固定的共起表現とその変化形

安武 満佐子　小山 泰男　吉村 賢治　首藤 公昭
福岡大学 工学部

1　はじめに

自然言語処理において文の基本単位をどう捉えるかが工学の立場から基本的に重要な問題である。筆者らは永くこの問題に関わっており、特に、構成的な（文解析・合成等の）ルール体系では扱い得ないと思われる表現の収集とその取り扱いに関する研究を行なってきた。[2][3]　表現の収集はルール体系を想定しつつ人手で行なってきており、今日、総合的なルール体系が確立している訳ではないため、収集規準に或る程度のゆれや恣意性が入る事はやむを得ない。しかし、できるだけ広範な収集を試みており、将来、想定されるルール体系と相補的に働かせる実験や試行錯誤を繰り返しながら、ルール体系と同時に全体システムを完成させて行くのが現実的であると考えている。近年、この種の表現を大量コーパスから統計量に基づく規準を用いて自動抽出する試みが種々行なわれているが、収集された表現の必要性、十分性、有用性等については多くの課題を残している。[4][5][6][7]　本稿では、筆者らが収集した、慣用表現を中心とする語の固定的共起表現の辞書の現状を報告する。また、これらの表現が慣用的意味を保存しつつどの様な変化形まで許されるかという変化形情報についても述べる。

2　固定的共起表現について

現在、固定的共起表現辞書には約３３，０００表現が収録されている。これらは実際の雑誌や市販の国語辞典中に現れる１０数万文を対象として、いわゆる慣用句、機能動詞結合などと呼ばれている現象を含め、語の特有な結合と思われる表現を網羅的に収集したもの、および、市販の慣用句辞典類から現在も使用頻度が高いと思われるものを抽出・総合したものである。辞書に収録されている表現は一語性表現と多語性表現の２つに大別することができる。以下ではその概要を述べる。

2.1　一語性表現

一語性表現とは語が隣接して強く結合しており、他語の挿入や語の交換が通常なされない表現である。これらは処理上一括して単語的に取り扱うことになろう。これらは大きく自立語的なものと付属語的なものに分かれる。

2.1.1　自立語的表現

自立語的な表現は下表のように、大きくは８種類に分類される。

	code	個数	例
名詞的表現	Mi	2,726	赤の他人、鶴の一声
サ変名詞的表現	Mis	356	貰い泣き、ラッパ飲み
動詞的表現	Yiv	4,694	かみ締める、煮詰める
形容詞的表現	Yia	812	怒りっぽい、注意深い
形容動詞的表現	Yik	1,073	一巻の終り、筋書通り
副詞的表現	Di	3,758	悪くすると、うっとりと
連体詞的表現	Ti	1,848	他愛の無い、断固たる
接続詞的表現	Ci	226	その結果、それはさておき

格言・諺やその他の決まり文句には自立語というより文、あるいは文に近い概念を与える表現が多い。これらは大きく次のように分類されている。

		code	個数	例
格言・諺	末尾が用言	P1,P21〜P27	1,164	急がば回れ、光陰矢の如し、時は金なり
	末尾が助詞	P31,P32	37	病は気から、鰯の頭も信心から
	末尾が名詞	Pm	740	鬼の霍乱、馬の耳に念仏
決まり文句		Pt	151	嵐の前の静けさ、春眠暁を覚えず

また、四文字熟語は現時点ではひとまとめに収録している。

	code	個数	例
四文字熟語	P4m	416	暗中模索、四面楚歌、無我夢中

2.1.2　付属語的表現

付属語的に扱い得る表現については、筆者らによって以前から収集・整理が行なわれており、拡張文節モデルの語彙レベルの資源としてまとめられている。[1]これらの表現は大きく助詞的なものと助動詞的なものに分けられ、それぞれ、関係表現、助述表現と呼ばれいる。

	code	個数	例
関係表現	R01～R98	1,055	について、によって、における、に基づいて、のように、でさえも、などといった、だけでなく
助述表現	A00～A90	1,988	なければならない、かもしれない、ている、たほうがよい、べきである、に違いない、おそれがある

2.2 多語性表現

一語性表現に比べると隣接結合の度合が弱く、表現中に他語の挿入も時として許されるものを多語性表現と呼ぶ。これらは概念語同志の共起によるものと付属語が大きく関与するものに分けて示す。(後者については十分な収集を終えていないため表現数を省く。)以下に示す表中のNは名詞、Pは用言を意味する。ただし、実際の辞書では、P(用言)は動詞(V)、形容詞(A)、形容動詞(K)のいずれかに細分類している。

2.2.1 概念語同志の共起

(1) N+助詞+P

	code	個数	例
N+を+P	Ia1	4,860	喧嘩を売る、骨を折る
N+が+P	Ia2	2,775	気が散る、荷が重い
N+に+P	Ia3	1,387	恩に着る、記憶にない
N+で+P	Ia4	62	顎で使う、肩で笑う
N+から+P	Ia5	27	口から出る、一からやり直す
N+まで+P	Ia6	5	骨までしゃぶる、心まで腐る
N+と+P	Ia7	27	面と向かう、藻屑と消える
N+へ+P	Ia8	16	横道へ逸れる、後へ引けない
N+には+P	Ia9	12	その手には乗らぬ
N+では+P	Iaa	19	人事では済まない
N+でも+P	Iab	4	梃でも動かない、何でもこい
N+は+P	Iac	24	その手は食わない
N+も+P	Iad	81	及びもつかず、あてもない
N+にも+P	Iae	31	気にも止めない
N+へも+P	Iaf	3	下へも置かない

(2) N+の+N+助詞+P

	code	個数	例
NのN+を+P	Ib1	121	肩の荷を下ろす
NのN+が+P	Ib2	57	目の色が変わる
NのN+に+P	Ib3	48	悪の道に走る
NのN+で+P	Ib4	6	疑いの目で見る
NのN+と+P	Ib7	3	不帰の客となる
NのN+へ+P	Ib8	2	元の鞘へ収まる
NのN+も+P	Ibd	11	猫の手も借りたい

(3) P(連体形)+N+助詞+P
(PN: Pの連体形+N)

	code	個数	例
PN+を+P	Ic1	53	危ない橋を渡る
PN+が+P	Ic2	29	生きた心地がしない
PN+に+P	Ic3	18	同じ土俵に上がる
PN+で+P	Ic4	2	長い目で見る
PN+から+P	Ic5	1	教わる側から忘れる
PN+へ+P	Ic8	1	出る所へ出る
PN+は+P	Icc	5	逃した魚は大きい
PN+も+P	Icd	4	為す術もない

(4) N+助詞1+N+助詞2+P

	code	個数	例
N+も+N+も+P	Id1	63	影も形もない
N+に+N+を+P	Id2	77	胸に手を当てる
N+を+N+に+P	Id3	44	心を鬼にする
N+に+N+が+P	Id4	32	身に覚えがない
N+が+N+に+P	Id5	14	手が後ろに回る
N+で+N+を+P	Id6	22	目で物を言う
N+を+N+で+P	Idf	2	恩を仇で返す
N+が+N+を+P	Id7	10	金が子を産む
N+から+N+が+P	Id8	8	目から鱗が落ちる
N+から+N+を+P	Id9	2	横から口を挟む
N+から+N+に+P	Ida	3	口から先に生まれる
N+から+N+へ+P	Idb	3	闇から闇へ葬る
N+に+N+は+P	Idc	5	金に糸目は付けない
N+とも+N+とも+P	Ide	13	夢ともうつつともつかない

(5) その他　<いろいろな形式>+P
(Pa: Pの連用形, Ps: Pの仮定形)

	code	個数	例
Pa+て+P	Ie1	38	割って入る
Ps+ば+P	Ie2	9	貧すれば鈍する
Pa+P	Ie3	39	都合良くいく
Pa+ても+Pa+ても+P	Ie4	3	居ても立っても居られない
N+を+Pa+て+P	Ie5	19	腹を抱えて笑う
N+が+Pa+て+P	Ie6	3	幕が切って落される
N+に+Pa+て+P	Ie7	2	尻に付いていく
N+N+を+P	Ie8	6	満面朱を注ぐ
N+N+が+P	Ie9	1	攻守所が換わる
N+N+に+P	Iea	3	自他共に許す
副詞+P	Ieb	1,328	あかあかと照らす、ぐるぐるまく

2.2.2 付属語が直接関与する共起

	例
付属語が直接関与する共起	決して～ない、もし～ば、～しか～ない、いったい～か、とかく～がちだ、既に～ている、どんなに～ても～ない

3 変化形について

多語性表現における語の共起の固定化の度合は千差万別であり、どこまで変化形が許されるかは、それ

それの表現によって異なる。例えば「足が棒になる」は「棒になった足」（疲れた足の意）と言い換えることが出来るが、「足が出る」（赤字が出るの意）という表現は「出る足」と言い換えても「出る赤字」の意には解釈しにくい。従って、各表現に対して慣用的な意味を変えずにどこまでの変化形が許されるかをあらかじめ調査・整理しておく必要がある。以下に変化の基本的な形式について要点を整理しておく。

A. 付加 基本形に何らかの語が付加されるか否か、付加される場合はどの様な語が許されるかが明らかにされなければならない。

A-1. 修飾要素の付加

A-1-1. 連体修飾語句 名詞を修飾する語句については、必須的に要求される場合、禁止される場合、その他の場合がある。また修飾句の形態に制約のある場合とない場合が考えられる。さらに、修飾語句に対する意味上の制約も規定しておく必要がある。

例：○ [必須]　＜感情＞＋顔をする
　　×　業を煮やす
　　○　＜人＞の＋手に余る
（＜X＞は概念Xの語を表す。）

A-1-2. 連用修飾語句 述語を修飾する語句についても、必須的に要求される場合、禁止される場合、その他の場合が考えられる。また、動詞が本来取らなかった格が新たに取られる場合や、その逆の場合もある。これらについても通常の動詞に対する格要素の規定と同様のことを行なっておく必要がある。

例：○ [必須]　＜人＞に＋バトンを渡す
　　○　＜人＞と＋手を切る

A-2. 助動詞等の付加 述語性の慣用表現の場合、特定の助動詞等が必須的に要求される場合や禁止される場合がある。

例：○ [必須]　手が付けられない
　　×　頭が切れている

A-3. 副助詞等の付加 副助詞等の付加も場合によって許されたり許されなかったりする。

例：○　気が気ではない
　　×　喉からは手が出る

B. 助詞の削除 含まれる助詞を取りさっても慣用表現として働くものがある。助詞の削除が可能であるか否かを明らかにする必要がある。また、削除した結果の表現と複合動詞、複合形容詞等との区分も、少なくとも取扱い上明確にしておくことが必要である。

例：○　気味が悪い

C. 助詞の置換 A-3と類似の現象として格助詞を副助詞で置き換え得るかどうかも表現ごとに異なる。

例：○　諦めを付ける⇔諦めが付く

D. 語順の変更 格要素等の順序を入れ換えてもよい場合と悪い場合がある。

例：○　横から口を出す⇒口を横から出す
　　×　足が棒になる⇒棒に足がなる

E. 受身化・使役化 受身化や使役化が許される場合と許されない場合がある。

例：○　金が物を言う⇒金に物を言わせる
　　×　へそが茶をわかす⇒へそに茶をわかさせる

F. 倒置による体言化 含まれる格要素としての名詞を末尾に回して連体被修飾語にかえる事が許される場合と許されない場合がある。

例：○　足が棒になる⇒棒になった足
　　×　足が出る⇒出る足
（「赤字が出る」の意）

4　調査結果と辞書

現在、3で示した変化形のうち連体修飾語の付加、連用修飾語の付加、助詞の削除、助詞の置換、倒置による体言化について調査が進んでいる。調査にあたっては各表現の基本形を変化させた例文を作りながら適否の判定を行なった。調査結果は、ある変化形項目に対して変化が可能な場合や不可能な場合等のケースに分け、コード化して辞書に収録した。図1の［2］が変化形のコード部である。調査結果から各表現には用法上固有の制約があることがわかる。また、変化形調査の一環として、「新聞でたたかれる」、「足を洗う」の

$nI0004830$: ___00___ :	あしを.なげだす	足を - 投 (げ) - 出す	$Ia1 - V$	***
$nI0004850$: 0__02_3_:	あしを.のばす	足を - (伸/延) ばす	$Ia1 - V$	***
$nI0004870$: 0__00_1_:	あしを.はこぶ	足を - 運ぶ	$Ia1 - V$	***
$nI0004890$: 0__00_3_:	あしを.ふみ - いれる	足を - 踏 (み) - 入れる	$Ia1 - V$	***
$nI0004910$: ___01___ :	あしを.みだす	足を - 乱す	$Ia1 - V$	***
[1]	[2] [3]	[4]	[5]	[6]

[1] 見出し語の固有番号。

[2] 変化形コード。3で示したように変化形の調査項目がAからFまで9項目あり、1桁が1項目に対応する。(予備の項目を加え、10桁で表す。) 調査結果は項目毎に"0"から"3"までの数字で表現されており、"_"はその見出し語について未調査であることを示している。

[3] かな書き見出し語。"."は多語性の単語のみにあり、この位置に他の単語 (名詞、副詞等) が挿入可能であることを表す。"."または" - "で区切られた文字列は、かなで表記する場合と漢字で表記する場合があることを示す。

[4] 仮名漢字まじり見出し語。漢字の後ろにある () 内のひらがなは送り仮名のゆれを表す。また漢字を囲った (/) は漢字表記のゆれを表す。" - "で区切られた文字列は、かなで表記する場合と漢字で表記する場合があることを示す。

[5] 種別コード。一語性表現は8種、多語性表現は151種に分類している。また、格言・諺は一語性の自立語表現であるが、別にまとめて14種に分類している。2で示した表中の code を参照。

[6] 辞書の整理上付けた編集コード。

図 1: 固定的共起表現辞書の形式

ように語句本来の意味として使われ得る表現と、「事を構える」、「話の腰を折る」などのように慣用的意味としてのみ使われる表現に分類する作業も行なっている。機械処理を行なう上で、変化形データをいかに有効に利用するかは今後の重要な課題である。

5 おわりに

本稿では変化形の情報をのせた固定的共起表現辞書の現状を報告した。固定的共起データは形態素、構文レベルの処理でも、例えば、かな漢字変換の精度の向上 [8] や構文解析結果の曖昧さを減らすため等に有効であると考えられる。近年盛んに行なわれている大規模コーパスからの慣用的な表現の自動抽出の研究ではその抽出結果に対する評価が難しいが、本稿で報告した固定的共起表現辞書がその評価に一つの目安を与えると思われる。また、変化形情報は入力文字列が慣用的表現として用いられているのか、標準的用法の表現として用いられているのかを判別することにもある程度利用できるであろう。筆者らは言語処理における非標準的な用法からのアプローチとして今後も変化形情報を含めたコロケーションの調査を続け、ルールの確立とあわせて辞書の充実を図っていきたいと考えている。

変化形情報の整理を開始するにあたっては、元電気通信大学教授岡本哲也先生の御指摘に触発された所が大きい。ここに記して謝意を表したい。

参考文献

[1] 首藤, 楢原, 吉田:日本語の機械処理のための文節構造モデル, 電子通信学会論文集,vol.62-D,No.12,1979.

[2] 首藤, 吉村, 武内, 津田:日本語の慣用的表現について, 情報処理学会研究報告,88-NL-66,1988.

[3] 首藤, 吉村:日本語における語の固定的共起, 電子情報通信学会　文法的知識と意味的知識の蓄積、管理シンポジウム論文集,1989-1.

[4] 内野, 池原, 白井:弱抑制による連鎖共起表現の抽出とそれに基づく離散共起表現の抽出, 言語処理学会第2回年次大会発表論文集,1993-3.

[5] 延澤, 堤, 孫, 佐野, 佐藤, 大森, 中西:自然言語における有繋文字列の抽出, 言語処理学会第2回年次大会発表論文集,1993-3.

[6] 尾本, 北:距離反比例型スコアを導入したコロケーションの自動抽出, 情報処理学会研究報告, 96-NL-112,1996.

[7] 新納, 井佐原:片方向の共起性による述語型定型表現の自動抽出, 自然言語処理,vol.2,No.3,1995-7.

[8] 小山, 安武, 吉村, 首藤:かな漢字変換における固定的共起表現, 言語処理学会第3回年次大会発表論文集,1997-3.

[9] 宮地裕編:慣用句の意味と用法, 明治書院,1982

[10] 村木新次郎:慣用句・機能動詞結合・自由な語結合, 日本語学,vol.4,No.1,1985