コンピュータによる自由回答の処理方法

髙橋　和子*

Computerized Data Processing of Open-Ended Questions

Kazuko TAKAHASHI

In conducting research with questionnaires, two main methods can be taken: a multiple-choice questionnaire in which respondents are obliged to choose from among multiple-choice answers presented by the researcher, and an open-ended questionnaire in which respondents may freely structure their own answers. Traditionally, multiple-choice questionnaires have been used rather than open-ended ones in large-scale surveys where statistical processing of quantitative samples is needed. The reason for this is that not only is complex after-coding necessary in open-ended questions, but also the reliability of the codings is not so easily guaranteed.

However, there is clearly information that cannot be obtained except through open-ended questions. Furthermore, it is

＊たかはし・かずこ：敬愛大学国際学部講師　情報処理論
Lecturer of computer science, Faculty of International Studies, Keiai University; information processing.

not desirable that response styles should be restricted by data-processing techniques. Active discussions about data-processing of open-ended questionnaires should be held, and methods for allowing applications to make use of the abundant information that questionnaires can provide must be developed.

In this paper, methods of processing open-ended questions that require statistical treatment of quantitative samples will be examined. Here, two strategies will be taken to solve the problems mentioned above. First, due to a lack of control by the researcher that inevitably results in different styles, information is classified and sorted into certain types instead of being treated collectively. Secondly, computers are actively adopted enabling higher efficiency of the processing and to assure consistency of results. Following these strategies, effective data-processing methods according to the types of open-ended questions will be suggested.

1. はじめに

社会調査を始めとする質問紙調査法においては、代表的な回答の形式として、分析者の枠組みによる選択肢をあらかじめ提示しておき、その中から強制的に選ばせる「選択肢法」と、被調査者自身の枠組みにより自由に記述させる「自由回答法」の2種類が存在する。従来から、大量サンプルで統計処理が必要な本調査においては選択肢法が用いられることが多く、自由回答法は全く用いられないか、用いられたとしても典型的なものが事例としてそのまま示される程度で、回答全体が統計処理や分析の対象とされることはほとんどなかった(1)。

この理由はいくつかあるが(2)、回答の処理過程に存在する次のような問題によることが大きい。まず、自由回答法により得られるデータ（自由回答）を統計的に処理するための定式化された方法が確立されていないこと。現状では、各調査ごとに分析者が独自に処理しているために、データに対する依存性が高く方法論としては汎用性に欠ける。次に、一般に統計処理

を行うにはデータをコード化（コーディング）する必要があるが、自由回答法においては「アフター・コーディング」と呼ばれるようにデータ収集後にしかこれを行えないため、作業が非常に煩雑となり、多大の人手と時間を要すること。さらに、コーディングの結果に対しても信頼性や妥当性が保証されにくいことなどである。

しかし、選択肢法にも欠点がないわけではなく（安田 1970、林 1975、浅井 1987）[3]、そもそも両者により収集されるデータの性質は異なっているとする立場もある（小嶋、1975）[4]。また、それらの議論を別にしても、自由回答法でなくては得られない種類の情報が存在することは明らかで[5]、回答の形式がデータ処理技術の面から制約を受けるのは望ましいことではない。これまでなおざりにされていた感のある自由回答の処理・分析方法に対するより活発な議論が行われ、自由回答のもつ豊かな情報の活用が図られるべきである[6]。

ここで、自由回答における処理方法の研究を主張するもう1つの理由として、昨今の調査環境における悪化の問題がある。現在、調査に対する拒否または不在などによる調査不能のケースはますます増大しており、質のよいデータの収集が困難な状況となっている。この対策の1つとして、研究者間における収集データの公開および共有化が図られるべきであり、このためには、個々の分析者の枠組みに依存しない自由回答の共同利用が有効であると考えられる。しかしこれを可能とするには、自由回答における処理方法がある程度は確立されていることが条件である。

このような認識に基づき、本稿では統計処理を想定した大量サンプルの自由回答に対する処理方法についての検討を行う。その際、前述した問題を解決するために、次の2つの戦略を取ることとする。1つは、分析する側からのコントロールが不能なためにさまざまな形態を取り得る自由回答に対して、一括して扱うことはせずに、最初にいくつかのタイプに分類を行うこと。もう1つは、作業の効率化を図り、結果の一貫性を保証するために、コンピュータの援用を積極的に取り入れることである。

以下、次節で自由回答における一般的な処理過程について述べた後、3

節で自由回答の分類を行い、4節でそのタイプごとに処理方法を検討する。

2. 自由回答の処理過程

ここでは、自由回答全体に共通する問題を検討する。本稿で対象とする自由回答は、大量サンプルとして収集され、最終目標として集計を含む種々の統計処理を想定するものである。これには前述したようにさまざまな問題が存在するが、これらは主にコーディングの際に生じる問題であると考えられる。したがって、本稿では自由回答の処理をコーディングに絞って検討することとする。

自由回答のコーディングを、「回答のもつ意味内容を理解して、妥当なカテゴリー(7)に分類しそのコードを付けること」と捉えると、次の2つの処理過程に分解できる。

（A）回答のもつ意味内容を理解する

（B）妥当なカテゴリーに分類しそのコードを付ける(狭義のコーディング)

ところで、探索的な性格をもつ調査の場合には、特にカテゴリーを定めないまま調査が実施されることが多く、その場合には、コーディングの前に、収集されたデータに基づいたカテゴリーを生成しておく必要がある。したがって、カテゴリーが存在しない場合には、

（C）カテゴリーの生成

が必要である。

この他、自由回答の処理にコンピュータを利用するためには、回答を生データのままコンピュータに入力して電子化しておく必要がある。したがって、コーディングを行う前に必ず、

（D）データ入力

をしておかなければならない。

以上より、コンピュータを利用して自由回答のコーディングを行うには、第1図に示す処理過程を経る。

各処理過程は完全に独立したものではなく互いに関連しているが、説明

第1図　自由回答の処理過程

カテゴリーがある場合　（D）→（A）→（B）

カテゴリーがない場合　（D）→（A）→（C）→（A）→（B）

の都合上、以下では、各処理ごとに検討を行う。ただし、説明の順序は、処理を行う順序に従っていないことに注意する。

1.（A）回答の意味理解

人間にとって簡単に理解できるような文の意味でも、コンピュータには理解できないことが多い。特に自由回答のように、人間が日常使用する言語で語られ（書かれ）、省略があったり、文法的にも正しくない可能性のある文の意味をコンピュータに理解させることは簡単ではない[8]。コンピュータには作業の効率化や結果の一貫性を保証する利点がある反面、人間のように、状況に応じて言語的および非言語的な知識を過不足なく柔軟に活用させることができないという欠点があるからである。

ここで、コンピュータにとって「意味」とは何であるかを定義をしておく必要があろう。言語に関する情報科学的な研究である自然言語処理や、より広範な人工知能、認知科学においては、「意味とは関係である」（松本他 1997）とするのが標準的な立場である。したがって、ここでは、文の意味を理解するとは、「文の要素である単語間の関係を何らかの構造として表現できること」であるとする。したがって、自由回答をどのような構造により表現すればよいかが明らかになれば、回答を理解する枠組みができたことになる。

なお、文の意味を単語間の関係として捉える以前に、当然、単語そのものの意味を理解する必要があるが、このとき、われわれには次のような日本語特有の事情が存在する。すなわち、文字種の多さによる表記の揺れ[9]、単語間の切れめのなさ（膠着語）、複合語・造語の多さ、語の位置の不定などの問題である。これに対する一般的な方策として、コンピュータに言語的な知識を記述した「辞書」を準備しておくことで、ある程度の解決がで

きる。

2.（B）狭義のコーディング

回答の意味を理解することができたと仮定しても、それを妥当なカテゴリーに分類することは、コンピュータのみならず人間にとっても困難な処理である。実際には、適切なカテゴリーが準備されてないことが原因となる場合もあるが、カテゴリーの定義が明確かつ具体的になされていたとしても、どのカテゴリーに分類すればよいかの判断は難しい。

コーディングに対する評価の観点としては、信頼性と妥当性の2つがあるが、妥当性については、分析の目的とも関係するために単純に論じることはできない。ここでは、妥当性に関する議論は避けて、信頼性についての検討を行う。

コンピュータと人手による場合を比較すると、後者には種々の問題がある。例えば、人間は知識を十全に働かせて柔軟に判断できるが、この弊害として、コーダーの恣意性が入り込む余地が大きく、判断理由が明確にされないまま決定される危険性がある。また、サンプルが非常に大きい場合には複数のコーダーが作業を分担することが多く、コーダー間で判断に揺れが生じる可能性が大きい。もしコーダーが1人でも、作業に長時間を要する場合には同様の傾向がある。これらの問題を防ぐには、少なくとも、詳細なコーディング用手引きを用意して、厳密にチェックする必要があるが、しばしば経験されるように、途中で内容の変更・追加があったときに、それをコーダー全員に徹底させるのは困難であるし、新しい枠組みでコーディングを再度やり直すのは大変である。

この点に関しては、コンピュータの方が優位にある。コーディング用手引きに相当する判断ルールをあらかじめ「知識」としてコンピュータに持たせることができれば、前述したような問題は生じない。また、途中で変更・追加があっても、その時点で新たな情報としてコンピュータに入力して、再度、最初から実行しなおすことは簡単である。

したがって、狭義のコーディングをコンピュータを利用して自動的にで

きるようにすれば、人間の作業量が軽減されるだけでなく、判断ルールの明示化が行われ、コーディングの一貫性が保証される。ただし、コンピュータはいわゆる常識を持たないし、蓄えられた知識以外のものは使えないため、どのような知識をどのように形式化すれば最も効率がよいかについてを十分に検討する必要がある。

3.（C）カテゴリーの生成

カテゴリーの生成は、従来から人手で行われてきた。したがって、時間や手間の制約から、通常、「全サンプル数の1/2から1/10程度の調査票をランダムに抽出して、自由回答の抜き書きを行い、その内容を通読して、分析の目的に沿って定める」（杉山 1984）ことが多い。しかし、この方法によれば、抽出されなかった調査票の中に、新しくカテゴリーを生成せざるを得ないような回答が出現する可能性があり、「（B）狭義のコーディング」過程において、新カテゴリーのもとでコーディングを最初からやり直すか、または無理に既存のカテゴリーのどれかに分類するかという判断を迫られる事態が生じる。したがって、この過程を人手で行っている場合には判断が揺れる危険性が高くなり、信頼性が下がる結果となる。

ここでもコンピュータの積極的な利用を考えたい。ただし、カテゴリーの生成においてはすべてをコンピュータにさせるのではなく、人間と協調する方向で役割分担をさせる必要がある。なぜなら、収集された回答からどのようなカテゴリー群が必要にして十分であるかという判断を下すことは、最も高度な知的作業であり、かつそこにこそ分析者の視点が発揮されるために、コンピュータには無理であり、人間（分析者）にしかできない。現状におけるコンピュータの能力を考えると、コンピュータはあくまで分析者が最終的な判断を下すための補助的な役割を果たすものとして位置づけられるべきであろう。結論に至る過程をコンピュータが支援することで、分析者は余分な時間や手間から解放され、より適切なカテゴリーの生成に専念できるはずである。言い換えれば、コンピュータは人間の知能の増幅機械（Intelligence Amplifier）としての役目をするべきである。

分析者は一度に適切なカテゴリー群を用意することは困難で、作業は試行錯誤的に行われることが多い。したがって、コンピュータの役割はこの「試行錯誤」をうまく支援して、分析者が満足のいく結論を引き出すことができるようにすることである。例えば、コンピュータの記憶容量が人間をはるかに上回る点を利用して、すべての回答を記憶させておき、それを表示させた画面を眺めながら作業を行うことができるようにすることが可能である。

カテゴリーの生成をコンピュータに支援させる利点をもう1点挙げると、カテゴリーの生成が完了した時点で、各カテゴリーの最終的な定義内容をコンピュータに記憶させておけば、次に行う「(B) 狭義のコーディング」過程において、そのまま利用できることである。これについては、3節で述べる。

ところで、コーダーが人間であるにせよコンピュータであるにせよ、カテゴリーを生成・定義する分析者とは別のものである。したがって、「(B) 狭義のコーディング」過程における処理を円滑に行うためには、カテゴリーの意味内容を具体的に明確にしておくことが重要である。すなわち、カテゴリーの概念の本質を抽象的に記述する（内包的定義）のではなく、具体例の列挙（外延的定義）をできるだけ多く行うことが有効である。

4. （D）データ入力

データ入力のためには、エディタまたはワープロ用ソフトウェアが利用できる。1サンプル分を1レコードとしておくと扱いやすい。注意点としては、電子化された回答をどのコンピュータでも利用できるように、テキスト形式で保存を行うことである。その意味では、最近試み始められてきたインターネットを利用した調査においては、回答が最初から電子化されているために、データ入力の手間が省けるという利点がある。

データ入力の際に簡単な事前編集を行っておけば、1で述べた問題点のいくつかを避けることが可能である。例えば、膠着語については、単語の区切りをスペースなどで明らかにしておく[(10)]、表記の揺れに対しては、

表記を統一しておくなどの対応を取ることができる。また、「など、等、一般」などの単語は、単に回答の冗長性を増す語であると考えられるため、特に分析の目的に関係がない場合は省略する方がよい。

3. 自由回答の分類

前述したように、自由回答の形態はさまざまである。処理を検討する際には、一律に扱わずに、回答の適切な分類を行ってタイプ別に行うことが有効であると考えられる。実際、データをあるタイプに限定してその形態を特徴づけることができれば、他分野において扱われる類似したデータとの比較を行うことができるために、そこでの研究成果を参考にすることが可能になる。

分類を行う基準としては、前述した「(A) 回答の意味理解」と、「(B) 狭義のコーディング」の処理過程から、次の2つの視点を導入するのが自然である。すなわち、

(a) 回答の意味内容が理解しやすいかどうか

(b) 分類するためのカテゴリーがあらかじめ用意されているかどうか

まず、回答に関する視点である (a) について述べる。コンピュータにとっては、回答の構造が形式的に単純であればあるほど理解しやすいことは明らかである。したがって、ここでは、(a) を「回答の構造が単純な形式かどうか」と捉え、最小の分析単位である単語の構成状況により次の3種類に分ける。ここで、各タイプに示した例は、いずれも髙橋 (1992b) における自由回答法の質問「セミナーを一言でいうと」(Q10) に対する回答の例である。

(a 1) 1つの単語により構成されるもの

例　あそび　オアシス　スタート　パーティ

(a 2) 複数の単語 (1つの文) により構成されるもの

例　ふれあいの場　心の遊び場　ありのままの自分が分かる

（a３）複数の文により構成されるもの

例　現在に至るまで、給料の大半は自己啓発のための投資につぎ込まれてきた。その最初のきっかけがセミナー。

各々の回答は、コンピュータ処理の点からは1レコードである。したがって、この分類はコンピュータからみると、1レコードの構成がどのようになっているかという問題である。

次に、カテゴリーに関する視点である（b）について述べる。これは処理工程に関わることで、コーディング前にカテゴリーを生成する過程の必要性の有無である。次の2種類に分ける。

（b１）カテゴリーがあるもの

（b２）カテゴリーがないもの

以上より、自由回答は（a）と（b）の組み合わせにより、形式的に第1表に示す6つのタイプに分けることができる。このうち、a３タイプについては、自由回答というよりも自由記述文と呼ばれることが多く、通常の調査においては、統計的な処理が想定される可能性がほとんどない。

ここで注意すべき点は、同じ質問に対する回答であっても、同一タイプの処理だけで対応できるとは限らないことである。例えば、(a）に関しては、質問の意図と異なるタイプの回答が混在する可能性がある。質問文に「一言でお答え下さい」との指示を添えていても、1つの単語（a１タイプ）ではなく複数の単語（a２タイプ）で回答されることは珍しいことではない（前述の例を参照のこと)。(b）に関しても、あらかじめカテゴリーが用意されていたとしても、該当しない回答が多数出現した場合には、新しくカテゴリーを追加する必要が生じることがある。または、カテゴリーの定義が抽象的にしかなされていないときには、現実に出てきた回答がどれに該当するのかがわかりにくいために、コーディングの過程で具体的に明確化していく作業が生じることがある。すなわち、b１タイプとされていても、b２タイプに変化する可能性がある。

したがって、第1表の分類はあくまでも基本的なものでしかなく、実際の処理に際しては、例えばタイプが混在する場合にはより複雑な方のタイ

第1表　自由回答のタイプ

回答の形式	カテゴリーあり（b1）	カテゴリーなし（b2）
1つの単語から構成されるもの（a1）	a1b1	a1b2
複数の単語から構成されるもの（a2）	a2b1	a2b2
複数の文から構成されるもの　（a3）	a3b1	a3b2

プとして扱うなどの柔軟さが必要である。

4. タイプ別の検討

ここでは、第1表に示した6つのタイプごとに、コンピュータを利用した具体的な処理方法を検討する。

1. a1b1タイプ

（回答が1つの単語から構成され、カテゴリーがあるもの）

全タイプ中、回答の構造が最も簡単であり、カテゴリーも用意されているために、作業が楽である。また、信頼性の点からも、単語が1つしかないために最も安全である。

このタイプにおいては膠着語の問題が生じないために、「(A) 回答の意味理解」については、単語の意味を回答の意味であるとしてよい（ただし、複合語の場合は複数個ある場合もある）。コーディングを行わなくても単語別の頻度集計ができるが、これはそのまま回答の頻度集計となる。注意する点は、表記の揺れのために、同一の内容であっても別の単語として扱われることがないように、データ入力の際の事前編集を徹底させることである。または、表記の揺れを吸収できる「辞書」[11]を用意しておくことが必要である。

「(B) 狭義のコーディング」については、各カテゴリーの定義内容が具体的に明らかにしてあれば、比較的容易に行える。すなわち、あらかじめ、各カテゴリーとそれに分類されると予想できる回答の集合を対にした「カ

第２図　カテゴリー対応辞書（コンピュータ上に実現できる形態）

（カテゴリー１　単語11　単語12　……）

（カテゴリー２　単語21　単語22　……）

第２表　カテゴリーと定義内容の例

カテゴリー番号	内　　容	カテゴリー番号	内　　容
1	能力　実力	26	金権体質
2	クリーン　清潔　汚職	27	年齢
3	理想論　現実論	28	穏和　笑顔　きつそう
14	意欲的　積極的	41	誠実　正直　まじめ
	バイタリティ		責任感

テゴリー対応辞書」と呼べるものを作成しておけばよい（第２図参照）。これは最初から完全なものでなくもよいが、辞書にない回答が出現するたびに、コンピュータにより簡単に（できれば自動的に）単語の追加がなされるようにしておくことが必要である。

カテゴリー対応辞書を作成するための具体的なデータ例を第２表に示す。ただし、これは、髙橋（1990）において、「与野党の政治リーダーに対する好感と嫌悪感」を尋ねる質問（問28－問31）に対する自由回答からカテゴリーを生成したもので、あらかじめカテゴリーの内容を定義していたわけではない点で、厳密には次節で述べるタイプ（a１b２タイプ）の例である。なお、この段階ではカテゴリーにはまだ名称が付けられていないために、カテゴリー番号とされている。

ところで、回答には類義語が多数出現する可能性がある。それらをある程度客観的に解釈してもよい場合には、すべての単語を辞書に登録しておくよりも、代表的な単語のみ登録しておいて、他の語については既に電子化されている日本語シソーラス(12)を利用すると便利である。ここで、シソーラスとは辞書の一種で、単語を木構造状に分類配列したものであり、意味的に近い語ほど近くに配置されるという構造をもつ（長尾 1996）。もし、分析者が独自の解釈を行いたい場合には、それに従ったシソーラスを独自に作成して利用すればよい。いずれにしても、類義語の出現可能性が高い場合には、シソーラスによる対応が必要である（第３図参照）。

第３図　シソーラス（コンピュータ上に実現できる形態）

（単語11　類義語111　類義語112　……）

（単語12　類義語121　類義語122　……）

（注）ここに示す単語の番号と第２図に示す単語の番号が同じものは，同一のものとする．

結論として、このタイプにおいては、あらかじめ「カテゴリー対応辞書」とシソーラスを作成しておき、回答を読んでは、「シソーラス」→「カテゴリー対応辞書」を検索するプログラムを作成すればよい。コーディング結果に影響を与えるのはシソーラスを含めた辞書の充実度であり、プログラム自体は複雑なものとはならない。もちろん、シソーラスが不要であれば、それを作成・検索する必要はない。

2. a1b2タイプ

（回答が１つの単語から構成され、カテゴリーがないもの）

ここでも、a1b1と同様に、単語をそのまま回答とみなしてよいために、「(A) 回答の意味理解」や「(B) 狭義のコーディング」は単純であるが、カテゴリーを生成する過程が必要となる。以下では、「(C) カテゴリー生成」について述べる。

まず、単語を五十音順にソート（並べ替え）する。その際、単語別の頻度も付けることができる[13]。分析者はこの結果を眺めて、回答全体の傾向をつかんでおく。

次に、回答をまとめあげながらカテゴリーの生成を行っていく。これは、回答の分散傾向が明確な場合にはカテゴリーの生成が直ちに終了する可能性もあるが、一般的には試行錯誤的な作業となることが多い。したがって、分析者が画面を見ながら、試行錯誤的にカテゴリーを生成していくことを支援するソフトウェアが必要である。このとき、どの単語をまとめてどのカテゴリーを生成したかを記録できること、すなわち第２図に示した「カテゴリー対応辞書」を作成できる機能が必要とされる。Windows など多重ウィンドウの表示が可能なものであれば、カテゴリー生成用とカテゴリー

対応辞書用のウィンドウ、さらには回答を表示したウィンドウを常時開いておくことができ、ウィンドウ間でのカット＆ペーストが自由にできるために便利である。

最終的にカテゴリーの生成が決定した時点における「カテゴリー対応辞書」は、コーディングの際に辞書として用いられるものである。このとき、もし、すべての回答にカテゴリーとの対応付けができていれば、残された処理は単にコードを付けるだけである。この場合は、カテゴリーの生成と平行して実質的にコーディングも行ったことになる。カテゴリーの生成を別の方法で行った場合には、「カテゴリー対応辞書」に相当するものがないために、「(B) 狭義のコーディング」過程の前にこれを作成しておく必要がある。

3. a2b1タイプ

(回答が複数の単語から構成され、カテゴリーがあるもの)

このタイプは、最もよくみられる自由回答の形式である。回答の意味が複数の単語により表現されるために、a1タイプのように、単に単語の頻度を集計しても回答の意味を理解したことにはならず、文の構造を解釈する必要がある。ただし、分析の目的によってはその必要がない場合もあり、さらには、a1タイプとみなすことができる場合には、形式的にはこのタイプであっても、実際にはa1タイプとして処理してもさしつかえない(第3表参照)。したがって、以下では、各処理過程について、それぞれ第3表に示す3つのタイプ別に検討することとする。

第3表　a2b1タイプにおける分類

タイプ	データの扱い方	回　答　例
a2b1Ⅰ	文の構造を意識する必要がある	レタスを作る，小学校で教える
a2b1Ⅱ	a1b1タイプと同様に扱える	営業等，事務の仕事
a2b1Ⅲ	上記2タイプの中間	ふれあいの場，心の遊び場

(注)　回答例は，a2b1Ⅰ，同Ⅱタイプは髙橋（1998a），同Ⅲタイプは髙橋（1992b）による．

（1）「（A）回答の意味理解」について

まず、a2b1Ⅰタイプ（文の構造を意識しなければならない場合）について述べる。これは、分析の目的により、回答の意味を最も生かす構造を考案する必要がある。ここでは、汎用性の点から自然言語処理による方法について検討する。

自然言語処理により1つの文を理解するには、通常、次の3つの処理、すなわち形態素解析、構文解析（統語解析）、意味解析を行う必要がある。ここで、形態素解析とは語の構造を同定するもので、文を単語に分けて、各々について品詞を決める。形態素解析においては、精度を高くするために大容量の辞書を構築する必要があることや、登録されていない複合語は複数の語に分割されてしまうことなどの問題もあるために、単語の分割だけを字面処理で簡単に行う方法もある。字面処理とは、日本語がさまざまな文字種により構成され、特に名詞は漢字やカタカナ、助詞などは平仮名で表記されることを利用するもので、文字種が変化するところ[14]を語の区切りと考えて同定する処理を言う。

構文解析とは、形態素解析で分けた単語間の関係により、文の構造を同定するものである。

意味解析とは、これらから文の意味を理解するものであるが、これには単語の意味を理解することと、単語間の意味（関係）を理解することの両方がある。意味理解の形式としては、格フレーム[15]、述語論理[16]、意味ネットワーク[17]などが用いられるが、これらはいずれも互換性がある。回答の意味内容を最もよく表現できる形式を選択することが重要である。例えば、髙橋（1998a）において、「本人の仕事内容」を尋ねる質問に対する自由回答の意味表現として、格フレームの形式を用いた（第4図参照）。

一方、回答を妥当なカテゴリーに分類する必要上、カテゴリーの意味も

第4図　回答の意味表現例（格フレームによる）

回答		回答の意味構造
レタスを作る	→	（作る　を　レタス）

第５図　カテゴリーの意味表現例（格フレームによる）

	カテゴリーの定義		カテゴリーの意味構造
599	農耕・養蚕作業者		
	野菜を栽培する	→	（599　栽培　を　野菜）
521	小学校教員		
	小学校で教える	→	（521　教える　で　小学校）

（出所）　髙橋（1998a）より.

何らかの形式により表現されている必要がある。ここで、カテゴリーの意味を表現するとは、カテゴリーの定義内容だけでなく、そのカテゴリーに分類するための必要な知識を表現することでもある。したがって、カテゴリーの意味表現は、知識の表現にも適した形式であることが条件である。

このとき、カテゴリーの意味を表現する形式は、回答の意味形式と共通または簡単に変換できるものにしておくことが重要である。例えば、第５図は髙橋（1998a）で用いられたカテゴリーの例であるが、その意味内容が回答と同じ格フレームにより表現されている。

回答やカテゴリーの意味を既存の形式ではなく、独自に表現してもよいが、このときも、両者の形式については同様のことが要求される。

次に、ａ２ｂ１ Ⅱタイプ（形式的にはこのタイプに分けられても、実際にはａ１ｂ１タイプと同様の扱いができるもの）について述べる。このタイプは、分析の目的によっては、片方の単語の意味を特に考慮する必要がないこともある。例えば、第３表の回答例において、「営業等」における「等」、「事務の仕事」における「仕事」なる単語は、仕事の内容を尋ねるというこの調査の目的からは、いずれも不要な語であるとしてよい。

この場合は、データ入力の際に、分析に深く関連する単語のみを取り上げて入力することにしておけば、ａ１ｂ１タイプとして処理を行うことができる。データ入力の際の事前編集を行いたくないときは、次のようにする。すなわち、データ入力は回答の通りに行うが、あらかじめ、「冗長語」として無視してもよい単語の一覧を作成しておき、入力後にこれにより除去する。実際、髙橋（1998a）においては、この方法によりデータの整備を

行っている。

最後に、a2b1Ⅲタイプ（a1b1Ⅱタイプより多少a2に近いが、a2b1Ⅰタイプほど回答の構造を意識する必要のない場合）について述べる。例えば、第3表の回答例において、分析の目的からみて、「心の遊び場」のすべての単語が重要であるとは思えないけれども、現段階ではどれを無視してよいかを簡単には判断ができないというような場合である。

これは、とりあえずそのままのデータを入力しておき、「（A）回答の意味理解」処理を次のように行う。まず、単語別の頻度集計をとり（これをワードリストと呼ぶ）、回答全体の傾向を眺める。その際、頻度を知りたい単語をこちらがキーワードとして指定することもできるし、コンピュータに任せることもできる。

このとき、KWIC（Key Word In Context；用語索引）を行って、単語の意味を確認しておくことが必要である。なぜなら、単語別の頻度集計は、単語が文脈から切り離されて処理されており、そのまま解釈を行うと単語の意味を取り違える危険性があるからである。ワードリストやKWICについては、利用可能なソフトウェアがある(18)。

これにより、重要であると考える単語と無視できる単語の峻別を行うことができれば、これ以後は基本的にはa2b1Ⅱタイプと同様の扱いが可能となる。

（2）「（B）狭義のコーディング」について

ここでは、a1b1タイプと同一視できるa2b1Ⅱタイプを除く2つのタイプについて述べる。

まず、a2b1Ⅰタイプにおいては、回答とカテゴリーの意味は、共通の形式により表現されているはずである。両者の構造を比較して、カテゴリーの意味を構成する各要素の単語について、回答側の対応する各構成要素の単語が類似または包含されるかどうかをチェックする。

例えば、第4、5図はいずれも格フレームに基づいた表現が行われているが、両者の構造としては、（作る　を　レタス）と、（599　栽培　を　野菜）が等しい。したがって、「栽培」と「作る」が類似することと、「野菜」に

第6図　回答のキーワード例

回　答		キーワード
ふれあいの場	→	ふれあい
心の遊び場	→	心　遊び

「レタス」が包含されることが示されれば、「599」（農耕・養蚕作業者）なるコードを付けることができる。この単語間の関係は、それぞれ既存のシソーラスを利用するか、またはあらかじめ作っておくことで解決できる。髙橋（1998a）においては前者の述語については、『分類語彙表』（国立国語研究所 1964）、後者の名詞については自作のシソーラスを用いた。

一般に、カテゴリー側の単語の方が回答側に比較して抽象度が高い傾向があり、単語レベルにおいては、a1b1の項で述べた「カテゴリー対応辞書」を利用する方法と同様である。

次に、a2b1Ⅲタイプにおいては、回答ごとに重要だと考える語をキーワードと定めて、a1b1タイプと同様に処理を行えばよい。このとき、2つ以上のカテゴリーにコード付けしてもよい場合には、1つの回答に対して複数のキーワードを定めてもよいが、それが許されない場合には、1つに絞る必要がある（第6図）。

結局、a2b1タイプは、分析の目的の違いや回答の傾向により、文の構造を強く意識しなければならないもの（a2b1Ⅰタイプ）、a1b1と同様に扱ってよいもの（a2b1Ⅱタイプ）、その中間のもの（a2b1Ⅲタイプ）の3種類があり、各々で処理方法が全く異なるために、タイプの見極めが重要である。

4. a2b2タイプ

（回答が複数の単語から構成され、カテゴリーがないもの）

a2b2タイプにおいても、前節における分類Ⅰ、Ⅱ、Ⅲをそのまま踏襲すると、a2b2Ⅰタイプ、a2b2Ⅱタイプ、a2b2Ⅲタイプに分類される。ここでは、a1b2タイプを踏襲するa2b2Ⅱタイプを除く2つのタイプについて、「(C) カテゴリーの生成」過程を検討する。

前述したように、カテゴリーの生成こそ分析者の視点が生かされるとこ

ろであり、分析者が異なれば、同じ回答群からでも異なったものが生成される可能性がある。a 2 タイプはa 1 タイプに比べると、含まれる情報が多いためにその傾向はさらに強い。したがって、回答が与えられれば、コンピュータによりカテゴリーが自動的に一意に決まるということはむしろ不自然で、生成の過程において、分析者自身によりさまざまなアイディアが試されるべきである。

これに対する有効な方法として、KJ 法がある[19]。KJ 法は、そもそも人類学において、野外調査で収集される非構造的で混沌としたデータを分類する目的のために開発されたものである。データをそれぞれのカードに書いておき、視点を変えてはいろいろに並び替えてグループを作り、最終的に決定されたグループに最もふさわしい名称を付けて内容を要約し、グループ間の関係図を作成することでデータの構造化を行う。現在は、人類学の分野に限らずさまざまな場において、発想を支援する方法として用いられることが多いが、自由回答の処理においては、カード 1 枚を 1 つの回答、グループをカテゴリーと考えれば、妥当なカテゴリーの生成を行う方法として利用できる。

KJ 法は手作業を想定しているが、パソコン上でこれを部分的に実現しようとするソフトウェアもある。例えば Ohiwa（1990）による KJ エディタはその 1 つである。

いずれも、得られた情報を体系化して図解し、最終的には問題解決を行う目的をもつものであるが、初期の処理は、十分にカテゴリー生成の支援となり得る。ここでも意志決定を行うのは人間であり、コンピュータはあくまでもそれを支援する道具としての役割を果たす。

カテゴリーの生成を終了したら、その定義内容を明確にしておく必要がある。このとき、第 4、5 図で示した例のように、回答の意味表現と同じ形式にするのが必ずしも容易なわけではないが、コンピュータによる処理を行うためには、回答と共有できる構造を見つけ出す必要がある。

5. a3b1タイプ

（回答が複数の文から構成され、カテゴリーがあるもの）

a3タイプは、回答の構造としては最も複雑であるが、意味を理解するための情報が最も多く盛り込まれている。したがって、それを利用した方法を検討することが有効である。

ここでは、コミュニケーションの分野で用いられることの多い内容分析の手法を参考にする。内容分析とは、「データをもとにそこから（それが組み込まれた）文脈に関して反復可能でかつ妥当な推論を行うための1つの調査技術である」（クリッペンドルフ 1989）。古くは新聞や小説などのテキストにおいて、ある一定の範囲内における単語の出現傾向を調べるという量的な分析により、その内容を把握することから始まったが、現在では言語やその他のシンボルを用いたコミュニケーション・データから推論を導き出す科学的方法へと発展している。内容分析が対象とするものは映像なども含めさまざまであるが、ここではテキストに絞って考える。

内容分析における記録単位としては、単語が最小のものであり、信頼性に関する限りは最も安全である。その他、指示対象となる特定の事物、事象、人物、行為、国や思想（これらを言及単位と呼ぶ）、さらに命題をも記録単位とすることができる。一般に、記録単位を大きくすればするほど、表層的なものからより抽象的なものへと分析を深めることができるが、信頼性は低下する傾向にある。特に命題を記録単位とするときは一定の構造をもつことを要求されるが、これはa2タイプにおいて述べた表現形式と関連する。

ここで、コンピュータ利用の可能性から、一般的な方法ではないが、伊藤（1987）において独自に開発された「対象評価分析」を採り上げる。これは、各国の教科書から国のイメージを分析するために、文章を「評価語」と「評価対象」の2つの視点により構造化するものである。評価語としては、広い意味の形容詞、形容動詞を考え、評価対象としては、国を4つの側面（政府など、政府以外の組織や人間集合など、文化など、自然など）から捉

えたものと国そのものの計5つを考える。作業は、文章中から評価語と評価対象を取り出し、両者の結びつき方をあらかじめ準備された3つの基本文型に書き直すことから始まる[20]。「対象評価分析」においては対象となる文章の性質が格フレームで扱えるものと異なるものの、構造を捉えようとする点に類似性がある。この方法における手順のコンピュータ化が成功すれば、自由回答の処理方法として有効であると思われる。

ここで、本稿の中心課題であるコーディング処理とはややはずれるが、コーディングがなされた後の内容分析の統計処理は多彩で、記録単位として定めたもの各々に対する頻度、2つのものの関連性・相関・クロス表などの単純な解析の他に、判別分析やクラスター分析などの多変量解析、コンティンジェンシー（随伴）分析[21]、文脈的分類法[22]がある。

したがって、このタイプにおける「(A) 回答の意味理解」については、1つの回答を1つの文書として扱えば、記録単位を定めてそれに対するいずれかの技法を用いることで、各回答の内容を把握することができる。ただし、全部の回答に対してこれを行うことは、時間と手間がかかるために、コンピュータによる自動化が必要な条件である。これについては、次節でも述べる。

「(B) 狭義のコーディング」については、このタイプにおいては、回答とカテゴリーの意味表現の形式を共通化することが困難であり、両者を自動的に関係付ける方法が不明であるために、現状では人手で行うしかない。ただし、単純にカテゴリーを特徴付けるキーワードを設定して、回答における頻度を求めるような場合には、前述したa1b2タイプと同様の扱いができる。

6. a3b2タイプ

（回答が複数の文から構成され、カテゴリーがないもの）

a3タイプは、自動化の点からみると他のタイプと異なり、カテゴリーがあるものよりないものの処理の方が容易である。その理由は、カテゴリーがないものにおいては、カテゴリーがある場合に必要なカテゴリーと回答

との形式的な関係付けが不要になるからである。

ここでは、最近、自然言語処理の分野で盛んになってきた文書の自動分類の手法を参考にする。この研究が生まれた背景としては、インターネットなどにより収集される電子化テキストの量が膨大になり、そのままでは人間の処理能力を超えるため、事前にある程度の分類の必要性が生じてきたという状況がある。

自動分類は、方法的には、内容分析における技法をコンピュータ上に実現したものであると捉えることができる。現在のところ、一部の技法しか自動化されていないが、例えば、人手またはコンピュータにより重要語を定めて、コンピュータによりそれに対する頻度調査や文脈的分類法を行って文書の内容を把握し、分類することが試みられている。

１つの回答を１つの文書として扱えば、この方法はそのまま a３b２タイプの処理に適用できる可能性がある。コンピュータからみれば、自由回答もテキストという点で文書の一種であり、この分野の成果は非常に参考になる。

結論として、一般に a３タイプは統計処理を想定されることが少ないが、コンピュータ処理の立場からみれば、内容分析など他分野で用いられているさまざまな技法が使えるだけの豊富な情報量をもっているために、今後より多くの方法が提案される可能性がある。

5. おわりに

本稿では、統計処理を想定した大量サンプルの自由回答に関する処理方法、特にコーディング方法について検討した。その際、戦略として用いたのは、自由回答を形式的に６つのタイプに分類することと、コンピュータを積極的に利用することである。これにより、これまで漠然としていた自由回答の処理方法を、部分的にではあるが具体的に提案できたものと評価できる。今後の課題としては、各タイプごとに実際の分析例を増やしながら蓄積を行い、その中からより洗練された方法を生み出していくことであ

る。その際、他分野における成果の適用性についても検討されるべきである。

自由回答の利用は、今後ますます高まるものと予想できるために、処理方法に関する研究の活発化が強く望まれる。

(注)

(1) 一般には、原・海野（1984）で指示されるように、自由回答法を予備調査で探索的に用いて、出現する可能性のある回答を得ておき、それに基づいて選択肢を作成して本調査を行うという方法が採られる。

(2) 自由回答法の欠点としてしばしば指摘されるのは、質問や被調査者の属性によっては回答率が低くなる傾向があること、見当違いの回答がなされる場合があることなどである。

(3) 安田（1970）の選択肢法と自由回答法を比較したワーディングの実験によると、選択肢法ではしばしば選択肢が網羅的でないこと、また選択肢の提示により正しい（とされる）回答を促す作用をもつことがある。林（1975）によると、ある事実を提示して選択肢法によりその知識の有無を確かめると、一般に肯定的応用が誘発されやすいとの指摘がなされている。浅井（1975）によると、選択肢の配列順序による影響があること、無理強いした回答を求めることが挙げられる。

(4) 小嶋（1975）によれば、商品の広告効果の測定やマインドシェア測定において、選択肢法は再認知率、自由回答法は再生知名率にあたる。例えば、前者はブランド名の浅い記憶ではあるが、その広がりを測るもので、後者はブランド名の記憶の深さを測るものと考えられる。

(5) 例えば、最近の企業や行政による調査にみられるように、消費者や住民の本音や意識を探る場合には、自由回答法が用いられることが多い。また、理由を尋ねる場合は、自由回答法でなければうまく聞き出せないことが多い。

(6) 最近、ようやくデータ解析に関する研究者の間において、自由回答の処理・分析方法に対する関心が高まりつつある。例えば、日本行動計量学会（1997年９月６日、於：仙台）において、特別セッション「自由回答のコーディング法と分析法」が組まれた。

(7) カテゴリーとは、定性的標識（属性）に関して個体がいずれかに分類される区分けのことである（安田・原 1982)。

(8) コンピュータが理解可能なのはプログラム言語のように形式が整っており、文法が簡単な人工的な言語である。しかし、日常的に用いられる言語は文法が複雑な上、省略表現、並列表現などが多用されたり、文法的に正しくない文（非文）も意味が分かれば認められるなど、形式化されにくい。

(9) 日本語における文字種には、平仮名、カタカナ、漢字がある。例えば、「りんご」、「リンゴ」、「林檎」は、いずれも通常、同じものを指していると考えてよい。また、送りがなの違いによる表記の揺れもある。

(10) 例えば、「ふれあいの場」という回答の場合、「ふれあい　の　場」と入力する。

(11) 例えば、(りんご　リンゴ　林檎）をひとまとまりとする辞書を用意しておく。

(12) 国立国語研究所の『分類語彙表』（約３万3,000語）と『角川類義語新辞典』（約６万語）の２つがある。前者はフロッピーディスク、後者は CD-ROM の媒体で提供される。

(13) 志村拓・大池浩一（1990）により可能である。

(14) コンピュータにおける文字コードの割り当て方はいくつか種類があるが、いずれも同じ文字種のコードは連続しているために、コードを調べれば文字種が判明する。

(15)　格フレームは、C. Fillmore の格文法に基づくもので、単語と単語の間の意味関係を述語中心にとらえる。名詞は述語との関係で格が決まり、例えば、対象格、場所格などがある（長尾 1997）。
(16)　語論理は、述語と定義域という概念を導入した記号論理（文全体をそれを構成する基本的な文の論理的結合により表現し、記号の上でその真偽を論じる）をいう（田中・辻井 1988）。
(17)　意味ネットワークは、人間が物事を認識する際に、物事と物事との関係という観点からとらえることが多いとして、物事の関係をネットワークという形で表す（田中・辻井 1988）。
(18)　例えば、長瀬（1988）が利用できる。原版は英、独、仏などの 1 バイト文字に対応するが、日本語版マニュアルの完成時に、 2 バイト文字である日本語にも対応できるように改良された。
(19)　KJ 法は川喜多二郎氏により考え出されたもので、名前の頭文字を取って名付けられた。
(20)　実際には、それほど簡単ではなく、コーダー達は10ページに及ぶコードブックに従って作業を行った。また、コーダーの訓練に少なくとも丸 1 日は要しており、この中から試験に合格した者だけがコーダーとして採用された（伊藤 1987）。
(21)　コンティンジェンシー分析は、メッセージにおけるシンボルの共起のパターンから、ある情報源における関連性のネットワークを推測することを目的とする（クリッペンドルフ 1989）。
(22)　文脈的分類法は、データに含まれるある種の冗長性を取り除き、それによって根底にある概念を抽出するための多変量解析の手法である（クリッペンドルフ 1989）。

（参考文献）　アルファベット順

1995年 SSM 調査研究会『SSM 産業分類・職業分類（95年版）』、1995年。
1995年 SSM 調査研究会『1995年 SSM 調査コード・ブック』、1995年。
浅井晃『調査の技術』、日科技連、1987年。
原純輔「子どもは日本をどう見ているか」、辻村明他編『世界は日本をどう見ているか　対日イメージの研究』、日本評論社、1987年、205–218ページ。
原純輔「社会学における自由回答データ・文章データとコーディング」『第25回日本行動計量学会報告要旨集』、1997年、164–165ページ。
原純輔・海野道郎『社会調査演習』、東京大学出版会、1984年。
林英夫「質問紙の作成」、村上英治編『心理学研究法 9　質問紙調査法』、東京大学出版会、1975年、107–146ページ。
伊藤陽一「世界の歴史教科書に見られる自国イメージと他国イメージ――韓国、中国、日本の場合を中心に」、辻村明他編『世界は日本をどう見ているか　対日イメージの研究』、日本評論社、1987年、168–186ページ。
国立国語研究所編『分類語彙表』、秀英出版、1964年。
小嶋外弘「質問紙調査法の技法に関する検討」、村上英治編『心理学研究法 9　質問紙調査法』、東京大学出版会、1975年、224–270ページ。
クリッペンドルフ、クラウス（三上俊治他訳）『メッセージ分析の技法――内容分析への招待』、頸草書房、1989年。
松本裕治他『日本語形態素解析システム JUMAN 使用説明書 version3.0』、奈良先端科学技術大学院大学情報科学研究科松本研究室、1996年。

松本裕治他『岩波講座　言語の科学１　言語の科学入門』、岩波書店、1997年。
宮崎哲夫他「分類視点の学習機構を持つ情報自動分類システム」『自然言語処理研究報告』97（4）、1997年、91–98ページ。
長尾真『自然言語処理』、岩波書店、1996年。
西垣通『思想としてのパソコン』、NTT 出版、1997年。
野崎進他「アンケートにおける日本語自由文の情報分析」『第47回情報処理学会全国大会講演論文集』（３）、1993年、165–166ページ。
Ohiwa, H., K. Kawai and M. Koyama, Idea Processor and the KJ Method, *Journal of Information Processing,* 13（1）, 1990, pp. 44–48.
大野晋・浜西正人『角川類義語新辞典』、角川書店、1989年。
Oxford University Press（長瀬真理訳）『Micro–OCP 文章解析プログラム　マニュアル』、沖田電子技研、1988年。
志村拓・大池浩一『MS–DOS SOFTWARE TOOLS 基本セット32』、アスキー出版局、1990年。
塩見隆一他「シソーラスを用いた文書データの自動分類」『自然言語処理研究報告』97（4）、1997年、99–104ページ。
杉山明子『社会調査の基本』、朝倉書店、1984年。
髙橋和子「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法の分析」、『茨城大学人文学部紀要（社会科学）』、23、1990年、21–51ページ。
髙橋和子「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」、原純輔編『非定型データの処理・分析法に関する基礎的研究』、1992年、137–164ページ（1992a）。1991年度文部省科学研究費補助金総合研究（a）研究報告書（課題番号01301018）。
髙橋和子「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」、井上芳保編『苦悩する自己啓発セミナーの研究』、1992年、79–102ページ（1992b）。1991年度文部省科学研究費補助金奨励研究（a）研究報告書（課題番号03851034）。
髙橋和子「SSM 職業データにおける自由回答の分析―― SSM 職業コーディング支援エキスパートシステム構築のために」『千葉敬愛短期大学国際教養学論集』、5、1995年、73–109ページ（1995a）。
髙橋和子「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」『第50回情報処理学会全国大会講演論文集』（１）、1995年、323–324ページ（1995b）。
髙橋和子「自然言語処理による SSM 職業コーディング・システムについて」『第25回日本行動計量学会報告要旨集』、1997年、166–167ページ。
髙橋和子「自然言語処理による SSM 職業コーディングの自動化システム」、盛山和夫編『現代日本の社会階層に関する全国調査研究』、1998年（予定）（1998a）。1997年度文部省科学研究費補助金特別推進研究（１）研究報告書（課題番号06101001）。
髙橋和子「コーディングと内容分析」、宮野勝他編『新社会調査ハンドブック』、新

曜社、1998年（予定）（1998b）。
田中穂積・辻井潤一『自然言語理解』、オーム社、1988年。
堤豊他「電子メールを用いた日本語文による質問応答システムにおける類似質問の抽出について」『自然言語処理研究報告』、97（4）、1997年、161-166ページ。
辻新六・有馬昌宏『アンケート調査の方法』、朝倉書店、1987年。
安田三郎・原純輔『社会調査ハンドブック［第3版］』、有斐閣双書、1982年。