格フレームによる自由回答のコーディング自動化

職業コーディングの自動化システム

髙橋　和子*

Automatic Coding System for Open-Ended Questionnaires by Case Frame

Kazuko TAKAHASHI

Traditionally, open-ended questionnaires have been hardly used on large-scale surveys, where statistical processing of quantitative samples is needed. Although there are some reasons for this, it is not desirable that response styles should be restricted by data-processing technics. In this paper, an automatic coding system for open-ended questionnaires that require statistical treatment of quantitative samples by "case frame" will be proposed. That is, it will conduct a morphological and semantic analysis of occupational data in social stratification and social mobility surveys, which are important in social science, and will automatically choose an appropriate one from about 200 occupation categories. This system has not been completed yet, but precision is about 50% and estimated to increase up to 70–80% with relatively easy improvement.

＊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.

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強制的に選ばせる「選択的回答法」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 種類が存在するが、統計処理を目的とした本調査においては選択的回答法が用いられることが多く、自由回答法はほとんど用いられない。この理由は、自由回答法においては、データ収集後に各回答に分類用カテゴリーのコードを付ける「アフター・コーディング」が必要なため、作業が煩雑になり、多くの人手と時間を要することや、コーディングの結果に対する信頼性が保証されにくいためである（原・海野 1984)。しかし、選択的回答法にも欠点がないわけではなく（林 1975、小嶋 1975、安田・原 1982)、また、自由回答法でなくては得られない情報が存在することは明らかである（浅井 1987）ために、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。

このような問題を解決する 1 つの方法として、本稿では、大量に収集された自由回答のコーディング方法として、コンピュータの利用により回答に対する形態素解析と格フレームによる簡単な意味解析を行って、自動的に妥当なカテゴリーにコーディングするシステムを提案する。今回対象としたのは、アフター・コーディングの説明でしばしば取り上げられる職業データ(1)（原・海野 1984）である。これを選んだ理由は、回答がそれほど複雑ではないことと、1995年SSM調査研究会（1995a）において明らかなように、個々のカテゴリーの定義が明確で、その内容を比較的形式的に扱うことができると判断したためである。

最初に、一般的な自由回答のコーディング処理について簡単に説明した後、職業データの自動コーディングシステムについて述べる。

2. 自由回答のコーディング

自由回答のコーディングを行う際に、コーダー（人間）は通常、次の 2 つの処理を行う。

⑴　回答のもつ意味内容を理解する。

⑵　妥当なカテゴリー(2) に分類してそのコードを付ける（狭義のコーディング）。

ただし、前提条件として、コーダーは、

⑼カテゴリーの定義内容を知っている。

すなわち、カテゴリーに関する知識をもっている必要がある。

したがって、ここでコンピュータによる自由回答のコーディングを、回答からカテゴリーの定義内容を探す「情報検索」と捉えたとき、単なるキーワード検索（佐藤 1992、都築 1992）ではなく、構造をもった検索を行えることが望ましい。なぜなら、人間は回答やカテゴリーの意味を理解する際に、単語だけでなく単語間の関係（構造）まで捉えているからである。

以下では、本稿において自由回答の具体的なコーディング例として取り上げた「職業コーディング」に対して、⑴、⑵の処理と前提条件である⑼を実現する方法を示す。

3. 職業コーディングの場合

1. 職業コーディングと問題点

職業コーディングとは、階層移動研究において重要な「職業」を変数として扱えるようにするため、SSM（Social Stratification and social Mobility）調査により収集された「職業データ」を分析者が総合的に判断して、あらかじめ定められた「職業のカテゴリー」のいずれかのコードにコーディングすることを言う。ここで、職業データとは、狭義の職業を意味する「本

人の仕事内容」*に加えて、「従業上の地位」、「従業先の名前」*、「従業先事業の種類」*、「従業員数」、「役職名」の計6種類のデータを総称したものである（*を付けたものは自由回答）。職業コーディングの中心となるのは、自由回答である「本人の仕事内容」である。職業データの回答例を資料1に示す。

SSM調査は全国調査であり、コーディングの対象となる職業データの個数は毎回数万個に達しており、職業のカテゴリーも少ない時でも約200個程度あることから、作業は非常に煩雑で多くの人手と日数を要するという問題がある。また、そのために、コーディングの一貫性に関する問題が生じる可能性もあり、結果として、信頼性の保証がなされにくいという問題点も抱えている。これらは、いずれも大量サンプルの自由回答に共通の問題である。

2. 回答の意味表現

回答の意味を捉えるために、回答がどのように表現されているかを考察する必要がある。ここでは、最も新しい1995年調査（約7,000サンプル）におけるA票のうち、地区番号が001〜133の約1,000サンプル（無職と学生を除いた有効763サンプル）の問4（回答者の現職を尋ねる質問）に対する回答を用いて、表現形態の傾向を調べた。

「本人の仕事内容」を中心に分析した結果、回答は1例を除いて、すべて1語（「看護婦」など）または1文（「ベランダの木製デッキの製作（ホームリゾート、MG建設ともに）」など）から成っており、比較的単純な構造であった。また、事実を述べるためか、曖昧な表現がなく、肯定の平叙文が多い。出現する品詞は名詞と動詞が多く、形容詞や副詞による修飾はほとんどない。特に、回答の末尾にある語は、不要な語（後述）を除くと、動詞（6%）、サ変名詞（51%）、普通名詞（39%）で計96%になった。時制は現在形であり、主語は省略されているが「本人」（私）であるのは明らかである。

質問によっては、時制が過去形になったり、省略された主語が、配偶者や本人（配偶者）の父（母）などに変化することもあるが、職業のコーディ

ング自体には影響しない。

以上より、回答は比較的制限された形式をもち、その意味表現を「格フレーム」により行える（78%）と判断した。その際、回答の末尾の語を格フレームの述語とすることができる。格フレームにより表現できないもの（22%）は、単に「全般」などのような回答や、役職（「代表取締役社長」など）、職場名（「××係」など）による回答である。格フレームによる回答の意味表現例を図1に示す。

図1　格フレームによる回答の意味表現

回答が「レタスを作っている」の場合	回答が「中学校教員」の場合
述　語：作る	述　語：教員
対象格：レタス	場所格：中学校

この他の特徴として、今回の意味表現においては不要であると判断される語（等、一般、いるなど）の使用（22%）や、並列表現（「住宅の設計・建築」、「米、野菜作り」など）（17%）が比較的多かった。また、論文や新聞記事などと異なり、省略されたり、誤字も含め文法的に正しくない表現もみられた。

3. カテゴリーに関する知識の表現

カテゴリーである職業は、国勢調査における職業分類に基づいて作成された1995年SSM調査研究会（1995a）に定義されて、各々に「職業小分類コード」と呼ばれるコードが付けられている（資料2参照）。これによると、一般に職業は大まかには動作（述語により表現される）の違いにより分類され、さらに、動作の対象や動作を行う場所などにより細分類される傾向がある。したがって、職業に関する知識も、図2のように格フレームにより表現できると考えられる。

図2　格フレームによる職業に関する知識の表現（数字は職業小分類コードを表す）

599　農耕・養蚕作業者[3]の場合

述　語：栽培

対象格：野菜

522　中学校教員[4]の場合

述　語：教える

場所格：中学校

実際のコーディングにおいては、この他に1995年SSM調査研究会（1995a）に明記されていないヒューリスティックな知識が用いられる場合もある（1995年SSM調査研究会 1995b）。例えば管理的職業の1つである「548会社役員」は、「従業上の地位」や「従業員数」、「役職」をチェックする必要がある。これについては、例えば「従業上の地位格」のように格の概念を拡張したものを新たに作り出して追加することで対応することができよう。

以上より、職業コーディングにおいては、回答、カテゴリーともに格フレームにより適切に意味表現を行えると判断できる。

4. 辞書とシソーラス

回答からカテゴリーを検索するためには、3.3で述べたカテゴリーに関する知識をまとめた辞書を用意する必要がある。また、回答とカテゴリーに出現する語を関連づけるためのシソーラスも必要である。

1. 辞　書

図2に示したような職業に関する知識をまとめたものを「職業小分類辞書」と呼ぶ。対象格を「を」、場所格を「で」で示すことにすると、職業小分類辞書においては、各々の職業、例えば「501　自然科学系研究者」（図3）の定義内容は図4のように表現される。

図3 「501　自然科学系研究者」の定義例

501　自然科学系研究者

研究所、試験場、研究室などの研究施設において、専ら理学、工学、農学、医学、薬学など自然科学に関する研究、試験、検定、分析、鑑定、調査などの専門的、科学的な業務に従事するものをいう。

（以下略）

図4　職業小分類辞書の例

職業小分類

コード　述語　対象格の格要素　場所格の格要素

↓　↓　↓　↓

（501　研究（を 理学 工学……自然科学）（で 研究所 試験場 研究室 研究施設））

（501　試験（を 理学 工学……自然科学）（で 研究所 試験場 研究室 研究施設））

（以下略）

職業によっては、対象格や場所格の両方が必要なものもあるが、どちらか1つでよいものや、両方とも不必要なもの（述語のみで決定できる）もある。また、この他にヒューリスティックな知識として、「従業上の地位格」や「従業員数格」などが必要になるものもあるが、図4に示すように職業ごとに必要な格を列挙すればよい。なお、図3においては、職業を定義する述語が1つではない（研究、試験、……）が、その場合は各述語ごとに職業を記述していく（図4）。

本システムにおいては、最終的には職業小分類辞書をそのまま使用せずに、職業の分類上、同一視してよいと判断できる述語はまとめて扱うこととし（例えば、「組立て」、「構成」、「組合わせ」は同じ述語であるとみなす）、さらに、検索を容易にするために、図5に示すように、述語ごとにまとめた辞書に作成し直す。このとき、述語を語そのものではなくコードに変換して扱うが（例えば、「組立て」、「構成」、「組合わせ」をいずれも「132 2」なるコードとする）、この理由については4. 2で述べる。図5に示すように、職業

に関する知識を述語のコード（述語コード）別に記述し直した辞書を、「述語コード別職業小分類辞書」と呼ぶ。

図 5　述語コード別職業小分類辞書

```
述　語　職業小分類
コード　コード　　格要素
  ↓　　↓　　↓
((132 2)(631　(を 鉄塔 橋梁 鉄骨 タンク））
　　⋮
　　(664　(を 和傘 提灯 うちわ 扇子））　）
```

2. シソーラス

回答と職業小分類辞書に出現する語は、類似語や表記の問題、抽象度レベルでの違いがあるため、格フレームにおける述語である動詞や名詞と、格要素である名詞に対して、各々、次のようなシソーラスを用意する必要がある。

- 述語　　類似した述語を同一視するための「述語シソーラス」(図 6)
- 名詞　　「職業小分類辞書」における格要素を代表語とし、回答における格要素を用例とする「名詞シソーラス」(図 10)

(1)　述語シソーラス

述語シソーラスは、図 6 に示すように述語に読みと述語コードを付けたものである。述語自体が異なっていても、述語コードが等しければ同じものとして扱うこととする。例えば、図 6 によると、「製する」と「製造」は語自体は異なるが、述語コードが等しいために同じものとみなされる。

図6 「述語シソーラス」の例

読み	述語	述語コード
↓	↓	↓
（せいする	製する	386 1）
（せいする	征する	367 3）
（せいぞう	製造	386 1）

述語コードは、図8や図9に示すように、『分類語彙表』（国立国語研究所 1964）（図7）で用いられている分類番号の小数部分と、グループ番号をセットにしたものとする。ここで、グループ番号とは同一の分類番号内における語のグループの順番を示す数値である。

図7 『分類語彙表』における記述例

1.386 ←分類番号

製造 製作 作製 作成 制作 調整 調進 手製 手作り 細工 工作

人造 人工 加工

新造 新調 創製 複製 模造 偽造 変造 贋造 試作

2.386 ←分類番号

作る 作り上げる 作り出す 作り直す 作り替える 製する こしらえる

こさえる

つくろう

（注） 分類番号の整数部分が1は体，2は用を表す．

図8 「製造」の述語コード

分類番号の小数部分	グループ番号
↓	↓
386	1

図9 「製する」の述語コード

分類番号の小数部分	グループ番号
↓	↓
386	1

今回、述語シソーラスに『分類語彙表』を利用した理由は次の通りである。まず、『分類語彙表』には日常的に用いられる基本的な語が大量に収集されて分類されているが、職業を定義する述語に関しては、この分類とそれほど大きな隔たりがないと判断したこと。次に、これは述語を分類番号に基づいたコードで扱うことにした理由でもあるが、体や用の細分（分類番号の小数部分）がなるべく平行するようにしてあり、内容上関係のある語は同じ分類番号（少なくとも小数点第2位まで）を与えてあるために、小数部分が等しければ、動詞、名詞の違いに関係なく、同じ意味をもつものとして扱うことができて便利なこと（図8、図9）である。このようなコードを述語とする利点は、将来的に、「等しい」を拡張して「類似性」まで考慮できる可能性があることである。

なお、述語シソーラスに読みを付けた理由は、回答が平仮名による表記であっても対応できるようにするためである。

（2） 名詞シソーラス

名詞シソーラスについては、今回は『分類語彙表』を利用できない。なぜなら、前述したように、職業はもしそれを表す述語が等しくても、対象格や場所格の要素の違いにより細分類されるために、格要素となる名詞に関しては『分類語彙表』のような粗い分類では対応できないものが多いからである。

名詞シソーラスは図10のように記述される。用例には、述語シソーラスと同様に、平仮名表記にも対応できるようにしておく。代表語と用例は、最初は1995年SSM調査研究会（1995a）を利用して作成し、その後、データの蓄積により用例を増やしていくこととする。

図10 「名詞シソーラス」の例

代表語　　用例
↓　　↓
（野菜　　レタス　キャベツ　きゃべつ）
（穀物　　米　こめ　麦　むぎ）

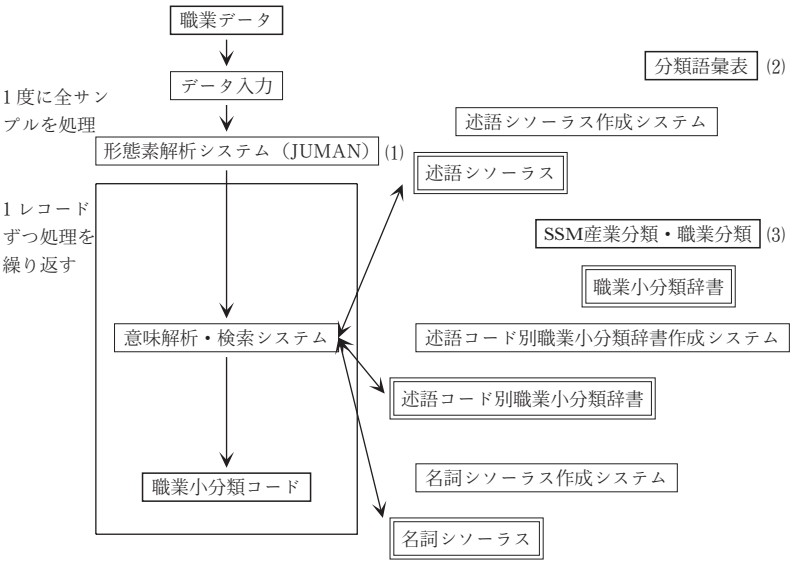
5. システムの概要

1. システムの全体構成

システムは、図11に示すように、回答（職業データ）を入力した後、形態素解析を行うシステム、意味解析・検索を行ってカテゴリーのコード（職業小分類コード）を出力するシステム、そこで使用する辞書やシソーラスを作成しておくシステムから構成される。形態素解析システムとしては京都大学で開発されたJUMANを使用することとし、最も重要な部分である意味解析・検索システムを自作する。また、辞書やシソーラスの作成は、既存の表計算ソフトとワープロ用ソフトを組み合わせて行うこととした。

以下では、紙面の都合上、意味解析・検索システムを中心に述べる。

図11　システムの全体構成



（注）(1)松本 1996,(2)国立国語研究所 1964,(3)1995年SSM調査研究会 1995a.

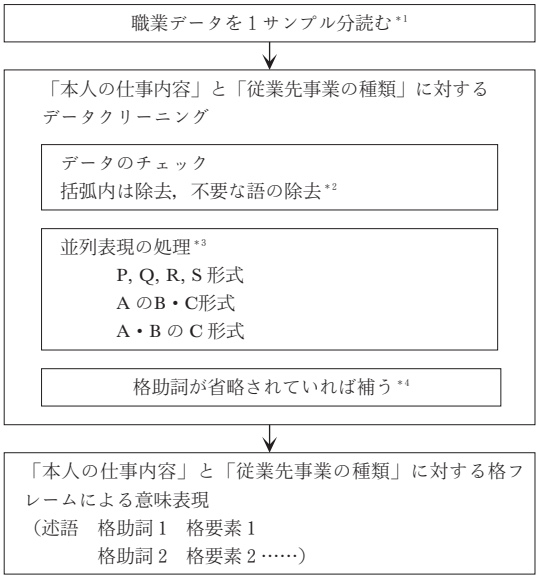
2. 意味解析・検索システム

意味解析・検索システムは、回答の意味解析処理と職業小分類コード検索処理から構成される。これは、それぞれ、2で述べた⑴、⑵の処理、すなわちコーディング処理に相当する。今回は対象を職業コーディングに限定しているが、1文の自由回答に一般化できるよう汎用性を意識している。

（1） 回答の意味解析処理

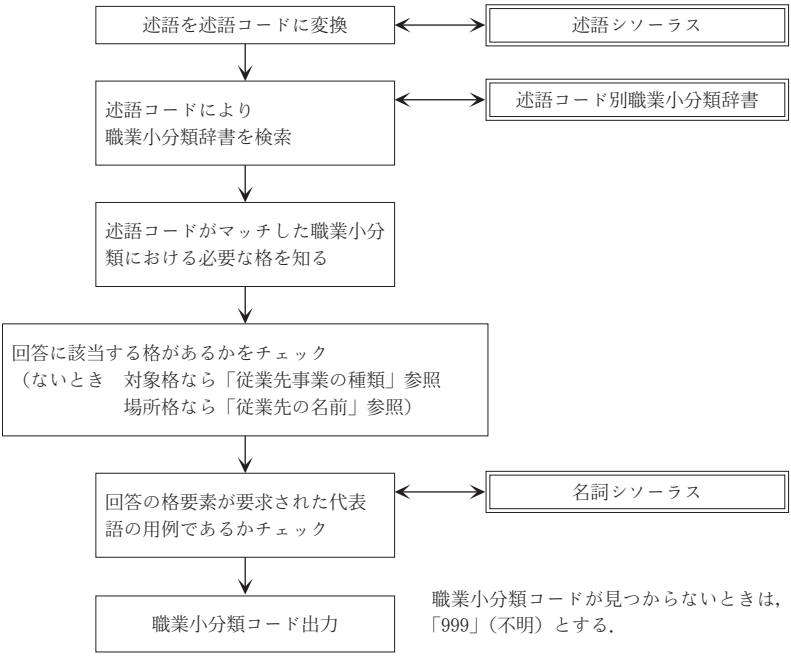
回答の意味解析処理を図12に示す。ここでの目的は、回答の中心である「本人の仕事内容」を格フレームの形式により表現することである。しか

図12　回答の意味解析処理



（注）＊1「従業上の地位」が，8（学生）または9（無職）のときは，これ以降の処理を行わず，職業小分類コード（学生は「988」～「990」，無職は「986」）を付けて終了する．
＊2 データ例は，「レタス等を作っている」，「営業一般」など．
＊3 データ例は，「営業，販売，布団打ち直し」，「住宅の設計・建築」，「米・麦を作る」など．
＊4 データ例は，「建具制作」，「中学校教員」など．

図13　職業小分類コード検索処理



し、3. 2で述べたように、回答の中には、職業の分類には直接関係しないと思われる不要な語が含まれていたり、逆に格助詞などの必要な情報が省略されていたりする。また、並列表現がなされるものも多いため、最初にデータの整備を行っておく必要がある。例えば、「米作りなど」であれば、不要な語「など」を削除し、格助詞「を」や「で」を補って「米をで作る」に修正する。ここで、格助詞として「をで」を補う理由は、文の意味まで考慮しなければ、対象格と場所格のいずれの格助詞を補えばよいのかが判断できないために、両方の格助詞を補うことで、表層的に両方の処理を行うことができるようにするためである。

（2）　職業小分類コード検索処理

職業小分類コード検索処理を図13に示す。

意味解析・検索システムのプログラム開発言語は、リスト処理に向くLISP（Allegro Common Lisp for Windows）を用いた。現段階における処理機

能は表1に示す通りで、プログラムの大きさは約600ステップ程度である。ところで、本システムは使いやすさの点からパソコン上で稼働したために、ワークステーション上で稼働するJUMAN（形態素解析システム）とは日本語のコード体系が異なる。したがって、JUMANの出力結果に対して、コード変換（EUCコードからシフトJISコード）を行う必要がある。

3. 述語コード別職業小分類辞書作成システム

述語コード別職業小分類辞書作成の手順は、次の通りである。

⑴ 1995年SSM調査研究会（1995a）に基づいて、職業小分類辞書を作成する。
⑵ 職業小分類辞書における述語に注目して、並べ替えを行う。
⑶ 各述語に対して、『分類語彙表』の分類番号とグループ番号による述語コードに変換を行う。
⑷ 述語コードに注目をして、並べ替えを行う。
⑸ 同一の述語コードについては最初のものだけ残して、述語コード別職業小分類辞書の形式にする。

作成された述語コード別職業小分類辞書はテキスト形式で、容量は44Kバイトであった。一部を資料3に示す。

4. 述語シソーラス・名詞シソーラス作成システム

シソーラスの作成については、今回、述語シソーラスは既存の『分類語彙表』を利用したのに対して、名詞シソーラスは独自に作成した。この理由は、前述したように、職業の分類においては、動作を表す述語は通常の分類でよいが、動作の対象または場所を表す名詞（格フレームにおける各要素）に関しては、日常的な分類と異なった視点によりなされるという事情による。このように、研究目的により分類の視点が一般的でない場合には、既存のシソーラスを利用せずに独自に作成する必要がある。

作成された述語シソーラスと名詞シソーラスはいずれもテキスト形式で、容量はそれぞれ678Kバイトと26Kバイトであった。一部を資料4､5に示す。

表1　現段階で処理可能な内容

処　理　内　容	可能・不可能の別
「本人の仕事内容」の処理	
・処理の対象とする部分	
主要な部分（括弧外の文）	○
追加部分（括弧内の文）	×
・並列表現（4個まで）の処理	
P, Q, R, S形式	○　①
AのB・C形式	○　②
A・BのC形式	○　③
その他の形式	×
・格の処理	
対象格を必要とするもの	○　④
場所格を必要とするもの	○　⑤
その他の格（道具格など）を必要とするもの	×
複数の格を必要とするもの	×
格助詞が省略されたもの（対象格または場所格と解釈する）	○　⑥
・不要な語の除去	○　⑦
・「本人の仕事内容」がないもの（内職，無職，学生など）の処理	△
「従業先事業の種類」も参照（対象格が必要なのに，抽象的にしか記述されていない〔例えば，部品など〕または省略されているときの対応）	△　⑧（両者の述語が同じときのみ処理）
「従業先の名前」も参照（場所格が必要なのに省略されているときの対応）	×
「従業員数」，「役職名」も参照（管理職の場合に必要）	×

（注）　○はできる，△はほぼできる，×はできないことを表す．①〜⑧の具体的な回答例については、次の通りである．①営業，販売，布団打ち直し，②住宅の設計・建築，③米・麦を作る，④レタスを作る，⑤小学校で教える，⑥建具製作　小学校教員，⑦レタス等を作っている　営業一般，⑧部品の製造（「本人の仕事内容」）　コンデンサ製造（「従業先事業の種類」）．

6. 結果と考察

システムは現在のところ表1に示したように完成されているわけではないが、実際のデータ（前述のA票における地区番号001〜019の141サンプル中、無職や学生を除く有効103サンプル）を本システムに従って人手により処理した[(5)]結果、現状のままで約50%が正しくコーディングされた。以下で詳細を述べる。

結果について、正しくコーディングされたもの（A）、誤って別のもの

表2　コーディング結果の分類（サンプル数計103）

コーディング結果	サンプル数（割合）	評　価
A 正しく決定されたもの	48（47%）	○　Ⓐ
B 誤って決定されたもの	14（14%）	
・辞書の知識不足	11	
管理職関係	6	△　Ⓑ−1a
自営業関係	3	△　Ⓑ−1b
保険代理人・外交員	1	△　Ⓑ−1c
その他	1	△　Ⓑ−1d
・プログラムの機能不足	1	
括弧内の処理	1	△　Ⓑ−2
・回答の情報不足	2	×　Ⓑ−3
C 決定できなかったもの	41（40%）	
・形態素解析の失敗	8	×　Ⓒ−1
・辞書間での単語の整合性の問題	5	×　Ⓒ−2
・格フレームによる意味表現不能	2	×　Ⓒ−3
・述語シソーラスの不備	8	
単語がない	2	×　Ⓒ−4a
述語のグルーピング不備	6	×　Ⓒ−4b
・プログラムの機能不足	13	
括弧内の処理なし	2	△　Ⓒ−5a
職業データの他項目処理なし	9	△　Ⓒ−5b
並列表現処理の不足	2	×　Ⓒ−5c
・辞書の知識不足	3	△　Ⓒ−6
・回答記述欄誤り，情報不足など	2	×　Ⓒ−7

（注）　○は改良の必要なし，△は比較的容易に改良できる，×は改良が困難またはできないことを表す．評価欄における記号（Ⓐなど）は，表 3，4 の種類欄の数値と一致する．割合の合計は，丸め誤差のために100%にはならない．

にコーディングされたもの（B）、決定できなかったもの（C。「999」にコーディングされる）の 3 種類に分類し、B、Cについてはさらにその理由を調査したものを表 2 に示す。回答が並列表現の場合には、複数のコーディング結果の中に正しいものが含まれているときにAと判定した。A、B については、回答例、コーディング結果、正解を表 3、Cについては、回答例、解析状況、正解を表 4 にそれぞれ示す。ここで、サンプル番号は地区番号 3 桁の後に対象番号 2 桁を加えた計 5 桁である。

前述したように、現システムにおける正解率（○印）は約 5 割である。表 2 によれば、比較的容易に改良できるもの（△印）として、辞書の知識

表3　回答例とコーディング結果（AとBの場合）

種類	サンプル番号	回答（太字）→コーディング結果	正解
Ⓐ	00110	**営業・経理等**→557，559	557
Ⓐ	00215	**電話交換手**→617	617
Ⓐ	01914	**看護婦**→514	514
Ⓐ	00500	**会計等の事務**→682	682
Ⓐ	00516	**営業事務（車が売れた時の手続き，ナンバー，車庫証明をとるなど）**→557	557
Ⓐ	00623	**保険外交員**→574	574
Ⓐ	00720	**足場を組んだり，建材を運んだり男と同じ労働作業**→679，999	679
Ⓐ	01510	**牛の飼育**→601	601
Ⓐ	01512	**ホタテ加工**→645	645
Ⓐ	01613	**新聞配達**→686	686
Ⓐ	01623	**社長室炊事係**→578	578
Ⓐ	01910	**圧延の作業**→628	628
Ⓐ	01922	**せんべいを作っている**→644	644
Ⓑ−1a	00504	**営業，外回り（町の得意先回り）**→557 ただし，「従業上の地位」が 2 「役職」が　課長　4 「従業員数」が　6 選択肢の意味は図8を参照のこと	550
Ⓑ−1b	00805	**配達・店番**→569 ただし，「従業上の地位」が 5　同上	566
Ⓑ−1c	00312	**営業**→557 ただし，「従業先の名前」が××**生命**	574
Ⓑ−1d	01707	**融資（担当）**→559 ただし，「従業先の名前」が××**銀行**	557
Ⓑ−2	01720	**事務（窓口）**→558	555
Ⓑ−3	01300	**営業**→557 ただし，「従業先事業の種類」が**製麺製造**	573

不足すなわち述語コード別職業小分類辞書にヒューリスティックな知識の不足があるが、これを追加すれば、正解率が約60%に上昇する計算になる。さらに、括弧内の情報も処理するなどによりプログラムの機能を上げると約70%となる。

逆に、本システムの限界（×印）としては、JUMAN 辞書の不備による形態素解析の失敗や、JUMAN 辞書と分類語彙表における単語の切り出し方の違い、述語シソーラスの不備などがあるが、これらを完全に解消することは困難である。また、格フレームによる回答の意味表現ができない

表4　回答例と解析状況（Cの場合）

種類	サンプル番号	回答（上段太字） 解析状況の説明（中段以下）	正解
Ⓒ–1	00100	**とそう工施工** 「と(格助詞)そう(動詞)工(普通名詞)施工(サ変名詞)」［JUMAN］と誤って解析されてしまう.	661
Ⓒ–2	00609	**商品仲立人（魚市場のせり）** 仲立人［JUMAN］に対して，仲立［分類語彙表］のみ	572
Ⓒ–3	01003	**北海道電力から料金集金の委嘱を受ける** （受ける（を　委嘱））と解析されるが，（集金（を　料金））でなければ決定できない.	561
Ⓒ–4a	00201	**SE** SE　なし［分類語彙表］（システムエンジニアもなし）	506
Ⓒ–4b	00702	**実際に火を消す仕事** （消す（を　火））と解析.（鎮圧（を　火災））なら決定できる.「消す」と「鎮圧」の述語コードが異なる.	595
Ⓒ–5a	01021	**ホルモン焼飲食店（材料仕込係）** 括弧内の情報より（仕込（を　材料））が解析できれば，（飲食店（を　ホルモン焼））より決定できる.	645
Ⓒ–5b	01415	**指導員** 「従業先事業の種類」が**精薄者更正施設**も参照すれば決定できる.	538
Ⓒ–5c	01702	**酒，食のはんばい，小売り** 現在は，A・BのC・D形式の処理を行っていない.	566
Ⓒ–6	01500	**農産物生産全般** （生産（を　農産物））と解析.　職業小分類コード599にこの知識も追加すれば決定できる.	599
Ⓒ–6	01504	**育成牛の管理** （管理（を　牛　育成））と解析.　　同上	601
Ⓒ–7	00306	**管理指導員（施設使用）** 「従業先の名前」が**市スポーツ振興事業団しょくたく職員**を参照しても，決定するにはなお情報不足.	592

ものや回答の情報不足については、人間の知識によってしか理解できず(後者については、人間でも困難であるが)、本システムでは解決できない。これらは、全部で約20%を占める。結局、本システムに比較的容易な機能アップを行うことで、約70〜80%の回答が処理可能であると予想できる。

ここで、失敗の主な原因はあるが比較的容易に改良できると判断した辞書の知識不足について具体的に述べておく。今回は図11に示したように、1995年SSM調査研究会（1995a）における記述のみを知識としたが、実際

の職業コーディングにおいてはヒューリスティックな知識も用いられる。例えば、管理職にコーディングされる職業（「548　会社役員」、「550　会社・団体等の管理職員」など）は、「本人の仕事内容」が「○○の管理」であるだけでなく、次のような条件すなわち、「従業上の地位」が「1」（役員）または「4」（自営業主）の場合には、「従業員数」が「5」（30人）以上であること、「2」（常時雇用されている一般従業者）または「6」（家族従業者）の場合には、「役職」が「4」（課長）かつ「従業員数」が「5」（30人）以上であることが必要である。逆に、「従業上の地位」や「役職」、「従業員数」が前述の条件を満たしているときは、「本人の仕事内容」に管理という語がなくても、管理職にコーディングされることになっている。Ⓑ–1aはその例で、「従業上の地位」が2、「役職」が課長、「従業員数」が6であったために、「550 会社・団体等の管理職員」にコーディングされるのが正しい。今回は、このようなヒューリスティックな知識を記述していないために、それを必要とする職業、例えば管理職以外では自営業関係や保険代理人・外交員においても、コーディング結果を誤ったり（Ⓑ–1b、c、d）、決定できなかったり（Ⓒ–6）した。

7. おわりに

本稿では、職業データのコーディングを例として、これまで人手で行っていた自由回答のコーディングを格フレームによる意味解析を行うことで自動化する方法を提案した。これにより、これまで問題であったコーディング作業の軽減化、コーディングルールの明示化、コーディングの一貫性の保証ができる。今回は職業データを対象としたが、本システムでは、対象とするデータ領域の特徴や分析者による分類の視点を辞書やシソーラスの内容に反映できるために、これらを変更することで職業データ以外の自由回答（格フレームにより適切に意味表現ができるもの）に対しても適用が可能である。

今後の課題としては、まず、日本語コードの問題を解決するために

Windows 版 ACL を早急に Linux 版に移植することである。次には、プログラムの処理機能の向上、辞書やシソーラスの充実により職業コーディング自体の正解率を高めることである。最後に、本システムを他の自由回答においても適用することができるように一般化を進めることの 3 点である。

（注）

（1） 職業は社会学の階層移動研究（直井・盛山 1990など）で重要な役割を果たす変数で、SSM調査（social stratification and social mobility survey）により職業データとして収集されるが、中心となる「本人の仕事内容」（狭義の職業データ）が自由回答であるため、分析に入る前に職業小分類コード（約200種類）にコーディングされる必要がある。これは膨大な人手と時間を要する作業で、「職業コーディング」と呼ばれる。

（2） 自由回答には、あらかじめカテゴリーが用意されているものと、回答から生成する必要があるものがある。今回は前者の場合を扱う。

（3） 穀物、野菜、果樹その他の作物の栽培、収穫などの作業および蚕の飼育、収繭、蚕種の製造などの作業に従事するものを言う。

（4） 中学校において、生徒の中等普通教育および養護に従事するものを言う。ただし、次の業務に従事するものは本分類に含まれない。（以下略）

（5） 今回使用したパソコン版のACL（Allegro Common Lisp for Windows）においては、日本語コードがシフトJISコードであるために、一部の日本語（例 義、形、ーなど）に対してプログラムが誤動作する問題があった。この問題を解決するために、現在、プログラムや辞書、シソーラスをすべて、パソコンで稼働するUNIX（日本語コードはEUCコードである）として代表的なOSであるlinux上で動くACL（Allegro Common Lisp for linux）環境に移植中である。これにより、JUMAN による形態素解析の結果をコード変換する手間も不要となる。

（参考文献）

1995年SSM調査研究会、1995a、『SSM産業分類・職業分類（95年版）』。
同上、1995b、『1995年SSM調査コードブック』。
浅井晃、1987、『調査の技術』、日科技連。
原純輔・海野道郎、1984、『社会調査演習』、東京大学出版会。
林英夫、1975、「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、107−146ページ。
国立国語研究所、1964、『分類語彙表』、秀英出版社。
小嶋外弘、1975、「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、224−270ページ。
松本裕治他、1996、『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。
長尾真、1996、『自然言語処理』、岩波書店。
直井優・盛山和夫、1990、『現代日本の階層構造①社会階層の構造と過程』、東京大

学出版会。
佐藤嘉倫、1992、「職業コーディング支援システムの構築」、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　平成３年度文部省科学研究費補助金（総合A）研究成果報告書』、199–204ページ。
髙橋和子、1997、「自然言語処理によるSSM職業分類システム」、『第25回日本行動計量学会報告要旨集』、166–167ページ。
髙橋和子、1998a、「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　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』。
髙橋和子、1998b、「コンピュータによる自由回答の処理方法」『敬愛大学国際研究』1号、259–284ページ。
髙橋和子、1998c、「格フレームよる自由回答のコーディング自動化システム」『情報処理学会研究報告』、Vol. 98、No. 82（98–NL–127）、87–94ページ。
髙橋和子、1998d、「格フレームよる自由回答のコーディング自動化システム」『第57回情報処理学会講演論文集⑵』、247–248ページ。
田中穂積・辻井潤一、1988、『自然言語理解』、オーム社。
都築一治、1992、「職業コーディングの自動化システムの試験的構築」、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　平成３年度文部省科学研究費補助金（総合A）研究成果報告書』、205–214ページ。
安田三郎・原純輔、1982、『社会調査ハンドブック第３版』、有斐閣。
安田三郎、1970、『社会調査の計画と解析』、東京大学出版会。

資料１　職業データの回答例

問４〔回答票２〕あなたの現在のご職業について、お聞きします。

ａ　あなたのお仕事は大きく分けてこの中のどれにあたりますか。

〔以下同様にｂ～ｆまで聞く〕

ａあなたのお仕事は大きく分けてこの中のどれにあたりますか。	従業上の地位	該当するものに○をつける	1（ア）経営者、役員 2（イ）常時雇用されている一般従業者 (3)（ウ）臨時雇用・パート・アルバイト 4（エ）派遣社員 5（オ）自営業主、自由業者	6（カ）家族従業者 7（キ）内職→eのみ聞いて問７へ 8（ク）学生 9（ケ）無職 → 問７へ 19　わからない
ｂさしつかえなければ勤め先の名前を教えてください。	従業先の名前	△△会社○○支店（出張所）と、事業所単位で記入すること	〔具体的に記入〕〔派遣社員は派遣会社を勤め先とする〕 スーパー○○	
ｃそこは、どのような事業をいとなんでいますか。	従業先事業の種類	野菜の販売、自動車の製造、薬品の卸売、衣服の小売、旅館経営等と具体的に	〔具体的に記入〕 食料、衣料、雑貨その他の販売など （ふつうのスーパーの仕事） □□　99　わからない	
ｄ従業員（働いている人）は、会社全体で何人ぐらいですか。〔家族従業者も含める〕	従業員数	該当するものに○をつける	1（ア）1人 2（イ）2～4人 3（ウ）5～9人 4（エ）10～29人 (5)（オ）30～99人 6（カ）100～299人	7（キ）300～499人 8（ク）500～999人 9（ケ）1,000人以上 10（コ）官公庁 19　わからない
ｅあなたは勤め先でどのような仕事をしていますか。	本人の仕事の内容	経理、運搬、仕入れ、○○組み立て等と、職種が分かるように詳しく	〔具体的に記入〕 惣菜をつくる（販売はしない） □□□　999　わからない	
ｆ何かの役職についていますか。〔ついている場合〕具体的な名称を教えてください。また、それはこの中ではほぼどれに相当しますか。	役職名	具体的に記入、該当するものに○をつける	具体的な役職名を記入	(1)（ア）役職なし 2（イ）監督、職長、班長、組長 3（ウ）係長、係長相当職 4（エ）課長、課長相当職 5（オ）部長、部長相当職 6（カ）社長、重役、役員、理事 9　わからない

資料2　SSM職業分類項目表（一部）

番号	SSM新分類項目	No	国勢調査小項目
501	自然科学系研究者	1	自然科学系研究者
502	人文科学系研究者	2	人文科学系研究者
503	機械・電気・化学技術者	4	金属製錬技術者
		5	機械・航空機・造船技術者
		6	電気・電子技術者
		7	化学技術者
504	建築・土木技術者	8	建築技術者
		9	土木・測量技術者
505	農林技術者	3	農林水産業・食品技術者
506	情報処理技術者	10	情報処理技術者
507	その他の技師・技術者	11	その他の技術者
508	医師	12	医師
509	歯科医師	13	歯科医師
510	薬剤師	15	薬剤師
511	助産婦	17	助産婦
512	保健婦	16	保健婦
513	栄養士	23	栄養士
514	看護婦、看護士	18	看護婦、看護士
515	あん摩・はり・きゅう師、柔道整復師	24	あん摩マッサージ指圧師、はり師、きゅう師、柔道整復師
516	その他の保健医療従事者	19	診療放射線・エックス線技師
		20	臨床・衛生検査技師
		21	歯科衛生士
		22	歯科技工士
		25	その他の保健医療従事者
517	裁判官、検察官、弁護士	28	裁判官、検察官、弁護士
518	その他の法務従事者	29	その他の法務従事者
519	公認会計士、税理士	30	公認会計士、税理士
520	幼稚園教員	31	幼稚園教員
521	小学校教員	32	小学校教員
522	中学校教員	33	中学校教員
523	高等学校教員	34	高等学校教員
524	大学教員	35	大学教員
525	盲・ろう・養護学校教員	36	盲・ろう（聾）学校・養護学校教員
526	その他の教員	37	その他の教員
527	宗教家	38	宗教家
528	文芸家、著述家	39	文芸家、著述家
529	記者、編集者	40	記者、編集者
530	彫刻家、画家、工芸美術家	41	彫刻家、画家、工芸美術家
531	デザイナー	42	デザイナー
532	写真家、カメラマン	43	写真家、カメラマン
533	音楽家（個人に教授するものを除く）	44	音楽家（個人に教授するものを除く）
534	俳優、舞踊家、演芸家（個人に教授するものを除く）	46	俳優、舞踊家、演芸家（個人に教授するものを除く）
535	職業スポーツ家（個人に教授するものを除く）	50	職業スポーツ家（個人に教授するものを除く）
536	獣医師	14	獣医師
537	保母、保父	26	保母、保父
538	社会福祉事業専門職員	27	その他の社会福祉専門職業従事者

資料３　述語コード別職業小分類辞書（一部）

```
((104 3) (514(を 診療 医師 歯科医師 看護))
         (516(を 検疫 解剖 ))
         (518(を 特許出願 特許登録 ))
         (519(を 税務 ))
         (572(を 商品売買 ))
         (574(を 保険 ))
         (575(を 不動産売買 不動産貸借 不動産交換))
         (577(を 有価証券 ))
         (578(を 雑用 調理 洗濯 応対))
         (581(を 調理 ) (で 学校 飲食店 旅館 病院)))
((1131 1) (612(を 車輌 ))
         (652(を 布地 ))
         (684(を パイプ ボイラー )) )
((123 1) (659(を ゴム製品 プラスチック製))
         (667(を 洋傘 ))
         (669(を 玩具 )) )
((125 1) (594(を 安全 秩序 ))
         (596(を 秩序 )(で 工場 病院 学校 事務所))
         (613 (で 船舶 ))
         (620 (で 採鉱場 採炭場 )) )
((1251 2) (687(を 害虫 )) )
((130 1) (529(を 記事 取材 新聞 雑誌 )) )
((132 2) (631(を 鉄塔 橋梁 鉄骨 タンク ))
         (633(を 機械器具 機械部品 ))
         (634(を 半導体製品 電球 真空管 ))
         (635(を 自動車 自動車部 ))
         (636(を 機関車 車輌 ))
         (638(を 航空機 航空機部 ))
         (639(を 自転車 ))
         (640(を 輸送機械 輸送装置 ))
         (641(を 時計 ))
         (642(を 光学器具 計測器 眼鏡))
         (654(を 木材 ))
         (655(を 和船 ボート ヨット ))
         (664(を 和傘 提灯 うちわ 扇子 ))
```

資料4　述語シソーラス（一部）

(せいさく	制作	320	1)
(せいさく	政策	3084	5)
(せいさく	製作	386	1)
(せいさく	制作	386	1)
(せいさん	生産	3802	2)
(せいさん	成算	3083	1)
(せいさん	精算	3064	3)
(せいさん	清算	3064	3)
(せいし	制止	1563	1)
(せいし	制止	367	3)
(せいし	静止	1513	1)
(せいし	製紙	382	2)
(せいじ	政治	360	10)
(せいじか	政治家	233	4)
(せいじゃ	聖者	234	2)
(せいしょ	清書	3151	2)
(せいしょう	斉唱	3231	4)
(せいしょくしゃ	聖職者	241	15)
(せいじん	聖人	234	2)
(せいず	製図	3153	3)
(せいする	製する	386	1)
(せいする	制する	367	3)
(せいせい	生成	123	1)
(せいせい	精製	386	4)
(せいせい	精製	382	4)
(ぜいせい	税制	3082	1)
(せいせん	精選	3063	6)
(せいそう	政争	3501	3)
(せいそう	正装	3333	2)
(せいそう	清掃	3844	1)
(せいそう	盛装	3333	2)
(せいぞう	製造	386	1)

資料5　名詞シソーラス（一部）

（理学　物理学 数学 統計 数理 気象 地質 天文
）

（工学　機械 電気 土木 建築 繊維 金属 材料 通信 製糸 鉱山
）

（農学　林学 畜産 獣医 動物 植物 水産 林業 食品
）

（医学　解剖
）

（機械　船舶 車輌 航空機 機器 電子管 通信 紡績 染色 測定 写真機 工具 回路 配電盤 半導体
）

（器具　船舶 車輌 航空機 機器 電子管 通信 紡績 染色 測定 写真機 工具 回路 配電盤 半導体
）

（装置　船舶 車輌 航空機 機器 電子管 通信 紡績 染色 測定 写真機 工具 回路 配電盤 半導体
）

（設備　船舶 車輌 航空機 機器 電子管 通信 紡績 染色 測定 写真機 工具 回路 配電盤 半導体
）

（電気施設　発電 送電 照明
）

（通信施設　有線 無線 変電
）

（化学製品　薬品 肥料 繊維 油脂 塗料 樹脂 医薬品 発火物 香料 化粧品 石油製品 ゴム プラスチック 染料 コークス アルミナ 塩 インキ 火薬 歯磨き アンプル ゼラチン ろうそく クレンザー 墨 かとり線香
）

（建築物　橋梁 水道 空港 庭園 河川 水路
）

（農業技術　育種 栽培 土壌 病害虫 養蚕 繭 蚕種 増殖 飼養 養鶏 人工授精
）

（作物　米 稲 麦 雑穀 豆類 芋類 野菜 園芸 飼料 たばこ もやし 茶 きのこ しいたけ　草花　芝
）