68. テキスト型データの多次元データ解析

——Web 調査自由回答データの解析事例

【概要】 定性情報の利用法があらためて注目されている．とくに，社会調査，市場調査等においては，自由回答・自由記述をはじめ，さまざまなテキスト型データの取得方法や解析手法の実用研究が期待されている．また，情報技術の進歩に伴い，多数のテキストマイニングの手法が登場している．

とくに，インターネットの普及に伴い，Web 調査等が多用され，自由回答データの取得方法や利用法も様相が変わりつつある．しかも，長年の実績に裏づけされた従来からの科学的調査法の援用を受けた，新たな自由回答取得方法，解析手法の研究が必要な時期にある．

ここでは，言語情報処理技法（形態素解析，とくに分かち書き処理）と統計解析との諸要素技術を適切に組み合わせることで，個々の方法論だけでは解決できなかった調査分野のテキスト型データ解析手法の新たな方向を示すと同時に，その適用可能性を事例解析により紹介する．

現象解析に当たって重要なことは，問題とする対象の説明のために，最も適したデータ取得法はどうあるべきか，その取得データの解析法はどうあるべきかを，データ科学の視点から実証的に検証することであり，ここでの分析例をその一例とする．

[キーワード：自由回答，テキスト型データ解析，分かち書き処理，WordMiner，Web 調査，対応分析，クラスター化法（ハイブリッド法）]

68.1 定性的調査とテキスト型データ

68.1.1 テキスト型データ解析の方向と適用の範囲

日本語の電子的処理がさまざまな形で実現可能となったこと，言語情報処理分野の諸研究が進んだこと等から，テキスト型データ（textual data）の取得方法や解析手法への関心が高まっている．とくに，社会調査（意識調査，態度調査等）や市場調査等の分野における自由回答・自由記述（open-ended question，free answer）の取得方法や取得後に必要とされる多次元データ解析の具体的な方法論の登場が期待されている．

とくに，調査の品質があらためて問われているこの時代にあって，さまざまな要因から満足できる調査の実施が極めて困難になっている．主として質問紙を用いた従来からの調査（面接調査，留置調査等）の実施困難性やさまざまな問題の提起，たとえば住民基本台帳の閲覧制限，情報公開法の実施等に関連した調査情報取得環境の変容がある．また，伝統的な標本調査法に従ったサンプリング操作を経ておこなわれる調査（たとえば従来の調査の中心であった面接調査，留め置き調査等）が，経済的にも労力の面からも負担が大きく，一方それに見合った成果が次第に期待できない状況にあり（たとえば回収率の低下），定性調査に次第に関心が移行する傾向が見られる．

一方，インターネットの普及により，電子的調査情報取得手法（CASIC；computer assisted survey information collection）の研究や実用化が進み，テキスト型データの取得が，内容の質の適否にかかわりなく，容易に，しかも大量取得が可能となっている．とくに，インターネットを用いる調査（Web 調査，電子メール調査等）が多くなった．

ここでは，データ科学の理念に従い，独自の調査計画に基づき実施された Web 調査（Web-based surveys）で取得した自由回答データの分析の一部を紹介することで，多次元データ解析手法のこの分野への適用可能性への事例とする．したがって，ここで述べる方法論は，言語学や言語情報処理における従来の研究やその延長線上にある種々のアプローチとは視点が異なる．筆者らの主張は，探索的多次元データ解析手法を日本語テキスト型データの解析に取り入れたとき，どのような分析が可能であるのか，どこに問題があるかを，データ科学の考え方で解決を図るための１つの実証研究を示すことである．

取得データの多様化と膨大な情報の電子化やデータベース技術の進歩に伴い，データアーカイブ，データウェアハウス等の関連技法の技術的要素を背景に，データマイニングの方法論が登場している．とくにナレッジマネジメント，知識の組織化や知識発見に関連して，テキストマイニングをデータマイニングツールの一機能として組み入れる傾向があるが，目覚ましい成果が上がっているかは現時点では即答できない．

従来の言語学研究の方向は，大きく自然言語処理，計量言語学的な研究，計量文献学的アプローチ，言語情報処理や全文検索技術等に分けられる．とくに人工知能研究を基盤にして，たとえば「発想支援ソフト」として，「DEFACT（デファクト）」（電通）やその元となった HIPS（Hybrid Idea Processing System，富士通研究所），AIDE（Augmented Informative Discussion Environment，ATR），VextSearch（コマツソフト），SurveyAnalyzer（NEC）等がある．この種のソフトの一部はすでに商用化されている．

68.1.2 テキスト型データの解析上の留意事項

自由回答文の解析を考えるとき，① 考えたことがないことには答えにくい（「白紙に何かを書くように」といわれても即座には思いつかない），② 予想しなかった回答や知見が得られるという期待もある，③ 無記入が多くなる傾向があるとされる，④ 調査法や標本抽出法との関連性が明らかにできないとされる（妥当性の問題），⑤ 通常の選択肢型設問の選択肢の影響を受けるとされる（回答誘導の懸念），⑥ 適切なデータ解析法がないといわれてきた，⑦ 内容の再現性や信頼性に欠ける，等が指摘されてきた．

こうした指摘は当然でありこの種の研究課題の複雑さを示すものである．とくに，調査における自由回答データ取得上の重要な考慮事項は，選択肢型設問調査と異なり，数量として定量的かつ客観的な評価が困難であるとされてきたことにある．現時点では，定性調査を定量調査との優劣を比較するという観点ではなく，“自由回答データ”に基づく定性情報の取得過程において，従来の選択肢型設問形式による調査との“併用”が妥当であるとの観点からの議論が肝要である．テキスト型データ解析システムの開発に際しても，こうした発想が反映された設計指針を設ける．これは，林や大隅が主張してきた「データ科学」の理念を具体的な実証例として実現することでもある（林，2000；Ohsumi, 2000）．

68.2 テキスト型データの解析上の課題

68.2.1 日本語の特徴と形態素解析

言語類型学的には，言語を孤立語，膠着語，屈折語等と分類する（加賀野井，1995, 1999）．日本語が欧米の言語と大きく異なることの１つに，文章・テキスト型データが「べた書き」（膠着言語）であって「分かち書き」されていないということがある．

さらに，日本語は漢字，カタカナ，ひらがな，それに外来語（それに充てられた漢字や仮名等）が混在しているという特徴もある．欧米語が単語という単位で区切られた言語であることから，これを単位として扱うことができるという処理の容易性がある．日本語はこれが困難であるだけでなく，複数の語が連結されて複合語を形成することが多い．また，現実の現象として，ワードプロセッサの登場による表記法の変化，さまざまな分野の専門用語，さらに E-mail 語やチャット語，携帯電話用語（ケータイ語）の登場と，日本語の様相はさまざまである．

このために，日本語の処理をおこなうには，まずある要素単位に文章を分解する「分かち書き処理」が必要となる．これを含めていくつかの処理過程を「形態素解析」（morphological analysis）という．形態素（morpheme，morphology）とは，表記された文章を「最小の有意義な意味ある単位，意味を持つ最小の単位」（池上，1993）あるいは「単語や接辞等，文法上，最小の単位となる要素のこと」（長尾他，1998）をいう．形態素とは絶対的な概念ではなく，あくまでも１つの便宜的な約束事であり，しかも元来は日本語独自の考え方でもない．いずれにしても，日本語データ解析処理においては，何らかの意味で文章をある単位に分けねばならない．このため形態素解析が必要であり，コンピュータ処理機能の向上のおかげで，いろいろな解析方式が提案されている（全文検索システム協議会，1997）．

68.2.2 テキスト型データの多次元データ解析と解析用ソフトウェア

ここでは，テキスト型データ解析を，探索的な多次元データ解析のパラダイムの中で考える．つまり，従来からの形態素解析（とくに分かち書き処理）と多次元データ解析の諸要素技術を適当に組み合わせることで，個々の方法論では解決できなかったデータ解析手法としての新たな方向への展開を図ることを試みる．とくに問題を単純化して，次のように考える．

まず，電子ファイル化した自由回答文・テキスト型データ等を，「分かち書き」により，分析が可能な「構成要素」(fragments) に分解する．「構成要素」とは，データ解析上の処理単位を表し，一般にいう単語・語，文節等より緩やかな意味で用いる用語である．緩やかな決まりとする理由は，元々の取得データ自体があいまいかつ多様な表現であるから，厳密な定義を避けて，分かち書きの若干の不具合は許容し，むしろ分析を容易にする方向で考えるという意味である．とくに，ここでは日本語の精密な言語学的研究が目的ではないこと，得られるデータにあいまい性があること，分かち書きだけでなく，従来から自由回答の処理法として利用されてきたアフターコーディング処理の併用や比較検証も必要となること，等々の理由がある．

次に，構成要素の出現頻度のパターンの探索的多次元データ解析の方法として考える．通常は「出現頻度の高い語は重要である」あるいは「頻度の近い位置にある語は関連性が高い」といった経験的ルールを用いることが多い．しかしここでは，分かち書き処理で得た「構成要素」の並び（パターン）の特徴抽出と考える．

ここで，自由回答データを解析する主たる目的は，「類型化による規則性の探査と個別意見・回答別意味の把握」にある．つまり，集積した自由回答・自由記述データに潜在する構造の類似・差異や規則性等を知ることにある．とくに，個々の回答・記述の意味内容や特徴，意見の規則性や典型を知ることが必要となる．同時に，解析から得た「類型・典型」に含まれる「個々の回答データの特徴」を読み取ることや，得られた典型や大勢の回答傾向だけでなく，少数例・少数意見の特徴も知りたい．つまり，単なる文章要約や分類では十分ではなく，意見の客観的な類型化とその内容の解釈が必要となる．

次に必要な要件は，従来の定量的調査法の方法論の援用を考えることである．自由回答データの特徴の1つに，選択肢型設問や属性等で得た数値データのように定量的に統計値として評価できないということがある．通常の選択肢型設問を例にとれば，回答比率データを算出し，統計的操作により標本誤差を検討し，設問間の差異を比較検証することが可能である．一方，自由回答データの場合，こうした操作が難しい．しかし調査結果に何らかの客観的な保証を与えるためには，間接的ではあっても従来の標本調査の理論や知識の援用を得ること，あるいは比較可能であること，つまり定量的操作との併用が，自由回答の解析を妥当なものとする付加措置として必須である．このことから，従来型の選択肢型設問項目や属性項目等と自由回答設問とを併用し，自由回答の分析結果に加えて，これら設問項目との相互関連性の検証を可能にする評価機能が重要となる．つまり，定性的調査と定量的調査の併用法を考えねばならない．

以上の主旨に沿った解析を達成するためには，それに適した統計システムの開発が求められる．フランスの研究者を中心に開発された SPAD.T (Système Portable pour l'Analyse des Données Textuelles) を基本エンジンとし，これに分かち書き処理ほかの日本語解析に必要な機能を追加した WordMiner® はそうしたソフトであり，日仏共同研究の成果の1つである．WordMiner は，調査環境下におけるテキスト型データの分析時に発生する諸事象を想定した記述的多次元データ解析を設計指針とするもので，とくに，選択肢型設問・属性項目等を併用する自由回答型を含めた調査データの解析に適している．その主な機能は，① 各種変数の編集機能，② 分かち書き処理による構成要素の生成，③ 構成要素の編集機能，④ 対応分析による処理，⑤ クラスター化によるサンプルと構成要素の類型化，⑥ 追加処理機能の応用，⑦ 構成要素，クラスター化生成情報，サンプル相互間の有意性の検証，等からなる（大隅他，2000；Lebart *et al.*, 1998)．

68.3 事例解析

68.3.1 調査方法と調査の特徴

インターネット環境下でおこなわれる Web 調査を，複数の調査機関との共同研究として実施してきた（大隅，2000, 2001)．Web 調査の特質の1つは，自由回答の取得が容易であり，しかも豊富な内容の情報が取得できることといわれてきた．しかし，これの確たる論証はなく，したがって Web 調査の研究目的の1つとして，自由回答の取得方法の研究を挙げてきた．とくに，Web 調査における有効回答率は極めて低く，これが Web 調査の特徴の1つでもある．また回答者の基本属性に，① 回答者の年齢区分が 20 歳代から 40 歳代に偏る傾向にある，② 性比構成の差異が顕著で男性回答者が女性のそれより多い，③ 調査をおこなうサイト上の回答者集団の登録方法により結果に差異がある，④ 回答者の居住地域が都市圏に集中するが，しかし，国内のあらゆる地域からの回答があるという Web 調査特有の地理的距離の消滅現象も見られる等，従来型調査とは異なるさまざまな特徴が見られる．これらの特徴はインターネット利用者数の増加に伴い，次第に緩和される傾向にあるといわれている．

とくに住民基本台帳や選挙人名簿に基づく標本抽出操作を前提とし，調査員による面接法や留置自記式法等で得られた従来型調査の結果とはかなり異なる傾向にあることから，回答者の代表性や信頼性を疑問視する意見も多い．つまり調査法としての基本特性の評価や検証は今後の研究を待たねばならない．このことから，あとに述べる解析結果の解釈には，ここで指摘した回答者特性を念頭に対応することが必要である．

68.3.2 調査データの概要と分析に用いる設問

Web 調査は，1997 年以来数年にわたって実施してきた．解析に用いる例は，3 つの調査機関で，1999 年度，2000 年度に実施した中の 4 回分の調査結果であり，その主な内容は表 1 のようになっている．詳しい集計値は省略するが，この種の Web 調査における自由回答の回答行動の傾向として，自由回答設問への記入率，選択肢型設問の「その他」(自由記述部) への記入率が，従来型調査に比べて高いこと（多くの場合，80％以上の記入率となる)，女性の回答者数が少なく，年齢区分に偏りがある，回答者の地域の偏りがある，等の特徴がある．

前述のように Web 調査の研究目的の1つに，自由回答取得方法の検証がある．実際に Web ページ上の調査票の中での自由回答の設問方法には，従来の質問紙と異なるさまざまな工夫をおこなっている．たとえば，回答記入欄の工夫（大きさ，枠組み，罫線有無等)，設問の配置，選択肢ボタンの工夫，前後の設問との文脈関係，箇条書き方式の有無，

表1 分析対象としたWeb調査の概要

	電通リサーチ社（1999年調査）				NTTナビスペース社（1999年調査）			
調査期間	1/28/1999～2/4/1999				2/16/1999～2/23/1999			
	回答数	男性	女性	無回答*	回答数	男性	女性	無回答*
全体	1045	831	202	12	1258	953	296	9
	%	79.5%	19.3%	1.1%	%	75.8%	23.5%	0.7%
分析対象としたサンプル数	1039	834	203	2	1250	951	296	3
	%	80.3%	19.5%	0.2%	%	76.1%	23.7%	0.2%

	リクルートリサーチ社（1999年調査）				電通リサーチ社（2000年調査）			
調査期間	3/1/1999～3/8/1999				4/20/2000～4/27/2000			
	回答数	男性	女性	無回答*	回答数	男性	女性	無回答*
全体	679	394	275	10	2773	1556	1192	25
	%	58.0%	40.5%	1.5%	%	56.1%	43.0%	0.9%
分析対象としたサンプル数	678	396	278	4	2591	1492	1075	24
	%	58.4%	41.0%	0.6%	%	57.6%	41.5%	0.9%

* 無回答：性別に対して無回答であったが，自由回答を含むので，除外せず分析対象とした．

表2 「日本人の国民性研究」調査で用いる自由回答設問

（質問1.1）あなたにとって，一番大切と思うものはなんですか．一つだけあげてください（どんなことでもかまいません）．
（質問1.2）では，この他に大切なものとして，何がありますか．いくつでもあげてください．

ページネーション等についての配慮をおこなった．このようなことで，用いた調査票の中には多数の自由回答設問が含まれるが，ここでは次に挙げる2つの設問を例として用いる．

表2の（質問1.1）は，統計数理研究所が5年おきに実施している「日本人の国民性研究」調査で用いられてきた自由回答設問を若干変えたものである．また，（質問1.2）はこのWeb調査において新たに加えた設問である．

68.3.3 分かち書き処理と構成要素の観察

分析のはじめに，原文（解析対象とする自由回答データ）に対して，分かち書き処理をおこなう．分かち書き処理によって得られる構成要素として，分かち書き文の中のすべてを選ぶ「分かち書き抽出機能」と，その中から特徴的な語を選ぶ「キーワード抽出機能」がある．ここでは形態素解析による品詞特定等をおこなわないので，ここでいうキーワードとは，品詞を限定した内容とはなっていない．また，各回答文の分かち書き後の構成要素数（分かち書き数，キーワード数）のそれぞれの数値ファイルも生成し，必要に応じて分析に用いる．

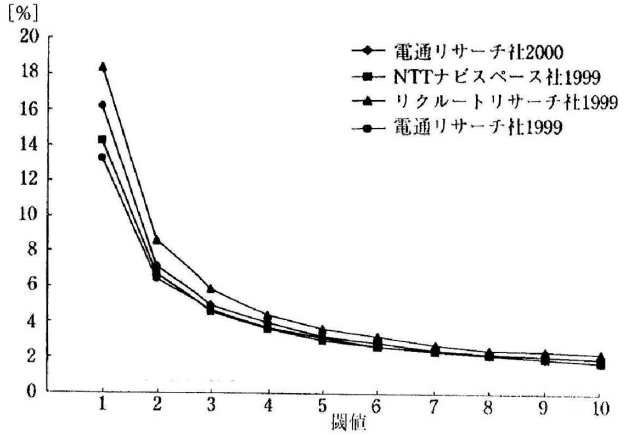

図1 異なり構成要素数と閾値の関係

a. 構成要素の頻度分布と構成要素異なり率の分布—4回調査の比較

あとの分析では，前述の2つの設問（「一番大切なもの」，「次に大切なもの」）を併合して用いる．もちろん，一番大切なものと次に大切なものとは若干意味が異なるし，またそうであるから，設問を分けてあり，個々の設問について個別に分析をおこなうこともできるのであるが，ここではあえて2つを併合している．したがって，回答の順序効果（つまり一番か次にか）という情報は以後の分析では考慮されない．まず4つの調査サイトにおける自由回答文データの分かち書き処理で得た情報を要約する．

分かち書き処理結果と構成要素編集機能を用いて，「総構成要素数(率)」，「異なり構成要素数（率）」，「編集後の異なり構成要素数（率）」，「構成要素の出現頻度区分別集計」等の諸量が得られる．これらのすべてを示すことはできないので主な指標について，4回の調査結果を比較する．

図1は「異なり構成要素率」と「閾値」の関係を示している．「異なり構成要素率」とは，総構成要素数（分かち書き後の総語数）に占める異なり構成要素数（同一表記の構成要素を1語と数えたときの語数）の割合である．また，「閾値」とは，指定したその値以上の頻度の構成要素数のことをいう．たとえば，閾値＝2とは，出現頻度が2語以上の構成要素数という意味である．図に見るように異なり構成要素率は閾値が増えると急速に低減する．しかも1語が圧倒的に多く，2語以上は急速に低減することが見える．また，この例だけでなく，総構成要素数が大きいほど，こうした共通した特徴がある．

一般に，異なり構成要素率が少ないと，用いられた用語が特定の内容に集中しており，一方この数値が大きいことは，表記の自由回答の内容が多岐にわたり発散していることと考えられている．数理的な根拠があるわけではないが，常識的に考えて，異なり構成要素率の推移を探査することは，自由記述の内容のまとまりの程度（意見の発散度）を知る指

表3　キーワードの出現頻度の分布

順位	キーワード	構成要素数*	異なり構成要素数	異なり構成要素率**[%]	順位	キーワード	構成要素数*	異なり構成要素数	異なり構成要素率**[%]
1	家　族	1347	1305	60.5	16	事	153	125	5.8
2	自　分	793	715	33.1	17	心	153	145	6.7
3	健　康	589	572	26.5	18	子　供	147	142	6.6
4	お　金	544	542	25.1	19	思いやり	130	129	6.0
5	仕　事	494	481	22.3	20	環　境	123	116	5.4
6	人	447	422	19.6	21	家　庭	120	120	5.6
7	友　人	440	438	20.3	22	一　番	118	117	5.4
8	生　活	370	350	16.2	23	信　頼	112	112	5.2
9	趣　味	354	354	16.4	24	社　会	103	100	4.6
10	時　間	316	296	13.7	25	会　社	100	100	4.6
11	関　係	300	233	10.8	26	幸　せ	90	85	3.9
12	人　間	267	204	9.5	27	自　身	90	90	4.2
13	友　達	265	262	12.1	28	恋　人	90	87	4.0
14	大　切	200	191	8.9	29	今	86	85	3.9
15	気持ち	183	179	8.3	30	両　親	86	85	3.9

* 構成要素数でソートした.

** 異なり構成要素率[%]＝(異なり構成要素数÷総構成要素数)×100，総構成要素数＝2157

標として重要と考えられる（たとえば，Lebart (1998) は語彙の潤沢度とし，村上 (1994) は語彙量の指標としている）.

しかし同時に，サンプル数（自由回答数）が増え，それに応じて総構成要素数が増えると，異なり率が次第に低減する（つまり，設問の内容に対する表記内容が限られ，また特定の用語の出現頻度がならされ，指数的に頭打ちとなる）．ここに挙げた例でも，回答者数の大きさにあわせて，「リクルートリサーチ＞電通リサーチ 99＞NTT ナビスペース＞電通リサーチ 2000」のように順に順位が下がっている．いずれにせよ以後の解析で，閾値を決めて何語以上の構成要素を解析に採用するかは重要な選択肢となる.

b.　キーワードに見る構成要素の出現頻度の特徴

4 回分の調査データに見られる特徴を比べる前に，どのような構成要素が自由回答を特徴づけているのかを，キーワードを使って探索的に調べる．用いるデータセットは，電通リサーチ（2000 年調査分）とし，これの「大切なもの」に相当する自由回答から抽出されたキーワードを出現頻度順に検索し要約した（表 3）．上から順にみると，「家族，自分，健康，お金，仕事，人，友人，…」と続く．これらが自由回答に含まれる主要な語であり，実はこの結果は「日本人の国民性の研究」調査（統計数理研究所，1998 年第 10 次全国調査）における集計結果と極めて類似している.

たとえば，「日本人の国民性の研究」の集計結果では「家族（約 40%）」，「生命・健康・自分（約 22%）」，「愛情・精神（17 %）」の順になっている．この調査は調査員による「面接調査」であり，回収結果に基づいて分析者のアフターコーディング処理により整理したものである．また，両者は調査法も異なり，サンプル属性にもかなりの違いがある．とくに，Web 調査の場合の属性は，20〜40 歳代に回答者の分布が偏っている．しかしそれでも「家族」，「子供」，「生命・健康・自分」，「愛情・精神」等に回答が集まる傾向は類似している.

c.　構成要素の類似性，同義語等

ところで，キーワードの中に「友人，友達」，「お金，金」，「子供，こども，子ども」，「親，両親，母親，父親」，「幸せ，幸福」，「思いやり，おもいやり，思い遣り」のように，同義語（シノニム；synonym）あるいは類語の別表記や，シソーラス（分類語彙表；thesaurus），コーパス（corpus）等を考慮すべき内容が表れる．またこれとは別に，分かち書きの不具合の調整（再編集）や分かち書きして欲しくなかった語の併合（再結合）等も必要となる.

68.3.4　構成要素の出現頻度と順位の特徴

a.　構成要素の編集—削除，置換等

では 4 回分の自由回答データの分かち書き処理そのもので得た情報を用いるとどうであろうか．この場合，キーワードを用いるときに比べて登場する語やその連なりがやや複雑な様相を示す．とくに，不要語の削除，類語・同義語の併合と置換，誤記の訂正等，分かち書き後の再編集が必要となることが多い．こうした操作を一切おこなわず，分かち書き処理後に，ゴミの除去（特殊文字，記号等を除く）程度で直ちに解析をおこない，観察を進めることで十分な場合もある．とくに語の除去は，分析の意図に応じてさまざまである．たとえば句読点や括弧類の削除，助詞の削除，特殊記号の削除等がある．句読点や助詞の出現頻度に意味があると考えてこれらを活かす場合もあるし，括弧の利用が意味表現の強調であるとか，カタカナ表記に意味を持たせる，あるいは助詞の利用方法に意味があると考える場合もある．置換の対象の検討でも同様のことが起こる．たとえば，「主人」，「夫」と「旦那さま」とは意味が異なると考えたい等である．ここまで考えると，含意や意味論的な視点に重きをおいた解析となるが，ここではこれとは異なるアプローチで分析を進めることが目標である．むしろ，自由回答設問では，回答の表記の多様性に対してどう編集操作により対処するかがある.

たとえば，ここで扱った自由回答の編集内容のうち，削除や置換のごく一部を表 4 に挙げた．ここに見るように，自由回答設問では回答者は極めて多種多様な書き方をする．これが特徴であり，したがって，これをどのように要約するかの工夫が必要になる．筆者等が分析手法や解析ソフトだけでなく，自由回答の設問方式の研究も必要と考える根拠の 1 つである.

削除の例：　表 4 にあるように，句読点，助詞，特殊記号等のほかに，「特になし」，「わかりません」等の回答も削除対象としている．「特になし」，「別にありません」等は多くの場合，外れ値となることが多いからであるが，これが意味を持つような設問もありうる．たとえば，機器類の故障の状況を自由記述するような場合には，「特になし」等は無視できない語である.

表4　削除と置換の例

削除対象	置換前*	置換後**
とくになし，特になし，特にありません，とくにありません，とくにない，別にありません，別になし	Free	自由
の	jibun，じぶん，自分，私，私自身，自分自信，自分自身	自分 自分自身
を	kazoku，か ぞく	家族
に	あい，LOVE，愛情	愛 愛情
と	あそび，あそび	遊び
が	あっきぽい	あきっぽい
な	いきがい，生き甲斐，生甲斐，生きがい	生甲斐
で	いきて，生きる，いきる	生きて
や	いのち，命，生命	いのち 生命
も	うち，家，持ち家	家
は	おもいやり，おもいやる，思い 遣やる	思いやり
か	お金，かね，おかね，カネ，金銭，金品，御金	金
は	お稽古，お稽古事	お稽古事
か	お付き合い，つき合い，人付き合い	付き合い
へ	か けず	かけず
かも	か たち，かたち	形
そう	かん きょう	環境
ような	きずな，つながり，絆	絆 つながり
から	こども，子ども，子供，子どもたち，子供たち，子供達	子供 子供達
だけ	た ～ く さん	たくさん
して	だんな，夫，ダンナ，主人	夫 旦那
でも	女房，妻，家内，奥さん	妻
では	シアワセ，幸せ，幸福感	幸福
ので	アウトソーシング，アウトソ シング，アウトソ-シング	アウトソーシング
には	ケータイ，携帯 電話	携帯電話
こそ	親，兄弟，親兄弟，両親，父母，母．母親，父，父親	親 両親 兄弟
さえ	Health，けんこう	健康
しか	友人，友達，友だち，仲間，親友達，親友，友，ともだち	友人 仲間
ばかり	人々，人たち，人達，ヒト，ひと，ひとたち	人

* 置換，併合，誤記訂正，分かち書き不良訂正，等々無数にあるがそのごく一部を示した．
** 置換後の語は，それ自身を含む場合は，置換前には記載していない．
電通リサーチ，2000年調査結果を用いた．

置換の例：　例としたサンプル数（2591サンプル）程度の自由回答でも，493語もの置換をおこなっている．この中の特徴的な語の一部を示した（表4）．ここで，右の2つの

欄の左側の語群を右側の1語で置換することを表す．このような平易な設問でも，自由回答の表記が実に多種多様であることがわかる．しかし，構成要素の再編集をおこなうことで，通常は異なり構成要素数がさらに整理され，後述のように実際の解析対象となる語数はそれほど多くはならず，分析の見通しが改善される．

b．調査回間の比較

次に4回分の調査データについて，上のような編集を経て得られた構成要素群の特徴をキーワード（表3）にならって出現頻度順に整理する（表5）．ここで出現頻度とは重複を許す総出現構成要素数である．これを見るとあらためて説明を待つまでもなく，各調査回で登場する語が極めて類似している．

たとえば，「家族，自分自身，友人，金，健康，仕事，生活，人間関係，親一両親一兄弟，子供一子供達，余裕一ゆとり，…」等は，若干の順位の上下はあっても，どの調査回にも共通に現れている．また総構成要素数に占める各構成要素の割合の低減傾向にも類似性がある．調査回，回答者集団が異なるにもかかわらず，自由回答の結果にこうした類似性があることに注目すべきである．

この結果に限らず，解析全体を通じて4回の調査回間には類似性が見られるのだが，同時にそれぞれの調査機関に固有の特徴もある．とくに，ここで取り上げた設問については，3つの機関の4回の調査結果は大変によく似た傾向を示している．これを網羅的に述べることは無理があるので，以下では電通リサーチ2000年調査における取得データによる分析例を示す．要は，自由回答設問であっても，反復調査あるいはパネル調査，継続調査等をおこなうことで，回答の類似性や差異性がより明らかになるということである．ここらは従来，自由回答・自由記述取得に関する研究が十分にはおこなわれていなかった部分である．

68.3.5　多次元データ解析による分析と結果の解釈

a．データ表の構成

対応分析の特質を活かして，分析対象とする2元データ表の表側と表頭に充てる項目を，利用目的にあわせて選択する．たとえば，以下のような分析をおこなう．

1)　(抽出した構成要素)×(属性，たとえば性別，性年齢区分別) の分析
2)　(抽出した構成要素)×(回答者のクラスター化情報) の分析
3)　(抽出した構成要素)×(回答者，サンプル) の分析

このとき，行列の大きさ（次元数）がさほど大きくないデータ表から，大きさが大きくかつ行列要素が非常に疎なデータ表までさまざまな形態が起こりうる．上記の1)，2)は前者の例であり，後者の例が3)のデータ表である．このための計算処理上の工夫を必要とし，データ表によってアルゴリズムを使い分けねばならない．

b．「構成要素×性年齢区分」のデータ表の分析

多くの場合，抽出した構成要素群が，他の情報，たとえば選択肢型設問・属性あるいはそれに代わる何らかの定性情報とどう関連するかを知りたい．たとえば「性別」，「年齢区分」はどう関係するのか，あるいは用意した選択肢型設問のどれが，どのような構成要素

表5　構成要素の出現頻度分布（調査4回分の比較）

順位	電通リサーチ-2000(2591サンプル) 単語	出現頻度	(出現頻度/3351)×100	電通リサーチ-1999(1039サンプル) 単語	出現頻度	(出現頻度/1596)×100	NTTナビスペーズ-1999(1250サンプル) 単語	出現頻度	(出現頻度/2401)×100	リクルートリサーチ-1999(678サンプル) 単語	出現頻度	(出現頻度/1223)×100
1	家　族	1371	40.9	家　族	520	32.6	家　族	668	27.8	家　族	350	28.6
2	自分—自分自身	1034	30.9	こと—事	344	21.6	こと—事	537	22.4	こと—事	235	19.2
3	友人—仲間	814	24.3	友人—仲間	241	15.1	自分—自分自身	384	16.0	友人—仲間	195	15.9
4	金	658	19.6	金	223	14.0	友人—仲間	300	12.5	自分—自分自身	191	15.6
5	健　康	617	18.4	自分—自分	216	13.5	生　活	287	12.0	金	177	14.5
6	仕　事	580	17.3	健　康	215	13.5	健　康	275	11.5	健　康	127	10.4
7	人	479	14.3	仕　事	188	11.8	仕　事	242	10.1	仕　事	118	9.6
8	生　活	393	11.7	生　活	156	9.8	金	238	9.9	時　間	118	9.6
9	で　す	368	11.0	時　間	151	9.5	時　間	195	8.1	生　活	117	9.6
10	趣　味	363	10.8	趣　味	109	6.8	人	187	7.8	人	71	5.8
11	時　間	353	10.5	人	104	6.5	環　境	153	6.4	環　境	65	5.3
12	あ　る	296	8.8	環　境	89	5.6	あ　る	147	6.1	あ　る	64	5.2
13	人間関係	288	8.6	あ　る	88	5.5	大　切	117	4.9	趣　味	56	4.6
14	す　る	268	8.0	親—両親—兄弟	70	4.4	す　る	106	4.4	親—両親—兄弟	53	4.3
15	親—両親—兄弟	224	6.7	す　る	65	4.1	心	105	4.4	余裕—ゆとり	49	4.0
16	大　切	220	6.6	余裕—ゆとり	59	3.7	し　て	97	4.0	す　る	42	3.4
17	できる	207	6.2	し　て	56	3.5	いる	94	3.9	できる	42	3.4
18	子供—子供達	198	5.9	人間関係	55	3.4	できる	91	3.8	愛　情	41	3.4
19	気持	197	5.9	いる	54	3.4	余裕—ゆとり	90	3.7	心	39	3.2
20	心	190	5.7	心	54	3.4	趣　味	89	3.7	思いやり	33	2.7
21	いる	170	5.1	愛　情	49	3.1	親—両親—兄弟	79	3.3	いる	31	2.5
22	思いやり	153	4.6	自　然	49	3.1	子供—子供達	77	3.2	して	29	2.4
23	環　境	138	4.1	できる	47	2.9	ない	75	3.1	自　由	28	2.3
24	くれる	135	4.0	安　定	39	2.4	など	75	3.1	気持ち	27	2.2
25	ゆとり—余裕	133	4.0	ない	38	2.4	思　う	62	2.6	幸　福	26	2.1
26	愛—愛情	127	3.8	家　庭	38	2.4	幸　福	61	2.5	自　然	25	2.0
27	次	127	3.8	生きがい	37	2.3	人間関係	58	2.4	ない	24	2.0
28	いう	126	3.8	恋　人	36	2.3	思いやり	56	2.3	人間関係	24	2.0
29	家　庭	123	3.7	子供—子供達	34	2.1	愛　情	55	2.3	恋　人	24	2.0
30	ない	121	3.6	幸　福	34	2.1	自　然	53	2.2	など	23	1.9

1999年調査3回分と2000年調査1回の，再編集に用いた条件はやや異なっている．前者は分析にInfoMinerを用い，後者はWordMinerを用いた．

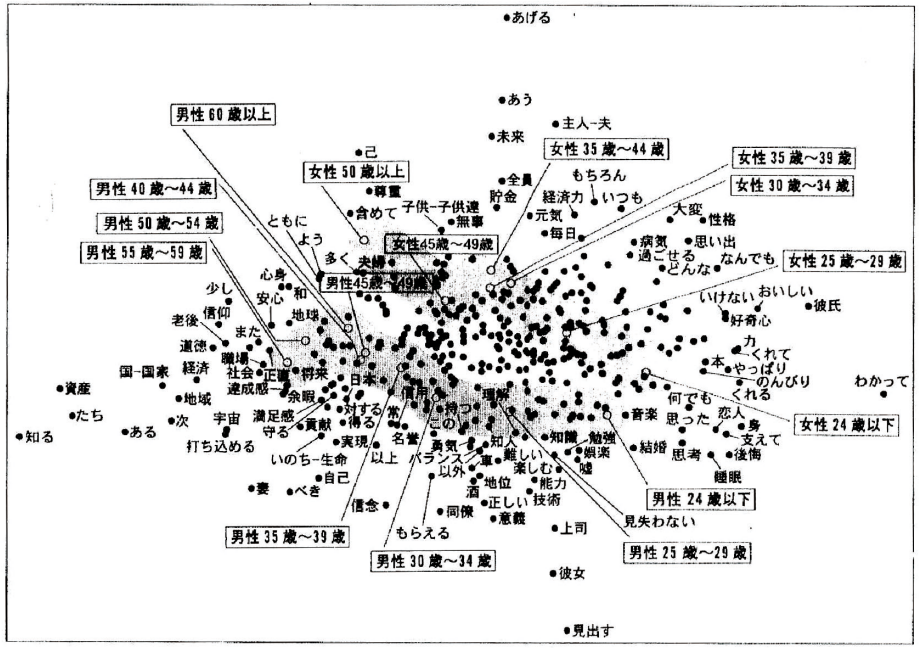

図2　構成要素と性年齢区分との関係（同時布置図）

と関連があるのか，その程度は，といったことである．

1つの例として，性別と世代間の類似差異を調べるため，新たに「性年齢区分」変数をつくり，これと利用語の間の関係を見る．年齢区分は男女とも5歳間隔とした．なお以後の解析で，女性については50歳以上のサンプル数が極端に少ないので，50歳以上をくくった．

次に対応分析を用いて，構成要素群と性年齢区分との関係を調べる．この結果は布置図として観察できる．ここでは，用いた構成要素数が409語（閾値8語以上），性年齢区分が16区分となっている．図2はこれを示すものだが，構成要素をすべて表示すると煩雑になるので，周辺に位置する主な語を示し，これと性年齢区分の同時布置図とした．これを見ると，①男女の性差が明らかに見られる，②加齢に伴う意識変化が見える（図の右下から左上に向かう年齢変化が見られる），③男女とも，30歳あたりを境にして変化が大きく，40歳以上ではあまり年齢差が出ていない，等の特徴が見えてくる．

また，各性年齢区分のカテゴリーの近傍にある語の関係から，およその特徴が見える．たとえば，特徴的な語として「妻」，「主人—夫」，「彼氏」，「彼女」等，あるいは「子供—子供達」，「健康」，「親—両親—兄弟」，「家族」等の語と性年齢区分の関係がおおよそ浮かび上がってくる．しかし，これは限られた次元（空間）の中で眺めたものであり，これだけではデータ構造の探査としては十分ではない．

表6　性年齢区分別の構成要素の有意性テスト（男性）

性年齢区分	24歳以下	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50～54歳	55～59歳	60歳以上
区分内サンプル数	137	239	279	333	244	149	91	27	48
1	彼女	時間	妻	己	家族	家族	安心	ことに	社会
2	支えて	自己	打ち込める	見出す	健康	社会	道徳	こと	現在
3	金	ため	技術	存在	仕事	最低限	家族	遊び	なり
4	もの	金	信頼	宇宙	趣味	有る	老後	社会的	また
5	時	貢献	出来る	欺かない	対する	健康	秩序	家庭	意識
6	時間	ゆとり-余裕	趣味	欺かれない	社会	行動	健康	付き合い	健康
7	友人-仲間	趣味	夢	信条	将来	家庭	夫婦	仕事	充実
8	必要	らしく	家族	即ち	貢献	命	自然	とは	した
9	正しい	自由	車	べき	安定	常	社会	資産	必要
10	地位	恋人	満足感	たち	環境	しない	平和	勉強	相手
11	目標	プライベート	娘	知人	対して	生活	趣味	有る	目的
12	思う	自分-自分自身	家庭	意義	やりがい	他人	地域	健康	安全
13	何でも	彼女	もらえる	全て	打ち込める	趣味	暮らし	いのち-生命	それ
14	感じる	バランス	見つける	守る	より	守る	少し	信仰	できる
15	常識	人脈	金	知る	地球	努力	和		生活
16	名誉	精神的	評価	価値	される	貢献	個人		財産
17	恋人	勉強	信用	世界	収入	等	職場		よう
18	経験	友人-仲間	なく	心	妻	取入	生活		明るい
19	睡眠	知識	面	地域	経済	責任	安全		地域
20	勇気	見失わない	コミュニケーション	家族	希望	正直	生甲斐		関係
20	あう	なおかつ	その	これ	主人-夫				
19	ペット-猫-犬	まぁ	したい	いう	あう				
18	収入	スムーズ	くれる	やはり	上司	あう			
17	上司	主人-夫	おもいます	おもいます	性格	上司			
16	性格	一番	つければ	つければ	大変	性格			
15	大変	あう	なおかつ	なおかつ	難しい	大変			
14	難しい	上司	まぁ	まぁ	己	難しい			
13	己	性格	スムーズ	スムーズ	労働	己			
12	こと	大変	主人-夫	あう	気持	労働			
11	やりたい	難しい	あう	上司	人	いく			
10	人間関係	己	性格	性格	これ	金			
9	家庭	労働	大変	大変	やはり	いう			
8	したい	これ	一番	難しい	やりたい	次			
7	心	やはり	己	いく	したい	いる			金
6	です	したい	労働	次	友人-仲間	理解			自分-自分自身
5	次	大切	大切	やりたい	です	くれる	いう		人間関係
4	健康	健康	です	くれる	恋人	その	次		家族
3	子供-子供達	次	心	です	次	一番	くれる		こと
2	仕事	子供-子供達	人間関係	人	くれる	気持	人間関係		時間
1	家族	です	次	人間関係	人間関係	自分-自分自身	時間		仕事

表7　性年齢区分別の構成要件の有意性テスト（女性）

性年齢区分	24歳以下	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50歳以上
区分内サンプル数	239	321	287	209	81	42	13
1	次	いる	子供-子供達	あげる	健康	健康	健康
2	人間関係	いい	主人-夫	子供-子供達	主人-夫	正直	皆
3	です	人	親-両親-兄弟	健康	よい	時	病気
4	おもいます	今	思い出	います	心	持ち	国-国家
5	つければ	しても	毎日	主人-夫	一人	常	あって
6	なおかつ	悪い	する	こと	豊かな	子供-子供達	モラル
7	まぁ	ペット-猫-犬	こと	なく	家族	希望	含めて
8	スムーズ	周り	幸福	気	なれる	暮らし	環境
9	上司	思います	確保	実感	夫婦	暮らせる	地域
10	性格	優しさ	もちろん	元気	あれば	自身	持って
11	難しい	できない	大事	教育	モラル	地球	世界
12	あう	主人-夫	未来	平和	親-両親-兄弟	仲良く	生甲斐
13	大変	前向きな	暮らせる	希望	不自由	生き方	食べ物
14	労働	友人-仲間	困らない	使える	友人-仲間	主人-夫	おいしい
15	これ	時間	旅行	未来	全員		尊重
16	したい	常識	もっと	平穏	病気		個性
17	やりたい	素直	色々	かな	子供-子供達		和
18	やはり	気持	全員	ペット-猫-犬	仲良く		経済力
19	くれる	する	時間	あって	人間		
20	理解	なの	やって	生きている	日々		
20	平和	やりたい	社会	恋人			
19	幸福	労働	もの	親戚-親族-親類			
18	安定	したい	これ	妻			
17	できる	会社	性格	その			
16	親-両親-兄弟	社会	したい	楽しく			
15	主人-夫	仕事	労働	理解			
14	己	その	妻	おもいます			
13	ゆとり-余裕	妻	やはり	つければ			
12	家庭	おもいます	おもいます	なおかつ			
11	時間	つければ	つければ	まぁ			
10	する	なおかつ	なおかつ	スムーズ			
9	心	まぁ	まぁ	上司			
8	金	スムーズ	スムーズ	難しい			
7	社会	あう	あう	己			
6	友人-仲間	上司	上司	労働			
5	ある	性格	大変	会社			
4	子供-子供達	難しい	難しい	やはり			
3	趣味	己	己	やりたい	人間関係		
2	生活	これ	人間関係	人間関係	次	仕事	
1	健康	次	次	次	仕事	大切	自分-自分自身

c. 構成要素の性年齢区分別有意性テスト

ここで，性年齢区分別の有意性テストを試みる．各年齢層に有意に働く語群と，逆にその層にはあまり寄与しない語群の20語ずつを表6, 7として要約した．また，寄与の程度を判定する検定値ほかの数値はここでは省略した．検定値とは，ある構成要素（の総利用数，つまりコーパス）が，ある分類基準（ここでは「男性」というカテゴリー）に占める割合（出現頻度）が有意となるか否かを正規近似で判定する1つの指標である（詳細はLebart他，1998；大隅，2000）．この表にある情報の解釈はとくに説明を要しないだろうが，たとえば以下の特徴がある．

1) 男性の若年層（24歳以下，25歳～30歳未満）では，彼女，金，時―時間，友人―仲間，地位，恋人，自分―自分自身，プライベート等が上位を占める．
2) 30歳以上になると，妻，技術，信頼，趣味，家族，信用等がある．
3) さらに，40歳～55歳未満あたりでは，家族，健康，仕事，趣味，行動，環境，社会，安定等々の身近の要素に次第に関心が移る．
4) 男性も55歳以上になると，仕事，資産，健康，いのち―生命，社会，安全，財産，平和，地域等，個人のことだけでなく，社会性のある語が混在する．
5) 女性は，若年層（24歳以下，25歳～30歳未満）に「人間関係」，「ペット―猫―犬」，「優しさ」，「主人―夫」，「友人―仲間」，「彼氏」等が現れる．
6) 30歳～40歳未満では，「子供―子供達」，「健康」，「主人―夫」，「親―両親―兄弟」，「思い出」，「幸福」，「元気」，「希望」等がある．どちらかというと自分の身近な生活観に関心があるように見える．
7) 女性40歳以上では回答数がかなり減るが，それでも「健康」，「主人―夫」，「家族」，「子供―子供達」，「親―両親―兄弟」そして「モラル」，「暮らし」，「病気」，「地球」，「国―国家」，「環境」のように，身辺の様子だけでなく社会性のある語が登場し，この点で男性に類似している．

一方，寄与しない側に表れる語も考慮することで，性年齢区分の特徴がより顕著になる．また，性年齢間の差異を同じ語の現れ方から見ることができる．たとえば「家族」は全体に共通した重要語であるが，性年齢区分別に見ると重要度の順位が層により異なる．なおここで，判定はあくまでも平均的な頻度の特徴を表したものということである．

このように見てくると，分析の方向として，性別や年齢区分だけでなく，Web調査に前向きに回答する者のプロフィル，とくに職業や所得，あるいは他の選択肢型設問との関連性を吟味することも必要であることが予想される．ここでは省略したが，実際にそのような多面的な分析から興味ある知見が得られるのである．

68.3.6 回答者と選出語の関係

a. 類型化による探査

上のように，事前情報（属性，選択肢型設問等）の区分（選択肢等）のいずれが選出した構成要素群の意味づけに有用かを知ることは，たとえば，マーケティングリサーチにおける消費者の類型化は重要な操作である．一方，事前情報なしに，クラスター化（自動分

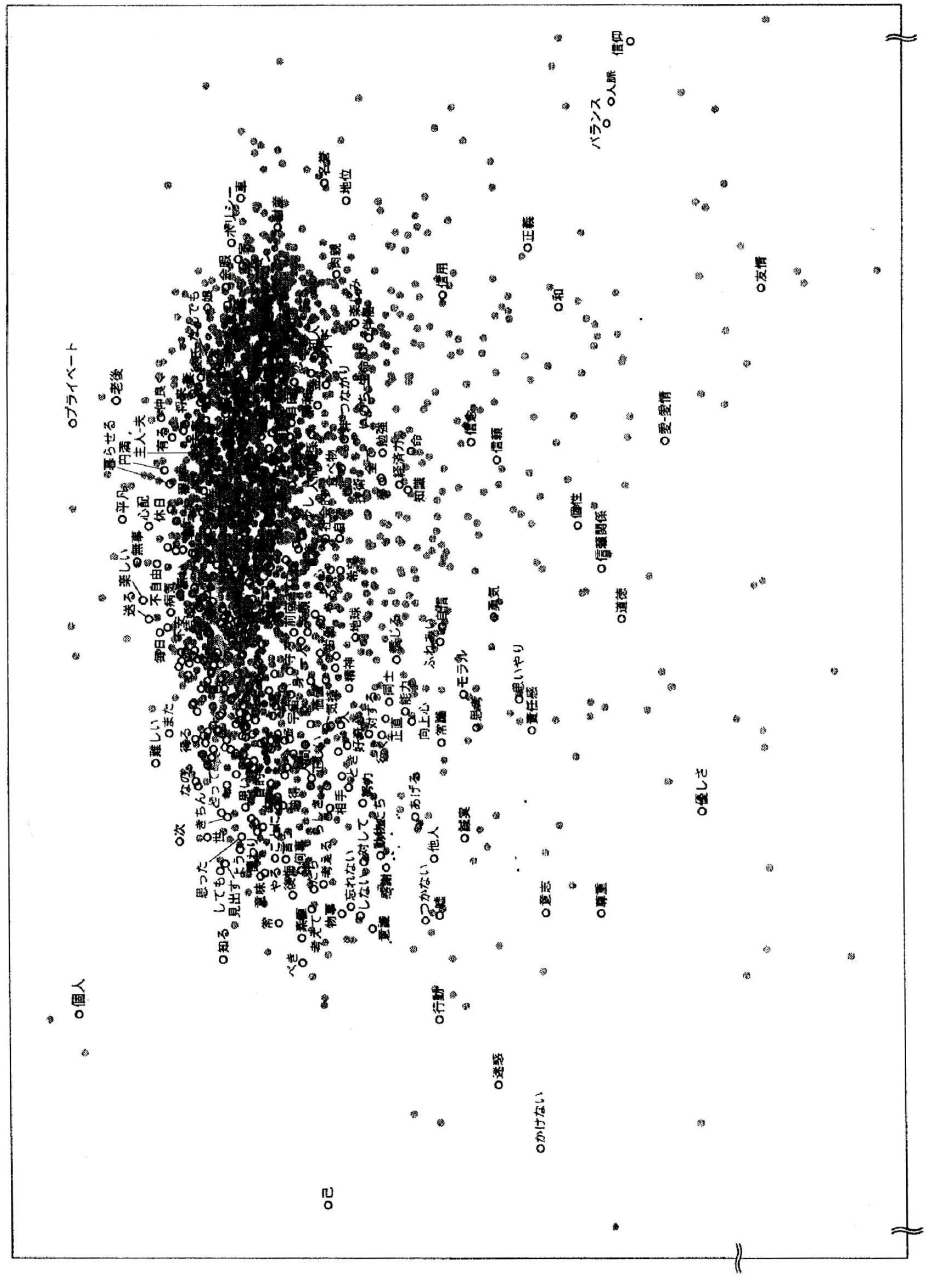

図3 回答者と構成要素の関係（同時布置図）

類）により回答者や構成要素群の類型化をおこない，潜在的にどのような特徴があるのかどうかを知ることも重要である（たとえばマーケットセグメンテーション）．これは，事前に分類基準が与えられていない場合に，回答と構成要素の出現パターンの情報を基に類似したグループを自動的に生成することである（古典的ないい方では「教師なし分類」をおこなったことになる）．さらに，必要に応じて，生成したクラスターをあらためて他の項目（選択肢型設問，属性情報等）と突合分析することも必要となる．

ここでは「抽出した構成要素群と回答者の回答パターンの関係」だけをみる．回答者によって用いる語にどんな類似や差異があるのか，あるならばその特徴は何か，といったことである．このためには構成要素，回答者の自由回答の記述の関係を類型化し対比することが目標となる．さらに回答者をクラスター化し，同時に構成要素のクラスター化や，それら相互の関係を探査し，回答者の個々の回答内容が分類で得た類型とどう関係するかを分析する．分析の出発行列は（抽出した構成要素）×（回答者）のクロス表であり，これに対応分析を適用する．構成要素，回答者の特徴は数量化スコアとして算出され，その関係が布置図として得られる（図3）．ここでは，閾値が8語以上の409語の構成要素を用いた．なお，ここで用いるクラスター化法は，筆者等が独自に開発した階層的分類法と非階層的分類法を併用する方式（ハイブリッド法）を用いている（詳細は大隅他，1994；Lebart 他）．これを用いる理由に，①大量データの分類操作が必要となること，②数量化スコアの分布の特徴として，明示的なクラスターの存在があまり期待できず，しかも外れ値が頻出すること，③したがって，こうしたスコアの分布のクセを考慮した分類法が必要とされること，等がある．

b. 回答者のパターン解析

回答者のスコアの布置図に，それぞれの自由回答の原文あるいは構成要素を表示して観察することは有効である．この操作は図を煩雑にし，視認性が悪くなるという欠点もあるが，同時に図の周辺に位置する外れ値あるいはスコアの分布の端にある回答の特徴を探査するには適している（あるいは，図の中央に位置するサンプルや語は，図の示す次元の中では平均的なパターンであるといういい方もできる）．図3がこの例である．図中の薄い灰色の点は回答者のスコアを，また構成要素を白丸で，その一部の語を表示してある．

図の右方向に「名誉」，「財産」，「地位」等，右下に「信用」，「信念」，「信頼」，左に移動して「勇気」，「思考」，「ふれあい」，「自信」，「誠実」，「常識」…，横軸左方向に「物事」，「意識」，「忘れない」，「後悔」，「素直」，…等とある．縦軸上方向には「プライベート」，「老後」，「平凡」，「無事」，「円満」，「仲良く」，「心配」，「病気」，…となる．それぞれがある意味で類似した要素と見えるが，いかにも煩雑であるし，この2次元内だけでは解釈には限界がある．

それは，出発データ行列の大きさ（次元数）が極めて大きく，また行列の各要素内の頻度が極めて疎ということがあるからである．つまり少数次元の中に射影したスコアの視認と解釈にはおのずと限界がある．そこで客観的に探索するための別のアプローチ，すなわち，クラスター別に，出現構成要素の頻度分布を有意検定し，また同時に元の回答文を，

その有意性の結果を利用して序列化して表示するという操作を用いる．

なお，クラスター数は25群としたが，その根拠は多分に探査的である．実際の分析では，①閾値を変えて解析に用いる選出構成要素数を変える，②クラスター数を変えてクラスター化生成過程を追跡する，③クラスタリング時に用いるスコアの成分数を変える（通常はかなり多数の成分数を用いる），④同時にクラスター評価基準（クラスター内分散等）の変化を追跡する等々の探査を繰り返しおこなう．

c. 回答者クラスター化情報と構成要素の関係

ここでは回答者のクラスター化で得た25群のクラスターと構成要素との関係，つまり，クラスター内の個々の回答の特徴が構成要素とどう関連するかを有意性テストにより探査する．このときの主なクラスターのサンプル構成，構成要素数（率）ほかの一覧は表8となる．

構成要素の有意性検定（出現頻度のテスト）の考え方は「ある構成要素の全出現頻度に対して，あるクラスター内に含まれた同じ構成要素の頻度が有意であるかどうかを検定する」ことである．つまり，各クラスターを特徴づける構成要素と，それとは逆にそのクラスターの説明に寄与しない構成要素とを，頻度検定をおこなって一覧として要約する（表9）．ここではそれぞれ寄与の高いあるいは低い順に20語を表とした．20語に満たないクラスターは判定基準値に満たなかった場合か，クラスター内のサンプル数（クラスターサイズ）が少ない場合である．なおここで，クラスターサイズの小さい群（15サンプル未満）は省略した．しかしクラスターサイズの小さい“外れ値的なクラスター”は，大勢から離れた特殊な意見を示すので，それの解釈も必要に応じて別途に考察することが肝要である．

表9の解釈はそれほど厄介ではない．たとえば，クラスター1に有意な構成要素として「生活，安定，ゆとり—余裕，収入，環境，社会，安全，老後…」等が現れる．一方，このクラスターの説明に寄与しない構成要素として「自分—自分自身，人，大切，親—両親—兄弟，…」等がある．他のクラスターについても同じように，その特徴を想起させるような類似した構成要素が並んでいる．

d. 用語検索と回答パターンの意味

これをより具体的に知るための探索的な工夫は可能である．その1つがコンコーダンス（用語検索機能）である．コンコーダンスは，内容分析（コンテンツアナリシス）やKWIC（Key-Word-In-Context）の基本操作で，指定した語を基準に原文を検索・ソートし出力する．コンコーダンスは原始的な操作ではあるが，重要語を知ったあとや，意味不明の語があるとき等，この機能を用いてデータ探査することは有効である．ちなみに上に出た語のうち，「友人」，「友達」を含む回答文（原文）を検索してヒットした回答文の一部を示したものが表10である．こうした探索的機能で，回答文がどのような表記がなされたかを容易に観察することができる．なお，実際の解析では，表4に見るように，「友人」，「友達」等は「友人—仲間」と置換されて用いられる．

しかし，これだけではそもそもの回答文（回答者の書いた原文）の内容の特徴までは見

表8　クラスターの構成とそれを特徴づける主な構成要素

クラスター番号	クラスターサイズ	クラスター構成比[%]	異なり構成要素数	異なり構成要素[%]	構成要素数	構成要素率[%]	解析対象構成要素数	解析対象構成要素[%]	クラスターを特徴づける主な構成要素(単語)
1	352	13.6	287	11.3	2622	15.9	128	11.6	生活，安定，ゆとり-余裕，収入，環境，社会
2	216	8.3	188	7.4	1177	7.1	86	7.8	仕事，やりがい，生き甲斐，充実感，家庭，趣味
4	24	0.9	53	2.1	111	0.7	10	0.9	やる気，能力，思考，職場，生き甲斐，楽しい
7	118	4.6	134	5.3	577	3.5	33	3.0	子供-子供達，妻，親-両親-兄弟，教育，主人-夫
8	37	1.4	64	2.5	194	1.2	13	1.2	名誉，地位，金，社会的，夢，命
9	613	23.7	191	7.5	2588	15.7	130	11.8	家族，友人-仲間，健康，金，恋人，趣味
12	236	9.1	334	13.2	2935	17.8	160	14.5	大切，大事，思う，今，何，一番，主人-夫
13	96	3.7	197	7.8	814	4.9	88	8.0	常，行動，自己，実現，努力，後悔
14	159	6.1	148	5.8	896	5.4	73	6.6	時間，自由，彼女，過ごす，時間，金，使える
15	369	14.2	299	11.8	2708	16.4	148	13.4	人，自分-自分自身，人達，支えて，周り，好きな，生き方
16	22	0.8	27	1.1	68	0.4	14	1.3	信頼関係，道徳，自信，人間関係，思いやり，人，会話
18	125	4.8	162	6.4	736	4.5	54	4.9	思いやり，優しさ，他人，責任感，金，人，誠実，気持ち
19	31	1.2	58	2.3	139	0.8	24	2.2	意志，いのち-生命，地球，自然，環境，国-国家
20	82	3.2	84	3.3	319	1.9	23	2.1	愛-愛情，友情，信頼，伴侶，金，平和，経済力
22	17	0.7	45	1.8	94	0.6	12	1.1	和，平穏，安全，達成感，程度，生活
24	26	1.0	62	2.4	172	1.0	18	1.6	迷惑，かけない，他人，人，生める-生きて，正直
合　計	2523	97.4	2333	91.9	16150	97.9	1014	92.0	

解析に用いたサンプル数は2591（人），構成要素数は409（語）

表9　回答者クラスターの構成要素の一覧（有意性テストの結果）

クラスター番号	クラスター1	クラスター8	クラスター9	クラスター18	クラスター19
クラスターサイズ	352	37	613	125	31
有意な構成要素数	128	13	130	54	24
1	生活	名誉	家族	思いやり	意志
2	安定	地位	友人-仲間	優しさ	いのち-生命
3	した	金	健康	他人	地球
4	ゆとり-余裕	社会的	金	責任感	自然
5	収入	夢	恋人	命	環境
6	環境	命	趣味	対する	国-国家
7	社会	秩序	親-両親-兄弟	心	世界
8	できる	信用	財産	人	言う
9	ある	家	親戚-親族-親類	誠実	平和
10	安全	幸福	ペット-猫-犬	気持	好奇心
11	老後	安全	家	愛-愛情	宇宙
12	充実		プライド	等	他人
13	暮らし		仕事	道徳	自由
14	精神的		パソコン	勉強	知識
15	健康		信用	自然	安全
16	暮らせる		含む	ふれあい	人間
17	程度		思い出	尊重	暮らし
18	未来		絆-つながり	知識	ない
19	安心		命	対して	支え
20	経済		食べ物	行く	あり
20	性格		これ	安定	
19	大変		やはり	やはり	
18	己		社会	いう	
17	労働		したい	やりたい	
16	理解		ない	したい	
15	いく		気持	くれる	
14	やはり		いく	大切	
13	一番		生活	仕事	
12	思いやり		した	健康	
11	人間関係		その	恋人	
10	友人-仲間		もの	会社	
9	金		ある	その	
8	愛-愛情		いう	できる	
7	くれる		次	時間	
6	親-両親-兄弟		できる	次	
5	次		くれる	です	
4	大切		です	人間関係	
3	人		する	自分-自分自身	
2	です		大切	親-両親-兄弟	
1	自分-自分自身	自分-自分自身	人	生活	仕事

表 10　コンコーダンスの例

通番	回答文の例							
1				自然体	友人	お金	家族	やりがい
2				趣味	友人			
3			自分に正直に生きること	家族	友人	健康	経済力	
4			家族の幸福	プライド	友人			
5				感動	友人			
6					友人	ポリシー	お金	便利さ
7				楽しく仕事をすること	友人			
8		信念	自信	家族	友人			
9				家族	友人	健康	そこそこのお金	
10	人間関係	自分の気持ち	仕事	家族	友人			
1				家族	友達	自分自身		
2				家族	友達			
3					友達	仕事	生活環境	趣味
4					友達	親，健康な体		
5					友達	家族		
6				家族	友達	親類		
7				好きな人と一緒にいること	友達			
8			家族がいつまでも健康で仲良く暮らせること	恋人	友達	生きがい	お金	
9		健康	仕事	お金	友達	趣味		
10					友達	楽しく仕事をすること		

一部を抜粋した，また，回答文が短い例をとりあげた．

えてこない．そこで，それぞれのクラスター内の代表的な回答（原文）がどのように分類されたかを観察する．文字情報量が膨大となるので，一例として，いくつかの特徴的なクラスターについて，回答文の一部を示した（表 11）．

一覧（表 11）の右端にある数値は構成要素有意性テストで得た語の検定値から求めた値を回答文の検定値として付与してある．この数値が大きいほど，そのクラスターに有意な構成要素が多数用いられた回答文となる．個々のクラスターを特徴づける構成要素の意味解釈は省略して，表に挙げたクラスターのいくつかについて，出力結果の回答文の特徴を読み取る（ここで，表 9，表 5，表 8 をあわせて参照）．

クラスター 1：　このクラスターに含まれる回答者数（クラスターサイズ）は 352 名（全体の約 13.6%）で「生活，安定，ゆとり一余裕，収入，…」といった語を含む身近な「暮らし感」を表す回答文が多い．

表 11　クラスター内の回答典型例

通番	クラスター 1 の回答例(原文)	検定値	クラスター 8 の回答例(原文)	検定値	クラスター 9 の回答例(原文)	検定値
1	生活	19.17	金銭 名誉，地位	10.02	家族 規律，品格，知性	23.39
2	生活 多すぎて分からない	19.17	命 名誉	8.23	家族 彼	23.39
3	安定した生活．	15.32	家族 金，地位，名誉	7.81	家族	23.39
4	安全な生活 無し	12.89	家族 お金，健康，地位，名誉，宗教	6.46	家族	23.39
5	ゆとりある生活 インターネット	12.33	健康 金，名誉，異性	6.39	家族	23.39
6	安定した生活 家族	12.20	友人 金，名誉，実力，人徳，容姿	6.04	宗教 家族	23.39
7	生活の安定とゆとり 家族	12.05	幸せを感じる瞬間 お金，地位，名誉健康	5.70	家族	23.39
8	不快感の無い生活	11.74	友人 お金，暇，社会的地位	5.37	人間性 家族	23.39
9	ゆとり	11.04	家族 友達，お金，地位，名誉，信頼	5.21	家族	23.39
10	家族 生活のゆとり	11.01	家族 家，お金，地位，名声	5.11	家族 社会学	23.39
11	家族との豊かな生活 生活レベル	11.00	家族の幸せ 金・名誉・人望・酒の強さ	4.54	イエス・キリスト 家族	23.39
12	穏やかにすごす生活 家族	11.00	命 お金	4.07	家族	23.39
13	安定した暮らし	10.96	人間関係(家族含む) お金，名誉	4.07	家族	23.39
14	家族 生活環境	10.43	自分の命 恋人，友人，親，自尊心，名誉，地位，金	3.92	家族	23.39
15	思いやり 安定した生活，環境保護，	10.18	家族，名誉，お金，将来，人並みの生活	3.86	家族	23.39
16	ゆとり 環境	10.17	達成感 名誉，収入，信望，健康	3.85	家族	23.39
17	健康 生活の安定安心して暮らせる社会	9.83	愛です 信用・お金・社会的地位・向上心・向学心・親孝行・夢	3.48	家族	23.39
18	心の充足 ゆとりのある生活，健康，社会の安定	9.75	家族及び自分の幸せ 努力の結果として，収入増，地位，名誉等，有形及び無形の財産	3.47	家族	23.39
19	夫婦 政治，社会生活	9.75	家族 友人，お金，家，名誉，趣味	3.43	家族 友達	21.67
20	現在の生活をどうして維してゆくか，家族，充実した生活	9.58	命	3.24	家族 仲間	21.67
21	老後の保障 豊かな生活	9.50	ネコ 命	3.24	家族 友達	21.67
22	健康 安住，世間並みの生活の出来る収入	9.46	家族 友人，信用，財産，名誉，秩序，道徳	3.22	家族 友人	21.67
23	精神的な豊かさ 快適な生活経済的安定	9.28	家族 知識，友達，安全，休暇，お金，衣，食，住，名誉（肩書）	3.21	家族や友人	21.67
24	ひびの暮らしの安定 老後のこと	9.24	なんでも話せる友達 お金 地位 彼氏 パソコン 家	3.14	家族 友達	21.67
25	のんびりとした生活 特に無い	9.06	子供が幸せになってくれること 健康であること，お金を持っていること，地位，名誉があること	3.00	友達 家族	21.67

クラスター 8：　このクラスターは，サイズは小さいが（37 名，約 1.4%），回答文には極めて顕著な特徴がある．「名誉，金，地位，社会的…」といった語を含む，特別な意

表 11　つづき

通番	クラスター 18 の回答例(原文)	検定値	クラスター 19 の回答例(原文)	検定値
1	思いやり	19.61	地球環境 自然	5.08
2	おもいやり 言葉	19.61	志 自然との一体感	4.97
3	マイペース 思いやりの情	19.61	仕事場の環境・周辺環境(都心からの距離)	4.88
4	思いやり 礼儀	19.61	家族 地球・自然・国家	3.75
5	思いやり	19.61	平和	3.48
6	やさしさ おもいやり	14.42	生命 自由，理想，楽しみ	3.46
7	命，思いやり	12.87	家族 自然・地球	3.45
8	人に頼らない・頼られない思いやり	12.34	家族 生命	3.37
9	思いやり やさしさ，尊重	10.47	自分の意志 他人の意思	3.32
10	家族 思いやり，協調性	9.81	人生 自然 地球 宇宙 人間 世界	3.20
11	家族 思いやり	9.81	強い意志 暴力のない暮らし，もっと，理性をおさえること．	3.14
12	家族 思いやり	9.81	平和を追求すること．地球環境 貧困や飢え食料問題	3.04
13	社会のルール 思いやり	9.81	Sence of wonder(好奇心) Sence of humour	2.93
14	家族 思いやり	9.81	生命 家族，財産，環境	2.90
15	強さ 優しさ，運の良さ	9.22	国家，家族・家庭・生命	2.85
16	思いやり，友人．	9.19	空 自分の世界，他人の世界，緑	2.82
17	思いやり 健康	8.87	家族 住む環境，広く考えれば地球であり宇宙	2.66
18	思いやり 健康	8.87	地球環境の保護 戦争のない，民族共存が出来る状況を作り上げること．	2.66
19	前向きな心 思いやり，いたわり，まごころ，など	8.34	命を引き継ぐ，自分の子どもたち，生命をはぐくむ，この地球の環境．	2.61
20	子供 思いやりの心	8.34	信頼感の持てる人づきあい，地球，自尊心，平和	2.54
21	命の尊さ，家族，他人へ思いやり	8.30	自然 お金	2.48
22	人への思いやり．家族	8.23	お互いの触れ合い 心のこもった(俗に言うハートフル)対話，応対．	2.47
23	家族の輪 人への思いやり	8.23	自分と家族の幸福 地球環境，世界平和，自由，豊かな暮らし	2.34
24	家族との触れ合い，他人に対する思いやり．共生．	8.21	生きようと思う強い意志 他人を思い遣る心 自分の気持意思 知識寛容さ謙虚さ聞く耳	2.31
25	人を思いやる心 お金	7.52	自分の意志 精神的支えとなる人間好奇心・知的欲求 国家の安全親食	2.30

検定値は実際に解析に用いた構成要素に対して付与されるものであり，ここでは原文を表記したので，記述内容が異なるのに同じ検定値がある．
原文は，ここで用いた 2 つの設問の併合となっている．

表 12　回答文があいまいな例

クラスター 12 の回答文の一部（原文）
●大切なものを守り通すこと．
●大切なものといっしょに暮らすこと．婚約者．
●孤独でないこと　気持ち良く暮らせること
●今は両親です．特に死に関わる病気を思っている母ですね．
●思いやりと，それを実行する行動力．今，大切だと思うのは「エコな心構え」です．便利・快適・早いなどを理由に使ってきたもの，それは環境えの影響を考えても，まだ使わなくてはいけないものか？？？と，思うのです．
●子供です子供できる前は，何も怖いものなどなかった…死ぬのも子供がよろこぶ顔を見られるのがシアワセです　お金　今の時代これがなくては生きていけませんお金持ちまでは　なろうとは思いませんが…まったくなければ　困る物です
●今のままでよい．休みの時は，大事にすること．
●向上心だと思います．とりたててありませんが，安穏とした日々といったところです．
●まだ夫婦 2 人暮らしですが，やはり家庭が一番大切です．もちろんお金があったら，あっただけいいですね．あとは，人間同志の信頼関係が必要で大切かな．
●時間　何もないような気もするし，何もかもが大切な気もする．
●家族です．今の暮らしを大事にしたいと思います．
●人と人との，コミニュケーションを大切にすることです．いくら，機械化が進んでも，いちばん重要なことと思います．健康です．
●現在妊娠中なので子供のことが一番大切．主人の健康手に入れたばかりのマイホーム
●やはり，子供です．子供との生活，子供の将来などが一番大切です．でも自分自身のことも大切です．親です．父親はすでに他界しているのですが，自分が親になって母親の存在が本当に大切だと思ってます．
●人です．どんなにお給料のいい会社でも人間関係が悪いと苦痛で，多少悪くても人間関係がいいと続いたりします．今の住んでいる所は特にいいところだとも思いませんが，周りに意地の悪い人などいなく，とても快適です．お金も大事です．貧乏は気持がすさみます．適当にあったほうがいいですね．
●生き甲斐　人々の幸せ，笑顔，喜び，生き甲斐のある生活．スペースシャトルが撮影したハイビジョンカメラから眺める美しい地球の地形美．地球の美しい景観デザインはかけがえのない大切なものと感じています．こんなに美しい感動的な創造は誰にも真似できない大切なものです．
●何か難しいことをやり遂げた達成感が重要だと思います．自分の趣味の世界を大事にすることも大切だと思います．
●時間でしょうねぇ一瞬一瞬を大事にしようと思っています．前向きな気持ちとか…
●自分の時間を，大事にし有効に使うこと　※ストレス解消のドライブ※レース観戦※ボーとすること以上の 3 点です
●人間としての美学誇りというほどたいしたものではないがプライドという言葉が持つほど高飛車ではない生きていく上での指針．自分を確立すること．あとは欲しいものが買える程度の金．あればあるにこしたことはない．
●生きていて良かった…と死ぬとき思えたらいい人生を送りたいですね．大切なのは，人間関係です．家族．生きがい．

見を持つグループである．同時に「家族，友人，子供」等の語も挙げているが，どちらかというと，権益志向に関心の高いグループである．

クラスター 9：　クラスターサイズが 613 名（約 23.7%）と，もっとも大きく，全体の

4分の1を占めるグループである．また，現れる語は，表5の一覧情報において上位に高頻度で現れる語を多数含むグループである．つまり，平均的な回答者像を代表するグループと解釈される．「家族，友人—仲間，健康，金，趣味，親—両親—兄弟，財産，…」等が含まれ，クラスター7に類似するが（表8），ここでは子供，妻，主人—夫，夫婦といった語は登場しない．

クラスター18：「思いやり，優しさ，いたわり，命，共生，コミュニケーション，…」等，人とのつながり，他人や家族への対人的な意識との間で見られる「気持ち，感情」を重要とみるグループである．生活，人間関係，自分自身，親—両親—兄弟…，等の語が薄いグループでもある．回答者数は125名（約4.8%）程度を占めている．

クラスター19：設問の「大切なもの」の意味をまったく別の視点から考える回答者群である．主要語は「意志，いのち—生命，地球，自然，環境，国家，世界，平和，…」等であり，実際に回答文も表11にあるように，地球環境，自然，周辺環境，国家，人間世界，平和の追求，といった社会性の高い意見を述べている．回答数は少ないが（32名，約1.2%），特徴あるクラスターの1つである．

このほかのクラスターについては，表8に「クラスターを特徴づける主な構成要素」として欄に記したが，これを見るだけでも個々のクラスターの特徴が想起できる．このように見てくると，類型化操作と有意性テストによる利用語とそれを含む回答文の検定結果の探索的な観察によって，分析対象としたデータの中に潜在する構造の特徴抽出がそうおかしなものではないことがわかってくる．

e．特徴があまり明らかでないクラスターの解釈

一方，クラスター12, 15等のように（表8），サイズは大きいのであるが回答内容の解釈がやや難しいグループもある．これらのクラスターの特徴を少し吟味する．まず，顕著な特徴として，これらのクラスターは，表8にあるように構成要素数（率）や異なり構成要素数（率），あるいは実際に解析に用いた構成要素数（率）が，クラスターサイズに比して大きいことがある．つまり，他のクラスターに比べて，内容の表現が多岐にわたる（発散している）ことが予想される．実際に表12にクラスター12の回答文（原文）の一部を示した．これを見ると，「大切」，「両親，親」，「子供」等の他のクラスターで説明力のある語も含まれてはいるが，総じて文章が長めの上，記述内容がさまざまな方向に広がり，何かに特化して記述していないことがうかがえる．自由回答である限りこうしたクラスターがあることは自然なことであり，設問「大切なものは？」に対する姿勢（回答行動）が，他のクラスターにあるように「家族，友人」等の単純な表現では表しきれないと考える人が存在することを示唆している．こうした現象は他の事例解析でもたびたび遭遇することで，したがって一度の分類操作で得た類型群のいくつかをあらためて取り出して，さらに詳細な分類をおこなう「再分類（であり細分類）」の操作が必要とされるだけでなく，同時に設問内容や調査票の設計方法との関連研究が必要とされる．

ここで解析に用いたデータセットは，ここ数年にわたり産学協同プロジェクトとして進めてきた実験調査の成果の一部である．個々の社名は省略するが，ご協力いただいた各調査機関に，紙面をもって謝意を表したい．［大隅　昇・Ludovic Lebart］

参考文献

林　知己夫（2000）．これからの国民性研究 — 人間研究の立場と地域研究・国際研究から計量的文明論の構築へ．統計数理，**48**(1)，33-66.

加賀野井秀一（1999）．日本語の復権（講談社現代新書1459）．講談社.

加賀野井秀一（1995）．20世紀言語学入門（講談社現代新書1248）．講談社.

小池清治(編)（1997）．日本語キーワード事典．朝倉書店.

Lebart, L., Salem, A. & Berry, L.（1998）. *Exploring Textual Data*, Kluwer Academic Publishers.

村上征勝（1994）．真贋の科学 — 計量文献学入門（行動計量学シリーズ6）．朝倉書店.

西本一志・角　康之・門林理恵子・間瀬健二・中津良平（1998）．マルチエージェントによるグループ思考支援．電子情報通信学会論文集，D-I, **J81-D-I**(5)，478-487.

長尾真編（1996）．自然言語処理（岩波講座「ソフトウェア科学」第15巻）．岩波書店.

大隅　昇，Lebart., L., Morineau, A., Warwick, K.M., 馬場康維(1994)．記述的多変量解析法．日科技連出版.

大隅　昇（2000 a）．「調査環境の変化に対応した新たな調査法の研究」報告書．文部省科学研究費特定領域研究，ミクロ統計データ，公募研究（研究課題番号：09206117）.

大隅　昇（2000 b）．調査における自由回答データの解析 — InfoMinerによる探索的テキスト型データ解析．統計数理，**48**(2)，339-376.

Ohsumi, N.（2000）. From Data Analysis to Data Science. In H.A.L. Kiers *et al.* (Eds.), *Data Analysis, Classification, and Related Methods* (pp. 329-334), Springer-Verlag.

大隅　昇（2001）．電子調査，その周辺の話題 — 電子的データ取得法の現状と問題点．統計数理，**49**(1)，199-211.

渡部　勇，三木和男，新田　清，杉山公造（1995）．ハイブリッド発想支援システム：HIPS，計測自動制御学会第17回システム工学部会研究会資料.

全文検索システム協議会編（1999）．全文検索システムとは何か？，第1部，1-63.

第10次日本人の国民性調査委員会編，国民性の研究第10次全国調査．1998年全国調査，統計数理研究所研究リポート83（1999）.