統計数理 (2000)
第 48 巻 第 2 号 339–376

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—

統計数理研究所　大隅　昇
ENST (École Nationale Supérieure des Télécommunications)*　Ludovic Lebart

(受付 2000 年 7 月 6 日；改訂 2000 年 9 月 26 日)

要　旨

日本語の電子的処理が可能となったことや，言語情報処理分野の諸研究が進んだことから，テキスト型あるいは文章型データの取得法や解析手法への関心が高まっている．とくに，社会調査や意識調査・態度調査，あるいは市場調査等の各種調査における自由回答・自由記述データの取得方法や取得後の統計的データ解析の具体的な方法論の登場が期待されている．本報告では，初めに，調査分野における筆者等の経験に基づき，自由回答データ取得において見られる諸問題やその取得方法のあり方について述べる．次に，従来の日本語文章・テキストの解析方法の方向とここで主張する統計的データ解析との関係について議論する．また，我々の主張を具現化したテキスト型データ解析システム InfoMiner with WinAiBASE (あるいは InfoMiner と略す) の主な特徴を紹介する．InfoMiner は，日本語処理で必要となる分かち書き処理機能，キーワード抽出機能，それらの編集機能，さらに多次元データ解析機能 (対応分析，クラスター化等) を含む独自に開発されたテキスト型データ解析システムである．さらに，データ科学の観点にたった独自の調査計画に基づき実施されたインターネット調査 (ここでは Web 調査) 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，InfoMiner やそれに含まれる多次元データ解析手法の利用可能性や適用性への事例とする．

キーワード：自由回答の分析，テキスト型データ解析，InfoMiner，インターネット調査，形態素解析，分かち書き処理，テキスト・マイニング，データ科学．

0.　問題の背景と研究の目的

ここに報告する内容は，従来の言語学や言語情報処理における研究，あるいはそれらの延長線上にある種々の方法論におけるアプローチとは若干視点を変えた方向で考えようという意図がある．このような立場をとることの適否については，多々議論があろうが，我々がここで主張することは，ある探査的統計データ解析手法を日本語のテキスト型データの解析に取り入れたとき，どこまで分析が可能であるのか，どこに問題があるのかを実験検証的に進め，また我々が主張するデータ科学の方向で解決を図るための一つの実証研究と

*Département Economie, Gestion, Sciences Sociales et Humaines: 46, rue Barrault, 75634 Paris Cedex 13, France.

することである．

後述のように，各種の調査の実査環境が大きく変化する中で，定性調査や定性的アプローチで取得した日本語文字情報あるいは文章型テキスト・データの取得法や解析方法を求める声が高い．こうした状況を背景として，筆者等は，主にマーケティング・リサーチの分野において様々なタイプのテキスト型データの解析を体験してきた．こうした経験や実験調査を通じて得た知識に基づいて，本報告では以下のような事項についての検討と報告を試みる．

(1) 各種調査における自由回答データ取得方法のあり方．
(2) 日本語言語情報処理との関連，日本語データ解析をどう考えるか．
(3) 日本語の特性を活かした本報告で述べるアプローチの特徴を述べること．
(4) テキスト型データ解析システム: InfoMiner の設計指針を示すこと．
(5) InfoMiner の事例紹介として，Web 調査で取得した自由回答の分析の一部を示すこと．

山本夏彦は近著「完本文語文」の中で，口で語って耳で分るのが言葉である，と述べ，また近年の日本語の乱れを指摘している．また最近の新聞紙面，雑誌等でも「日本語のあり方」が改めて (あるいは今もってと言うべきか)，あれこれと議論されている．国語審議会における議論や多くの研究報告を待つまでもなく，日本語自体が言語として完成された形にあるものではない．むしろ，言語は流動的にたえず変化するものであり，またそれであるからこそ言語であるとの指摘も多々ある．話題を少し絞り込んで，いわゆる調査 (態度，意見等) あるいはアンケート調査と総称される分野で用いられる自由回答方式 (open-ended answer: OA，free answer: FA 等) の調査に限って考えてみても，定性的アプローチの決定的な方法論があるわけでなく，未だ模索構築の途上にあると考えられる．

1. はしがき —定性調査における自由回答の役割—

1.1 定性調査

調査環境の急速な変化，とくにその環境悪化が言われてから久しい．調査の質がもっとも問われているこの時代にあって，様々の原因から満足のゆく内容の調査の実施がきわめて困難になってきたとされる．とくに従来からの定量的調査の実施の困難性や様々な問題の提起，例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．そして新たな調査法の登場や，従来法の見直し，例えばクォータ・サンプリング，郵送法，電話調査等のあり方が改めて問われている．

こうしたことは，従来とは異なる質的調査への関心の高まりという現象として現れている．サンプル数が大きく，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行われる量的な調査 (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，郵送調査等) 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ることから (例えば回収率の低下)，質的調査や定性調査に関心が移行する傾向が見られる．

こうした傾向の一つとして，マーケティング・リサーチの分野等では，グループ・インタビュー (GI) あるいはフォーカス・グループ，モチベーション・リサーチなどが改めて注目されている．また，少数のサンプルや，条件を限定した回答者を相手に，インターネット調査 (とくに Web 調査) などで，自由回答や自由記述の設問を多用し，ここで取得したデータの質的解析を試みる等の例が多くなってきた．

インターネット環境の普及により，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，文章型データ，とくに自由回答に代表されるテキスト型データの取得が，内容の質の適否に関わりなく，容易に，しかも大量取得が可能となったことから，これを多用する調査 (とくに消費者動向調査，Web マーケティング) が多くなった．

同時に，こうした自由回答・自由記述のデータを解析するための方法論の研究も見られるようになった．またこれとは別に，従来からの定性調査手法として，面接法 (深層，集団など)，投影法 (言語連想，文章完成，略画完成，絵画解釈) 等があったが，こうした方法でも，取得データが電子化されてテキスト型データとしての利用が容易になってきた．このほか，CRM (Customer Relationship Management) との関連で，企業のコールセンターや顧客相談窓口における取得データの定性解析など多種多様な試みがあり，また具体的方法論の開発への期待も高い．このように，今後は，調査環境の多様化に伴う，文章型・文字型によるデータ取得や解析の機会の増大が考えられる．

1.2 通常の調査法の条件と特徴

ところで，従来型の調査あるいは一般的な調査 (面接法や留置法) で，たとえば質問紙調査票による選択肢型設問を用いた調査法による調査実施における要件として，妥当性，信頼性，客観性，再現性等の保証が挙げられる．一方，定性情報データ，とくに自由記述文・テキスト型データの取得・分析では，これらの保証が得られにくいとされてきた．たとえば，再現性一つを考えてみても，これをどう扱うのかがあまり明らかではない．

しかし，自由回答方式による質問形式を定性情報取得の有力手段として確立するには，自由回答の取得環境，調査設計・抽出，妥当性をどう考えるか，という調査法としての本質的問題の検討からが重要なことは言うまでもない．同時に，現時点における研究成果の未成熟から，自由回答のみから得られる情報やその分析処理にも限界があることも自明である．そこでここでは，下記の観点から研究を進めることを考えた．

① 当面の課題は，日本語の精密な内容解析や意味解析の解答 (結果) を得るという事ではない．
② 自由回答の取得において，従来の選択肢型設問や属性データの定量的分析との比較検証方法の確立が必要かつ重要な操作となる．
③ しかし，細かい初動探査，基礎集計，コーディング技術が必須となる．
④ 完全な自動化を目指すことには未だ無理がある．
⑤ しかし，何を行って何を得たかの解析手順の具体的かつ明示的な提案が必要となる，つまり少なくとも同じ内容がいつでも追試再現実験できること．

とくに，自由回答データの分析だけでは意味解釈が恣意的となりやすいことは自明であるから，従来の調査手法に関連した方法論の援用，支援も必要となる．たとえば，次のような理由から従来の調査技術を併用することが必要となる．

⑥ 母集団とサンプルの関係，サンプリングという概念の設定が難しい．
⑦ 当然ながら，回答比率といった定量的な評価ができない．
⑧ 調査の再現性や客観性が問題とされる．

一方データ収集の技術面に目を向けると，データの性格が多様化し，扱うタイプも非数値的・質的，量的と様々である．また，膨大な量の情報の電子化，データベース技術の進歩に伴い，メタ・アナリシス (集積情報の横断的相互利用等)，データアーカイヴ，マイクロ・リンケージ，そしてデータウェアハウス等の関連技法が登場してきた．これら技術的

要素を背景にして，急速にデータ・マイニングの方法論が登場し，この中にも多数のテキスト・マイニング手法が登場している．コンピュータの処理能力に期待し，データベース上の大量データ処理を通じて，知識の組織化や知識発見を図る構想であるが，かけ声程に見合った高い成果が上がっているかは，今後を見ないと即断はできない．

2. 日本語文章・テキストの解析の方向

2.1 通常の解析方法として何があるか？

ここで，本報告の位置づけを明らかにする意味から，従来から行われてきた諸研究の方向を，筆者の観点から以下のように要約しておく．

(1) 自然言語処理

ここでは，文章を形態素解析による要素分解に始まり，品詞に分解，分類する操作，さらに文脈解析，構文解析，意味解析・意味理解，内容分析 (コンテンツ・アナリシス) などを行う．これをデータベースと併用して，いわゆる全文検索やキーワード検索を行ったり，あるいは個々の成果要素が音声認識，自動翻訳，その他の言語情報処理技法へと広く応用展開されてきた．

(2) 計量言語学的な研究

一方，日本語を構造的に見て，その文法や用法の面から考察し，語彙や語句等の計量的分析，単語，各種品詞の使用率，頻度分布の検討，語彙分布の法則性の探査を行う (例: ジフ分布，パレート分布，大野の法則，樺島の法則など)．さらには語彙調査，シソーラス・分類語彙表の作成などを計量的に行うなどの方法がある．

(3) 計量文献学的アプローチ

また，古典・古文や小説等の著者推定や，文体の研究などを行う方向がある．ここでも，計量的な評価方式として，助詞，語句長，句読点等の統計分布，品詞分類やその頻度分布，などの計量的方法が用いられる．また，多変量解析手法 (因子分析，数量化法，判別分析法など) も多用されてきた．例えば，村上 (1994) による精力的な古典文学の分析研究とその成果にみられるような方向である．

(4) 言語情報処理，全文検索他

また，形態素解析や検索ツール，Web ブラウザ等の進歩により，Web ページやデータベース上の集積情報としてのテキスト型データから，全文検索やキーワード検索を効率的に行ったり，検索結果を効果的に表示したり内容を要約化して理解を容易にするような多数のツールが登場している．たとえば，ワードプロセッサに組み入れられた文章要約機能や，その発展型と見られるテキストデータ検索要約ソフトがある．具体的には，ConceptBase (CB: ジャストシステム) やその関連ツールである CB Classifier, CB Summarizer がある．また米国のソフトを日本語対応とした VextSearch (コマツソフト) 等もある．インターネット上での検索結果の分類配信システム，知識データベース化技術等の応用も盛んである．

(5) 人工知能的なツールの開発

言語解析研究の中の人工知能的研究を基盤にして，いわゆる「発想支援ソフト」との関連で,「デファクト」(電通・富士通) やその元となった HIPS (Hybrid Idea Processing System,

図 1. 自由回答解析の位置付け —— 概念図 ——.

富士通研究所)，AIDE (Augmented Informative Discussion Environment, ATR) などがある．この種のソフトの一部は既に商用化されている．

なお，上記の (4), (5) に挙げたソフトウェアに共通のこととして，商品としての技術要素の守秘事項や公開できないノウハウ等の制約から，ユーザがその処理内容を正確に理解できないことがある．この他，いわゆるナレッジ・マネジメントに関連して，テキスト・マイニングをデータマイニング・ツールの中の機能として組み入れる傾向がある．このように，研究対象の多様化に限らず様々のソフトも登場しており，ここでの分類はあくまでも筆者の私的な観点から行ったものであるが，本報告の議論の位置づけを知るには十分と考える．

また，自由回答・自由記述文に関連した研究と，以上に要約した研究分野の相互の関連と位置づけを，筆者の視点から表したものが図 1 である．ここにみるように，調査における自由回答の取得法や取得データの解析は，従来の調査法との接合面を保持しながら，また従来の言語情報処理の諸研究とは異なる方向から，実験・実証的に検討すべきことと考えている．

2.2 テキスト型データ解析上の留意点

以上を事前情報として，調査における自由回答文の解析を考えるとき，以下の特徴・事項に留意した客観的な方法論が必要と考えられる．まず，調査における自由回答文・自由記述文の取得にあたって，従来から次のようなことが指摘されてきた．

① 考えたことがないことには答えにくい，いきなり設問を受けても答えることが難しい，いわば「白紙を出されて何かを書くように」といわれてもなかなか思いつかない．
② 予想しなかった回答や知見が得られるという期待がある．
③ 無記入が多くなる傾向があるとされてきた．
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言われている (妥当性の問題)．
⑤ 通常の他の設問 (選択肢型) の選択肢の影響を受けるのではないか (回答誘導の懸念)．

⑥ 適切なデータ解析法がないと言われてきた.
⑦ 客観的な統計解析手法の確立が難しいとされる.
⑧ 内容の再現性がない，あるいは信頼性に欠ける.
⑨ 集計の手間がかかると考えられている.
⑩ 回答に均一性を欠くと言われている.
⑪ 設問文の意図がどう反映されたか，自由回答のみからは読みとり難い.

こうした指摘はもっともであり，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，選択肢設問型調査と異なり，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．たとえば，回答比率で比較する，標本誤差を推定するなどの操作に相当する方法が未だない．また，標本抽出との関連や調査票や設問の設計をどう考えるか，さらには，調査の反復可能性や再現性の問題をどう捉えるか (通常の回答変動とは異なる事象が現れるであろう) 等々，様々の検討要件がある.

一方，今まで述べたように，従来型の調査の実施環境の急激な変化から，とくに実査の困難性が指摘される中で，従来とは異なる意味で，定性調査への期待が高い．とくに，インターネット調査等の電子的情報取得手段の普及により自由回答の電子的取得がきわめて容易となった事から，その利用法の可能性の十分な検証のないままに，急速に利用されるようになっている．こうした事情までを考慮して，筆者等は自由回答の取得法や分析法の研究，さらにはその解析システムの開発を進めるには以下のような対応が必要と考えてきた.

(1)“ 実験調査 ” による事例検証の積み重ねが必要であること.
(2) テキスト型データの異なるタイプの事例検証が重要と思われること.
(3) 現場の要請 (例えばマーケティング・リサーチにおける利用) を整理し，システム開発に必要な要件の要因分析を行うこと.

とくに，定性調査を定量調査との優劣を比較するという観点から捉えるのではなく，また従来は定性調査を定量調査とは別の視点で捉えるという傾向もあったが，これを改めて，ここでは，調査における“ 自由回答データ ” にもとづく定性情報の取得において，従来の選択肢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議論を進めることとする．つまり，従来の選択肢設問形式を，蓄積のある調査手法に関連した方法論に裏付けされた定量的評価に用い，一方，自由回答形式はその定性情報の計量化を図る方法論を新たに考え，それら両者の併用を工夫することで，より客観的な情報要約を図る方向からデータ解析を考える．また，テキスト型データ解析の統計システム開発に際しても，こうした発想が反映された設計指針で検討する．これは，林や大隅が主張してきた「データ科学」の具体的な実例とする意味をも含んでいる (Hayashi (1998)，林 (2000)，Ohsumi (2000)).

2.3 文章型・テキスト型データの解析の方向と適用範囲

文章・テキスト型データが現れる場面は様々であり，その利用分野は多彩である．ここでは，筆者が今までの研究の中で何らかの形で接点のあった主要な課題とその適用場面を拾い出して列記した.

[文章型・テキスト型データの利用場面]
① 一般的なアンケート調査の中で自由回答形式を設けて取得する
　適用場面：消費者行動調査分析，自由回答方式の研究
② 発想法，KJ 法等の文字データ解析

適用場面：商品モニターによる意見・評価データの類型化，相互関連性の検証
③ 電子調査での取得データ
適用場面：インターネット調査 (Web 調査，電子メール調査等)
④ Web ページによる製品ユーザの意見聴取，電子メールによるモニタ回答収集
⑤ 製品に添付の意見葉書の自由回答
⑥ コールセンターで収集した聞き取りデータ，その電子化情報
⑦ 集団面接法，グループ・インタビュー等による記録データの解析 (記録や録音をテキストとして書き起こしてこれを解析)
⑧ 医学分野での応用
適用場面：診断カルテの分析 (電子カルテ等)，歯科矯正治療に関する患者−医師双方の自由回答意見の比較，患者の聞き取り追跡調査
⑨ 一般書籍の文章解析，小説・文芸作品，新聞・雑誌記事等の分析
⑩ 小中学校における論述形式問題の評価分析
⑪ TV ドラマの番組紹介記事説明文 (あらすじの解説文) の分析
⑫ 野外調査で取得した観察記録データの分析
適用場面：エゾ鹿の行動観察記録等
⑬ グループインタビュー取得データの分析
⑭ インターネット上でのチャット，電子掲示板上の対話データ等の分析

これらの課題の多くは，従来の自然言語処理，言語情報処理技法だけではカバーできないことは明らかであり，従って別の観点からのアプローチが必要となる．このほか，最近の傾向として，データベース検索との関連で，文書，書籍等の電子化の加速化やオンデマンド出版，多様なテキスト型データの蓄積，テキスト・データベースの普及に伴うテキスト・コーパスの利用環境の変化，従来は分析を諦めていた大量の文字型データの電子化 (の実現容易性)，これらのコンピュータ処理の可能性の増大 (全文検索，キーワード検索などのツール) 等がある．

2.4 テキスト型データの取得段階の類型化

ところで，上に例示でみたように，テキスト型・自由記述文型データの分析技法が求められる場面は非常に多様化している．こうした適用場面から，経験的に取得データの様相を以下のように要約した．

(1) 単に集めただけのテキスト・データ
サンプル・調査対象の背景やデータ取得状況があまり明らかでないデータ，例えば調査票の設問の最後に「何かご意見を自由に」式の取得情報，電子掲示板やチャット等の蓄積情報．
(2) 元来が文字情報であるとき
古典，文学書・文芸書，新聞・雑誌類，各種の記録文書等，比較的文体が整った文字情報として提供される場合．
(3) 過去の蓄積データの見直し・再評価
“ 再発掘 ” 等の過程を経て取得したデータ，付帯情報やデータ取得履歴が整理可能なデータ，蓄積した定性情報データベース等をいう．
(4) 通常の定量型調査との併用
選択肢型設問等に組み入れられた，あるいは組み合わせた自由回答，もっとも多いと思われるタイプである．
(5) 計画的に設計された環境下での収集データ

テキスト型データの取得を主目的とした場合，自由回答取得を主目的として設計された調査，特定の商品ユーザの日記形式調査，モニター制による追跡調査等での自由記述文等.

3. データ処理，解析上の課題

3.1 日本語の特徴と形態素解析

言語類型学的には，多くの場合言語を，孤立語 (語形変化せずに，文法関係が語順で示される言語)，膠着語 (文法関係が助辞や接辞によって示される言語)，屈折語 (文法関係が語形変化によって示される語) と分類する．これに従うと日本語は「膠着語(膠着言語)」である．また，主な欧米語は屈折語である．膠着語の特徴は「自立語」と「付属語」が膠着していることにある.「付属語」(あるいは「辞」)とは，助詞，助動詞，接尾辞，用言の活用語尾をいう．また,「自立語」(あるいは「詞」)とは「付属語」に付くものをいう．これはいわゆる「詞−辞」構造という (時枝誠記による「詞−辞」論) (加賀野井 (1995, 1999)).

このように，日本語が欧米の言語と大きく異なることの一つに，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．もっとも，中国語など，アジア圏で利用される言語にもべた書きという共通した特徴がある.

分かち書きされていないということの他に，日本語は漢字，カタカナ，ひらがな，それに外来語 (や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．また，古くは万葉仮名，カタカナの誕生，漢字の読み替え・当て字による造語，そして明治期 (以降) における造語現象 (主として日本にはなかった諸概念の表現のための新造語，例えば社会，自由，経済，生産，科学，真理等々多くの現代用語)，明治以降の言文一致体の登場等々，歴史的にみてもきわめて流動的である.

また，欧米語が「単語」という単位で仕切られた言語であることから，その処理系として単語を単位として扱うことができ，結果として個々の単語を抽出できるという容易性がある．一方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.

このために，日本語の処理を行うには，まずある単位に文章を分解する「分かち書き処理」が必要となる．これを含めて幾つかのコーディング処理を形態素解析 (morpho-logical analysis) という．形態素 (morpheme, morphology) とは，表記された文章 (日本語とは限らない) を「最小の有意義な意味ある単位，意味を持つ最小の単位」と定義されている (池上 (1993))．これは池上によると Bloomfield によって唱えられた概念であるという．また，長尾 編 (1996) によると，形態素とは「単語や接辞など，文法上，最小の単位となる要素のこと」としている．ここで，両者には明らかに考え方 (解釈) にわずかな相違がある．いずれにせよ，形態素とは絶対的な概念ではなく，あくまでも一つの便宜的な約束事である．たとえば，池上によれば，形態素が示す矛盾を説明する概念として「語彙素」(lexeme) を挙げて両者の特徴を指摘している．いずれにしても，形態素，語彙素ともに「語」(word) とは必ずしも一致はしないし，ここにみるように言語学的にも確定的な概念があるわけではない.

しかしながら，日本語の処理，とくにデータ解析処理においては，何らかの意味である単位に分けねばならない．そこで，通常は分かち書き処理で得た単位要素の候補を辞書 (形態素辞書) と照合し，次にそれを解釈可能な候補に絞り込み，文字・文章の文法的な接続関係 (word connection) を検証する．そのうえで，その分かち書き単位についての品詞の同定を行い，続いて辞書にはない語の処理を行う等の手当をする．このような一連の過程が「形態素解析」である．従ってその処理にはかなり発見的あるいは経験的な要素が含まれるこ

ととなり，実際にいろいろな解析方式が提案されてきた．たとえば，最長一致法，字種区切り法，文節数最小法，接続規則法等がある (全文検索システム協議会編 (1999))．いずれにせよ，日本語処理の初めの処理過程として「分かち書き処理」と「辞書照合」の操作が不可欠となる．

以上のことから明らかなように，どのような方式を用いてもノイズの混入は避けられない．最近はコンピュータの処理機能の進歩のおかげで，力仕事でこの処理がかなり可能となってきてはいる．しかしながら，調査における自由回答文の場合は，設問 (説明文) の内容や主題をかなり絞り込んでも，得られる記述の内容や記法・表記が乱れることが一般的であり，現状の技術力では単語や語を確実に同定できるか十分には期待できない．ちなみに，形態素解析を行うとどのような情報が得られるかを例で示す．

例．形態素解析の例

例題．後述の事例で用いるデータセットとして，Web 調査における取得データがある．用いる設問は，日本人の国民性調査にある自由回答に合わせた以下の 2 問である．用いた方法は「QJP(リコー)」ならびに「茶筅」である．解析結果はなるべくオリジナルの出力に合わせて表記した (表 1)．なお，我々の開発した InfoMiner with WinAiBASE で用いる分かち書き処理機能については後述の解析例で現れるので，ここでは省略する．

(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.

(注 1) 日本人の国民性調査は面接法による調査であり，また，設問文も若干表現が異なる．我々の Web 調査で用いた設問は 2 つに分かれるが，国民性調査では第 8 次調査までは 1 問のみ，第 9 次調査からは 2 問を用いている．
(注 2) 我々が用いている WinAiBASE による結果は，後述の事例にある．

この例の形態素解析からも分かるように，自然言語処理的には，まず表記の構造を形態素解析，構文解析により確認し，続いて意味的なアプローチから意味解析，意味理解といった操作が行われる．いずれにしてもコンピュータ処理の支援は避けられない．しかも一般には相当量の計算処理時間を要する．

このように，いわゆる自然言語処理的な観点に立つと，その要素技術は言語学的というよりも，きわめて工学的な考え方や研究が多い．またこのような処理形態が，実際に人が行う言語処理行動 (回答行動) に合っているか否かは，現時点の研究だけでは説明できるものではないし，ここでの報告の方向とは異なるものである．一方，言語学的観点からは，日本語は未熟あるいは流動的な変化や変容が日常的であり，その意味で言語 (学) 研究そのものが発展過程にある．

この他，日本語の曖昧性 (本当に曖昧かどうかの議論があるが)，デノテーション (語の明示的な意味，表向きの意味) とコノテーション (語の言外の意味，含意),「テニヲハ」の考慮，カテゴリー論との関係，最近話題となっている認知科学的なアプローチからのメタファの重要性等,「日本語の構造的な特徴」を巡る諸研究や議論がある．さらに，単純に電子的操作・処理法との関係で見ても，ワードプロセッサの登場による表記法の変化やインター

表 1. 形態素解析の例.

（回答例１）自分が安心して毎日の暮らしを送っていけること.
（回答例２）単に、食べて眠って、という動物的なことだけではなく、生きていくだけではなく、生き甲斐を持って生きていくこと。

下記のように，形態素解析の結果の分かち書き文と単語の品詞特定化の結果とが表記される.

【例１：QJPの場合】

（回答例１）
【形態素分割】
|自分.が|安心.し.て|毎日.の|暮.らし.を|送-っ.て|い-け-る.こと..

【単語リスト】

*[3](10)自分	(自分)	(41)名詞
.[4](14)が	(が)	(51)ガ=格助
*[5](16)安心	(安心)	(89)サ変名詞
.[6](20)し	<する>	(60.2)する=用
.[7](22)て	(て)	(55)テ=接助
*[8](24)毎日	(毎日)	(48)時副詞名詞
.[9](28)の	(の)	(51)ノ=格助
*[10](30)暮	(暮)	(41)名詞
.[11](32)らし	<らしい>	(60.26)ラシイ=ク幹
.[12](36)を	(を)	(51)ヲ=格助
*[13](38)送-っ	<送う|送つ|送る>	(15.22)動:五ワ用b|動:五タ用|動:五ラ用b
.[14](42)て	(て)	(55)テ=接助
*[15](44)い-け-る	<いける>	(12.3)動:下カ終
.[16](50)こと	(こと)	(54)コト=終助
.[17](54).	(.)	(91)句点

（回答例２）
【形態素分割】
|単に.、|食-べ.て|眠-っ.て.、.と.い-う|動物-的-な|こと.だけ.で.は|な-く.、|生-き.て.い-く.だけ.で.は|な-く.、|生-き.甲斐.を|持-っ.て|生-き.て.い-く.こと.。

【単語リスト】

*[3](10)単に	(単に)	(31)副詞
.[4](14)、	(、)	(92)読点
*[5](16)食-べ	<食べる>	(12.2)動:下バ用
.[6](20)て	(て)	(55)テ=接助
*[7](22)眠-っ	<眠る>	(15.22)動:五ラ用b
.[8](26)て	(て)	(55)テ=接助
.[9](28)、	(、)	(92)読点
.[10](30)と	(と)	(51)ト引用=格助
.[11](32)い-う	<いう>	(75.4)イウ=五ワ体
*[12](36)動物-的-な	<動物的だ>	(22.4)形動:ダ体
*[13](44)こと	(こと)	(43)形式名詞
.[14](48)だけ	(だけ)	(53)ダケ=副助
.[15](52)で	<だ>	(60.23)ダ=用c
.[16](54)は	(は)	(52)ハ=係助
*[17](56)な-く	<ない>	(21.22)形:ク用b
.[18](60)、	(、)	(92)読点
*[19](62)生-き	<生きる>	(11.2)動:上カ用
.[20](66)て	(て)	(55)テ=接助
.[21](68)い-く	<いく>	(75.4)イク=五カ体
.[22](72)だけ	(だけ)	(53)ダケ=副助
.[23](76)で	<だ>	(60.23)ダ=用c
.[24](78)は	(は)	(52)ハ=係助

表 1.　(続き).

```
*[25](80)な-く      <ない>          (21.22)形:ク用b
.[26](84)、         (、)            (92)読点

*[27](86)生-き      (生き)          (41)名詞
.[28](90)甲斐       (甲斐)          (41)名詞
.[29](94)を         (を)            (51)ヲ=格助

*[30](96)持-っ      <持う|持つ|持る> (15.22)動:五ワ用b|動:五タ用|動:五ラ用b
.[31](100)て        (て)            (55)テ=接助

*[32](102)生-き     <生きる>        (11.2)動:上カ用
.[33](106)て        (て)            (55)テ=接助
.[34](108)い-く     <いく>          (75.3)イク=五カ終
.[35](112)こと      (こと)          (54)コト=終助
.[36](116)。        (。)            (91)句点
```

【例２：茶筅の場合】

（回答例１）

```
自分        じぶん      自分        普通名詞
が          が          が          格助詞
安心        あんしん    安心        サ変名詞
して        して        する        動詞          サ変動詞        タ系連用テ形
毎日        まいにち    毎日        時相名詞
の          の          の          名詞接続助詞
暮らし      くらし      暮らす      動詞          子音動詞サ行    基本連用形
を          を          を          格助詞
送って      おくって    送る        動詞          子音動詞ラ行    タ系連用テ形
いける      いける      いける      動詞          母音動詞        基本形
こと        こと        こと        形式名詞
.           .           .           句点
EOS
```

（回答例２）

```
単に        たんに      単に        副詞
、          、          、          読点
食べて      たべて      食べる      動詞          母音動詞        タ系連用テ形
眠って      ねむって    眠る        動詞          子音動詞ラ行    タ系連用テ形
、          、          、          読点
と          と          と          引用助詞
いう        いう        いう        動詞          子音動詞ワ行    基本形
動物的な    どうぶつてきな 動物的だ  形容詞        ナ形容詞        ダ列基本連体形
こと        こと        こと        形式名詞
だけ        だけ        だけ        副助詞
で          で          で          格助詞
は          は          は          副助詞
なく        なく        ない        形容詞        イ形容詞アウオ段 基本連用形
、          、          、          読点
生きて      いきて      生きる      動詞          母音動詞        タ系連用テ形
いく        いく        いく        動詞          子音動詞カ行促音便形 基本形
だけ        だけ        だけ        副助詞
で          で          で          格助詞
は          は          は          副助詞
なく        なく        ない        形容詞        イ形容詞アウオ段 基本連用形
、          、          、          読点
生き甲斐    いきがい    生き甲斐    普通名詞
を          を          を          格助詞
持って      もって      持つ        動詞          子音動詞タ行    タ系連用テ形
生きて      いきて      生きる      動詞          母音動詞        タ系連用テ形
いく        いく        いく        動詞          子音動詞カ行促音便形 基本形
こと        こと        こと        形式名詞
。          。          。          句点
EOS
```

ネットの利用下における E-mail 用語 (専門語)，E-mail 語，チャット語，さらには携帯電話用語 (ケータイ語) の登場と，日本語の様相は様々である．

3.2 統計的データ解析の観点からのアプローチ

ところで，自然言語処理や言語情報処理で行われてきたような発見的，計算アルゴリズム的なアプローチによる処理方法から少し離れて，これを統計的データ処理のパラダイムの中で考える．つまり，従来からある形態素解析と統計解析 (とくに多次元データ解析) の諸要素技術の部品 (手法) を適当に組み合わせることで，従来の個々の方法論では解決できなかった調査分野のデータ解析手法としての新たな方向に向かう可能性の検証とすることを試みる．まず，問題を単純化して，次のように考える．

(1) テキスト・ファイル化した自由回答文・テキスト型データ等を「構成要素」(fragments) に分解する．これを「分かち書き」処理により，例えば「単語や文節」に分解する．つまり，形態素解析の要素技術のうちの「分かち書き処理」の機能だけの援用を受ける．
(2) これから導かれる構成要素の出現頻度のパターン等の解析を行う方法として考える．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」といった経験的な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」の並びという程度に考える．ここで，構成要素とは以下のようなことから単語・語と区別するために用いる，ある曖昧な概念である．

① 前述のように日本語には「分かち書き」の考え方はない．また形態素解析の確定的な方式は未だあるとは言えず，それだけに流動的である．そこで，この操作はむしろ事前処理・中間処理として利用する．
② テキスト化された日本語文章を何かの意味で「分かち書き」した各単位，つまり「構成要素」に適当に分解するという程度の緩やかな約束でよいと考える．
③ しかし，分かち書き処理をどう行ったかの過程が明示的に分かるように努める．
④ 構成要素を複数結合した場合を「文節」と呼ぶことにする (文法で言う文節より緩やかな意味)．

分かち書きを緩やかな決まりとする理由は，元々の取得データ自体が曖昧かつ多様な表現であるから，むしろそれを許容して，厳密な定義や拘束を避ける方向で分析を進めるという視点に立つという意味である．たとえば，以下のような理由がある．また，このことが具体的な解析システム開発時の設計指針に反映されている．

(1) 日本語の精密な言語学的研究が目的ではないこと

研究対象とする内容は，元来がノイズが多い自由回答・自由記述等の解析を目標としている．自由回答をいかに科学的に取得し統計的な処理を可能とするかに焦点があり，個々の記述内容の意味論的な分析や言語学的な構造の研究が主たる目標ではない．また，現時点でこれに関わることは，そもそも「(近代)日本語」とは何か，その解釈の根幹に関わることでありきわめて難しい課題である．しかも，近代日本語の歴史自体が浅く，言語学的，日本語文法的にも明らかでないことが多すぎる．加えて，外来語や新造語の混用が特徴であり，こうした範囲までを言語学的手法でアプローチできるとは限らない．

(2) 得られるデータに曖昧性があること

更に重要なこととして，アンケート調査等で取得される自由回答データは，そもそもその

表記法や記述内容に曖昧性 (ambiguity) があり，整った文章が得られるとは限らない．表現の豊かさ，柔軟性があるという言い方もあるが，それだけ表記内容に自由度・曖昧性が高く意味を捉えにくいとも考えられる．とくに，Web 調査などではカタカナ語，欧米語の氾濫現象がある．また，続々と新語やカタカナ語が増える傾向にある．回答者の表記法・表現法もまちまちである．

(3) 他の利用方法との関連

分かち書き処理を行った文章の解析だけではなく，キーワード抽出で得られた「語の列記」のデータ等も扱うことがある．また，従来から自由回答処理の方法として利用されてきたアフターコーディング処理などとの併用や比較検証も必要となることがある．さらに，“ 意図的に ” (目的に応じて) テキスト・データを再編集して，解析結果を相互に比較するという利用方法も考えられる (あるデータセットから得られる答えは一つとは限らない)．

(4) 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

集積した自由回答・自由記述データの中に潜在する構造の類似性・差異性や規則性等を知ることは重要な目的である．このために，探索的な多次元データ解析手法が有効である．とくに，個々の回答・記述の意味内容や意見の規則性や類型を知ることが必要となる．しかし同時に，解析から得た「類型」に含まれる「個々の回答データの特徴」を読みとることや，類型で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だけでは十分ではなく，意見の類型化とその内容分析が必要となる．

(5) 従来の定量的調査法の理論の援用を受けること

自由回答データの特徴の一つに，選択肢型設問や属性など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たり統計的な検定の操作により標本誤差を検討したり設問間の差異を比較検証することが可能である．しかし，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，あるいは比較可能となっていること，つまり定量的操作との併用が，自由回答の解析に妥当性を付加する措置として必須であると考えられる．このことから，従来型の選択肢型設問項目や属性項目などと自由回答設問とを併用し，これらのデータの相互関連性の分析が重要である．自由回答の分析結果に加えて，これらの項目との相互検証を可能にする集計評価機能が重要と考えられることがある．つまり，定性的調査と定量的調査の併用が必要と考えられる．

4. テキスト型データ解析システム: InfoMiner with WinAiBASE

次に必要とされることは，以上のような目的 (解析の指針) を達成するためにはそれに適したコンピュータを用いた統計システムの開発である．これのために開発されたシステムが InfoMiner with WinAiBASE である．これは，日仏他の研究者を中心に開発された SPAD.T (Système Portable pour l’Analyse des Données，Donnée Textuelles) を基本エンジン部とし，これに分かち書き処理他の日本語解析に必要な機能を追加した統計システムであり，長

期にわたる日仏共同研究の成果の一つである．

(注) InfoMiner は商標登録第 4387759 号 (第 9 類: 電子応用機械器具及びその部品) を取得している．

従来の類似ソフトが形態素解析に始まる一連の言語情報処理的な視点から開発されてきたことと異なり，調査環境下において取得したテキスト型データに発生しうる状況を考えたデータ重視・実践型機能を実装した統計的記述解析を設計指針とすることが特徴である．とくに，選択肢型設問・属性データを併用する自由回答型を含めた調査データ解析に適している．なお，InfoMiner と類似の機能を備えた，とくに調査データの分析に特化したソフトウェアはほとんど例がないと思われる．その一つの例として，フランスで開発された “Sphinx Survey: Plus2 & Lexica” (Sphinx Development (1998)) があるがこれには当然日本語処理機能は含まれない．

4.1 システムの設計指針

SPAD.T から InfoMiner に至る開発経緯とそのシステムの詳細を述べることは別の機会に譲るが，現状の InfoMiner となるまでに約 10 年を経ていることだけを指摘しておく．ここでの目標は，前述の日本語の特徴 (現時点での利用体系) を考慮したうえで，分かち書き処理機能と統計解析機能を違和感なく接続利用できる利用環境を実現することにある．回答間，構成要素間それぞれの間の回答パターンの類似性や，回答と構成要素の間の関連性 (対応) の理解に役立つ知見をうるためにはどうあるべきかという考え方が背景にある．

(1) システムの主な特徴

基本操作は，元のテキスト型データを構成要素に分解し，構成要素 (単語や文節) と回答 (たとえば被験者，回答者，著者，検体)，その類型化情報との相互の関係をパターンとして表現し分析することにある．これに対して，計算処理上の工夫が必要とされるので，これへの手当を行う．

① 日本語独自の事前処理を必要とするのでその機能が含まれる．
② 一般の文章データの分析も可能である．
③ 平易な多次元データ解析手法を使っている．既に統計的方法論としての実績のある，対応分析法，クラスター化法およびそれに関連した基本的な統計処理を用いる．これは解析内容の透明化を図るためである．
④ 通常の選択肢型設問・属性データとの併用分析を行う．
⑤ 数値計算処理上の工夫が必要となる．
扱うデータ行列の要素がきわめて「疎」となる事，はずれ値への手当が必要である事，データ表の寸法が不定である事等への対策 (単語数が確定しないと行列の大きさが確定しない)，大量データの分類操作が必要となる事，辞書の再編集，不要文字の削除，類似文字，類語の再編集機能等の手当を要する事などがある．

(2) システムの動作環境の概要

全文検索等を行うソフトの大半が，ワークステーションや汎用大型コンピュータ，あるいは場合によっては (例えばテキスト・マイニングのツール)，並列コンピュータなどの利用を必要とする．しかし，調査データの解析ではなるべく可搬性を考慮して PC で利用できることが重要である．InfoMiner は Windows98 あるいは 2000 対応の既存の PC で十分に

利用可能である.

4.2 InfoMiner の基本機能

(1) ファイル設定機能

① テキスト型データの解析対象原文データの登録
② 設問・属性のラベルファイル
③ 設問・属性に対応する数値データ・ファイル

(2) 分かち書き処理機能

日本語文章・テキストを電子化したファイルに基づき，分かち書き処理を行い，統計解析に必要なデータセット・ファイルを生成する．InfoMiner では分かち書き処理機能として Happiness (平和情報センター) を InfoMiner 用に改良した WinAiBASE を採用している．これを用いて自動的に分かち書き処理が行われる．これにより以下の 3 種のファイルが作成される.

① 分かち書きファイル (テキスト・ファイル)
② キーワード・ファイル (テキスト・ファイル)
③ それらの計数ファイル (回答別の分かち書き数，キーワード数の記録)

また，パラメータ変更による処理条件の設定，検索・置換，編集，更新等の機能がある.

(3) 辞書の作成機能

① 単語辞書作成機能：単語・文節の辞書生成
② 辞書の編集機能
　・除外する文字・記号などの指定
　・単語の最小出現度数の閾値設定
　・解析から外す単語の指定
　・置換・読み替え等の単語の指定編集
③ 辞書の更新
　・編集指示に従って辞書を更新
④ 使用する単語辞書の指定
　・「標準」「編集・更新」の指定
⑤ 文節辞書作成
　・保存した文字列の文節としての相互リンケージを確認するモジュール
　・文字列間の隣接引用関係の表示

(4) 解析部機能

①「(回答)×(単語あるいは文節)」表の分析

最も標準的な分析を行うモジュールである．抽出した単語 (文字列) と回答 (たとえば回答者，サンプル) の関連表の多変量解析による情報の縮約を行う.

(注) 前述のように，以下で「単語，文節」とは，分かち書きで得た緩やかな意味の「構成要素」のことを言う.

(注)「追加処理」(supplementary treatments) を指定するオプションがある．これについては大隅 他 (1994) を参照.

②「(単語・文節)×(生成クラスター)」のクロス表の有意性テスト

クラスター化で生成したクラスター番号は一種のカテゴリー変数である．これを用いた有意性テストを行う．またクラスター化で，単語・文節のクラスターあるいは回答者のクラスターが生成される．

- 「単語出現頻度」を用いたクラスター別の出現単語頻度の有意検定機能 (クラスターに有意な単語を知る)
- 「カイ二乗距離」を用いた有意検定の機能 (生成した各クラスターに寄与する回答パターンを知る)
- 有意として選出の回答 (回答者，サンプル) のリストを出力する機能

(注) 自動分類による「教師なし分類」に相当する．分類結果の情報から二次分析を行うこともある．

③「(単語・文節)×(設問・属性)」表の解析

抽出した構成要素 (単語，文節等) と，予め用意した「選択肢型設問・属性」データセットの個々の変数とのクロス表の出力，有意性テスト，生成クロス表の対応分析，単語および回答のクラスター化等を行う機能である．つまり「単語と設問・属性間との関連を知りたいとき」に用いる機能である．たとえば，抽出単語と「性別・年齢区分」の変数とがあって，これらがどう対応するかを知りたいとき等に用いる．比較する設問がそれぞれ類似の内容であれば，回答間の相似性や回答の均一性の有無についての探査も可能である．

④「(単語・文節)×(設問・属性)」クロス表の有意テスト

- 抽出した構成要素 (単語，文節等) と「選択肢型設問・属性」データセットの個々の変数との関連表の出力，有意検定，生成データ表の多変量解析等を行う機能
- 「(単語・文節)×(設問・属性)」表の有意テスト，どの設問・属性が構成要素の説明に寄与しているかの検証.

(注) これは自動分類に対して，選択肢や属性を分類尺度として利用する一種の「教師あり分類」と考えることができる．

⑤ 単語の用語検索 (concordance) 機能

コンテンツ・アナリシス (内容分析) や KWIC の基本操作である用語検索 (コンコーダンス：ある指定した単語が与えられた文章・回答の中でどう使われたか) の一覧を，指定した単語を基準に検索・ソートし出力する．

(注) KWIC: Key-Word-In-Context の略．KWIC リストの一部としてコンコーダンスを用いる．この他，cross-reference, KWOC (Key-Word-Out-of-Context), 出現単語頻度表などを用いる．

(注) ②〜④ でいう有意性テストについては，後述の事例ならびに Lebart et al. (1998), 大隅 他 (1994) を参照.

表 2. データ表の構成.

＜表側項目：I ＞	＜表頭項目：J ＞
(単語，キーワード，文節)	(回答者)
(単語，キーワード，文節)	(クラスター・メンバーシップ情報)
(単語，キーワード，文節)	(選択肢型設問・属性)

回答者を (一般的な分析対象) と置き換えた場合の適用可能.
表側と表頭を入れ替えても解析結果は同等である (対応分析の持つ双対性から).

4.3 データ表の構成

前述の機能に合わせて種々のデータ表を対象とした分析が可能であるが，その基本形は「二元のデータ表」である．たとえば“ クロス表 ” を考えればよい．いま 2 つの項目を I, J とし，二元のデータ表 (表側項目 I) $\times$ (表頭項目 J) を以下のように行列 $\underset{m\times n}{\boldsymbol{F}}$ で表す.

$$\underset{m\times n}{\boldsymbol{F}} = (f_{ij}) \qquad (f_{ij} \geq 0, i \in \boldsymbol{I}, j \in \boldsymbol{J})$$

ここで，$\boldsymbol{I}$ と $\boldsymbol{J}$ はそれぞれ出発時のデータ表の行 (表側) と列 (表頭) の項目のカテゴリ (選択肢) の集合である.

$$\boldsymbol{I} = \{1, 2, \ldots, i, \ldots, m\}, \quad \boldsymbol{J} = \{1, 2, \ldots, j, \ldots, n\}$$

このとき，表 2 のようにデータ表の解析を任意に指定して行うことができる．このような構成を取ることから，InfoMiner の中では，以下の分析をオプションとして利用可能となる.

① 回答者の回答パターンと単語の関係を分析.
② 回答者のクラスター化情報と単語の関係を分析.
③ 回答者のクラスターを意味づける単語群を知り，典型的な回答者例を有意性チェックのもとに表示する.
④ 選択肢型設問や属性情報のうち，どれが単語と関連しているかを知る．また，その選択肢別の単語の有意性を知り，その典型的な回答者例を有意性チェックのもとに表示する.
⑤ 回答者のクラスター化情報と単語のクラスター化情報の関連を知る (2 次分析).

5. 事例解析

InfoMiner を用いた解析事例や適用例は既に何例もある．また，幾つかの企業ではこれを基本エンジンとした定性情報統合解析システムの構築も試みている (博報堂 (2000) による WordMap®)．ここでは，筆者が調査会社の協力を得て行った Web 調査から取得した自由回答の分析例を取り上げた.

5.1 調査方法と調査の特徴

インターネット環境下で行われる Web 調査の実験調査を，複数の調査機関との共同研究として実施してきた．ここで行われた Web 調査の目的の一つに，自由回答取得の方法論の研究が含まれる．従って，積極的に自由回答設問を取り入れ，しかもそれを目的に合わせて意図的に設計して，自由回答の表記の現れ方や実査で見られる諸現象の観察と分析を進めた．なお，調査全体の説明は報告書に譲り (大隅 (2000a))，ここでは以下の分析に必要な最小限の情報を要約する.

表 3. Web 調査の概要.

		電通リサーチ社			NTT ナビスペース社			リクルートリサーチ社	
		第 1 回調査			第 1 回調査			第 1 回調査	
		計画標本	登録情報	回答者数	計画標本	登録情報	回答者数	登録情報	回答者数
	総数	5,000	1,045	1,045	10,000	1,258	1,258	3,969	679
	回答率			20.9%			12.6%		17.1%
性別	男性	82.0	80.5	79.5	80.4	76.5	75.8	55.7	58.8
(%)	女性	18.0	19.5	19.3	19.6	23.5	23.5	44.3	41.2
	NA	—	—	1.1	—	—	0.7	—	—
年齢	～19 歳	1.5	0.9	0.8	2.7	2.8	2.1	1.0	0.9
(%)	20～24 歳	13.7	10.0	10.1	13.9	11.2	8.6	24.9	14.3
	25～29 歳	21.3	18.9	18.0	22.9	21.1	19.8	27.9	26.4
	30～34 歳	23.0	22.9	23.3	22.5	22.4	22.0	24.1	28.1
	35～39 歳	18.4	22.7	23.3	17.3	20.1	22.4	12.7	16.8
	40～44 歳	10.1	11.4	11.3	9.7	10.9	12.0	5.8	8.8
	45～49 歳	6.3	6.8	6.7	6.3	6.8	7.2	1.9	2.5
	50～54 歳	3.0	3.4	3.2	2.5	2.2	2.9	1.0	0.6
	55～59 歳	1.6	1.5	1.5	1.2	1.3	1.4	0.3	0.4
	60 歳以上	0.9	1.6	1.6	1.0	1.3	1.2	0.3	0.9
	NA	—	—	0.3	—	—	0.3	0.2	0.2

まず，Web 調査はある一定期間に，相異なる調査機関が設けている Web サイト上に置かれた共通の設問からなる調査票を用いて行われた．調査の概要は表 3 に示した．

これにみるように回収率がきわめて低いことが Web 調査の特徴の一つである．さらに，別の特徴 (とくに回答者の基本属性) として以下のようなことがある (表 3 参照).

(1) 回答者の年齢区分が 20 歳台から 40 歳台前半に偏っている (計画標本，回収標本ともにその傾向がある).
(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ずっと多い.
(3) しかし，回答者集団の登録方法により差異がある (電通リサーチ社, NTT ナビスペース社とリクルートリサーチ社で傾向が異なる).
(4) 回答者の居住地域が都市圏に集中する．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.
(5) 計画標本と回収標本の間に系統的なずれが生じている.

このように，住民基本台帳や選挙人名簿に基づく標本抽出の操作を経て，また調査員による面接法や留置自記式法などで得られた従来の調査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に疑問があると指摘する意見が多い．こうした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が，ここでは指摘するにとどめる．しかし，後に述べる解析結果の解釈時にはここで指摘した回答者の特性を十分に念頭に入れて対応することが必要である．なおこの調査分についての自由回答取得状況を 2 つの設問について集計すると表 4 のようになった.

5.2 用いた設問と回答の傾向

前述のように，Web 調査における自由回答取得方法の研究が目的の一つであったので，Web ページ上の調査票の中での自由回答の設問方法には様々な工夫を行った．たとえば，記入欄のスペースの大きさ，レイアウト，前後の設問との文脈関係等についての配慮を行った．このようなことで，自由回答設問は多数あるが，ここでは次に挙げる 2 つを例として用いる.

表 4. 自由回答の取得状況.

	電通リサーチ社			NTT ナビスペース社			リクルートリサーチ社		
	回答数	男 性	女 性	回答数	男 性	女 性	回答数	男 性	女 性
全 体	1,045	841	204	1,258	962	296	679	399	280
分析対象とした	1,039	836	203	1,250	954	296	678	398	280
サンプル数		80.5%	19.5%		76.3%	23.7%		58.7%	41.3%
質問 1–1									
自由回答	1,037	834	203	1,247	951	296	674	396	278
あり	(99.2%)			(99.1%)			(99.3%)		
回収の構成比	(100%)	80.4%	19.6%	(100%)	76.3%	23.7%	100%	58.8%	41.2%
質問 1–2									
自由回答	1,002	801	201	1,198	910	288	654	377	277
あり	(95.9%)			(95.2%)			(96.3%)		
回収の構成比	(100%)	79.9%	20.0%	(100%)	76.0%	24.0%	(100%)	57.6%	42.3%

表 5. 各サイトの特徴的な単語.

電通リサーチ社	家族，友人・仲間，金，健康，自分自身， 仕事，生活，時間， 趣味，人，環境，人間関係，…
NTT ナビスペース社	家族，自分自身，友人・仲間， 生活，健康，金，人， 時間，仕事，心，大切，環境，…
リクルートリサーチ社	家族，友人・仲間，自分自身，金， 健康，時間，仕事，生活，人， 環境，趣味，心，愛情，…

(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.

この設問のうち (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．表 4, 表 5, 表 6 によると，以下の特徴がみえる.

① いずれの調査サイトでも回答記入率が高い (90%以上である).
② サイト別にみると，表 5 のような単語の出現頻度が高い．また表 5 にみるように比率の若干の高低はあるものの，高頻度の出現単語は各調査サイトできわめて類似している.
③ 女性の回答率が低い．また偏りがある (表 4).

この他,「日本人の国民性調査」と比較したとき，選択される出現単語の特徴として，属性とくに「性・年齢区分」の間でかなり異なることが分かっている.

④ 生活環境，ゆとり，自分自身，生活等,「日本人の国民性調査」ではあまり現れない又は出現頻度の少ない単語が見られる.

表 6. 性別にみた利用単語の特徴.

	＜男性に多い単語＞	＜女性に多い単語＞
電通リサーチ社	家族，家庭，仕事，妻，趣味， 環境，健康，自然，平和，ゆとり， 財産，…	主人・夫，人，愛情，生きる・生きて， 大切，ペット・犬・猫，私，…
NTT ナビスペース社	仕事，家族，環境，健康， 生活，収入，社会，安定，…	心，自分自身，大切，人，恋人，私， 今，両親，主人・夫，…
リクルートリサーチ社	生活，金，仕事，家族，時間， 環境，趣味，ゆとり，…	友人・仲間，思いやり，人，心，気持， 両親，親，主人・夫，…

(注) 各サイトの自由回答の回答者数は以下のようである.
電通リサーチ社 (有効回答数: 1,037 名，男性: 834 名，女性: 203 名)
NTT ナビスペース社 (有効回答数: 1,247 名，男性: 951 名，女性: 296 名)
リクルートリサーチ社 (有効回答数: 674 名，男性: 396 名，女性: 278 名)

(注) 実はこの Web 調査では (大隅 (2000a))，4 回の調査を行い，第 1 回と第 4 回はまったく同じ内容の調査を行った．この設問についても同様で，同じ自由記述設問を 2 度行った．ここでは，このうちの第 1 回分の結果を示した．今後の検討が必要ではあるが，第 4 回の結果は第 1 回と大変に類似した結果が得られている．たとえば，第 1 回調査で「家族」「友人・仲間」を記述すると第 2 回でも「家族」「友人・仲間」を挙げる傾向にある．しかし，単語によっては，必ずしも同じとならない例もある (自分自身，思いやりなど)．ここらは，さらに慎重な検討が必要と思われる．なお，参考として，以下に「日本人の国民性調査」(1998 年実施分) の結果も併せて示した.

＜参考＞

1998 年に実施された「日本人の国民性調査」(1998 年) では以下の設問の調査を行っている.

Q2-7 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 (なんでもかまいません).
(自由回答：)

Q2-7a2 では，あなたにとって二番目に大切なものはなんですか.
(自由回答：)

ここで，Q2-7a2 はこの回に初めて設けられた設問である．前述の我々の用意した設問 Q1-4 に類似しているが設問文がやや異なっている．集計結果の一部は表 7 となっている.

なお,「日本人の国民性調査」は，調査員による面接調査であり，回収結果に基づいて，アフターコーディング処理により整理されたものである．両者は調査法も異なり，また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 歳台前半に回答者の分布が偏っていること注意せねばならないが，それでも「家族」「健康」「自分」「愛情」等に回答が集まる傾向は類似している.

5.3 分かち書き処理の結果

(1) 出現単語の頻度分布他

ここでは，2 つの設問(「一番たいせつなもの」「次にたいせつなもの」)を併合して分析を行った例を示す．まず 3 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.

分かち書き処理の結果得られた「総単語数」「異なり単語数」「編集済み異なり単語数」

表 7. 「日本人の国民性調査」の結果.

	回答者数	
	一番大切なものは	二番目に大切なものは
	1,339	1,339
1 生命・健康・自分	22.4(%)	16.0(%)
2 子供	8.5	4.9
3 家族	39.5	22.1
4 家・先祖	0.8	1.0
5 金・財産	3.4	16.5
6 愛情・精神	16.7	18.1
7 仕事・信用	2.8	7.6
8 国家・社会	1.8	3.6
9 その他	1.3	1.8
10 DK，特になし	2.8	8.4

表 8. 総単語数他の一覧.

閾　値	1	2	3	4	5	6	7	8
総単語数	9,914	8,954	8,522	8,216	7,928	7,733	7,529	7,382
異なり単語数	1,595	635	419	317	245	206	172	151
編集済み異なり単語数	1,504	562	358	262	197	163	129	110
異なり単語率	16.1	7.1	4.9	3.9	3.1	2.7	2.3	2.0

表 9. 頻度区分別単語数とその累積頻度.

語　数	1	2	3	4	5	6	7	8
頻度区分別単語数	942	204	96	65	34	34	19	16
累　積	942	1,146	1,242	1,307	1,341	1,375	1,394	1,410

「異なり単語率」「頻度区分別単語数」等の一覧を調査サイト別に求めた．ここでは電通リサーチ社の例を詳しくみる．

表 8, 9 は，上記の諸量の一覧である．ここで閾値とは，指定したその値以上の頻度の単語数のことをいう．たとえば，閾値=2 とは，出現頻度が 2 語以上の単語という意味である．表 9 の情報をグラフとしたものが図 2 である．ここで異なり単語数が急速に少なくなっている．さらに頻度区分別に単語数を集計すると，すなわち 1 語のみ，2 語のみ, . . . とそれぞれの語数別に集計し，累積度数と併せてこれをグラフとすると図 3 となる．ここで，1 語が圧倒的に多く，2 語以上は急速に低減することがみえる．こうした特徴は，この例だけでなく，経験的に共通した特徴である．従って以後の解析で何語以上の単語を採用するかは重要な選択肢となる．

(注) 少なくとも，ここで分析対象とした 3 つの調査サイトの例については，ほとんど類似の特徴がみられる．

(注) 図 3 はパレート図のようにみえるが，そうではない．ここで横軸は単語数，縦軸はその頻度と累積頻度である．1 語がもっとも頻度が多く，以下 2 語，3 語と続く．これと頻度の順位と一致しているということであるが，必ずこうなるということではないだろう．しかし，今までの分析例ではすべて同じ傾向を示している．これは興味ある特性であるがさらに検証が必要である．

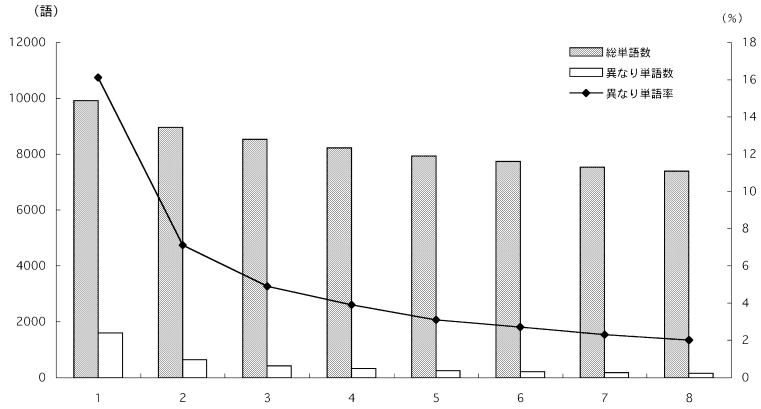


図 2. 単語の出現頻度分布.

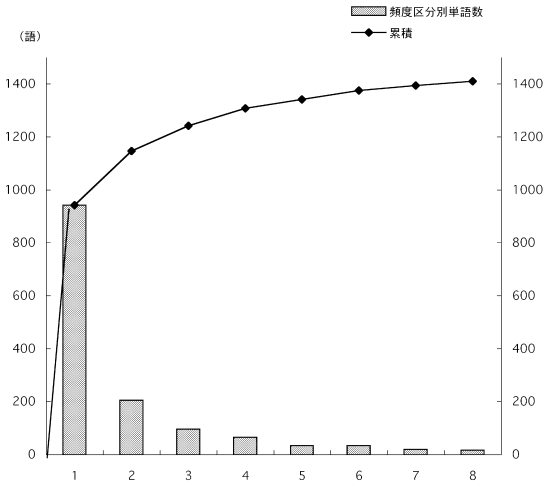
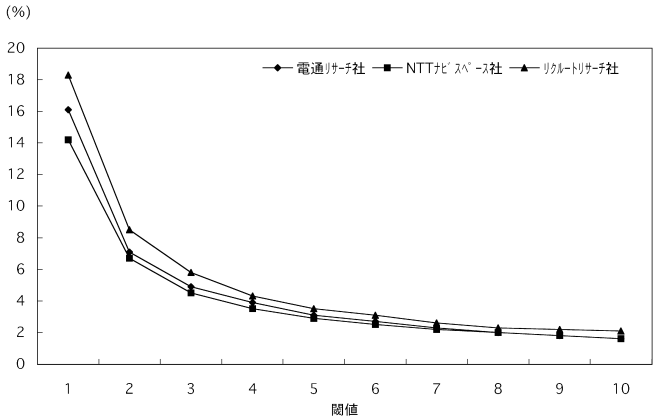


図 3. 頻度区分別単語数とその累積グラフ.

(2) 異なり単語率

次に，図 4 は「異なり単語率」と「閾値」の関係を表す図である．ここで，異なり単語率とは，総単語数に占める異なり単語数 (同一表記の単語を一語としたときの単語数) の割合である．図 4 に示した 3 サイトの異なり単語率の推移を比較するとかなり類似性がある．

通常，今までの経験則ではあるが，異なり単語率が少ないと用いられた用語 (単語) 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る．経験則では，8〜15%程度となることが多いが，グループインタビュー等のデータでは 40〜50%を越えることもある．また，反対に数% (5%以下) にとどまる場合も見られたがこれは記載の内容がきわめて限られたある話題に限定された場合であった．経験的には，総単語数の増大に伴って異なり単語数は減少する傾向があるが，それは必らずしも比例的には増加しない．数理的な根拠がないままに，異なり単語率の変化 (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



閾値	1	2	3	4	5	6	7	8	9	10
電通リサーチ社	16.1	7.1	4.9	3.9	3.1	2.7	2.3	2	1.8	1.6
NTT ナビスペース社	14.2	6.7	4.5	3.5	2.9	2.5	2.2	2	1.8	1.6
リクルートリサーチ社	18.3	8.5	5.8	4.3	3.5	3.1	2.6	2.3	2.2	2.1

図 4. 調査サイト別の異なり単語率.

る指標として重要と考えられてきたが，ここらは再考の余地がある (例えば，Lebart et al. (1998) では語彙の潤沢度: richness of vocabulary とし，村上 (1994) では語彙量の指標としている).

(3) 回答者別単語数の分布

InfoMiner では，各回答者の回答の分かち書きの結果，それぞれが何語からなるかの頻度も算出する．設問の内容やリーディングの内容の影響，調査票の記入欄の大きさ等によって文章の長さに差異が生じるからである．ここでいま言えることは，実験例を重ねることで調査方式 (設問内容，調査票のフォーマット等) の影響等が確かに見られたということだけであるが，ここらの研究はほとんど例がないので，算出情報をどう利用するかは今後の課題であるとだけ指摘しておく．また，この数値例はここでは省略した.

(注) 今までの実験例では，回答あたりの単語数の分布のバラツキ (分散)・変動係数や歪み (歪度) が非常に大きく，裾の長い分布となる．つまり，非常に長い回答文を書く少数の回答者が含まれるということを意味している．このことから，後述の対応分析による分析時には，こうした回答者データが，全体の大勢から外れた特異なはずれ値となる傾向が現れ，これに対処する機能が必要となる.

5.4 単語の編集

次に行う操作として，不要単語の削除，類語・同義語の併合，誤記の訂正等の分かち書き結果の再編集を行う．もちろんこうした操作を一切行わないで，分かち書き処理の後に直ちに解析を行って，特徴の初動探査を行うこともある．とくに不要単語の除去は，例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．もっとも，句読点や助詞の出現頻度に意味があると考える場合もあろうし，括弧の利用が意味の強調であったり，カタカナ表記に意味を持たせる，あるいは助詞・テニヲハの利用方法に意味があるとする

表 10. 削除の例.

， ． ・ ： ？ ！ ＾ ｜ … ‥ ”( ) 「
」 － ＝ ＃ ★ ● ← １ ２ ３ ４ ｅ ｔ ｃ の
と を が に な で や は か も へ し こと ある する いる だ
できる した もの よくわからない わかりません さ でも です
思いつきません 特にありません 特に無し 特になし 特にない なし
特に無いです 等

視点に立つと，ここでの操作はひどく乱暴なものとなる．ここではこうした含意的あるいは意味論的なアプローチとは異なる観点から分析を進めることに主眼がある (立場が異なると言った方がよい)．ここでの事例で扱ったデータセットの削除対象や置換処理の一部を表 10 に例として挙げた．ここにみるように，自由回答ではきわめて多種多様な書き方をするということに注意せねばならない．このことが，筆者等が分析手法だけでなく，自由回答の設問方式の研究も必要と考える根拠の一つである．

＜削除の例＞

表 10 にあるように，句読点，助詞，特殊記号等の他に,「特になし」「わかりません」等の回答も削除対象とした.

＜置換の例＞

置換については，表 11 にみるように，誤記の置換，類語・同義語と考えた語の統合等を指定する (ここでは一部を示した)．ここで，各行の等号の左側の語で，右側の複数の語で置換することを表す.

この簡単な例をみただけでも，自由回答の表記が実に多彩であることが分かる．実は，このような再編集を行うことで，異なり単語数がさらに減って，後述のように実際の解析対象となる単語数はそれほど多くはならないことに注意すべきである (分析がそれだけ容易になる)．再編集後の単語数の分布例は図 3 にある.

なお，上記の単語数の分布に限らず，解析全体を通じて，3 つの調査サイトの傾向には類似性が多々見られたが，同時にそれぞれのサイト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 つのサイトは大変によく似た傾向を示している．これを網羅的にここで述べることには無理があるので，以下では電通リサーチ社における Web 調査取得データを中心に分析例を示すことにする.

5.5 解析結果の解釈

InfoMiner による分析の方向としては，前述のように出発時の二元データ表の表側と表頭に充てる項目によって種々の結果が得られるが，ここでは以下のような分析を行う.

(1) (抽出単語) × (属性，とくに性別，性年齢区分別) の分析
(2) (抽出単語) × (回答者) の分析
(3) (抽出単語) × (回答者のクラスター化情報) の分析

(1) 属性と単語の関係

多くの場合，抽出した単語が既存情報，例えば選択肢型設問や属性あるいはそれに代わる何らかの情報とどう関連するかを知りたい．例えば「性別」「年齢区分」はどう関係するのか，あるいは何らかの設問選択肢との関係の有無などである.

表 11. 置換の例.

思いやり＝思い遣り おもいやり 思いやれる 思いやる
思って＝おもって
面倒＝めんどう めんどうみ
つながり＝つながち 繋がり
必要＝ひつよう
自体＝事体
電子メール＝ E–mail
暮らし方＝くらし方
暮らす＝くらす
友人–仲間＝友人 ともだち 友 友人たち 友達 なかま 仲間
生きる＝生きて 生きる 生きて いきている 生きている
思う＝思える
主人–夫＝主人 夫 ダンナさま
親戚–兄弟＝親戚 兄弟
ペット–犬–猫＝犬 猫 ペット
すべて＝全て
すべき＝瀬部器
不幸＝不等
ない＝なく
生きがい＝生き甲斐
コミュニケーション＝コニュニケーション コミニケーション コミニュケーション
コンピューター＝コンピューター
なにより＝ naniyori
家族＝ FAMILY family kazoku かぞく
満たされる＝満たされてる
愛情＝愛
子供＝子供たち 子ども
妻＝家内 嫁 嫁さん
団欒＝団らん だんらん
金＝かね かねぇ おかね お金
人付き合い＝つきあい
規則＝きそく
便利＝便利さ
自然＝しぜん
仕事＝しごと
食事＝しょくじ
時間＝じかん
自分自身＝自分 じぶん
友達つきあい＝友人つきあい
充実感＝充実感って 充足感
幸福＝ Happy 幸せ shiawase しあわせ
ヤマハ＝ YAMAHA
音楽＝ music
心＝こころ

① 性別を用いた場合

今の例について，性別について「男性，女性」で，それぞれ用いられた単語について，InfoMiner が出力する有意性テストの一覧を挙げる (表 12)．ここでは有意の程度に従って検定値の大きさの順に並べ替えた．例えば「男性」の初めに「家族」がある．この語は全体で 520 回 (語) 使われたが，そのうち男性が 417 回も用いたので，結果として男性を特徴付ける単語と判断される．またその有意確率は 0.001 であったとなる．

ここで検定値とは，ある単語 (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

判定する一つの指標である (詳細は Lebart et al. (1998) を参照).「女性」を見ると,「主人−夫」「いい」「私」等が上位にあるが,ここでの総単語数がきわめて少ないので,検定値の近似の程度が悪いことを考慮のうえで,解釈せねばならない.また,「家族」については,女性の利用頻度も高いのであるが (520−417=103 語となる),ここでは男性を特徴付ける語として現れる.こうして,男女間の差異を比較できる.

表 12. 性別からみた単語の有意性テスト.

［男性の場合(834 名)］

単語	出現比率 (%) 内部	出現比率 (%) 全体	出現頻度 (語) 内部	出現頻度 (語) 全体	検定値	有意確率
1 家族	12.20	11.35	417.	520.	3.125	0.001
2 家庭	1.05	0.83	36.	38.	3.001	0.001
3 自然	1.29	1.07	44.	49.	2.459	0.007
4 妻	0.64	0.50	22.	23.	2.315	0.010
5 経済的	0.41	0.31	14.	14.	2.134	0.016
6 精神的	0.38	0.28	13.	13.	2.014	0.022
7 平和	0.85	0.70	29.	32.	2.013	0.022
8 仕事	4.45	4.10	152.	188.	1.967	0.025
9 友情	0.35	0.26	12.	12.	1.888	0.030
10 趣味	2.63	2.38	90.	109.	1.873	0.031
11 環境	2.17	1.94	74.	89.	1.797	0.036
12 自己	0.29	0.22	10.	10.	1.615	0.053
13 彼女	0.29	0.22	10.	10.	1.615	0.053
14 暮らせる	0.26	0.20	9.	9.	1.466	0.071
15 教育	0.41	0.33	14.	15.	1.440	0.075
16 緑	0.23	0.17	8.	8.	1.307	0.096
17 信用	0.23	0.17	8.	8.	1.307	0.096
18 ゆとり	0.94	0.83	32.	38.	1.196	0.116
19 社会	0.47	0.39	16.	18.	1.142	0.127
20 老後	0.20	0.15	7.	7.	1.134	0.128

［女性の場合(203 名)］

単語	出現比率 (%) 内部	出現比率 (%) 全体	出現頻度 (語) 内部	出現頻度 (語) 全体	検定値	有意確率
1 主人−夫	0.60	0.15	7.	7.	3.818	0.000
2 いい	0.52	0.13	6.	6.	3.464	0.000
3 私	0.69	0.22	8.	10.	3.313	0.000
4 大切	1.12	0.48	13.	22.	3.155	0.001
5 ような	0.52	0.15	6.	7.	2.976	0.001
6 ペット−犬−猫	0.86	0.37	10.	17.	2.700	0.003
7 なる	0.60	0.22	7.	10.	2.667	0.004
8 楽しい	0.52	0.17	6.	8.	2.606	0.005
9 一緒	0.43	0.13	5.	6.	2.577	0.005
10 くれる	0.60	0.24	7.	11.	2.398	0.008
11 思いやり	1.20	0.68	14.	31.	2.225	0.013
12 毎日	0.43	0.15	5.	7.	2.204	0.014
13 して	1.89	1.22	22.	56.	2.170	0.015
14 必要	0.60	0.26	7.	12.	2.158	0.015
15 人	3.09	2.27	36.	104.	2.020	0.022
16 生きる−生きて	1.12	0.68	13.	31.	1.849	0.032
17 何	0.52	0.24	6.	11.	1.789	0.037
18 上	0.34	0.13	4.	6.	1.755	0.040
19 から	0.60	0.33	7.	15.	1.547	0.061
20 思う	0.60	0.33	7.	15.	1.547	0.061

② 性年齢区分の場合

上の例で性差があると分かっても，さらに年齢差を知ることも必要となる．そこで，新たに「性年齢区分」の変数を生成し，これと利用単語の間の関係を見る (表 13)．年齢区分は男女とも 5 歳間隔としたが女性については 45 歳以上の頻度が極端に少ないので，40 歳以上を括った．ここでは，各年齢層に有意に働く単語群と，逆にその層にはあまり寄与しない単語群とを，上位 10 語づつを列記した．また，検定値他の数値は省略した．この表にみる情報の解釈はとくに説明を要しないであろうが，主な特徴をひろうとつぎのようなことがある．

(ⅰ) 男性の若年層 (25 歳〜30 歳未満) では,「友人−仲間」,「彼女・恋人」,「人間関係」などが上位を占める．
(ⅱ) 一方，男性も 30 歳以上になると,「家族」,「経済」,「安定」等に意識が移る．
(ⅲ) さらに，40 歳以上では,「財産」,「健康」,「夫婦」,「環境」などの身近の要素に関心が移る．
(ⅳ) 一方女性では,「自分」,「思いやり」,「好き」,「愛情」といった語が若年層にある．
(ⅴ) 女性 30 歳以上では,「夫婦」,「子ども」,「主人−夫」，等，どちらかというと自分の身近な範囲の人間関係や生活環境の要素に関心が向いているようにみえる．

これらは，寄与しない側の語も考慮して解釈することで，より性年齢区分の特徴が顕著になる．また，同じ語，たとえば「家族」の現れ方などに性年齢間の差異を見ることができる．ただし，ここで留意することは，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．

(2) 回答者と選出単語の関係

上のように，事前情報 (属性，設問等) の区分 (選択肢) のいずれが選出単語の意味づけに有用かを知ることは，例えばマーケティング・リサーチにおける消費者類型では重要な操作である．しかし，事前情報なしに，クラスター化 (自動分類) による類型化を行って，潜在的にどのような回答者群があるかどうかを知ることも必要となる (例えばマーケティング・リサーチにおけるセグメンテーション)．これは分類基準が事前に与えられていない場合に回答パターンと単語の出現情報をもとに (自動的に) 類似したグループを生成することである (古典的な言い方では「教師なし分類」を行ったことになる)．従って，生成したクラスターを改めて他の項目 (選択肢型設問，属性情報等) と突合分析することが必要となるかもしれない．

ここでは単純に「抽出単語と回答者の回答パターンの関係」をみる．回答者によって用いる単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．このためには単語，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，同時に単語のクラスター化や，それら相互の関係を探査し，また回答者の個々の回答の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 (抽出単語) × (回答者) のクロス表であり，これに対応分析を適用する．従って (抽出単語)，(回答者) への数量化スコアが算出され，またその布置図が得られる (図 5)．なおここでは，今までと同様に対象として閾値が 5 以上の 197 語の単語を選んだ (表 4 も参照)．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 (ハイブリッド法と呼称) を用いている (詳細は大隅 他 (1994), Lebart et al. (1998))．これを用いる理由は，(i) 大量データの分類操作が必要となること，(ii) 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

表 13. 性年齢区分別にみた単語の有意性テスト.

男性25歳未満 (69人)	男性25歳以上–30歳未満 (131人)	男性30歳以上–35歳未満 (185人)	男性35歳以上–40歳未満 (205人)	男性40歳以上–45歳未満 (111人)	男性45歳以上–50歳未満 (68人)	男性50歳以上 (65人)
1 友人–仲間	1 など	1 妻	1 充実感	1 仕事	1 健康	1 健康
2 恋人	2 恋人	2 家族	2 家族	2 家族	2 財産	2 老後
3 金	3 人間関係	3 安らぎ	3 安定	3 職	3 家族	3 社会
4 情報	4 欲しい	4 趣味	4 精神的	4 環境	4 他人	4 教育
5 事	5 とき	5 遊び	5 実現	5 経済	5 夫婦	5 家庭
6 彼女	6 人	6 車	6 経済的	6 施設	6 平和	6 友情
7 音楽	7 彼女	7 楽しさ	7 余暇	7 地球	7 交流	7 人間
8 いろいろな	8 程度	8 空間	8 幸福	8 安全	8 特	8 自然
9 親友	9 努力	9 親戚–兄弟	9 教育	9 家	9 自然	9 仲良く
10 考える	10 しない	10 両親	10 緑	10 公共	10 地域	10 空気
10 趣味	10 経済的	10 健康	10 自然	10 人生	10 夢	10 財産
9 子供	9 ない	9 何	9 好きな	9 生きがい	9 友人–仲間	9 家
8 生活	8 安全	8 人生	8 親戚–兄弟	8 夢	8 充実	8 金
7 平和	7 充実感	7 必要	7 ペット–犬–猫	7 財産	7 子供	7 思いやり
6 家庭	6 子供	6 人	6 今	6 友人–仲間	6 思いやり	6 生きる–生きて
5 安定	5 健康	5 など	5 自分自身	5 心	5 恋人	5 家族
4 環境	4 親戚–兄弟	4 教育	4 両親	4 人間関係	4 安定	4 恋人
3 健康	3 家	3 恋人	3 人	3 金	3 時間	3 時間
2 仕事	2 幸福	2 ペット–犬–猫	2 命	2 事	2 金	2 友人–仲間
1 家族	1 家族	1 心	1 友人–仲間	1 人	1 して	1 自分自身

女性25歳未満 (43人)	女性25歳以上–30歳未満 (57人)	女性30歳以上–35歳未満 (55人)	女性35歳以上–40歳未満 (37人)	女性40歳以上 (11人)
1 私	1 なる	1 主人–夫	1 いい	1 事
2 楽しい	2 思いやり	2 相手	2 子供	2 収入
3 能力	3 ような	3 上	3 家	3 心
4 して	4 大切	4 目標	4 毎日	4 夫婦
5 夢	5 思う	5 状態	5 優しさ	5 空間
6 便利	6 人	6 日々	6 生きがい	6 思いやり
7 くれる	7 何	7 思いやり	7 命	7 ような
8 含む	8 愛情	8 生きる–生きて	8 人	8 社会的
9 大事	9 いう	9 仲良く	9 ない	9 安定
10 好きな	10 交流	10 食べ物	10 くれる	10 地位
10 妻	10 時間	10 今	10 妻	
9 環境	9 生活	9 人生	9 関係	
8 子供	8 会社	8 収入	8 思いやり	
7 思いやり	7 親戚–兄弟	7 事	7 幸福	7 人間関係
6 平和	6 財産	6 家族	6 家族	6 して
5 親	5 仕事	5 妻	5 友人–仲間	5 家族
4 家庭	4 家族	4 命	4 家庭	4 金
3 趣味	3 事	3 自然	3 ゆとり	3 仕事
2 家族	2 自然	2 人	2 心	2 趣味
1 健康	1 環境	1 家庭	1 時間	1 生活

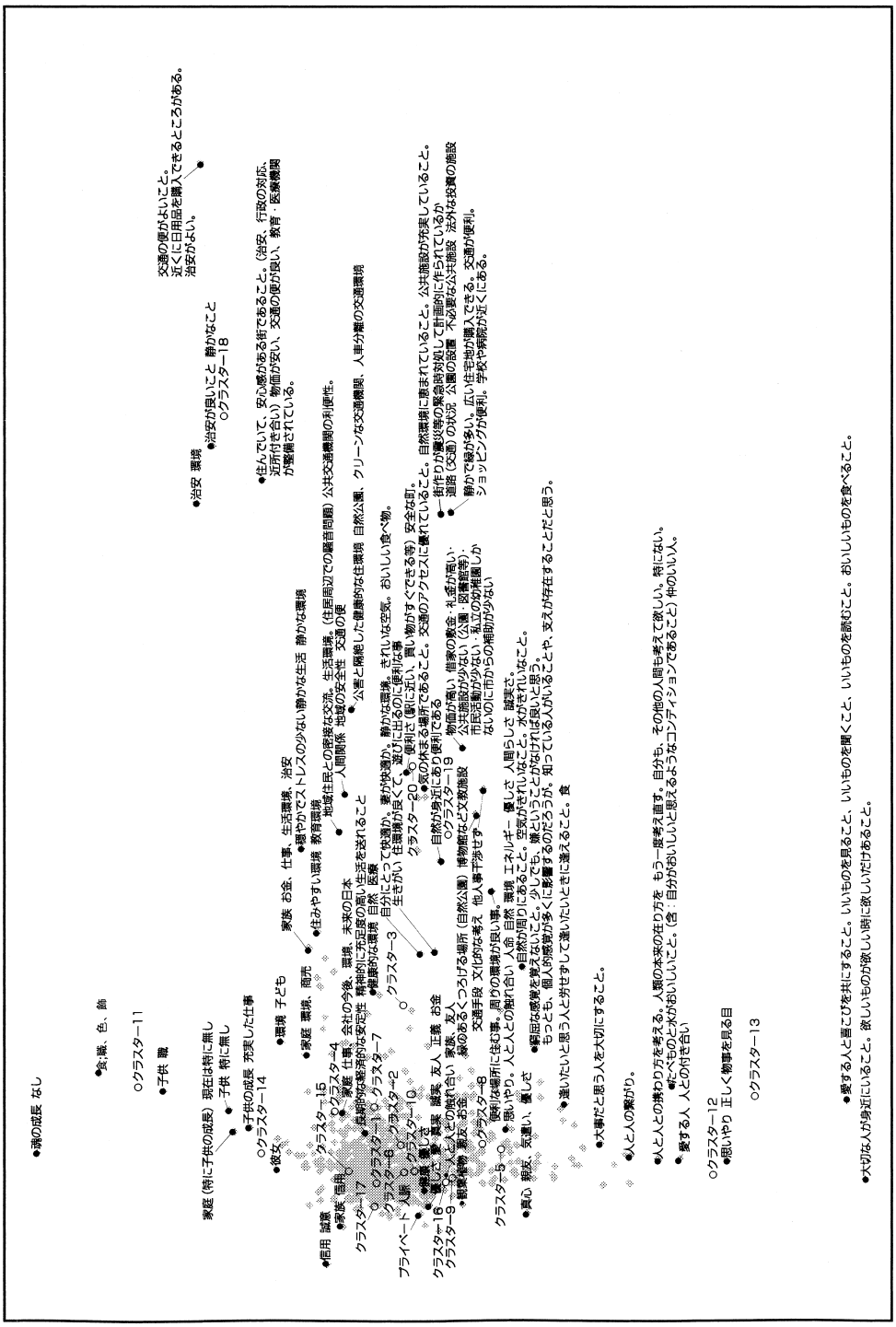


図 5.　回答者スコアの布置と自由回答.

かもはずれ値が頻出すること，(iii) 従って，こうしたスコアの分布のクセを考慮した分類法 (クラスター化操作) が必要とされたこと，などがある．

① 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を貼り付けて観察することが意味があるが，この操作は図を乱雑にし視認性が悪くなるという欠点もある．しかし，主にはずれ値あるいは点の分布の端の方にある回答の特徴を探査するには適している．図 5 がこの例である．またこの図にはクラスター化で得たクラスター重心スコアも書き入れてある (クラスター数は 20 群とした)．

図の右方向に「地域環境，生活上の安全，住みやすさ，住居環境と利便性」に関連した意見，図の下方には「思いやり，大切な人, . . .」などが特徴として見える．しかしこれではいかにも視認性が悪い．また，出発行列の性質上，行列のサイズ (次元数) がきわめて大きく，また各セル内の頻度がきわめて疎ということがある．したがって，数次元の少数次元の中に射影したスコアの視認観察には自ずと限界がある．そこで後述のように，クラスター別に，出現単語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 20 群としたが，これとする根拠は多分に主観的である．実際の分析操作としては，(i) 閾値を変えて選出単語数を変える，(ii) クラスター数を変化させてクラスター化生成過程を追跡するなどの探査を繰り返し行うことが必要となる．

(注) クラスター化処理ではクラスター数をいくつと指定するかが常に問題とされる．ここでは 20 群となっているがこれを採用した理由は特にない．InfoMiner では，クラスター数を変えながらそのクラスター化の履歴を追って，ある分類基準を目安に決めることができるとだけ指摘しておく (大隅 他 (1994) を参照)．

② 回答者クラスター化情報と単語の関係

次に，解析に用いた単語と回答者の関係を検証するが，これも布置図としては情報が煩雑となりかえって理解が困難となるので，ここでは回答者のクラスター化で得た 20 群のクラスター重心と単語との関係を布置図として表した (図 6)．

こうしてクラスターと回答者が用いた単語との関係がおおよそ分かる．個々のクラスターの意味を特徴付ける単語がその近傍にあるようには見えるが，これでは十分とはいえない．つまりクラスター内の個々の回答の特徴が単語とどう関連するかはこれだけでは曖昧である．理由は前述のように疎なデータ行列を扱うことにある．

次の操作として，前に用いた単語の有意性検定 (出現頻度のテスト) をここでも適用する．基本は「ある単語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単語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付ける単語と，それとは逆にそのクラスターの説明に寄与しない単語とを，頻度検定を行って一覧とした情報を要約する (表 14)．ここではそれぞれ寄与の高いあるいは低い順に 30 語を表とした．30 語に満たない例は判定基準値に満たなかった場合 (オプションで指定) である．また，クラスター・サイズの小さい群 (サイズが 15 サンプル未満) は省略した．実はクラスター・サイズの小さいはずれ値的なクラスターは，大勢から離れた特殊な意見を代表するので，実用上はそれなりに意味がある．

前と同様に，この表の解釈はそれほど厄介ではない．また，このクラスター別の検定結果に対して，それぞれのクラスター内の代表的な回答 (元の自由回答) がどのように分類され

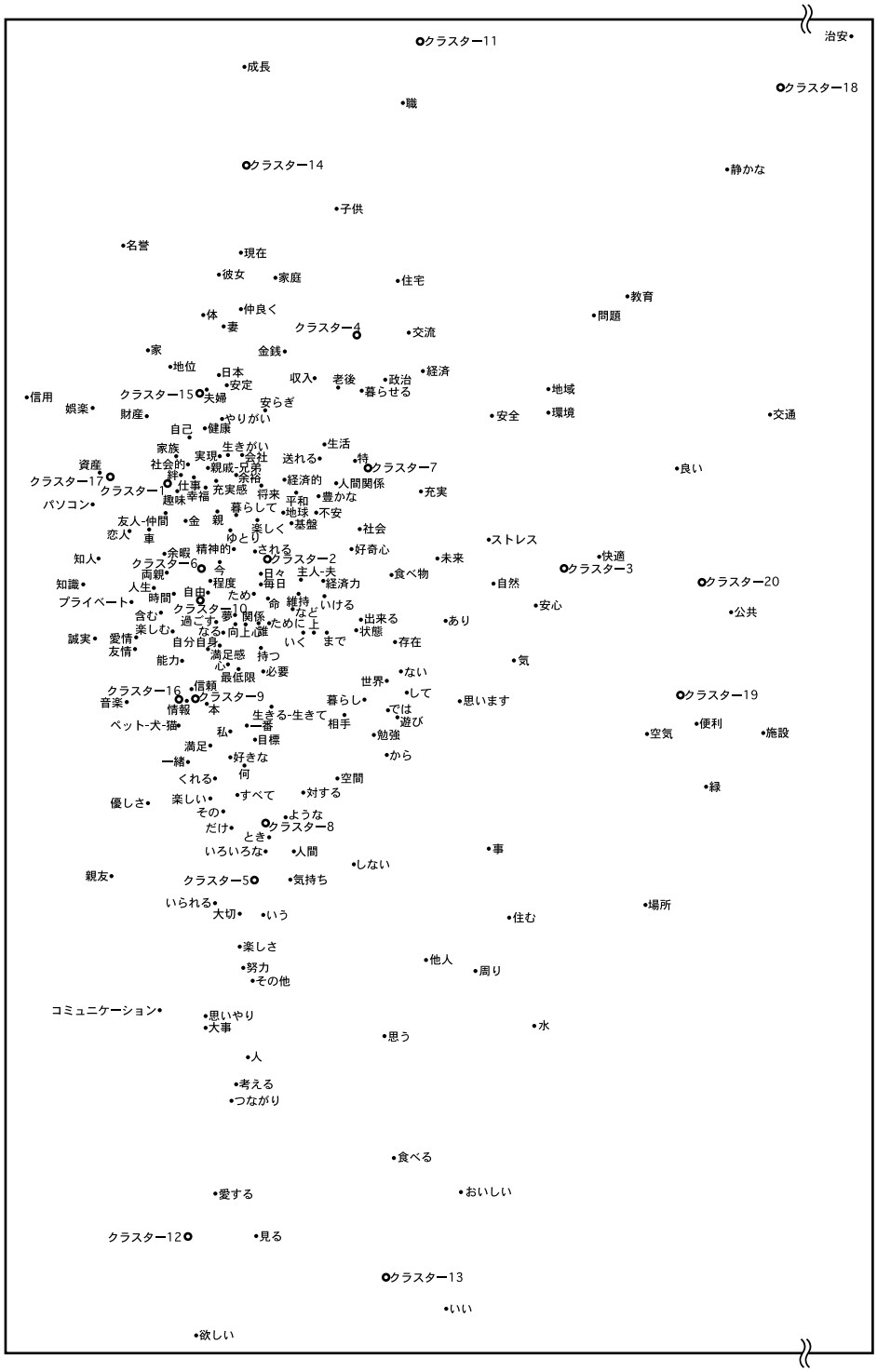


図 6. 回答者クラスター重心と単語の関係 (同時布置図).

表 14. 回答者クラスターの単語の一覧.

クラスター 1 (314人)	クラスター 2 (461人)	クラスター 4 (25人)	クラスター 5 (45人)	クラスター 6 (20人)	クラスター 9 (64人)	クラスター 15 (19人)	クラスター 17 (18人)
1 家族	1 生活	1 家庭	1 人	1 人生	1 信頼	1 彼女	1 信用
2 友人–仲間	2 ゆとり	2 教育	2 つながり	2 楽しむ	2 愛情	2 名誉	2 信頼
3 仕事	3 時間	3 安らぎ	3 いう	3 将来	3 思いやり	3 地位	3 誰
4 恋人	4 心	4 豊かな	4 情報	4 社会	4 友情	4 社会的	4 誠実
5 家	5 生きがい	5 環境	5 愛する	5 余暇	5 関係	5 今	5 一番
6 パソコン	6 安定	6 体	6 大事	6 自分自身	6 努力	6 金	6 愛情
7 車	7 余裕	7 住宅	7 目標	7 含む	7 対する	7 日本	7 人間関係
8 金	8 事	8 ゆとり	8 場所	8 まで	8 気持ち	8 仲良く	8 家
9 趣味	9 して	9 対する	9 その	9 充実	9 他人	9 知識	9 金
10 親戚–兄弟	10 自由	10 心	10 住む	10 世界	10 人	10 目標	10 家族
11 健康	11 平和	11 生活	11 大切	11 遊び	11 娯楽	11 自然	11 友人–仲間
12 親	12 好きな	12 社会	12 思う	12 幸福	12 幸福	12 いく	
13 両親	13 子供	13 今	13 空間	13 すべて	13 財産	13 人間関係	
14 知人	14 持つ	14 会社	14 満足	14 ために	14 音楽	14 車	
15 音楽	15 ない	15 妻	15 コミュニケーション	15 知識	15 から	15 収入	
16 会社	16 だけ	16 充実	16 生きる–生きて	16 人間	16 つながり	16 両親	
17 やりがい	17 経済的	17 親	17 充実感	17 いく	17 金	17 財産	
18 ペット–犬–猫	18 いける		18 考える	18 健康	18 コミュニケーション	18 命	
19 娯楽	19 暮らせる		19 食べる	19 充実感	19 相手	19 子供	
20 財産	20 命		20 とき	20 今	20 いられる	20 親	
21 本	21 夢		21 最低限	21 夢	21 金銭	21 生活	
22 勉強	22 維持		22 一番	22 大切	22 両親	22 事	
23 食べ物	23 充実		23 人間関係	23 財産	23 能力		
24 自分自身	24 社会		24 両親	24 命	24 地球		
25 知識	25 収入		25 して	25 関係	25 夢		
26 余暇	26 毎日		26 その他	26 生きる–生きて	26 すべて		
27 主人–夫	27 老後		27 満足感	27 平和	27 絆		
28 地球	28 暮らし		28 絆	28 家族	28 好奇心		
29 コミュニケーション	29 妻		29 知人		29 知識		
30 含む	30 人間関係		30 良い		30 自己		
30 教育	30 誠実		30 車		30 充実感		
29 だけ	29 娯楽		29 親戚–兄弟		29 ない		
28 思う	28 経済		28 収入		28 安定		
27 充実感	27 プライベート		27 余裕		27 社会		
26 つながり	26 施設		26 会社		26 車		
25 関係	25 名誉		25 夢		25 人生		
24 社会	24 金		24 財産		24 会社		
23 信頼	23 いう		23 命		23 余裕		
22 生きがい	22 楽しむ		22 妻		22 親戚–兄弟		
21 人生	21 音楽		21 家		21 趣味		
20 余裕	20 優しさ		20 愛情		20 妻		
19 収入	19 信用		19 友人–仲間		19 家		
18 安定	18 家		18 時間		18 人間関係		
17 平和	17 思いやり		17 心	17 自由	17 子供		
16 命	16 両親		16 子供	16 生きがい	16 充実		
15 時間	15 仕事		15 充実	15 恋人	15 生きる–生きて		
14 ない	14 便利		14 思いやり	14 仕事	14 平和		
13 ゆとり	13 彼女		13 など	13 ゆとり	13 時間		
12 して	12 車		12 平和	12 ない	12 友人–仲間		
11 充実	11 人生		11 親	11 環境	11 恋人		
10 子供	10 人		10 自由	10 安定	10 環境		
9 思いやり	9 パソコン		9 信頼	9 事	9 ゆとり		
8 自由	8 恋人		8 生きがい	8 時間	8 家庭	8 愛情	
7 人間関係	7 交通	7 して	7 恋人	7 人	7 仕事	7 心	
6 家庭	6 友情	6 健康	6 趣味	6 自然	6 事	6 して	
5 事	5 つながり	5 友人–仲間	5 環境	5 趣味	5 健康	5 友人–仲間	
4 心	4 愛情	4 人	4 金	4 人間関係	4 して	4 環境	
3 環境	3 信頼	3 自分自身	3 生活	3 生活	3 家族	3 仕事	3 趣味
2 生活	2 友人–仲間	2 時間	2 健康	2 友人–仲間	2 自分自身	2 人	2 生活
1 人	1 家族	1 家族	1 家族	1 金	1 生活	1 趣味	1 自分自身

たかを観察する．情報量が膨大となるので，一例として，クラスター番号 9 に含まれる 64 名の一部，クラスター番号 15 に含まれる 19 名すべてを出力した (表 15)．なお，一覧の中の左端にある数値は単語有意性テストで得た単語の検定値から求めたクラスター基準値を付与してある．つまり，この数値が大きいほど，そのクラスターに有意な単語が多数用いられたことになる．個々のクラスターを特徴付ける単語の意味の解釈は省略して，ここで 2 つのクラスターについて，出力結果の回答文を例示する．

(i) クラスター 5 について

このクラスターは「人とのつながり，人間関係，人とのコミュニケーション」など，どちらかというとパーソナル・コミュニケーションに関連した発言が多いクラスターである．

表 15.　2つのクラスター内の回答典型例.

クラスター番号：5　（はじめの32名分）

クラスター基準値			特徴的な回答／サンプル
10.290	--	1	人
	--	1	
5.799	--	2	人 と 人 の 繋がり 。
	--	2	
4.142	--	3	人 と 人 と の つながり
	--	3	両親
4.137	--	4	愛する 人
	--	4	人 と の 付き合い
3.251	--	5	人 と の つながり
	--	5	健康
3.202	--	6	愛する 人
	--	6	金 、 時間 、
3.118	--	7	人 と の つながり
	--	7	・ 自然
3.092	--	8	人 と 人 と の つながり
	--	8	専門的 な 知識 ・ 情報
2.507	--	9	時間
	--	9	情報
2.424	--	10	人 と の 繋がり
	--	10	ちょっと した 贅沢 できる 程度 の お金 。 愛する 人 。
1.973	--	11	大事 だ と 思う 人 を 大切 に する こと 。
	--	11	
1.648	--	12	人 と 人 と の 触れ合い
	--	12	家族 、 友人 、 仕事 、 趣味
1.601	--	13	人 と の つながり 。
	--	13	お金 、 住む 場所 、 着る もの 、 食べる もの 、 生きる ため の 目標 みたいな
	--	13	もの 等 。
1.599	--	14	友達 関係 、 広く いう 人 と の 関わり 。
	--	14	今 の 仕事 。
1.540	--	15	お金
	--	15	住む 場所
1.442	--	16	自分 で いる こと 。
	--	16	安定 家族 人 と の つながり 愛
1.393	--	17	家族 の つながり 。
	--	17	ゆとり 。
1.303	--	18	充実感
	--	18	目標 達成
1.286	--	19	人間関係
	--	19	仕事 （ 技術 ） 、 人 の 笑顔
1.264	--	20	自分 と いう 個 を 大事 に する こと
	--	20	
1.257	--	21	家族 と の 絆 。
	--	21	人 と して きちん と 生きる と いう こと 。
1.223	--	22	家族 を はじめ と した 人 の つながり
	--	22	精神的 な やすらぎ 物質的 な 満足 コミュニケーション 金銭的 な 満足
1.216	--	23	人間関係 です 。 もう少し 詳しく 言う と 、 人 と 人 と の つながち 、 男女
	--	23	を 問わず 友人 や 知人 （ 当然 家族 も ） です 。
	--	23	大切 な 物 と いう より は 必要 な 物 に なります が 、 現実的 には 、 お金
	--	23	です 。 その他 は 、 その 人 その 人 の 生きて きた 歴史 です 。
1.143	--	24	人 を 信じさせる こと 嘘 を つかない こと
	--	24	両親
1.086	--	25	家族
	--	25	情報

表 15. (続き)

1.076	--	26	逢いたい と 思う 人 と 労せず して 逢いたい ときに 逢える こと 。
	--	26	食
1.016	--	27	人 と の つながり でしょう か 。
	--	27	家庭 や 仕事 、 健康 で いる こと 。 基本的 です ね 。
1.011	--	28	自分 を 支えて くれる 仲間 ・ 友人 。
	--	28	人 と の つながり 。 最低限 生活 できる 環境 。
0.935	--	29	人人 と の つながり
	--	29	自然 に 対する 謙虚 さ
0.906	--	30	友人 （ 人脈 ）
	--	30	住む 場所 仕事
0.900	--	31	目標 設定 と その 達成 。
	--	31	こころ の 充足感 。
0.881	--	32	自分 。
	--	32	生き方 ・ ・ ・ と いう か 、 目標 と いう か 「 やりたい こと 」 が 有る と
	--	32	云う こと 。

クラスター番号：15 （19名分すべて）

クラスター基準値			特徴的な回答／サンプル
8.775	--	1	彼女
	--	1	
5.016	--	2	金
	--	2	名誉 、 地位
3.691	--	3	子供
	--	3	彼女 お金
3.009	--	4	自然
	--	4	お金 地位 名誉 プライド
2.978	--	5	家族
	--	5	金 ・ 名誉 ・ 地位 ・ 彼女
2.975	--	6	家族
	--	6	お金 、 名誉
2.137	--	7	彼女
	--	7	日本 、 車 、 今 の 生活
1.866	--	8	家族
	--	8	お金 時間 友人 社会的 地位 知識
1.785	--	9	才能
	--	9	時間 、 お金 、 名誉 、 学歴
1.753	--	10	家族
	--	10	両親 、 今 の 地位
1.680	--	11	彼女 と 仲良く 暮らす こと
	--	11	健康
1.632	--	12	彼女
	--	12	家族 仕事 自分 の 目標
1.505	--	13	健康 です ね 。
	--	13	人間関係 。 お金 。 肉親 。 社会的 地位 。
1.384	--	14	命 。
	--	14	彼女 。 お金 。 親 。 家族 。 宝物 。 飲食物 。 自然 。
1.207	--	15	家族
	--	15	社会的 評価
1.105	--	16	家族
	--	16	友人 人間関係 財産 教養 名誉
1.038	--	17	今 の 自分 を 変えて いく
	--	17	景気 回復 彼女 を つくる
0.731	--	18	地 に 足 を 着けた 生活 。
	--	18	彼女 ・ ・ 健康 ・ 収入
0.488	--	19	学校 を 卒業 する 事 （ 大学 4年 です ）
	--	19	彼女 、 自分 の 生活 スタイル （ ポリシー ）

表 16. コンコーダンスの例.

	家族 など と の 関係)
精神的 な 安らぎ	家族
暖かい	家族 生活 できる 程度 の お金
	家族 仕事
	家族
	家族 お金
	家族 の 交流 お金 家 愛情
	家族 仕事
考えた こと も なかった が こども が 生まれて 初めて「	家族 」だ と 思った
自分 の 能力 (これ が ない と 何 も 出来 ない から)	家族 友達 お金 時間
	家族
	家族 の 健康 と 幸福
	家族 お金
	家族 皆 が 健康 で ある こと
	家族 仕事
自分	家族
	家族 の 健康
等々) ・ 人間関係 (	家族
いのち	家族
	家族 だ と 思います
人 と 人 と の 触れ合い	家族
	家族 親友
生き甲斐	家族
	家族 人 から の 信頼
安定 した 生活	家族 ・ お金 (財産)
	家族 仕事
	家族 人 と 人 と の 信頼 関係
	家族 友人
	家族
	家族 健康 と 両親
いのち	家族
	家族
	家族 の 健康 今後
時間	家族
	家族 が 毎日 楽しく 暮らして いける こと だ と 思います
	家族 と の 時間 プライベート な 時間
	家族 の 健康 仕事 の 安定
	家族 平和 健康 目的 趣味 自然 心
くれる 人 が いる こと … つまり 友達 や 会社 の 同僚 や	家族 や 恋人 など 信頼 しあえる 人間 が いる こと が 一番 大切
	家族 友人 お金 健康
	家族 人的 ネットワーク
	家族 親 兄弟
余暇 を 好きな 趣味 を して 過ごす 時間	家族 友人 お金
	家族 大切 な 物 を 維持 する 収入

(ii) クラスター 15 について

一方，このクラスターは「社会的地位，名誉，彼女」といった語が中心となっている.「人間関係」は，ここにもあるが，語の使われ方が少し異なることが見える．他のクラスターについても，同様な考察が可能となる．単に類型化の数値結果を示すだけでなく，元の自由回答の内容に遡及して解釈を行うことが InfoMiner の特徴である.

(3) 用語検索と回答パターンの意味

今までにみた多次元データ解析的なアプローチに限らず，探索的にデータの特徴を知る操作の工夫は可能である．その一例がいわゆるコンコーダンス (用語検索機能) である．コンコーダンスは，コンテンツ・アナリシスや KWIC の基本操作で，指定した単語を基準に原文を検索・ソートし出力する．ある特定の語が，自由回答文の中でどう用いられたかのパターンを知るにはいままで述べた方法で，その大要を知ることができる．コンコーダンスはこれに比べて原始的な操作ではあるが，重要語を知った後や，意味不明の語があるときなど，この機能を用いてデータ探査することは有効である.

表 16 の例は，単語「家族」で検索してヒットした 520 例の一部を示したものである．こうした探索的機能により，回答文がどのように表記されたかを容易に観察することができる．

6. むすび

既述のように，既存の言語処理技法 (とくに形態素解析) と統計解析の諸要素技術を適切に組み合わせることで，従来の個々の方法論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としての新たな方向への示唆を示すとともに，簡単な事例解析を通じてその適用可能性を検証した．

インターネットの普及に連動して，自由回答等の文章型・テキスト型データの適切なデータ解析手法の登場が期待されている．データマイニングに関連してテキスト・マイニングの手法が重用される事が良い例であるが，これらの技法はあまりにも調査の現場から乖離している．ここに，従来型の調査の長い歴史に裏付けされた科学的手法の援用を受けた新たな自由回答取得方法の研究が必要とされる大きな理由がある．しかしながら，調査における自由回答の取得方法や，そのデータ解析技法の方法論の研究はそう進んでいるとはいえないことも事実である．現実の現象解析にあたってもっとも重要なことは，問題とする対象を説明するためにもっとも適したデータ取得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ると考えてきたが，この報告をその提案と事例としたい．

参 考 文 献

(株)博報堂インタラクティブカンパニー編 (2000). 『インターネットマーケティング』，日本能率協会マネジメントセンター.

Hayashi, C. (1998). What is data science? Fundamental concepts and heuristic examples, *Data Science, Classification, and Related Methods,* 40–51, Springer Tokyo.

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.

池上嘉彦 (1993). 『意味論 ── 意味構造の分析と記述 ──』，大修館，東京.

Iversen, G. R. (1991). *Contextual Analysis*, Sage Publications, Newbury Park, California.

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，東京.

加賀野井秀一 (1999). 『日本語の復権』，講談社現代新書 1459，講談社，東京.

北原保雄 (1997). 『概説日本語』，朝倉書店，東京.

小池清治 他 編 (1997). 『日本語キーワード事典』，朝倉書店，東京.

小池清治 (1993). 『日本語はいかにつくられたか』，ちくまライブラリー 25，筑摩書房，東京.

Lebart, L., Salem, A. and Berry, L. (1998). *Exploring Textual Data*, Kluwer, Dordrecht.

村上征勝 (1994). 『真贋の科学 ── 計量文献学入門 ──』，朝倉書店，東京.

NTT コミュニケーション科学基礎研究所 監修 (1999). 『日本語語彙大系』(池原 悟，宮崎正弘，白井 諭 他 編)，岩波書店，東京.

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 Vol. J81-D-I No.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，東京.

長尾 真，黒橋禎夫，佐藤理史，池原 悟，中野 洋 (1997). 『言語情報処理』，岩波講座「言語の科学」，第 9 巻，岩波書店，東京.

松本祐治，今井邦彦，田窪行則，橋田浩一，郡司隆男 (1997). 『言語の科学入門』，岩波講座「言語の科学」，第 1 巻，岩波書店，東京.

McEnery, T. and Wilson, A. (1997). *Corpus Linguistics,* Edinburgh University Press, Edinburgh.

Oakes, M. P. (1998). *Statistics for Corpus Linguistics,* Edinburgh University Press, Edinburgh.

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

大隅　昇 他 (1996). テキスト型データの解析について，第 10 回日本計算機統計学会シンポジウム.

大隅　昇，丸岡吉人 他 (1997). 自由回答データの解析法についての提案──実験調査におけるいくつかの試み──，第 25 回日本行動計量学会大会.

大隅　昇，Lebart, L. 他 (1997). テキスト型データの統計解析システム──SPAD.T/J──，第 11 回日本計算機統計学会シンポジウム.

大隅　昇 (1997). 定性情報の客観的分析法，第 17 回 JMRA トピックスセミナ資料，1997.6.25.

大隅　昇 (2000a). 「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 (2000b). 定性情報のマイニング──自由回答データの解析──，*ESTRELA,* **74,** 5 月号，14–26.

Ohsumi, N. (2000). From Data Analysis to Data Science, *Data Analysis, Classification, and Related Methods,* 329–334, Springer, Heiderberg.

佐藤武義 (1997). 『概説日本語の歴史』，朝倉書店，東京.

Sphinx Development (1998). *Sphinx Survey*: *Plus2 & Lexica Editions, Software for Surveys, Statistics and Text Analysis, Reference Manual Version 2 for Windows,* SCOLARI, Sage Publication Software, London.

山本夏彦 (2000). 『完本文語文』，文藝春秋社，東京.

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.

Analyzing Open-ended Questions: Some Experimental Results for Textual Data Analysis Based on InfoMiner

Noboru Ohsumi

(The Institute of Statistical Mathematics)

Ludovic Lebart

(Département Economie, Gestion, Sciences, Sociales et Humaines, ENST)

Interest in methods for the acquisition and analysis of textual data is increasing, as electronic processing of the Japanese language becomes possible, and research efforts advance in the analysis of natural languages and in related studies. Objective and reliable techniques are required for the impartial analysis of responses to open-ended questions, particularly in surveys of social attitude and public opinion.

First, we discuss what should be solved during data acquisition, based on our experience in analyzing open-ended survey questions. Second, we summarize comparisons between the statistical data analysis thus evaluated and more conventional approaches for Japanese textual data analysis. Third, we introduce the statistical system, "InfoMiner", developed to analyze textual and related data obtained from quantitative questionnaires. InfoMiner is based on the "data science" paradigm, which re-asserts the priority of data collection methodologies in any data analysis. In particular, InfoMiner includes some functions developed specifically for analyzing the agglutinative aspect of Japanese textual data. For example, there are functions for parsing any Japanese sentence as a set of linguistic units using a morphological analysis technique. These sets of data are used as dictionaries for statistical analyses. They are then amenable to analysis by multidimensional procedures such as correspondence analysis and clustering procedures. Finally, we use InfoMiner to illustrate a partial analysis of textual data, which were acquired from some actual surveys that were originally designed and conducted for data science based on the Internet. We provide a few examples to illustrate the practicality of this type of multidimensional data analysis.

Key words: Analyzing responses to open-ended questions, textual data analysis, InfoMiner, Internet survey, morpho-logical analysis, word segmentation, text-mining, data science.