ISSN 1348－7345

製表技術参考資料 1

形態素解析等の言語処理手法を用いた生活行動分類自動格付システムの開発

－平成 13 年社会生活基本調査－

横　内　　宏　至

N S T A C

Working Paper No. 1

平成 15 年 8 月

独立行政法人　統計センター

製表技術参考資料は，独立行政法人 統計センターの職員がその業務に関連して行った製表技術に関する研究の結果を紹介するためのものである。

本資料の内容は，原則として職員が個人として執筆しており，機関の見解を示すものではない。

形態素解析等の言語処理手法を用いた
生活行動分類自動格付システムの開発
－平成 13 年社会生活基本調査－

横内　宏至*

要　　旨

本資料は，平成 13 年社会生活基本調査で初めて導入されたアフターコーディング方式による生活時間に関する調査の集計のために開発した生活行動分類自動格付システムについてまとめたものである。

同システムは，２種類の自動格付方法を採用した。一つは，試験調査結果等を基に具体的な生活行動とこれに対応する生活行動分類符号を予め登録しておき，これと完全に一致する調査回答に対して該当符号を付与する完全一致方式，もう一つは，自然言語処理手法によりキーワードの抽出，表現のゆれの軽減処理などを行ってから，同様のマッチングを試みる単語分割方式である。

平成 13 年社会生活基本調査に適用した結果，完全一致方式で７割を超える格付件数となった。単語分割方式は，完全一致方式よりも 3 ポイント程度多く格付することができた。生活行動は，「睡眠」のように少数の出現頻度の多い行動が一定割合を占めるため，完全一致方式でも格付できる割合が高くなるが，異なりパターン数でみると，格付できた割合は完全一致方式の 6.2%に対し単語分割方式は 19.6%と向上しており，自然言語処理手法の導入による表現のゆれの軽減効果が明確に現われている。

単語分割方式による自動格付の処理時間は，完全一致方式の約５倍かかる。このため，完全一致方式で格付できなかったものについて単語分割方式を適用する併用型が望ましいと考えられる。

本資料の構成は，Ⅱで平成 13 年社会生活基本調査におけるアフターコーディング方式調査の内容，Ⅲで自動格付システムの概要，Ⅳで単語分割方式で用いた形態素解析法について簡単に触れ，Ⅴで生活時間に関する調査における単語の出現頻度など予備的考察の結果，Ⅵ，Ⅶで単語分割方式の詳細，処理の流れを説明している。Ⅷ～Ⅸでは，本調査のデータに適用した結果を示し，Ⅹで適用結果について詳細に分析している。さらに，今後の課題

についてXI に提示している。

*　統計センター研究センター
E-mail: research@nstac.go.jp

目　次

Ⅰ　背景と目的

平成 13 年社会生活基本調査における生活時間に関する調査では，従来のプレコード方式の調査票（調査票 A）に加え，アフターコード方式の調査票（調査票 B）の導入を行った。この新たな調査票には 1 日の生活行動が 15 分刻みで記入されるため，集計に際し，生活行動を記述した文字情報をデータ入力した上で生活行動分類の符号格付を行う必要があった。

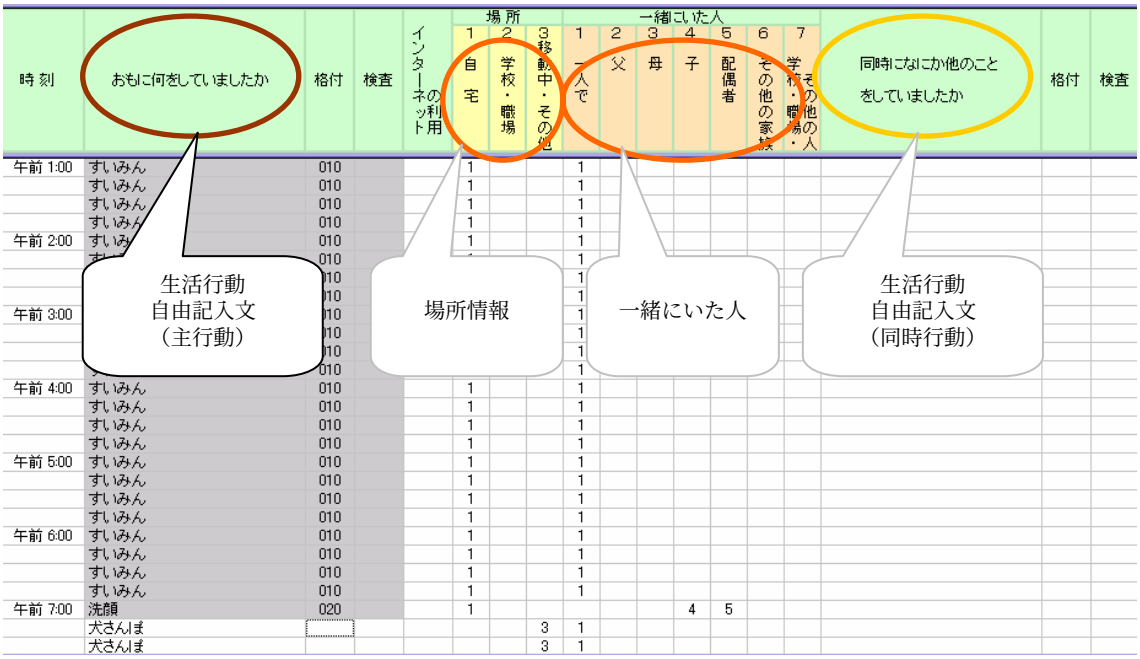
符号格付にあたって，試験調査等で格付された生活行動の文字情報をコンスタント*1 として用いて，これと完全に一致するものに生活行動分類符号を付与するシステム（「完全一致方式」）と，技術的な研究を目的として，自然言語処理手法を用いたシステム（「単語分割方式」）の開発を並行して行うこととした。

Ⅱ　社会生活基本調査　調査票 B について

1　日本語の文章と調査票について

文章の基本構造として，「いつ，だれが，どこで，なにを，だれと，どうする」が考られる。調査票 B は，図 1 のような形式となっている。今回の社会生活基本調査の調査票 B「生活時間について」において，自由記入文が正しく記入されているとすれば，「いつ」は調査日と時刻であり，「だれが」は記入者であり，「どこで」は「場所」，「だれと」は「一緒にいた人」ということになるので，自由記入文から得るべき情報としては，「なにを」「どうする」の二つであることが分かる。

図 1　調査票 B のイメージ



*1　コンスタント　・・・　文字情報と該当する生活行動分類符号を登録したデータベースの意味で用いる。

2　研究条件

今回の自動格付システムでは，自由記入文のみを利用することとした。

自由記入文のみということは，本来調査票に存在する情報の一部しか利用しないため格付に必要な情報が不足することが考えられる。

例えば，行動内容によっては「場所」や「一緒にいた人」に影響を受け，格付符号が変わってくるものが存在するので，「なにを」「どうする」のみの記入では情報が不足する，また，「なにを」「どうする」の形になっていない「学校」などの記入では，「学校」に移動中なのか，授業を受けているのか，給食を食べているのかは分からない。それを判断するには，前後の状況や「時刻」「場所」「一緒にいた人」など自由記入文以外の情報が必要になってくる。

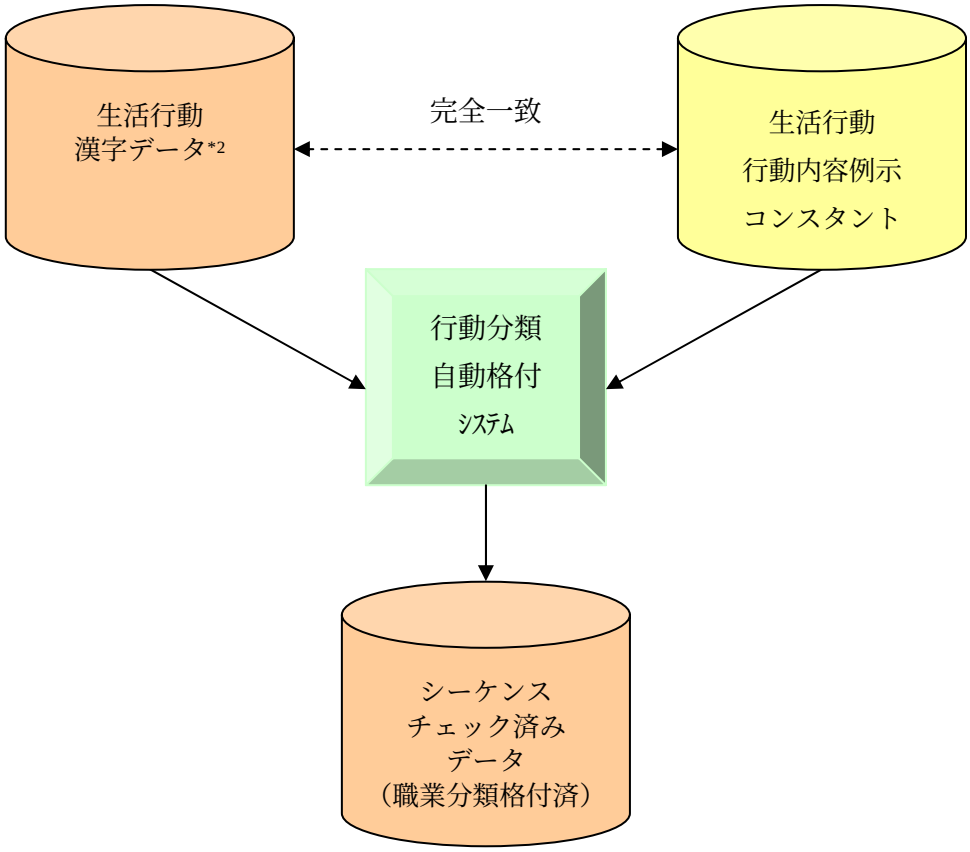
III　自動格付プログラムについて

1　完全一致方式

完全一致方式とは，自由記入文とコンスタントを「てにをは」を含め完全に一致しているかの比較を行い格付する方式である。

システムの構成を，図 2 に示す。

図 2　完全一致方式のシステム構成

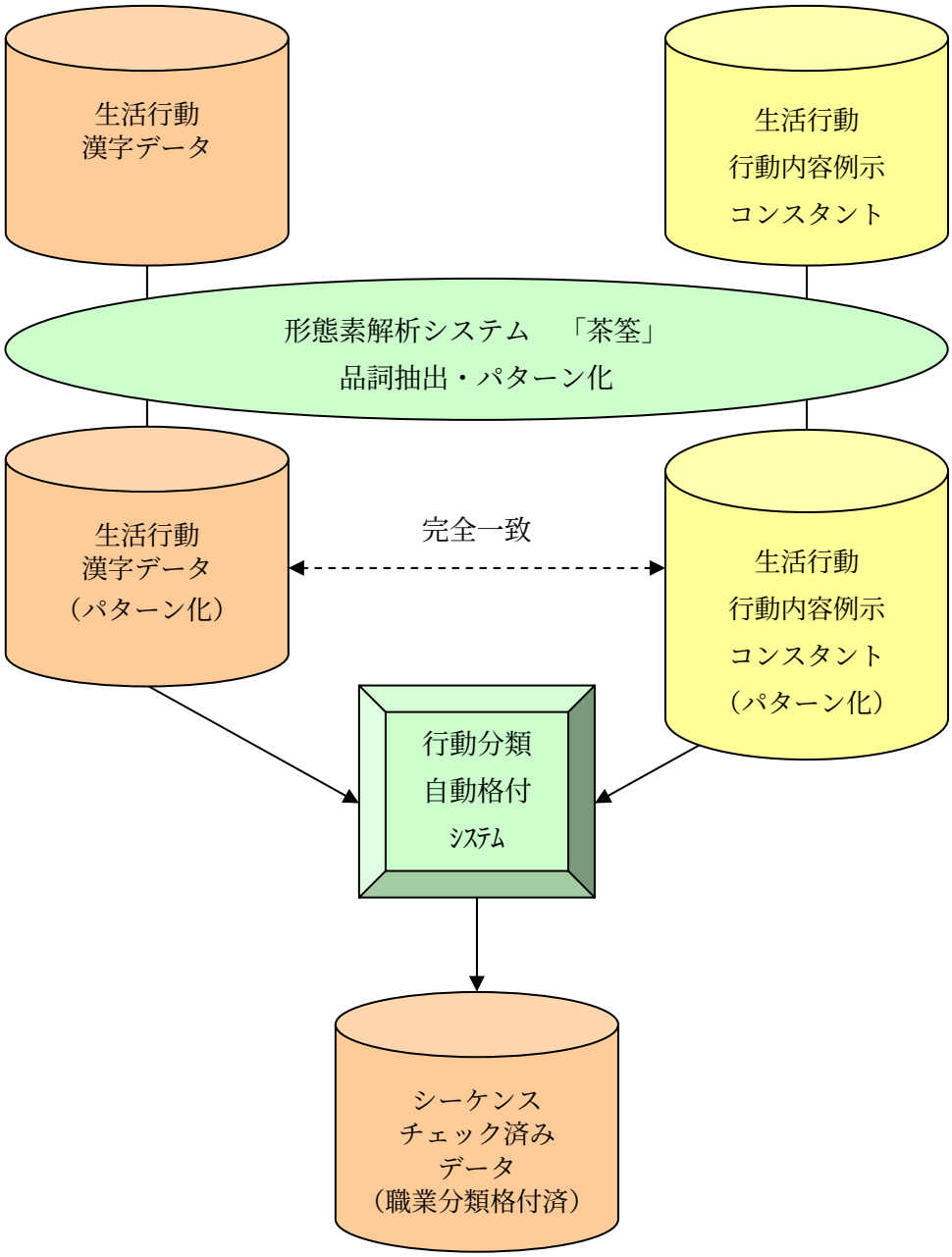


2　単語分割方式

単語分割方式とは，形態素解析などの自然言語処理（次項参照）を利用することにより，自由記入文からキーワードとなる単語を抽出して格付する方式である。

システムの構成を図 3 に示す。

図 3　単語分割方式のシステム構成



*2　生活行動　　・・・　　「主行動」,「同時行動」欄の自由記入文をテキスト化したデータ

漢字データ

IV　自然言語処理 [1]

対象となる手書きの調査票は，自由記入方式であり自然言語を取り扱うこととなる。コンピュータ側からみれば文字列は単なる記号列にすぎず，文字列を分類符号格付する場合には自然言語処理が必要となる。

自然言語処理（1996 年・岩波書店・長尾　真編）によると，自然言語は以下に示すような階層型構造をもっている。

音素	人間の意味（意思）伝達において音声をどのように使っているかを基に考えた音の単位
形態素	意味を持つ最小の言語単位。一つ以上の音素から成る。
語	一つの意味のまとまりをなし，文法上一つの機能をもつ最小の言語単位。一つ以上の形態素から成る。
文	あるまとまった内容を持ち，形の上で完結した言語単位。一つ以上の単語から成る。
文章・テキスト	あるまとまった内容を表現するための文の順序づけられた集まり。隣接する文相互間にある種の関係性が存在する。

言語解析は，この要素を解析することによりコンピュータ上で自然言語を取り扱う。
言語解析の種類としては，以下のものが挙げられる。

- 形態素解析
- 構文解析
- 意味解析

以下では今回単語分割方式で利用する形態素解析について説明する。

1　形態素解析 [2]

形態素解析とは，文を適切な形態素に分割する処理のことである。この形態素とは，意味をもつ最小の要素のことで，文はこの形態素で構成されている。例えば，「私は本を読む」という文では，「私」「は」「本」「を」「読む」という形態素に分割することで，それぞれの形態素に意味を与え，構文解析や意味解析（後述 XI章 3 項参照）などへと処理をつなげる。

英語などは，単語の間にスペースが含まれているので，比較的容易に解析が可能だが，日本語などは，そのスペースがないので解析は困難である。形態素解析の問題点としては，例えば，その解析時間の問題や適切な分割ということが挙げられる。

膨大な辞書を用意してどんな文でも解析できるようにしても，時間がかかれば非効率であり，時間がかからないからといって適切でない分割をしてしまうのは本末転倒である。この解析をより効率的に行うため，より曖昧さを減らすために，さまざまな研究が行われている。さらに，形態素解析には辞書が必要不可欠であるが，すべての語を網羅

することは不可能である。その辞書に記載されていない単語が文に含まれている場合の処理なども問題点の一つである。

2　形態素解析アルゴリズム[1]

日本語文において，2 語が連続して文中に現われるかどうかの条件を連接可能性の規則（制約）として与える。この連接可能性に従い入力文を語の並びに変化させる方法が最も基本的な形態素解析の方法である。

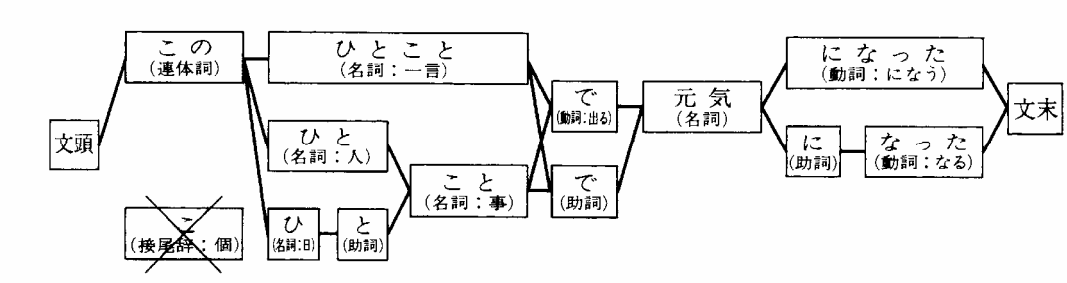
この処理は，

1．辞書を参照して入力文中の各位置から始まる語を取り出し，

2．連接可能性をチェックしながら取り出された語をつないでいく

という二つの処理を繰り返し行うことによって実現される。この処理によって得られる解析結果の例を図 4 に示す。

図 4　2 単語間の連接規則による形態素解析の結果



こ 1 の 2 ひ 3 と 4 こ 5 と 6 で 7 元 8 気 9 に 10 な 11 っ 12 た 13

3　形態素解析システム『茶筌』[3]

『茶筌』は，奈良先端科学技術大学院大学松本研究室において，計算機による日本語の解析の研究を目指す多くの研究者に共通して使える形態素解析ツールを提供するために開発された形態素解析システムである。

V　予備実験

今回の実験においては，形態素解析システム『茶筌（Chasen Version 2.0 for Windows)』を用いて，調査票 B「生活時間について」の行動において，どのような記入の偏りがあるかの分析を行った。

1　予備実験用プログラムの概要

予備実験においては，試験調査のデータを利用した。まず，主行動と同時行動の漢字データから日本語形態素解析システムにより，形態素の品詞を調べる。抽出単語分析・漢字データ分析では，品詞情報をもとに，文の意味を直接左右しないであろうと思われる品詞を削除し，単語の抽出を行う。次に，各単語の異なり出現数を計算し出現割合を求める。語の出現の偏りについて，「どのような回答（語・文）が多く出現するか」について，分析を行った。

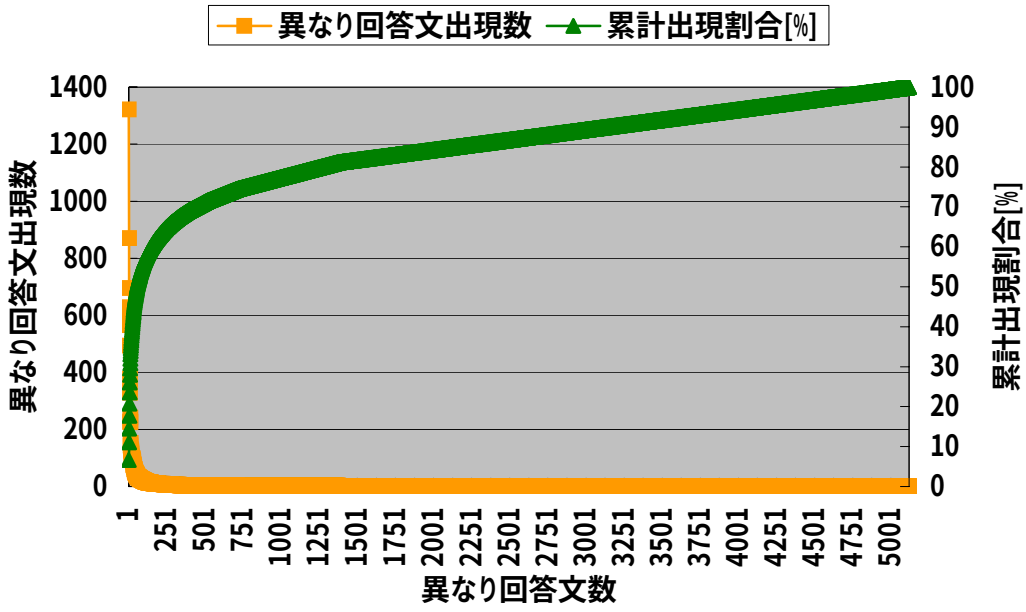
2　予備実験結果

（1）　漢字データ分析

生活行動漢字データの主行動を，異なり回答文毎に出現数を求めた結果を図 5 に示す。異なり回答文数 150 ぐらいまで急激に上昇し以降緩やかになっている。異なり回答文数 5,120 のうち出現数が 3 以上のものは 714, 2 以上のものは 1,376 であり，複数回現われない文が 3,744 で 73%という結果が得られた。

同時行動については，上昇は緩やかであるが同様の傾向が見られた。

図 5　漢字データ分析（主行動）



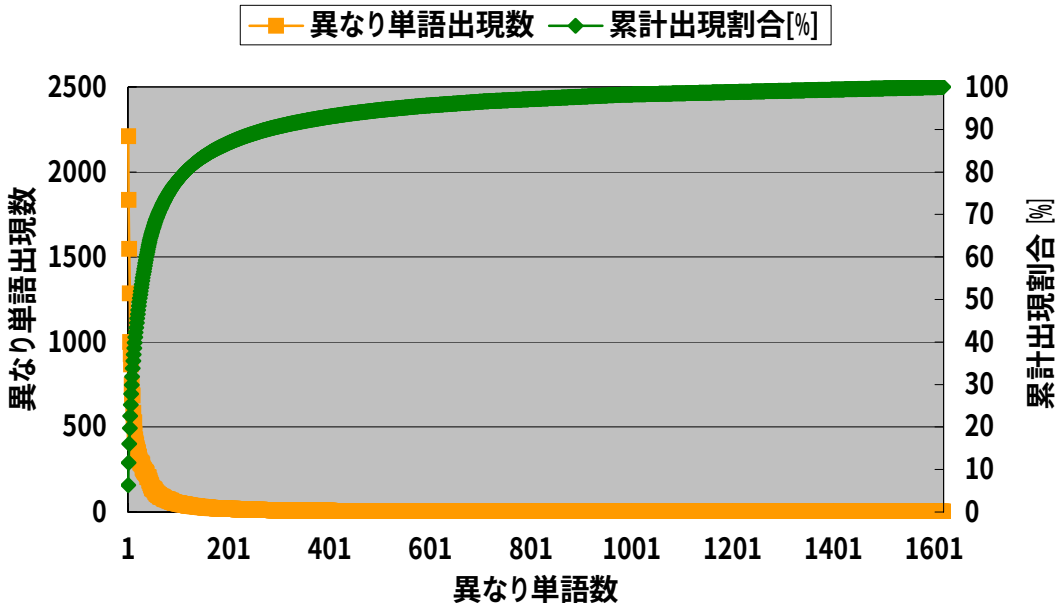
（2）　抽出単語分析

形態素解析をした結果から単語抽出を行い出現数・累計出現割合をグラフにまとめたものを図 6 に示す。異なり単語数 200 ぐらいまで急激に上昇し以降緩やかになっている。このことから，調査票に記入される単語には大きな偏りがあることが分かった。図 5 と図 6 を比較すると図 6 の方が急激な上昇となっており，頻繁に使わ

れる語とそうでない語の差が現われている。

出現数の多い単語には，「睡眠」・「テレビ」・「仕事」・「買物」・「入浴」・「食事」・「洗濯」・「勉強」などがあり，基本的な人の行動パターンが伺われる結果となっている。上位の単語の中でも，「テレビ」と「ＴＶ」，「風呂」や「フロ」などが存在し，同じ単語でも表現のゆれが生じていることが分かる。逆に，出現数の少ない単語には，基本的な人の行動パターンに当てはまらない趣味，地名や固有名詞などの表現が目に付いた。

図 6　抽出単語分析（主行動と同時行動の合計）



３　表現の統一化

形態素解析の辞書に含まれない単語は，正確な語の分割がされなかったり，未知語として処理されることになる。表現のばらつきは極力おさえて統一化し，正確に解析されるようにする必要がある。また，動詞には活用形が多く存在するので同様に表現の統一化を行い，コンピュータ上で処理しやすくする工夫が必要となる。

今回の試験データにおいて，形態素解析により正しく分割されない単語の例の一部として以下のようなものがあった。

「そうじ」「かぞく」「ふとん」「着がえる」「したく」「すいみん」「かたずけ（かたづけ）」「仕たく」「じかん」「じゅんび」「ちょうさひょう」「よう い」「だんらん」等が存在した。

前項で挙げたように，日本語には漢字・ひらがな・カタカナ，送り仮名などにより同一の言葉にも表現が複数存在することが多く，このような表現のゆれも統一化する必要がある。

VI　システムの設計

1　格付ルールの設定

（1）　品詞抽出

Ⅱ章１項で，自由記入文から得るべき情報は「なにを」「どうする」であることは述べた。「なにを」は名詞，「どうする」は動詞であると考えられる。形態素解析では，辞書に含まれない単語は未知語として取り扱われるが，専門用語や固有名詞等が多いので名詞，動詞，未知語を抽出対象とした。

予備実験において，名詞，動詞の中でも削除できる品詞や抽出されない品詞の中でも必要なものがないか検討した結果，表 1 に挙げるものが選択された。アンダーライン部は，まとめられている品詞に対応する。例えば，「新築中」は「新築（名詞-サ変接続）」と「中（名詞-接尾-副詞可能）」となっている。

品詞別に抽出の対象とするものとしないものについて表 2 にまとめた。

表１　品詞内容例

名詞-接尾-副詞可能
・新築中　・ベッド上　・就寝中　・朝食後　留守中　・仕事中　・休憩中 ・終了後　・睡眠中　・夕食後　・移動中　・帰宅中　・運転中
名詞-接続-人名
・ヘルパーさん　・芸者さん　・舞子さん　・本屋さん　・クリーニング屋さん ・歯医者さん　・看護婦さん　・姑さん
動詞-非自立
・送っていく　・連れて行く　・帰ってくる　・買ってくる　・来てくれる ・しておく
動詞-接尾
・させる　・帰られる
名詞-非自立-副詞可能
・帰宅のため　・風邪の為　・4 時頃
名詞-数
・1 時限目　・六時間目
副詞-助詞類接続
・ゴロゴロする　・のんびり　・ゆっくり　・ゆったり　・ぶらぶらする

表 2　対象とする品詞・対象としない品詞

品詞	対象としない	対象とする
名詞	名詞-接尾-副詞可能 名詞-接尾-人名 名詞-非自立-副詞可能 名詞-数	名詞
動詞	動詞-非自立 動詞-接尾 動詞-自立 （「する」・「ある」・「なる」）	動詞
副詞		副詞-助詞類接続
未知語		未知語

（2）　パターン化

漢字データに形態素解析処理を行い，単語の組を抽出する。例えば，パターン化では，「お風呂に入る」は，「風呂，入る」とする。

パターン化の際，動詞は基本形にして処理を行う。例として，「入浴。」「入浴する」「入浴した」を挙げパターン化処理について述べる。

形態素解析による解析結果

表層語，基本形，品詞で表示され形態素ごとに改行される。

入浴,入浴,名詞-サ変接続
。,。,記号-句点

入浴,入浴,名詞-サ変接続
する,する,動詞-自立

入浴,入浴,名詞-サ変接続
し,する,動詞-自立
た,た,助動詞

品詞の絞込みにより，名詞・動詞が抽出される。削除される品詞を＜＞で，動詞を［］で表現すると

「入浴。」→「入浴＜。＞」⇒「入浴」
「入浴する」→「入浴［する］」⇒「入浴［する］」
「入浴した」→「入浴［し］＜た＞」⇒「入浴［する］」

次に，品詞抽出で対象としない語に［する］が含まれていると，

「入浴。」→「入浴＜。＞」⇒「入浴」⇒　「入浴」
「入浴する」→「入浴［する］」⇒「入浴［する］」⇒「入浴」
「入浴した」→「入浴［し］＜た＞」⇒「入浴［する］」⇒「入浴」

となり，全てが等しい語としてマッチングされる。

（3）　生活行動漢字データとコンスタントのパターン化

今回の実験においては，生活行動漢字データとコンスタントのパターン化（Ⅲ章 2 項参照）ではルールを変更している。生活行動漢字データは，文章に括弧が存在した場合，括弧内は前記の事象を詳しく述べているものと見なして削除して取り扱うように処理を行っている。休憩や外出などで括弧内の内容に格付符号が左右されるものについては，括弧内を削除しないよう必要に応じてルールを見直している。

コンスタントのパターン化においては，どのような場合でも文章に括弧が存在した場合は括弧内の情報をいかすために，括弧のみを取り除いて処理を行っている。

生活行動漢字データパターン化例（括弧内の削除）

「読書をする（小説）」→「読書＜を＞[する]＜（小説）＞」
　⇒「読書」・・・括弧の内容が重要ではない：OK
「休憩（お茶を飲む）」→「休憩＜（お茶を飲む）＞」
　⇒「休憩」・・・括弧の内容が重要：NG

コンスタントパターン化例（括弧のみ削除）

「読書をする（小説）」→「読書＜を＞[する]＜（＞小説＜）＞」
　⇒「読書，小説」
「休憩（お茶を飲む）」→「休憩＜（＞お茶＜を＞飲む＜）＞」
　⇒「休憩，お茶，飲む」

（4）　日本語の語順

日本語の特徴として，語順が入れ替わることが可能であることが挙げられる。例えば，「子どもと本屋に行く」という文は「本屋に子どもと行く」のように書き換えができる。そこで，パターン化した単語を五十音順に並び替えて登録することを試みる。漢字データとコンスタントを同じようにパターン化し並べ替えておけばマッチングする可能性が高くなると思われる。

並び替え例

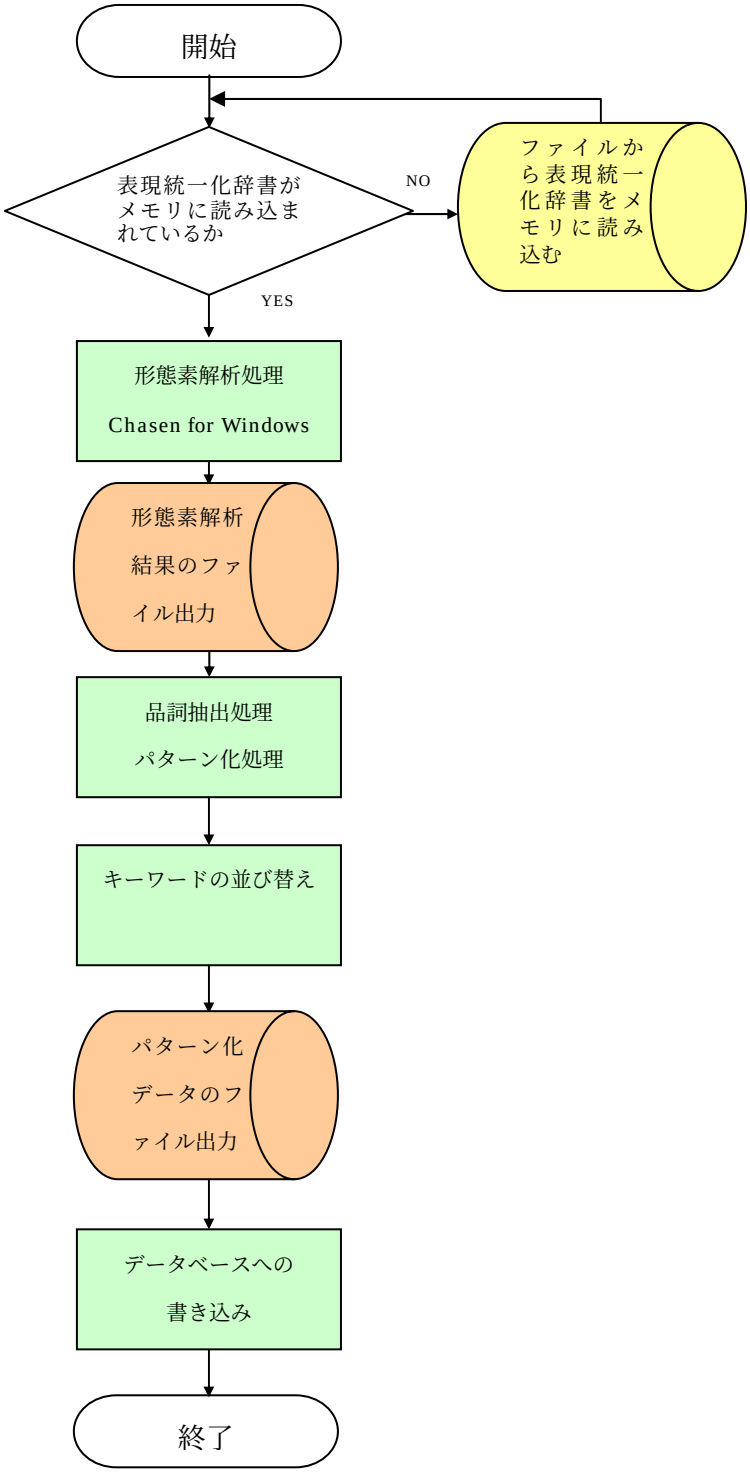
「子どもと本屋に行く」⇒「子ども，本屋，行く」
　≫「行く，子ども，本屋」
「本屋に子供と行く」⇒「本屋，子供，行く」
　≫「行く，子ども，本屋」

Ⅶ　システムの実装

1　システム処理の流れ

今回試作したプログラムのシステム処理の流れを図 7 に示す。

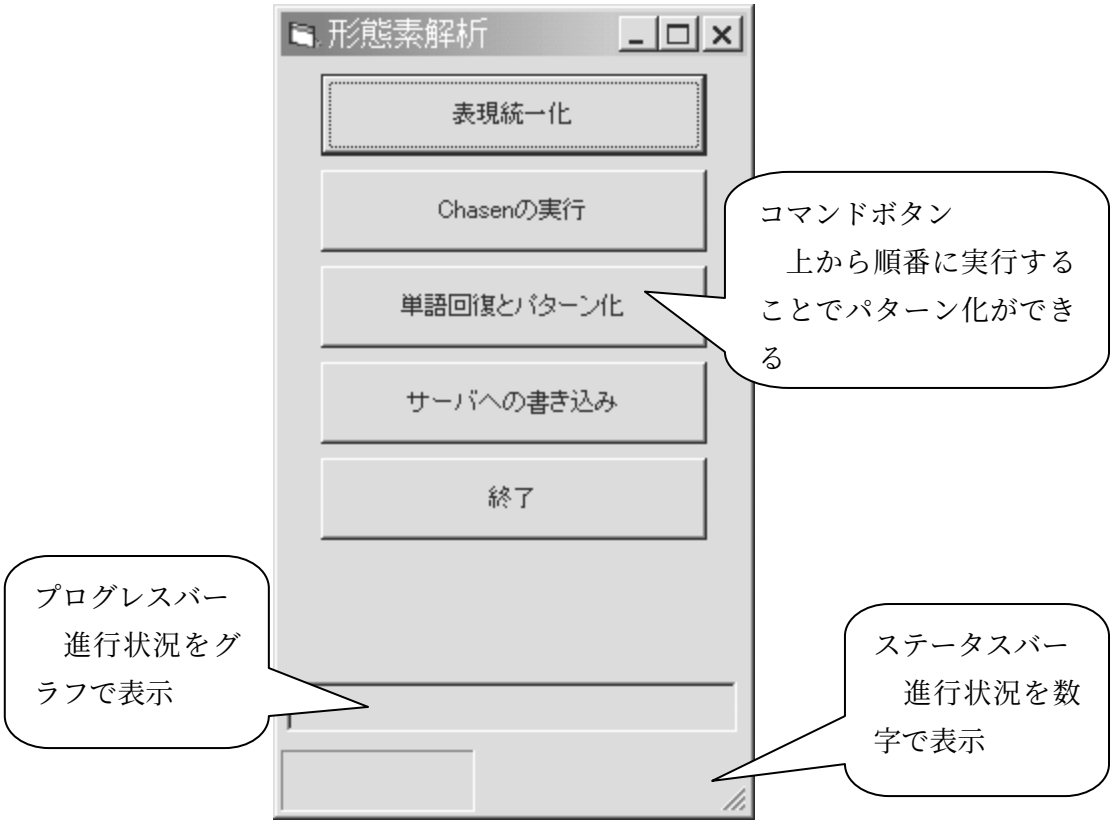
図 7　システム処理の流れ



2　パターン化プログラム画面構成

パターン化プログラムの画面構成を図 8 に示す。

図 8　パターン化プログラム



Ⅷ　プロトタイプでの結果

本調査のデータを用いて，３県（以下，A県，B県，C県と記す。）について完全一致方式と単語分割方式による自動格付状況を比較した。「単語分割でのみ格付」の欄で，A県で 4.0%，B県で 4.7%，C県で 5.7%で，単語分割方式で約 5%新たに符号格付できることが分かった。

結果を，表 3，表 4，表 5 に示す。

「総レコード数」とは，
　世帯員数×4 コマ（1 コマ：15 分刻み）×24 時間×2 日間のレコード数
「格付対象レコード数」とは，「総レコード数」から「白紙レコード数」を引いた数
「共に格付，一致」とは，完全一致方式と単語分割方式で格付符号が同じになった数
「共に格付，不一致」とは，完全一致方式と単語分割方式で格付符号が同じにならなかった数
「共に格付不能」とは，完全一致方式と単語分割方式で格付符号が共に得られなかった数（記入のないレコードを含む。）

白紙という表現は，社会生活基本調査の調査票 B において 2 日分の調査を 4 頁に分けて記入するので，その一部の頁について記入がされてない調査票を指している。記入がされていない部分については，漢字データがないので格付対象レコード数から除いている。

表 3　A県自動格付状況

調査区数	6
世帯数	77
世帯員数	198
総レコード数	38,016
格付対象レコード数	38,016
白紙世帯人員	0
白紙記入数	0

	完全一致方式と単語分割方式			
	主な行動		同時行動	
	件数	割合	件数	割合
格付対象レコード数	38,016	100.0%	38,016	100.0%
共に格付，一致	28,267	74.4%	1,591	4.2%
共に格付，不一致	2	0.0%	0	-
単語分割でのみ格付	1,514	4.0%	148	0.4%
完全一致でのみ格付	0	-	0	-
共に格付不能	8,233	21.7%	36,277	95.4%

表 4　B県自動格付状況

調査区数	7
世帯数	91
世帯員数	208
総レコード数	39,936
格付対象レコード数	38,784
白紙世帯人員	6
白紙レコード数	1,152

	完全一致方式と単語分割方式			
	主な行動		同時行動	
	件数	割合	件数	割合
格付対象レコード数	38,784	100.0%	38,784	100.0%
共に格付，一致	28,398	73.2%	1,745	4.5%
共に格付，不一致	0	-	0	-
単語分割でのみ格付	1,813	4.7%	296	0.8%
完全一致でのみ格付	0	-	0	-
共に格付不能	9,725	25.1%	37,895	97.7%

表 5　C県自動格付状況

調査区数	5
世帯数	65
世帯員数	141
総レコード数	27,072
格付対象レコード数	26,880
白紙世帯人員	1
白紙レコード数	192

	完全一致方式と単語分割方式			
	主な行動		同時行動	
	件数	割合	件数	割合
格付対象レコード数	26,880	100.0%	26,880	100.0%
共に格付，一致	19,824	73.8%	869	3.2%
共に格付，不一致	6	0.0%	0	-
単語分割でのみ格付	1,544	5.7%	62	0.2%
完全一致でのみ格付	3	0.0%	0	-
共に格付不能	5,503	20.5%	25,949	96.5%

Ｂ県については，パターン化されたデータが幾つの単語で構成されているかについて，さらに分析を行った。

パターン化回答文数では，図 9，図 10 共に単語分割数 2 のものが最も大きくなっている。

単語分割数が 6 以上のものについては，5 以下のものと比べとても少なくなっている。図 9 においては，単語分割数が上がるごとに出現割合が減っているが，図 10 では単語分割数 2 のものが最も多くなっており，異なっている。

図 9 において単語分割数 1 のものが 7 割近い割合を占めているが，図 10 においては単語分割数 2 のものが 7 割近い割合を占めている。その記入内容をみると，主行動では，「睡眠」「朝食」「昼食」「夕食」「仕事」「入浴」などの行動が多く，同時行動では，「テレビを見る」「ラジオを聞く」「新聞を読む」「ＣＤを聞く」などの行動が多く，出現頻度に偏りがみられる。

漢字データとそのパターンの上位 20 を，主行動・同時行動それぞれについて，表 6，表 7，表 8，表 9 に示す。

図 9　B県：単語分割パターン分析（主行動）

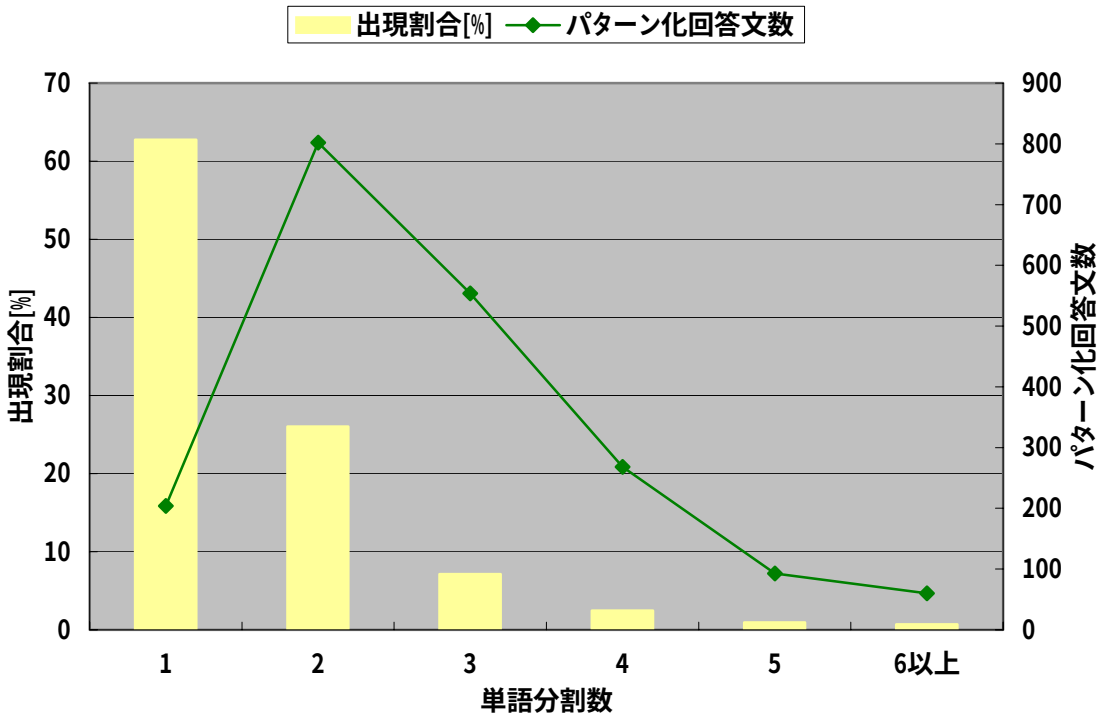


図 10　B県：単語分割パターン分析（同時行動）

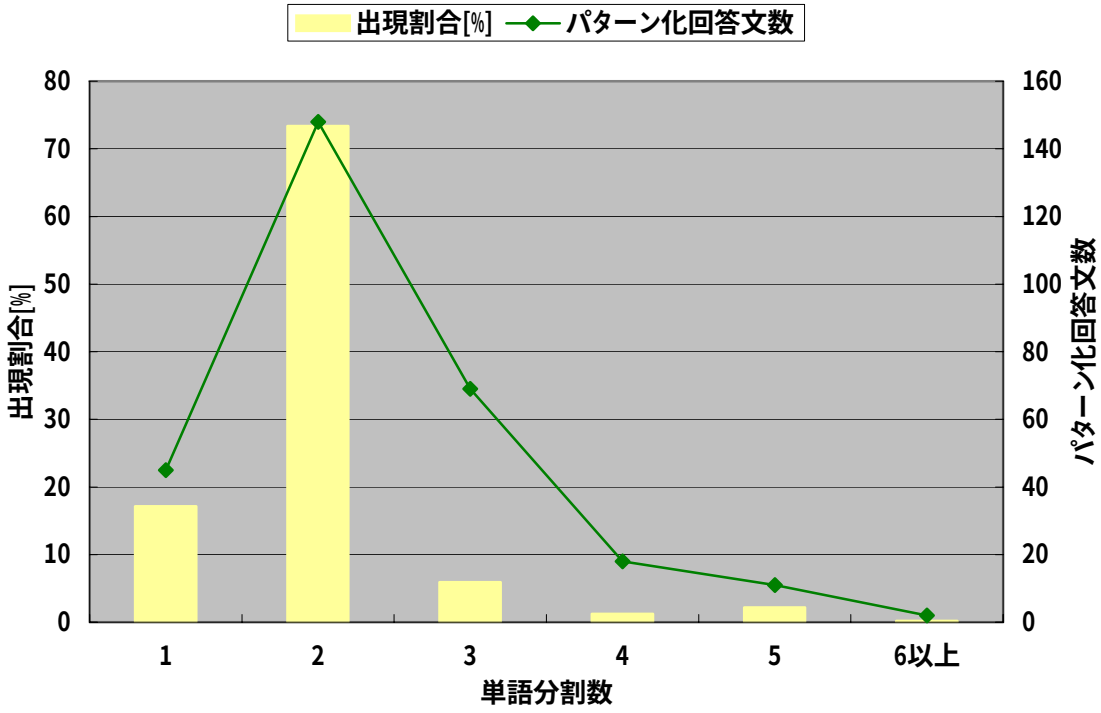


表 6　主行動の漢字データ上位 20

漢字データ	件数
睡眠	8,710
すいみん	3,264
テレビを見る	2,642
仕事	2,222
夕食	922
昼食	821
朝食	614
空欄	591
就寝	486
入浴	402
ＴＶを見る	287
寝る	262
テレビ見る	260
勉強	192
睡眠中	187
洗顔	185
休憩	165
テレビ	157
洗濯	148
夕食のしたく	147

表 7　主行動の漢字データのパターン上位 20

漢字データ	単語抽出数	件数
睡眠	1	12,161
テレビ，見る	2	3,243
仕事	1	2,375
夕食	1	971
昼食	1	850
朝食	1	634
空欄	1	591
就寝	1	534
入浴	1	416
寝る	1	347
休憩	1	311
買物	1	288
勉強	1	251
仕度，夕食	2	249
テレビ	1	239
洗濯	1	215
昼寝	1	214
みる，テレビ	2	208
入る，風呂	2	189
洗顔	1	185

表 8　同時行動の漢字データ上位 20

漢字データ	件数
テレビを見る	693
ラジオを聞く	180
店番しながら	138
テレビをみる	133
テレビ見る	80
ＴＶを見る	69
テレビ	55
みんなでおしゃべり	48
ラジオ	43
新聞を読む	42
テレビをみた	34
テレビを見ながら	32
ＴＶ	28
ＣＤを聞く	24
読書	24
ラジオ聞く	23
練習を見ながら他のお母さんとおしゃべり	18
ＴＶをみる	18
運動会のお手伝い	16
ＣＤをきく	15

表 9　同時行動の漢字データのパターン上位 20

漢字データ	単語抽出数	件数
テレビ，見る	2	894
ラジオ，聞く	2	203
みる，テレビ	2	190
店番	1	138
テレビ	1	83
おしゃべり，みんな	2	48
ラジオ	1	43
新聞，読む	2	42
聞く，ＣＤ	2	28
読書	1	28
セット，機，洗濯	3	28
おしゃべり，他，母，練習，見る	5	18
聞く，音楽	2	16
票，記入，調査	3	16
お手伝い，運動会	2	16
子供，遊ぶ	2	15
きく，ＣＤ	2	15
おしゃべり，知人	2	15
っ放し，テレビ，付ける，聞く，見る	5	14
きく，音楽	2	13

IX　実装結果

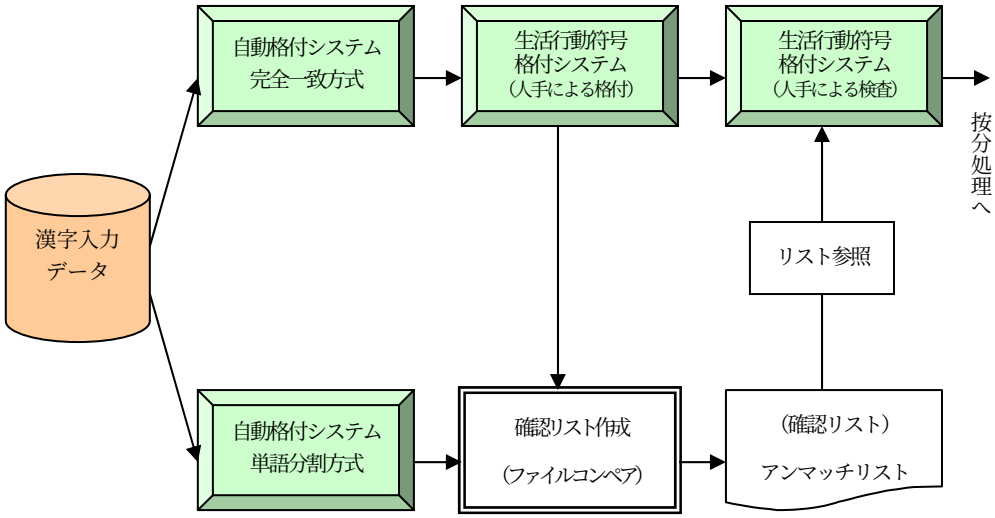
1　単語分割方式の業務適用の流れ

漢字データに対し，「完全一致方式」により自動格付を行い，「人手による符号格付」・「人手による格付検査」，「按分処理[*3]」をし，「データチェック」の処理へとつなげる。

「人手による符号格付」を行った結果と「単語分割方式」により格付された結果を比較し，確認リストを作成した。そのリストを「人手による格付検査」で参照した。

この業務の流れについて，図 11 に示す。

図 11　業務適用の流れ図



業務で使用したコンスタントが，運用中に何回か更新された。単語分割方式のコンスタントは，業務で使用したコンスタントの初期バージョンから作成したものを最後まで利用したので，実際の業務における結果を直接比較することができなかった。

このため，同じ条件で比較を行うため最終のコンスタントを用いて，完全一致方式と単語分割方式の研究用に再度自動格付プログラムを実行し，結果の比較を行った。

---

*3　按分処理　　・・・　調査票の回答において，複数の行動が含まれる場合，男女年齢階級別行動者平均時間から行動時間の時間配分を行う。
例えば，「洗顔，朝食，着替え」が 60 分の行動として記入されていた場合は，洗顔が 15 分，朝食が 30 分，着替えが 15 分といったように時間を割り当てる処理。

2　完全一致方式と単語分割方式の比較

それぞれの方式の比較に当たっては，主行動を対象として行った。

社会生活基本調査の調査票 B の世帯員数が 9,922 人 で，レコード数は 9,922 人×24 時間×4 コマ（1 コマ：15 分刻み）×2 日分=1,905,024 となる。

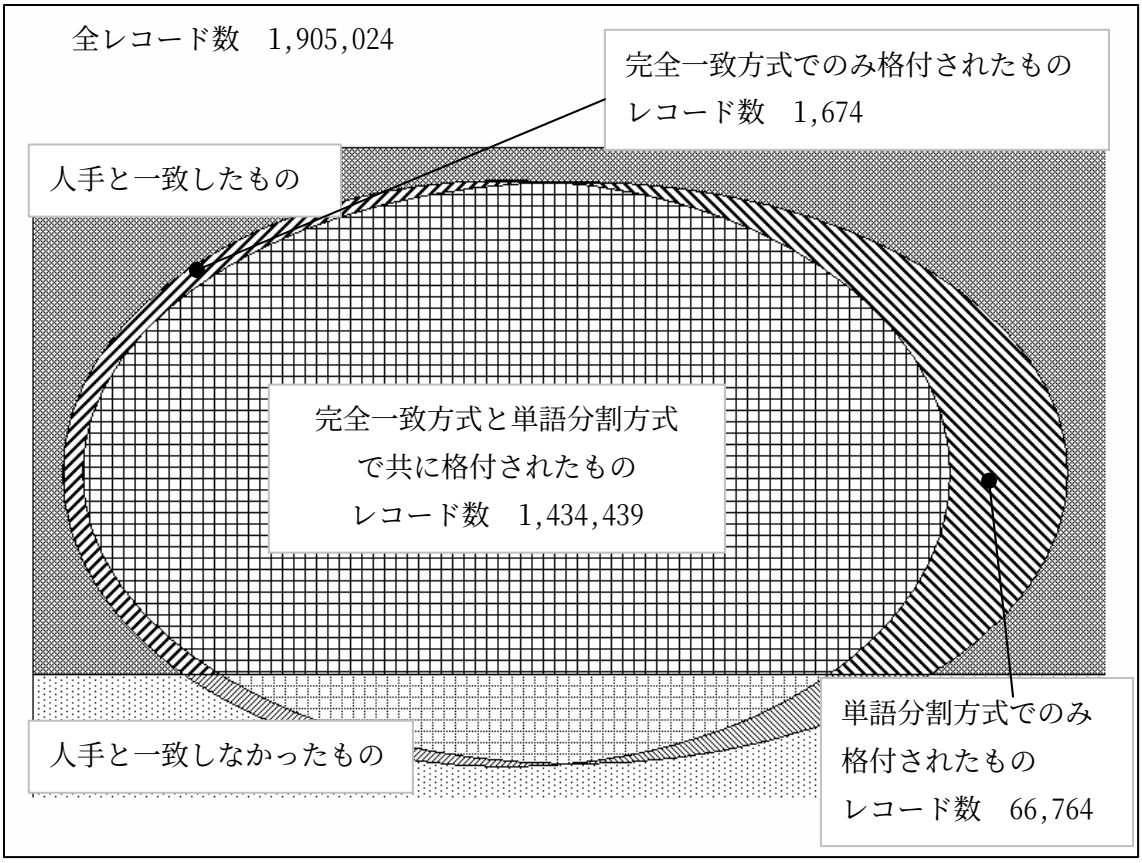
今回は，人手による格付を正としてカウントを行っている。

レコード数で結果を表示しているのは，正確な格付本数の算出が困難なためである。その理由としては，取り消し（「139　その他の休養・くつろぎ」と他の行動は，『他の行動を主行動とし，「139　その他の休養・くつろぎ」は削除して集計しない』など，主行動・同時行動に漢字データがあっても符号が取り消されることがある）や主行動と同時行動の入れ替え（「092　乳幼児の付き添い等」と他の行動は，『他の行動を主行動とし，「092　乳幼児の付き添い等」は同時行動として集計する』など，主行動と同時行動の符号が入れ替えられることがある），前後の状況による符号訂正（複数行動が含まれている漢字データで按分処理が必要なものや，前後の記入や主行動・同時行動の記入などにより同じ漢字データが連続していても異なった符号付けがされることがある）などが挙げられる。

完全一致方式及び単語分割方式と人手による格付の関係をベン図に表わしたものを図 12 に示す。

白紙のレコード数は，746（頁数）×48（調査票 1 頁あたりのレコード数）=35,808 で図 12 の白い部分であり，格付対象のレコード数は全レコード数から白紙のレコードを除いた 1,869,216 となり，人手と一致したものと人手と一致しなかったものを合計したものとなる。割合は，格付対象のレコード数を 100 として算出している。

図 12　完全一致方式と単語分割方式と人手による格付の関係



「完全一致方式による格付と人手による格付」「単語分割方式による格付と人手による格付」の件数と全レコード数に対する割合を表 10 に示す。共に格付されたものについてみてみると，「主行動，一致」の割合は，完全一致方式が 75.0%に対して，単語分割方式は 77.7%（+2.7ﾎﾟｲﾝﾄ）となっている。

同様に，「主行動，不一致」の割合は，完全一致方式が 1.8% に対して，単語分割方式は 2.6%（+0.8ﾎﾟｲﾝﾄ） となっている。

不一致は，前述の「取り消し」，「入れ替え」，「前後の状況による符号訂正」によるものが大部分を占める。

「主行動，格付」の割合の差は，+3.5ﾎﾟｲﾝﾄとなっており，プロトタイプでの結果の+4.8ﾎﾟｲﾝﾄと差があるが，これはコンスタントの追加訂正を行っており完全一致方式で符号格付される件数が増えていることが大きいと思われる。

表 10　件数と割合

都道府県	全県
世帯員数	9,922
全レコード数	1,905,024
白紙レコード数	35,808
格付対象レコード数	1,869,216

		完全一致方式と人手(A)		単語分割方式と人手(B)		差(B)-(A)	
項目名		件数	割合	件数	割合	件数	割合
主行動，格付		1,436,113	76.8%	1,501,203	80.3%	65,090	+3.5 ﾎﾟｲﾝﾄ
内訳	一致	1,401,703	75.0%	1,451,685	77.7%	49,982	+2.7 ﾎﾟｲﾝﾄ
	不一致	34,410	1.8%	49,518	2.6%	15,108	+0.8 ﾎﾟｲﾝﾄ

		完全一致方式と単語分割方式	
項目名		件数	割合
主行動，格付		1,434,439	76.7%
内訳	一致	1,432,847	76.7%
	不一致	1,592	0.1%
完全一致方式のみ格付		1,674	0.1%
単語分割方式のみ格付		66,764	3.6%

格付精度は，完全一致方式で，1,401,703／1,436,113×100=97.6%，単語分割方式で，1,451,685／1,501,203×100=96.7%となった。ここで，件数の差について注目してみると単語分割方式で，完全一致方式と比較して多く格付ができたレコード数は 65,090 であり，そのうち 49,982 レコードが一致した。

完全一致方式と比較して単語分割方式で格付精度が落ちている原因として，単語分割方式で新たに格付ができたものは，調査票の「場所」欄の記入等によって格付される符号が影響を受けるものが比較的多かったためと考えられる。

X　考察

1　人手による格付結果

「休憩」,「休養」は格付符号として「131: 仕事中，学校での学習（学業）中の休息（以下では職場・学校における休憩）」,「139: その他の休養・くつろぎ」,「191: 療養」のどれかになると思われるが，人手で格付された結果の表 11 をみてみると,「120: 新聞・雑誌」,「121:テレビ・ラジオ」,「155:ＣＤ・カセットテープ・ビデオ」のほかさまざまな符号が存在する。

休養は最も消極的な行動と捉えられるので，同時行動に積極的な行動があると，同時行動が取り消され主行動に符号が上書きされる。このような取り消しのほかに，主行動と同時行動の入れ替え，前後の状況による符号変更があり，結果を分析する上で注意が必要である。

表 11　漢字データ（主行動）が「休憩」の格付符号

格付符号	件数	割合（%）
139＜その他の休養・くつろぎ＞	**8012**	**67.9**
131＜職場・学校における休憩＞	**2229**	**18.9**
121＜テレビ・ラジオ＞	777	6.6
050＜主な仕事＞	229	1.9
130＜軽飲食＞	200	1.7
120＜新聞・雑誌＞	79	0.7
191＜療養＞	61	0.5
180＜人と会って行う交際・つきあい＞	39	0.3
010＜睡眠＞	38	0.3
154＜読書＞	18	0.2
030＜食事＞	15	0.1
132＜家族とのコミュニケーション＞	15	0.1
155＜ＣＤ，カセットテープ，ビデオ＞	14	0.1
051＜主な仕事中の移動＞	8	0.1
061＜学校の宿題＞	6	0.1
182＜電話による交際・つきあい＞	6	0.1
X＜疑義＞	6	0.1
020＜個人ケア（自分自身や家族が行うもの）＞	5	0
072＜衣類等の手入れ＞	4	0
100＜買い物＞	4	0
150＜娯楽・教養＞	4	0
153＜ゲーム＞	4	0
160＜エアロビクス系＞	4	0
9○○＜按分対象（901～996）＞	4	0
-＜取り消し＞	3	0
071＜住まいの手入れ・整理＞	3	0
091＜乳幼児と遊ぶ＞	3	0
140＜学習・研究（学業以外）＞	3	0
152＜趣味＞	3	0
252＜園芸・ペットの世話＞	3	0
201＜社会生活基本調査に関連する行動＞	2	0
075＜公的サービスの利用＞	1	0
183＜メールや手紙などを書いたり，読んだりする交際・つきあい＞	1	0
総数	1,1803	100.0

2　頻出する漢字データ

結果において,格付精度で完全一致方式は 97.6%,単語分割方式は 96.7%であった。

高い値が得られる要因として，全レコード数の約 30%を占め，格付精度の高い「睡眠」などの貢献も考えられる。

表 12 のように単語分割方式により「睡眠」のキーワードが抽出された場合，99.8%の割合で「010」の符号となったことが分かる。「010」以外の 0.02%について，同時行動の格付符号が「-＜取り消し＞」となっているものが多く，主行動の格付符号に同時行動の格付符号が上書きされていると思われる。

表 12　単語分割方式により「睡眠」のキーワードが抽出されたものの格付符号

格付符号（主行動）	同時行動	格付符号（同時行動）	件数	総数に対する割合
010＜睡眠＞			559,783	99.8%
020＜個人ケア（自分自身や家族が行うもの）＞			5	0.02%
020＜個人ケア（自分自身や家族が行うもの）＞	トイレ	-	2	
020＜個人ケア（自分自身や家族が行うもの）＞	起床	-	1	
080＜乳幼児の介護・看護＞	子供にセキ止めの薬を貼る	-	1	
090＜乳幼児の身体の世話と監督＞			4	
121＜テレビ・ラジオ＞	テレビをみる	-	1	
121＜テレビ・ラジオ＞	テレビを見る	-	1	
130＜軽飲食＞	一時起きておやつを食べる	-	1	
154＜読書＞	文庫本を読む	-	4	
154＜読書＞	夜中に目がさめ本を読む	-	5	
191＜療養＞			156	
191＜療養＞	遠赤外線にあたる	-	4	
191＜療養＞	腰が痛くなりあんま機にかかる	-	4	
191＜療養＞	湿布をはる	-	1	
201＜社会生活基本調査に関連する行動＞	アンケート	-	1	
9○○＜按分対象（901～996）＞			8	
9○○＜按分対象（901～996）＞	トイレ	020	1	
X＜疑義＞*4			667	
X＜疑義＞	15日夜半より発熱のため行動しておりません	-	1	
X＜疑義＞	オムツ交換をしてもらう	X	1	
X＜疑義＞	お風呂に入る	X	1	
X＜疑義＞	すいみん	X	48	
X＜疑義＞	トイレ	X	1	
X＜疑義＞	トイレに起きる	X	1	
X＜疑義＞	トイレに行く	X	1	
X＜疑義＞	ラジオ聞く	X	16	
X＜疑義＞	ロシアモスクワ上空から見る	X	2	
X＜疑義＞	酸素吸入	X	5	
		総数	560,727	100.0%

*4　X（疑義）　・・・　生活行動の符号格付において，状況や主行動と同時行動の関係について精査する必要のあるものについて，疑義として取り扱っている

ちなみに,「睡眠」は，全格付対象レコードに占める割合が 560,727／1,869,216 ×100=30.0%，格付精度は 559,783／560,727=99.8%となる。

次に，表 13 の上位 20 までの漢字データと件数をみてみると，頻出する漢字データに非常に偏りがある。上位 20 までの合計で 1,052,374 となり，全レコード数のうちの約 55%を占めることになる。

また，単語または単純な文章で書かれた漢字データが上位にきていることが分かる。行動としては,「睡眠」・「食事」・「仕事」・「学業」・「テレビ」・「身の回りの用事」となっており，睡眠→身の回りの用事（洗顔等）→朝食→仕事（学業）→昼食→仕事（学業）→夕食→テレビを見る／入浴→睡眠　といった基本行動を伺わせる。

3　異なりパターン数

表 10 に示されるように,単語分割方式において,新たに格付できた漢字データは,全レコードに対する割合で，3.5%にとどまる。これは,「睡眠」等の決まった行動に漢字データの分布が非常に偏っているためなので，ここでは，漢字データの異なりのパターン数（以下異なりパターン数）を考えてみる。

異なりパターン数では，漢字データにおいて,「睡眠」という表現が何レコードでてきたとしても，１とカウントする。同じ行動を表わす漢字データでも表現が異なれば，カウントする。例えば,「睡眠」「すいみん」「スイミン」はそれぞれカウントされ，三つの異なりパターン数となる。

表 13　漢字データ（主行動）上位 20

漢字データ	件数
すいみん	389783
睡眠	165181
仕事	140031
テレビを見る	107348
夕食	37159
昼食	31611
テレビ	27490
朝食	25038
就寝	23815
入浴	18219
テレビをみる	15790
食事	9259
テレビ見る	8807
授業	8551
洗顔	8120
ねる	7843
帰宅	7782
夕食のしたく	7128
勉強	6835
休憩	6584
合計	1,052,374

主行動における全世帯員の格付対象レコード数は，1,869,216 で，異なりパターン数は，65,737 となった。

完全一致方式のコンスタントに登録してあるパターン数は，7,292 で実際に利用されたのは 4,076 となった。単語分割方式のコンスタントに登録してある単語の組のパターン数は，5,225 で実際に利用されたのは 2,972 であった。

完全一致方式のコンスタントからキーワードを抽出することでパターン化を行っているので，下の表のように単語分割方式のコンスタントのパターン数は少なくなる。

完全一致方式コンスタント の漢字データ		単語分割方式コンスタント のパターン化後
洗濯	→	洗濯
洗たくする		
せんたくをする		
洗濯をする		
・・・		・・・
パターン数 7,292(4,076)		パターン数 5,225(2,972)

単語分割方式では，キーワードでのマッチングなので完全一致方式と比べ多くの漢字データと対応している。以下は「洗濯」についての例であるが，全体についてみると完全一致方式で異なりパターン数 4,076，単語分割方式で異なりパターン数 12,889 となる。

漢字データ	完全一致方式	単語分割方式
おせんたく		○
せんたく		○
せんたくした		○
せんたくする		○
せんたくなど		○
せんたくをする	○	○
洗たく	○	○
洗たくしかける		○
洗たくする		○
洗濯	○	○
洗濯する		○
洗濯をした		○
洗濯をする	○	○
洗濯中		○
	・・・	・・・
	4,076	12,889

単語分割方式は完全一致方式の 3 倍以上（12,889／4,076　3.2）の異なりパターン数となる。したがって，表現のゆれを解消できたことが分かる。

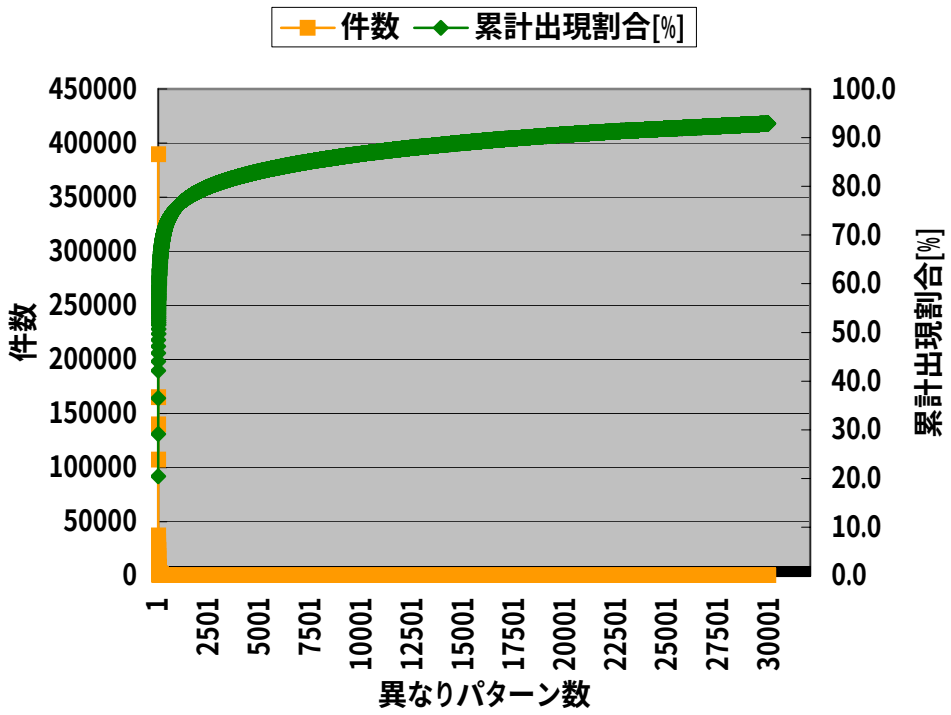
4　完全一致方式，単語分割方式の長所と短所

（1）　完全一致方式の長所と短所

完全一致方式の長所として，コンスタントに登録されたデータと漢字データを比較し「てにをは」や記号を含め，全く同じ物だけを格付するという明確なルールで，分かりやすい。また，システムも単純なものでよく，プログラムの実行時間も少なくてすむ。

短所として，語順や「てにをは」の違い，動詞の活用などを考慮して全ての組合せをコンスタントに登録することは難しいと考えられ，格付率向上に限界があると思われる。今回得られた異なりパターン数 65,739 をそのままコンスタントに利用したとしても，次回どれだけが利用されるかは疑問である。前述したとおり，試験調査等を基に作成された完全一致方式のコンスタントに登録してある異なりパターン数が，7,292 で実際に利用されたのは 4,076 で，登録した 44.1%のコンスタントが利用されなかった。

図 13　件数の多いものからソートした異なりパターン数（件数 3 以上）



異なりパターン 65,739 のうち，出現頻度の高いものから何番目で累積割合がどのくらいになるかをみてみると，10 で 50%，42 で 60%，253 で 70%，757 で 75%，2,488 で 80%，7,219 で 85%，17,751 で 90%となっている。図 13 で異なりパターン 2000 付近で急激に曲がっていることから，出現する漢字パターンには非常に偏りがあることが分かる。

（2）　単語分割方式の長所と短所

単語分割方式の長所として，キーワードによりパターン化を行うのでコンスタントのレコード数を少なくでき，語順や「てにをは」の違い，動詞の活用など表現のゆれを解消することができる。単語単位で自由に加工することが可能（語順の入れ替えやシソーラスの活用が可能）で，重要語[*5]や共起確率[*6]を導入することで格付率向上が期待できる。単語分割方式を利用することにより期待される格付の改善方法については，XI の章で後述する。

短所として，完全一致方式に比べシステムが複雑なため，プログラムの実行時間が多くかかる。完全一致方式 47 県分の格付では，11 時間 34 分 54 秒で自動格付を行うことができ，1 県の平均 14 分 47 秒である。単語分割方式は概ね 5 倍程度であるので，1 時間強ということになる。

プログラムの実行には人手がかからないので，プログラムの書き方やパソコンの性能の向上等で時間の短縮は可能である。

---

*5　重要語　　・・・　単語分割方式による単語抽出では，文章を特徴付ける単語を漏れなく抽出することが必要となる。格付符号と文章の関係にどれだけその単語が密接に関わっているかについて重み付けを行い，重要度の高い単語を重要語として取り扱う。

*6　共起確率　　・・・　漢字データの単独の語にだけ注目するのではなく，どの語とどの語が組合せとして現われるのかに注目し，分析の結果求められる確率。
例えば，「風呂」という語は「入る」「掃除」「わかす」などの語と共起しやすい。

XI　今後の課題

1　格付精度の向上

（1）　単語分割方式用コンスタント

今回単語分割方式で利用したコンスタントは，完全一致方式のコンスタントからプログラムにより作成した。人手によらずプログラムで作成したことにより，次に挙げるパターン化の不具合を内包していた。

完全一致方式のコンスタントには，ほぼ同じ意味の文章でも「てにをは」や語順などを考慮し登録する必要がある。完全一致方式のコンスタントは 7,292 レコードであったが，パターン化後の単語分割方式のコンスタントは 5,225 レコードとなった。

パターン化前とパターン化後の差（7,292-5,225=2,067）には文意が異なり符号も同一にならないのに，同じパターンとなるもの（子どもと食事（食事）・子どもの食事（家事）等→「子ども，食事」）が存在した。このケースは想定外であり，システムの仕様により今回は，一番初めに読み込まれたパターンの符号が適応された。このことは，格付精度を下げる要因となる。

文の係り受けも単語分割方式のパターン化で問題として挙がってくる。「学校へ行く準備」を，パターン化すると「学校，行く，準備」となり，パターン化後の結果からでは，係り受けの関係の微妙なニュアンスが無視されてしまう。「学校の準備をして行く」「準備して学校へ行く」などのパターンが考えられ，文意が「準備をしているのか」,「学校へ向かっているのか」が曖昧になる。

また,「ひらがな（あいろんかけ等）・カタカナ（ヒゲソリ等）」の漢字データは，形態素解析処理で正しく単語分割されずキーワードが取得できないために，単語分割方式では格付ができないケースがあった。

「お休み」という言葉には,「お休みなさい」等の睡眠を表わす際に用いられる場合と,「お休みです」等の休暇・欠席を表わす場合があり，これは単語の意味的多義で格付の際には注意が必要であることが分かった。

格付精度向上のため，次回は，単語分割方式の格付ルール，単語分割方式用コンスタントの整備が必要である。

（2）　人の判断に近づけるために

今回は，漢字データのみで符号を与えることを試みたが，漢字データ以外の情報（場所・一緒にいた人・時刻・他の世帯員等や，生活時間の項目以外）も活用することで，格付精度の向上が期待できる。

先に例を挙げた「休憩」,「休養」について考えると，場所の情報で,「自宅」,「学校・職場」，が記入されているので，自宅ならば「139：その他の休養・くつろぎ」が，学校・職場であれば「131：職場・学校における休憩」の確率が高いと考えられる。

表 14，表 15 の「131」「139」の符号に注目してみると，表 14 において,「131」は 6.2%,「139」は 80.2%となっており，表 15 において,「131」は 45.7%,「139」は 39.9%となっており，場所が「自宅」であるか「学校・職場」であるかに大きく影響されていることが分かる。

表 14　漢字データ（主行動）が「休憩」で，場所が「自宅」のもの

格付符号	件数	割合（%）
139＜その他の休養・くつろぎ＞	**5693**	**80.2**
121＜テレビ・ラジオ＞	677	9.5
131＜職場・学校における休憩＞	**437**	**6.2**
120＜新聞・雑誌＞	62	0.9
130＜軽飲食＞	61	0.9
191＜療養＞	61	0.9
010＜睡眠＞	30	0.4
154＜読書＞	15	0.2
155＜ＣＤ，カセットテープ，ビデオ＞	12	0.2
132＜家族とのコミュニケーション＞	11	0.2
050＜主な仕事＞	5	0.1
072＜衣類等の手入れ＞	4	0.1
180＜人と会って行う交際・つきあい＞	4	0.1
-＜取り消し＞	3	0
020＜個人ケア（自分自身や家族が行うもの）＞	3	0
091＜乳幼児と遊ぶ＞	3	0
140＜学習・研究（学業以外）＞	3	0
153＜ゲーム＞	3	0
252＜園芸・ペットの世話＞	3	0
071＜住まいの手入れ・整理＞	2	0
201＜社会生活基本調査に関連する行動＞	2	0
9○○＜按分対象（901～996）＞	2	0
030＜食事＞	1	0
152＜趣味＞	1	0
182＜電話による交際・つきあい＞	1	0

表 14-2　同時行動の記入のないレコードについて

格付符号	件数	割合（%）
139＜その他の休養・くつろぎ＞	**5008**	**90.4**
131＜職場・学校における休憩＞	**413**	**7.5**
191＜療養＞	59	1.1
121＜テレビ・ラジオ＞	23	0.4
154＜読書＞	11	0.2
010＜睡眠＞	9	0.2
130＜軽飲食＞	7	0.1
050＜主な仕事＞	4	0.1
140＜学習・研究（学業以外）＞	3	0.1
020＜個人ケア（自分自身や家族が行うもの）＞	1	0.0

表 15　漢字データ（主行動）が「休憩」で，場所が「学校・職場」のもの

格付符号	件数	割合（%）
131＜職場・学校における休憩＞	**1601**	**45.7**
139＜その他の休養・くつろぎ＞	**1399**	**39.9**
050＜主な仕事＞	203	5.8
130＜軽飲食＞	112	3.2
121＜テレビ・ラジオ＞	78	2.2
180＜人と会って行う交際・つきあい＞	29	0.8
120＜新聞・雑誌＞	17	0.5
030＜食事＞	14	0.4
051＜主な仕事中の移動＞	8	0.2
010＜睡眠＞	8	0.2
061＜学校の宿題＞	6	0.2
X＜疑義＞	6	0.2
182＜電話による交際・つきあい＞	5	0.1
160＜エアロビクス系＞	4	0.1
020＜個人ケア（自分自身や家族が行うもの）＞	2	0.1
100＜買い物＞	2	0.1
152＜趣味＞	2	0.1
154＜読書＞	2	0.1
9○○＜按分対象（901～996）＞	2	0.1
155＜ＣＤ，カセットテープ，ビデオ＞	1	0
183＜メールや手紙などを書いたり， 読んだりする交際・つきあい＞	1	0

表 15-2　同時行動の記入のないレコードについて

格付符号	件数	割合（%）
131＜職場・学校における休憩＞	**1549**	**52.7**
139＜その他の休養・くつろぎ＞	**1226**	**41.7**
050＜主な仕事＞	148	5.0
051＜主な仕事中の移動＞	8	0.3
130＜軽飲食＞	5	0.2
121＜テレビ・ラジオ＞	2	0.1
180＜人と会って行う交際・つきあい＞	1	0.0

データチェックにおいて，漢字データが「休憩」で，場所が「学校・職場」の場合は，要確認データとして抽出しているので，データチェック済み後は「131」の割合が増えている可能性が高い。

このことから，条件について詳しく分析し格付ルールを定めることによって，格付精度の向上が期待できる。自動格付システムの格付率が向上したとしても，格付精度が犠牲になってしまったのでは，検査等に手間が発生し業務軽減には結びつかない。逆に，格付精度が高い水準にあれば，符号の入れ替えなどの処理も含めて自動化できると思われる。

（3）　構文解析と意味解析

今回，自然言語処理で用いた形態素解析よりも，高度な解析システムが存在する。この分野は，現在も盛んに研究が行われ進歩の速い部分であり今後の導入等も考慮に入れ更なる研究が必要と思われる。

ア　構文解析 [1]

自然言語処理（1996 年・岩波書店・長尾　真編）によれば,「構文解析とは，文法規則（制約）および種々の優先規則によって入力文を解析し，その構造を明らかにすることである。文の構造は多くの場合，意味的・文脈的情報がなければ一意に決定することはできない。そこで構文解析には二つの状況がありえる。一つは構文解析を意味解析・文脈解析の前処理と考えて，あらゆる可能な構造をつくり出すことを目的とする場合である。この場合には，一文に対して途中あるいは最終的に得られる構造が膨大なものになる可能性があるので，これをどう扱うかが問題となる。もう一つの立場は，構文解析の段階でできるだけ解を絞り込み，場合によっては一意の解を得ようという立場で，この場合にはどのような優先規則を用いればよいかということが問題となる。」と説明されている。

イ　意味解析[4]

情報検索と言語処理（1999 年・東京大学出版・徳永　健伸著）によれば，「語と語の文法的関係を手がかりにして，意味解析では文全体の意味構造を抽出する。意味構造を決定するためには個々の語が表わす概念を同定する処理と，それらの概念同士の関係を同定する処理が必要となる。このためには語が表わす概念としてどのようなものを認定するかを明確にしなければならない。このような情報を整理したものは，概念体系あるいはオントロジーと呼ばれる。」と説明されている。

高度な自然言語処理システムは，翻訳や要約，テキストの分類等に利用されるので民間各社で研究・開発が行われている。

手軽に使えるシステムではないが語同士の係り受け（主格，目的格，所有格など）等の情報を解析できるので，格付精度の向上が期待される。

2　格付率の向上

（1）　表現の統一化

コンピュータ側からみれば文字列は単なる記号列にすぎない。例えば，「すいみん」と「スイミン」で異なった語として処理される。これは，自動格付処理では不都合があるので，「すいみん」と「スイミン」は語を統一して表現する必要がでてくる。

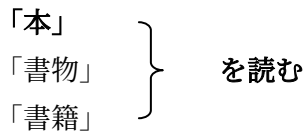
漢字データには，送り仮名等のゆれが存在するので形態素解析で取り扱いやすいように文章の自動校正がなされるべきである。そこから，同義語・類義語（シソーラス）を利用した処理を行い，表現の統一化を行う。

例えば，

「本を読む」を形態素解析して，「本」「を」「読む」となり，

「本」＝「書物」「書籍」等を関係付けるシソーラスを活用することにより，

「書物を読む」「書籍を読む」といった表現も「本を読む」に統一することが可能となり，格付率の向上が期待できる。



一般的に，同義語の中でも，「本」と「書物」，「単語と語」などはほとんどどのような文の中で使われても置換可能であり，「会合と集会」，「社会と世の中」，「自書と自筆」などは，場合によっては置き換えると不都合が生じることが知られているので，表現の統一化を行う際には注意が必要である。

（2）　重要語

より重要なキーワード（重要語）を抽出し，その語と他の単語の組合せにより与える符号の絞込みを行うような手法を提案する。

例えば,「風呂」が重要なキーワードとすると，組み合わさる「入る（身の回りの用事)」や「洗う（家事)」などにより与えられる符号が特定されると考えられる。

重要なキーワードの例として，一つの語でも,「洗濯」は「072」と高い関係を持つ。表 16 の「072」は 97.9%であり，表 17 についても 82.8%と高い数値を示す。

表 16　単語分割方式の抽出単語「洗濯」のみ

格付符号	件数	割合（%）
072＜衣類等の手入れ＞	**6971**	**97.9**
X＜疑義＞	130	1.8
9○○＜按分対象（901～996）＞	9	0.1
050＜主な仕事＞	6	0.1
076＜商業的サービスの利用＞	4	0.1
010＜睡眠＞	3	0
071＜住まいの手入れ・整理＞	1	0

表 17　単語分割方式の抽出単語「洗濯」を含み複数の単語のもの

格付符号	件数	割合（%）
072＜衣類等の手入れ＞	**10782**	**82.8**
9○○＜按分対象（901～996）＞	1540	11.8
X＜疑義＞	524	4.0
071＜住まいの手入れ・整理＞	67	0.5
050＜主な仕事＞	34	0.3
070＜食事の管理＞	25	0.2
076＜商業的サービスの利用＞	11	0.1
020＜個人ケア（自分自身や家族が行うもの）＞	10	0.1
110＜家事関連に伴う移動＞	7	0.1
121＜テレビ・ラジオ＞	5	0
030＜個人ケア（個人サービスの利用）＞	4	0
081＜乳幼児以外の家族の介護・看護＞	4	0
010＜睡眠＞	2	0
090＜乳幼児の身体の世話と監督＞	2	0
111＜家事関連以外の無償労働に伴う移動＞	2	0
130＜軽飲食＞	2	0
202＜礼拝・読経等＞	2	0
040＜通勤・通学＞	1	0
091＜乳幼児と遊ぶ＞	1	0
119＜その他の移動＞	1	0
132＜家族とのコミュニケーション＞	1	0
150＜娯楽・教養＞	1	0
201＜社会生活基本調査に関連する行動＞	1	0

自動格付の利用形態としては，現在のところ単語分割方式では，語の係り受け等の細かいニュアンスまで拾い上げることが難しいので，まず完全一致方式で格付を行い，格付されなかったものについて，単語分割方式を適用することが望ましい。パターン化したものと完全一致させ，さらに格付できなかったものについてパターンとの部分一致を行う，次に重要語の部分一致を適用することにより，格付率向上が期待できる。

例えば，

子供と洗濯物を干す	完全一致方式
↓	×洗濯物を干す
子供，洗濯，物，干す	単語分割方式
↓	×洗濯，物，干す
子供，**洗濯，物，干す**	単語分割方式（部分）
↓	○洗濯，物，干す
子供，**洗濯**，物，干す	単語分割方式（重要語）
	○洗濯

小↑曖昧性↓大

3　記入方法の整備

（1）　自動格付が難しい記入について

Ⅱ章 1 項で述べたように，自由記入文から得るべき情報としては，最低限「なにを」「どうする」があり，正しく自動格付するには，例えば，「子供と食事」は「子供と食事（をとる)」,「子供の食事」は「子供の食事（を準備する)」等「どうする」の部分が省略されないことが望ましい。

15 分刻みの生活行動を記入してもらっているが,「起床」などの行動は瞬間を表す表現で, 15 分継続する行動ではないので判断が難しい。また,「帰宅」は「会社から出た瞬間」「家に帰ってきた瞬間」「家に向かっている間」などさまざまな場合に用いられる。瞬間を表す表現で用いられた場合には「起床」と同じく注意が必要である。

人は，経験や知識から知らず知らずのうちに言葉を補完して理解しているが, コンピュータにはできない。

また,「学校へ」という表現が半日程度続く記入があったが，これは,「学校へ行っていて自宅にはいないという状態」を記入しているものと思われる。「学校でなにをしているのか」の詳細が記入されていないので，原則本人記入ということになっているが，親が記入している場合も考えられる。

「子供が帰ってくる」「主人が帰宅」など，調査票記入者が行動主体ではないものは，格付対象となる情報が存在せず人手でも格付は難しいと思われる。

（2）　格付の必要条件

前述のように，漢字データの記入にはさまざまなバリエーションが存在する。格付に必要な情報について，まとめると表 18 のようになると，考えられる。

表 18　漢字データと格付の条件

漢字データに含まれるキーワードのみで格付が可能なもの					単純↑　条件　↓複雑
漢字データに含まれるキーワードのみでは格付が難しいもの					
	漢字データのみで格付が可能なもの				
	漢字データのみでは格付が難しいもの				
		「場所」「一緒にいた人」等の情報を用いて格付が可能なもの			
		「場所」「一緒にいた人」等の情報を用いても格付が難しいもの			
			人手で格付が可能なもの		
			人手でも格付が難しいもの		

ア　漢字データに含まれるキーワードのみで格付が可能なもの

「なにを」「どうする」の「なにを」が重要。

「洗濯物を」　+「干す」　072＜衣類等の手入れ＞
　　+「取り込む」　072＜衣類等の手入れ＞
　　+「たたむ」　072＜衣類等の手入れ＞

イ　漢字データのみで格付が可能なもの

「なにを」「どうする」の両方が揃うことで，分類可能。

「車で／を」　+「ドライブ」　156＜ドライブ＞
　　+「掃除する」　074＜乗り物の手入れ＞
　　+「買う」　100＜買い物＞

ウ　「場所」「一緒にいた人」等の情報を用いて格付が可能な場合があるもの

「なにを」「どうする」のほかに，「場所」等の情報が関係する。

「休憩（をとる）」+「学校・職場」　131＜職場・学校における休憩＞
　　+「自宅」　139＜その他の休養・くつろぎ＞

エ　人手で格付が可能なもの

漢字データのみでは情報が不足しているが，後の状況や他の世帯員から情報を補完することが可能なもの

「子供が帰ってくる」
「主人が帰宅」　｝　記入者の行動は不明

両親が「新宿へ」という記入があったとする。この時点では，「新宿へなにをしに行った」かは，不明。子どもに「新宿で食事」，「一緒にいた人」に「親」の記入があった場合に，両親も，子どもと一緒に行動していたと判断し，詳しい記入のあった子どもから情報を補完。

オ　人手でも格付が難しいもの

漢字データで情報が不足しており，前後の状況や他の世帯員から情報を補完することが不可能なもの

「パソコン」　インターネットの閲覧・ゲーム？
　　メール？
　　インターネットでショッピング？

「パソコンを使ってなにをしていたか」が詳細に書かれていないために，格付が困難。

今回は，漢字データのみを対象として自動格付を試みた。格付が可能な場合について分析を行い，コンスタントを含めた格付ルールの見直しが必要である。

（3）　自動格付システムにおける漢字データ

情報の過不足がない漢字データは，自動格付システムに適している。

例えば，

情報が過剰なもの

洗濯物を取り込もうと思ったが昼寝→昼寝

ベッドで寝る→睡眠

簡潔に過不足のない記入がされることで表現のゆれも軽減する。

情報が不足しているもの

パソコン　・・・　ネットサーフィン・ゲーム？
　　メール？
　　インターネットでショッピング？

学校　　・・・　授業？
　　休み時間？
　　給食？

パソコンで「なにをしていたか」，学校で「なにをしていたか」の部分が不足している。

4　業務の軽減

（1）　コンスタントの自動生成

今回得られた結果を基に，漢字データに対する符号の関係分布をプログラム的に解析し，最も関係の強い格付精度 100%～95%：コンスタント 1，95%～70%：コンスタント 2 など（例えば先に挙げた，「睡眠」は 99.8%となるので，コンスタント 1 に登録）のように，コンスタント 1 の場合には，符号格付・格付検査を行わないなどと決めることによって人手による業務が軽減される。

単語分割方式で，パターン化したデータに対しても同様である。

（2）　単語分割方式のコストと実際の業務

自動格付システムにおいては，漢字データがテキスト化されていることが前提となっているが，実際の業務では，手書きの調査票から人手で入力するなどのテキスト形式に変換するためにコストがかかっている。将来的に OCR 技術[*7]の向上やインターネットを利用した調査を行うことなどによって，テキスト化のコストを少ないものにできると考えられる。

5　自然言語処理技術の適用範囲の拡大

今回は，社会生活基本調査の調査票 B に適用を行った。自然言語処理の技術は，産業分類自動格付システム[6]でも利用されており，入力コスト等の問題がクリアされるならば自由記入である職業分類や家計調査の収支項目分類などの格付に応用できるものと考えられる。

XII　結論

言語処理手法を用いた生活行動分類自動格付システムの開発を行った結果，「睡眠」等の出現頻度の高いものが存在し，完全一致でも 7 割を超える格付件数となることが分かった。単語分割方式を用いることにより表現のゆれを軽減し，異なりパターン数での格付を 6.2%から 19.6%に改善することができた。単語分割方式において，格付率・格付精度を向上させることが可能であると考えられる。また，完全一致方式に比べ単語分割方式で自動格付に要する時間は，約 5 倍となるのが分かった。

以上のことから，表現のゆれが少ない「睡眠」「入浴」「テレビをみる」等の漢字データに対しては，格付時間が短くて済む完全一致方式を用い，完全一致方式で格付できなかったものについて，単語分割方式を適用する併用型が，より望ましいと考えられる。

---

*7 OCR 技術　・・・　OCR とは，optical character recognition／reader：光学式文字認識／読取装置のこと[5]で，スキャナなどを用いて得られた画像から，文字部分を切り出して文字を認識し，テキストデータ等の再利用可能なデータを得ること。従来は明朝体やゴシック体など特定のフォントで印刷された文字を対象とした製品がほとんどであった。最近になって，パソコンの性能が向上し手書き文字に対応する製品も供給されるようになってきた。

XII 参考文献

[1]　自然言語処理　長尾　真編　岩波書店　1996 年

[2]　東京農工大学　西村小谷研究室
(http://www.tuat.ac.jp/~nisimura/Research/Gengo/Process/keitaiso.html)

[3]　日本語形態素解析システム『茶筌』　Version 2.0 使用説明書　第二版
奈良先端科学技術大学院大学　松本研究室

[4]　情報検索と言語処理　徳永　健伸著　東京大学出版　1999 年

[5]　アスキーデジタル用語辞典（http://yougo.ascii24.com/gh/13/001373.html）

[6]　「産業分類自動格付システムの 7 年間（平成 4～10 年度）の研究について」
統計局研究彙報 59　米澤　哲一　2000 年