コーパスを用いた名詞と接辞の形態的分類**

丸元聡子† 乾 裕子† 荻野孝野‡
†計量計画研究所 ‡日本電子化辞書研究所

1. はじめに

接辞は、「それだけで単独に用いられることのない語構成要素」と定義され（大辞林より）、品詞分類の上では、問題視されることが少ない。しかし実際には接辞か否かの判別が難しく、辞書によって名詞・接辞の品詞記述に違いがある。「場（ジョウ）」のように名詞と同形である接辞がその例である。また、言語処理においては、接辞の基準が不明確なために名詞・接辞の両方に登録されている語があり、形態素解析時の曖昧性が増すという問題がある。

そこで、本研究では名詞を作る接辞もしくは名詞性接辞の可能性が高い語を含む実例文中での実態を調査し、名詞と接辞の判別基準を提案する。具体的には、ＥＤＲコーパスから用例を収集し、１）単独で格に立たない、２）必ず複合語として用いられる、という基準に基づき調査を行う。接頭語・接尾語の双方を対象とするが、本稿では名詞性の接尾語の分析結果を中心に報告する。

作業の流れは、図１の通りである。

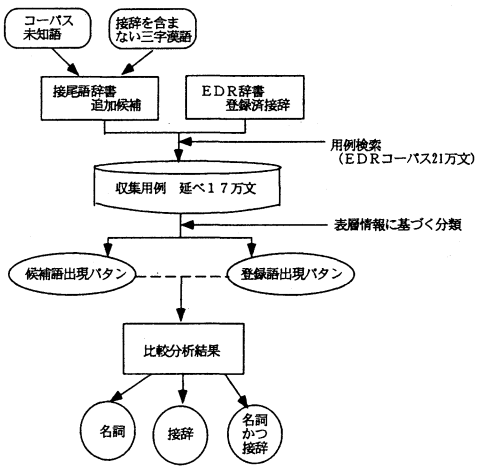

図１ 表層情報を基準にした語の分類作業の流れ

2. 背景

2.1 関連研究

どのような語を接辞と扱うかは、研究者によって判断が異なる。造語成分（特に一字漢語）を語基とするか接辞とするかの判断は難しく[10][12]、辞書によって、付与された品詞にかなり差異がある[3]。

多くの研究では、語基と結合して形式的な意味を添えるか語の品詞性（文法的性格）を決定し、単独では語を構成することはできない語（例：不-・無-・-的・-式）を接辞と扱っている[5][9][11]。本稿では、これらを狭義の接辞と呼ぶ。

しかし、造語成分・接辞と品詞分類されているものでも実際には自立した用法を持つことがある[3]。また、国研の調査報告[4]における「接辞の表」には、語構成要素としてのみ出現したものが挙げられているが、ここには他の研究で語基と扱われるものが多く含まれる。

また、[1]では、仮名漢字変換や未知語処理などの自然言語処理の観点から、分かち書きされる可能性のない複合語[*1] を派生語[*2] と扱い、この広義の派生語における語構成を自立語と接尾語の係り受け関係として辞書に記述することで、処理精度の向上を図ることが出来るとしている。

以上の通り、辞書の目的に応じて接辞の位置づけは異なることが分かる。

2.1 EDR辞書接辞登録状況

ＥＤＲ辞書に接辞として登録されている語は、実質的な意味を担っているものが多く、狭義の接辞に留まらない（例：兄[アニ]、狩り[カリ]、機[キ]、海[カイ]）。しかし、語構成要素にしかならないと考えられるものが名詞に登録されている場合もあり、登録基準が不明確である。

表１ EDR辞書での品詞付与基準

品詞コード	品詞	説明	例
JN1	普通名詞	格助詞「が、を」が後接するもの	太陽、山
JN7	形式名詞	必ず連体修飾句に修飾されて成り立つ名詞。	こと、もの
JB1	接尾語	他の単語の語末について、複合語を形づくる。	上、別
JUN	単位	計測する対象の、計測基準量を表す。数値がなくても（単独でも）、指示している実質概念がある点で、助数詞とは異なる。	メートル ＰＰＭ ダース 時
JN6	助数詞	数や順序を示すのに、数えられるものを特徴づける接尾語。	回、章

** 本研究は、ＥＤＲ日本語単語辞書の保守検討作業に基づく。

*1 いわゆる自立語（または自立語に準ずるもの）を組み合わせて出来た語。

*2 狭義の接辞と自立語からなる語。

ＥＤＲ電子化辞書仕様説明書[7]に示された品詞付与基準は表1の通りである。

しかし、実際には、下記のように登録されており（表2）、文中でどのように働くものを接辞として登録すべきかという指針が必要である。

表２ EDR辞書登録語の例

見出し	読み	登録品詞	複合語の例
後	ゴ	JN1	数日-後、夕食-後
室	シツ	JN1	特別-室、電話-相談-室
場	ジョウ	JN1	歌劇-場、野球-場
員	イン	JN1・JB1	組合-員、銀行-員
園	エン	JN1・JB1	野菜-園
系	ケイ	JN1・JB1	革新-系、理工-系
車	シャ	JN6	新型-車
者	モノ	JN1	不届き-者
者	モン	JN7	うっかり-者
者	シャ	JB1	協力-者、有力-者

2.3 問題点の整理

現在は、接辞の認定基準が曖昧であるため接辞の機能を持つ語が名詞に登録されている可能性がある。

狭義の接辞以外で語構成要素にしかならないものを名詞として登録してある場合でも複合語は生成できるが、接辞として登録することで、読みの異なる語、すなわち現れ方の異なる語を品詞の異なりとして整理でき、解析の曖昧性を減らせる可能性が高い。

例えば、造語力・新型車・防音壁といった複合語において、力[リョク]・車[シャ]・壁[ヘキ]が接尾語として登録されていれば、名詞である力[チカラ]・車[クルマ]・壁[カベ]と区別できる。現在、左記のうちＥＤＲ辞書でJB1登録されているのは力[リョク]だけである。

先行研究から、接辞は、少なくとも語構成要素としてのみ働くという条件を満たす必要がある。これにより、A）単独で格に立たない、B）必ず複合語として用いられる、ことを基準とする。この基準に基づき、登録済接辞・接辞の可能性のある語の実例文中での出現状況を調査し、接辞としての登録可否を検討する。

3. 接辞候補の選出

3.1 選出方法

下記、二種の方法で得た語群を検討対象として人手で接辞候補を選出する。候補とするべきかの判断に迷う場合は広く採り、調査対象に加える。

（1）接辞に分類されたコーパス未知語

ＥＤＲ辞書に登録されておらずＥＤＲコーパスで未知語扱いとなった語のうち、手作業で接辞に分類（タグ付与）されている語群

接頭語候補検討対象：　162語
接尾語候補検討対象：　645語

ここから、区切り誤り・品詞付与誤りや助数詞（登録済の語の異表記が多い）、また明らかに名詞と考えられる語を除き、接辞候補を選出する。

（2）接辞の辞書を使って分割できなかった三字漢語

コーパスから抽出した三文字漢字列のうち、ＥＤＲ辞書に登録済の接辞（接頭語・接尾語）辞書を用いて分割できなかった語群(6002語）。

ここから、1）接頭語候補、2）接尾語候補、3）新語（三字漢語）候補を選出する。

接頭辞候補の場合は(A)、接尾辞候補の場合は(B)を第１条件とし、かつa)～c)でない語を上記1)2)の候補とする。

(A) １文字-２文字に分割できる語　　(例：誤-動作)
(B) ２文字-１文字に分割できる語　　(例：地元-民)
a)一語登録してよいと考えられる語（例：火砕-流)
b) ２文字語が未知語であるために抽出された接辞（例：霊能-力)
c)複合語でない語　（例：気前良、一部略)

3.2 選出結果

接辞候補（用例調査の対象）として選出した語数は下記の通りである。（１）の語数のうち括弧内は、（２）の方法によって選出した語と重複しない語の語数である。また、例に挙げた語に付記した数字は、その接辞候補が出現した頻度である。

表３ 接辞候補選出結果

選出方法	候補種別	語数	例
（1）	接頭	87（79）語	新(1567)、最(280)、無(104)
	接尾	174（123）語	民(27)、心[シン](20)
（2）	接頭	18語	異(6)、過(2)、好(3)
	接尾	156語	率(283)、料(150)、量(235)
	(新語)	56 語(他要検討：37 語)	英数字、恒等式、懐風藻、暗順応、親水基

4. 用例調査

4.1 調査対象

ＥＤＲ辞書に既に登録されている接辞、また、「３．接辞候補の選出」によって選出した接辞候補を用例調査の対象とする。

なお、登録済接尾語に関しては名詞性接辞を検討対象とすることから助数詞（JUN・JN6）は除き、JB1を対象とする。また調査対象の語数を絞り込むため、下記の方法で語を抽出する。

・接辞の辞書で分割できなかった三字漢語に含まれる登録済接辞（結合した二字漢語が未知語であったため。

57語)
・ランダムに抽出した接尾語（10分の1標本。45語）
実際の調査語数は、表5の通りである。

4.2 調査方法

EDRコーパス全文（約21万文。タグ付き）から、4.1で挙げた調査対象の語が出現する用例を全て収集し[*3]、人手で分類する。調査項目を表4に示す。

分類の際には、実例から確実に抽出できる用法だけを採用する。実例に現れていなくても用法を想定できる場合があるが備考として記述するにとどめる。表層情報による分類を目的とするため内省による補完は一切行わない。

また、表記・読みの二点を組みとし、これらが共通でないものは別語として扱う。同形異語の情報を残すため、この別語が接尾語候補でない場合にも接辞候補と同様の調査を行う。

表4 調査項目

検討項目	調査項目	備考
格に立つか	単独でガ・ヲが後接	
派生するか	単独でナ・ノ・スルが後接	
複合語を作るか	二字以上の漢語または和語・外来語の語基と結合するか	前接・後接の両方を調査する。
その他の用法	1)略語 2)狭義の接辞との結合	1)例：大(=大学) 2)狭義の接辞と結合して一語を形成しているか

なお、格に立つ用例がある場合、補助情報として下記の区別をする。複合語としてだけ、格に立っている場合も記録に残す。

1：全く独立して、格に立つ。　例）日が暮れる
1'：連体修飾語を受けて格に立つ。　例）その日が来る
1''：連体修飾句を受けて格に立つ。　例）夢に見た日が来る
＊：複合語として格に立つ。　例）検査日が来る

4.3 調査結果

ここでは接尾語の調査結果について述べる。調査対象語数の合計（重複して選出されたものを除く）とEDRコーパスから収集できた用例数は表5の通り

表5 用例抽出結果

接尾語種別	調査対象語数合計	用例文数
登録済	102	74,469
候補語	279	103,769
合計	381	178,238

である。

これらの語について、表4の調査項目のうち、ガ格・ヲ格に立つか、語に後接して複合語を形成するかの調査項目で分類した結果は表6の通りである。ナ・スルによる派生や略語の情報は、今回の分類には用いていない。なお、接辞候補語ではないが候補語の異音語（別語扱い）として用法を調査した語の例を最右列に示す。

表6 用法分類結果

	コード	連接	説明	登録済 割合(%)	登録済 語例	候補語 割合(%)	候補語 語例	cf) 別語 語例
（独立した）名詞	10	ガ,ヲ,ノ	ガ格に立つが複合語の用例なし	2.0	格好	2.9	玉[タマ],箱[ハコ],柵	会社[カイシャ],数[カズ],音[オト],額[ヒタイ],暦[コヨミ],卵[タマゴ]
	12	ガ,ヲ		0	—	0.6	報	品[シナ]
	15	ガ		1.0	糞[クソ]	0.6	因	癖[クセ]
名詞	20	ガ,ヲ,ノ,前接	ガ格に立ち、複合語の用例がある	11.0	軍,語,式,(州),弁,用,(便[ビン]),(分[ブン])	34.2	案,液,駅,法,卵[ラン],票,社,(室),欲,(説),(術)	板[イタ],場[バ],壁[カベ],村[ムラ],雪[ユキ]
	22	ガ,ヲ,前接		1.0	感	2.3	宴,評,作,訳	
	25	ガ,ノ,前接		4.0	官,券,主[ヌシ],派	4.0	格,柄,具,瓶	札[フダ]
	27	ガ,前接		0	—	2.9	益,科	
名詞／名詞かつ接辞／接辞	30	ヲ,ノ,前接	複合語の例があるが、ガ格に立たない	3.0	系,糖	6.9	材,札[サツ],美,項,商	
	32	ヲ,前接		3.0	院,回,審	4.0	財,難,比,圧,域	市[イチ]
	35	ノ,前接		6.0	医,圏,性,部,別	4.6	倍,種[シュ],連,学	
	37	前接		44.0	化,画,頭[ガシラ],国[コク],然,人[ジン],的,力[リョク],帯[タイ],費,所[ショ],庁	26.3	界,会社[ガイシャ],剤,技[キ],車[シャ],卒,壁[ヘキ],場[ジョウ],料,博	
—	40	ヲ,ノ	ガ格に立たず複合語の用例なし	0	—	1.1	億	
	42	ヲ		2.0	軒[ノキ]	0.6	題	
	45	ノ		1.0	書院	0.6	あて	
	47	—		22.0	丘[キュウ],研,司,色[ショク]	8.6	癖[ヘキ],腫,暦[レキ]	

また、語例のうち括弧内に記したものは、格に立った例が、全て連体修飾されていた（単独では格に立っていなかった）ものである。

表層情報を整理した結果、登録済の接尾語は、格に立つ用例が少ない。例えば、コード37に分類された画・国[コク]は実質的な意味を担っているが接辞としての用法しか現れないことから、接尾語だけの登録でよいことが分かる。このコードの語が名詞としても登録されていた場合には、名詞を削除する検

[*3] コーパスでは複合語は分割されており、検索語とコーパスでの区切りが一致しない場合は検索されない可能性がある。

討対象としてよい。不要な重複登録を削除することで、解析の曖昧性を減らすことが出来る。なお、名詞（10 番台・20 番台）に分類された語群は、全て、名詞・接尾語の両方が登録されている。

また、候補語の用法を見ると、登録済の語に比べると格に立つ用例も多いが、四分の一以上はコード 37（接尾語）に分類されている。これらは、単独では格に立たず、また、必ず複合語として用いられることから、接尾語として登録してよいと考えられる。

上記の分類を品詞に対応させると、下記の通りになる。

10 番台・20 番台：名詞
30～35：名詞／名詞かつ接尾語／接尾語
　　のいずれか（各語の語義に依存）。
37：接尾語
40～45：名詞
47：（接尾語）

なお、40 番台の語は、用例が少ないために必要な接続情報が抽出できなかったものを含む可能性が高い。これらの語については、辞書情報を修正する際に、他のコーパスから用例を収集するなどの対処が必要である。

5. 品詞修正による効果の予測

コード 37 に分類された接尾語候補が、接尾語として登録されていないために、適切に解析できないと考えられる例がある。なお、以下の解析結果の例は JUMAN による。

本来ならば、接尾語として用いられているか否かで、下記、#1・#2 のように区別ができることが望ましい（者[シャ]接尾語登録あり）。

#1 他の大学へ流れる者が多い。
　→ 他の/大学/へ/流れる/者モノ/が/多い/。/
#2 若い医学者がいた。
　→ 若い/医学/者シャ/がいた/。/

しかし、車[シャ]は接尾語(JBI)としての登録がないため、下記のように解析される。

#3 小型車では世界一ですが、普通車市場では外車が強い。
　→ 小型/車クルマ/で/は/世界一/です/が/、/普通(副詞)/車クルマ/市場/で/は/外車/が/強い/。/

このような、読みが適切でない例は多い（都道府県-版[ハン]・防音-壁[カベ]など）。他に、区切り誤りの例も見られる。

#4 大学卒の初任給が4000バーツ。
　→ 大/学卒/の/初任給/が/4000/バーツ/。/
#5 短大卒女子も昨春と同じ。
　→ 短/大卒/女子/も/昨春/と/同じ/。/
#6 全体会議場に集まった人。
　→ 全体(副詞)/会/議場/に/集まった/人/。/

また、下記で未定義語となっている、剤[ザイ]はＥＤＲ辞書でも登録がないが、コード 37 であり接尾語として登録して良い。

#7 妊娠促進剤を服用していた。
　妊娠/促進/剤(未定義語)/を/服用/して/いた/。

接尾語登録されていないコード37の語の解析結果に不適切なものが多く見られることから、これらの語を登録していくことで、解析精度の向上を図ることが出来ると考える。

6. おわりに

以上、コーパスに現れた用例を表層情報で分類した結果に基づき、名詞と接尾語の境界に位置する語について接尾語としての登録可否を検討した結果を述べた。今後は、語義は同じでありながら品詞が重複して登録されていた語や登録が不足していた語について、追加・削除などの辞書の保守を行う。また、ここで収集した接続情報に基づき、個々の語の連接情報を検討する予定である。

謝辞：接辞候補選出の検討対象データを提供して下さった(株)日本電子化辞書研究所の小林正博さん、村上孝也さんに感謝致します。

参考文献

[1] 稲永紘之・新谷隆之(1986)「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」『自然言語処理』57-3
[2] 野村雅昭(1974)「三字漢語の構造」『電子計算機による国語研究VI』（国立国語研究所報告51）
[3] 山下喜代(1995)「形態素と造語成分」『日本語学』Vol.14
[4] 国立国語研究所報告42(1973)『電子計算機による新聞の語彙調査（Ⅲ）』秀英出版
[5] 野村雅昭(1987)「複合漢語の構造」『朝倉日本語新講座１文字表記と語構成』朝倉書店 pp130-144
[6] 石井正彦(1987)「複合名詞の構造と機能」『朝倉日本語新講座１文字表記と語構成』朝倉書店 pp145-157
[7] 株式会社日本電子化辞書研究所(1993)『ＥＤＲ電子化辞書仕様説明書』
[8] 小川泰嗣・望主雅子・別所礼子「複合語キーワードの自動抽出法(1993)『自然言語処理』vol.97(15) pp.103-110
[9] 野村雅昭(1977)「造語法」『岩波講座 日本語９語彙と意味』岩波書店,第7章
[10] 山本清隆(1995)「単純語・複合語・派生語」『日本語学』Vol.14
[11] 影山太郎(1993)『文法と語形成』ひつじ書房
[12] 石井正彦(1989)「語構成」『講座日本語と日本語教育第六巻』明治書院