非定型データの処理・分析について

―― 自由回答を中心として ――

高橋 和子

Although open-ended questionaires are effective in social surveys, there is a problem in analyzing them statistically. As a result in large-scale surveys researchers tend to avoid them.

In this paper, I propose two computer systems which support the construction of a data matrix from open-ended questionaires necessary for statistical analysis. One is a category-generating system, and the other is a coding system. The statistical methods for analyzing them are discussed herein.

1．はじめに

さまざまな社会調査の進め方のうち，意識調査や世論調査の特徴は，①大量回答者，②斉一的データ収集，③統計処理の3点に要約できる(原 1992)[6]．これらは，調査データを分析単位と変数から構成されるデータ行列(注1)に変換することが必須であるが，非定型データとは，「容易にデータ行列を作ることができないようなデータ」をいい，何らかの理由によりアフターコンディングが必要な「不定型データ」と，情報自体は定型であるが各サンプルにより長さが一定しない「定型不定長データ」の二つが含まれる（原 1988)[5]．

前者の例としては，調査における自由回答（被調査者の自由な回答をその言葉通りに記入したもの）や，これまで内容分析が研究対象としてきたような文章データがあり，後者には職業経歴データ（被調査者が現在までに携わってきた職業の変遷）(注2)がある．この他にもさまざまな非定型データが存在する中で，本稿では，特に自由回答に関する処理・分析方法について検討する．その理由は次の通りである．

これまでの状況を見る限り，自由回答は調査票の段階では用いられていても，処理・分析の過程で結論に生かしきれず，結果的に，眺められただけで終わってしまうケースが非常に多い．被調査者に負担をかけて収集したデータであるにもかかわらず，そのまま捨て去られるというのは残念なことである．

特に，調査環境が悪化しつつある現在においては，収集されたデータの共同利用が考えられるべきであり，その意味で，分析者側ではなく，被調査者側の枠組みで産出された自由回答は貴重な存在となる．これを積極的に利用できるような方法論が，早急に検討されるべきではないだろうか．

以下，2節で自由回答を選択回答と比較しながら考察し，その有効性を示す．3節で自由回答を処理・分析する過程で生じる問題点を明らかにし，4，5節でそれらの解決法について検討する．最後に6節でまとめを行う．

2．自由回答と選択回答

調査データには，自由回答の他に選択回答

Kazuko TAKAHASHI 千葉敬愛短期大学国際教養科

がある(注3). これは, 調査票にあらかじめいくつかの選択肢を記載しておき, 被調査者にその中から適当なものを選ばせて回答してもらうもので, 自由回答に比べると, 収集後の処理が楽なのため, うまく選択肢を用意しておけば, 質のよい情報を手軽に得ることができる.

従って, 一般的な傾向として, 大量のサンプルを必要とするような本格的な調査には選択回答が用いられ, 自由回答は, その準備として探索的に少数のサンプルを収集して分析する場合にのみ用いられる. もし本格的な調査に用いられた場合でも, 統計的に処理することはせず, 代表例をそのままの形で提示する程度である.

しかし, 調査データが大量サンプルとして収集される場合, 自由回答も選択回答と同様またはそれ以上に大きな意味をもつはずであり, 収集後の処理の煩雑さにより放置されるべきではない. 選択回答なら統計処理, 自由回答なら予備調査または事例提示という堅固な図式にこだわらず, 自由回答を統計的に処理するための方法論について, 活発な議論が行われるべきではなかろうか.

ここでは, 自由回答に関連するこれまでの議論について検討する中で, サンプリングされた自由回答の有効性を示し, 統計処理のための方法論の必要性を強調したい.

(1) 多段分析法に対する批判

最初に, 少し古くなるが, 見田(1965)[14]による多段分析法を取り上げよう(ただし, 用語的には質的データ, 量的データとしている).

ここでは, 質的データはおもしろいが, 代表性のなさと恣意性の点でたしかではなく, 量的データはおもしろ味に欠けるがたしかである(注4)とすることから, これらのデータを処理・分析するときには, 質的データを無理に数量化したり, 中間的な妥協形態を見出したりせずに, 両者をうまく使い分けて, 多段階に分けた分析を行うべきであると結論づけている.

前述した回答形式の違いによる分析の図式も, 単に作業の煩雑さのためだけでなく, この多段分析法による影響が大きいものと思われるが, 彼のいう質的データは日記や投書などを対象としており, 正しくサンプリングされた自由回答を含まないことに注意する必要がある(表1). すなわち, ここには, 本稿の議論の中心である「代表性をもった」自由回答というものが存在していないのである.

とはいえ, 両者の本質をうまく突いていることも事実である. 確かに, 自由回答は被調査者自身の枠組みで述べられており, 分析者側の枠組みで設定された選択回答にはないおもしろさをもってはいるが, これを処理する場合に, 恣意性の入り込む余地が十分にある. すなわち, 次節で述べるように, 妥当性と信頼性の問題が常に存在して, 大きな障害となっているのである.

(2) 安田, 小嶋による研究

次に, しばしば言われることであるが,「可能な限り多くの選択肢を用意しておけば, 自由回答は必要ない」とする議論について検討してみよう. これに対しては, 次のような問いすなわち, 同一の質問に対して回答形式を変えた場合に, 得られる情報の種類に違いはないのかということで考えてみる.

これに関する研究例は少ないが, 安田(1970)[30]の質問紙におけるワーディング実験によると, 理由を尋ねる質問において, 選択回答は, ①(選択肢が)網羅的でない, ②

表1 サンプリングされた自由回答の位置づけ

データの種類 代表性	質的データ	量的データ
あり	サンプリングされた自由回答*	調査データ
なし	日記, 投書など	——

*は多段分析法の議論に含まれない.

(規範となるような)正しい回答を促すという欠点をもつことが確かめられている（注5）.

小嶋(1975)[11]は，商品の広告効果の調査において，自由回答は再生知名率(記憶の深さ)，選択回答は再認知名率（記憶の広がり）という異なる側面を測定しており，両者は優劣を問題にするのではなく，その補完的効能を検討するべきであるとする．これによると，たとえ選択肢を網羅的に用意できたとしても，研究目的によっては，不適当なデータを収集する可能性があることになる．

これらの結果から見る限り，自由回答と選択回答は決して互換可能なものではなく，分析の目的に合致した情報を得るためには，回答形式は慎重に決定されなくてはならない．自由回答の存在理由が十分に認められるのである．

(3) 職業データの場合

次に，階層移動研究における職業データを取り上げよう．ここでは，職業を直接被調査者に尋ねることはせず，自由回答（仕事の内容，従業先の事業内容，事業先名称）と選択回答（従業先の規模，地位，役職）の計6種類のデータで収集した後，専門家がそれらを総合的に判断して決定する．

判断は主として仕事の内容を中心に行われるが，これに自由回答を用いる理由は，選択回答を網羅的に作ることができないということではなく，被調査者自身の分類能力に対する信頼性の問題があるからである．階層移動研究においては，職業は重要な調査項目のため，非常に細かく（約300種類）また正確に分類される必要がある．そこでは高度に専門的な知識や複雑な判断が要求されるために，このような方法が必要なのである．

(4) データの共同利用

最後に，データの共同利用という観点から自由回答の有効性を強調したい．

前述したように，近年，調査拒否などの増加により調査自体行いにくい状況になりつつある．今後，さらにこの傾向が強まることが予想されるため，データの収集はますます困難になるものと思われる．従って，収集できたデータの十全な活用が講じられるべきで，その最も有効なものはデータの共同利用であろう．

例えば，ミシガン大学政治学センターに設置されているNES (National Election Studies)では，アメリカの大統領選ごとに候補者に対するイメージを自由回答で収集しており，研究者は各々の目的に合わせてそれを利用し分析を行なっている(注6)．ここに見られるように，データの共同利用においては，自由回答の欠点とされる非構造性が逆に長所となり，さまざまな研究目的の違いに耐え得るのである．

以上で明らかなように，サンプリングされた自由回答は処理の煩雑さなどの問題点はあるものの有効性があり，調査の目的によっては不可欠な存在である．従って，統計処理のための方法論についての検討が是非とも行われるべきである．この目的のために，まず次節で自由回答の処理・分析過程を概観し，それぞれの過程における問題点を明らかにしておこう．

3．自由回答の処理・分析過程と問題点

(1) 自由回答の処理・分析過程

自由回答に対して統計処理を行うには，他のデータと同様に，データ行列を作成する必要がある．すなわち，(広義の)コーディングを行わなければならないが，これは，①分類カテゴリー(以下，カテゴリーと略す)の設定，②コードの決定，③回答のコード化の順に行う(原1984)[4]．ただし，あらかじめカテゴリーがあるものは，①，②を行う必要はない．

データ行列を作成できれば，次のような統計処理を行うことができる．まず，カテゴリー

ごとの単純集計を行ってデータの概要をつかみ，必要に応じて他の質問とのクロス集計を行う．

従来はこの頻度分析までしか行われなかったが，ダミー変数によりデータ行列を１－０行列に変換すれば，属性相関が計算できるため，各カテゴリー間の関連の度合がわかる．さらに，属性相関に基づいて適当な多変量解析が適用できる．

結局，自由回答の処理・分析もデータ行列作成と統計処理の二つの過程を経るわけだが，両者は性質が異なるために，今後は独立に検討することとする．

(2) データ行列作成における問題点

自由回答における最大の欠点はデータ収集後の作業の煩雑さにあった．また恣意性の問題も存在したが，これらはいずれもデータ行列作成過程で生じるものである．

① 作業の煩雑さ

回答形式がどのようなものであっても，統計処理を行うためにはデータを構造化する必要がある．自由回答の場合は，調査票を作成する時点で緩いコントロートしかかけない分（図１），後の処理が煩雑になるといえよう．選択回答の場合は構造化されたデータを収集できるが，選択肢を決定する手間がかかる．

作業全体で考えれば特に自由回答だけが煩雑であるともいえないが，何よりもデータの形態が自然言語であること，回答の内容が分析者の枠組みを越えたものである可能性が高いことから，その程度は比較にならない．

② 恣意性

恣意性は妥当性と信頼性の問題に分けられるが，社会調査が客観的なものであるためには重要な問題となる．これらは，カテゴリーの設定と回答のコード化の過程でその意味内容が異なるため，別々に検討することとする．

まず，カテゴリーの設定においては，設定されたカテゴリーの妥当性が問題になる．収集された回答を分類するのに，本当にそのカテゴリーが最適であるか（相互排反的）．回答から必要で十分な情報を吸収するのに，そのカテゴリーで尽くせるか（包括的）．

しかし，もしこれらを満足できても，同じ回答群に対するカテゴリーの設定の仕方は唯一ではなく，研究目的や分析者の視点の違いによりいく通りもあり得る．従って，絶対的に正しいものは存在するはずがなく，妥当性の評価は非常に困難である．最終的に優れた分析であることが認められたときに，初めてカテゴリーの妥当性が証明されたことになろう．

カテゴリーの妥当性は，本来，自由回答だけの問題ではなく，選択回答における選択肢についても議論されなければならない．すなわち，図１におけるコントロールの仕方が妥当なものであるかをチェックする必要があるが，実際にはこれについても評価が困難であり，また議論される機会もほとんどない．データ行列の作成過程は，回答が収集された時点

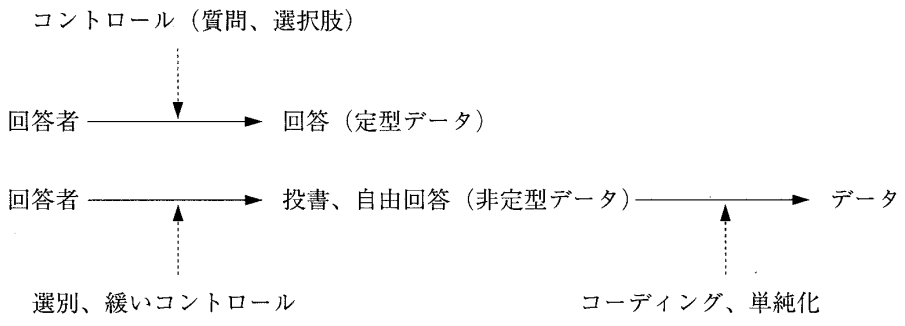

図１　データの作成過程（原 1992，pp.10）

から始まるのではなく,「手つかずの状態にある情報を収集して構造化する」過程の一部であることが認識されるべきであるが，ここではこの問題についてこれ以上深入りしない.

次に，回答のコード化においては，妥当性と信頼性の問題がある．基本的にはカテゴリーの妥当性と独立で，もし，妥当性の高いカテゴリーが設定されたとしても，コード化において妥当性と信頼性が低ければ，結果はでたらめなものとなる．回答のコード化における妥当性と信頼性に関しては，前述した2の(3)の場合を除いて，選択回答の方が高いことを認めざるを得ない.

一般に，設定されたカテゴリーの意味を熟知していれば，妥当なコード化ができるはずである．従って，分析者自身がカテゴリーを設定し，コード化を行う場合はそれほど問題はないことになるが，実際には判断に迷う回答は多いし，サンプル数がふえるにつれ，判断にゆれが出てくる可能性もある.その結果，一貫性が保証されなくなる恐れがあるが，それ以前に，分析者がコード化まで行うことについては，信頼性の点から問題がないわけではない.

大規模な調査の場合，回答のコード化は通常，複数のコーダーを訓練して行うが，コーダー達にカテゴリーの意味を徹底させるのは困難である．その結果，コーダーによる判断のばらつきが生じ，妥当性や信頼性の問題が累積されてしまう.

これを少しでも回避するためにコーディング・マニュアルが必須であるが，自由回答は意味の問題まで立ち入ることが多いため，詳細な内容のものを作成する必要がある.また，コーダーの訓練の手間もかかることから，結局，作業の煩雑さを増すことになる.

なお,カテゴリーの個数が非常に少ないか，または意味範囲が広い場合は，コード化の過程にそれほど大きな問題が生じないことは明らかであろう.

(3) 統計処理における問題点

さまざまな問題点を含んではいるものの，とにかくデータ行列を作成できれば頻度分析は行える．ここには何も問題はなく，場合によっては，むしろ「その他」の回答が少ないという利点があるほどである.

統計処理における問題は，多変量解析を行うときに生じるもので，1－0行列における1の出現頻度が低いこと，言い替えれば0の出現頻度が高いという現象である．これにより，必然的に相関係数が低くなり，多変量解析がうまく適用できなくなる．もちろん，選択回答においても，データの質によっては希望する解析手法が適用できるわけではないが，1－0行列におけるこのような現象は自由回答特有のものといえるだろう．理由としては，次の二つが考えられる.

一つは，自由回答の場合は，被調査者の関心の高いことがらが質問されたときはよいが，そうでない場合には回答率が非常に低くなる傾向が顕著なこと(高橋 1992 d[25], 佐藤 1993[19], Glass 1985[2]) (注7).もう一つは，カテゴリーを細かく設定する傾向があるために，回答が分散されることである.この場合，安易にカテゴリーの個数を減らしても，各カテゴリーのもつ意味範囲が拡散してしまい，自由回答で収集した価値がなくなってしまう恐れがある.

次では，それぞれの過程における問題点を解決法について検討する.

4．データ行列作成における解決法

データ行列作成過程における問題は，作業の煩雑さと恣意性であった．ここでは，解決法を検討するために，次の二つを基本方針とする.

まず，データ行列作成過程をカテゴリーの設定と回答のコード化に分解する．この理由は，自由回答には，階層移動研究の職業データのようにあらかじめカテゴリーが設定され

ているものとそうでないものの2種類があり，両者で作業を開始する過程が異なるためである．

次に，コンピュータによる支援を積極的に取り入れ，支援システムとして位置づける．これにより，作業の煩雑さを減らすこと，一貫性を保つことで信頼性を高めることが期待できる．

以上より，ここでの解決法は，「カテゴリー設定支援システム」と「回答コード化支援システム」を開発することとなる．注意することは，これらはあくまで「支援」システムであり，データを入力すれば自動的に結果が得られるようなものではなく，最終的な意思決定は分析者が行なうものである．

なお，使用するコンピュータは，ウィンドウ対応のパーソナル・コンピュータを想定した．

(1) カテゴリー設定支援システム

分析者は，研究目的に合わせて包括的かつ相互排反的なカテゴリーを設定しなければならないが，一般的にこれは試行錯誤的にしかできない．従って，ここでは，分析者の試行錯誤的な作業を助けることを目的とする．

これまで，分析者が紙と鉛筆を使用して(もしくは頭の中でだけ) 行ってきた作業を，コンピュータの特性を生かした形でどのように支援すればよいか．最低限，次の機能が必要であるとした((　)内は，使用するソフトウェアの種類を表す)．

① カテゴリー設定用サンプルの抽出
② カテゴリー設定用ファイル作成（エディタ）
③ カテゴリー設定（ウィンドウ）
④ カテゴリー決定（データベース）

① サンプル数が多いとき，カテゴリー設定のために全部の回答を対象とするのは大変である．10%から50%程度を無作為に抽出して設定すればよい．ただし，この場合，抽出されなかった回答から設定され得るカテゴリーの可能性がある（注8）．

② ①で抽出された回答を入力して，カテゴリーを設定するためのファイルを作成する．その際，多少の事前編集を行う必要があろう．ここでは，これを生データとよぶ．

③ 画面上に，「生データ」，「カテゴリー」，「コメント」の3つのウィンドウを開き，生データを見ながらカテゴリーを設定していく．その際，KJ法なども有効であろう．

このとき，設定したカテゴリーと生データの関係も保存しておく必要がある．また，コメントはカテゴリーを設定したときのメモや注意を書いておくもので，後でコーディング・マニュアルを作成するときに役立つ．

ここでは，カテゴリーをどんどん設定していけばよく，最終的には④で見直しを行って決定する．

④ カテゴリーが出揃ったところで，生データとカテゴリーの間を行ったり来たりしながら，場合によってはカテゴリーと生データの関係を変更したりして，最終的なカテゴリーを決定する．

そのためには，あるカテゴリーにどのような生データが含まれているかまたはその逆が簡単にわかることが必要である．すなわち，カテゴリーから生データを，生データからカテゴリーを簡単に引き出してこれるようになっていなければならない(図2)．具体的には，あるカテゴリーをクリックすると，それ

	生データ1
カテゴリーA	生データ3
	生データ8
	生データ2
	生データ5
カテゴリーB	生データ6
	生データ7
生データ1	カテゴリーA
生データ2	カテゴリーB
生データ3	カテゴリーA
⋮	⋮

図2 カテゴリーと生データの関係例

に含まれる生データがすべて表示されたり，生データをクリックすると対応するカテゴリーが表示できる機能が必要である．

カテゴリーはどうしても単純化された情報となるために，意味内容を正確に把握するためにはコメントが必要であり，例えばカテゴリーをクリックすれば，随時読めるようになっていなければならない．

決定したカテゴリーに分析者が満足した時点で，作業を終了する．

(2) カテゴリー設定支援システム（簡便形）の適用例

カテゴリー設定支援システムは，現在，まだ完成していない．本システムの構想を基本とし，③，④を簡便化したものを「参加者からみたセミナーのイメージ」調査（調査期間1991年11月から1992年1月，サンプル数約550）に適用したので，その概要と結果を示す（高橋1992ｃ）[24]．調査項目のうち自由回答で収集したのは，セミナーのイメージや参加動機を尋ねたもの（注9）などである．

カテゴリーの設定は次のような手順によった．

① カテゴリー設定支援システムと同様である．カテゴリー設定用サンプルとして10%の回答を用いた．

② カテゴリー設定支援システムと同様である．入力時の事前編集は，一つの回答中に複数の内容が含まれるものがあったため，各内容ごとにブランクなどで区切りをつけて入力した．これを自作プログラムにより，１内容／１レコードに変換後，カテゴリー設定用ファイルを作成した．

③ ウィンドウ機能は使わず，次のようにした．ただし，（ ）内は，志村（1992）[20]におけるMS-DOSのコマンドである．

- 内容を50音順にソートする（sort）
- 全く同じ内容は一つにまとめて頻度を付ける（uniq）

この結果は，１画面ですべてを眺めることができないので，印刷する．

④ 生データとカテゴリーの往復運動のためには，キーワードとなりそうな用語が回答中でどのように用いられているか知る必要がある．いわゆるKWIC（Key-Word-In-Context）であるが，次のようにした．

- 文字列の検索を行う（grep）．

この結果も印刷して一覧できるようにする．

簡便形を本システムと比較すると，特に③，④の機能が連動していないために，分析者が個々に結果を求めなくてはならない点が不便である．すなわち，簡便形ではウインドウを使用しないために，試行錯誤がダイナミックに行えない．

しかし，③で回答の整理ができ，全体の見通しがつくために，コンピュータの支援なしに行ったカテゴリー設定作業（高橋1992ｂ）[23]に比較すると，非常に楽であった．また，今回，回答のコード化は人手によったが，④の回答中でのキーワードの用いられ方の一覧が役立った．

設定したカテゴリーの妥当性の評価については，分析自体が未完成のため，設定されたカテゴリーと意味内容，頻度分布を掲げるにとどめる（表2，表3）．

なお，機能的に本システムと簡便形の中間に位置するものとして，文章解析用ソフトウェア（注10）があり，カテゴリー設定支援システムの一部として利用可能である．

(3) 回答コード化支援システム

カテゴリーが設定されると，次には回答のコード化を行う．ここでの問題点は，コード化の妥当性と信頼性であり，さらにコーダー達の訓練の問題もあった．これらを解決するために，コンピュータの直接的な利用，すなわちコンピュータにコーダーの役目をさせる方法を考える．うまく実現できれば，コード化の一貫性が確実に保証されて信頼性が高まる上に，人間のコーダーが不要となり，訓練

表2 セミナーのイメージのカテゴリーと頻度

カテゴリー名称	意味内容（例）	頻度（%）
気づき・自己啓発	気づき，自己発見，自己啓発，鏡，眼の前が開ける	26.4
やすらぎ	やすらぎ，楽な，家庭，母，ふるさと，息抜き	14.9
楽しい	楽しい，ワクワクする，天国，遊園地，おまつり	11.3
リフレッシュ	リフレッシュ，垢落とし，心の洗濯をする	6.4
さわやか	さわやか，クリアーな空気，大草原	1.5
癒し	癒し，病院，お助け村	3.8
解放・自由	解放，ストレスの発散，心開ける，自由	3.8
ふれあい	ふれあい，出会い，サークルの合宿	4.0
学校	学校，勉強，道場	6.9
人生の通過点	人生の通過点，線路のポイント，きっかけ，出発点	4.0
実験室	実験室，非現実，別世界，異次元	5.5
逃避	逃避，慣れ合い，保育園，自己満足	2.7
怖い	怖い，緊張，苦しい，刑務所，不安，異様	3.8
事実を述べる*	集団療法，人生の縮図，模擬社会	2.7

表3 セミナーの参加動機のカテゴリーと頻度

カテゴリー名称	意味内容（例）	頻度（%）
悩み・救済	(……に) 悩んでいた，行き詰まっていた，手助けを必要	19.5
問題を解決したい	(……の) 問題を解決する，現実を脱出する	4.7
自分を知る・自己啓発	自分を知る，可能性を見いだす，自己啓発	26.6
自分を変える	自分を変える，自分を成長させる	12.0
役に立つ	仕事の役に立つ，知識が増える	7.3
友人が欲しい	友人が欲しい，ネットワークづくりをする	2.2
好奇心・新しいもの	好奇心，面白そう，新しいものを体験する，気分転換	16.5
誘い・経験者をみて	信頼する人からの誘い，先に参加した人の様子をみて	14.6
仕方なく	仕方なく，義理で	4.6

の手間が省ける．

回答のコード化をカテゴリー生成設定システムに続いて行う場合は，そこで得られた生データとカテゴリーの関係（データベース）を利用するのが有効である．しかし，あらかじめカテゴリーが設定してある場合はカテゴリー設定支援システムを経由していないため，このようなデータベースが存在しない．カテゴリーの意味内容を明らかにし，生データとの関係を明示的にしてこれを作成する必要があろう．

しかし，本稿ではこれに関する検討を行わない．なぜなら，両者の関係をデータベースとして作成できれば，その後のコード化過程に関して特に議論の対象となるものがあるとは思われないからである．

ここでは，回答のコード化自体に問題がある場合，すなわちカテゴリーの個数が非常に多かったりまたはコード化の際の判断が複雑なものなどについて検討する．その際，両者の関係を単なるデータベースではなく「知識ベース」として捉え，それを利用したエキスパート・システムとして開発することを検討してみたい．すなわち，コンピュータに専門

家（ここでは分析者）と同等の知識を保有させて，非専門家（ここではコーダー）の意思決定を専門家に代わって行ったり，決定の過程が複雑な場合には専門家自身をも助けるものを回答コード化支援システムとして考えたい．

本システムを実用的なものとするためには，知識ベースや推論機構だけでなく，非専門家(ユーザ)でも使いやすくするためのユーザ・インターフェイス・モジュールや，結果を導きだした過程を説明する推論過程説明モジュールも重視されなければならない．

さらに本システムに特有の機能として，回答をいったん適当な用語に翻訳する辞書が必要である．この理由は，自由回答は自然言語であるために，同じ意味を表すにもさまざまな表現の仕方があり，それはまたカテゴリーを定義する用語とは抽象度のレベルが異なることが多いためである（注11）．

以上より，回答コード化支援システムは図3のような構成をとる．

ところで，回答のコード化は，対象の状況(各サンプルの回答)を分析対象の属性や事実と捉え，あらかじめ用意された仮説やカテゴリーの中からこれと照合するものを選ぶ過程であると考えられる．従って，本システムをタスクの型で分類すると，エキスパート・システムの中では比較的簡単な分析型（注12）で，しかもその中で最も典型的な診断型（注13）として開発できる．

エキスパート・システムにおいては知識は重要で，根本的な原理や原則である深い知識と，ヒューリスティックや教科書知識の浅い知識に分けて扱われる．本システムにおいては深い知識は必要なく，浅い知識のみを用いる．

従って，知識表現としては浅い知識の表現に適した「IF 条件 THEN 行動」なる

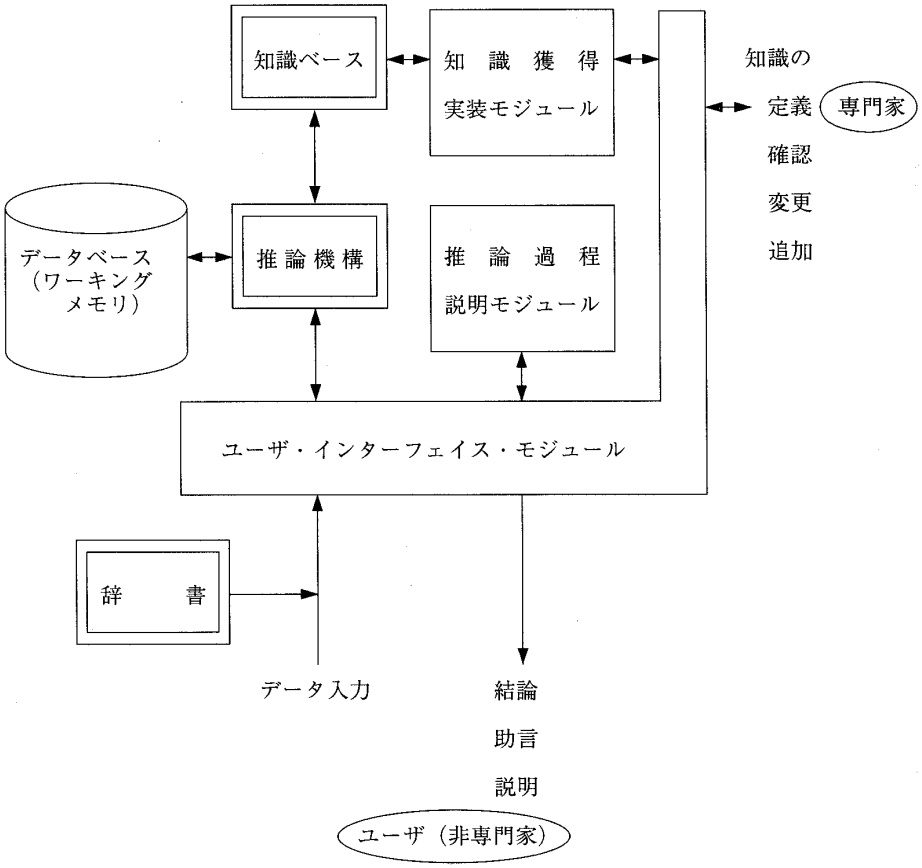

図3 回答コード化支援システムの構成（高橋 1994 b）

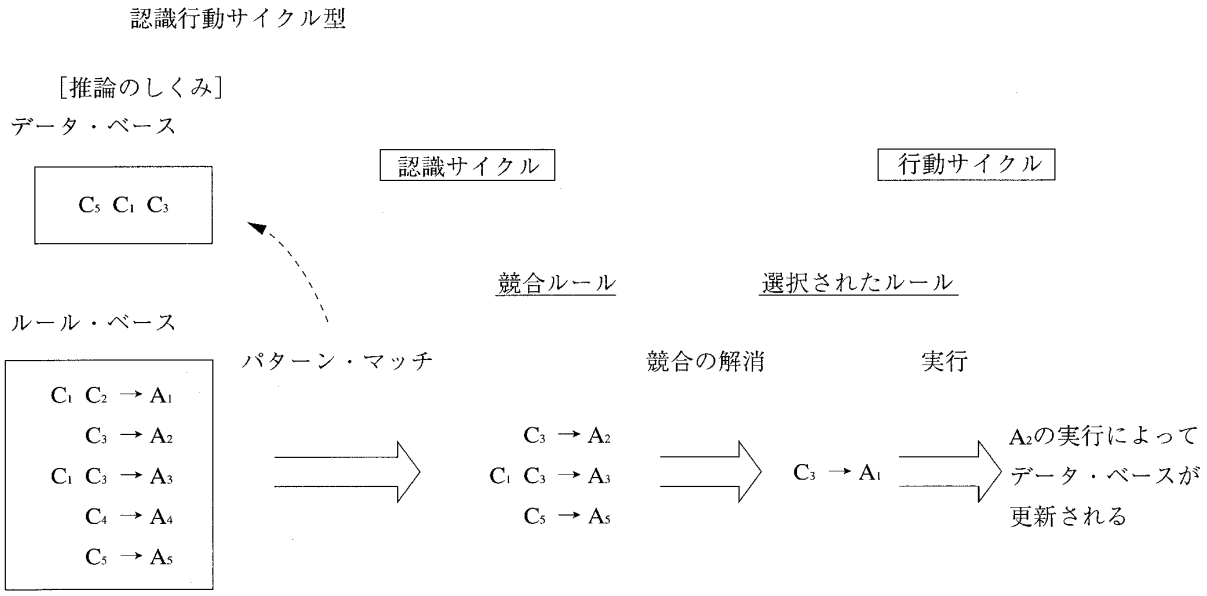

図4 プロダクション・システムにおける推論のしくみ（上野・小山 1988）

ルールで表現するプロダクション・システムを採用する．すなわち，「もし ～（回答） ならば ～（カテゴリーまたはそこに至る中間的な仮説） である」として，次々に推論を進めていく．プロダクション・システムにおける推論のしくみを図4に示す．

開発は，従来型の言語に比べて汎用性は劣るが，効率性が優れている知識表現言語OPS83を用いることとする．これは，特にプロダクション・システムの表現に向いており，他言語とのインターフェイスを取りやすいという利点をもつために，本システムを複数の言語で構成することができて都合がよい．

本システムの概要は以上の通りであるが，これを階層移動研究における職業データのコーディングに適用する例を示す．

(4) 回答コード化支援システムの適用例（SSM職業コーディング支援システム）

前述したように，階層移動研究においては職業分類は必須の作業であり，特に「SSM (Social Stratification and Social Mobility) 職業コーディング」と名付けられている．ここでは，293種類のカテゴリーが『SSM職業分類』（原 1993）[7] に明確に定義されており，6種類で収集されたデータを総合的に判断して，いずれかのカテゴリーに対応付ける作業が行われる．

ここでの問題点は，カテゴリーの種類が多すぎてコーダー達が理解しきれないこと，対象とするデータの個数が非常に多いこと（1985年の場合約4,000サンプル×10個）などによる作業量の多さである（注14）．また，判断の際用いられる知識が非常に複雑で，ごく少数の専門家しか知らないことから，作業全体が非効率な状況にあることが報告されている（佐藤 1992）[18]．

以下で，SSM職業コーディング支援システムについて述べる（詳細については高橋 1994 b [27]）．

一般に，エキスパート・システムを成功させるためには，高級な推論機構を備えるよりは知識ベースの充実がカギになるとされており，知識獲得は最も重要な過程となる．本システムでは，ヒューリスティックは専門家の経験則，教科書知識は『SSM職業分類』から獲得することとした．前者については専門家からのヒアリングを計3回行い，後者については，その内容を形式的に扱えることが確認できた段階である（高橋 1994 a）[26]．

推論機構は，競合するルールが複数ある場

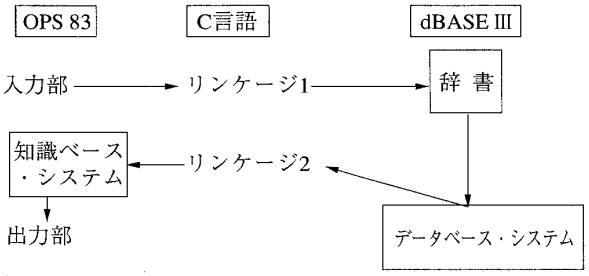

図5 SSM職業コーディング支援システムの言語構成（高橋 1994 b）

合は，明らかに解決できるものを除いた後は無理に一つに絞らず，残りをすべて提示して，分析者自身が最終的な判断を行う方針を採る．これは，分析者からの要望によるものであるが，システムとしてもこの方が容易である．

この他，本システムにおいてはデータベースによるマッチングの可能性もあるため（注15），これを併用することとする．従って，知識ベース・システムとデータベース・システムの混合型となる．

以上より，本システムは図5に示すような言語構成をとる．すなわち，知識ベース・システムは知識表現言語OPS 83，データベースはdBASE IIIで開発し，両者のリンケージはC言語で取る．

一般に，エキスパート・システムの開発は，一つの過程を固めてから次の過程に進む従来型の方法より，ラピッド・プロトタイピング（注16）による方が効率がよいとされる．従って，本システムもこの方法によることとする．

5．統計処理における解決法

ここでは，統計処理における問題の解決について検討する．単純集計やクロス集計などの頻度分布までは特に問題がなく，多変量解析を適用するために作成される1－0行列において，1の出現頻度が少ないというのが問題であった．ここでは，この問題を中心に検討する．

(1) 属性相関について

変数間の関連を表すものとして，通常はピアソンの積率相関係数を用いるが，これは変数がともに正規分布またはそれに近似できるものを仮定しているため，1－0のような2値変数の場合には使えない．測定値は2値でも，もとの分布に連続性を仮定できれば，4分相関係数を用いることができる（Miller 1986）[13]．しかし，高橋(1990)[21]においては.1以上のものがほとんどなく，いずれも非常に小さい値であった．

一般に，自由回答から作成した1－0行列は，変数に連続性を仮定せず，もともと2値であるとした方が自然であろう．従って，表4のような2×2表に示した数値により算出される係数を用いた方が適切である．ここで，表中a，b，c，dは変数Xi，Xjの特性のあるなしにより，両方ともありの比率をa，なしをd，どちらか一方がありでもう一方がなしをb，cとしたものである(0.00≦a，b，c，d≦1.00)．

表4において，注目する現象（マッチまたはノンマッチ）（注17）によりさまざまな係数が計算されるが，一般に，次式で定義されるϕ係数（四分点相関係数）が用られることが多い．

$$\phi=\frac{ad-bc}{(a+c)(b+d)(a+b)(c+d)}$$

ϕ係数はピアソンの積率相関係数の2値変数版ともいえ，－1と1の間の値をとるが，自由回答の場合には小さい値となることが多い．この対応として，次のような方法をとる．すなわち，

・表4の周辺比率を変えないで，ϕ係数を最

表4 2値変数の場合の組合せの比率

		変数 X_i		計
		あり	なし	
変数 X_j	あり	a	b	a＋b
	なし	c	d	c＋d
計		a＋c	b＋d	1.00

大にする．すなわち，次式により ϕ_{max} を定義する．

$a+c \leqq a+b$ の場合

$$\phi_{max}=\frac{(a+b)(b+d)}{(a+c)(c+d)}$$

・ϕ 係数を1ではなく，ϕ_{max} との関係でみるために，ϕ/ϕ_{max} により修正した ϕ 係数を修正 ϕ 係数として用いる．

修正 ϕ 係数を用いれば，相対的に係数の値は大きくなる．さらに，ϕ 係数をそのまま用いるより，撹乱要因となる因子を持ち込む可能性が減るという利点もある．しかし，自由回答の場合はこれを用いても，なお値が小さ過ぎることがある．

2×2表を用いた係数は，これ以外にも数多く存在する．属性相関の値を上げる次の策としては，それぞれについて算出の根拠も考慮しながら，適切なものを見つけることが重要である．例えば，佐藤(1993)[19]においては，マッチに注目したJaccardの類似性測度を用いている．

$$\text{Jaccardの類似性測度}=\frac{a}{a+b+c}$$

算出された類似性測度を順序尺度として扱い，非計量多次元尺度法を適用し，回答者を6つのグループに分類している．

さらに，属性相関の値を大きくする方策として，1－0行列を作成する際にカテゴリーの選別を行うことが必要である．その方法は一律ではなく，実際にいろいろなケースを試行して決めるしかない．一つの方法としては，頻度の多い順に並べて上位から見て行き，次順位との頻度差が大きいところまで残す．または，全サンプルの5～10%程度の頻度があるものまで残すなどの方法が採られる．

自由回答の場合は苦労してデータ行列を作成するためか，なかなか実行しにくいが，不必要にデータ空間を大きくしたままではうまくいかないことが多い（一般の多変量解析の失敗例を見よ）．

最後に，より基本的なことであるが，回答率を上げるために，例えば回答者が答えやすいような質問内容にすること，見当はずれの回答が出るのを避けるために，ある程度枠を狭めるなどの工夫が必要である．

(2) 多変量解析の手法について

属性相関に関連して，自由回答の分析に適した多変量解析の手法について触れておこう．現在のところ議論がほとんどなされていない状況であるが，これまでの研究例によると，因子分析により回答の解空間を縮約する(Miller 1986[13]，高橋1990[21])，多次元尺度法により被調査者の分類をマッピングする(佐藤1993[19])，クラスター分析により回答や被調査者の分類を行う（大井1992[16]）などの手法が適用されている．

適用される手法が用いたデータの水準に合っていれば，基本的には問題はない．妥当性については，カテゴリーと同様，得られた結果から評価するしかない．今後，研究例が増えた時点で，検討の対象としたい．

6．おわりに

本稿では，非定型データの代表的な存在である自由回答の処理・分析方法について述べた．支援システムについてはまだ構想段階にとどまっているが，今後は早期に実装を目指し，実際のデータに適用して評価を行いたい．また，統計処理については，属性相関に関するより深い研究が必要である．

なお，本稿では文章形式をとるような複雑な内容の自由回答に関する考察を行わなかった．これについては，今後，内容分析の研究と関連して行う．そこで，データにより方向や程度の強さなども情報として扱えることができれば，データの水準も上がり，適用できる統計処理の手法も広がるであろう．

また，自由回答など自然言語を扱う場合は，表層的な表現だけで深層の意味までつかみきれないのではないかという指摘がなされる

が，これについては，選択回答も含めた質問紙調査法に関する限界であるとして受け止めたい．

自由回答に関しては，汎用的な処理・分析方法というものが存在するのかという疑問も依然としてないわけではないが，個別に対応しながらも，そこでの知見を生かした積み重ねを行って何らかの共通点を見つけ，洗練された方法論といえるものを確立していきたい．

謝辞 カテゴリー設定支援システムの構想において小山照夫学術情報センター教授より貴重な御助言をいただき，SSM職業コーディング支援システムにおける知識獲得において原純輔東北大学教授により多大な御協力をいただいたことを記して感謝いたします．

注

（注1）データ行列は，次のものをいう（直井1983, pp.37）．ただし，Ui（i＝1，……．N）は分析　の単位，Vj（j＝1，……，M）は変数，Xijは　各変数に対する各単位の測定値を表す．

$$\begin{array}{c} \\ U_1 \\ \vdots \\ U_i \\ \vdots \\ U_N \end{array} \begin{array}{c} \begin{array}{cccccc} V_1 & \cdots & V_j & \cdots & V_M \end{array} \\ \left[\begin{array}{ccccc} X_{11} & \cdots & X_{1j} & \cdots & X_{1M} \\ \vdots & & \vdots & & \vdots \\ X_{i1} & \cdots & X_{ij} & \cdots & X_{iM} \\ \vdots & & \vdots & & \vdots \\ X_{N1} & \cdots & X_{Nj} & \cdots & X_{NM} \end{array}\right] \end{array}$$

なお，このデータ行列は，Shepardによる分類によればプロフィール行列である．

（注2）職業経歴データは，一般に，被調査者により個数が等しくないために，長さが一定ではない．

（注3）この他，回答自体は自由回答によるが，その場であらかじめ用意されたリスト（カテゴリー）に従って，調査員がコード化するプリコーディッド自由回答法と，調査票に記載された選択肢に序列を付けて回答する序列回答法がある．

（注4）量的データのおもしろくなさとは，追体験的な了解可能性の希薄，総合的・多次元的な把握の困難さ，変化のプロセスや可能性に関する動的な把握の困難さをいう（見田1965：pp.168）．

（注5）例えば，「あなたは何故現在の学科に進学しましたか」という質問に対して，自由回答では「就職に有利だから」や「他に適当な学科がないから」とするものが多いが，選択回答では「その学問に興味があるから」が断然多く，前2つの回答は少ない（安田1970, pp.51）．

（注6）Glass（1985），Miller（1986）など．

（注7）一般に，自由回答は選択回答に比較すると，文字にして表現するという積極的な行動をとるために，関心の低い質問に対しては無答が多くなる傾向があるとされる．高橋（1992 b）では最も関心の低いと思われる質問に対しては16%，高いものに対しては90%であった．佐藤（1993）では全体で10%であった．一方，Glass（1985）では，「カーター氏に投票したい訳が特にありますか（それは何ですか）」または逆に「カーター氏に投票したくない訳が特にありますか（それは何ですか）」という質問に対して，無回答率が非常に低かったとしている．

（注8）新しいカテゴリーは，回答のコード化過程の段階で設定されるために，③，④の入力方法を柔軟にしておく必要がある．

（注9）具体的な質問は次の通りである．セミナーのイメージ：「ずばり一言で表現するとセミナーはあなたにとってどんな場所ですか．」．セミナーの参加動機：「あなたがセミナーへの参加を決意した動機として最も強かったものは，今振り返ってみるとどんなことですか．自由にお答え下さい．」

（注10）「Micro-OCP文章解析プログラム」（沖電子技研）で，日本語版もある．

（注11）例えば佐藤（1992）によれば，回答では「海であわびを採っている」が，SSM職業分類では「海草・貝採取人」である．

（注12）分析型の他には設計型がある．これは結論の集合が確定していないために，要求仕様に合わせて構成要素を組合わせていかなければ

いけない．分析型と比較すると，非常に困難である（田中他 1987）．

（注 13）他に，制御型がある．推論や動作の実時間性が要求される点で，診断型より困難である（田中 1987）．

（注 14）1985 年調査の場合，約 20～30 人が 1 週間泊り込みで行った（原 1993）．

（注 15）SSM 職業分類においては，職業を定義した後に具体的な職名を記載するものが多いため，仕事の内容が職名で回答されている場合には，うまくマッチングする可能性がある．もちろん，これだけで職業分類を決定すると誤ってしまう．

（注 16）動作するだけの基本知識の用意ができたら，早期にプロトタイプを作り上げ，あとはその挙動を調べながら徐々にシステムの能力を高めていく方法のことをいう．

（注 17）マッチは，2×2 表において特性がともにありまたはなしの場合をいい，ノンマッチは片方がありでもう一方がなしの場合をいう．

参考文献

(1) Creecy R.H., et al.: Trading Mips and Memory for Knowledge Engineering, *Comm. ACM*, Vol.35, No.8, pp.48－63 (1992).

(2) Glass, D.P.: Evaluating Presidential Candidates: Who Focuses on Their Personal Attributes?, *The Public Opinion Quarterly*, Vol.49, No.4, pp.517－534 (1985).

(3) 福田収一他：OPS83 プログラミングテクニック，パーソナルメディア（1990）．

(4) 原　純輔・海野道郎：社会調査演習，東大出版会，(1984)．

(5) 原　純輔：非定型データの処理・分析，数理社会学の展開，数理社会学研究会，pp.461－471（1988）．

(6) 原　純輔：定型データと非定型データ，非定型データの処理・分析法に関する基礎的研究，文部省科学研究費補助金研究成果報告書研究（代表者　原　純輔），(1992)．

(7) 原　純輔：SSM 職業分類（改訂版），文部省科学研究費補助金研究成果報告書，(1993)．

(8) 伊藤陽一：原子力発電をめぐる論点の出現パターン，原子力発電に関する新聞論調の研究，社会工学研究所，(1975)．

(9) 伊藤陽一：世界の教科書にみられる自国イメージと他国イメージ，世界は日本をどう見ているか，日本評論社，(1987)．

(10) 石田　雄：内容分析による田中耕太郎最高裁長官の観念構造の究明，社会科学研究，Vol. 1，No.22，東京大学社会科学研究所，(1970)．

(11) 小嶋外弘：質問紙調査法の技法に関する検討，心理学研究法 9　質問紙調査，東大出版会，(1975)．

(12) クラウス・クリッペンドルフ：メッセージ分析の技法，三上俊治他訳，勁草書房，(1989)．

(13) Miller, H., et al.: Schematic Assessments of Presidential Candidates, *American Political Science Review*, Vol.80, No.2, pp.521－540 (1986).

(14) 見田宗介：現代日本の精神構造，弘文堂，(1965)．

(15) 直井　優（編）：社会調査の基礎，サイエンス社，(1983)．

(16) 大井　紘（編）：大都市に住む人々の生活環境に関する意識自由記述文の分析，国立環境研究所，(1992)．

(17) ラザーズフェルド：質的分析法，西田春彦他訳，岩波書店，(1984)．

(18) 佐藤嘉倫：職業コーディング支援システムの構築，非定型データの処理・分析法に関する基礎的研究，文部省科学研究費補助金研究成果報告書（研究代表者　原　純輔），pp.199－204，(1992)．

(19) 佐藤　裕：部落問題に関する「表現」の構造，解放社会学研究 7，pp.63－86，(1993)．

(20) 志村拓他：MS-DOS SOFTWARE TOOLS 基本セット，アスキー出版局，(1990)．

(21) 高橋和子：質的データの解析に関する一考察―政治意識調査における自由回答方の分析―，茨城大学人文学部紀要（社会科学），Vol.23，pp.21－50，(1990)．

(22) 高橋和子：自由回答における構造化支援システムの開発について，茨城大学人文学部紀要

(社会科学), Vol.25, pp.103－124, (1992 a).
(23) 高橋和子：日本人における政治リーダーのイメージ―自由回答法によるデータの処理・分析―, 非定型データの処理・分析法に関する基礎的研究, 文部省科学研究費補助金研究成果報告書（研究代表者 原 純輔）pp.137－164, (1992 b).
(24) 高橋和子：参加者からみたセミナーのイメージ―自由回答法によるデータの分析から―(文部省科学研究費補助金研究成果報告書 研究代表者 井上芳保), pp.79－101, (1992 c).
(25) 高橋和子：自由回答法に関する考察, 現代日本経済社会研究, Vol.12, pp.21－40(1992 d).
(26) 高橋和子：職業分類における自然言語処理について―職業コーディング・エキスパートシステム構築のための予備的考察―, 千葉敬愛短期大学紀要, Vol.16, pp.127－140, (1994 a).
(27) 高橋和子：社会調査におけるエキスパート・システム構築について―SSM職業コーディング支援エキスパート・システムの構想―, 千葉敬愛短期大学国際教養学論集, Vol.4, (1994 b).
(28) 田中博他：エキスパートシステム構築の方法, パーソナルメディア, (1987).
(29) 上野春樹・小山照夫：エキスパートシステム, オーム社, (1988).
(30) 安田三郎：社会調査の計画と解析, 東大出版会, (1970).