◆ 質的データのマイニングのための対応分析法 ◆

簡単な分析例 —対応分析法とクラスター化法で何がわかるのか —

これは，JMRA セミナー「**質的データのマイニングのための対応分析法**」紹介のために，講師が作成したメモです．このセミナーの目的は，多次元データ解析手法の 1 つである「対応分析法」（数量化法 III 類，コレスポンデンス分析）とその特性を利用した独自の「クラスター化法」に的を絞って，これらの基本的な事項を丁寧に説明，紹介することにあります．

周知のように，対応分析法は，従来，さまざまな分野で広く利用されてきました．この手法はかなり完成度の高い“古くて新しい手法”です．データ解析のいろいろな場面で登場する“質的データ・定性情報の分析”に適しており，多様な形態のデータ表に柔軟に対応できるという特徴があります．

対応分析法は質的データから得た成分スコアの布置図・同時布置図を観察し解釈する“視覚化ツール”であるという限られた使い方として認識されているようです．しかし，実際の対応分析の利用場面は多様です．とくに，数理的特性をうまく利用した対応分析法とクラスター化法の併用は，テキスト型データの分析などで，非常に有効に機能します．

このメモは，意図的に小規模の典型的な実例データを用いて，対応分析とクラスター化による分析で得られる情報の要点を，「分析例」の形でまとめてみました．いずれの分析例も，分析過程の流れを抜粋して具体的に記しました．やや冗長的な説明となりましたが，“なにができそうか”はわかっていただけると思います．ぜひご一読いただき，セミナー参加をご検討いただく参考情報としていただけると幸いです．

要旨

対応分析法の適用例を小規模の実際データで確かめよう．ここでは，**【分析例 1】～【分析例 3】の 3 例**を用意した．

【分析例 1】 典型的な 2 元データ表の分析
【分析例 2】 自由回答質問の分析（1） −ウェブ調査データの構造探査の基本−
【分析例 3】 自由回答質問の分析（2） −クラスター化による仮説発見の効用を知る−

いずれも対応分析，クラスター化の典型的な分析例である．【分析例 2】と【分析例 3】は，同じデータセットに対して“視点を変えて分析を行う”ことで，データに内在する特徴を探査する“**探索的アプローチ**”の例となっている．いずれも，元となる“2 元データ表”（two-way data table）の作り方と結果の説明・解釈に注意して読んでいただきたい．

ここでは小規模のデータ表を扱うが，一般に，データ表の寸法はこの程度では済ま

ないことが多い．とくに，テキスト型データ，たとえば自由回答質問や自由記述，日記型式，インタビュー・データ，患者の聴取調査，看護・介護聴取記録，さらにはソーシャル・メディアで表れる非定型データなどは，データ量が多く，ゴミも多く含まれ，処理が容易ではない．【分析例 2】，【分析例 3】は，そうした自由回答質問の回収データを用いた教科書的な分析の抜粋である．

なお，いずれの例も，JMP®（ジャンプ），JMP スクリプト，WordMiner®（ワードマイナー），エクセルを用いて得られた分析結果を抜粋，再編集して用いた．

【キーワード】

セミナーの紹介内容に関連する，キーワードを拾い出した．こうした語句が，データ解析のどのような場面で，どのように機能し，用いられるのかを知っていただくことも，セミナーの目標である．

質的データと量的データ，定型データと非定型データ，マイニング，ソーシャル・メディア，ウェブ調査，ウェブ・パネル，調査方式（モード），対応分析法，プロファイル，ストレッチ・プロファイル，分布の同等性，成分スコアとその相関，双対性，固有値・特異値と寄与率，寄与度（絶対寄与度，相対寄与度），布置図・同時布置図と情報の視覚化，自動分類法とクラスター化法，階層的分類法と非階層的分類法，デンドログラム（樹形図），（重み付き）ウォード法，データ圧縮化，（ピアソンの）カイ二乗統計量，独立性の検定，ユークリッド距離とカイ二乗距離，ワードクラウド，自由回答質問と選択肢型質問，テキスト型データ，辞書編集，コーパス，シソーラス，テキスト･マイニング，調査誤差（観測誤差と非観測誤差），測定誤差，データの品質，外部情報源，探索的・探査的，発見的，帰納的，理解と洞察

【本セミナーの目標】

- ここにあげた「分析例」のようなデータ解析を進める際に必要となる，“対応分析法とクラスター化法”の考え方（基本的な数理と仕組み）を知ること．
- 基本となる“2 元データ表”の考え方，作り方の要点を知ること．
- 対応分析法の視覚化ツールとしての効用と限界を知り，同時に正しい使い方を知ること．
- とくに，布置図，同時布置図の解釈における誤用や勘違いを避ける方法を知る．
- 対応分析法とクラスター化法は，セットで利用することが効果的であることを知る．とくに“大規模データの自動分類の効用”を知ること．
- 統計ソフトウェアの性能向上で，多種多様な情報が得られる．対応分析・多重対応分析，クラスター化も例外ではない．統計ソフトウェアの出力情報を読み解く方法を知ること．

【分析例 1】 2 元要約表の分析 ーメディア接触・情報源とその評価

調査の概要：

- 調査課題：情報に関する調査
- 調査対象（標本抽出枠）：あるウェブ・パネル（非公募型）に登録の首都 40 ㎞圏に在住，15 歳以上 69 歳未満の男女（パネル構成の詳細情報は省く）
- 調査方式（モード）：ウェブ調査
- 計画標本の大きさは 766（人），有効回収標本の大きさは 347（人），参加率[1]は 45.3（%）

分析の目標：

あるウェブ・パネルの登録者を対象に，上のような課題で調査を行った．多数の質問のうち，**“日ごろの情報入手手段である「情報源」とその「評価」の関係”について**たずねた質問について，分析を行ってみる（図 1 は質問文の一部，表 1 は用いた選択肢）．

実は，この質問は，調査後の分析で対応分析法の適用を想定して設計されている，典型的な例である．ここで，回答者は 23 の「情報源」と 9 の「評価項目」について，「あてはまる」場合を選ぶ．表 2 が得られた回答分布である．データ表の各セルの度数（回答者数）に何か傾向があるようにみえる．「情報源」と「評価項目」にはどのような関係があるのだろうか．

結果とその観察 −ステレオタイプに分析すると−

1) 対応分析法では，データ表の行側（ここでは「情報源」）と列側（ここでは「評価項目」）のそれぞれの選択肢に**成分スコア**が与えられる[2]．図 3 は「評価項目」について描いた第 1，第 2 成分スコアの**布置図**である．ここでは 2 つの成分で**寄与率**は約 76%で，全情報の約 8 割がこの 2 つの成分で説明されている．

2) （第 1 成分スコアを使って）元のデータ表の行と列を並べ替え，両者の関連の程度（相関）を観察する（図 2）．これをみると，「情報源」と「評価項目」にはかなりの相関があるようにみえる（つまり得られた成分スコア間の**相関**を観察する“意味がありそうだ”）．

3) またこの視覚化情報から「評価項目」はおよそ 4 つのグループに分かれているように“みえる”．ここで，クラスター化を行うと，**デンドログラム（樹形図）**が得られた（図 4）．ここでも，このグループ化の傾向は読める．

G1＝｛情報が詳しい，役に立つ，生活に欠かせない｝
G2＝｛商品を選び購入する｝
G3=｛情報量が多い，世間の話題や流行を知る｝
G4=｛情報が正確，信頼できる，古くさい｝

4) ここで，「情報源」を加えて，「評価項目」との対応を，図 5 の**同時布置図**で観察しよう．ここで両者の関係が観察される“はず”である．図中の各要素はやや煩雑になっていてみに

[1] ウェブ調査では，多くの場合，非確率的パネルを用いるので，的確な回答率（response rate）の算出が難しい．そのため，参加率（participation rate）を用いる．
[2] （対称型の一般的な）対応分析法では，2 元データ表の行と列を入れ替えて分析しても同じ結果がえられる．

くい（しかしこれが一般的である）．“主観的に”あるいは“見栄え”でこれを読み取ろう．簡単にわかることは離れた位置にある，「9．パンフレット・カタログ・ダイレクトメール」が，G2＝｛商品を選び購入する｝に関連するように“みえる”．

5） では他はどうであろうか，図の左上には「15．ツイッター（Twitter）:「17．ミクシィ(mixi)」「18．フェイスブック（Facebook)）「19．グリー（GREE)」「「20．モバゲータウン」「22．ニコニコ動画」が集まり，これが G3=｛情報量が多い，世間の話題や流行を知る｝に関連するように“みえる”．バブルプロットの円の大きさは回答数に合わせてあるが，これでこのあたりの回答数が多いことがわかる（円が大きいほど，回答者数が多い）．

6） とくに表 2 に照らすと，図 5 の左上「19．グリー（GREE)」「「20．モバゲータウン」「22．ニコニコ動画」の回答比率傾向（**プロファイル**）が良く似ていることもわかる．

7） さらに，図の中央あたりは，つまりおおかたの**平均的意見**は，G1＝｛情報が詳しい，役に立つ，生活に欠かせない｝に対して，「1．テレビの番組」「2．ケーブルテレビ・衛星放送の番組」「5．新聞の紙面広告（電子版を含む)」「6．書籍（漫画・コミック以外)」「7．一般の雑誌・週刊誌（漫画・コミック以外）の記事」にある“ようだ”と読める．

8） また，G4=｛情報が正確，信頼できる，古くさい｝には，「3．ラジオの番組」「4．新聞の記事（電子版を含む)」「8．各分野専門の情報誌の記事」「10．都・県や市・区など自治体の広報誌紙」「11．所属する会や組織の会報・同人誌・ニュースレター」が“対応する“ようである”．

9） とくに，評価用語の「情報が詳しい」と「情報が正確」に注目すると，対応するメディアへの“見方”が違っていて，“詳しい”と“正確”とは異なるニュアンスで捉えられている“ようだ”．

何が検討課題か　–もう一歩踏み込んで，仮説発見的に探索するには–

上の読み方は，なんとなく，それらしく感じられる．ここまでの解釈は，対応分析法の紹介記事や適用例などで，**“よくある記述”**であり，**“良くみる説明”**である．つまり，あえて**ステレオタイプな説明**としてみた．こうした解釈がおかしいわけではない．しかし，“このような推論だけ”でよいのであろうか．上では，あえて，“はず”，“主観的”，…，“みえる”，“ようだ”，“…ようである”，…などと書いた．どこか客観性に欠けているようにもみえる．

“見た目”での，あるいは人の感覚的な解釈は無駄ではないし，ときには鋭い洞察となる場合もある．しかしここで，上にみた出力情報に，より客観性を与える**帰納的で仮説発見的な観察**を行うための“仕組み”があればありがたい．

たとえば，上の記述では表だって登場しないが，対応分析法では，結果の解釈をすこしでも客観的に評価するための，さまざまな**評価指標**がある（相関と特異値・固有値，慣性と寄与度，相対寄与度，絶対寄与度など）．その多くは，対応分析法に特有の符丁であるのだが，結果の誤った解釈を避けるには，仕組みをよく理解することである．たとえば，以下に挙げるような事柄の意味をよく知ることで，対応分析法はより強力な**マイニング・ツール**となる．

- なぜ，“**2 元データ表**”なのか，他のデータ表の形もあるのか．
- **“プロファイル”**とは何か，（図に示した）“**成分スコア”**とは何か．

- （図に書き入れた）“**固有値，寄与率”とは**なにか，その大きさは何を意味するのか.
- 多次元情報を含む 2 元データ表を，限られた少数次元の“**布置図，同時布置図”で読むこと**は正しいのか.
- 何をもって，成分スコアの特徴・傾向を説明するのか.
- **成分スコア間の“相関”**とはなにか，またなぜそれが必要なのか.
- 布置図で（評価項目，情報源の）各点の類似・差違はどう評価するのか（指標はなにか）.
- また，遠い，近いなどの“**距離”は何を意味する**のか.
- **質的データ**であるのに，どうして“**分類”（クラスター化）が可能**なのか.
- たとえば，テキスト型データであっても，どうして分類可能なのか.
- 他の質問項目，調査項目，とくに人口統計学的要因などの影響はどう測るのか.
- たとえば，この課題では，性別や年齢区分により傾向が異なるのではないか．表 2 のデータ表を性別や年齢区分でブレイクダウンしたとき，どのように分析すればよいのか.

こうした疑問や課題に，どのように応えたら（答えたら）よいのであろうか.

Q17　現在私たちは、情報を入手できる手段として数多くの情報源に囲まれており、
それらの情報源についていろいろな意見が言われています。
さて、以下でAからDの４つのことばがあてはまる情報源にはどのようなものがあるでしょうか。
あなたが「あてはまる」と思われるものをすべてお選びください。（それぞれいくつでも）

	A 情報が正確 ↓	B 情報が詳しい ↓	C 情報量が多い ↓	D 信頼できる ↓
1. テレビの番組	☐	☐	☐	☐
2. ケーブルテレビ・衛星放送の番組	☐	☐	☐	☐
3. ラジオの番組	☐	☐	☐	☐
4. 新聞の記事（電子版を含む）	☐	☐	☐	☐
5. 新聞の紙面広告（電子版を含む）	☐	☐	☐	☐
6. 書籍（漫画・コミック以外）	☐	☐	☐	☐
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	☐	☐	☐	☐
8. 各分野専門の情報誌の記事	☐	☐	☐	☐
9. パンフレット・カタログ・ダイレクトメール	☐	☐	☐	☐
10. 都・県や市・区など自治体の広報誌紙	☐	☐	☐	☐
11. 所属する会や組織の会報・同人誌・ニュースレター	☐	☐	☐	☐

図１　質問文の一部（グリッド形式）

表１　「情報源」と「評価項目」

情　　報　　源	
1. テレビの番組	12. パソコンでみるインターネットサイト
2. ケーブルテレビ・衛星放送の番組	13. 携帯電話・PHS、スマートフォンでみるインターネットサイト
3. ラジオの番組	14. インターネットブログ、ブログ
4. 新聞の記事（電子版を含む）	15. ツイッター（Ｔｗｉｔｔｅｒ）
5. 新聞の紙面広告（電子版を含む）	16. 電子書籍（電子書籍端末や電子ブックリーダーで読む）
6. 書籍（漫画・コミック以外）	17. ミクシィ(mixi)
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	18. フェイスブック（Facebook）
8. 各分野専門の情報誌の記事	19. グリー（GREE）
9. パンフレット・カタログ・ダイレクトメール	20. モバゲータウン
10. 都・県や市・区など自治体の広報誌紙	21. YouTube
11. 所属する会や組織の会報・同人誌・ニュースレター	22. ニコニコ動画

評　価　項　目
情報が正確
情報が詳しい
情報量が多い
信頼できる
生活に欠かせない
役に立つ
世間の話題や流行を知る
商品を選び購入する
古くさい

(*) 分析では「23. 1～22 の中にはひとつもない」は除外した．また，無回答も除外した．

表 2　質問への回答者の頻度分布[分析対象とする 2 元データ表]

このデータ表には，どのような特徴があるのだろうか．
数字を見慣れた人にはおよその規則性や特徴が読み取れるであろう．対応分析法は，これを効率的かつ客観的に評価する情報を提供する．

		評価項目								
		情報が正確	情報が詳しい	情報量が多い	信頼できる	生活に欠かせない	役に立つ	世間の話題や流行を知る	商品を選び購入する	古くさい
情報源	1. テレビの番組	100	114	235	90	226	166	248	51	20
	2. ケーブルテレビ・衛星放送の番組	43	91	103	44	42	94	74	30	17
	3. ラジオの番組	65	68	86	54	50	103	79	9	82
	4. 新聞の記事（電子版を含む）	140	153	131	131	135	167	124	17	24
	5. 新聞の紙面広告（電子版を含む）	36	59	90	28	29	62	74	59	16
	6. 書籍（漫画・コミック以外）	41	98	108	39	57	100	71	33	16
	7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	19	84	139	12	27	74	136	46	12
	8. 各分野専門の情報誌の記事	88	163	89	86	24	137	63	47	10
	9. パンフレット・カタログ・ダイレクトメール	31	90	90	18	12	76	59	155	29
	10. 都・県や市・区など自治体の広報誌紙	125	70	38	132	41	148	25	7	79
	11. 所属する会や組織の会報・同人誌・ニュースレ	45	73	50	46	11	102	25	5	67
	12. パソコンでみるインターネットサイト	26	118	274	24	186	204	200	182	0
	13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	18	63	166	15	77	105	107	44	0
	14. インターネットブログ、ブログ	6	59	150	5	34	67	135	14	1
	15. ツイッター（Twitter）	6	31	143	9	21	43	117	5	1
	16. 電子書籍（電子書籍端末や電子ブックリーダー	22	39	103	19	9	53	61	7	1
	17. ミクシィ(mixi)	9	36	128	11	25	43	107	8	8
	18. フェイスブック（Facebook）	12	17	115	12	14	35	94	7	9
	19. グリー（GREE）	6	20	99	5	6	18	79	3	6
	20. モバゲータウン	8	20	100	4	7	20	76	2	6
	21, YouTube	16	42	135	10	27	88	120	4	1
	22. ニコニコ動画	7	22	122	6	10	46	94	2	5

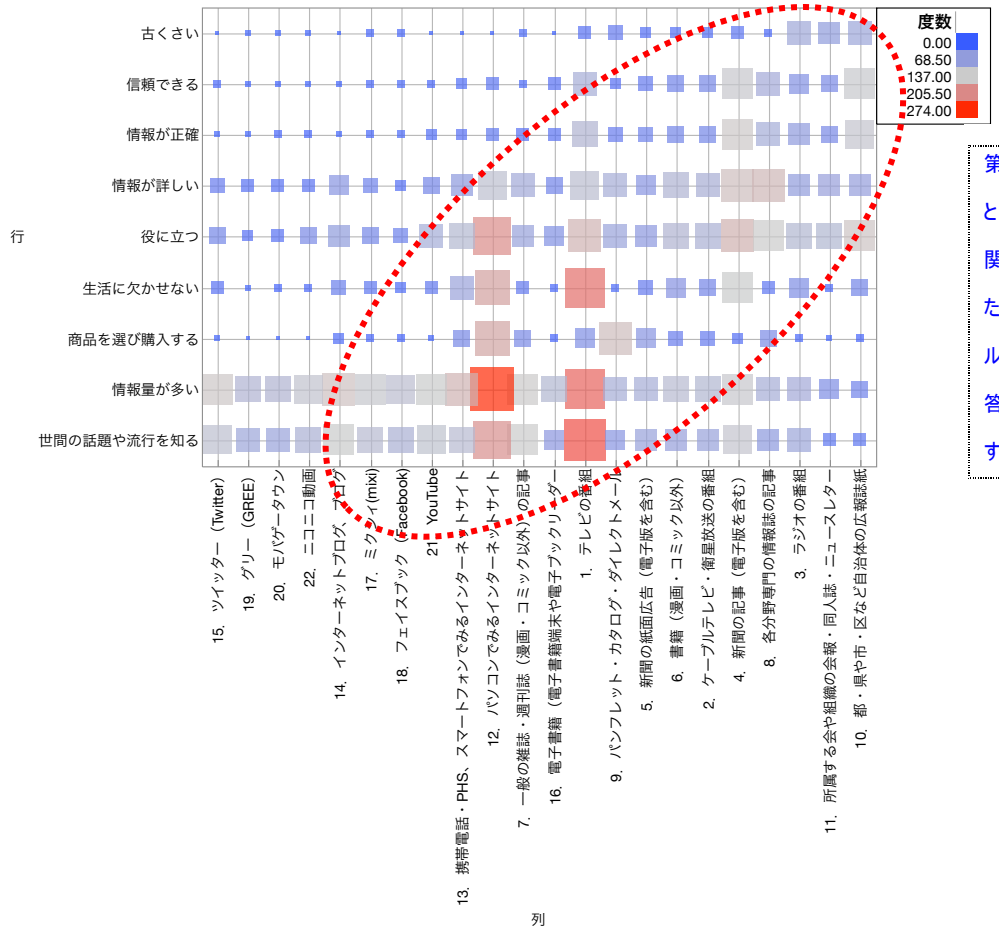


図２　データ表の特徴観察（第１成分スコアについて表２の行と列を並べ替え）

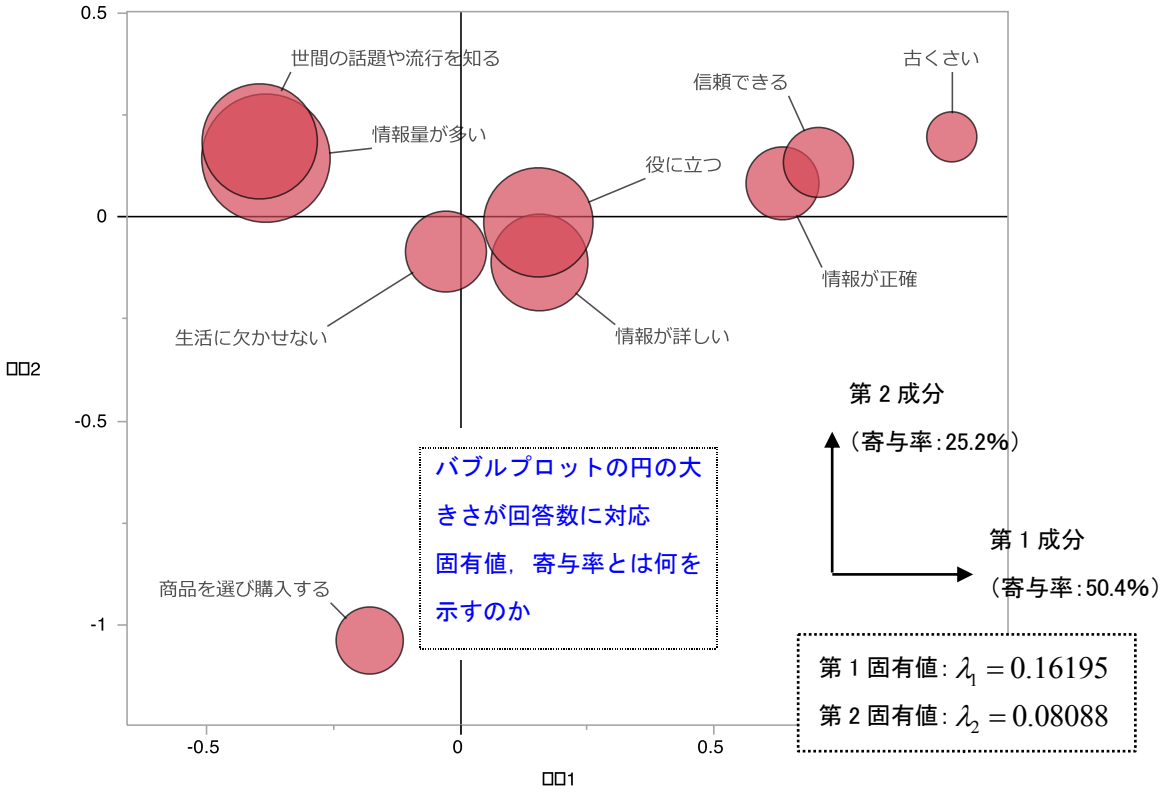


図３　「評価項目」の布置図（第１，第２成分）

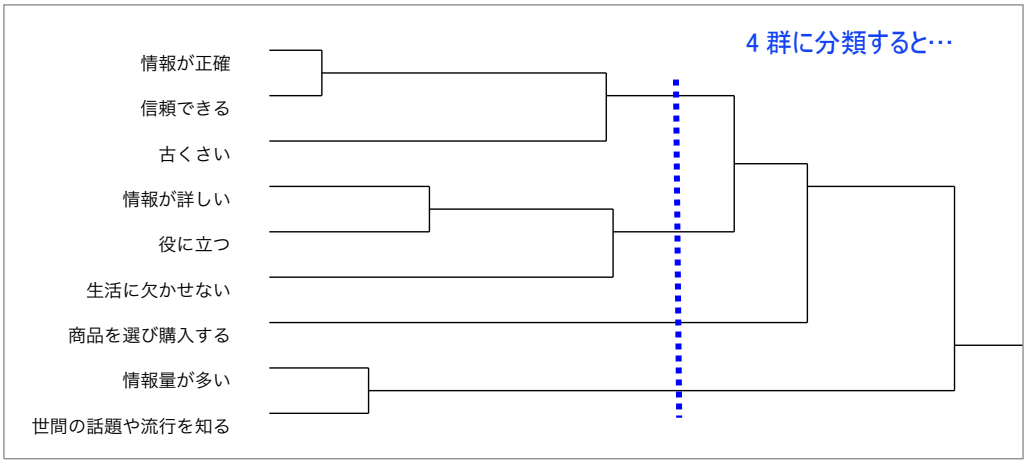


図 4　評価項目の分類（デンドログラム）

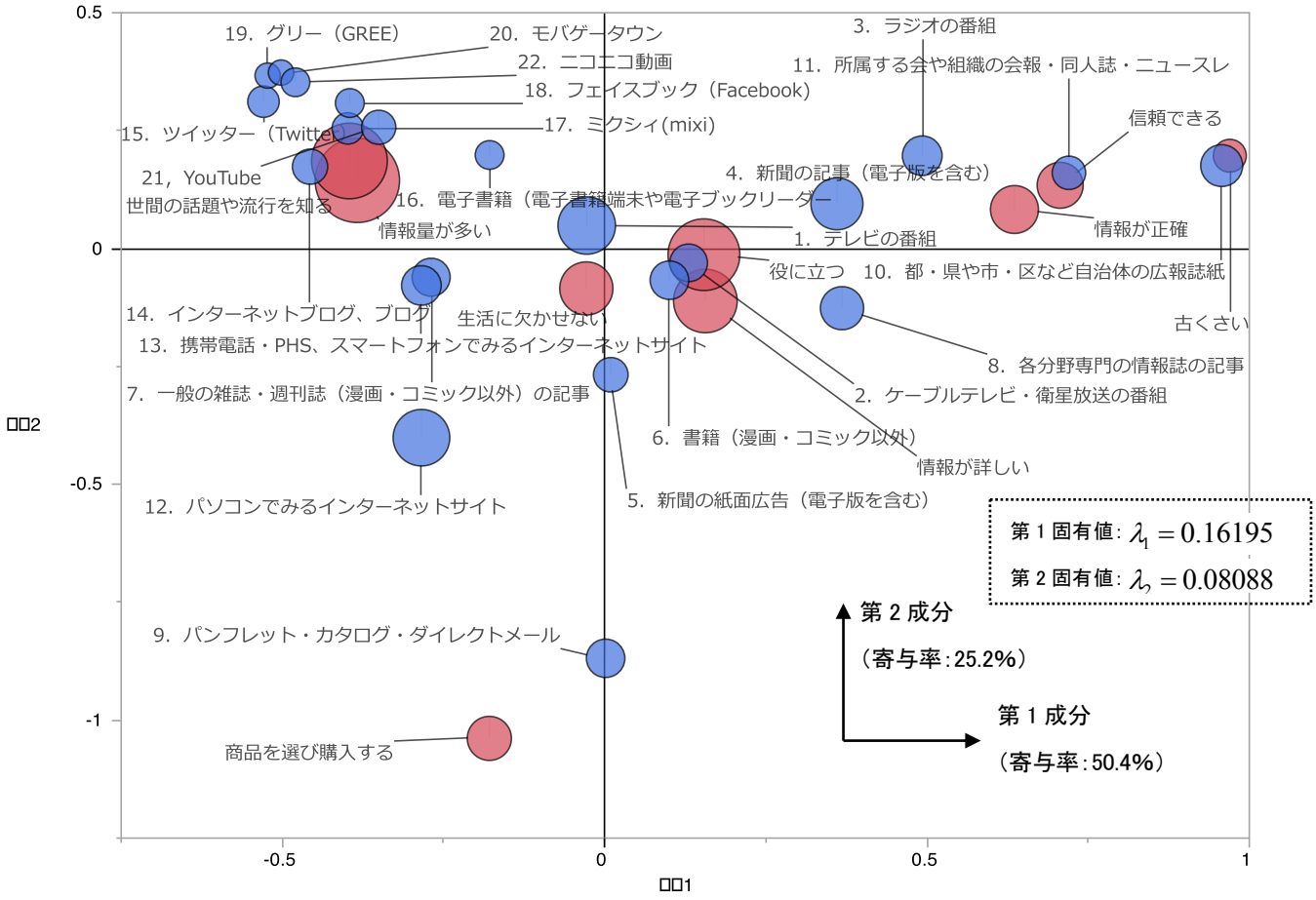


図 5　「情報源」と「評価項目」の同時布置図

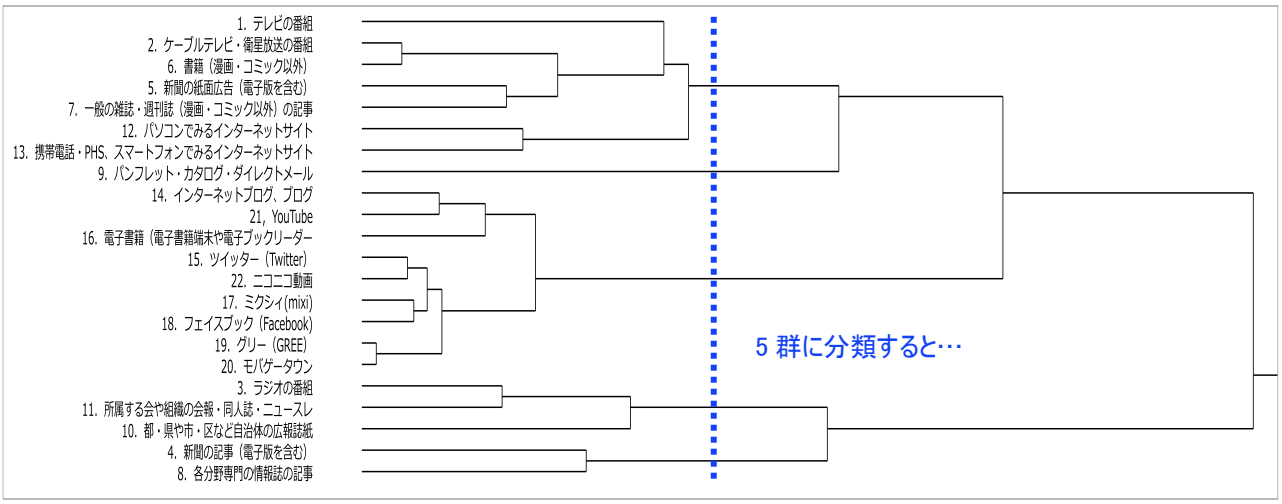


図 6　「情報源」の分類（デンドログラム）

＜情報源の分類結果（5 群）の例＞

C1=｛1．テレビの番組／2．ケーブルテレビ・衛星放送の番組／6．書籍（漫画・コミック以外）／5．新聞の紙面広告（電子版を含む）／7．一般の雑誌・週刊誌（漫画・コミック以外）の記事／12．パソコンでみるインターネットサイト／13．携帯電話・PHS、スマートフォンでみるインターネットサイト｝

C2=｛9．パンフレット・カタログ・ダイレクトメール｝

C3=｛14．インターネットブログ、ブログ／21，YouTube／16．電子書籍（電子書籍端末や電子ブックリーダー｝

C4=｛15．ツイッター（Twitter）／22．ニコニコ動画／17．ミクシィ(mixi)／18．フェイスブック（Facebook）／19．グリー（GREE）｝

C5=｛3．ラジオの番組／11．所属する会や組織の会報・同人誌・ニュースレ／10．都・県や市・区など自治体の広報誌紙／4．新聞の記事（電子版を含む）／8．各分野専門の情報誌の記事／20．モバゲータウン｝

ここで「情報源」（メディア）を 5 群に分類し，また「評価項目」を 4 群に分けた（図 4）．このクラスター化では以下のことに注意しよう．

- このデンドログラムは対応分析と“どういう関係”にあるのか（どのように作れるのか）．
- クラスター化で成分スコアはどのように用いるのか（機能するのか）．
- クラスター数はどのように決めたのか．デンドログラムをみて恣意的に行ってはいけない．
- 「情報源」と「評価項目」は，それぞれ“別々にクラスター化”していること．
- 図 5 をみて，この図の上にある点の布置で「情報源」と「評価項目」を同時にくくって分類してはいけない．それはなぜか．

こうした，一見すると単純にみえる手順や疑問にどのような回答があるのだろうか．

寄与度による解釈　－図４の解釈の例

成分スコアの解釈，次元（軸）の解釈などを客観的に行ういろいろな指標が提案されている．１つの例として，「評価項目」の側の“絶対寄与度”（どの軸にどの程度寄与するかを測る）を作ってみた．元の 2 元データ表の寸法は，表 2 にあるように「22（行）×9（列）」である．よって，対応分析の性質から，えられる次元数は 8 である．図 3 に対応するはじめの 2 次元（2 軸）について，絶対寄与度をグラフとした．

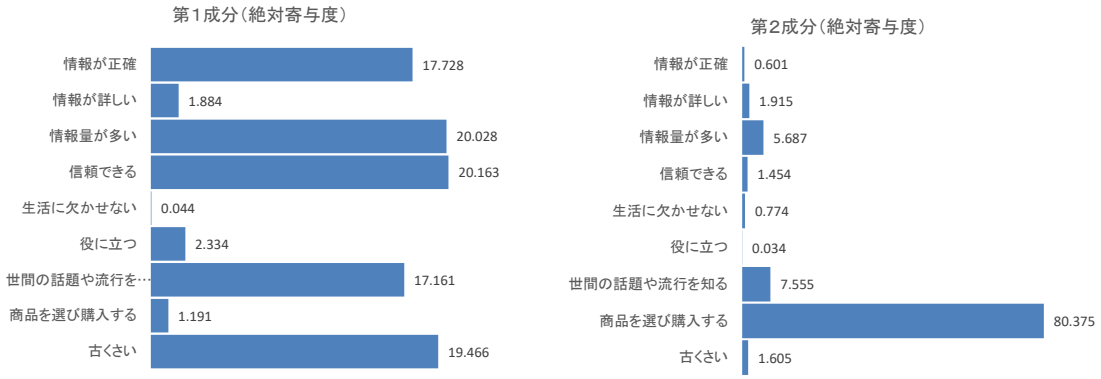


図７　「評価項目」の絶対寄与度のグラフ

まず，第 1 成分軸の「評価項目」の各項目に与えられた絶対寄与度と図の第 1 軸の方向に向かって各項目を観察すると，重心（平均的位置）から，左右に遠い項目で値が大きいことが読める．第 2 軸方向では，「商品を選び購入する」という項目だけが（この 2 つの軸の範囲では）顕著に大きいことがわかる．

また，（この 2 次元内では）「役に立つ」「生活にかかせない」「情報が詳しい」が重心（平均）のあたりに布置し，つまりこれらはこの 2 つの次元ではさほど有意に寄与していないと読める．

このように，寄与度をもちいて，布置図内の個々の点（ここでは評価項目）の各成分軸に対する関係を“客観的に調べる”ことが大切である．布置図を見て，“恣意的な印象で解釈を行う”ことは控えるべきである．同様の観察は，「情報源」の 22 の項目についても可能である（ここでは略す）．

絶対寄与度の他に，相対寄与度（ある項目がどの軸でどの程度説明されるか，近似の程度を測る）ほか，いろいろな評価指標が考えられている．このほかの諸指標を用いて，総合的かつ客観的に情報を評価することが必要である．対応分析は，多次元情報を単に布置図，同時布置図でグラフィカルに観察するだけのツールではない．

こうした対応分析の正しい使い方を紹介することが，本セミナーの目標である．

【分析例 2】自由回答質問の分析(1) ―データ表の構造探査の基本―

0. 調査の概要

まずここで扱う調査データの概要を示す．このデータを［分析例 2］［分析例 3］で用いる．

- 調査課題：普段の生活やインターネットなどについて
- 調査対象（標本抽出枠）：あるウェブ・パネル（非公募型）に登録の首都 40 ㎞圏・近畿 20 ㎞圏に在住の 12 歳以上 65 歳未満の男女（パネル構成の詳細は省く）
- 調査実施期間：2005 年 03 月 16 日　17:00　～　2005 年 03 月 23 日　17:00
- 調査方式（モード）：ウェブ調査
- 計画標本の大きさは 857（人），有効回収標本の大きさは 529（人），参加率 61.7（%）

メモ:

- 調査実施時期が，だいぶ前であることに注意する．つまり，自由回答の中に，最近のソーシャル・メディアに関する語句など（例：ブログ，ツイッター，フェイスブックなど）が用いられるチャンスは少ない．
- 多くのウェブ調査データの特徴として，回答者の人口統計学的変数，とくに性別，年齢区分に偏りがある[1]．この調査データも例外ではなく，うしろに示したように回収された回答者の属性分布にはかなり偏りがある（図 13）．分析を行ううえで得られた結果解釈時には注意が必要である．

1. 用いる自由回答質問と課題

図 1 にある 2 つの自由回答質問のうち「(インターネット利用が) プラスになると思われること」(以下，「プラスになると思うこと」と記す）に回答者が記入した自由記述データを分析する．ここで「プラス」面と「マイナス」面に分けて質問したことに注意しよう．

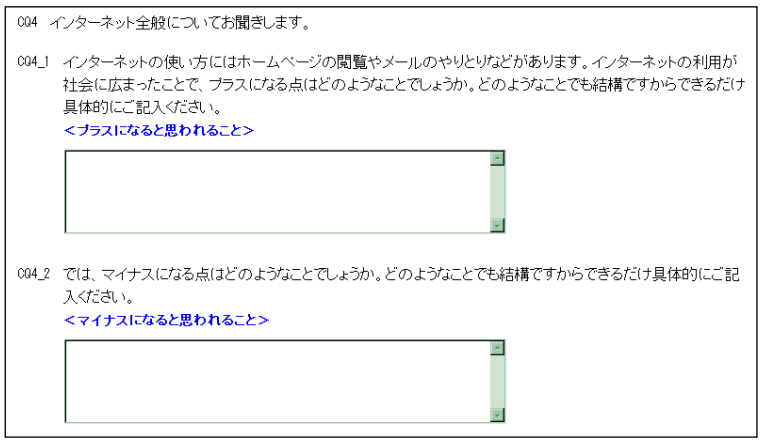


図 1　分析に用いる自由回答質問

[1] ウェブ調査では，多くの場合，ボランティア・パネル（非確率標本）を用いる．パネルの人口統計学的変数とくに性別，年齢区分の分布構成には偏りがある．パネルを標本抽出枠と考えて，調査対象者を選ぶときに，国勢調査情報を参考に，割付法で性年齢を割り付けて計画標本を無作為抽出したとしてもこの偏りは消えない．また，回収された回答者の属性分布は計画標本の属性分布からさらに「ずれ」を生む（例：若年層が回収できないなど）．これを補填するために追加標本を抽出するなどを行うことがあるが，これでさらに別の偏りが介入する．つまり，調査対象者や回答者が何を代表するのかかがきわめて曖昧になる．

2.　自由回答に登場する主要語句は何か？

まず，この調査の回答者が自由回答のなかで用いた「語句の頻度分布」を調べる．続いて，「検索機能」を用いて語句が回答文のなかでどのように用いられたかを観察する．

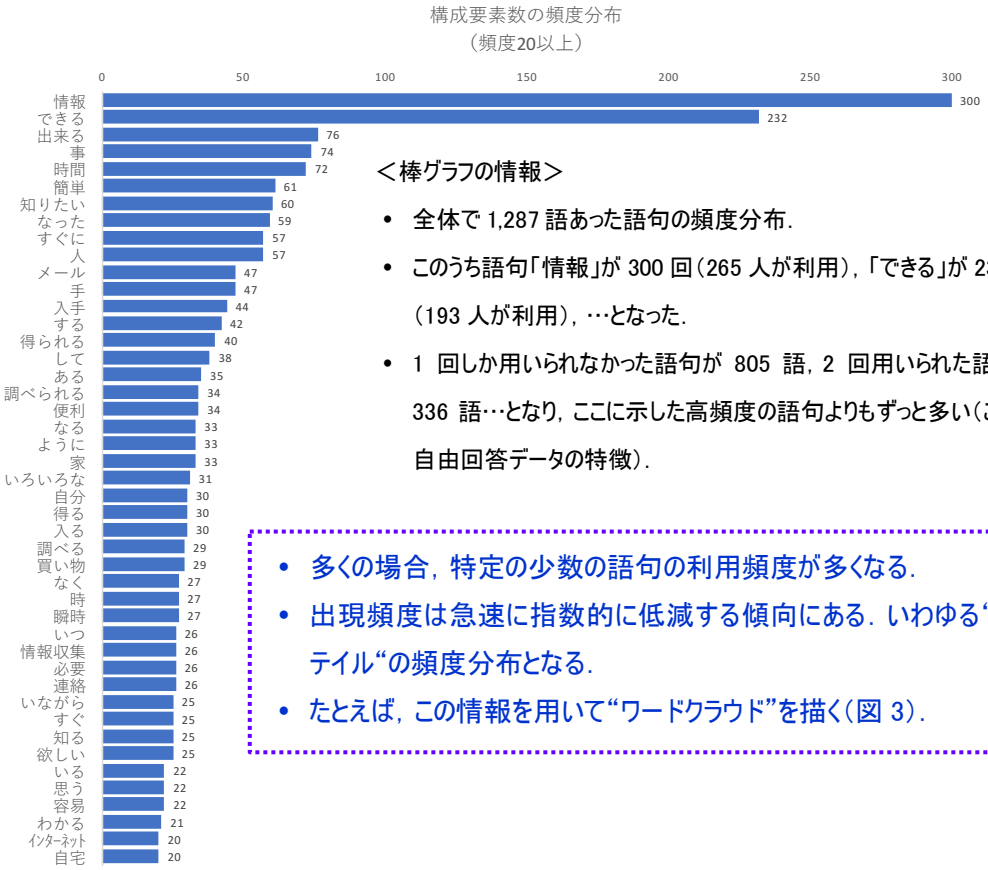
語句の利用頻度の観察 － 予備分析 1 －

「プラスになると思うこと」で，回答者はどのような語句を用いたのだろうか．得られた自由回答データを分かち書き処理し，若干の編集を行ったあとのデータを用いて，どのような語句が用いられたか，その頻度分布を観察する．これはテキスト型データ分析の基本操作である．

ここで，出現頻度の特徴を頻度グラフとワードクラウドとして描画した（図 2）．“ワードクラウド”は，最近，ウェブ上の多様なテキスト型データの視覚化ツールとして，ひろく利用されている．

メモ：ワードクラウドの利点，欠点

利点は，単語群の出現頻度など，用いる指標（の値の大小）に応じて，その単語群の様子を可視化できること．フォント色やサイズで，指標のサイズを表現できること．欠点として，描画方法が一意に決まらないこと，描画方法（アルゴリズム）がさまざまあって曖昧なこと，単語数が多いとが図内に収まりきれないこと，同値の頻度の処理の方法が曖昧なことなどがある．あくまでも一つの“目安”（初動探査ツール）として用いるとよい．ここでは，JMP®のテキストエクスプローラと Office アドイン（E2D3: Excel to D3.js）を用いた．



＜棒グラフの情報＞

- 全体で 1,287 語あった語句の頻度分布．
- このうち語句「情報」が 300 回（265 人が利用），「できる」が 232 語（193 人が利用），…となった．
- 1 回しか用いられなかった語句が 805 語，2 回用いられた語句は 336 語…となり，ここに示した高頻度の語句よりもずっと多い（これが自由回答データの特徴）．

- 多くの場合，特定の少数の語句の利用頻度が多くなる．
- 出現頻度は急速に指数的に低減する傾向にある．いわゆる“ロングテイル“の頻度分布となる．
- たとえば，この情報を用いて“ワードクラウド”を描く（図 3）．

図 2　用いられた語句の頻度分布（構成要素変数の分布）

図 3　ワードクラウドによる主要語句の観察

コンコーダンスによる観察 － 予備分析 2 －

簡単な分析として，検索機能を用いた語句検索を行ってみよう．さまざまな検索機能があるが，自由回答のようなテキスト型データの場合は，内容分析などで用いる**“コンコーダンス”**あるいは“KWIC”（key word in context）が効果的である．利用頻度の多いいくつかの語句「情報」．「情報収集」「情報入手」を選んで，検索を試みる．

検索例 1：「情報」のみで検索すると，369 件がヒットした．その一部を下に示した．「情報」という語句が，どのように自由回答内で用いられたかが，わかる．

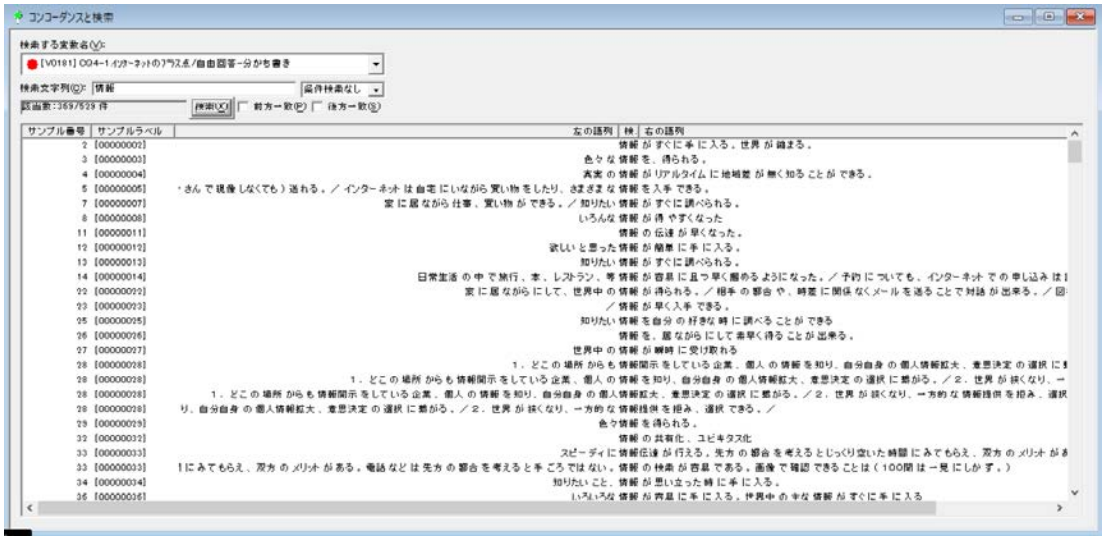


図 4　コンコーダンスの例（語句「情報」をどのように用いたか）

見にくいので数人分を抜き出してみると以下のような自由回答となっている．
「情報がすぐに…」「情報がリアルタイムに…」「情報の伝達が…」「情報が容易に…」「…知りたい情報が…」「…さまざまな情報が…」といった語句の使われ方が見えてくる．

[00000002]		**情報**	が すぐに 手 に 入る 。 世界 が 縮まる 。
[00000003]	色々 な	**情報**	を 、 得られる 。
[00000004]	真実 の	**情報**	が リアルタイム に 地域差 が 無く 知る こと が できる 。
[00000005]	メール は 相手 の 時間 や 状況 を 気 に せず 、 こちら の 意思 を 手早く 伝える 事 が 出来る 。 郵便 と 異なり 写真 も 簡単 に （ 写真 や さん で 現像 しなくても ） 送れる 。 ／ インターネット は 自宅 に いながら 買い物 を したり 、 さまざま な	**情報**	を 入手 できる 。
[00000007]	家 に 居 ながら 仕事 、 買い物 が できる 。 ／ 知りたい	**情報**	が すぐに 調べられる 。
[00000008]	いろんな	**情報**	が 得 やすく なった
[00000011]		**情報**	の 伝達 が 早く なった 。
[00000012]	欲しい と 思った	**情報**	が 簡単 に 手 に 入る 。
[00000013]	知りたい	**情報**	が すぐに 調べられる 。
[00000014]	日常生活 の 中 で 旅行 、 本 、 レストラン 、 等	**情報**	が 容易 に 且つ 早く 掴める ように なった 。 ／ 予約 について も 、 インターネット で の 申し込み は 割引 が あり 、 便利 に なった 。

検索例 2:「情報」と「情報収集」の 2 つを「AND」条件で検索する．ここでは，55 件は選ばれた．その自由回答文の一部が以下である．

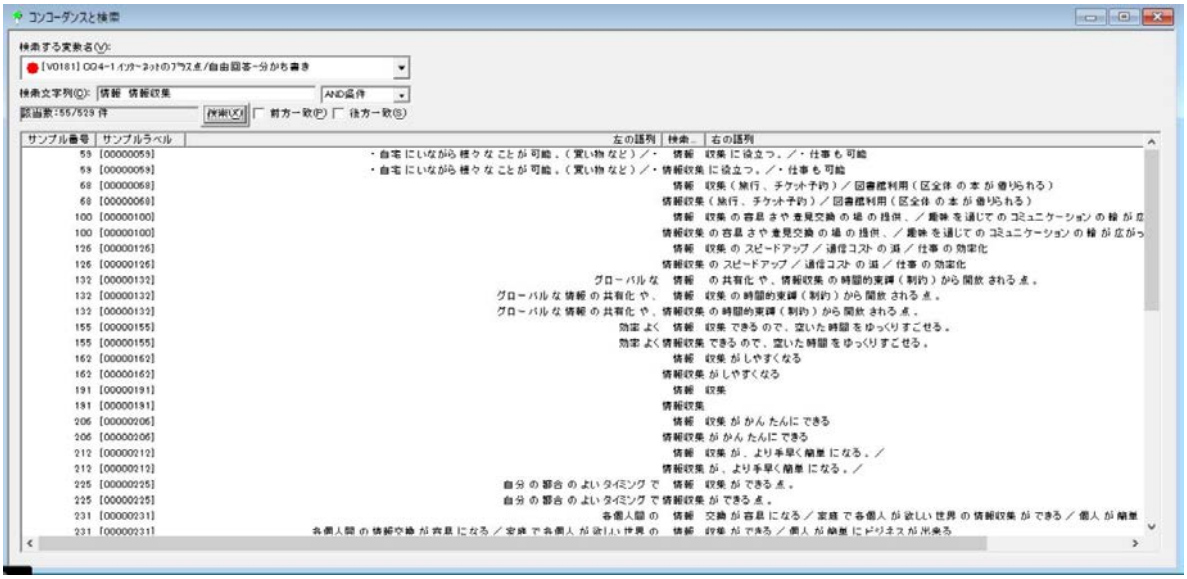


図 5 コンコーダンスの例 2（語句「情報」「情報収集」を同時に用いた自由回答）

これも数例を取り出してみる．ここでは，同じ回答者が「情報」「情報収集」のいずれも用いていること（図４の「サンプルラベル」を参照），そのつながりが読み取れる．

[00000059]	・ 自宅 に いながら 様々 な こと が 可能 。（ 買い物 など ） ／ ・	**情報収集**	に 役立つ 。 ／ ・ 仕事 も 可能
[00000068]		**情報収集**	（ 旅行 、チケット予約 ） ／ 図書館利用 （ 区全体 の 本 が 借りられる ）
[00000100]		**情報収集**	の 容易 さ や 意見交換 の 場 の 提供 、 ／ 趣味 を 通じて の コミュニケーション の 輪 が 広がって いく 楽しさ 。
[00000126]		**情報収集**	の スピードアップ ／ 通信コスト の 減 ／ 仕事 の 効率化
[00000132]	グローバル な **情報** の 共有化 や 、	**情報**	**収集** の 時間的束縛 （ 制約 ） から 開放 される 点 。
[00000162]		**情報収集**	が しやすく なる
[00000231]	各個人間 の 情報交換 が 容易 に なる ／ 家庭 で 各個人 が 欲しい 世界 の	**情報**	**収集** が できる ／ 個人 が 簡単 に ビジネス が 出来る
[00000338]	検索 に よる **情報収集** 、 遠方 へ の	**情報**	**伝達** や コミュニケーション 等 が 、 すぐに 出来る点 。 ／

この小さな例でもわかるように，コンコーダンスを用いる利点は，以下のような傾向探査が可能となることである．

- 回答者が“どういう自由回答文を記した”のかを調べること．
- これらの語句が，自由回答文の中でどう使われるか．
- とくに，検索語句の前後の回答の表現を観察する．
- 複合検索で，複数の語句の“係り受け”の傾向を知る．

さらに検索を続けると，
「情報」「情報入手」を「AND」条件で検索すると　→　10 件が該当
「情報」「情報収集」「情報入手」を「AND」条件で検索すると　→　該当なし
「情報」「情報収集」「情報入手」を「OR」条件で検索すると　→　400 件
となる．

とうぜん，「AND」検索で指定語句を増やせば，該当する件数は低減するが，検索語句数が少なくても該当なしのチャンスが意外と高い．回答者は“さまざまな表現，書き方”をし，“意見はいろいろ”ということである．

- 特定の語句の利用頻度が多く，集中して登場するかどうか
- 利用語句の種類が多様で，個々の語句の利用頻度は散らばるかどうか
- 全体の語句の種類が限られ，変化が少ないかどうか

こうした特徴は，“質問文の作り方”（尋ねたいこと）に関連する．質問文の問い方が曖昧であるほど，自由回答の表現は**多様で曖昧な内容**となることは，多くの研究でわかっている．実は，ここの例は，やや問い方（内容）が曖昧かもしれない．

一度尋ねた自由回答質問の内容を傾向探査し,「インターネットの利点, プラスと思うこと」を聴きだすには，どのような質問文が適切であるか，を調べることが必要であり，そのためのツールとして対応分析，多重対応分析，そしてクラスター化が有効である.

分析のポイント＜1＞

① コンコーダンスなどで事前探査のあと，自由回答の“利用語句の出現頻度”を頻度グラフやワードクラウドで視覚化する.
② 視覚化の限界を知る，利用ツールやアルゴリズムの違いによる結果の差違に注意,
③ 用いる“指標”による視覚化の違いを知る.“錯視”に注意しよう.
④ 質問文の作り方と，自由回答の語句（構成要素）の客観的観察が重要である.
⑤ データの“探索発見的”,“インサイト力”の効用を知る.

視覚化ツールによる観察 － 予備分析 3 －

回答者は「プラスになると思うこと」をどう捉えているか.「性別，年齢による意見（自由回答）の類似，差違はあるのか」,「利用語句の違いはあるのか」，これを単純に“利用語句の出現頻度”で比べてみよう[2].

ここで性年齢別の語句の出現頻度の特徴をワードクラウドとして描画した．あえて多数の図を示し，性別（男性，女性）ごとに，年齢区分（20 代，40 代，60 代）の回答者の利用語句の特徴（とくに差違の有無）を比べる.

- 図 6，図 7 は,「語句の出現頻度の多少」を文字列の大きさと色に合わせて描いたワードクラウドである.
- 図 2，図 3 の語句全体の分布と比べると，高頻度語句には差違がなさそうである.
- いずれも「情報」「できる」「時間」が多い．当然，予想されることである.
- これに続いて「簡単」「入手」「知識」などが多いようにみえる.
- 文字列の大きさが急速に小さくなる．出現頻度の大きい語句は“限られている”こと.
- 主要な特徴は“少数の語句に限られて”おり，性差や年齢区分の違いがあるようだが，具体的には“その違いがみえない”こと.

男性 20 代：**「情報」「できる」**「知識」「する」「簡単」「手」…などが主要語句
男性 40 代：**「情報」「できる」**「情報収集」「必要」「得られる」「入手」時間」…
男性 60 代：**「情報」「できる」**簡単」「知りたい」「得られる」「入手」「早く」…
女性 20 代：**「できる」「情報」**「人」**「出来る」**「知りたい」「事」「メール」「自分」…
女性 40 代：**「情報」「できる」**「時間」「瞬時」「出来る」「買い物」「手」すぐに」…
女性 60 代：**「情報」「出来る」「できる」**「得られる」…「インターネット」などと続く

[2] 語句の出現頻度で比べる理由はなにか．数個の語句からなるある文章の一部をぬき出して検索すると，かなり長い文章であっても，検索で該当する頻度はほとんど１回である．つまり，語句の使われ方は非常に多様である.

男性 20 代

男性 40 代

女性 20 代

女性 40 代

図 7　男性・女性（20 代，40 代）のワードクラウド

男女とも，いずれの年齢層も，きわだって高頻度の利用語句は，よく似ており差違がなさそうで，（違いがみえにくい．本当だろうか，別の見方があるのではないか．これを調べることが目標であ

3. 視点をかえて探査する

語句の出現頻度ではなく，“別の指標”（検定値：ある検定統計量の実現値）を用いてワードクラウドを作ってみる（図 8～図 10）．いずれの図も，前の結果とはだいぶ異なる．たとえば，性別，年齢区分の特徴語句は以下のようになり，上でみた結果とは異なる．なお，年齢が 30 代，50 代ももちろん確認できるがここでは省略した．

ワードクラウドの観察で得られる特徴

男性 20 代：**「様々」「商品」「判断」「迅速」「情報」「知識」「調べ」「取得」「使用」**…
男性 40 代：「情報収集」「必要」「時間的」「程度」「なくなる」「迅速」「自宅」「世界」…
男性 60 代：上のいずれとも異なり，「選択」「情報」「世界」「早く」「入手」「得られる」「容易」「誰」「個人」「場所」など．
女性 20 代：**「友人」「相談」「人」「参加」「思った」「興味」「やすい」「ネット」「自分」**…
女性 40 代：「瞬時」「時間」「タイムリー」「関係」「入れる」「ほしい」「資料」「活用「手間」「利用」「価格」「なく」…とある．
女性 60 代：「本」「出来る」「できる」「予約」「場合」「知識」「関係」「して」「購入」「物事」「得られる」「社会」…とある．

20 代（男女）の発語の比較

まず，20 代（男女）について，比較の指標とした検定値の情報（一部）を棒グラフとワードクラウドとして示した（図 8，図 9）．この検定値の特徴は，性年齢区分ごとに，全体の語句の分布（図 1）からみて，それぞれの区分では“どの語句がより利用されたか”（上位語句），あるいは逆に“利用されていないか”（下位語句）を，示している．

図中に示した情報（検定値の大きさ，その対応語句，ワードクラウド）から，20 代男女の自由回答に登場した語句の使い方には，だいぶ違いがあることがわかる．

いずれも前の観察では頻出した「情報」「できる」「出来る」などは，主要語としては（フォントサイズの大きい文字列としては）出てこない．“なぜ，登場しない”のだろうか．実は目にはみえないだけで，用いた検定値に，そのような識別する機能があることがワードクラウドに反映されているのである．

同じデータを用いても，選んだ指標によりワードクラウドも描画結果の印象も違ったものとなる．用いる指標が重要という示唆である．2 つのワードクラウド情報から受ける印象（みえるもの，視覚化情報）にはずいぶん違いがある．しかし，“いずれもが正しい”のである（視覚化の特徴であり限界）．

その他の年代の特徴

40 代，60 代の男女の発語の傾向を，ワードクラウドで観察しよう．ここでも上と同じように，語句に対する検定値を用いて図 10 の 4 つのワードクラウドを作った．それぞれの図に現れた語句の特徴は，こまかい説明を必要とはしないであろう．

なお，回答者数が多い，30 代の男女の傾向については，うしろで対応分析の結果とあわせて観察する．

男性 20 代

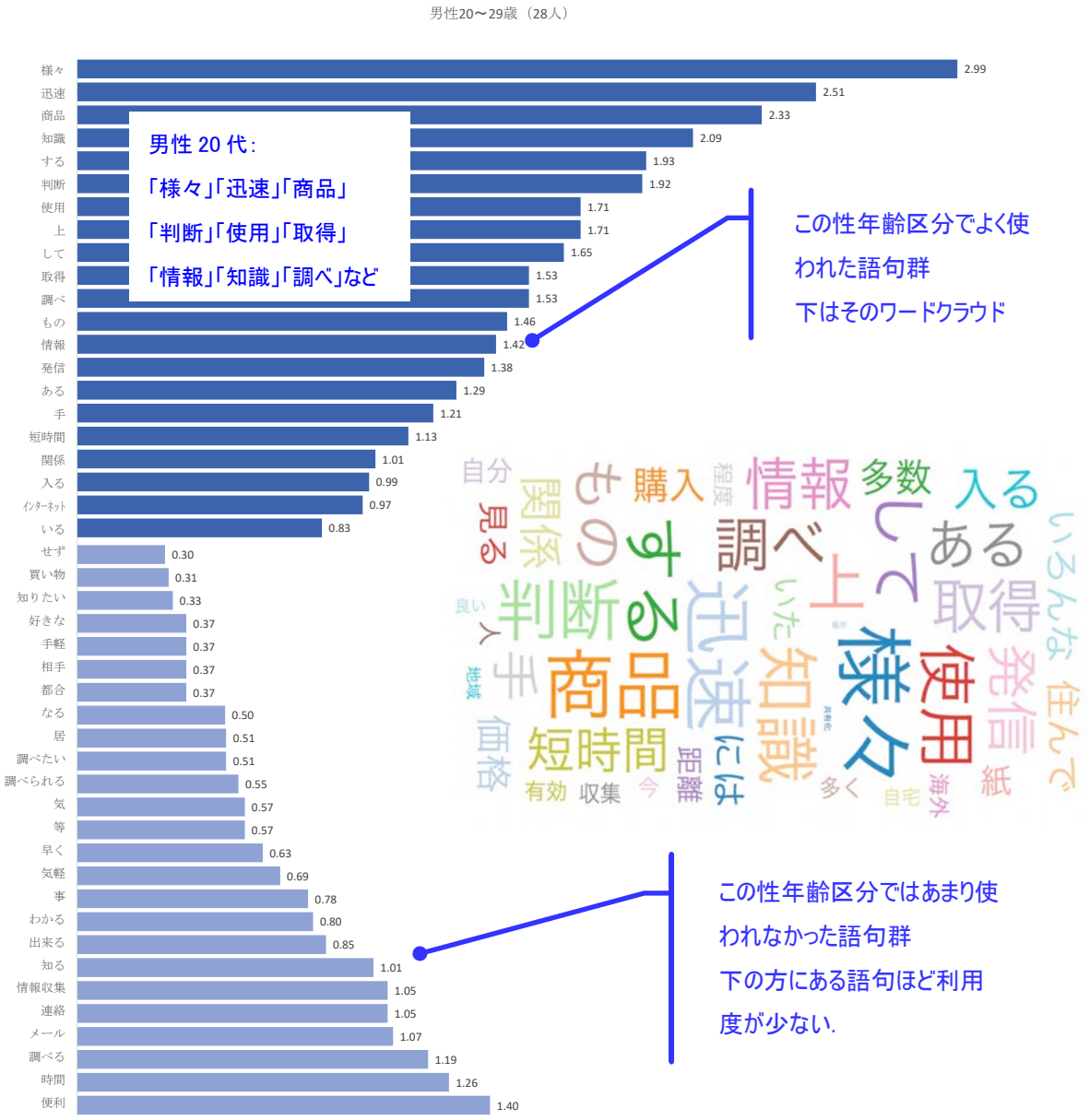


図 8　男性 20 代の利用語句，検定値，ワードクラウド

女性 20 代

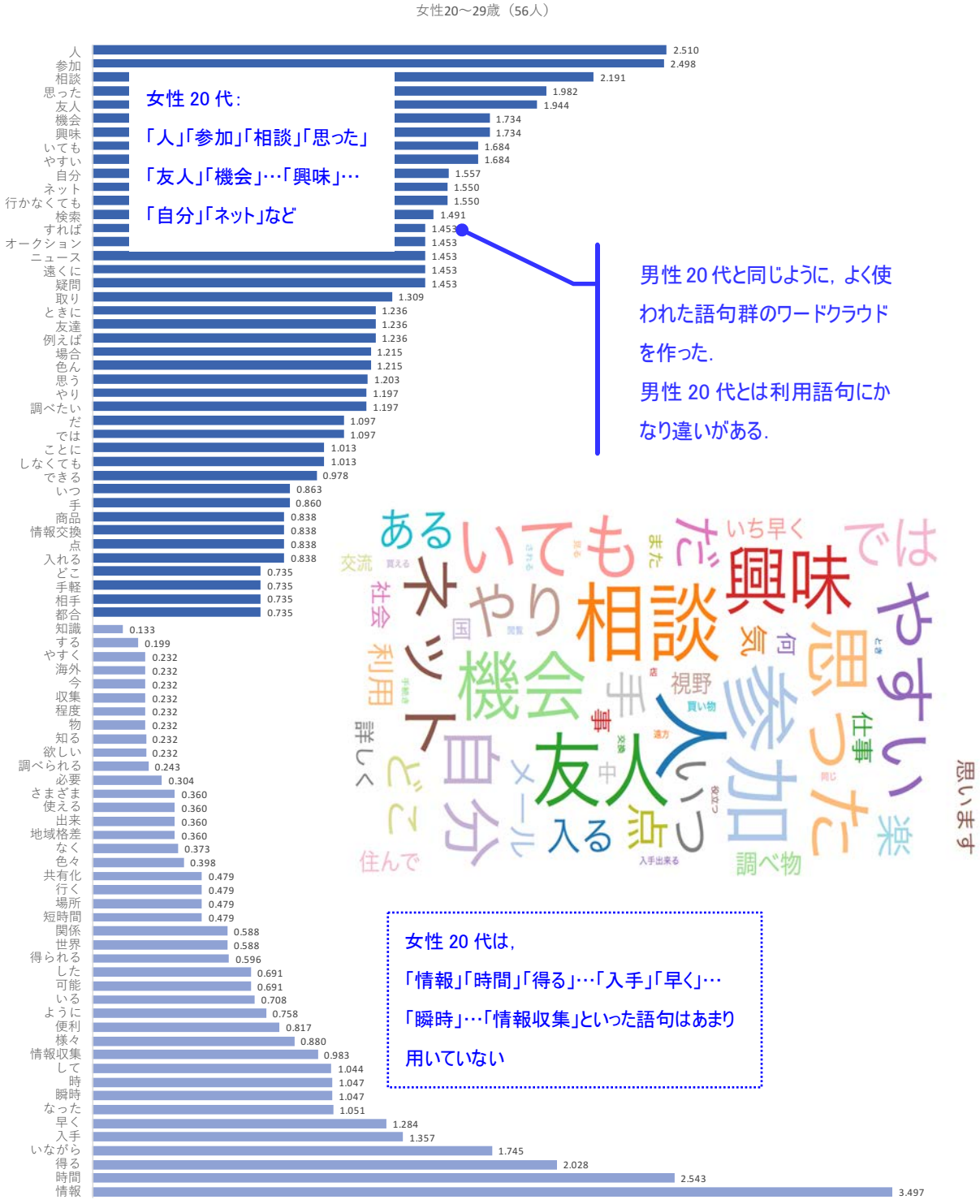


図 9　女性 20 代の語句群，検定値の特徴とワードクラウド

男性 40 代

男性 60 代

女性 40 代

女性 60 代

図 10　検定値を用いて描いたワードクラウド（男女，40 代，60 代）

4.自由回答の処理と対応分析による探査

- **「分かち書き処理」**[3]のあと，句読点・記号類他の除去を行い，出現頻度が 4 語以上の単語・語句[4]を抽出した[5].
- ここでは，あえてこれ以上の**辞書編集**などは一切行わないこととした.
- これで得られた総語句数は 3,501（語），このうち重複語句として計数しない分析対象とする異なり語句数（異なり構成要素数[6]）は 235（語）となった.
- これはデータ処理量としては，少ない方である．通常はさらに量が多い.

分析の目標 － 何を知りたいか

ここの目標は，調査で用いた質問「プラスと思うこと」の自由回答からえた語句群と，調査票にある「質的変数」（人口統計学的変数や選択肢型質問）との関係を探査的に検証することである．たとえばここで，質的変数として人口統計学的変数の「性年齢区分」を選ぼう.

分析のポイント＜2＞

① 自由回答データから得られる情報だけでは，十分な知見が得られるとはかぎらない.
② 自由回答（非定型データ）から得た語句群（構成要素変数）の情報と，定型情報である「質的変数」（人口統計学的変数や選択肢型質問）との関係を探査すること.
③ 質的変数として人口統計学的変数（とくに属性）との関連を調べること.
④ これは，「(構成要素変数）×（質的変数)」の“2 元データ表の分析”に帰着できること.
⑤ 具体的な手法として，対応分析とクラスター化法を用いること.
⑥ 分析のシナリオは 1 つではないこと．たとえば，質的変数である選択肢型質問の選び方を変えて（例:「あなたは，現在の生活にどの程度満足していますか［現在の生活満足度］;「あなたの現在の暮らしの経済状況は余裕のある方だと思いますか［経済の余裕度］」，これとの関係を調べるなど.
⑦ どの質的変数が構成要素変数と高い有意性があるかを“客観的に”知る（数値を与える）という意味で“探索発見的”かつ“記述的・帰納的”であること.

自由回答でえた語句群と，回答者の性年齢を 2 元データ表として要約した表 3 を分析対象とする．つまり，得られた自由回答と，性年齢区分別の間にどのような特徴，傾向があるかを探査する.

通常，データ表の寸法はかなり大きくなるので，一部を切り取って示した（表 1)．この程度の寸法のデータ表でも各セル内の度数は“**かなり疎**である”．一般にデータ表の寸法は非常

---

[3] 日本語のテキスト型データは，分かち書き処理を行って区分化（segmentation）することが必要である．①分かち書き処理は用いるソフトによって結果が異なること，②分かち書き処理のあとに形態素解析まで行うと，さらに結果に違いが生じること，③処理のオプションの指定によっても違いが生じるなどがある．つまり．同じデータセットであっても分析の出発点で行った処理によって，すでに内容に違いがあることに注意しよう（先の分析を進めてその結果だけを比較するなどは要注意である).

[4] 形態素とは，単語を細分析してえられる意味を持った最小の言語単位のこと．ここでは，単語や語句というやや曖昧な言い方を用いる．単語は最小の言語単位ではない.

[5]通常，出現頻度は“1 語”（1 回のみ使用）がもっとも多く，あと指数的に低減する．よって，4 語以上というとかなり少なくなっている.

[6] ある語句が複数回用いられるので，これを 1 と計数したときの語句数.

に大きく，また疎なデータ表となる．事前情報として，このデータ表の見方を簡単に説明しよう．

- 表の「構成要素」とは，自由回答の分かち書きでえた“分析対象とする語句の単位”である．また「行和」の欄はその構成要素の出現頻度，つまり何回使われたかを示す．また，この大きさで降順にソートしてある．たとえば「情報」はもっとも利用頻度が高く（当然だが，実は使われ方が異なる），全体で 3,501（語）ある中で 300（語）となることを示している．
- すでにみたように語句「情報」に注目すると，総じて男女ともに利用頻度が高い（しかし，おおむね 20 代から 40 代あたりが多い）．ただし，性年齢区分の各層の回答者数の大小に注意しよう．（女性の 30 代がもっとも多い）．回答者数は計画標本と回収標本を手にした段階で決まってしまう値である（その意味で属性は“説明変数的”である）．うしろにあげた図 13 の有効回答者の性年齢区分の構成分布を確認しよう．
- 他の語句も，（表示の範囲で）何か特徴があるようにみえるが，はっきりした“構造”は確かではない．
- ここで別の事象に注目する．いま「できる」「出来る」の 2 語が上位のほうにある．さらに「出来」「出来る点」もある．“これらは同じ意味”であろうか．厳密には“個々の原文（自由回答）にあたって”調べねばならない．かりにこれらを“同じ意味，使い方である”とするなら，“1 語にまとめて**“同義語”**として”扱う必要がある．これが，**“辞書編集”**を必要とするもっとも単純な例である．
- 前述のように，ここでは“こうした同義語などの辞書編集”は**一切行わない**で，つまりほとんど前処理なく分析を進めた（手間を省いた）ことを記憶しておこう．
- また，はじめに出現頻度の少ない語句は**切り捨てたこと**も覚えておこう．つまり，1〜3 回しか登場しない語句，とくに 1 回しか自由回答原文に登場しない語句は，ここでは分析には登場しない（図 11）．しかし実際は少数頻度の語句，とくに 1 回登場の数が圧倒的に多いのである．また利用語句の頻度分布は“必ず峰のない”指数的に低減する典型的な“ロングテイルの分布”となる（図 2）．

メモ:

参考までに，語句の利用頻度情報を確認しておこう．上述のように，利用語句の頻度分布は“必ず峰のない”分布となることは調べた（図 2）．この分布情報から，頻度区分ごとの語数の分布を作ると，図 11 のようになる．この 2 つの図の情報から，ほとんどの語句の利用頻度はわずか数回であること，特徴的な利用頻度の高い語句の割合は少ないことなどがわかる．このことは，この調査データに限ったことではなく，自由回答質問や一般のテキスト型データの分析で**“共通にみられる特徴”**である．

この，分析対象とする語句の拾い出しやスクリーニングの操作，さらに辞書編集処理などは，分析結果を大きく左右するので（結果がさまざま），テキスト型データの分析では非常に重要なことであるが，多くのテキスト・マイニング・ツールはこうした基本操作の説明や情報提供が十分ではない．

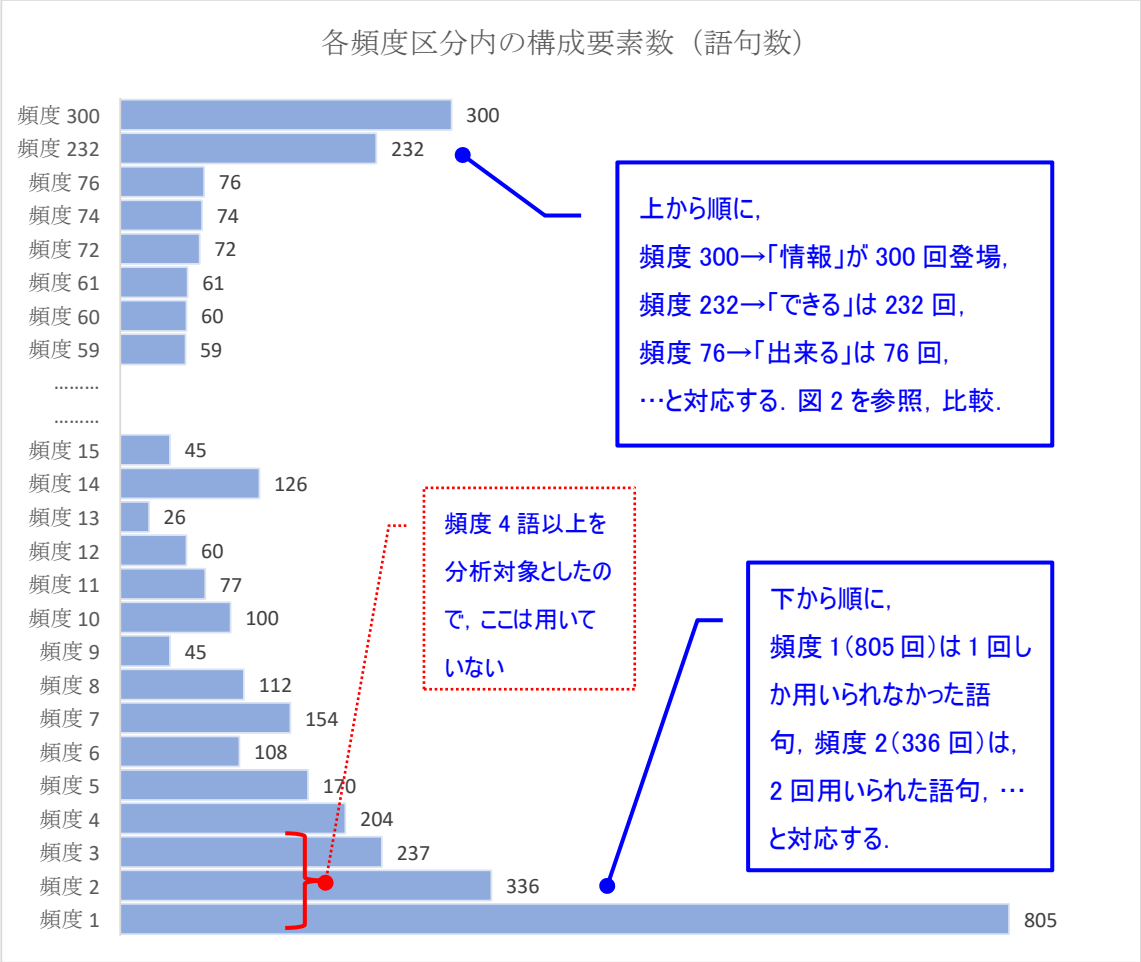


図 11　分析に用いる語句の利用頻度情報

結果と観察 ― 何がみえるか

1) 対応分析法では，データ表の行／列の各要素にいわゆる**“成分スコア”**が与えられる．これを“布置図”として視覚的に観察する．図 12 は性年齢区分（12 区分）の布置図である．図に書き入れた点線の囲みと矢印から，①性別により回答傾向が異なること，②とくに，年齢が高くなるほど（50 代，60 代），性差が顕著であること，③（2 次元までの情報では）男性が，女性にくらべて（成分スコアの）変動が少ないようであること，④男女「10 代」の傾向は他の年齢区分とすこし異なる“ようである”，などがみえる．

2) 語句の傾向はどうであろうか．年齢区分に重ねて“同時布置図”を描いてみる（図 13）．ここでは“わずか”235（語）の語句の布置が加わっただけで図はかなり煩雑になり**特徴が見にくくなる**（視覚化の限界）．図の中心（重心）つまり平均的な回答から離れて周辺に散在する語句の特徴は観察できるが（こういう観察は意味があるが），他の多くの語句は難しい．これは客観的ではないし恣意的である．

3) ではどうすれば，性年齢区分と語句との関係を合理的に調べることはできるのか．

表 1（構成要素）×（性年齢区分）のデータ表（寸法 235 語句 × 12 区分から一部を切り出し）

SEQ	登場した語句（一部）	行和	男性 12〜19 歳	男性 20〜29 歳	男性 30〜39 歳	男性 40〜49 歳	男性 50〜59 歳	男性 60〜69 歳	女性 12〜19 歳	女性 20〜29 歳	女性 30〜39 歳	女性 40〜49 歳	女性 50〜59 歳	女性 60〜69 歳
	回答者数→	503	9	28	55	48	33	28	20	56	119	60	36	11
	列和→ 利用語句（↓）	3,501	47	249	306	245	182	191	99	420	979	416	267	100
148	情報	300	5	28	43	29	17	27	7	18	64	36	17	9
33	できる	232	5	19	13	20	14	14	5	33	62	29	14	4
138	出来る	76	0	3	6	2	5	2	6	9	21	8	8	6
118	事	74	1	3	4	3	7	1	1	10	24	5	11	4
121	時間	72	1	2	5	5	3	3	0	2	30	15	5	1
84	簡単	61	1	6	3	2	6	6	1	7	21	5	2	1
181	知りたい	60	0	3	2	4	2	5	3	7	24	5	5	0
44	なった	59	2	3	11	4	3		1	4	16	5	4	3
22	すぐに	57	2	4	4	3	1	0	3	6	22	9	3	0
156	人	57	2	5	2	1	1	1	5	14	15	8	2	1
69	メール	47	1	1	4	2	1	0	1	7	16	4	7	3
130	手	47	1	6	6	3	0	0	1	8	11	9	1	1
207	入手	44	2	2	4	6	6	6	0	2	8	5	2	1
23	する	42	0	7	4	3								0
203	得られる	40	0	2	2	5								3
19	して	38	0	6	5	2								4
2	ある	35	0	5	3	3								2
190	調べられる	34	0	1	1	0								0
219	便利	34	0	0	2	1	0	1	3	2	12	5	7	1
46	なる	33	1	1	3	4	6	4	2	4	5	2	1	0
＜途中略＞														
167	前	4	0	0	2	0	0	0	0	0	2	0	0	0
173	速く	4	0	0	0	0	1	1	1	0	1	0	0	0
175	多数	4	0	1	1	0	1	0	0	0	0	1	0	0
176	大量	4	0	1	2	1	0	0	0	0	0	0	0	0
177	誰	4	0	0	1	0	1	2	0	0	0	0	0	0
186	中	4	0	0	1	0	0	1	0	1	0	0	0	1
189	調べたり	4	0	0	0	0	0	0	0	1	2	0	1	0
197	伝える	4	0	0	0	1	0	1	0	0	1	1	0	0
212	判断	4	0	2	1	0	1	0	0	0	0	0	0	0
215	幅広い	4	0	0	0	1	0	0	0	0	3	0	0	0

- 多くの場合，セル内の頻度が非常に少ない 2 元データ表となる．
- データ表の寸法も，かなり大きくなる．
- 初動探査として対応分析とクラスター化による“視覚化”が有効だが，それだけでは不十分．

(*) 自由回答内に出現の語句をソートし，235 語句のうち，出現頻度の大きい単語から順に並べてある（出現頻度が 4 まで）．

(**) 有効回収標本数 529（人）のうち，503（人）がこの自由回答分析の対象となった．

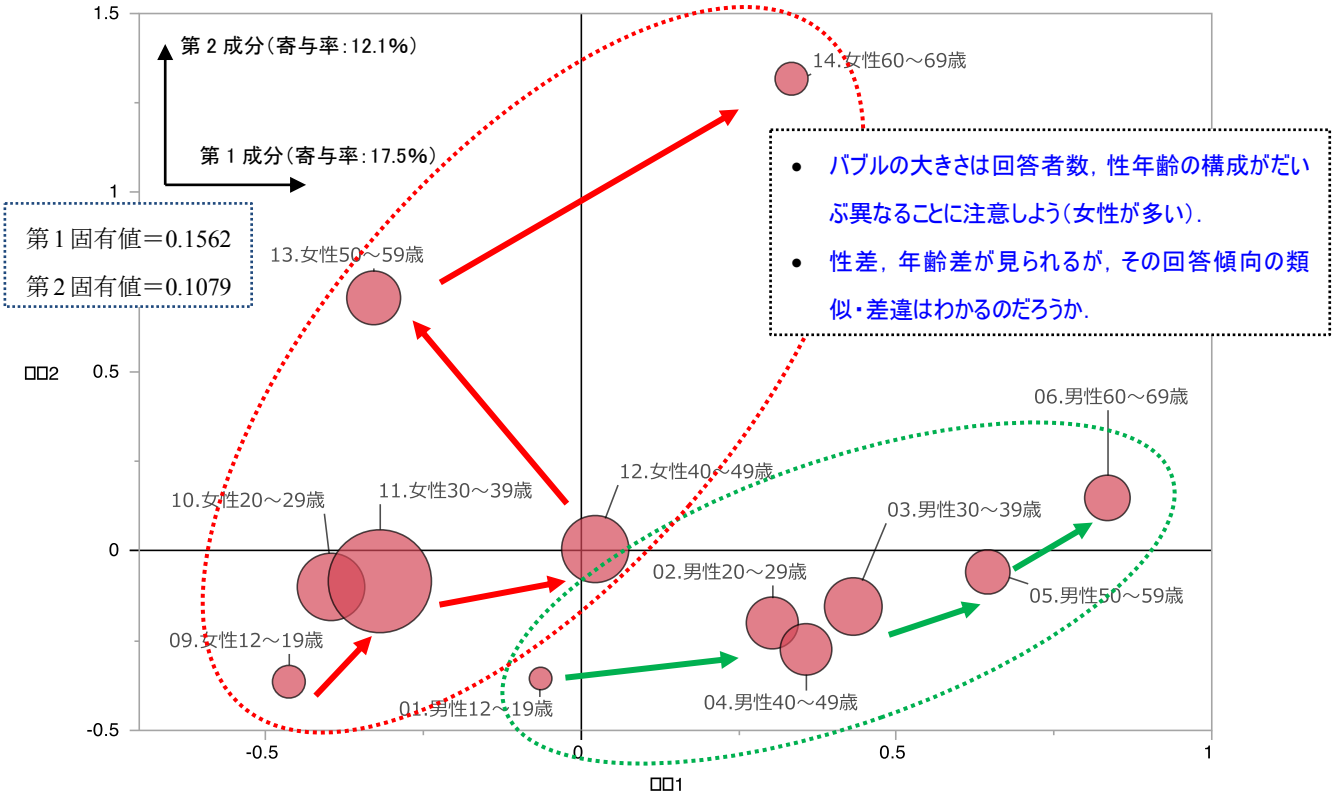


図 12　性年齢区分（12 区分）の成分スコア布置図

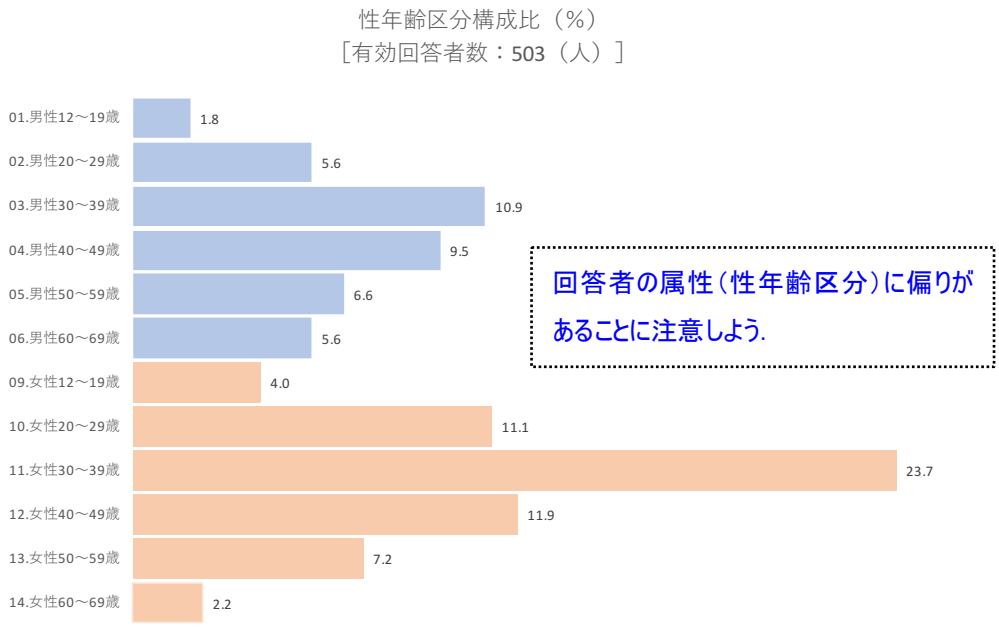


図 13　回答者の性年齢区分（12 区分）の分布

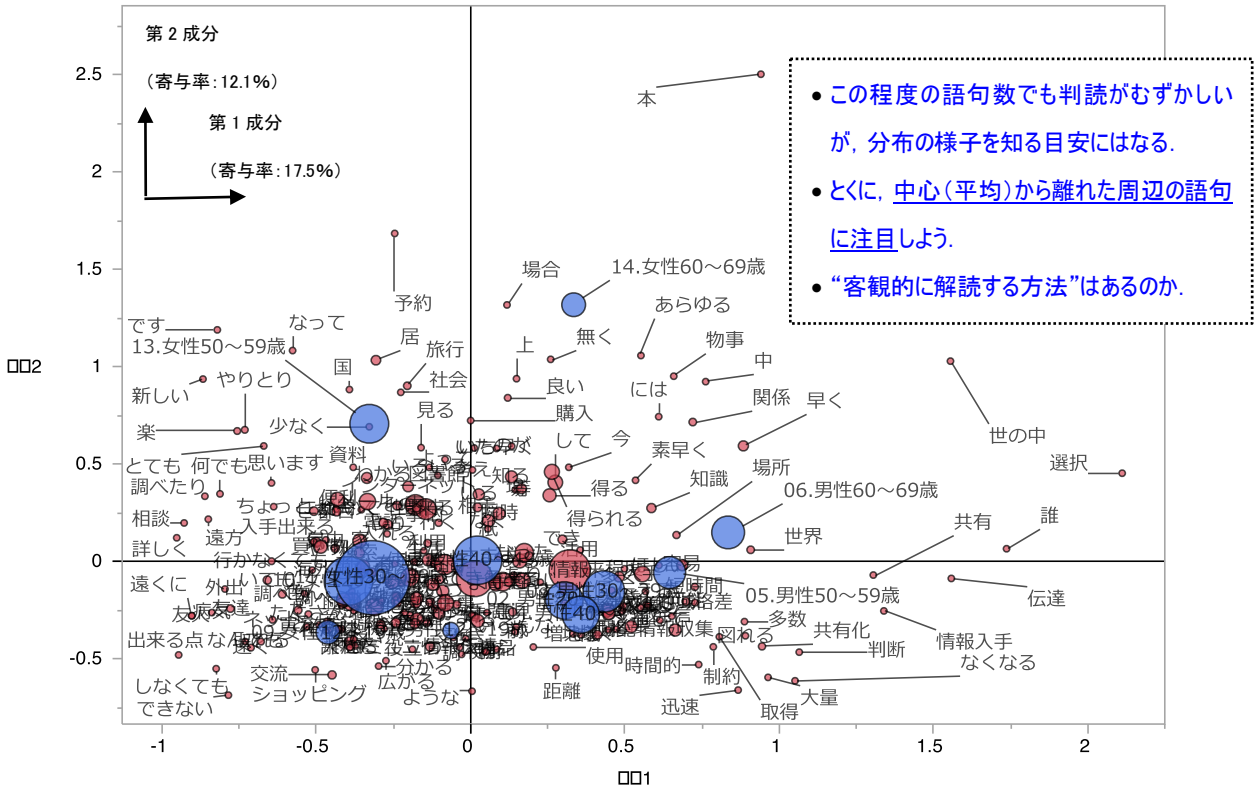


図 14　性年齢区分と語句の同時布置図（青いバブルが図 12 の性年齢区分に対応する）

5.　何が検討課題か　―データ構造を探査的に調べる

【分析例 1】でみたような問題点，課題は，ここでも同じように検討せねばならない．これに加えて，以下の諸要素の検討も必要である．

- 【分析例 1】と比べて，（図に書き入れた）固有値，寄与率に違い（値が小さい）があるがなぜか．またどう読むか．
- 布置図，とくに同時布置図は，この程度の点（語句数，性年齢区分の選択肢数）に増えただけで，すでに観察はむずかしい．**視覚化には限界がある**（この例に限らず多くの視覚化，可視化ツールの性質）．一般にはデータ表の寸法がさらに大きくなるが，どのように対応できるのか．
- 自由回答のテキスト型データを扱うことと，他の質的データ（質的変数とする選択肢型質問，人口統計学的変数，そしてクラスター化で生成されるクラスター変数）の関係をどう考えるのか．
- 一般に自由回答・自由記述情報だけを調べても，情報は十分でない．ではどのような情報を**自由回答データと比べればよいのか**．その調べ方はあるのか．
- テキスト・マイニングでは，対応分析法に必要なコア計算部分があるだけでは，不十分である．事前・事後の“探索的ツール”をいろいろ用意することが必要である（実用上はこちらが重要）．ではその探索ツール，マイニング・ツールとして，なにがあるのか．

きわめて簡単な“例”を挙げる．表 4，表 5 は，統計的なある操作で（“有意性テスト”と呼んでおく），ここで用いたデータ表の，性年齢区分別という属性の回答者が“どのような語句を積極的に用いたか”（上位），あるいは逆に“あまり使っていないのか”（下位）をランキングした情報である．

この解釈を詳しく示すことはやや面倒なので，1 つの例で示そう．いま，もっとも利用頻度の高かった語句「情報」に注目する（質問文からこれは想定されたことである）．これは 2 つの表 4，表 5 の“どこに表れている”だろうか．要点を列記しよう．

- もっとも利用頻度の高い語句「情報」はあちこちに散らばって登場するが，順位が異なる（性年齢区分により利用の重要度が違う）．たとえば，男性の「20 代」では上位 11 に，「30 代」では上位 1（先頭）にあり，「40 代」では上位 6，「60 代」では上位 4 である．
- 一方，女性は，（男性とは逆に）いずれも下位にある．「20 代」では下位 1 に，「30 代」では下位 2 に，そして「50 代」では下位 3 にある．
- これは何を意味するのか．「情報」がもっとも自由回答内に利用頻度が多いが，その“使われ方”が性別と年齢層で異なることを示唆している．
- しかも，男性がすべて「上位」にあるのに，女性は「下位」にある．これは，男性と女性では，回答した自由回答の中での“「情報」の使われ方”（重み）が異なることを示唆している．かりに「情報」が男女ともに似たような使われ方をしたならば，有意とはならない，ということである．これは“本当だろうか”．
- この分析例では記さないが，これを原文（つまりもとの自由回答文）に戻って，検索・比較・分類することで，これは確認できる．（注：【分析例 3】で別の例を示した）
- このようにみてくると，表に示した性年齢区分別の各層の上位，下位にある語句が，その層を“特徴付ける語句”であることも示唆している（実際にそうである）．
- ここでは，“プラスと思われること”という問いへの（もとは非構造的な）自由回答を整理（編集）して，これと“性年齢区分”という構造的な選択肢型質問（質的変数）とを比べてみた．
- では，この他の選択肢型質問や人口統計学的変数を組み合わせるとどうなるか．あるいは，他のアプローチはあるのか，たとえばクラスター化で回答者の類型化を行うとどうであろう．このような仮説発見的，探査的にマイニングを行う意味がある．

6. 再び検定値とワードクラウドを観察する ― 取得情報の併用の効用

実は，すでにこれと同じ操作で，20 代の男女について，この検定値の上位，下位の一部を棒グラフで示した．また，その情報からワードクラウドを描いてみた（図 15）．これを含め，前半でみたワードクラウドの元となった情報は，どこにあるのだろうか．この情報源（の一部）は，表 4 と表 5 である．

表 4 の「男性 20 代」「男性 40 代」「男性 60 代」の各列の上位と判定された語句に対し

て，その検定値[7]の大きさを用いてワードクラウドを作ると図 8，図 10（の上 2 つ）となる．一方ここで，単純に各性年齢区分層の語句の利用頻度を用いて描くと図 7 となる．

同じように，表 10 で「女性 20 代」「女性 40 代」「女性 60 代」の列にある上位の語句に対する検定統計量を用いて描いたワードクラウドが図 9，図 10（の下 2 つ）である．

ちなみにここで，男女の回答者数が，もっとも多い年齢層「30 代」について，男女でどのような違いがあるか，上位にある語句を比べてみた（図 15，図 16）．「女性 30 代」については下位にある語句もワードクラウドで示した．これらと，表 4，表 5 を併せて観察することで，「プラスと思うこと」への回答傾向，とくに人口統計学的変数（性別，年齢別）の傾向をより客観的に知ることができる．

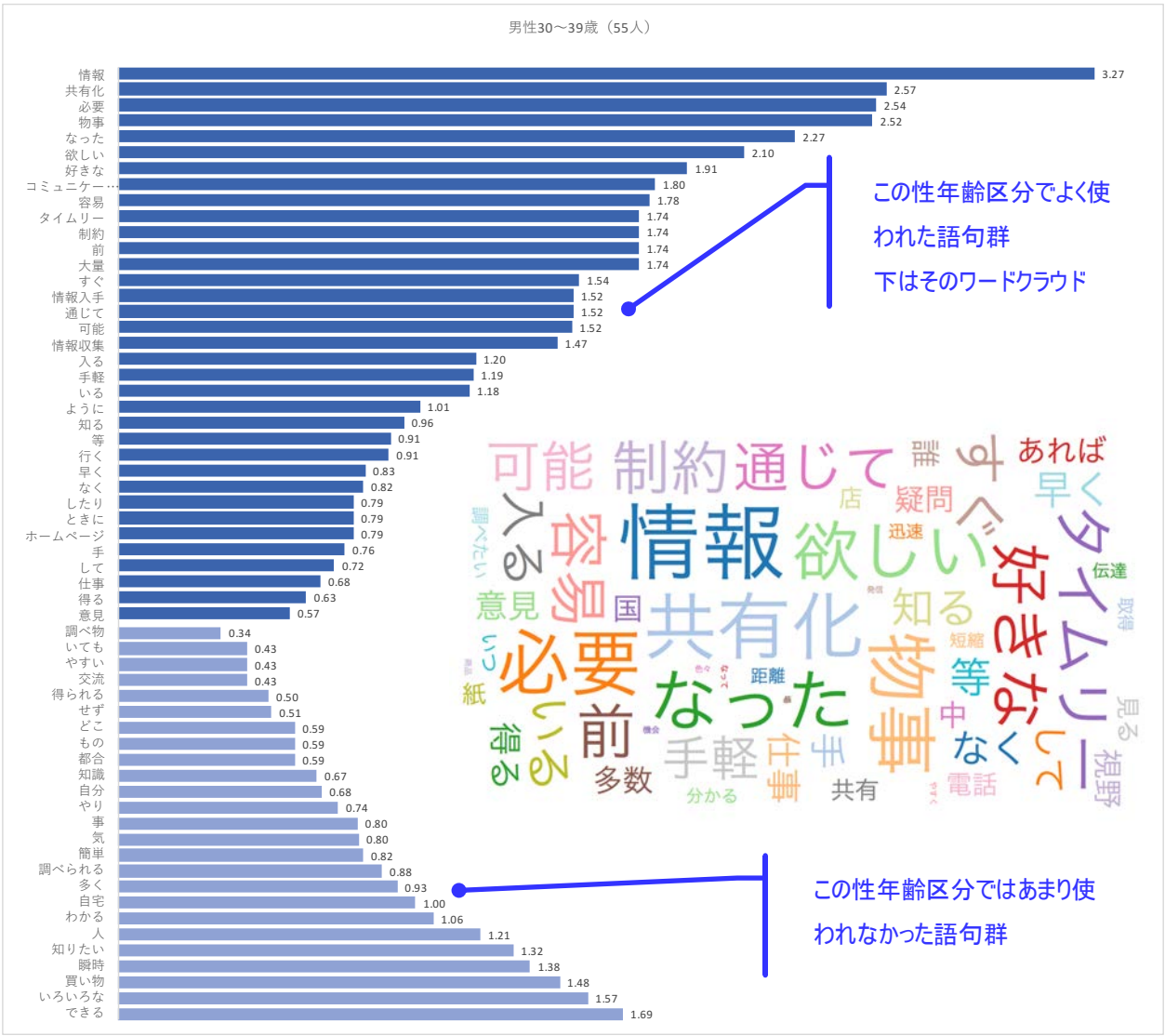


図 15　男性 30 代の場合

（上位，下位の主な語句と上位語句のワードクラウド）

[7] 正確には，語句ごとに得られるある検定統計量の実現値のこと．これの一覧表，たとえば表 3 の各列の語句にたして，客観的な指標が付与される（その一覧表が得られる）．この大きさを用いてワードクラウドを描いている．

図 16　女性 30 代の場合

（上位，下位の主な語句とそれらのワードクラウド）

表 4　性年齢区分別にみた自由回答における特徴的な語句（男性：上位 20 位，下位 10 位まで）

性年齢区分	男性 12～19 歳	男性 20～29 歳	男性 30～39 歳	男性 40～49 歳	男性 50～59 歳	男性 60～69 歳
語句数	36	111	112	115	94	85
上位 1	できない	様々	情報	情報収集	知識	選択
上位 2	情報伝達	迅速	共有化	必要	なる	世界
上位 3	交換	商品	必要	物	世の中	早く
上位 4	しなくても	知識	物事	なくなる	閲覧	情報
上位 5	外出	する	なった	時間的	共有	誰
上位 6	距離	判断	欲しい	情報	図れる	容易
上位 7	無く	使用	好きな	迅速	入手	入手
上位 8	機会	上	コミュニケーション	広がる	伝達	伝達
上位 9	遠く	して	容易	程度	情報交換	得られる
上位 10	入手	取得	タイムリー			
上位 11	した	調べ	制約			
上位 12	やすい	もの	前			
上位 13	意見	情報	大量			
上位 14	交流	発信	すぐ			
上位 15	すぐに	ある	情報入手			
上位 16	人	手	通じて	多く	簡単	思う
上位 17	なった	短時間	可能	得られる	容易	可能
上位 18	できる	関係	情報収集	世界	情報収集	情報収集
上位 19	早く	入る	入る	自宅	でき	でき
上位 20	多く	インターネット	手軽	仕事	相手	あらゆる
＜一部，省略＞						
下位 10	容易	事	調べられる	事	いながら	時
下位 9	様々	わかる	多く	わかる	いつ	出来る
下位 8	欲しい	出来る	自宅	いる	買い物	調べる
下位 7	利用	知る	わかる	簡単	入る	買い物
下位 6	離れた	情報収集	人	連絡	すぐに	入る
下位 5	旅行	連絡	知りたい	なく	人	人
下位 4	良い	メール	瞬時	調べる	家	事
下位 3	例えば	調べる	買い物	出来る	便利	メール
下位 2	知りたい	時間	いろいろな	調べられる	する	手
下位 1	出来る	便利	できる	人	手	すぐに

- 主要語句が，性年齢区分によって“利用の程度”に違いがあることがわかる．
- たとえば，「情報」や「情報収集」の使われ方をみよう．
- 「調べる」「調べられる」はどうか，男性はあまり発語していない（下位にある）．
- 対応する性年齢区分のワードクラウドと比べてみよう．

(*)「情報」と「情報収集」「情報入手」「情報伝達」…とさまざまある．さてどう使われたか．

(**)「人」「交流」「友人」「便利」「出来る」「買い物」「調べる」「調べられる」…などは女性で特徴的．

(***) 個々の語句が，各性年齢層でどう表れるか，ある傾向がみえるが，これをさらにどう探査するか．

表 5　性年齢区分別にみた自由回答における特徴的な語句（女性：上位 20 位，下位 10 位まで）

性年齢区分	女性 12～19 歳	女性 20～29 歳	女性 30～39 歳	女性 40～49 歳	女性 50～59 歳	女性 60～69 歳
語句数	61	160	192	145	115	59
上位 1	交流	人	調べられる	瞬時	なって	本
上位 2	色々	参加	病気	ほしい	です	場合
上位 3	気軽	相談	時間	時間	居	予約
上位 4	出来る	思った	趣味	関係	便利	出来る
上位 5	地域	友人	好きな	なく	やりとり	して
上位 6	人	機会	家	入れる	楽	関係
上位 7	調べ物	興味	どんな	タイムリー	事	知識
上位 8	便利	いても	知りたい	何	国	居
上位 9	できない	やすい	友達	価格	新しい	等
上位 10	なんでも	自分	たくさん	活用	旅行	早く
上位 11	ような	ネット	取れる	資料	見る	得られる
上位 12	ショッピング	行かなくても	せず	手間	少なく	あらゆる
上位 13	出来る点	検索	時	手	瞬時	には
上位 14	詳しく	すれば	すぐに	したり	メール	購入
上位 15	図書館	オークション	気軽	利用	わかる	社会
上位 16	速く	ニュース	思う	交換	とても	図書館
上位 17	人たち	遠くに	いろいろな	手続き	無く	世の中
上位 18	離れた	疑問	何でも	色ん	せず	中
上位 19	すぐ	取り	幅広い	通じて	得られる	インターネット
上位 20	しなくても	ときに	気	買い物	いる	いた
＜一部，省略＞						
下位 10	入る	して	瞬時	好きな	多く	自分
下位 9	情報	時	地域	情報収集	できる	入る
下位 8	事	瞬時	良い	必要	人	いろいろな
下位 7	ある	なった	なる	コミュニケーション	思う	時間
下位 6	できる	早く	得られる	気	容易	なる
下位 5	して	入手	関係	事	簡単	調べられる
下位 4	得られる	いながら	世界	検索	情報収集	する
下位 3	する	得る	多く	家	手	できる
下位 2	入手	時間	情報	容易	情報	すぐに
下位 1	時間	情報	早く	すぐ	自分	知りたい

【分析例 3】自由回答質問の分析（2） ―クラスター化による“仮説発見”の効用―

1. 分析の目標

【分析例 2】で用いたデータセットについて，“別の視点から”分析を試みる．同じデータセットであっても，見方を変えると別の情報が浮かびあがる，という例である．換言すると，データの解析とは，いくつもの**“筋道（シナリオ）”**を想定し，帰納的，探査的，仮説発見的に進めることが肝要ということである．

【分析例 2】では，人口統計学的変数とくに性別，年齢区分，性年齢区分などの既知の定性情報が自由回答の意見の中に登場する語句群（構成要素群）とどう関連するかを調べた．ここでは見方を変えて，回答者の自由回答パターン，語句の使い方の類似性を**“対応分析とクラスター化法”**を同時的に用いて探査する．ここのポイントは，両手法の特性をうまく利用してセットで用いることである．

分析のポイント＜1＞

① 具体的には回答者の語句の使い方の類似性・非類似性を調べる．

② 回答者ごとの語句（構成要素）の使い方を“2 元データ表”の要約する．つまり「(回答者)×（語句群)」の要約表をデータ表とする．

③ 対応分析とクラスター化により，回答者と語句群を，同時的に“自動分類”する．回答者のクラスターと語句のクラスターを作る．

④ 回答者の分類情報，つまり回答者クラスターをあらたな質的変数として，各クラスターの特徴付けを行う．

⑤ 回答者の語句の使い方の類似性・非類似性を調べる．具体的には語句や自由回答文に比較評価のための**“数値を付与”（数量化）**し，客観的に分析する．

⑥ 対応分析の数理的特性として，**“はずれ値”**が生じやすいことがある．とくに，疎な大規模データ行列では，頻度の少ない周辺和（行和，列和）があると，これが起こりやすい．この外れ値の手当を工夫する必要がある．

対応分析法を適用する際の 2 元データ表として，「(回答者）×（語句群)」の形式に作り替え，このデータ表の対応分析を行う．同時に，データ表の行（回答者）と列（語句群）のそれぞれについて（同時的に）クラスター化を行う．このことは，データ表の作り方の工夫で，あとはいままでの分析手順とほぼ同じ操作を行うことを意味する．ここの目標は，このデータ表の“回答者と，回答者が用いた語句との関連性”を探査することであり，さらにクラスター化によりあらたな仮説を引き出すこと，つまり**“仮説発見”**にある．

まず，分析対象となる“2 元データ表”の形を示す（表 1)．ここで行側は，前に【分析例 2】の表 1 の行と同じく，分析に用いた語句群である，列側は回答者に対応する．ここでは，行と列とを列和・行和の大きさで並べ替えてある（観察しやすくしただけ)．たとえば，もっとも利用頻度の多い語句「情報」は 300 回登場し，これを自由回答内で用いた回答者に対応するセル内に，その回答者が用いた回数（頻度）が入る．たとえば，回答者[051]は 1 回，回答者［477]は 3 回，

回答者[458]は5回，…それぞれ「情報」を用いたことになる．また「0」は，それを使っていないことを示すが，これが多いこともわかる（非常に疎なデータ表）．以下，同じように用いられた語句と回答者の回答パターンが2元データ表の形で作られる．

【分析例2】と同じ辞書編集の結果を用いたから，用いた総語句数（総構成要素数）は，同じ「3,501（語）」となる．また，各行の語句（構成要素）の行和も同じになる．上に断ったように，ここで同じ語句数を用いる必要はない．ここでは，前の2元データ表との対応関係を分かりやすくするために，あえてそろえた．こうて得られた2元データ表は，寸法が「235（語句）×503（人）」というさほど大きくない行列である．

メモ:

一般に，こうして作る2元データ表には，次のような特徴がある（一部，前述の指摘と重複）．

- データ表の寸法（行・列の大きさ）が，数千から数万，さらに大きくなることも珍しくはない．
- データ表のセル内の度数は非常に少なく，度数ゼロのセルも無数にある，つまり非常に疎（スパース）なデータ表となる．
- データ表の寸法が，はじめからわかっていない．通常は，ある程度のデータ加工を経た後でないとこの寸法は定まらない．
- こうした性質のデータ表（非常に疎な大規模データ行列）に対応分析法を適用する際には，数値処理上のアルゴリズムの工夫が必要となること．
- クラスター化法も，扱うデータ数が非常に多いので，一般的な自動分類法（階層的分類，非階層的分類）を，そのまま利用することがむずかしい．データ圧縮化の方法や階層分類・非階層分類を併用する**ハイブリッド方式**など，工夫が必要である．

2. 結果と観察:回答者の類型化と回答語句の傾向探査

ここからの説明では，対応分析の結果は表だってはみえない．これまでの例のように，布置図で成分スコアの傾向を視覚化的に観察する，といった操作はあまり登場しない．対応分析を質的情報の視覚化ツールと考えがちだが（実際そういう指摘や適用例が多いが），それは対応分析の特性のごく一部にすぎない．むしろこれから述べる対応分析の性質を巧みに用いる**“クラスター化法”との併用**や類型化といったことに利用の効用がある．細かい議論は横に置き，表1のデータ表に対応分析とクラスター化法を適用して得られる情報を観察する．ここでは，以下の要領で分析した[1]．

- 「語句群（構成要素）×（回答者）」の2元データ表（表1）に，対応分析を適用し，構成要素，回答者それぞれの成分スコア，その他の指標を求める．
- 必要に応じて，布置図，同時布置図なども併用するが，ここでは，これが主たる目的ではない．
- 回答者の分類と類型化を行う．ここでは「15群」に分ける．

[1] 分析の手順は【分析例2】にほとんど同じである．要点は，用いる2元データ表の作り方の違いにある．それと，成分スコアをクラスター化で用いる際の数理的特性に注意することである．

- クラスター情報を調べる．クラスターに含まれる回答者がよく用いた語句（上位語句），とあまり使わなかった語句（下位語句）を，調べる．
- 具体的には語句に数値情報（検定値）を付与し（数量化し），利用語句の特徴を探査する．
- クラスター内の回答者属性（例：性年齢区分）を調べる．
- クラスター内の回答者の自由回答文も数値化した上で（検定値を付与し），各クラスターの自由回答文の傾向を調べる．
- “語句群”（構成要素変数）についても，語句のクラスター化，語句の数値評価などを行う．

以上の操作は，すべてソフトウェア（WordMiner，JMP，JMP スクリプト）でほとんど自動的に処理されるが，出力情報の選択やアレンジを要領よく行う手順を知ることが肝要である．この他，さまざまな分析手順を用いることができる．たとえば，布置図，同時布置図の観察，成分スコアの有意性テスト（種々の寄与度の利用），デンドログラムによる分類結果の観察とあるが，ここでは上述のことに絞り説明する．

対応分析とクラスター化法を用いる際の注意事項

対応分析で得た成分スコアにクラスター化法を適用するとき，いくつかの注意が必要である．成分スコアにクラスター化を適用するという操作は日常的に用いられているが，**誤った使い方**が多いので，留意事項をここに記しておく．機械的に（勝手に）対応分析とクラスター化法をつなげて“利用してはいけない”ということを強調しておこう．

- 対応分析で得た成分スコア（たとえば，回答者や語句に付与された成分スコア）に対して，そのままクラスター化法[2]を適用してはいけない．対応分析の数理的性質を考慮し，重み付きの平方ユークリッド距離を用いる．基本は**“（重み付きの）ウォード法”**となる．
- クラスター化で用いる成分スコアの成分数（次元数）に依存して，クラスター化の結果が変わる．用いる次元数と固有値（あるいは特異値），寄与率などの間にはある関係があり，用いる成分数で情報量が変わる（分類結果が変わる）．
- 同時に，**“クラスター数”**をいくつとするかがある（難問）．クラスター数の決め方にはこれといった決定的な方法がない．幸いなことに，対応分析／クラスター化法の数理的特性を用いた**“発見的な判断基準”**が適用できる．

ここまで述べたことは，“数理的な説明が必要”であるが，これがセミナーで扱う課題となるのでここでは詳細は省略する．

[2] ここでいうクラスター化法とは，一般に利用されている自動分類法（階層的分類，非階層的分類）を言う．こうしたクラスター化法をいきなり成分スコアに対して適用はできない，ということ．

表 1 （構成要素）×（回答者）のデータ表［寸法（235 語句×503 人）から一部を切り出し］

語句総数 (3,501)	回答者 番号	[477]	[506]	[051]	[458]	[394]	[379]	[453]	[249]	[131]	[470]	[184]	[005]	[323]
利用 語句（↓）	列和（→） 行和（↓）	31	29	27	27	26	25	25	24	22	22	20	19	19
情報	300	3	0	1	4	0	1	1	1	0	2	2	1	0
できる	232	3	1	6	2	2	2	0	0	1	3	0	1	0
出来る	76	0	3	0	0	3	0	1	3	0	0	0	1	2
事	74	0	0	0	0	3	0	0	1	0	0	0	1	0
時間	72	0	0	0	0	0	1	0	1	0	0	0	1	0
簡単	61	1	0	1	0	2	0	0	0	0	0	0	1	0
知りたい	60	0	0	0	2	1	0	0	0	0	0	0	0	0
なった	59	0	2	0	0	0	0	0	0	0	0	1	0	1
すぐに	57	0	0	0	0	2	1	0	0	0	0	0	0	0
人	57	0	0	1	0	0	0	0	0	0	1	1	0	1
メール	47	0	1	0	0	0	1	0	0	1	0	0	1	0
手	47	0	0	0	0	0	0	3	0	0	0	0	0	0
入手	44	1	0	0	0	0	0	0	0	0	0	0	1	0
する	42	3	0	0	2	0	0	0	0	0	1	1	0	0
得られる	40	0	0	0	0	0	0	0	0	0	1	1	0	0
して	38	1	1	0	0	0	0	0	0	0	1	1	0	0
ある	35	1	0	0	1	0	0	1	0	0	0	0	0	1
調べられる	34	0	0	0	0	0	0	0	1	0	0	0	0	0
便利	34	0	1	0	0	0	0	0	0	0	0	0	0	0
なる	33	0	0	0	1	0	0	0	0	0	0	0	0	0
ように	33	0	0	0	1	0	0	0	3	0	0	0	0	0
家	33	0	0	0	0	0	0	0	0	0	1	0	0	0
いろいろな	31	1	0	0	0	0	1	0	0	0	0	0	0	0
自分	30	0	0	2	0	0	0	1	0	0	1	0	0	0
得る	30	0	0	0	1	0	1	0	0	0	0	1	0	1
入る	30	0	0	0	0	0	0	0	0	0	0	0	0	0
調べる	29	0	0	0	0	1	1	0	0	0	0	0	0	0
買い物	29	1	1	0	0	0	0	0	0	0	0	0	1	0
なく	27	0	0	0	0	0	0	1	0	0	1	0	0	0
時	27	1	0	0	0	0	0	0	0	0	1	0	0	0

一般に，2 元データ表はこのような非常に疎で，寸法の大きな行列となる．

分類結果の要約

“回答者”を 15 群に分類し，得られたクラスターの大きさ（“クラスター・サイズ”という）を要約し図 1 を得た．図をみると，クラスター・サイズには大小があり，比較的サイズの大きいクラスターから小さいものまでさまざまである．成分スコアの分布に限らず，たいていは教科書にあるような塊状のクラスター（房状）は存在しない（例：うしろの挙げた図 7 を参照）．

クラスタリングによる分類操作の目標は，クラスターを探すのではなく，“クラスターを生成すること”つまり**“クラスター化”**することにあると考えるほうがよい．また，多くの場合**“はずれ値”**の影響を受けやすい（図 1 のサイズの小さいクラスターは外れ値の傾向あり）．これについては，後ろで例を示す．

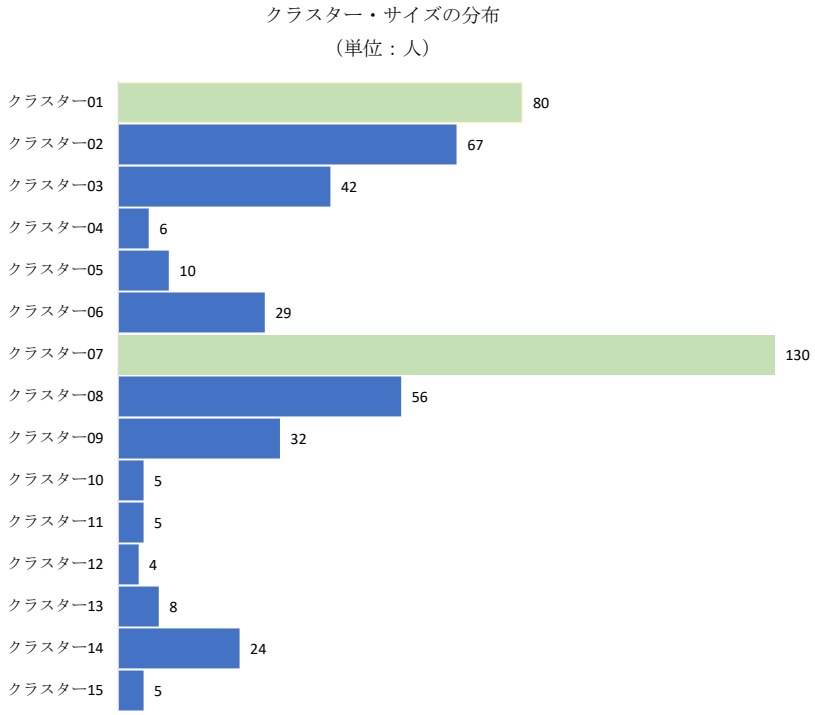


図 1　回答者（503 人）の分類でえたクラスター・サイズの分布
（*）サイズの大きい 2 つのクラスターの色を変えてある．

3.　自動分類の効用を知る　―いくつかのクラスターの観察と傾向探査

得られた回答者のクラスターのうち，サイズの大きい 2 つのクラスターと，サイズの小さい 1 つのクラスターを選び，これらのクラスターの回答者と回答傾向を調べよう．もちろん，通常はすべてのクラスターについて同様の吟味を行い，このデータセットについて総合的な評価分析を行うのだが，ここは紙幅の都合で一部を取り上げる．

クラスター01 の観察：

クラスター・サイズが 2 番目に大きい「クラスター01」（80 人）から得た基本情報を図 2 に要約した．人口統計学的変数として属性「性年齢区分」を調べると，男性も混じっているが（約 23%），女性が多い（約 77%）．とくに「女性の 30 代」が多い．

つぎに，このクラスター内の回答者がよく用いた語句（上位語句）と，あまり使わなかった語句（下位語句）に，有意な程度を数値として付与し，これの一部を選んで棒グラフとした．上から「調べられる」「すぐに」「事」「わかる」「調べたい」「調べる」「知りたい」といった語句が上位に登場する．一方，下位には「情報」「入手」「人」「入る」「時間」…「情報収集」「欲しい」…と続く．

これだけでもおよその回答傾向がみえるのだが，さらにこのクラスターに分類された回答者の自由回答文を有意な順に上から一部を選んでみよう（もちろん，ここは自動的に分類される）．図 2 にある＜自由回答の例＞がそれであり，ここで検定値とは個々の自由回答文に扶養された有意の程度を示す値である（つまり，ある規則で，自由回答文のランキングを行ったことになる）．80 人から選ばれた自由回答文と（棒グラフにある）上位の語句の関連は，説明するまでもなく，非常にわかりやすい．このように，対応分析とクラスター化を用いて，**“客観的な結果”**がほぼ“自動的に提供”される（恣意的な解釈とならない）ということである．これは定性情報の分析では重要なことである．

さらに，棒グラフに一部を示した語句の検定値情報を用いて作ったワードクラウドが図 3 である．用いる指標の工夫で，単純な棒グラフだけでなくワードクラウドが補助情報として非常に有効に機能することもわかる．

クラスター07 の観察：

クラスター・サイズがもっとも大きい「クラスター07」を調べよう．ここでも得られた基本情報を 1 つの図に要約した（図 4）．このクラスターは，あきらかに「クラスター01」とは異なる．ここでは男女がほぼ折半した構成になっている（男性が約 55%，女性が約 45%）．年齢区分は，男性が各年齢層にばらついているのに対して，女性は 30 代，40 代あたりが多い．

では，このクラスターの回答者はどのような語句を用い（あるいは用いず），またどのような自由回答を記したのであろうか．上位の利用語句は「得られる」「情報」「いながら」「して」「入手」「短時間」「得る」…と続く．一方，下位の語句は「調べられる」「なった」「すぐに」「入る」「手」「情報収集」…と続く．これらの下位の語句の中には，上でみたクラスター01 内の“上位の語句”があることに注意しよう．

クラスター01, クラスター02 のいずれのクラスター内の回答者も,「プラスになると思うこと」として.「必要情報の入手」という点で共通するものの“自由回答文の語句の用い方，表現”が異なることもわかる（それをクラスター化で自動的に分類された，ということ）．たとえば，主な回答文を挙げてみると，以下のようになる．

＜クラスター01 から引用＞
すぐに調べられる／知りたいことがすぐにわかる／知りたい情報がすぐに調べられる，…

＜クラスター07 から引用＞
いろいろな情報が得られる／多くの情報が得られる／様々な情報が得られる，…

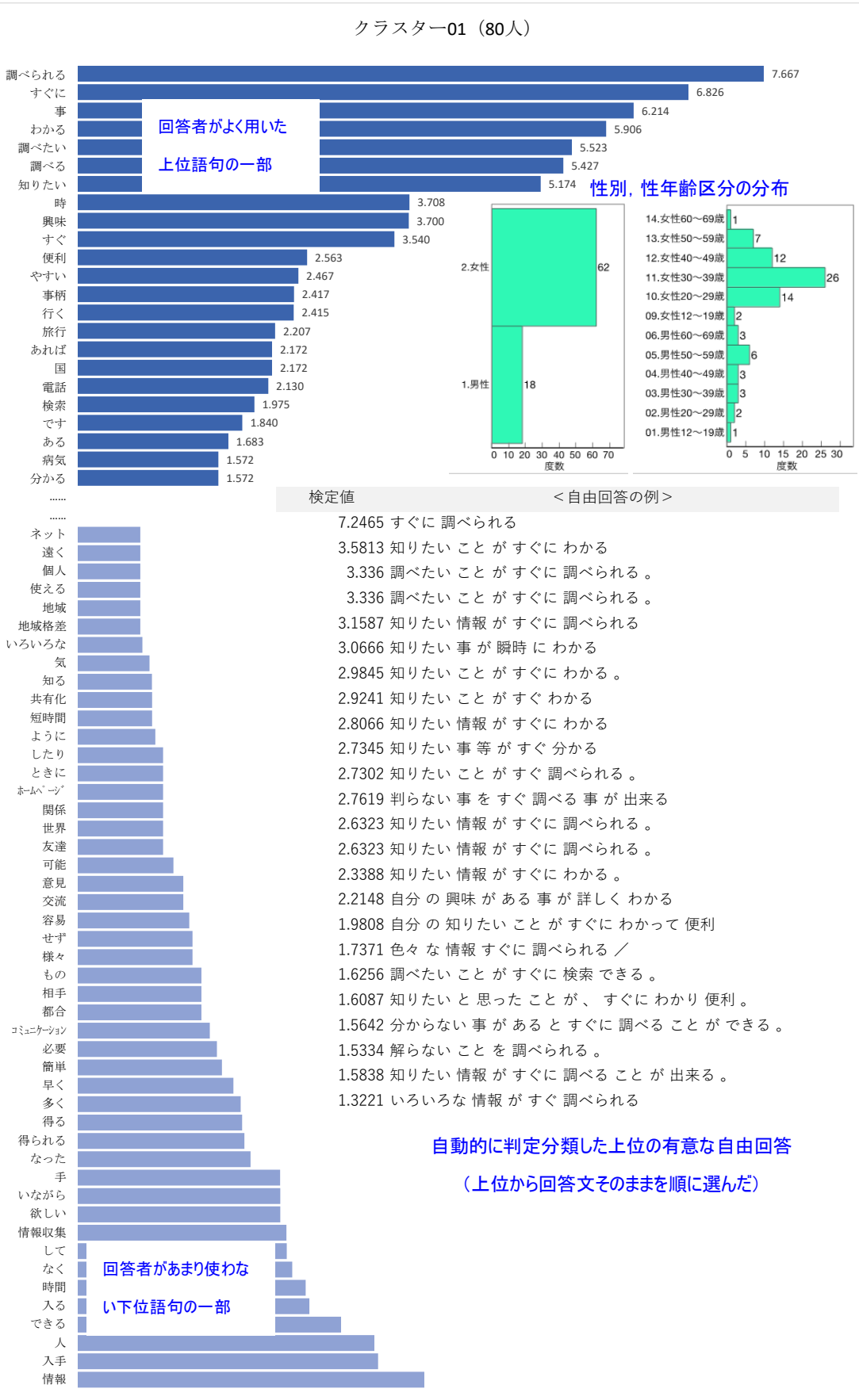


図 2　クラスター01 の基本情報

図 3　クラスター01 の上位語句から得たワードクラウド

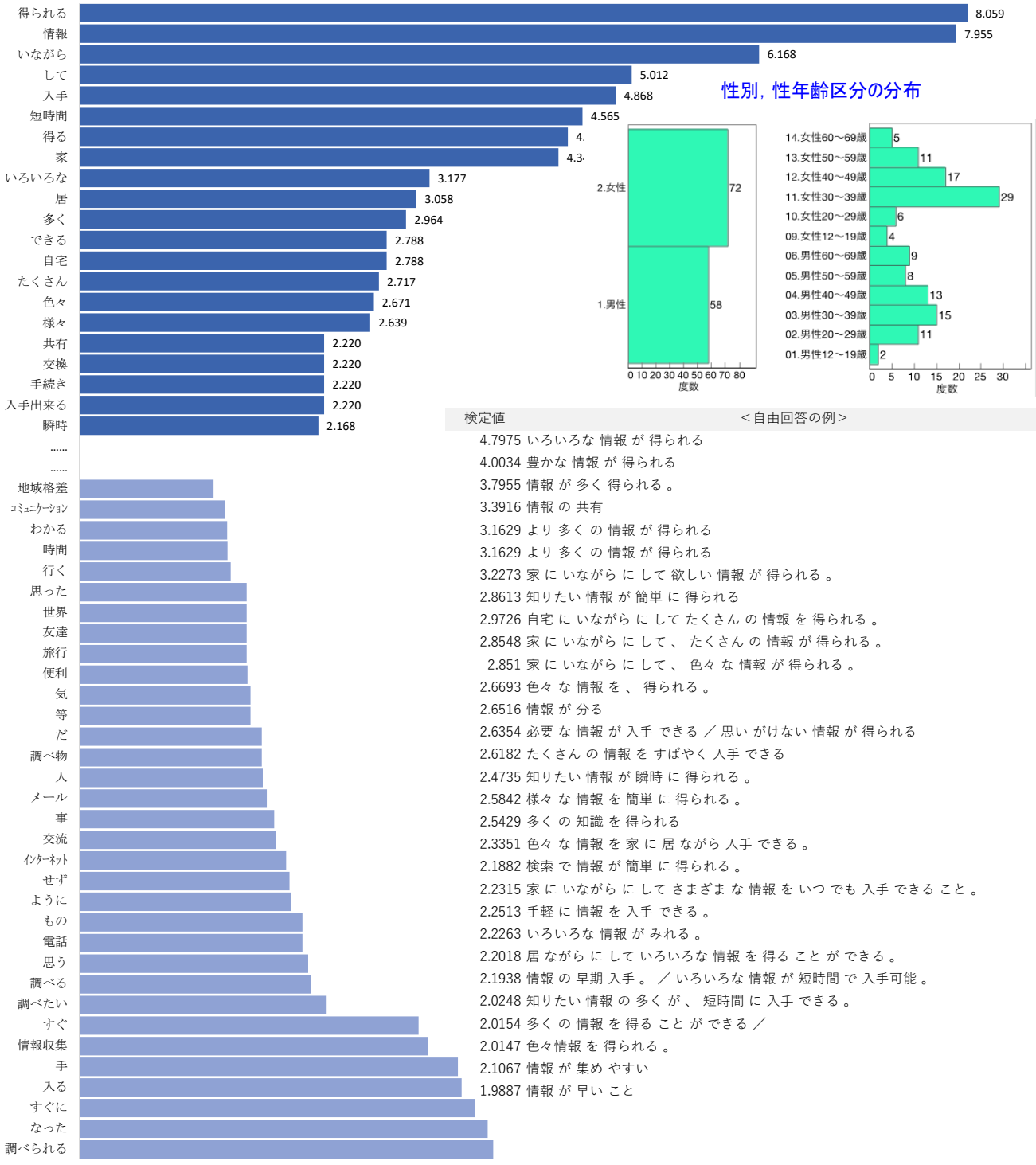
かつての自由回答文の処理方法として，コーダーが（作為が入らぬよう指示書，簡易辞書に従って）自由回答文から語句を選ぶ，あるいは類似表現を別の要約語句としてまとめるなどを行ってきた．たとえば，上のクラスター01 にあるような語句群を 1 つに括って「迅速な情報取得」とする，あるいは同じようにクラスター07 から「取得情報の多様性」とする，などである．こうした操作を行うことは，コーダーに起因する偏り（バイアス）を生じることも知られている．

さらに，自由回答文が長くなるほど，とうぜん，記述内容は次第に曖昧になり発散する．実際，クラスター01 の自由回答のうち，検定値の値は小さく，有意からは遠い例をいくつか挙げてみよう（表 2）．

表 2　クラスター01 から：自由回答の内容が長く，曖昧になった例（記入回答のままを引用）

検定値	自由回答の例（検定値の小さい方から 10 名を選んだ）
0.250	メールは手紙を書いて投函しに行くよりとても便利です。／自宅で沢山の情報を手に入れることができて便利です。／旅行の時の宿の予約でいるのもありがたい。
0.232	時間に拘束されずに、自分のペースでできる。／わからないことを、すぐに調べることができる。
0.226	幅の広い情報が家に居てもわかる
0.222	どこでも世界中の情報を瞬時に集められる。今知りたいことをインターネットひとつで調べることができる。
0.215	・交友などは深まり仲間を作りやすい。／・自己アピールの機会が増えた。／・普通に私生活で使うわからないことがわかるようになる／電車やバスの利用時間など。
0.199	自分の知りたい情報が即時に取り込むことが出来る。／
0.198	気軽に世の中の事柄が知ることが出来る。／生活の時間短縮になる。
0.155	／知識が豊富になる。／分からないことを調べることができる。
0.023	いろいろな情報が簡単に調べられる。気兼ねなくメールのやりとりができ、各県・各国とやり取りできることはプラス。仕事の幅が広がり、独立可能性ができる。
−0.090	知りたい情報が電話やその場所に行かなくても得ることができる。／

クラスター07（130人）



検定値　＜自由回答の例＞

4.7975 いろいろな 情報 が 得られる
4.0034 豊かな 情報 が 得られる
3.7955 情報 が 多く 得られる 。
3.3916 情報 の 共有
3.1629 より 多く の 情報 が 得られる
3.1629 より 多く の 情報 が 得られる
3.2273 家 に いながら に して 欲しい 情報 が 得られる 。
2.8613 知りたい 情報 が 簡単 に 得られる
2.9726 自宅 に いながら に して たくさん の 情報 を 得られる 。
2.8548 家 に いながら に して 、 たくさん の 情報 が 得られる 。
2.851 家 に いながら に して 、 色々 な 情報 が 得られる 。
2.6693 色々 な 情報 を 、 得られる 。
2.6516 情報 が 分る
2.6354 必要 な 情報 が 入手 できる ／ 思い がけない 情報 が 得られる
2.6182 たくさん の 情報 を すばやく 入手 できる
2.4735 知りたい 情報 が 瞬時 に 得られる 。
2.5842 様々 な 情報 を 簡単 に 得られる 。
2.5429 多く の 知識 を 得られる
2.3351 色々 な 情報 を 家 に 居 ながら 入手 できる 。
2.1882 検索 で 情報 が 簡単 に 得られる 。
2.2315 家 に いながら に して さまざま な 情報 を いつ でも 入手 できる こと 。
2.2513 手軽 に 情報 を 入手 できる 。
2.2263 いろいろな 情報 が みれる 。
2.2018 居 ながら に して いろいろな 情報 を 得る こと が できる 。
2.1938 情報 の 早期 入手 。 ／ いろいろな 情報 が 短時間 で 入手可能 。
2.0248 知りたい 情報 の 多く が 、 短時間 に 入手 できる 。
2.0154 多く の 情報 を 得る こと が できる ／
2.0147 色々情報 を 得られる 。
2.1067 情報 が 集め やすい
1.9887 情報 が 早い こと

図 4　クラスター07 の基本情報

図 5　クラスター07 上位語句のワードクラウド

これらの例にみるように，図 2 に示した自由回答文の例に比べ，記述内容がかなり長くなる，あるいは，用いる語句の種類が多くなっている．つまり，それだけ書き方が多様になった，あるいは表現が豊富になった（曖昧になる），ということである．言い方をかえれば，こういう自由回答を行った回答者たちは，与えられた質問文（という刺激）からこれだけの文章を想起したということになる．こうしたことから，自由回答質問，とくにナラティブ型の自由に書き込む質問を用いることに懐疑的である人たちのもいる．一方，ここにみたように，ほとんどの処理をソフトウェアに任せることで，ほぼ**“自動的”に**客観的な結果がえられるという利点もある．

クラスター15 の観察：

対応分析の特徴の 1 つに，初期の 2 元データ表の（行和や列和の）出現頻度の少ない場合に，成分スコアの値が大きく外れることがある．いわゆる“はずれ値”が生じることがある．これに対する手当の方法もいろいろ提案されている．要は，（よく内容をみないで）外れ値を機械的に除外することは避けるべき，ということである．

上の 2 つのクラスターで行った手順と同じ方法で，このクラスター15 の基本情報を要約した．これが図 6 である．

クラスター15 のクラスター・サイズは 5，つまり回答者数は 5 人（男性）だけである．用いられた上位語句と 5 人の回答者の自由回答，そしてされらに付与した検定値を比べてみれば，内容はあきらかである．

実はこのグループに属する 5 人だけが語句「情報入手」を用いている．対応分析を行って，は

じめの 2 成分について描いた布置図を図 6 に付けた．語句「情報入手」は成分 1（1 軸）の方向に大きくずれて“はずれ値”となって現れる（1 軸の変動を大きく支配している）．

ちなみにここで，成分スコアの布置図を次元 2，次元 3 で描画すると図 7 となる．「情報入手」は左下に位置し，他の語句が散布している．中心から遠くにある（この 2 つの次元の平均的な位置から離れた）語句のラベルを引出線で付けた．これらの周辺にある語句が，この次元における特徴的な語句であり，これらの語句が，どのクラスターで有意に意味があるか，また自由回答文とどう関係するかを，ここで述べてきた要領で調べればよい（探査を深める）．

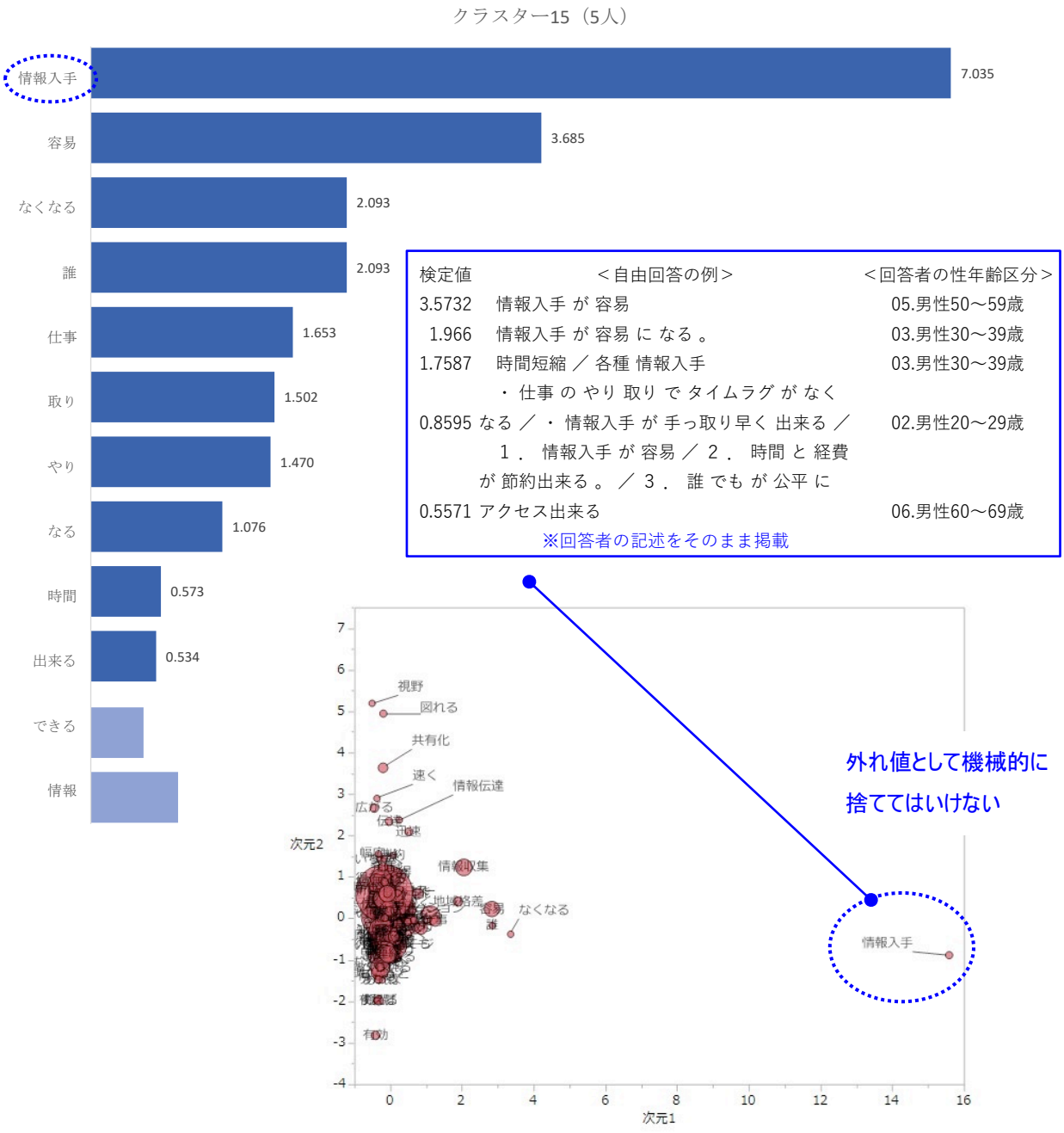


図 6　クラスター15 の基本情報

(*)外れ値と布置図，利用語句，自由回答文の関係を示した例

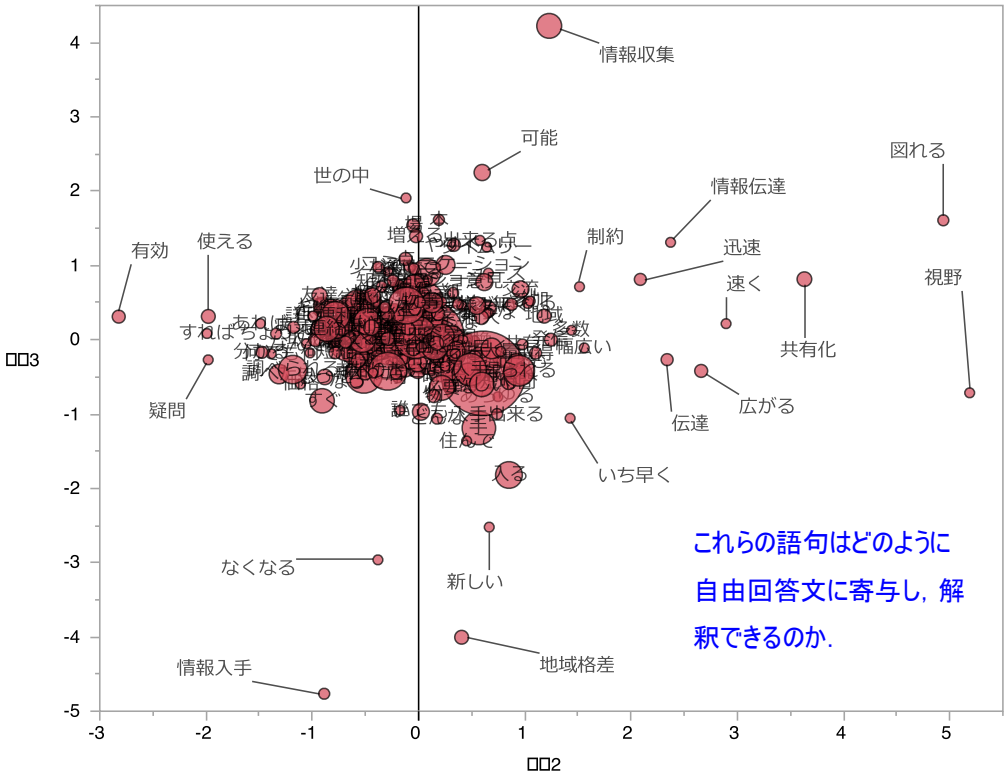


図 7　対応分析で得られた布置図（第 2 次元，第 3 次元の布置）

クラスター化の結果を，性年齢区分という人口統計学的変数（であり質的変数）を取り上げ，これと自由回答文との関係を調べているが，それで十分ということではない．ここで，この自由回答質問と関係のありそうな別の選択肢型質問項目や人口統計学的変数を選んで分析を行うと，また別の特徴や傾向が観察される可能性もある（分析の筋道は，いくつもある）．こういうことがいわゆる探査的であり，（仮説）発見的なアプローチなのである．

4.　いくつかの課題　―同義語や誤記について―

ここで，【分析例 2】と【分析例 3】のはじめに，分かち書き処理の結果はなるべく編集加工を行わずに，簡単な編集，たとえば句読点や記号類の削除に留めたとした，ことを思い出そう．つまり，得られた自由回答文には，さまざまな同義語や類語，そして誤記やゴミ情報などが混在しているはずである．

また，「判らない，分からない，わからない」「事，こと」「出来る，できる」「「等，など」「見られる，みられる」「いろいろな，色々な，いろんな」「調べる，しらべる」などの同義語がすぐに見つかる．これらは同じ意味であり，同じ使い方をしている．しかしこのままでは，分析上は異なる語句として扱われる（注：同義語を調べ，判断し括るなどの処理はむずかしいだろう．最近の人工知能的ソフトでは部分的に解決可能かもしれないが，問題の本質は別のことにある）．

場合によっては「情報が入手できる」，「情報が手に入る」「情報入手」「情報取得」…は同義としたい，と考える分析者もあるだろう．こうなると，単に同義語，類語を括るだけで不十分で，いわゆる**“語句の辞書編集”**が必要となる．これも細かいことを言い出すと，際限のない作業と

なり，分析時間の浪費となる．語句を整える程度と，扱うテキスト型データの特性，分析者の要求度との兼ね合い（トレードオフ）の問題となる．

たとえば，ある作家の書いた文章についての分析を行うことと，自由回答質問で取得のテキスト型データの分析では，かなり様相が異なる．前者は，場合によっては，その作者の著作物を丁寧に調べ詳しい“コーパスを作る”など行えば，かなり詳しい分析が可能である．しかし，自由回答質問では，ときどきの状況で質問の内容や聞き方も変わるので，整った（画一的な）コーパスやシソーラスの利用が適わないことが多い．ただ，パネル調査として，経時的に同じ内容の自由回答質問を繰り返すなどの場合には，簡易コーパスの作成，あるいは語句の**“共通辞書の作成”**は，手間はかかるが検討する意味がある（実際にそういう例もある）．

われわれが扱う非定型・非構造型のテキスト型データでは，こうした事象は日常的に頻出することである．これらをどう扱うか，は分析者の要求する分析の緻密性・厳密さによる．ゴミやノイズが多少はあっても，まず“データのおおまかな傾向や特徴”を知ることを優先するか，反対に，語句をきれいに整え，語句の使い方や記述内容の関係，文脈の流れなどまでを“詳しく分析”したいのか，その要求度に応じて，分析のレベルを変えることが求められる．

社会調査であれば自由回答質問などへの回答記入を依頼する際に，回答者に十分に注意することを促すことも必要である．これは，自由回答質問の設計そのものの問題でもある．回答者にとって分かりにくく，回答しにくいような質問文は避けねばならない．

ここで示した例では，インターネットの長所・短所といったことを，漠然と尋ねることはせずに，「インターネットのプラスになると思われること」「インターネットのマイナスになると思われること」と分けて尋ねている．状況によっては，これでもなお，回答者が自由回答を記述する際に，迷いや判断に戸惑うことがあるかもしれない．実際，分析でそういう傾向にある自由回答も観察された．また，年齢によっては，“インターネットの理解の程度”が異なるかもしれない．

要は“調査者（実施側）あるいは分析者が，一体何を知りたいのか”調査の構成概念を明らかにし，それを反映させた質問文の設計が必要だ，ということである．こうした課題は，調査誤差とくに**“測定誤差の問題”**としてデータの品質に影響を及ぼす（注：調査誤差の低減と調査データ・内容の品質向上はトレードオフである）．

別の検討要素として，分析時に用いる“二次情報あるいは**外部情報源**”（external information）の選択がある．自由回答というテキスト型データに潜在する特徴を，他の定型情報（構造型情報）と対比分析することで探り出す効用と言い換えてもよい．

たとえば．【分析例 2】では，語句群と別に得た（集めておいた）人口統計学的変数である「性年齢区分」という情報を用いて分析した．また，【分析例 3】では，まず回答者の回答傾向の類似性に従って，クラスター化により分類を試みた（古典的な言い方であると，教師なし分類で類型を生成）．そのあとに，得られたクラスター（つまりある程度の等質性を確保したグループ）について，別の要因，たとえば【分析例 3】では，性年齢区分がどのように関与するかを調べた．さらに，場合によっては，当該調査では収集していないが，他にある情報源から得た情報を対比分析することもあるだろう．

上の例だけでも，単純に対応分析法の計算処理を行うだけでは，探索的なマイニングが可能とはならないことがわかるであろう．**分析のシナリオを整える**という細かいチューニングが必要で

あり，いかに**探査的に帰納的推論を行えるかを考えること**が鍵である．対応分析法の考え方，特性をこうした場面にいかに活かすか，を紹介することが本セミナーの狙いである．

5. テキスト・マイニングにおけるいくつかの課題 ―思いつくまま―

いわゆるテキスト・マイニングは，依然として無数の問題を抱えている．コンピュータや周辺処理機器の性能向上で，大量データが高速に処理できる時代といわれている．人工知能的なアプローチが有効であるという提案も多くみられる．しかし，基本操作である“分かち書き”処理一つを考えても，信頼できる結果をみせてくれるツールはあまりないだろう（それに単に分かち書きや形態素解析ができればよいでは済まない）．この他，経験的にみて解決すべき無数の課題がある．こういう事柄は理屈や数理で解決される問題ばかりではない．

- 上でも少し触れたが，辞書編集をどう扱うか．日本語特有のさまざまな問題がある．
- （欧米語にくらべて）同義語・類語（狭義のシソーラス）あるいはコーパスなどの利用環境が十分ではない．特定の分野での専門用語などは整備されつつあるが，問題は，整っていない**“普通の言葉，文章”**の解読である．［例 1 ：WordMiner では，看護･介護などの専門語，擬態語いわゆるオノマトペ語の辞書を利用できる／例 2：無数の電子化類語・専門語辞典類もあるがそのまま利用可能かは疑問］
- 通常，身近にみるような文章や発話情報には，“ゴミ”や“ノイズ”が含まれる．これをどう扱えばよいのだろう．“ゴミ”はかならず除去がよい，とはならない．
- ゴミやノイズの介入は，調査であれば**“質問文の設計”**に深く関与することがある．これは，**“データの集め方”**に関わることでもある．かつて，データ・マイニングの話題で必ず始めに登場した決まり文句に“GIGO（Garbage in, Garbage out）“があった（最近はすっかり聞かなくなった）．
- テキスト型データの分析では，たいていは**出現頻度の高い語句**だけを拾い出すこと，あるいは注目するのが一般的である．しかし，その抽出や取捨選択はどう行うのか，閾値をどう決めのるか，分析への影響はどう表れるのか．また，出現頻度の低い語句はどう扱えるのか，除外してもよいのか．一般には，かなり不透明である．
- これらは，分析で扱う語句数の大きさが一意に決まらないことにつながる．その決まらないことで，分析結果は，多くの場合 **“一意的な解とはならない”** のが普通である（答えはまちまち）．多次元データ解析などを用いるときに注意せねばならない事柄である．
- 具体的な分析上の課題として，“寸法のわからない，しかも非常に大きい，しかも疎なデータ表”は，どのように処理すればよいのだろうか．たとえば，対応分析法やクラスター化はどう行えるのか（ソフトによっては実行できないことがままある）．
- さらに，欧米の，あるいは日本語以外の言語で動作するテキスト・マイニング・ツールがそのままでは適用できないということもある．

こうした事象を書き出すと際限がない．要は，何かできそうな期待感だけが先走って，本当は何が，どこまでできるのか，不透明なことが多々あることに注意すべきである．