1H-6

標準科学技術用語と多言語シソーラスシステム

伊藤俊男、霜山友肖、大保信夫、藤原譲
(筑波大学)

1．はじめに

情報の記述・理解・流通には、正確な用語の利用は必須のものである。特に、国際的な情報流通を考えた時、その要求はますます大きなものとなる。

これまで、このような問題に対して、複数言語間での相互情報流通を促進する道具としての多言語シソーラスの作成を目的として、文部省学術用語集を対象として取り上げ、用語の持つ関係性の解析を行ってきた。今回、対象とする用語の分野数を増やして行った統計解析結果と共に、多言語間の多段階変換における用語の対応関係についても報告する。

2．統計解析結果

ここで言う学術用語ファイルとは、冊子体の文部省学術用語集の内容を磁気テープ上に実現したものである。現在、このファイルに収録されているのは２５分野の用語約１０万で、それぞれ和英の部と英和の部とに分かれており、日本語の用語に対してローマ字表現（以下ローマ字という）が与えられている（表１参照）。

異言語間でキーワードの変換を行う場合に問題となるのは、それらキーワード間に存在する多対多の対応関係である。学術用語ファイルにおいては、ローマ字と日本語、日本語と英語の間に多義性を含んだ対応関係がある。

(1)ローマ字と日本語の間の多義性

まず、一つのローマ字に対しては平均1.0824個の日本語が対応している。これを、分野毎にみると、一つのローマ字に対して２５分野平均1.0159個の日本語が対応し、分野を限定することによって多義性の減少するのがわかる。

逆に、一つの日本語に対するローマ字は、平均1.0355個存在している。これを分野毎にみると、２５分野平均1.0207個となり、分野限定による多義性の減少がみられる。

(2)日本語と英語の間の多義性

一つの日本語に対しては、２５分野全体で平均1.49個の英語が対応しているが、分野限定を行うことにより、２５分野平均で1.13個に減少する。反対に、一つの英語に対する日本語は平均1.37個存在する。これを分野別に見ると２５分野平均で1.20個となり、やはり分野限定による多義性の減少がみられる。

(3)分野間の相関

異なる分野に共通して現れる用語の多少は、それらの分野の関連の度合を示すものと考えられる。そこでは予想されたように、内容的に類似していたり、関係の深い分野間では重なりが大きく、そうでないものでは重複が少ない。また、多くの分野に関連を持つ基礎的な分野と、他の関連性の少ない専門性の高い分野とが定量的に区別されている。

3．多言語間における用語の対応関係

ｎ言語間での多言語シソーラスを作成しようとすると、ｎ個の用語ファイルと共に、用語の関連を記述したn(n-1)個のファイルが必要となり、収録する言語の数の増大につれて非常に大きな記憶容量を必要とするようになる。この時、一つの言語を変換の軸として他のn-1言語との間に用語の関連を記述したファイルを持たせるようにすれば、記憶容量を節約することができる。但しその場合、出発言語から多段階の変換によって目的言語の用語を得ようとすると、それぞれの変換段階に存在する多義性のために、変換時に不要の意味的拡散が引き起こされる可能性がある。例えば、学術用語ファイルを用いて、日本語→英語→日本語の多段階変換を行うと、「基数」という用語はノイズを含んだ９個の用語に展開・変換されてしまう。

Analysis of Standard Term for Science and Technology and the Multilingual Thesaurus System

Toshio ITOH,Tomonori SHIMOYAMA,Nobuo OHBO,Yuzuru FUJIWARA
University of TSUKUBA

基数（電気工学）

↓　日本語→英語変換

ｂａｓｅ（電気工学、情報処理）
ｒａｄｉｘ（電気工学、情報処理）

↓　英語→日本語変換

基数（電気工学、情報処理）、ベース（図書館学、機械工学、計測工学、物理学）、塩基（化学、機械工学）、基剤（化学）、基地（航空工学）、脚（機械工学）、口金（建築学、電気工学）、礎盤（建築学）、素地(化学)

このような状況をもたらす用語の多義性は、次のように分類して考えることができる。

①一義　　全体として一対一の対応関係を持つもの
②準一義　全体としては一対多であるが、分野を限れば一義となるもの
③多義　　分野を限定してもなお多義であるもの

これらのうち、①は変換時には問題とならず、②についてもあらかじめ分野を限定することで不要の変換を避けることができる（上の例では出発用語の分野＝電気工学に限定して変換することにより、ノイズを消すことができる）。最も問題となる③については、少なくとも準一義性を持つようになるまで分野の細分化を行うか、もしくは意味的な取り扱いをすることが必要となる。

4．ＣＤ－ＲＯＭ上の多言語シソーラス

多言語シソーラスは、膨大な用語と、その間の多対多の関連を含んだ大容量のデータベースとなる。ユーザに対する高速なレスポンスを確保するためには、このような辞書は端末に付随させることが必要である。

このような目的に対しては、小型大容量のＣＤ－ＲＯＭの利用が最適と考えられる。ＣＤ－ＲＯＭの場合、ビット単価が安いことに加えて、その名の通り読み取り専用であることから、多対多の関連表現に十分な冗長性を持たせてもデータ間の矛盾の発生する可能性がないので、高速な検索に適したデータ構造を取り入れ易いという利点がある。

現在、日本語－英語－ドイツ語の科学技術用語に対する相互変換用辞書をＣＤ－ＲＯＭ上に実現し、パーソナルコンピュータとの接続で十分な変換速度を得ている。実現方式としては、日本語、英語、ドイツ語の相互変換ファイルに多段の索引によりアクセスする方式を採用している。

5．おわりに

情報化社会における分散配置データベースへのアクセスによる情報の国際的流通は、そこで用いられる用語の整備・標準化と、それを盛り込んだＣＤ－ＲＯＭ等による多言語シソーラスによって円滑化される見通しが得られた。しかしながら、現在利用できる用語と分野は本格的利用には不充分であるので、今後はＪＩＳ用語及び特許用語などの追加入力によって、学術面のみならず、技術面での利用性の拡大を図るとともに、分野の細分化による精度の高い変換方式の実現を図る。

	分野名	英語	日本語
1	動物学	2169	2048
2	原子力工学	3607	3707
3	航空工学	3195	3120
4	植物学	2667	2579
5	化学	9977	9848
6	電気工学	10116	10030
7	図書館学	4846	3963
8	数学	1644	1624
9	計測工学	2684	2542
10	物理学	4033	3902
11	地震学	2475	2451
12	船舶工学	9232	8412
13	分光学	2108	2098
14	歯学	535	596
15	気象学	1841	1751
16	機械工学	10272	9204
17	建築学	5841	6373
18	論理学	803	735
19	海洋学	2414	2378
20	土木工学	6027	5761
21	地理学	1720	1857
22	遺伝学	1846	1821
23	天文学	2214	2177
24	採鉱冶金学	4812	4404
25	情報処理	7427	7089
	合　計	104505	100468

表１　学術用語ファイル収録語彙数

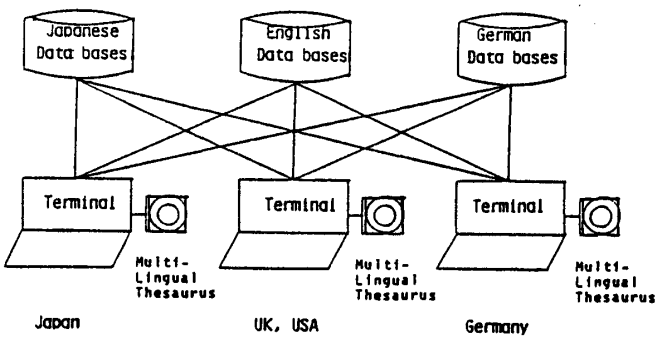

図１　多言語シソーラスシステム