国語辞典に基づくシソーラスの計算機処理

1G-2 横山 晶一 ・ 荻野 孝野 ・ 荻野 綱男

(電子技術総合研究所) (日本電子化辞書研究所*) (筑波大学)

□はじめに 国語辞典の語義文が、人間用のものであるとはいえ、語の間の種々の関係を示すものであることはよく知られている[1]。この語義文から、人手によってこれらの関係を抽出すると有効であることはすでに述べた[2]。ここでは、人手によって抽出されたこれらの関係に種々の計算機処理を施すことによって、自然言語処理や知識処理のためのシソーラスに「加工」する手順とその有効性を示す。

□資料・統計 電子化された新明解国語辞典(第2版、三省堂、1974年)の主として重要語からなるデータのうち、名詞とその語義文に対して、種々の関係を人手によって入力した。このデータは、親見出し語で全体の約10分の1、語義文で全体の6分の1に相当する。結果的には親見出し語 3,702語、子見出し語 2,375語に関係が付加されている。計算機処理の詳細は次節以下に述べることとして、まず表1に各関係に関するデータ数を示す。この表から分かるように、親見出し語については1語あたり約4.28語、子見出し語については1語当たり約2.77語の関係付けがなされている。上下関係(特に見出し語の側が語義に対して下位概念になるもの)、同義語、関連語が多く抽出されている。

表1 意味関係の抽出データ数

関係名		親	子	計
上下1 (上位)	>	1,029	337	1,366
上下2 (下位)	<	5,024	2,461	7,485
全体部分1	(	215	42	257
全体部分2	)	73	27	100
同義	=	3,097	711	3,808
反義	+	441	157	598
和集合1 (上)	〉	41	19	60
和集合2 (下)	〈	6	2	8
例示 (値) 1	≧	17	13	30
例示 (値) 2	≦	45	19	64
関連	R	5,096	2,562	7,658
兄弟 (類義)	G	772	230	1,002
合計		15,856	6,580	22,436

□基本処理 人手で抽出した関係語の計算機データには、見出し語の行(レコード)番号、関係語の見出し行からの相対位置、関係、該当語が含まれている。ただし、反義語は、もともと国語辞典に明記されているものであるから、人手ではデータ抽出を行っていない。そこで、最初の処理としては、見出し語と関係語との結合、および反義語の自動抽出が行われた。もとの国語辞典では、子見出しはたとえば、親見出し語の「積極」に対して「−的」のような省略形式がとられているが、ここでは子見出しを自動的に埋めこんだ見出し語データと照合することによって対応をとった。通常親見出し語はカナで記載されているが、ここでは後述する逆関係の生成の際に、なるべく同じ語が同じ場所にまとまるようにするために、漢字表記のある見出し語は、そちらを採用するようにしてある。図1に、見出し語「あいだ(間)」に対する関連語の一覧を示す。なおこの図では、後に述べる処理は行っていない。また、図の@は、関係語抽出者によって付加された一般的な概念であり、もとの辞書にはのっていない。

```
>  すきま、限り、中間、仲、仲間同士
<  空間、時間、人間関係
=  絶えま、あい、@間柄
R  非連続部分、大多数
```

図1 「あいだ(間)」に対する関係語

そのほかの処理は、次のようなものである。

(a)類義語・語義番号の埋め込み 国語辞典では、対比される同音語、語源の異なる外来語などを一まとめにした見出し語で記述し、㊀、㊁(計算機ではⅠ、Ⅱなど)のように区別している。また、上の図1のような関係を作成すると、KWICなどで参照する場合を除いては、語義の違いによる区別が明確でなくなる。たとえば、見出し語「グループ」の中には、「(1)似通った点によって分けた・人(物)の集まり。」と、「(2)行動をともにする集団。仲間。」という2つの語義があり、「集まり」、「集団」は、いずれも上位語(<)としてとられている。コンテクストをはずすと、

* 当時、(財)計量計画研究所

Automatic Construction of a Thesaurus Based on a Japanese Dictionary

Shoichi YOKOYAMA[1], Takano OGINO[2*], and Tsunao OGINO[3]

1. Electrotechnical Laboratory, 2.Japan EDR Institute Ltd., 3.University of Tsukuba
* In this study, she was working for the Institute of the Behavioral Sciences.

2つの語はきわめて似たものとなってその区別が曖昧である。そこで、語義番号を付加した形で、「グループ(1) ＜ 集まり」、「グループ(2) ＜ 集団」という関係を作ってもとの情報の一部を保存したシソーラスを作る。

(b)省略部分の埋め込み　子見出しの省略は、すでにその部分を埋めこんだデータと照合しているので問題はないが、語義文中でもこのような省略表現が用いられており、しかも省略表現を含んだままの形で下位語などの指定がなされている場合がある。たとえば、見出し語「器官」においては、「呼吸－」という派生語的なものが下位語として関係づけられているが、これはもちろん「呼吸器官」という形に直した上でシソーラスに登録しなければならない。国語辞典では、通常カナの見出しに対して、それを漢字で表現した表記部分がついているので、それがある場合には、上の例のように漢字部分を埋めこんだ。そうでない場合にはカナ見出しを直接埋めこんである。

(c)不要データの除去　関係語の部分にも、照合される見出し語の部分にも、シソーラスを整備する上では直接必要のない文字（列）が含まれていることがある。たとえば、図1に示した、関係語抽出者によって付された「@」は、KWICの作成には必要であるが、語と語の関係を問題にする場合には不要である。また、（特に）子見出しなどで、読みがなをつけてあることが多い。たとえば、「知識＊人＜じん＞」は、＜＞で囲まれた部分が＊のあとの部分の読みを表わしている。シソーラス作成の際には、この読みがな部分を完全に除去した。この例ではさらに、子見出しの埋め込みと切れ目を表現した＊も除いてある。なお、カッコについては、種々の問題点があり、すべての事柄を自動的に処理することは困難であるので今回は扱っていない（これについては問題点のところで論じる）。

□逆関係の自動生成　表1から明らかなように、番号を付した関係（上下関係1など）は、互いに逆の関係になっている。また、そのほかの同義、反義などの関係は、対称的な関係であり、逆に並べても同じ関係が成立する。そこで、表1に示した各々の関係について、語を入れ換えた後に、上記の基準に従って改めて関係付けを行った。これによって表1に示したもののちょうど2倍にあたる44,872組の関係が得られたことになる。

次に、これをコード順にソートする。これによって、見出し語と、抽出された関係語とが（語義番号などを除いて）同じ場所に集まる。図2に「反応」という語に対して得られた関係語の一覧を示す。語義番号のついているのが見出し語部分からとられたもの、そうでないのが関係語として抽出されたものを示している。なお、この図を図1のような形に表わすこともちろん可能である。

反応	＋	刺激(3)
反応	＞	拒絶反応
反応	＞	酸性反応
反応	＞	反射(2)
反応(1)	＋	刺激
反応(1)	＝	手ごたえ
反応(1)	＜	動き
反応(1)	R	働きかけ
反応(2)	＜	化学変化
反応(2)	R	物質
反応(3)	＜	現象
反応(3)	＞	生体反応

図2　「反応」と関係語のシソーラス

このようなシソーラスを作成すると種々の利点がある。まず、語と語の関係が明確になり、内容をさらに詳しく分析することにより、意味分類や、語のグルーピングが可能になる。また、単純なグルーピングであれば、上下関係を次々にたどるなどによって、自動的に階層的な関係が抽出される。また適宜語義番号などによってもとの語義を参照することにより、共起関係の解析にも役立つ。

また、これらの階層化された関係を組み合わせてネットワークを作成し、オブジェクト指向型概念辞書[1]による知識表現や、自然言語処理などに応用することもできる。

□問題点　このシソーラスにはまた、問題点もいくつか残されている。まず第一に、図2からも分かるように、逆関係を抽出した場合には、通常語義文の部分には、関係語の語義番号まで記載されていないから、どの語義を表わす語同士が対応するかが明確でない。ただし、反義語のように、図2の2つの部分を統合することによって、「反応(1) ＋ 刺激(3)」といった対応が自動的にとれる場合もある。いずれにしても、ある程度の細かい語義区分までを問題にする場合には、人手による作業がさらに必要になる。

次にデータの処理の問題がある。ここで最も大きい問題はカッコの処理である。国語辞典の中ではカッコは種々の意味で用いられており、人間には明確であっても計算機処理の上からは問題になる場合が多い。たとえば、「スター」の語義、「人気・俳優（歌手・選手）。」は、この辞典独特の木構造を表わすもので、「人気俳優。人気歌手。人気選手。」と展開されなければならない。また「渦巻（き）」のように、送り仮名の有無が任意であることを示すカッコもある。これは、「渦巻」、「渦巻き」の2種類に展開すればよいのであるが、「組（み）合（わ）せ」のようにカッコの数が多くなると、組合わせの数も増大する。また、そのほかに「越える（越す）こと」といった表現もあって、これを前記のカッコと区別することは自動的には不可能である。現在そのほかにどのようなカッコの用い方があるか分析を進めており、最も手間の少ない方法でカッコの自動処理を行うことを検討中である。

さらに、上とも関連するが、表記の問題がある。抽出された44,872組の関係のうち、平仮名で書かれているものは2,667語で、数としてはあまり多くはないが、これの漢字表記との突き合わせは必要である。

□今後の課題　最初に述べたように、このシソーラスは、主として重要語の名詞を対象にしている。そのほかの語に対する作業は、すでに人手による部分は完了しており、計算機によるデータ作りが進行中である。これらのデータについても全く同じ手順でシソーラスを作成することができるので、これらの新しいデータを組み込んだシソーラスの作成を計画中である。また、別の辞典（三省堂国語辞典）に対する同様のシソーラスがすでに作成されており[3]、電子化された形で実用に供されているので、2つのシソーラスを比較対照したり、あるいは統合してさらに研究を進めることを考えている。

□参考文献　1．横山・羽中田・鈴木：情処全大37, 3B-1 (1988.10).　2．荻野(孝)・横山・荻野(綱)：本大会(1989.10).　3．荻野(綱)：文部省特定研究報告集(1989.3).