JGSS-2001 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-2001

Kazuko TAKAHASHI

"The automatic occupational/industrial coding system" understands the occupational/industrial data that consists of mainly open-ended questionnaires, with the concept of case frame in Natural Language Processing, to give a proper occupational/industrial code for each data. Using this system together with human coding has three merits; shorter working, consistency in results, and checking human errors. The purpose of this study is to evaluate of this system in applying to JGSS-2001 data. The precision is from 75% to 80% in occupational coding and from 90% to 95% in industrial coding, while the recall is from 60% to 70% and from 70% to 75% respectively. These numbers are similar to the previous report of JGSS-2000 data. As morphological analysis is changed from a unix version to a windows version, each process can be performed on a personal computer.

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

職業・産業コーディング自動化システムは、自由回答を中心とした職業や産業に関するデータを自然言語処理における格フレームの概念により理解し、該当する職業や産業の分類コードを付けるシステムである。従来から行われてきた人間によるコーディングにシステムを併用することで、時間や労力が軽減されるだけでなく、コーディング結果における一貫性の保証や人間が犯しがちな単純ミスの防止ができるという利点もある。本稿の目的は、システムを JGSS-2001 に適用した結果について報告することである。システムの精度と再現率は前回（JGSS-2000）とほぼ同様で、職業は 75%～80%と 60%～70%、産業は 90%～95%と 70%～75%であった。また、今回、システムは形態素解析処理を unix 版から windows 版に移行したが、これにより、データの入力からコーディング結果の出力まですべての処理をパソコンでできるようになった。

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

１．はじめに

社会調査において自由回答で収集されたデータを統計的に処理する場合には、事前に個々のデータをコード化（コーディング）しておく必要がある。JGSS においても、分析の目的上、自由回答で収集されるデータが存在し、それらに対するコーディング作業が行われるが、その中で最も労力を要するものは職業や産業に関するデータのコーディング [1)]である。ここでは、１サンプル中に、職業だけでも、「本人の現在の職業（無職の場合は最後に就いた職業）」（本人現職または本人最後職）、「本人が最初に就いた職業」（本人初職）、「配偶者のある場合は配偶者の職業」（配偶者職）、「本人が 15 才のときの父親の職業」（父職）の計４種類が収集され、さらに、父職以外は従業先の事業内容すなわち産業のデータも収集されるため、煩雑で膨大な量の作業が発生する。

「職業・産業コーディング自動化システム」は、この作業を行う人間（コーダー）の支援を目的に開発されたプログラム群で（高橋 2000）、これまで、「健康と階層」調査や JGSS-2000 などに適用されてきた（高橋 2001、2002a、2002b）。後者の場合、システムの適用方法は次の通りである。まず、システムとコーダーが別々にコーディングを行う。次に、コーダーはシステムの結果を参考にしながら見直しを行い、必要ならば訂正を行う。さらに、これらすべてを専門家が見直し、判断に迷う場合には複数の専門家による協議を行って、最終的な決定を行う（詳細については西村・石田 2001 を参照のこと）。

コーディング作業にシステムが適用された結果、作業時間の短縮化と労力の軽減化をはかるという当初の目的が達成された。また、内容の面においても、システムはコーダーと異なり、判断に揺れがなく一貫性がある上に、結果に至った理由を明確に説明できる。コーダーが犯しがちな単純ミスや思いこみによる間違いを犯すことがないという利点をもつことから、コーダーはシステムが出した結果を参照することで、効果的なチェックを行うことができた。

しかし、システムはコーダーがもっている「常識」や柔軟さに欠けており、例えば、入力されたデータに誤字や文法的な誤りがある場合は意味を理解することができないし、データの内容が不十分な場合に、一部を除き [2)] 他の情報で補うことができないという欠点もある。実際、前回のデータによれば、システムの性能は、精度が職業で約 75～80%（コーダーは 70%～80%）、産業で約 90～95%、再現率が職業で約 60～70%（同 70%～80%）、産業で約 70～75%であり、コーダーと比較すると、精度がやや優れているものの再現率はかなり劣るという結果であった。

このように、システムの能力には限界があるものの適用については有効性が認められ、今回 JGSS-2001 のデータに対して再度、適用されることとなった。本稿の目的はその結果について報告することである。ここで、システムは、今回、形態素解析を行うためのコンピュータ環境は変えたが、システムの精度や再現率に大きく関係する自動コーディング部においては、辞書類の整備を行っただけで、プログラムの変更はほとんど行われていない。従って、今回の結果を前回と比較することにより、システムの安定性についてのチェックも行うことができる。すなわち、今回の結果が前回と同様またはそれ以上の値であればシステムの安定性が確認され、

大きく下まわるならば、システムはデータに対する依存度が高く、安定性に欠けるという問題を認識する必要があることになる。

以下では、まず、次節で今回のシステムの概要と前回からの変更点を述べる。３節でシステムにおけるコーディング結果をコーダーによる結果とともに報告し、４節で考察を行う。最後にまとめと今後の課題を述べる。

２．職業・産業コーディング自動化システムの概要

2.1　今回のシステムの概要

システムは、図１に示すように職業・産業データを一括して処理する。今回は通常のコーディングと異なり、コーダーもシステムと同様に表計算ソフトにより入力されたデータから作業を開始しため、システム独自の処理は、図１において波線で囲まれた部分、すなわち、次の（１）から（３）の３つである。システム構築当初、（１）（２）はともに unix（ワークステーション）上で処理されていたが、システムを改良するたびに linux や windows（パソコン）に移行した結果[4)]、現在では、入力から出力まですべてを windows 環境で処理できる。

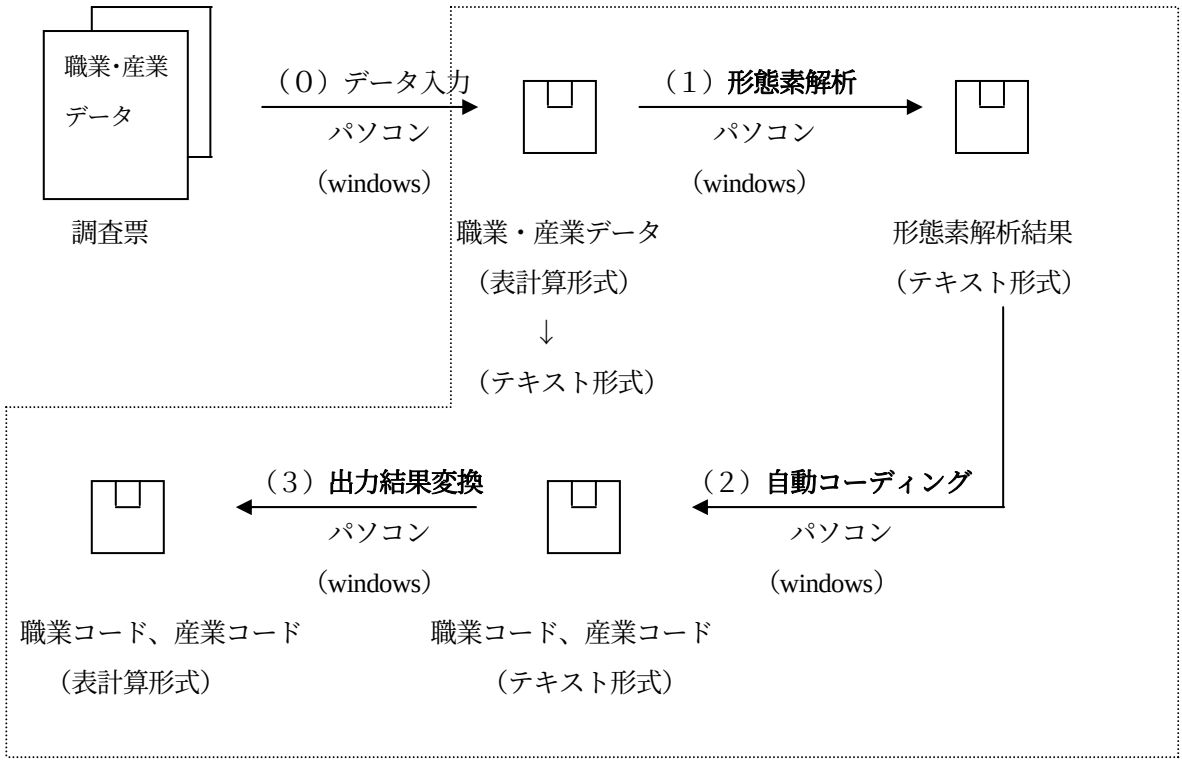

図１　コーディング自動化システムの処理概要

（１）形態素解析部

仕事の内容や従業先の事業内容を記述した自由回答に対する語の切り出しと品詞の認定を行う。京都大学長尾研究室で開発された形態素解析システム JUMAN（黒

橋・長尾 1999）を用いる。

（2）自動コーディング部

自然言語処理における格フレームの概念を用いた意味解釈[3)]により、職業コードと産業コードの決定を行う。職業の場合は、自由回答である「仕事の内容」から決定されたコードに対して、従業上の地位・役職や従業先事業の規模のデータも調べ、管理職、自営、建設関係のチェックを行って最終的に決定する。

（3）出力結果変換部

コーディング結果をみやすくするために、表計算形式により出力表示する。

図1には示されないが、（2）の自動コーディング部においては、職業や産業を決定する際に必要な知識を記述した「職業辞書」と「産業辞書」、さらに言語表現の多様さを吸収するために、格フレームにおいて重要になる述語と述語が格としてとる可能性のある名詞をそれぞれグループ化した「述語シソーラス」と「名詞シソーラス」を開発し[5)]、利用している。

2.2 前回システムとの相違点

前回からの主な変更点は、（1）の形態素解析を行う JUMAN の操作環境を、unix 版（ワークステーション）から windows 版（パソコン）に移行したことであり、システムの精度や再現率の値に深く関係する（2）の自動コーディング部においては大幅な変更を行っていない（図1参照）。JUMAN において windows 版は unix 版の限定版であるために、一般的にはこのような移行により機能の上で制限が生じるが[6)]、本システムにおいて問題になるものではなく、身近にワークステーションをもたない筆者のコンピュータ環境においては、むしろ処理作業を容易にすることができた。この結果、次に述べる2つの効果があった。

1つは、システムは図1に示すように形態素解析以外の処理をすべてパソコンで行うため、JUMAN の処理もパソコンでできれば、すべてが同一のパソコンで連続して行えるようになり、作業の大幅な効率化がはかられることである。これは、システムとしてまとまりや日本語コードの点[7)] からも望ましいものといえる。また、実際に、形態素解析は一度だけでなく何度も行うことが多いため（例えば、形態素解析終了後になって入力データにミスが発見され、やり直しが必要になる場合や、システムのメンテナンスのためにデータを変えて何度も実験を繰り返したい場合など）、これをパソコンで手軽に行うことができれば、時間や労力を大きく節約できる。パソコンを用いる場合にしばしば問題となる処理時間については、JGSS 程度の規模（約3000 サンプル）であれば1秒程度でワークステーションとの差はほとんどなく、問題とはならなかった。

もう1つは、本システムの場合には、JUMAN システムにある「形態素辞書」のメンテナンスも、パソコンの方が便利に行えることである。システムの性能を高めるには、自動コーディング部の改良だけでなく、その出発点となる語の切り出しと品詞の認定を行う形態素解析部とのインターフェイスをよくすることが重要である。それには、JUMAN の辞書、特に JUMAN

により切り出される語やその品詞などが記載された「形態素辞書」を本システムに適合するように改良することが有効であるが、その際、システムのもつ辞書やシソーラスと連動させる必要がある。従って、本システムの場合には、「形態素辞書」のメンテナンスは、ワークステーションではなくシステムが存在するパソコンで行う方がはるかに効率的である。

ここで、本システムにおける「形態素辞書」のメンテナンスの必要性についてより詳細に説明しよう。これは、次の３つの場合に必要となる。すなわち、1)JUMAN が形態素解析を失敗する場合　2)システムにおいて重要な語が複合語である場合　3)これと関連するが、JUAMN により切り出された語とシステムのシソーラスや辞書に登録された語が一致しない場合である。

まず、1)については、職業・産業の場合には、データの末尾が「業」であることが多いが、この場合にしばしば生じる失敗として、「業」が分離されず、直前の語と結びついて切り出されることである。例えば、「左官業」は「左」と「官業」、「不動産業」は「不動」と「産業」、「経営業」は「経」と「営業」のようになる。これを解決するには、「業」まで含めた語を新たに登録すればよく、例えば、「左官業」「不動産業」などの語を登録すると、その語が切り出される。

次に、2)については、JUMAN においては、カタカナ表記（ほとんどが未定義語なる品詞が付けられる）のものを除き、比較的短い単位で語を切り出す傾向があるために、複合語は基本となる語に分割されて切り出される場合が多い。例えば、「製造加工」は「加工」と「製造」と２語のサ変名詞として切り出されるが、これらは、現システムでは２つの語が並列であると見なされない限り、意味を理解することができない。これに対して、「製造加工」なる１語のサ変名詞として切り出されれば、解析に成功する（図２参照）。もちろん、この場合には、述語シソーラスに「製造加工」が妥当な述語コードとともに登録されていることや、述語がとる名詞と職業（または産業）コードが知識として職業（または産業）辞書または名詞シソーラスに登録されていることが前提になる。このような複合語の扱い方は名詞においても同様で、例えば、「場」のように抽象的な名詞が直接的に対象格や場所格となる場合は、システムは意味を理解することができないことが多いが、直前の語まで含めて１語とみなせば意味のある名詞となり、解析に成功する（図３参照）。実際、JUMAN においても、名詞の場合は分野により長い語（例えば、「特別養護老人ホーム」など）が登録されていることもあり、文法上からは正しくなくても、職業・産業に関するデータとして意味を理解するには１語として扱う方が都合のよいものはそのように扱っても差し支えないものと思われる。

最後に、3)については、特に述語の場合に重大であり、最初に述語シソーラスを検索する段階で述語コードが付かないために、それ以降の職業辞書や産業辞書を参照する処理に進むことができない。この問題を解決するには、JUMAN により切り出された語が述語シソーラスや名詞シソーラスに存在するか、または、逆に、シソーラスにある語を JUMAN が切り出すようにすればよく、両者の辞書やシソーラスをうまく連動させながらメンテナンスを行う必要がある。

しかし、以上のような方法により「形態素辞書」のメンテナンスを行っても、うまくいかない場合がある。例えば、「自営業」なる語は、「自」（普通名詞）と「営業」（サ変名詞）に分離

JUMAN 形態素辞書改良前

［JUMAN による解析結果］

○○	普通名詞
製造	サ変名詞
加工	サ変名詞

［自動コーディングプログラムによる解析］

述語	加工
対象格	なし

↓

職業コード　不明

JUMAN 形態素辞書改良後

［JUMAN による解析結果］

○○	普通名詞
製造加工	サ変名詞

［自動コーディングプログラムによる解析］

述語	製造加工
対象格	○○

↓

職業コード　該当するコード

図２　述語が複合語の場合

JUMAN 形態素辞書改良前

［JUMAN による解析結果］

ゴルフ	未定義語
場	普通名詞
で	格助詞
接客	サ変名詞

［自動コーディングプログラムによる解析］

述語	接客
場所格	場

↓

職業コード　不明

JUMAN 形態素辞書改良後

［JUMAN による解析結果］

ゴルフ場	普通名詞
で	格助詞
接客	サ変名詞

［自動コーディングプログラムによる解析］

述語	接客
場所格	ゴルフ場

↓

職業コード　該当するコード

図３　名詞が複合語の場合

されて切り出されるために、「自営業」（サ変名詞）なる語を新たに登録したが、結果は変わらなかった。この理由は、JUMAN においては、語の品詞情報だけではなく、品詞間の連接関係を接続コストとして定めて加算していき、文全体で最も低いコストになる語の切り出し方を解とするためである。この対策としては、現在は便宜的に、判明しているものに対して、形態素解析を行う前にまとめて置換処理を行っている（例えば「自営業」を「自営」と置換するなど）。

この方法のもう一つの問題点としては、現実には、すべての語を職業や産業として意味をもつように長い単位で登録し直すことが不可能なことである。従って、プログラムによる解決方法も検討する必要がある。例えば、現システムにおいては、例えば「部品」や「製品」のように職業や産業に多出する抽象度の高い名詞に対しては、自動コーディング部のプログラムでその直前の語もみているが、この方法を適用する語を増やすべく検討する必要がある[8)]。また、述語の場合もプログラムによる何らかの方法を検討する必要がある。

ところで、JUMAN においては「形態素辞書」は品詞の小分類ごとに用意されており[9)]、メンテナンスも品詞ごとに対応する辞書に対して行うのが自然である。しかし、システムの汎用性を考慮した結果、オリジナルの辞書はそのまま残し、新たに職業・産業コーディング専用のユーザー辞書（名称：occindadd.dic）として新たに１つ作成することとした。この理由は、一般に、対象とするドメインやシステムの目的により、細かい粒度で扱いたい語が異なるはずであるために、コーディング・システムごとに対応するユーザー辞書を作成しては取り換える方がよいと考えたためである。今回は、職業・産業コーディング用のユーザー辞書として実験的に数十語程度で作成したが、うまく機能した。

最後に、今回、JUMAN システム以外に行った変更点を述べておく。１つは自動コーディング部のプログラムに職業コードの最終チェックに建設関係を追加したことであり、もう１つは前回のデータより得られた知識をできる限り辞書やシソーラスの内容に反映させたことである。

３．結果

3.1　システムにおける正解個数と決定個数

専門家の協議による最終決定を「正解」としたときのシステムによるコーディング結果を表１、表２に示す。ここで、正解個数とはシステムの結果が正解と一致したものの個数、決定個数とはシステムにより意味のあるコード、すなわち不明（職業コード「999」）や非該当（職業コード「998」）以外のコードが付けられたものの個数をいう。また、該当サンプル数とは、全2790 サンプルから非該当と正解が不明であるものを除いたもの、すなわち、職業コードとして「998」と「999」以外のコードが付くべきものの個数である。

表１　職業コーディングの結果（全 2790 サンプル）

	本人現職	本人最後職	本人初職	配偶者職	父職
正解個数	1035	599	1755	818	1737
決定個数	1334	758	2126	1028	2070
該当サンプル数	1683	906	2565	1238	2452

表２　産業コーディングの結果（全 2790 サンプル）

	本人現職	本人最後職	本人初職	配偶者職
正解個数	1286	691	2009	944
決定個数	1392	733	2095	1008
該当サンプル数	1697	912	2577	1246

3.2　精度と再現率

職業コーディングと産業コーディングにおける精度と再現率をそれぞれ表３、表４に示す。

ここで、精度と再現率は情報検索における性能を示す指標で、それぞれ次式により計算される。

精度　＝　正解個数　／　決定個数

再現率＝　正解個数　／　コーディングされるべき個数（該当サンプル数）

比較のため、前回の結果（表中*を付けたもの）と今回との差も表中二重線の下段に示し、職業の場合はコーダーによる結果も示した。

表３　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	
	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	77.6	61.5	79.0	66.1	82.5	68.4	79.6	66.1	83.9	70.8
コーダー	82.5	82.1	87.3	85.8	85.7	85.1	81.7	79.8	86.6	84.5
両者の差	-4.9	-20.6	-7.7	-19.7	-3.2	-16.7	-2.1	-13.7	-2.7	-16.1
システム*	80.0	66.5	81.0	68.3	84.8	68.9	77.9	64.2	76.4	61.1
前回との差	-2.4	-5.0	-2.0	-2.2	-2.3	-0.5	1.7	1.9	7.5	9.7
コーダー*	78.7	78.1	73.1	72.1	81.2	79.0	70.7	68.8	75.7	70.7
前回との差	3.8	4.0	14.2	13.7	4.5	6.1	10.0	11.0	10.9	13.8

職業コーディングにおける精度は、システムが 77.6%～83.9%で平均 80.5%、コーダーは 81.7%～87.3%で平均 84.8%であった。再現率は、システムが 61.5%～70.8%で平均 66.6%、コーダーは 79.8%～85.8%で平均 83.5%であった。両者の平均の差は、精度が-4.3%、再現率が-16.9%である。また、システムにおいて前回との差は、精度が-2.4%～7.5%で平均 0.5%、再現率は-5.0%～9.7%で平均 0.8%であり、前回よりよいものと悪いものの両方があり、平均ではややよい程度でしかなかった。

産業コーディングにおいては、前回と同様に初回のコーダーの結果が残っていないため、システムの結果のみ示すが、精度は 92.4%～95.9%で平均 94.1%、再現率は 75.8%～78.0%で平均 76.4%である。前回との差は、精度が 0.7%～2.9%で平均 1.9%、再現率は 0.9%～4.2%で平均 2.5%であり、すべてよくなっていた。

表４　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	92.4	75.8	94.3	75.8	95.9	78.0	93.7	75.8
システム*	90.4	74.5	92.3	74.9	93.0	74.4	93.0	71.6
前回との差	2.0	1.3	2.0	0.9	2.9	3.6	0.7	4.2

3.3 処理時間

処理時間はコンピュータの性能に依存するが、筆者の場合、自動コーディング部については平均すると約2.0秒／1サンプルを要した。従って、１つのファイル（約3000サンプル）を処理するには、約 100 分（＝2.0 秒×3000）程度かかる。これに、自動コーディング部の前後の処理である形態素解析部（約1秒）と出力変換部（約1秒）を加算すると、約2時間で１つのファイルを処理できることになる。従って、JGSS のすべてのファイル（5つ）を処理するためには10時間程度が必要であるが、実際には、この他に、入力データのチェック[10)]や、前述した形態素解析前の置換処理、もともと１つのファイルである JGSS-2001 のデータを種類別に5つのファイルに分ける作業があるために、計3時間程度かかり、結局、全部で約13時間を要した。

JGSS のように、サンプル数が多く、ファイルの個数も多い場合には、処理時間の短縮化が要請される。このためには、自動コーディング部を高速化することが最も有効であるが、これは辞書・シソーラスに対するアクセスの高速化を実現すればよい。

４． 考察

4.1 職業コーディング

今回は前回と異なり、システムはコーダーよりすべて低い結果であった。これは、システムが精度・再現率ともに、前回とほぼ同程度でしかなかったのに対して、コーダーがいずれも大幅に上昇したためである。コーダーは前回より悪いものは全くなく、その差は最小でも 3.8%（本人現職の再現率）、最大では 14.2%（本人最後職の精度）で、平均すると精度、再現率がそれぞれ 7.3%と 12.6%上昇した。全 10 個あるうちの 7 個が 10%以上の上昇である。今回、コーダーの結果が非常によかった理由は、メンバー8 人のうち 7 人が前回の経験者であることから、前回の教育や学習が活かされ、能力が高まったためであると考えられる。システムにおいても、前回に得られた知識は辞書やシソーラスに反映されているが、人間における効果の方がはるかに大きかったといえよう。

一方で、システムが最終的な数値の上で前回とほとんど変わらなかった理由として、いわゆる管理職[11)]の判定に関する問題が大きい。JGSS が依拠する『SSM 調査　コード・ブック』（1995SSM 調査委員会 1995）によれば、管理職は、「仕事の内容」だけでなく、「従業上の地位」や「役職」、「従業先事業の規模」が一定の条件を満たしている必要がある[12)]（ただし、専門管理職は専門の方を優先する）。システムはこの原則に従い、「仕事の内容」から専門職（職業コード「501」～「544」）と決定されたもの以外のすべてのものに対して管理職チェックを行い、条件を満たすものはすべて該当する管理職に再コードした。これに対して、正解の方は、前回以上に「仕事の内容」に記述された情報を活かしたいとし、できる限り管理職にはしない方針をとった。このような見解の相違により、正解はシステムに比較して管理職の個数が極端に少なく、システムが管理職と判断したものの約6割がまちがいとなった[13)]。ここで、システ

ムが管理職としたもの以外について計算すると[14]、例えば、精度については、順に 80.3%、80.8%、83.1%、83.3%、87.6%で平均83.0%となり、表３より2.5%程結果がよい。単純な比較はできないが、前回と同じ条件の下では、システムの性能は数%程度よくなるものと思われる。

正解における管理職は「仕事の内容」の情報が重視されたとはいえ、「従業上の地位」や「従業上の規模」が考慮されたものもある。両者の違いを確認しても明確なルールはなく、明示化が困難なため、人間のもつ高度な判断によるものとするしかない。今後のシステムの方向性を考えると、システムの性能を高めるためにはむしろ管理職チェックを外す方がよく、その作業は容易であるが、人間に対する「よりよい支援」という観点からは、システムの性能は下がっても、これまで通り、『SSM 調査　コード・ブック』の原則によるチェックを行って提示する方が親切であると考えられる。なぜなら、人間が最終決定を行う際に、すでにチェックされた結果があれば参考にできて便利であるからである。ただし、チェックの方法について、専門職とコードされたもの以外にもチェックを除外する方がよいものがあるかを再検討する必要がある。

ところで、システムの結果を個別に見ると、精度、再現率のいずれにおいても、本人に関するものはすべて前回より悪く、本人以外のもの（配偶者職と父職）はよい。結果を良い順に並べると、精度、再現率のいずれも、前回は「本人初職→本人最後職→本人現職→配偶者職→父職」で、本人に関するものの方が本人以外より高かったのに対し、今回は「父職→本人初職→配偶者職→本人最後職→本人現職」（一部に同率あり）で、どちらかといえば逆の傾向を示す。従って、前回の説明である「本人以外については、本人ほど十分な情報が得られないであろうために、本人より結果が悪くなる」が、今回は当てはまらないことになる。しかし、これには次のような事情があった。すなわち、今回、システムが処理したデータのうち、本人に関するデータのみ、「従業上の地位＋役職」と「規模」に関する情報が100サンプル分誤っており（コーダーが処理する際には訂正された)、システムにおける管理職、自営関係、建設関係のチェックが誤った結果を出してしまった。ただし、本人の中でも、本人初職だけは管理職が少ないために、管理職チェックによる被害が少なかった[15]。

なお、データが誤りなく処理された配偶者職と父職においては、精度と再現率が前回より平均でそれぞれ4.7%と5.8%よくなっていた[16]。

以上より、今回のシステムによる結果は、表３に示された数値よりは高目であるとの解釈が許されるが、この点を考慮しても、前回と同様に再現率が低いのは明らかである。コーダーとの差を比較しても精度より再現率の方が大きく、前回は約10%、今回は約15%も低い結果であった。システムにおいて再現率が精度を大きく下回る理由は、正確さを重視するために、曖昧なものはすべて不明とするためである。不明の個数を減らすことができれば再現率はよくなるが、一般に両者はトレードオフの関係があるために、精度が悪くなる可能性が高い。現在の精度を保ちながら再現率を上げる工夫が必要である。

4.2　産業コーディング

表４より、産業コーディングはすべてにおいて前回より上昇している。これは、前回で得られた知識を産業辞書に追加した効果が現れたものと解釈できる。前回、今回ともに正常に処理された配偶者職における職業・産業コーディングの結果を比較すると、特に再現率の上昇の程度に違いが見られ、産業（4.2%）は職業（15.6%）の約４倍であった。これは、前回検討したように、産業は職業と異なり、新しい種類の産業が出ず、また、新しい表現や新しい語が用いられにくいという傾向があるためであると考えられる。

以上、職業・産業コーディングのいずれにおいても、平均すれば前回の結果とはほぼ同じかそれ以上であり、システムの安定性を確認できたとしてよい。

５．おわりに

本稿では、JGSS-2001 に職業・産業コーディングの自動化システムを適用した結果について、コーダーによる結果や前回 JGSS-2000 の結果との比較を行いながら報告した。その結果、システムの性能は、職業では精度が約 80%、再現率が約 70%程度、産業では精度が約 95%、再現率が約 80%であった。この値は前回との差がほとんどないことから、システムはデータに依存することなく安定していると評価してよい。しかし、コーダーと比較すると、特に職業において再現率の低さが目立っており、これを解決する必要がある。

今回は前回で課題とした３つの点、すなわち、1)形態素解析において切り出される語が辞書やシソーラスと同じ語であるように、JUMAN の形態素辞書を職業・産業用に特化させること 2)形態素解析が手軽にできるように、現在は unix 上で稼働する JUMAN を windows 上に移植すること　3)自動コーディングの性能を高めるために、システムの辞書・シソーラスをより充実させること　について、2)は解決し、1)と 3)の作業を開始できた。当面の課題は、この作業を進めることであるが、その他に、1)職業コーディングにおいて、精度を保ちながら再現率をあげる　2)自動コーディング部における処理の高速化を図る　の２点を課題とし、システムの精緻化を行いたい。

[謝辞]

ＳＳM職業分類の使用に当たり、東北大学文学部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) JGSS においては、職業はＳＳM調査（社会階層と社会移動全国調査）とほぼ同様に約 200 種類、産業はこれより細かく約 20 種類のカテゴリーに分類される。

(2) 「仕事の内容」に必要な情報が不足している場合、それが格フレーム（注の(3)参照）における

対象格に該当する名詞であれば、「従業先事業の種類」を参照し、その述語が「仕事の内容」における述語と同じ場合に限って、その対象格（もしあれば）の名詞を充てる。

(3) 格フレームによる意味理解によると、文の主要な意味は述語が担い、文中の名詞は表層的には助詞により示される述語との関係によりその役割が決まるとする。

(4) システム開発当初（高橋 2000）は図１における（３）がなく、（１）（２）とも unix 環境であったため、システムの使い勝手はよいとはいえなかった。「健康と階層」調査（高橋 2001）の際に、（０）をワープロ用ソフトから表計算ソフトに変更すると同時に（３）を開発し、（２）を linux に変更した。これによりシステムとしての形が整い、さらに、前回 JGSS-2000（高橋 2002b）の際に（２）を windows に変更したことで、使いやすさが向上した。

(5) 当初、述語シソーラスは『分類語彙表』、名詞シソーラスは『ＳＳM産業分類・職業分類（９５年版)』に基づいて作成したが、システムの改良を重ねるうちにかなり異なったものとなっている。

(6) JUMAN システムが持つことのできる辞書の個数や場所（ディレクトリ）に制約がある。

(7) 日本語コード体系は、一般的に windows と unix で異なっており、それぞれシフト JIS、EUC コードが用いられる。従って、使用するＯＳを変えるたびにコード変換が必要であった。

(8) ただし、JUMAN の「形態素辞書」に登録されている語の長さには揺れがあり、語の選択には検討が必要である。

(9) 例えば、名詞の場合、Noun..dic だけでなく、Noun.hukusi.dic、Noun.keishiki.dic、Noun.koyuu.dic、Noun.suusi.dic、Noun.sahen.dic 、Noun.time.dic がある。

(10) 半角文字（(　) が半角になっている場合が多い）を全角に置換したり、助詞がカタカナ表記の場合にひらがなにするなどの事前処理を行っている。

(11) ここで管理職とは、職業コード「545 管理的公務員」「546 国会議員」「547 地方議員」「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」「551 駅長、区長」「552 郵便局長、電報・電話局長」「553 その他の管理的職業従事者」の 9 つの職業をいう。

(12) 『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。

(13) 本人現職、本人最後職、本人初職、配偶者職、父職においてシステムが管理職と判断した個数

と正解が管理職とした個数は、それぞれ 72 個（38 個)、41 個（32 個)、16 個（1 個)、90 個（48 個)、176 個（102 個)（(　) 内の数字は正解が管理職ととした個数）であり、このうち、システムが正解と一致した個数は、順に 22 個、20 個、1 個、37 個、78 個であった。

(14) システムが管理職としたものを正解個数と該当サンプル数から除いて精度を計算し直した。

(15) 本人現職、本人最後職、本人初職においてシステムが管理職と判断したものが全体に占める割合は、それぞれ 5.4%（＝72/1334)、2.6%（＝20/758)、0.8%（＝16/2126）で、本人初職だけ非常に低いことが明らかである（(　) 内の分母はシステムの決定個数を示す)。

(16) ただし、父職は前回、プログラムに一部ミスがあったため、精度、再現率ともに数値が低めに出た点を考慮する必要がある。

[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，2000,「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，2001,「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―」,『敬愛大学国際研究』, 8.31-52.

高橋和子, 2002a,「職業・産業コーディング自動化システムの活用」,言語処理学会第 8 回年次大会発表論文集，491-494.

高橋和子，2002b,「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)，171-183.