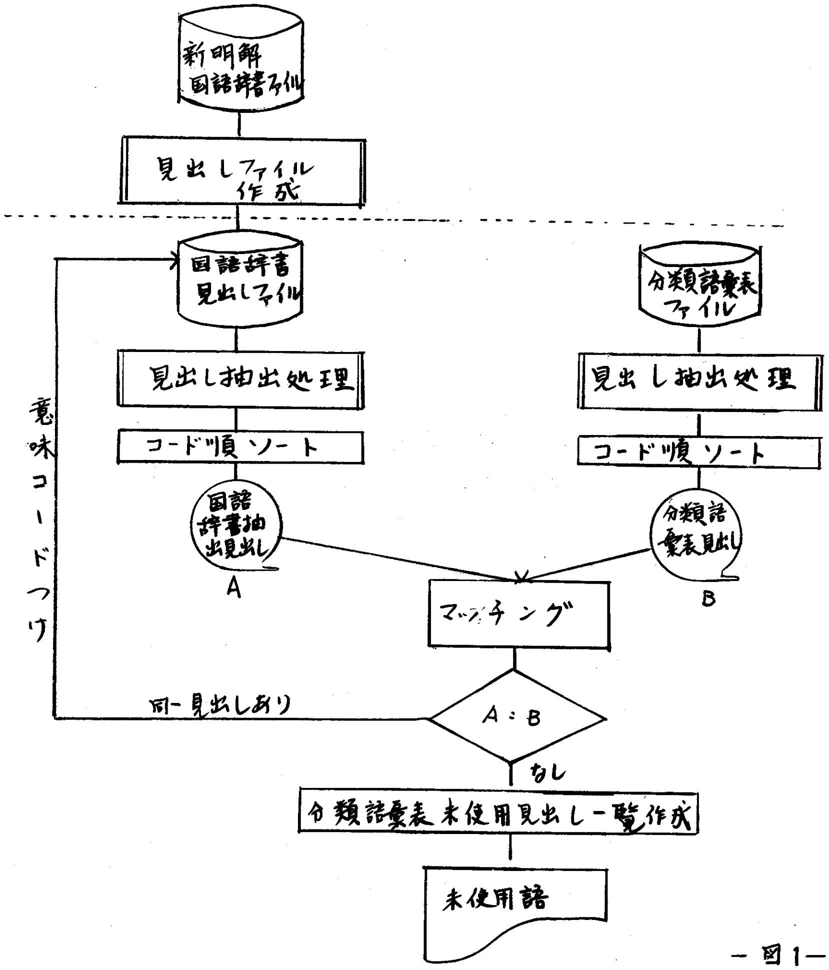
自然言語処理 27－1
（1981. 9. 24）

分類語彙表による国語辞書ファイルの意味コード付け

荻野綱男（東大）、荻野孝野（計量計画研究所）
淵一博、田中穂積、横山晶一（電総研）

1. 初めに

電総研では、レコード長以外何ら制約なしで入力された国語辞書ファイルから見出し部分を自動的に認定し、見出しファイルを作成した。これは、まず問題ごとにチェックプログラムを作成、データの特性を調べ、手がかりになりうる特定の符号を決めて、アクセント・活用語尾・品詞など見出し部分の要素を識別し、所定の形式に収めたものである。更に、こうしてでき上がった見出しファイルに、ファイル化された分類語彙表を用いて自動的に意味コードづけを行なった。これは構文解析等において意味が参照できることを目標にして、見出しファイルの充実をはかったものである。ここでは、この見出しファイルへの意味コードづけの作業内容と問題点などについて報告する（破線以下の部分）。



ー図1ー 作業の流れ

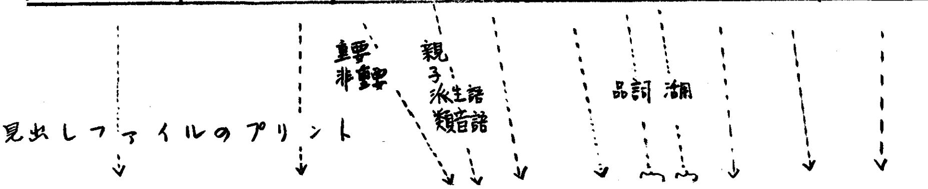
2. 各ファイルの形式とレコード数

2.1 見出しファイルの形式

—図2.1— 見出しファイル

かな見出し	正書法	区分1	区分2	ファイル名	行番号	品詞活用型	活用語尾	アクセント	分類番号	ポインタ
高千穂コード18文字	高千穂コード18文字					1品詞1活用	N個	N個	N個	
							N≦10	N≦5	N≦6	
9ワード	9	1	1	1	1	1	1	1	6	1

1レコード=32ワード
1ワード=36ビット



見出しファイルのプリント

```
 1 あ                           @OY 2  1 2000
 2 ア                           @OY 2  2  100
 3 あ                           @OY 2  6 1100       2
 4 あ             痾            @OY 2 10  100       2
 5 ああ                         #OY 1  1  900       1
 6 ああ                         @OY 2 11 1100       2
 7 ああ                         @OY 2 12 1100       2
 8 アーク とう    アーク燈      @OY 2 13  100       1  14600
 9 アーケード                   @OY 2 15  100       4
10 アース                       @OY 2 18  100       2  14620
11 アーチ                       @OY 2 21  100       2  14420
12 アーティスト                 @OY 2 25  100       2
13 アート                       @OY 2 26  100       2
14                ーディレクター @KO 2 27            6
15                ーペーパー    @KO 2 28            5
16 アーベント                   @OY 2 29  100       2
17 アーム                       @OY 2 31  100       2
18                ーホール      @KO 2 35            5  11840
19 アーメン                     @OY 2 36 1100       1
20 アーモンド                   @OY 2 37  100     204
21 アール                       @OY 2 39  100       2  11961
22 アール エイチ  Rh            @OY 2 40  100       5
```

2.2 見出し抽出テープの形式

これは、分類語彙表とのつき合せのため、見出しファイル図2.1の斜線部分を取り出して作ったテープである。図2.2の見出しは図2.1のかな見出しあるいは正書法の部分のいずれかか、それらの組み合せから作られる。

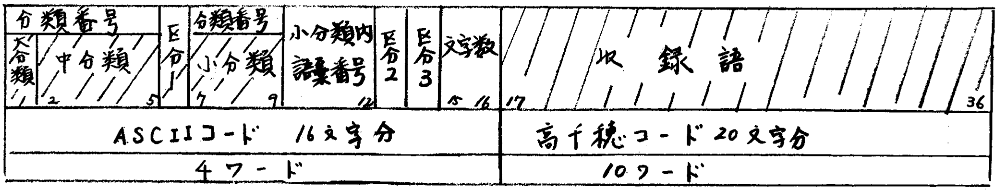
見出し 18文字	区分2	辞書見出しファイルのレコード番号
9ワード	1	1

図2.2 国語辞書見出し抽出テープ

国語辞書抽出見出しのプリント例

```
ア         あ          あ          痾     ああ          ああ          ああ          アークとう
アーク燈   アーケード  アース      アーチ アーティスト  アート        アートディレクタアートペーパー
アーベント アーム      アームホール アーメン アーモンド アール        アールエイチ  Rh
Rh因子     Rhマイナス  あい        愛     アイ          アイシェード  アイシャドー  アイバンク
```

2.3 分類語彙表ファイル



分類番号		区分1	分類番号	小分類内語彙番号	区分2	区分3	文字数	収録語
大分類 1	中分類 2 5		小分類 7 9	12			15 16	17 36
ASCIIコード 16文字分								高千穂コード20文字分
4ワード								10ワード

—図2.3 分類語彙表ファイル—

大分類・中分類 --- 分類語彙表記載の番号
区分1 --- 標題（0）か収録語（1）かの区分
小分類 --- 各分類番号内の段落番号
区分2 --- 元から入っている語（0）か校正による追加（1）か
区分3 --- 標題（0）、重要語（*）、それ以外（空白）

分類語彙表のプリント

```
0000810    11010100400030*01物（もの）
0000820    11010100400040 01物（ぶつ）
0000830    11010100400050 02物件
0000840    11010100400060 02物心
0000850    110200000000000002事項
0000860    11020100100010*02事項
0000870    11020100100020*02条項
0000880    11020100100030*02項目
0000890    11020100100040 02費目
0000900    11020100100050*01項
0000910    11020100100060 02前項
0000920    11020100100070 02別項
0000930    11020100100080 01条
0000940    11020100100090 03くだり
```

原本の、同義異表記、略語の並記、接頭・接尾のついた見出し語は、レコードを分け1レコード1語形で収録した。

予・余 → 予
余

停年（定年） → 停年
定年

（お）じいさん → じいさん
おじいさん

2.4 分類語彙表抽出ファイル

見出し 高千穂コード 18文字	分類番号 整数型	分類語彙表ファイルのレコード番号
9ワード	1	1

—図2.4 分類語彙表抽出テープ

辞書見出し抽出ファイルとのつき合せのため、図2.3の斜線部分を取り出して作ったテープである。

2.5 各ファイルのレコード数

表2.5 ファイル別語数

ファイル名	語数	備考
国語辞書見出しファイル	72747	
国語辞書見出し抽出ファイル①	129288	見出しファイルのよみ・表記を各々1見出しにたてたもの
〃 ②	124985	第1回マッチングで対応しなかった見出しについてまぜ書きを追加
分類語彙表ファイル	36776	
分類語彙表抽出ファイル ①	35954	元のファイルから標題語822語を除いたもの
〃 ②	8698	第1回マッチングで未使用のもの

3. 作業詳細

図1作業の流れに従って、順に作業内容を説明する。

3.1 国語辞書見出し抽出処理

これは、国語辞書見出しファイル（図2.1）の斜線部分を使って、分類語彙表とのマッチング用辞書テープ（図2.2）を作るものである。見出し部分のひらがなと漢字コード以外の特殊符号を除いたりして、通常の表記形式の見出しを作るのが主な作業である。見出し区分別に以下のような処理を行なう。

① 親見出し

(1.1) 親見出しのかな見出し部分を取り出す。

かな見出し部分に含まれる特殊符号（・（活用語幹を示す）空白（語の区分を示す））を除いた部分を見出しとする。

(1.2) 親見出しの正書法部分を取り出す。

正書法部分とは辞書中［ ］で囲まれた部分である。送りがなを略記したかっこは規則を作って展開する。（小文字はひらがなを示す）

主なパターン	展開見出し
A（a）	→ A，Aa
A（a）B	→ AB，AaB
AB（b）	→ AB，ABb
A（a）B（b）	→ AB，ABb，AaBb

特殊パターン（千例）

A(a)B(b)C(c)D(d)

→ AaBbCcDd，ABCD

例.

入(れ)代(わ)り立(ち)代(わ)り

(1.3) 親見出しのかな見出しと正書法部分を組み合せて、まぜ書き見出しを作る。

分類語彙表には、まぜ書きの見出しがあるが、これは(1.1)かな見出し部分や(1.2)正書法部分で作った見出しとは対応しない。そこでかな見出し部分と正書法部分を組み合せた見出しを作る。

例. かな見出し　ふき　さらし
正書法　吹（き）晒（し）→（（）とり）吹き晒し → （まぜ書き）ふき晒し　吹きさらし

② 子見出し

子見出しで［一　］で表現されているものについて一に親見出しの正書法部分（なければかな見出し部分）を代入して見出し語を作る。

例. 親　あいだ［間］
子　［一柄〈がら〉］ →（合成）間柄〈がら〉 →（不要連系除く）間柄

③ 派生語

(3.1) 派生語1は辞書で　一さ　の形になっているので一に親見出しの正書法部分で語幹にあたる部分（見出しの・以降の連系を正書末尾から除いた部分）を代入して見出し語を作る。

例. 親　あたたか・い［暖（か）い］→（語幹取り）暖（か）→（合成）暖(か)さ →（（）展開）暖かさ　暖さ
派生1　一さ

(3.2) 派生語2はそのまま取り出す

例. 親　きり　はな・す［切（り）放（す）］
派生2　切放せる → 切放せる

④ 類音語

親見出しと表記の異なるものだけ取り出す。

例　親　かん　かい　→　かんかい
　　類音語　官界　→　官界
　　　　　　官海　→　官海

以上①～④のうち(1.3)については一回目のマッチングで分類番号のつかなかった辞書見出しに対して行なったものである。

次に、①～④の処理をして一つの見出しがどのように展開するか、具体例でみてみよう。

例．たま　むかえ［魂迎(え)］　　たべ　もの［食べ物］
　→　たまむかえ（(1.1)かな見出し）　→　たべもの（(1.1)）
　　　魂迎え（(1.2)（）とり）　　　　　食べ物（(1.2)）
　　　魂迎　（　〃　）　　　　　　　　たべ物（(1.3)）
　　　たま迎え（(1.3)まぜ書き）　　　　食べもの（(1.3)）
　　　魂むかえ（　〃　）　　　　　　　　［一　屋］
　　　　　　　　　　　　　　　　　→　食べ物屋（②）

3.2　分類語彙表抽出見出し

3.1同様、分類語彙表ファイル（図2.3）を使って、斜線部分から必要情報を取り出すとともに、収録語から特殊符号など不要連系を除いて、国語辞書見出しファイルとのマッチング用テープを作るものである。

分類番号は大分類から小分類までとる。収録語は特殊符号によって次の処理をする。

① 接辞の類についた 一 は除く。

② 活用語についた ／ は除く。見出しのとり方は語尾によって異なる。

- 動詞、形容詞は ／ の後ろ部分を取る。

例．組み替え／組み替える　→　組み替える

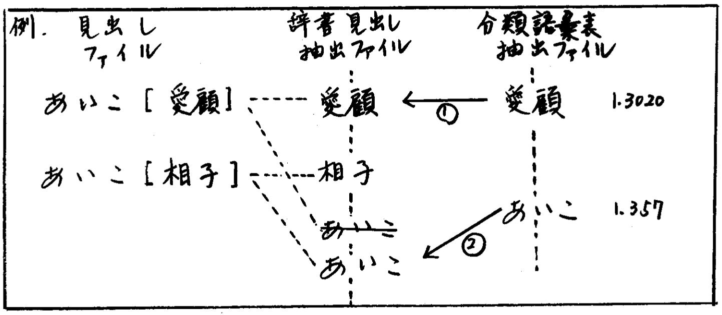
- 副詞、形動語幹、サ変・下一段動詞は ／ の前後を取る。

例．思い思い／思い思いに　→　思い思い
　　　　　　　　　　　　　　　思い思いに
　　四角／四角い／四角な　→　四角
　　　　　　　　　　　　　　　四角い
　　　　　　　　　　　　　　　四角な
　　存じ／存ずる　　　　　→　存じ
　　　　　　　　　　　　　　　存ずる

3.3　ソート

3.1、3.2で作成した見出しテープをソートし、マッチング用テープを作成する。

分類語彙表では漢字表記のものは、大体、漢字で表記されている。一方、国語辞書の方は、よみ（見出し部分相当）も見出しとしてたてたため、ひらがな表記の同音異語が多い。そこで、ソートは漢字コードが先にくるような降順で行なった。これによって、マッチングの時は、漢字コードで始まる見出しが先に対応づけられる。更に一度対応づけられた見出しと同一レコード番号の見出し（国語辞書見出しは1レコードから複数見出しに展開）をマッチング対象から外すと、ひらがな表記の同音異語は減ることになる。



①のマッチングによって［愛顧］のよみにあたる「あいこ」が消えて、二番目の「あいこ」だけになって一対一対応となった。一応、辞書見出しファイルの二つの見出しに意味コードがついた。ただし、これは、あくまでも分類語彙表の方に同語異表記なしという仮定に立つにすぎない。

3.4　マッチング

漢字コードを降順ソートした辞書見出し抽出テープと分類語彙表テープをつき合せ、一対一に対応づけられたものについては辞書見出しファイル（図2.1）の分類番号の欄に書き込む。1見出しにつき6個までつけられる。

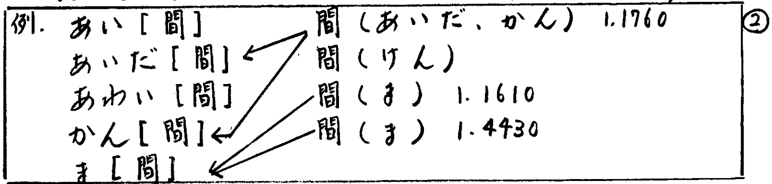
なお、このマッチングは二度行なった。二度目は一度目の結果を検討し、抽出テープ作成手順及びマッチングプログラムの一部に変更を加えたものである。

主な改良点は次のようなものである。

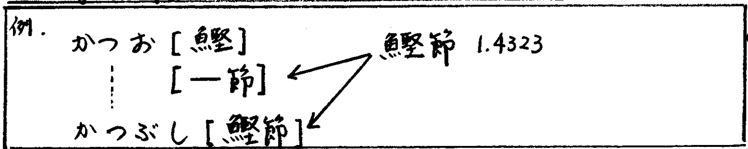
① 見出し抽出処理に3.1-⑬のまぜ書き処理を入れた。

② マッチングする時、分類語彙表の方によみがついていれば、よみも使った。これは、辞書見出しが漢字表記の同形異語の時、分類語彙表見出しの方によみがあれば辞書見出しのよみとのつき合せをするというものである。

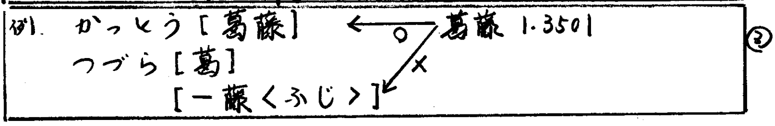
③ 見出し区分（親・子・派生など）による段階づけを使った。
親見出し（あるいは類音語）と子見出し（あるいは派生語2）が同形異語の時、同じ語とみなし一つの分類番号を二つの辞書見出しにつけたものである。



②よみを使ったマージ



③で正しくついた例



③でまちがった例

3.5　作業上の問題点

3.1〜3.4の作業内容において、特に問題となった事項について述べる。

① 子見出しの作成

子見出し・派生語1は3.1-②、③で述べたように、代入処理を行なうが、親・子・類音語の出現順によってはどれを代入するかが問題になった。見出し区分の異なるものの並び方のパターンを出力してランクづけしてみたところ、同一形式でも画一的に決定できないものもあった。

例. ④やさしい　　　④′ かって[勝手]　　　⑤ おろか[愚か]
　Ⅱ[優しい]　　　　Ⅱ　台所　　　　　ϕ－さϕ
　②[易しい]　　　　②－な－に　　　　[－しい]
　ϕ－さϕ　　　　　ϕ－さϕ　　　　　ϕ－さϕ

親→子→派生
↘類音語↗

② 複数見出しの扱い

一対一に対応したものをマッチングしたものとみなすので、辞書あるいは分類語彙表見出しのいずれかで複数個の同形見出しができたものが問題になる。

表3.5　同形見出し

	分類語彙表ソート		辞書見出しソート	
	1回目マッチング用	2回目マッチング用	1回目マッチング用	2回目マッチング用
同形語なし	29480語	6853	95008	103623
2	2655組	663	4568	6263
3	301	118	879	996
4	47	27	301	306
5	9	7	124	120
6	2	1	51	46
7	0	0	28	24
8	1	1	20	17
9	1	1	18	12
10	0	0	5	4
⋮	⋮	⋮	⋮	⋮

分類語彙表の同形見出し例

34336	あどけない	330401002
35048	あどけない	334201001
26990	あばく	212101003
28234	あばく	215531005
604	あべこべ	111301003
31942	あべこべ	311201006
25608	いか	156601004
31720	いか	310001004
30578	いかれる	235301013
31638	いかれる	258301005
31725	いかん	310001004
32335	いかん	313301004
6080	いくら	119501007
36663	いくら	431501003
27157	いける	213101001
27523	いける	215161002
29019	いたむ	230001002
29242	いたむ	230201013

動物
いかー (にくなど)

1 いか (名)
~~2 いか(以下) (造語)~~
~~2 いか(医科) (名)~~
2 いか(医家) (名)
2 いか(異化) (名)ーする
2 いか(易化) (名)ーする

品詞×
表記×

連形
1 いかん(名)
1 いかん(名)
2 いかん[衣冠] (名)
⋮

表記×
品詞×

辞書見出しあるいは分類語彙表見出しで複数見出しになるものの関係は次のいずれかである。

④ 辞書に複数見出し ―― 分類語彙表に複数見出し
⑤ 〃 複数見出し ―― 〃 1見出し
⑥ 〃 1見出し ―― 〃 複数見出し

⑥の場合は、辞書見出しが複数個の意味を持つとみなし、辞書に複数個の分類番号をつけた。④、⑤が問題になる部分であって、特にひらがな表記のものについては、正書法部分から作った見出しかかな見出し部分から作った見出しかというような表記の型や、品詞等の情報を使ってもあまり有効なきめ手にならない。この部分については分類語彙表にあっても分類番号がつかないまま残った。

4. 結果の検討

4.1 数量的結果

表4.1 マージの結果

項目	語数	元になる見出しに対する割合
辞書見出しファイルに分類番号がついた語数	26834	37 % (72747)
〃 のうち親見出しについた語数	25241	43 (58409)
分類語彙表のうち使われた語数	29291	81 (35954)
〃 使われなかった語数	6820	19 (35954)

4.2 分析

分類語彙表の未使用見出しを検討してみると以下のような理由があげられた。

㋑ 辞書見出しが複数見出しになって一対一対応にならない。

例. にせ 1.1040 ---→ にせ [贋]
にせ [二世]

㋺ 辞書見出しに対応するものがない。

例. これら 1.1000
どれどれ 1.1000

㋩ 辞書見出しと分類語彙表で表記が異なる。

例. 椿事 1.1010 珍事
奇蹟 1.1010 奇跡

㋥ 辞書見出しの子見出し、派生語によみがついていないので対応できない。

例 かかわり 1.1110 拘わり (派生2)
かかりあい 1.1110 掛かり合い (子見出し)

分類語彙表で未使用のものの例.

```
1100010010010*01こ       辞複
1100010010030 03これら   辞ナ
1100010010040*01そ       辞複
1100010010050*02それ     〃
1100010010060*02あれ     〃
1100010010070*01か       〃
1100010030020 01の       〃
1100010030040*01こ       〃
1100010040010*02ほう     〃
1100010050020 04どれどれ 辞ナ
1100010050030 03何くれ   辞表記[なにくれ]
1100010060010*04おのおの 分[各]が辞の表記とマッチ
1101010010170 02一端     不明
1101010020010*01件       辞複
1101010020040*02本件     辞ナ
1101010030110 02椿事     辞表記 [珍事]
1101010030130*02奇蹟     辞表記[奇跡]
1101010040060 02物心     辞複
1103010010060*01真       辞複分複
1103010010100*01実       辞複
1103010020070 02枢密     辞ナ
```

一部について、左のようなチェックをしてみたところ、

㋑複数見出し 62%
㋺辞書にない 17%
㋩・㋥ 表記の違い 15%
その他 6%

であった。㋑、㋺、㋩については自動的処理では不可、㋥については更に処理可能(5-②)である。

5. 今後の処理

見出しファイルに分類番号のつかなかったものについて、4.2のような理由があげられたが、更に精度を上げるとすれば次のようなことが考えられる。

① 派生語で分類番号のついてないものでも親についていれば、分類番号の1けた目(品詞相当)を書き替え、流用する(分類番号2けた目以降の一致仮定)。

② 子見出し・派生語などに自動的によみをつけて、マッチングの処理を行なう。

③ 自動的分類番号づけの誤り部分及び4.2−㋑,㋩によるもれを人手によってチェックする。

なお、分類語彙表のすべての語が使用されたとしても、辞書見出しのすべての見出し(子見出しなど含む)の49%しか分類番号がつかない。