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

高橋和子

（敬愛大学国際学部）

An applying the automatic occupational/industrial coding system to JGSS-2000

Kazuko TAKAHASHI

As occupational/industrial data including open-ended questionnaires need to be taken after-coding, they show some specific problems such as requirement of quantity of complicated work or inconsistency of the results. To solve them, an automatic occupational/industrial coding system based on the concept of “case frame” in Natural Language Processing had been constructed (Takahashi 2000b), and effectively applied to Health and Social Stratification survey (Takahashi 2001b). This paper reports the outcome of applying this system to JGSS-2000, which has five/four kinds of occupational/industrial data. In the results, the precision of the system ranges from 75.7% to 84.3% in occupational, from 89.4% to 93.0% in industrial coding, while the recall ranges from 61.1% to 68.8% in occupational, from 71.6% to 74.9% in industrial coding. These numbers are similar to the result from Health and Social Stratification survey. For the future improvement, it is necessary to extend the dictionaries and thesauruses of the system and the morphological dictionary of JUMAN (a software for morphological analysis), and to discussion the actual usage of the system.

Key words: JGSS, occupational coding, industrial coding, automatic coding, open-ended questionnaire, case frame

職業・産業データは自由回答を含むためにアフターコーディングが必要になるが、作業量の多さや煩雑さに加えてコーディング結果の非一貫性などさまざまな問題が存在する。これ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、自然言語処理における格フレームの概念に基づいたコーディングの自動化システムが開発され、「健康と階層」調査の職業データに適用された（高橋 2000b、2001b）。今回、この経験を踏まえて、システムを JGSS-2000 における「本人の現職」など５種類の職業データと４種類の産業データに適用したので、本稿で報告する。システムの精度と再現率は、職業において 75.7%～84.3%と 61.1%～68.9%、産業において 89.4%～93.0%と 71.6%～74.9%で、「健康と階層」調査とほぼ同様の結果を示した。今後、システムのもつ職業・産業辞書やシソーラスと形態素解析を行うソフト（JUMAN）のもつ形態素辞書の改善をはかることにより性能の向上が期待できるが、システムの使いやすさについての検討も必要である。

キーワード：JGSS、職業コーディング、産業コーディング、自動コーディング、自由回答、格フレーム

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・産業分類の個数と内容が人間に記憶できないほど多岐にわたるため[1)]、作業量の多さや煩雑さの問題だけでなく、コーディング結果の一貫性が保証されにくい問題がある。

高橋（2000b）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自然言語処理における格フレームによる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。ここで、格フレームとは、文の主要な意味は述語が担うとして文中の名詞が述語に対して果たす意味役割を深層格で表し、その表現形（表層格）や取り得る意味内容（選択制限）とともに述語を中心にフレーム形式で表現したものをいう[2)]。

システムを 1995 年 SSM 調査（約 1000 サンプル）や JGSS 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、高橋 2000c）。この後、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）に活用されたが、職業コーディング全体の中で、人間が行ったコーディングに対して別の観点からチェックを行うことができたと評価できる（高橋 2001a）。すなわち、ここでは、人間とコンピュータが別々にコーディングを行って結果を比較し、両者が一致したものは正解とみなし、一致しなかったものに対してのみ再度、職業コーディングの専門家数人が検討して正解を決定するという方法がとられた。システムを適用したことにより、チェックを含めた人間の作業時間が軽減化され、コーディングの一貫性が保証されたが、それ以外にも人間が犯しがちな見落としや勘違いによるミスがなくなる効果があった。反面、システムには人間がもつような常識がなく、表層的な意味解釈しか行えなかった点は否めない。

これらの経験を踏まえて、システムは今回、JGSS-2000 に適用されたが、調査の目的上、これまで以上に正確なコーディングが要求されるため、「健康と階層」調査と同様に適用された後、専門家による全体の検討作業が追加された。本稿では、これについて報告する。以下、次節と３節で JGSS-2000 におけるデータとコーディングの方法について述べた後、4 節、5 節でシステムを適用した結果と考察を行う。最後に今後の課題について述べる。

２．データ

システムが対象としたデータは、JGSS-2000 における職業・産業データで、本人現職、本人最後職[3)]、本人初職、配偶者職、父職の５種類である。本人現職から配偶者職までは、「従業上の地位＋役職」、「従業先事業の種類」、「仕事の内容」、「従業先の規模」のデータが収集されて職業・産業コーディング、父職は「従業先事業の種類」を除くデータが収集されて職業コーデ

ィングが行われた[4)]。データのうち、自由回答は、産業や職業を決定する際にそれぞれ中心となる「従業先事業の種類」と「仕事の内容」の２つで、他は選択回答である。

職業・産業コーディングの成否は、人間、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。

まず、産業を表す「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。これに対して、職業を表す「仕事の内容」では情報が不足する回答が目立った（表１）。これは、産業が大分類しか行わないのに対して職業は小分類までを行うためにより詳細な情報が必要になることと、質問の順番が職業の方が後にあるために回答が省略されやすくなったのではないかと思われる。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（注 12 参照）、そこでも必要な情報が得られない場合は決定することが不可能である[5)]。この場合は人間もコーディングが困難なことが多いため、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するかまたは回答欄を工夫するなどの対策を講じる必要がある[6)]。

表１　「仕事の内容」における情報不足の回答例（本人現職の場合）

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。

３．方法

3.1　コーディング自動化システムの位置付け

今回行われたコーディングの全過程を①から④に示す[7)]。①から③までは「健康と階層」調査とほぼ同様で[8)]、④が今回追加された作業である。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、１つの回答に対して 2 人 1 組とした。

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人間とシステムの結果を比較し[9)]、一致したものはそのままにし、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、必要ならばコーディングし直す。

3.2　コーディング自動化システムの概要

3.2.1　システムの処理の流れ

システムは本来、図１の（１）から（４）に示す独立した４段階から構成されるが、今回は人間も調査票のデータが入力された段階からコーディングを開始したために[10]、（２）形態素解析部、（３）自動コーディング部、（４）出力結果変換部の３段階（波線枠内）がシステムの中心的な処理となる。図１より明らかなように、システムは相異なるＯＳ上で稼働するために日本語コードが異なるが[11]、そのコード変換は容易である。

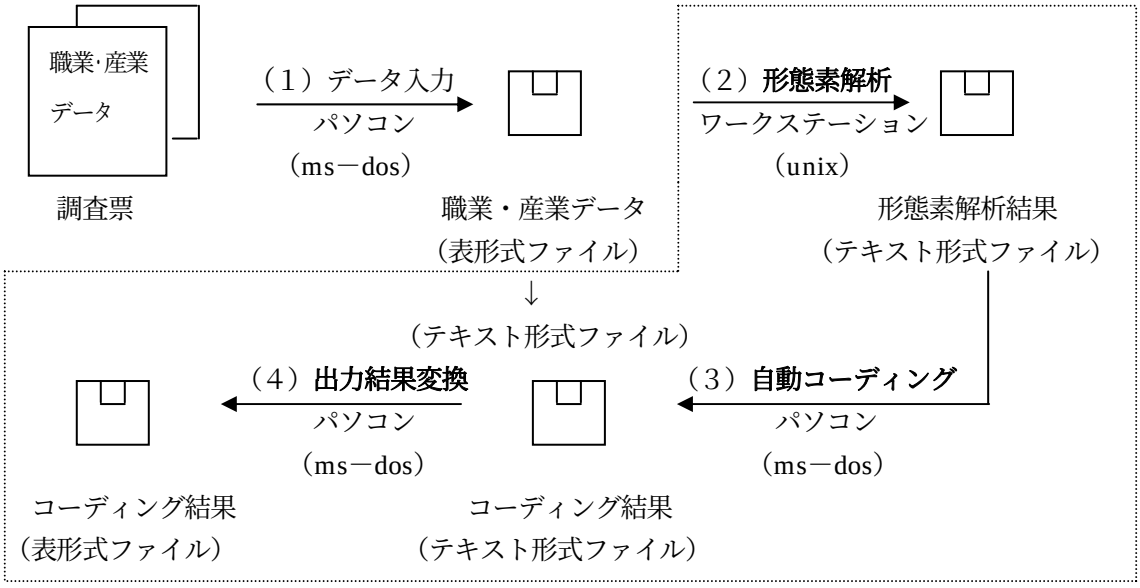

図１　コーディング自動化システムの処理の流れ（コンピュータ下の（　）内はＯＳの種類を示す）

3.2.2　システム各部の処理概要

紙面の都合上、（２）形態素解析部と（３）自動コーディングの処理概要について述べる。

3.2.2.1（２）形態素解析部

（３）の自動コーディング部における意味解析が語や品詞を単位とするために、（２）では、データを形態素(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(図２)。形態素解析は、形態素解析用ソフト JUMAN（黒橋・長尾 1999）を利用した。

```
表記　　読み　　原型 品詞 品詞コード 品詞細分類 （以下略）
 ↓　　↓　　↓　↓ ↓　　↓
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
```

図２　JUMAN Ver.3.1（ｅオプション指定）による形態素解析の結果例

3.2.2.2（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・産業データに対して該当する職業・産業コードを付け、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３は職業における自動コーディングの処理概要を示すが、産業についてもほぼ同様である。

まず、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」）。

次に、システムのもつシソーラスと辞書について述べる。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、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（図４）と名詞シソーラス（図５）を作成した。前者においては、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が付けられる。後者においては、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。

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４　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５　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の形式で記述したもので、それぞれ職業・産業辞書とよぶ。図６に示すように、これらの辞書においては、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が用いられる。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語である。述語によっては、複数の職業が対応するが、職業の違いにより必要な格にくる名詞が異なる。

述語コード	職業コード	必要な格	（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）　・　・　・
・　・　・			

図６　職業辞書（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）

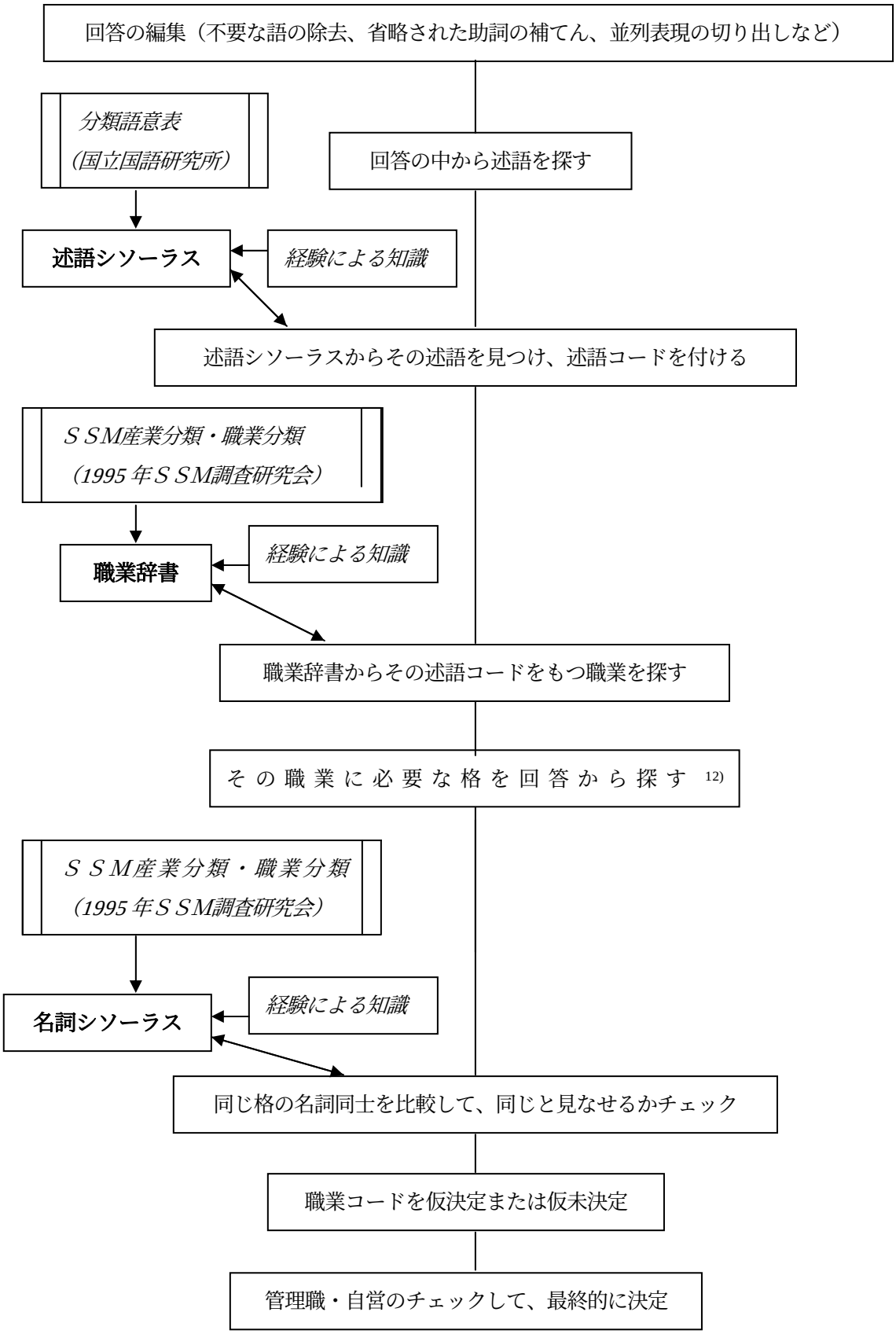

図３　自動コーディング部の処理概要（職業の場合）

最後に、職業の自動コーディングの場合は、「仕事の内容」に基づいて決定された職業に対して、必ず管理職[13]や自営[14]のチェックが行われた後にシステムとしての最終決定がなされる。すなわち、「従業上の地位」、「従業先事業の規模」、「役職」が各職業における一定の条件[15]を満たすかどうかを調べ、満たす場合にはそのまま最終決定とし、そうでない場合には「未決定」に変更する。例えば、「仕事の内容」が「会社の管理」と回答された場合、「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、最終的には管理職のチェックを行ってから該当するものに決定される。場合によっては、「仕事の内容」からは管理職以外の職業や「未決定」となっていても、このチェックにより管理職に決定せざるを得ないこともある。これら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LISP 言語により開発したが、約900 ステップ（42KB）の大きさとなった。

４．結果

4.1 精度と再現率

３人の協議による最終決定を「正解」としたときの職業・産業データに対する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、人間による結果との比較も示した。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で、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[16]

式から明らかなように、いわゆる正解率と呼ばれるもの（「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・産業コーディングの場合、再現率で示される。

表２　職業コーディングの結果（単位：%）

	本人現職		本人最後職		本人初職		配偶者職		父職[17]	
	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
人間	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：%）

	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

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、人間と

比較すると、すべての職業で精度が高く再現率が低い（表２）。また、産業コーディングにおけるシステムの精度は90.4%～93.0%で、再現率は71.6%～74.9%であった（表３）。

なお業コーディングにおいて、システムと人間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

システムと人間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。そこで、システムにおける傾向を分析するために、本人初職においてシステムだけが正解だったものと、システムだけが非正解だったもの（未決定を含む）を調べた（表４）。ここで、本人初職のデータを用いたのは、非該当を除いたサンプル数（何らかの職業にコーディングされたもの）が最も多かったためである（注16参照）。

表４　システムによる職業コーディングの傾向（本人初職の場合）

	システムだけが正解	システムだけが非正解
サンプル数（未決定も含む）	２３４（８．５%）	６４１（２３．２%）
サンプル数（未決定を除く）	１８１（６．５%）	２８５（１０．３%）
出現した職業の個数（全１５８個）	５４（３４．２%）	８９（５６．３%）
出現頻度の高い職業 （頻度５以上の職業コード）	503 554* 555 557 559* 569 573 607	523 554* 559* 560 629 630 633 634 642 645 648 649 653 659 682 686

*の付いた職業は、正解・非正解の両方に出現するもの

なお、システムと人間の両方が非正解だったものは183サンプル（6.6%）（システムが未決定のものを除くと86サンプル（3.1%））であった。このうち、システムと人間が同じ間違いをしているものは37サンプル（1.3%）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処理するが、その時間は形態素解析部、自動コーディング部、出力結果変換部に事前の処理（注10参照）を加えても人間より速かった。すなわち、今回、システムは延べ約14,500サンプル（＝約2,900×5種類）をすべて処理するのに、自動コーディング部[18)]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」のオーダーで、これ以外に自動コーディング部において生じたトラブル（注８参照）の解決時間を

含めても6日で処理を完了した。一方、人間は3.1の②と③で63日（＝9日×7人）かかっており、同じ全サンプルを処理するのに平均31.5日を要したことになる。従って、システムの適用により短縮された時間を単純に計算すると、25.5日（＝31.5日－6日）となる。

５．考察

5.1　職業コーディング

表２においてシステムが示した精度や再現率、また人間との一致率はいずれも、「健康と階層」調査と全く同様の数値であった[19)]。さらに、人間と比較して精度が高く再現率が低かった点も同様である。これより、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割弱で人間より劣るものの、正確さにおいては人間より優れていてコーディングした中の約8割は正解であるといえる。またこのとき、システムと人間は約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のではないかと考えられる。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、１回を人間に代わってシステムに行わせる方が、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。

ここで、システムのコーディング傾向を考察する前に、システムと人間の両方に共通して、本人に関する職業の方が本人以外（配偶者や父）のものより精度・再現率ともに高かった点に注目したい。当然、自分自身に関する事柄の方が自分以外のものよりも詳細に回答できるはずで、これは、両者における回答の質に大きな差があったためであると考えられるが、それがコーディング方法の違いを上回るほどであったものと解釈できる。従って、ここでも、「質のよい（過不足のない情報をもつ）回答を収集すること」がコーディングを成功させるカギとなることの確認ができた。

また、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には、辞書の知識やシソーラスの登録語を充実させることが重要である。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[20)]。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。

さて、表４よりシステムによる職業コーディングの傾向をみると、未決定のものを除くと、システムだけが正解なものは全体の約7%で、これは人間だけが正解である場合の約6割（＝181／285）である。そこに出現する職業の種類も人間だけが正解である場合の約6割（＝54／89）で、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

どである。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。その点、人間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は全体の約10%で、その多くは製造作業者(「629」～「659」)[21](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)。このような場合、精度の点からは未決定とした方がよいが[22]、少しでも回答の情報を活かすには追加されたコードを利用した方がよいわけで、分析時におけるこれらのコードの扱いを含めた議論が必要である。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。

最後に、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ことが挙げられる(表４)。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、その他に、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が相違する場合があることも原因である。例としては、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に、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。産業コーディングにおいては本人と本人以外の結果に差がみられなかったが、これは前述の２番目の理由により、職業の場合ほど情報の質に差が出

なかったものと考えられる。興味深いのは、システムは産業コーディングにおいても職業コーディングと同様に、本人の中で初職、最後職、現職の順に結果がよかったことで、これも過去の回答ほど辞書やシソーラスが活用されやすかったためであると考えられる。しかし、職業コーディングほどには差がなく、特に再現率がいずれも等しいと見なせる値であったのは、職業に比較すると、産業は新しい分類や新しい表現が出現しにくいためではないかと考えられる。

６．おわりに

本稿では、JGSS-2000 における職業・産業コーディングに、自動化システムを適用し支援した結果について述べた。今後の課題は次の３点である。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が Windows 上で稼働できるようにする（現在 unix 版である JUMAN を Windows 版に変更する）。

今後も、本システムが職業・産業コーディングの有効な支援を行うべく検討し、改良を重ねていきたい。

[謝辞]

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。

[注]

(1) 社会調査において、職業は約200種類、産業は約20種類のカテゴリーに分類されることが多い。

(2) 本来、格フレームは下図（述語「買う」を格フレームで表現した例（松本 1998））に示すように、述語が取る名詞の種類を限定することで構文解析における曖昧性を解消する目的を持つが、本稿では意味解析に応用した。

```
買う
    が：[Agentive]：human
    を：[Object]：physical_object
   （で）：[Locative]：shop
     ↑        ↑         ↑
  表層格   深層格    選択制限
```

なお、本稿では、意味解析の点から述語を広義に解釈し、サ変名詞（「製造」など）や職業名として認知されている名詞（「弁護士」など）も述語に含めた。

(3) 本人最後職とは、現在仕事をしていない人が最後に就いた職業をいう。

(4) 対応する質問は次の通りである。本人現職（問１(4)～(7)）、本人最後職（問５(5)～(8)）、本

人初職（問９(1)～(4)）、配偶者職（問１１(4)～ (7)）、父職（問２５(1)～(3)）。

(5) 例えば、回答が「製造」のみの場合に候補となる職業は、「陶磁器工」「石工」「ガラス・セメント製品製造作業者」「その他の窯業・土石製品製品作業者」など多数ある。

(6) 回答例の提示については、JGSS-2000 では、「小学校教員、農作業、バスの運転、自動車の修理、スーパーのレジ、銀行の経理の仕事、塾の講師、コンピュータのプログラマー」が例示されたが、JGSS-2001 ではさらに「営業事務、外回り営業」が追加されることになった。回答欄の工夫については、最低限、述語と対象格または場所格を収集するために、例えば、[　　]で[　　]を[　　]する　という形式が考えられる。

(7) 作業過程の詳細は西村・石田（2001）を参照のこと。

(8) 「健康と階層」調査では、③を専門家が行って④を省略した点と、図１（３）自動コーディング部を linux 上で稼働させた点が異なる。今回、自動コーディング部を ms－dos 上に移行したことによりいくつかの文字コード（「.」など）でトラブルが生じ、解決はしたものの対策時間を要した。

(9) 結果はいずれも表形式のファイルに保存されるために、両者の比較は容易である。

(10) 今後はこの方法がとられることになると思われるが、システムは、データが入力された表形式ファイルをテキスト形式ファイルに変換する作業から始まる。その際、形態素解析で失敗する可能性のある語を置換する事前編集を行う(例えば、「経営業」→「経営」など。詳細は高橋(2001b)を参照のこと)。

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

(12) 回答に必要な格が欠如している場合には「従業先事業の種類」を参照する。欠如するものが対象格であれば、述語が同じ場合に限ってそこでの対象格（もしあれば）を対象格とする。

(13) 特に、「545 管理的公務員」「548 会社役員」「549 その他の法人・団体の役員」「550 会社・団体等の管理職員」「553 その他の管理的職業従事者」のチェックを行う。

(14) 特に、「566 小売店主」「567 卸売店主」「568 飲食店主」のチェックを行う。

(15) 『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、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。

(16) 非該当のサンプルを除いたもので、本人現職（1935 サンプル）、本人最後職（801 サンプル）、

本人初職（2769サンプル)、配偶者職（1264サンプル)、父職（2615サンプル）であった。

(17) 父職における自営チェックの箇所で、単純なプログラムミスがあった。

(18) 処理時間はコンピュータの性能に依存するが、今回は約1.5秒／1サンプルを要した。

(19) 「健康と階層」調査におけるシステムの精度と再現率はそれぞれ81.7%と69.4%で、人間との一致率は60.0%であった。人間の精度と再現率はそれぞれ81.4%と80.0%であった。

(20) 人間においても同様の傾向があるが、システムに比較すると、想像力があり融通が利くために、システムほど明確な差としては現れなかったものと思われる。

(21) 「629 化学製品製造作業者」（以下、名称を省略。『ＳＳM産業分類・職業分類（９５年版)』参照)、「630」、「633」、「634」、「642」、「645」、「648」、「649」、「653」、「659」の10種類。

(22) 「704」と「703」を未決定にすると、システムだけ非正解のものは190個となり、システムだけ正解の個数とほぼ等しくなる。

[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，2000d,「自由回答の分析―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―」，『テキスト型データの取得から活用まで　資料集』，日本分類学会・日本行動計量学会共催シンポジウム，於統計数理研究所.

高橋和子，2001a,「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―」,『第32回数理社会学会大会研究報告要旨集』，38-41，於群馬大学.

高橋和子，2001b,「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.