自由回答のコーディング自動化システム

「健康と階層」調査における職業コーディング

髙橋　和子*

An Automatic Coding System for Open-Ended Questionnaires

—Occupational Data Coding in “Health and Social Stratification” Survey—

Kazuko TAKAHASHI

The automatic coding system for occupational data, which has been proposed by Takahashi (2000b), basically understands occupational data by the concept of "case frame" in natural language processing and classifies to the appropriate occupation-codes by using an occupational dictionary and verb and noun thesauruses constructed for this system.

This paper reports the outcome of applying this system to the practical case of “health and social stratification” survey conducted in November 2000.

In application, occupational data were first coded by the system and a person independently, then the results were compared. The pairs of agreement results were considered as correct codes, whereas the disagreement ones were discussed and settled with appropriate codes by three persons.

＊たかはし・かずこ：敬愛大学国際学部助教授　情報処理論
Associate Professor of Computer Science, Faculty of International Studies, Keiai University; information processing.

In this survey, precision and recall for the system were 80% and 70%, respectively, while both for the person were 80%. The fact that the degree of agreement between results by the system and the person was 60%, implies that they coded in different ways.

1．はじめに

本稿は、これまで実験段階にあったコンピュータによる職業データのコーディング自動化システムを、初めて本格的に活用した結果について報告するものである。

一般に、社会調査における代表的な回答形式には、被調査者にあらかじめ提示した選択肢の中から適当なものを選ばせる選択回答と、自由な記述を許す自由回答の 2 種類がある。前者はデータ収集後の処理が比較的容易で、分析手法も数多く研究されているために、特にサンプル数の大きい本調査において多用されてきた。しかし、例えば、探索的な分野で選択肢を明確に定めにくい場合や、選択肢（カテゴリー）の名称と通常の解釈との間に相違があり誤って回答される可能性がある場合などには、自由回答で収集したものを分析者が後で分類（アフターコーディング）する必要がある。今回、処理の対象とする職業は主に前述 2 番目に挙げた理由によりアフターコーディングが必要とされる典型的なデータで、通常、自由回答である「仕事の内容」（狭義の職業）と、選択回答である「従業上の地位」、「従業先事業の規模」、「役職」の計 4 種類のデータ（以後、職業データと総称する）により収集される(1)。調査終了後、職業データはすべての分析に先立って、分析者達により総合的に判断され、数百にわたる職業(2) の中から該当する職業コードを決定される必要があるが、これが「職業コーディング」と呼ばれる作業である。

髙橋（2000b）は、職業コーディングに存在する作業量の多さや煩雑さ、コーディング結果の非一貫性などさまざまな問題を指摘し、コーディングを支援するためにコンピュータによる自動化システムを提案した。システ

ムの概要は次の通りである。まず、コンピュータは回答の中心である「仕事の内容」を読み、自然言語処理における格フレームの概念に基づいて述語を探し出すと同時に、他の語、特に名詞を述語との役割関係で捉えることで回答の意味を解釈する。次に、コンピュータは解釈した意味に従って「職業辞書」（あらかじめ各職業に関する定義内容を知識としてコンピュータ上に構築したもの）を調べ、該当する職業があればそのコードを付け、なければ未決定とする。コードを決定する際は、人間と同様に、「従業上の地位」や「役職」、「従業先事業の規模」も考慮する。従って、全サンプルの職業データを電子化しておけば、システムにより一貫性のあるコーディングが自動的に行われ、前述したような問題は生じないはずである。

システムを、人間によりすでにコーディング済みの1995年 SSM 調査（Social Stratification and Social Mobility survey；社会階層と社会移動全国調査）における「本人の現職」のうち、無作為に抽出した約1,000サンプルに対して実験的に適用し、人間による結果を正解として比較したところ、有効性が示唆された（高橋 2000b）。この後、システムは、1999年に実施された JGSS（Japan General Social Surveys）[3] による「生活と意識についての国際比較調査　第 2 回予備調査」（以下、JGSS 予備調査と略する）において、職業だけでなく産業のコーディング[4] も行うべく機能が拡張されたが、これも人手によるコーディング済みのデータ（約500サンプル×5種類）[5] を用いた実験的なものであった（高橋 2000c）。

システムがより完成度の高いものとして実際に活用されるためには、精度や再現率など狭い意味での性能を高めるだけでなく、人間にとってシステムの使いやすさという観点を取り入れる必要がある。また、職業コーディングの全過程において、システムをどの時点でどのような位置付けにより適用するのが効果的であるかについての検討も必要となるが、今回、システムを実際の職業コーディングに適用する機会に恵まれたため、これらの検討を行うことが可能となった。本稿の目的は、そこで得られた結論と問題点を明らかにすることであるが、システムを今回のデータに適用した結果についても簡単に報告する。

以下、次節でデータと方法における検討結果について述べ、3節でシステムを適用して得られた結果と考察について述べる。最後に今後の課題を中心にまとめる。

2．データと方法に関する検討

1．データに関する検討

今回用いたデータは、2000年11月に実施された「健康と階層」調査における全1,236サンプル[6]である。本研究において処理の対象とするのは、「本人の現職」に関するデータで、「従業上の地位＋役職」（従業上の地位と役職のデータを同時に収集）（Q26）、「仕事の内容」（Q27）、「従業先事業の規模」（Q28）から成る[7]。

ここでは、職業データの中心である「仕事の内容」（自由回答）に注目し、回答の傾向を分析する[8]。まず、形式的にはこれまでと同様に、主語は「私」、時制は現在形で肯定文に限定されており、長さは1文以下のものが多かった。内容的には、表1に示すように文字通り仕事の内容を回答するものが多く、それ以外のものが誤って回答されることは比較的少なかった。誤った回答の中には、産業のデータである従業先事業の種類や、従業上の地位、役職などのように厳密には職業ではないが、日常生活においてはし

表1　「仕事の内容」における回答の傾向

回答の内容	回答例	出現率（%）
仕事の内容	ＮＴＴ営業　プログラマー　市場の食堂で配膳の仕事 建築の配管工事　経理事務員　部品組立　設計	92.5
仕事の内容以外		7.5
従業先事業の種類	パチンコ店　会計事務所　金融関係　市立図書館	3.2
従業先の名前	（略）	0.2
従業上の地位	市の臨時職員　内職　派遣技術者　フリーター	0.3
役職	行政部長　会社重役　団体役員	2.3
生産物・製品名	ソフトウェア　毛織物　住宅の外壁	0.2
職場名	清掃局　市役所土木課	0.2
その他	公務員（のみで他に説明なし）	1.1

ばしば職業であると認識されているものが多い[9]。

この傾向を例えばJGSS予備調査における傾向と比較すると、調査員が異なるなど条件が同じではないが、従業先事業の種類は約10倍、役職、生産物・製品名、職場名についてはほぼ同様の出現率であった。今回、従業先事業の種類が多かった理由としては、JGSS予備調査と異なり産業が尋ねられなかったために、両者を混同しても気づきにくかったためではないかと思われる。また、JGSS予 備調査においてはSSM調査と同様に、従業上の地位や、「公務員」だけで他に何も説明のない回答が出現することはなかったが、今回は少数ながら出てきており、今後は質問文で注意を促す、または調査員に指示を与えるなどの対策が必要であろう。しかし、これらの点を除外すると、これまでの調査と同様の傾向を示しており、今回もシステムの有効性が期待できる。

次に、内容的には正しく回答されていても職業を決定するためには情報が不足するものがあり、これを情報不足が明らかなものに限定してみても、各々少数ではあるが出現していた（表2参照)。表2において、対象格や場所格とは、格フレームにおいて中心的な意味を担う述語に対して他の語が果たす意味役割（深層格）の種類をいう。ここで、格フレームは、本来、図1に示すように述語の取る表層格、深層格、選択制限が結合された表現で、選択制限により構文解析における曖昧性の解消に用いられるが、システムでは、この概念を用いて意味解析を行う（その際、システムは述語の範囲を広く捉えて、動詞だけでなくサ変名詞〔例えば「製造」など〕や、職業名とし

表2　明らかに情報不足の回答例

回答例	不足する情報	候補となる職業
製造	対象格	陶磁器工　石工　ガラス・セメント製品製造作業者 その他の窯業・土石製品製造作業者　…
設計	対象格	機械・電気・化学技術者　建築・土木技術者 農林技術者　情報処理技術者　その他の技師・技術者 …
教員	場所格	幼稚園教員　小学校教員　中学校教員　高等学校教員 大学教員　盲・ろう・養護学校教員 その他の教員　…

図1　格フレームの例（「買う」の場合）

（出所）（松本 1998）

て認知されている名詞〔例えば「教員」など〕も含めている）。

システムを格フレームの概念に基づいて構築した理由は、職業分類が、基本的に、「『作る』や『教える』など動作の違い（述語）により行われ、さらにその動作が『何を』（対象格）や『どこで』（場所格）を表す名詞とどのように結びつくかにより細分化される」と解釈できると考えたためであり（高橋 2000b）、回答の中にこれらの情報が欠如するとシステムは有効に機能しない(10)。また、このような場合には、人間もコーディングが困難な場合が多い。

情報不足の回答をなくすためには、調査員が各職業に関する定義を熟知し、どのような述語が出たらどのような格が必要かということを知っておけばよいが、そのような状況を仮定することは不可能である。現実的には、質問文に付ける例を増やす、正しい例だけでなく情報不足の例も付けて注意を促すという対策が考えられよう。

以上に述べた事柄は、システムだけでなく人間によるコーディングにおいても問題となるものであった。最後に、システムにおいてだけ問題となるものについて述べておく。

代表的なものは、並列を表す記号「、」と「・」の使われ方(11)である。今回、「、」は124サンプル（約10%）、「・」は24サンプル（約２%）の回答に出現したが、表３に示すように、並列以外の用途に誤用される場合もあった。この傾向はこれまでの調査と同様で、一般的な誤りであると考えられるが、これがシステムにおいて問題となるのは、人間は前後にある語の関係から常識を用いて正しく解釈できるのに対して、システムは常識を持た

表3 「、」や「・」の用法

種類	回答例
並列	ケーキの販売・製造　　声優、タレント
助詞「の」の代用	公用車、運転業務
助詞「で」の代用	新聞社・製作　　有限会社、塗装　　製造業、現場の仕事
同上（後の語に付く）	看護婦、市立病院

ないために、これらの記号を並列表現を切り出す際の手がかりとして表層的に扱ってしまうためである。従って、例えば、「ケーキの販売・製造」のように並列表現として用いられた場合には、システムは、「ケーキの販売」と「ケーキの製造」の2つを正しく切り出すが、「公用車、運転業務」と誤記されると、「公用車の運転業務」と解釈できずに、「公用車」と「運転業務」の2つの回答があると解釈してしまう。

対策としては、調査員がこれらの記号を正しく用いて記入すること、データの入力作業者に、これらの記号が誤記された場合には、常識の範囲で適当な助詞に置き換えるよう要請しておくことが考えられる。

2．方法に関する検討

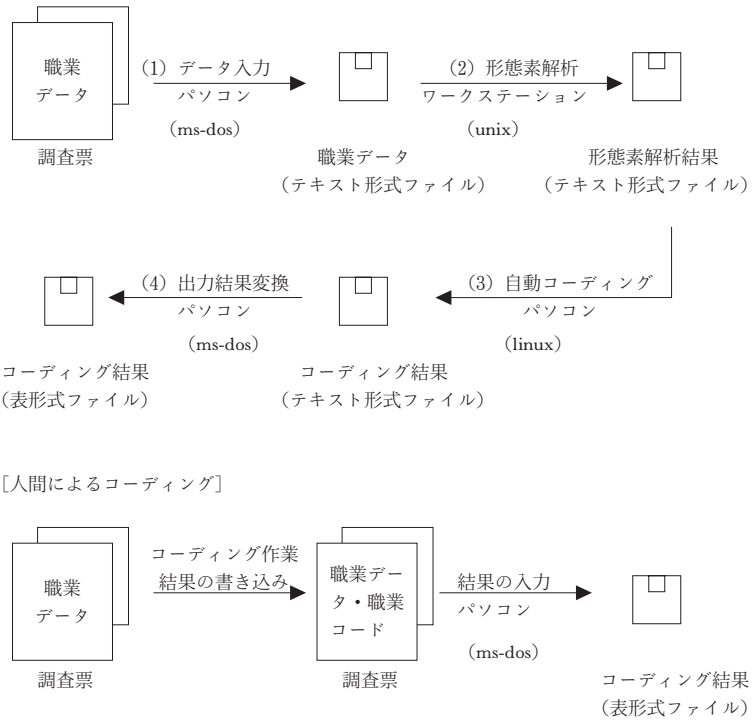
（1） システムの概要

システムは次の4段階、すなわち（1）データ入力部、（2）形態素解析部、（3）自動コーディング部、（4）出力結果変換部から構成される（図2参照）[12]。このうち、（4）はこれまでの実験段階にはなく、今回実際に適用するに当たって追加された処理である。

図2より明らかなように、システムは相異なる2種類のコンピュータと3種類のＯＳ上で稼働する。すなわち、人間とのインタフェースとなる入出力部分の（1）と（4）はパソコン上の ms-dos（windows. 以下同様）、（2）はワークステーション上の unix、システムの基幹部分である（3）はパソコン上の linux である。ここでの問題は、日本語コードが OS 間で異なることであるが[13]、その変換作業は容易である。

図2には人間が行うコーディングも示したが、システムにおいても人間においても、出発点は職業データが記載された調査票、到着点はコーディ

図2　コーディング処理



ング結果が入った表形式のファイルである点は同じである。

（1）データ入力部

ここでの中心的な作業は、職業データを入力した後、システムの要求する順序に並べ直し、（2）に渡すために、各質問に対する回答を空白（全角）で区切ったテキスト形式のファイルを作成することである（図3参照）。

調査票の段階で職業データとして収集されるデータの種類や順序は、国勢調査、SSM調査、JGSS予備調査、今回のいずれも異なっており、一律ではない。従って、当初はデータ入力にワープロ用ソフトを想定し、入力データの種類や順序を指定していたが、今回から表計算ソフトを用いて、調査票の質問通りにデータを入力することとする。表計算ソフトを利用すれば、列の移動が容易なために簡単にデータの入れ替えができ、必要な種

図3　テキスト形式にしたデータ例
（サンプル番号、Q26、Q28、Q27の順）

209␣ 8␣ 9␣ 製薬会社事務員
210␣ 8␣ 4␣ 市場の食堂で配膳の仕事
713␣ 9␣ 3␣ 声優、タレント
…

（注）␣ は空白（全角）を表す。

図4　形態素解析の失敗例

不動産業 → 不動 ＋ 産業
経営業 → 経 ＋ 営業
工事業 → 工 ＋ 事業
左官業 → 左 ＋ 官業

類のデータのみ使うことが可能である。また、データ入力の担当者にとっても、職業データも他のデータと同様に入力していけばよいこと、質問ごとに列がそろっているために作業が行いやすいというメリットがある。ただし、表計算ソフトにより入力した場合は、最終的にはデータをテキスト形式に変換する必要があるが、その作業は容易である。

ところで、データ入力において入力ミスは避けられないものであるが、システムにおいては他の場合に比べてこの問題が大きいために[14]、入力ミスのチェック作業が重要である。さらに、ここでは、（2）の形態素解析における失敗を避けるための作業、すなわちある特定の語に対する変換作業をしておくことも必要である。具体的には、文字「業」が末尾にくる語のうちの一部は、「業」と直前の文字が結びついて次のように誤って認定されてしまうため[15]（図4参照）、文字「業」を削除するような置換（例えば「不動産業」を「不動産」と置換）を行う。

このほか、1文中に半角文字が混在すると形態素解析が正しく行われないために、半角で入力されやすい「(　)」（カッコ）やカタカナの語をチェックし、半角であれば全角に置換する必要がある。

日本語コードについては、データを（2）に渡す前に、シフトJISコードからEUCコードに変換する必要がある。

（2）形態素解析部

ここでは、テキスト形式であるデータに対して形態素（日本語の場合は語と考えてよい）に区切って品詞を付ける作業を行う。これは、（3）の自動コーディング部で行われる意味解析が語を単位とし、品詞も調べるためで

図5　JUMAN Ver.3.1による形態素解析の結果（eオプション指定の場合）

表記	読み	原型	品詞	品詞コード	品詞細分類	（以下略）
↓	↓	↓	↓	↓	↓	
210	210	210	未定義語	15	その他	1 * 0 * 0
			特殊	1	空白	6 * 0 * 0
8	8	8	未定義語	15	その他	1 * 0 * 0
			特殊	1	空白	6 * 0 * 0
4	4	4	未定義語	15	その他	1 * 0 * 0
			特殊	1	空白	6 * 0 * 0
市場	しじょう	市場	名詞	6	普通名詞	1 * 0 * 0
の	の	の	助詞	9	接続助詞	3 * 0 * 0
食堂	しょくどう	食堂	名詞	6	普通名詞	1 * 0 * 0
で	で	で	助詞	9	格助詞	1 * 0 * 0
配膳	はいぜん	配膳	名詞	6	サ変名詞	2 * 0 * 0
の	の	の	助詞	9	接続助詞	3 * 0 * 0
仕事	しごと	仕事	名詞	6	サ変名詞	2 * 0 * 0
EOS						

ある。

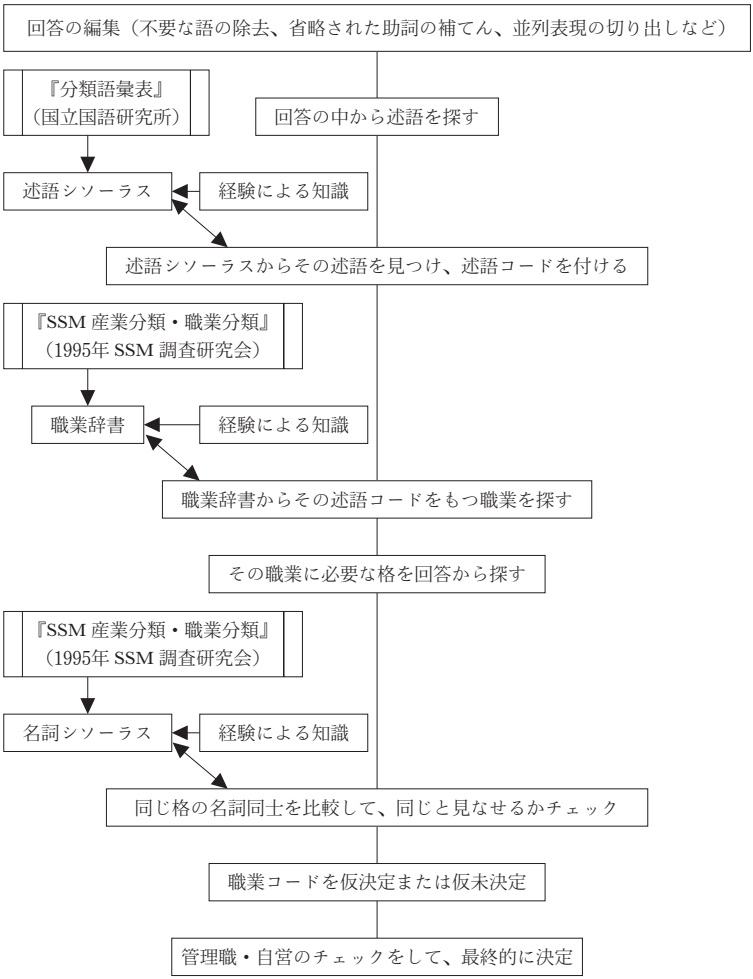
形態素解析は、当初から形態素解析用ソフトJUMAN（黒橋・長尾 1998）を利用してきた。図5にJUMANによる結果を示すが、（3）で用いるのは、語の「原型」と「品詞」または「品詞細分類」である。日本語コードについては、（3）は（2）と同じEUCコードであるために変換を行う必要がない。

（3）自動コーディング部

システムの最も重要な部分で、職業データに対して、数百ある職業の中から該当するコードを1つ選んで付け、該当するものがない場合には未決定のコード「999」を付ける（図6）。その際、「仕事の内容」中に不要な語（例えば、「等」、「こと」など）や品詞（例えば、形容詞や副詞）があれば除去し、助詞が省略されていれば補って（例えば、「建具製作」→「建具を製作」「建具で製作」）[16]、回答の内容と形式を自動的に整備する。また、並列表現により複数個回答されている場合には最大4個まで切り出して、各々に職業コードを付ける。

自動コーディングを効率よく行うために、システムは次の4つの特徴をもつ。

図6　自動コーディング部



(1)　コンピュータに職業の定義内容や回答の意味内容を理解させるために、格フレームの概念を利用する。

(2)　コンピュータが語の意味を柔軟に解釈することができるように、述語と名詞に対して、それぞれ述語シソーラスと名詞シソーラスをもつ。ここで、シソーラスとは、語と語を意味的な上下関係や類似関

係に注目して関係付けて整理するものである。

⑶　職業の定義内容をまとめた辞書（職業辞書）をもつ。

⑷　「仕事の内容」により決定された職業コードに対して、「従業上の地位」、「従業先事業の規模」、「役職」を調べ、管理職や自営関係のチェックを行って最終決定とする。

⑴は前述したように、文の意味を形式的にコンピュータで扱うことを可能にするためにシステムで採用した基本的な戦略である。⑵において、2つのシソーラスは図7、図8に示すように形態的には異なる。述語シソーラスでは、職業を理解する上で同じ意味をもつと考えられる述語（例えば、「製造」と「作る」）に対して、品詞が異なっていても同一の述語コードが付けられる。名詞シソーラスでは、職業の定義内容を表現する語と回答に出現する語の抽象度レベルの相違（例えば、「果樹」と「ミカン」）や、日本語に特有の表記のゆれ（例えば、「蜜柑」「みかん」「ミカン」）が吸収される。

⑶において、職業辞書では各職業の定義内容を格フレームの形式で表現するが、述語シソーラスとの関連から述語そのものではなく述語コードを用いる（図9参照）。また、必要な格にくる名詞は、名詞シソーラスにおける代表語レベルの話である。述語によっては複数の職業が対応するものもあるが、職業の違いにより必要な格にくる名詞が異なる。

⑷は、特に管理職や自営に関して、「仕事の内容」の情報のみでは誤って決定される可能性があるために設けられたチェックである。例えば、「仕事の内容」が「会社の管理」である場合、この情報からは、「548　会社役員」か「550　会社・団体の管理職員」のいずれかの管理職であると判断されるが、管理職（職業コード「545」～「553」）は、「従業上の地位」、

図7　述語シソーラス

述語	述語（ふりがな）	述語コード
↓	↓	↓
製造	せいぞう	386　1
製作	せいさく	386　1
作る	つくる	386　1
…		

図8　名詞シソーラス

代表語	例
↓	↓
果樹	蜜柑　みかん　ミカン
	林檎　りんご　リンゴ
	…

図9　職業辞書

```
述語コード　職業コード　必要な格と名詞　（以下、もしあれば繰り返し）
    ↓          ↓          ↓                    ↓
386  1          599  （を  穀物  野菜  果樹）    623  （を  陶磁器）…
…
```

599は農耕・畜産作業者、623は陶磁器工・絵付け作業者の職業コードを表す。

「従業先事業の規模」、「役職」における一定の条件[17]を満たす必要があり、最終的な決定の前にこれらのチェックを行う必要がある。逆に、「仕事の内容」により管理職以外の職業や「未決定」と判断されていても、このチェックにより管理職に決定される場合もあり得る。これは、自営の場合も同様である。

自動コーディング部のプログラムは、リスト処理に適したLISP言語により開発したが、約900ステップの大きさとなった。

（4）出力結果変換部

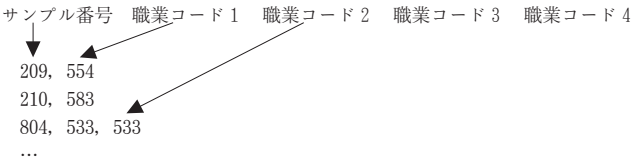
システムが実験ではなく実際に活用されるためには、コーディングの結果が多数の人間が検討できるように見やすいものでなくてはならない。また、これを例えば人間による結果と比較するなど他の処理にリンクすることができると便利であり、この点についても考慮する必要がある。

以上より、今回、システムの出力結果を表計算ソフトで読み込める形式（CSV形式）（図10参照）に変換するためのプログラムを作成した。出力結果を表計算ソフトで見ることができれば、人間による結果との比較が簡単に行えるだけでなく、各種の統計処理も容易になる。

（2）　システムの位置づけ

一般に、職業コーディングは次のように人間により3回行われる。すなわち、1度すべてのコーディングが行われた後、2回目に別人によるチェック作業が行われ、3回目に職業コーディングに熟練した専門家数人により正解が協議されて、最終的に決定される。今回、システムはデータ入力が完了したファイルを受け取り、人間によるコーディング（1人）と並行して独立に実行された。その後、両者による結果を比較して、一致するものはそのまま正解とし、不一致のものに対してのみ3人による協議が行われ

図10　最終出力結果（CSV 形式）



て正解が決定された。これより、職業コーディング全体におけるシステムの位置づけとしては、従来において 2 回目に行われる「別人によるチェック作業」に相当するものと考えられる。

このように、チェック作業を人間ではなくシステムで行う場合のメリットとしては、第一に作業量の軽減化があるが、その他に、システムは人間のように不注意による見落としや勘違いをしないこと、また、深読みをしないことなど人間とは別の見方で情報を解釈するという点も大きい。デメリットとして、浅い意味解析しか行わない点があるが、互いに性格を異にする両者によりコーディングが行われることで、より速くより正確な結果を得ることが期待できる。その際に、両者の結果が一致した場合にはそのまま正解とすることは妥当な判断であると考えられる。

3．結果と考察

1．結果

（1） 精度と再現率

表 4 に、3 人による協議により最終的に決定された職業を「正解」としたときのシステムと人間によるコーディング結果を示す。ここで、精度と再現率は情報検索において性能を示す指標あり、職業コーディングにおいては、それぞれ次式により計算される。

精度　＝正しく決定された個数／決定された個数

再現率＝正しく決定された個数／コーディングされ得る個数（＝サンプル数）

表4　システムと人間によるコーディング結果の比較（n'＝1221の場合）

	決定された個数	正しい個数	精度（%）	再現率（%）
システム	1037	847	81.7	69.4
人間	1199	976＊	81.4＊	80.0＊

（注）＊：人間によるコーディングでは「一般事務」をすべて「558　その他の一般事務員」と決定していたが、これを「554　総務・企画事務員」に訂正し、正しい個数として計算した。

式から明らかなように、いわゆる正解率と呼ばれるもの、すなわち「全体のどのくらいが正しくコーディングできたのか」については、職業コーディングの場合、再現率で示される。ここで、今回は、最終的にn＝1236サンプルすべてに対してコードを付けることができず、15サンプルが不明（「999」）とされたため、コーディングされ得る個数をn'＝1221（＝1236－15）サンプルとして計算した。もし、「999」も1つのコードとみなしてn＝1236サンプルで計算すると、システムは「999」を含めて861個を正しくコーディングしているために、精度81.9%、再現率69.7%となり、人間は「999」がなく正しい個数が同じであるために、精度は81.4%で、再現率は79.0%となる。なお、システムと人間による結果の一致率は60.0%であった。

（2）　システムにおけるコーディングの傾向

今回最終的に決定された職業の種類は、未決定（999）を含めて133種類で、職業全体（194種類）[18]の68.6%を占めた。職業別にシステムの正解率（再現率）を算出すると、100%のものが35種類（26%）で、0%のものが24種類（18%）であった（表5参照）。

職業ごとの正解率を、紙面の制約上、職業の出現数が高い（12個〔＝全サンプルの10%〕以上）ものについてのみ表6に示す。また、この中で正解率が50%を下回る職業については、表7にシステムが付けたコードを示す。

（3）　処理時間

コーディングに費やされた時間は、システムでは2時間（データ入力の時間を除く）、人間では8時間×1人（結果入力の時間を除く）であった。これより、1サンプルの処理にかかる平均時間を単純に計算すると、システムでは5.8秒、人手では23.3秒となる。

表 5　職業別にみた正解率の度数分布

正解率	種類（カッコ内%）	累積度数（%）
100%	35（26）	26
90%～	6（5）	31
80%～	20（15）	46
70%～	7（5）	51
60%～	14（11）	62
50%～	6（5）	66
～50%	21（16）	82
0%	24（18）	100
計	133（100）	－

前述したように、システムと人間によるコーディング結果の一致度は、741サンプル（60.0%）であるために、決定できなかったものも含め、残り495サンプル（40.0%）のデータに対して正解を協議する必要がある。作業は 3 人で延べ26.5時間かかり、 1 サンプル当たり平均3.2分（＝1.1分× 3 人）を要した。

2．考察

表 4 より、システムによる結果を人間と比較すると、精度がほぼ等しく、再現率が10%程度低いことがわかる。これは、システムの開発に当たって精度の高さを優先したために、トレードオフの関係にある再現率が犠牲となったためである。しかし、システムの利用者から見れば、確実な結果を得る方が望ましいものと思われるため、今後もこの方針は継続する。精度をより高めるためには、辞書やシソーラスの充実をはかる必要があるが、特に、出現頻度が高く正解率が低かった職業における間違い方（例えば680番台の職業など）に注目することが効果的であると思われる。

処理時間については、システムの方が人間より速いことは明らかである。また、システムの適用により、人間による 2 回目のコーディングの必要性がなくなることと、最終的な協議の際にすべてのサンプルをチェックしなくて済むために、これらの時間も軽減される。コーディング 3 回分の処理

表 6　職業ごとの正解率（職業の出現数が12以上のもの）

職業コード	職業分類	正しい個数	正解の出現数	正解率(%)
701*	スーパーなどのレジスター係員・キャッシャー	16	16	100.0
548	会社役員	24	25	96.0
514	看護婦、看護士	15	16	93.8
599	農耕・養蚕作業者	54	58	93.1
687	清掃員	21	23	91.3
652	縫製工、裁断工	19	21	90.5
607	自動車運転者	31	35	88.6
566	小売店主	22	25	88.0
579	理容師、美容師	14	16	87.5
557	営業・販売事務員	30	34	85.7
539	個人教師	10	12	83.3
581	料理人	10	12	83.3
578	女中、家政婦、家事サービス従事者	17	21	81.0
569	販売店員	54	67	80.6
568	飲食店主	12	15	80.0
559	会計事務員	41	53	77.4
550	会社・団体管理職員	10	13	76.9
506	情報処理技術者	11	15	73.3
554	総務・企画事務員	137	190	72.1
645	味噌・醤油・缶詰食品・乳製品製造工、飲食料品製造作業者	5	12	41.7
686	運搬労務者	6	16	37.5
682	土工、道路工夫	9	25	36.0
630	金属工作機械工、めっき工、金属加工作業者	5	18	27.8
688	その他の労務作業者	3	13	23.1
704*	製品製造作業者	1	27	3.7

（注）　701*：SSM 職業分類「559　会計事務員」から分離して今回追加した職業コード.
704*：対象格が無い製品製造作業者（表 2 参照）を未決定（999）としないために今回追加した職業コード.

時間を単純に計算すると、システムを利用する場合は13.2時間（＝ 8 ＋ 2 ＋ 8 × 0.4)、しない場合は 2 4 時間（＝ 8 ＋ 8 ＋ 8 ）で、軽減される時間は10.8時間（45%）となる[19]。

表７　正解率が低かった職業（職業の出現数が12以上のもの）

職業コード	正しい個数	未決定の個数	間違えた個数	間違えた職業コード（個数）	精度（%）
645	5	5	2	566（1）　569（1）	71.4
686	6	3	7	569（1）　607（3）　672（3）	46.2
682	9	2	14	678（1）　679（2）　684（4） 688（5）　702（2）	39.1
630	5	9	4	627（2）　628（1）　674（1）	55.6
688	3	6	4	550（1）　684（2）　685（1）	42.9
704	1	20	6	550（5）　633（1）	14.2

処理時間は結果が出るまでにかかる時間であるが、システムを利用する場合は人間による場合と異なり、その間に人間が作業をする必要のない時間帯も発生する。従って、処理時間とは別種の時間、すなわち人間の作業時間についても考察しておく必要がある。作業時間を単純に計算すると、システムを利用する場合は11.2時間（＝８＋０＋８×0.4）、しない場合は処理時間と同じ24時間で、軽減される時間は12.8時間（46.7%）となる。ここで、システムを利用する場合の２回目のコーディング作業は、コンピュータ操作があるために厳密には無ではないが、数分程度のため四捨五入して０時間とみなした。

ところで、システムは、当初から人間による職業コーディングの支援を行うものとして開発され、利用のされ方としては次の２つが想定されていた（高橋 2000b）。１つは今回のようにチェック用としての適用であり、もう１つは、まずシステムを適用し、システムが未決定としたものに対してのみ人間がコーディングを行うというものである。前者の場合は、１回は全サンプルに対して人間によるコーディングが不可欠であるが、後者の場合は、人間は一部のサンプルに対してのみ行えばよいために作業の軽減化が著しい。しかし、システムの精度が80%程度にしか達していない現状では、後者のような利用は無理があり、前者の利用が現実的である。また、今回、システムと人手による結果の精度がともに80%程度であったにもかかわらず、両者の一致率が60.0%でしかなかったことを考慮すると、システムと人間とは異なる観点によりコーディングを行っているものと考えら

れる。従って、従来のように人間だけで3回行う方法よりも、今回のように両者が独立にコーディングを行って結果を比較し、一致したものはそのまま決定とし、一致しないものに対して深く検討する方法の方が、より信頼性の高い結果を得ることができるものと判断できる。

4．おわりに

今回、システムを初めて実際の調査に活用し、システム自身の評価を行うと同時に、職業コーディングの過程においてどのような適用の仕方がよいかを検討した。その結果、システムの精度が80%程度である現時点においては、人間とシステムの両方によるコーディングを独立に実行して結果を比較し、両者が不一致のものに対してのみ正解を協議するという方法が妥当であるとの結論を得た。また、実際に活用するに当たっては、結果の見やすさや他の処理とのリンクの取りやすさについても考慮した結果、システムの出力結果を表計算ソフトで読み込める形式に変換する機能の追加を行って、支援システムとしての機能を高めることができた。

今後の課題は、システムの性能、特に精度を上げることであるが、一般の研究者が容易にシステムを扱えることができるように、使いやすいものにすることも必要である。具体的には、次の2つを課題とする。

⑴　JUMAN の形態素辞書を職業コーディング用に改良すること。

⑵　データ入力部や出力変換部だけでなく、形態素解析や自動コーディングもパソコン上の ms-dos で稼働させること。

また、現在、システムの各段階は分断されており、フロッピーディスクによるデータの受け渡しを行っているが、システムの利用者にとっては、（2）～（4）が連続して稼働すると便利である。これは将来の課題としたい。

システムは、今回の適用直後に、2000年11月に実施された JGSS-2000[20]（約3,000サンプル×5種類）における職業・産業コーディングにも活用され、今回とほぼ同様の位置づけによる支援を行ったが、これについては稿を改

めて報告したい(21)。

〔謝辞〕
システムを「健康と階層」調査に実際に活用する機会を与え、システムの位置づけや使いやすさについての検討をして下さった東京大学社会科学研究所石田浩教授に大変感謝いたします。また、SSM職業分類の使用に当たり、東北大学大学院文学研究科原純輔教授に快諾していただいたことについて感謝いたします。
なお、「健康と階層」調査は、文部科学省科学研究費（基礎研究 A(2)「福祉社会の価値観に関する実証的研究　1999―2001年度」）（研究代表　東京大学・大学院人文社会系研究科武川正吾）の一環として行われたもので、武川正吾教授に感謝いたします。

（注）
(1)　例えば、国勢調査や SSM 調査（本文後述）では、これらに、自由回答である「従業先事業の種類」と「従業先事業の名前」（いずれも産業データ）を加えた計 6 種類、JGSS 調査（注 3 参照）では「従業先事業の種類」のみを加えた 5 種類で収集される。
(2)　例えば、1970年国勢調査で用いられた職業分類をもとに作成された SSM 職業小分類は、85年版が288種類、95年版が189種類（「501　自然科学系研究者」～「688その他の労務作業者」）（『SSM 産業分類・職業分類（95年版）』）であった。
(3)　日本版 GSS（JGSS）は、大阪商業大学比較地域研究所が、文部科学省から学術フロンティア推進拠点としての指定を受けて（1999―2003年度）、東京大学社会科学研究所と共同で実施している研究プロジェクト（研究代表：谷岡一郎、仁田道夫）で、時系列分析が可能な継続的かつ総合的社会調査のデータを蓄積し、データの二次的利用を目的として公開する。
(4)「従業先事業の種類」を読んで、20種類の産業大分類コード（「10　農業」～「91　卸売」、「92　小売業」、「93　飲食店」～「180　公務」）を付ける。
(5)「本人の現職」、「本人の最後職」、「本人の初職」、「配偶者の職業」については産業データと職業データ、「父親の職業」は職業データが収集された。
(6)　被調査者の属性は次の通りである。
　性別：男580人、女656人
　年齢：20代120人、30代211人、40代224人、50代259人、60代254人、70代141人、80代以上27人
　教育歴：（新）中学（旧）小・高小270人、（新）高校（旧）中学614人、（新）短大111人、（新）大学（旧）高専大234人、不明 7 人
(7)　具体的には次のような質問がなされた（Q26とQ28については、選択肢も示す）。
　Q26．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。1 つだけお答え下さい。現在お仕事をしておられない方は、最後にしていたお仕事についてお答え下さい。
　1（ア）経営者・役員
　2（イ）常時雇用の一般従事者（役職なし）
　3（ウ）常時雇用の一般従事者（職長、班長、組長）
　4（エ）常時雇用の一般従事者（係長、係長相当職）
　5（オ）常時雇用の一般従事者（課長、課長相当職）
　6（カ）常時雇用の一般従事者（部長、部長相当職）
　7（キ）常時雇用の一般従事者（役職はわからない）

8（ク）臨時雇用・パート・アルバイト
9（ケ）自営業主・自由業者
10（コ）家族従業者
11（サ）内職
12（シ）一度も仕事についたことはない
13　　わからない

［Q27〜Q28は Q26で 1〜11と答えた人に］

Q27. あなたは通常、どのようなお仕事（あるいは最後にしていたお仕事）をしていますか。仕事の内容を具体的にお聞かせください（例えば、小学校教員、農作業、〔以下略〕）。

Q28. あなたが働いている場所（あるいは最後に働いていた場所）の従業員は、会社・組織全体で何人くらいですか

1（ア）1人　　7（キ）300〜499人
2（イ）2〜4人　　8（ク）500〜999人
3（ウ）5〜9人　　9（ケ）1000人以上
4（エ）10〜29人　　10（コ）官公庁
5（オ）30〜99人　　11　　わからない
6（カ）100〜299人

(8) 今回は調査の性質上、無職がなく、サンプルすべてに何らかの回答がある点が特徴である。

(9) ただし、「農業」や「林業」などのように、産業でもあり職業でもあるものがあり、単純には区別しにくい場合もある。

(10) 実はこのような回答であっても、システムは「従業先事業の種類」など産業データを参照することにより、条件付きではあるが、不足する情報を補うことが可能である。しかし、今回については産業データが収集されていないためにその機能を用いることができない。

(11) 「、」は、厳密には並列の記号ではなく、語句の区切りに用いられるが、職業データの場合には、並列を表現する記号として用いられることが多い。これまでの調査と同様に、今回も「、」と「・」に意味の差が特に見られないために同一に扱う。

(12) システムの詳細については高橋（2000b）を参照のこと。

(13) ms-dos ではシフト JIS コード、unix と linux では EUC コードが用いられる。

(14) 自動コーディングは形態素解析の結果に基づいて行われるため、ミスがあった場合は再度、形態素解析からやり直さなければならない。

(15) これを避けるために、現在、形態素解析を行うソフト JUMAN の形態素辞書を改良中である。

(16) 対象格と場所格のどちらが妥当であるかの判定はせず、単純に両方を補っている。

(17) 『1995年SSM調査　コード・ブック』によると、管理職については次のようにコードする。

①従業上の地位が役員または自営業主の場合
　規模5人未満…必ず管理的職業以外の仕事の内容でコードする。
　規模30人未満…管理的職業以外の仕事の内容を優先してコードする。
　規模30人以上…原則としていずれかの該当する管理的職業でコードするが、それ以外の仕事の内容が書いてあれば、それに従ってコードする。
②従業上の地位が一般従業者や家族従業者である場合
　役職が課長以上……①と同様。
　役職が課長補佐以下…必ず管理的職業以外の仕事の内容でコードする。
③専門的管理職（設計技師長、病院長、学校長など）は「専門」の方を優先する。

(18) 今回は、SSM 職業小分類である188種類に、「701」（表 6 注参照）など新たに 6 種類を追加した職業分類を用いた。

(19)　実際には、3節1．(3)で述べたように、最後の検討を行う3回目の処理時間は、システムを利用した場合（サンプルの40%を検討）でも延べ26.5時間を要しており、利用しない場合（全サンプルを検討）の時間を想定することができない。従って、単純計算の結果は参考程度でしかない。

(20)　JGSSにより2000年から2003年にかけて計4回実施予定の「生活と意識についての国際比較調査」本調査の中の第1回目調査を意味する。

(21)　大阪商業大学比較地域研究所・東京大学社会科学研究所編『日本版 General Social Surveys 研究論文集JGSS-2000で見た日本人の意識と行動』(2002年3月発行予定）に掲載。

(参考文献)

1995年SSM調査研究会（1995)、『SSM産業分類・職業分類（95年版)』。

1995年SSM調査研究会（1995)、『1995年SSM調査　コード・ブック』。

国立国語研究所（1964)、『分類語彙表』、秀英出版社。

黒橋禎夫・長尾真（1998)、『日本語形態素解析システムJUMAN Version 3.61』、京都大学大学院情報研究科。

松本裕治（1988)、「意味と計算」『言語の科学4　意味』、岩波書店。

髙橋和子（1999a)、「自然言語処理に基づく自由回答の処理支援について」、日本行動計量学会計量社会学研究会（於立教大学)。

髙橋和子（1999b)、「自然言語処理に基づく自由回答のコーディング支援――格フレームによるSSM職業コーディング自動化システム」、日本行動計量学会行動科学研究会、第6回（於東北大学)。

髙橋和子（2000a)、「格フレームによる職業コーディングの自動化支援システム」『言語処理学会第6回年次大会発表論文集』(於北陸先端科学技術大学院大学)、155－158ページ。

髙橋和子（2000b)、「自由回答のコーディング支援――格フレームによるSSM職業コーディング自動化システム」、数理社会学会論文誌『理論と方法』Vol.15, No.1、149－164ページ。

髙橋和子（2000c)、「日本版General Social Surveys（JGSS）の調査方法論上の問題について（4）産業・職業コーディング自動化支援システム」、日本社会学会、第73回（於広島国際学院大学)、25ページ。

髙橋和子（2000d)、「自由回答データの分析――格フレームによる産業・職業コーディング自動化システムを中心として」、日本分類学会・日本行動計量学会共催シンポジウム資料集『テキスト型データの取得から活用まで』(於統計数理研究所)。

髙橋和子（2001)、「職業コーディング自動化システムの実用化――『健康と階層』調査における活用例」『第32回数理社会学会大会研究報告要旨集』(於群馬大学)、38－41ページ。