

定型データの処理・分析法に関する基礎的研究』(平成 3 年度文部省科学研究費報告書) : 205-214.

上野晴樹. 1985. 『知識工学入門』 オーム社

上野晴樹・小山照夫(編). 1988. 『エキスパートシステム』 オーム社

付 録

職業分類における職業コードと職業名一覧

(小分類のみを一部抜粋)

- 1 自然科学系研究者
- 2 人文・社会科学系研究者
- 3 鉱山技術者
- 4 金属製練技術者
- 5 機械技術者
- ⋮
- 54 管理的公務員
- 55 会社役員
- ⋮
- 61 一般事務員
- 62 会計事務員
- ⋮

- 70 小売店主
- 71 卸売店主
- ⋮
- 83 農耕・養蚕作業者
- 84 畜産作業者
- ⋮
- 97 採鉱員
- 98 採炭員
- ⋮
- 105 電気・ディーゼル・蒸気機関士
- 106 電車・気動車運転士
- ⋮
- 121 制工、製鋼工
- 122 非鉄金属製練工
- ⋮
- 269 自衛官
- 270 警察官、海上保安官、鉄道公安官
- ⋮
- 274 家事手伝い(住込みの女子)
- 275 家政婦
- ⋮
- 293 分類不能の職業

もある。

- 8) 「採鉱員」や「石切出作業者」において、「ここで」に続いて、それぞれ鉱物や石材についての説明がされている。
- 9) それぞれ「軽くよりをかける作業」、「体裁をよくするために」のように用いられる。
- 10) 「権限に属させた」（「裁判官、検察官、弁護士」、「飲食(を)させる」（「バーテンダー」、「その他の個人サービス職業従事者」）のように用いられる。
- 11) 「の」の用例を示す。
……研究、試験、検定、分析、鑑定、調査などの専門的、科学的な業務に従事するものをいう。（「自然科学系研究者」）
- 12) 「を」の用例を示す。
……各機関の課（課相当を含む）以上の部署の業務を管理、監督する業務に従事するものをいう。（「管理的公務員」）
- 13) 箇条書の文を「情報処理技術者」における例で示す。
 - (1) 電子計算機の操作に従事するものは「電子計算機等操作員69」に分類される。
 - (2) カード穿孔機及びデータ・エントリー装置の操作に従事するものは「せん孔機等操作員68」に分類される。
- 14) 「会計事務員」における例で示す。
郵便貯金、郵便為替、窓口為替、窓口受払い、原簿登記、証書発行などの貯金現業業務に従事するものも本分類に含まれる。
- 15) その他に、「修理作業」、「調整作業」、「草刈作業」、「紡織作業」、「点検作業」、「代理業務」などがある。
- 16) 役割格 (role) (「として」) も 2 例存在する。

[参考文献]

- 原 純輔編. 1993. 『S S M職業分類 (改訂版)』
- 原 純輔・海野道郎. 1984. 『社会調査演習』
東京大学出版会
- 岩本秀明・野村浩郷. 1991. 「法律文の自然言語処理について」『情報処理学会研究報告 NL-83』：7-14
- 情報処理振興事業協会. 1987. 『計算機用日本語基本動詞辞書 I P A L (Basic Verbs) - 解説編一』
- 長野 馨・永井秀利・中村貞吾・野村浩郷. 1993. 「動詞の機能に基づく法律文の制限言語モデル」『情報処理学会研究報告 NL-95』：25-32
- 長尾 真. 1986. 『機械翻訳はどこまで可能か』
岩波書店
- 中岩浩巳他. 1993. 「J E I D A 機械翻訳システム評価基準 (開発者編) —— 開発者のための評価項目の作成と検討——」『情報処理学会研究報告 NL-96』：73-80
- 佐藤嘉倫. 1992. 「職業コーディング支援システムの構築」原 純輔 (編) 『非定型データの処理・分析法に関する基礎的研究』(平成3年度文部省科学研究費報告書)：199-204.
- 総務庁統計局・統計センター『昭和60年国勢調査に用いる職業分類』
- 高木 朗・伊藤幸宏. 1987. 『自然言語の処理』
丸善
- 高橋和子. 1992. 「自由回答における構造化支援システムの開発について」『茨城大学人文学部紀要 (社会科学)』：103-124.
- 都築一治. 1992. 「職業コーディング自動化システムの試験的構築」原 純輔 (編) 『非

システム構築のために、自然言語処理による職業分類の構造化を行った。現在のところ、並列構造の解釈などは人手に依存しており、知識の獲得をすべてコンピュータが行っているわけではない。今後は、語レベルの並列構造など比較的解決しやすいものについては、可能な限り、コンピュータによる方法を考えたい。また、教科書知識だけでなくヒューリスティックな知識の獲得についても「専門家」の協力を得て検討し、カテゴリー側となる知識ベースを充実したものにしていきたい。

職業分類における知識獲得がうまく行われれば、職業コーディング・エキスパートシステム構築の準備が整うばかりでなく、他の解説書一般にもその成果が応用できるものと期待できる。

注)

- 1) 職業コードと職業名については付録を参照のこと。一般の調査におけるものと比較すると非常に細かい分類となっている。
- 2) ①従業上の地位、②従業先の名称、③従業先事業の種類、④雇用者数、⑤本人の仕事内容、⑥役職名を尋ねるが、このうち①、④、⑥が選択肢形式で、②、③、⑤が自由回答形式となっている。
- 3) 佐藤 (1992) は職業分類をそのまま文章型データベースとして既存のデータベース用ソフトを利用する。検索するためのキーワードの設定に専門的知識が必要である。都築 (1992) は統計パッケージ S A S (Statistical Analysis System) によりあらかじめ各職業分類における文章を 2 語ずつに切出しておき、回答においても 2 語ずつ切出したものとのヒット回数により機械的に判断を行う。検索

のアルゴリズムが明示化される。

4) 職業分類の一例を示す。

1 自然科学系研究者

研究所、試験場、研究室などの研究施設において、専ら理学、工学、農学、医学、薬学など自然科学に関する研究、試験、検定、分析、鑑定、調査などの専門的、科学的な業務に従事するものをいう。

ただし、専門的・科学的知識と手段を生産における規格、管理、監督、研究など科学的な業務に従事するものは「技術者 3 ~ 12」に分類される。

○物理学研究員、(途中略)、

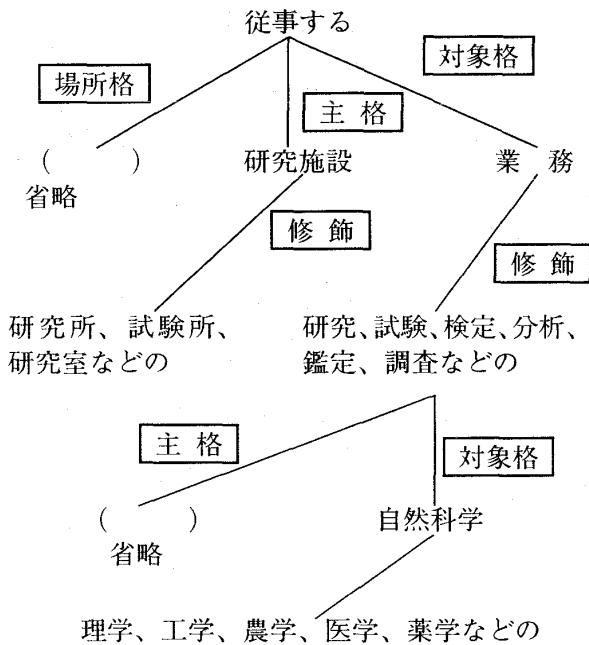
○解剖学研究員

×化粧品製造技術者 [7]、(途中略)、

×試験工

- 5) 「芸者、ダンサー」における解説文は、「芸者及びダンサーをいう。」のみである。
- 6) 「分類不能」における解説文は、「いずれの項目にも含まれない職業が分類される。これは主に調査票の記入が不備であって、いずれの項目に分類すべきか不明の場合又は記入不祥で分類しえないものである。」である。
- 7) 例えば、「郵便局長、電報・電話局長」において、「郵便局長として郵便その他の郵便現業業務を統括するもの及び電報局、電話局又は電報電話局長として電報、電話に関する通信現業業務を統括するものをいう」のように用いられるように、「及び」が節同士を結合することが多いのに対して、「又は」は語同士を結合することが多い。しかし、「及び」は「化学技術者」における「……………油脂・油脂製品 (動植物油脂及び食用油脂製品を除く) ……………」のように語同士を結合する場合

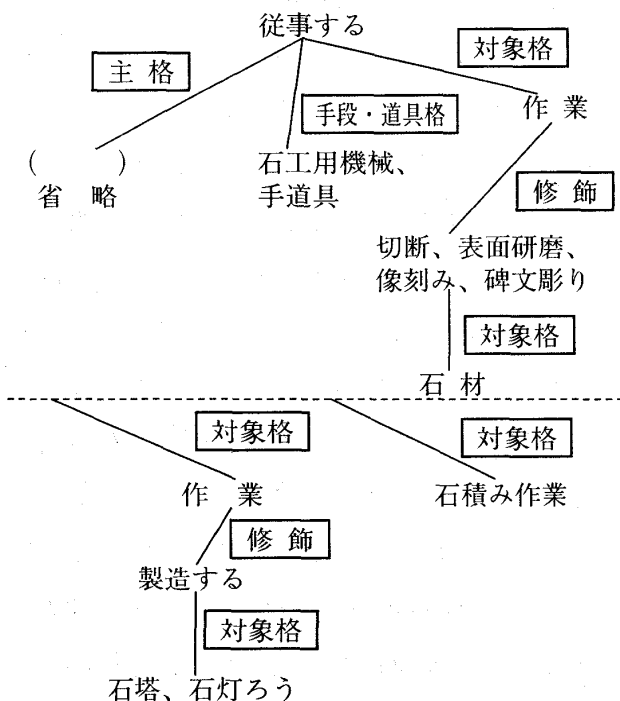
図 6



注) 修飾部における形容動詞 (相当語) に注目すると何らかの格を与えることが可能であると考えられるが、今回は省略した。

もう一例、「従事する」が手段・道具格をもち、「作業」が並列構造であるものの例として、「石工」の職業定義文を構造化したものを図7に示す。

図 7



注) 修飾部の主格は紙面の都合上、省略した。

「従事する」が用いられた職業定義文は285個存在するが、上記の例で示したもの以上に複雑な構造のものは存在しない。

最後に、「行う」と「総括する」が用いられた職業定義文についても、格文法による木構造で表現する(図8、図9)。ただし紙面の都合上、修飾部までは構造化していない。

図 8

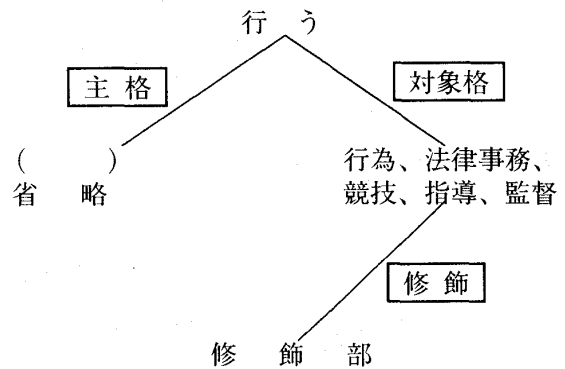
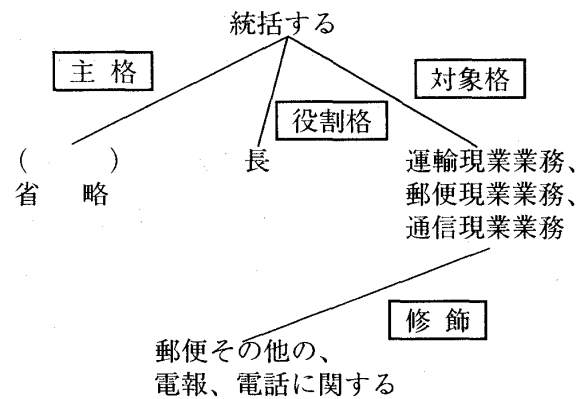


図 9



以上でわかるように、格文法に基づいて修飾部まで構造化すると、職業の意味内容が形式的に表現できる。すなわち、職業分類に存在する知識を形式的に取扱うことが可能となる。

5. おわりに

本稿では、職業コーディング・エキスパート

ができる。ただし、この動詞（相当語）は、前述したように非常に種類が多い上に頻度が少ないため、今回は各々の動詞（相当語）についての結果は示さず、いくつかの例を示すこととする。

まず、修飾部に出現する主な任意格に注目して動詞（相当語）をおおまかに分類してみる。ただし、これは動詞が本来的にもつ（とされる）格スロットによるものでなく、職業定義文における用いられ方による分類である。

対象格のみで、任意格をとらないものが約65%程度あり、最も多い。任意格は、「を用いて」、「を利用して」、「で」で表現される手段・道具格（tool）と「のため（に）」や「を目的として」で表現される目的格（purpose）がほとんどで、それぞれ修飾部の約15%、約20%のものに存在する。両者が同時に用いられることはほとんどない。「に関する」や「についての」は、修飾部においては対象格であるとみなせる。これより、修飾部における動詞（相当語）は一応、次のように分類できる。

- 対象格のみをとるもの
- 対象格+手段・道具格をとるもの
- 対象格+目的格をとるもの

表 5

対象格のみをとるもの 研究、販売、企画、販売、保守、採取する
対象格+手段・道具格をとるもの 製造、作る、製作する、曲げる、組み立てる
対象格+目的格をとるもの 代理、裁断、梱包する
対象格+その他の格をとるもの 搬出する

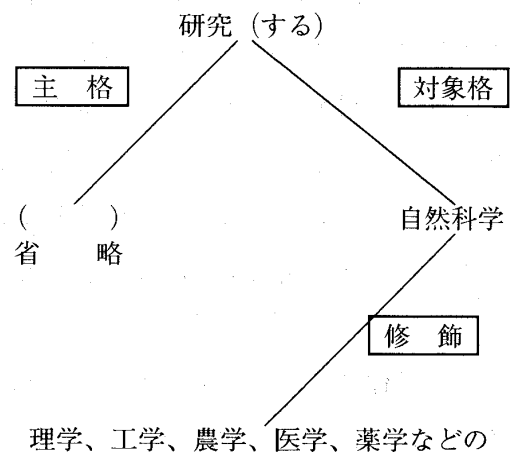
- 対象格+その他の格をとるもの¹⁶⁾

表5に、それぞれの代表例を示す。「製造」や「彫刻」など同一の動詞（相当語）でも、職業により注目する点が異なるために複数のタイプに分類されることがある。

職業により特徴づけられ方がさまざまに異なるが、それを格関係から読み取ろうとすると、例えば手段・道具格や目的格をとるものは（半・非）熟練的な職業に多く、その内容の違いが職業を決定づけることが多いことがわかる。従ってこのような職業においては、これらの格となるものの中身を具体的に示すことが重要となる。

次に、修飾部を格文法による木構通で表現する。修飾部の動詞も、「従事する」の場合と同様に主格はすべて省略されている。例として、「自然科学系研究者」における一つの動詞（相当語）である「研究」に注目したものを図5に示す。

図 5



これを図4の修飾部に組込んで、「自然科学系研究者」の職業定義文を構造化したものを図6に示す。

構文解析の方法には、句構造分析や格文法などさまざまなものが存在するが、本稿で用いる格文法は、Fillmoreにより1968年に提唱されたもので、まず文中の主要な動詞に注目し、名詞句をその動詞との格関係を中心に分析していく方法である。句構造分析と比較すると、意味をある程度扱うことができるために都合がよい。また語順が問題とならないために、日本語の解析には適しているとされる。

職業定義文において主要な動詞となるのは、前述したように「従事する」や「行う」、「統括する」などであった。最初に極めて頻度の高かった「従事する」を中心に分析を行い、その後、「行う」や「統括する」についても簡単に触れる。

職業定義文における動詞「従事する」は、対象格 (object) として「作業」、「業務」「仕事」をとることが多い (約72%) が、その他「操作」、「事務」、「教育」などもとる。対象格はこの動詞にはなくてはならないもので、必須格である。「従事する」はまた約20%の文において、「において」や「で」で表現される場所格 (space) をとる。文法的には場所格が欠けていても文が構成できるので、これは任意格である。また前述したように、職業定義文においては常に主部が省略されているために、主格は存在しない。以上より、動詞「従事する」が用いられている職業定義文は、格文法により図4のような木構造で表現される。

対象格が「作業」などの形式的な名詞相当語である場合には、修飾部が重要であり、この部分に対する分析が必要となる。形式的な名詞相当語を直接修飾してその職業内容の意味を限定するものは、形容動詞を除外すれば、次の4種類のうちのいずれかの形態をとる。

- ・名詞
- ・名詞+格助詞「の」
- ・名詞+動詞「する」の連体形
- ・動詞の連体形

名詞のみで修飾するものは、「整備作業」や「補助業務」のように複合語として存在し得るものである¹⁵⁾。ここで、「整備」は職業内容をある程度規定するサ変名詞であるが、「補助」は形容動詞 (「補助的だ」) に相当するような意味しかもたないため、形容動詞と同様に扱うこととし、ここでは前者のような名詞を分析の対象とする。

この名詞には格助詞「の」をつけることができるが、それは次の名詞+格助詞「の」の形となる。名詞+各助詞「の」は、例えば「修理の作業」、「研究の業務」、「販売の仕事」などである。また、名詞にそのまま動詞「する」をつければ、名詞+動詞「する」の連体形の場合になる。これは、「修理する作業」、「採取する作業」、「販売する仕事」などである。動詞の連体形は、「組み立てる作業」や「整える業務」などである。

形態の違いについては、基本的には動詞が用いられるが、その動詞の名詞形や複合名詞が語として存在する場合には、文の前後関係から適当なものが選ばれていると説明できる。実際、「修理」という語は、「修理作業」(2)、「修理の作業」(19)、「修理する作業」(3) (()内は出現頻度を表す) の3種類の形態で用いられるが、意味的には同一のものである。

これより、「作業」などの形式名詞相当語を直接修飾する語は、形態的には異なっても、動詞に相当する語であると考え、これらを修飾部における主要な動詞であるとみなすと、この部分に対しても格文法による構造化を行うこと

けられる。このうち、定義除外文は除外する場合が単数か複数かにより、表現形式が異なる。単数の場合は、「ただし」以下にそのまま続けて適用が除外される場合を説明するが、複数の場合は、2.3で述べた

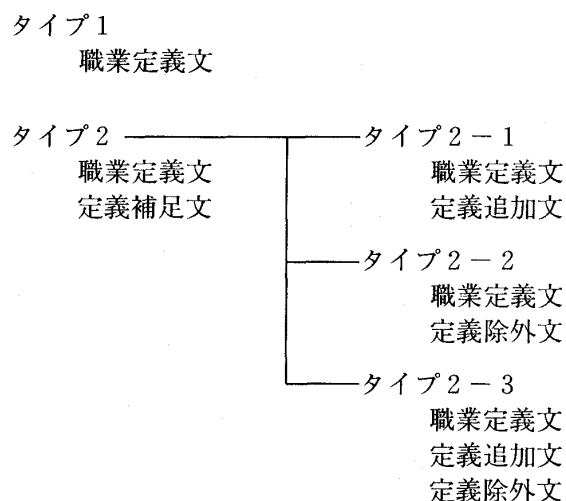
表 4

解説文の種類	文のある位置又は形式
職業定義文	最初にある文
定義追加文	文末が「も本分類に含まれる」である文
定義除外文 (単数の場合)	文頭が「ただし」、文末が「分類される」である文
定義除外文 (複数の場合)	文頭が「ただし」、文頭が「含まれない」である文 + 箇条書の文

ように、まず定型文で除外する場合があることを予告をした後に、箇条書の文で除外される場合を説明する。

ここで職業定義文は必須であるが、定義補足文は任意である。解説文を論理構造の上から分類したものを図3にまとめる。

図 3



これらは表4にまとめたように、文の位置または形式から特定できるため便利である。すなわち、論理構造が表層的な情報により表現できるために、コンピュータによる処理が行いやすいという利点がある。

ところで、職業定義文と定義追加文は見出しに提示された職業の

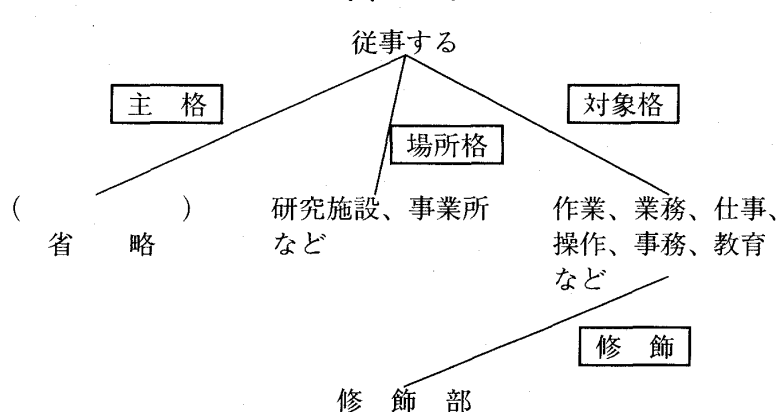
適合範囲を規定し、定義除外文はそれとは別の職業の適合範囲を規定すると考えられる。すなわち、定義除外文の内容と同様のものは別の場所でも述べられており、見出しに提示された職業に関しては新たな情報をもたないことを考慮すると、当面はこの文を分析する必要がないと考えられる。結局、職業分類を知識ベースとする場合には、職業定義文と定義追加文を中心に構築すればよいことがわかる。

ここで、定義追加文の定型的な文末「も本分類に含まれる」を除外して、その直前の「従事するもの」を文末と解釈すると、これは形式的にも職業定義文と同様となるために、以下では、職業定義文についてのみ検討することとする。

4. 格文法による構造化

言語情報と論理情報より得た結論から、ここでは職業定義文を格文法に基づいて分析する。

図 4



2.3 構文について

職業分類における解説文は、文の位置や形式から次の3種類に分けられる。

A：解説文の最初に登場する文

B：「ただし」で始まる文

C：主部の最後に副助詞「も」があり、述部が「本分類に含まれる」である文

いずれも接続部や独立部はないものの、修飾部が長く、また並列構造が語、句、節などさまざまなレベルで含まれるために複雑な構文となっている。まず、解説文の種類別に分析を行う。

Aは、注4の例における「研究所」から「従事するものをいう」までの文であり、どの解説文にも必ず存在する。常に主部が省略され、述部のみからなるという特徴をもつ。ここで、述部は動詞の項で述べたように、「ものをいう」ではなく、その直前の「従事する」などとした。修飾部はすべて述部にかかるが、語や句レベルの並列構造を含む場合が多い。また、「……従事するもの、……従事するものをいう」のように述部自体が複数個存在するものが約10%ある。この場合は節レベルの並列構造と考えられる。

Bは約50%の職業に存在するが、形式的に2種類に分けられる。一つは、注4の例のように、主部（「従事するものは」）が並列構造を含む長い修飾部をもち、述部が常に定型的で「」に分類される（「」内には職業名と職業コードが入る）であるもの。もう一つは、まず「ただし、次の業務に従事するものは本分類には含まれない」なる定型的な文があった後、箇条書による1～4個の文が続くものである。箇条書の文は、基本的には先のものと同様に、修飾部をもつ主部と定型的な述部からなる¹³⁾。両者

の出現度数は全体のそれぞれ23%と27%であり、ほぼ等しい。いずれも修飾部の並列構造は、語や句のレベルである。

Cは約18%の職業に存在する¹⁴⁾。比較的簡単な構文のものが多く、修飾部がある場合は必ず主部にかかり、語や句レベルの並列構造を含むものもある。

並列構造は、名詞または名詞句によるものが多い。連用中止形のもの動詞の項で述べた「有し、」が代表的で、その他はほとんどみられない（合計で約7%程度）。また、埋め込み文は存在しない。

以上より、構文の分析においては並列構造を解決することが重要であることがわかる。しかし前述したように、接続詞の違いにより機械的に範囲を決定することができない場合も多いために、意味まで考慮した分析が必要となる。

3. 論理構造分析

ここでは職業分類における解説文の論理情報について述べる。

解説文は通常、複数個の文から構成されるが、一つの文が論理的には一つの意味を表す。すなわち、それぞれの文が各意味要素と考えられるために、論理的な構造は文単位に扱うことができる。

解説文は、論理的には大きくは職業を定義するもの（以下、職業定義文とよぶ）とそれに続けて補足を行うもの（以下、定義補足文とよぶ）に分けられる。定義補足文には、定義内容に追加される場合を説明するもの（以下、定義追加文とよぶ）と、定義内容から除外される場合を説明するもの（以下、定義除外文とよぶ）に分

約50% (143) 近くの職業において、2番目以降の解説文の先頭に出現する。最初の文で定義した職業の内容に追加する形で、その職業が除外される場合を説明する文には必ず用いられる。

なお、「ただし」が出現しない職業のうち、特定の語の説明を行うものとして「ここで」が用いられているものが2つある。⁸⁾

④ 形容詞

形容詞は非常に少なく、「軽い」と「よい」がいずれも1ずつ出現するだけである⁹⁾。職業分類においては、形容詞が名詞句を修飾することはほとんどなく、意味的にも分析の対象とする必要はないと考えられる。

⑤ 副詞

副詞は「専ら」と「主として」のみで、出現頻度もそれぞれ5と2で非常に少ない。副詞が動詞句を修飾することはほとんどなく、形容詞と同様に分析の対象とする必要がないといえる。

⑥ 形容動詞

形容動詞も少なく、最も多い「技術的だ」は約4% (11) であり、用法はすべて「技術的な」、「専門的な」、「科学的な」などのように連体形に制限されている。しかし、これらは職業の内容をある程度規定しており、語の意味を考慮する必要がある。

⑦ 助動詞

助動詞により、解説文における動詞をヴォイス(態)情報、テンス・アスペクト(時相)情報、ムード(法)情報の観点からまとめてみる。

ヴォイスは、ここでは動詞自体に本来的に有する語彙的なものではなく文法的なものを

いう。「ラレル」形である直接受動が約90%程度 (275) の職業において存在するが、「含まれる」、「含まれない」、「分類される」のように用いられるために多くなっている。その他は「サセル」形である使役が1% (3)¹⁰⁾ あるだけで、可能、自発、尊敬はない。テンス・アスペクトにおいては、すべてが現在で、未来はない、また単純状態、進行、結果依存、経験・完了もない。ムードにおいては、命令、願望、意志・勧誘、推量はない。

助動詞においても、非常に制限されていることがわかる。

⑧ 助詞

助詞は非常に多いが、格助詞と副助詞に制限され、接続助詞や終助詞はない。格助詞は「の」、「に」、「を」、「が」のみで、副助詞では「など」と「も」に限られている。「の」、「に」、「を」はほとんどの職業に出現するが、「が」は約10% (27) である。また「など」は72% (212)、「も」は18% (53) である。

格助詞「の」は、名詞に付いてそれ以前の語がある場合にはそれもまとめながら、形式的な名詞に相当するとした「作業」、「業務」、「仕事」に対する修飾句を作る¹¹⁾。「に」はそれらの語に付いて、「従事する」にかかる。「を」は前述したように文末「ものをいう」により頻繁に出現するが、これを除外すると、前述した形式的な名詞相当語に対する修飾句の中で、連体形の動詞とともに用いられる¹²⁾。「が」も同様である。修飾句については4でさらに検討する。

副助詞「など」は、必ず読点とともに用いられて、並列構造を表す。「も」は必ず、その後「本分類に含まれる」が続く。

まるが、この修飾句には名詞だけでなく連体形の動詞も用いられることが多い。名詞の場合と同様に、出現頻度は少ないが多岐にわたるものが多く、「掘進する」などのように特殊な語も含まれる。これについては4で検討する。

「分類する」は、実際にはすべて受身「分類される」として用いられる。出現頻度は約50% (145) である。「含む」は、受身「含まれる」(約20% (53))か受身の否定「含まれない」(約25%(77))のいずれかである。これらはいずれも解説文の最初の文には出現しない。「含まれる」の場合は必ず2番目の文の文末である。「分類される」または「含まれない」が文末にくる文の先頭は、必ず、接続詞「ただし」で始まるという共起関係がある。

この他、「有する」が常に連用形「有し」+読点(「、」)という形で、資格を必要とする職業(「医師」など)に出現する。これは格助詞「を」をとって、必ず、名詞「免許」とともに用いられる。出現頻度は約4% (13) である。

「関する」と「おく」は格助詞「に」に続いて、「に関し(て)」、「におい(て)」や「に関する」、「における」のように連用形や連体形で出現する。「用いる」は、格助詞「を」に続いて、「を用い(て)」という副詞句のかたちでのみ出現する。頻度はそれぞれ39、49、24でいずれも約10~20%程度である。

「する」はサ変名詞を動詞化する場合に用いられ、同様の意味を表すものとして格助詞「を」をともなって「をする」と用いることもできるが、職業分類においては前者の用法がほとんどである。ほとんどが連体形だが、

連用形も少数存在する。

③ 接続詞

接続詞の出現頻度を表3に示すが、内容的に2種類のものに分けられる。

表 3

接 続 詞	出 現 頻 度	%
及 び	178	60.8
た だ し	143	48.8
又 は	86	29.4
並 び に	11	3.8
も し く は	5	1.7

一つは、「及び」や「並びに」(並立・累加)と「又は」や「もしくは」(対比・選択)で、職業分類においてはすべて並立の関係(以下、並列構造とよぶ)を表すものとして用いられている。これを出現頻度の高い「及び」と「又は」に注目してみると、一般的には「及び」は主として小さな要素同士、「又は」は大きな要素同士を結合するというように、ある程度の範囲を定めるものであるが、ここではむしろ逆の用いられ方がされることが多かったり、規則性がみられなかったりする⁷⁾。「並びに」や「もしくは」まで含めると、さらに不規則なものとなり、言語情報だけから並列構造を解釈することは不可能である。

ここで、接続詞ではないが、並列構造を表すために読点が多用されている(注4の例を参照のこと)ことに注目し、これを接続詞に相当するものとして、単なる読点と区別して扱うこととする。これも含めると、並列構造は「分類不能の職業」以外のすべてのものに見られる。

もう一つは、「ただし」(説明)であるが、

「作業」が約70% (204)の職業で用いられていて最も多い。「業務」も約25% (72)でやや多いが、「操作」、「仕事」、「事務」からは10%未満に急減している。この後、「教育」や「保育」、「指導」などが続いた後、出現頻度1のものが非常に多くの種類で出現する。この中には、「鑄造」、「心出し」のような各職業独特の専門用語も多く含まれている。

ここで「作業」、「業務」、「仕事」の3語に注目すると、これらは相互に類似する意味をもつが、職業によって用いられ方が異なる。主に「作業」は職業コード121以降の(半・非)熟練的な職業、「業務」は同79以前の専門的または管理的な職業、「仕事」はその間のコードで事務的または販売的な職業に出現しており、かなり明確に区別される。これらはいずれも格助詞「に」が付いて、動詞「従事する」に続く場合が多いが(それぞれ145、42、15で合計202(約70%)となる)、これだけでは職業の具体的な内容が特定できないため、解説文においてはこれらの語は形式的な名詞に相当するものと考えられる。従って、これらを前から修飾して意味を作り出している他の語やそれとの関係について調べる必要がある。この修飾句には、前述したような少頻度で多種類にわたって出現する傾向のある名詞が多く含まれる。これに関しては4で分析を行う。

なお、「もの」は、ほとんどの解説文の末尾に「・・・ものをいう」なる表現があるために多く出現するが、形式名詞でもあり、特に分析の対象とする必要がないと判断できるために除外した。

名詞に関連して代名詞について述べておくと、「これ」のみが6% (18)の職業に出現

するが、数としては少ない。

② 動 詞

ここでは述部との関係から文末にくる動詞を中心として述べる。ここで、文末に必ず出現する「いう」については、名詞の項で述べた「もの」と同様に分析の対象とはしないことから、「ものをいう」が文末となるものについては、その直前の動詞を文末のものとして扱うこととする。

表 2

動 詞 ^{注)}	出現頻度	%
従事する	285	97.3
分類する	145	49.5
含 む	130	44.4

注)動詞は基本形に直してある。

表2に、解説文の文末にくる動詞のうちで出現頻度の高いものを示す。

「従事する」が文末にくる可能性が非常に高いことがわかる。解説文の最初の文において、「従事する」以外では、「行う」と「統括する」が2ずつ、「教授する」と「執行する」が1つずつで、それ以外にはなく、この文については動詞が非常に制限されて用いられていることがわかる。また解説文全体でみると、「従事する」、「分類する」、「含む」のうちのいずれかが必ず文末にくる(「分類不能の職業」⁶⁾を除く)ことから、全体でも制限された動詞が用いられることがわかる。

「従事する」は格助詞「に」をとって、前述したように「作業」、「業務」、「仕事」などとともに用いられる。これらの語は名詞の項で述べたように、前からの修飾句で意味が定

図 2

職業コード	職業名
解説文	
該当職業例	
非該当職業例 ^{注)}	

注) 非該当職業例は存在しない場合もある。

文ら成り、各職業はすべて図2のような構成で説明される。

職業分類の例を注4に示す⁴⁾。全体では1つの文は平均59.3文字から成るが、職業コードと職業名からなる見出し項目を除くと71.8文字となり、さらに(非)該当職業例を除いた解説文部分においては134.4文字となる。そこでの最短文は「芸者・ダンサー」における12文字⁵⁾、最長文は「農林技術者」における307文字で、バラツキが大きい傾向がある。

解説文は平均で2.3個の文からなり、最小は1個、最大は6個である。大まかには、解説文は平均130文字程度からなる文2～3個から構成されるといえよう。

ここで注4に示した例より明らかなように、(非)該当職業例は、職業名の先頭に○または×を付けただけのもので、文とはいえない。従って、本稿では解説文を中心に分析を行う。

ところで、各職業には必ず該当職業例が付いているが、職業分類をデータベースとしてキーワード検索を行う場合は、この部分にヒットする可能性が高いものと考えられる。これは今後、新システムを開発する際にも活用するべきであろう。

2.2 語彙について

まず出現傾向を品詞別に調べると、名詞(サ変名詞を含む)、動詞、接続詞が多く、形容詞や副詞はほとんどない。形容動詞も少なく、かつ制限されたものしかない。また職業分類の性質から当然であるが、感動詞は全くない。連体詞も「その」が頻度1で出現するだけである。助動詞は受身と否定以外はなく、助詞も格助詞と副助詞に制限されている。すなわち、自立語においては名詞、動詞、接続詞の3種類に制限されており、付属語においては制限された意味でしか用いられていない。以下で、これをより詳細に示す。なお、ここで出現頻度というのは語が登場した職業の個数を表す。

① 名 詞

サ変名詞(用言性名詞)は「する」を付けると動詞化する名詞のことであるが、これも含めて、出現する語の種類が多岐にわたること、しかし限られたいくつかの語については頻繁に出現するという傾向がある。表1に出現頻度の多いものを示す。

表 1

名 詞	出現頻度	% ^{注)}
作 業	204	69.6
業 務	72	24.6
操 作	28	9.6
仕 事	20	6.8
事 務	18	6.1

注) %は出現頻度を全職業(293)で割ったものを表す(以下同様)。

職業分類における自然言語処理について

——職業コーディング・エキスパートシステム

構築のための予備的考察——

高橋 和子

Natural Language Processing in the Occupation Category

Kazuko Takahasi

1. はじめに
2. 言語分析
3. 論理構造分析
4. 格文法による構造化
5. おわりに

1. はじめに

一般に社会調査における回答者の職業は、性別や年齢と共に重要な属性の一つである。1955年から10年毎に実施されてきた「社会階層と社会移動全国調査」(Social Stratification and Social Mobility Survey—以下SSM調査と略称する)においては、その研究目的上、特にこの傾向が顕著で、非常に詳細な職業データを必要とするために、まず、現在までの職歴が自由回答形式により得られ、最終的にはあらかじめ用意された293種類の職業(総務庁統計局・統計センター『国勢調査に用いる職業分類』における小分類—以下職業分類と略称する¹⁾)にコーディングされるという方法が採られている。

SSM調査においては、職業データは比較的構造化されており²⁾、カテゴリー側である職業

分類も明快に定義されている。しかし、主として有効回答数の多さ(約4000)やカテゴリーの細かさから生じる判断のむつかしさなどの理由により、自由回答特有の問題とされる「コーディング段階における作業の煩雑さ」が存在し、例えば1985年調査の場合、「20~30人のメンバーの泊まり込み作業でほぼ1週間を要した」(原、1993)ほどの膨大な作業量の問題があり、作業内容においても、コーダーは各人が随時、ごく少数の「専門家」に尋ねながら行うという非効率な状況があったことが報告されている(佐藤、1992)。

これを改善するために、最近、コーディング段階にコンピュータを利用しようとする考え方が生まれた(佐藤、1992、都築、1992)。特に佐藤(1992)のシステムはその後改良が重ねられて、現在ではほぼ実用の域に達しているとされる(原、1993)。両者はかなり異なったアプローチをとっている³⁾が、いずれも職業分類をデータベース化して、キーワードによる検索を行う点で共通性がある。すなわち、職業分類を「データ」の集まり(データベース)であるとしてコンピュータによる検索を行い、コーダーは提示