# テキスト CR 分析の語数調整法と単語の選択
## －専用プログラムの開発－

Number Adjustment and Selection of Words for Textual Analyses
with Correspondence Analysis: development of a special program

福井正康，渡辺清美
福山平成大学経営学部
Masayasu FUKUI and Kiyomi WATANABE
Fukuyama Heisei University

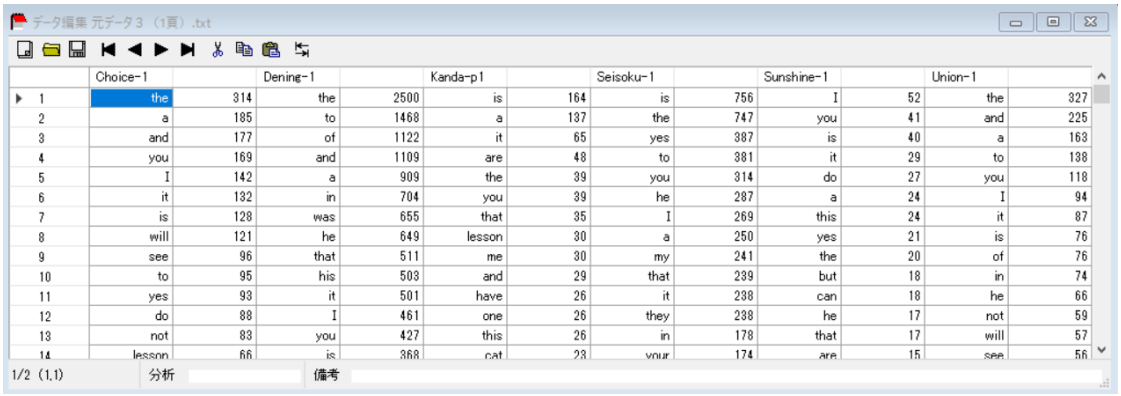
## １．はじめに

文書内の出現単語を行、文書名を列として、単語の出現数の 2 次元分割表を作り、コレスポンデンス分析（CR 分析）を用いて、文書を分類する分析が行われることがあるが、我々はこれをテキスト CR 分析と呼ぶことにする。テキスト CR 分析は、通常の CR 分析に比べて以下のような特徴がある。１つは単語の出現数をそのまま使うかどうか、もう１つは出現単語のすべてを取って分析するのか一部を利用するのかである。これらの問題に対して著者らは参考文献[1]で、ほぼ以下のような結論を得た。前者に対しては文書の長さを変えると単語数も変わり、分析結果も変わることから、単語数は文書ごとにある一定の数に標準化して利用する方がよい。また、後者に対してはある程度安定的な答えが出る必要性から、分割表の中で 0 の占める割合の 0 比率という指標を考えて、これが、0.2 程度以下がよいと結論した。また、同じ文献の中で新しい標準化の方法である 2 段階標準化法も提案した。これらの結果を元に、我々はこのテキスト CR 分析に特化した分析プログラムを College Analysis の中に組み込むことにした。この報告ではそのプログラムの特徴と利用法について説明する。また最後の章で、これまでの汎用的な CR 分析を使った利用法とテキスト CR 分析を使った利用法の違いをもう一度まとめて説明する。

## ２．単語比較ツール

複数の文書から単語の数を取り出した後、CR 分析には、テキスト間で共通する単語について１つにまとめ、すべての文書の語数の合計順に並べ替えるという前処理が必要である。この処理を簡単に行うために、ここではまず以前に作成し、今回少し改良したツールについて紹介する。

単語比較のためには、図 1 に示されるようなデータが必要である。これは、１つの文書につき、College Analysis の 2 列を使い、2 列ごとに単語とその出現数を表示したものである。左側に単語、右側はその頻度である。元のバージョンのツールでは 1 頁が 1 文書の設定になっていたが、新しい方法としてこのようなデータ形式も導入した。単語の並びについては、図では降順になっているが、特に指定はない。



| | Choice-1 | | Dening-1 | | Kanda-p1 | | Seisoku-1 | | Sunshine-1 | | Union-1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | the | 314 | the | 2500 | is | 164 | is | 756 | I | 52 | the | 327 |
| 2 | a | 185 | to | 1468 | a | 137 | the | 747 | you | 41 | and | 225 |
| 3 | and | 177 | of | 1122 | it | 65 | yes | 387 | is | 40 | a | 163 |
| 4 | you | 169 | and | 1109 | are | 48 | to | 381 | it | 29 | to | 138 |
| 5 | I | 142 | a | 909 | the | 39 | you | 314 | do | 27 | you | 118 |
| 6 | it | 132 | in | 704 | you | 39 | he | 287 | a | 24 | I | 94 |
| 7 | is | 128 | was | 655 | that | 35 | I | 269 | this | 24 | it | 87 |
| 8 | will | 121 | he | 649 | lesson | 30 | a | 250 | yes | 21 | is | 76 |
| 9 | see | 96 | that | 511 | me | 30 | my | 241 | the | 20 | of | 76 |
| 10 | to | 95 | his | 503 | and | 29 | that | 239 | but | 18 | in | 74 |
| 11 | yes | 93 | it | 501 | have | 26 | it | 238 | can | 18 | he | 66 |
| 12 | do | 88 | I | 461 | one | 26 | they | 238 | he | 17 | not | 59 |
| 13 | not | 83 | you | 427 | this | 26 | in | 178 | that | 17 | will | 57 |
| 14 | lesson | 66 | is | 368 | cat | 23 | your | 174 | are | 15 | see | 56 |

図１　単語比較のための１頁データ形式

メニュー［ツール－単語比較ツール］を選択すると、図 2 のような「単語比較ツール」実行画面が表示される。

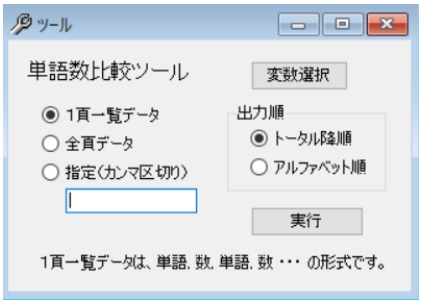
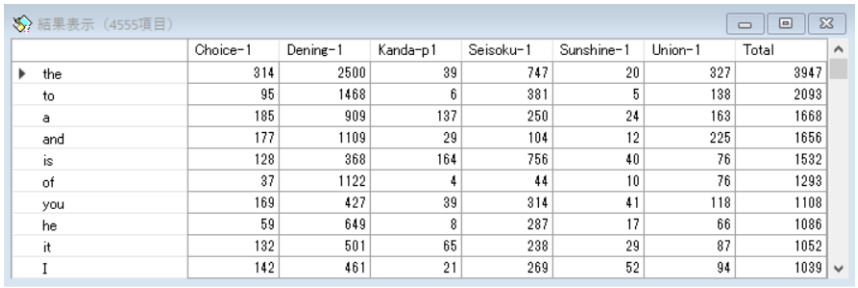



図２　単語比較ツール実行画面

図１のようなデータなら、「1 頁一覧データ」を選択する。文書の選択は変数選択で行う。また、これまでのような 1 頁 1 文書の形式のデータならそれ以外を選択する。1 頁 1 文書形式で、すべての文書について単語をそろえるなら、「全頁データ」ラジオボタンを、指定されたページだけを用いるなら、「指定」ラジオボタンを選択し、そのページ番号を下のテキストボックスにカンマ区切りで入れておく。出力は、選択文書全体の語数合計の降順の「トータル降順」か「アルファベット順」が選べる。通常、ページ指定は「１頁一覧データ」、出力順は「トータル降順」がよい。

この後「実行」ボタンをクリックすると図３に示す実行結果が表示される。この結果は単語が頻度順に並べられている。



| | Choice-1 | Dening-1 | Kanda-p1 | Seisoku-1 | Sunshine-1 | Union-1 | Total |
|---|---|---|---|---|---|---|---|
| the | 314 | 2500 | 39 | 747 | 20 | 327 | 3947 |
| to | 95 | 1468 | 6 | 381 | 5 | 138 | 2093 |
| a | 185 | 909 | 137 | 250 | 24 | 163 | 1668 |
| and | 177 | 1109 | 29 | 104 | 12 | 225 | 1656 |
| is | 128 | 368 | 164 | 756 | 40 | 76 | 1532 |
| of | 37 | 1122 | 4 | 44 | 10 | 76 | 1293 |
| you | 169 | 427 | 39 | 314 | 41 | 118 | 1108 |
| he | 59 | 649 | 8 | 287 | 17 | 66 | 1086 |
| it | 132 | 501 | 65 | 238 | 29 | 87 | 1052 |
| I | 142 | 461 | 21 | 269 | 52 | 94 | 1039 |

図３　単語比較ツール出力結果

## ３．テキスト CR 分析プログラム

我々のテキスト CR 分析プログラムは、図 3 の形式のデータを用いるが、単語数の合計を表す「Total」の欄は、分析に不要である。しかし、後に変数選択の中で落とすことができるので、あっても問題はない。このデータは新規に作成されたデータとしても、既存のデータの最後に追加して使うこともできる。後者の場合は、グリッド出力メニュー「編集－エディタ頁追加」を利用すると便利である。

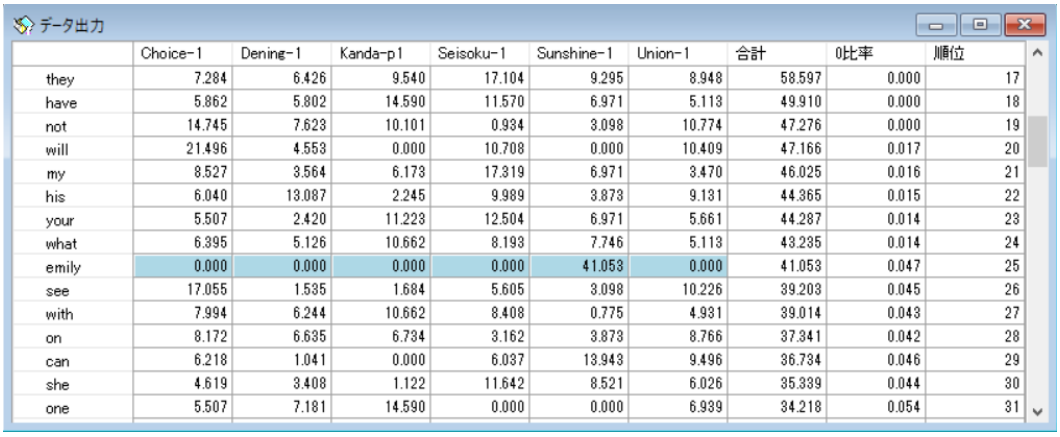
メニュー「分析－多変量解析他－分類手法－テキスト CR 分析」を選択すると図 4 のテキスト CR 分析実行画面が表示される。


図 4　テキスト CR 分析実行画面

この中の、「単語比較ツールへ」ボタンからも、図 2 のメニューを表示することができる。

テキスト CR 分析では単語数の調整を行うが、このプログラムでは、単語の頻度をそのまま利用する「実数」、単語の頻度をそろえる「１重調整」、単語の頻度をそろえた上で分析に利用する単語数を設定し、再度頻度をそろえる「２重調整」の方法を扱うことができる[1]。利用する単語数は「すべて」か、後ろに語数を指定した「指定」を選択できる。このメニューではデフォルトとして、調整法は「１重調整」、語数は「指定」100 語にしている。語数の「調整数」は分析に直接影響を与えないが、「データ出力」の際には値が変わってくるので、見た目が良い程度で記入しておく。デフォルトは 1000 になっている。

「変数選択」で Total を除くすべての変数を選択し、図 4 の「データ出力」ボタンをクリックすると、図 5 のような出力結果を得る。



データ出力

| | Choice-1 | Dening-1 | Kanda-p1 | Seisoku-1 | Sunshine-1 | Union-1 | 合計 | 0比率 | 順位 |
|---|---|---|---|---|---|---|---|---|---|
| they | 7.284 | 6.426 | 9.540 | 17.104 | 9.295 | 8.948 | 58.597 | 0.000 | 17 |
| have | 5.862 | 5.802 | 14.590 | 11.570 | 6.971 | 5.113 | 49.910 | 0.000 | 18 |
| not | 14.745 | 7.623 | 10.101 | 0.934 | 3.098 | 10.774 | 47.276 | 0.000 | 19 |
| will | 21.496 | 4.553 | 0.000 | 10.708 | 0.000 | 10.409 | 47.166 | 0.017 | 20 |
| my | 8.527 | 3.564 | 6.173 | 17.319 | 6.971 | 3.470 | 46.025 | 0.016 | 21 |
| his | 6.040 | 13.087 | 2.245 | 9.989 | 3.873 | 9.131 | 44.365 | 0.015 | 22 |
| your | 5.507 | 2.420 | 11.223 | 12.504 | 6.971 | 5.661 | 44.287 | 0.014 | 23 |
| what | 6.395 | 5.126 | 10.662 | 8.193 | 7.746 | 5.113 | 43.235 | 0.014 | 24 |
| emily | 0.000 | 0.000 | 0.000 | 0.000 | 41.053 | 0.000 | 41.053 | 0.047 | 25 |
| see | 17.055 | 1.535 | 1.684 | 5.605 | 3.098 | 10.226 | 39.203 | 0.045 | 26 |
| with | 7.994 | 6.244 | 10.662 | 8.408 | 0.775 | 4.931 | 39.014 | 0.043 | 27 |
| on | 8.172 | 6.635 | 6.734 | 3.162 | 3.873 | 8.766 | 37.341 | 0.042 | 28 |
| can | 6.218 | 1.041 | 0.000 | 6.037 | 13.943 | 9.496 | 36.734 | 0.046 | 29 |
| she | 4.619 | 3.408 | 1.122 | 11.642 | 8.521 | 6.026 | 35.339 | 0.044 | 30 |
| one | 5.507 | 7.181 | 14.590 | 0.000 | 0.000 | 6.939 | 34.218 | 0.054 | 31 |

図 5　データ出力結果

この結果は一度 1000 語に調整を実行して、その中で頻度の上位から指定語数を選択して表示されている。これが分析に使うデータである。この中には、参考のために、調整後の単語の合計数や 0 比率などが表示されている。ここでは例として、総頻度が 17 位から 31 位までを表示しているが、この中で水色の網掛けの単語がある。これは 1 つの文書以外では頻度が 0 の単語である。0 比率が低いところの単語では、本来利用しない固有名詞などが残っている場合があり、そのような場合にはデータから削除する。データの削除にはエディタのメニュー［ツール－検索］で表示される検索画面で、「行名検索」機能を用いるとよい。

分析実行画面で「CR 分析」ボタンをクリックすると、指定された調整法で、指定された語数で CR 分析を実行する。但し、単語数は文書数より多くないといけない。実行結果を図 6 に示す。

コレスポンデンス分析結果

| 0比率: 0.112 | 群 | 第1成分 | 第2成分 | 第3成分 | 第4成分 | 第5成分 | 重み1成分 | 重み2成分 | 重み3成分 | 重み4成分 | 重み5成分 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 固有値 | | 0.182 | 0.146 | 0.079 | 0.057 | 0.016 | | | | | |
| 相関係数 | | 0.427 | 0.383 | 0.281 | 0.239 | 0.127 | | | | | |
| 寄与率 | | 0.379 | 0.305 | 0.164 | 0.119 | 0.034 | | | | | |
| 累積寄与率 | | 0.379 | 0.684 | 0.848 | 0.966 | 1.000 | | | | | |
| Choice-1 | 2 | 0.161 | 0.286 | 0.203 | 1.787 | -1.140 | 0.069 | 0.109 | 0.057 | 0.427 | -0.145 |
| Dening-1 | 2 | 1.268 | 0.731 | 0.775 | -1.513 | -0.995 | 0.541 | 0.280 | 0.218 | -0.361 | -0.126 |
| Kanda-p1 | 2 | -1.856 | 0.604 | 0.472 | -0.453 | 0.141 | -0.792 | 0.231 | 0.133 | -0.108 | 0.018 |
| Seisoku-1 | 2 | 0.094 | 0.052 | -2.134 | -0.331 | 0.004 | 0.040 | 0.020 | -0.600 | -0.079 | 0.001 |
| Sunshine-1 | 2 | -0.019 | -2.245 | 0.449 | -0.230 | 0.031 | -0.008 | -0.859 | 0.126 | -0.055 | 0.004 |
| Union-1 | 2 | 0.841 | 0.502 | 0.442 | 0.515 | 2.008 | 0.359 | 0.192 | 0.124 | 0.123 | 0.255 |
| the | 1 | 0.908 | 0.691 | -0.112 | 0.008 | -0.132 | 0.388 | 0.264 | -0.031 | 0.002 | -0.017 |
| is | 1 | -1.441 | 0.049 | -0.620 | -0.631 | 0.231 | -0.615 | 0.019 | -0.174 | -0.151 | 0.029 |
| a | 1 | -0.945 | 0.574 | 0.798 | -0.148 | 0.403 | -0.403 | 0.220 | 0.224 | -0.035 | 0.051 |

図 6　CR 分析結果

同じ処理を通常の CR 分析のメニューで実施すると、最初に単語が表れるようになっているが、ここでは文書の類似性の方が重要であるので、文書名が最初に並ぶように設定している。内容については、コレスポンデンス分析の章を参照してもらいたい。

CR 分析の結果を用いてクラスター分析を行い、すべての次元を参照して分類することも可能である。その際、行成分と列成分に付けられる係数の相関係数によって次元の重みを付ける処理が行われるため、クラスター分析では相関の重み付き成分を利用する方が現実

的である。これらのことを考えて、「クラスター用データ」ボタンをクリックすると図 6 の四角で囲んだ部分を出力するようにしている。結果を図 7 に示す。

クラスター分析用CRデータ

| | 重み1成分 | 重み2成分 | 重み3成分 | 重み4成分 | 重み5成分 |
|---|---|---|---|---|---|
| Choice-1 | 0.069 | 0.109 | 0.057 | 0.427 | -0.145 |
| Dening-1 | 0.541 | 0.280 | 0.218 | -0.361 | -0.126 |
| Kanda-p1 | -0.792 | 0.231 | 0.133 | -0.108 | 0.018 |
| Seisoku-1 | 0.040 | 0.020 | -0.600 | -0.079 | 0.001 |
| Sunshine-1 | -0.008 | -0.859 | 0.126 | -0.055 | 0.004 |
| Union-1 | 0.359 | 0.192 | 0.124 | 0.123 | 0.255 |

図 7　クラスター用データ出力

これをクラスター分析のプログラムのデータとしてデンドログラムを描くことになるが、距離測定法は重み付けをしたことを考慮して、「平方ユークリッド距離」、クラスター構成法は標準的な「ウォード法」が適していると考える。これらの設定での結果を図 8 に示す。

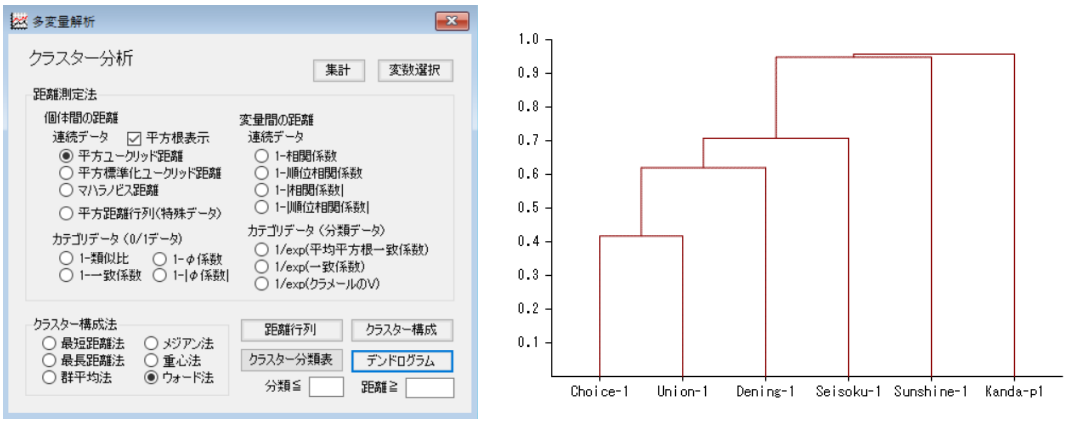



図 8　クラスター分析の実行画面とデンドログラム

次に、x 軸を第 1 成分に、y 軸を第 2 成分にし、「相関重み」を加え、その他の設定をデフォルトの設定にして、「散布図」ボタンをクリックした結果を図 9 に示す。

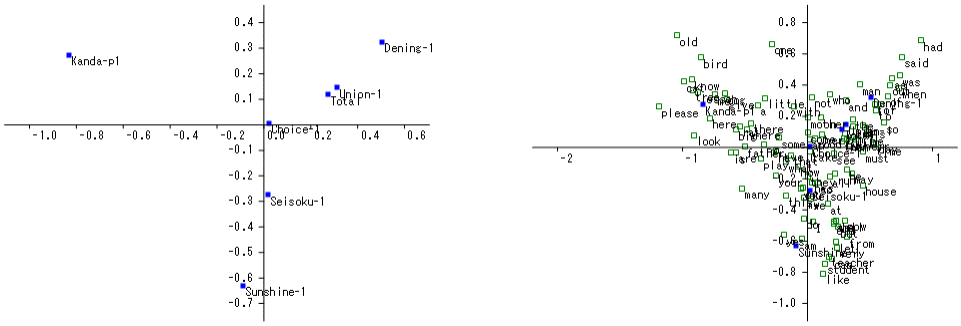



図 9　CR 分析による散布図

ここに、左が「列」成分だけの表示、右が「行」成分も含めた表示である。

同様に、「3D」チェックボックスをチェックし、z 軸を第 3 成分にして、その他の設定を図 9 と同じにした散布図を図 10 に示す。但し、分かりにくいので「列」成分だけにしている。

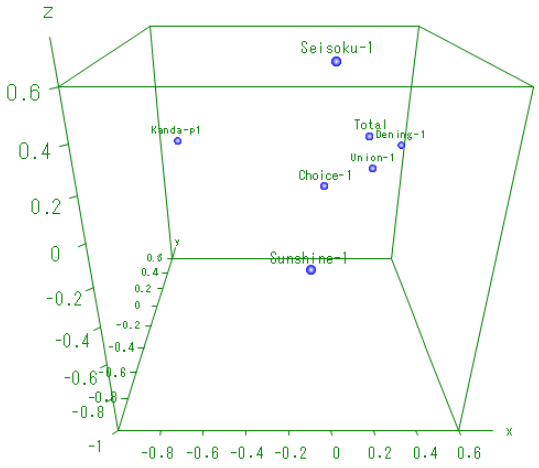



図 10　CR 分析による散布図（3 次元表示）

我々は利用する語数を 100 語に固定してこれまでの計算を行ってきたが、これは 0 比率の値を参考にしながら決めた。語数を決定するとき結果の安定性は重要である。そこで、結果が語数によっていかに変化するかをアニメーションで表示する試みを思い付いた。これは指定された最大語数から、徐々に選択語数を減らして行き、最終的に指定された最小語数まで、散布図が変わって行く様子をアニメーションのように表示する機能である。この動きは紙面上で表現できないが、変化の過程の文書名の配置の連続性によって CR 分析の正当性を確認する方法である。図 11 にその過程を簡単に示す。実際に動かしてみると大変興味深いのでぜひ試してもらいたい。

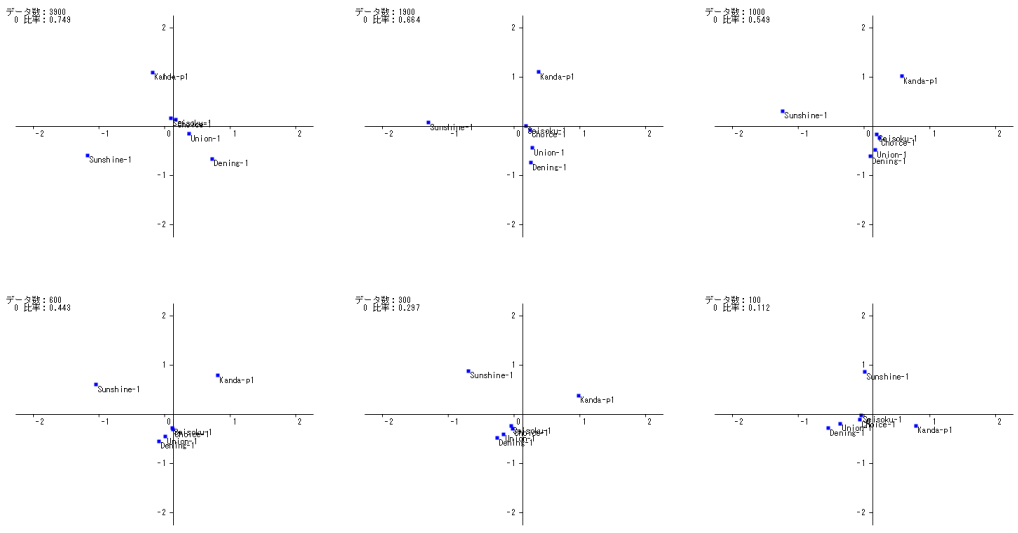


図 11　アニメーション表示の例

## ４．おわりに

最後に、これまでのプログラムを使った分析法と今回のプログラムを使った分析法の違いをまとめておく。

まず、単語数比較ツールについて、今回の機能追加は「１頁一覧データ」への対応である。これまで１頁１文書でデータを入力していたが、頁を追加しながらの入力や、どの文書を使うかなどの選択でかなり手間がかかった。しかし、今回の改訂で、データは１頁にまとめておき、一回のコピーで貼り付けを行い、簡単に文書を選んで単語数比較ができるようになった。

次に、CR 分析の機能についての改善点を説明する。比較のために、一般の CR 分析の実行画面を図 12a、テキスト CR 分析の実行画面を図 12b に示す。

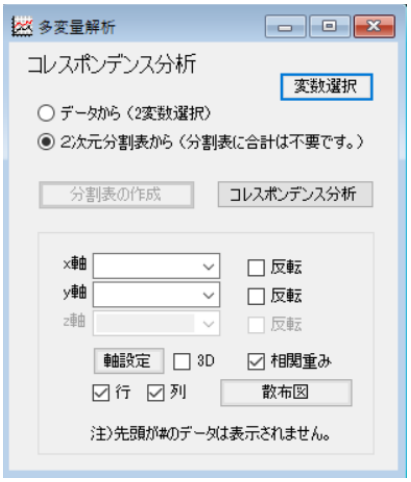



図 12a　CR 分析実行画面


図 12b　テキスト CR 分析実行画面

CR 分析の画面では、分割表以外にアンケートデータなどから直接分析を実行する「データから」の選択肢があるが、テキスト CR 分析では、データがそのまま分割表とみなされるので、「データから」の機能はない。しかしその他の機能はすべてそのまま揃っている。

またテキスト CR 分析では、分析特有の語数調整や語数選択の機能も持っている。一般の CR 分析を使った分析では、まず語数調整を Excel などで実行しなければならない。次に、その中から必要な語数をコピーしてデータとするが、２重調整法ではさらに語数調整する必要があり、かなり手間のかかる作業であった。しかし、テキスト CR 分析では、ボタン選択と語数入力だけでこの処理ができるようになった。

次に、我々が注目した 0 比率については、以前は Excel などで別に計算しなければならなかったが、今回は「データ出力」ボタンで、図 5 のように表示できるようになった。また、

その際、クリーニングし忘れた可能性がある 1 つの文書だけに含まれる単語を、水色の網掛けで目立つようにしている。

CR 分析の結果を散布図に示した場合、2 次元の成分だけの表示になるが、結果をすべて使って、クラスター分析でデンドログラムによる表示も可能である。CR 分析のプログラムでは図 6 のように必要な部分をコピーしてデータとする必要があったが、テキスト CR 分析のプログラムには専用の機能があり、「クラスター用データ」ボタンをクリックすると図 7 のような結果が出力される。これを貼り付ければクラスター分析は容易である。

アニメーションについては、もちろん元の CR 分析のプログラムにはない。これは新しく追加された機能である。

**参考文献**

[1] コレスポンデンス分析を用いた英文テキスト分類における語数調整法と単語の選択基準，福井正康，渡辺清美，福山平成大学経営研究，第 15 号，(2019) 掲載予定.