職業コーディング支援システム（NANACO）の開発と JGSS-2003 における適用

高 橋 和 子　　**須 山 敦**　　**村 山 紀 文**
（敬愛大学国際学部）　（東京工業大学大学院総合理工学研究科）
高 村 大 也　　**奥 村 学**
（東京工業大学精密工学研究所）

Applying the occupation coding supporting system for coders (NANACO) in JGSS-2003

Kazuko TAKAHASHI
Atsushi SUYAMA　Norifumi MURAYAMA　Hiroya TAKAMURA　Manabu OKUMURA

We have developed the new system called NANACO, which directly supports coders' works in the Occupation Coding. Conducting the occupation coding manually is a time-consuming and complicated task and sometimes leads to inconsistent coding results when coders are not experts. For this reason, a rule-based automatic method called ROCCO has been developed and used in JGSS. Moreover a machine learning method, which is Support Vector Machines (SVMs), has also been applied to JGSS. However, for coders, the task is still hard. We apply NANACO system to the occupation coding in JGSS-2003 and show that the system contributes to increase an accuracy of coders and to reduce working hours. And we also show that most coders think the system is good.

Key words: JGSS,Occupation Coding, Rule-based method, Support Vector Machines

JGSS においては､ルールベース手法により職業コーディングを自動的に行うシステム（Rule-based Occupation Coding；ROCCO システム）が利用されてきた。さらに、最近では、機械学習の一つであるサポートベクターマシンによる自動コーディングの適用も行われている。これらの自動コーディングシステムは性能が安定しており、コーダの作業を減らす効果がある。しかし、コーダの作業負担は依然として軽いものではなく、作業そのものを支援する必要があると考えられる。そこで、コーディング作業を直接支援するシステム（NANACO）を開発し、JGSS-2003 に適用した。その結果、コーダの正解率が高まり、作業時間が大幅に減る効果があった。また、コーダによるシステムの評価も好評であった。

キーワード：JGSS、職業コーディング、ルールベース手法、サポートベクターマシン

１．はじめに

職業コーディングとは、社会調査において職業に関するデータ（職業データ）を該当する職業コードに分類する作業をいう。これは職業データが自由回答を中心に収集され、このままでは統計処理を行うことができないために、データ分析の前に必ず行われる作業である。しかし、職業コーディングでは、複数の質問から構成される職業データを総合的に判断する必要があることや、職業コードのカテゴリとして約 200 個もの小分類コードを用いるため、作業は煩雑で特に初心者のコーダにとっては困難なものとなっている。また、サンプル数が多い場合には作業が長期にわたるため、熟練したコーダでもコーディングの結果に揺れが生じ、一貫性が欠如する傾向がある。

これらの問題を解決する目的で、コンピュータによる職業コーディング自動化システム（Rule-based OCcupation Coding；ROCCO）が開発された（高橋 2000）。JGSS では、第１回本調査（JGSS-2000）から ROCCO システムの利用を行っており、その結果については、高橋（2002a、2003、2004）において報告されている。そこでは、ROCCO システムの性能はいずれの回でも安定しており、付与したコードの正確さ（精度）は、熟練したコーダには劣るものの初心者コーダよりは優れた結果を示している。

職業コーディングでは、コーディング結果の妥当性や正確さをできるだけ高くするために、通常、同一のデータに対してコーダを変えて複数回のコーディングが行われる。JGSS では、ROCCO システムはコーダの作業１回分の役割を果たすことができると考え、一般コーダが行うべき作業のうちの１回分を ROCCO システムに行わせてきたが、その利用方法には多少の変化が見られる。例えば、JGSS-2000 においては、コーダは ROCCO システムと全く独立にコーディングを行ったが（高橋 2002a）、JGSS-2002 においては、ROCCO システムによる結果を参考にしながら行った（高橋 2004）。すなわち、JGSS-2000 の場合は、ROCCO システムはコーダの１回分を代行したと考えられるが、JGSS-2002 の場合は、ROCCO システムはコーダのコーディング作業の支援を行ったと考えられる。いずれにしても、ROCCO システムはコーダの行うべき作業の絶対量を減らしたり、コーダ自身が作業する際に参考となる情報を提示したりすることで、コーダの負担を軽減する効果があったといえる。

しかし、職業コーディングの性質上、コーダの作業がなくなる可能性はあり得ず、また現状においてはその負担が依然として軽くないことも事実である。そこで、今度はコーダのコーディング作業そのものに注目し、これを積極的に支援することを目的とするシステムの必要性を考え、開発を行った（システムの名称は NANACO である（高橋他 2004c）。NANACO システムは、コーダを「直接」支援するという意味で、JGSS-2002 における ROCCO システムの用途をより進めたものであるといえる。

NANACO システムの目標は、コーダが短時間で容易に正確なコーディングを行うことができるような支援を行うことである。従って、主要な機能として、コーダが正しいコードを思い付くためのヒントとして、ROCCO システムを始めとする自動コーディングシステムに

よる結果をわかりやすく表示すること、必要に応じて職業コードの内容を確認することができるように、職業の定義文を随時閲覧できること、同じサンプルにおける種々の職業（例えば、本人現職、本人初職、配偶者職など）が参照しやすいように、これらをサンプルの属性とともにまとめて表示することなどが考えられる。

NANACO システムは、今回 JGSS-2003 において初めて適用された。そこでは、コーダの正解率が平均で 90%を超えるという非常によい結果が得られ、作業時間もこれまでと比較すると、大幅に短縮された。また、コーディング終了後にコーダによるシステムの評価を行ってもらったが、その結果も概ね良好であった。ただし、問題点や希望する機能も指摘されており、改善の余地がある。

ところで、ルールベース手法である ROCCO システムは、ルールの作成やルールセットの継続的なメンテナンスの手間などルールベース手法に特有の欠点をもつため、JGSS-2002 からは、新たな自動コーディングの方法として、機械学習の一つであるサポートベクターマシン（SVM）による方法が検討された（高橋他 2004b、高橋 2004）。その結果、SVM を単独に適用するのではなく、新たに ROCCO システムと組み合わせる手法（4 種類）が提案され、実験の結果、両者を組み合わせた手法はいずれも SVM 単独の場合よりも高い正解率を示すこと、その中でも特に ROCCO システムによる結果を SVM の素性とする手法が最も高い正解率を示すことが明らかになった（高橋他 2004b）。従って、これは JGSS-2002 に適用した手法とは異なるが、JGSS-2003 ではこの手法を適用した。

本稿の目的は次の２つである。すなわち、NANACO システムの概要とシステムを適用した結果について報告すること、および JGSS-2003 における自動コーディングシステムの結果についても、これまでと同様に報告することである。

以下では、次節で、これまでに引き続き基本的な報告として、JGSS-2003 における職業コーディングに ROCCO システムを適用した結果を報告する。また、これと SVM と組み合わせた手法を適用した結果についても報告する。これらの結果は、今回、次節で述べる NANACO システムにおいて、コーダに対する参考情報として提示された。3 節では、NANACO システムの概要について述べる。4 節で NANACO システムを JGSS-2003 に適用した結果をコーダの正解率と作業時間の点から報告し、さらにコーダの評価についても簡単に報告する。最後にまとめと今後の課題を述べる。

２．JGSS-2003 における自動コーディングの適用結果

JGSS-2003 においては 3,663 サンプルが収集され、コーディングの対象は、本人現職、配偶者職、父職の 3 種類であった。職業の種類はこれまで収集された 5 種類[(1)]と比べ少ないが、のべサンプル数では約 1,000 サンプル程度多い。

2.1 ROCCOを適用した場合

ROCCOシステムによる結果を表1に示す。ただし、今回もJGSS-2002と同様に、SVMによる方法との比較を考慮し、複数個出力した場合は最初の1個（プログラムの都合上、回答においては最後に記述されたもの）のみを見る。ここで、精度と再現率は、それぞれ、「どのくらい正確にコードを付与できるか」、「全体の中でどのくらい正確にコードを付与できるか（正解率）」の指標となるもので、次式により計算される。

精度　＝　正解であったサンプル数／未決定以外のコードを付与したサンプル数

再現率　＝　正解であったサンプル数／総サンプル

結果は、3種類の職業の平均で、精度が79.8%、同再現率が64.1%であった。この中で精度・再現率ともに最もよいのは父職で、本人現職と配偶者職はほぼ同様である。

表1　ROCCOシステムによる職業コーディングの精度と再現率（単位：%）

本人現職		配偶者職		父職	
精度	再現率	精度	再現率	精度	再現率
77.8	61.5	77.0	61.2	84.7	69.5

今回の結果を前回（JGSS-2002）と比較すると、精度についてはやや高くなっているか（本人現職で約1%）ほぼ等しいが、再現率については約1.8～5.1%ほど低下した。

JGSS-2003における精度と再現率の平均を過去3年間の結果とともに表2に示す。４年間の平均を単純に計算すると、精度は79.9%で再現率は66.1%である。精度・再現率ともに変動が小さいことから、ROCCOの性能が安定していることがわかる。

表2　ROCCOシステムによる職業コーディングの精度と再現率の平均（単位：%）

JGSS-2000		JGSS-2001		JGSS-2002		JGSS-2003	
精度	再現率	精度	再現率	精度	再現率	精度	再現率
80.0	65.8	80.5	66.6	79.4	67.7	79.8	64.1

2.2 ROCCOとSVMと組み合わせた手法を適用した場合

高橋他（2004b）では、ROCCOとSVMと組み合わせた手法として、次の4種類を提案した。ここで、基本素性とは、「仕事の内容」に出現する単語(2)（原形と品詞)、「従業先事業の種類」に出現する単語（原形と品詞)、「授業上の地位」(14 個の選択肢）をいう。

・「ROCCOにより出力された職業コード」をSVMの素性として新たに基本素性に追加してSVMを適用する手法（以下、SVM(add-code)と呼ぶ）

・「ROCCO により用いられたルール」を SVM の素性として新たに基本素性に追加して SVM を適用する手法（以下、SVM(add-rule)と呼ぶ）
・「ROCCO により出力された職業コード」と「ROCCO により用いられたルール」の両方を SVM の素性として新たに基本素性に追加して SVM を適用する手法（以下、SVM(add-code-rule)と呼ぶ）
・ROCCO が職業コードを１つに決定できない場合に SVM による結果を用いる手法（以下、SVM(seq)と呼ぶ）

JGSS-2002 に適用した手法は SVM(seq)であるが（高橋 2004）、今回 JGSS-2003 に適用した手法は、JGSS-2002 データによる実験の結果、正解率が最も高かった SVM(add-code)である。

表 3 に JGSS-2003 に SVM(add-code)を適用した結果の正解率を示す。ここで、正解率は ROCCO システムにおける再現率を意味する。表 3 より、SVM(add-code)はいずれの職業においても ROCCO より約 17%程度高い値を示す。JGSS-2003 における正解率の平均は 80.7%で、JGSS-2002 データによる実験時より約 6%程度高い。また、厳密には対象とするデータが異なるが、JGSS-2002 における SVM(seq)と比較しても約 8%程度高い(3)。

表 3　ROCCO と SVM を組み合わせた手法による職業コーディングの正解率（単位：%）

本人現職	配偶者職	父職
78.9	77.4	85.8

今回、SVM(add-code)の正解率が前回より高かった理由としては、訓練データ数の増加が考えられる。JGSS-2002 においては、訓練データとして JGSS-2000 データと JGSS-2001 データが用いられたが、今回はこれに JGSS-2002 データが追加されたため、訓練データ数は前回の約 1.5 倍となった(4)。訓練データ数と正解率の関係は、高橋他（2004b）における実験結果より明らかなように、訓練データ数が大きいほど正解率が高いため、前回よりもよい結果が得られたのではないかと思われる。

３．NANACO システムの概要

NANACO システムの目的は、コーダの作業を軽減し、コーダができる限り一貫性のあるコーディングを行うことができるように支援することである。従って、NANACO システムを利用することにより、コーダは短時間で正確なコーディングを行うことが可能になるはずである。さらに、忍耐が必要なコーディング作業が少しでも快適にできるように、ユーザインタフェースについても考慮した。以下では、職業コーディングにおける NANACO システムの位置づけと主要な機能について述べる（操作方法については須山(2004)を参照のこと）。

3.1 職業コーディングにおける NANACO システムの位置づけ

職業コーディングにおける NANACO システムの位置づけを図１に示す。これにより明らかなように、NANACO システムの役割は、職業データや、ROCCO システムなどの自動コーディングシステムによる出力結果、職業分類の定義内容をなどの情報を取り込んでコーダに提示したり、コーダが決定したコードを受け取ってファイルに保存したりすることである。すなわち、職業コーディングに関係する情報の「入出力」処理が中心で、システム自身が（狭義の）コーディングを行うわけではないことに注意された。言い換えれば、NANACO システムで扱われる情報の「中身」は、NANACO システムとは独立した別の（人間を含めた）システムやファイルにより作成される。現在は、NANACO システムは、自動コーディングを行うものとして ROCCO システムや SVM(add-code)を利用し、職業分類定義としては『SSM 職業分類・産業分類（９５年版)』(1995 年 SSM 調査研究会 1995）ファイルを用いているが、これらは NANACO システムとの間受け渡し形式が同じであれば、別のものに取り替えが可能である。

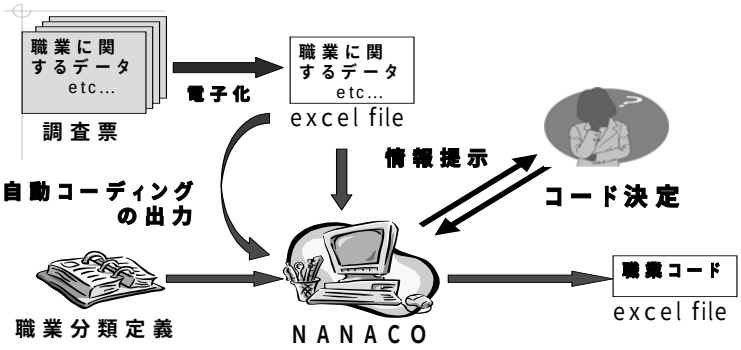

図１　職業コーディングにおける NANACO システムの位置づけ

NANACO システムは、コーディング作業時における処理の高速化をはかるために、現状では時間を要する可能性のある処理については前処理としてコーディング作業時に行う処理と切り分けている。例えば、コーディング作業時に提示する情報はすべて、前処理の段階で入力され整理される。

結局、NANACO システム全体における入力情報は、調査票に記述されたデータ、自動コーディングによる結果、『SSM 職業分類・産業分類（９５年版)』の記述内容の３つで、最終的な出力情報は、コーダの決定した職業コードである。この中でデータ分析に直接関係するファイル（既存の『SSM 職業分類・産業分類』ファイル以外）はいずれも EXCEL ファイルの形式で扱えるようにすることで、他のシステムとの連携を容易にしている。

3.2 NANACO システムにおける主な機能

NANACO システムの主な機能は次の 4 つである。このうち、（1）（2）が前処理として行われ、（3）（4）がコーディング作業の場で行われる。

（1）ファイルの読み込み

（2）情報提示に必要な処理

（3）情報提示

（4）結果ファイルに出力

以下、順に説明する。

（1）ファイルの読み込み

前処理として NANACO が読み込むファイルは、「データファイル」、「職業分類定義ファイル」、「ROCCO システムおよび SVM（add-code）による出力ファイル」の 3 つである。いずれも他のシステムとの連携を考慮し、CSV 形式にした。

「データファイル」は図 2 に示す項目から構成される。「職業分類定義ファイル」は『SSM 産業分類・職業分類（９５年版)』における記述内容で、職業コードが定義されている。

・属性（性別、年齢、本人学歴、配偶者学歴、父学歴など）
・職業データ１（例えば、本人現職の「仕事の内容」、「従業上の地位」、
「従業先事業の種類」、「従業先の規模」、「役職」など）
・職業データ２（例えば、配偶者職の「仕事の内容」、「従業上の地位」、
「従業先事業の種類」、「従業先の規模」、「役職」など）
（職業データの種類だけ続く）

図 2　「データファイル」の項目（1 サンプル分）

「ROCCO システムおよび SVM（add-code）による出力ファイル」は、職業データを自動コーディングした結果で、内容は、ROCCO により決定された職業コードおよび機械学習である SVM（add-code）により決定された職業コード（現在は第 1 位から第 5 位まで）と各職業コードに決定されたときの分離平面からの距離である。分離平面からの距離の値は、次の（2）における確信度の計算に用いられる。

（2）情報提示に必要な処理

ここでの主な処理は次の 2 つである。1 つは、（3）において、類似性の高いデータを提示することができるように、データ間の類似度の計算を行っておくことであり、もう 1 つは、（1）における SVM（add-code）の結果（分離平面からの距離情報）から確信度を計算することである。

まず、類似度の計算は、ベクトル空間法により Cosine 類似度 (5) を用いて行った。その結果、値が閾値以上のペアを「似ている」とする。今回は、職業データの中の「仕事の内容」、「従業先事業の種類」、「従業上の地位」によりベクトルを作成し、閾値を 50 に設定し

たが、用いる素性や閾値は自由に変更することができる。

次に、確信度の計算は、SVM においては、分離平面からの距離が大きいほど正解率が高いことを利用し、距離の値が大きいほどそのカテゴリであるとする確信度が高いと考えることができる。これにより、分離平面からの距離の値が大きいほど確信度が高くなるように計算した。ただし、100%や 0%という数値を表示するのを避けるため、計算結果が 1 になった場合は 99%とし、0 になった場合は 1%に変更するなど多少の調整を行っている。

（3）情報提示

NANACO システムが提示する主要な情報は次の 5 種類である。

・コーディングの対象とするサンプルのデータ内容（図 3 参照）

・職業コードの候補（図 3 参照）

・類似データのリスト（図 3 参照）

・職業分類の定義内容（図 4、図 5 参照）

・データの検索結果（図 6 参照）

図 3 に作業画面の例を示す。上から順に、本人現職（最後職）、配偶者職、父職の職業データが表示される。今回は職業の種類が少ないために 1 画面ですべてを表示できたが、多い場合でも画面をスクロールさせて見ることができる。このように、同一サンプルのデータを一覧の形で見ることができるため、ある職業データに不足する情報があっても、他の職業データにより補完することが容易である。例えば、家族内で同じ職業に就いている場合や、（今回は該当しないが）職歴を収集されている場合には有効であると思われる。画面下には、属性が表示される。

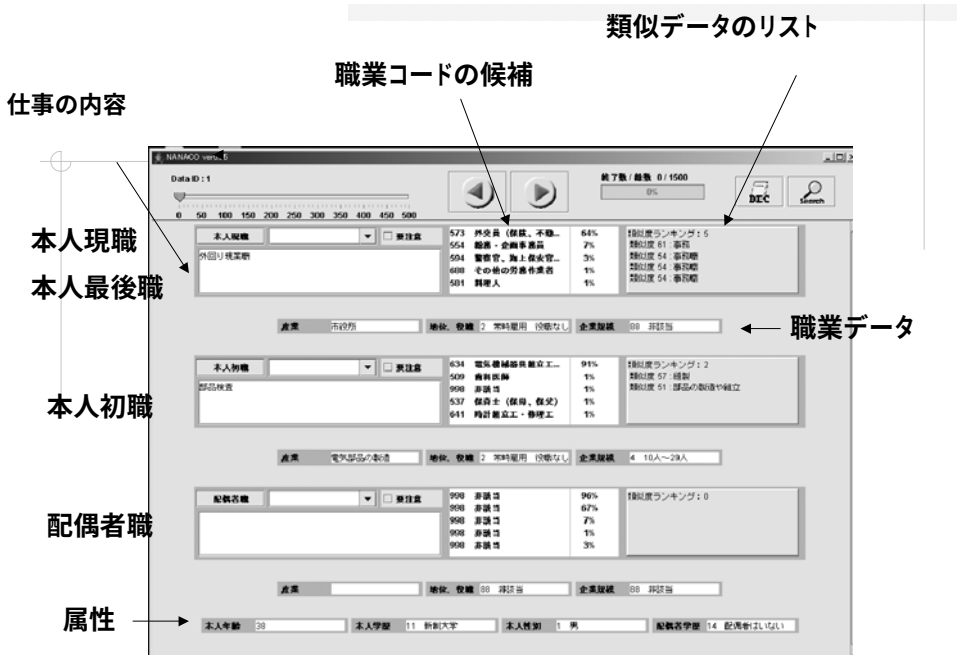

図 3　作業画面例

職業コードの候補には、SVM（add-code）により決定された第1～第5位までの職業コードと職業分類が表示され、その中でROCCOシステムの結果と一致するものがあれば先頭に○印が付与される。コーダはこれを参考にしながら、妥当な職業コードを考えることができる。

類似データのリストは、現在表示されているデータと「似ている」と判断できるデータが表示される。類似データのリストには、コーダが現時点でコーディングを行っているデータと（閾値以上の値で）類似したデータが表示される。このとき、リストされたデータについては、さらに詳しく個別情報を見ることが可能である（紙面の都合上、画面例は省略した）。ただし、現状では、類似データのリストの中にそれほど似ていないものが出現する場合も多い。この理由としては、作業効率の点から、類似データのリストに対して一括して職業コードを付与できるようにしたため、別のコーダが担当するデータは表示されないようになっている。このため、類似データは1人のコーダが担当する範囲内でしか見つけられないため、効果的なリストになりにくいと思われる。しかし、類似度計算の方法や素性の選び方についても再検討する必要があり、今後の検討課題である。

図4、図5はいずれも職業分類の定義内容の提示例であるが、図4は職業コードまたは用語をキーワードとしてマッチする全文を表示させた場合であり、図5は作業画面上に表示されている職業コードや職業分類名をクリックし、別ウィンドウとして表示させた場合である。このように、職業分類の定義内容は、コーダが作業中に随時閲覧できるよう工夫を行った。

554 の定義を表示

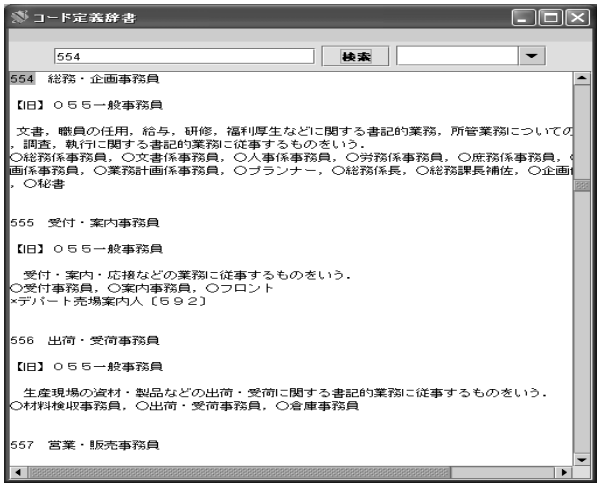

図4　職業分類の定義内容の表示例（職業コードや用語による全文検索）

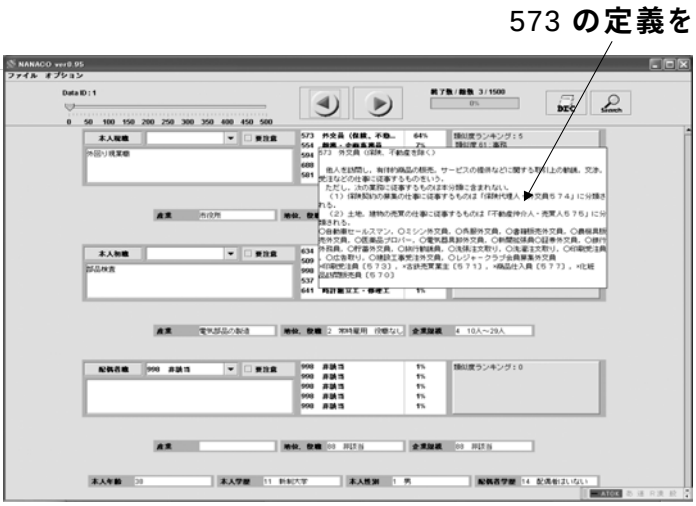

図 5　職業定義分類の表示例（作業画面上における別ウィンドウによる表示）

図 6 はデータの検索結果例である。職業コードや用語または両方をキーワードとして、データを検索できる。ここで表示されたデータに対してまとめてコードを付与することができるために、効率よくまた一貫性のあるコーディングが行うことができる。

573 または営業で検索

図 6　データの検索結果

その他として、作業画面においては上部中央にあるボタンがあるが、左が直前のサンプル、右が次のサンプルに進む（戻る）ためのものである。もし離れたサンプルにジャンプ

したい場合には、画面左上の移動式のボタンを使用すればよい。これらにより、コーダは処理を行いたい（または行った）サンプル間を自由に移動することができるため、すでにコーディングを終えたデータに関する処理の内容を容易に確認することができる。また、画面右上のボタンは、左が職業分類の定義内容を閲覧する（図4参照）ためのもので、右はデータ検索（図6参照）のためのものである。さらに、職業コードの候補の左には、コーダが結果に自信がもてない場合にチェックを入れることのできる「要注意チェック」欄が用意されている。これは、一般コーダの後で最終的な判断（コーディング）を行う専門家コーダにとって有用であると思われる。また、作業画面右上には作業の進行状況が表示されるため、コーダが現時点以降の作業時間を見積もる目安として、達成感を感じたりすることができるものと思われる。

（4）結果ファイルに出力

ユーザにより決定された職業コードは結果ファイル(CSV形式)に出力され､通常のEXCELファイルとして扱うことができる。これにより、結果ファイルは、他のシステムへの入力が容易である。

４．NANACOシステムの適用結果

ここでは、NANACOシステムを適用した結果として、ここではJGSS-2003におけるコーダの正解率と作業時間について報告する。また、NANACOシステムについての評価票をコーダに回答してもらった結果についても簡単に報告する。

4.1　コーダの正解率

JGSS-2003におけるコーダの正解率を表4に示す。比較のため、対象とするデータやコーダが異なるが、JGSS-2002、JGSS-2001、JGSS-2000における結果をそれぞれ表5、表6、表7に示す。これより、NANACOシステムを利用した場合の正解率は、平均や各職業（JGSS-2002における本人現職以外）において、最も高いことがわかる。

ここで、各回におけるコーディング時の状況を比較する。まず、コーディング時に利用された参考情報については、最近ほど自動コーディングの結果を利用する度合いが強い。例えば、JGSS-2003では、SVM（add-code）による第1から第5位までの結果およびROCCOシステムによる結果の一部（(SVM（add-code）と一致したものがある場合に表示される)、またJGSS-2002およびJGSS-2001ではROCCOシステムによる結果が参考できるのに対し、JGSS-2000ではコーダは自動コーディングの結果を全く見ずに独自に行った(6)。JGSS-2000を除く3回では自動コーディングの結果が利用できたが、特にJGSS-2003では、1つのサンプルに対して常に職業コードが5個表示され、別のシステムによる結果との整合関係まで示されたという優位性は大きいと考えられる。また、同じ参考情報であっても利用形態の違いについては、例えば、職業データは、JGSS-2000では調査票を繰っていたが、

JGSS-2001 からは EXCEL ファイルとしてパソコン画面でみることができた。さらに、JGSS-2003 では属性や他の職業と一緒に整理された情報としてみることができた。また、『SSM 職業分類・産業分類』ファイルの利用は JGSS-2000 から行われているが、JGSS-2003 のように､コーディング中の画面と同じ画面上で随時確認ができるような使い方ではなく、『SSM 職業分類・産業分類』ファイルを利用できるパソコンに移動して操作する必要があった。

次に、コーダの熟練度については、毎回、熟練者が 2～3 人含まれるものの基本的には初心者がほとんどであるが、JGSS-2002 の場合だけは熟練者と JGSS-2001 を経験したコーダがほとんどであり、熟練度が相対的に高かったといえる。

以上のように、コーディング時の条件がさまざまに異なるために単純には比較ができないが、今回、熟練度が高かった JGSS-2002 を上回った理由としては、NANACO システムが特に初心者コーダの正解率を上げる効果があり、その結果、全体の正解率が高まったのではないかと考えられる。

表 4　NANACO システムを利用した場合のコーダの正解率（JGSS-2003）
（SVM（add-code）および ROCCO の結果を参考）平均：95.0%

本人現職	配偶者職	父職
93.2%	95.0%	96.7%

表 5　NANACO システムを利用しない場合のコーダの正解率（JGSS-2002）
（ROCCO の結果を参考）　平均：93.7%

本人現職	本人最後職	本人初職	配偶者職	父職
95.7%	94.3%	92.3%	91.9%	94.3%

表 6　NANACO システムを利用しない場合のコーダの正解率（JGSS-2001）
（ROCCO の結果を参考）　平均：83.5%

本人現職	本人最後職	本人初職	配偶者職	父職
82.1%	85.8%	85.1%	79.8%	84.5%

表 7　NANACO システムを利用しない場合のコーダの正解率（JGSS-2000）
（自動コーディングの結果を参考にしない）　平均：73.7%

本人現職	本人最後職	本人初職	配偶者職	父職
78.1%	72.1%	79.0%	68.8%	70.7%

4.2 作業時間

今回、NANACO システムを利用して「職業」のコーディングに費やした時間は、のべ 114 時間で 3 日間であった。収集されたサンプル数がこれまでより多いため、収集された職業の種類が減ったにもかかわらず、職業コーディングの対象となる有職者ののべサンプル数は、約 7,700 サンプルで、これまでより約 1000 サンプル程度多い。これより、処理時間を単純に計算すると、1 サンプルを処理するのに平均で 0.9 分(=54 秒)かかったことになり、速い。

職業・産業コーディングに必要な作業時間としては、毎回、2 週間（実働 10 日間）が予定されているが、今回は、現時点では NANACO システムを利用できない産業コーディングやコーディング作業前のデータの読み合わせ（2 日間）を行っても十分な日程であった。

また、コーダ（計 11 人）1 人当たりの平均作業時間を計算すると、約 10 時間という結果であった。

4.3 コーダによる評価

NANACO システムの改善を目的とする評価表を作成し、コーディング終了時に、コーダ(11 人）による評価を行ってもらった。その結果について、紙面の都合上、ごく簡単に報告する。コーダの内訳は、学部生が 3 人、院生が 8 人である。職業コーディングの経験者は 6 人、なしは 4 人、不明は 1 人である。評価はいずれも 5 段階評価である。

まず、NANACO システムの有用性について、総合的な場合と個別の機能に分けて尋ねた。総合的には「大変有用」が 7 人、「やや有用」が 4 人と評価された。主な機能についての評価を表 8 に示す。

表 8　NANACO システムにおける機能別の評価（単位：人）

機能	大変有用	やや有用	どちらでもない	あまり有用ではない
職業名の表示	5	4	2	0
職業内容の表示	5	6	0	0
職業コードの階層表示	0	3	6	1
職業定義内容の表示	9	1	1	0
回答の検索	4	1	3	1

次に、NANACO システムが提示する情報に関する評価を表 9 に示す。

表 9　NANACO システムにより提示される情報に関する評価（単位：人）

評価項目	大変有用/わかりやすい	やや有用/わかりやすい	どちらでもない	あまり有用ではない/わかりやすくない
職業コードの候補	2	7	2	0
類似データのリスト	0	4	4	3
提示のわかりやすさ	2	8	1	0

最後に、操作性についての評価について、全体と個別の操作を分けて尋ねた。全体では「大変わかりやすい」と「ややわかりやすい」がいずれも 5 人、「どちらでもない」が 1 人であった。個別の評価を表 10 に示す。

表 10　NANACO システムにおける操作性の評価（単位：人）

評価項目	大変わかりやすい	ややわかりやすい	どちらでもない	あまりわかりやすくない
システム開始作業	1	5	4	1
システム終了作業	0	5	4	2
画面の切り替え	5	2	3	1

以上より、簡単に考察を行うと、いずれの項目においても、5 段階評価において最も低いものがなかった。そこで、5 段階の中間評価を低い評価と合わせて、高い評価（ポジティブ）と低い評価（ネガティブ）の 2 つにまとめると、ほとんどの項目においてポジティブな評価の方が多いが、「職業コードの階層表示」機能、「類似データのリスト」情報（前述）、「システム終了作業」の 3 項目においては、ネガティブな評価の方が多かった。今後の課題としたい。

評価票では、この他に、改善や追加を希望する機能や操作について自由回答で収集されているが、この後に適用された他の調査[7]における評価と併せ、稿を改めて報告したい。

5．　おわりに

本稿では、次の 2 つについて報告した。すなわち、まず、JGSS-2003 における職業コーディングに対する 2 種類の自動コーディングの適用結果を報告し、次に、コーダによるコーディング作業そのものを支援する目的で開発された NANACO システムに関する報告を行った。自動コーディングの適用結果においては、ROCCO システムを適用した結果（精度と

再現率）をこれまでの 4 年間の報告とも併せて報告し、ROCCO システムの安定性を明らかにした。また、機械学習である SVM に ROCCO システムを組み合わせた手法を適用した結果（正解率)についても報告し、この手法が ROCCO システムをはるかに上回ることを示した。次に、NANACO システムについては、概要説明を行った後、今回初めて実際のコーディング作業に適用された結果について、コーダの正解率と作業時間においていずれも高い効果があったことと、コーダによるシステムの評価も概ね良好であったことを報告した。

今後の課題は、まず、NANACO システムにおいてあまり評価の高くない類似データのリストについて検討することである。また、これと関連するが、コーダに対する参考情報としてより有効であると思われる「過去のデータと実際にこのデータに付与されたコードのペア」を容易に検索できる機能を追加することを検討中である。

NANACO システムは、JGSS-2003 以降 2 つの調査で利用され、2005 年 SSM 調査においても利用が決定されている。今後も、コーダに対する効果的な支援とは何かということについての検討を重ね、よりよいシステムの構築を目指していきたい。

[Acknowledgement]

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（研究代表：谷岡一郎・仁田道夫、代表幹事：佐藤博樹・岩井紀子、事務局長：大澤美苗)。東京大学社会科学研究所附属日本社会研究情報センターSSJ データアーカイブがデータの作成と配布を行っている。

NANACO システムの開発に当たっては、構想段階から実用に至るまで、石田浩教授・篠崎武久助手を始めとする東京大学社会科学研究所スタッフの多大なご協力を得ましたことを記して感謝いたします。

[注]

(1) JGSS-2000 から JGSS-2002 までは、今回の 3 種類以外に「本人最後職」、「本人初職」も収集されていた。

(2) 形態素解析ソフト JUMAN（黒橋・長尾 1999）により切り出された語の「原形」と品詞を用いた。

(3) 同一のデータ（JGSS-2002）において SVM(add-code)と SVM(seq)の差は約 2%程度であった（高橋他 2004b）ことを考慮すると、この差は大きいと判断できる。

(4) 訓練データとして用いたデータ数は、JGSS-2000（6,848 サンプル)、JGSS-2001（6,448 サンプル)、JGSS-2002 (6,770 サンプル）であることから、JGSS-2002 では 13,296 サンプル、JGSS-2003 では 20,066 サンプルである。

(5) ベクトル空間法は、データの両方をベクトルとして統一的な空間の中に表現し、その間に

類似度を定義することにより、類似したデータを見つける方法である。Cosine 類似度は、次の式で表される内積であり、類似度の定義としてしばしば用いられる。

$$x \cdot y = |x||y|\cos\theta$$

ここで、$|x|$、$|y|$はそれぞれベクトルの長さ、θは2つのベクトルのなす角を表す。

(6) ただし、コーダはコーディングした後でROCCO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ったという意味では、ROCCO システムによる結果を参考にしていないわけではない。しかし、ここで用いられたデータは、ROCCO システムと独立にコーディングを行ったときのものである。

(7) NANACOシステムは、「要介護状態及び健康の形成過程における社会経済的要因の役割に関する実証的研究」(基盤研究(A)(2)研究代表者武川正吾)および「2003年SSM予備調査」における職業コーディングにも適用された。いずれも、コーダから今回と同じ評価表を用いて評価を行ってもらった。

[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』.

工藤拓, 松本裕治, 2002,「Support Vector Machineを用いたChunk同定」,『自然言語処理』9(5), pp.3-22.

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.

長尾　真・黒橋禎夫・佐藤理史・池原　悟・中野　洋, 1998,『岩波講座言語の科学９　言語情報処理』.

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.

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), 2003,『日本版General Social Surveys JGSS-2001　基礎集計表・コードブック』東京大学社会科学研究所.

須山敦,2004,『NANACO system 操作マニュアル(WINDOWS用)』.

高橋和子, 2002a,「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-2001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), pp.171-184.

高橋和子, 2002b,「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』, pp.491-494.

高橋和子, 2003,「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」,『日本版General Social Surveys研究論文集(２)　JGSS-2001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), pp.179-192.

高橋和子，高村大也，奥村学，2004a,「機械学習とルールベースによる職業コーディング」，『情報処理学会第 159 回自然言語処理研究回報告 NL-159-9』，pp.53–60.

高橋和子，高村大也，奥村学，2004b,「機械学習とルールベースの組み合わせによる職業コーディング」,『言語処理学会第 10 回年次大会発表論文集』，pp.737-740.

高橋和子，須山敦，村山紀文，高村大也，奥村学，2004c,「職業コーディング支援システム（NANACO）の開発」，『第 37 回数理社会学会大会研究報告研究報告要旨集』,pp. 20-23.

高橋和子，2004,「職業コーディングにおける ROCCO システムと SVM の組み合わせ」,『日本版 General Social Surveys 研究論文集（３） JGSS-2002 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），pp.163-174.