目次

表紙・CD-R デザイン：篠塚朋子

1．テキストマイニングとは

心理学の質問紙調査では、リッカート法(例えば、「私は、心理学が好きである。」という記述に対して「非常にあてはまる」から「全く当てはまらない」まで５段階の数字のどれか一つに○をつける場合)など、個人の態度や行動傾向、評価などを数量的に計量する方法が多く取られる。しかし、個人の持つ素朴な意見や、尺度では捉えきれないものを捉えるために自由記述を使いたいと思う人も多いだろう。

従来、このような自由記述に対する分析方法としては、カテゴリの定義を決めて、それぞれのカテゴリに該当する記述の数をカウントする方法や、自由記述を事例としてそのまま提示する方法などが取られてきた。

これに対して、テキストマイニングという方法を使うと、自由記述文の単語の出現頻度、あるいは共起や係り受けなどの関係に基づき、その記述分の特徴を抽出することで、テキストを相対的に客観的・機械的に処理することができる。特に、元にする自由記述のデータが、非常に多くて読み込むことが不可能な場合に、テキストマイニングは、特にその効力を発揮し、その自由記述のデータの特徴を効率的に要約することができる。

他方で、テキストマイニングは、文章の意味を理解しての分析ではなく表層的な言葉遣いの違いや表記の揺れによって大きな影響を受けることなどの限界も指摘されている。

2．TEXTORVA について

TEXTORVA とは、テキストマイニングを簡易に体験するためのエクセルアドインです。

本アドインを使うことで、テキストマイニングで用いられる以下の機能を行うことができます。

- 品詞ごとの出現頻度の分析
- 特定タグ(属性)とのクロス集計
- 数量化Ⅲ類による分析とそのグラフ化
  （数量化Ⅲ類の計算については、青木(2001：下記)に基づく
- 文脈において各語彙がどのように使われているかの検討補助

以下の条件を備えている環境下での使用を想定しています。

- Windowsパソコン(XP、Vista、および７で動作確認)
- Microsoft Excel(2003および2007で動作確認)
- Microsoft Access(2003および2007で動作確認)
- 茶筌(奈良先端科学技術大学院大学松本研究室で開発された形態素解析ソフト)

機能は、高性能なテキストマイニングソフトに劣りますが、本アドインでは、簡単な操作で、テキストマイニングを体験することができます。

青木(2001)
http://aoki2.si.gunma-u.ac.jp/lecture/stats-by-excel/vba/html/qt3.html

3．TEXTORVA のセットアップ

3.1　ɳ ɸ ȣ ɱ ş ɤ Ƶ ɱ ɬ œ ɸ Ã Ȓ

ȸ　の　　ŏ Ȓ ɳ ɸ ȣ ɱ ş ɤ Ƶ ɱ ɬ œ ɸ Ã è Ã Ƴ
ǰ こに移すか迷った場合は、デスクトップに移しておくとよいで Ã Ô Ă Ƴ

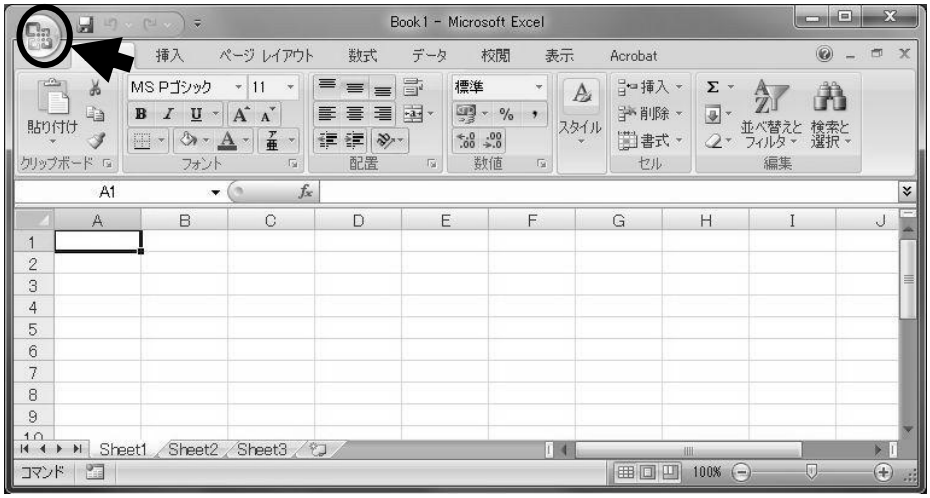
3.2　茶筌のセットアップ

(1) 茶筌を ɤ Ƶ ɱ ɬ œ ɸ Ã ǫ インストールする
「Cha21244sp5.exe」を実行し、画面の指示に従います。下記の画面が出たら、矢印の部分をクリックしてインストールします。ディレクトリは変更しないでください。

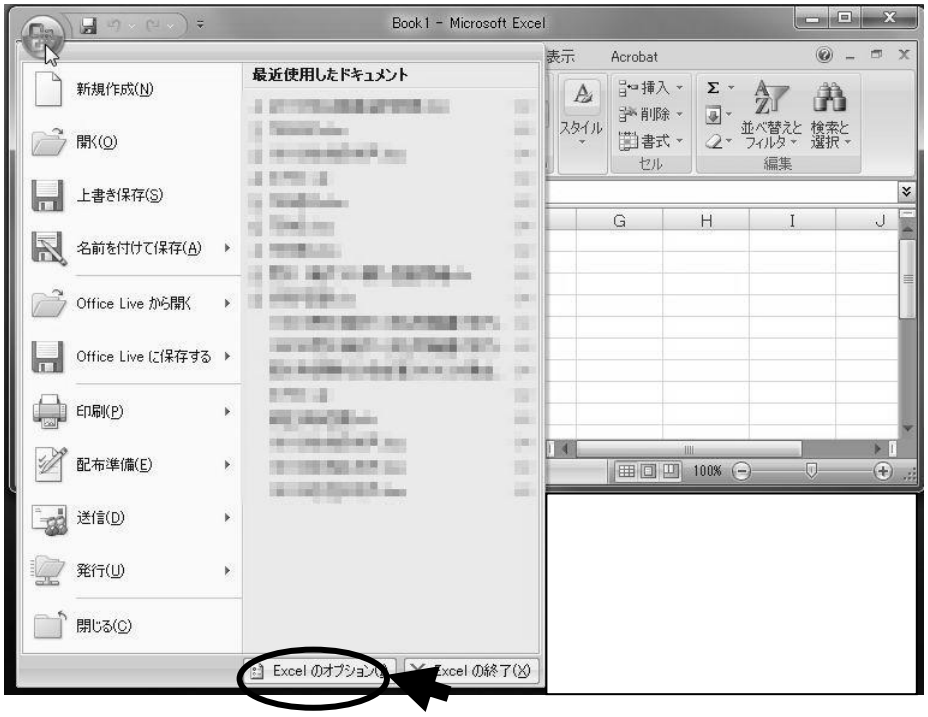
Tips：付属 CD-R がない場合、下記の URL からダウンロードできます。
http://www.k2.dion.ne.jp/~kokoro/mivurix/textorva/
茶筌については
http://chasen.naist.jp/hiki/ChaSen/?茶筌の配布

3.3　アドインの設定

⑴Excel を開いたら Office ボタンを押して、メニューを出します。

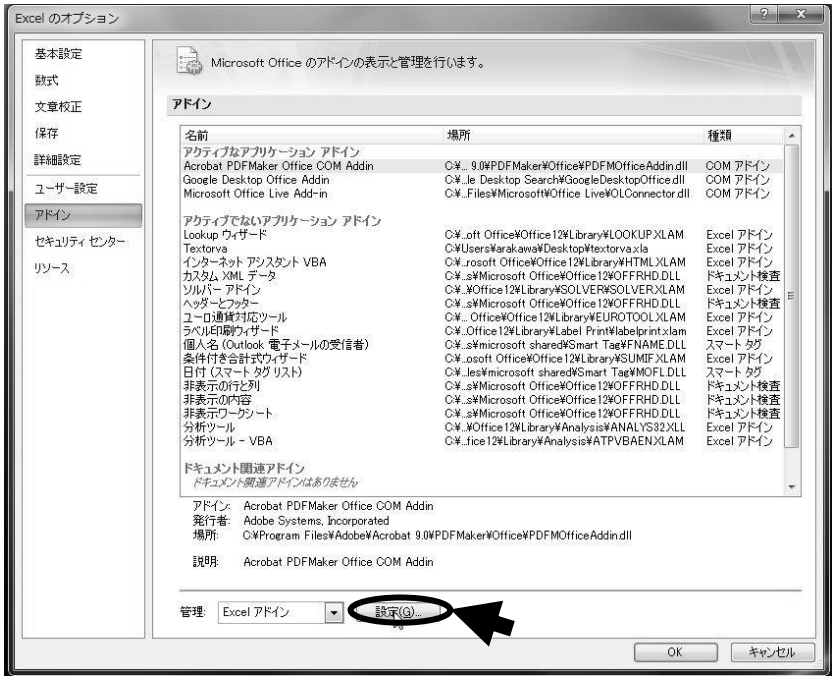


⑵Excel のオプションを選びます。

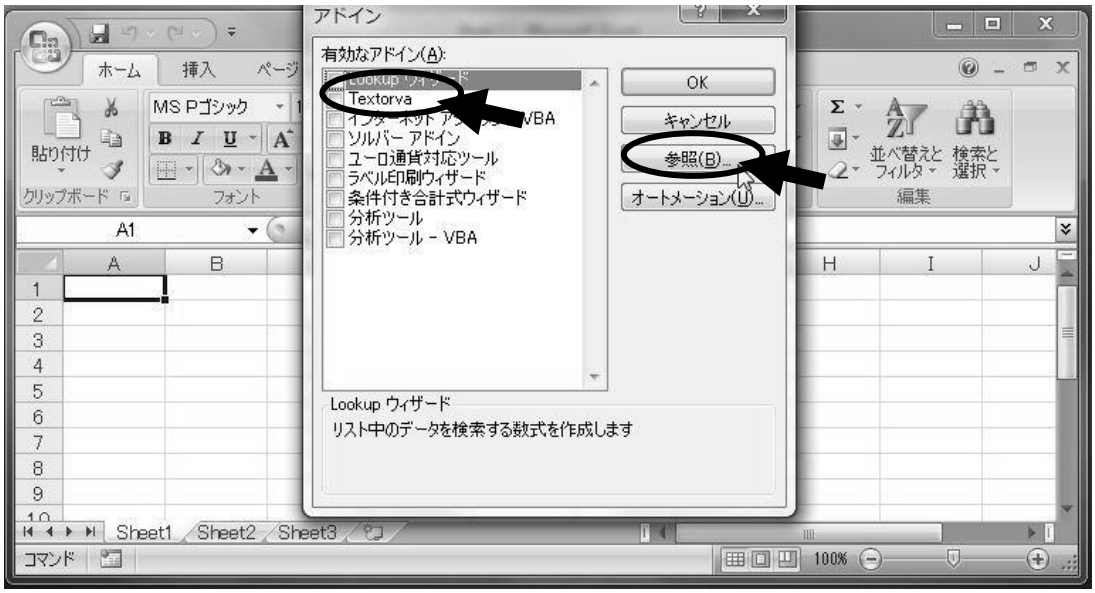


⑶アドインを選びます。

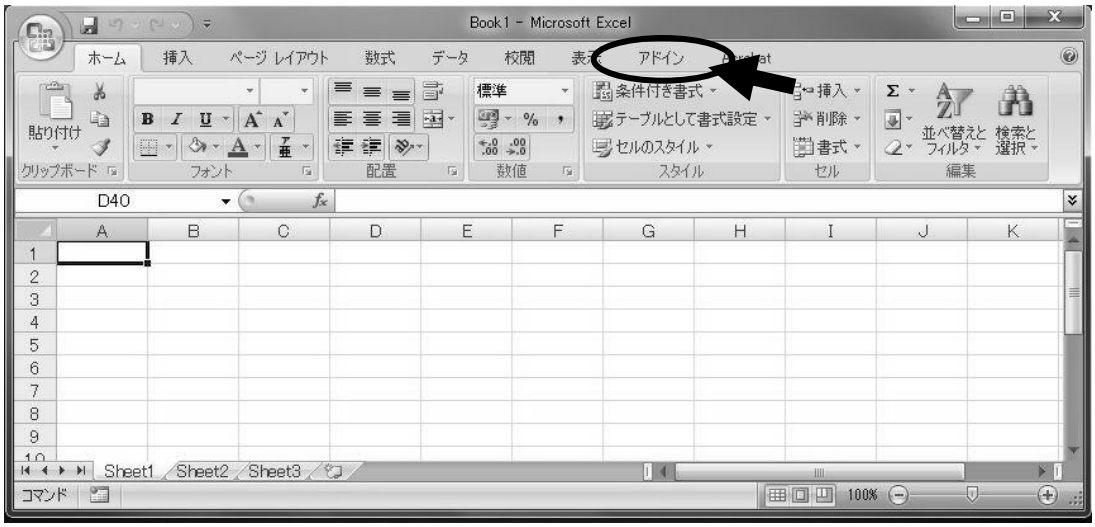
⑷設定を選びます。



⑸選択肢にあれば、「Textorva」を選ぶ。選択肢になければ、「参照」をクリックして、保存した場所を指定する。



⑹アドインというコマンドが増えている。



3.4データの加工

本アドインで分析するためには、縦に個々のデータ、横に変数が並ぶようにデータを配置しておきます。

例：顔文字の使用に関するグループインタビューの模擬データ。

	A	B	C
1	話者 ID	世代	発言(自由記述)
2	1	40 代	娘からのメールについているけど、自分からは使わないかなぁ
3	2	30 代	友達から来た時はつけないと
4	3	20 代	友達に送る時は毎回顔文字使いますよ。入れないとさみしいし・・。
5	4	40 代	後輩からきたらちょっとどうしようかと思う。
6	5	30 代	たしかにあまり知らない人から送られてくるとちょっと驚く。
7	6	20 代	でも逆に知らない人と仲良くなるために顔文字を使うこともあるよ。
8	1	40 代	顔文字をつけずに返したら冷たいといわれたことはあるなぁ。
9	2	30 代	顔文字つけて送ったのに、文字だけのメールが来たらさみしいかも。
10	3	20 代	顔文字があったほうが、かわいいし！！

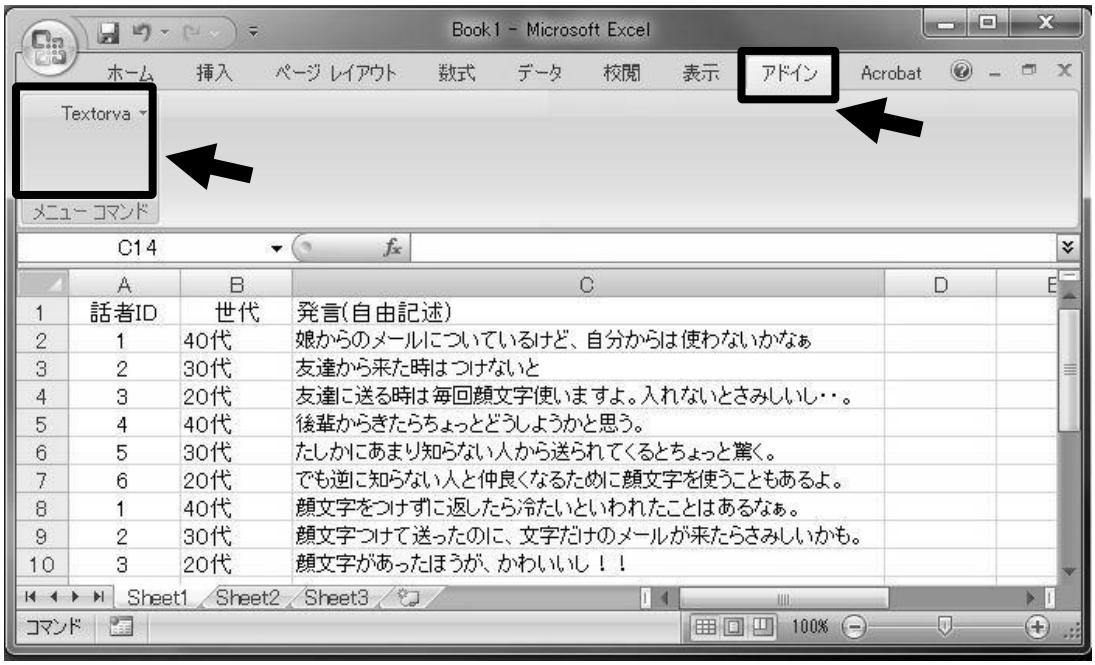
１行目にはラベルを入れておきます。２行目以降が１行ずつ個々のデータを並べます。

この模擬データでは、B 列に分析の際に比較するカテゴリデータを、C 列に分析対象になる自由記述データが入っています。

4．分析１：どんな単語がどれくらい使われているか調べてみよう

4．１　TEXTORVA を起動しよう。

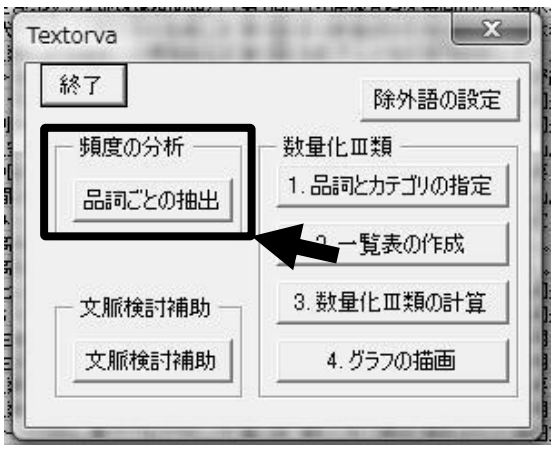
「アドイン」のタブにある「Textorva」をクリックして、アドインを起動する。



	A	B	C	D
1	話者ID	世代	発言(自由記述)	
2	1	40代	娘からのメールについているけど、自分からは使わないかなぁ	
3	2	30代	友達から来た時はつけないと	
4	3	20代	友達に送る時は毎回顔文字使いますよ。入れないとさみしいし・・。	
5	4	40代	後輩からきたらちょっとどうしようかと思う。	
6	5	30代	たしかにあまり知らない人から送られてくるとちょっと驚く。	
7	6	20代	でも逆に知らない人と仲良くなるために顔文字を使うこともあるよ。	
8	1	40代	顔文字をつけずに返したら冷たいといわれたことはあるなぁ。	
9	2	30代	顔文字つけて送ったのに、文字だけのメールが来たらさみしいかも。	
10	3	20代	顔文字があったほうが、かわいいし！！	

4．２　「品詞ごとの抽出」を選ぼう。

起動されたフォームのボタンの中の「品詞ごとの抽出」を選んでクリックします。



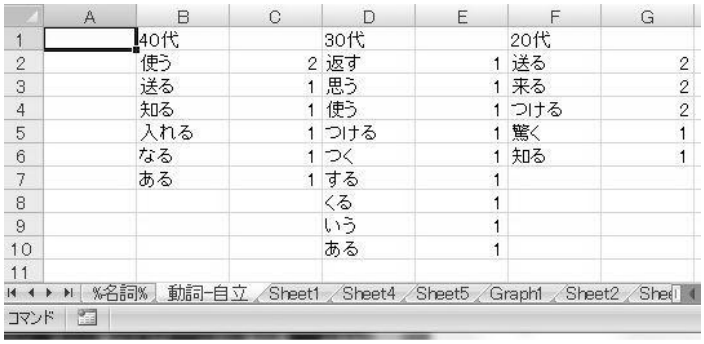
４．３　変数を設定しよう

①分析したい自由記述が含まれた行をデータ行として指定します。②分析する品詞を選択します。③主要なものだけ見ればいいときは、特定の度数以上出現した単語を拾うことができます。その場合は最小値のところに数値を入力します。④また必要に応じてカテゴリ行を設定します（カテゴリ行を指定すると、カテゴリごとに度数が算出されます。）設定したら、⑤の「分析」を押して下さい。

例：模擬データの場合、世代によって、顔文字について説明する言葉の出現度数を検討する場合には、世代をカテゴリ行に指定します。

４．４　出力の読み方

品詞ごとにエクセルのタブが追加されます。またカテゴリごとに集計が示されます。

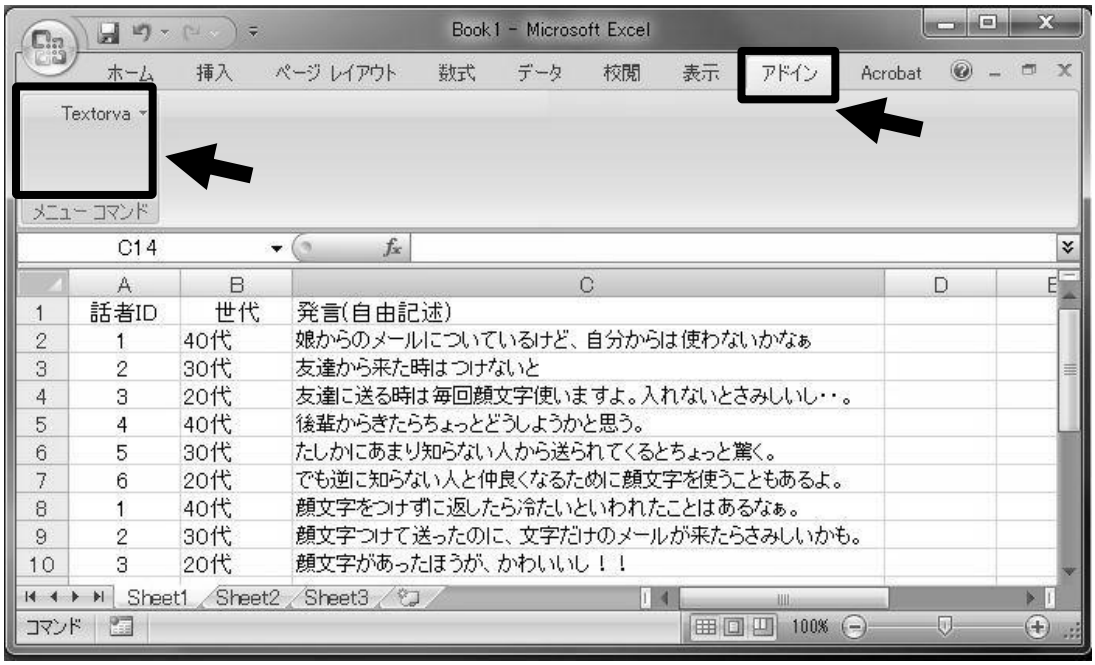


	A	B	C	D	E	F	G
1		40代		30代		20代	
2		使う	2	返す	1	送る	2
3		送る	1	思う	1	来る	2
4		知る	1	使う	1	つける	2
5		入れる	1	つける	1	驚く	1
6		なる	1	つく	1	知る	1
7		ある	1	する	1		
8				くる	1		
9				いう	1		
10				ある	1		
11							

5．分析２：単語間のつながりを見てみよう

5．１　TEXTORVA を起動しよう。

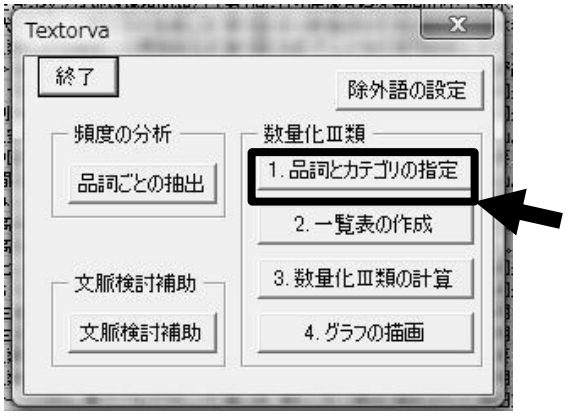
「アドイン」のタブにある「Textorva」をクリックして、アドインを起動する。



	A	B	C	D
1	話者ID	世代	発言(自由記述)	
2	1	40代	娘からのメールについているけど、自分からは使わないかなぁ	
3	2	30代	友達から来た時はつけないと	
4	3	20代	友達に送る時は毎回顔文字使いますよ。入れないとさみしいし・・。	
5	4	40代	後輩からきたらちょっとどうしようかと思う。	
6	5	30代	たしかにあまり知らない人から送られてくるとちょっと驚く。	
7	6	20代	でも逆に知らない人と仲良くなるために顔文字を使うこともあるよ。	
8	1	40代	顔文字をつけずに返したら冷たいといわれたことはあるなぁ。	
9	2	30代	顔文字つけて送ったのに、文字だけのメールが来たらさみしいかも。	
10	3	20代	顔文字があったほうが、かわいいし！！	

5．２　「品詞とカテゴリの指定」をしよう。

起動されたフォームのボタンの中の「品詞ごとの抽出」を選んでクリックします。

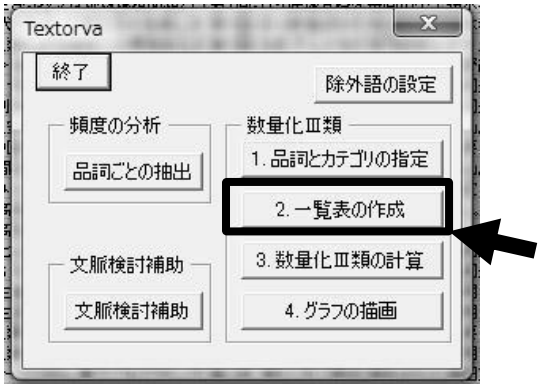


５．３　オプションを指定しよう

①分析する対象行を指定し、②分析する品詞、③カテゴリ行を指定します。④さらに必要があれば出現度数の下限を設定し、決定をクリックします。すると出現度数以上の度数で現れた単語とカテゴリ行にあるラベルが、新しいタブで現れます。不要なカテゴリなどを削除して分析することもできます。

５．４　「一覧表の作成」をしよう。

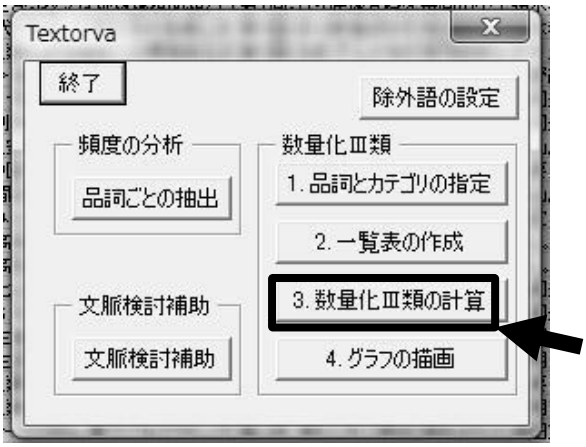
「一覧表を作成」をクリックします。先の操作でできた語のリストについて共起行列が作成されます。



５．５　数量化Ⅲ類の計算をしよう。

数量化Ⅲ類とは、林知己夫によって開発された分析法で、共起関係(textorva の場合、ある単語とある単語が同じ文の中に生起する頻度)から、どのような反応とどのような反応とが類似性のある出現傾向があるのかを把握することができます。

量的分析における主成分分析や因子分析に該当する方法です。



５．６　数量化Ⅲ類の表の見方

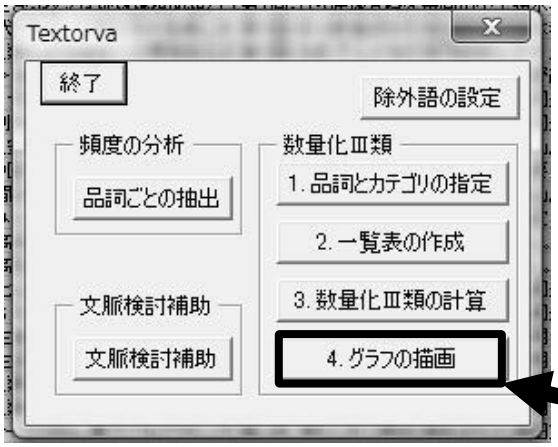
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	table	解		固有値		カテゴリ	軸1	軸2	軸3			ケース1	1.024518	-1.61035	-0.40675
2			1	0.666667		40代	0.83652	-1.31484	-0.28486			ケース2	0.807915	0.647071	1.38873
3			2	0.666667		30代	0.57123	1.44992	1.37543			ケース3	-1.83243	0.963276	-0.98198
4			3	0.490468		送る	-1.31932	-1.05666	0.68771			ケース4	0.108302	-1.12871	0.46291
5			4	0.333333		使う	0.74809	-0.39326	0.56972			ケース5	0.699614	1.775781	1.963961
6			5	0.176198		つける	-1.49618	0.78651	-0.68771			ケース6	-1.61583	-1.29414	0.981981
7						ある	0.65966	0.52833	-1.66029			ケース7	0.916217	-0.48164	-0.65465
8												ケース8	-0.1083	1.128709	-0.46291
9												ケース9	-1.72413	-0.16543	-1.3E-15
10												ケース10	0.807915	0.647071	-2.37071

上記の例では、カテゴリ数が６つなので、今回のデータのばらつきを説明するのに、それより１つ少ない５つの解が得られています。**固有値**は、それぞれの解が各データのばらつきをどの程度説明できるかを示しています。

また、各カテゴリの内容に、上位３つの解の固有値で重みづけたものが中央にある３つの軸の値になります。さらに、右側は、元の１０の会話を同じく重みづけたものです。ただし数量化Ⅲ類の結果は、次のページにあるグラフで検討するのが一般的です。

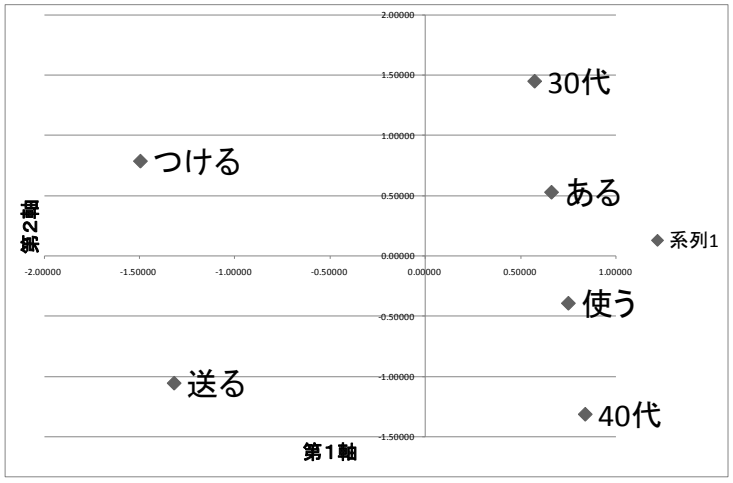
５．６　「グラフの描画」をしよう。

　数量化Ⅲ類の結果が表示されている場面で「グラフの描画」を押すと、結果を図的に表現することができます。



５．７　グラフの見方

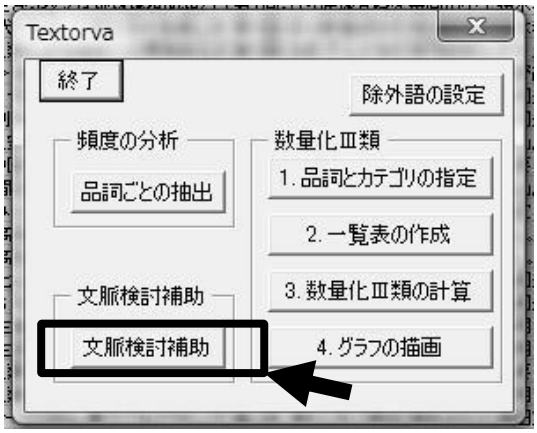
　第１軸と第２軸の値をもとにプロットされます。この図の中で近くに位置しているものは、共起している割合が高いことを示しています。



６　分析３：各単語がどのような文脈で使われているか見てみよう

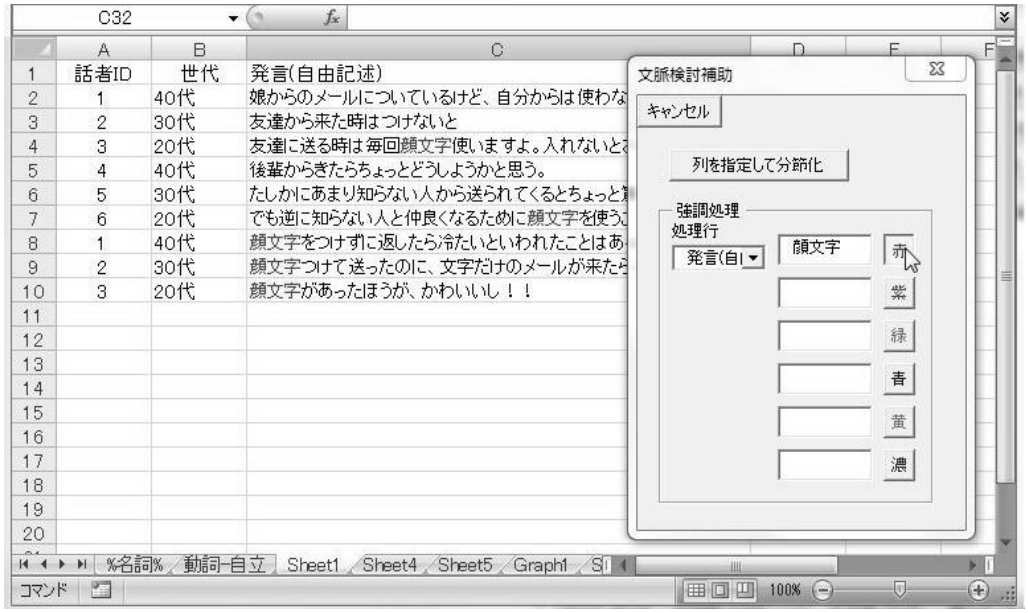
６．１　「文脈検討補助」を選択しよう。

Textorva を起動し、「文脈検討補助」をクリックします。



６．２　文脈検討補助の操作の仕方

検索する文字列を記入して、色のボタンを押すと、その単語を色付けして検索しやすくすることができます。

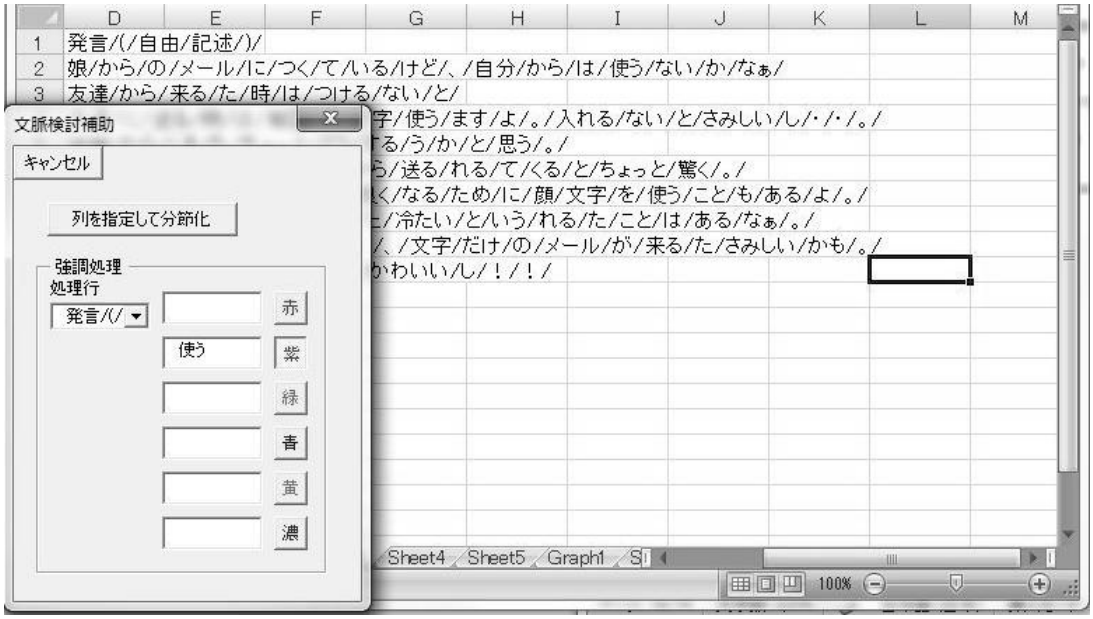


６．３　語の原型での分析

動詞や形容詞の場合(例：使わない)、先のような形式(例：使う)ではうまく検索できません。そのため原型の形式におきかえる必要があります。「列を指定して分節化」をクリックすると、任意の行に、原型で、分かち書きして表記することができます。

６．４　分かち書き結果に基づいて分析する。

分かち書きすると下記のように、「使う」でも、検索できるようになります。



7．よくあるエラーと対応

Q1：最初の段階でエラーが出てまったくうまく動かない。
A1：Access、茶筌がインストールされているかなど必要な条件は、満たしているかを確認してください。

Q2：エラーが出る。/分析結果がおかしい。
A2:エクセルのタブの移動を行った際などに、ソフトが参照すべきデータを見失って、エラーを起こしたり、分析結果がおかしくなったりすることがあるようです。お手数ですが、元のデータに戻って、分析をやり直してください。また、データの入力の仕方をご確認ください。

Q3：茶筌について
A3：茶筌については、茶筌のウェブページをご参照ください。
http://chasen.naist.jp/hiki/ChaSen/?茶筌の配布

Q5：ここに書かれていないエラーなど
A5：下記 URL をご参照いただき、下記 URL でも解決しない問題については、お手数ですが開発者までお尋ねください。対応可能な範囲内で回答いたします。
http://www.k2.dion.ne.jp/~kokoro/mivurix/textorva/

8．おわりに

TEXTORVA でできる機能は、限られた機能しか持たない。無償で利用可能なソフトウェアのなかには、TEXTORVA より高性能なものも少なくない。TEXTORVA でテキストマイニングに関心を持った人は、他のテキストマイニングソフトを試してみるとよいだろう。一般的に、市販されている有償のテキストマイニングソフトの方が高価である分、高性能である。

無償で利用できるテキストマイニングソフトの例

・KHcoder
  http://khc.sourceforge.net/
・TTM: TinyTextMining
  http://mtmr.jp/ttm/

有償で利用できるテキストマイニングソフトの例

・野村総合研究所 TRUE TELLER
・ジャストシステム トラスティア
・SPSS SPSS Text Analysis for Surveys (STAfS)

テキストマイニングについて勉強するには

・松村 真宏・三浦 麻子　人文・社会科学のためのテキストマイニング　誠信書房
・石田 基広　R によるテキストマイニング入門　森北出版

数量化Ⅲ類について勉強するには

・内田治　数量化理論とテキストマイニング　日科技連出版社

関連するエクセルマクロについて学ぶには

・高橋信（2005) Excelで学ぶコレスポンデンス分析　オーム社
・青木繁信　（2001）　数量化 Ⅲ 類
  http://aoki2.si.gunma-u.ac.jp/lecture/stats-by-excel/vba/html/qt3.html

９．著作権その他

本アドインの開発者は、荒川歩(武蔵野美術大学)ですが、本アドインはフリー・ソフトウェアであり、使用、改変、再配布は、自由です。
ただし、本アドインの使用に関して生じたあらゆる損害に゙Ą Ă ő Ü　　Hu ˆ ş Nj ません。また改変したものを再配布する場合には、その旨を記載するようにしてください。

なお、本アドインの制作にあたり、東洋大学平成 21 年度特別研究（教育システム開発）「授業に対する意見・質問・感想、及び授業カリキュラム・シラバスのテキストマイニングによる内容分析」（研究代表者　東洋大学教授　杉山憲司）の援助を受けた。また、杉山憲司先生、佐藤史緒さんをはじめ、同研究会のメンバーの皆さまには、さまざまな示唆をいただきました、ありがとうございました。

Textorva の最新版および良く聞かれる質問は下記の URL にあります。
http://www.k2.dion.ne.jp/~kokoro/mivurix/textorva/