

付録資料 : 「テキストマイニングのための1・0データファイルの作成手順」

# テキストマイニングのための1・0データファイルの作成手順

東北農業研究センター 総合研究部

磯島昭代

## 目次

1．はじめに	・・・・・・・・	1
2．本書の解説内容		
(1) 本書の解説範囲と取り扱うデータの種類	・・・・・・・・	2
(2) 1・0データファイル作成手順の概要	・・・・・・・・	4
(3) マクロの内容	・・・・・・・・	5
3．データファイル作成手順		
(1) 文章を形態素解析ツールで分解する	・・・・・・・・	7
(2) 形態素に文章番号を付与する	・・・・・・・・	10
(3) 品詞情報を利用してキーワード候補の抽出を行う	・・・・・・・・	12
(4) 基本形の「ひらがな」を「カタカナ」に変換する	・・・・・・・・	15
(5) 形態素番号と出現数を割り当てる	・・・・・・・・	18
(6) 出現数からキーワードを絞り込む	・・・・・・・・	21
(7) 「1・0データ化」の準備	・・・・・・・・	26
(8) マクロの実行	・・・・・・・・	30
(9) 1・0データファイルの完成	・・・・・・・・	34
4．補足説明		
(1) 2つ以上のキーワードをまとめる場合	・・・・・・・・	38
(2) マクロを使わずに1・0データファイルを作成する方法	・・・・・・・・	42
5．むすび	・・・・・・・・	50
付録    マクロのフローチャート	・・・・・・・・	51
付録    マクロのコード	・・・・・・・・	52
付録    「茶坊主くん.txt」からVBファイルを作成する	・・・・・・・・	56

## 1. はじめに

アンケート調査などで得られる自由記述回答文やクレームデータなどに含まれる文章データは、非定型の定性的データであるためこれまでは計量的な分析手段がなく、活用されることもほとんどなかった。近年、こうした大量の文章データを電子化して計量的に分析するための手法としてテキストマイニング<sup>注(1)</sup>が注目されるようになったが、同手法は未だ開発途上にある上、市販されている専用のソフトウェアは非常に高額であるため、誰もが利用できる状況にあるとはいえない。

こうした現状を踏まえ、磯島(2002)はフリーソフトの形態素解析システム「茶筌(ちゃせん)」<sup>注(2)</sup>と表計算ソフト「Excel2002」および統計処理ソフトを用いて、低コストで簡便にできるテキストマイニングの手法(以下、本手法)を紹介している。

本書では、このテキストマイニングの過程で必要なデータファイル作成の手順を説明する。上記の通り、本手法はフリーソフトと一般的に使用されている表計算ソフトおよび統計処理ソフトを利用しているため、新たな投資を必要とせず、低コストで分析を行うことができる。一方、高度な専門知識を要さないシンプルな手法であるため、簡便ではあるがその作業手順は若干複雑である。基本的にはExcelのコピー&貼り付け機能、オートフィルタ機能、データの並べ替え機能、簡単な関数などでほとんどの作業を行い(一部Wordを利用)、単純作業の繰り返しが多くなる「1・0データ化」の部分だけは簡単なマクロプログラム(以下、マクロ)を組んでいる<sup>注(3)</sup>。

本手法は試行段階であるため、最小限の作業しかマクロを組んでおらず、汎用性に乏しい。従って、その前段階であるキーワードの抽出において、データの位置やシートの名前などは既定のものとしている。そのため、「とりあえずやってみよう」という方は、これからお示しする手順に沿って作業をしていただきたい。一方、マクロの知識をお持ちの方は、ご自由に使い勝手の良いように書き換えていただければと考えている。

なお、本マクロにおいて不具合を発見された方は、ご一報いただければ幸いであるが、筆者はプログラミングの初心者であり、必ずしも対処できるとは限らないのでその点はご了承ください。また、本マクロを使用して発生したいかなる損害に対しても責任を負いかねることをここに明記しておく。

注(1) テキストマイニングの概要については、市村・長谷川・渡部・佐藤(2001)を参照のこと。

注(2) 奈良先端科学技術大学院大学自然言語処理学講座が提供しているフリーソフトウェア。詳細は <http://chasen.aist-nara.ac.jp/index.html.ja> を参照のこと。

注(3) 筆者が使用したパソコンの環境は以下の通りである。OS: Microsoft Windows XP SP1、メモリ: 1GB、Excel2002、Word2002。

## 2．本書の解説内容

### (1) 本書の解説範囲と取り扱うデータの種類

筆者が想定しているテキストマイニングのプロセスと、本書において解説する範囲について説明する(図1)。テキストマイニングのプロセスは、大きく、文章データの入手、1・0データファイルの作成、データの分析、の3段階に分けることができる。このうち、本書で扱う範囲は、の「1・0データファイルの作成」を中心とする部分である。

文章データの入手および分析については、本書では解説の対象外とするが<sup>注(4)</sup>、取り扱う文章データの種類について若干触れておく。

文章データについては、アンケート調査の自由記述回答文、クレームデータの他、Web上の掲示板など、様々な入手方法が考えられる。また、データの種類も、「についてご自由にお書き下さい」という非定型的自由記述文から、「は、()なので、()だと思ふ」などのように、文章の一部を空白にしたフォーマットを用意し、その空白を埋めてもらう形の定型自由文など様々である<sup>注(5)</sup>。

本書で取り上げるテキストマイニング手法では、分析する文章データとして、

文章データの1単位(例えば、1人の発言、1クレームなど)がExcelの1つのセルに入力できる程度の分量であること、

上記単位の文章が、統計的処理を行うに相当する分量だけあること、

を原則としている。そのため、論文1本、議事録の1議題などの長文は想定外としているが、比較的短い文章もしくは単語が大量に存在しているような文章データであれば、定型・非定型によらず基本的に適用可能であると考えている(図2)。

注(4)データの分析については、磯島(2002)、磯島(2004)、磯島・野中・清野(2004)を参照のこと。

注(5)林(2002)は定型自由文へのテキストマイニングの適用事例を紹介している。

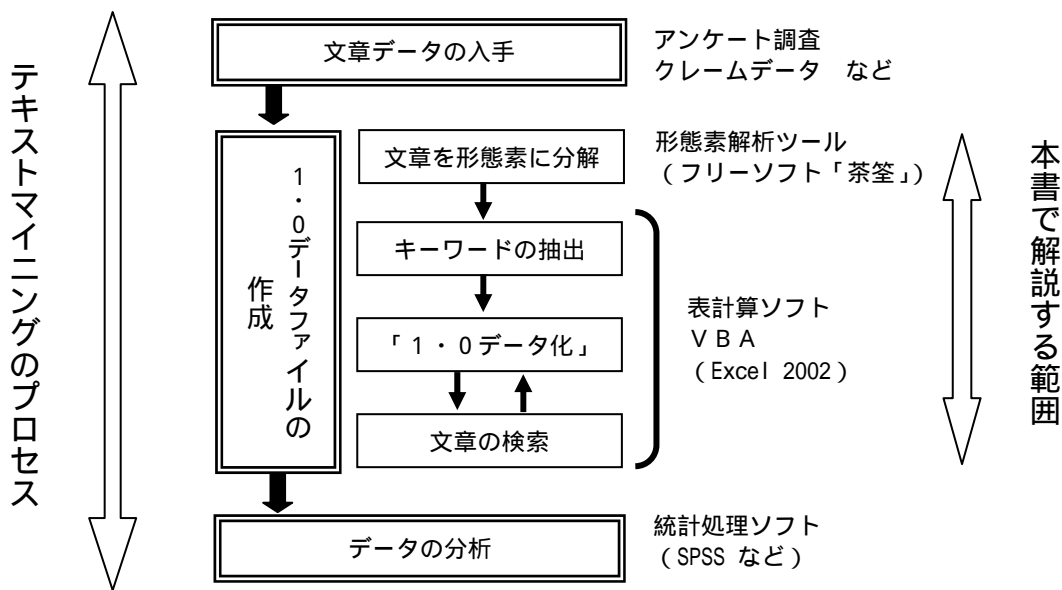


図1 テキストマイニングのプロセスと本書の解説範囲

	A	B	C	D	E
1	No.	code1	code2	code3	記述文
2	1	2	1	2	同じ物でも、とても高く感じる時は手頃な青菜にしています。
3	2	3	4	2	安く、新鮮で安全なもの
4	3	3	6	2	旬の時期になるべく購入し旬でない時は売っていても他の野菜で
5	4	3	3	1	新しい種類が売っていても、利用のしかたが分からなくて買わない
6	5	3	3	2	生協にある青菜ならば(野菜全般にも思っていることだが)安くで
7	6	2	4	2	青菜類は100円程度の価格の時に購入するので、冬はあまり購入
8				2	新鮮でなるべく県産品を購入
9				2	家族が少ないと一束が多すぎてすぐ葉が萎びてしまうので、少量、
10				1	ほぼ毎日、必ず食べる様にしてはいる。
11				4	袋に入っているけれど、葉のほうは空気にふれているので、傷み
12				2	時期にあったものを購入しています。その方が美味しく、また、安
13				2	行者にんにくを店においてほしい
14	13	4	3	2	新鮮で無農薬であること
15	14	3	5	2	安さ 安全 地物
16	15	2	4	2	栄養価のことを気にして、意識して購入しています。(カリウム、カ
17	16	3	4	2	特にありません
18	17	2	4	2	やはり、気になるのは、農薬。もっと、はっきりと、分かりやすい、

図2 基本となるデータファイル

## (2) 1・0データファイル作成手順の概要

1・0データファイルの作成手順の詳細については次章で述べることにし、ここでは簡単に概要を説明する(図3)。

- (A) まず、文章データを形態素に分解する。これには、フリーソフトウェアの形態素解析ツール「茶筌」を用いる。
- (B) 次に、分解された形態素が、どの文章に含まれていたかを明らかにするために、各形態素に文章番号を付与する。
- (C) 続いて、各形態素の品詞情報から、キーワードの候補となる形態素を抽出する。
- (D) 抽出した形態素は、基本形が同じである語をまとめて、形態素番号と出現数を割り当てる。
- (E) この出現数を基準にして、キーワードの絞り込みを行う。
- (F) 抽出したキーワードと各文章番号とを関連づけるために、マクロを使用して「1・0データ化」を行う。
- (G) 最後に、元の文章データと完成した1・0データを結合する。

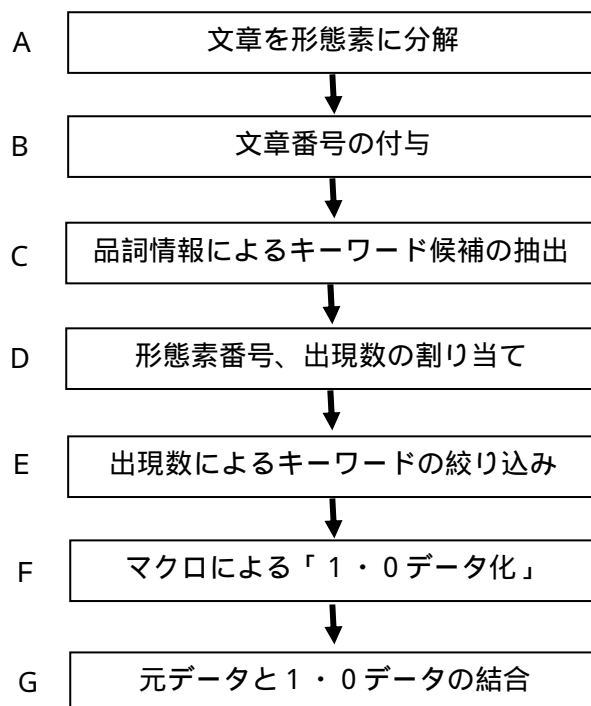


図3 1・0データファイルの作成手順

### (3) マクロの内容

本書で紹介する1・0データファイルの作成手順では、抽出したキーワードが各文章に出現するか否かを示す「1・0データ化」の部分にのみマクロを使用している。ここでは、そのマクロの内容について簡単に説明する。なお、巻末にマクロのフローチャートとコードを掲載しているので、詳細についてはそちらをご覧ください。

前述のように、マクロを実行する前段階として、各形態素にはどの文章に出現した語かを示す文章番号を付与しており、さらに同じ基本形をもつ形態素には、同じ形態素番号を割り当てている。従って、それぞれの文章に含まれる形態素の番号と、キーワードとして抽出した形態素番号（以下、「キーワード番号」とを照合することにより、各文章における各キーワードの出現の有無を調べることができる。

本書で用いるマクロでは、まず文章番号の小さい順に1つの文章を選び、その文章に出現する形態素番号を小さい順に1つずつキーワード番号と照合していく(図4)。キーワード番号と出現形態素番号が等しい場合には1を出力し、次のキーワード番号および出現形態素番号に移動する。キーワード番号が出現形態素番号よりも大きい場合には、次の出現形態素番号と照合する。出現形態素番号がキーワード番号と等しくならずキーワード番号よりも大きい数値を示したら、そのキーワード番号に対応するセルには0を出力し、次のキーワード番号に移る。

これを、もう少し具体的に説明する。例えば文章1に出現した形態素番号が「1、2、3、5、7、8、12」、抽出したキーワード番号が「2、4、8、10」であったとする(図5)。

まず、形態素番号「1」とキーワード番号「2」を照合する。

「形態素番号」<「キーワード番号」なので、形態素番号「2」に移る。

形態素番号「2」とキーワード番号「2」を照合する。

「形態素番号」=「キーワード番号」なので「1」を出力する。

形態素番号「3」、キーワード番号「4」にそれぞれ移る。

形態素番号「3」とキーワード番号「4」を照合する。

「形態素番号」<「キーワード番号」なので、形態素番号「5」に移る。

形態素番号「5」とキーワード番号「4」を照合する。

「形態素番号」>「キーワード番号」なので、「0」を出力する。

キーワード番号「8」に移る。

形態素番号「5」とキーワード番号「8」を照合する (以下略)。

以上の手順を、全ての文章番号について行う。

Microsoft Excel - 形態素.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

読んdel!!ココ(K)

E2 f 1

	A	B	C	D	E	F	G	H
1	表層語	基本形	カタカナ	品詞	記述No	形態素No	出現	
2	高く	高い	高イ	形容詞-自	1	12		文章に出現した 形態素の番号
3	し	する	スル	動詞-自立	1	29		
4	感じる	感じる	感シル	動詞-自立	1	49		
5	青菜	青菜	青菜	名詞-一般	1	189	1	15
6	手頃	手頃	手頃	名詞-形容	1	235	1	1
7	安く	安い	安イ	形容詞-自	2	10	1	12
8	安全	安全	安全	名詞-形容	2	229	1	5
9	新鮮	新鮮	新鮮	名詞-形容	2	236	1	17
10	し	する	スル	動詞-自立	3			
11	する	する	スル	動詞-自立	3			
12	なる	なる	ナル	動詞-自立	3			
13	売っ	売る	売ル	動詞-自立	3			
14	購入	購入	購入	名詞-サ変	3			
15	代用	代用	代用	名詞-サ変	3			
16	句	句	句	名詞-一般	3			
17	句	句	句	名詞-一般	3	174	2	2
18	他	他	他	名詞-一般	3	199	1	2

文章番号

文章1には、  
12、29、49、189、235  
の5つの形態素が  
含まれている

Microsoft Excel - 形態素.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

読んdel!!ココ(K)

MS Pゴシック

A94 f

	A	B	C	D	E	F	G	H
1	1	7	9	10	12	13	15	18
2	2	0	0	0	1	0	0	0
3	3							
4	4							
5	5							
6	6							
7	7							
8	8							
9	9							
10	10							
11	11							
12	12							
13	13							
14	14							
15	15							
16	16							
17	17							

文章番号

キーワード番号と、  
各文章中に出現した  
形態素番号を照合し、  
同じ番号の場合は1、  
それ以外は0とする。

1つの文章が終わったら、  
次の文章の形態素番号を照合する

キーワード番号

図4 マクロの内容



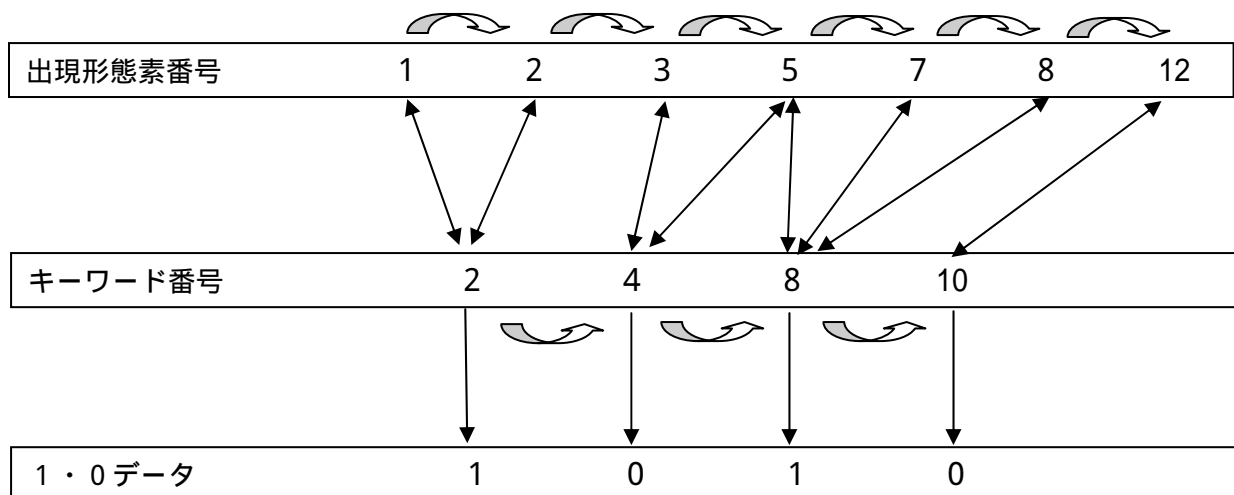


図5 形態素番号とキーワード番号の照合手順

### 3. データファイル作成手順

#### (1) 文章を形態素解析ツールで分解する

本手法は、形態素解析ツール「茶筌」を利用している。「茶筌」は奈良先端科学技術大学院大学自然言語処理学講座で開発された日本語形態素解析システムのフリーソフトであり、以下のサイト（画面1）で管理、配布されている。

<http://chasen.aist-nara.ac.jp/>

なお、本書では茶筌の説明は割愛する。詳しい内容や使用方法については、上記サイトおよび林（2002）を参照すること。

分析する文章データは、あらかじめ Excel ファイルに入力しておく。ここでは、アンケート調査の結果（回答数 92 件）を事例としているが、1つの行につき1人分の文章番号（回答者番号、記述番号）回答者属性、文章データを入力している（画面2）。クレームデータなども同様に、1行に文章データとそれに関連するデータを入力する。このとき、文章データは原則として1つのセルに収める。

また、この文章データでは、「半角カタカナ」を使わないように注意したい。なぜならば、「茶筌」で形態素解析を行う際に「半角カタカナ」があるとバグが発生し、一見すると普通に解析しているようであっても、以降の手順に大きな障害が残るからである。既に入力されているデータに半角カタカナが含まれている場合には、Word の「文字種の変換」機能（後述）などを用いて、全角カタカナに変換しておくことが大切である。

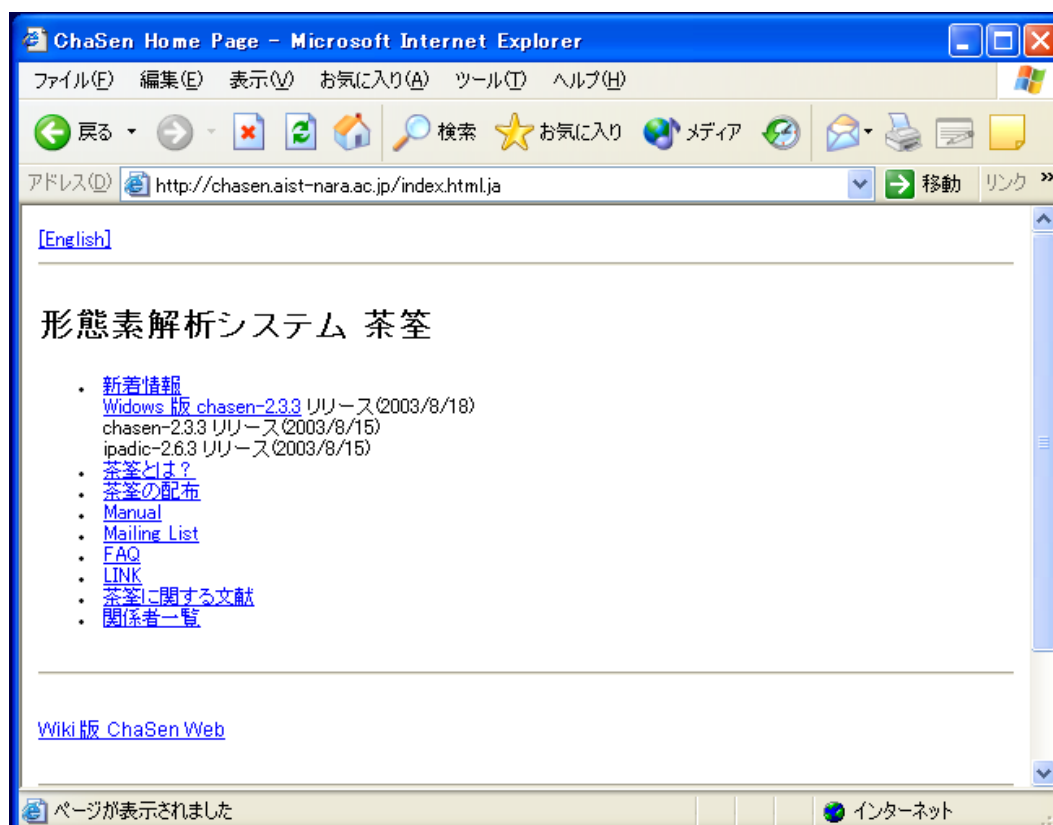
まず、「茶筌」の初期画面を立ち上げる（画面3）。「クリア」ボタンをクリックして「文エリア」を空にし、そこに画面2の文章データを一括してコピー＆貼り付けする。「文エリア」の下にある

「表層語」「基本形」「品詞」にチェックを入れ、最後に「全文解析」をクリックする。

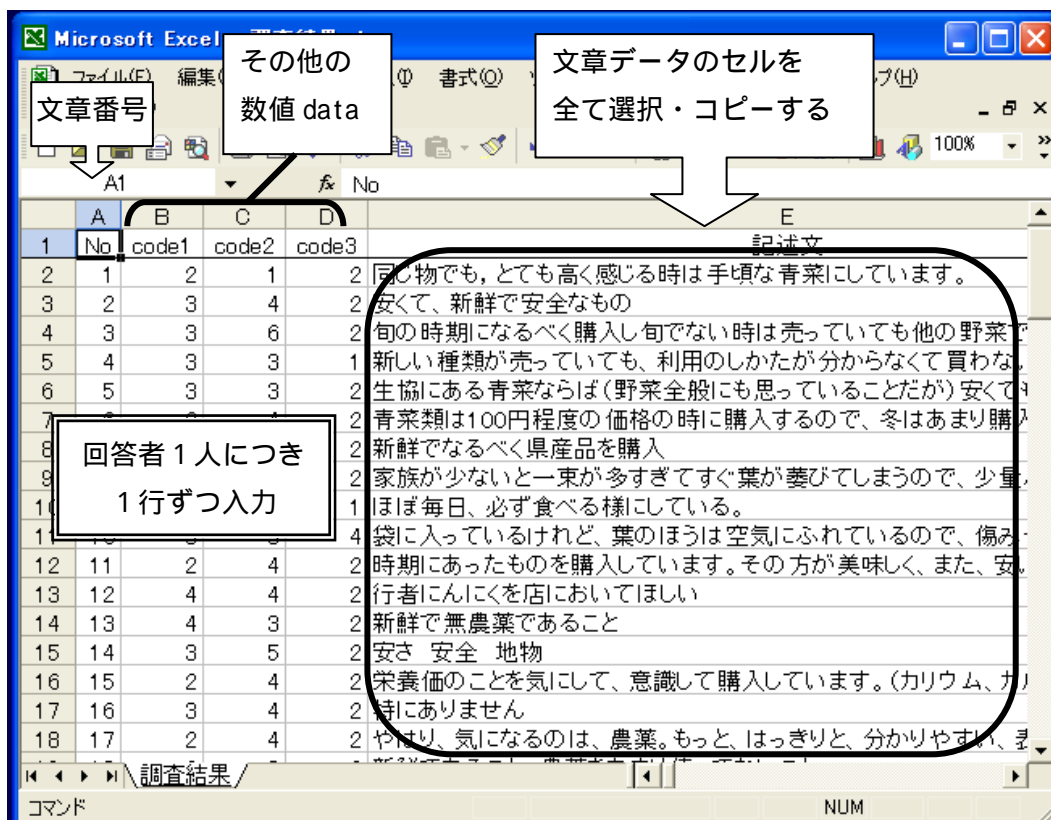
ここで、通常は画面下部の「形態素エリア」に解析結果が表示されるが、テキストマイニングを行うほどの文章量になると「解析結果が大きすぎて全体を表示できません。今すぐ結果を保存しますか?」と聞かれるので、「OK」をクリックしてテキストファイルで保存する。

次に、保存したテキストファイル(事例では「形態素.txt」)を Excel で読み込む。形態素解析の結果はテキスト保存してあるので、Excel でファイルを開く場合は「ファイルの種類(T)」を「すべてのファイル(\*.\*)」にしてファイル名を選択する。さらに、「データファイルの形式を選択して下さい」では、「カンマやタブなどの区切り文字によってフィールドごとに区切られたデータ(D)」にチェックを入れる。

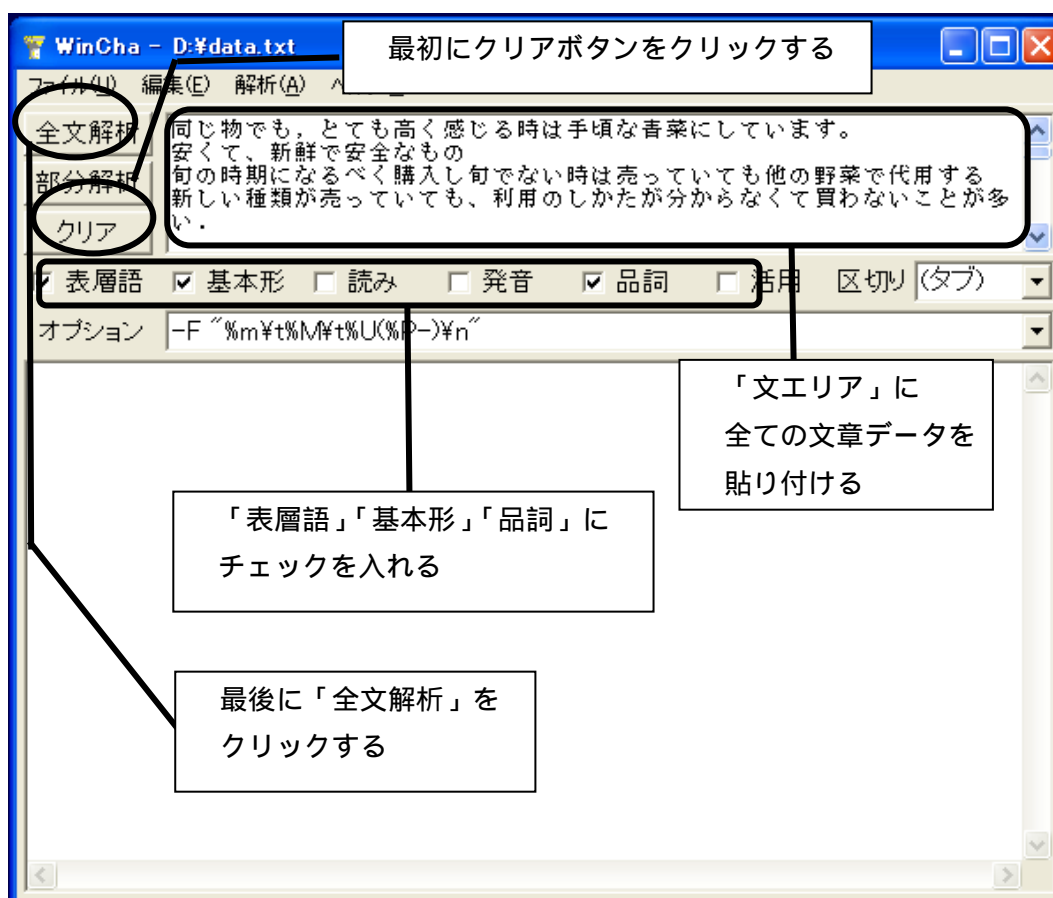
読み込んだデータは、A列が文章を形態素に分解した「表層語」(これを縦につなげると元の文章と等しくなる)を、B列は「基本形」を、C列はその「品詞」を示している。各列の名前を付けるために一番上に1行挿入し、それぞれの列に「表層語」「基本形」「品詞」と入力する(画面4)。



画面1 「茶筌」ホームページ



画面2 文章データのファイル



画面3 「茶筌」による解析手順

	A	B	C	D	E	F	G	H
1	表層語	基本形	品詞					
2	同じ	同じ	連体詞					
3	物	物	名詞-非自立-一般					
4	で	だ	助動詞					
5	も	も	助詞-係助詞					
6	,	,	記号-読点					
7	とても	とても	副詞-助詞類接続					
8	高く	高い	形容詞-自立					
9	感じる	感じる	動詞-自立					
10	時	時	名詞-非自立-副詞可能					
11	は	は	助詞-係助詞					
12	手頃	手頃	名詞-形容動詞語幹					
13	な	だ	助動詞					
14	青菜	青菜	名詞-一般					
15	に	に	助詞-格助詞-一般					
16	し	する	動詞-自立					
17	て	て	助詞-接続助詞					
18	い	いる	動詞-非自立					

画面4 形態素解析結果

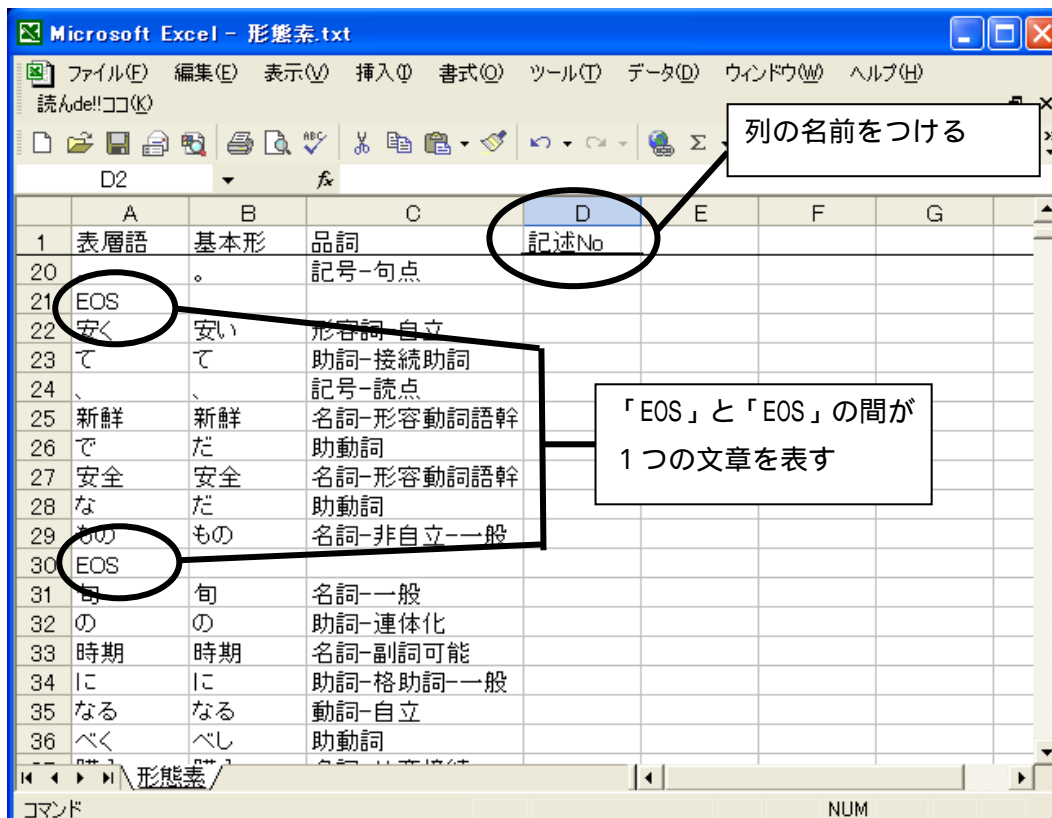
## (2) 形態素に文章番号を付与する

ここで作成しようとしている1・0データファイルは、「どの文章にどのキーワードが含まれているか」を明示するためのものである。そのため、いったん分解された形態素がどの文章において出現した語であるかを明確にしておく必要がある。ここでは、各々の形態素に元の文章と対応する文章番号（回答者番号、クレーム番号など。以降、説明画面では「記述No」と表示）を付与することにする。

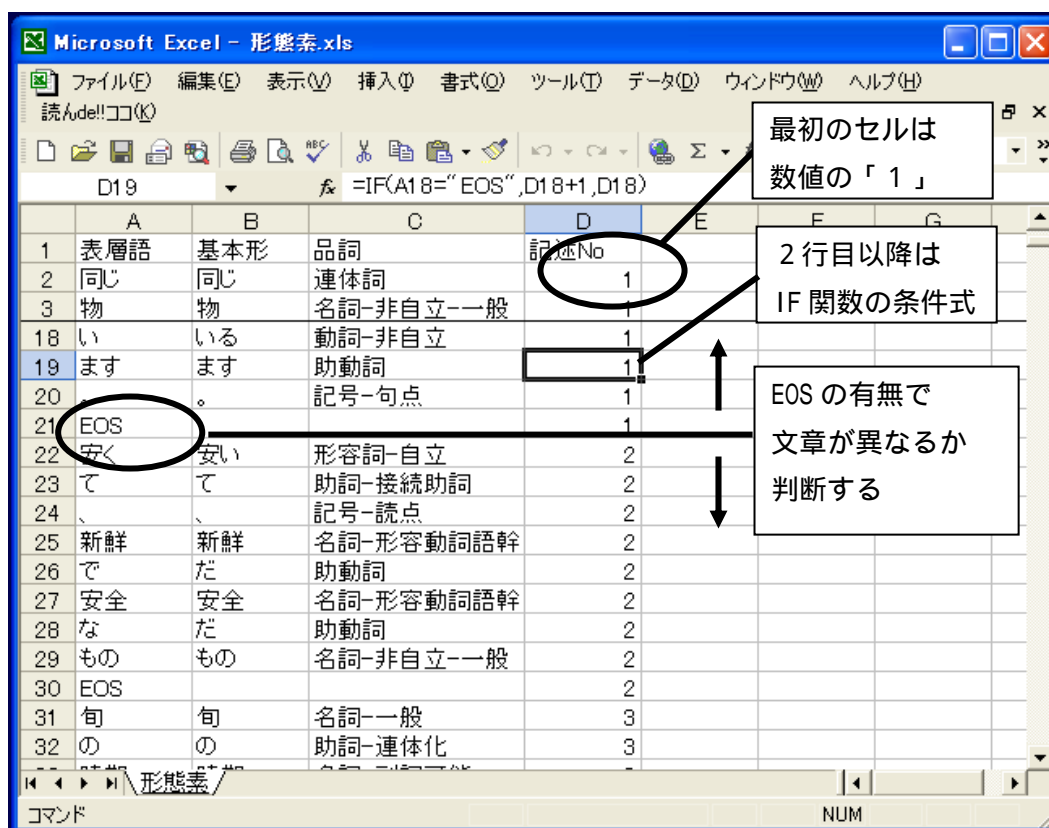
「茶釜」で分解された形態素は、表層語として縦1列に連なっており、これを上から読んでいくと元の文章と等しくなる。そして、1つの文章が終了すると、その末尾に「EOS」と表記される。従って、文章の区切りは、この「EOS」を用いることで判断することができる（画面5）。

文章番号の付与は「IF関数」を利用する（画面6）。まず、1行目のセル（D1）には、文章番号を示す列であることがわかる名前をつける（事例では「記述No」）。次の行、すなわち最初の形態素の行のセル（D2）には、数値の「1」を入力する。そして次の行のセル（D3）には、「もし上の行の表層語がEOSならば上の行の数値に1を加える。EOSでなければ上の行の数値と等しい」という条件式を入れる。[ D3の場合：IF(A2="EOS", D2+1, D2) ]

この条件式を最後の行までコピーする。ここで、最後の数値と文章の総数とが一致しているかを必ず確認する。もし、総文章数よりも大きい数値になっている場合には、「半角カタカナを使ったことによるバグ」などで、途中で不要な「EOS」が紛れている可能性がある。



画面5 文章番号の付与



画面6 文章番号の付与

### (3) 品詞情報を利用してキーワード候補の抽出を行う

文章を形態素に分解すると、膨大な数の形態素が検出される。これらの中には、句読点や括弧などの記号や、「が」「を」などの助詞、「です」「ます」などの助動詞など、キーワードとして意味を成さない語も多く含まれている。キーワードの抽出に際してこれらの語を1つ1つ検証するには大変な労力を要するので、不必要な語はある程度機械的に切り捨ててしまった方が効率的である。そこで、形態素の品詞情報を利用してキーワードの候補になりそうな語を抽出することにする。

まず、条件式で付与した文章番号が作業中に変化しないように、「数式」から「値」に変換しておく。画面7では、シートをコピーして新たに設け<sup>注(6)</sup>、「記述 No」の列を選択してコピー、「形式を選択して貼り付け(S)」の「値(V)」にチェックして上書きしている。

続いて、オートフィルタ機能を利用する(画面8)。データのある適当なセルをクリックし、「データ(D)」-「フィルタ(F)」-「オートフィルタ(F)」コマンドを選択する。1行目の各セルに「フィルタ矢印」がつくので、「品詞」の列の「フィルタ矢印」をクリックする。

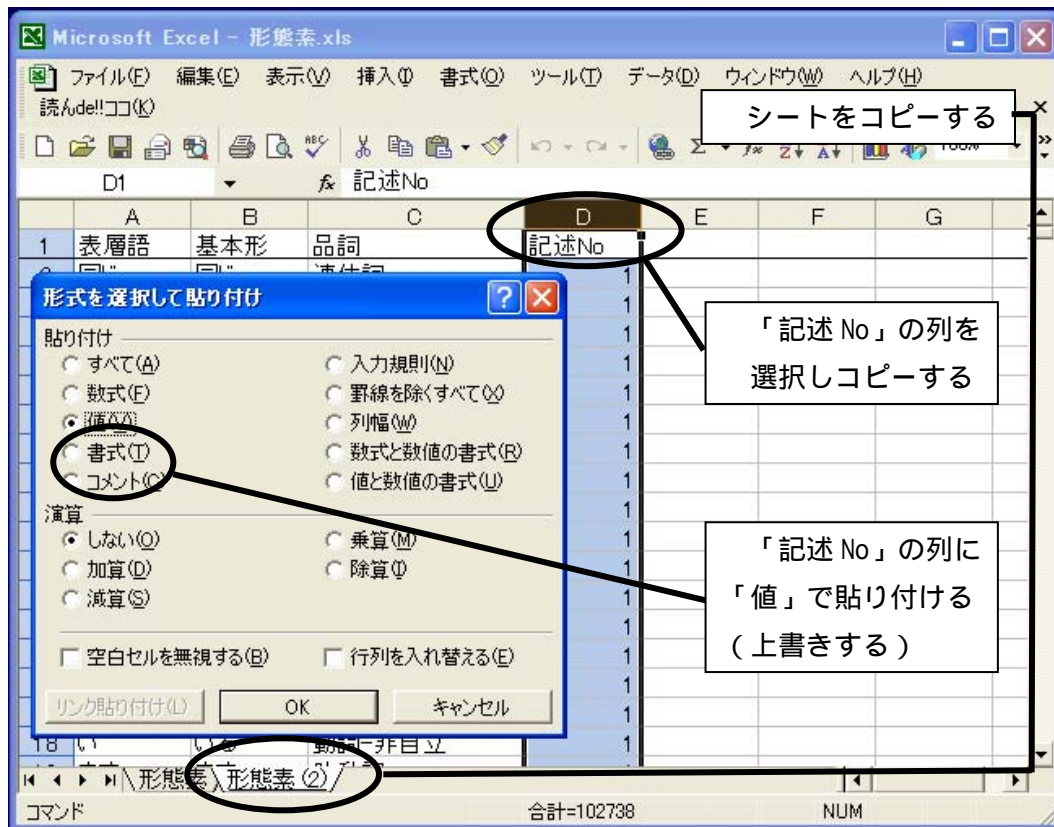
リストには様々な品詞名が表示されるが、その中からキーワードになりそうな語を多く含んでいる品詞を選択する。ここでは、「形容詞 - 自立」「動詞 - 自立」「名詞 - サ変接続」「名詞 - 一般」「名詞 - 形容動詞語幹」を選んでいるが、データの内容や分析の目的によっては他の品詞が必要になることもあるので、個別に判断する<sup>注(7)</sup>。未知語の利用も考慮する必要がある。

ここで新たに「抽出語」シートを設ける。オートフィルタでキーワードとして使いたい品詞を表示し、全ての行を選択してコピーし、「抽出語」シートに貼り付ける(画面9、画面10)。その際、1つ目の品詞(例では「形容詞 - 自立」)は1行目(「表層語」「基本形」などの見出しの行)も含めてコピー&貼り付けをし、2番目の品詞からは、上の品詞に続けて形態素の行を貼り付けていく。

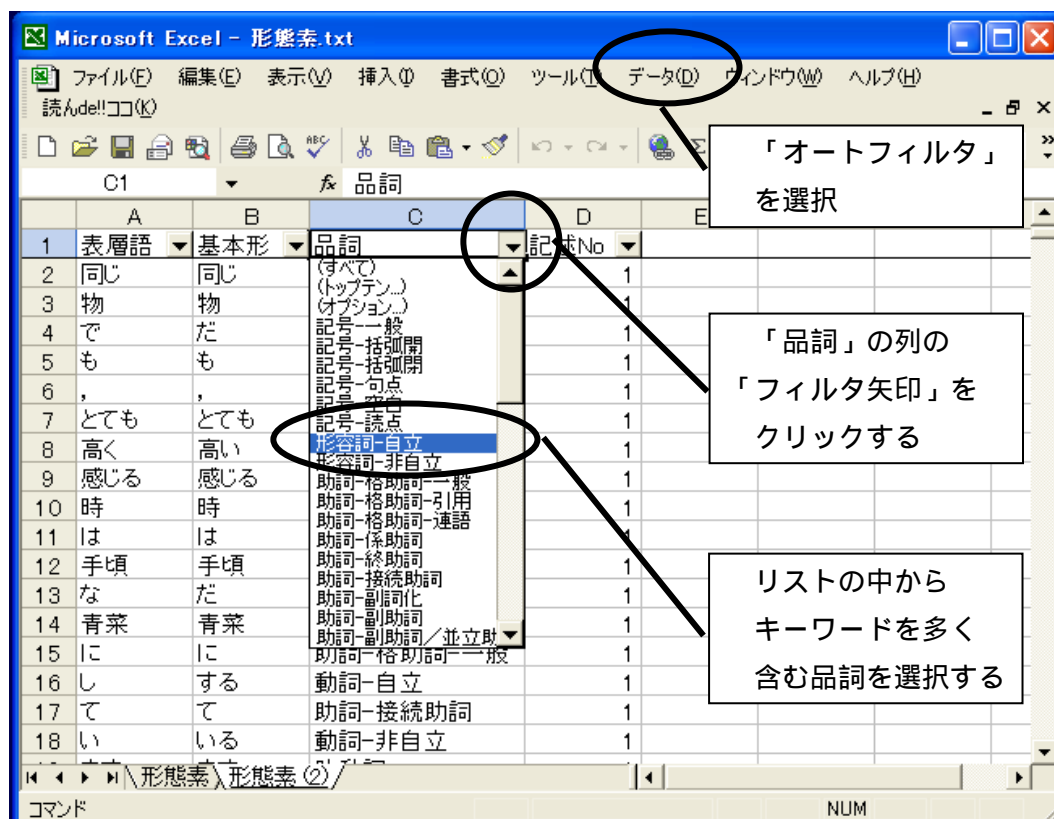
注(6)以降、作業途中もしくはファイル完成後にファイルを保存する際には、以下の点に注意されたい。

テキストファイルを読み込み、新たなシートを設けた後でそのまま保存しようとする「選択したファイルの種類は複数のシートを含むブックをサポートしていません」というエラーメッセージが出される。その場合、「ファイルの種類(T)」において「Microsoft Excel ブック(\*.xls)」を選択し、Excel形式で保存すること。

注(7)林(2002)は、「品詞ごとにキーワードとして適切かどうか」を一覧表にして示している。

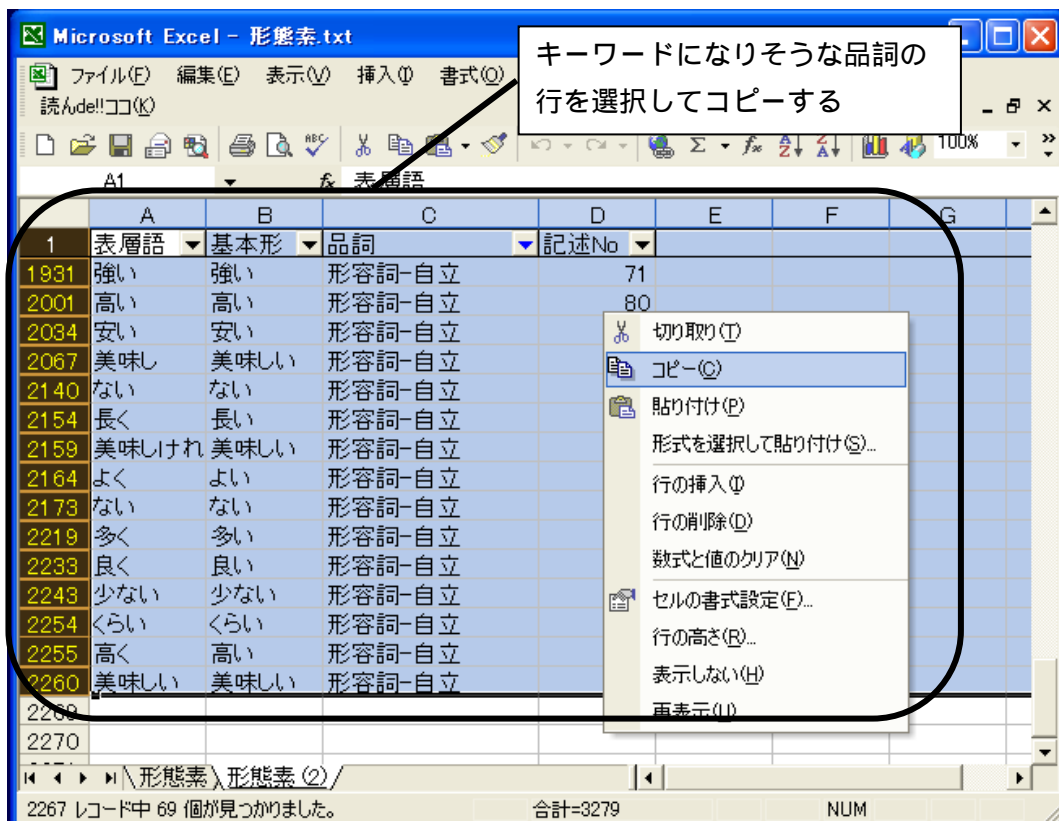


画面7 キーワード候補の抽出

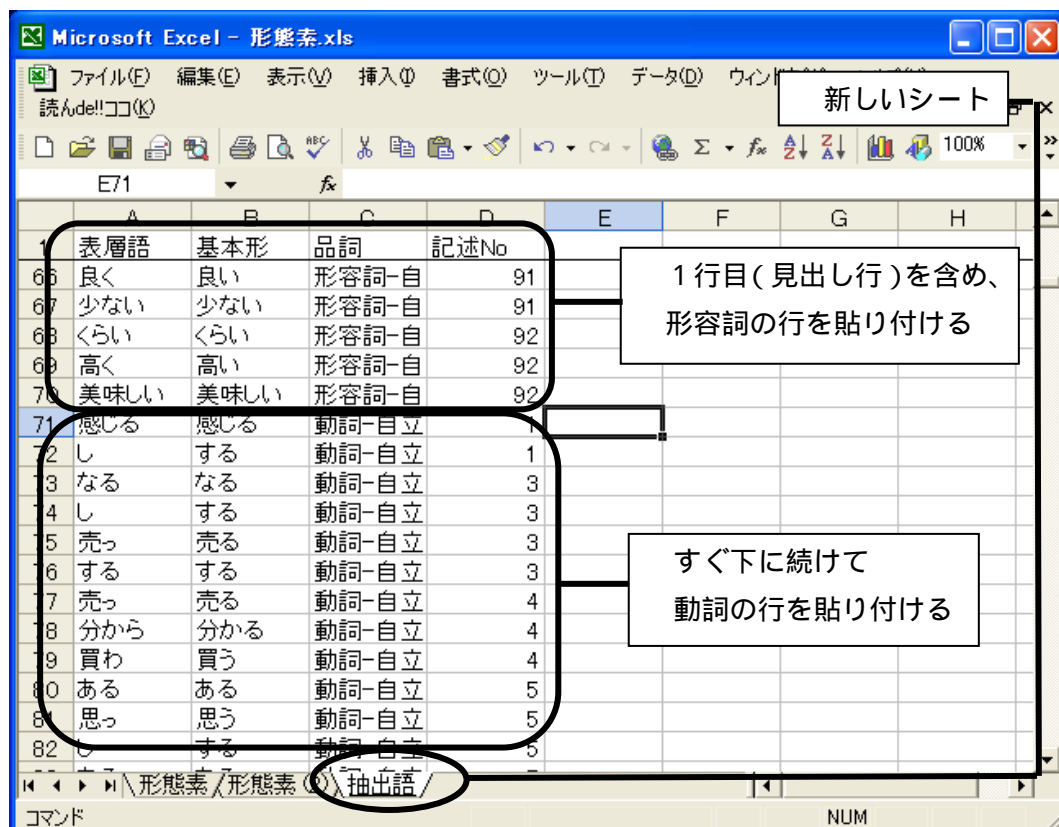


画面8 キーワード候補の抽出





画面9 キーワード候補の抽出



画面10 キーワード候補の抽出

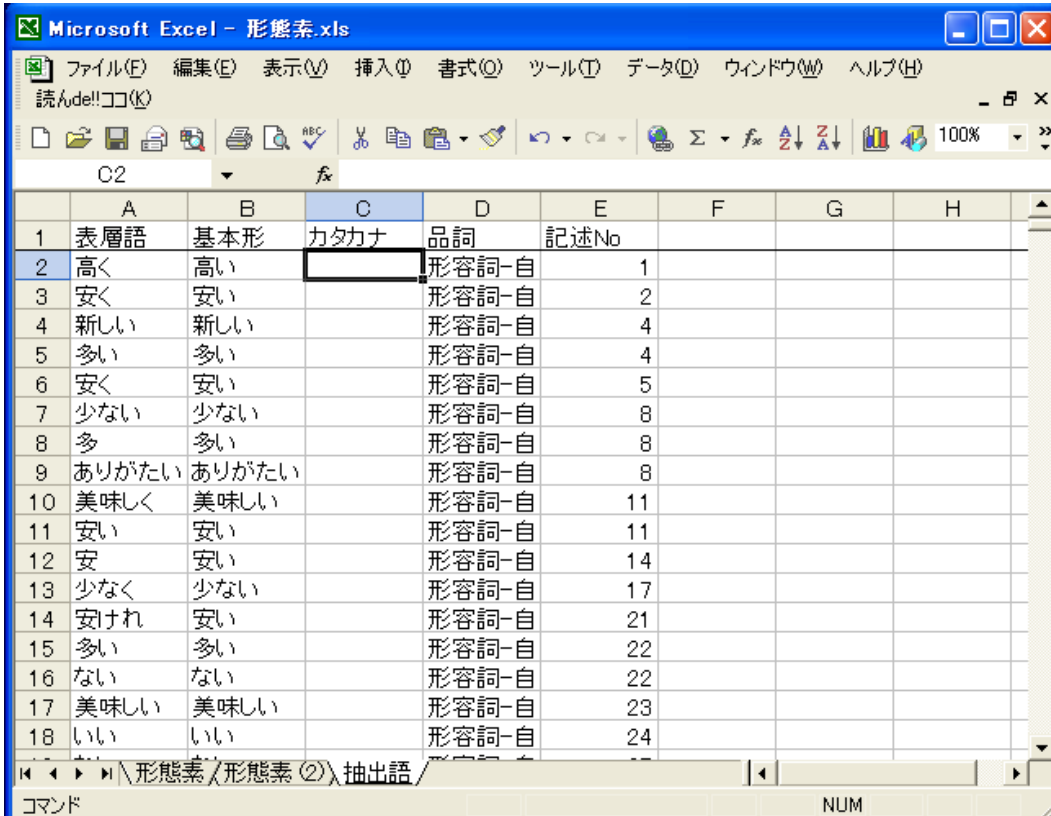


#### (4) 基本形の「ひらがな」を「カタカナ」に変換する

ここで、基本形のひらがな表記をいったん全てカタカナ表記に変換する。これは、形態素の基本形が何種類あるかを数える過程で、同じ語でもひらがな表記の語とカタカナ表記の語は、別の語としてカウントしてしまうからである。従って、もとの文章データの段階でそうした表記の揺れを修正している、または「ひらがな語」と「カタカナ語」は区別してカウントしたいという場合には、この過程は無視してよい。ただし、後ほど行う「1・0データ化」の過程で、「記述No」は5列目、「形態素No」は6列目と定義しているため、マクロを実行する際には、1列挿入するか、もしくはマクロの一部を書き換える必要がある。

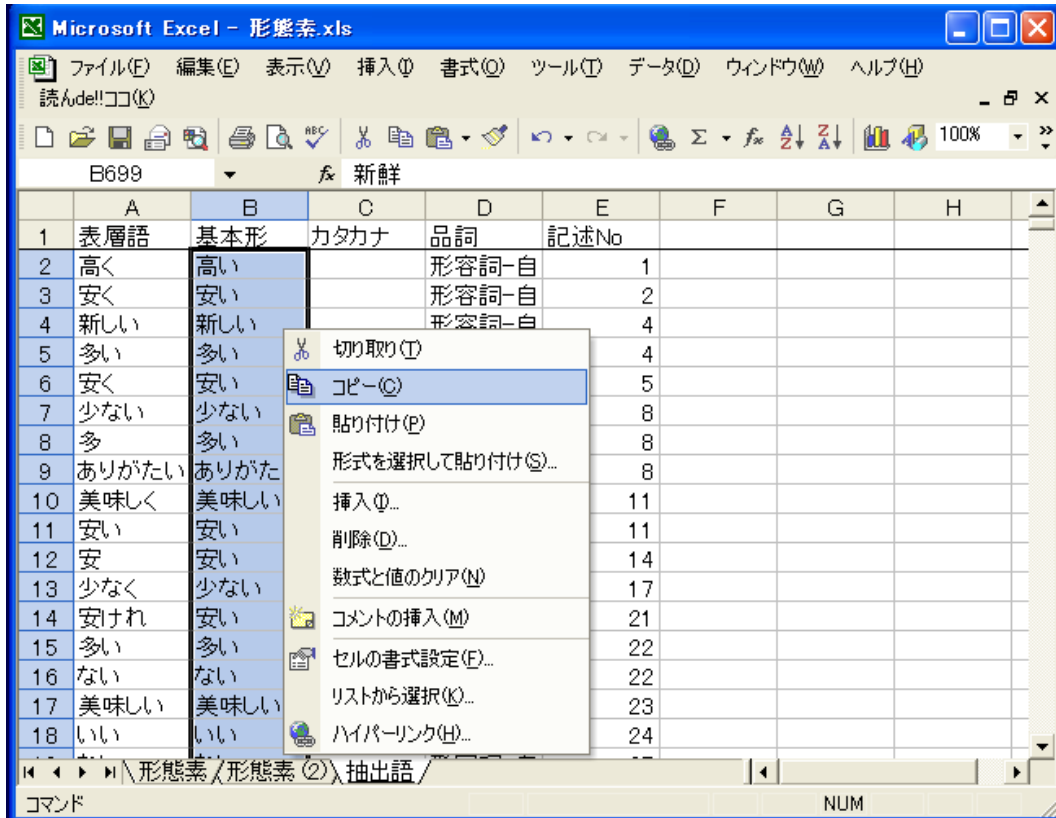
まず、B列とC列の間に1列挿入し、表頭に「カタカナ」と入力する(画面11)。「基本形」の列の語を全て選択してコピーし(画面12) Word を立ち上げ、「形式を選択して貼り付け(S)」 - 「テキスト」で貼り付ける(画面13)。

Word 上で、「編集(E)」 - 「すべて選択(L)」で文字をすべて選択し、「書式(O)」 - 「文字種の変換(E)」 - 「カタカナ(K)」にチェックを入れて「OK」をクリックする(画面14)。カタカナに変換した語を再びコピーして Excel の「カタカナ」列に「テキスト」で貼り付けをする(画面15)。

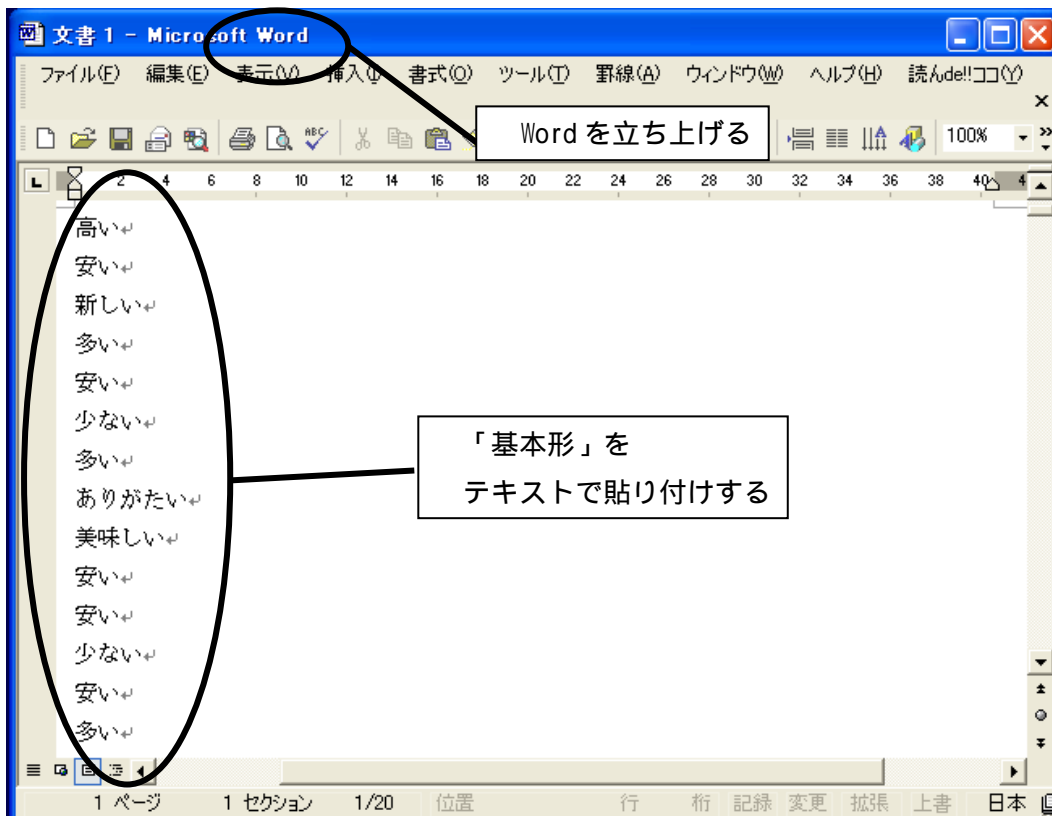


	A	B	C	D	E	F	G	H
1	表層語	基本形	カタカナ	品詞	記述No			
2	高く	高い		形容詞-自	1			
3	安く	安い		形容詞-自	2			
4	新しい	新しい		形容詞-自	4			
5	多い	多い		形容詞-自	4			
6	安く	安い		形容詞-自	5			
7	少ない	少ない		形容詞-自	8			
8	多	多い		形容詞-自	8			
9	ありがたい	ありがたい		形容詞-自	8			
10	美味しく	美味しい		形容詞-自	11			
11	安い	安い		形容詞-自	11			
12	安	安い		形容詞-自	14			
13	少なく	少ない		形容詞-自	17			
14	安けれ	安い		形容詞-自	21			
15	多い	多い		形容詞-自	22			
16	ない	ない		形容詞-自	22			
17	美味しい	美味しい		形容詞-自	23			
18	いい	いい		形容詞-自	24			

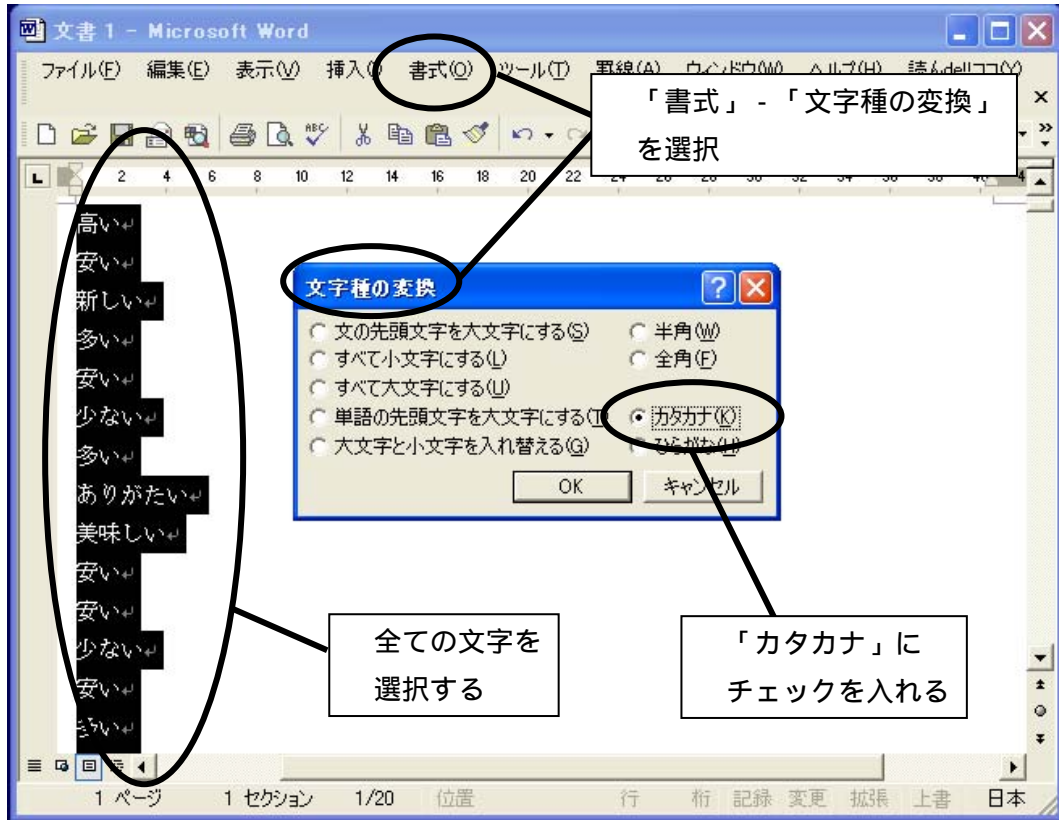
画面11 「基本形」のカタカナ変換



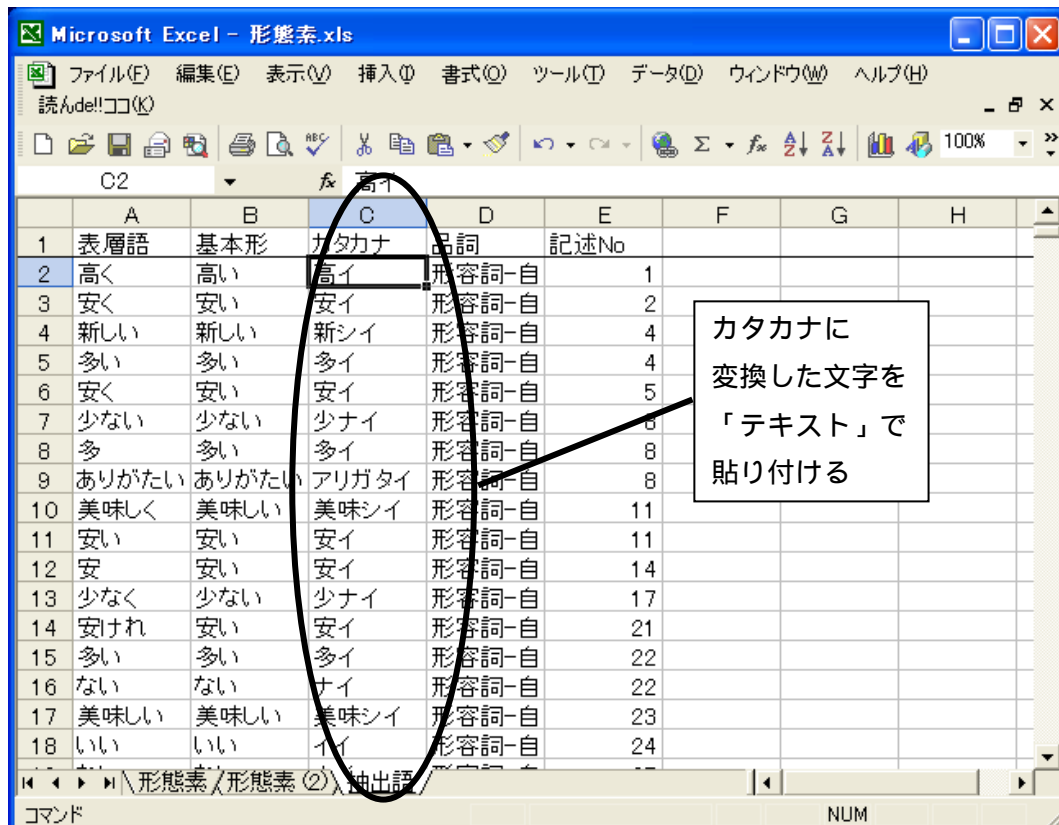
画面 12 「基本形」のカタカナ変換



図形 13 「基本形」のカタカナ変換



画面 14 「基本形」のカタカナ変換



画面 15 「基本形」のカタカナ変換

## (5) 形態素番号と出現数を割り当てる

品詞情報を用いて抽出した語には、基本形の等しいものが多数存在する。ここでは、この中に何種類の形態素(基本形)が存在するかを明らかにし、後の「1・0データ化」の過程で必要となる形態素番号(以降、説明画面では「形態素 No」と表示)の割り当てを行う。また、それらの出現数の算出も行う。

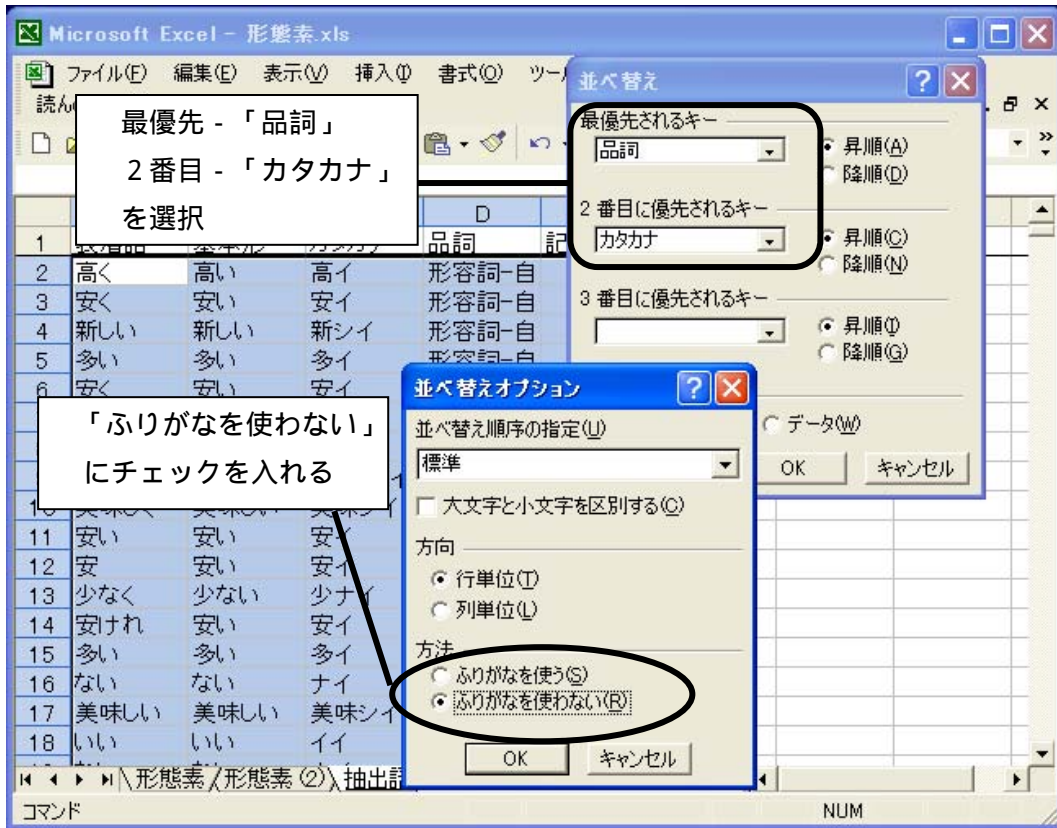
まず、適当なセルを選択して、「データ(D)」 - 「並び替え(S)」を表示する(画面 16)。「最優先されるキー」に「品詞」を、「2番目に優先されるキー」に「カタカナ」を選択する。また、「オプション(O)」をクリックして、「方法」 - 「ふりがなを使わない(R)」がチェックされていることを確認してから、並び替えを実行する。

次に、F列、G列、H列の1行目に、それぞれ「形態素 No」「出現数」「出現数 2」と入力する(画面 17)。

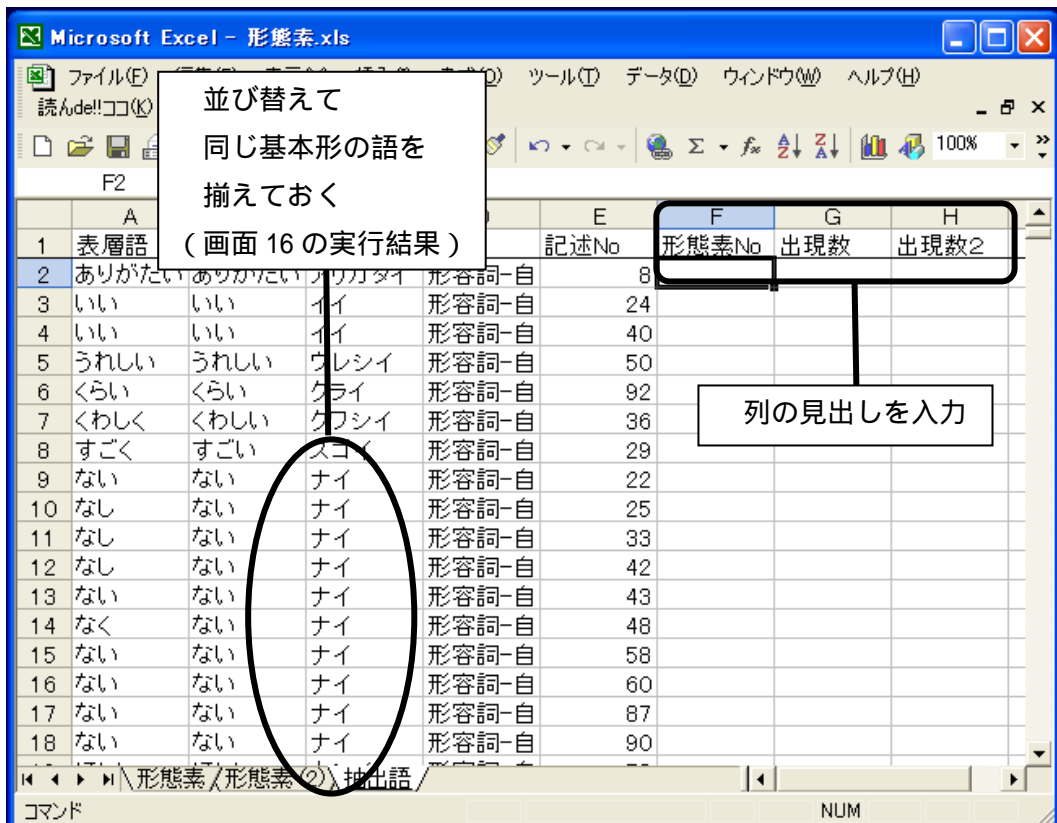
「形態素 No」の入力は、最初の形態素の行のセル(F2)には数値の「1」を入力する(画面 18)。次の行のセル(F3)には、「もし『カタカナ列』のセル(C3)の文字が、上のセル(C2)の文字と同じであれば、上のセル(F2)と同じ数値にする。異なる場合には上のセルの数値に1を加える」という条件式を入れる[F3の場合: IF(C3=C2、F2、F2+1)]。残りの行にはこの関数をコピー&貼り付けする。

次に、「出現数」の入力を行う(画面 19)。最初の形態素の行のセル(G2)は、同様に「1」を入力する。次の行のセル(G3)には、「もし『形態素 No』列のセル(F3)の数値が、上のセル(F2)の数値と同じであれば、上のセル(G2)の数値に1を加える。異なる場合には、数値の1を入力する」という条件式を入れる[G3の場合: IF(F3=F2、G2+1、1)]。残りの行にはこの関数をコピー&貼り付けする。

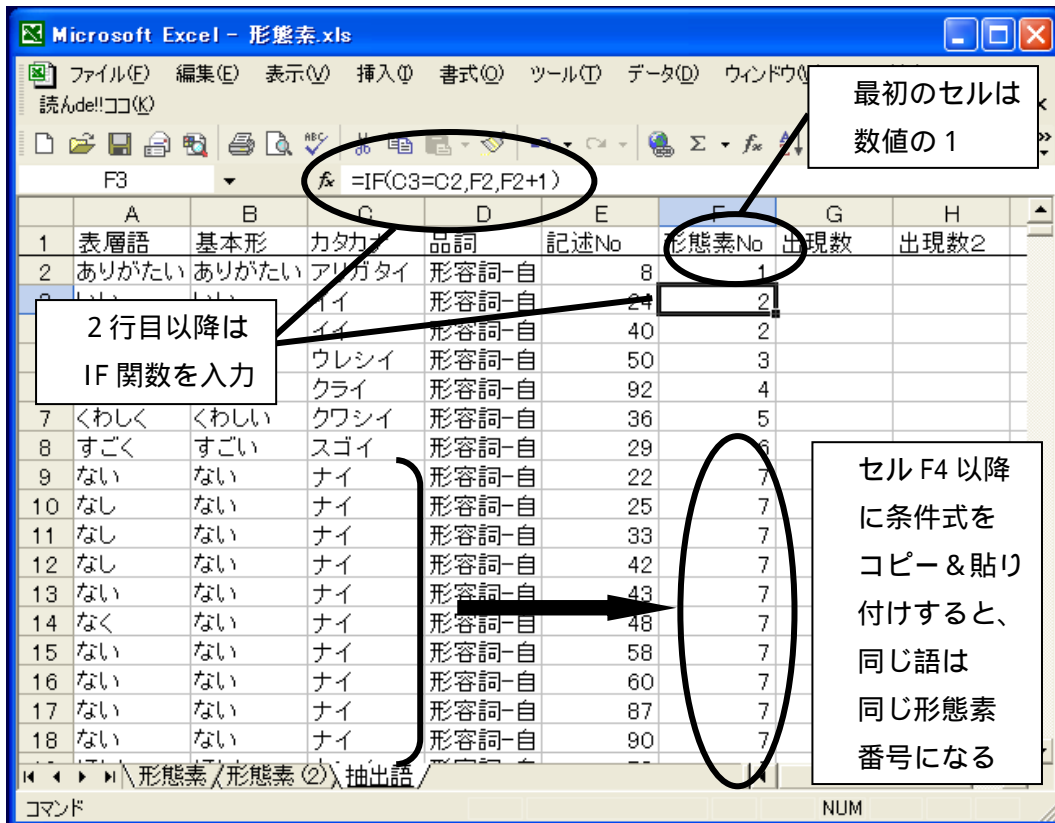
続いて、「出現数 2」であるが(画面 20)、これは、「出現数」の入力と同じ作業を下(最後の行)から行っている。これにより、「出現数」列で数値が1の行は、「出現数 2」列でその形態素の総出現数(以下、出現頻度)を示すことになるからである。まず、最後の行のセル(セル番号はデータにより異なる。事例では H699)に数値の「1」を入力する。続いて、下から2行目のセル(事例では H698)に、「もし『形態素 No』列のセル(F698)の数値が、下のセル(F699)の数値と同じであれば、下のセル(H699)の数値に1を加える。異なる場合には、数値の1を入力する」という条件式を入れる[H698の場合: IF(F698=F699、H699+1、1)]。そして、それより上の行にはこの関数をコピー&貼り付けする。



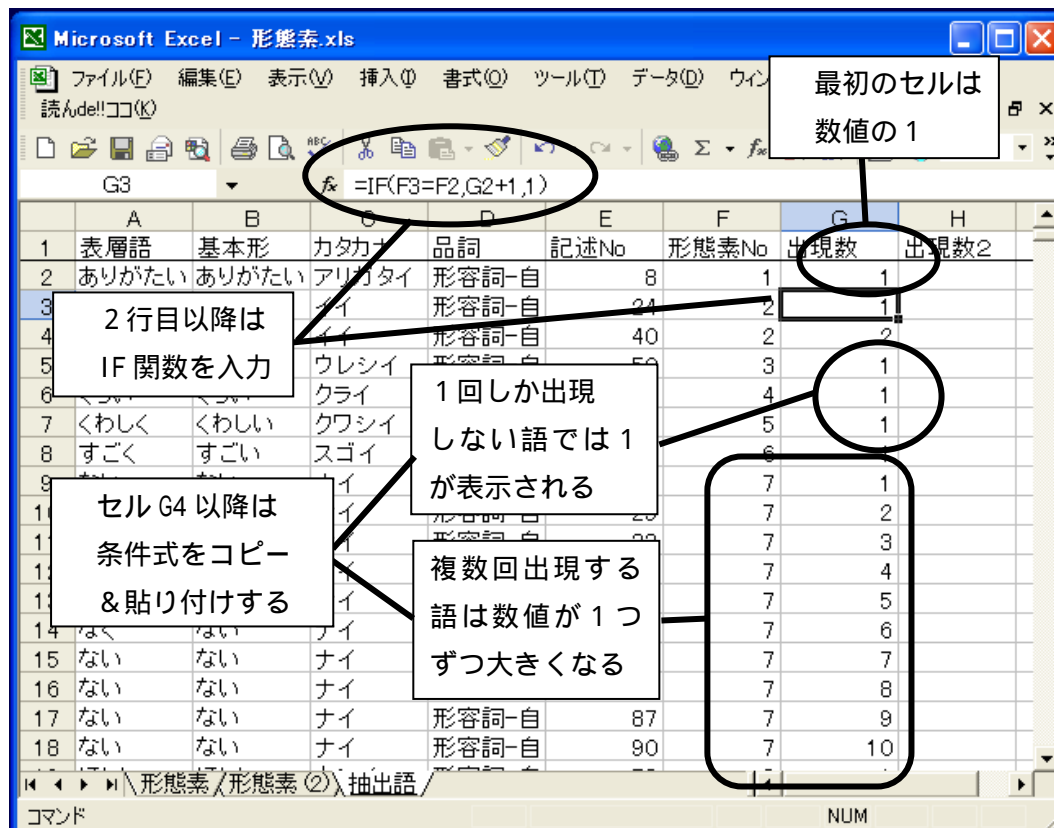
画面 16 形態素番号と出現数入力



画面 17 形態素番号と出現数の入力

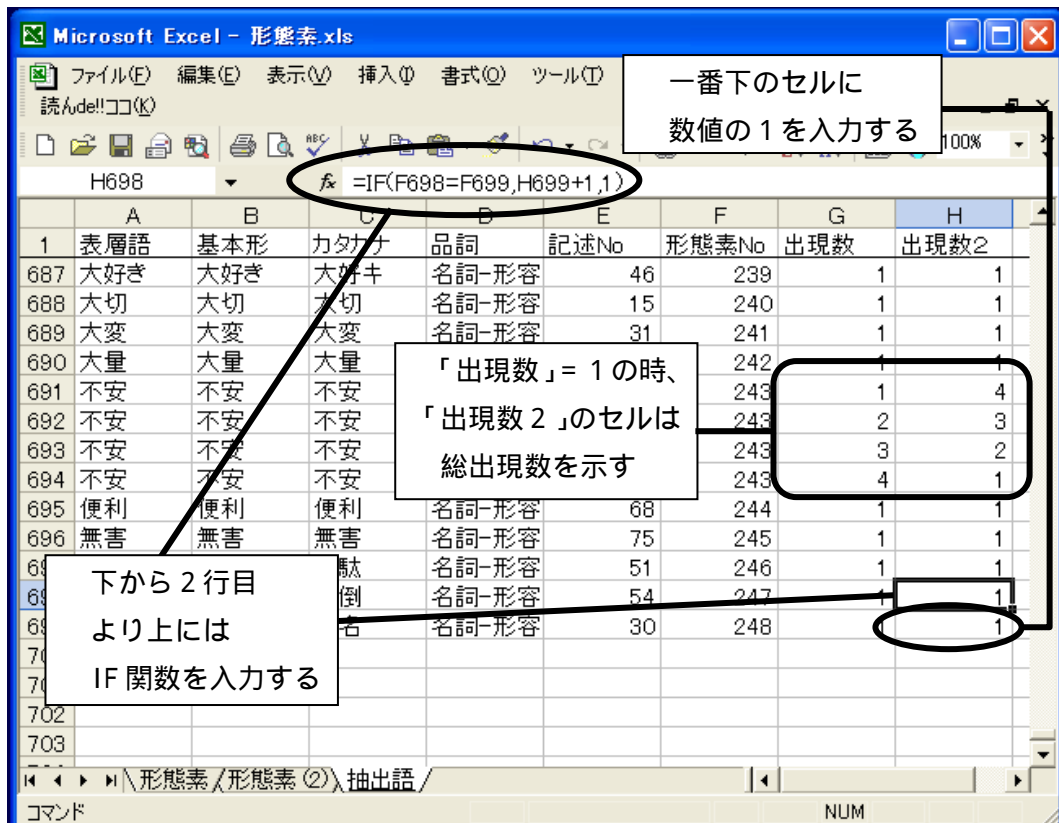


画面 18 形態素番号と出現数の入力



画面 19 形態素番号と出現数の入力





画面 20 形態素番号と出現数の入力

(6) 出現数からキーワードを絞り込む

「形態素 No」「出現数」「出現数 2」の入力が全て終了したら、並び替えなどで数値が変化しないように、新しいシート（シート名「data」）にコピー & 「形式を選択して貼り付け(S)」 - 「値(V)」で貼り付けをする（画面 21、画面 22）。なお、ここで指定したシート名「data」は、後で用いるマクロに組み込まれているので、他の名前にする場合にはマクロの書き換えが必要である。

抽出語データを「data」シートに貼り付けたら、「データ(D)」 - 「フィルタ(F)」 - 「オートフィルタ(F)」を選択して、1 行目にフィルタ矢印をつける。次に、「出現数」列のフィルタ矢印をクリックして、1 を選択する（画面 23）。

出現数が 1 の行だけを表示させたら、新たに設けた「出現数」シートにコピーして貼り付ける（画面 24）。前述したように、「出現数」列の数値が 1 の行は、「出現数 2」列にその抽出語の出現頻度が示されているので、「出現数」シートには各形態素の出現頻度を示した行だけが抽出されたことになる。

ここで、「データ(D)」 - 「並び替え(S)」を選択し（画面 25）、「出現数 2」列の値が大きい順に並び替える。「出現数 2」列は各抽出語の出現頻度を示しているため、より多く出現する語から順番に示される（画面 26）。

ここから、一定以上の出現頻度がある語をキーワードとして抽出する。これは、例えば出現頻度が 1 というのは、たった 1 回しか記述がなかった言葉であり、全体の集計をする上でこれを取

り上げる意義は少ないと考えるからである。そこで、「最低何回出現したか」という基準でキーワードを絞り込むことにする。

この最低ラインについては、データのサイズや求める分析の精度によって異なる。データの量が少ない場合に出現頻度の下限を高く設定すると、一般的な語（例えば、青果物に関する記述で「野菜」と「購入」だけ）しか拾えない場合がある。また、なるべく多くキーワードを用意した方が、類義語をまとめるなど分析に幅を持たせることができる。一方、大量のデータがあるのに下限を低く設定すると、膨大な量のキーワードの扱いに苦慮することになる。また、出現頻度の低い語を統計的に扱うことの意義についても問われることになるであろう。

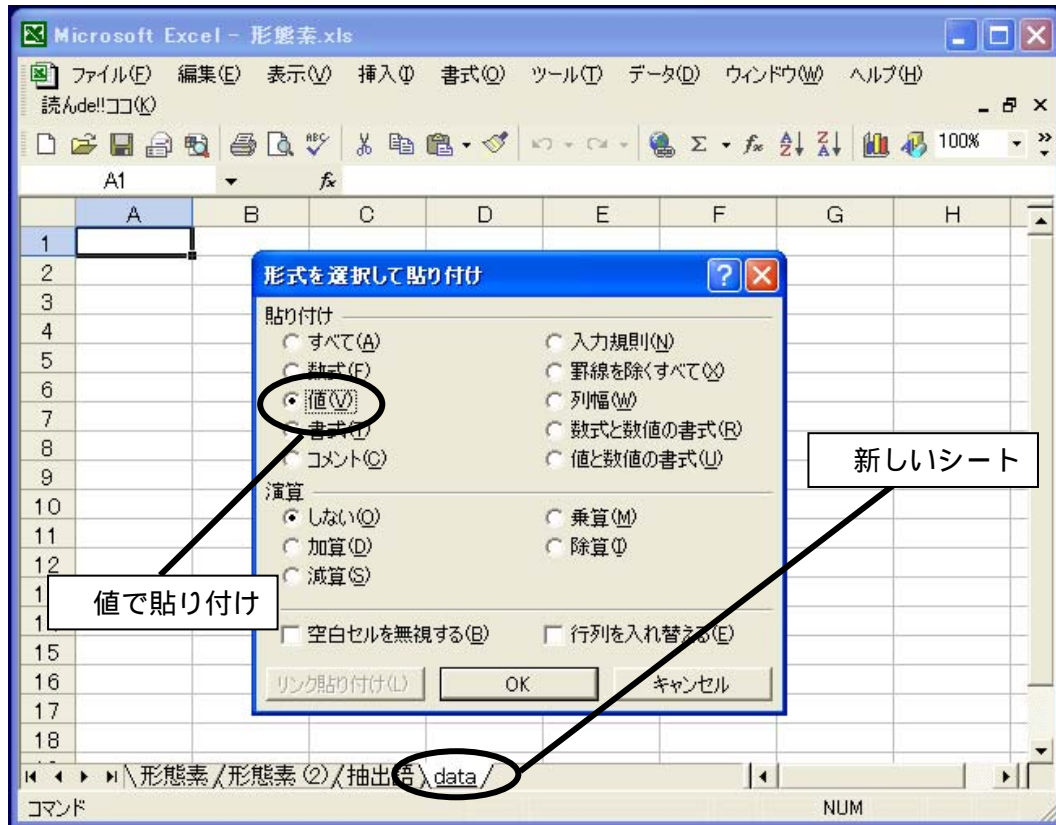
ただし、本手法では「1・0データ化」の過程で、Excelの各列にキーワードを割り振る。すなわち、1回の作業で扱えるキーワード数は「Excelの列数の上限(256)」-「文章番号を示す列(1)」=255個である。これに回答者属性などのデータを加えることを考えると、扱いやすいキーワード数の上限は250個程度であるといえる。もちろん、システム上の上限であるため、250が適当なキーワード数であるという根拠はない。また、データを分割していくつかシートを作成すれば、255を上回るキーワードを抽出することも可能である。しかし、適当な出現頻度の下限を設定できない場合には、1つの目安として、キーワード数が250を超えない程度に抑えられる出現頻度を挙げることができる。

ここで示す事例は説明用の簡易データなので、出現頻度3という低い数値で抽出している(画面27)。「出現数」シートの「出現数2」列の数値が3以上の行をコピーし、新たに設けたシート(事例ではシート名「3以上」)に貼り付ける。このシートにある語が、1・0データ化を行うキーワードとなる。

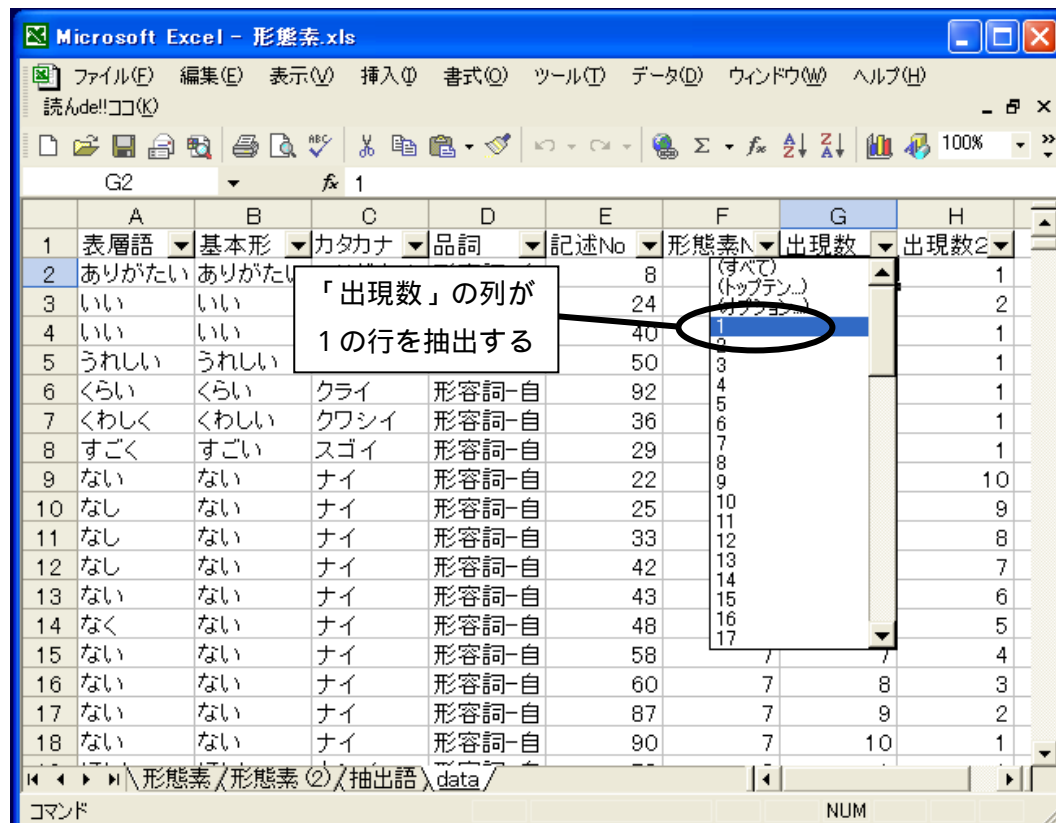
	A	B	C	D	E	F	G	H
	表層語	基本形	カタカナ	品詞	記述No	形態素No	出現数	出現数2
1	ありがたい	ありがたい	アリガタイ	形容詞-自	8	1	1	1
2	いい	いい	イイ	形容詞-自	24	2	1	2
3	いい	いい	イイ	形容詞-自	40	2	2	1
4	うれしい	うれしい	ウレシイ	形容詞-自	50	3	1	1
5	くらい	くらい	クライ	形容詞-自	92	4	1	1
6	くわしく	くわしく	クワシク	形容詞-自	36	5	1	1
7	すごく	すごく	シゴク	形容詞-自	29	6	1	1
8	ない	ない	ナイ	形容詞-自	22	7	1	10
9	なし	なし	ナシ	形容詞-自	25	7	2	9
10	なし	なし	ナシ	形容詞-自	33	7	3	8
11	なし	なし	ナシ	形容詞-自	42	7	4	7
12	なし	なし	ナシ	形容詞-自	43	7	5	6
13	なく	なく	ナク	形容詞-自	48	7	6	5
14	ない	ない	ナイ	形容詞-自	58	7	7	4
15	ない	ない	ナイ	形容詞-自	60	7	8	3
16	ない	ない	ナイ	形容詞-自	87	7	9	2
17	ない	ない	ナイ	形容詞-自	90	7	10	1

画面 21 キーワードの絞り込み

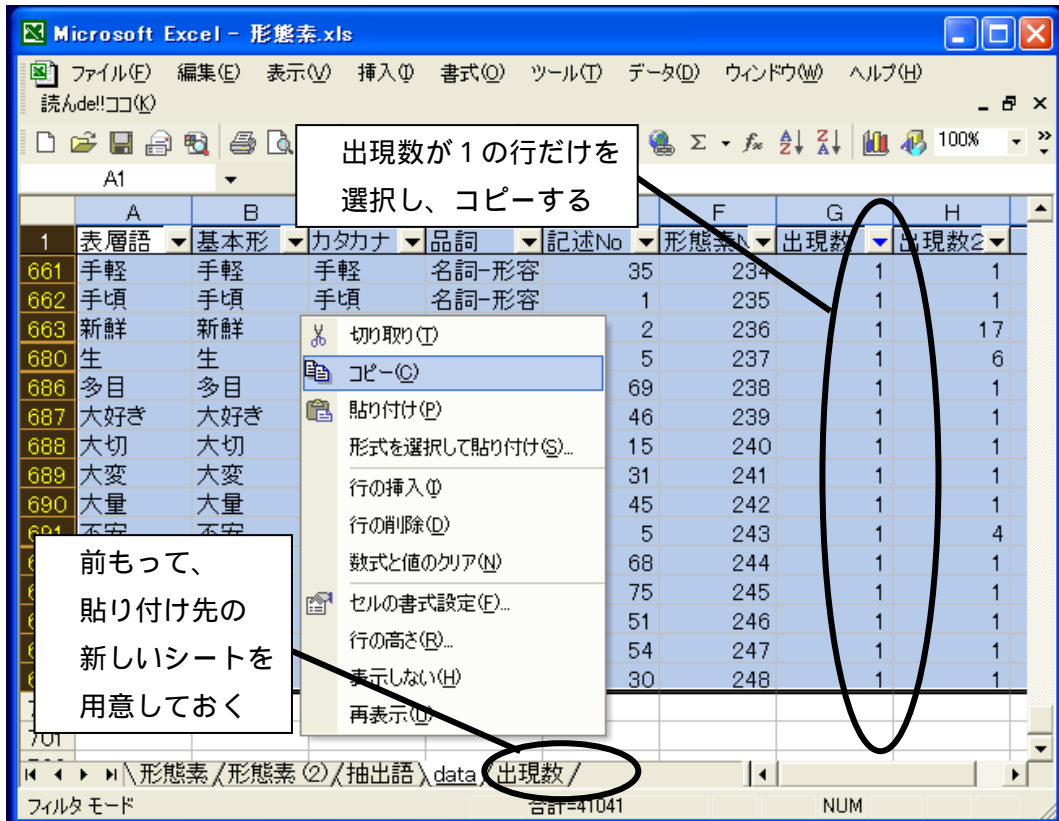




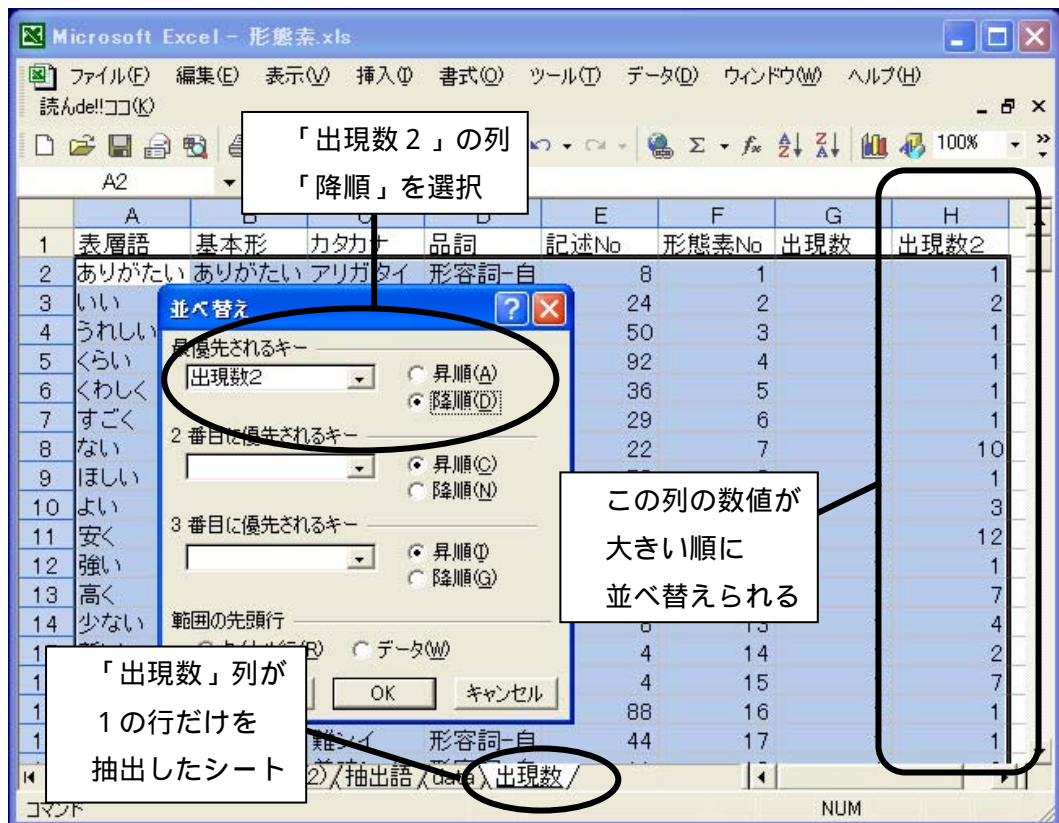
画面 22 キーワードの絞り込み



画面 23 キーワードの絞り込み



画面 24 キーワードの絞り込み



画面 25 キーワードの絞り込み

Microsoft Excel - 形態素.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

読んdel!!ココ(K)

H2 70

	A	B	C	D	E	F	G	H
1	表層語	基本形	カタカナ	品詞	記述No	形態素No	出現数	出現数2
2	し	する	スル	動詞-自立	1	29	1	70
3	購入	購入	購入	名詞-サ変	3	96	1	25
4	思っ	思う	思ウ	動詞-自立	5	62	1	21
5	買わ	買う	買フ	動詞-自立	4	79	1	21
6	ある	ある	アル	動詞-自立	5	21	1	18
7	農薬	農薬	農薬	名詞-一般	5	214	1	18
8	新鮮	新鮮	新鮮	名詞-形容	2	236	1	17
9	食べる	食べる	食ベル	動詞-自立	9	69	1	16
10	青菜	青菜	青菜	名詞-一般	1	189	1	15
11	なる	なる	ナル	動詞-自立	3	35	1	13
12	安く	安い	安い	形容詞-自	2	10	1	12
13	ない	ない	ナイ	形容詞-自	22	7	1	10
14	分から	分かる	分カル			83	1	10
15	野菜	野菜	野菜			223	1	10
16	美味しく	美味しい	美味シイ			18	1	9
17	虫	虫	虫			204	1	9
18	高く	高い	高イ	形容詞-自	1	12	1	7

コマンド NUM

各抽出語の出現頻度がわかる

画面 26 キーワードの絞り込み

Microsoft Excel - 形態素.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

読んdel!!ココ(K)

A1

	A	B	G	H
47	知り	知る	76	3
48	入っ	入る	77	3
49	売っ	売る	81	3
50	生産	生産	56	3
51	調理	調理	20	3
52	減	減	27	3
53	子供	子供	15	3
54	手	手	31	3
55	旬	旬	3	3
56	体	体	40	3
57	袋	袋	10	3
58	店頭	店頭	39	3
59	日持ち	日持ち	45	3
60	いい	いい	24	2
61	新しい	新しい	4	2
62	良い	良い	70	2
63	冷たく	冷たい	31	2
64	かかる	かかる	22	2

コマンド 合計=8616 NUM

出現頻度がいくつ以上の語をキーワードとするか決めてコピーする

貼り付け先の新しいシートを用意しておく

3以上

画面 27 キーワードの絞り込み

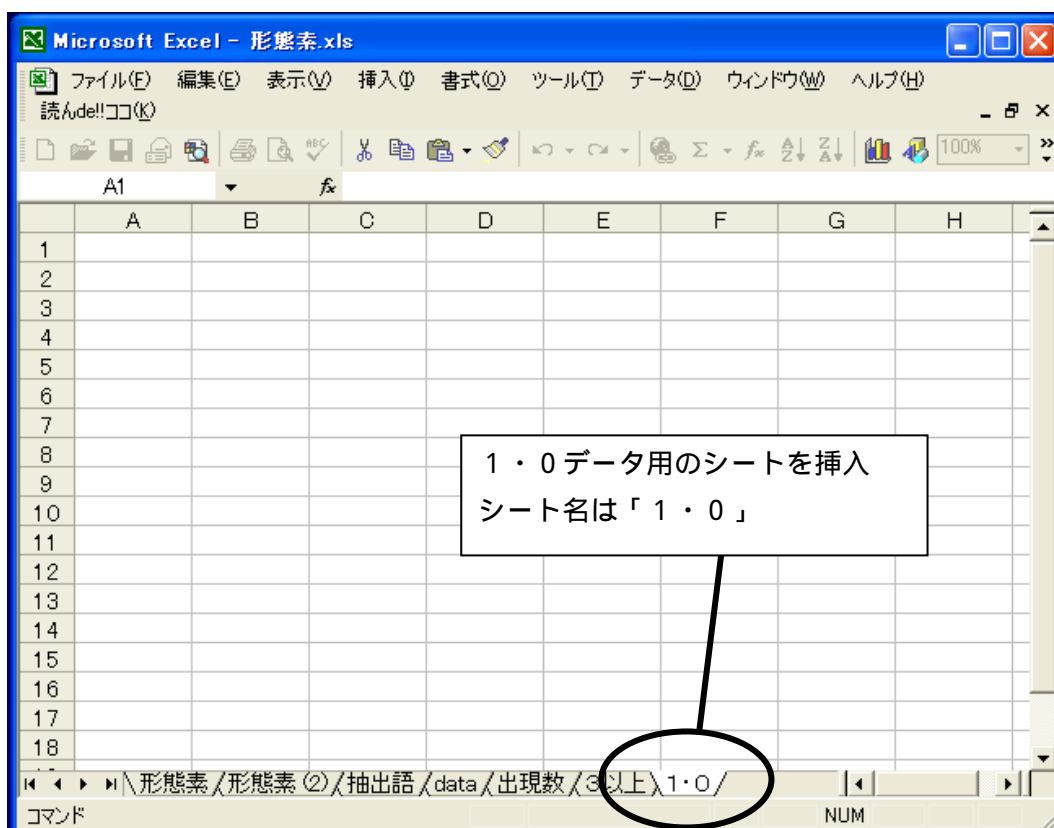
### (7)「1・0データ化」の準備

まず、1・0データ用のシートを新たに設ける(画面28)。シート名は「1・0」(全て全角)である。このシート名は「data」シートと同様、マクロに組み込まれているので、他の名前にする場合には、マクロの書き換えが必要である。

次に、(6)で絞り込んだキーワードのシート(この事例ではシート名「3以上」)に戻り、「形態素 No」の数値が小さい順に並べ替えをする(画面29)。これが、キーワードとして扱う形態素の番号である。この番号を選択し(画面30)、「1・0」シートの1行目B列(セル「B1」)から横に貼り付けをする(画面31、32)。さらに、セル「A2」を選択し(画面33)縦に1から全文章数(事例では文章数92)までの文章番号を入力する。

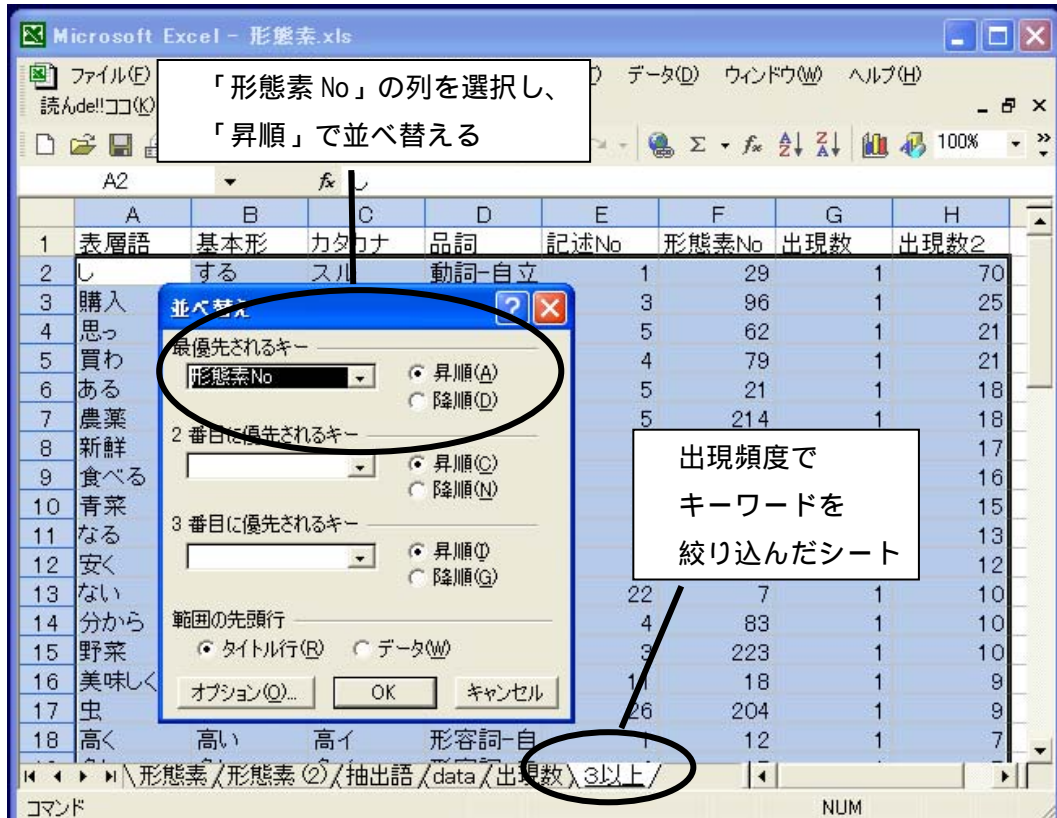
ここで、「data」シートに戻る(画面34)。出現数によるキーワード絞り込みの過程でオートフィルタ機能を使用したままの状態にあるので、「出現数」列のフィルタ矢印をクリックし、「すべて」を選択して全てのデータを表示させる。

次に、「記述 No」と「形態素 No」で並べ替えをする(画面35)。「データ(D)」-「並べ替え(S)」を選択し、「最優先されるキー」として「記述 No」を、「2番目に優先されるキー」として「形態素 No」を選択し、どちらも「昇順」にチェックを入れる。これにより、文章番号の小さい順に並び、同じ文章番号の場合には形態素番号の小さい順に並ぶことになる。

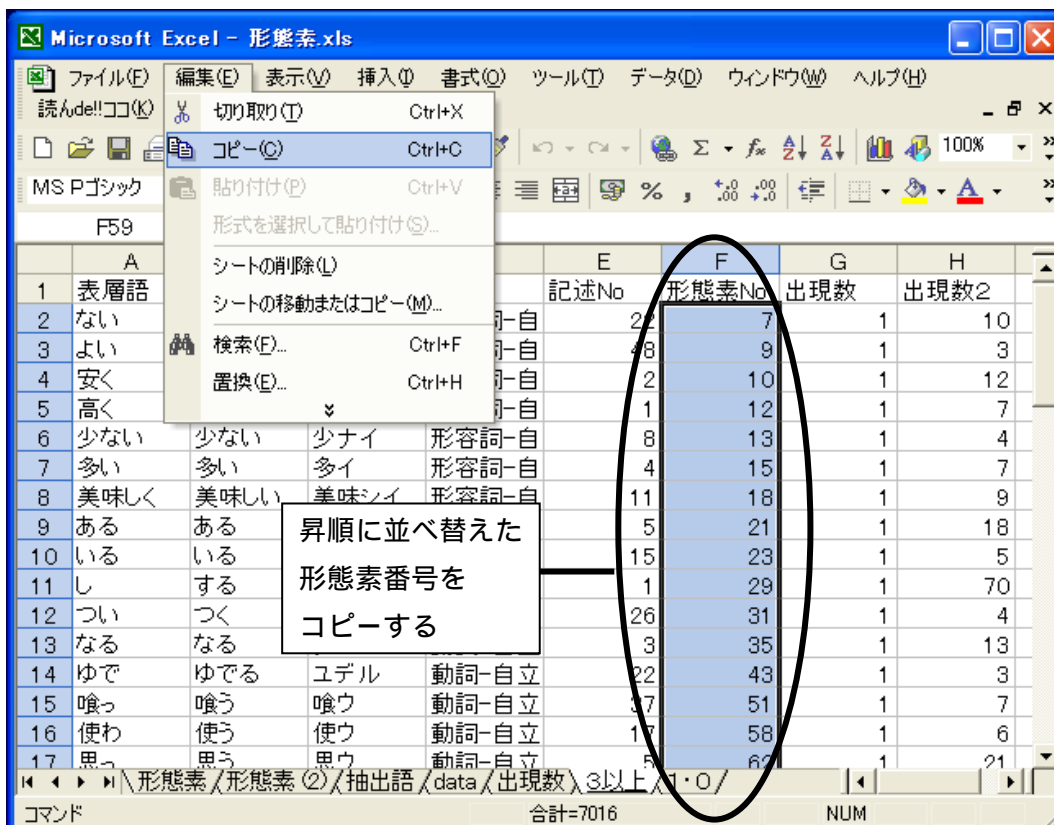


画面28 「1・0データ化」の準備

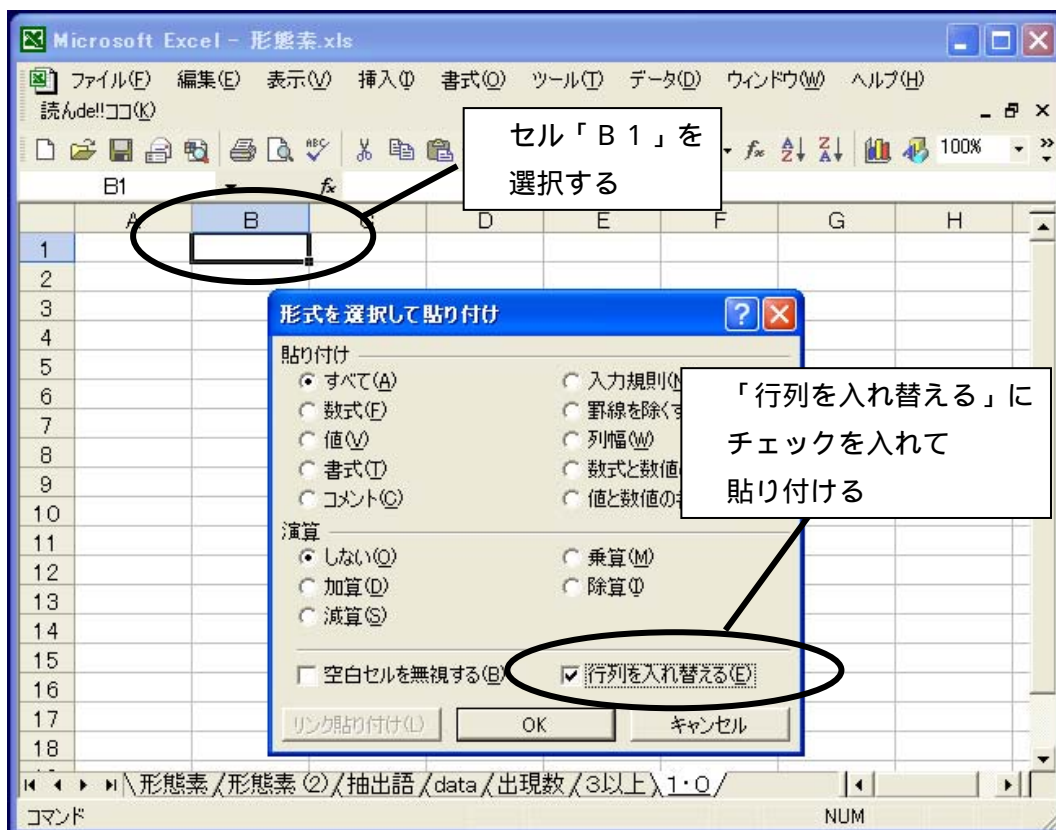




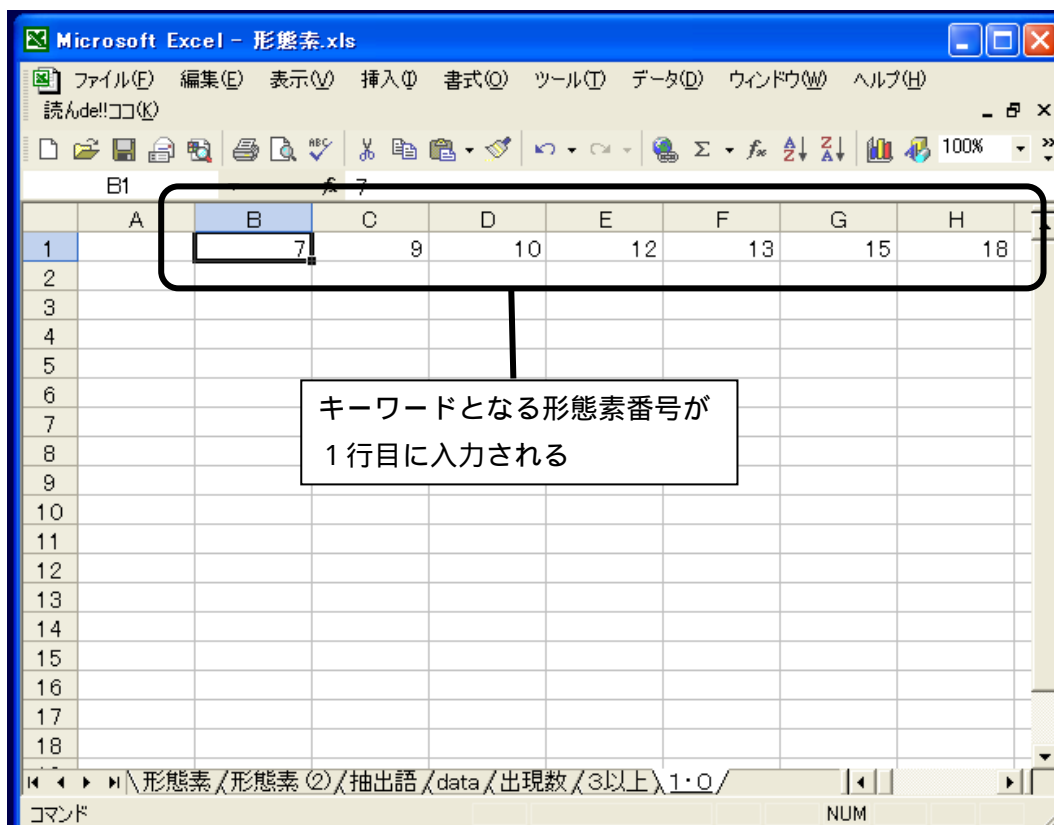
画面 29 「1・0データ化」の準備



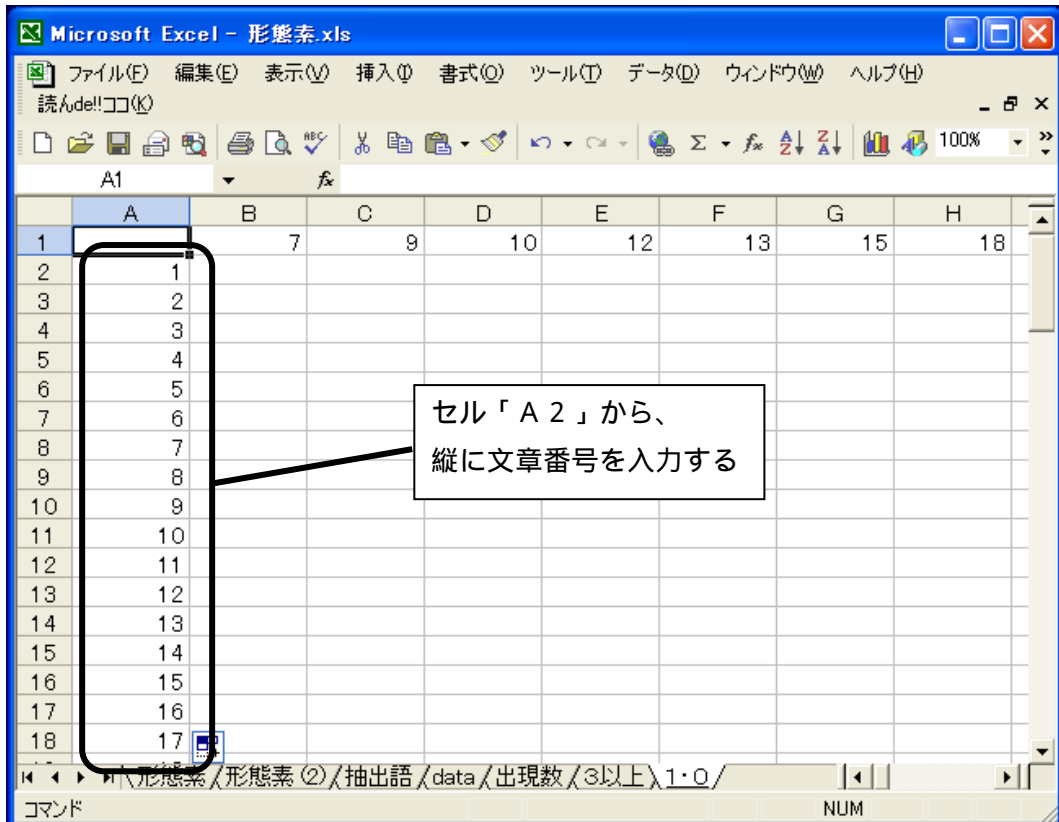
画面 30 「1・0データ化」の準備



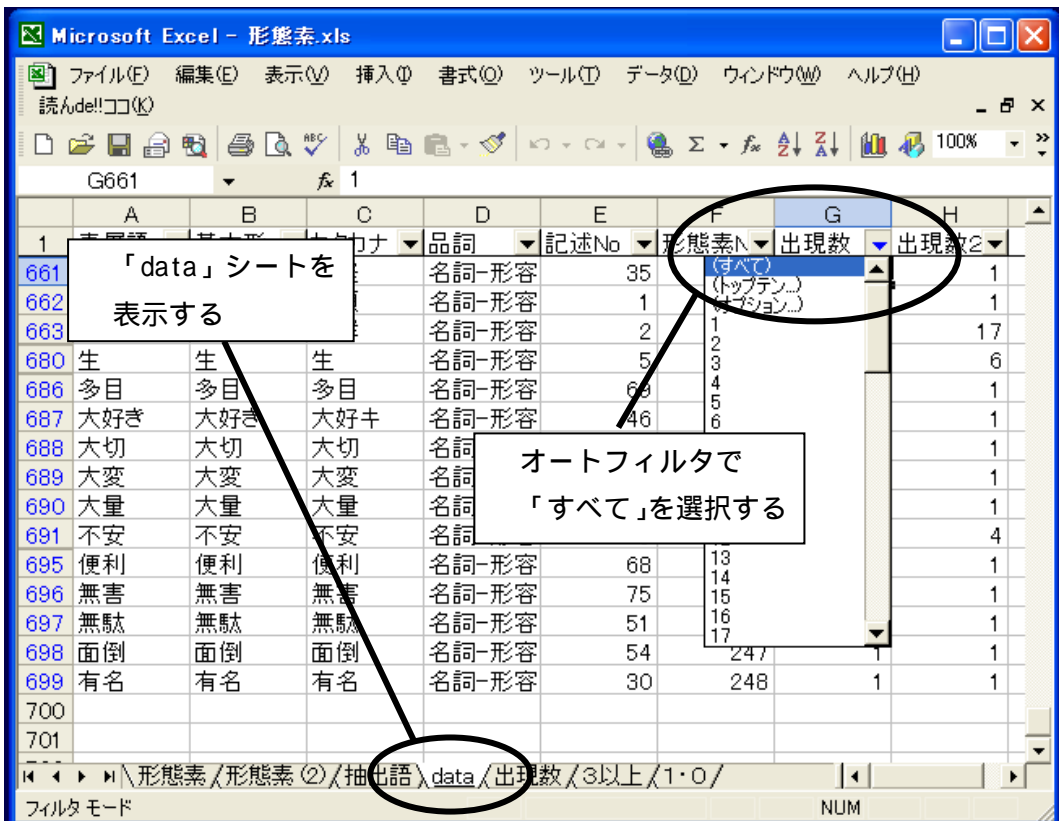
画面 31 「1・0データ化」の準備



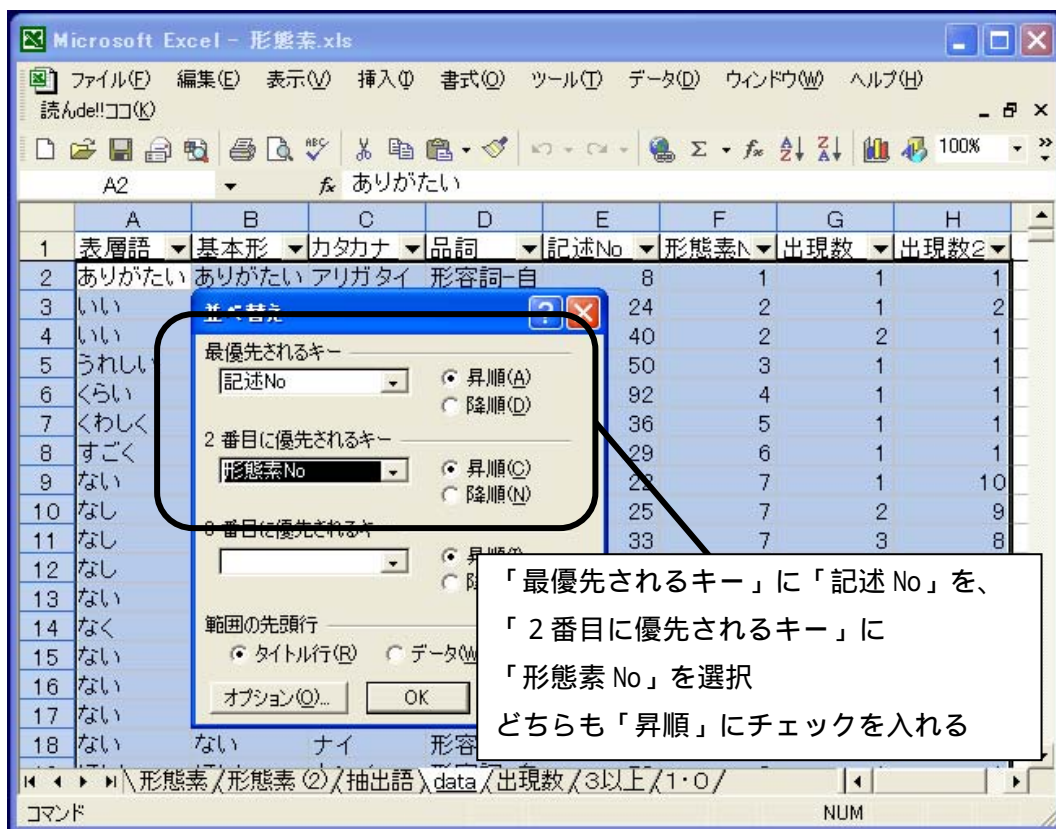
画面 32 「1・0データ化」の準備



画面 33 「1・0データ化」の準備



画面 34 「1・0データ化」の準備



画面 35 「1・0データ化」の準備

## (8) マクロの実行

まず、プログラムをインポートする<sup>注(8)</sup>。「ツール(T)」 - 「マクロ(M)」 - 「Visual Basic Editor(V)」を選択し、VBEを起動する(画面36)。VBEの「ファイル(F)」メニューから「ファイルのインポート(I)」を選択し、「1・0データ化」のマクロ「茶坊主くん」(仮称)をインポートする(画面37)。

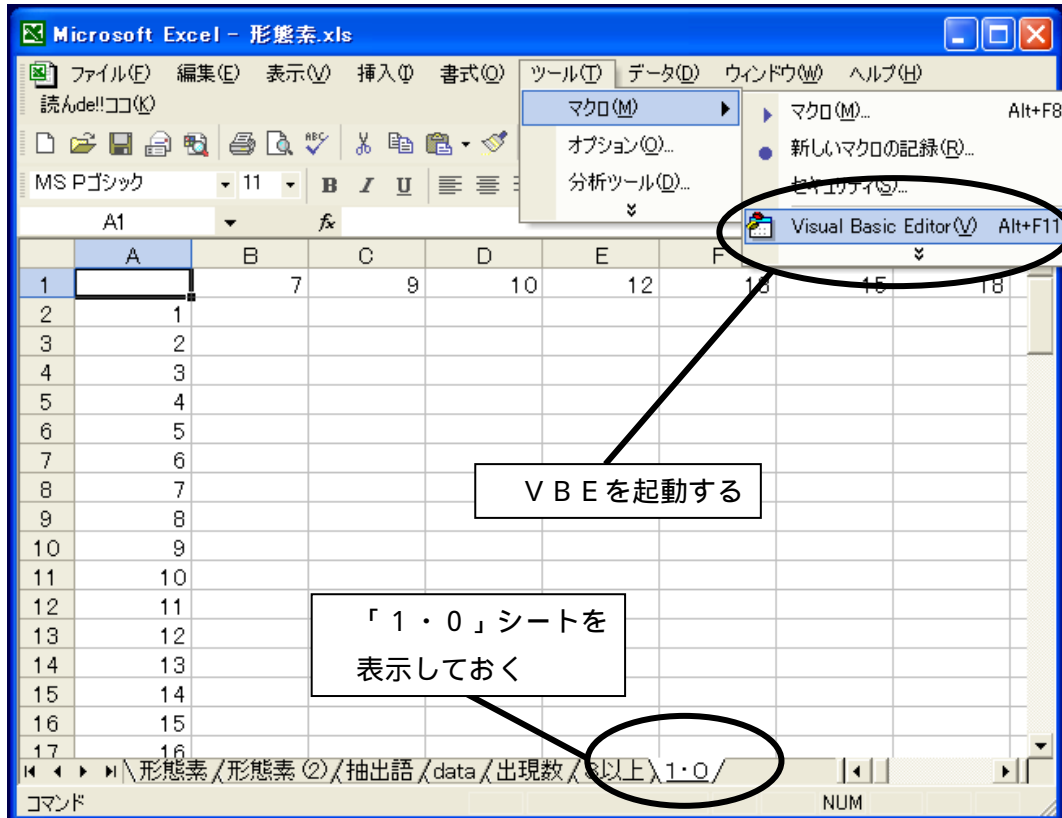
Excelの画面に戻り、「1・0」シートが表示されていることを確認する(画面38)。「ツール(T)」 - 「マクロ(M)」 - 「マクロ(M)」を選択すると、マクロ名「茶坊主くん」が表示されるので、これを実行する。すると、「抽出したキーワードの数」と「1・0データ化する文章の数」がそれぞれ表示されるので(画面39、40)数に間違いがないことを確認して「OK」をクリックする。

「1・0データ化」は、文章番号1から形態素番号の小さい順に行われる。この事例のように文章数92×キーワード数58という少ないデータ数の場合には数秒で終了するが、もっとデータの規模が大きい場合には、終了まで数分を要することがある<sup>注(9)</sup>。そのため、どの程度まで処理が進行しているか、現在処理中の文章番号がセル「A1」に表示されるようになっている(画面41)。このセルの値が総文章数(事例の場合92)と等しくなったら、「1・0データ化」は終了である。

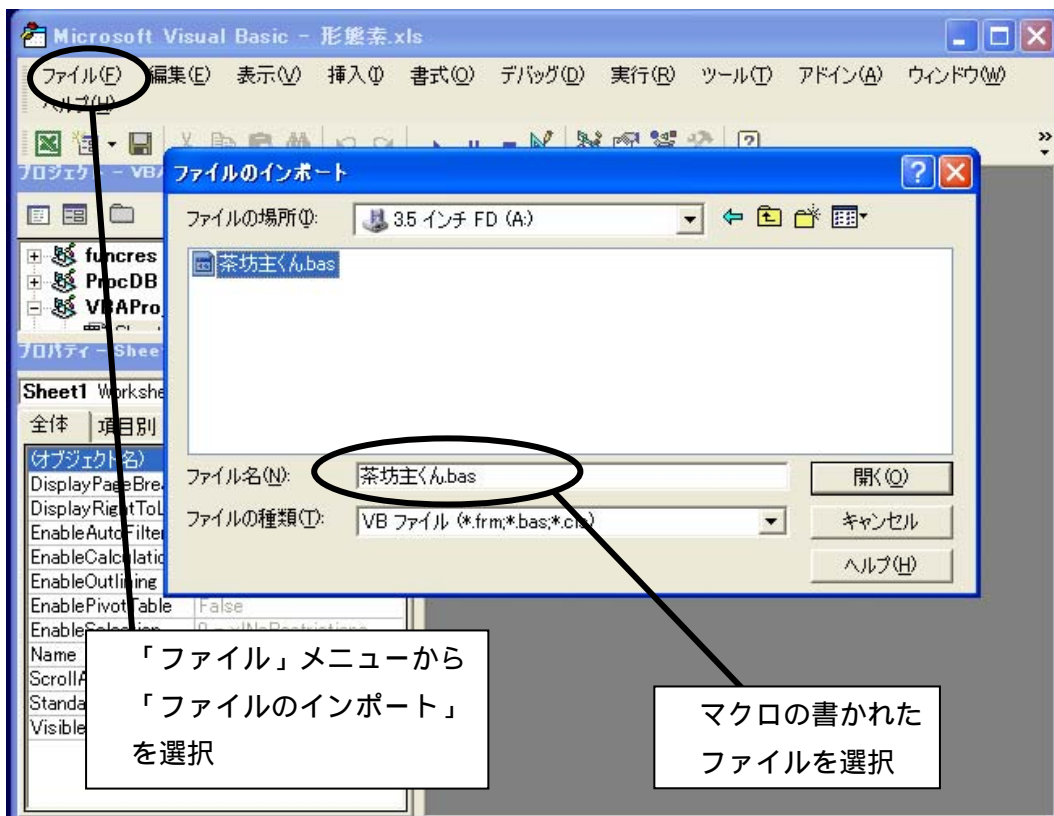
注(8)ここでは、マクロのファイルが作成済みであることを前提としている。マクロの作成については本書でコードを公表しているので、専用の解説書などを参照の上、各自で対応していただきたい。

注(9)データ規模が大きい場合には、処理速度を上げるために、「data」シートおよび「1・0」シートのみを他のブックに移してマクロを実行することをお勧めする。

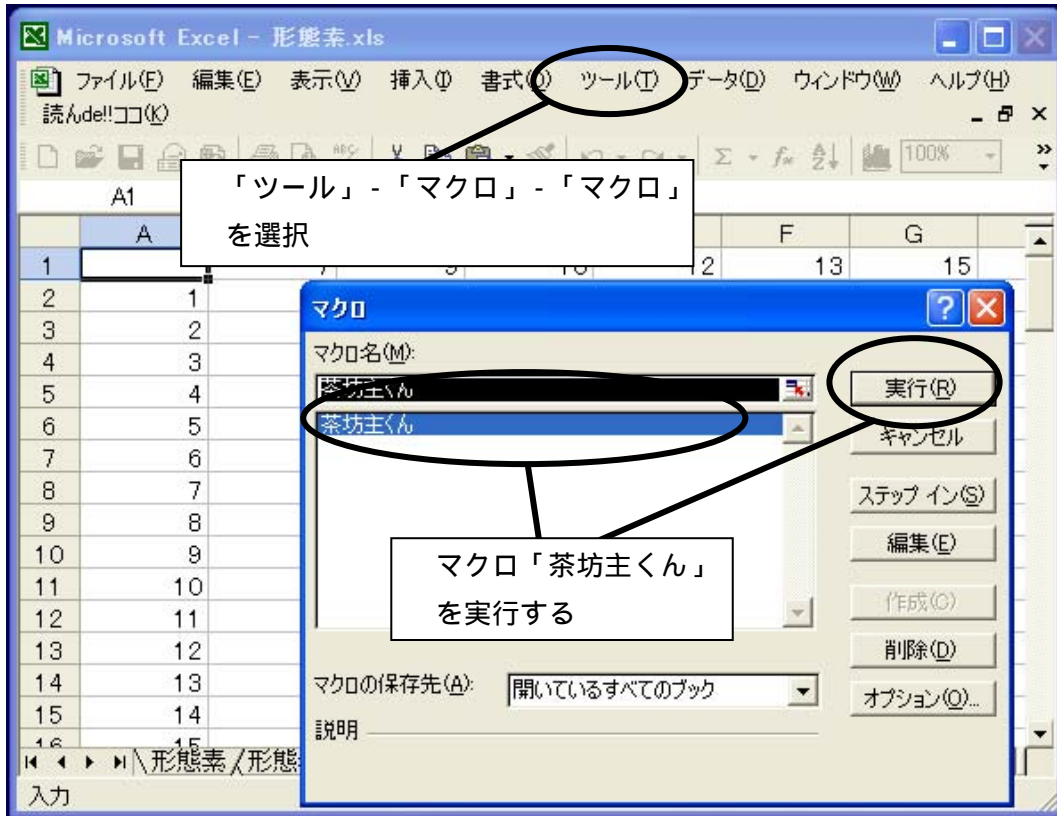




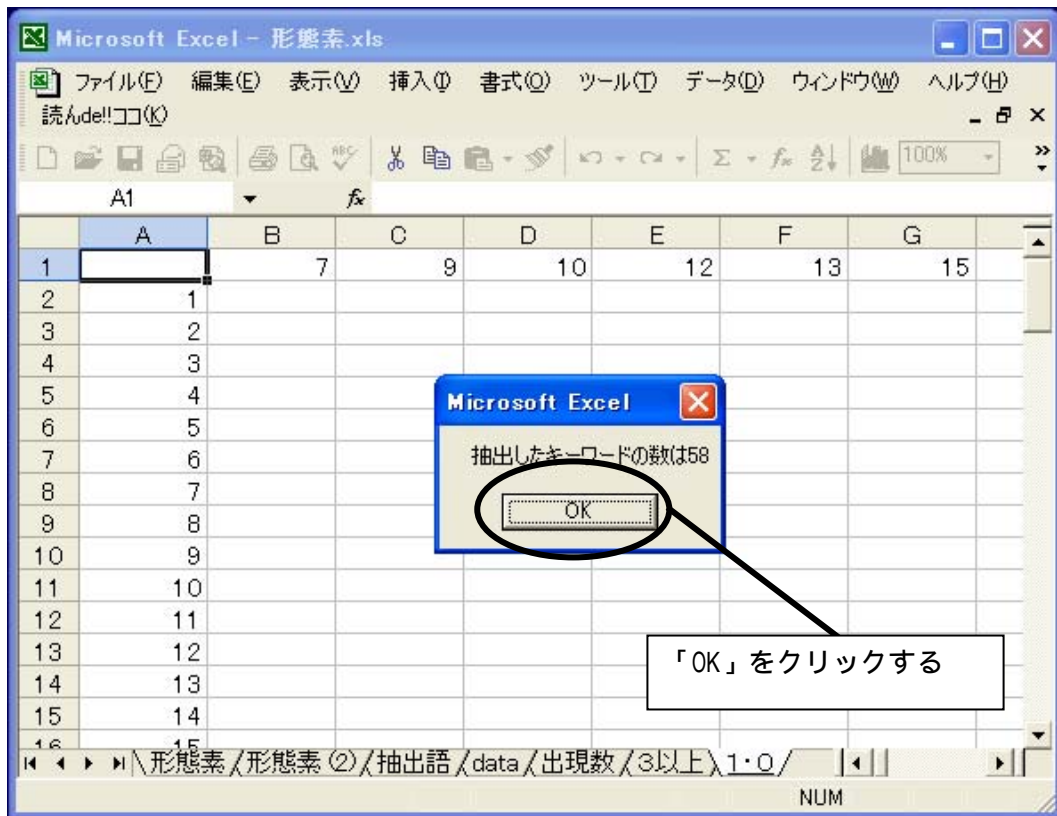
画面 36 マクロの実行



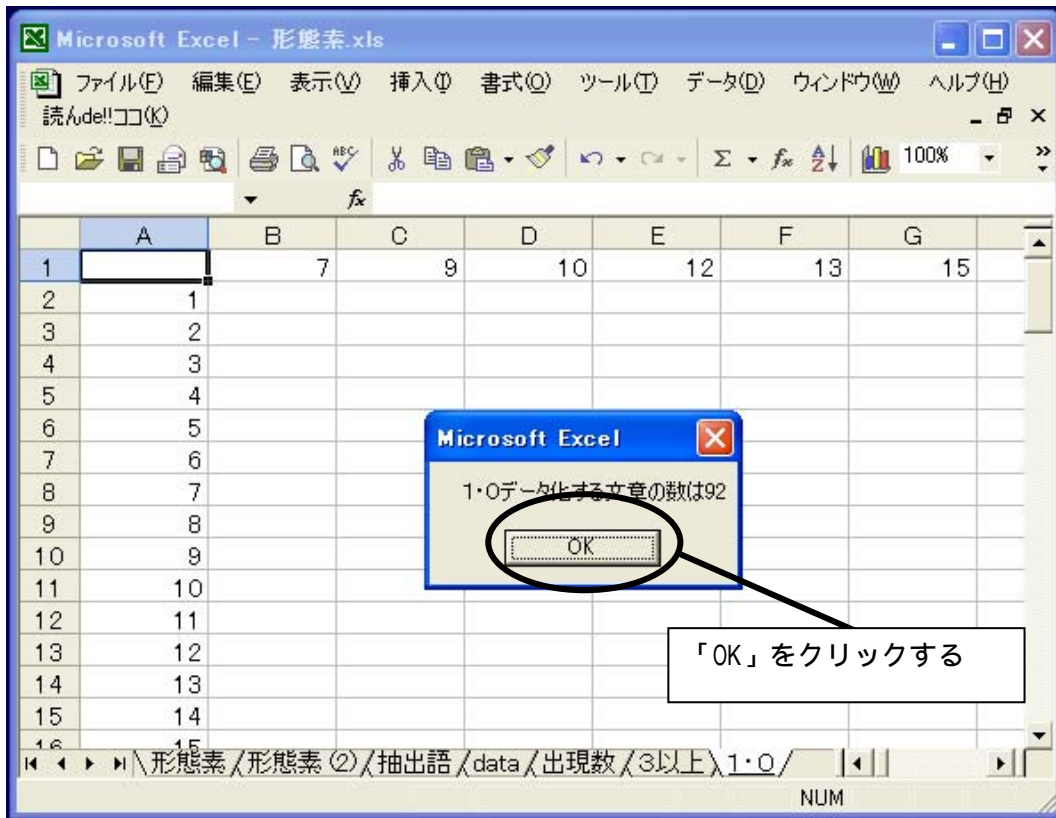
画面 37 マクロの実行



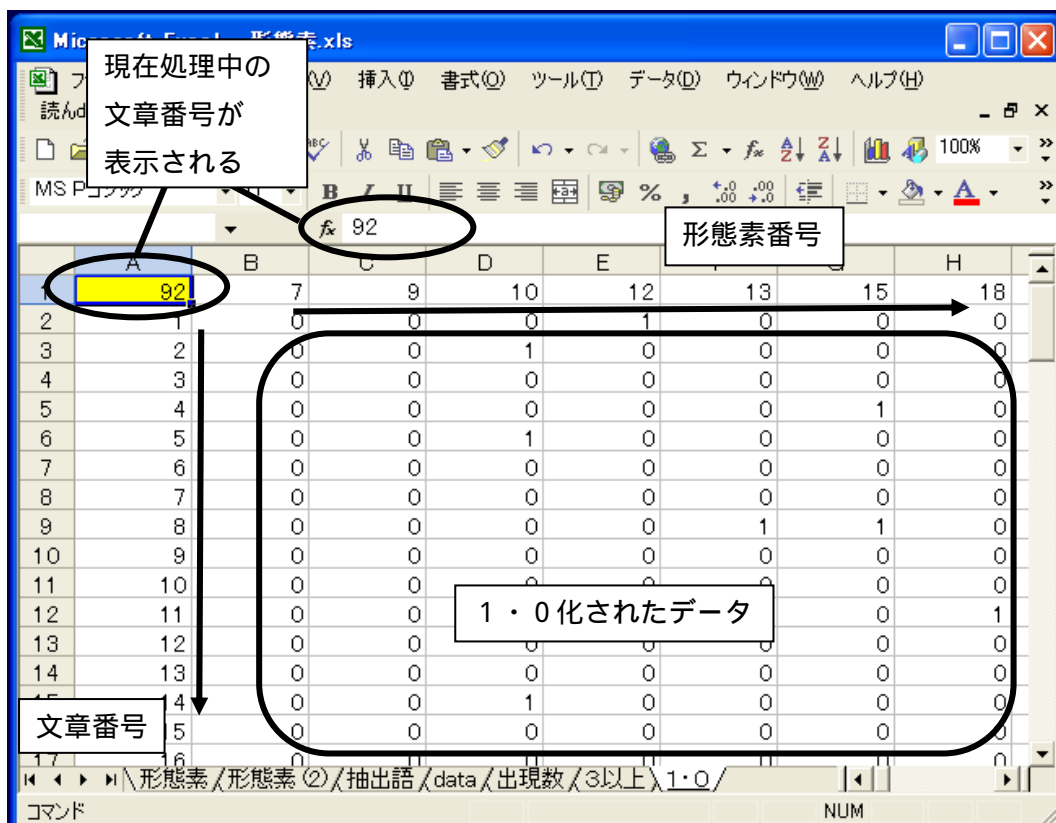
画面 38 マクロの実行



画面 39 マクロの実行



画面 40 マクロの実行



画面 41 マクロの実行

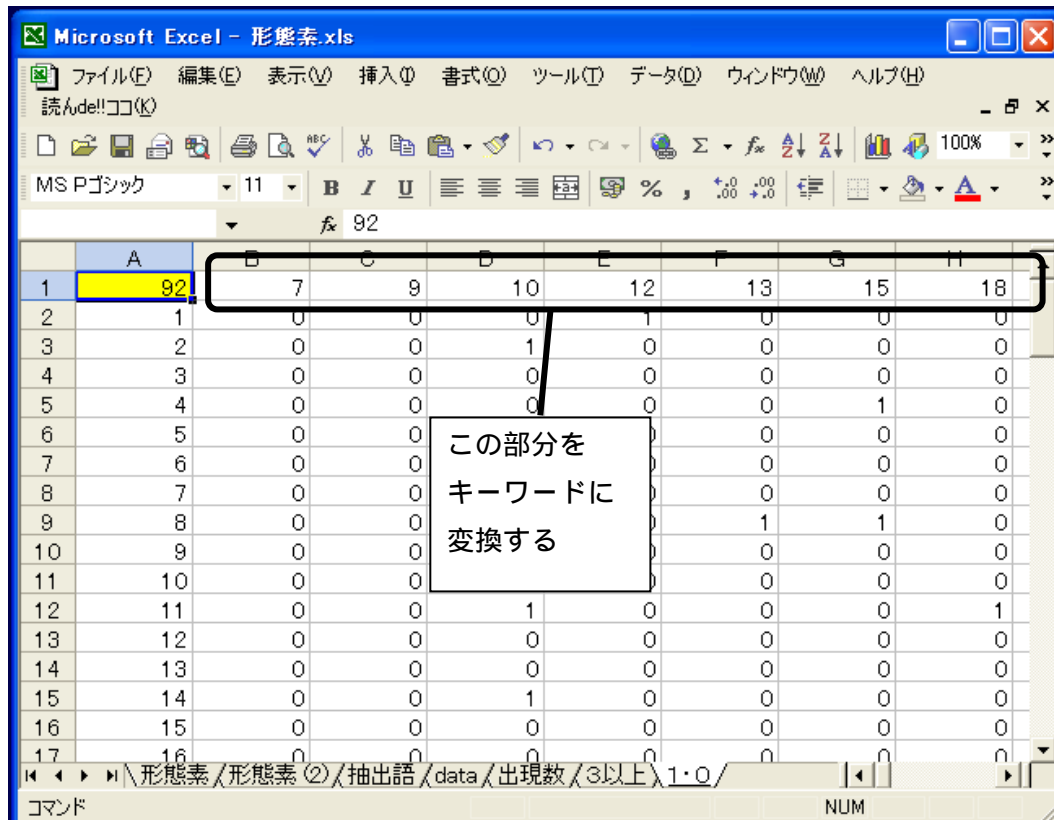
### (9) 1・0データファイルの完成

「1・0データ化」が終了したら、最後に元の文章データや属性コードのあるデータファイルと結合させる。

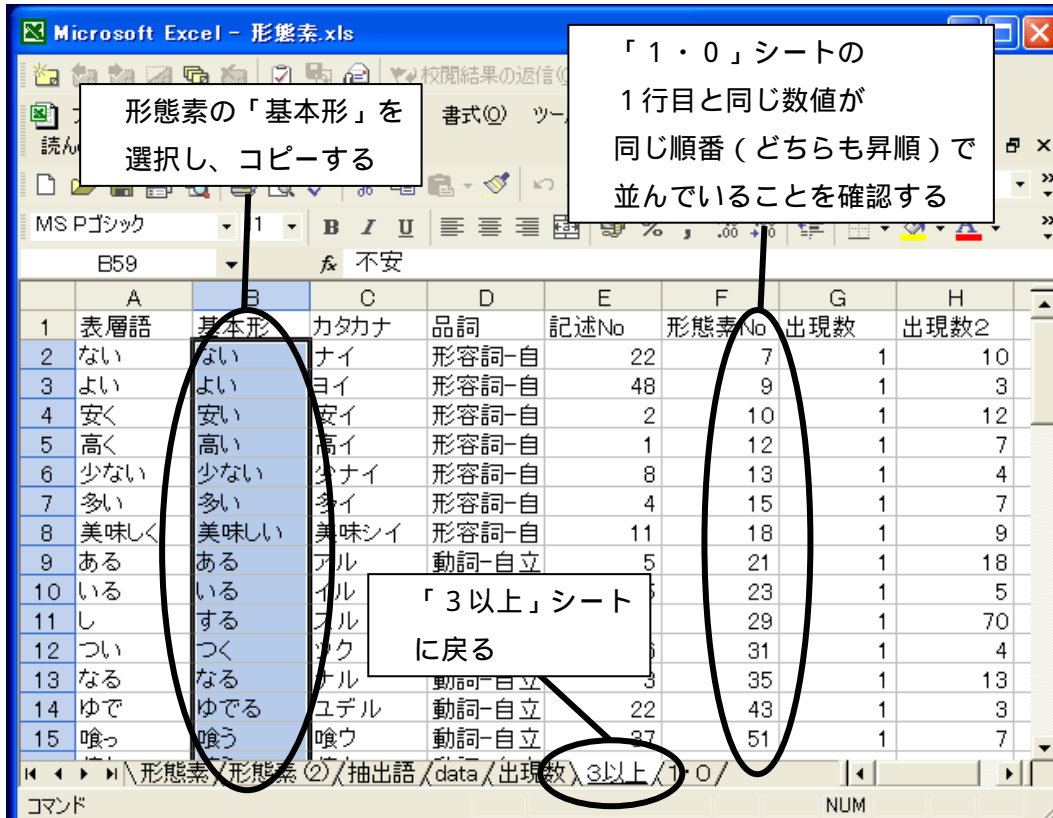
まず、1・0データの1行目にある形態素番号のセルを、数値から形態素(基本形)に書き換えて、どのキーワードに関する出現状況を示しているかわかるようにする(画面42)。先にキーワードを絞り込んだシート(事例ではシート名「3以上」)に戻り(画面43)「基本形」列にあるキーワードを選択してコピーする。「1・0」シートに戻ってセル「B2」を選択、「形式を選択して貼り付け(S)」で「行列を入れ替える(E)」にチェックを入れて貼り付ける(画面44)。

最後に、キーワードの出現状況を示した列を全て選択し(画面45)文章データが入力してある元のファイルに貼り付け(画面46)、1・0データファイルを完成させる(画面47)。

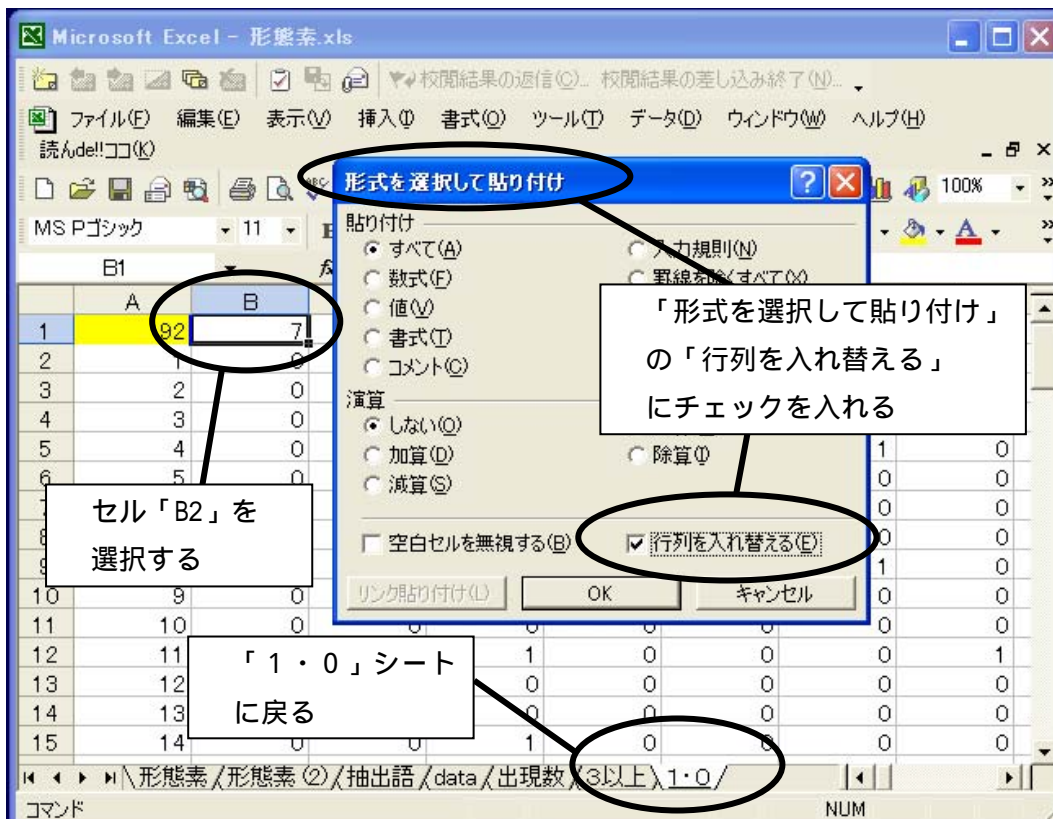
なお、キーワードから元の文章を検索する場合には(画面48)「データ(D)」-「フィルタ(F)」-「オートフィルタ(F)」を選択して1行目にフィルタ矢印をつけ、検索したいキーワードのフィルタ矢印をクリックして1を選択すると、求めている元の文章を抽出することができる。



画面42 1・0データファイルの完成

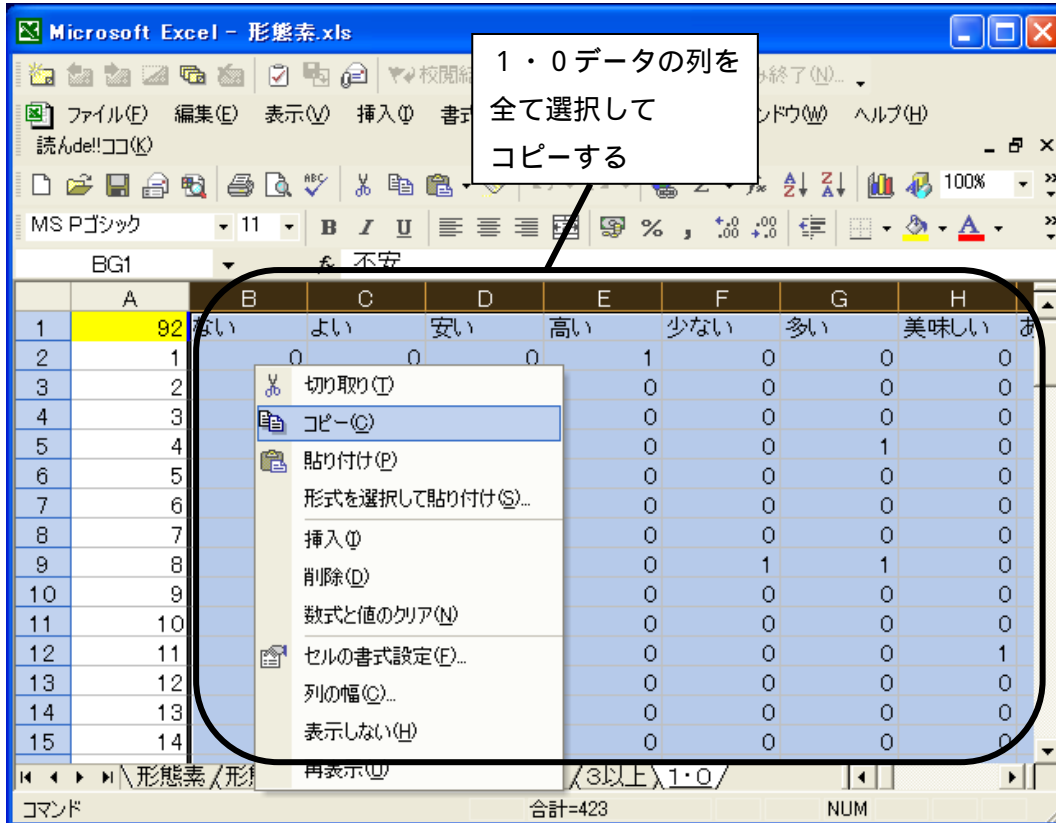


画面 43 1・0 データファイルの完成

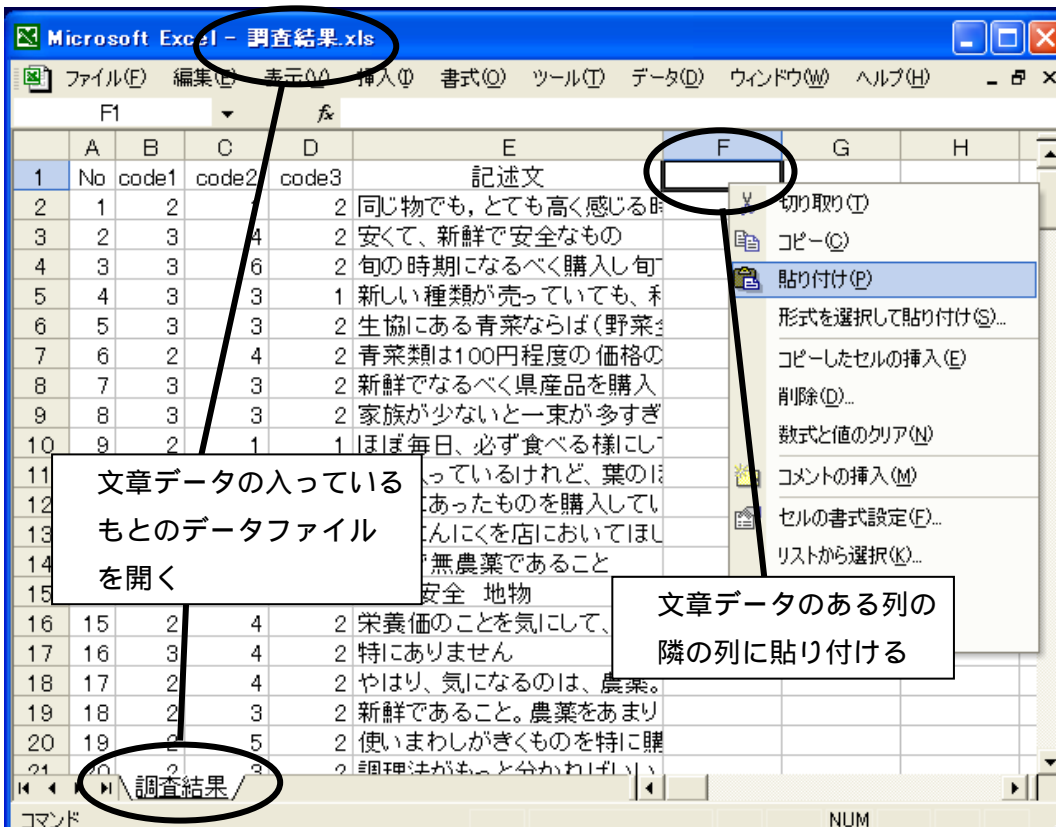


画面 44 1・0 データファイルの完成





画面 45 1・0データファイルの完成



画面 46 1・0データファイルの完成

Microsoft Excel - 調査結果.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

E179

	A	B	C	D	E	F	G	H
1	No	code1	code2	code3	記述文	ない	よい	安い
2	1	2	1	2	同じ物でも、とても高く感じる時	0	0	0
3	2	3	4	2	安くて、新鮮で安全なもの	0	0	1
4	3	3	6	2	旬の時期になるべく購入し旬	0	0	0
5	4	3	3	1	新しい種類が売っていても、希	0	0	0
6	5	3	3	2	生協にある野菜ならば(野菜)	0	0	1
7	6	2	4	2	野菜類は100円程度の価格の	0	0	0
8	7	3	3	2	新鮮でなるべく県産品を購入	0	0	0
9	8	3	3	2	家族が少ないと一束が多すぎ	0	0	0
10	9	2	1	1	ほぼ毎日、必ず食べる様にし	0	0	0
11	10	3	3	4	袋に入っているけれど、葉のは	0	0	0
12	11	2	4	2	時期にあったものを購入してい	0	0	1
13	12	4	4	2	行者にんにくを店においてほし	0	0	0
14	13	4	3	2	新鮮で無農薬であること	0	0	0
15	14	3	5	2	安さ 安全 地物	0	0	1
16	15	2	4	2	栄養価のことを気にして、意識	0	0	0
17	16	3	4	2	特にありません	0	0	0
18	17	2	4	2	やはり、気になるのは、農薬。	0	0	0
19	18	2	3	2	新鮮であること。農薬をあまり	0	0	0
20	19	2	5	2	使いまわしがきくものを特に購	0	0	0
21	20	2	3	2	調理法がもっと分かれればい	0	0	0

調査結果/

入力 NUM

画面 47 1・0 データファイルの完成

Microsoft Excel - 調査結果.xls

ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) ツール(T) データ(D) ウィンドウ(W) ヘルプ(H)

読んだ!!ココ(K)

L1

検索したいキーワードの  
フィルタ矢印をクリックし、  
1を選択する

	A	B	C	D	E	F	G	H
1	No	code1	code2	code3	記述文	ない	よい	安い
12	11	2	4	2	時期にあったものを購入しています。その方が美味	0	0	1
24	23	3	*	5	家で作っているホウレンソウは甘みがあって美味し	0	0	1
29	28	5	1	4	埼玉産直の「美味しい」時々は道の駅めぐりをして	0	0	1
31	30	3	3	2	地元がホウレンソウが有名なので、なるべく地元で	0	0	1
64	63	3	3	2	八百屋さんの野菜は新鮮で安く「美味しい」です。派	0	0	1
66	65	3	3	2	青菜はやはり新しい方が美味しいので、いつも新鮮	0	0	1
84	83	5	4	2	見た目が生き生きして美味しそうだとつい、買って	0	0	1
89	88	4	4	2	虫喰いがあっても丈が長ても安全で美味しいければ	0	0	1
93	92	3	3	2	ちょっとくらい高くても新鮮で美味しいものであれば	0	0	1
94								
95								
96								
97								
98								
99								
100								
101								
102								
103								

調査結果/

92 レコード中 9 個が見つかりました。 NUM

求めるキーワードを  
含む文章が抽出される

画面 48 1・0 データファイルの完成

#### 4．補足説明

##### (1) 2つ以上のキーワードをまとめる場合

1・0データファイルの作成を終了し、文章の検索や分析を進めていくうちに、同じような意味で使われている語など、いくつかのキーワードをまとめて1つの語として扱う必要が生じてくることがある<sup>注(10)</sup>。その場合、それらのキーワードの出現状況をあわせて1・0で示す列を新たに加えることになる。以下、その手順を説明する。

まず、合成したい語(ここでは「調理」と「料理」をあわせる事例を示す)の列を、別のシートにコピー&貼り付けする(画面49、50)。新たに、「合計」列をつくり、2列の数値をそれぞれに行ごとに合計する(画面51)。これにより、「合計」列では、どちらか一方の語が出現した文章には1、両方の語が出現している場合には2、どちらの語も出現していない場合には0が表示される<sup>注(11)</sup>。

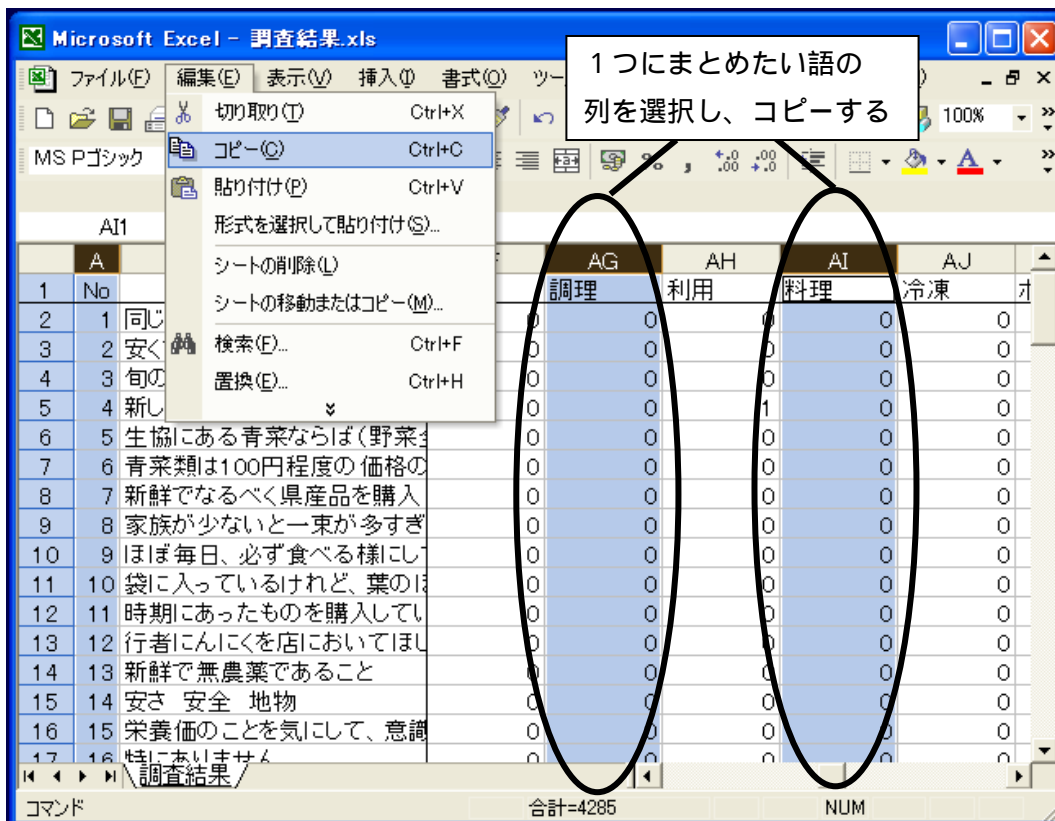
次に、「合計」列の隣に、2つの語を合成した項目であることを示す名前を付ける(事例では、「調理・料理」)(画面52)。そして、2行目以降には、「『合計』列が0の場合は0、それ以外は1」というIF関数を入力する。[E2の場合:IF(D2=0、0、1)]

最後に、合成項目の列を選択・コピーし(画面53)、元のデータの最後尾に貼り付ける。この際、「数式」のまま貼り付けると隣の列と同じ数値になってしまうので、必ず「値」で貼り付けるよう注意が必要である(画面54、55)。

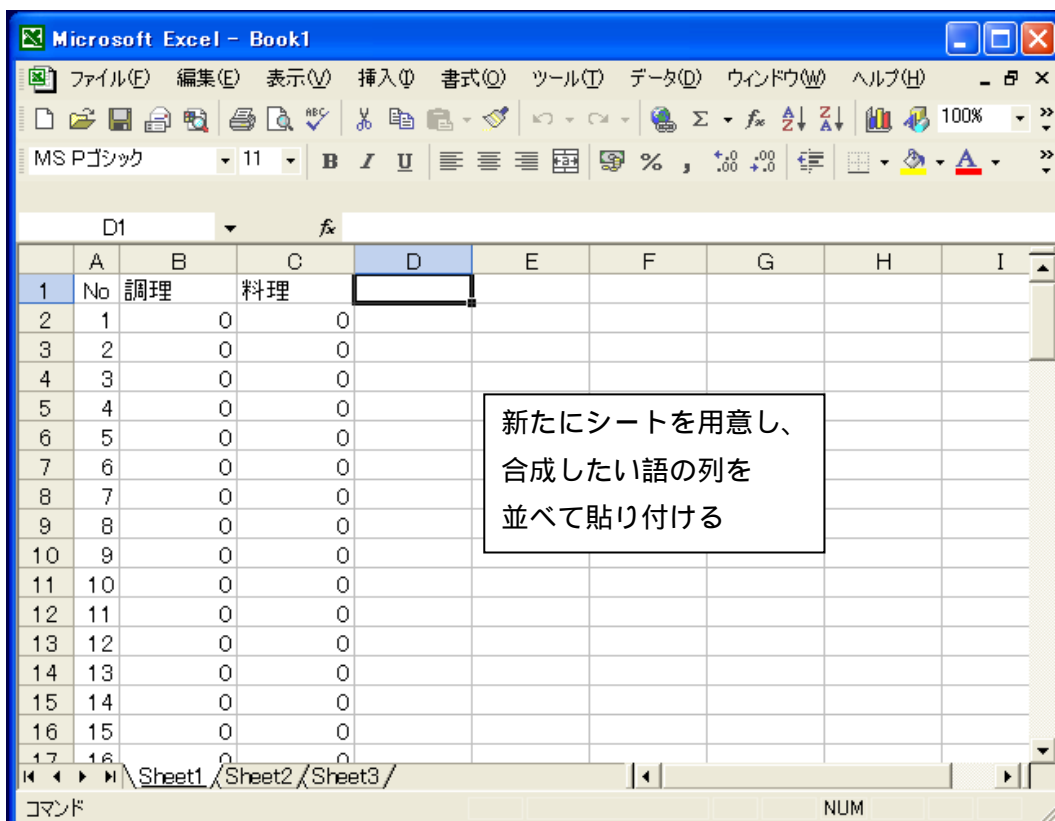
注(10)同じ意味で使われる語の他、全く反対の意味を持つ語も1つにまとめることによって分析が容易になる場合がある。例えば、「美味しい」と書かれた文章と、「美味しくない」と書かれた文章は、全く反対の内容であるにも関わらず、形態素で抽出すると「美味しい」という語になる。一方、「不味い」と書かれた文章は、「美味しくない」とほぼ同じ意味であるにもかかわらず、別の語として抽出される。そこで、「美味しい」も「不味い」も1つにまとめて、「味」に関する記述と捉えるのである。

注(11)事例のように合成する語が2つの場合は、行ごとの合計を出さずに、直接「2列とも0の場合は0、そうでなければ1」という条件式を使うことも可能である。ただし、合成する語が3語以上の場合には、いったん合計した方が容易である。

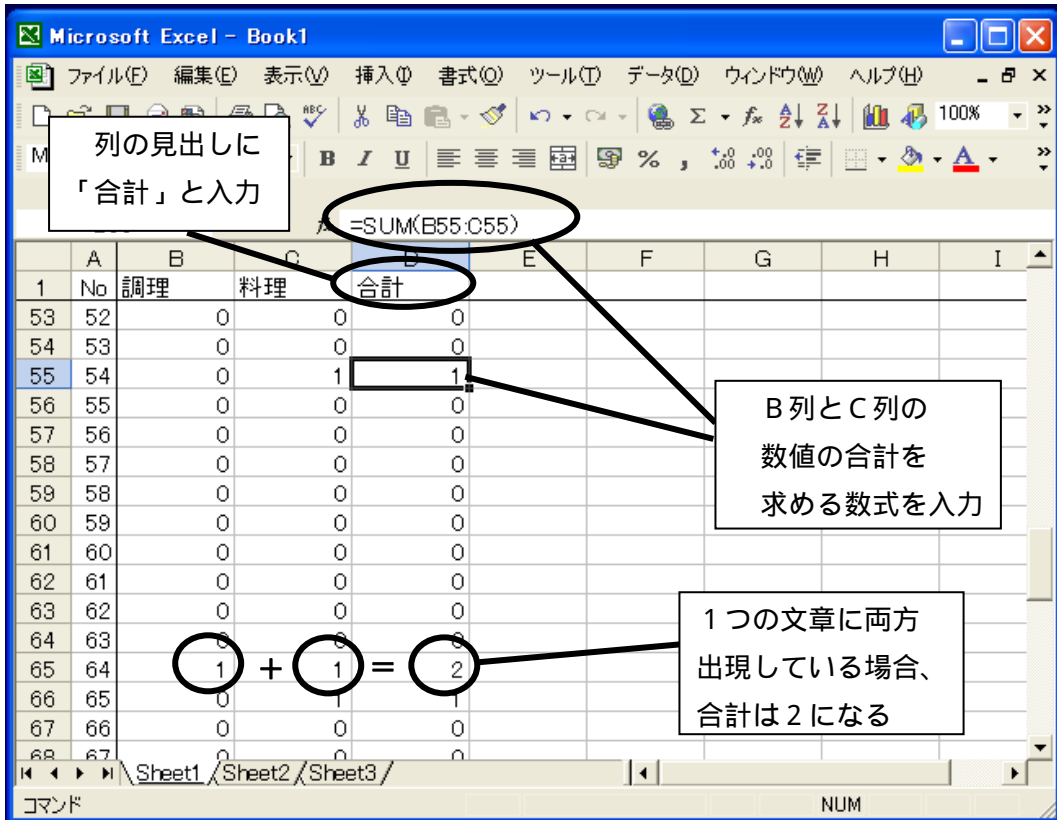




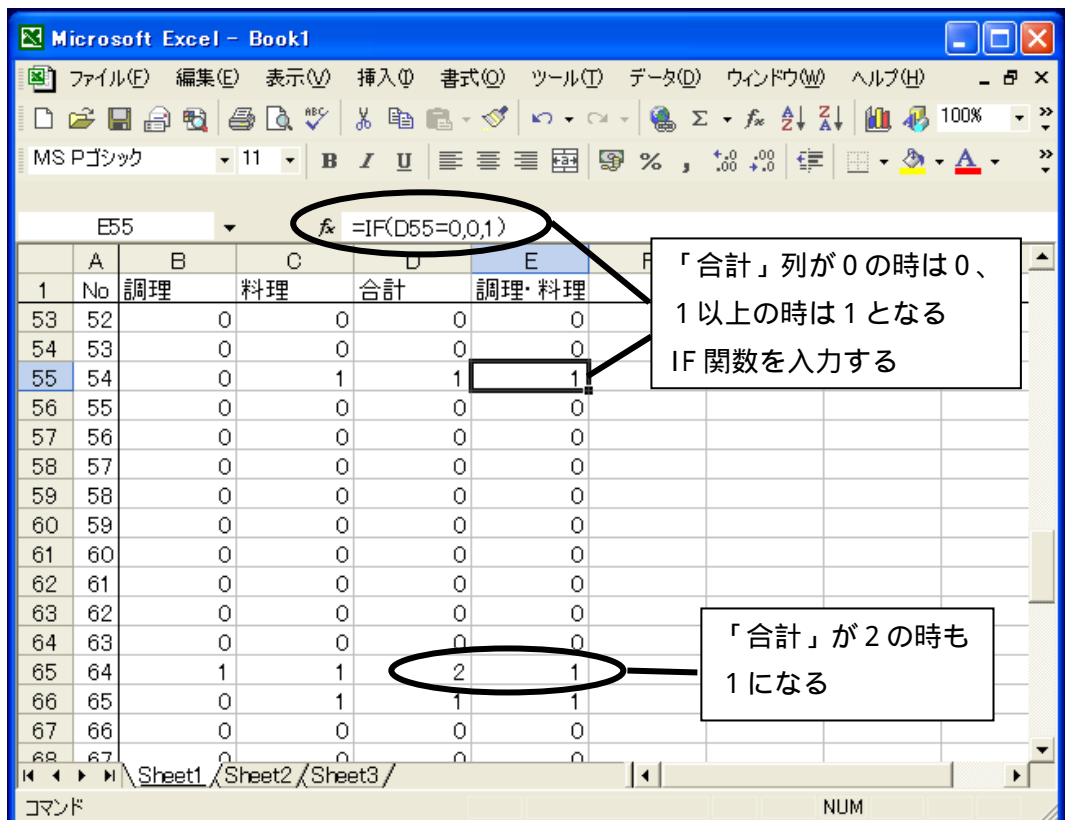
画面 49 語の合成



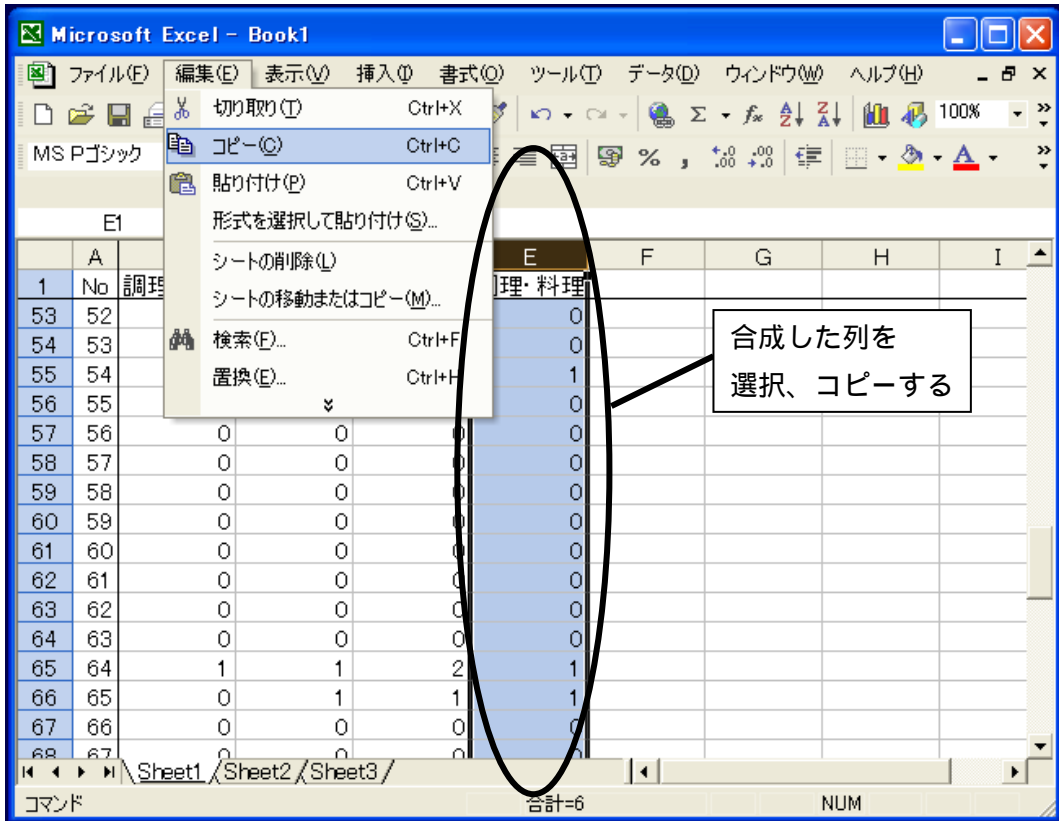
画面 50 語の合成



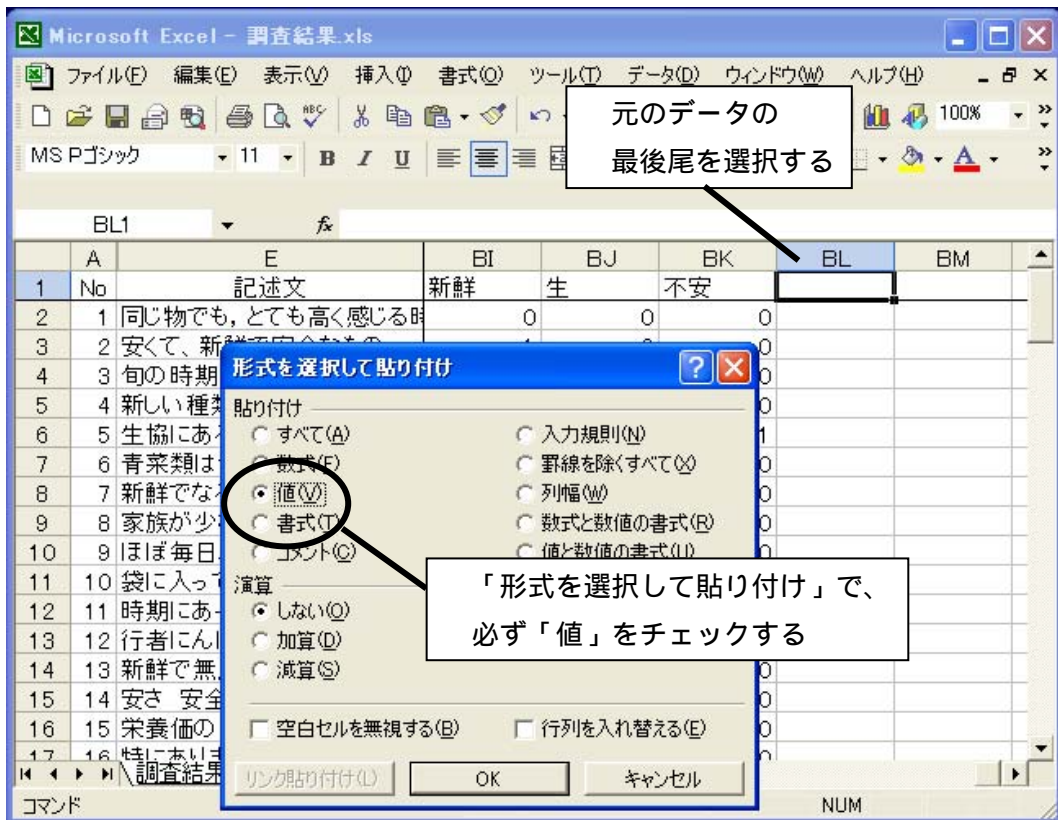
画面 51 語の合成



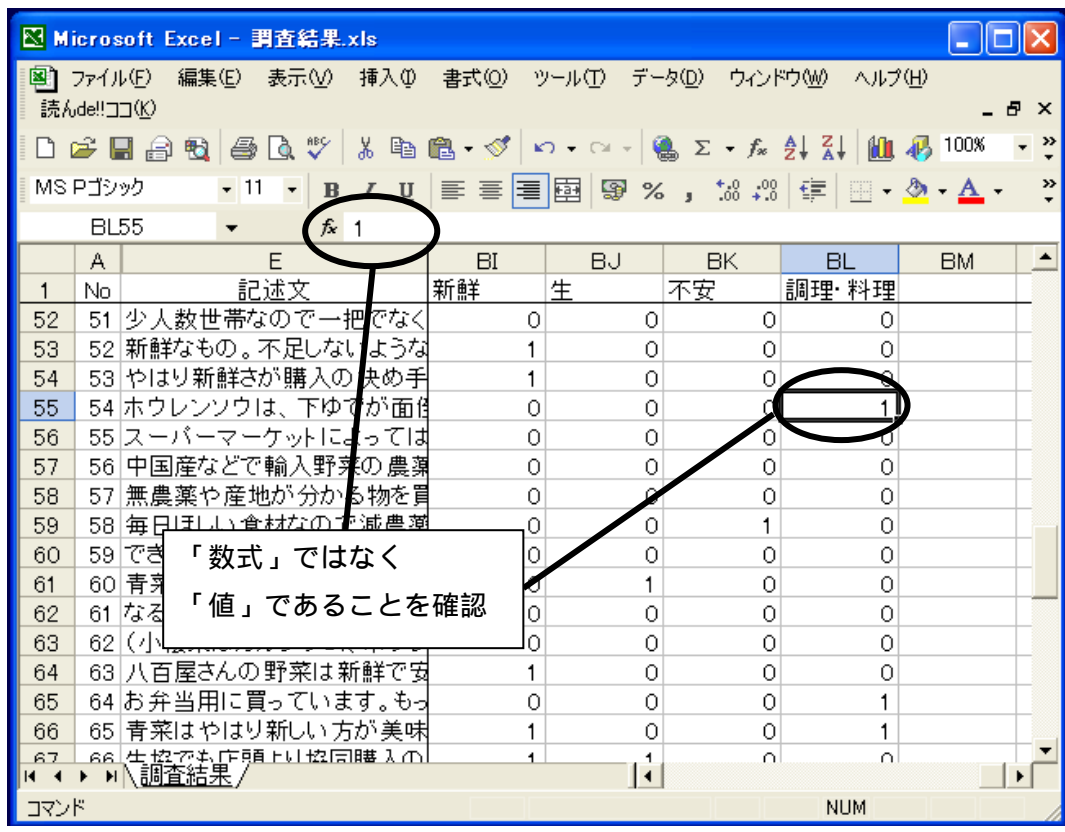
画面 52 語の合成



画面 53 語の合成



画面 54 語の合成



画面 55 語の合成

(2) マクロを使わずに 1・0 データファイルを作成する方法

最後に、マクロを使わずに 1・0 データファイルを作成する方法を紹介する。本手法においてマクロは、単純ではあるが膨大な作業量を要する部分でのみ用いている。従って、この部分を手作業で行うには、相当の労力を要する。しかしながら、マクロの実行において不具合が生じた場合など、何らかの事情でマクロを利用できない事態を想定して、手作業で「1・0 データ化」を行う場合の手順を述べておく。

まず、キーワードの絞り込み (画面 27) までは、前章で述べた手順と同様である。ここから、各キーワードがどの文章において書かれたものであるかをリストアップする。

抽出したキーワードを昇順に並べ替え (画面 56)、キーワードを示す番号の列 (「形態素 No」列) とその基本形の列 (「基本形」列) をそれぞれコピーし、新たに設けたシート (シート名「出現語」) の 1 行目、2 行目に、「形式を選択して貼り付け (S)」 - 「行列を入れ替える (E)」で貼り付ける (画面 57)。

ここで、「data」シートに戻る (画面 58)。「出現数」列のオートフィルタがかかった状態にある場合には、そのフィルタ矢印の「すべて」を選択して全データを表示させておく。そして、「形態素 No」列のフィルタ矢印をクリックし、先に挙げたキーワード番号の 1 つを選択する。ここに示された文章番号 (「記述 No」列) が、そのキーワードが出現した文章であることを示すので、これをコピーし、先に作成した「抽出語」シートに貼り付ける (画面 59)。この手順を、抽出した全てのキーワード番号について行う。

次に、1・0 データ用のシートを用意し (シート名「1・0 data」)、1 列目 2 行目 (セル A2)

から下に向けて文章番号を入力する（画面 60）。

さらに、「新規作成」で、新たに作業用のファイル（以下、「作業ファイル」）を用意する。ここにも、1 列目 2 行目（セル A2）から下に向けて文章番号を入力しておく（画面 61）。

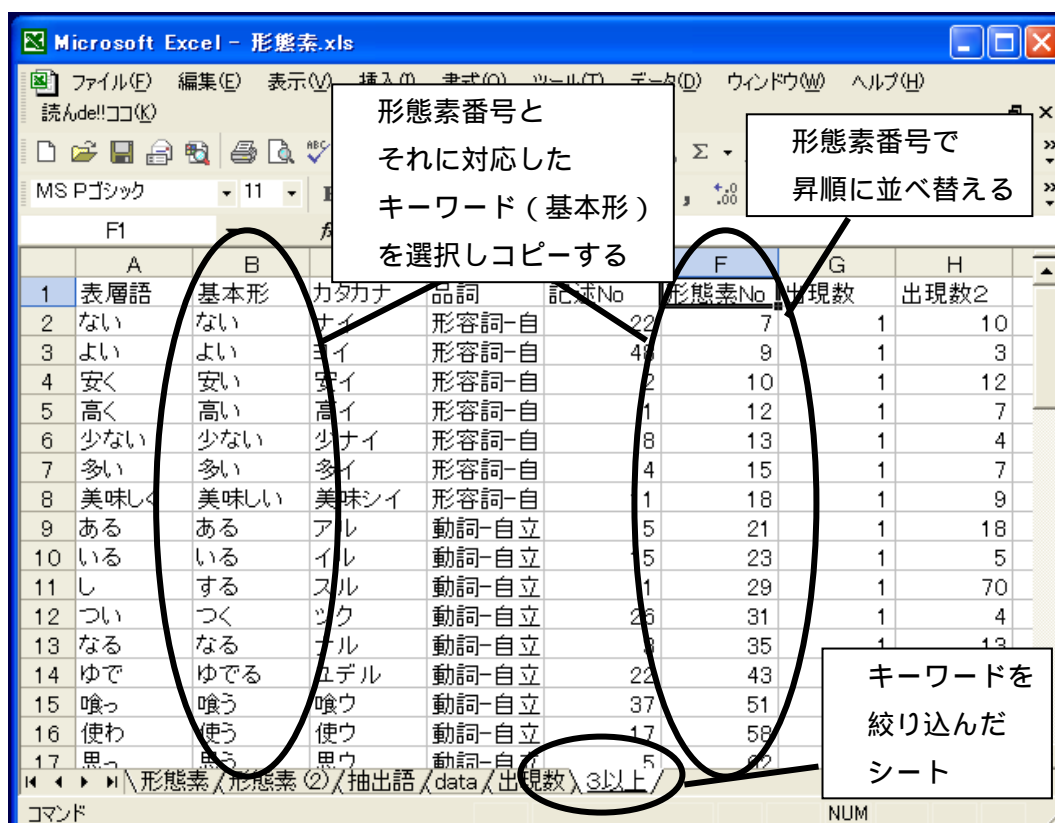
「出現語」シートに戻り、1 つめのキーワードについて、「基本形」とその下に入力されている文章番号をコピーし（画面 62）、「作業ファイル」の 1 行目に「形式を選択して貼り付け(S)」 - 「行列を入れ替える(E)」で横に貼り付ける（画面 63）。

これにより、A 列縦方向に全体の文章番号、1 行目横方向にキーワードの出現した文章番号が入力されたことになる。そこで、各セルについて、表側の番号と表頭の番号が等しい場合には 1、異なる場合には 0 とする IF 関数を入力する（画面 64、65）。なお、IF 関数のセルをコピー & 貼り付けする場合には、表側と表頭の参照がずれないように、絶対参照「\$」をつける必要がある。

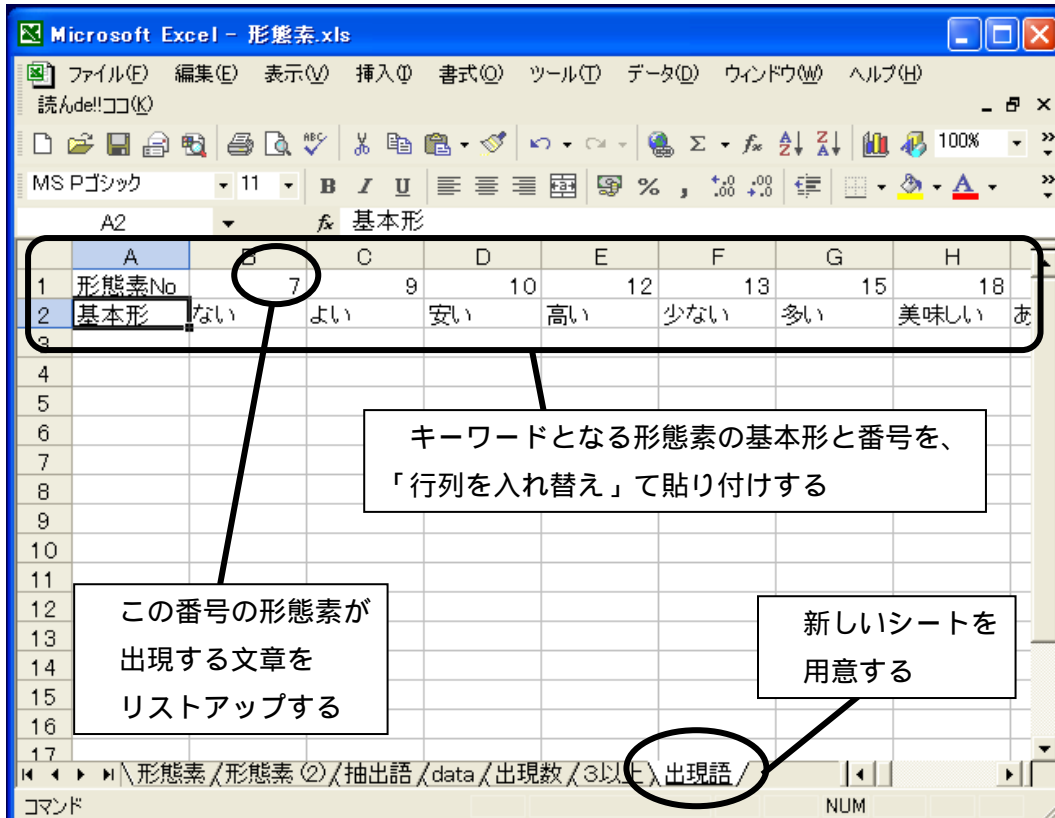
[ B2 の場合：IF(\$A2=B\$1、1、0) ]

全てのセルに条件式を入力したら、最後の列にカーソルを移動させ、新たに「合計」列を作る（画面 66）。ここに、各行の 1・0 の合計を入力する。もし、1 つの文章に 2 回以上同じキーワードが出現している場合、この合計は 2 以上になるため、更に 1・0 への変換を行う（画面 67）。「合計」列の隣に、現在 1・0 への変換を行っているキーワード名を入力し（画面では「ない」）、「『合計』列の数値が 0 の時は 0、それ以外は 1」という IF 関数を入力する。[ M2 の場合：IF(L2=0、0、1) ]

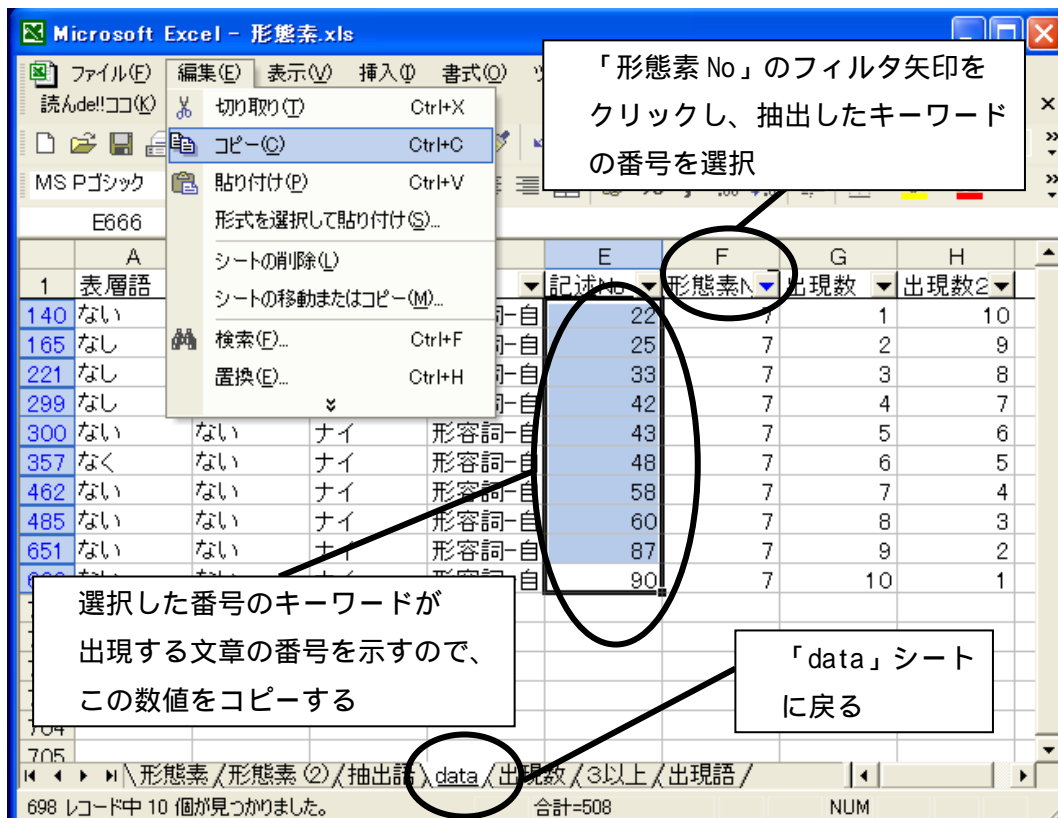
最後に、この列（「ない」列）をコピーし、「1・0 data」シートに「形式を選択して貼り付け(S)」の「値(V)」にチェックを入れて貼り付ける（画面 68）。この手順を、全てのキーワードについて行う。



画面 56 マクロを使わず「1・0 データ化」

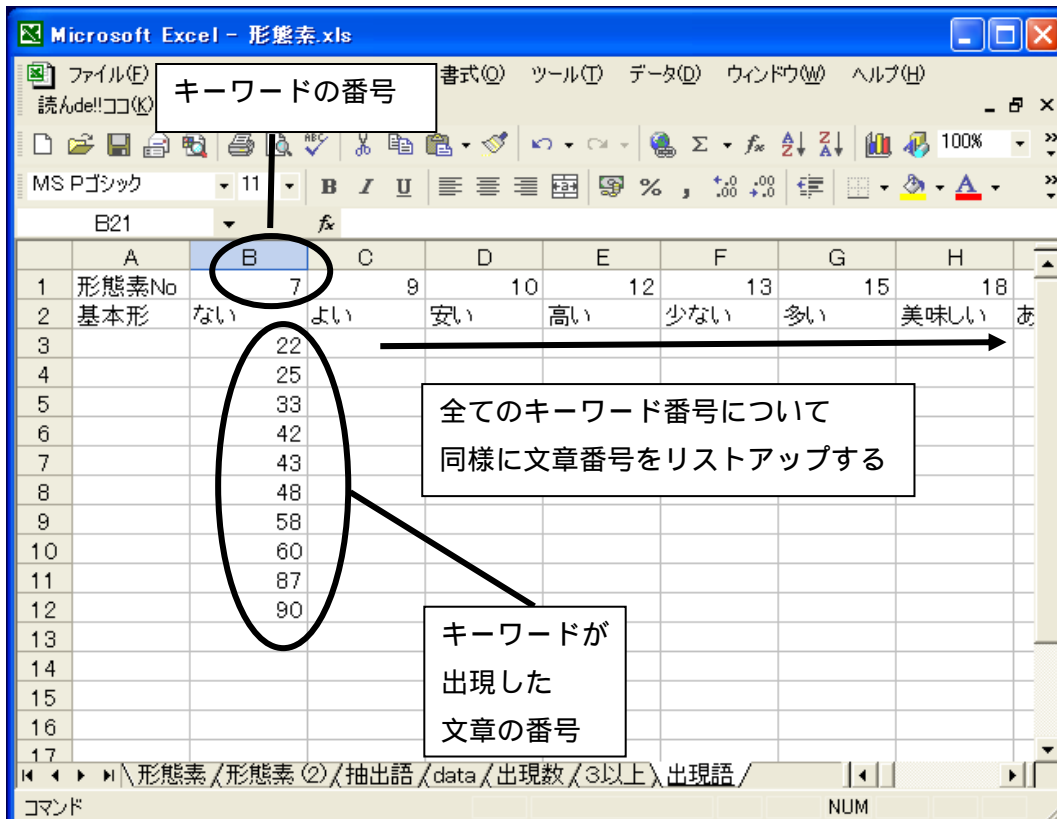


画面 57 マクロを使わず「1・0データ化」

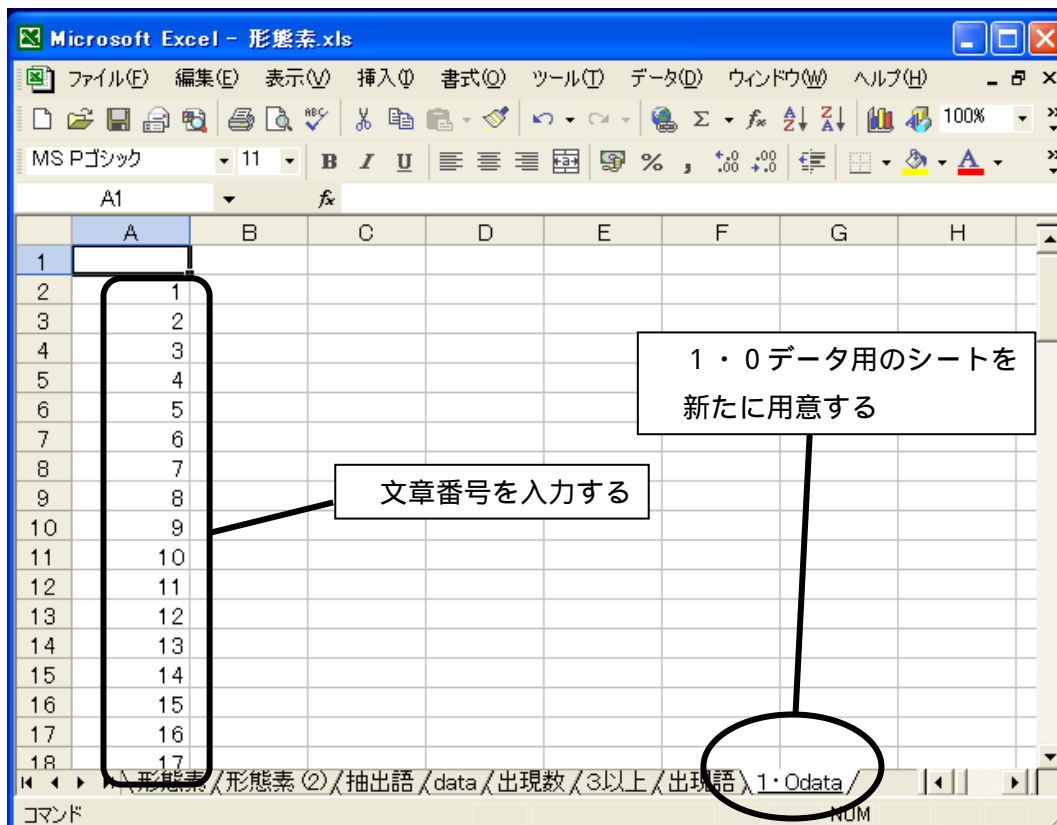


画面 58 マクロを使わず「1・0データ化」

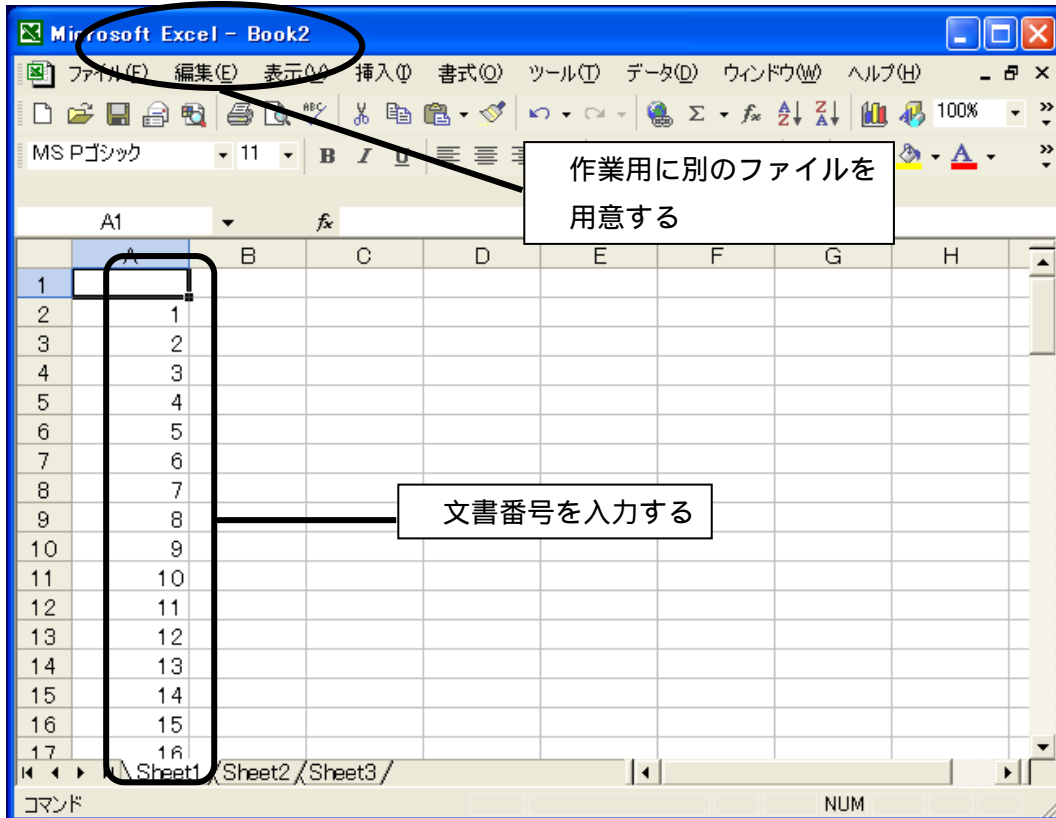




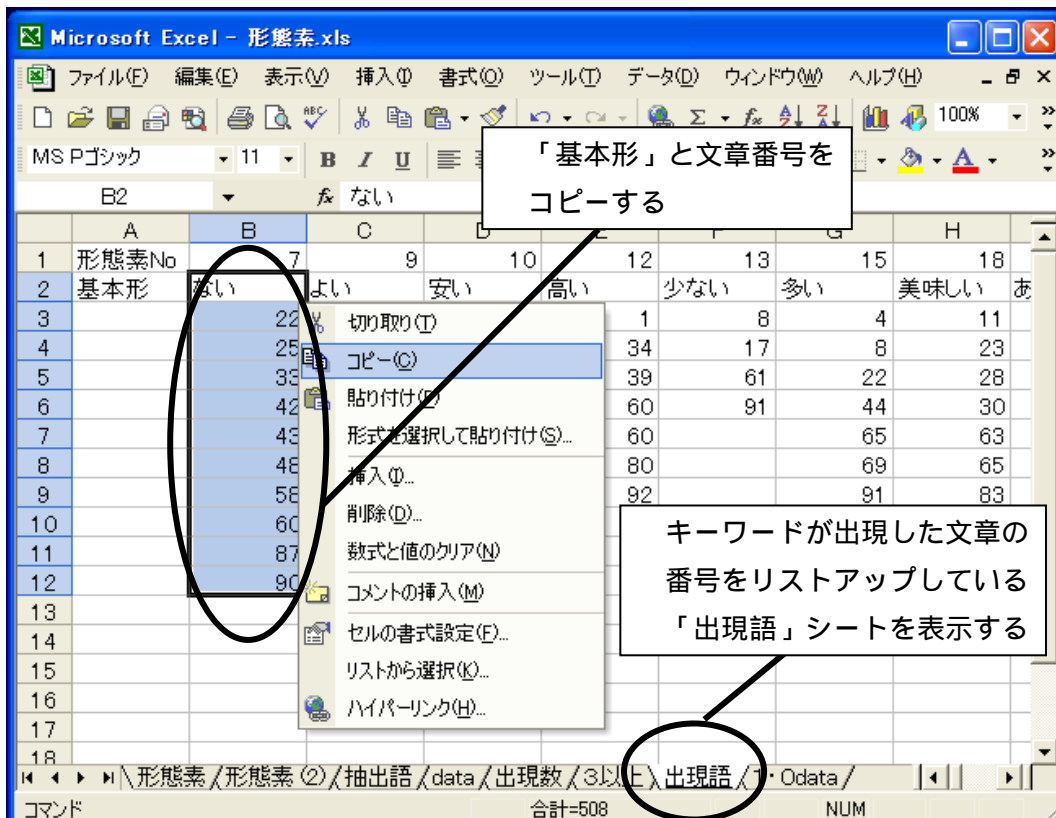
画面 59 マクロを使わず「1・0データ化」



画面 60 マクロを使わず「1・0データ化」

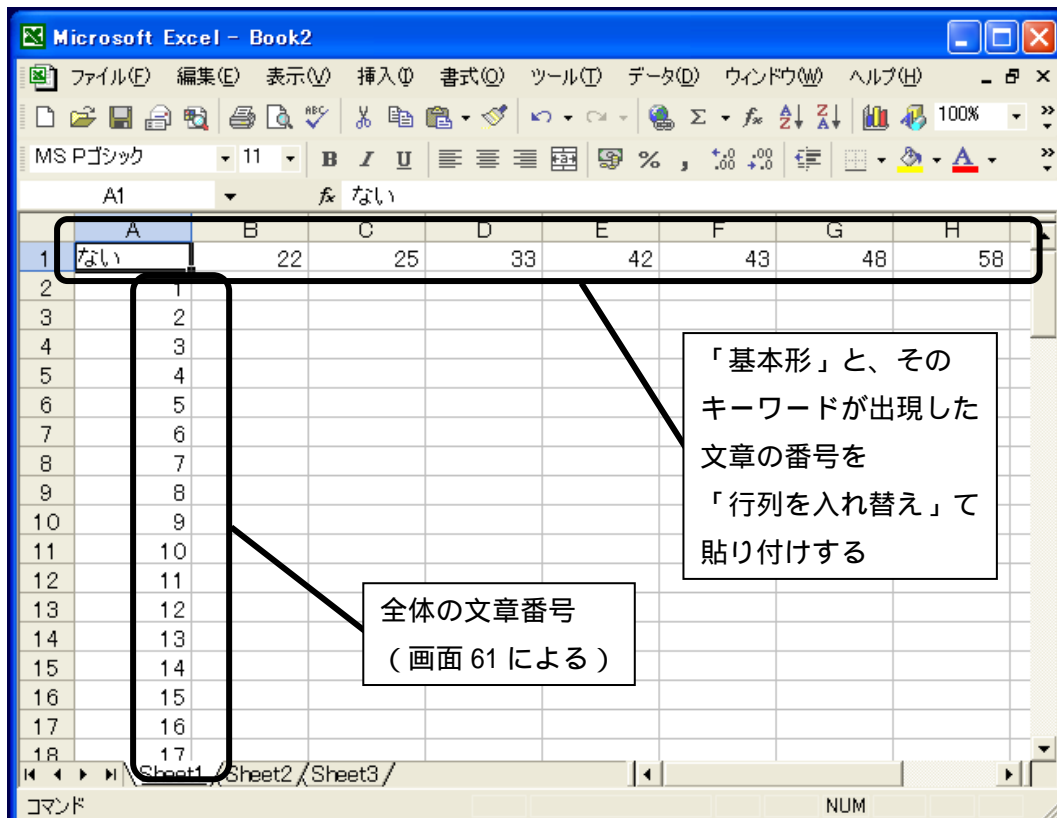


画面 61 マクロを使わず「1・0データ化」

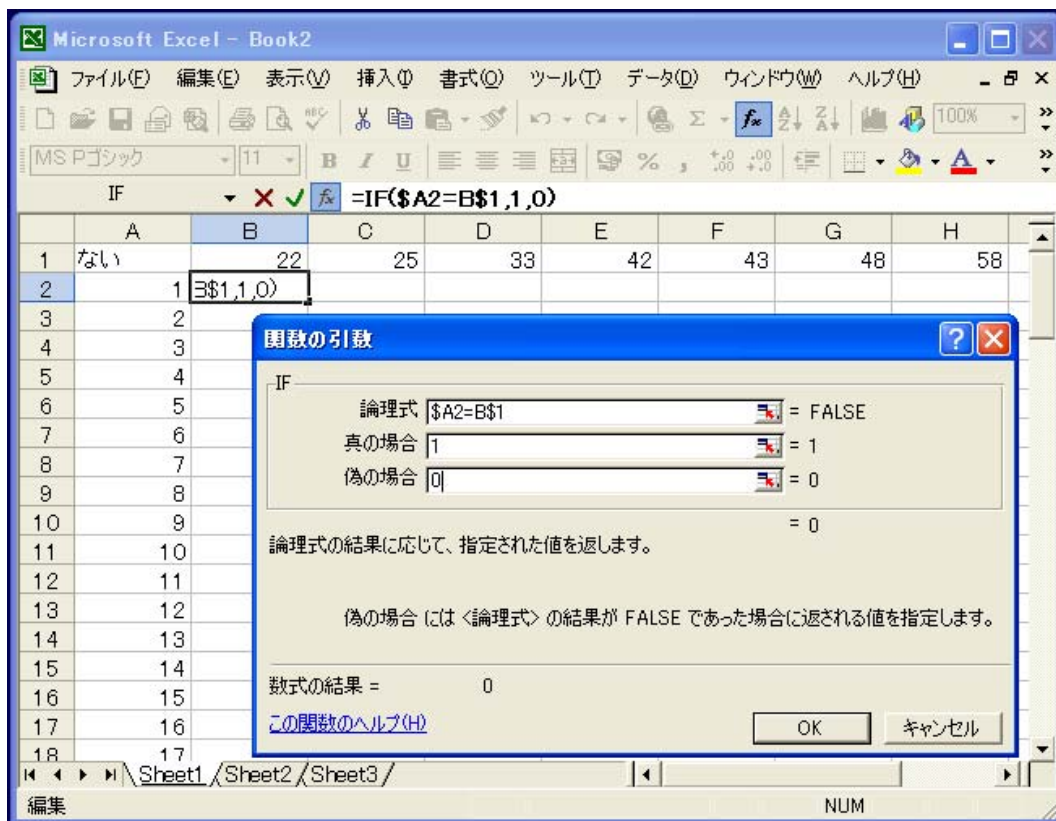


画面 62 マクロを使わず「1・0データ化」

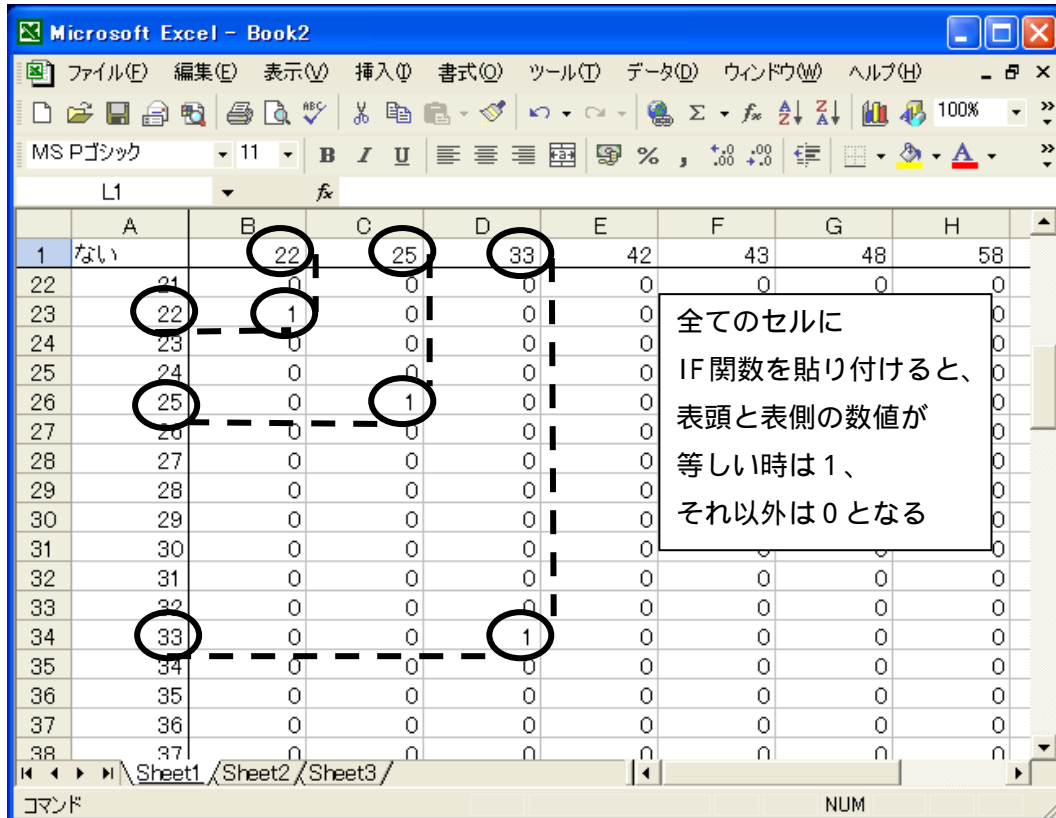




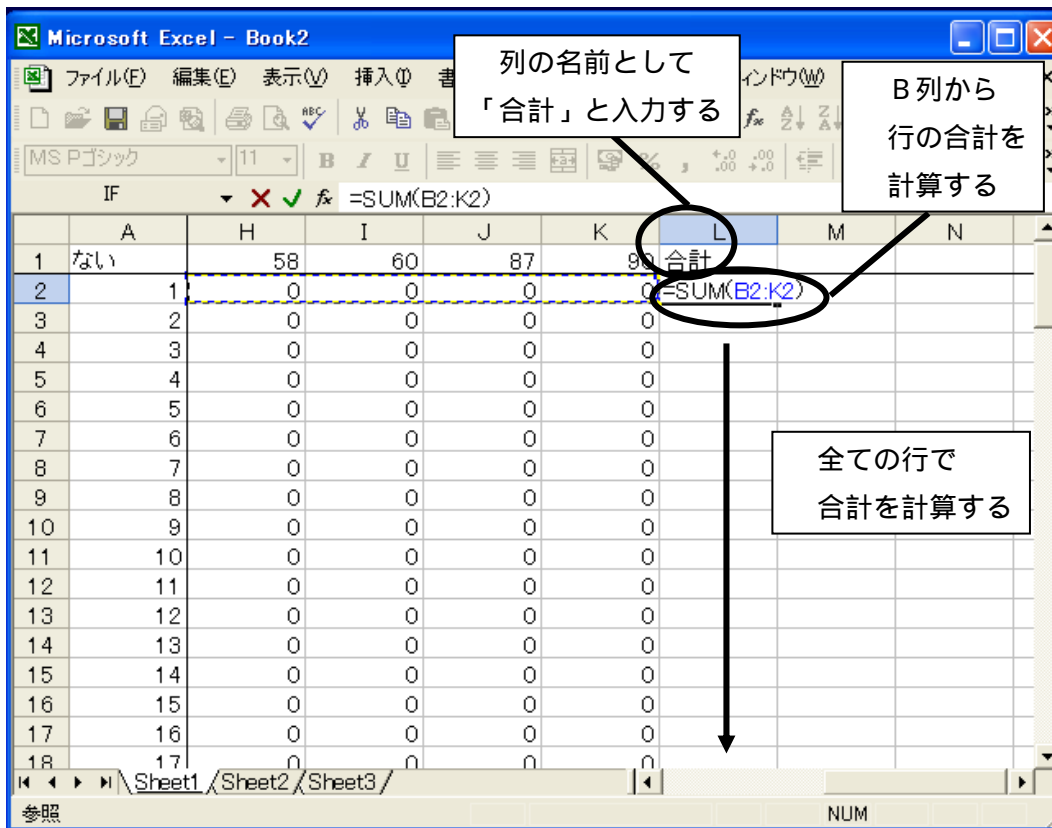
画面 63 マクロを使わず「1・0データ化」



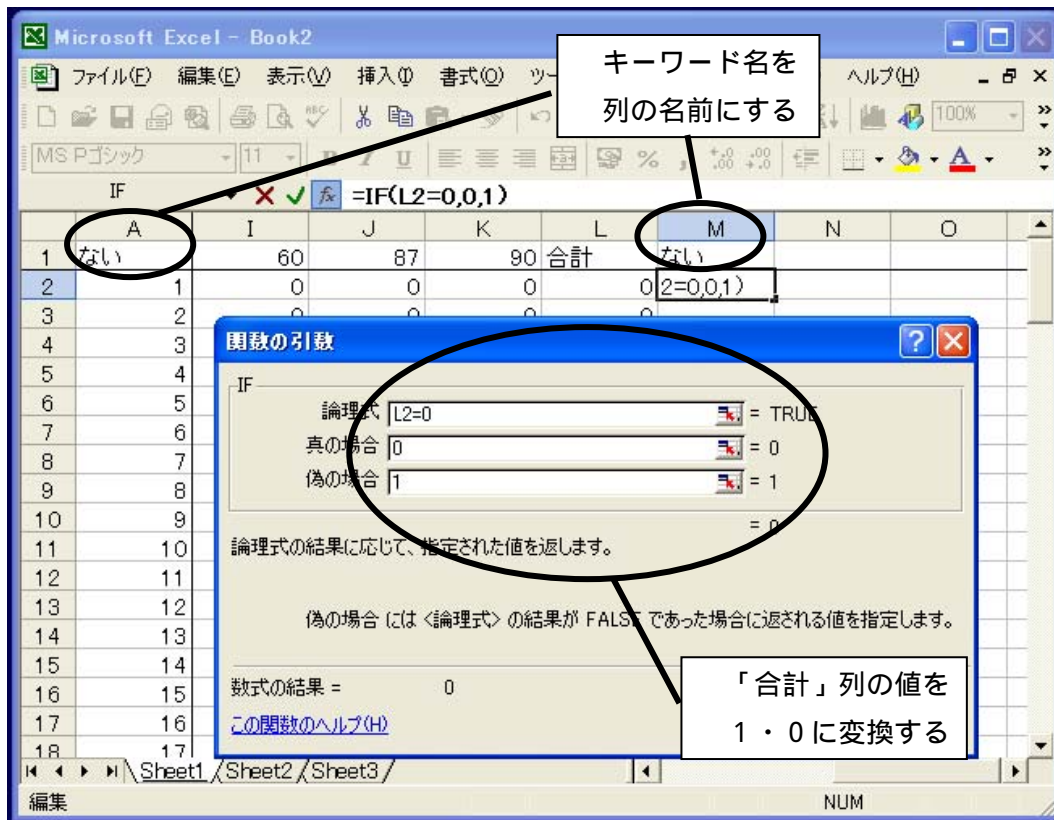
画面 64 マクロを使わず「1・0データ化」



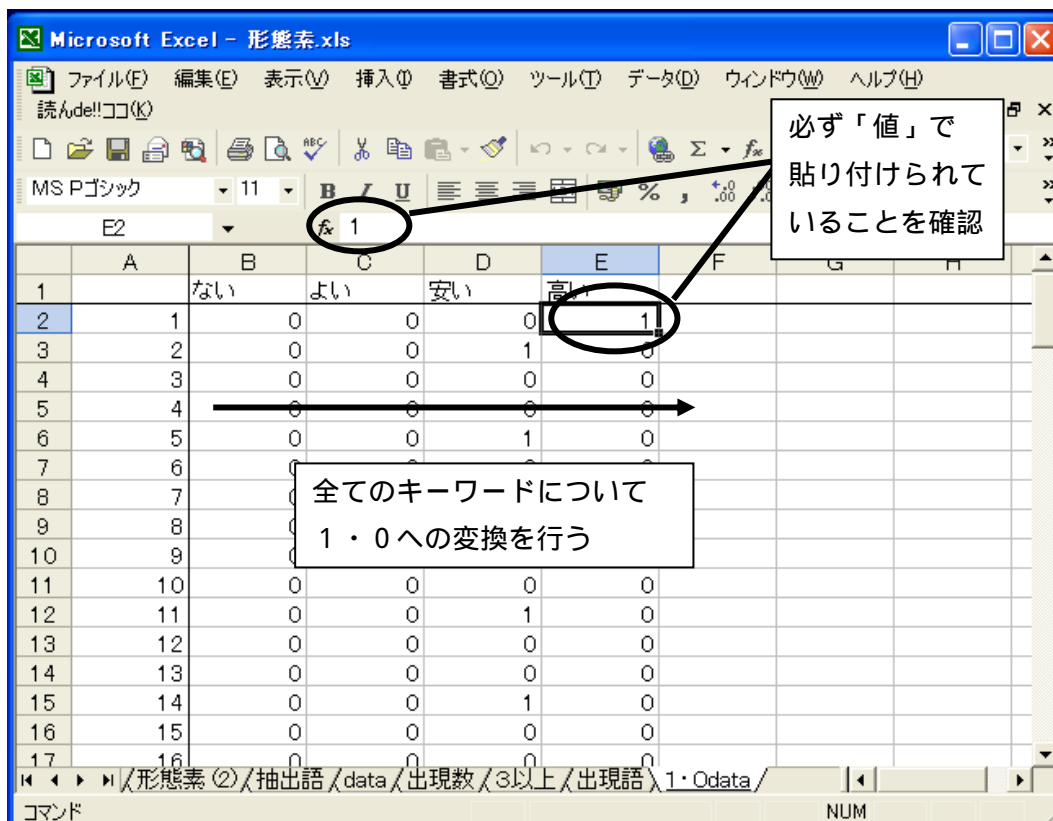
画面 65 マクロを使わず「1・0データ化」



画面 66 マクロを使わず「1・0データ化」



画面 67 マクロを使わず「1・0データ化」



画面 68 マクロを使わず「1・0データ化」

## 5 . むすび

本書では、低コストで簡便にテキストマイニングを行うために必要な1・0データファイルの作成手順を解説した。主な作業の流れは、文章を形態素に分解する、品詞情報と出現数でキーワードの絞り込みを行う、各文章におけるキーワード出現の有無を1・0で示す、というものである。

本書で示した手順は未だ試用段階であり、マクロによる作業の自動化はごく一部でしか行っていない。市販のテキストマイニング用ソフトウェアと比較すると、複雑な作業を要するため、膨大な量の文章を日々の業務として分析する場合には作業効率の点で不向きであると言わざるを得ない。

しかし、本手法は新たな投資を必要としないため、気軽に取り組むことができる。また、ほとんどの工程を手作業で行っているため、市販の専用ソフトウェアに見られるようなブラックボックスとなる部分がない、というメリットもある。そのため、「テキストマイニングがどのようなものか、試しにやってみたい」という方、「たくさんの文章データを持っているがどう整理したらよいかわからない」という方にとっては、テキストマイニングに触れる良いきっかけになるのではないかと考える。

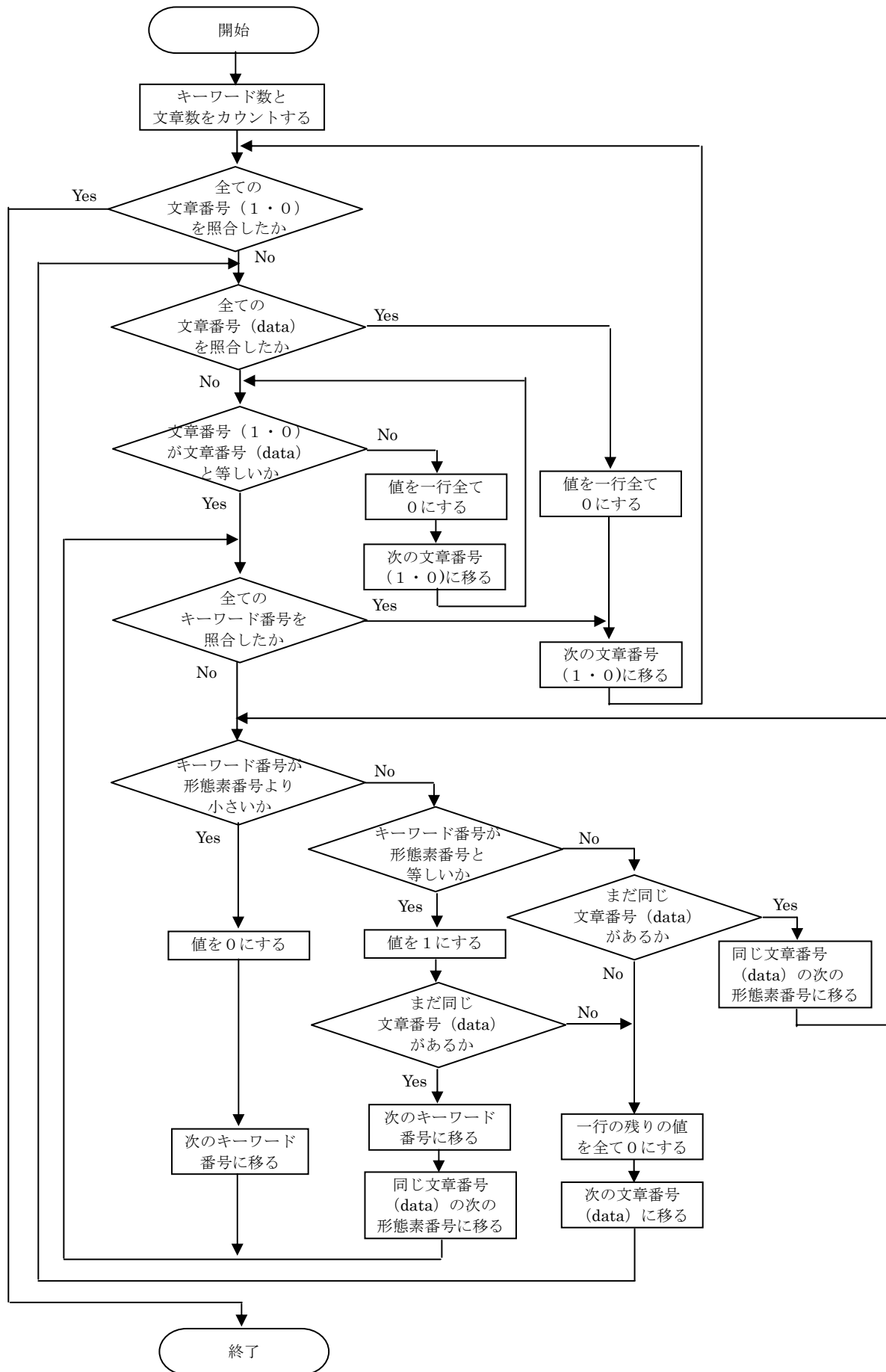
本書の末尾には、本手法で使用したマクロのコードを紹介している。また、本書では、マクロを使用しない場合の1・0データファイル作成手順についても補足説明を行っているので、必要に応じて参考にしていただきたい。

本手法の適用により、これまで放置されていた大量の文章データから、有益な情報が抽出されるようになれば幸いである。

## 参考文献

- 1) 市村由美・長谷川隆明・渡部勇・佐藤光弘(2001): テキストマイニング - 事例紹介、人工知能学会誌、16巻2号、192-200
- 2) 磯島昭代(2002): 米に関する自由記述回答文の分析、農業経営通信 214、26-29.
- 3) 磯島昭代(2004): 定性情報分析の方法 - テキストマイニング -、東北農研総合研究(A)第15号、10-14.
- 4) 磯島昭代・野中章久・清野誠喜(2004): テキストマイニングによるクレームデータの分析、農業経営研究、第42巻第1号(通巻120号) 148-152
- 5) 林俊克(2002): Excel で学ぶテキストマイニング入門、オーム社

付録 I マクロのフローチャート



## 付録Ⅱ マクロのコード

Sub 茶坊主くん()

```
Dim gyo As Integer, retu As Integer
Dim goku As Integer, keygyo As Integer
Dim bun As Integer, bunretu As Integer
Dim cnt As Integer, cnt2 As Integer
Dim kaz As Integer, bunkazu As Integer
Dim i As Integer, j As Integer
Dim motodata As String, ichizerodata As String
```

```
retu = 5 ' motodata の文章番号がある列
goku = 6 ' motodata の形態素番号がある列
bunretu = 1 ' ichizerodata の文章番号がある列
keygyo = 1 ' ichizerodata のキーワード番号がある行
motodata = "data" ' 元のデータのあるシート
ichizerodata = "1・0" ' 1・0 データを作成するシート
```

```
gyo = 2 ' motodata の文章番号の行：初期値
bun = 2 ' ichizerodata の文章番号の行：初期値
kaz = 0 ' キーワード数：初期値
bunkazu = 0 ' 全文章数：初期値
```

```
Do While Worksheets(ichizerodata).Cells(keygyo, kaz + bunretu + 1) <> ""
    kaz = kaz + 1
    ' キーワードの数をカウント
Loop
MsgBox "抽出したキーワードの数は" & kaz
```

```
Do While Worksheets(ichizerodata).Cells(bunkazu + keygyo + 1, bunretu) <> ""
    bunkazu = bunkazu + 1
    ' 全文章数をカウント
Loop
MsgBox "1・0 データ化する文章の数は" & bunkazu
```

```
Cells(keygyo, bunretu).Select
' セル (1, 1) をアクティブにする
Cells(keygyo, bunretu).Interior.ColorIndex = 6
' セル (1, 1) に色を付ける
```

```
Do While bun <= bunkazu + keygyo
' ichizerodata の文章番号の行を全て参照するまで行う。
```

```
Worksheets(ichizerodata).Cells(keygyo, bunretu).Value = bun - 1
' icizerodata のセル (1, 1) に、現在操作中の文章番号を記入
```

```
If Worksheets(motodata).Cells(gyo, retu) <> "" Then
' motodata の文章番号のセルが空白でなければ
```

```

Select Case Worksheets(ichizerodata).Cells(bun, bunretu)
' ichizerodata の文章番号が

Case Is < Worksheets(motodata).Cells(gyo, retu)
' motodata の文章番号より小さいとき
    For i = 2 To kaz + 1
        Worksheets(ichizerodata).Cells(bun, i).Value = 0
        ' 値は一行全て 0
    Next i

, -----

Case Is = Worksheets(motodata).Cells(gyo, retu)
' (ichizerodata の文章番号が) motodata の文章番号と等しいときは
    cnt = bunretu + 1
    ' ichizerodata の 2 列目 (キーワードのある列) から

    Do While cnt <= kaz + bunretu
    ' ichizerodata の全てのキーワード番号の列を参照するまで繰り返す

        Select Case Worksheets(ichizerodata).Cells(keygyo, cnt)
        ' ichizerodata のキーワード番号が

        Case Is < Worksheets(motodata).Cells(gyo, goku)
        ' motodata の形態素番号より小さいときは

            Worksheets(ichizerodata).Cells(bun, cnt).Value = 0
            ' 値を 0 にして
            cnt = cnt + 1
            ' 次のキーワード番号の列へ
            If cnt > kaz + bunretu Then
            ' もし, ichizerodata の全てのキーワード番号を参照してしまったら

                Do While Worksheets(motodata).Cells(gyo, retu) = _
                    Worksheets(motodata).Cells(gyo + 1, retu)
                ' motodata の文章番号が変わるまで行を進める
                gyo = gyo + 1
            Loop

            End If

        Case Is = Worksheets(motodata).Cells(gyo, goku)
        ' motodata の形態素番号と等しいときは

            Worksheets(ichizerodata).Cells(bun, cnt).Value = 1
            ' 値を 1 にする

            If Worksheets(motodata).Cells(gyo, retu) = _
                Worksheets(motodata).Cells(gyo + 1, retu) Then
            ' motodata の次の文章番号が変わらない場合

```



```

gyo = gyo + 1
' 同じ文章番号の次の形態素番号の行に移り
cnt = cnt + 1
' ichizerodata の次のキーワード番号に移る

If cnt > kaz + bunretu Then
' もし、ichizerodata の全てのキーワード番号を参照してしまったら

    Do While Worksheets(motodata).Cells(gyo, retu) = _
        Worksheets(motodata).Cells(gyo + 1, retu)
    ' motodata の文章番号が変わるまで行を進める
        gyo = gyo + 1
    Loop

End If

Else
' motodata の文章番号が変わった場合
    For i = cnt + 1 To kaz + 1
        Worksheets(ichizerodata).Cells(bun, i).Value = 0
    ' 残りの値を 0 にして
    Next i
    Exit Do
' ループを抜ける
End If

Case Is > Worksheets(motodata).Cells(gyo, goku)
' motodata の形態素番号より大きいとき

    If Worksheets(motodata).Cells(gyo, retu) = _
        Worksheets(motodata).Cells(gyo + 1, retu) Then
    ' motodata の次の文章番号が同じ場合

        gyo = gyo + 1
        ' 同じ文章番号の次の形態素番号の行に移り

    Else
    ' motodata の文章番号が変わった場合

        For i = cnt To kaz + 1
            Worksheets(ichizerodata).Cells(bun, i).Value = 0
        ' 残りの値を 0 にして
        Next i
        Exit Do
        ' ループを抜ける
    End If

End Select

```

```

Loop
' ichizerodata の最後のキーワード番号の列を参照するまでループ

gyo = gyo + 1
' motodata の次の文章番号の行に移る

-----

Case Is > Worksheets(motodata).Cells(gyo, retu)
' (ichizerodata の文章番号が) motodata の文章番号より大きいときは

End Select

Else
' motodata の文章番号が空白になったら

For i = 2 To kaz + 1
Worksheets(ichizerodata).Cells(bun, i).Value = 0
' 値は全て 0 にする
Next i
End If

bun = bun + 1
' ichizerodata の次の文章番号について行う
Loop
End Sub

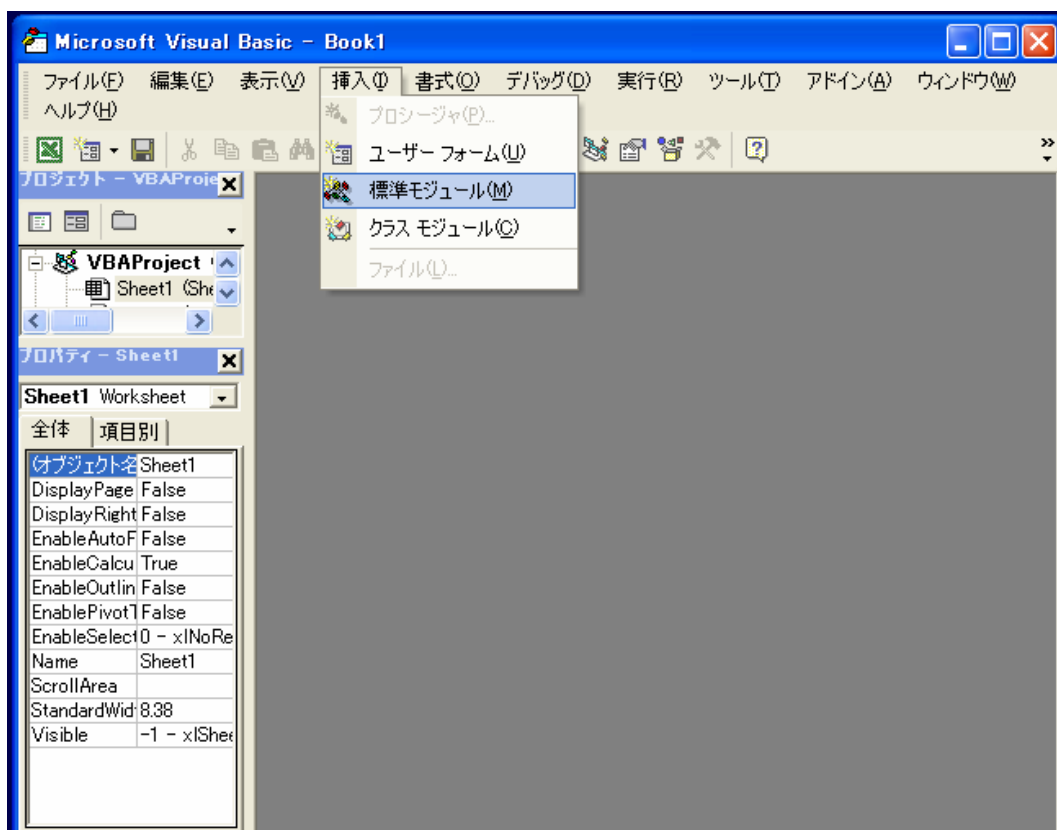
```

### 付録Ⅲ 「茶坊主くん.txt」 からVBファイルを作成する

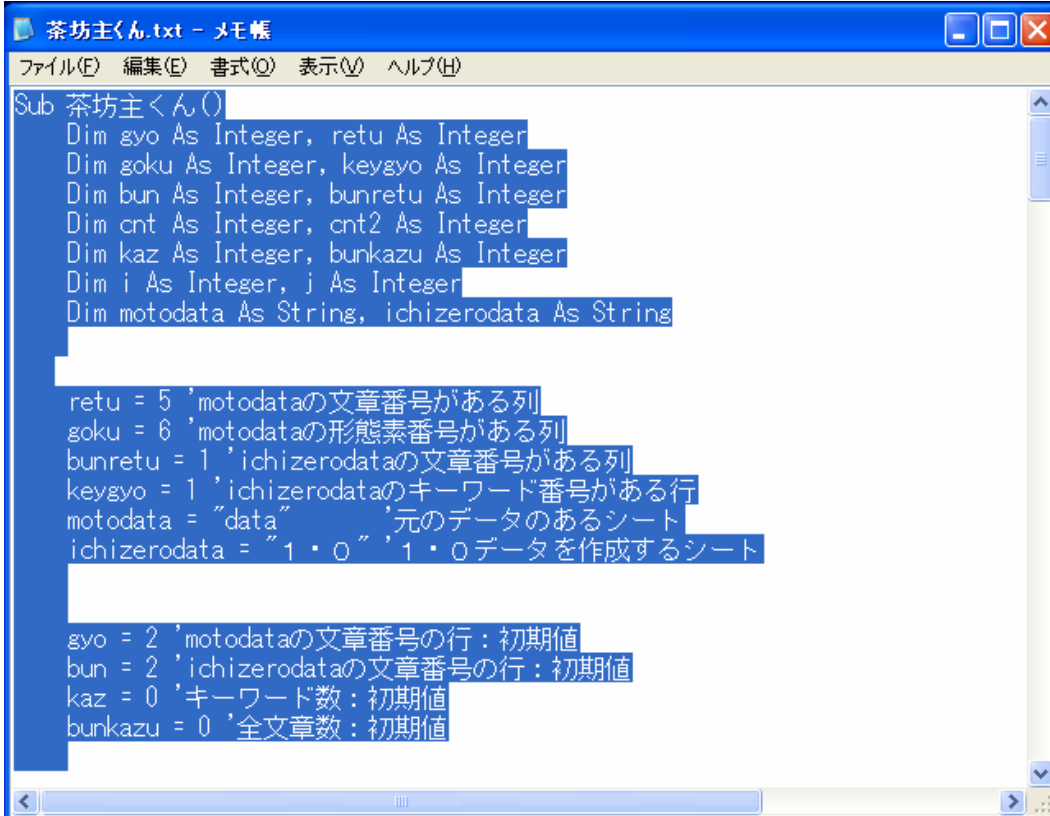
本文中では、「茶坊主くん」のマクロは作成済みという前提で解説しており、公開しているマクロのコードを参考に各自で対応していただくことにしている。しかし、実際にコードを入力する作業は面倒と感じられる方も多いと思われるので、テキスト形式でマクロのコードを配布することにした。テキスト形式としたのは、セキュリティ上の問題に配慮したためであるが、テキスト形式のファイルでは、そのままマクロとして実行することはできない。

そこで、以下では、テキスト形式のファイル「茶坊主くん.txt」から、マクロが実行可能なVBファイル「茶坊主くん.bas」を作成する手順を紹介する。なお、ここではVBファイルの作成について述べるが、もちろん、1・0データファイル作成の過程で、テキストのコードをVBE (Visual Basic Editor) のコードウィンドウに直接コピー&ペーストして実行していただいても一向に構わない。

1. Excel の「ツール(T)」 - 「マクロ(M)」 - 「Visual Basic Editor(V)」を選択してVBEを起動し、「挿入(I)」 - 「標準モジュール(M)」をクリックする。

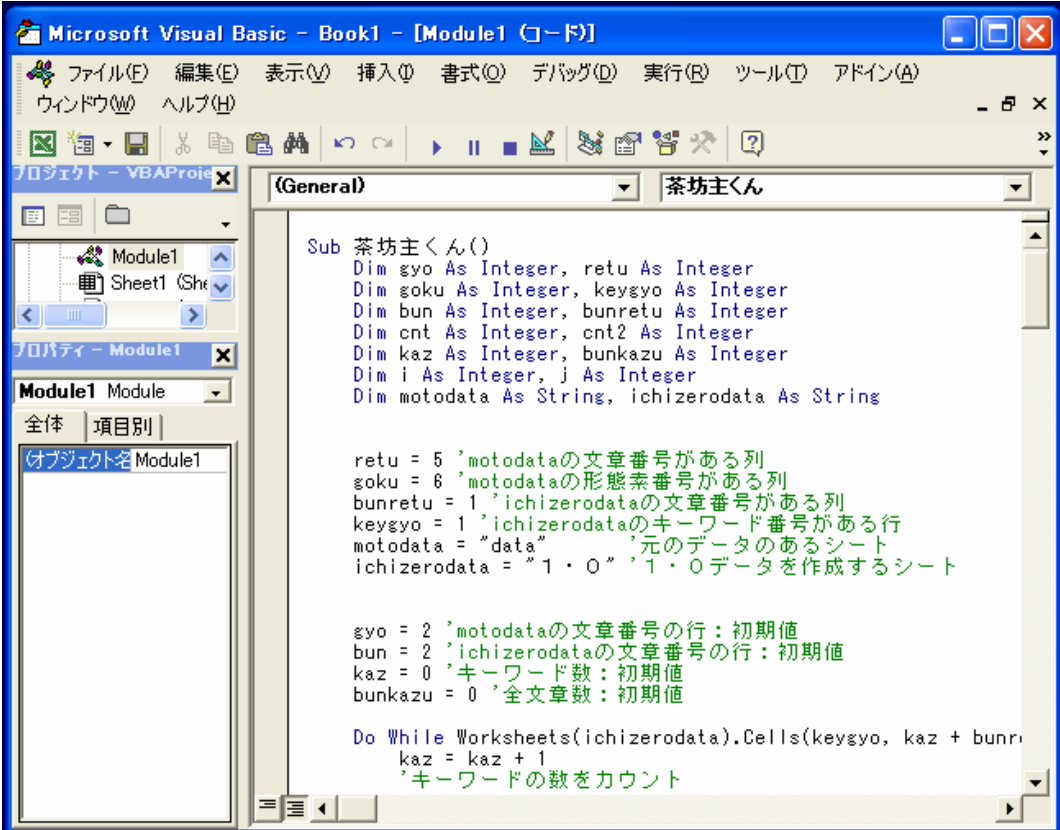


2. 「茶坊主くん.txt」のファイルを開き、コードをコピーする。



```
Sub 茶坊主くん()  
Dim gyo As Integer, retu As Integer  
Dim goku As Integer, keygyo As Integer  
Dim bun As Integer, bunretu As Integer  
Dim cnt As Integer, cnt2 As Integer  
Dim kaz As Integer, bunkazu As Integer  
Dim i As Integer, j As Integer  
Dim motodata As String, ichizerodata As String  
  
retu = 5 'motodataの文章番号がある列  
goku = 6 'motodataの形態素番号がある列  
bunretu = 1 'ichizerodataの文章番号がある列  
keygyo = 1 'ichizerodataのキーワード番号がある行  
motodata = "data" '元のデータのあるシート  
ichizerodata = "1・0" '1・0データを作成するシート  
  
gyo = 2 'motodataの文章番号の行：初期値  
bun = 2 'ichizerodataの文章番号の行：初期値  
kaz = 0 'キーワード数：初期値  
bunkazu = 0 '全文章数：初期値
```

3. VBEのコードウィンドウにコードを貼り付ける。



```
Microsoft Visual Basic - Book1 - [Module1 (コード)]  
ファイル(F) 編集(E) 表示(V) 挿入(I) 書式(O) デバッグ(D) 実行(R) ツール(T) アドイン(A)  
ウィンドウ(W) ヘルプ(H)  
プロジェクト - VBAProject  
Module1  
Sheet1 (She)  
プロパティ - Module1  
Module1 Module  
全体 項目別  
オブジェクト名 Module1  
(General) 茶坊主くん  
Sub 茶坊主くん()  
Dim gyo As Integer, retu As Integer  
Dim goku As Integer, keygyo As Integer  
Dim bun As Integer, bunretu As Integer  
Dim cnt As Integer, cnt2 As Integer  
Dim kaz As Integer, bunkazu As Integer  
Dim i As Integer, j As Integer  
Dim motodata As String, ichizerodata As String  
  
retu = 5 'motodataの文章番号がある列  
goku = 6 'motodataの形態素番号がある列  
bunretu = 1 'ichizerodataの文章番号がある列  
keygyo = 1 'ichizerodataのキーワード番号がある行  
motodata = "data" '元のデータのあるシート  
ichizerodata = "1・0" '1・0データを作成するシート  
  
gyo = 2 'motodataの文章番号の行：初期値  
bun = 2 'ichizerodataの文章番号の行：初期値  
kaz = 0 'キーワード数：初期値  
bunkazu = 0 '全文章数：初期値  
  
Do While Worksheets(ichizerodata).Cells(keygyo, kaz + bunr  
kaz = kaz + 1  
'キーワードの数をカウント
```

4. 「ファイル(F)」－「ファイルのエクスポート(E)」を選択し、ファイル名を「茶坊主くん」として保存する。

