語義の特異性を利用した慣用表現の自動抽出

新納浩幸[†] 井佐原均[††,☆]

本論文では「水をあける」「目を盗む」といった述語型慣用表現をコーパスから自動抽出する手法を提案する．慣用表現を自動抽出する従来手法の多くは，本質的に，その表現の頻度や名詞動詞間の共起性の強さに基づいて慣用表現かどうかの判定を行っている．しかし，慣用表現はコーパス中での頻度が低く，更に強い共起性は慣用表現の１つの特徴でしかない．このため頻度や共起性の観点だけからでは抽出が困難である．本手法は慣用表現中の語義の特異性に注目する．概略，述語型慣用表現中の名詞はその語義の特異性のために類義語と置換されることがないという性質を利用する．例えば「水をあける」「目を盗む」は「真水をあける」「まぶたを盗む」とは言えないが，一般の表現「穴をあける」「現金を盗む」では「穴」や「現金」の類義語と置き換え可能である．本手法では共起性が弱い慣用表現や頻度の低い慣用表現を取り出せる．また反例を探すという戦略になっていることからコーパスの質や量の問題を避けることができる．最後に新聞記事１か月分のコーパス（約９Mbyte）と分類語彙表を用いて実験を行った．この詳細についても述べる．

Automatic Acquisition of Idioms on Lexical Peculiarity

HIROYUKI SHINNOU [†] and HITOSHI ISAHARA [††,☆]

Idiomatic expressions are useful in many natural language systems and in linguistics, but collecting of them is costly work. There have been some attempts to extract idioms from a corpus automatically. However, for predicative idioms, their methods don't succeed fully, because they use frequency or collocation essentially. In general many idioms have too small frequency in a corpus. Some idioms don't have strong collocation. This paper presents a technique for extracting predicative idioms from a corpus. We use lexical peculiarity for the noun in an idiom. The noun in a predicative idiom have particular lexical. So, the noun in the idiom can not be exchange by a similar noun. By confirming existence of a phrase exchanged by a similar noun, we can judge whether a phrase is idiomatic or general. This technique can extract predicative idioms which have small frequency in corpus or don't have strong collocation. The technique has an advantage to be apply for a small corpus. We report the experiment performed on the newspaper articles of one month.

1. はじめに

本論文では，日本語コーパスから述語型慣用表現を自動抽出する手法を提案する．述語型慣用表現とは，例えば「水をあける」「目を盗む」のように，

＜名詞＞＜格助詞＞＜動詞＞

の品詞列となっている慣用表現をここでは指す．

† 茨城大学工学部システム工学科
Department of Systems Engineering, Faculty of Engineering, Ibaraki University
†† 電子技術総合研究所知能情報部自然言語研究室
Natural Language Section, Machine Understanding Division, Electrotechnical Laboratory
☆ 現在，郵政省通信総合研究所関西先端研究センター知的機能研究室
Presently with Intelligent Processing Section, Kansai Advanced Research Center, Communications Research Laboratory, Ministry of Posts and Telecommunications

慣用表現は一般に個々の構成語の意味からその表現全体の意味を作り出すことが困難になっている[1]．このため機械翻訳などの自然言語処理システムでは，慣用表現に対して例外的な処理を施す必要があり，それら表現をあらかじめ収集，整理しておくことが重要になる．

慣用表現の中でも，特に述語型慣用表現を収集することは重要である．なぜなら述語型慣用表現は慣用表現の多くの部分を占め[2]，しかもこれら表現は一般の表現との境界が特に曖昧になっているからである．このため述語型慣用表現をある基準をもって収集することは言語学的にも興味深い．また解析処理の面からも，述語型慣用表現は通常の名詞動詞間の共起による解析との整合性が必要になるため[3),4]，述語型慣用表現を収集し，語の持つ非標準的な語義の扱いを明確にしておく必要がある．近年研究の盛んな用例による翻訳でも

慣用表現のような特異な表現への対策が必要である[5]．また慣用表現は個々の語間に強い共起関係を持つという特徴もある．共起性の強い表現を記憶しておくことは外国語習得の面でも効果的であるし，音声認識，OCRにも，共起性の強い表現を記憶しておくことが，そこでの曖昧性の解消に役立つことが知られている[6),7]．さらに近年コーパスから名詞のクラスタリングを行う研究も盛んであるが[8]，クラスタリングを行う際には，慣用表現中の特殊な語義を持つ語をあらかじめ別枠で考慮しておくことが有効である．

慣用表現を収集，整理することは言語学的にも，機械処理の面からも有益であるが，その収集は困難である．なぜならそれら表現の定義は曖昧なため個々の表現に対して人間の判断が必要となり，その収集には膨大な時間と手間がかかるからである[9]．また人手による収集では，その網羅性，一貫性などの問題点もある．

これらの点から慣用表現（あるいは定型表現）の自動抽出の試みがなされている．新納の研究[10),11]では抽出対象を付属語的表現に限定し，その文法上の性質や字面上の特徴を用いているために，そこでの手法を述語型の表現に単純に応用することはできない．また北の研究[12]では本質的に表現を構成する単語の数とその表現の頻度を抽出の基準としているが，述語型慣用表現は一般に頻度が小さいために，頻度に重点をおく手法では述語型慣用表現の抽出はできない．SmadjaやChurchの研究[6),13]は相互情報量などを用いて単語間の共起の強さを測り，慣用表現を抽出する試みである．これはある種の慣用表現では有効であるが，強い共起性は慣用表現の1つの特徴でしかなく，その特徴を外れた表現も多く存在する．また相互情報量でも出現頻度が小さいとその値の信頼性が低くなり，現実的なコーパスの大きさでは抽出が困難である．

本論文では述語型慣用表現を対象にして，語義の特異性に注目した抽出法を提案する．概略，述語型慣用表現中の名詞はその名詞の類義語と置換されることがないという性質を利用する．例えば，「水をあける」という慣用表現の「水」という語は，通常の『酸素と水素の化合物である物理的な水』という語義ではなく，この表現に特有の語義『競っている両者の間の差』という語義として使われている．前者の語義を持つ類義語として「真水」という語があるが，この語を用いた「真水をあける」という表現は意味をなさない．更に後者の語義を持つ類義語として「差」という語があるが，この語を用いた「差をあける」という表現では，その意味は通じるが一般に用いられず，通常は「差を広げる」という表現が使われる．この性質に注目すると，「をあける」と共起する名詞をコーパスから取り出し，類義語になっているものどうしを省いていくことによって，慣用表現を形成する名詞「水」が残されることが予想される．

ここでの当面の目的は従来まで人手で行われていた慣用表現の収集作業を軽減し，より効率的に慣用表現を収集することである．実験として，「を」格だけを対象に，朝日新聞1か月分のコーパス（テキスト部分約9 Mbyte）から「AをBする」の形の共起データを抽出し，そこから本手法を用いた慣用表現の抽出実験を行う．また従来手法である相互情報量を用いた抽出と比較し，抽出できる範囲が広がることを示す．

2. 述語型慣用表現の自動抽出

2.1 語義の特異性と置き換え不可能性

述語型慣用表現と一般の表現とを区別する1つの基準は，そこで利用されている名詞の語義と動詞の語義をある概念として一般化できるかどうかである．これは突き詰めれば，その名詞や動詞が類義語に置き換え可能かどうかに帰着する．

例えば，「水を飲む」の名詞「水」の語義はwaterという概念に置換でき，動詞「飲む」の語義はdrinkという概念に置換できるために一般の表現と判断できる．当然，ここで概念に一般化できるかどうかの基準が必要であるが，それが類義語への置き換え可能性である．「水を飲む」と言った場合の名詞「水」の語義を持つ類義語として「真水」「お湯」「蒸留水」などがあるが，それらは「水を飲む」の「水」と置換できるし，「水を飲む」の動詞「飲む」は類義語「飲み込む」「飲み干す」などに置換できる．一方，先の「水をあける」の例では，それが不可能であるために，結果として慣用表現と判断できる．

この類義語への置き換え不可能性は慣用表現の定義によく用いられる構成不可能性「その表現を構成する個々の語の意味からは，全体の意味が構成できない」[1),2)]とほぼ同等のことを述べている．今，「AをBする」という表現を考えてみる．この表現が慣用表現であるとすれば，構成不可能性から以下のことがいえる．

> 『名詞Aのある語義と動詞Bのある語義が「を」格によって共起し，結合されることでは，「AをBする」の全体の意味が構成されない．』

これには様々な原因が考えられるが，その1つの原因として，「AをBする」を慣用表現たらしめている名詞Aの語義あるいは動詞Bの語義あるいはその両方が存在しない，あるいはそれらのいずれかが特異であることが挙げられる．存在しない場合，全体の意味を

構成できないのは当然であるが，特異である場合も構成できない．なぜなら，仮に名詞Aの語義が特異であるにもかかわらず，全体の意味が構成できるとしたら，その語義と動詞Bの語義との共起関係が存在することになる．共起関係の要素どうしは概念なので，名詞Aの特異な語義はある概念に対応することになる．しかし特異な語義が概念に対応できれば，それはもはや特異ではなく矛盾である．結局，ここでは概念に対応できない語義を特異な語義と呼んでいるわけである．そして「特異な語義」の定義を「類義語への置き換えが不可能であること」とすれば，構成不可能性とほぼ同等な性質が類義語への置き換え不可能性となる．語義が特異であることを類義語への置き換えが不可能であることと考えるのは極めて自然であり，慣用表現の1つの性質として類義語への置き換え不可能性を挙げてもよいと考える．また名詞Aや動詞Bの語義が存在しないような場合（通常はこのように考えるのが自然である）は，類義語自体が存在しないために，当然，置き換え不可能である．

具体的な例として「目を盗む」を考える．この表現は市販の慣用表現辞典[14]にも記載されており，主観的に慣用表現と認められたものである．またこの表現の意味『気づかれないようにこっそり行動する』は明らかに「目」や「盗む」の持つ意味からは構成できず，構成不可能性からも慣用表現だと判定できる．「目を盗む」の慣用的意味は構成不可能だが，これはその慣用的意味を構成するための「目」の語義がないか，あるいはこの表現に限定された特異な「目」の語義が利用されているからだと考えられる．どちらも場合でも「を盗む」と共起する類義語は存在しないと予想でき，置き換え不可能性からも慣用表現だと判定されるであろう．

整理すると，構成不可能性の1つの側面が類義語への置き換え不可能性であり，この性質は慣用表現の一側面を担っている．本論文では類義語への置き換え不可能性つまり語義の特異性に注目して慣用表現かどうかを判定する．

最後に注記として，構成不可能性も慣用表現の客観的な定義にはなり得ないことを記しておく．通常，市販の慣用表現辞典に記載されている表現が人手により主観的に集められた慣用表現と考えられるが，それら表現のすべてが構成不可能性を満たしているかどうかは明らかではない．これは語義をどのくらい細かく捉えているかに依存している．

2.2 自動抽出処理

全体の抽出処理手順は以下のようになっている．

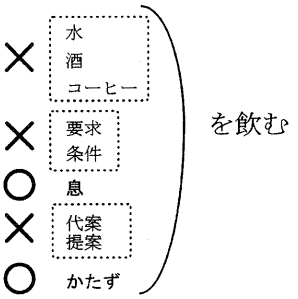


図1 類義語の削除
Fig. 1 Elimination of similar words.

step 1 コーパスから共起データを収集する．
step 2 共起データから格助詞と動詞を固定した場合の名詞の集合を作成する．
step 3 step 2で作成した各名詞の集合から，類義語の関係にあるものを取り除く（**図1**参照）．
step 4 残った名詞と先に固定した動詞と格助詞を組み合わせて慣用表現とする．

step 1～step 4の中で実際に問題となるstep 1とstep 3を以下に説明する．

2.2.1 共起データの収集（step 1）

コーパス中で名詞Aが格助詞Bを介して動詞Cと共起した場合に，[A，B，C]の3組を取り出す．この3組を共起データと呼ぶ．例えば「雨が降っている」からは共起データとして[雨，が，降る]が取り出せる．

コーパスから共起データを収集することは一般に困難である．これは解析の曖昧性の問題（省略も含む）があるからである．このため手作業により収集することや[15]，曖昧性のない共起データだけを収集することが行われる[16]．

本論文でも曖昧性のないデータだけを対象にする．基本的に名詞，格助詞，動詞が以下のように連続して現れた場合のみを対象とする．

名詞(A)＋格助詞(B)＋動詞(C)

このデータからは[A，B，C]を取り出す．ただし以下の点に注意する．

（1） 副詞の挿入

格助詞と動詞の間に副詞が入った場合，副詞を無視して[A，B，C]を取り出す．

（2） 代名詞

名詞の部分が代名詞になっているものは共起データを作成しない．

（3） 複合名詞

名詞の部分が複合名詞になっている場合は複合名詞のままで共起データを作成し，「AのB」になっているものは「B」の形で共起データを作成する．

（4） 連用句の挿入

連用句が挿入されている以下の形の場合，

名詞(A)＋格助詞(B)＋名詞(C)＋
格助詞(D)＋動詞(E)＋句読点

曖昧性なく[A, B, E]，[C, D, E]を認識できるが，ここでは[C, D, E]のみを取り出す．これは実際には[A, B, E]，[C, D, E]が組み合わさって意味をなすような5項関係のものも多く存在するからである．例えば「損を覚悟で売る」，「彼女をキャリア・ウーマンと呼ぶ」，「株式市場を研究テーマに選ぶ」などから[損，を，売る]，[彼女，を，呼ぶ]，[株式市場，を，選ぶ]を取り出すのは妥当ではない．

（5） 「を」格に対する使役の助動詞

通常，動詞に助動詞が付随した場合には，助動詞を取り除いた形で共起データを作成する．しかし格助詞が「を」であり，しかも使役の助動詞が使用されている場合には，その助動詞は取り除かない．これは「を」格の場合，使役の助動詞を取り除くと意味をなさないものが生じるからである．例えば「波長を合わせる」の場合，助動詞を外して「波長を合う」とは言えないので，[波長，を，合う]ではなく，[波長，を，合わせる]を取り出す．

（6） 数量詞移動の現象

数量詞移動の現象に対しては，数量詞も別個に取り出す．数量詞移動とは，

「3匹の子豚が住んでいた」

という表現が

「子豚3匹が住んでいた」
「子豚が3匹住んでいた」

という表現にそれぞれ互いに置き換えることができるという言語現象である[17]．いずれの形態で現れても，置き換え可能であることが解析システムで判断できた場合，[子豚，が，住む]，[N匹，が，住む]を取り出す．本論文では上記の場合，「子豚」と「N匹」に類似関係が認められ，「子豚が住む」を慣用表現として取り出さない効果が現れることを期待している．

2.2.2 類義語の削除（step 3）

助詞Bと動詞Cを固定して，step 1によって収集した共起データから，[＊，B，C]の形のデータを取り出し，＊にあたる名詞の集合を作成する（step 2）．次にこの名詞の集合から，類義語の関係になっているものを取り除く．

ここで類義語の判定方法が問題になる．本論文では基本的に分類語彙表[18]を用いる．分類語彙表は単語を6レベルの木構造体系にまとめており，その上位4レベルまでが一致した場合にそれらの名詞を類義語だと判定する．

ただし複合名詞など分類語彙表に記載されていない名詞も存在する．述語型慣用表現の名詞の多くは一般の名詞であり，そのような名詞が分類語彙表に記載されていないとは考え難い．このため分類語彙表に記載されていない名詞をもつ共起データは述語型慣用表現の候補にはしない．分類語彙表に記載されていない名詞に関しては，その字面情報により述語型慣用表現の候補との類似性を調べるだけに用いる．つまり候補を棄却するためだけに用いる．字面情報による類似性の判定については，ある単語Aの文字列αが単語Bの文字列βの部分文字列になっている場合に，単語Aは単語Bと類似性があるとした．例えば「協力」は「経済協力」の部分文字列になっているので，「経済協力」は「協力」と類似性があり，その結果類義語だと判断する．これは反射律になっていないが，実際に長い文字列の方が候補になることは稀であるので実害はない．また，これは後方一致ではなく包含関係を条件としている．後方一致の場合は，下位概念をチェックすることに相当し，より適切にも思えるが，ここでは類似性を認められれば良いので包含関係の方が妥当である．

頻度の問題を注記しておく．類義語の関係を調べる際，調べる対象の名詞の集合が小さいと，信頼性のある結果が得られない．このためある程度の要素数を持つ名詞の集合だけを対象にせざるをえない．本論文の実験ではこの要素数を10としている．またある名詞がその集合内に類義語がなく取り除かれなかったとしても，その名詞の頻度が2以下のものは抽出しない．つまり収集した共起データのうち頻度が2以下のものは述語型慣用表現の候補にしない．これはノイズの排除である．一般に完全な解析は不可能なので，共起データにノイズが入ることは避けられない．しかも本手法の場合，ノイズによるデータは特異になる可能性が高く問題が大きい．ここでは頻度2以下のものはノイズの可能性が高いと考え，候補としない戦略をとる．

2.3 多義語の問題

多義語の問題について述べる．通常，動詞，名詞の両方ともに多義語になっていることが考えられる．ここでは，基本的に上記の手法だけで多義語の問題に対処している．

例えば，step 2によって作成される名詞の集合内のある名詞Aの語義としてn_1，n_2，n_3，n_4があり，対応する動詞Bの語義としてv_1，v_2，v_3がある場合を考えてみる．もし可能な共起の組合せが，常に1通り（例えば，n_1とv_1）であれば基本的に問題はない．この場合，名詞の方の語義の中で1つでも類義語を見つけら

れれば，慣用表現でないと判断すれば良い．仮に名詞Aと動詞Bによる表現が一般表現であれば，n_1の語義を持つ名詞Dが先の集合に含まれることは十分考えられるために，この候補を取り除けるし，逆に名詞Aと動詞Bによる表現が慣用表現であれば，n_1の語義を持つ名詞Dが先の集合に含まれることはない．n_1の語義を持つ名詞Dが先の集合に含まれないのは，「水をあける」と「差をあける」の例で述べたように，名詞Aを名詞Dと置換することが通常行われないからである．

しかし残念ながら，共起関係が意味をなしえる組合せは1通りとは限らず，逆に慣用表現では複数の組合せが生じる場合が多い．例えば「油を売る」「うでをあげる」などは慣用表現としての『さぼる』や『技量を高める』という意味の他に，『(物理的な) 油を販売する』や『腕を上方に移動させる』という意味も持つ．

本手法ではここが限界であり，この問題には対処できない．たとえ人間であっても，共起関係に裏の意味がある（慣用表現になっている）のかどうかを共起データだけで判断することはできない．ただし結果的に対処できる可能性もある．これはコーパスの分野性である．例えば，「油を売る」という表現の場合，[*，を，売る]の*にあたる名詞の集合から「油」と類似の言葉が現れなければ結果的に正しく抽出できる．コーパスの分野を限定した場合，名詞の集合はより限定されたものになるため，「油」という単語自身が特異になっていくと考える．

実際の処理としては，名詞に複数の語義がある場合（語義の数kとする），それぞれの語義に対しての類義語を探し，類義語を持たない語義がkの過半数以上であれば，慣用表現とみなすことにした．

3. 抽出実験

3.1 実　　験

本手法の有効性を確認するために，朝日新聞の記事1か月分(テキスト約9 Mbyte)をコーパスとして述語型慣用表現の抽出実験を行った．ただし本実験では「を」格だけを対象にした．これは共起データを収集する際の処理と実験の評価が膨大な作業になることを避けたためである．また比較実験として，作成した共起データだけから名詞，動詞間の相互情報量を用いて慣用表現の抽出も行った．

まずコーパスから格助詞が「を」である共起データを収集した．その結果45,070組，31,899種類を取り出した．次にその共起データから頻度が10以上の動詞を取り出した．この結果，728種類の動詞を選択した．各々の動詞に対して，その動詞と共起している名詞の集合を作成した．このときその集合の要素数は10以上になっている．次にその名詞の集合中で頻度が3以上の名詞を選び，その名詞と対応する動詞および格助詞「を」から構成される表現を述語型慣用表現の候補とした．候補となった表現を合計すると2,030種類であった．それら候補の名詞に対してだけ，その名詞の類義語が対象となる名詞の集合内に存在しているかどうかを調べ，存在していない場合に述語型慣用表現として抽出した．結果として349種類の表現を抽出した．抽

表1 抽出結果
Table 1 Results of extraction.

〈正　解〉		〈誤　り〉	
つじつまを合わせる	ふたをあける	たばこを吸う	つながりを持つ
影を落とす	音頭をとる	みそを作る	わいろを受け取る
顔をつぶす	顔を合わせる	イメージを与える	スタートを切る
顔を立てる	顔色をうかがう	テープを聞く	デザインを考える
気を配る	気勢をあげる	トンネルを掘る	バランスをとる
犠牲を払う	脚光を浴びる	ボールを投げる	握手を交わす
胸を張る	群を抜く	圧力を加える	安全を維持する
血を引く	肩を並べる	安全を守る	威嚇を伴う
口をそろえる	口火を切る	威信をかける	意見を交わす
姿を消す	思いをはせる	意見を交換する	意見を聴く
耳を傾ける	実を結ぶ	意識を変える	意欲をみせる
手をつなぐ	手を引く	違いを見せる	運航を開始する
手を携える	手を合わせる	運転を中止する	運命をたどる
手を伸ばす	手を貸す	影を投げかける	影響を心配する
首をかしげる	心を打つ	映画を撮る	映像を流す
……	……	……	……
(合計　55種類)		(合計　294種類)	

出できた表現の一部を**表1**に示す．

次に比較実験として名詞と動詞間の相互情報量を用いて，慣用表現の抽出を行った．相互情報量の定義は以下の式である．

$$I(x, y)=\log_2\frac{p(x, y)}{p(x)p(y)}$$

ただし$p(x)$はコーパス中にxが現れる確率，$p(x, y)$はこの場合，「を」格によって名詞xと動詞yが共起する確率である．

定義式からもわかるように，これは頻度が小さいと，非常に大きな値になり意味をなさない．このため頻度が5以上の共起データを対象にする．対象となる表現は820種類であった．上位約10%の80種類を抽出結果とする．また相互情報量の計算は厳密には，動詞，名詞のそれぞれの出現頻度に「を」格をとらないものも含めなくてはいけないが，ここでは最初に収集した共起データだけで計測する．実際はこのような近似をしても影響は少ないと考えられる．抽出結果の上位20種類を**表2**に示す．

3.2 評　価

実験の評価は市販の慣用表現辞典[14]との差異を調べることにより行う．市販の辞典こそが人間が主観的に収集した慣用表現であるため，市販の辞典との差異を調べることは，従来人手で行っていた収集をどの程度自動化できるかの目安になると思われる．

表2 相互情報量からの抽出
Table 2 Extraction by mutual information

順位	表　現	相互情報量
1	腎臓を摘出する	15.460
2	臨時国会を召集する	15.237
3	警鐘を鳴らす	15.197
4	上告を棄却する	14.460
5	一線を画す	14.122
6	端を発する	14.061
7	第一歩を踏み出す	14.044
8	神経をとがらせる	13.974
9	緊張を緩和する	13.957
10	たばこを吸う	13.857
11	身柄を拘束する	13.723
12	鉱業権を引き継ぐ	13.694
12	自国を防衛する	13.694
12	立候補を届け出る	13.694
15	国交を樹立する	13.685
16	幕を閉じる	13.666
17	一歩を踏み出す	13.652
18	武力行使を伴う	13.628
19	事情を聴く	13.612
19	足並みをそろえる	13.612
20	顔色をうかがう	13.586

候補となった共起データ2,030種類に対して，それぞれの表現が市販の慣用表現辞典に記載されているかどうかを調べた．その結果88種類の慣用表現が含まれていた．この88種類を正解と考えると，自動抽出した349種類の中で55種類が正解であった．再現率，適合率は以下のようになる．

再現率＝55/349＝0.15

適合率＝55/88＝0.62

3.2.1 未抽出の原因

未抽出の表現は33種類であった．これら表現のうち「しのぎを削る」「一線を画す」「知恵を絞る」「背を向ける」の4種は，名詞の語義が分類語彙表に記載されていないために，候補として外されたものである．

また表記のゆれ（ここでは，ひらがな表記と漢字表記）によって，名詞の集合内に同一の単語が別要素として混在してしまったために抽出できなかった表現が1種あった（「あと（後）を絶つ」）．本実験ではコーパスとして新聞記事を使い，表記が統一されていたので，特に表記のゆれへの対処は行わなかったが，本手法を行う上では重要なポイントである．

残りの28種類は本手法の置き換え不可能性を破る表現であった．それぞれ置き換えられる前の表現と置き換えた後の表現の一部を**表3**に示す．

「気を使う」が「気候データを使う」に置き換えられたように，字面情報による類似関係からの誤りが3種類あった．分類語彙表の見出しが増えれば，この問題はなくなるが，その他の原因によるものは，問題になりそうな部分の語義の分類を細かく見ていくことで対処は可能である．例えば，分類語彙表では体の部分を表す名詞は上位4レベルの分類番号では差が出ない

表3 置き換えられた表現
Table 3 Phrases exchanged by a similar noun.

元の表現	置き換えられた表現
けじめをつける	区切りをつける
めどをつける	理由をつける
圧力をかける	攻勢をかける
火をつける	かじをつける
顔をそろえる	口をそろえる
顔を見せる	笑顔を見せる
顔を出す	口を出す
気をつける	景気をつける
気を使う	気候データを使う
穴をあける	風穴をあける
口を開く	窓口を開く
視線を浴びる	注目を浴びる
手をつける	見通しをつける
手を広げる	指を広げる
……	……

ことが多い（例えば，「口」と「顔」，「手」と「ひざ」など）．慣用表現では体の部分を名詞にした表現が多いため，この点を細かく見てゆくことにすれば若干改良できるはずである．

3.2.2 誤抽出の原因

誤抽出の表現は294種類であった．誤抽出の原因は6つに分類できる．

原因1 類義語が分類語彙表に記載されていない（80種類）

例えば「英語を教える」という表現が抽出されている．「～を教える」の～の部分には「日本語」という名詞が現れているが，分類語彙表には「日本語」の見出しがないので，「日本語」が「英語」の類義語だとは判断されなかった．このような表現は対象となった類義語が分類語彙表に記載されていないための誤抽出と判断する．

原因2 分類語彙表の不備（34種類）

例えば，「立場」という名詞を含んだ以下のような表現が抽出されている．「立場をとる」「立場を改める」「立場を強調する」「立場を堅持する」「立場を支持する」「立場を表明する」．ここでの立場の語義は『見方，考え方』であるが，この語義は分類語彙表の「立場」には記載されていない*．

また「～をとる」「～を改める」「～を強調する」「～を堅持する」「～を支持する」「～を表明する」の～の部分には，いずれも「考え」「主張」「見解」などの『見方，考え方』を語義として持つ名詞が現れている．このような表現は分類語彙表の不備による誤抽出と判断する．

原因3 類似の捉え方の違いによる誤り（47種類）

例えば「みそを作る」という表現が抽出されている．「～を作る」の～の部分には「料理」「米」「弁当」「うどん」など食べ物に関するものは頻出しているが，それらの語義は「みそ」の語義とは異なり，類義語だと判断されていないために誤抽出されている．それらの名詞は食べ物という観点からは類義語だと判断できるが，物理的な形状や構成要素の観点からは類義語だと判断できない．このような表現は類似の捉え方の違いによる誤りと判断する．

ただしこのケースの誤りは，原因2の分類語彙表の不備による誤りとの区別が微妙なものもある．例えば，「仕事を休む」という表現をここでの分類としたが，「～を休む」の～の部分には「会社」「工場」「練習」といった名詞が現れている．例えば「会社」という名詞に『仕事』のような語義があるものと考えればこの表現は原因2の誤りとなる．この判断は筆者の主観的な判断で行った．

また原因1で分類番号の記述がないための誤抽出は，本質的には原因2あるいは原因3が原因であるが，原因1をこれ以上，分類することは行わなかった．

原因4 コーパスの不備（95種類）

例えば「日程を調整する」という表現が抽出されている．「～を調整する」の～の部分には「日程」と類似の語はなかった．しかし，「過程」や「スケジュール」という名詞を想定してみると，これらは「日程」と置き換え可能である．またこれらの単語はコーパスに出現しても不思議ではない．つまりコーパスを拡大し，より多くの共起データを利用することで，避けられると予想できる誤りをコーパスの不備による誤りと判断する．

ただしこれは先の原因2，原因3と区別できない．コーパスを拡大して下位分類に当たるような語が現れれば，誤抽出は避けられるためである．ここでは集合内に類義語だと判定できるものがあった場合は，原因2あるいは原因3のいずれかにすることにした．

原因5 適当な言い換えができないもの（35種類）

類義語が現れず，さらに適当な言い換えが思いつかない表現をこの分類とした．比較的定型性が認められる表現がこの分類に入る．**表4**にこの一部を示す．これらの表現は慣用表現との区別も微妙であると考えられる．

原因6 その他（3種類）

これは「手を振る」「手を組む」「金を稼ぐ」の3種類である．いずれも「右手」「腕」「外貨」といった分類語彙表上でも類義語となる名詞が出現していたが，「手」や「金」は多義語であるために，その他の語義が

表4 言い換え困難な表現
Table 4 Phrases without similar nouns.

顔を見合わせる	顔を上げる
疑問を呈する	興味を覚える
敬意を払う	思いをさせる
自殺を図る	手続きを踏む
傷をつける	心をとらえる
身を置く	人質をとる
足を止める	大台を割る
大台を突破する	弾みをつける
展望を開く	波紋を広げる
不信を買う	舞台を移す
舞台を踏む	面倒をみる
……	

* 「立場」の語義としては『位置』に相当するような語義だけが与えられている．

他の名詞からでは棄却できず，結果的に類義語を持たない語義が過半数を占めたためである．

3.2.3 相互情報量による抽出の評価

次に相互情報量を用いた抽出結果について述べる．候補となった表現は820種類である．それぞれの表現が先の慣用表現辞典に記載されているかどうかを調べた．その結果58種類の慣用表現が含まれていた．これらを正解と考えると，抽出結果の80種類中には14種類の正解があった．再現率，適合率は以下のようになる．括弧内の数値は本手法を用いた実験による値である．

再現率＝14/80＝0.18 (0.15)

適合率＝14/58＝0.24 (0.62)

また上位40種類を選んだ場合，正解は9種類に減少し，再現率，適合率は以下のようになる．

再現率＝9/40＝0.23 (0.15)

適合率＝9/58＝0.15 (0.62)

再現率は上げることができるが，それ以上に適合率が下がる．

3.3 結　　論

未抽出の原因のうち，分類語彙表に見出しがなかった4種は手法の問題ではない．また誤抽出の原因1の一部，原因2，原因4も手法自体の問題点ではないと考えることができる．原因1の誤りの半分を手法自体の問題点と考えると，改善される見込みのある誤抽出は169 (40＋34＋95) 種類である．これらから本手法は以下の値にまで改善可能であると言える．

再現率＝(55＋4)/(349－169)

＝0.33(0.15)

適合率＝(55＋4)/88

＝0.67(0.62)

値自体は良好な結果ではないが，定型表現の自動抽出では再現率を高くすることは難しいし[13]，この程度の値ならツールとしての利用法は可能だと考える．また誤抽出の原因5の扱いによっては，適合率も向上する．

結論として，以下の2点から，本手法は述語型慣用表現の抽出手法として有益であるといえる．

- 実験で用いた類義語の判定法は精度が粗く，そのために抽出の精度が落ちていた．この点の改良により，妥当な数値に改善の見込みがある．
- 相互情報量を使った従来手法と比較して，現段階でも，再現率が同程度の場合，適合率で2倍以上の値を得ている．

4. 考　　察

本手法は動詞を固定して名詞の語義の特異性から，慣用表現を取り出した．同様にして，名詞を固定し動詞の語義の特異性から慣用表現を抽出することや，両者を組み合わせることも考えられる．しかし，いくつかの実験を行ってみたが，名詞の語義の特異性から判定する以上の良好な結果は得られなかった．動詞の語義の特異性から慣用表現を判定する手法では，以下のような問題があると感じられた．

- 分類語彙表の動詞の分類番号をどのレベルで見ても，本論文の利用目的に限れば，動詞の類義語を適切に判定できない*．
- 動詞の場合，類似の捉え方が難しく，また字面の情報もうまく利用できない．
- 慣用表現でなくても置換えが起こり難い表現が多い．

動詞の類似性を適切に知る方法があれば，組み合わせて利用することで，本実験よりも良好な結果が得られると考える．

本手法は類義語の判定法の精度に大きく依存する．本実験の判定では分類語彙表の上位4レベルの一致を確認したが，この部分で更に効果的な方法を考案できれば精度が向上する．例えば詳細なシソーラスを利用することも改良になると考える．分類語彙表でも上位下位関係は摑めるが，上位3レベルあるいは上位5レベルの一致では，本実験以上の精度は出なかったことも注記しておく．また誤抽出の原因3は視点を考慮した分類の多次元化の問題であり，この点での対処はさらなる改良に継り，今後の研究発展を期待する[19]．

コーパスのスパース性について述べる．コーパスからの知識獲得の研究ではコーパスのスパース性の問題(頻度の小さい単語や表現が大部分を占めるという問題)が深刻である[20]．本手法の場合はこの点に関して，有利な手法になっている．それは候補の表現が慣用表現であるかどうかを判定するのに反例を見つけるという戦略をとっているからである．反例は1つ見つければ十分であり，小さなコーパスからでも発見が可能である．また慣用表現は分野に依存しない表現である．反例を探す場合には，別分野のコーパスも利用可能であり，コーパスの質を均等に保たなくても良い．

本手法は基本的に共起の強さを考慮していない．共起の強さに注目した相互情報量による基準だけでは有効な抽出ができないことからも単純に共起の強さを判定に組み入れることはできないことがわかる．ただし，慣用表現には一般に構成語の間に挿入語句が入る場合

* 分類語彙表にはサ変動詞の見出しがほとんどなく，サ変名詞の分類番号で代替させると通常の動詞とは必ず分類番号が異なってしまうという問題もある．

が少ないという性質もある．この点から共起データを収集する際に，副詞などの挿入がおきるものに関してはマイナスのポイントを与えておき，そのポイントが慣用表現の判定に役立つと予想している．またこのマイナスポイントは，共起データを収集する際に取り除いた以下のパターンの[A，B，E]にも与えることができる．

名詞(A)＋格助詞(B)＋名詞(C)＋
格助詞(D)＋動詞(E)＋句読点

5. おわりに

本論文では名詞の語義の特異性を利用して，述語型慣用表現をコーパスから自動抽出する手法について述べた．

述語型慣用表現は頻度も少なく，強い共起性を持たないものも多い．このため従来の手法では抽出が困難な表現であった．本手法は基本的に動詞と共起する名詞の語義は特異であるという性質を利用している．このため頻度が小さい，あるいは，共起性が弱いような慣用表現も抽出できる．

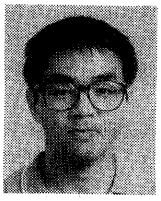
実験として，「を」格の共起データを対象に，本手法の有効性を確かめた．従来手法と比べると，抽出の再現率は同程度であったが，従来手法では取り出せない慣用表現を数多く抽出できることが確認できた．また類義語の判定法を改良することにより更に精度が良くなることも述べた．

今後は別種の格による実験，類似性の考察，本手法で得られた結果を利用しての名詞のクラスタリングなどを行いたい．

参考文献

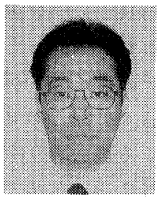
1) 国広哲弥ほか：特集—慣用表現—，日本語学，Vol. 4, 1月号（1985）．
2) 宮地　裕：慣用句の意味と用法，明治書院，東京（1982）．
3) 奥　雅博：日本語文解析における述語相当の慣用的表現の扱い，情報処理学会論文誌，Vol. 31, No. 12, pp. 1727-1734（1990）．
4) 鈴木克志，太田　孝：日英機械翻訳における共起表現の扱い，情報処理学会自然言語処理研究会資料，82-9（1991）．
5) 野見山浩：事例の一般化による機械翻訳，情報処理学会論文誌，Vol. 34, No. 5, pp. 905-912（1993）．
6) Church, K. W. and Hanks, P.: Word Association Norms, Mutual Information, and Lexicography, *Proc. ACL-89*, pp. 76-83（1989）．
7) Kita, K., Omoto, T., Yano, Y. and Kato, Y.: Application of Corpora in Second Language Learning — The Problem of Collocational Knowledge Acquisition, *Proc. Second Annual Workshop on Very Large Corpora*, pp. 43-56（1994）．
8) 平岡冠二，松本裕治：コーパスからの動詞の格フレーム獲得と名詞のクラスタリング，情報処理学会自然言語処理研究会資料，104-11, pp. 79-86（1994）．
9) 首藤公昭，吉村賢治，武内美津乃，津田健蔵：日本語の慣用表現について，情報処理学会自然言語処理研究会資料，66-1, pp. 1-7（1988）．
10) 新納浩幸，井佐原均：コーパスからの関係表現の自動抽出，情報処理学会論文誌，Vol. 35, No. 11, pp. 2258-2264（1994）．
11) 新納浩幸，井佐原均：疑似Nグラムを用いた助詞的定型表現の自動抽出，情報処理学会論文誌，Vol. 36, No. 1, pp. 32-40（1995）．
12) 北　研二，小倉健太郎，森元　逞，矢野米雄：仕事量基準を用いたコーパスからの定型表現の自動抽出，情報処理学会論文誌，Vol. 34, No. 9, pp. 1937-1943（1993）．
13) Smadja, F.: Retrieving Collocations from Text: Xtract, *Computational Linguistics*, Vol. 19, No. 1, pp. 143-177（1993）．
14) 井上宗雄：例解慣用句辞典，創拓社，東京（1994）．
15) 田中康仁，吉田　将：語と語の関係について—‘に’について—，情報処理学会自然言語処理研究会資料，73-5, pp. 33-42（1989）．
16) 中島弘之，梶　博行：テキストからの共起関係の自動抽出の試み，第38回情報処理学会全国大会論文集，2 E-6, pp. 325-326（1989）．
17) 井上和子：日本語文法小辞典，6-3 数量表現—「3匹の子豚」，大修館書店，東京（1989）．
18) 国立国語研究所：分類語彙表，秀英出版，東京（1994）．
19) 片桐康裕，宮崎正弘：シソーラスの多次元化のための観点の半自動抽出法，第49回情報処理学会全国大会論文集，3 G-9, pp. 3-151-3-152（1994）．
20) Dangan, I., Pereira, F. and Lee, L.: Similarity-Based Estimation of Word Cooccurrence Probabilities, *Proc. ACL-94*, pp. 272-278（1994）．

（平成6年12月20日受付）
（平成7年4月14日採録）



新納　浩幸（正会員）

1961年生．1985年東京工業大学理学部情報科学科卒業．1987年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，1993年4月より茨城大学工学部システム工学科助手，現在に至る．自然言語処理の研究に従事．人工知能学会，言語処理学会，ACL各会員．



井佐原　均

1954年生．1978年京都大学工学部電気工学第2学科卒業．1980年同大学大学院工学研究科電気工学専攻修士課程修了．同年通商産業省電子技術総合研究所入所．1995年より郵政省通信総合研究所関西先端研究センター知的機能研究室長．京都大学博士（工学）．主たる研究テーマは，自然言語処理，知識表現，機械翻訳など．日本認知科学会，人工知能学会，言語処理学会，ACLなど会員．