自然言語処理 57－3
（1986. 9. 19）

意味的なつながりを考慮した接尾語辞書の作成について

稲永紘之　　新谷隆之

（九州芸術工科大学）　（日本ユニバック）

カナ漢字変換をはじめとするカナ文字文の機械処理システムにおいては、未登録語の処理は同形異義語の処理とならんで重要な研究テーマである。

どのように辞書の見出し語を増やしても、派生語の全てをカバーすることは不可能であり、接頭語辞書と接尾語辞書の量的・質的充実を図る以外に未登録派生語の処理方法は考えられない。

日本語においては、文節間のみならず単語を構成する成分の間においても係り受けの存在が認められる。

本稿では、接尾語の中でも圧倒的に多い名詞の接尾語を中心として、受けの成分である接尾語に対して、係りの語をその漢字表記の意味で分類し接続の条件として与えた接尾語辞書の作成手順について述べる。

A SUFFIX DICTIONARY UTILIZING THE MEANINGS OF THE STEM OF DERIVATIVES AS THE CONNECTION RULE TO PROPER SUFFIXES

Hiroyuki INANAGA and Takayuki SHINTANI

Kyushu Institute of Design, Nippon UNIVAC Kaisha
4－9－1, Minami－ku, Fukuoka－shi, 815, Japan

Along with the homonym processing, it is crutial to study how to handle unknown words like derivatives in machine processing of Japanese sentences including Kana－Kanji translation. Even massive addition of headwords does not ensure the full coverage of derivatives. Nothing but an improvement of prefix/suffix dictionary both in quality and quantity could solve the problem.

Applying Kakariuke relation to the intrastructure of derivatives, this paper describes the way to construct a suffix dictionary which utilizes the meanings of the stem of derivatives as the connection rule to proper suffixes.

1．まえがき

近年仮名漢字変換をはじめとする日本語処理システムは急速な勢いで普及しており、入力方式も単漢字変換から文節変換を経て、今や連文節変換、文章一括変換へと進歩してきた。

当面は、この市販のワープロ、パソコンに見られるように、大容量の辞書によらないシステムが発展すると思われるが、辞書の内容が十分とはいえないために文節間、複合語内の論理的意味的つながりを考慮しているとはいえず、特に、派生語を含んだ文節の変換精度の悪さが目立つ。

将来的にはメモリーをはじめハードウェアの低価格化が進み、数十万の見出し語を有する辞書がこれらのシステムでも搭載可能になると思われるが、その意味で本格的な辞書の整備は急務である。

未登録語としての派生語の処理のためには、その派生語を直接自立語辞書に登録するのが最も簡単な解決策であり効果も大きい。しかし派生語はその数が膨大である。しかも基本的な単語としての地位が確立しておらず、単語というよりはむしろ言語表現の一形態と考えられるような語が少なくない。このため派生語をすべて登録しようとするのは無理であり無駄である。

未登録派生語（名詞）の処理方法の一つとして接尾語とそれが係る名詞あるいは用言を標準の格助詞と共に標準形で係り受けの句として辞書に登録しておいてこれを利用する方法についてはすでに報告した[1]のでここでは述べない。

ところで派生語を構成する末尾の語は、それ以前の単語（群）の接尾語として機能していることが多く、両者の間には一種の係り受け関係が成立している。

そこで、この係り受け関係を分類整理して予め辞書に登録しておけば、同形異義の単語群と接尾語群の中から正しい組み合わせを選ぶのに役立ち派生語の漢字変換の精度の向上を図ることができる。

本稿では、まず我々のいう派生語を定義し、未登録語となりやすい派生語のタイプを示した後、派生語を含んだ１０万７千余語の自立語辞書から上記のような係り受け関係の存在する派生語名詞を機械的に抽出し、それから目的の接尾語辞書を作成した手順について述べる。また、係りの語の品詞／意味分類別にどのような接尾語が付きやすいかの統計も取ったので併せて報告する。

2．用語の定義と派生語の構造

国文法用語の定義は解釈の違いから諸説があり定まっているわけではない。

ここでは、まず本稿で用いる派生語及び接尾語の定義を行ない、次に派生語の構造について考える。

2．1．　用語の定義

国文法では、意味の面からも形態の面からも独立していて言葉の基本単位となるものを自立語という。名詞、代名詞、動詞、形容詞、形容動詞、副詞などである。

一方独立して一語を形成せず、造語成分として作用するものを接辞といい、いわゆる接頭語・接尾語がこれに相当する。

この接辞と自立語からなる二次語を派生語といい、例えば「ご指導」、「美しさ」をあげることができる。これに対して、自立語（または自立語に準ずるもの、例えば形容詞の語幹）を組み合わせてできた二次語を複合語という。
「社会運動」、「望み薄」、「飛び跳ねる」などである。

本稿では派生語の定義を拡大して、上に述べた複合語のうち、分かち書きされる可能性がなく、また、発音するときには一まとめに発音されて間にポーズが入らない複合語も派生語という。そして、派生語の末尾の自立語または接辞をその派生語の接尾語といい、派生語のうちで名詞であるものを派生語名詞という。

接尾語の付いた派生語では「判断力」のように自立語「判断」が接尾語「力」を修飾している場合が少なくない。これはいわば単語内の係り受けである。そこで、このような修飾部を接尾語に係るという意味で係り語という。

派生語「合格者名」では接尾語「名」に対して「合格者」が係り語である。ところが、「合格者」の他にも「失格者」、「応募者」など多数の自立語が「名」と結び付いて派生語名詞となる。これらの派生語に共通している末尾の語「者」はまた一つの接尾語であり、かつ、接尾語「名」に係る真の係り語であるということができる。

「合格者」、「失格者」、「応募者」などをすべて接尾語「名」の係り語として登録せずに、真の係り語

「者」（名詞の意味分類としては人間のコードを与える）のみを接尾語辞書の見出し語「名」に登録することによって派生語内での係り受け関係が記述でき、かつ派生語処理用辞書群のトータルのサイズも小さくできる。

2．2．　派生語の構造

次に本稿で扱う接尾語の付いた派生語を語構成の面から考えて、未登録語となりやすい主なものを列挙すると次のようになる。

（1）　狭義の接尾語が付いたもの
　　感情的、金次第、若さ、先生達
（2）　サ変動詞語幹＋名詞→名詞
　　担当者、終了後、点滅器、連絡船
（3）　名詞＋名詞→名詞
　　郵便局、西洋館、人事部、有名校
（4）　名詞＋動詞連用形→名詞、サ変動詞
　　水洗い、山登り、刑務所帰り、北向き
（5）　動詞連用形＋名詞→名詞
　　貸し料、降り口、気取り屋、上げ幅
（6）　動詞連用形＋動詞→動詞
　　飲み続ける、歩き回る、言い兼ねる
（7）　動詞連用形＋形容詞→形容詞
　　歩き易い、聞き苦しい、聞き取り難い

このうち、普遍的にある特定の品詞に付く接尾語である6と7、また、1と2の大部分の接尾語は我々の名詞の意味分類には直接関係しないため、残りのタイプの接尾語とは違った辞書構成がなされる。

3．　接尾語辞書の構造

接尾語辞書作成に当たっては辞書のデータ構造を考えるのは当然のこととして、その他に

（1）　どのような語をどれくらい辞書に登録しておけばよいか（自立語辞書に派生語に含まれた形で入れておいた方がよいものもある。）
（2）　他の機械辞書との関わり方（使用の優先順位など）をどうするか

を考慮しなければならない。

ここで、1については、2．2．で述べた接尾語のうち、普遍的にある特定の品詞に付くものは、その品詞や活用形ごとに接尾語をまとめた辞書群を作る。また、特定の語にしか付かないものは接尾語辞書の見出し語とはせず、派生語そのものを自立語辞書に登録する。その他、使用頻度の高いと思われる派生語は、その接尾語が接尾語辞書の見出し語であろうとなかろうと自立語辞書の見出し語として登録する。

2については、我々のシステムでは漢字情報を用いた係り受け辞書との競合がある。これについては、直接本稿の論議とは関係がないので、稿を改めて述べたい。

個々の接尾語については、それが付く自立語の品詞や活用形、またそれが付くことによってできる派生語の品詞はどうなるかということの他に、

- 意味的にどのような範疇の語に付くことが多いのか
- その接尾語が付くことによってできる派生語の意味はどうなるか

ということなどを考慮しなければならない。このためには、自立語辞書の各見出し語の意味分類がなされている必要がある。

我々の自立語辞書では、名詞についてはそれを３５種類に分類し、更に各名詞を表記する漢字の意味による細分類を行なった情報が利用可能であるので、それを用いることにした。

4．自立語辞書

接尾語辞書作成に当たって使用した辞書は芸工大で作成した自立語辞書（ＫＩＤ－Ｊ８６）で、１９８６年８月現在、登録語数１０７，５２２語、そのうち名詞が９３，３８１語であり、旧版のＫＩＤ－Ｊ８２を大幅に増補改定したものである。ＪＩＳの第１水準以外、第２水準の漢字も用いて漢字表記の充実を図っており、多数の派生語を含んでいる。

自立語辞書は１レコード１２８バイトの順編成ファイルであり、カナ見出しの降順に配列されている。

各見出し語には、カナ見出しと、漢字表記、漢字表記に対する切れ目という欄を設け、漢字に読み仮名を与える働きをさせている。

用言については、その語幹（上一段、下一段動詞にあっては活用語尾の不変化部分を含む）のみをカナ見出しに与えている。

表1は、自立語辞書で用いている名詞の分類コードとその意味である。

コード	意　味
1	他の３４のいずれにも属さないもの
2	時、方向など、副詞的用法のあるもの
3	形容動詞的に使われるもの
4	人間、神、仏、霊魂、妖精など
5	動物名、動物、細菌
6	植物名、植物、植物の一部分
7	鉱物名、鉱物、液体、自然物など無生物
8	飲食物、料理など
9	人間や動物の体、体の一部分
A	衣服、織物、装身具、履き物など
B	家、ビル、橋、堤など建造物
C	道具、日用品、簡単な機構の器具類
D	書籍、評論、文章、詩歌など
E	乗り物、交通機関
F	病気、怪我、傷
G	弾く楽器、楽器一般
H	吹く楽器
I	打楽器
J	生物、法律（～法）と表記されるもの
K	貨幣、金銭、給料、財産、経済関係の語
L	建物の一部分や部品
M	複雑な機構の機械類
N	気体、雲、霞、自然現象の語
O	音楽、曲、歌、芸能関係の語
P	時間、期間
Q	衣服、織物、装身具の一部分
R	スポーツや遊びの道具、用具
S	スポーツ、遊び
T	時刻、時期
U	人間が自然物に手を加えてできたもの
V	部屋、室など建物の内部の場所
W	世界や活動分野など抽象的な場所
X	団体、組織、人の集まり、展覧会など
Y	光景、様子、状態、形式
Z	具体的な場所

表1　名詞の意味分類とコード

５．接尾語辞書の作成

接尾語辞書は、そのプロトタイプが、４章で述べた自立語辞書から自動生成される。最終的には、やはり辞書の常として人手による確認とチエックを経なければならない。

５．１．接尾語辞書の自動作成

名詞の意味分類を用いた、派生語名詞処理用の接尾語辞書を作成するために、次の条件を適用して、自立語辞書からデータとしての派生語を抽出した。

（１）名詞であること
（２）当て字でないこと
（３）形容動詞的に使われる名詞ではないこと
（４）末尾の語が目的格になっていないこと
（５）全体の漢字数が３文字以上であること
（６）係りの語の漢字数が２文字以上であること
（７）末尾の語の漢字数が２文字以下であること

次に、抽出された派生語の係りの語それぞれについて、品詞／意味分類情報を付加するため自立語辞書を検索した。抽出された派生語が接頭語＋自立語の組み合わせであった場合、自立語辞書に接頭語は登録されていないので、係りの語は見つからず、そのような派生語は、この段階で除かれた。

ただし、係りの語が用言の場合、自立語辞書は用言の語幹のみを持つので係りの語が自立語辞書に見つからないこともあり得る。従って、自立語辞書に同一の漢字コードが見つかり、かつ仮名見出しが異なる場合には、動詞なら連用形、形容詞なら連体形の活用語尾を自立語仮名見出しに付けて、係りの語の仮名見出しと等しければ、その自立語の品詞情報を係りの語の品詞情報とした。

また、自立語辞書に見つかった係りの語が同じく派生語である場合は、その派生語の接尾語部分を元の接尾語の真の係り語と見なし、抽出した派生語内の係りの語を真の係り語の情報で置き換えた。

次に、このようにして収集した派生語を、接尾語および係り語をソート・キーとして分類した。

最後に、接尾語ごとに、それを含む派生語の出現頻度と、係り語の品詞／意味分類コードの出現頻度を取

り、各接尾語に対する複数個の同一係り語は、出現頻度を付して一つレコードにまとめた。

作成した派生語名詞処理用の接尾語辞書は、接尾語ごとに、

（1）その接尾語自体の辞書情報を持つレコード

（2）係り語のレコード（一つ以上）

（3）係り語の品詞／意味分類の出現頻度を示すレコード

からなる。

図1は、派生語名詞処理用の接尾語辞書のフォーマットである。

作成結果をまとめると表2のようになる。

辞書	レコード項目	レコード数
自立語	名詞	93381
	接尾語付き派生語	31164
接尾語	接尾語	2090
	平均異なり係り語	12.31

表2　派生語名詞処理用接尾語辞書作成結果

5．2．品詞別接尾語辞書の作成

2．2．で述べたように、接尾語の中には、名詞に対する、「次第」、「だらけ」のように、ある特定の品詞の語に、普遍的に付く、あるいは付くことが非常に多いものがある。。また、動詞の連用形に付いて複合動詞を作る動詞など、名詞の意味情報が使えないものがある。それらは、品詞別に接尾語辞書を作成した。また、各接尾語に対しては、用例を参考にして、接続の強さを3段階に分けて与えた。

係り語の品詞と活用形	見出し語数
名詞	1770
動詞連用形	1270
サ変動詞語幹	300
形容動詞語幹	50
形容動詞的に使われる名詞	43
形容詞語幹	36

表3　品詞別接尾語辞書

タイプ1（接尾語）レコード

見出しID	仮名見出し	見出し切れ目	漢字コード																	
CNSFX1	接尾語 仮名見出し	接尾語 仮名見出し 切れ目	接尾語 漢字コード	小文字数	同形順位							名詞分類1	代表漢字数	名詞分類2	固有名詞	助数詞	人称代名詞	指示代名詞		レコード番号

タイプ2（係り語）レコード

見出しID	仮名見出し	見出し切れ目	漢字コード																			
CNSFX2	係り語 仮名見出し	係り語 仮名見出し 切れ目	係り語 漢字コード	小文字数	同形順位	サ変動詞	その他動詞	形容詞	形容動詞	副詞	連体詞	名詞分類1	名詞分類2	名詞分類3	名詞分類4	代表漢字数	名詞分類5	固有名詞	助数詞	人称代名詞	指示代名詞	係り語出現頻度	レコード番号

タイプ3（品詞／意味分類）レコード

見出しID	仮名見出し	漢字コード			品詞出現頻度						名詞出現頻度								レコード番号
CNSFX3	接尾語 仮名見出し	接尾語 漢字コード	小文字数	接尾語出現頻度	サ変動詞	他の動詞	形容詞	形容動詞	副詞	連体詞	普通名詞意味分類 A～Z 0～9				固有名詞	助数詞	人代名詞	指代名詞	

図1　接尾語辞書フォーマット

6．実験結果と考察

ここでは自立語辞書から自動作成された名詞の接尾語辞書を中心に、その特性と問題点について述べる。

表4は、抽出された接尾語を、その出現頻度順に並べたものである。個々の接尾語の係り語は、意味的なコードで分類したくらいでは、重なりが多く、その特徴が捉え難い。

ところで、名詞は普通、漢字表記され、漢字は意味を担っている。このことに注目し、我々はすべての名詞を、その語構成（語末の漢字が意味の主体をなすかどうか）と語末から数えて何文字の漢字が意味の主体をなすかで分類している。そこで、このような漢字情報を用いれば、係り語に対して、更に詳しい分類を行なうことができる。

No	出現頻度	接尾語	係り語品詞／意味分類
1	979	者	サ変，3，4，D，E
2	480	性	サ変，形動，7，Z
3	366	品	動詞，形動，3，Z
4	352	機	動詞，固名，X，Z
5	345	室	サ変，4
6	344	法	サ変，形動，X，Z
7	333	部	サ変，O，S，Z，9
8	332	地	動詞，Z
9	313	費	サ変，X
10	281	日	動詞，T

表4　出現頻度の高い接尾語

No	接尾語	係り語	例
1	費	送，造，務	輸送費，建造費
2	名	員，者，長	社員名，合格者名
3	品	送，蔵，製	郵送品，愛蔵品
4	者	学，論	科学者，筋論者
5	室	長	学長室

表5　出現頻度の高い係り語と接尾語の組合せ

No	頻度	係り語	係り語品詞／意味	例
1	40	長	名詞（4）	院長室
2	10	事	名詞（4）	理事室
3	8	員	名詞（4）	職員室
4	7	務	名詞（1）	事務室
5	6	官	名詞（4）	教官室
6	6	議	名詞（1）	会議室
7	6	等	助数詞	一等室
8	6	理	サ変動詞	管理室
9	5	児	名詞（4）	乳児室
10	4	習	サ変動詞	実習室

表6　接尾語「室」への係り語の品詞／意味分類

No	頻度	係り語	係り語品詞／意味	例
1	13	造	サ変動詞	改造費
2	13	送	サ変動詞	移送費
3	13	務	名詞（1）	医務費
4	11	料	名詞（8、C）	食料費
5	10	備	サ変動詞	軍備費
6	8	築	サ変動詞	建築費
7	7	入	サ変動詞	加入費
8	7	査	サ変動詞	検査費
9	6	助	サ変動詞	援助費
10	6	理	サ変動詞	修理費

表7　接尾語「費」への係り語の品詞／意味分類

表5は、係り語の語末の漢字情報を基にして、出現頻度の高い係り語と接尾語の組み合わせを並べたものである。更に表6及び表7には接尾語「室」と「費」について、それに係る語の品詞と意味分類を出現頻度順に示した。

例えば、接尾語「室」は、人を表わす「長」（名詞の分類コード4）に連接して、自立語辞書の中で、40回現われたことを意味する。「長」と名の付くような人は、専用の部屋を持っていることが多いことを表わしている。

また、動詞性の強いサ変動詞の語幹に接続する語と

して「費」があるが、これは「造」に連接して１３回現われている。物を造るのには金が要ることを表わしていると言えそうである。この他「費」は、「送」、「務」、「料」、「備」、「築」などに連接することが多いがこのことは、係り語と接尾語の漢字の意味的なつながりを反映している。

そこで、我々の接尾語辞書のうち、漢字情報が利用できる派生語名詞処理用の接尾語辞書では、各接尾語に対して、それの係り語の語末の漢字表記と意味分類コードを与えて、きめの細かい接尾語処理ができるようにした。

結局我々の接尾語辞書は、係り語の品詞情報のみを用いるものと、更に係り受け情報を用いるものの二本立てである。

前者については、すでに人手によるチェックを済ませている。係り語の性質が文法情報のみで決まるものなどからなり、後者の接尾語辞書を補うものである。

後者については、例えば、「曜」と「日」の関係のように、接尾語辞書の見出し語「日」に対する係り語として「曜」を登録するより、むしろ「～曜日」という語は、せいぜい１１の限定された語しかないため、派生語そのものを見出し語として、すべて自立語辞書に登録した方がよい場合もある。このため、自動作成された接尾語辞書をそのまま使うことは問題があり、自立語辞書との兼合いを考えて、現在チェック作業を行なっている。

７．あとがき

自然言語の機械処理には辞書が重要な働きを示すことは当然であるが、中でも仮名漢字変換などカナ文字文の機械処理では、同形異義語と未登録語の処理が中心課題となり、その処理のためには予想以上に膨大な容量の辞書が必要となる。

ここで、例として、未登録語の中で大きな割合を占める漢語系の派生語名詞について考えてみよう。

これは、大雑把に言えば、三文字漢字語の処理をどうするかということになる。

二文字漢字語は、それが、同じ格調、同じレベルの文中では、そのまま他の語に置き換えることが困難なことにより、単語としては不動の地位を占めている。これはまた、市販の国語辞書を見ても、見出し語は、二文字漢字語については、日常使用される語は網羅されており、同じ規模同士の辞書であれば、見出し語はほぼ一致していることからも言うことができる。

一方三文字漢字語は、一般に二文字漢字語からできた派生語である。ということは、確定した二文字漢字に一文字漢字を添えればすぐにできあがる簡便性を持っている。このため、造語成分としての漢字の意味を知った人間なら簡単に新たな表現を持った語を作り出せるということになる。

このような訳で三文字漢字語は、その潜在的な数は二文字漢字語の比ではなく、それをカバーしようという機械辞書はまた、語構成の仕組を取り入れた柔軟性を持ったものが要求される。

我々は、基本的には、可能な限り派生語は自立語として取り込むべきだと考え、使用頻度の高い派生語は逐次辞書に登録している。

本稿で述べたことは、要するに網に漏れた派生語をいかに処理するかということであり、辞書に派生語の合成の仕方を記述しておくことにより、辞書に登録されていない語を分析して、そのうちもっともらしいものを選ぼうというのである。

本稿では取り上げなかったが、接頭語が付いた語も派生語であり、これの処理は避けて通れない問題である。文法的に処理できる一部の接頭語以外、本稿で述べたような簡単な係り受けで解決できるものは少なく、新たな観点からの研究が待たれる。

最後に、日頃ご指導ご鞭撻賜わる吉田将九大教授に深甚の謝意を捧げる。

参考文献

（１）稲永・小西：「カナ文字文のための機械辞書の構成について」電子通信学会技報ＡＬ７６－３９、１９７６

（２）松村明編：「日本文法大辞典」明治書院、１９７１

（３）稲永・吉田：「日本語処理のための機械辞書」情報処理２３－２、１９８２

（４）稲永・橋本・吉田：「係り受け辞書のカナ漢字変換への応用」情報処理学会第３０回全国大会１９８５

（５）稲永・橋本：「漢字の意味を利用した係り受け辞書によるカナ漢字変換システムについて」日本ユニバック技報、Ｎｏ．１０、１９８６