職業・産業コーディング自動化システムの活用

高橋　和子

敬愛大学国際学部

１．はじめに

社会調査においては、職業や産業は通常、調査票から得られた生データがそのまま分析に用いられることはなく、自由回答法と選択回答法からなる複数個の質問により収集されたデータをコーダーが総合的に判断して決定したものを用いる。この作業は職業・産業コーディングと呼ばれ、調査終了後すべての分析に先立って行われる必要があるが、カテゴリーである職業・産業分類の個数（職業は約200種類、産業は約20種類）と内容が多岐にわたるため、作業量の多さや煩雑さだけでなく、コーディング結果の一貫性が保証されにくいという問題がある。

高橋（2000a）は、これらを解決するためにはコンピュータによる支援が必要であるとして、コーディングの自動化システムを提案した。システムの特徴は、職業や産業は基本的に格フレームによる意味表現が可能であるとして、回答や職業・産業分類に対し格フレームに基づいた意味解釈を行う点にある。システムを1995年SSM（Social Stratification and social Mobility）調査（約1000サンプル）やJGSS（日本版General Social Surveys）[1]第２回予備調査（1999年11月実施、約800サンプル×5種類）の職業や産業データに実験的に適用した結果、いずれも有効性を示した（高橋2000b、2000c）。

この結果を踏まえて、今回、システムは実際に「健康と階層」調査（2000年11月実施、約1200サンプル）、続いて「JGSS第1回本調査（JGSS-2000）」（2000年11月実施、約2900サンプル×5種類）に活用されたので、この結果について後者を中心に報告する。

２．データ

今回対象としたデータは、「健康と階層」調査では職業（本人の現職）、「JGSS-2000」では職業と産業（本人現職、本人最後職、本人初職、配偶者職、父職の５種類）である。調査により多少の違いはあるが、通常、職業は自由回答である「仕事の内容」を中心に選択回答である「従業上の地位」、「従業先事業の規模」、「役職」を組み合わせた計４種類、産業は自由回答である「従業先事業の種類」と「従業先事業の名前」の計２種類のデータから構成される。

職業・産業コーディングの成否は、人手、システムのいずれにおいても、自由回答に記述された内容の質に強く依存するために、ここでは、回答の内容が、職業・産業を決定するのに十分な情報であるかどうかを検討する。

自由回答で記述された語の個数は平均６個／１サンプルで、産業である「従業先事業の種類」については、従業先の名前や生産物、製品名のみの回答もあったが、比較的適切な情報が提供されていた。一方、職業である「仕事の内容」では情報が不足する回答が目立った（表１）。例えば、回答の記述が「製造」だけの場合、候補となる職業は「陶磁器工」、「石工」、「ガラス・セメント製品製造作業者」など多数あり決定できない。職業は産業が大分類しか行わないのに対して小分類まで行うためにより詳細な情報が必要であるにもかかわらず、質問の順番が後にあるためか情報が省略されやすい傾向がある。

情報不足の場合には、システムも人間と同様に「従業先事業の種類」を参照するが、そこでも必要な情報が得られない場合は決定することは不可能である。従って、回答に十分な情報が記述されるように、質問文に適切な回答例を提示したり、回答欄を工夫するなどの対策を講じる必要がある。

表１　「仕事の内容」における情報不足の回答例（JGSS-2000）

不足する情報	回答例
格フレームにおける対象格	事務、オペレータ、工事、メンテナンス、設計、製造、加工、技術指導、検品、整備、組立、指導員、工員、仕分け、検査、育成、管理、研究、調査 部品の製造*、製品検査*
同　場所格	現場、現場作業、現場監督、教師、非常勤講師
その他	一般、ウチの仕事で作業、作業員、印刷機械を受け持つ、ノーコメント

＊　対象格を有していても、「部品」や「製品」のように名詞が具体化されない場合は情報不足となる。

1　日本版General Social Surveys(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である。

3．方法

3.1 コーディング自動化システムの位置付け

システムを利用した場合、職業・産業コーディングは次の過程で行われた（①～④）。ここで、「人間」とは職業コーディング経験者を含む大学院・学部学生の計７名で、２人１組とした（西村・石田2001）。

① システムがコーディングを行う。
② 人間がコーディングを行う。
③ 人手とシステムの結果を比較し、一致しないものに対してシステムの結果を参考にしながら人間が再コーディングを行う。
④ 専門家によりすべてを見直して、必要ならばコードを付け直す。

3.2 コーディング自動化システムの概要

3.2.1 システムにおける処理の流れ

システムにおける処理の流れは次の通りである。

（1） データ入力
（2） 形態素解析
（3） 自動コーディング
（4） 出力結果変換

処理の中心は（3）自動コーディングで、回答の編集を行って、職業・産業データに対して該当する職業・産業コード（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。

回答の編集は、回答中に不要な語（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」）、回答の内容と形式を自動的に整備する。また、並列表現がある場合は、最大４個まで切り出す（例えば、「野菜の生産・販売」→「野菜の生産」と「野菜の販売」）。

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シソーラスを作成した（図２、図３）。前者では、職業を理解する上で同じ意味を持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を付ける。後者では、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、表記のゆれ（例えば、「蜜柑」「みかん」「ミカン」）を吸収する。

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

図１ 述語シソーラス

代表語	用語例
↓	↓
果樹	蜜柑 みかん ミカン
林檎	りんご リンゴ

図２ 名詞シソーラス

辞書はカテゴリーである職業・産業の定義内容を格フレームに基づいた形式で記述したもので、それぞれ職業・産業辞書と作成した（図３）。辞書において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語である。

述語コード	職業コード	必要な格と名詞（以下、繰り返し）
↓	↓	↓
386 1	599	（を 穀物 野菜 果樹）
	623	（を 陶磁器） ・ ・ ・

図３ 職業辞書（産業辞書も同様）

その他、職業において管理職や自営等は「仕事の内容」だけでなく「従業上の地位」、「従業先事業の規模」、「役職」も含めて総合的に判断されるため、最終的な決定の前にこれらの条件を満たすかどうかのチェックを行う。

4．結果

4.1 精度と再現率

「正解」は３人の協議による最終決定をとした。システムのコーディング結果を表２、表３に示す。職業の場合は、人手による結果（3.1の②）との比較も示した。ここで、精度と再現率はそれぞれ次式により計算した。

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（＝有効サンプル数）

職業コーディングにおけるシステムの精度は 76.4～84.3%、再現率は 61.1～68.9%で、人手と比較するとすべてにおいて精度が高く再現率が低い。この傾向は「階層と健康」調査でも同様で、システムの精度は81.7%（人手81.4%）、再現率は 69.4%（人手80.0%）であった。産業コーディングにおいては、精度は90.4%～93.0%で、再現率は71.6%～74.9%であった。

なお職業コーディングにおいて、システムと人手による結果の一致率は、本人最後職、本人現職、父職、本人初職、配偶者職の順に高く、それぞれ 63.1%、62.3%、60.2%、59.1%、57.9%で平均60%程度であったが、「階層と健康」調査でも同様で60.0%であった。

4.2 システムによる職業コーディングの傾向

システムと人手によるコーディング結果の一致率がいずれも約60%程度でしかないことから、両者におけるコーディングの傾向は異なるものと考えられる。有効サン

表2　職業コーディングの結果（単位：%）（JGSS-2000）

	本人現職		本人最後職		本人初職		配偶者職		父職	
	精度	再現率	精度	再現率	精度	再現率	精度	再現率	精度	再現率
システム	80.0	66.5	81.0	68.3	84.3	68.9	77.9	64.2	76.4	61.1
人手	78.7	78.1	73.1	72.1	81.2	79.0	70.7	68.8	75.7	70.7
両者の差	1.3	-11.6	7.9	-3.8	3.1	-10.1	7.2	-4.6	0.7	-9.6

表3　産業コーディングの結果（単位：%）（JGSS-2000）　*父職は産業コーディングを行わない

	本人現職		本人最後職		本人初職		配偶者職	
	精度	再現率	精度	再現率	精度	再現率	精度	再現率
システム	90.4	74.5	92.3	74.9	93.0	74.4	93.0	71.6

ブル数が最も多かった本人初職（2767サンプル）において未決定を除くと、システムだけが正解だったものは全体の約7%（181サンプル）、非正解だったものは全体の約10%（285サンプル）であった。システムだけが正解だったものは、サンプル数、職業の種類のいずれにおいても人手だけが正解である場合の約6割（=181／285、54／89）であった（出現した職業の種類は158種類）。

なお、システムと人手の両方が非正解だったものは3.1%（86サンプル）であった。このうち、システムと人手が同じ間違いをしているものは1.3%（37サンプル）で、その約4割は正解が「558 その他の一般事務員」または「559 会計事務員」であるものを「554 総務・企画事務員」にコーディングしていた。

4.3　処理時間

システムは職業と産業コーディングを同時に行う。延べ約14,500サンプル（=約2,900×5種類）に対して、自動コーディング部と事前の処理は「時間」、形態素解析部と出力結果変換部は「分」単位の時間を要したが、これ以外に自動コーディング部において生じた日本語コードの問題の解決時間を含めても6日で完了した。人間は3.1の②と③で延べ63日（=9日×7人）、平均で31.5日を要した。従って、システムの処理時間は人手の1/5程度であるが、「健康と階層」調査でも同様であった。

5．考察

5.1　職業コーディング

表2と「健康と階層」調査の結果より、システムの現段階での性能は、正しくコーディングする個数は全体の7割弱で人手より劣るものの、正確さにおいては人手よりやや優れていて、コーディングした中の約8割は正解である。また、システムと人手は約6割程度しか結果が一致しておらず、両者は別の見方によるコーディングを行っているものと考えられる。従って、従来のように人間によるコーディングを3回繰り返すよりも、今回のように1回をシステムに代える方が、処理時間の短縮化だけでなく内容的にも有効であると判断できる。

システムは、本人の中では初職、最後職、現職の順に結果がよかったが、これは、古い情報を持つ回答ほど、これまでに「職業辞書」に蓄積された知識や「シソーラス」に登録された語がうまく活用できたためであると考えられる。従って、システムの性能を向上させるために、辞書の知識やシソーラスの登録語を充実させることが有効である。特に、最新情報である現職の再現率が最も悪い結果であったのは、辞書やシソーラスにない新しい職業やカタカナなどの未知語（新語）に対応できず未決定としたためであると考えられる。システムは毎回、処理結果を辞書やシソーラスに反映することでバージョンアップを図っているが、常に新たな情報が出現するために、現職における再現率の向上には限界がある。

ここで、システムと人手の両方において、本人と本人以外の間で精度・再現率ともに差があったが、これは両者で回答の詳しさに差があったためであると考えられ、十分な情報をもつ回答を収集することの重要さが裏付けられたと判断できる。

システムによる職業コーディングの傾向をみると、システムだけが正解の主なものは、大分類が「専門・技術」である「503 機械・電気・科学技術者」（12個）、「事務」の「555 受付・案内事務員」（7個）や「557 営業・販売事務員」（8個）、「販売」の「569 販売店員」（15個）や「573 外交員」（16個）、「運輸・通信」の「607 自動車運転者」（7個）などであった。これらに共通することは回答の形式や出現する語が定型的な場合が多く、ルールに従って処理するシステムにとって一貫性のある正しい処理を行うことが容易なことである。これに対して、人手では複数人がコーディングするためかバラツキが多く、例えば、前述した「503」は5種類、「555」は4種類、「557」は6種類、「569」は8種類の誤ったコードが付けられていた。

一方、未決定を除いてシステムだけが非正解であった

ものの多くは製造作業者（「629」～「659」）（146個）を「704 製品製造作業者」とコーディングしたことによる（84個）。特に「629 化学製品製造作業者」においてはすべてが「704」であった（6個）。ここで、「704」は、情報不足が予想される父職に多数出現すると思われる「未決定」を減らす目的で今回追加されたコードであるが、これにより、システムは、格フレームにおける述語が「製造」（３８６　１）で対象格が欠落したものすべてにこのコードを付けてしまった。同様に、「専門・技術」の「521 小学校教員」～「523 高校教員」のいずれかで場所格が欠落したものすべてが「703 教員」とした（11個）。

類似の失敗としては、「事務」の「560 郵便・通信事務員」も場所格で決まるが、これを捉えることができなかったものを「555 受付・案内事務員」など他の事務員としていた（6個）。なお、製造作業者で次に多い失敗は製造内の他の職業に誤ったもの（26個）で、3番目は「688 その他の労務作業者」としたもの（11個）であった。さらに、「建設作業者」の「682 土工、道路工夫」の失敗はすべて「労務」の「688 その他の労務作業者」（6個）で、「686 運搬労務者」の失敗のほとんどが「運輸・通信」の「607 自動車運転者」（5個）であったが、これらの区別は人間でも難しい面がある。

なお、システムのコーディングの傾向として、未決定とする場合が多いが、これはシステムの基本的な方針が、「不明確なものは無理にコード化せずに未決定とする」ことによるが、その他の理由として、シソーラスや辞書が未だ不完全であることや、JUMANにより切り出される語とシソーラスに登録された語に相違する場合があることが挙げられる。例えば、述語シソーラスでは1語である「穴あけ」が、JUMANでは「穴」と「あけ」の2語に切り出されるために述語コードが付かず、この時点で未決定となる。対策としては、JUMANのもつ形態素辞書を本システム用に改良する必要があり、現在作業中である。

5.2 産業コーディング

産業コーディングの自動化システムは職業コーディングの後に開発されたため、辞書の整備が遅れているにもかかわらず、今回、精度・再現率がともに高かった理由としては次の3つが考えられる。まず、職業が小分類まで行うのに対して産業は大分類でよいこと、次に、これと関連するが、回答に求められる情報が少なくてすむこと、さらに、質問の順番が職業より先にあるために情報が省略されにくいことである。

6. おわりに

本稿では、これまで実験段階にあった職業・産業コーディング自動化システムを、実際に「健康と階層」調査や「JGSS-2000」に適用した結果について述べた。システムは今後2003年まで毎年実施されるJGSSにおける職業・産業コーディングに適用される予定であり、当面の課題として次の3点がある。すなわち、システムの性能を高めるために、1)形態素解析において切り出される語がシステムの辞書やシソーラスの登録語と整合性を保つように、JUMANの形態素辞書を改良する。2)システムの辞書やシソーラスの充実をはかる。システムの操作性を向上させるために、3)システムにおけるすべての処理をWindows上で稼働させる。

[謝辞]

システムを実際の調査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と、ＳＳＭ職業分類の使用を快諾して下さった東北大学大学院文学研究科原純輔教授に感謝します。なお、「階層と健康」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します。

[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,『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, 1999,『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学研究科.

松本裕治, 1998,「意味と計算」,『言語の科学4　意味』, 岩波書店, 125-168.

西村幸満・石田浩, 2001,『JGSS-2000 調査（2000年11月）　職業・産業コーディングインストラクション』, 東京大学社会科学研究所.

高橋和子, 2000a,「格フレームによる職業コーディング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, 155-158, 於北陸先端技術大学院大学.

高橋和子, 2000b,「自由回答のコーディング支援について－格フレームによるＳＳＭ職業コーディングシステム－」,『理論と方法』, 15(1), 149-164.

高橋和子, 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」,『第73回日本社会学会大会報告要旨』, 28, 於広島国際学院大学.

高橋和子, 2001,「自由回答におけるコーディング自動化システムの適用－「健康と階層」調査における職業コーディング－」,『敬愛大学国際研究』, 8.

高橋和子, 2002,「JGSS-2000 における職業・産業コーディング自動化システムの適用」,『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』（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））.（予定）