計算言語学 13 － 2
（1978. 2. 24）

日本語の自立語辞書

坂本義行　（電子技術総合研究所）

0. まえがき

自然言語による科学技術テキストの分析を行なうための基本的な構成要素の1つとして，語彙の抽出すなわち自立語の認定に用いる辞書が必要である。その情報源としては，学術論文，特許，科学技術雑誌の3種に大別される。これらの情報の計算機による管理は，質量の両面から重要性を増してきた。ここでは，分野を限定したテキストの語彙特徴を抽出することを目的として，特許公報の第12類（金属の加工）と計算機のプログラム分野におけるテキストが有する語彙の特性を抽出し，形態情報，構文情報，意味情報の3種の情報空間からなる自立語辞書の生成・検索システムを開発したので報告いたします。

とくに特許公報を選んだ理由として，検索といった利用面から要求もあるが，日本語で記述された情報が言語面からみたテキスト（特許公報1件を単位とする）として形態上明確な形式化が行なわれている点に着目した。またプログラム分野については，自然言語によるプログラミングの研究の基礎資料[1]を得ることを目的として，複数個のテキストから語彙を抽出し，自動プログラミング用の辞書を作成した。この2種類の辞書は，ほぼ同一の構造からなっており，1つのプログラムで辞書の生成と検索が行なえる。

1. テキストの語彙分析

特許公報をテキストとして，その構造を分析するには，公報が有するテキストの一般的な特性（形態的特性）と特定分野における語彙の特徴を抽出せねばならない。ここでは，語彙について分析を行ない，必要な語彙情報を抽出する方法について述べる。

公報全般にわたって出現語句を分類すると，

1) 無性格語（出現頻度が高く，機能的な役割をになっている）
　助詞，助動詞，接辞，etc
2) 接続・指示語（文間の係り受けを形態的に示す）
　接続詞，接続副詞，接続助詞，連体詞，指示代名詞
3) 修飾語（強意，限定を表わす）
　形容詞，形容動詞，副詞
4) 概念を表わす語
　物質名詞，属性名詞，抽象名詞，助数詞
5) 陳述を表わす語
　動詞

のように分けられる。

以上の項目のなかで4)，5)は，特定分野のテキストを特徴づける語彙が含まれていると考えられる。

金属の加工分野の特許公報12件とプログラミング関係の7種の本について名詞，動詞を中心に意味の分類を行なった。

1.1 金属の加工分野の語彙分類

分野をできるだけ特定化するため，分類に用いたテキストは，すでに報告したConcordanceの作成[2]に用いた特許公報1冊（延字数＝約25万，延べ語数＝9万，異なり語数。11,000）のなかを特許分類表[3]の類で分けて，第12類（金属の加工）に関する12件のみを用いた。抽出した自立語は，約6,000語となり，これらの見出しについて，構文情報，意味情報を与えた。

構文情報としては，既報の「文節辞書」[4]にそって品詞を与え，動詞については，活用と格支配の情報を格と意味分類コードとの組み合せの型として与えた。この格支配情報については，基本的な動詞に対する石綿氏の分類[5]があるが，この場合専門の用語が多いため，独自の型に分類した。

意味情報については，国研の分類語彙表[6]を検討したが，専門分野の語の分類という点で不十分であり，また特許分類表についても検討を行なったが，その使用目的が文献全体がどの分類に属するかである点，一般語の範ちゅうに属する語が多くなる点，JICSTのシソーラスについても検討を加えたが，詳細な分類を必要とする点から，個々の見出しについて手作業で独自の分類を行なった。

以上の作業により得られた語彙は，

見出し　2,230

（品詞）
名詞　1,458
動詞　591
形，形動，副　181

となり，第1表に示すような意味分類を行なった。

第1表　金属の加工分野のファセット分類表

A	行為・処理（人を主体とする動詞）
A1	一般語
A2	処理，操作に関するもの
A3	人および物が主体になるもの
B	現象，作用（物や性質を主語とする動詞
B1	通常の現象
B2	装置を主体とするもの
B3	一般語
C	物質名
C1	金属
C11	単体（元素名）
C12	合金
C2	非金属単体
C3	化合物
C31	無機化合物
C32	有機化合物
C33	その他
C4	混合物
C5	基，根，環など
C6	分子，原子，電子など
C7	その他
D	物質の役割
E	物質の形状
F	物の性質
G	条件，状況
H	装置
H1	完成品
H2	部品，ユニット
H3	部材
H4	型式
H5	モジュール
I	場所
I1	物の部位
I2	空間的位置
I3	方向
J	製法の特性
K	用途
L	その他の名詞
L1	一般語
L2	試験，評価の手段
L3	引用文献など
L4	メタ言語
M	形容詞

1.2 プログラム分野の語彙分類

自然言語（日本語）でプログラム仕様を記述するといった自動プログラミングに関する研究の基礎資料を作成する目的で，市販の本およびプログラム仕様書など7種の資料から以下のような語彙を中心に約600語を抽出した。

1) 処理，手続きを示す語，情報の単位，構成を示す語，入出力機器やフォーマットの指定に用いられる語は採用する。ただし，計算機のハードに関する語は除く

2) 問題に近い方の語は，一般的なものに限定し，個々の問題に固有のものはとらない。

3) プログラム以外の語は原則としてとらない。

構文情報は，1.1 の場合とほぼ同じ情報を与える。意味の分類は，1.1と異なり第2表に示すような分類を行なった。

第2表　自動プログラミング用語の意味分類表

0.	メタ言語
1.	（データ）
1.1	（ロジカルな単位）
：	：
1.111	データ（入力データ）
1.1111	データカード
：	：
1.3	情報内容
1.31	定数
：	：
1.5	（データ形式）
1.51	2進，10進，16進
：	：
1.6	（その他）
1.61	印（目印，マーク，記号）
：	
2.	場所
2.1	メモリ（記憶装置），コア
：	
2.5	位置
2.51	アドレス（番地）
：	
3.	入出力媒体
3.1	（入力専用）カード，紙テープ
3.2	（入出力両用）（磁気テープ）
：	
4.	（入出力編集）
4.1	ラベル
4.2	ヘッディング，見出し
5.	（抽象的な語）
5.1	順序（シーケンス，順）
5.11	アイウエア順，ABC順
5.12	上昇順（昇順，正順）
5.2	キー
5.3	単位〔処理の〕
5.4	範囲〔値の〕
5.5	エラー
5.6	条件
：	
6.	（データに関する述語）
6.1	レコード，アイテム
6.11	正しい，不正
6.12	古い，新しい
：	
6.2	値
6.21	正，負
：	
6.3	アドレス
6.31	連続
：	
7.	処理・操作
7.1	データ変更
7.11	四則演算
7.111	割る（割り算，除算）
：	
7.5	プログラム制御
7.51	反復
7.511	反復，くり返す

2. 辞書の情報空間

1節で行なった調査，分析の結果をもとにして，2種類（金属の加工，プログラム）について辞書の作成を行なった。

辞書の情報として，形態情報に見出しと読み（音形式），構文情報に品詞（活用型）と格支配，意味情報に意味分類コード，使用分野，区分コードを付加した。以後，とくに異なる部分を除いて，金属の加工の辞書について述べる。

2.1 形態情報

見出し—漢字仮名混り表記とする。複合語も見出しに採用する。用言は，その語幹（語形変化しない部分）のみとし，単品詞語は，1個の見出しとする。

読み—平仮名，片仮名，英字，数字，特殊記号からなる連糸は一括して「井」で表記し，漢字部分にのみ平仮名を与える。漢字1字ごとに区切り符号として，「/」を入れる。

2.2 構文情報

品詞—

1)	名詞	6)	連体詞
2)	動詞	7)	接続詞
3)	形容詞	8)	助数詞
4)	形容動詞	9)	接尾語
5)	副詞		

活用—

既報の「日本語の活用処理」[1]で分類した活用の型を動詞（35），形容詞（4），形容動詞（1）に分け，型の符号を与えた。

2.3 意味情報

使用分野— JICST シソーラスの主題分野を参考に，次の5つのカテゴリーを設けた。（対応するJICSTコードの頭文字）

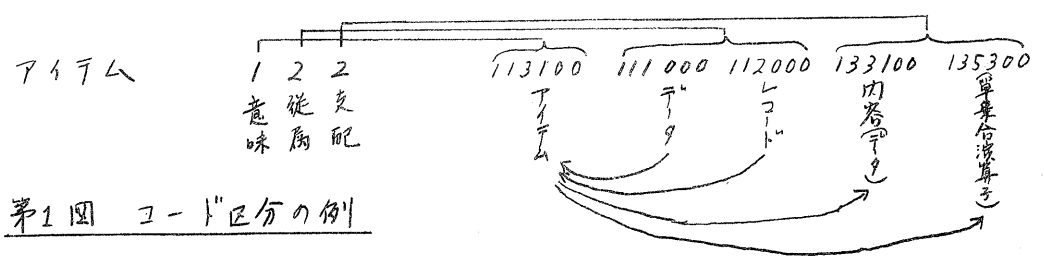
1)	化学	C	たゞし
2)	電気	E	MF04（材料力学等）→ 5
3)	金属	G	MF05（熱工学）→ 3
4)	機械	M	PA09（電磁気学）→ 2
5)	物理	P	とした。

コード区分— 意味分類コード欄の使用区分を3桁で表わした。

第1桁：見出し語の意味分類コードの数

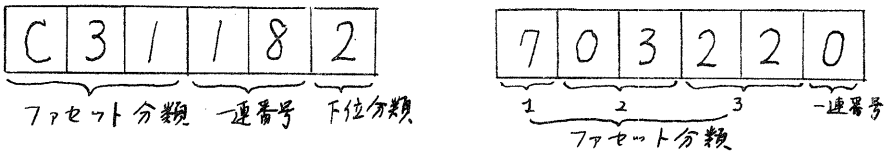
第2桁：見出し語の従属着たりうる語の意味分類コードの数

第3桁：見出し語の支配着たりうる語の意味分類コードの数



第1図 コード区分の例

意味分類コード — 6桁のコードからなり，辞書により構造が異なっている。複数個の意味分類に属すると考えられるものは，最大5個まで認める。その構造を第2図に示す。



(1) 金属の加工　　(2) プログラム

第2図　意味分類コードの構造

なお，プログラムの辞書では，前頁で示したように見出し語の従属者，支配者たりうる語の意味分類コードをその見出しの意味分類コードの右の余白を用いて与えた。

用言の格支配 — 見出しがとりうる格（最大3個）とその格がとりうる意味分類コードとの組み合わせを型で分類し，その分類番号を与えた．型は辞書と独立に型の表を作成する．格の表を第3表に，型の表の一部を第4表に示した。

第3表　格のコード表

1	無格	B	まで
2	係助詞	C	において
3	と	D	について，にかんして
4	が	E	にたいして
5	に	F	によって
6	を	G	とともに
7	で	H	として
8	へ	I	という
9	から	J	と
A	より		(31用，用言につく場合のみ)

第4表　用言の格支配の型の表

分類番号	型の数	支配しうる格の数	格	意味コード	格	意味コード	格	意味コード
1	1	1	4	N				
2	1	1	4	C, D, E				
30	2	2	4	B	5	A, B, F, G		
		2	4	H	5	H		
51	1	3	4	C	5	C	H	D
175	2	1	4	G0001, G0009				
		3	4	N	5	D, E, I	6	D, E
176	3	2	4	C, D, E	6	B, F		
		3	4	N	5	L4	6	A1032, C, D, E
		3	4	N	6	A1032, C, D, E	7	L4

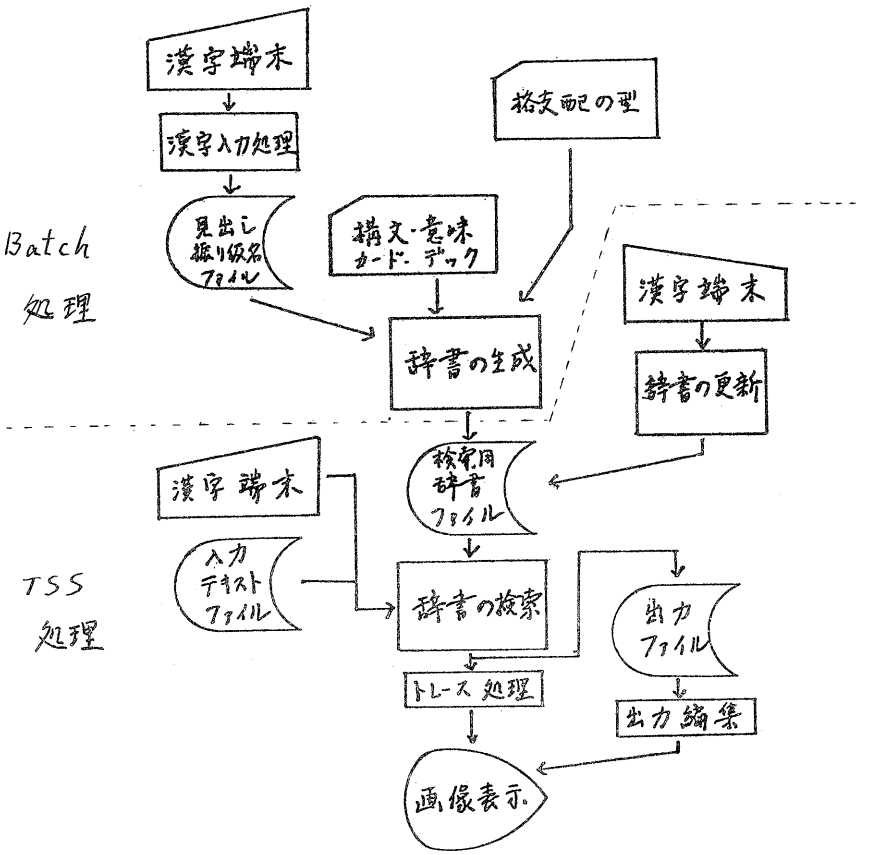
3. システムの構成

本システムは，第3図に示すように，辞書ファイルの生成，辞書の検索と表示出力および辞書ファイルの更新の3つの部分からなっている。生成は新しく辞書を作成する場合の1回だけであり，辞書データを大量に一括入力する。またカードイメージの入力といった点からBatch処理とした。検索と更新は，使用頻度が高い，入力テキスト，辞書ともにonlineのディスク・ファイルに記憶されている，操作を端末から行ない，結果を画像端末に表示させるといった点からTSS処理とした。なお，プログラミング言語は，いずれもFORTRANのBatch，TSSを用いた。また，漢字の表示には，PLOT10を用いている。

プログラムの大きさは，

	行
辞書ファイルの生成	264
〃 検索	522
〃 更新	523
表示（漢字）出力	465

である。



第3図　システムの構成図

3.1 辞書の生成

3.1.1 辞書データの入力

一連番号，見出し，読みは，漢字仮名混りであるため，TSS画像端末から漢字鍵盤入力プログラム（KGP）または，ローマ字による漢字入力プログラムのいずれかを用い，各項間に区切り符号を挿入した可変長レコードで入力する。漢字の内部コードは，電総研で共通に使用されている FONT 3000 符号である。この符号は，10進4桁で表現されており，その漢字パターンは，ストローク方式で表示される。なお，未登録漢字については，4,000 番以上のユニークなコードを与えた。その他の情報は，アルファニューメリックで記述されているので，一連番号とともにカードで入力する。

3.1.2 辞書の内部構造

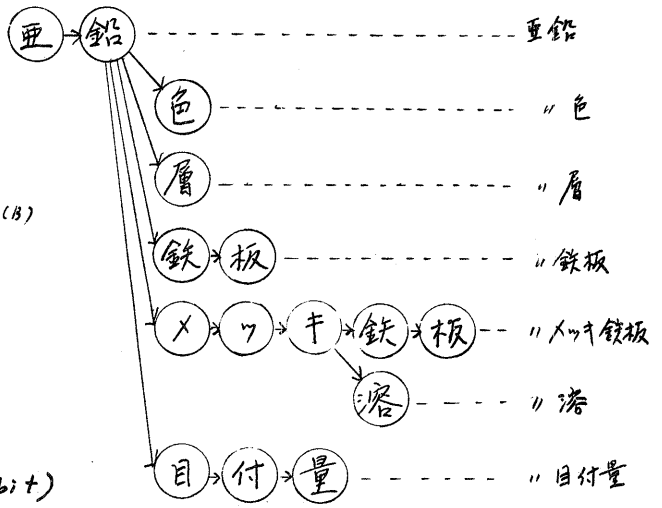
見出しは，第4図に示すように，文字を node とした Tree 構造で表現し，他の情報へのポインタを有する。

見出し以外の情報は chain table で表現した。Tree 部分と chain 部分は独立のファイルとし，ディスク上の online ファイルに蓄積されている。

各ファイルの大きさは，

	A (B)	B (B)
見出し	83	20
その他の情報	154	25
格支配ポインタ	2	2
格支配表	10	3

ただし
A：金属の加工
B：プログラム
(B)：Block = 320W (1W = 16 bit)



第4図　見出しの Tree 表現例

3.2 辞書の検索

3.2.1 テキストの入力形式

第5表で定義するような文節単位に区切り符号を有し，FONT 3000 コードで表現された online ファイル上に記憶されているテキストを用いる。

3.2.2 見出しの認定

文節を単位とし，語頭揃えによる完全一致および部分一致の全数検索を行ない，最長一致のものから順に，成功したものは，その情報を付加し，不成功のものは，見出しのみをファイルに出力する。

第5表　文節の構造

<文　　節>::＝<詞>|<文節><辞>
<　詞　　>::＝<第一種の詞>|<第二種の詞>
<第一種の詞>::＝<体言>|<用言>
<第二種の詞>::＝<副詞>|<連体詞>|<接続詞>
<体　　言>::＝<名詞>|<代名詞>
<用　　言>::＝<動詞>|<形容詞>|<形容動詞>
<　辞　　>::＝<助動詞>|<助詞>

3.3 検索結果の出力

3.3.1 検索表示

検索状況を逐次トレースして，TSS画像端末にアルファニューメリック（ASCII）で表示する。この例を第5図に示す。

3.3.2 ファイルへの蓄積

検索結果を online ファイルに蓄積する。

```
(  17)
      2160 2016
*******
      $MIDASHI--->2160/2016/
       FURIGANA-->@432.@409.@071.@436.
       IMICODE--->(M00000   )
       MINSHI=(01)     KATSUYO=(00)     SHIYO BUNYA=(0000)     KAKU SHIHAI=(0000)
      $MIDASHI--->%   /%   /
       FURIGANA-->
       IMICODE--->
       MINSHI=(02)     KATSUYO=(30)     SHIYO BUNYA=(0000)     KAKU SHIHAI=(0128)
                (  128)          (04)M                    (05)D,E,H,I              (06)D,E,H,I1

(  18)
       449
(  19)
      1554 1432 2487  962
*******
      $MIDASHI--->1554/1432/2487/0962/
       FURIGANA-->@413.@409.@071.@435.@486.@071.@468.@486.@071.@429.
       IMICODE--->(L40101   )
       MINSHI=(01)     KATSUYO=(00)     SHIYO BUNYA=(0000)     KAKU SHIHAI=(0000)
```

第5図　検索表示例

3.3.3 編集出力

出力ファイルとの検索結果に対し，TSS画像端末から編集パラメータを与えることによって，文字サイズ（3段階），横書き（標準），縦書き（文字を90°回転させて表示）等の処理をほどこし，漢字パターンに変換し，画面上に表示出力する。この例を第6図に示す。　§§§　検索結果出力　終わり　*§§§

さらに，検索された品詞数を各品詞ごとに第7図のように出力する。

第7図　検索された品詞数

名詞	= 0 4 6
動詞	= 0 3 4
形容詞	= 0 0 1
形容動詞	= 0 0 6
副詞	= 0 0 2
連体詞	= 0 0 7
接続詞	= 0 0 6
名詞兼助数詞	= 0 0 1

```
見出し     :  文字列
ふりがな   :
イミコード :  115100   030000   116000   135300   252000
品詞       :  名詞              活用     :  00
使用分野   :  104               格支配   :  000

見出し     :  文字
ふりがな   :
イミコード :  142000
品詞       :  名詞              活用     :  00
使用分野   :  100               格支配   :  000

見出し     :  文
ふりがな   :
イミコード :  115300   030000   116000   135300   252000
品詞       :  名詞兼助数詞      活用     :  00
使用分野   :  104               格支配   :  000

見出し     :  読込
ふりがな   :
イミコード :  702300
品詞       :  動詞              活用     :  26
使用分野   :  100               格支配   :  050
```

第6図　編集出力の表示例

3.4 辞書ファイルの更新

更新処理は，大量の場合，生成ルーチンにより一括処理を行なう。部分的な少量の更新では，TSS画像端末から会話形式で，見出し，その他の情報を追加，削除，変更の処理を行ない。ディスク上の検索辞書を更新すると同時に，マスター辞書である復元辞書の作成を行なう。

3.4.1 追加処理

この更新は，見出しが辞書に登録されていないとき，又は，情報のみを追加したい場合にこの更新処理を行なう。

3.4.2 削除処理

これは，見出しの削除を行なうもので，Treeファイル（見出し）およびchainファイル（その他の情報）から情報を削除する処理を行なうのではなく，Treeがもっているchainへのアドレスを削除することによって，この見出しを検索不可能とする。

3.4.3 変更処理

これは，Tree内に誤りが見出されたとき，あるいは変更を必要としたときに用いるもので，chainの変更は行なえない。

なお，onlineの更新では，ポインタの変更のみを行ない，garbage collectionは行なわない。これは，定期的に復元辞書を用いて，生成を行なう。また，各処理が完了したときに，その処理の前と後でのTreeとchainの大きさを表示出力する。これにより，ファイルの大きさ，定期処理の時期の判定が可能である。

5. あとがき

現在システムが完成したばかりであり，まだデータを流して，定量的な検索結果の評価はおこなっていない。この辞書への入力形式が分かち書きされたテキストでなければならないため，文節分かち書きのシステムを開発中である。このシステムは，本システムと結合して用いることを目的として，対象を金属の加工に限っている。また，検索で得られた自立語情報と既存の活用処理と結合することにより，付属語部分が認定され，文節の構造が明確となるだろう。さらに，格支配情報による構文分析，意味分類による類義語の判定，文間の結合関係の分析をすすめることにより，テキスト分析，自動抄録への応用をめざしている。

最後に，本研究の機会を与えて下さった石井ソフトウェア部長，島居宏次言語処理研究室長，また，辞書データの調査と作成，プログラムの作成を行なって下さった，IBSの木村睦子さん，佐々木幸博，松村有二の諸氏に感謝いたします。

参考文献

1) 坂本義行他：自然語プログラム仕様の計算機処理に関する基礎的考察，AL76-90
2) 坂本義行他：日本語のコンコーダンス，CL研究会資料2，1975.7.
3) 特許庁編：発明および実用新案の分類表
4) 坂本義行：辞書の構造－日本語の文節辞書，第11回情報技術科学研究集会，1974
5) 石綿敏雄：動詞を中心とした語彙の分類，国研報告51，1974
6) 国立国語研究所：分類語彙表，昭和40年
7) 坂本義行：日本語の活用処理，CL研究会資料，CL5-2，1976