情 報 学 基 礎　51－12
自 然 言 語 処 理　127－12
（１９９８．９．１７）

格フレームによる
自由回答のコーディング自動化システム

高橋　和子
敬愛大学　国際学部
〒285-8567　佐倉市山王１－９
Tel：043-486-6210，E-mail：PXK10076@nifty.ne.jp

従来から、統計処理を目的とした本調査においては、自由回答法が用いられることはほとんどなかった。自由回答は手作業による煩雑なアフター・コーディングが必要な上に、コーディング結果の信頼性が保証されにくいためである。しかし、自由回答でなくては得られない情報も存在しており、本来、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。本稿では、大量サンプルで収集された自由回答のコーディングを格フレームの適用により自動化する方法を提案する。今回は社会学の階層移動研究で重要な「職業データ」を対象として、形態素解析と格フレームによる簡単な意味解析を行い、カテゴリーである職業コードの定義を表現した「辞書」を検索することで、自動的に妥当な職業にコーディングするシステムを示したが、本システムの一般化は容易であると考えられる。

Automatic Coding System for Open-Ended Questionnaires by Case Frame

Kazuko　Takahashi
Faculty of International Studies, Keiai University

Traditionally, open-ended questionnaires have been hardly used on large-scale survey 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 statictical treatment of quantitative samples by "case frame" will be proposed. That is, it will conduct morphological and semantic analysis of occupational data in social stratification and social mobility survey which is important in social science, and will automatically choose an appropriate one from about two hundred occupation categories. This system will be generalized in easily.

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［4］。しかし、選択的回答法にも欠点がないわけではなく［5］［7］［18］、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［3］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。

このような問題に対して、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、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]［4］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、個々のカテゴリーの定義［1］が明確で、比較的形式化しやすいと判断したためである。

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、コーダーは通常、次の２つの処理を行う。

（１）回答のもつ意味内容を理解する。

（２）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、

（０）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。

従って、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［11］[16］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、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、「職業コーディング」における回答（主として「本人の仕事内容」）とカテゴリー（職業小分類）の内容を分析して、適切な意味表現の形式を定める。

3　職業コーディングの場合

職業データは、狭義の職業である「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）。

3．１　回答の意味表現

コーディングの対象となる回答の個数は、毎回数万個程度ある。今回は、1995 年調査

---

[1] 職業は社会学の階層移動研究(［10］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。

（約 7000 サンプル）におけるＡ票のうち、地区番号が 001 ～ 133 の約 1000 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。

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

図１　格フレームによる回答の意味表現

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。

３．２ 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［1］に各々定義されている（注3、4参照）。これらによると、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。従って、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。

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

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）

実際のコーディングでは、この他に［1］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［2］。例えば管理的職業の一つである「５４８　会社役員」は、

[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いう。

[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いう。
ただし、次の業務に従事するものは本文類に含まれない。（以下略）

「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。これに対しては、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追加することで対応する。

以上より、職業コーディグにおいては、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した。

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、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。

4．1辞書

3．2で述べた職業に関する知識（図2）をまとめたものを「職業小分類辞書」と呼ぶ。実際には、検索を容易にするために、これを同一とみなせる述語ごとにまとめ、さらにその述語をコード化した「述語コード別職業小分類辞書」（図3）に作成し直す。

```
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)))
```

図3　述語コード別職業小分類辞書

4．2 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、述語と格要素である名詞のそれぞれに対して、シソーラスが必要である。

・述語　述語をグループとして扱うための「述語シソーラス」（図4）

・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5）

述語シソーラスは、各々の述語に述語コードを付けて、これが等しければ同一のグループであるとする。ここで、述語コードは、『分類語彙表』[6]における分類番号の小数部分（体と用の区別をなくすため）とグループ番号をセットにしたものである。これにより、動詞だけでなく、名詞も述語として扱うことができる（例　「製する」と「製造」）。

```
述語     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)
```

図4　「述語シソーラス」の例

（野菜　レタス キャベツ きゃべつ）

図5　「名詞シソーラス」の例

5　システムの概要

5．1 システムの全体構成

システムは、回答（職業データ）の形態素解析を行うシステム、意味解析・検索を行ってコード（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（図6）。中心となるのは、意味解析・検索システムである。

5．2 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。

5．2．1 回答の意味解析処理

回答の意味解析処理は、2で述べた（1）の処理に相当する（図7）。

5．2．2 職業小分類コード検索処理

職業小分類コード検索処理は、２で述べた（２）の処理に相当する（図８）。

５．３　シソーラス作成システム

シソーラスの作成において、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、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。

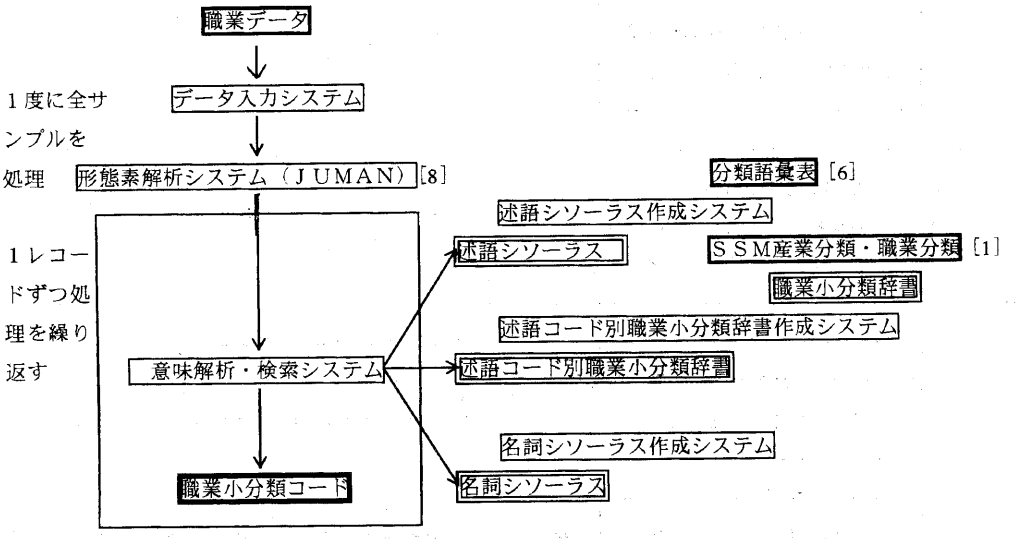
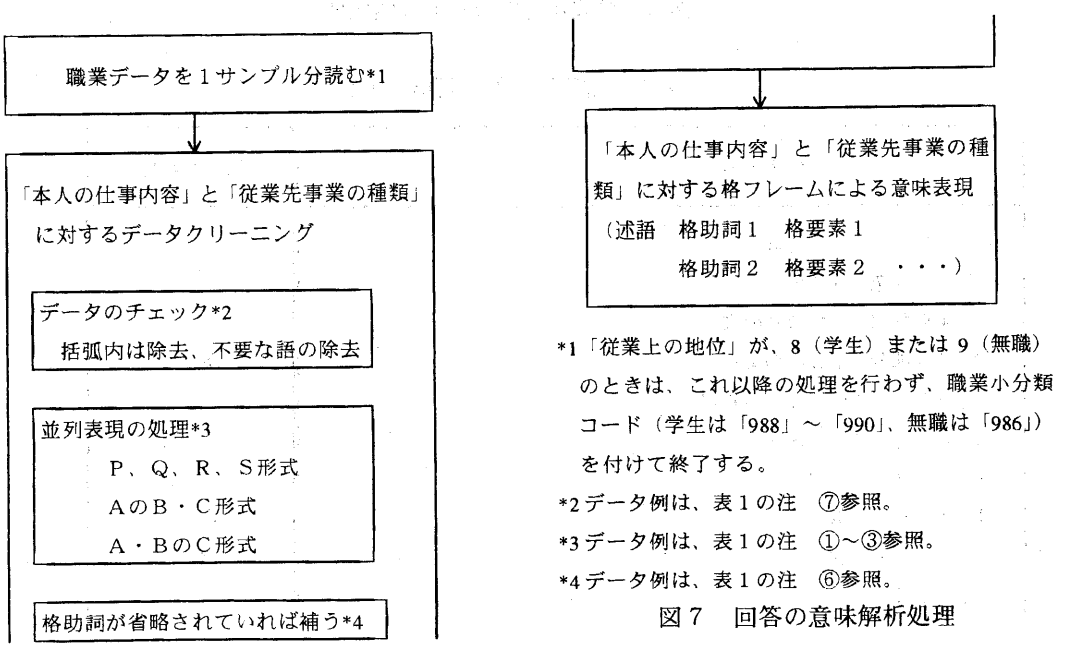


図６　システムの全体構成



*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、表１の注　⑦参照。

*3データ例は、表１の注　①～③参照。

*4データ例は、表１の注　⑥参照。

図７　回答の意味解析処理

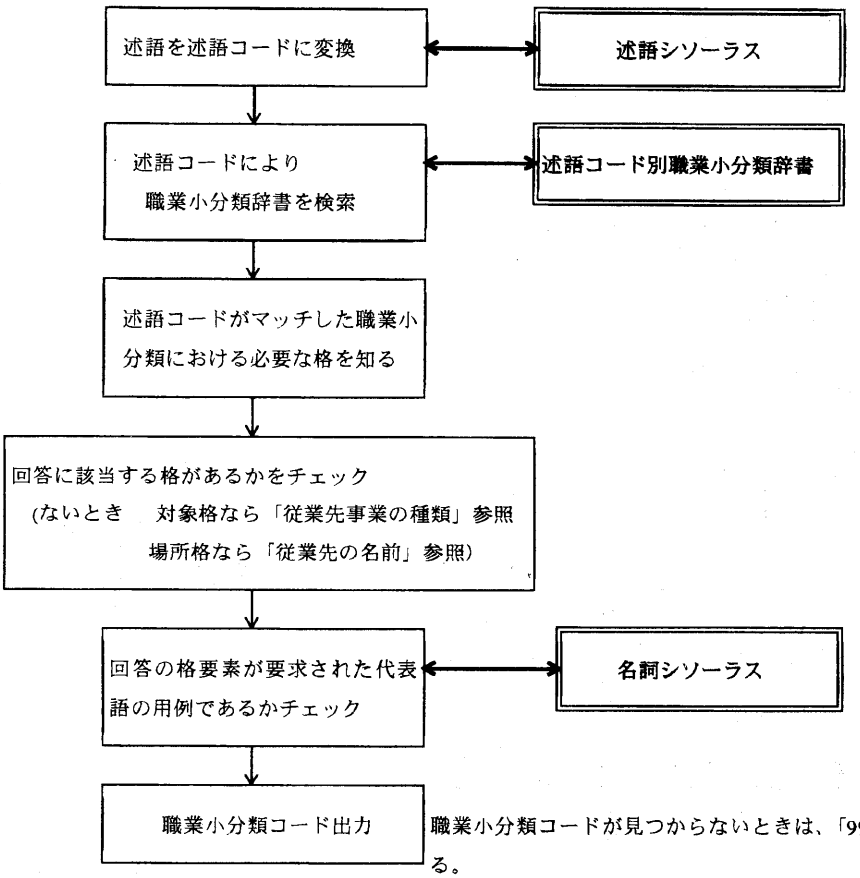


図８　職業小分類コード検索処理

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

システムは未完成である（表１）が、実際のデータ（前述のデータにおいて地区番号 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結果、現状のままで回答の約 50 %が正しくコーディングされた。

主な失敗の原因は、表２に示すように、職業小分類辞書におけるヒューリスティックな知識の不足や、プログラムの機能（職業データの他項目の処理）不足である。この他、シソーラスの不備などもある。このうち、比較的容易にできる改良を行うことで、システムの正解率を約 70 %程度まで高めることが予想できる。

表２ コーディング結果（サンプル数計 103）

コーディング結果	サンプル数
正しく決定されたもの	48
誤って決定されたもの	14
・辞書の知識不足	11
管理職関係	6
自営業関係	3
保険代理人・外交員	1
その他	1
・プログラムの機能不足	1
括弧内の処理	1
・回答の情報不足	2
決定できなかったもの	41

・形態素解析の失敗	8	
・辞書間の単語の不一致	5	
・格フレームによる意味表現不可	2	
・述語シソーラスの不備	8	
単語がない		2
グルーピング不備		6
・プログラムの機能不足	13	
括弧内の処理		2
職業データの他項目の処理		9
並列表現処理の不備		2
・辞書の知識不足	3	
・回答の誤り・情報不足など	2	

7 おわりに

本稿では、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、職業コーディング自体の正解率を高めることと、本システムの一般化を進めることである。

参考文献

[1] 1995 年 SSM 調査研究会. 1995.『SSM 産業分類・職業分類（９５年版）』.

[2] 同上. 1995.『1995 年 SSM 調査コードブック』.

[3] 浅井　晃. 1987.『調査の技術』. 日科技連.

[4] 原　純輔・海野道郎. 1984.『社会調査演習』. 東大出版会.

[5] 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』. 東大出版会. 107-146.

[6] 国立国語研究所. 1964.『分類語彙表』. 秀英出版社.

[7] 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』. 東大出版会. 224-270.

[8] 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.

[9] 長尾　真. 1996.『自然言語処理』. 岩波書店.

[10] 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』. 東大出版会.

[11] 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.

[12] 高橋和子. 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.

[13] 高橋和子. 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.

[14] 高橋和子. 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. 1：259-284.

[15] 田中穂積・辻井潤一. 1988.『自然言語理解』. オーム社.

[16] 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.

[17] 安田三郎・原　純輔. 1982.『社会調査ハンドブック第３版』. 有斐閣.

[18] 安田三郎. 1970.『社会調査の計画と解析』. 東大出版会.