第3回WordMiner活用セミナー
（2005/5/19-20）

構成要素変数の生成と編集

― WordMiner 辞書関連機能の紹介 ―

[本文]

1.構成要素と構成要素変数

2.分かち書き処理と構成要素変数の生成

3.構成要素変数の編集と編集辞書

[付録]
WordMiner Recipe 構成要素と構成要素変数

※本文中の RD_101 は、付録のWordMiner Recipeの関連するRecipe IDを示しています。

株式会社平和情報センター
保田　明夫
yasuda@hic.co.jp

1. 構成要素と構成要素変数

WordMinerにおける「構成要素」とは、テキスト型データの解析の対象となる基本の単位ことをいい、一般に分かち書き処理で区切られた文字や文字のならび（文字列）などを示す。
「構成要素変数」とは、この「構成要素」からなる変数のことであり、WordMinerの「構成要素変数の生成」機能により作成することができ、「構成要素変数の情報」、「データビュワー」、「構成要素の一覧と検索」など機能により、その内容を確認・観察することができる。

○ 構成要素変数の生成機能
① 分かち書きを行い、分かち書き及びキーワードの2つの構成要素変数を生成する。
② 空白区切りのデータ（変数）から構成要素変数を生成する。
③ 構成要素変数同士を併合して、新たな構成要素変数を生成する。

RD_101 RD_102 RD_103 RD_104

生成した構成要素変数は、WordMinerが解析するデータ表の基本形を構成する。

○ 解析データ表の基本形（二元のデータ表）の例
① <回答/サンプル> × <構成要素変数>
② <構成要素変数> × <質的変数（属性、デモグラフィック）>
③ <構成要素変数> × <質的変数（選択肢型設問、分類区分）>
④ <構成要素変数> × <サンプルのクラスター変数>

図1.に、WordMinerの操作の基本的な流れを示す。

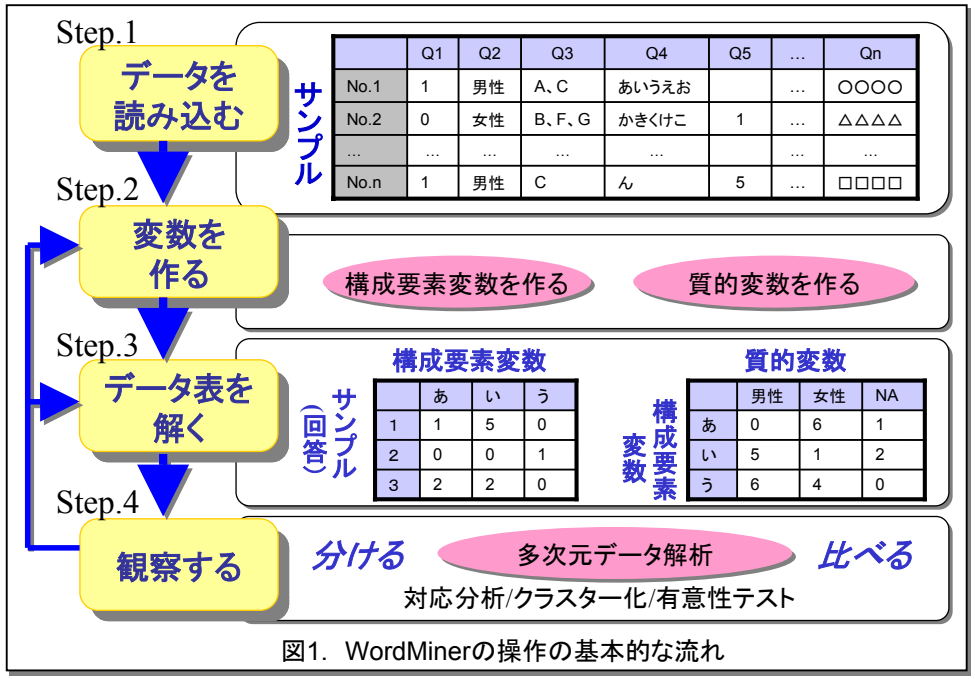

図1.　WordMinerの操作の基本的な流れ

構成要素変数は、解析の目的に応じて、様々な方法により置換・削除・抽出などの編集を行うことができる。

○ 構成要素変数の編集機能
① 編集辞書（置換辞書、削除辞書）による編集
解析に用いない構成要素の削除や標記のゆれ、言い回しの違いの編集
② 質的変数によるサンプルの抽出による編集
③ 閾値による抽出（ある出現頻度数以上による抽出）による編集 RD_501

○ その他の構成要素変数の編集方法
① 構成要素変数の生成時に、テキスト本文、または、分かち書き・キーワード抽出結果を修正する。
② 分かち書き・キーワード抽出結果（構成要素変数）をエクスポートし、適宜、エディタ等でそのファイルを修正する。修正後、再度、そのファイルを読み込み、構成要素変数を生成する。 RD_106

図2.に、構成要素変数（構成要素）の編集例を示す。

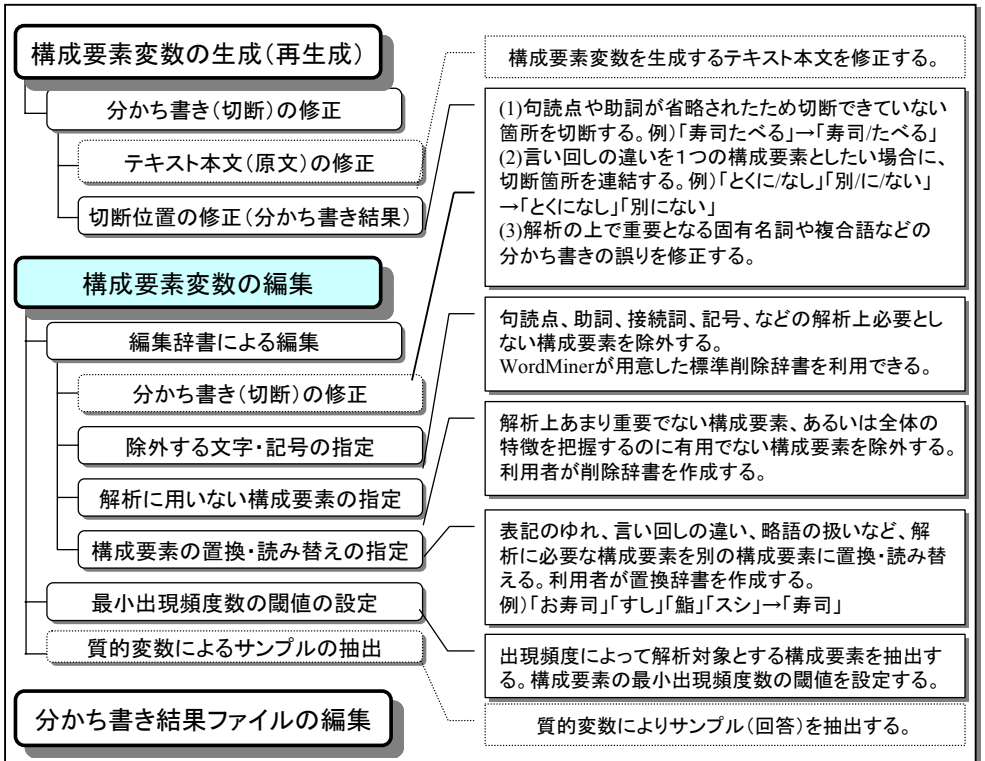

図2. 構成要素変数（構成要素）の編集例

生成・編集された構成要素変数は、その構成要素変数情報や構成要素の情報を検索・観察することができる。

○ 構成要素変数の情報
① 構成要素の出現頻度（閾値）による頻度別構成要素数、閾値水準による構成要素数
② 閾値による頻度別異なり構成要素数、閾値水準による異なり構成要素数
③ 編集辞書を適用した構成要素変数編集の削除・置換実行結果 RD_503
○ 構成要素の情報（構成要素の一覧と検索）
① 総出現頻度
② サンプル（回答）出現頻度
③ 文字列長
○ 構成要素の使われ方
① 文脈確認（コンコーダンス）
② サンプル（回答）検索
○ 解析データ表
① クロス表の出力
② その他、多次元データ解析結果

図3.に、構成要素変数の編集と解析・観察を示す。

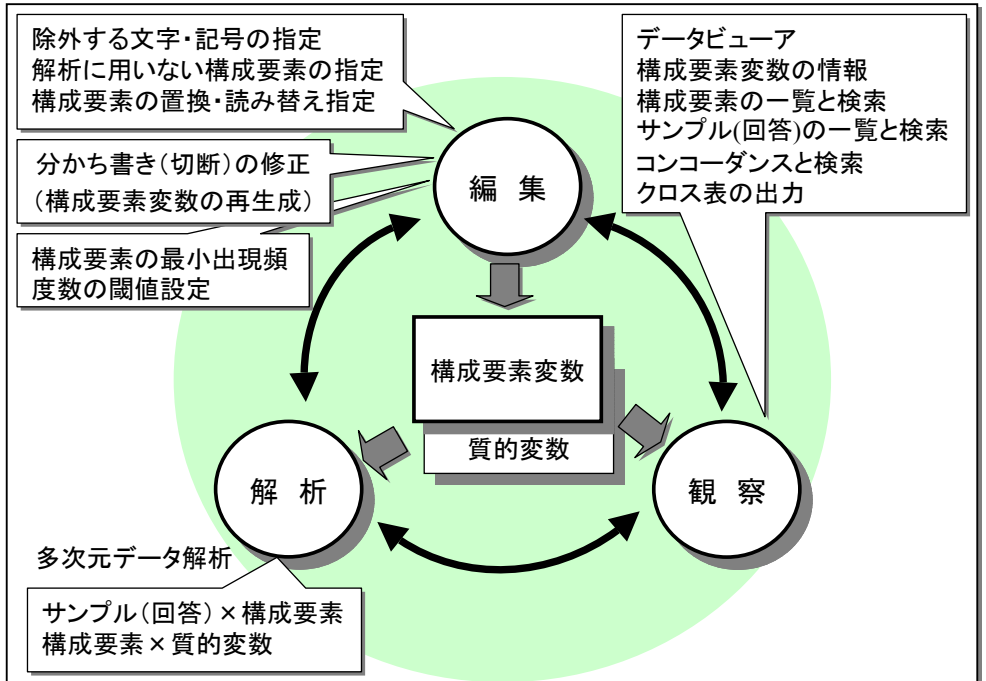

図3. 構成要素変数の編集と解析・観察

2. 分かち書き処理と構成要素変数の生成

WordMinerを用いてテキスト型データを解析する場合、一般に、まずテキスト型データを分かち書きして構成要素変数を生成する。
WordMinerでは、分かち書きによる構成要素変数の生成処理により、分かち書き結果とキーワード抽出結果からなる2つの構成要素変数が生成される。

○ 分かち書き処理による構成要素変数の生成
① 分かち書き処理結果による構成要素変数
原文を構成要素単位に分かち書きして得た構成要素変数
構成要素間の区切りは半角空白
② キーワード抽出結果による構成要素変数
分かち書き結果から、句読点や記号、用言などを削除し、主に名詞（名詞句）を抽出して得た構成要素変数
構成要素間の区切りは半角空白

RD_101

なお、分かち書き、キーワード抽出は、その処理方法をオプションで設定する。

○ 分かち書き、キーワード抽出処理のオプション
① 分かち書き、キーワードの最大文字数(構成要素の最大文字長)
② 指定した記号で括られた文字列は、分かち書きされず、キーワードとして抽出する
分かち書き回避記号（但し、組み合わせ範囲設定が優先）
③ 最短語または最長語による分かち書き処理
④ 最短語または最長語、もしくは、隣接する構成要素（用語）の組み合わせ範囲の設定によるキーワード抽出（最小語基と最大語基を設定する）

RD_105

図4.に、構成要素変数の生成（AiBASEによる分かち書き処理）を示す。

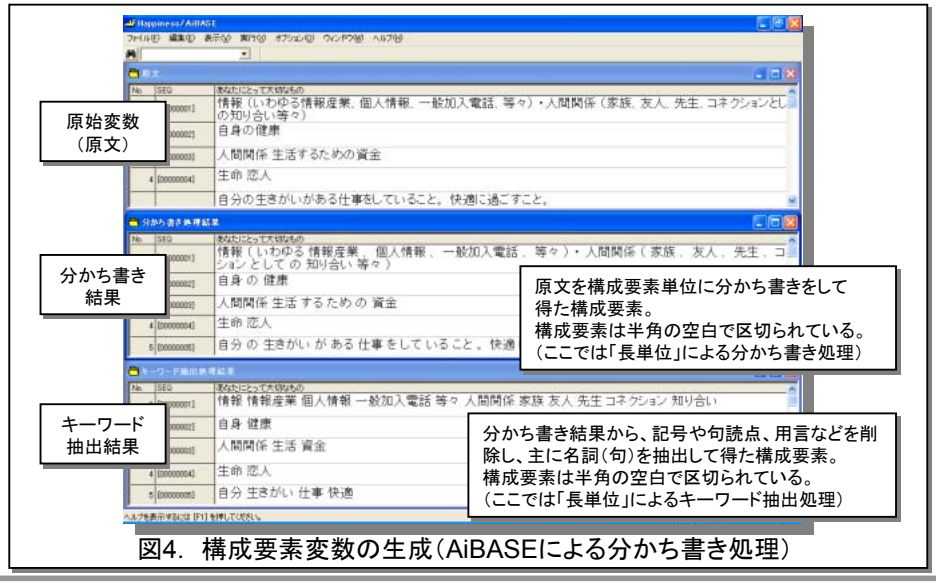

図4. 構成要素変数の生成（AiBASEによる分かち書き処理）

分かち書き処理は、分かち書き辞書と切断ルールにより、品詞による切断と複合語の切断を行い、最短語・最長語のオプションの設定により、構成要素を生成する。
最短語による場合は、複合語の切断を行い、最小単位の分かち書きによる構成要素が生成され、最長語による場合は、複合語の切断は行わず、品詞による分かち書きによる構成要素が生成される。

例えば、原文が「産学協同研究の成果」の場合、最長語方式では「産学協同研究」「の」「開発」の3つの構成要素が生成され、最短語方式では「産学」「協同」「研究」「の」「開発」の5つの構成要素が生成される。また、原文が「産学の協同による研究の成果」の場合は、最長語方式でも最短語方式でも、「産学」「の」「協同」「に」「よる」「研究」「の」「成果」の8つの構成要素が生成される。

また、分かち書き回避記号のオプションを設定することにより、回避記号で括られた文字列について、分かち書き（切断）を強制的に抑止することができる。
例えば、「『』」を分かち書き回避記号と設定した場合、「『産学の協同による研究』の成果」という原文からは、最長語・最短語のオプションに関わらず「『」「産学の協同による研究」「』」「の」「成果」の5つの構成要素が生成される。
ただし、括られた文字列内に空白（半角・全角とも）がある場合、空白は除外され、空白を詰めた文字列が構成要素として生成される。

RD_105

図5.に、分かち書き処理と構成要素を示す。

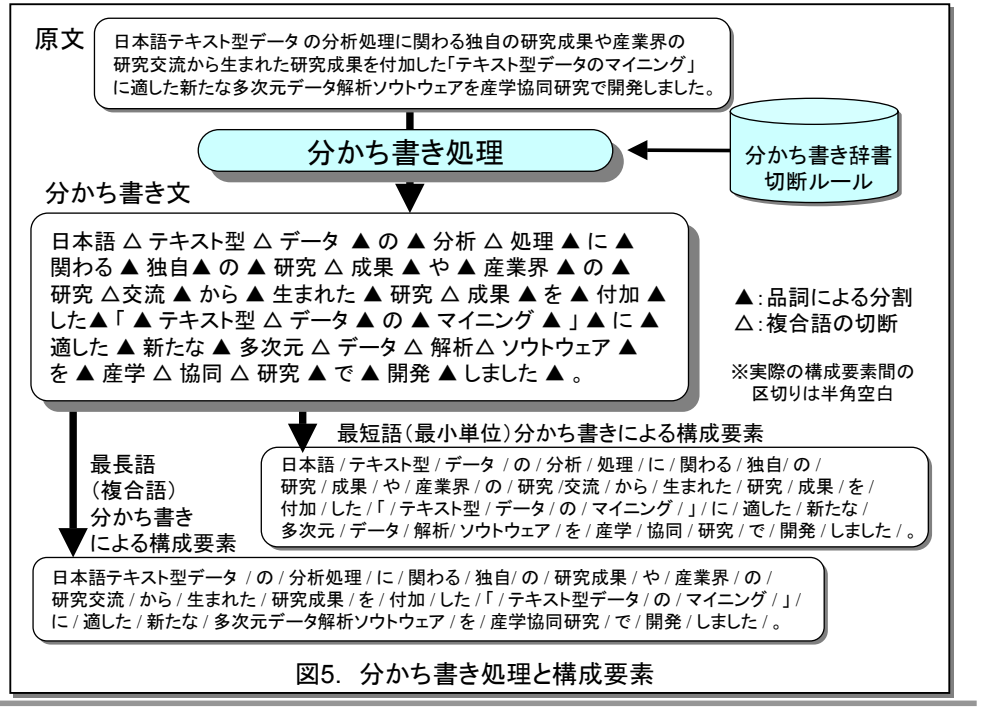

図5. 分かち書き処理と構成要素

キーワード抽出処理は、不要語辞書と抽出ルールにより、分かち書き文から、句読点や記号、用言などを除去し、最短語・最長語オプション、もしくは組み合わせオプションの設定により、構成要素を生成する。この方式を、辞書に登録された用語を抽出する「統制語方式」に対し、「不要語除去方式」と呼ぶ。「不要語除去方式」の利点は、新語や造語に強く、辞書登録・管理の手間が比較的容易なことにある。

ここで、組み合わせオプションとは、隣接する用語(複合語など品詞分割を超えない範囲)の組み合わせ範囲(最小，最大)を設定するものであり、最短語方式では組み合わせ範囲が(1，1)となる(組み合わせなし)構成要素を生成し、最長語方式では組み合わせ範囲が最大となる構成要素を生成する。

例えば、原文が「産学協同研究」の場合、組み合わせを(1，3)とすると、「産学」「協同」研究」「産学協同」「協同研究」「産学協同研究」の6つの構成要素が生成され、組み合わせ(2，2)では「産学協同」「協同研究」の2つの構成要素のみが生成される。組み合わせ(3，3)では「産学協同研究」のみ生成され、(4，4)の場合、この原文から構成要素は1つも生成されない。

また、分かち書き回避記号のオプションを設定すると、回避記号で括られた文字列はキーワードとして抽出される。(ただし、組み合わせ指定の最小語基が優先され、最小語基以上の組み合わせキーワードでなければ抽出されない。) RD_105

図6.に、キーワード抽出処理と構成要素を示す。

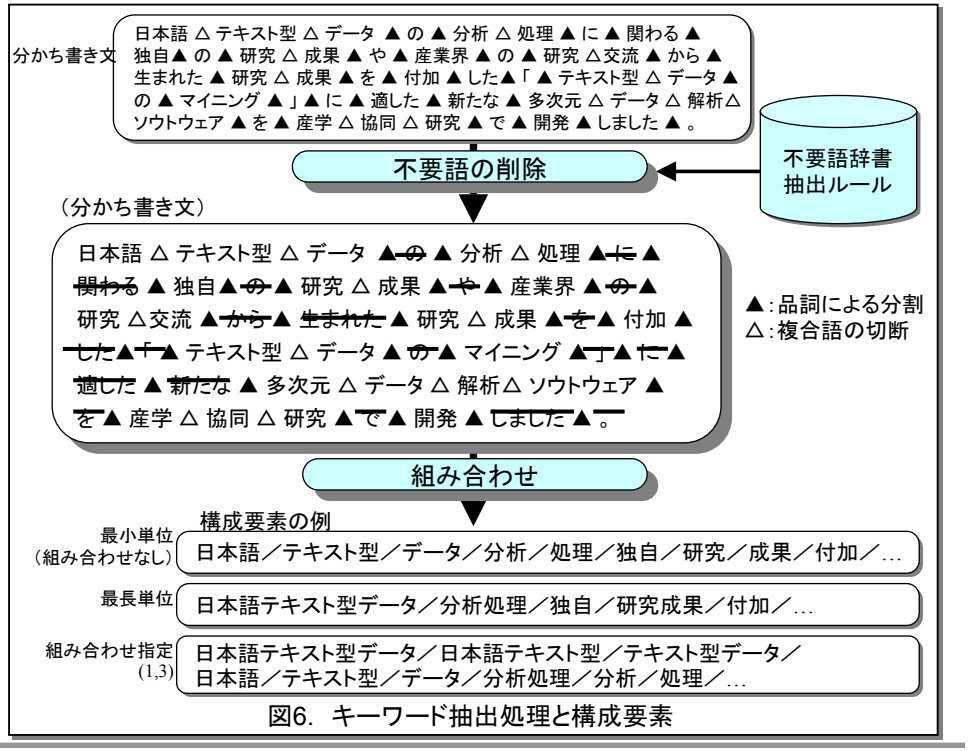

図6. キーワード抽出処理と構成要素

分かち書きによる構成要素変数には、句読点、記号、助詞類など、一般的に解析に用いない構成用語が多数含まれているので、編集辞書(削除辞書)により構成要素変数を編集する。

一方のキーワードによる構成要素変数は、構成要素が主に名詞(句)のため一見性には優れているが、形容詞や副詞、動詞などの用言が抽出されていない。

また、1つのサンプル(回答)で同一の表現が複数回出現した場合、分かち書きによる構成要素変数ではサンプル内の重複を認めカウントは出現回数となり、キーワードによる構成要素変数ではサンプル内での重複を認めず複数回出現してもカウントは1回となる。

最短語による場合は概念的・抽象的な構成要素が多くなる傾向にあり、逆に、最長語による場合は具体的・固有の構成要素が多くなる傾向がある。最長語の場合、組み合わせられた要素に同じものがあっても、構成要素としては、まったく別のものとして扱うことになる。

例えば、「研究開発」「委託研究」「委託開発」といった3つの原文について、最短語では「研究開発」と「委託研究」は「研究」で、「委託研究」と「委託開発」は「委託」で、「委託開発」と「研究開発」は「開発」で、それぞれ共通の構成要素を持つことになるが、最長語ではこれらの3つの原文に共通の構成要素は存在しないことになる。

分かち書きによるか、キーワードによるか、あるいは、最短語か、最長語か、構成要素変数の生成方法は、解析の目的やデータの取得法、収集データの特性などにより選択する。

図7.に、構成要素生成の例を示す。

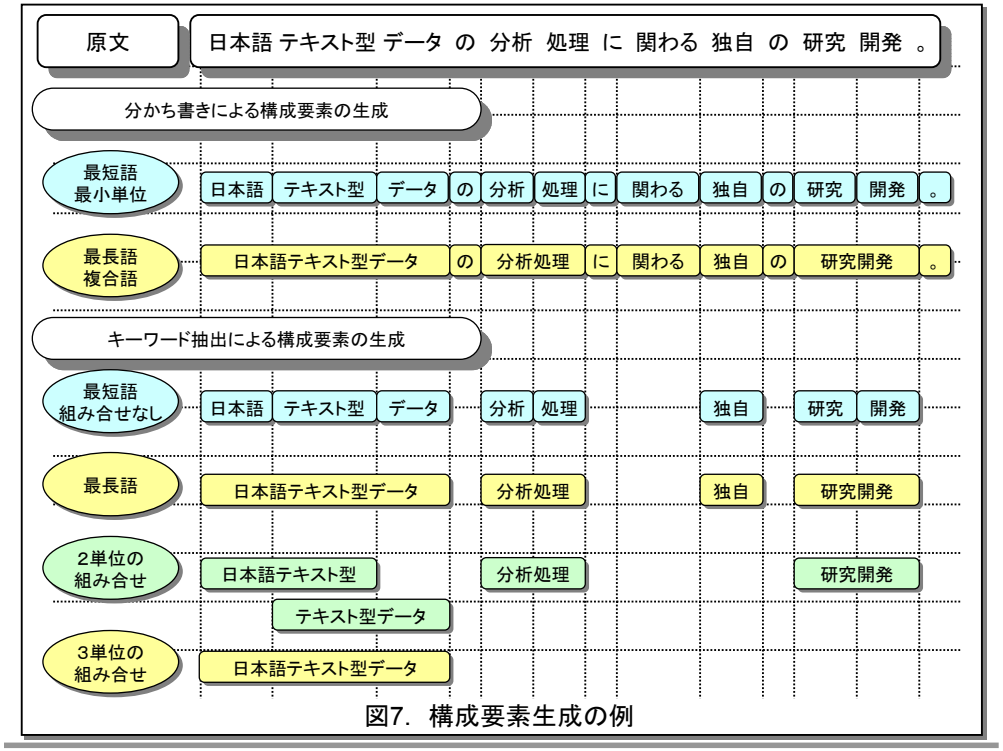

図7. 構成要素生成の例

表1.に、分かち書きオプションと構成要素生成の例を示す。

最短語・最長語の分かち書きオプションは、分かち書きによる構成要素変数の生成とキーワードによる構成要素変数の生成において、個々に独立して有効となる。また、組み合わせ設定のオプションは、キーワードによる構成要素変数の生成においてのみ有効となる。なお、組み合わせ設定の最小語基数は、分かち書き回避記号の設定より優先される。すなわち、回避記号により括られたキーワードでも、最小語基数以下であれば抽出されない。 RD_105

表1. 分かち書きオプションと構成要素生成の例

	分かち書き・キーワード抽出基準		組み合わせ語基数		分かち書き回避記号の設定:「」	生成される構成要素変数 【原文】「川べりの道」で文学界新人賞受賞、「駆ける少年」で泉鏡花文学賞を受賞した。 (△:半角空白を示す)	構成要素	
	最短語	最長語	最小	最大			総数	異なり数
分かち書き処理	ー	チェック	ー	ー	なし	「△川べり△の△道△」△で△文学界新人賞受賞△、△「△駆ける△少年△」△で△泉鏡花文学賞△を△受賞△した△ 。	18	15
					あり	「△川べりの道△」△で△文学界新人賞受賞△、△「△駆ける少年△」△で△泉鏡花文学賞△を△受賞△した△ 。	15	12
	ー	チェック オフ	ー	ー	なし	「△川べり△の△道△」△で△文学界△新人賞△受賞△、△「△駆ける△少年△」△で△泉△鏡花△文学賞△を△受賞△した△ 。	22	18
					あり	「△川べりの道△」△で△文学界△新人賞△受賞△、△「△駆ける少年△」△で△泉△鏡花△文学賞△を△受賞△した△ 。	19	15
キーワード	チェック オフ	チェック	ー	ー	なし	川べり△道△文学界新人賞受賞△少年△泉鏡花文学賞△受賞	6	6
					あり	川べりの道△文学界新人賞受賞△駆ける少年△泉鏡花文学賞△受賞	5	5
	チェック	チェック オフ	ー	ー	なし	川べり△道△文学界△新人賞△受賞△少年△泉△鏡花△文学賞	9	9
					あり	川べりの道△文学界△新人賞△受賞△駆ける少年△泉△鏡花△文学賞	8	8
	チェック	チェック	ー	ー	なし	川べり△道△文学界△文学界新人賞受賞△新人賞△受賞△少年△泉△泉鏡花文学賞△鏡花△文学賞	11	11
					あり	川べりの道△文学界△文学界新人賞受賞△新人賞△受賞△駆ける少年△泉△泉鏡花文学賞△鏡花△文学賞	10	10
	チェック オフ	チェック オフ	1	3	なし	川べり△道△文学界△文学界新人賞△文学界新人賞受賞△新人賞△新人賞受賞△受賞△少年△泉△泉鏡花△泉鏡花文学賞△鏡花△鏡花文学賞△文学賞	15	15
					あり	川べりの道△文学界△文学界新人賞△文学界新人賞受賞△新人賞△新人賞受賞△受賞△駆ける少年△泉△泉鏡花△泉鏡花文学賞△鏡花△鏡花文学賞△文学賞	14	14
	チェック オフ	チェック オフ	3	3	なし	□ □□ □□ □ □□ △□ □ □□ □□	2	2
					あり	□ □□ □□ □ □□ △□ □ □□ □□	2	2

3. 構成要素変数の編集と編集辞書

構成要素変数の編集辞書の作成は、変数の編集前にプロジェクト共有辞書として作成しておく方法と、変数の編集(の設定)時に編集固有の辞書として作成する方法がある。 RD_201 RD_202

新規に、「構成要素変数の編集の設定」を行うと、その時点で「構成要素の編集辞書の管理」で管理(表示)されるすべての辞書が、編集する構成要素変数の編集辞書としてコピーされる。すなわち、この時点で、編集辞書の名前が同じでも、互いに独立した辞書(群)となっており、編集辞書は個々の構成要素変数(編集名)に固有のものとなる。従って、いずれかの辞書を修正しても他の辞書が影響を受けることはなく、また、編集辞書を新規に作成しても、その辞書が自動的にコピーされることはない。構成要素変数の編集を行った編集辞書は、「構成要素変数の編集名」で管理される。

図9に、構成要素変数の編集と編集辞書を示す。 RD_105 RD_203 RD_204 RD_303

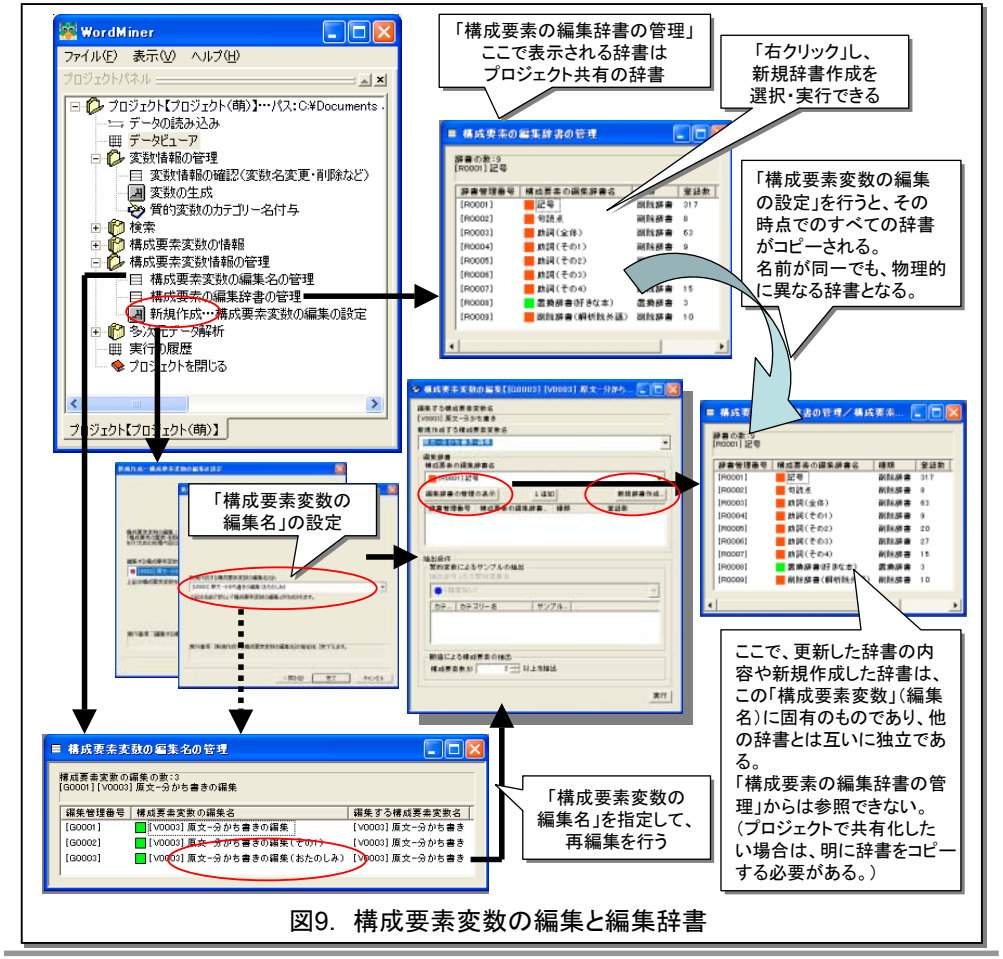

図9. 構成要素変数の編集と編集辞書

構成要素変数を編集する編集辞書には、置換辞書と削除辞書の2種類がある。
置換辞書は、単に構成要素を別の構成要素に置換する他、分かち書き結果の修正（切断位置の訂正）を行うことができる。
なお、WordMinerには、句読点、助詞、接続詞、記号などの標準削除辞書が装備されている。

○ 編集辞書
① 置換辞書
・表記のゆれ、言い回しの違い、略語の扱いなど、解析に必要な構成要素を別の構成要素に置換・読み替える。
例）「お寿司」「すし」「鮨」「スシ」→「寿司」

・句読点や助詞が省略されたため切断できていない箇所を切断する。
例）「寿司たべる」→「寿司/たべる」

・言い回しの違いを1つの構成要素としたい場合に、切断箇所を連結する。
例）「とくに/なし」「別/に/ない」→「とくになし」「別にない」

RD_401 RD_402

・解析の上で重要となる固有名詞や複合語などの分かち書き結果を修正する。
例）「回転/寿司」→「回転寿司」

② 削除辞書
・句読点、助詞、接続詞、記号、などの解析上必要としない構成要素を除外する。この場合、WordMinerが装備している標準削除辞書を利用できる。

・解析上あまり重要でない構成要素、あるいは全体の特徴を把握する上で有用でない構成要素を除外する。

編集辞書は、直接入力して作成する以外にも、「構成要素の一覧と検索」や別の編集辞書からコピーして作成したり、外部ファイルの内容を取り込んで作成することも可能である。

○ 編集辞書の主な作成方法
① 直接入力
② 「構成要素の一覧と検索」、その他、WordMinerの結果画面からのコピー
③ 既に作成した編集辞書からのコピー
④ 外部ファイルのデータの取り込み

RD_301 RD_302 RD_303 RD_304 RD_305

編集辞書は、置換及び削除辞書とも複数用意し、構成要素変数の編集時に複数適用することができる。複数の編集辞書を適用した場合には、設定した順序に従って辞書が適用される。
編集辞書の適用結果は、「構成要素変数の情報」で確認する。

RD_502 RD_503

なお、構成要素変数の編集は、編集辞書以外にも、「質的変数によるサンプルの抽出」と「閾値による構成要素の抽出」により、編集を行うことができる。

○ 構成要素変数の編集の適用順序
① 編集辞書（削除・知久）による編集
② 質的変数によるサンプルの抽出
③ 閾値による構成要素の抽出

RD_501