Ⅳ-49　情報検索システム用辞書を目的とする道路工学英用語データベース

北海道大学工学部　正員　上島　壮

1　はじめに

文献は構造化された概念の集合であり、我々がデータベース内の文献を探す作業はキーワードに概念の意味を托して意図する情報を選び出そうとする操作である。このキーワードは、一般に、文献から機械的に区切り記号で切り出されたものか、労力をかけて人為的に入力された統制語かのどちらかである。そして前者の場合にはキーワードはその抽出手順から言って必ずしも概念の単位ではない。特に道路工学用語については一般語の組合わせで作られていることも多くあり、キーワードを用いて望む情報を的確に探し出すということは大変難かしい技術である。つぎのことは大規模情報システムでは考え難いことであろうが、仮に論文などの内容が概念の集合である辞書にもとづいて記述され、また、辞書についても用語が多様な概念にもとづいて分類組織され、それが自由に検索・操作可能であれば、文献データのデータベースへの編入を機械的に行ないながら統制語方式以上に的確に内容を「読む」ことが可能になると思われる。

このような発想から、アスファルト研究者用文献蓄積システムに用いることを目的に、複合語を取り扱うことができ、用語間の関係を記述できる検索システム用英語辞書を試作した。用語は書物の用語辞典より採取したが、これをベースに文献データより切出した用語を逐次編入して行く予定である。なお、この用語集はＲＲＲ（Resources for Road Researchers）として北海道大学大型計算機センターにおいて公開しているが、現在のところ「辞書を引く」機能と関連語の検索機能のみ持つ。

2　用語集の内容

用語集は次の三つのブロックより成る。

(1) 道路工学和英用語集　資料は「道路用語辞典（日本道路協会編　1977）」によっており、見出し語約5400語（カタカナ）とそれに対応する英語を図－１のように階層化したものである。ここで、①和英、②等号による上位語への結合、③矢印による上位語への結合、などのように関係を定義した。

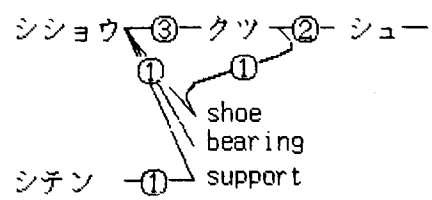
なお、日本語部については端末より対話形で利用する和英、あるいは英和辞典としてのみ有用である。

(2) 道路工学英用語シソーラス　資料は「IRRD THESAURUS 1979」の英語部によっており、語数は約3400語で50のグループに分類されている。図－２に階層関係の構築例を示す。ここで、①細分類グループの見出し語、②グループのメンバー、③複数の上位語を持つ結合、④SEE ALSOによる同格語への結合、⑤使用地域コードへの結合、などのように関係を定義した。

また、IRRDの分類コードは用語と同じレベルで取り扱われ xxyy は #08.0xxyyのように表示される。

```
してん　　英　support
くつ　　英　shoe　　⇨　ししょう
ししょう　英　support,bearing,shoe
シュー　　＝　くつ
```

イ）原典の表示

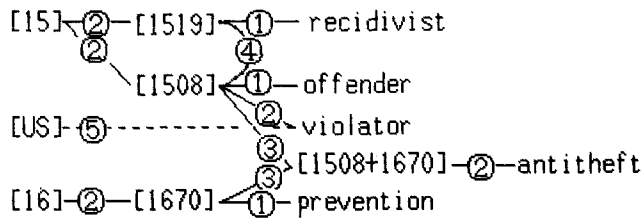


ロ）階層関係の設定

図－１　和英用語集の階層設定例

```
15 08 OFFENDER
     * VIOLATOR(US)
     * (SEE ALSO RECIDIVIST 15 19)
     * (ANTITHEFT=OFFENDER  15 08
              + PREVENTION 16 70)
```

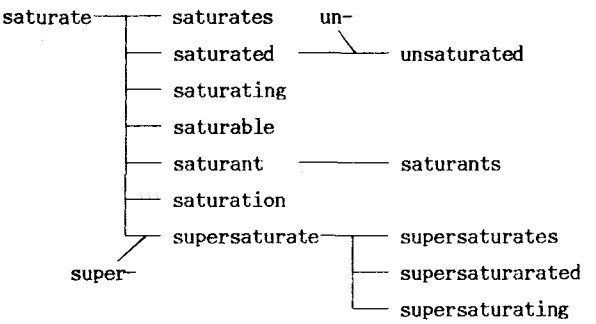
イ）原典の表示



ロ）用語辞書内の結合構造

図－２　IRRD Thesaurus用語の結合構造例

(3)一般英用語集　一般英語辞典より採取した約４万語にその変化形約３万語を加えたものである。接尾語などの規則利用による記憶領域の圧縮は行なっていない。上位、下位の関係付けコードとして イ）複数形または動詞の現在変化 ロ）動詞の過去、過去分詞 ハ）-ing形 ニ）比較級、最上級 ホ）略語 ヘ）別綴り ト）派生語などを与えている。図－３に階層関係の設定例を示す。

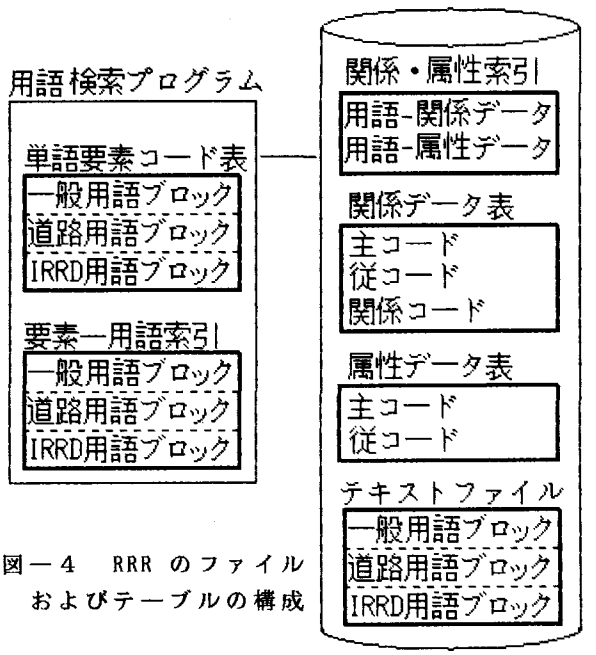


図－３　英一般用語の階層関係設定例

３　システムの構成

この辞書を用いて構築した用語検索システムＲＲＲのファイルおよびテーブルの構成を図－４に示す。コマンドあるいは入力文章（現在テキスト処理機能はない）は次の手順で処理される。

① 入力文章を単語要素の列に分解する。
② 単語要素コード表より要素コードを求める。
③ 要素－用語索引により要素を含む用語のコードの集合を求める。用語情報には構成要素数とその要素の位置があり、それを用いて複合語を抽出する。文献蓄積プログラムとして用いる場合は複数候補を処理の上、この用語コードをキーワードとして保存する。
④ 探索キーワード集合は関連語検索機能を用いて作成することが可能である。検索は用語コードに対して行なうことになる。
⑤ 関連語の探索は、関係・属性索引より、その用語の関係データ表上の番地を得て、関係データ表から従コードを読むことにより行なう。



図－４　ＲＲＲ のファイルおよびテーブルの構成

属性データ表は第３のコードが不要の場合に用いる。関係コードは用語の使用地域コード、分類コードなどとしても用いる。主コードは検索のキーとなる用語コードまたは関係コードである。したがって、関係付けられる対象は用語でも関係（属性）でもよい。

部分一致語探索、複合語探索などを高速処理するために、単語要素コード表および要素－用語索引をプログラムロードモジュール内に組込んだが、メモリーの使用量はそれぞれ 0.64MB および 0.73MB ある。

ＲＲＲ の検索機能は不完全な指定に対して単語あるいは複合語を表示させるあい昧指定探索に比重が掛っている。関連語の検索は上位語検索、最上位語検索、下位語検索、全下位語検索などを組み合せて行なう。

４　おわりに

この用語辞書のシステムで大きな課題は更新方法でありそのために ADABAS などのデータベースシステムを利用することも考慮している。計算機の主記憶は 2MB程度必要とし用語の増大、機能の拡張に伴って更に増大すると思われるがユーザーが使用できる容量が拡張しつつあるのでこれには楽観している。

最後に、道路工学用語データの作成とその分類（作業中）を担当していただいた元首都高速道路公団技師高橋稔氏の労に深く感謝いたします。