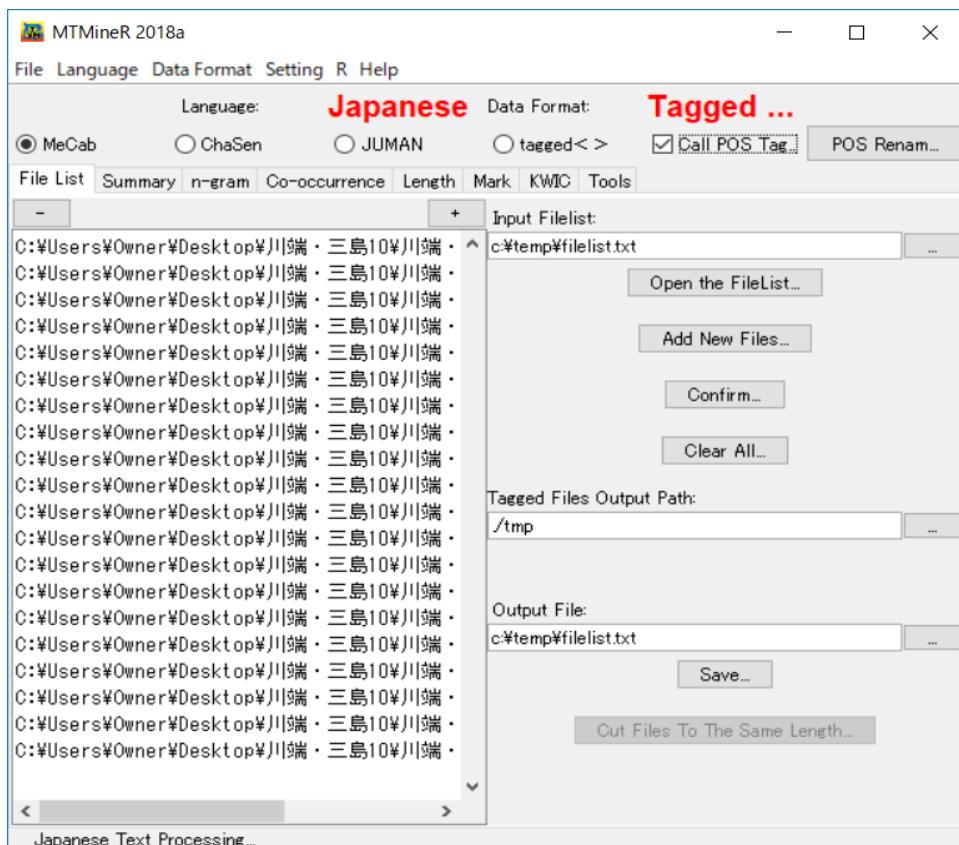


1. File List (データの読み込み)

MeCab、ChaSen、JUMAN によって形態素解析を行ったテキストを読み込み処理するときには、画面の上部の三種類の形態素解析器の名前にラジオボタンを押す。ただし、形態素解析結果のファイルを読み込んで用いるときには、形態素解析結果の形式は表 1 に示す通りとする。表 1 から分かるように JUMAN の出力結果の中の品詞は階層化されていないのでそのまま用いてもよいが、ChaSen と MeCab の品詞は階層化されている。たとえば、助詞

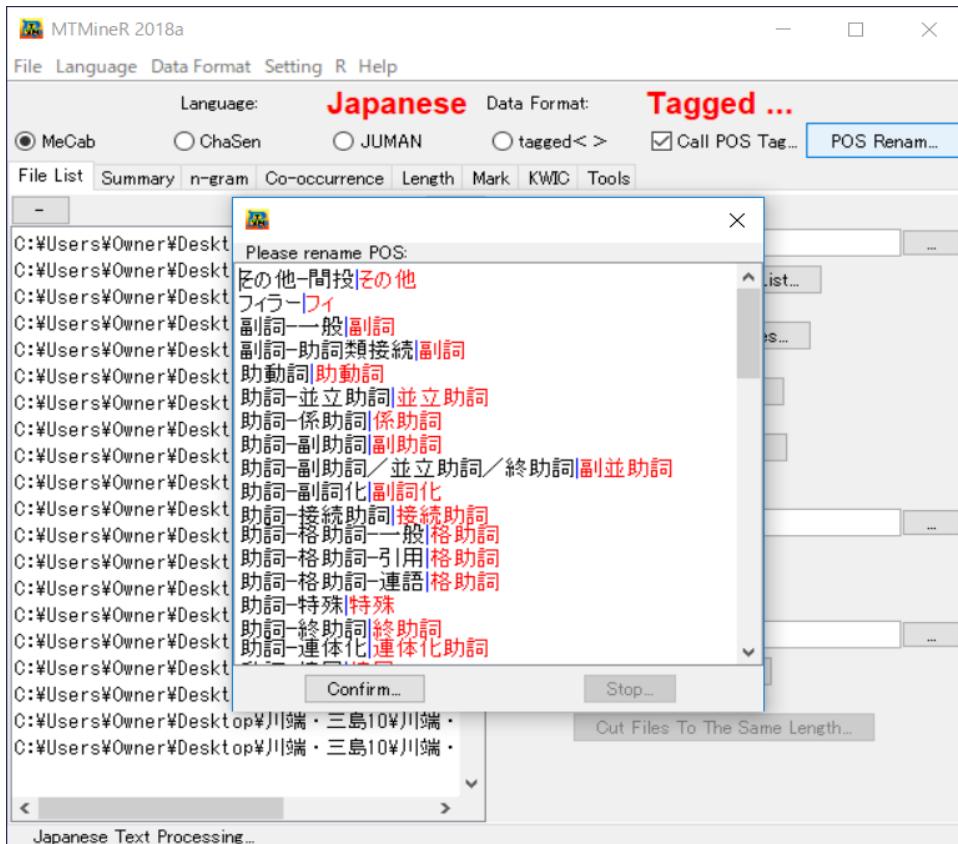
「の」の第1層は助詞で、第2層は連帯化助詞になっている。品詞タグをどのように付けて集計するかは集計者の考えによって異なるが、MTMineRでは、各自が自由にタグを命名するステップを置いている。

ボタン「Add New Files」を用いてファイルの読み込みを行う。ボタン「Add New Files」を押すとファイルが置かれている場所を指定する画面が開かれる。ドライブ、フォルダ、サブフォルダ順に選択し続け、読み取りたいテキストを選択し、画面上の「開く」ボタンを押すと選択されたテキストが MTMineR のタブ「File List」の左の窓にリストアップされる。形態素解析済みのテキストを読み込み、形態素解析器の名前のラジオボタンを選択すると、下図に示すメッセージボックスが開かれる。これは処理したテキストを MTMineR の中の tmp というフォルダに保存することを知らす。保存場所は画面上の「Tagged Files Output Path」で自由に指定できる。



ChaSen と MeCab の結果の場合は、さらに「POS Renaming」ボタンを押し、品詞の命名を行う。ボタン「POS Renaming」を押すと下図のような品詞を命名する窓が開かれる。黒字は形態素解析器の結果であり、青色縦棒の右の赤文字は自由に書き換えられる形態素

の属性である。属性の命名が終わったら確認ボタン「Conform」を押す。



平テキストを読み込み、MTMineR 上で形態素解析処理を行うためには、画面右上の「Call POS Tagger」の前にチェックを入れることが必要である。

「Save」ボタンを用いて、リストアップしたリストを保存しておくことができる。これにより、後日にリストアップしたファイルを利用する時、再びリストアップしなくて済む。保存する前にまず保存場所を「Save」ボタン上の窓に指定する。保存したファイルリストを読み込んで用いる際には、ボタン「Open the file list」を用いて保存しておいたファイルリストを指定する。

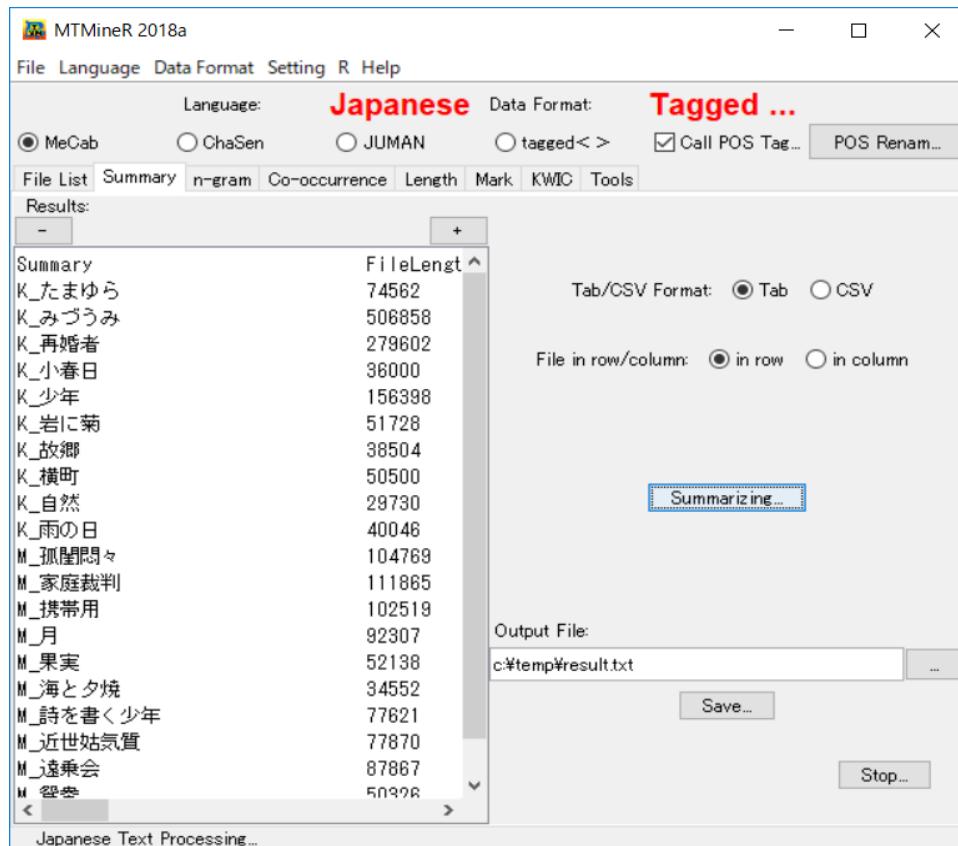
リストアップしたファイルについて処理を行う前に確認ボタン「Confirm」を押す。ファイルリスト画面のクリアはボタン「Clear All」を用いる。

2. Summary(データの要約)

タブ「Summary」では、半角と全角によるテキストのサイズ (File Length In Byte, Files Length In Char)、述べ語数 (Token Num)、異なり語数 (Token Type Num)、片仮

名語の数 (Katakana Token Num)、ローマ字語の数 (Romaji Token Num)、数値の数 (Number Token Num) を集計する。ボタン「Summarizing」を押すと集計結果が左側の窓に返される。

データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」はデータを行で、「File in column」はデータを列で表示する。集計したデータを保存する時、保存の場所とファイル名前を指定し、ボタン「Save」を押すと保存される。



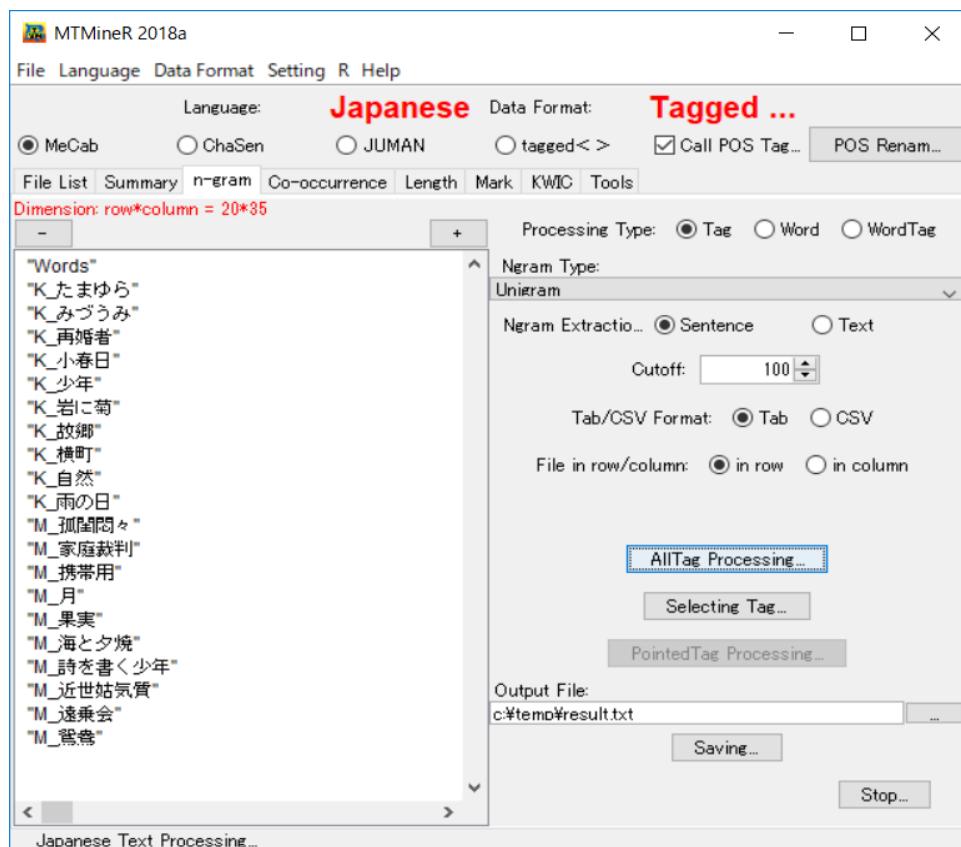
3. n-gram

タブ「n-gram」ではタグ(Tag)、形態素 (Word)、タグ付いた形態素 (Word Tag) の n-gram を集計する。また、すべての属性と指定した属性のみを分けて集計することもできる。まず、処理の種類 (Processing Type) の Tag、Word、WordTag から一つを指定する。そして、「Ngram Type」下の窓で n を選択する。中には Unigram(n=1), Bigram(n=2), Trigram(n=3), Fourgram(n=4), Fivegram(n=5), Sixgram(n=6) という 6 つの選択肢がある。次に「Cutoff」を用いて集計サイズをコントロールする。デフォルトは 100 に

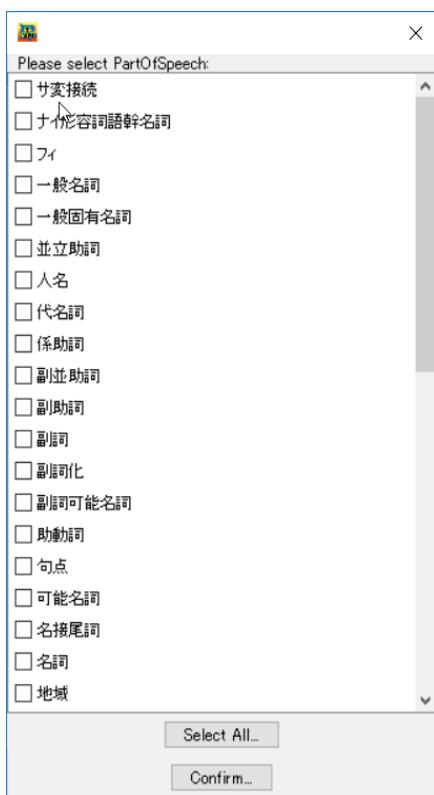
なっている。Cutoff 値が 100 の場合は、100 文字以下の文はすべて一つの項目にまとめて集計する。

データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」はデータを行で、「File in column」はデータを列で表示する。

ボタン「All Tag Processing」を押すと集計結果が左側の Results 窓に返される。下側の左図はタグを集計した画面である。右図はタグ付きの形態素を集計した画面である。処理の種類(Processing Type)を Word 或は WordTag に指定すると、項目数は下図と同じであるが、タグ付いていない形態素またはタグ付きの形態素形態素の集計結果を返す。



タグの種類を指定し、集計を行うためにはタグを指定するボタン「Selecting Tag」を押し、タグを指定することが必要である。ボタン「Selecting Tag」を押すと、次のようなタグ選択画面が表れる。



タグの前にチェックを入れ、確認ボタン「Confirm」を押し、画面上の「Pointed Tag Processing」を押すと集計結果が左側の窓に返される。

4. Co-occurrence (共起)

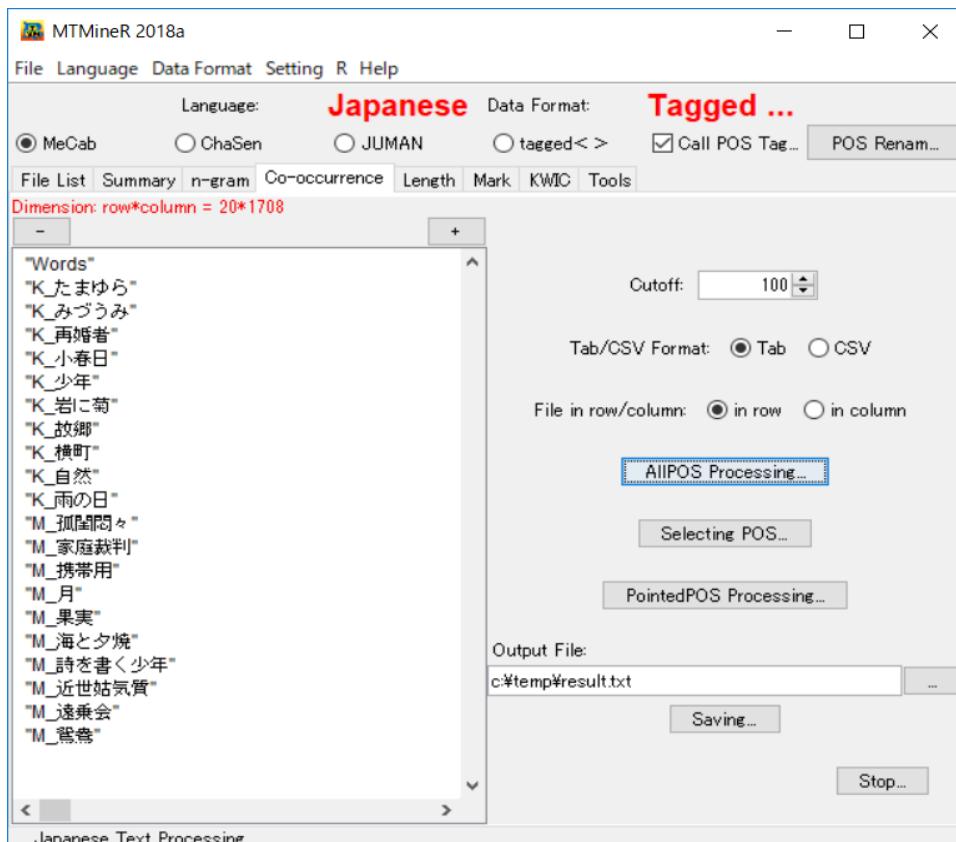
タブ「Co-occurrence」では、形態素の共起データを集計する。集計するのはタグに基づいた形態素の共起である。n-gram の場合と同じく、すべての形態素の共起と指定したタグのみの形態素の共起を集計することができる。

「Cutoff」を用いて集計サイズをコントロールする。デフォルトは100になっている。Cutoff 値が 100 の場合は、100 文字以下の文はすべて一つの項目にまとめて集計する。

データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」はデータを行で、「File in column」はデータを列で表示する。

ボタン「AllPOS Processing」を押すと、全ての形態素の共起についての集計結果が左側の Results 窓に返される。指定したタグのみの形態素の共起を集計したい時に、タグを指定するボタン「Selecting POS」を押し、タグを指定することが必要である。ボタン「Selecting POS」を押してタグの選択ができる。

タグの前にチェックを入れ、確認ボタン「Confirm」を押し、画面上の「PointedPOS Processing」を押すと集計結果が左側の窓に返される。



5. Length (文の長さ、形態素の長さ)

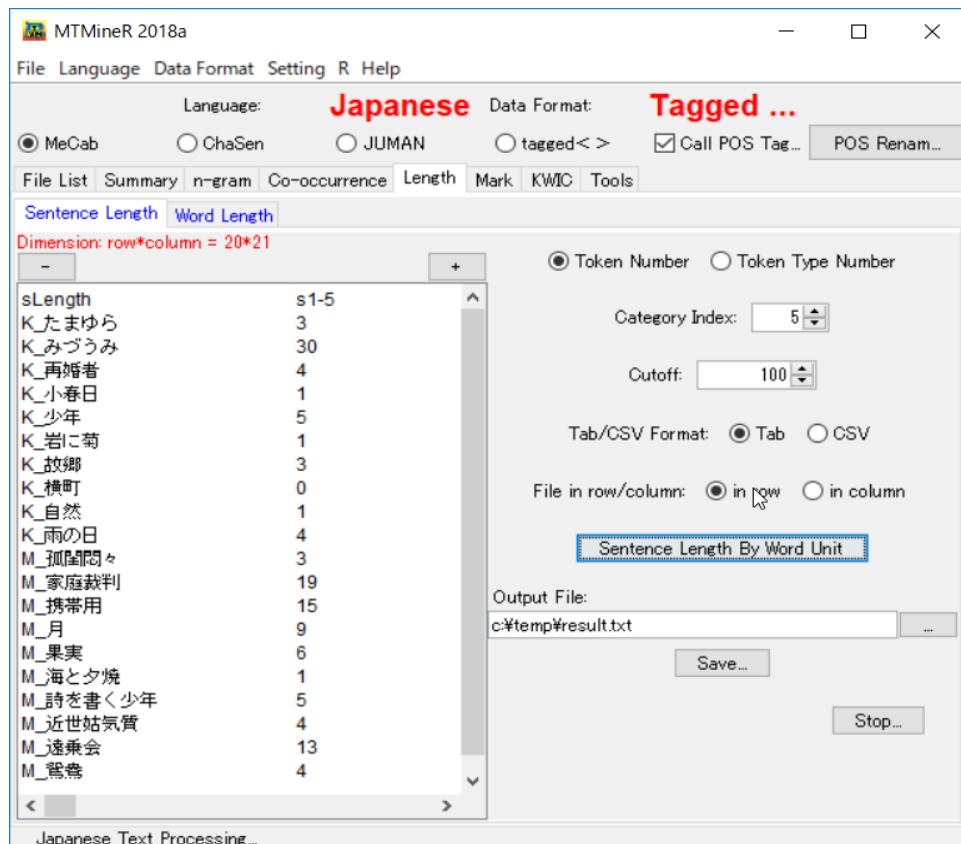
タブ「Length」には、「Sentence Length」、「Word Length」という二つのサブタブがあり、それにより、文の長さ、文字単位で形態素の長さを集計することができる。

文の長さを集計するとき、延べ語数「Token Number」で数えることができ、異なり語数「Token Type Number」で数えることも可能である。

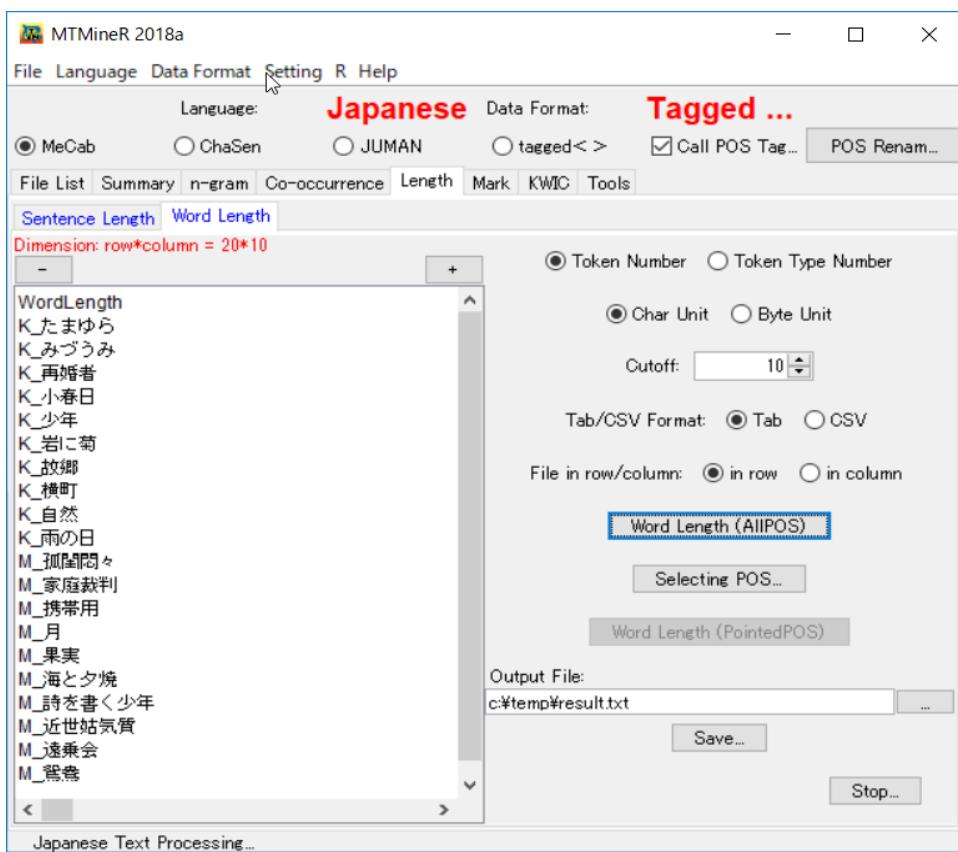
画面上の「Category Index」により、いくつの形態素を1つの項目にまとめて集計することを指定できる。k個の形態素をまとめて一つの項目にする時には、Category Indexの窓に数値kを指定してください。Cutoff値が100の場合は、100文字以下の文はすべて一つの項目にまとめて集計する。

データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」はデータを行で、「File in column」はデータを列で表示する。

ボタン「Sentence Length By Word Unit」を押すと、文字を単位とする文の長さについて集計結果が左側の窓に返される。



文字単位で形態素の長さを集計する時のやり方も大体同じである。文字を単位とする時は「Char Unit」を選択し、バイトを単位とする時は「Byte Unit」を選択する。そして、ボタン「Word Length(AllPOS)」を押すと、全部の形態素の長さについての集計結果が左側の窓に返される。特定したタグの形態素の長さを集計する時、タグを指定するボタン「Selecting POS」を押し、タグ選択画面が開かれる。対象タグの前にチェックを入れ、確認ボタン「Confirm」を押し、画面上の「PointedPOS Processing」を押すと集計結果が左側の窓に返される。

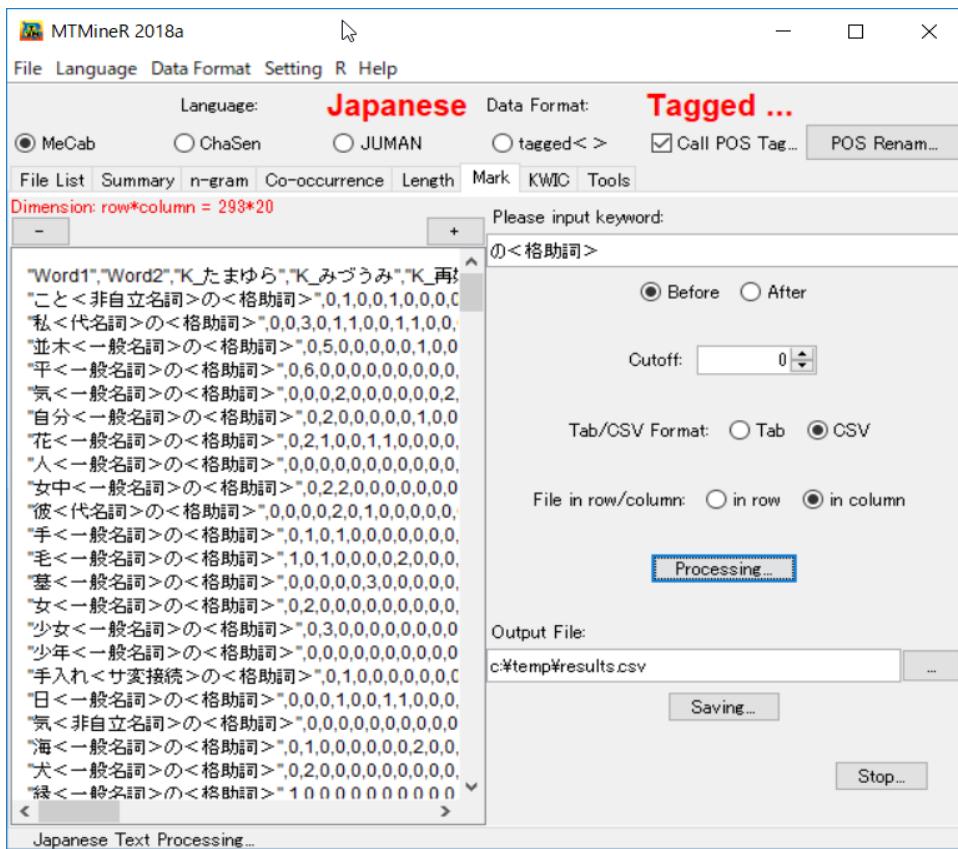


6. Mark (指定した形態素の前後の形態素)

タグ「Mark」では、ある形態素の前後のデータを集計する。たとえば、格助詞「の」がどの形態素の前に位置するかを集計する際には、キーワードを記述する「Please input Keyword」の窓に「の<格助詞>」のカギ括弧の中のものを入力する。画面上のボタン「Processing」を押すと、集計結果が左側の窓に返される。

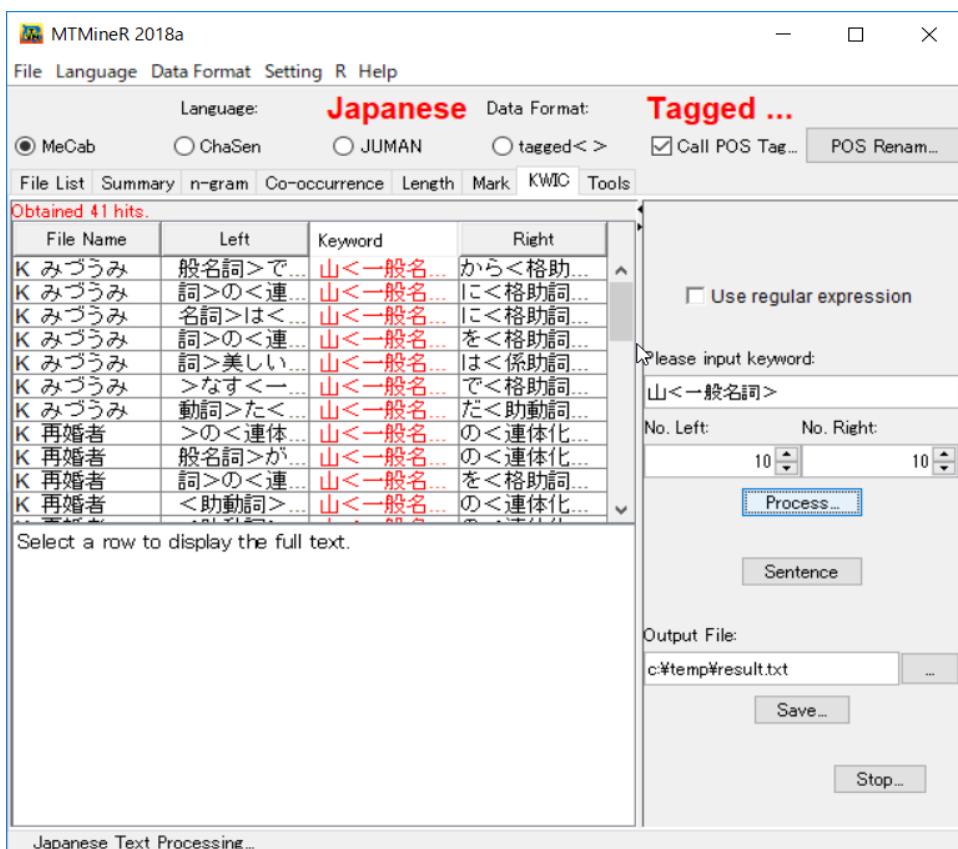
尚、「Cutoff」を用いて集計サイズをコントロールすることができる。Cutoff 値が 100 の場合は、100 文字以下の文はすべて一つの項目にまとめて集計する。

データの形式は画面の右側のラジオボタンで指定できる。「Tab format」はデータをタブで区切り、「CSV format」はデータをコンマで区切る。「File in row」はデータを行で、「File in column」はデータを列で表示する。



7. KWIC (タグ付きの KWIC 検索)

タグ「KWIC」では、タグ付きのテキストから指定したキーワードの前後を切り取り返す。たとえば、一般名詞「山」について全てのテキストから、その前後の文脈を一定の長さで切るとき、キーワードを記述する「Please input Keyword」の窓に「山<一般名詞>」のカギ括弧の中のものを入力し、画面上の「No. Left」と「No. Right」を用いて前後切り取る長さを指定し、ボタン「Process」を押すと、結果が左側に返される。返された結果は自由にソートすることができる。切り取った部分の前後を基準としたソートは、左側の画面上の「Left」或は「Right」の部分をクリックすると降順、昇順に入れ替わる。



返された結果の一行をクリックするとそれが含まれているテキストが左下側の空白欄に返される。また、画面の右側の「Use regular expression」の前にチェックを入れれば、キーワードを正規表現 (regular expression) で指定できる。

8. Tools

タグ「Tools」には、「Format Converter」と「Replacement」という二つのサブタブがある。サブタブ「Format Converter」では JUMAN、ChaSen、MeCab で形態素解析を行った結果をカギ括弧<>でタグ付けるなどの処理を行うことができる。「Replacement」では置き換え処理ができる。

構文解析

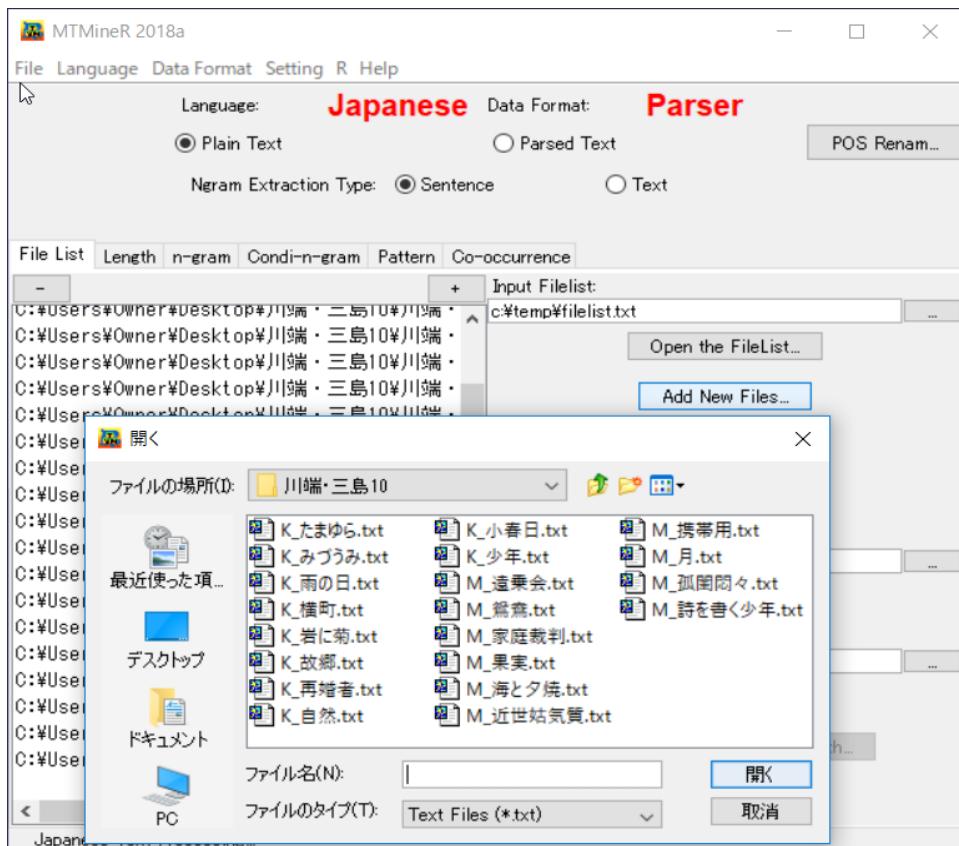
構文データ集計については 6 つのタブが用意されている：

- ●File List
- ●Length
- ●n-gram
- ●Condi-n-gram
- ●Pattern
- ●Co-occurrence

上にある「Ngram Extraction Type」は、「n-gram」と「Pattern」のところで n-gram を集計する時の設定である。「Sentence」を選択する場合は、文ごとに n-gram を集計する。「Text」を選択する場合は、n-gram は文章全体として n-gram を集計する。

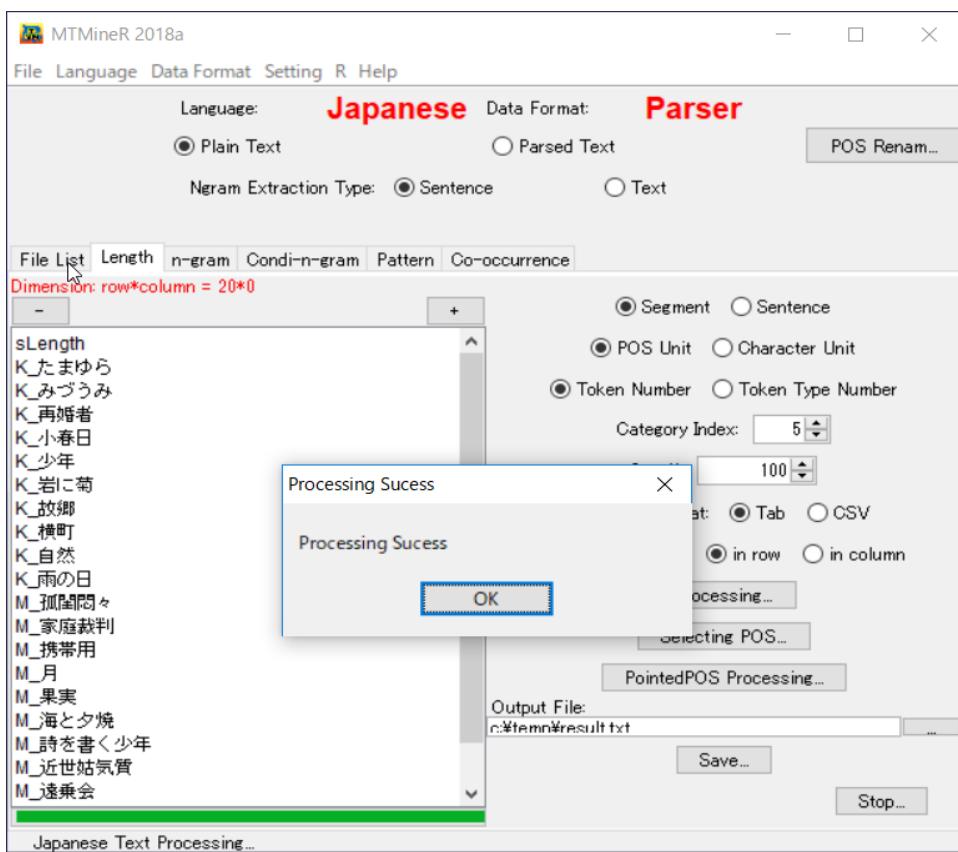
1 .File List (データの読み込み)

タブ File List ではデータの読み込みを行う。まず、MTMineR のメニューの「Data Format」から Parser を選択する。デフォルトは Parsed Text となっている。構文解析済みおよび整形したファイルを用いて集計する際は Parsed Text のまま行う。平テキストを用いて構文解析を行ったうえでデータを作成したいときには Plain Text を選択する。平テキストを用いる際には CaboCha がインストールされ、パスが通されている環境で行う必要がある。File List の右側にある二つ目の Add New Files をクリックしテキストファイルを読み込む。読み込みたいファイルを選択し、「開く」をクリックすると選択したテキストが File List の左の窓にリストアップされる。平テキストを選択した場合、POS Renaming のタブをクリックし、Confirm をクリックすると構文解析を行うことができる。この際に赤字で書かれたものを書き換えることで品詞タグをどのように付与するか各自で自由に決めることができる。



2.Length (文節を単位とした長さの分布)

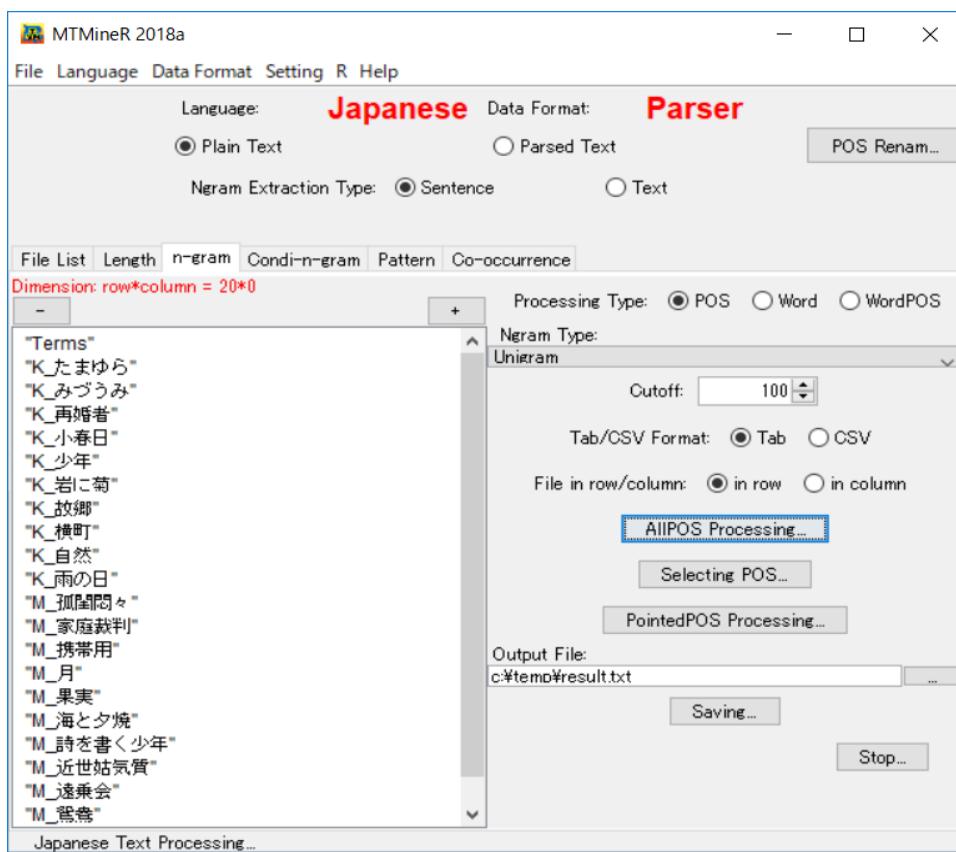
タブ Length では、文節を単位とした長さの分布(文節の長さ、文節単位の文の長さ)を集計することができる。下図の右側にある選択肢をクリックし、Processing を押すと結果が左側の窓に出力される。文節の長さは Segment、文節単位の文の長さは Sentence で出力することができ、長さの単位は形態素 (POS Unit)、文字 (Character Unit)と述べ語数 (Token Number)、異なり語数 (Token Type Number)が選択可能である。Category index には一つの項目にいくつの要素をまとめるかに関する単位を入力する。文字を単位としていないときには一般的には 1 にする。Cutoff 値は 100 に設定されている。これは長さが 100 以下のものは一つの項目にまとめることを意味する。これを変更したい場合は、Cutoff の数字を変えればよい。出力結果の形式は Tab/CSV Format と File in row/column を指定する。Tab format はデータをタブで区切り、CSV format はデータをコンマで区切る。File in row は個体を行で、File in column は個体を列で表示する。集計したデータを保存する際には、保存の場所とファイル名前を指定し、Save をクリックする。



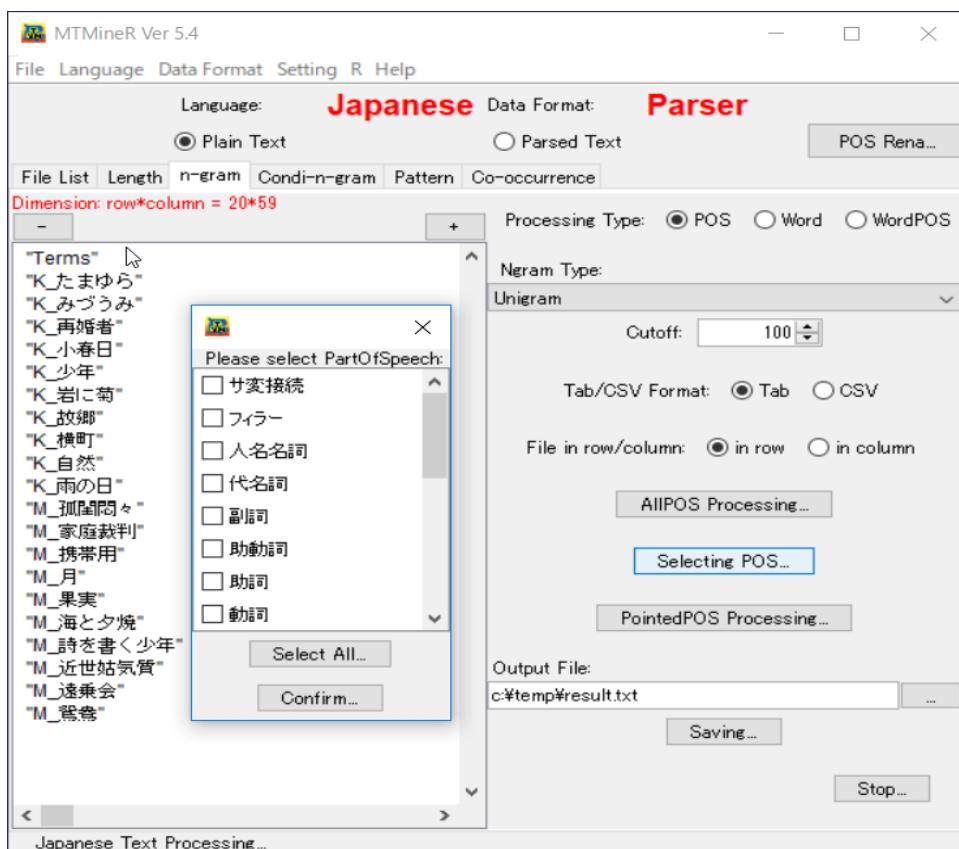
3. n-gram (文節の n-gram)

タブ n-gram では文節単位の属性の n-gram、属性付きの文節の n-gram、指定した属性のみを含んだ文節の n-gram を集計することができる。

集計したい形態を下図右側上部にある Processing Type の POS(文節単位の属性の n-gram)、Word(文節の n-gram)、WordPOS(属性付きの文節の n-gram)から指定する。そして、Ngram Type 下の窓で n を選択する。N-gram とは、集計する際の文節区切りの数である。Unigram($n=1$)の場合文節を一つずつ集計し、Bigram($n=2$)の場合は二つの文節の組み合わせを一つとして集計する。選択肢は、Unigram($n=1$)、Bigram($n=2$)、Trigram($n=3$)、Fourgram($n=4$)、Fivegram($n=5$)、Sixgram($n=6$)の六つがある。また、Cutoff を用いて集計サイズをコントロールすることができる。デフォルトは 100 となっており、100 以上の頻度のものを表示する。それ以下のものに関しては others して一つの項目にまとめられる。データの形式と保存方法は Length と同様である。All POS Processing をクリックすると集計結果が左側の Results 窓に返される。下側の左図は POS(文節単位の属性の n-gram)を集計した画面である。



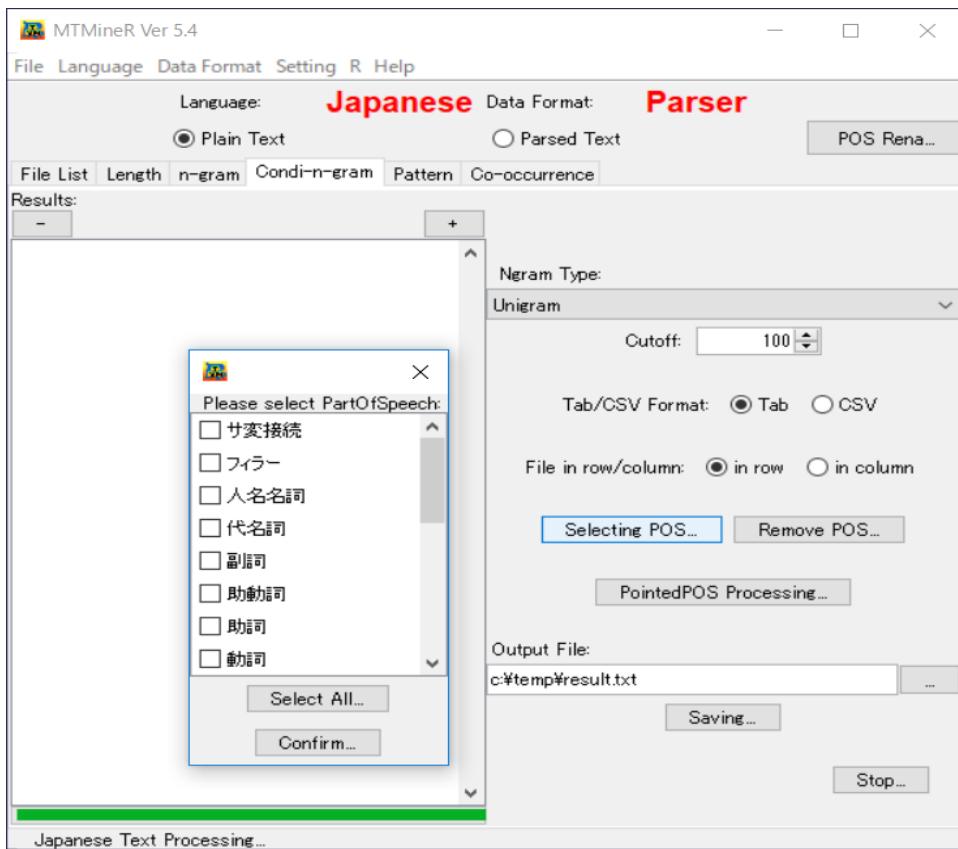
指定した属性のみを含んだ文節の n-gram を集計する際は、All POS Processing ではなく、Selecting POS をクリックする。下図のように指定したい属性の種類にチェックを入れ（下図では代名詞と副詞にチェックをいれている）Confirm をクリックした後に Pointed Tag Processing を選択すると集計結果が左側の窓に返される。



4. Condi-n-gram (条件付きの文節の n-gram)

タブ Condi-n-gram では、条件付きの文節の n-gram が集計できる。これは、指定した属性の n-gram の中から一部の属性データを除外したデータを集計する。下図は名詞を含んだ文節の中から、記号を取り除いたものである。

まず、Selecting POS をクリックし、指定したい属性を選択し、Confirm をクリックする。次に、Remove POS をクリックし、取り除きたい属性を選択し、Confirm をクリックする。最後に PointedPOS Processing をクリックすると、左窓に結果が表示される。結果の表示形式と保存は Length と同様である。また、デフォルトは、Cutoff が 100 となっており、100 以上の頻度のものしか表示されない。100 文字以下のものはすべて others にまとめられる。



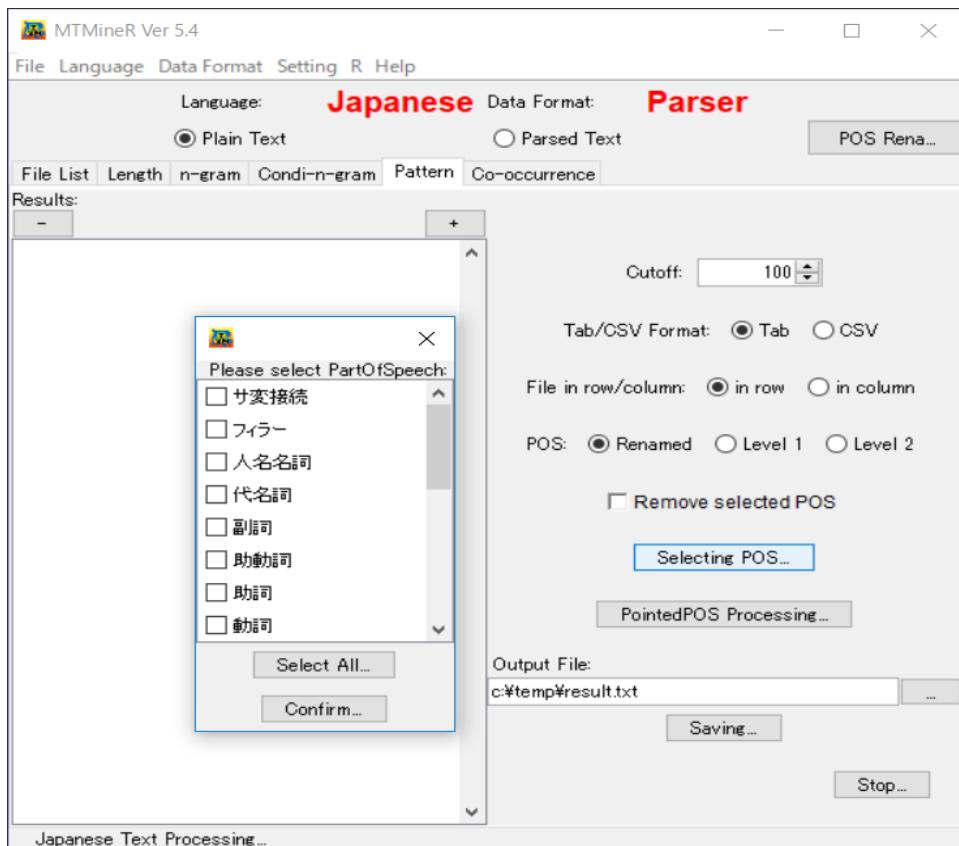
この機能は複合語を含む形態素より長い単位の語句の集計が可能である。例えば、mecab で「自然言語処理を行う」を形態素解析すると「自然」「言語」「処理」「を」「行う」に分解され、名詞を集計すると「自然」「言語」「処理」が集計される。一方 cabocha で文節分解を行うと「自然言語処理を」「行う」に分解される。名詞を含む文節の中から助詞を取り除くことにより「自然言語処理」が一つの項目として集計することができる。

5. Pattern (文節のパターン)

タブ「Pattern」では文節パターンを集計することができる。下図は助詞と記号は原型、それ以外は形態素を用いた文節パターンの集計である。

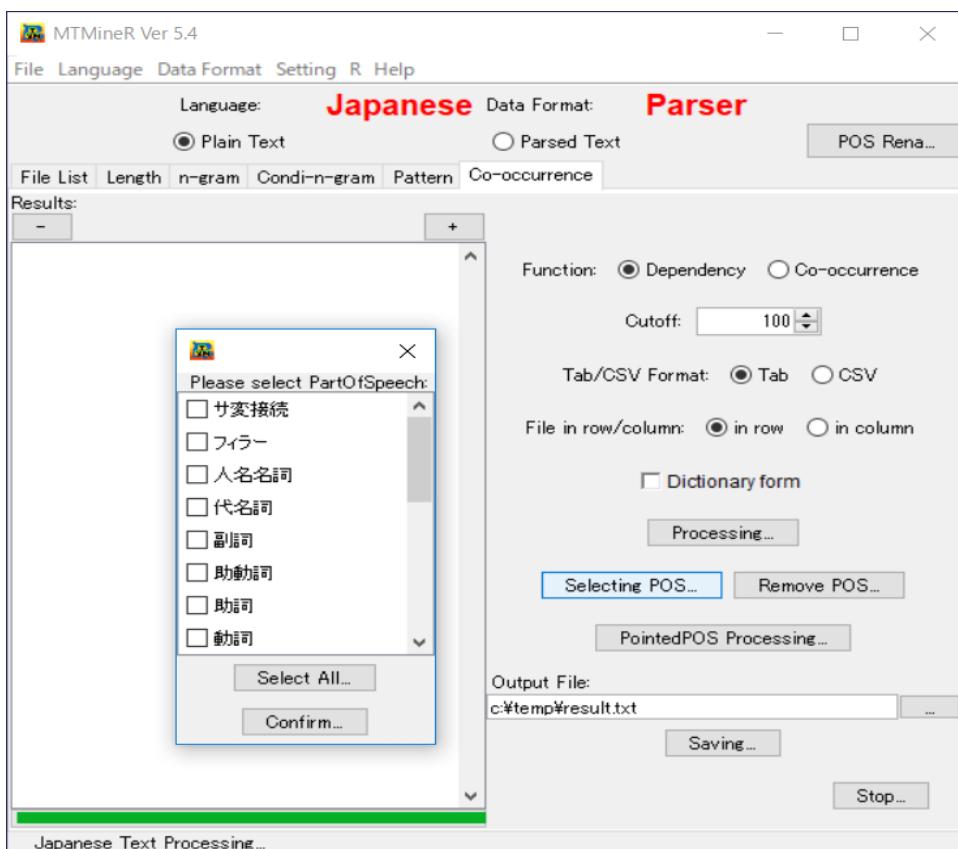
まず、Renamed を用いるか POS で形態素解析した結果の何層を用いるかを指定できる。Renamed を選択した場合、自分で自由に設定した赤字の部分の形態素の名前を用いることが可能である。Level1 は第1層、Level2 は第2層の情報を用いることができる。次に、Selecting POS をクリックし、品詞を選択すると、選択された品詞・タグのみが原型(助詞の場合であると、は、がのような単語)で集計される。また、Remove selected POS にチェックを入れると、Selecting POS で指定した品詞を除いた条件付のパターンを集計する

ことが可能である。



6 . Co-occurrence (文節の共起)

タブ Co-occurrence では係り受け先を考慮した共起と係り受け関係を無視した共起を集計することができる。下図の右側にある Function: の Dependency を選択すれば係り受け先を考慮した共起、Co-occurrence を選択すれば係り受け関係を無視した共起を集計することができる。これを選択した後に Processing をクリックすると集計結果が左窓に出力される。また、Condi-n-gram と同様に Selecting POS で指定した品詞を除く条件付のパターンを集計することが可能である。また、「Show Selected POS」で指定した品詞を表示できる。

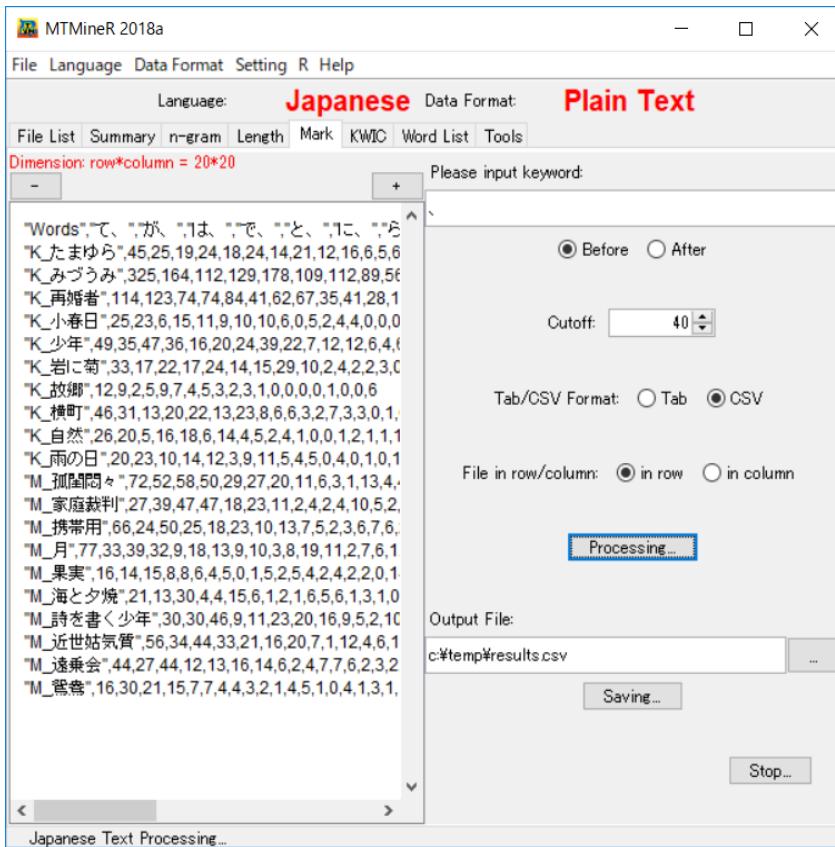


分析機能

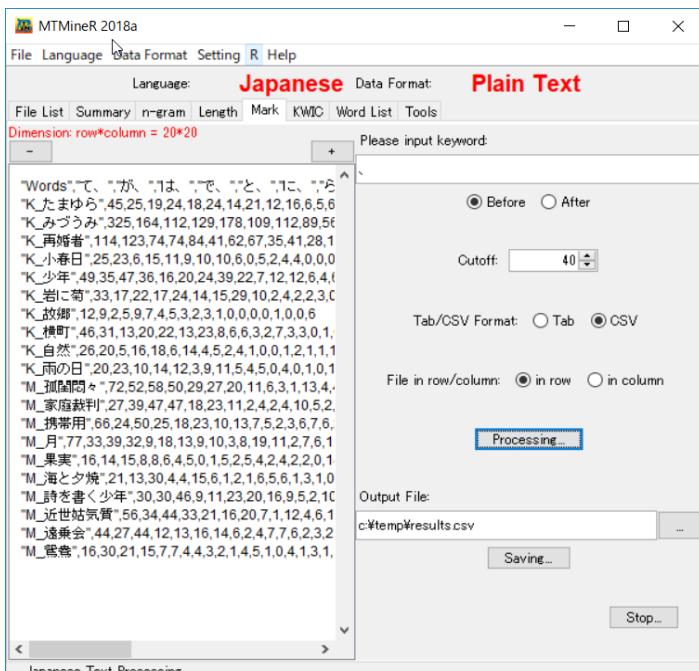
データの変換

データ解析 GUI の起動

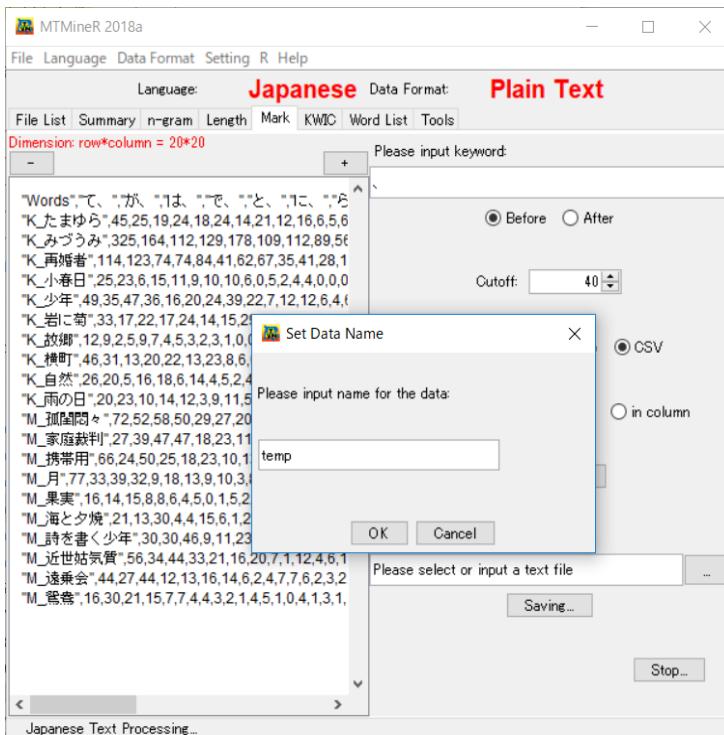
MTMineR に同梱されている sample フォルダの中の Japanese の中に用意されている川端康成と三島由紀夫の作品を読み込み、読点がどの文字の後に打たれているかに関するデータを用いて説明する。まず、次の GUI 画面の設定通りにデータを集計する。



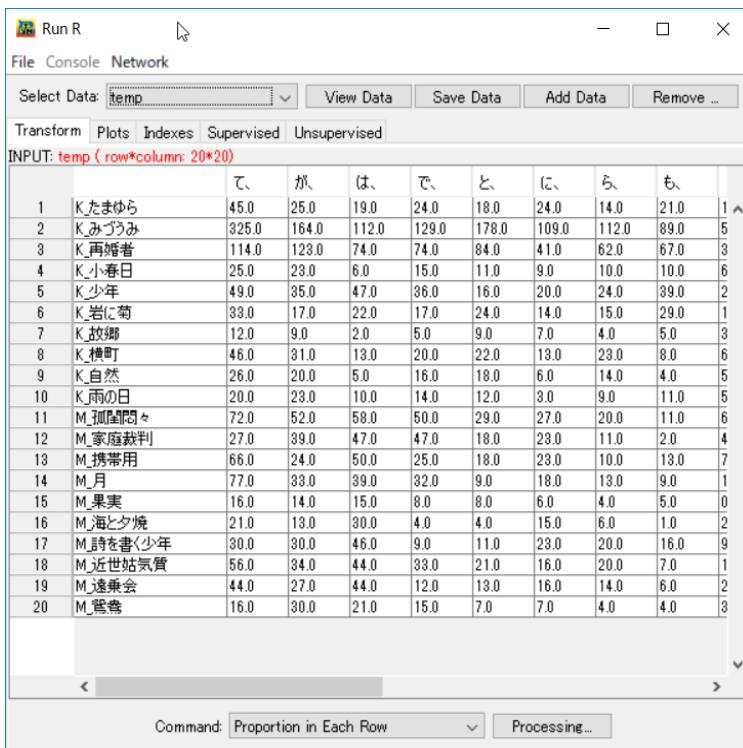
収集したデータを分析するため、メニュー[R]の中のもっともトップの項目 [ProcessingOutputsInthisTab]を選択し、クリックするとデータセットの名前を指定する窓が現れる。



現れた窓に temp が入力されている。これは仮のデータ名前である。temp を削除し、各自好みの名前を指定することができる。



名前を指定し[OK]ボタンを押すと次のような三つの画面が同時に現れる。



左上は R のコンソールのような役割を果たす CUI 環境である。左下は、データ解析を行う際の結果出力画面である。右の画面はデータをメニュー操作で分析を行う RunR という GUI である。メニューに用意されているデータ処理・分析は、このメニュー操作で行うことが出来る。データ処理タブは 5 つ(Transform、Plots、Indexes、Supervised、Unsupervised)が用意されている。

データ変換

タブ Transform には、行の総和を基準とした相対頻度(Proportion in Each Row)、列の総和を基準とした相対頻度(Proportion in Each Column)、列を基準とした標準化(Normlization)、データ行列の転置(Transpose)の機能が用意されている。これらの変換する項目は、画面の下部の Command の右にリストアップされている。下向きの▼を押して選択する。

	て、	が、	は、	で、	と、	に、	ら、	も、	
1	K_たまゆら	45.0	25.0	19.0	24.0	18.0	24.0	14.0	21.0
2	K_みづうみ	325.0	164.0	112.0	129.0	178.0	109.0	112.0	89.0
3	K_再婚者	114.0	123.0	74.0	74.0	84.0	41.0	62.0	67.0
4	K_小春日	25.0	23.0	6.0	15.0	11.0	9.0	10.0	10.0
5	K_少年	49.0	35.0	47.0	36.0	16.0	20.0	24.0	39.0
6	K_岩に菊	33.0	17.0	22.0	17.0	24.0	14.0	15.0	29.0
7	K_放郷	12.0	9.0	2.0	5.0	9.0	7.0	4.0	5.0
8	K_横町	46.0	31.0	13.0	20.0	22.0	13.0	23.0	8.0
9	K_自然	26.0	20.0	5.0	16.0	18.0	6.0	14.0	4.0
10	K_雨の日	20.0	23.0	10.0	14.0	12.0	3.0	9.0	11.0
11	M_孤闇問々	72.0	52.0	58.0	50.0	29.0	27.0	20.0	11.0
12	M_家庭裁判	27.0	39.0	47.0	47.0	18.0	23.0	11.0	2.0
13	M_携帯用	66.0	24.0	50.0	25.0	18.0	23.0	10.0	13.0
14	M_月	77.0	33.0	39.0	32.0	9.0	18.0	13.0	9.0
15	M_果実	16.0	14.0	15.0	8.0	8.0	6.0	4.0	5.0
16	M_海と夕焼	21.0	13.0	30.0	4.0	4.0	15.0	6.0	1.0
17	M_詩を書く少年	30.0	30.0	46.0	9.0	11.0	23.0	20.0	16.0
18	M_近世姑氣質	56.0	34.0	44.0	33.0	21.0	16.0	20.0	7.0
19	M_達兼会	44.0	27.0	44.0	12.0	13.0	16.0	14.0	6.0
20	M_鶯聲	16.0	30.0	21.0	15.0	7.0	7.0	4.0	4.0

Command: Proportion in Each Row Processing...

- Proportion in Each Row
- Proportion in Each Column
- Normalization
- Transpose

ProportioninEachCoulum を選択し、ボタン[Processing]を押すと変換されたデータの名前を知らすメッセージボックスが開かれる。データの名前は元の名前に [.tra] が追加されていることに注意してほしい。

		て、	が、	は、	で、	と、	に、	ら、	も、	
1	K_たまゆら	17.37	9.65	7.34	9.27	6.95	9.27	5.41	8.11	4 ^
2	K_みづうみ	20.43	10.31	7.04	8.11	11.19	6.85	7.04	5.59	3
3	K_再婚者	12.46	13.44	8.09	8.09	9.18	4.48	6.78	7.32	3
4	K_小春日	18.66	17.16	4.48	11.19	8.21	6.72	7.46	7.46	4
5	K_少年	11.53	8.24	11.06	8.47	8.76	4.71	5.65	9.18	5
6	K_岩に菊	13.92	7.17	9.28	7.17	10.13	5.91	6.33	12.24	4
7	K_故郷	17.39	13.04	2.9	7.25	13.04	10.14	5.8	7.25	4
8	K_横町	21.6	14.55	6.1	9.39	10.33	6.1	10.8	3.76	2
9	K_自然	20.0	15.38	3.85	12.31	13.85	4.62	10.77	3.08	3
10	K_雨の日	15.15	17.42	7.58	10.61	9.09	2.27	6.82	8.83	3
11	M_孤高閣々	17.39	12.56	14.01	12.08	7.0	6.52	4.83	2.66	1
12	M_家庭裁判	9.06	13.09	15.77	15.77	6.04	7.72	3.69	0.67	1
13	M_携帯用	20.5	7.45	15.53	7.76	5.59	7.14	3.11	4.04	2
14	M_月	21.04	9.02	10.66	8.74	2.46	4.92	3.55	2.46	2
15	M_果実	13.68	11.97	12.82	6.84	6.84	5.13	3.42	4.27	0
16	M_海ヒタ焼	14.69	9.09	20.98	2.8	2.8	10.49	4.2	0.7	1
17	M_詩を書く少年	10.45	10.45	16.03	3.14	3.83	8.01	6.97	5.57	3
18	M_近世姑氣質	18.01	10.93	14.15	10.61	6.75	5.14	6.43	2.25	0
19	M_遠乗会	16.48	10.11	16.48	4.49	4.87	5.99	5.24	2.25	0
20	M_鶴巣	11.43	21.43	15.0	10.71	5.0	5.0	2.86	2.86	2

変換されたデータの名前は、元のデータの名前に .tra が付けられる。メッセージボックス上の [OK] を押すとメッセージボックスが消される。変換されたデータを用いて分析を行うためには RunR 上の Select data から変換したデータを読み込む必要である。

グラフ

タブ Plots には、4 種類のグラフ作成機能を実装している。ここでは、様々な場面で多用される WordCloud を例として説明する。WordCloud はテキストデータを頻度の高い単語ほど大きな文字で表示するグラフのことである。この図は川端康成と三島由紀夫の作品における形態素の wordcloud を示した。この図では、川端康成は「の」を多く用いるのに対し、三島由紀夫は「は」を多用しているのが見て取れる。

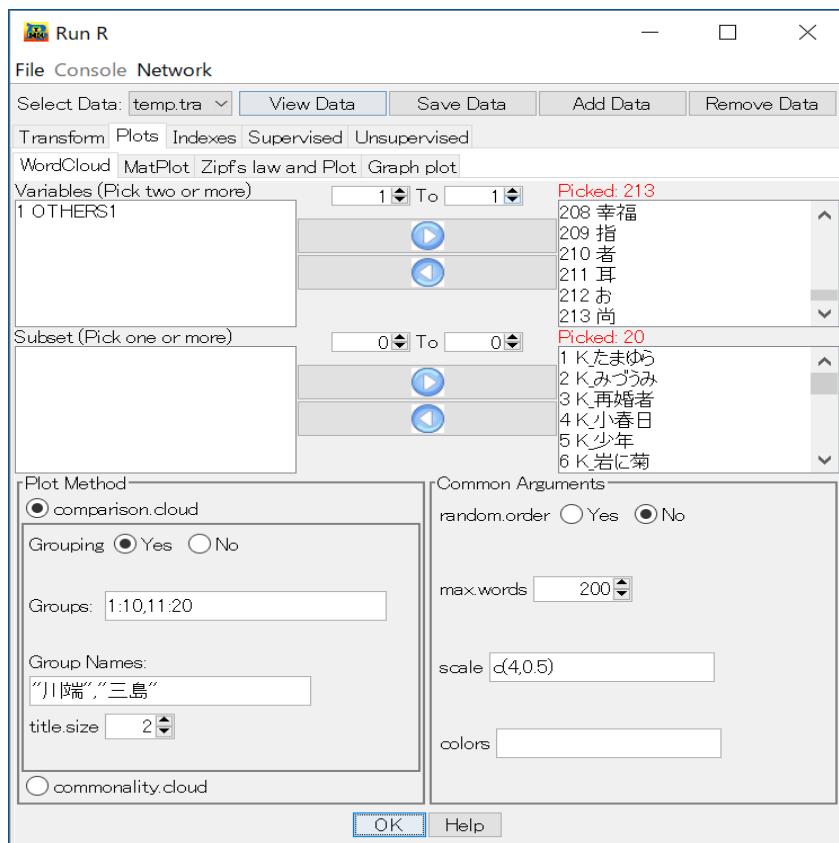
● WordCloud：文章の単語出現頻度を可視化する画像

● MatPlot : 多群の折れ線グラフ

● Zipf's low and Plot : Zipf 法則のグラフと対数回帰モデルの当てはめるグラフ

● Graph Plot : 語や文節のネットワークグラフを作成する

次に示す分析したい変数やテキストを右側のそれぞれの窓に読み込み、分析したい項目をチェックし、画面上の[OK]ボタンを押すと、処理結果が返される。



詳細の結果は、OUTPUT 画面に返される。



作成されたグラフはコピーあるいは保存して使用することができる。グラフ画面上にマウスポインターを合わせ、右ボタンを押すとコピーおよび画像の保存形式 (JPEG,BMP , PNG, EPS)のリストが表示される。Rconsole 上で R コマンドラインを作成し、[Enter]キーを押すとコマンドラインが実行される。Rconsole は簡易の R 言語の操作環境である。

特徴語抽出とグラフ作成

特徴語の抽出は複数のテキストの間、複数のテキストのグループ間に特徴となる要素の候補を検出する。複数のテキストのグループ間の特徴語抽出を行うためには、テキストをグルーピングすることが必要である。グルーピングの指定は、Please enter groups separating by commas の下の窓に記述する。連番の整数は a:b のように、始まる整数と終わる整数を半角のコロン記号「:」でつなげる。連番ではない場合は、c()の中に番号をカンマで区切り記入する。例えば、テキスト 1,3,5 が一つグループ、2,4,6 が一つグループの場合は、c(1,3,5), c(2,4,6)のように記入する。テキストの番号を用いてグルーピングを指定したら、それに対応するグループの名前を記述する。グルーピングの指定は、Please enter groups separating by commas の下の窓に記述する。連番の整数は a:b のように、始まる整数と

終わる整数を半角のコロン記号「:」でつなげる。連番ではない場合は、c()の中に番号をカンマで区切り記入する。テキストの番号を用いてグルーピングを指定したら、それに対応するグループの名前を記述する。

特徴語を抽出する方法は現段階では以下の方法が実装されている。それぞれの方法の詳細に関しては画面上の Help に説明されている。

● Chi-square(カイ二乗統計量)

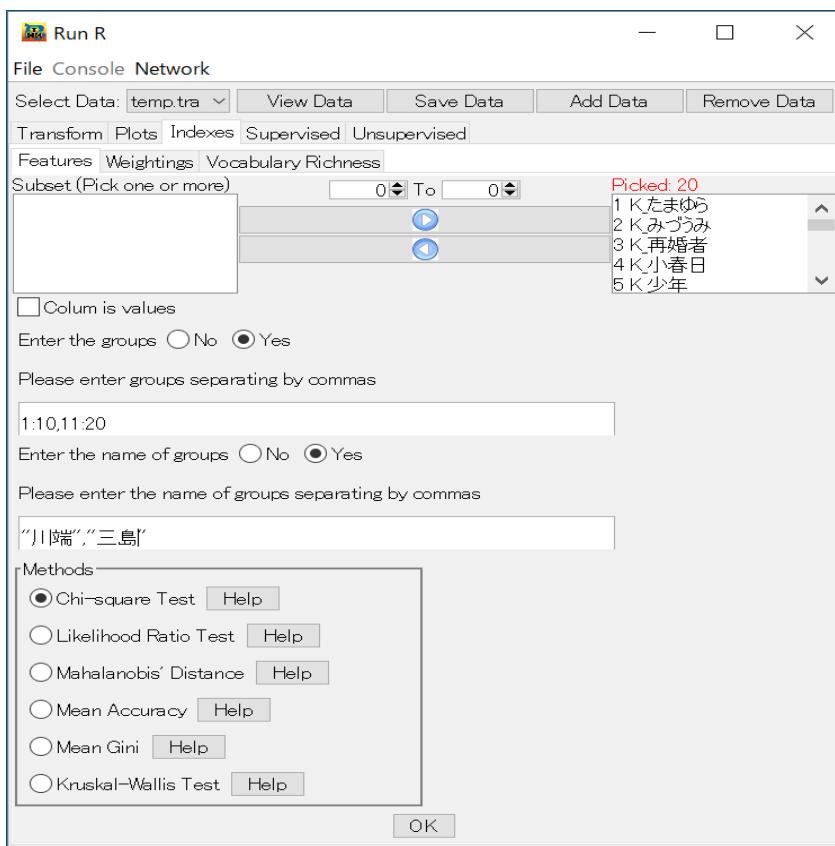
● Likelihood Ratio(尤度検定統計量)

● Mean Accuracy(RandomForest の中の MeanDecreaseAccuracy)

● Mean Gini(RandomForest の中の MeanDecreaseGini)

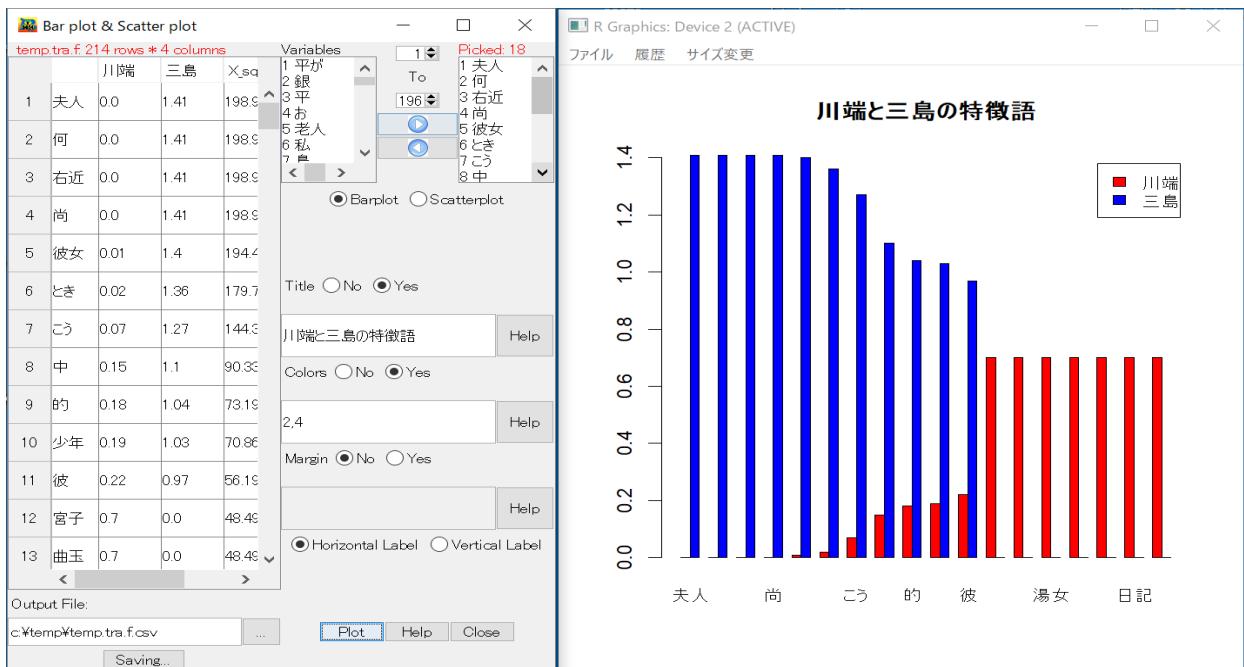
● Mahalanobis' Distance(マハラノビス距離)

● Kruskal-Wallis(クラスカル・ウォリス検定統計量)



上のグラフのようにラジオボタンを指定し、[OK]ボタンを押すと選択された方法で計算する。計算が終わると結果を保存するファイルの名前を知らせると同時にグラフの作成に関するメッセージボックスが開かれる。グラフ作成を選択すると下のような画面が開かれる。 グラフのタイトル、色、グラフの周辺のマージン、ラベルの文字列の方向を指定し

[OK]ボタンを押すと棒グラフが作成される。グラフ作成画面の Scatterplot を指定すると選択された変数の散布図、あるいは対散布図が作成される。



注： グラフのラベルを縦にし、全部表示できるような画面に入れ替えてください。

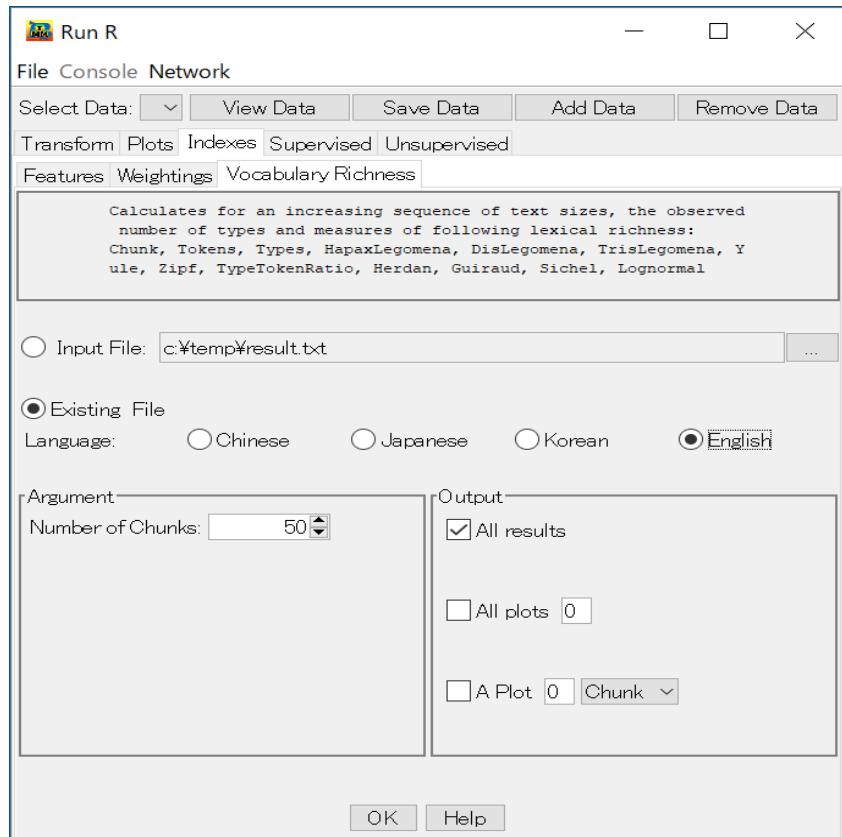
サブタブ Weightings では、エントロピー(entropy)や TF-IDF などの 10 種類の重みを計算する。

サブタブ Vocabulary Richness では、1 つのテキストの語彙の豊富さを計算する。日本語においてはタグ記号 <> により区切られているテキストを用いる。1 つのタグ付きテキストを読み込み、テキストを複数のチャンク (chunks) に区切り、チャンクを累積しながら、述べ語数、異なり語数、トーケン比(TTR)、Yule の K 特性値、Sichel の S 値などの 12 種類の語彙の豊富さの指標に関する値を計算する。

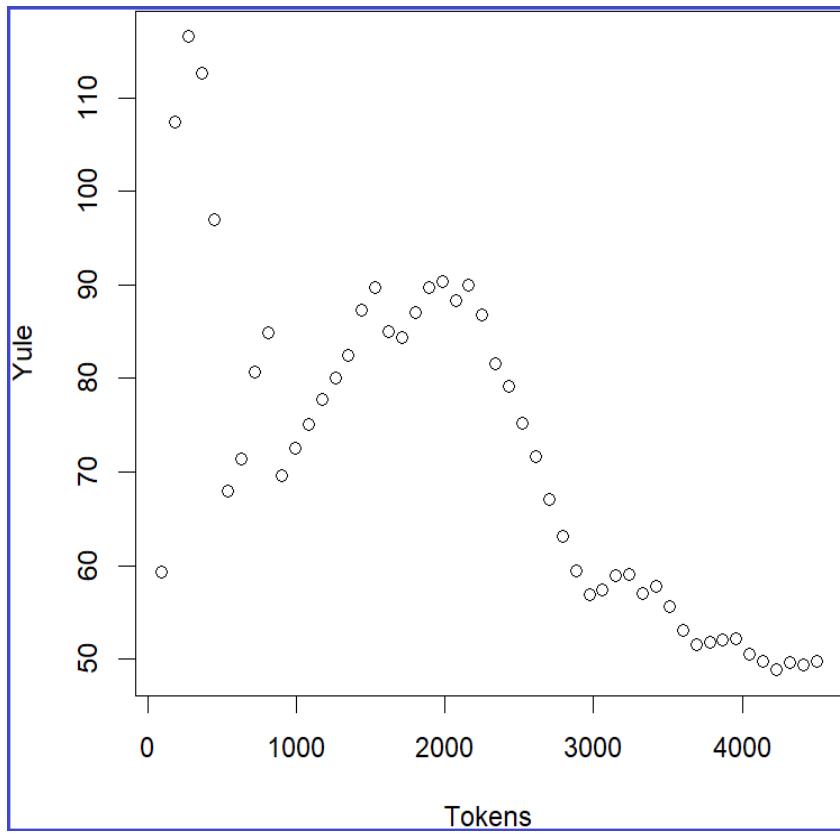
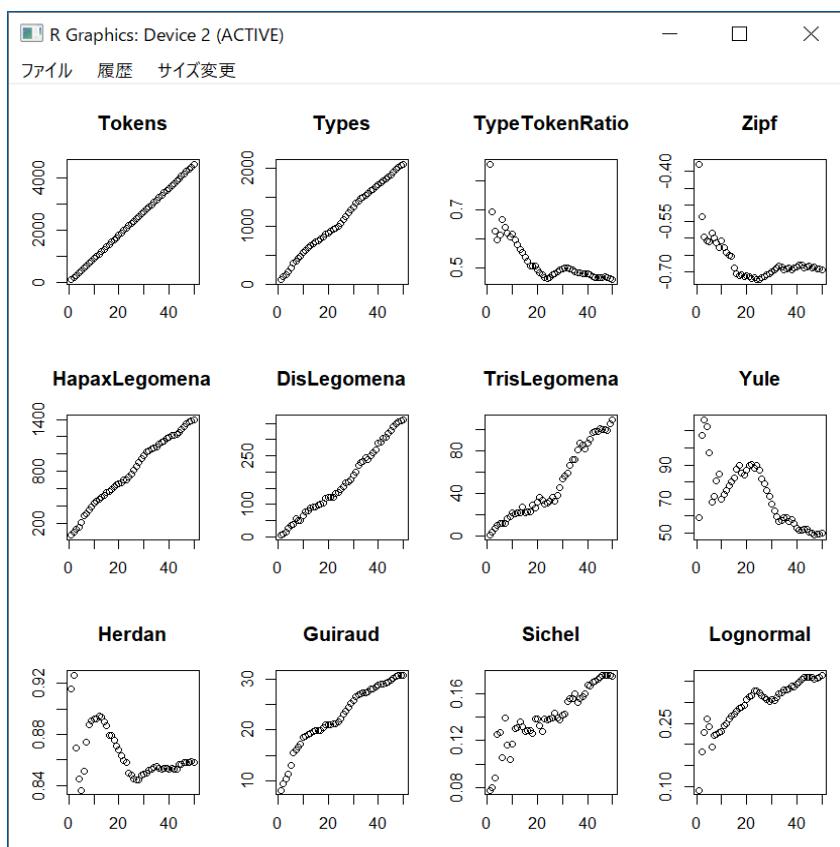
語彙の豊富さ

サブタブ Vocabulary Richness では、テキストの語彙の豊富さを計算する。日本語においてはタグ記号 <> により区切られているテキストを用いる。タグ付きテキストを読み込み、テキストを複数のチャンク (chunks) に区切り、チャンクを累積しながら、述べ語数、

異なり語数、トークン比(TTR)、Yule の K 特性値、Sichel の S 値などの 12 種類の語彙の豊富さの指標に関する値を計算する。二つの選択「Input File」と「Existing File」がある。「Input File」で一つのテキストを読み込むことでその語彙の豊富さを求める。「Existing File」は MTMineR に読み込んだすべてのテキストの語彙の豊富さを求める。デフォルトは「Existing File」である。



複数のテキストの語彙の豊富さを求める時、「All plot」で何番目のテキストの 12 種類の語彙の豊富さを表す指標のプロットを指定できる。更に、「A plot」で何番目のテキストに対して、各語彙の豊富さ指標を考察できる。



教師なし学習 (Unsupervised learning)

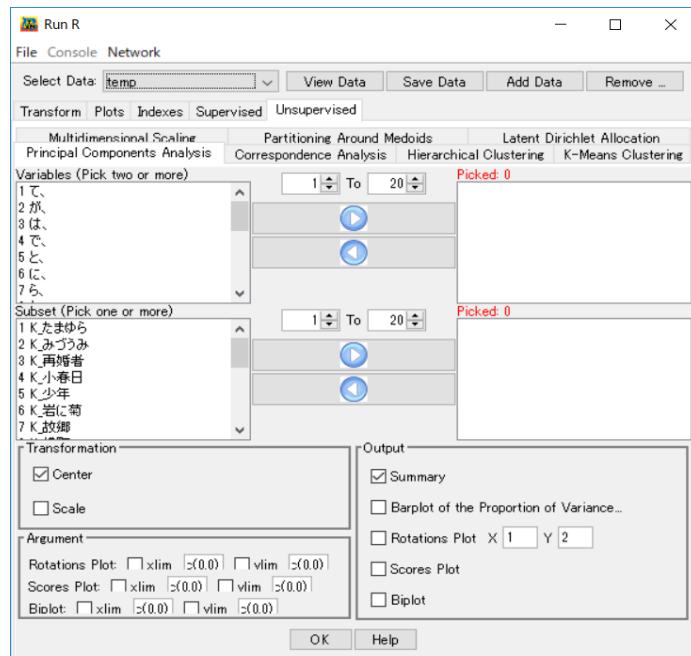
教師なし学習は、教師ラベルが与えられていないデータの分析方法である。

Unsupervised タブでは次に示した主な教師なしの方法を実装している。ここでは、川端康成と三島由紀夫の小説を例として各手法を説明する。

- [主成分分析\(Principle Components Analysis\)](#)
- [対応分析\(Correspondence Analysis\)](#)
- [階層的クラスター分析\(Hierarchical Cluster Analysis\)](#)
- [K-Means 法\(K-Means Clustering\)](#)
- [多次元尺度法\(Multidimensional Scaling\)](#)
- [PAM 法\(Partition Around Medoids\)](#)
- [LDA 法\(Latent Dirichlet Allocation\)](#)

主成分分析 (Principle Components Analysis)

主成分分析は、情報の損失をできる限りすくないように高次元データを低次元に圧縮する教師なしの方法である。川端康成と三島由紀夫の作品を読み込み、文字 bigram を集計し、主成分分析のタグの画面コピーを次に示す。



画面上の左側には、変数のリストとテキストのリストが表示されている。主成分分析を行うためにはまず用いる変数と用いるテキストを選択し、右側の Picked の窓に取り入れることが必要である。

1. 変数と個体の取り入れ

一つひとつ選択し矢印で取り入れ、または取り除くことが出来る。すべてを取り入れる場合はマウスポインターを変数あるいはテキスト窓に合わせ、キーボードの Ctrl+A キーを同時に押すとすべて選択され、矢印ボタンで一括取り入れ、取り除くことが出来る。また変数の数を連番で指定し取り込むことも可能である。

2. 分析のオプションを指定する

MTMineR には、分散共分散行列と相関係数行列を用いる方法が実装されている。デフォルトでは分散共分散行列を用いた主成分分析を行うになる。GUI 上のオプション[scale]にチェックを入れると相関係数行列を用いた主成分分析を行う。[center]は主成分得点をセンターリングするオプションである。一般的にはデフォルトの設定のとおりにセンターリンを行う。

3. 出力のオプションを指定する

GUI の右下側に出力のオプションがある。

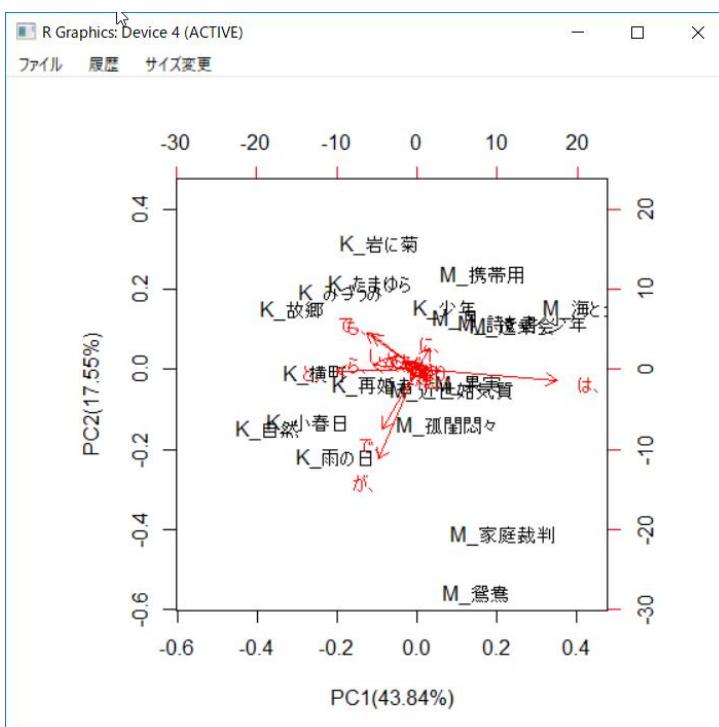
- Summary にはチェックボックスにチェックがついている。この環境で実行すると主成分分析要約（標準偏差、寄与率、累積寄与率）が OUTPUT 画面に出力される。
- Barplot of the Proportion of Variance を選択すると寄与率と累積寄与率の棒グラフが一つグラフ画面として出力される。
- Rotation Plot を選択すると固有ベクトル・主成分の散布図を作成する。X の窓には横軸に用いる主成分の番号、Y には縦軸に用いる主成分番号を指定する。

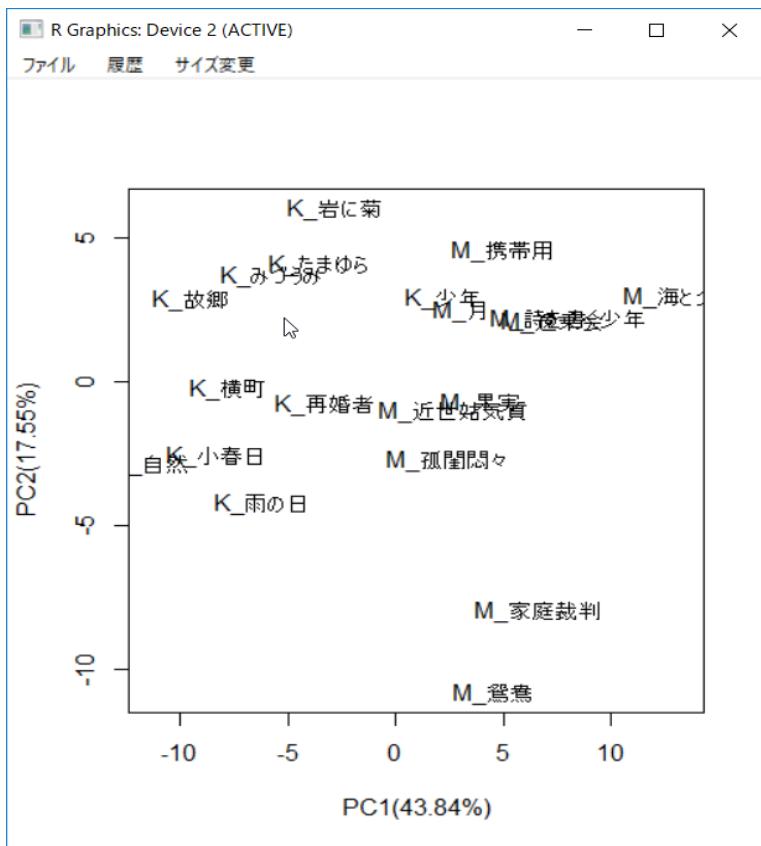
- Scores Plot を選択すると主成分の得点の散布図を作成する。
横軸と縦軸は Rotation Plot と一致する。
- Biplot を選択すると主成分のバイプロットを作成する。

4. 出力グラフの調整オプション

出力オプションの左側には出力するグラフの微調整オプションが用意されている。出力されている主成分分析のグラフは、文字列が枠組み線に切られたり、余白が大きいすぎたりする場合がある。その際には、軸の範囲を広めたり、狭めたりすることでグラフの完成度を上げることが出来る。初心者は無視してよい。

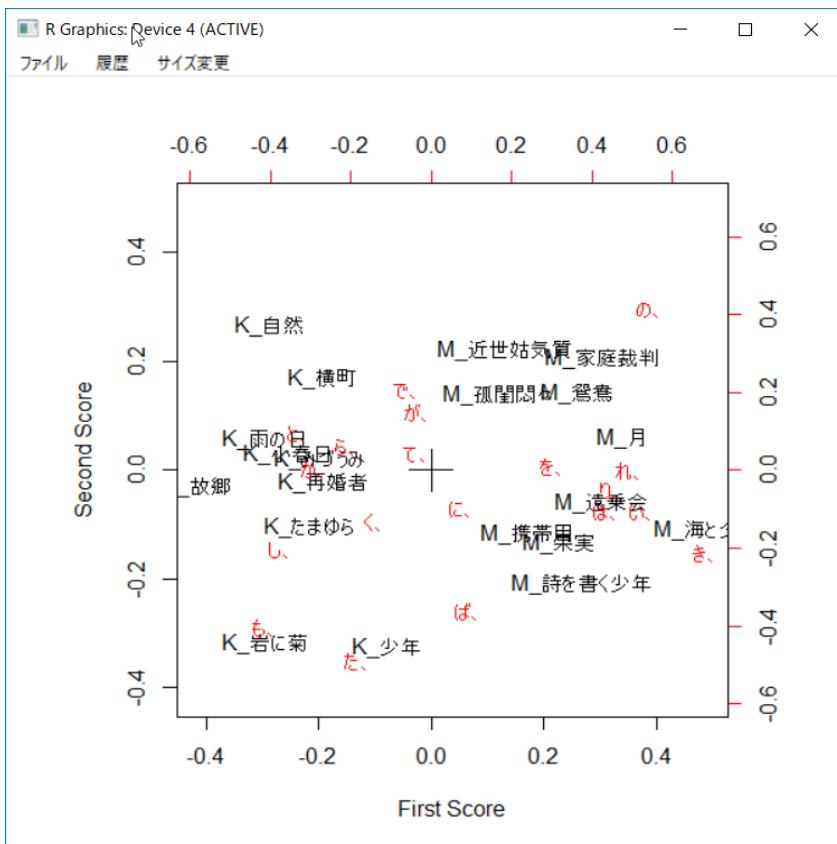
川端康成と三島由紀夫のそれぞれ 10 編の小説に読点がどの文字の後に打っているかに関するデータを主成分分析 GUI に読み込み、GUI 上の[OK]ボタンを押すと主成分分析の要約は OUTPUT 画面に出力され、第 1 と第 2 の主成分に関する二つの散布図と一つのバイプロットが作成される。変数が少ないとときには、主成分分析では個体と変数の関係をバイプロットで考察する。出力されているバイプロットに示したように、川端康成の文章は主に左側に集まり、「と、」、「し、」と「も、」の使用率が相対的に高い。テキスト分析の特徴量の次元数がは数百ないし数千に上るときには、バイプロットは大量の変数に覆いつぶされて考察できない。この場合 score plot で個体の位置関係だけを見る。





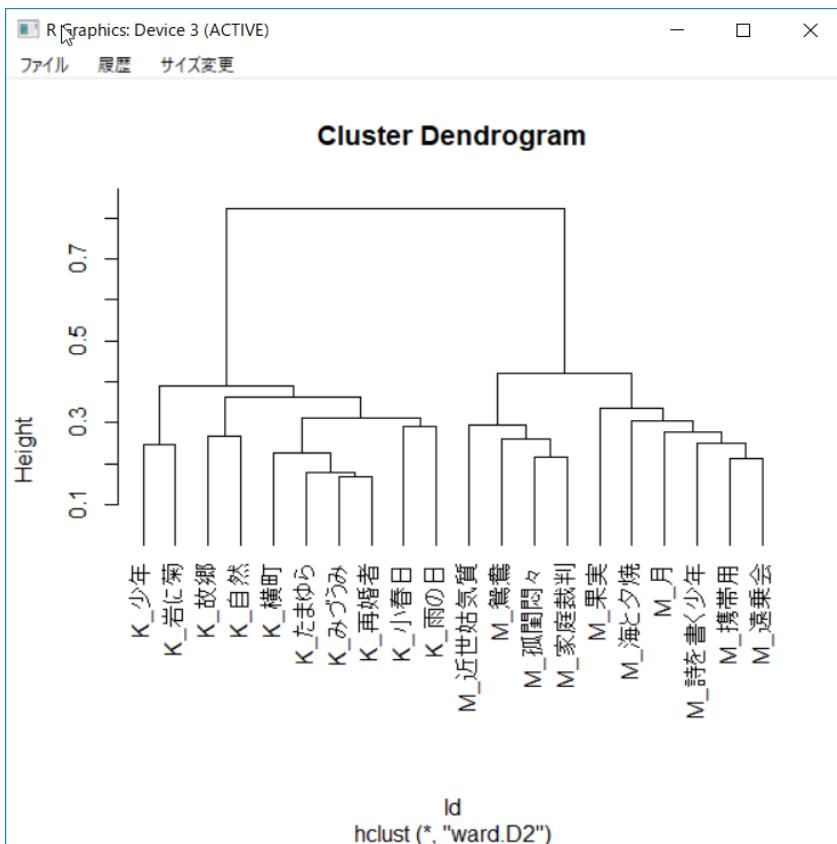
対応分析 (Correspondence Analysis)

対応分析は、高次元データを低次元に圧縮する考え方では主成分分析と同じである。対応分析は、分割表において行の項目と列の項目の相関が最大になるように、行と列の双方を並び替えている。個体を考察する場合 Rscore Plot、変数を考察する場合 Cscore Plot にチェックを入れる必要がある。



階層的クラスター分析 (Hierarchical Cluster Analysis)

データ分類の手法としてクラスター分析がよく用いられる。クラスター分析は、階層的クラスター分析と非階層的クラスター分析に大別される。階層的クラスター分析は、最も似ている組み合わせから順番にクラスターにしていき、最終的に樹形図（デンドログラム）でデータ間のグルーピングを表す手法である。この手法には(1)元のデータ行列から距離行列を求める(2)クラスターの結合方法を選ぶ(3)コーフエン行列を求める(4)樹形図作成の4つの手順が必要である。距離と結合方法は、それぞれ Distance Measures と Methods タブから選択できる。ここでは、クラスター内の分散が最小になる Ward 法と SKLD 距離を用いた分析結果を示す。川端康成と三島由紀夫の作品は大きく2つのクラスターに分かれていることが樹形図から見て取れる。



K-mean 法(K-Means Clustering)

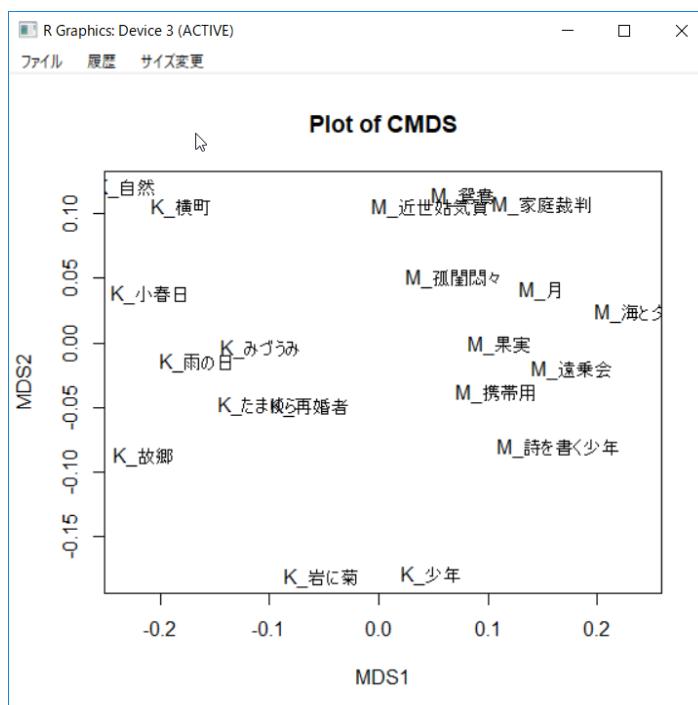
非階層的クラスター分析の代表的な手法は k-means 法である。k-means 法のアルゴリズムは次の通りである。(1)初期クラスターの中心を求める(2)すべてのデータと中心との距離を求め、データを距離の最も近いクラスターに分類する。(3)新しいクラスターの中心を求める。(4)(2)と(3)を繰り返し、クラスターの中心が前と全く同じか指定回になると終了する。MTMineR で、指定クラスター数の Num. of Cluster を 2 にし、指定最大計算回数 Max. Num. of iterations を 10 にする場合の結果を次に示す。各小説名の下にある数字の 1 と 2 はクラスマラベルで、1 は三島由紀夫、2 は川端康成をそれぞれ表す。川端康成の『少年』と『岩に菊』は間違って三島由紀夫を表す「1」に判別されたことが見て取れる。

OUTPUT				
Clustering vector:				
K_たまゆら	K_みづうみ	K_再婚者	K_小春日	K_少年
2	2	2	2	1
K_岩に菊	K_故郷	K_横町	K_自然	K_雨の日
2	2	2	2	2
M_孤闇問々	M_家庭裁判	M_携帯用	M_月	M_果実
1	1	1	1	1
M_海と夕焼	M_詩を書く少年	M_近世姑氣質	M_遠乗会	M_鶯聲
1	1	1	1	1

Output File: c:\temp\result.txt

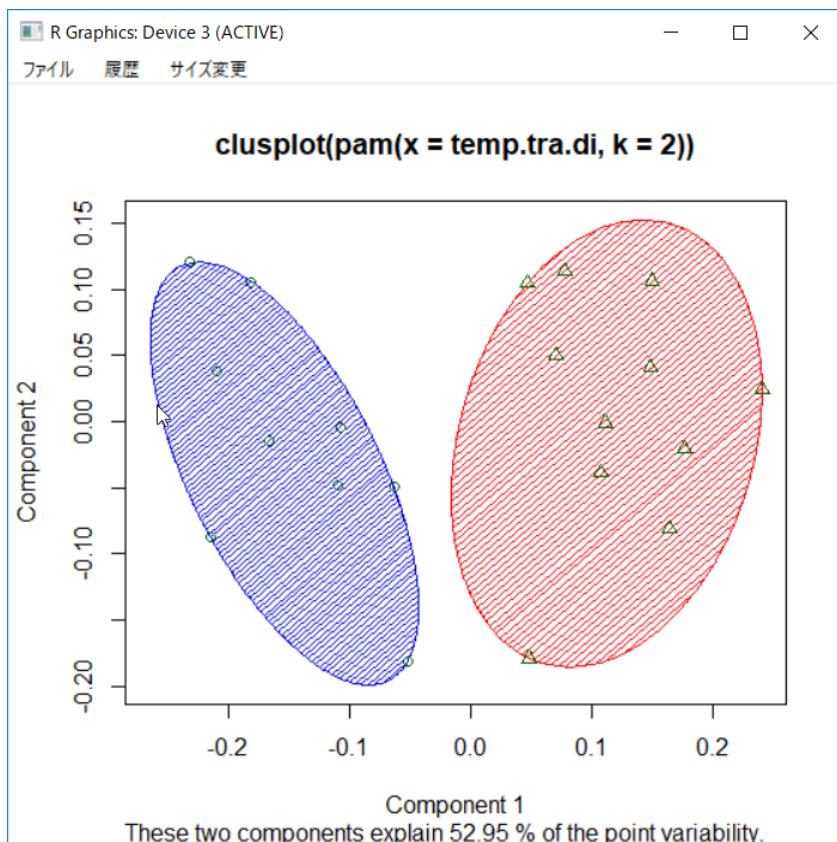
多次元尺度法 (Multidimensional Scaling)

多次元尺度法は個体間の親近性データを低次元（2 或いは 3 次元）空間に配置する方法である。この方法は、計量的多次元尺度法と非計量多次元尺度法）に大別される。計量的多次元尺度法は個体間の距離データに基づいたもので、MTMineR では Method タブの cmdscale で対応している。非計量的多次元尺度法は個体間の類似度や相関係数行列に基づいたもので、MTMineR では Method タブの isoMDS、sammon と metaMDS で対応している。cmdscale を用いた川端康成と三島由紀夫の分析結果では、両作家の作品は異なるグループを形成していることが見て取れる。



PAM 法 (Partition Around Medoids)

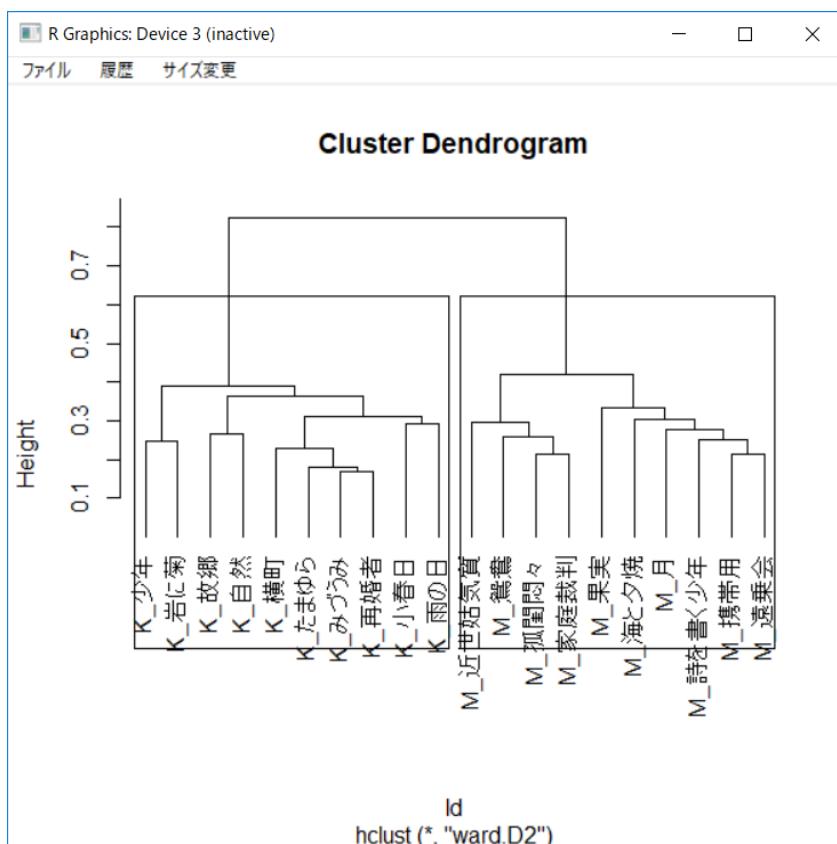
PAM 法は k-means 法に似ているが、クラスターの中心ではなく medoid で代表する点で異なる。Medoid 以外の対象はそれと最も近い medoid が代表するクラスターに分類される。その大まかなアルゴリズムは、(1)k 個の medoid を選択する(2)すべてのデータと中心との距離を求め、データを距離の最も近い medoid が代表するクラスターに分類する。(3)全てのデータに対して medoid の再計算を行う(4)(2)と(3)を繰り返し、最適な分類に近づいていく。次のグラフに示したように、川端康成の作品（○）と三島由紀夫の作品（△）は大まかに 2 群に分かれている。



LDA (Latent Dirichlet Allocation)

LDA はトピックモデルの一種 latent dirichlet allocation の略語、階層的ベイズモデルのアプローチで pLSA を拡張したモデルである。MTMineR に実装された LDA では、文章

の属するトピックの推定し、そのトピックごとのクラスタリングも行う。川端康成と三島由紀夫の作品における LDA の実行結果を次の図に示す。作品は 2 群に分かれていることが見て取れる。



教師あり分析方法

タブ Supervised には教師ありの方法を実装している。

- [CART](#)
- [C50](#)
- [k-Nearest Neighbour](#)
- [RandomForest](#)
- [SVM\(support vector machine\)](#)
- [LDA\(supervised LDA\)](#)
- [HDDA\(High-Dimensional Discriminant Analysis\)](#)

RandomForest に自作の機能があるので、ここで、RandomForest を例として教師あり方法の操作を説明する。

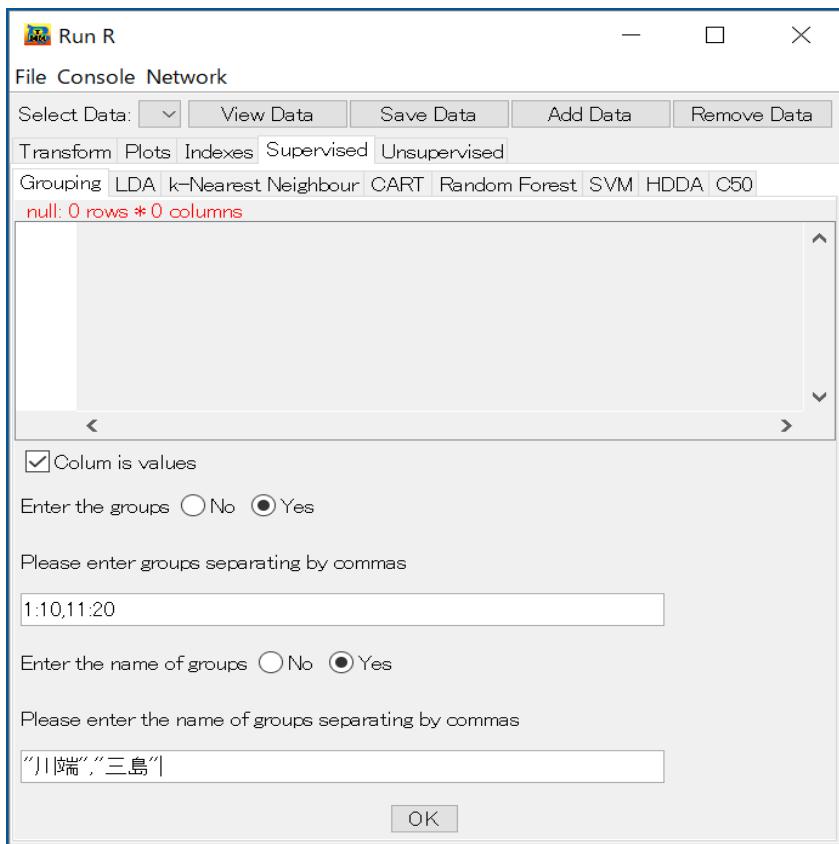
MTMineR の中のすべての関数表示は各自の R パッケージと同じである。関数の意味を各パッケージのサイトに確認してください。

これらの方法を用いるためには、まず教師となる外的基準を指定しなければならない。外的基準の指定はタブ Grouping で行う。

- [Grouping](#)
- [RandomForest](#)

Grouping

- ①各グループの範囲を指定する。
- ②各グループにラベルを付ける。
- ③OK を押すと、ラベル付きデータセットを作成する。作成したデータセットを Select Data の中に指定する。



Random Forest

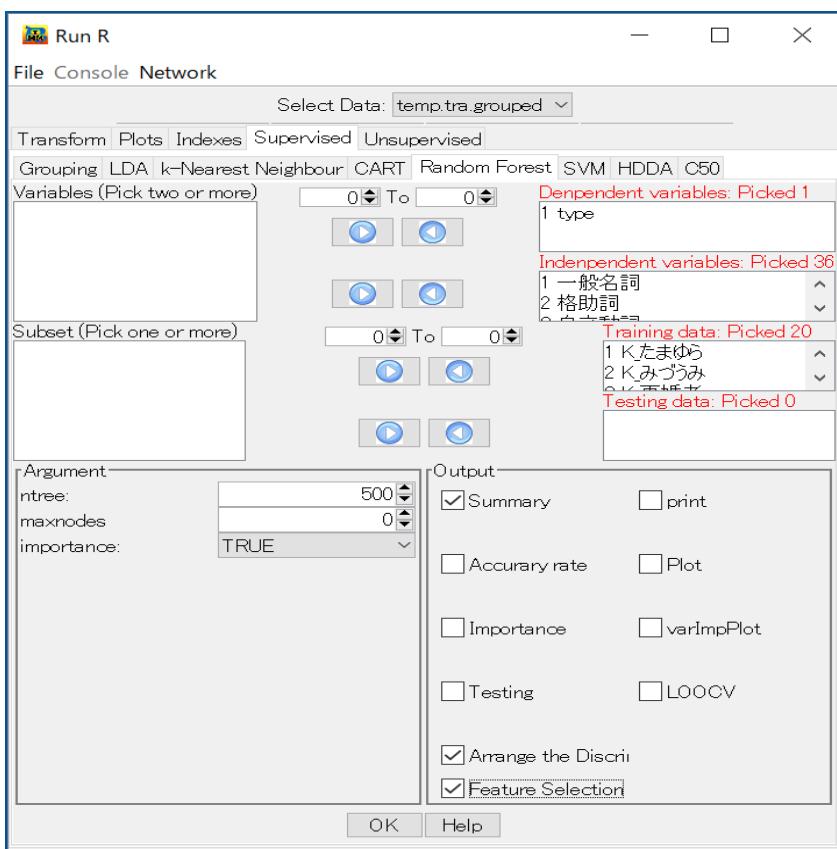
最初に目的変数①、説明変数②、学習データ③とテストデータ④を入れる。

●Arrange the Discrimination Maker:

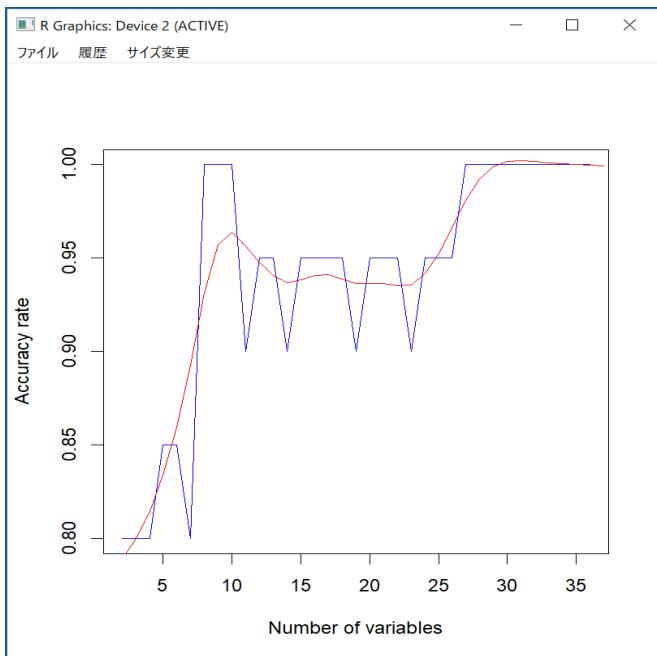
MeanDecreaseAccuracy で決めた変数重要度により、降順で変数を並べ替える。第一列は最も重要な変数である。結果を arrange に入れておく。"arrange"を左上の R Console に入力すると変数を並べ替えた結果を示す。

●Feature Selection:

並べ替えたベクトルを用いて、最初は一つの変数、一つずつを増やしてそれぞれの正判別率を計算し、結果を Smatrix に入れておく。



図に示しているように、前 26 位の変数を用いる場合は正判別率が最も高い。



●VarImpPlot:

「MeanDecreaseAccuracy」と「MeanDecreaseGini」二つのアルゴリズムで変数重要度を出す。

