計算言語学 25－5
（1981. 2. 6）

分類番号つけ支援システム

中野　洋（国立国語研究所）

1. はじめに

「分類語彙表」（国立国語研究所資料集6）は、我国で唯一の現代一般語のシソーラスである。これは、表現辞典，詞藻辞典としての役割、文体論、表現論的研究のための資料としての役割、基本語彙設定のための基礎データとしての役割を持っており、いろいろと利用されてきた。最近の日本語情報処理の発展，特に意味処理の必要性が高まるなかで、「分類語彙表」は新しい役割を持ったといえる。

「分類語彙表」の語彙量は次章に述べるとおりであるが、延べ語数36,263という数値は、小型の国語辞典の収録語数（三省堂の「新明解国語辞典」の58409、岩波書店の「岩波国語辞典」の5万7千余など）とくらべても少なく、実用には不便を感じることが少なくない。

そこで、我々は「分類語彙表」の増補を企画した。分類語彙表の番号つけの最終判断は人間（それも、「分類語彙表」の製作者である現国立国語研究所所長林大）によるのが好ましい。増補作業の中での機械処理の役割は、候補となる語の選択、この段階で最も可能性の高い分類語彙表の番号をつけること、増補版分類語彙表の作表である。

2.「分類語彙表」の構造，語彙量

大分類は品詞による分類である。1桁目の数値で表わされる。

1. 名詞の仲間 ― 体の類
2. 動詞の仲間 ― 用の類
3. 形容詞の仲間 ― 相の類
4. その他の仲間

また、各類の細分は、小数点以下の数値で示される。下1桁目の数値で示される分類は次のとおりである。

.1 抽象的関係（人間や自然のあり方のわく組み）
.2 人間活動の主体
.3 人間活動 ― 精神および行為
.4 人間活動の生産物 ― 結果および用具
.5 自然 ― 自然物および自然現象

小数点以下1，2桁で分類される語の量は左図のとおりである。

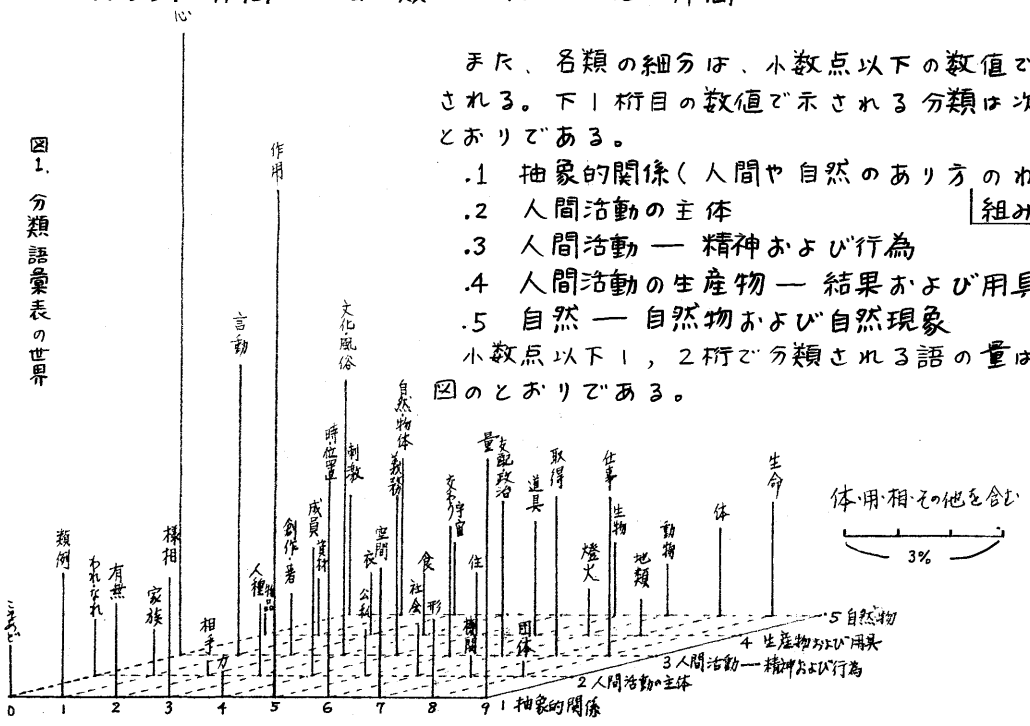


図1. 分類語彙表の世界

1桁目と下1桁目で分類された語の量は右表のとおりである。

これらの語は、語彙調査の結果得られた高頻度語、基本語彙に選ばれた語を含み、日常生活でより基本的な役割をはたしている語である。

表1. 分類語彙表の語彙量

	1. 体	2. 用	3. 相	4. その他
.1 抽象的関係	6641	2139	2192	99
.2 人間活動の主体	3183			
.3 人間活動 精神および行為	9804	2188	1774	263
.4 生産物および用具	3217			
.5 自然物および自然現象	3642	474	647	
計 (%)	26487 (73.0)	4801 (13.2)	4613 (12.7)	362 (1.0)

総計 36,263

3. 分類番号つけ支援システムの目的

先にも述べたように本システムの役割は、語の選択，作表，および自動分類番号つけである。このうち、自動分類番号つけの最終目標は、正しい番号をつけることにあるが、現段階では最も可能性の高い番号をつけられればよい。最終チェックは人間にまかせることになっている。

現在、我々は磁気テープ化された「分類語彙表」「新明解国語辞典」「コンサイス英和辞典」をもっている。これらを活用して、自動分類番号つけのプログラムを作ることを考えた。本報告はそのうちの一つ、現版の「分類語彙表」を辞書として、「新明解国語辞典」の意味記述文を解析して、分類番号をつけるプログラムである。

4. 「新明解国語辞典」の構造

もちろん、一般の辞典は人間が、ことばの意味を、ことばによって、人間に対して、説明するために作られたものであるから、機械処理むきにはできていない面が多い。長尾らは、辞書のデータベース化について、このような問題を指摘している。今回の処理においても、この種の問題は少なくなかった。

辞典の構造は次のようになっている。

**だいじ〔大事〕㊀③①㊀〔安危にかかわるので〕慎重な扱いを必要とする重大な・事(事件)。「国家の一」・「一を取る」〔＝慎重に取り扱う〕」㊁↔小事 なんとしても成しとげなければならない大切な事業。「一の前の小事」

見出し / 漢字表記 / アクセント / 語義分類 / 意味記述文 / 用例 / 用例 / 用例の意味 / 反対語 / 意味記述文 / 用例

*おだ・てる⓪〔く煽てる〕(他下一)〔何かをやらせる下心も有って〕しきり

品詞・活用型

*おせじ⓪〔《御世辞》〔「世事」の変化〕〔相手に取り入ろうとしたり、好意

補足的説明

これらは、次の4種に分類できる。

(1.) 見出し語に関するもの --- 見出し，漢字表記，アクセント，品詞
(2.) 意味記述に関するもの --- 意味記述文，いいかえ語，
(3.) 用例に関するもの --- 用例，用例の意味・説明
(4.) 補足的説明 --- 位相(雅・古・俗・方，〔野球で〕〔数学で〕等)

--- 注記（〔接尾語的に〕など），反対語，造語成分など

意味記述文の中は、意味を記述している部分とそれを文として成り立たせるための部分とにわかれる。後者によく用いられる語を意味記述用語と名づける。以下のような部分がそれである。

意味記述パターンの一例　表2. 意味記述用語

(1) ―― の意の老人語
(2) ―― の意の雅語的
(3) ―― すること
(4) ―― の一つ
(5) ―― に関する事柄
(6) ―― する職業の人
(7) ―― の様子
(8) ―― の形容
(9) ―― の変化
(10) ―― の略

意	漢語的表現	事柄	する
老人語	尊称	こと	いる
漢語	関する	もの	ある
雅語	様子	一つ	
造語	状態	無い	
造語成分	形容	なく	
雅語的表現	変化	いい	
強調表現	略	なる	

意味記述用語の「こと」以下は、意味記述用語というより、どのような文章にもよく用いられて、文表現として成り立たせるための「組み立て語彙」である。

辞書の意味記述の中には、次のようなものがある。

しらあえ --- 白ゴマと豆腐とをすり交ぜて味をつけ、これに野菜などを和えたもの。

「しらあえ」は料理の一種であるということは、「豆腐」が食品名であること、「味をつける」が料理の一つの動作を示していることによってわかるのだが、外人や機械にはむつかしい表現である。

このことは、辞書の解釈に多くの意味解釈上の前提があることを示している。辞書はそれだけで完結しているのではないようである。

意味記述文の表記は必ずしも辞書の見出し語の漢字表記を用いるとはかぎらない。むしろ、かな書きが多いといえる。これは読み手を意識してのことだろう。

かつらく〔滑落〕 --- ～すべり落ちること。

ぼうりゃく〔謀略〕 --- 相手を陥(おとし)れるための はかりごと。

はこう〔く跋行〕 ---～進み方に早いおそいが有って、つりあいがとれないこと。

かな表記が多いと、語分割や同音語の判別が複雑になり、機械処理むきとはいえない。

意味記述がいいかえ語だけであるものがある。次の例がそれだ。

くちづけ〔口付(け)〕接吻(せっぷん)。キス。

このような記述がどれほどあるかを調べた。10ページおきに1語ずつで100語を品詞の割合にしたがって選び出し、この調査をした結果が次のとおりであった。

	いいかえ語だけ	意味記述にいいかえ語を含むもの
新明解国語辞典	3	11
岩波国語辞典	6	25
新潮国語辞典	9	25

小調査ではあるが、「新明解国語辞典」の意味記述にいいかえ語が少ないのがわかる。

5. 自動分類番号つけの方法

分類語彙表の番号、あるいはそのような語の意味をあらわした番号をつける方法はいろいろ考えられる。

(1) 自然語の文章中の語用を解析して、意味分類をする方法
(2) 類義語辞典、反対語辞典を利用する方法
(3) 英和辞典、和英辞典を(2)のように考えて、利用する方法
(4) 国語辞典を利用する方法
　(ア) 見出し語の漢字表記を利用する方法
　(イ) 意味記述文を利用する方法
　(ウ) 用例・補足的説明を利用する方法

今回は(4)の(ア)と(イ)についての実験報告である。

意味記述文を解析するためには、日本語がもっている種々の問題、すなわち、分かち書き、複合語分割、表記のゆれの処理、活用語処理を解決しなければならない。この意味では、自動分類番号つけのプログラムを作るための土台として、日本語処理システムを筆者がすでに作成している一貫処理システムの上に構築する試みであるともいえる。

6. 漢字情報を利用する方法

漢字は表意文字であって、個有の意味を持っている。日本語の単語を漢字かな混りで表記すると、その漢字表記の部分が意味をもち、かな表記の部分が中国語にはない日本語個有の活用語尾、送り仮名、助辞・接辞部分をあらわす。このように考えることができれば、ある単語に漢字が一字だけ含まれている時の、その単語の意味番号は、漢字の意味番号であると考えられる。たとえば、次の漢字は次のような意味番号を持つ。

反　3.112, 1.1961, 2.112
　反　ハン　接辞　3.112, タン　助数詞　1.1961, ハンスル　動詞　2.112
博　1.3510, 1.234, 2.370
　博　ハク(ハクランカイの略) 1.3510, ハカ(ハカセの略) 1.234, ハクスル　動詞　2.370

右図の漢字意味辞書は、分類語彙表のデータに以上のような処理をほどこして作ったものである。もちろん、第2章で述べたように、1桁目の数値は品詞論的な分類を示しているから、ここでは、さほどの重要性をもたない。1桁目を無視すると、ほぼ値が同じくなる。右例で値の異なるのは、次の例である。

醜　1.1344(美醜・難易など), 3.502(色)
習　1.305(習慣…) 2.305(まね・学習・慣れ)
　　1.330(文化・歴史・風俗)

図2. 漢字意味辞書

脈	1・583		
妙	3・132		
矛	1・4550		
迷	1・3061	2・3060	2・3063
銘	1・3102	1・3154	2・3150
朱	1・502		
狩	1・3371		
趣	1・1302		
腫	1・586	2・585	
寿	1・336		
秀	2・190		
与	2・377		
醜	1・1344	3・502	
習	1・3051	1・330	2・305
賊	2・3392		
衆	1・202		
舟	1・466		

一つの単語の意味を決定するのに、以上のような漢字意味辞書がどれほどの有効性をもつのかを調べた。

「漢字・英字・単語・最後・辞書」などは、それぞれ最後の漢字がその単語の中心的な意味を表わしているようであり、

「切腹・説法・混信・遊山・読書」などは、前の漢字がその単語の中心的な意味を表わしているようであり、

「混入・滑落・少々・深遠・進出」などは、どちらもその語の意味を表わしているようである。

分類語彙表に含まれる語のうち、2字以上の漢字を含む語の分類番号を、漢字意味辞書を用いて生成したところ、その適合率は以下のとおりであった。ただし、下2桁までで、国語研漢テレ盤外字を含む語を除く。

表3. 漢字意味辞書の適合率

	最後の漢字の意味番号	語に含まれる全ての漢字の意味番号
正解が含まれる	8930	11352
含まれない	7923	6585
その漢字が辞書にない	1152	67
適合率	53%	63%

語中の位置に関係なく、どれか一つの漢字の意味番号を適用すると、適合率が10%ほど下がる。

語中の最後の漢字が、その語の品詞性(意味上の)を荷っていると考え、漢字に品詞を与える(図3)。ある漢字の品詞のうち、最も多く使用されたものをその漢字の代表品詞とする。代表品詞によって語の構造を決定する。「動詞+純名詞」の構造をもつ語について、漢字意味辞書の適合率を調べた結果は次のとおりであった。

表4. 「動詞+純名詞」の語における、漢字意味辞書の適合率

	最後の漢字(純名詞)の意味番号	前の漢字(動詞)の意味番号
正解が含まれる	2201	1640
含まれない	2508	3154
その漢字が辞書にない	283	198
適合率	47%	34%

この調査のかぎりでは、ここで述べた意味での、語構成分析処理の効果はない。効果はあがらなかったが、語構成分析は今後の処理においてその必要性はますと思われる。参考のために用いた辞書の例をあげておく。

図3. 漢字品詞・形態素処理用辞書

漢字	←直後のかなの3音・漢字 品詞	使用率(単独)	使用率(全体)
飛	1(名)	0000533	/0000533
飛	6(接)	0000649	/0000649
飛B1		0000000	/0000023
飛B2	(動詞)	0000000	/0000092
飛B3	(接)	0000000	/0000023
飛BE	(動詞)	000000	/0001176
飛NE		0000000	/0000000
飛漢1		0000000	/0003171
飛漢2		0000000	/0000069
飛漢3		0000000	/0000092

(動詞)

決	伝	当	盗	以
絶	押	下	走	選
操	送	干	対	感
痛	疑	詰	教	知
迎	滅	呼	動	働

(形容詞)

悪	深	遠	正	浅
激	古	堅	広	美
濃	厳	煩	薄	遍

図4. 漢字意味辞書 不適合データ

E1 (動詞+名詞)

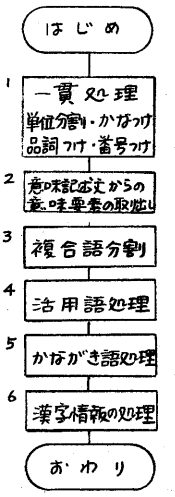
結核	結託	結局	結氷	結緑
当社	以外	以前	以後	以上
進呈	垂木	英遇	英雄	英霊
成功	成文	成句	成否	成仏
絶句	下男	下馬	下腹	下方
加熱	加算	加味	加害	加給
設計	訴因	改修	改心	改姓
送別	干草	脱皮	脱字	脱稿

7. 意味記述文を利用する方法

意味記述文は見出し語の意味を説明しているのだから、これを解析することによって意味番号を得る方法である。

日本語の文の構造は、多くの場合、文の成分の前のものが後のものを修飾して成り立っている。最終成分は、見出し語の上位概念や類概念をもっており、修飾を受けて見出し語の意味を示している。このように考えれば、意味記述文から意味記述用語を取り除いたあとの最終成分の意味番号が見出し語の意味番号と最も関連深い。文全体がどのような意味を表わしているか、それをどのように取り出し表現するかの問題は重要だが、今回は扱わない。

図5. 自動分類番号つけシステム



システムフローチャートを図5に示す。テストデータは、第4章に述べた方法で「新明解国語辞典」から100語抽出した。フローの1は筆者が作成したプログラムで、大きな辞書を用いないで漢字かなまじり文の処理をおこなう。単位分割は文節から助詞・助動詞を除いた長い単位を求める。かなつけでは、漢字に読みがなを与えるとともにローマ字表記に変換することもおこなう。ローマ字表記は4の活用語処理・形態素処理で用いる。番号つけは、一貫処理とは独立している。分類語彙表・新聞の語彙調査結果をまとめた総合辞書を用いて、分類語彙表の番号をつける。2では、1の結果得られた品詞情報によって助詞・助動詞・記号を除き、第4章で述べた意味記述用語を除いた最終成分を意味要素として取り出す。

1で分類番号がつかなかった語で、3字以上の語は、3で複合語分割処理がおこなわれる。方法は、分割のすべてのパターンについてためし切りし、それらすべてが総合辞書にあれば候補とする。候補のうち、分割して得られた語の平均語長の長いものを出力する。平均語長が等しい場合は、分割して得られた語の平均使用率（総合辞書には、新聞の語彙調査の使用率が書かれている。新聞に出現しなかった語は使用率0として計算する）の高いものを候補とする。平均使用率を優先すると、助詞・助動詞・接辞などと同じ文字列のものが取られやすい。下の例のように細く切られてしまう。

今し方 → 今 し 方　　かしこ → か し こ

出力例中の意味要素が分かち書きされたものは、このルーチンの処理結果である。1の一貫処理で間違って長くつけられた文字列が正しく分割されている例がある。複合語の最後の要素の分類番号をつける。

3の複合語処理の結果でも番号がつかなかった語は、4で活用語処理・形態素処理される。活用語の語尾を落し、辞書を検索する。終止形に変換はしない。辞書の方もこの処理によって作られている。ここで形態素処理と言ったのは、いわゆる語尾よりも長いものを用意しているからである。下記のように処理される。

(入力)	(ローマ字変換)	(出力)			
動く →	動KU →	動K	輝かす	→輝KASU	→輝K
高かっ →	高KAQQ →	高K	逃がす	→逃GASU	→逃G
高まる →	高MARU →	高M	高める	→高MERU	→高M

和語の複合語で、まぜ書き（前要素が漢字で後がかな書き）のものは、6章図3の辞書とこの処理を逆におこなうことででも分割できる。

（辞書）	（入力）	（ローマ字変換）	（分割）	（出力）
飛B	飛びばこ	→ 飛BIBAKO	→ 飛BI BAKO	→ 飛び ばこ

5ででも分類番号が得られなかった語は、その語のよみでよみ辞書を検索する。総合辞書は漢字かなまじり見出しがkeyになっているためである。同音語の判別はおこなっていない。

以上の処理でも分類番号が得られないときは、6の漢字情報によって分類番号を得る。

処理結果は次のとおりである。全体で86例である。

		正しく番号がついたもの
1で分類番号が得られたもの	44	31 (70.5)
3 〃	16	12 (75.0)
4 〃	7	2 (3.5)
5 〃	2	1 (50.0)
5までで分類番号が得られなかったもの	17	0
計	86	46 (53%)

この実験では53%と低い正解率だが、1や3では高く、6の処理を加えると64%にあがる。また、たった一つだけの番号しか出力していないので、多義語の多くが誤りとなる。総合辞書の整備などによっても正解率は高くなろう。結果的には70%台の正解率は得ることができると思われる。今後の課題としたい。「煙霧（煙と霧）」のような並立句の処理、「すく（問題・解決［処理］に、それを唯一の手段・方法とすること。）」のような文全体の解釈が必要なもの、先例の「しらあえ」のような文以外の知識を援用しなければならないものの処理は、この課題だけでなく、自然語情報処理に共通に必要なものであると思われる。

最後に入力データと処理結果を示しておく。

本報告は、文部省科学研究費試験研究(1)「言語辞書活用のための計算機プログラムの開発と言語辞書の解析」（課題番号58900２，研究代表者　長尾真，昭和55年度）を受けて行なわれた。

参考文献

1. 国立国語研究所資料集6「分類語彙表」（秀英出版，1964，1971.4.15 13版）
2. 金田一京助代表編者「新明解国語辞典 第二版」（三省堂，1974）
3. 田中穂積ほか「言語理解システム」（「大型プロジェクトパターン情報処理システム 研究開発成果発表会論文集」通産省工業技術院編，1980）
4. 暮しの手帖社「国語の辞書をテストする」（「暮しの手帖」1971.10）
5. 国立国語研究所報告37「電子計算機による新聞の語彙調査」（秀英出版 1970）
6. 中野洋「言語処理における一貫処理の研究」（「電子計算機による国語研究IX」秀英出版1978）
7. H. NAKANO. An Automatic Processing of the Natural Language in the Word Count System. COLING 80, 1980.

図6. 自動分類番号つけ 入力データ

オモウサマ／④ ←見出し語
〔母屋に居る人の意〕宮中で，父上の意の尊称．④ ←意味記述文
カート／④
手押し車．④
カマス／④
からだが細長く，口が長く突き出ている近海魚．④
カワオ／④
革で作ったひも．④
ガンセイヒロウ／④
目が疲れて頭痛を起し，本などが長く読めなくなる状態．④
キカカル／④
何かが行われているその場所へ，ちょうど来る．④
カイデン／④
師から奥儀をすべて伝えられ，弟子を取ってもよいと認められること．④
ガクギョウ／④
学生・生徒の本分として，それに励むことが要求される学校の授業．④
カシク／④
「かしこ」の意の雑語的表現．④

図7. 自動分類番号つけ 処理結果

(レコード番号) (見出し語)	(意味要素)	(品詞)(分類番号)
00160=オモウサマ	父上	11・212
00168=カート	車	11・4150
00185=カイデン	認め	E1・453
00208=ガクギョウ ⊔	授業	11・3640
00212=カシク ⊔	かしこ 」	11・1700
00231=カマス	近海 魚	11・564
00238=カワオ	ひも	11・4160
00254=ガンセイヒロウ	読め	E
00272=キカカル	来る	E2・1527
00278=キタン	遠慮 する	E2・3393
00293=キリガクレ ⊔	見え	E1・3422
00303=グアン	書く	M2・3150
00313=クチヌキ ⊔	器具	11・450
00323=クレーター	地形	11・524
00336=コウセイ ⊔	人	11・1960

品詞コード
1：名詞
E：動詞
M：形容詞