分類語彙表の増補とその利用

中　野　　洋（国立国語研究所）

１．はじめに

分類語彙表は、日本を代表するシソーラスである。多くの研究に用いられた。言語情報処理においても利用は古くからある。言語処理の高度化にともないその利用は増えているが、その意味を正しく理解されていないところがある。そこで、分類語彙表は、「①意味分類表である。②掲載してある語は例である。③多義語の語義すべてを配置したのではない。」ことを述べる。

『分類語彙表』を増補している。現在、増補の候補は82,828語である。「①増補の手順、②現在の語数分布」について述べる。

さらにこれまでになかったと思われる分類語彙表の１つの利用法として「①対照研究への利用 ②Semantic Count への利用」について述べる。

２．分類語彙表

国立国語研究所資料集６『分類語彙表』が昭和39年3月に刊行されていらい、現在29版をかさねる。研究所の刊行物の中ではもっとも発行部数が多い。一般の表現辞典としての利用が多いためだろうが、言語研究への利用も少なくない。宮島達夫・小沼悦(1992)は『分類語彙表』を言語研究に利用した論文１１９例を集めて解説している。この中には、たとえば日本語処理の道具（例えば辞書）としての使用は含んでいない。したがって、『分類語彙表』を直接間接に利用した研究はこの何倍、何十倍にのぼると思われる。

2.1 意味分類表である

分類語彙表は、語の意味分類表であって、事物の分類や概念の分類表ではない。また、これは人間用のシソーラスである。概念と意味の違いについては、柴田武(1988)を参照されたい。

言語処理でも概念と語の意味を区別して扱う方がよいことを示した報告がある。堺和宏ら(1988)は、意味処理用の辞書の構造を、言語に依存しない内容を記述した概念辞書と、言語に依存する情報の概念対応辞書、それに統語情報と概念に依存しない意味を記述する単語辞書の３層に分けることを提案している。

田中穂積ら(1987)は、分類語彙表などの「従来のシソーラスは、階層化されたもの相互が意味的にどの様な関係にあるかが不明確で曖昧なことが多い。たとえば階層化されたもの相互が上位／下位関係にあるのか、それとも部分／全体関係にあるのかがはっきりしない」ことを指摘し、これらが自然言語意味処理には不十分であり、階層関係を明確にしたシソーラスの作成が重要であると主張している。分類語彙表は言語処理用に作成したものではないのでただちに役に立たない面もあろうかと思う。しかし、下に示すように人間用には使いやすい。人間が自然言語を用いて機械を利用する場面などこの種のシソーラスが必要となる状況も生じると思う。

田中もいうようにシソーラス作成は、人間の長期にわたる注意深い作業となる。理想的には、人間にとって作成しやすい分類と作業、人間にも機械にも使える分類が望ましい。我々は従来の目的のために同じ方法で増補している。利用の側で工夫したり、作り替えたりできればよい。

分類語彙表は、人間用である。その分類は次のようになっている。たとえば、「におい(1.504)」の項目には、次の語例がある。

香(か・におい) かおり　芳香　香気　臭み
　臭気　異臭　悪臭　残り香　移り香
体臭　口臭　俗臭

この項目の中は、２つの段落に分かれている。「香(か・におい)」から始まる段落と、「体臭」に始まる段落である。前者は、においそのものを表す語であり、後者は何かのにおいを表す語である。この段落分けは形の上でも示してある。

ひとつの段落もいくつかに分かれる。前者は、「香(か・におい)」、「かおり」、「臭み」、「残り香」で始まる４つのグループに分かれそうである。後者は「体臭」と「俗臭」の２つのグループになろうか。これらのグループ分けの印はなにもない。しかし、利用者である我々には分かる

し、したがって検索も早くできる。

グループの中の語の並びにも意味がありそうである。たとえば、「臭み、臭気、異臭、悪臭」は、良くないにおいのグループだが、最初はより広い意味の語であり和語である。次に漢語で、さらに悪いにおいそのものを表す複合語で後ろ要素が「臭」となっている。これらの配列順序は必ずしも各項目に共通する規則にしたがって並んでいるのではない。しかし、それぞれを読むと分かる。もしこれを５０音順に並べると「悪臭、異臭、臭み、臭気」となってしまう。これと比べると人間には先の配列の方が検索が早いのが分かるだろう。

このような段落やグループ分け、またその並びに印がないから、あるいは明確な規則がないからといって、意味がないわけではない。さらに、機械処理に利用できないとも思わない。その分野の研究に期待したい。

2.2　語は例である。

『分類語彙表』の解説にあるとおり、語は例である。その項目に入る語すべてを掲載しているのではもちろんない。その項目をみて容易におもいつく語はのせていない。いくつかの語が組になっているような場合にもその一部をあげているだけである。「イギリス、アメリカ」はあるが、「スペイン」はない。「真北、真西」はあるが、「真南、真東」はない。

2.3 多義語

解説には、ある程度、多義を考慮したとある。増補においてはさらに考慮したが、完全ではない。したがって、多義語のすべての意味それぞれに番号をつけたのではない。もちろん、番号をつけなかったからといってその意味を認めないというわけではない。2.2のように語は例である。

３．分類語彙表の増補

解説によれば、収録語数はおよそ３万２千６百である（この数値は語彙表の延べ語数ではない）。これらの語は国立国語研究所報告21『現代雑誌九十種の用語用字』第一分冊の語彙表に掲げる使用率の高い語、さらに阪本一郎氏の『教育基本語彙』など日常生活でより基本的な役割をはたしている語である。これを研究に用い、あるいは言語処理に用いるには語が少ない。そこでこれを増補する。

3.1 増補の手順

増補は以下の手順で進めている。

①方針決定　体系の大きな変更は行わない。
　多義語を入れる。サ変語幹を用の類にも入れる。
②候補語の選択　全体に語を増やす。複合語・新語・多義語・慣用句・ｵﾉﾏﾄﾍﾟ・専門語など
③仮番号付け・段落および段落内の位置決定
④項目内の調整　語の追加削除・新段落作成
⑤項目間の調整　品詞分類間の調整ほか
⑥全体の見直し　機会あるごとに行う
⑦表記などの統一　ただし表記の基準を示すものではない。
⑧白表紙版公開　２分冊計 660頁。広く意見を聞く。
⑨公刊

現在は、③④⑤⑥の段階である。次回の白表紙版は、今年秋を目指している。

3.2 現在の語数分布

科研費を受けて作成した「『分類語彙表』形式による語彙分類表」(中野 1989、下表白表紙)の掲載延べ語数は５万２千弱だった。次の科研費「言語研究におけるｼｿｰﾗｽの利用法」（平成元～２年度）では 60,784語を得た。国語研内の課題「分類語彙表の増補」を経て、現在 82,828語（下表 671本）が増補の候補となっている。

この表の語数は、各項目に配置された語例の延べ語数である。多義語はいくつもの項目に配置されているからそれぞれ数えられている。たとえば、６７１本では、ある表記の語がひとつの項目だけに配置されたのは59,850語である。複数の項目に配置された語の項目数と語数は、それぞれ 2-8311、3-1455、4-294、5-80、6-31、7-12、8-8、9-4、10-5、18-1、21-1である。ちなみに 21の項目に配置された語は「する」である。

これらの掲載語の表記の異なり語数は、順に 33,472 47,826 70,052である。

さて、表の増補率をみれば、元の版からどれほど増補されたかがわかる。全体では白表紙が 1.42倍、６７１本が 2.25倍である。

品詞別に見れば、用の類が 3.91倍ともっとも多くなっている。相の類、体の類はそれぞれ1.9倍前後である。用の類が多い理由は、「サ変語幹＋する」の形の語を入れたからである。

分　類	掲載延べ語数			増補率	
	元版(FD)	白表紙	671本	白表紙	671本
体の類	26,984	40,227	54,591	1.49	2.02
抽象的関係	6,780	9,026	12,663	1.33	1.87
人間活動主体	3,272	5,020	7,127	1.53	2.18
人間活動	9,920	14,708	19,247	1.48	1.94
生産物	3,277	5,656	7,960	1.73	2.43
自　然	3,735	5,817	7,594	1.56	2.03
用の類	4,779	5,358	18,710	1.12	3.91
抽象的関係	2,153	2,380	8,145	1.11	3.78
人間活動	2,158	2,441	9,265	1.13	4.29
自　然	468	537	1,300	1.15	2.78
相の類	4,653	6,147	8,928	1.32	1.92
抽象的関係	2,212	2,899	4,316	1.31	1.95
人間活動	1,788	2,515	3,571	1.41	2.00
自　然	653	733	1,041	1.12	1.59
その他類	364	390	599	1.07	1.65
抽象的関係	99	105	127	1.06	1.28
人間活動	265	285	472	1.08	1.78
総計	36,780	52,122	82,828	1.42	2.25

最も増えたのは、2.32で 21語から 153語と7.29倍に増補された。相の類では 3.17で 19語から 103語と 5.42倍となった。体の類では 1.27で 142語から 546語と 3.85倍となった。 2.32(創作)の一部を次に示す。

［ＦＤ版の最初の段落］

1 *著わす 詠む 詠み込む 歌いあげる 焼き直す

［６７１本で上に直接対応する段落］

3　著す　著作する　著述する　撰述する
　執筆する[3152-1]　述作する　書く
　創作する　生む［傑作を～］　作る
　作文する　作詞する　詩作する　句作する
4　詠む　詠み込む　詠ずる　詠じる
　歌う［和歌を～］　歌い上げる
　即詠する　即吟する　偶詠する
　苦吟する　沈吟する　詠進する
5　焼き直す　翻案する　潤色する
　改作する　摸作する
　脚色する　劇化する　アレンジする
　文字化する[23150 1]

以上に示したものは、増補作業中のデータである。語数についても、語例、項目、表記なども変る可能性がある。

4．分類語彙表の利用

分類語彙表の解説には、このようなシソーラスの役割の一つとして情報処理での利用をあげている。実際、言語処理での利用は早かった。
現在でも大学や企業などのいろいろな研究機関で使われている。たとえば、手元にある奈良先端科学技術大学院大学の松本研究室年報 1993-94 Lingua vol.1-2 掲載の論文29件のうち参考文献に分類語彙表があげてあるものが12件ある。

以下では、最近発表者が行った語彙の対照研究での利用と計画中の雑誌の語彙調査でのSemantic Countについて述べる。

4.1 中国流行歌の日中対照研究

標題についての語彙の対照研究を行った。
この種の研究において何を手がかりとすべきかが問題となる。研究の単位である語の認定すらも共通となりえるかさえ疑問である。

この研究では、中国語とその逐語訳を用いた。次の例は、日本語訳で「会う」を用いた一聯と対応する中国語の一聯である。同じ漢字を用いているとはいえ、このように対訳でかつ同じ意味を表わす語の用例を集めなければ分析はすすみがたい。

中国語と日本語訳での用例

再過二十年我們来相会
二十年たったら我々はまた会いましょう

再来看望親愛的媽媽
また親愛なるママに会いに来る。

不見哥哥心憂愁
兄に会わぬと心が憂愁する。

我們再相逢
わたしたちは再び会おう

我們相約在那小木橋
私たちはあの小さい木橋で会うことを約束する

4.1.1 分類語彙表の意味番号による語彙の対照

日本語訳の語に分類番号を付ける。これを集計し、意味分類項目を語数順に並べたのが次表である。表をみると中国流行歌の内容が表れる。すな

わち、異なりの「地形・山野,植物名,川・湖」が多いことは自然現象が題材になっていることを、異なりの「対人感情」と延べの「われ・なれ・かれ,親・先祖」が多いことは人称代名詞や親が話題になっていることを示している。

日本語訳 異なり語数順			延べ語数順		
分類番号	項目名	異	分類番号	項目名	延
1.5240	地形・山野	25	1.2000	われ・なれ	497
1.5520	植物名	20	2.3420	行為	174
1.5250	川・湖	17	3.1000	こそあど	107
1.3020	対人感情	15	2.1527	往復	103
1.1950	一二三	14	2.1200	存在	100

実際、日本と中国の流行歌を比べると山野、川等を表わす語が日本のそれより多いことがわかる。

分類	異なり	延べ	中国	日本
山など	36	88	16.7	3.45
川など	17	62	11.7	2.33
海など	28	51	9.7	6.16

さらに日本語訳の分類番号によって、中国語の類義語を集めることができる。以下に語例を示す。

分類	語例
[波・潮]	海波,海浪,巨浪,驚涛,春潮,清波,波濤,碧波,浪,浪流,涛声
[海・島]	海,海峡,海上,海水,海風,海面,海洋,海疆,岸,重洋,大海,島,東海,南海,湾,戈壁灘,鼓波嶼
[川・湖]	黄河,河流,渓流,湖,湖水,湖面,江水,小河,清流,西湖,泉水,太湖,大江,長江,灘,澎湖

以上のように、対照の手がかりとして意味番号は有効である。他の言語の意味分類を利用してどう異なるかも興味深いテーマである。

4.2 語彙調査における Semantic Count

語彙の対照研究では、異なる言語間の対照だけでなく、同じ言語の異なる材料、たとえば異なる時代の言語材料、異なる分野、異なる作品の比較も重要な課題である。

人称代名詞の比較などのように品詞や語種などの比較はそれぞれにつけた品詞や語種情報によって分析できる。しかし、個々の意味分野の語の差異を知るには、意味情報をつける必要がある。今、我々は分類語彙表を約８万語にまで増補している。そのような語彙を配置できる意味分類体系を作成している。これを用いれば Semantic Count が可能になり、語彙分析がより深められる。

宮島達夫(1986)は、1906年から1976年までの雑誌『中央公論』８冊を比較して場所を表わす語が漢語から外来語に変ったことを明らかにした。たとえば「英国」を「イギリス」、「露国」を「ロシア」、「西洋」を「ヨーロッパ」という具合である。分析には分類語彙表の番号を用いている。

これまでにも、国立国語研究所の語彙調査『総合雑誌の用語（前編）』『雑誌９０種の用語用字』『高校教科書の語彙調査』『中学校教科書の語彙調査』(国語研 1957,62,83,86)には分類語彙表の番号を付けている。山崎誠(1989)は、高校と中学校の教科書のデータを用いて「意味別語彙集」を作成した。これらは見出し語に番号を付けて分析したものである。

Semantic Count とするためには、文脈中の語それぞれに分類番号を付けなければならない。しかし、『分類語彙表』に掲載されていない語が現れた場合、独自に番号を付与できるかどうか、また、慣用句や擬声語・擬態語などへの付与は難しく、さらにひとつの番号に決定できるかどうかが問題である。

困難な問題が山積しているが、得られる成果も大きいと考えられるので試行的な調査を計画しているところである。

参考文献

堺和宏,徳永健伸,奥村学,田中穂積(1988)「自然言語の意味処理用辞書の構成法」（情処技法 Vol.88,No.38 88-NL66）

柴田武(1988)「語の意味と概念と外界」（『日本語大百科事典』,講談社）

田中穂積,仁科喜久子(1987)「上位／下位関係シソーラスＩＭＩＭＡＰＩの作成（Ｉ）」（情処技法Vol.87,No.84 87-NL-64）

中野洋(1995)「中国における流行歌の語彙」(計量国語学19巻8号)

宮島達夫・小沼悦(1994)「言語研究におけるシソーラスの利用」（宮島達夫著『語彙論研究』,むぎ書房）

--(1986)『雑誌用語の変遷』(秀英出版)

山崎誠(1989)「意味別語彙集」（『高校・中学校教科書の語彙調査』,秀英出版）