# 日本語複合辞用例データベースの作成と分析

土屋雅稔[†1] 宇津呂武仁[†2] 松吉俊[†3]
佐藤理史[†4] 中川聖一[†5]

日本語には，複数の形態素がひとかたまりとなって，1つの機能語相当語として働く表現が多数存在する．このような表現は一般に複合辞と呼ばれ，日本語文の構造を理解するために非常に重要である．特に，複合辞と同一の形態素列が本来の意味で構成的に用いられている場合と，非構成的に複合辞として用いられている場合とがあり，これらの用法を適切に識別できることが不可欠である．しかし，既存の解析系においては，多くの場合において，それらの用法の識別を適切に行っておらず，複合辞の取扱いが不十分であった．そこで，本論文では，複合辞を研究するための基礎資料として，複合辞用例データベースを作成する手順を提案し，実際に用例データベースの作成を行った結果を報告する．また，作成された用例データベースを用いて，日本語複合辞についての分析を行った結果を述べる．

# Development and Analysis of An Example Database of Japanese Compound Functional Expressions

Masatoshi Tsuchiya,[†1] Takehito Utsuro,[†2]
Suguru Matsuyoshi,[†3] Satoshi Sato[†4] and Seiichi Nakagawa[†5]

The Japanese language has many compound functional expressions which consist of more than one words including both content words and functional words. Those compound functional expressions are very important for recognizing the syntactic structures of Japanese sentences and for understanding their semantic contents. Unfortunately, recognition and semantic interpretation of them are especially difficult because it often happens that one compound expression may have both a literal (in other words, compositional) *content* usage and a non-literal (in other words, non-compositional) *functional* usage. Even widely-used Japanese text processing tools often fail in resolving the ambiguities of their usages. Aiming at being used as an example database for training/testing a tool which properly recognizes and interprets them, this paper proposes how to develop an example database of those Japanese compound functional expressions. This paper also presents the details of the developoed example database and the results of analyzing the example usages of the compound functional expressions.

## 1. はじめに

日本語には，複数の形態素がひとかたまりとなって，1つの機能語相当語として働く表現が多数存在する．このような表現は一般に複合辞と呼ばれ，日本語文の構造を理解するために非常に重要である．

例として，以下の2つの文を翻訳する場合を考える．

(1) 私は彼について話した．→ I talked about him.
(2) 私は彼について走った．→ I run following him.

この2つの文には「について」という表現が共通して現れているが，文(1)の「について」は前置詞 about と翻訳されるのに対して，文(2)の「について」は動詞 follow の現在分詞に翻訳される．このように区別して翻訳するには，文(1)中の「について」をひとかたまりの複合辞として検出すると同時に，文(2)中の「ついて」は本来の動詞の意味で用いられていると区別して，それぞれの文の構造を正しく把握しておく

†1 豊橋技術科学大学情報メディア基盤センター
Information and Media Center, Toyohashi University of Technology
†2 筑波大学大学院システム情報工学研究科
Graduate School of Systems and Information Engineering, University of Tsukuba
†3 京都大学大学院情報学研究科
Graduate School of Informatics, Kyoto University
†4 名古屋大学大学院工学研究科
Graduate School of Engineering, Nagoya University
†5 豊橋技術科学大学情報工学系
Department of Information and Computer Sciences, Toyohashi University of Technology

必要がある．

しかし，既存の解析系はいずれも，そのような処理を行っていない．たとえば，形態素解析器 JUMAN [1] と構文解析器 KNP [2] の組合せは，形態素解析時には複合辞を検出しない．構文解析時に，解析規則に記述された特定の形態素列が現れると，直前の文節の一部としてまとめたり，直前の文節からの係り受けのみを受けるように制約を加えたりして解析を行うといった，複合辞を意識した処理を行う．文 (1)，(2) はいずれも，直前の文節からの係り受けのみを受けるように制約を加えて構文解析され，それぞれの文の構造は区別されない．区別して処理する必要がある複合辞は，少なくとも 111 種類あるが☆，JUMAN/KNP では 21 種類（約 19%）しか区別されていない．他の例として，形態素解析器 ChaSen [3] と構文解析器 CaboCha [4] の組合せを利用して，IPA 品詞体系（THiMCO97）の形態素解析用辞書[5] を用いて形態素解析を行い，さらに京都テキストコーパス[6] から機械学習したモデルによる構文解析を行った場合を考える．この場合，形態素解析用辞書に「助詞・格助詞・連語」と登録されている複合辞は，形態素解析時に検出される．また，「〜ざるを得ない」などの表現は直前の文節の一部としてまとめられる．文 (1) は「助詞・格助詞・連語」の「について」と正しく解析されるが，文 (2) も「助詞・格助詞・連語」の「について」と解析されてしまい，2 つの文の構造は区別されない．区別して処理しなければならない 111 種類の複合辞のうち，この組合せでは 14 種類（約 13%）しか区別されない．

以上より，複合辞は，日本語の文構造を把握するときに重要な役割を果たしているにもかかわらず，従来の自然言語処理における複合辞の取扱いは不十分であることが分かる．このような現状を改善するには，複合辞である可能性がある形態素列が現れたときに，非構成的な意味を持つ複合辞として用いられているか，その形態素列本来の意味で構成的に用いられているかを区別できる検出器が必要である．本論文では，そういった検出器を作成するための基礎資料として，複合辞と同一の形態素列が，非構成的な意味を持つ複合辞として用いられている用例だけでなく，その形態素列本来の意味で構成的に用いられている用例を含み，かつ，それぞれの複合辞について十分な数の用例を含む複合辞用例データベースを作成する手順を提案する．加えて，実際に用例データベースを作成した結果と，作成した用例データベースを用いて日本語複合辞について分析した結果を述べる．

まず，本論文では，用例データベースに収録する複合辞のリストを選定するにあたって，現代語複合辞用例集[7]（以下，**用例集**と呼ぶ）を基礎資料とし，この用例集でとりあげられている全 125 項目のうちの 123 項目の複合辞を収録対象とした．次に，複合辞用例データベースを設計するにあたって，その中心部分を占める用例の収集方針としては，以下の点に留意した．

(1) 複合辞と同一の形態素列が，その形態素列の本来の意味で構成的に用いられている場合と，非構成的な意味で複合辞として用いられている場合があるので，それぞれの場合に対応した用例が必要である．

(2) 複合辞は，既存の形態素体系や形態素解析器と整合的ではない場合もあると予想されるが，そのような用例も必要である．

(3) 通常のコーパスを作成する場合は，一定量のテキストを用意し，そのテキスト全体に斉一的にタグ付与作業を行う．しかし，この方法では，対象テキスト中の複合辞の出現頻度によって，それぞれの複合辞の用例の数が変わってきてしまう．出現頻度の低い複合辞も含めて，十分な量の用例を確保するには，大量のテキストにタグ付与作業を行わなければならなくなる．そこで，本論文では，複合辞の種類ごとに，複合辞と同一の形態素列を含む用例を均等に収集し，これらの用例に対して複合辞の用法のタグ付けを行う．

また，本論文では，作成された複合辞用例データベースが，研究用に広く利用できることを考慮して，用例収集の際の出典としては，すでに研究用に広く利用されている新聞記事（毎日新聞）を採用した．

本論文の構成は以下のとおりである．最初に複合辞用例データベースに収録する複合辞のリストを作成し（2 章），次に複合辞用例データベースの仕様と作成手順を説明する（3 章）．4 章では，作成した複合辞用例データベースを分析して，用例集で説明されている用法の出現率などの統計量について報告する．5 章では，作成した複合辞用例データベースに基づいて，新聞記事における複合辞の様相を簡単に示す．6 章では，関連するコーパスと研究について述べ，最後に結論を述べる（7 章）．

## 2. 複合辞リストの作成

### 2.1 複 合 辞

複合辞とは，いくつかの語が複合してひとかたまりの形となって非構成的な意味を持ち，辞的な機能を果たす表現である．用例集では，複合辞は以下の 5 つに

☆ 表 8 で，「自然な文が作れる」に分類されている複合辞に相当する．詳細は 4.5 節を参照．

◇ **A56〜にとって・〜にとり**

**接続**　名詞（名詞節を含む）に付く.

**意味・用法**

「A にとって B」という形で文の内容を規定する形で用いられ,「A にとって B」が係っていく文の内容として述べられる個別的な判断・とらえ方をする主体を表す.

**用例**

(1) 技術的な問題（拡大・縮小や，ゆがみ，雑音など）はいろいろありますが，コンピュータ**にとって**「原理的に不可能」とはいえません.（野崎昭弘「人工知能はどこまで進むか」）

...

**文法**

「にとり」という言い方も，いささかぎこちないがなお可能である．連体修飾の言い方としては,「にとる」とそのまま連体形にしては用いられないが,「にとっての」という形でなら可能である.「にとりまして」という丁寧の形も取れる．とらえ方をする主体という立場を強調した言い方として (17)(18) のように「〜にとってみれば」という形もある.

**図 1**　用例集の項目例
Fig. 1　An example of entries of “Gendaigo Hukugouji Youreishu”.

分類されている.

A. 基本的に活用しない「**助詞的複合辞**」

**接続辞類**　基本的に節を受けて，複文前件を形成するもの（「〜とはいえ」など 36 項目）.

**連用辞類**　基本的に名詞を受けて，述語にかかる成分を形成するもの（「〜について」など 45 項目）.

**連体辞類**　名詞や節などを受けて，連体修飾句を形成するもの（「〜といった」など 2 項目）.

**文末辞類**　文末に付加されて，話し手のコミュニケーション上の様々な気持ちを示すもの（0 項目，詳しくは 2.2 節を参照）.

B. 述語の部分に付加されて活用する「**助動詞的複合辞**」（「〜つもりだ」など 42 項目）

本研究では，この 5 分類に，前の文を後ろの文に関係付ける働きをする**接続詞類**を加えて，複合辞を 6 つに分類する.

### 2.2　既存の複合辞リスト

用例集は，**図 1** のような形式の 125 項目の解説からなっている．それぞれの項目は,「A56 〜にとって・〜にとり」というような見出しと,「接続」「意味・用法」「文法」「ノート」といった説明文，および用例を含む．また，用例集では，複合辞は，2.1 節で述べたように 5 種類に分類されている．ただし，以下の例文の下線部のように終助詞的な働きをする文末辞類の複合辞は，用例集には収録されていない.

毎年大量の雨が降っているではないか

用例集以外に，複合辞を列挙したリストとしては以下の 2 つがある.

- 日本語表現文型[8)]
  7 機能的分類，52 意味大分類，210 意味小分類，450 表現
- 日本語文型辞典[9)]
  115 意味分類，965 表現，2,169 用法

用例集に収録されている複合辞は，この 2 つのリストに収録されている複合辞の一覧対照表を作成したうえで，日本語表現文型に収録されている複合辞を基本とし，その中でも 1 つの複合形式として熟合度が高く，また一般性も高いと判断される複合辞が選ばれている.

用例データベースの作成にあたっては，網羅的なデータベースの作成よりも，データベース作成時に生じる問題点の洗い出しを優先する．しかし，検出器作成のための基礎資料として考えると，用例データベースは，少なくとも，日本語文において一般的に用いられる主要な複合辞を網羅している必要がある．用例集に収録されている複合辞は，他の既存の複合辞リストよりも少ないが，主要な複合辞は網羅されていると考えられる．また，用例集には，用例データベースの作成時の判定基準として有用な，丁寧に記述された解説文と比較的多数の用例（16.6 文/項目）が含まれている点も好都合である．この 2 点より，用例データベースの作成時に参照するリストとしては，3 つの複合辞リストの中で用例集が最も適当と考える.

ただし，用例集は人間が閲覧するためのリストであり，意味的・機能的に似通った，ある範囲の異形は 1 つの項目でまとめて説明されている．つまり，異形が明示的に列挙されていないので，そのままの形では計算機から利用する複合辞リストとして不完全である．次節では，この問題を解決し，用例集に収録されている複合辞を列挙する方法について述べる.

### 2.3　収録対象とする複合辞の体系化

用例集では，意味的・機能的に似通った，ある範囲の異形は項目として区別されず，1 つの項目で説明さ

れている．そのため，そのままの形式では，用例データベースに収録する複合辞リストとしては不十分であり，異形を明示的に列挙・体系化したリストが必要である．本節では，最初に，収録する項目について説明し，次に，各項目で説明されている複合辞をすべて明示的に列挙・体系化する方法を説明する．

用例集は 125 項目からなっている．そのうち，「A66 ～といい～といい」および「A67 ～といわず～といわず」は，（連続しない）複数の要素の呼応という特別の形をとっているため，これらの項目で説明されている複合辞の用例を収集するには，特殊な工程が必要になる．そこで，今回のデータベース作成にあたっては，これら 2 つの項目は対象外として，123 項目を対象とする．

用例集の 1 つの項目では，ある範囲の異形がまとめて，項目して区別されずに説明されていることがある．たとえば，図 1 の項目には，「～にとって」や「～にとり」などの異形が区別されずに，1 つの項目にまとめられている．このような複合辞もすべて明示的に列挙・体系化するために，項目を表記などに着目して細分した**小項目**という単位を設ける．この方法には，以下のような利点も存在する．第 1 に，表記の違いと複合辞の用法の違いの間に何らかの関係がある可能性がある．たとえば，「～にとって」「～にとり」を比べると，「～にとり」という表現の方が少しぎこちない．このような関係を明らかにするには，表記の異なる複合辞を区別しておく方が都合が良い．また，「～にとり」と，その丁寧な形の表現「～にとりまして」を区別しておくと，敬体から常体への言い換えなどにデータベースを再利用できる可能性がある．第 2 に，大量のテキストから用例を機械的に収集する過程では，実際上は，表記ごとに用例を集めることになるから，表記ごとに収集された用例をわざわざまとめる必要性は低いと判断した．

用例集の 1 つの項目に含まれている複数の小項目を区別するため，各桁が以下のような意味を持つ 4 桁の枝番号を設定し，この枝番号と項目 ID を組み合わせて小項目 ID とする．

1 桁目：助詞の挿入や脱落および交替，同意語の交替などによって，表記の一部が異なっている異形を区別．

2 桁目：文体を区別．
0 = 常体，1 = 敬体，2 = 口語体

3 桁目：以下の表現を区別．
0 = 基本形，1 = 連体修飾形，
2 = 否定の変化形，3 = 否定形

4 桁目：1 桁目～3 桁目がまったく同じである複数の小項目を区別するための一意な番号（0, 1, 2, . . .）．

例として，図 1 の項目（A56）を小項目に分割し，それぞれの小項目を区別するための 4 桁の枝番号と項目 ID を組み合わせた小項目 ID を付与した結果を以下に示す．

見出し：A56 ～にとって・～にとり
A56-1000：にとって
A56-1010：にとっての
（←「～にとって」の連体修飾形）
A56-1100：にとりまして
（←「～にとって」の丁寧形）
A56-2000：にとり
A56-3000：にとってみれば

ただし，1 つの小項目は，つねに 1 種類の表記にだけ対応するわけではなく，複数の表記に対応する場合もある．「～におうじ」と「～に応じ」のように平仮名と漢字の違いは，1 つの小項目にまとめている．また，「～てならない」が形容動詞語幹に後続して「～でならない」のように，複合辞の先頭が濁音に変化する場合も区別せずに，1 つの小項目にまとめている．

また，連用辞類に属する表現「～について」などは，助詞「は」「も」が後続することによって提題助詞的または副助詞的に働くことがあるが，これらの表現について個別の小項目は立てて区別することはしなかった．

最終的に，用例集に掲載されている 125 個の項目から，「A66 ～といい～といい」および「A67 ～といわず～といわず」を除いた 123 項目を 337 個の小項目に分割し，収集対象として選定した．

## 3. 複合辞用例データベースの設計と作成

### 3.1 用例データベースの仕様

用例データベースは，項目，小項目，用例の 3 つの単位から構成されている．

項目は，(1) 見出し語と，(2) 項目 ID および (3) 1 つ以上の小項目からなる．見出し語と項目 ID は，用例集の項目に完全に準拠している．たとえば，図 1 に準拠した項目では，見出し語は「～にとって・～にとり」，項目 ID は A56 である．

小項目は，(1) 小見出し語，(2) 小項目 ID および (3) 用例（複数）からなる．小見出し語は，この小項目の可能な表示（表記と読みの組）のリストである．多くの小見出し語には，少なくとも形式的には内容語と分類される形態素が含まれている．たとえば，図 1 の「～にとって」には動詞「とる」が含まれている．

**表 1** 判定ラベル体系
Table 1 A system of decision labeling.

| 判定ラベル | 判定単位 | 読み | 内容 vs. 機能 | 用法 | 複合辞 |
|---|---|---|---|---|---|
| B | 不適切 | | | | — |
| Y | 適切 | 不一致 | | | × |
| C | 適切 | 一致 | 内容的 | 内容的用法 | × |
| F | 適切 | 一致 | 機能的 | 用例集で説明されている用法 | ◎ |
| A | 適切 | 一致 | 機能的 | 接続詞的用法 | ○ |
| M | 適切 | 一致 | 機能的 | その他の機能的用法 | ○ or × |

用例 ID: A56–1000–003
収集元 ID: MNP–950115192–6
テキスト (下線部がターゲット文字列):
大阪・関西にとって試金石だと思う.
判定ラベル： F
備考： （なし）

**図 2** 用例データベース中の用例
Fig. 2 An entry of the example database.

そのため，ある小見出し語と同一の形態素列が，内容語の本来の意味で用いられている場合があるが，そのような区別を説明文だけで記述することは大変困難であり，具体的な用例を多数示すことが重要である．用例集では，平均すると 1 項目あたり 16.6 文の用例文が収録されている．本データベースでは，用例集で説明されている複合辞用法で用いられている用例と，それ以外の用法で用いられている用例の両方を収録するため，少なくとも 2 倍の数の用例が必要である．しかし，最初から大規模なデータベースを作ることは困難なので，データベース作成にあたっての問題点を明らかにするのに必要かつ十分な規模として，1 小項目あたり 50 個の用例を収集することにする．

用例は，(1) 用例 ID，(2) 収集元の記事 ID，(3) テキスト，(4) ターゲット文字列，(5) 判定ラベルおよび (6) 備考からなる．**図 2** に例を示す．用例 ID は，小項目 ID に用例を識別するための一意な自然数（3 桁）を加えたものである．収集元の記事 ID は，この用例のテキストを収集した記事を表す．本データベースの作成にあたっては，研究用に広く利用できること，および，大量のテキストが収集できることの 2 点を考慮して，毎日新聞からテキストを収集することにした．ターゲット文字列は，文字列のみに基づいて判断すると複合辞である可能性がある部分であり，テキストは，ターゲット文字列を含む文である☆．判定ラベルは，ターゲット文字列が文中において果たしている働きを人手で判定した結果を表す．

☆ 文は，句点を手がかりとして機械的に分割した．

### 3.2 判定ラベル体系

判定ラベルとは，ターゲット文字列が文中でどのような働きをしているかを表すラベルであり，本データベースでは**表 1** のとおり，6 種類のラベルを設定している．判定ラベル付与とは，用例 ID，収集元 ID，文およびターゲット文字列が与えられたときに，判定ラベルを確定する作業のことである．

任意の文とターゲット文字列が与えられたとき，ターゲット文字列の用法を判定することができる場合と，判定できない場合とがある．本データベースの作成にあたっては，ターゲット文字列が 1 個以上の語，複合辞または慣用表現からなる列であるとき，そのターゲット文字列は判定単位として適切であり，用法を判定することができるとする．IPA 品詞体系の形態素解析用辞書に登録されている形態素を語とし，複合辞リスト（2.3 節）に収録されている 337 個の小項目と，収録されていないが用例中に現れた 24 種類の表現を複合辞とした．また，慣用表現は，「気にかける」などのように複数の語がひとかたまりとなって非構成的な意味を持ち内容的に働いているような表現であり，用例中には 38 種類が現れた．

判定ラベル B は，ターゲット文字列が判定単位として不適切であることを表すラベルである．たとえば，文 (3) のターゲット文字列は助詞「に」と副詞「とりあえず」の一部からなっており，判定単位として不適切であるから，文 (3) には判定ラベル B を付与する．

(3) 震災直後にとりあえずスタッフを出動させることができ，速やかに救援活動に入れる

判定ラベル Y は，ターゲット文字列の読みが，判定対象となっている小項目の読みと一致していないことを表す．たとえば，「〜うえは（A14–1000）」の用例として文 (4) を判定する場合，ターゲット文字列の読みは「じょうは」であり，小項目の読み「うえは」と一致していない．このような文には，判定ラベルとして Y を付与する．

(4) 法律上は困難でも，もう少し組織的に救援活動に参加する道がないか考えたい

判定ラベル C は，ターゲット文字列に内容的に働い

ている語が含まれていることを表す．たとえば，文(5)のターゲット文字列中の動詞「とる」は本来の意味で内容的に働いているので，判定ラベルとして C を付与する．

(5) まな板にとっていねいに納豆のタタキを作りみそ汁の実にするのである．

判定ラベル F，A，M は，ターゲット文字列が機能的に働いているとき，その機能を区別するためのラベルである．判定ラベル F は，ターゲット文字列が用例集で説明されている用法で働いていることを表す．判定ラベル A は，ターゲット文字列が接続詞的に働いていることを表す．判定ラベル M は，これら以外の機能的な働きをしていることを表す．たとえば，「A ところで B」の形で逆接の意味に用いられる「〜ところで（A22–1000)」の用例として，文(6)〜(8)を判定する場合を考える．

(6) 受験などでは倍率が上がったところで入学金があがることはない．

(7) ところで，全国の桜の名所では近年，樹勢の衰えが目立ち，保護対策に頭を痛めているという．

(8) 浜ノ島はあと一歩のところで勝ち星に結び付かず負け越した．

文(6)のターゲット文字列は，用例集で説明されているとおりに逆接の働きをしているので，判定ラベルとして F を付与する．文(7)のターゲット文字列は，文頭にあって接続詞的に働いているので，判定ラベルとして A を付与する．文(8)のターゲット文字列に含まれる名詞「ところ」は，形式的に働いているので，文(8)には判定ラベルとして M を付与する．

ターゲット文字列が機能的に働いていることを意味する 3 つの判定ラベル F，A，M のうち，用例集で説明されている用法で用いられていることを表す判定ラベル F が，最も判定基準が明確な判定ラベルである．そのため，以後の論述にあたっては判定ラベル F が付与されたターゲット文字列と，その用例を中心に考察を行う．

### 3.3 作成手順

用例データベースの作成手順の概略は以下のとおり．

(1) 新聞記事から 50 文を収集．

(2) 作業者による判定ラベル付与．

(3) 別の作業者による判定ラベルの検証．

(4) 用例集の用法で用いられている用例が 10 個以上含まれているか調査．含まれていない場合は，用例集の用法で用いられている可能性が高い用例を補充収集．

(5) 用例集の用法で用いられていない用例が 10 個以上含まれているか調査．含まれていない場合は，用例集の用法で用いられていない可能性が高い用例を補充収集．

最初に，毎日新聞（1995 年）から複合辞が用いられている可能性があるターゲット文字列を含む文を収集する．このとき，既存の形態素体系と整合しないような文を含めて収集するために，文字列一致による収集と基本形を考慮した収集を組み合わせた．なお，文字列一致による収集の妥当性は 4.4 節で述べる．

**(a) 文字列一致による収集** 小見出し語を含む文を無条件に収集する．「〜として（A62–1000)」の収集例を以下に示す．

助手として働く
彼はきちんとしている
財布を落として困っている

**(b) 基本形を考慮した収集** 小見出し語の末尾形態素が活用して用いられている場合を収集する．以下に，「〜つつある（B35–1000)」の収集手順(1)〜(3)を示す．

(1) 形態素解析器 MeCab[10] を利用して，IPA 品詞体系の形態素解析用辞書に基づいた形態素解析を行う．
台風/は/本土/を/北上/し/つつ/あっ/た

(2) 文中の活用している語の 1 つだけを基本形に置き換えた文を生成．
台風/は/本土/を/北上/する/つつ/あっ/た
台風/は/本土/を/北上/し/つつ/ある/た

(3) 小見出し語「つつある」と一致し，かつ，一致部分の先頭と末尾の位置が形態素区切りとなっている部分が検出されれば，この文を収集する．
台風/は/本土/を/北上/し/つつ/ある/た

収集された文が 50 文以上になった場合は，均等に 50 文を取り出して，判定ラベル付与の対象とする．収集された文が 50 文に満たなかった場合は，文収集の対象とする新聞記事の範囲を毎日新聞（1991 年〜1999 年）に広げて，50 文を確保する．ただし，本データベース作成においては，まず 50 文が収集された小項目のみに集中して取り組むことにする．

次に，収集された文を対象として作業者による判定ラベル付与を行い，その結果を別の作業者によって検証する．判定ラベル付与作業における作業者間の一致度は，4.2.2 項で報告する．

続いて，検証された 50 個の用例に，判定ラベル F

が付与された用例が 10 個以上含まれているか調べる．含まれていなかった場合は，用例集の説明文に記述されている接続制約を利用して，判定ラベル F が付与される可能性が高い文に重点をおいて補充する．たとえば，図 1 の項目には「名詞につく」という接続制約が記述されており，この接続制約を満たすターゲット文字列は，満たさないターゲット文字列に比べて，用例集の用法で用いられている可能性が高いと予想される．そこで，接続制約を満たすターゲット文字列を含む文を 40 個，それ以外の文を 10 個，新聞記事から収集する．このようにして，判定ラベル F が付与された用例と，それ以外の判定ラベルが付与された用例が，なるべくバランス良く含まれるようにした 50 用例を追加し，作業者による判定ラベル付与と別作業者による検証を行う．

さらに，検証された 50 個の用例に，判定ラベル F 以外の判定ラベルが付与された用例が 10 個以上含まれているか調べる．含まれていなかった場合は，用例集の説明文に記述されている接続制約を利用して，判定ラベル F が付与される可能性が低い文を 30 個，それ以外の文を 20 個，新聞記事から収集する．このようにして，判定ラベル F の用例と，それ以外の判定ラベルの用例が，なるべくバランス良く含まれるようにした 50 用例を追加し，作業者による判定ラベル付与と別作業者による検証を行う．

## 4. 複合辞用例データベースの評価

### 4.1 基本的な統計

毎日新聞（1995 年）からの用例収集結果を**表 2** に示す．用例集の項目を単位とすると，114 項目について 50 個以上の文が収集された．小項目を単位とすると，187 小項目について，50 個以上の文が収集された．

毎日新聞（1995 年）から 50 文以上が収集された 187 小項目について，人手による判定ラベル付与と検証を行った．187 小項目の 9,350 用例のうち，判定ラベル F が付与された用例（用例集の用法と判定された用例）は 6,271 個（67.1%）だった．187 小項目を，判定ラベル F が付与された用例の出現率によって分類した結果を**表 3** に示す．

表 3 より，99 小項目は判定ラベル F 以外の判定ラベルが付与された用例が不足しており，33 小項目は判定ラベル F が付与された用例が不足している．そのため，これらの 132 小項目については，毎日新聞（1991 年〜1999 年）を対象として接続制約を考慮した補充収集を行った．補充収集された 50 文に対して，人手による判定ラベル付与と検証を行った結果を**表 4** に示す．

**表 2** 新聞記事から収集された文数
Table 2 Number of sentences collected from newspaper.

| | 項目数 | 小項目数 |
|---|---|---|
| 50 ≤ 文数 | 114 (93%) | 187 (55%) |
| 0 < 文数 < 50 | 9 (8%) | 117 (35%) |
| 文数 = 0 | 0 (0%) | 33 (10%) |
| | 123 | 337 |

**表 3** 判定ラベル F の出現率
Table 3 Occurrence ratio of label “F”.

| 出現率 $x$ | 小項目数 | |
|---|---|---|
| $x$= 100% | 61 | (33%) |
| 80% <$x$< 100% | 38 | (20%) |
| 20% ≤$x$≤ 80% | 55 | (29%) |
| $x$< 20% | 33 | (18%) |
| 計 | 187 | |

$$x = \frac{\text{判定ラベル F が付与された用例数}}{\text{用例数}}$$

**表 4** 補充収集した小項目における判定ラベル F の出現率
Table 4 Occurrence ratio of label “F” in sub-entries with supplemented examples.

| 出現率 $x$ | 小項目数 | |
|---|---|---|
| $x$= 100% | 40 | (31%) |
| 80% <$x$< 100% | 37 | (28%) |
| 20% ≤$x$≤ 80% | 43 | (33%) |
| $x$< 20% | 12 | (8%) |
| 計 | 132 | |

### 4.2 判定ラベル付与

#### 4.2.1 判定ラベル付与に要する作業量

筆者らが判定ラベルの検証作業を行ったところ，「〜をもって（A74–1000）」「〜ものだ（B1–1000）」「〜ことだ（B11–1000）」の 3 小項目は，特に判定が難しいことが明らかになった．

判定ラベル付与の作業対象となる 187 小項目から，これらの特に判定が困難な 3 小項目を除いた 184 小項目について，新規の作業者に判定ラベル付与作業を依頼し，判定ラベル付与に要する作業量を調べた☆．184 小項目は，補充収集の必要がなかった 55 小項目から特に判定が困難な 2 小項目を除いた 53 小項目と，補充収集を行った 132 小項目から特に判定が困難な 1 小項目を除いた 131 小項目からなる．補充収集の必要がなかった 53 小項目については，均等に収集された 50 用例を対象とし，補充収集を行った 131 小項目については，補充収集した 50 用例を対象として，判定ラベル付与作業を行った．

184 小項目に判定ラベル付与を付与する作業には，

☆ 作業者は，文学研究科で言語学を専攻している大学院生である．

**表 5** 作業者間の判定の一致度（全 184 小項目）
Table 5 Agreement ratio between annotators (184 Sub-entries).

| $X$ | $\bar{X}$ | 平均値 | | 小項目数 | | |
|---|---|---|---|---|---|---|
| | | $P_a$ | $\kappa$ | $0.8 < \kappa \leq 1$ | $0.67 < \kappa \leq 0.8$ | $\kappa \leq 0.67$ |
| B | Y or C or F or A or M | 0.97 | 0.77 | 126 (69%) | 11 (6%) | 47 (26%) |
| F or A or M | B or Y or C | 0.93 | 0.73 | 120 (65%) | 19 (10%) | 45 (25%) |
| F | B or Y or C or A or M | 0.96 | 0.85 | 144 (78%) | 14 (8%) | 26 (14%) |

37 時間を要した．平均すると，1 小項目（= 50 用例）あたり，12 分かかっていることになる．小項目単位でかかった作業時間は，記録していない．ただし，判定ラベル F の用例が極端に多い小項目はかなり判定が簡単で，もっと短時間で判定ができた☆．それに対して，判定ラベル F の用例とそれ以外の判定ラベルの用例が適度に混ざっている小項目や，形式名詞を含む小項目などは，かなり時間がかかった．

#### 4.2.2 判定ラベル付与作業の一致度

作業者がまったく独立に判定ラベル付与作業を行った場合に判定ラベルが一致する割合を検討する．判定が作業者間でどのくらい一致しているかを調べるには，次式によって求められる $\kappa$ 値がよく用いられる[11),12)]．

$$\kappa = \frac{P_a - P_e}{1 - P_e} \tag{1}$$

ここで，$P_a$ は 2 人の作業者の判定が実際に一致した割合，$P_e$ は 2 人の作業者の判定が偶然に一致する確率である．2 人の作業者が一致して判定ラベル $X$ を付与した用例の数を $a(X)$，2 人の作業者が一致して判定ラベル $X$ 以外を付与した用例の数を $a(\bar{X})$，すべての用例の数を $n$ とすると，$P_a$ は次式で求められる．

$$P_a = \frac{a(X) + a(\bar{X})}{n} \tag{2}$$

また，ある作業者が判定ラベル $X$ を付与した用例の数ともう 1 人の作業者が判定ラベル $X$ を付与した用例の数の和を $c(X)$ とすると，$P_e$ は，次式で求められる．

$$P_e = \left(\frac{c(X)}{2n}\right)^2 + \left(1 - \frac{c(X)}{2n}\right)^2 \tag{3}$$

$\kappa$ 値の最大値は 1 であり，値が大きいほど，2 つの判定結果の一致は偶然ではなく，その判定結果は信頼できる．Carletta[13)] は，$\kappa > 0.8$ の場合の判定結果は完全に信頼でき，$0.67 < \kappa < 0.8$ の場合の判定結果はおおむね信頼できると報告している．

筆者らが検証を行った判定ラベルと，4.2.1 項で述べたように新規の作業者によって付与された判定ラベルを対象として，以下の 3 つの条件で，小項目ごとに $P_a$ と $\kappa$ 値を計算した結果を**表 5** に示す．

- $X$ = B，$\bar{X}$ = Y or C or F or A or M：2 人の作業者の判定が，ターゲット文字列が判定単位として適切か否かを判定する段階の一致度．
- $X$ = F or A or M，$\bar{X}$ = B or Y or C：2 人の作業者の判定が，ターゲット文字列が機能的に働いているか否かを判定する段階の一致度．
- $X$ = F，$\bar{X}$ = B or Y or C or A or M：2 人の作業者の判定が，ターゲット文字列が用例集で説明されている用法で働いているか否かを判定する段階の一致度．

表 5 より，多くの小項目（75%〜85%）については，作業者による判定結果は信頼できるが，一部の小項目では信頼できないことが分かる．$\kappa \leq 0.67$ となった小項目について，判定結果が一致しない原因を人手で分析した．たとえば，「〜とはいえ（A2–1000）」は，ターゲット文字列が判定単位として適切か否かを判定する段階で，$P_a = 0.70$，$\kappa = -0.17$ と，作業者による判定結果と検証結果が大きく異なっていた．判定ラベルが一致しなかった用例はすべて，文 (9) のような「〜とはいえない」という形の表現だった．

(9) 地方自治が十分定着したとはいえない

この用例について，筆者らは動詞「言う」に助動詞「ない」が後続した表現として判定ラベル C を付与していたのに対し，作業者は「いえない」を 1 語の動詞として判定ラベル B を付与していた．したがって，「〜とはいえ（A2–1000）」については，判定ラベル付与の作業マニュアルに方針を明示することにより，一致度を改善することができる．他の小項目についても，一致度を下げている原因（慣用表現や用例集には掲載されていない複合辞など）を特定できており，判定ラベル付与の作業マニュアルに方針を明示することにより，一致度を改善することができるという見通しを得ている．

以上の考察に基づき，本データベースの用例に対する判定ラベルの付与について，以下のとおり結論する．判定単位として適切か否かの判定は，75%の小項目について安定して行うことができ，機能的に働いているか否かの判定も，75%の小項目について安定して行

☆ 5 分以内の場合もあった．

うことができる．用例集で説明されている用法として働いているか否かの判定（判定ラベル F の付与）は，85%の小項目に対して安定して行うことができる．残りの小項目については，判定が不安定になる要因を特定できており，安定した判定が可能であるという見通しが得られている．

### 4.3 用例集で説明されている用法以外の複合辞的用法

187 小項目について均等に収集した 50 用例を対象として，用例集で説明されている用法と，それ以外の機能的用法の割合を調査する．

187 小項目のうち，判定ラベル F が付与された用例（用例集で説明されている用法の用例）または判定ラベル M が付与された用例（用例集で説明されている用法および接続詞的用法以外の機能的用法の用例）が存在する小項目は，48 小項目（27%）である．そのうち，用例集に説明されている意味と異なる非構成的な意味が存在する小項目は 19 小項目（10%）である．つまり，用例集で説明されている意味・用法は，それらの複合辞の意味・用法の大部分をカバーしているといえる．

### 4.4 用例収集方法の妥当性

本研究では，複合辞として用いられている可能性がある候補部分を 2 通りの方法で収集している．第 1 の方法は文字列一致による収集であり，第 2 の方法は基本形を考慮した収集である．ここでは，特に文字列一致による収集が必要であるかを検討する．

複合辞は，複数の語がひとかたまりとなって辞的な機能を果たす表現と定義されるから，何らかの形態素解析器を利用し，候補部分の先頭と末尾が形態素境界となっている場合だけをターゲット文字列として検出することが考えられる．しかし，このような制約を追加して収集を行うと，その形態素解析器の形態素体系と矛盾するような小項目について，用例の取りこぼしが発生する．たとえば，「〜までもない（B17–1000）」という小項目について，文字列一致によって文を収集すると，以下のような文が発見される．

大阪の発展につながることは言うまでもない.

しかし，この文を IPA 品詞体系の形態素解析用辞書に基づいて形態素解析すると，末尾部分は「言うまでもない」という形容詞 1 語と解析されるため，「までもない」の先頭は形態素境界とならず，用例の取りこぼしが生じることになる．

文字列一致によって収集された用例を IPA 品詞体系の形態素解析用辞書に基づいて形態素解析したところ，以下の 7 個の小項目について，判定ラベル F が付与されているにもかかわらず，ターゲット文字列の先頭・末尾が形態素境界とはなっていない用例が見つかった．

〜うが（A10–2000），〜に比べ（A45–2000），〜からすると（A77–3000），〜ほかない（B7–1000），〜までもない（B17–1000），〜てはいけない（B29–3000），〜て仕方がない（B33–5000）

これらの小項目については，形態素解析結果を利用した収集を行うと，用例の取りこぼしが発生することになる．

逆に，文字列一致によって用例を収集すると，用法を判定するのには不適切な文までも収集してしまう可能性がある．187 小項目を対象として均等に収集された 50 用例中の，用法を判定する単位として不適切だと判定された用例（判定ラベル B が付与された用例）の割合を**表 6** に示す．15 個（8%）の小項目では，用法を判定する単位として不適切と判定された用例が過半数を占めており，文字列一致による文収集が悪影響を与えている可能性がある．そこで，この 15 小項目について，用法を判定する単位として不適切と判定された理由を人手で調査した．理由は，以下の 3 通りに分類できる．

- 判定ラベル B が付与された用例の大多数は，他の複合辞と重なっていることが原因である ··· 4 小項目
  「ては（A29–1000）」の例を以下に示す．「について（A53–1000）」と重なっている．
  教育扶助**について**は，学校が休校していても支給を継続するなど弾力的に運用してきた．
- 判定ラベル B が付与された用例の大多数は，他の語と重なっていることが原因である ··· 8 小項目
  「える（B39–1000）」の例を以下に示す．
  地方自治を**考える**「列島ロジー」へのご意見，情報をお寄せ下さい．
- 両方の原因がある ··· 3 小項目

他の複合辞と重なっているために判定ラベル B が付与された場合は，データベース作成前には予見しえなかったものである．他の語と重なっているために判

**表 6** 判定ラベル B の出現率
Table 6 Occurrence ratio of label “B”.

| 出現率 $x$ | 小項目数 |
|---|---|
| 80% $\leq x \leq$ 100% | 3 |
| 50% $\leq x <$ 80% | 12 |
| 0% $\leq x <$ 50% | 172 |

**表 7** 内容語を含む小項目/含まない小項目
Table 7 Sub-entries with/without content words in morpheme sequencies.

| 判定ラベル F の出現率 $x'$ | 内容語を含む小項目 | 内容語を含まない小項目 |
|---|---|---|
| $x'=100\%$ | 72 (41%) | 2 (18%) |
| $80\% \leq x' < 100\%$ | 34 (19%) | 2 (18%) |
| $50\% \leq x' < 80\%$ | 26 (15%) | 2 (18%) |
| $5\% \leq x' < 50\%$ | 34 (19%) | 2 (18%) |
| $0\% \leq x' < 5\%$ | 10 (6%) | 3 (27%) |
| 計 | 176 | 11 |

$$x' = \frac{\text{判定ラベル F が付与された用例数}}{\text{判定ラベル F，A，M または C が付与された用例数}}$$

定単位が不適切と判定された場合は，形態素解析結果を利用することによって，そのような文を収集することを避けられたかもしれない．しかし，そのような小項目は 8 個であり，多くはない．

このように，文字列一致による収集を行うと，多少の悪影響はあるが，用例を取りこぼさずに収集できるという利点がある．本研究では，対象となる用例を取りこぼさずに収集することを重視し，文字列一致による候補部分の収集を行うことは妥当と考える．

### 4.5 新聞における用法の偏り

新聞記事から均等に収録された 50 用例において，判定ラベル F の出現率が非常に大きい小項目と非常に小さい小項目を対象として，このような用法の偏りが本当に新聞上で生じているか，このような小項目について様々な用法が適度に含まれるように用例を収集することができるかを検討する．

毎日新聞（1995 年）から 50 個以上の用例が収集された 187 小項目を，小見出し語に内容語が含まれているか否かによって分類した結果を**表 7** に示す．

#### 4.5.1 もっぱら用例集の用法で用いられている場合

新聞記事から均等に収集された 50 用例において判定ラベル F の出現率が非常に大きい小項目について，その小項目の内容的用法の文が自然に作れるかどうかを検討する．まず，小見出し語に内容語を含む 176 小項目を，その内容語が本来の意味で内容的に用いられている文を内省によって作り出すことができるかどうかによって，以下の 3 種類に分類する．

(1) 内容語が内容的に用いられている自然な文を作ることができる場合
例：「〜について（A53–1000）」
彼についていく

(2) 内容語が内容的に用いられている文を作ることはできるが，不自然な言い回しになってしまう場合
例：「〜に至るまで（A50–1000）」

**表 8** 自然な内容的用法の文が作れる小項目/作れない小項目
Table 8 Sub-entries with/without content usages.

| 判定ラベル F の出現率 $x'$ | 自然な文が作れる | 不自然な文しか作れない | 不自然な文も作れない |
|---|---|---|---|
| $x'=100\%$ | 21 (19%) | 10 (59%) | 41 (85%) |
| $95\% \leq x' < 100\%$ | 13 (12%) | 1 (6%) | 2 (4%) |
| $80\% \leq x' < 95\%$ | 13 (12%) | 3 (18%) | 2 (4%) |
| $50\% \leq x' < 80\%$ | 23 (21%) | 1 (6%) | 2 (4%) |
| $5\% \leq x' < 50\%$ | 32 (29%) | 1 (6%) | 1 (2%) |
| $0\% \leq x' < 5\%$ | 9 (8%) | 1 (6%) | 0 (0%) |
| 計 | 111 | 17 | 48 |

あの山の頂上に至るまでもう少しだ

(3) 内容語が内容的に用いられている文を作ることができない場合
例：「〜についての（A53–1010）」

分類結果を**表 8** に示す．

含まれている内容語が本来の意味で内容的に用いられている文が自然に作れるにもかかわらず，新聞記事から均等に 50 文を取り出した場合には，判定ラベル F の用例しか発見できなかった小項目が 21 個，判定ラベル F の用例が 95%以上を占めていた小項目が 13 個あった．この 34 小項目のうち，接続制約を考慮した補充収集対象となっている小項目は 33 小項目である☆．補充収集を行った結果，22 小項目については，判定ラベル F 以外の判定ラベルが付与された用例が見つかった．接続制約を考慮した補充収集を行っても，判定ラベル F 以外の判定ラベルが付与された用例が見つからなかった小項目は，以下の 11 小項目である．

〜に対する（A52–1011），〜にわたり（A58–2000），〜によっては（A60–1000），〜によれば（A61–1000），〜を問わず（A68–1000），〜ことになる（B14–1000），〜ほうがいい（B27–1000），〜ばいい（B28–3000），〜といけない（B29–2000），〜ても構わない（B30–6000），〜ても仕方がない（B31–3000）

これらの 11 小項目は，新聞上ではかなり偏って用いられているようである．したがって，これらの小項目を対象として，様々な用法の用例が適度に含まれるように用例を収集するには，用例集で説明されているよりも厳しい接続制約を利用するか，新聞以外の収集元を利用するなどの対策が必要である．

☆ 判定ラベル F の用例の比率 $x'$ が 95%以上であるにもかかわらず，補充収集対象となっていない小項目がある．これは，この小項目には，判定ラベル B を付与された用例があり，判定ラベル B を含めた判定ラベル F の用例の比率 $x$ は 80%以下となっているからである．

#### 4.5.2 もっぱら用例集で説明されている用法以外の用法で用いられている場合

表 7 より，10 小項目については，判定ラベル C の用例（内容的用法の用例）の比率が高く，判定ラベル F の用例が 5%未満となっており，判定ラベル F の用例が不足している．

これらの小項目は，用例集に記述されている接続条件を利用した補充収集の対象となっている．補充収集の結果，7 個の小項目は，判定ラベル F の用例の割合が 5%以上になった．しかし，「〜にかけて（A43–2000）」「〜ことだ（B11–1000）」と「〜に限る（B20–1000）」は，接続制約を考慮した補充によっても，判定ラベル F の用例が 5%未満だった．ただし，いずれの小項目も，判定ラベル F の用例を少なくとも 1 つは含んでいる．

したがって，これらの 3 小項目を対象として，判定ラベル F の用例が適度に含まれるように用例を収集するには，用例集で説明されているよりも厳しい接続制約を利用するか，新聞以外の収集元を利用するなどの対策が必要である．

## 5. 複合辞の統計分析

### 5.1 新聞記事における複合辞の出現頻度

用例データベースから，新聞記事（毎日新聞）上において用例集の用法で用いられている複合辞の出現率が推定できる．この推定された出現率を用いて，用例集の用法で用いられている複合辞の出現頻度を推定した結果を**表 9** に示す．毎日新聞（1995 年）から 50 個の用例が収集された 187 個の小項目については，出現頻度の分布に，あまり目立った偏りはないことが分かる．平均出現頻度は約 1400 回，最頻出小項目は約 45,000 回の「〜という（A82–1000）」と推定された．

毎日新聞（1995 年）は約 130 万文からなっている．この 187 個の小項目に限って考えると，用例集の用法で用いられている複合辞が，平均して 5 文に 1 つ現れていると推定される．言い換えると，複合辞の検出をまったく行わない場合には，20%の文について，その構造を正しく理解できないことになる．

### 5.2 新聞記事と話し言葉の比較

日本語話し言葉コーパス[14] は，学会講演を中心として，現代日本語の自発音声を研究用付加情報とともに大量に格納したデータベースである．話し言葉コーパスには，短単位・長単位の 2 種類の粒度の形態論的情報が付与され，用例集に収録されている複合辞の一部は長単位の付属語として認定されている．そのため，複合辞として用いられている可能性があるターゲット文字列を含む文を取り出し，そのターゲット文字列が 1 つの長単位の付属語と認定されている比率を求めることによって，話し言葉コーパスにおける複合辞の出現率を求めることができる．毎日新聞（1995 年）から 50 個以上の用例が収録された 187 小項目のうち，話し言葉コーパスで長単位の付属語として認定されている 49 小項目について，新聞記事における判定ラベル F の出現率と，話し言葉コーパスにおける複合辞の出現率を比較した．

この 49 小項目について新聞記事から均等に収録された 50 個の用例は，約 84%の割合で判定ラベル F が付与されている．それに対して，話し言葉コーパス中では，約 95%の割合で複合辞と認定されている．つまり，この 49 小項目に限ってみると，新聞記事よりも話し言葉の方が複合辞として用いられている比率が高くなっている．なお，新聞記事と話し言葉で，複合辞の出現率が大きく異なる小項目の例を**表 10** に示した．

**表 9** 新聞記事における複合辞の出現頻度（推定値，全 187 小項目）
Table 9 Frequency of compound functional usages in newspaper.

| 出現頻度 $f$ | 小項目数 | 例 |
|---|---|---|
| $5{,}000 \leq f$ | 10 (5%) | 〜べきだ（B41–1000） |
| $1{,}000 \leq f < 5{,}000$ | 35 (19%) | 〜にとって（A56–1000） |
| $500 \leq f < 1{,}000$ | 25 (13%) | 〜とはいえ（A2–1000） |
| $100 \leq f < 500$ | 63 (34%) | 〜かもしれません（B37–1100） |
| $50 \leq f < 100$ | 30 (16%) | 〜ても仕方がない（B31–3000） |
| $0 < f < 50$ | 24 (13%) | 〜にかけては（A44-1000） |

$f$ = 収集された文数 × 判定ラベル F の出現率 $x$

**表 10** 新聞記事と話し言葉の比較
Table 10 Comparison between newspaper and spoken language.

| 例 | 複合辞出現率 | |
|---|---|---|
| | 新聞記事 | 話し言葉 |
| 〜てもいい（B30–1000） | 98% | 59% |
| 〜なくてはならない（B42–4000） | 95% | 57% |
| 〜をめぐる（A73–1011） | 90% | 41% |
| 〜に比べ（A45–2000） | 82% | 44% |
| 〜にあって（A38–1000） | 60% | 12% |
| 〜に応じた（A42–1011） | 42% | 98% |
| 〜に従って（A51–1000） | 12% | 82% |

## 6. 関 連 研 究

RWC テキストデータベース[15] は，通産省報告書，日本電子工業振興会報告書および毎日新聞（1991 年〜1994 年）のテキストを，THiMCO95 体系の形態素に分割したコーパスである．このうち，毎日新聞（1994 年）を形態素分割したコーパスでは，59 種類の複合辞が「助詞・格助詞・連語」という形態素として扱われ

ている．59 種類の複合辞のうち，本データベースに収録されている複合辞は 43 種類（約 73%）である．本データベースの複合辞リスト（337 小項目）を基準とすると，この 43 種類の複合辞は 33 小項目（約 10%）に相当する．ただし，「助詞・格助詞・連語」という形態素に分割する作業と，用例集を基準として用法を認定する作業では，作業基準が自ずから異なってくる．たとえば，「〜に際して（A48–1000）」の場合，本データベースでは，以下のように動詞に「〜に際して」が後続している表現も，用例集で説明されている複合辞であると判定している．

防空演習を論評するに際して，その専門的知
識において驚くべき無智を表白した

それに対して，RWC テキストデータベースでは，このような表現は助詞「に」，動詞「際す」の連用タ接続形および助詞「て」と分割されている．

日本語話し言葉コーパス[14] は，現代日本語の自発音声を研究用付加情報とともに大量に格納したデータベースである．研究用付加情報には，音声の書き起こしテキストに対する短単位・長単位の 2 種類の粒度の形態素情報が含まれている．たとえば，「〜に際して（A48–1000）」は，短単位列としては助詞「に」，動詞「際す」の連用タ接続形および助詞「て」の 3 短単位に分割され，長単位としては助詞「に際して」という 1 長単位に分割されている．付属文書[16] によると，172 種類（助詞型 80 種類，助動詞型 92 種類）の複合辞が対象となっている．このうち，本データベースに収録されている複合辞は 73 種類（約 44%）である．本データベースの複合辞リスト（337 小項目）を基準とすると，この 73 種類の複合辞は 69 小項目（約 20%）に相当する．主な違いは，口語体や丁寧形などの異形である．

京都テキストコーパス[6] は，毎日新聞（1995 年）から取り出した 2 万文を対象として構文解析情報を付与したコーパスである．付与されている情報は，基本的に構文解析器 KNP によって出力される情報と同じであり，「〜ざるを得ない」などの 37 種類の決まり文句については，内容語が現れても文節区切りをしないことによって，通常の内容語としての用法と区別されている．加えて，コーパスの一部（5,000 文）には，格関係情報も付与されている[17]．格関係情報には，複合辞によって表現されている 25 種類の格関係が含まれている．このうち，本データベースに収録されている複合辞は 12 種類（約 48%）である．本データベースの複合辞リスト（337 小項目）を基準とすると，この 12 種類の複合辞は 20 小項目（約 6%）に相当する．

EDR 日本語単語辞書[18] には，86 種類の助詞相当語と 266 種類の助動詞相当語が登録されている．このうち，本データベースでは，54 種類の助詞相当語と 52 種類の助動詞相当語が収録対象となっている．これは，本データベースの複合辞リストを基準とすると，64 小項目（助詞型 38 小項目，助動詞型 26 小項目）に相当する．

首藤ら[19)〜22] は，複合辞や慣用表現を含む複数の形態素からなる定型的表現をできるだけ網羅的に収集し，複合辞間に類似度を定義して，複合辞の言い換えや機械翻訳に利用することを提案している．兵藤ら[23)〜25] と伊佐治ら[26] は，日本語の文構造の解析を容易にするため，通常よりかなり長い文節を単位として解析を行うことを提案し，複合辞を含む大規模な長単位機能語辞書を作成している．しかし，これらの先行研究における日本語処理系においては，複合辞と同一の形態素列が内容的に振る舞う可能性が考慮されていない．また，単一の形態素列について複合辞用法と内容的用法の両方を考慮して用例を収集したデータベースも整備されていない．

## 7. おわりに

本論文では，現代語複合辞用例集に収録されている複合辞を対象として，複合辞用例データベースの作成手順を提案した．複合辞とは，複数の形態素がひとかたまりとなって，1 つの機能語相当語として働く表現である．データベースの仕様と作成手順の設定にあたっては，複合辞と同一の形態素列が本来の意味で構成的に用いられている用例と，非構成的に複合辞として用いられている用例の双方が適度に収集されるように配慮した．さらに，実際に複合辞用例データベースを作成して，このようなデータベースが作成可能であることを示した．加えて，作成したデータベースを利用して，新聞記事における複合辞の出現頻度の推定を試み，新聞記事と話し言葉での複合辞の出現率の違いを調べた．

新聞記事は，研究目的に広く利用できるテキストを大量に収集できるという点で優れている．しかし，一部の小項目については，新聞上では用法が偏っているため，様々な用法の用例を適度に含むように収集することはできなかった．新聞記事以外の言語資源を利用して，そのような小項目の用例を収集することは今後の課題である．また，作成したデータベースを利用して，複合辞を適切に取り扱う検出器を実現することを計画している[27),28]．

本研究で作成したデータベースは，筆者らのウェブ

サイトで公開する予定である．

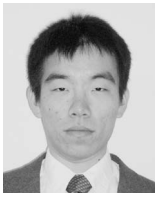
**謝辞** 本研究の一部は，次の研究費による：文部科学省科学研究費基盤研究（A）「円滑な情報伝達を支援する言語規格と言語変換技術」（課題番号 16200009），京都大学–NTT コミュニケーション科学基礎研究所共同研究「グローバルコミュニケーションを支える言語処理技術」．

## 参 考 文 献

1) 黒橋禎夫，河原大輔：日本語形態素解析システム JUMAN version 5.1 使用説明書 (2005). http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman/juman-5.1.tar.gz
2) 黒橋禎夫，河原大輔：日本語構文解析システム KNP version 2.0 使用説明書 (2005). http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp/knp-2.0.tar.gz
3) 松本裕治，北内 啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：形態素解析システム ChaSen version 2.3.3 使用説明書 (2003). http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.3.3-j.pdf
4) 工藤 拓，松本裕治：チャンキングの段階適用による係り受け解析，情報処理学会論文誌，Vol.43, No.6, pp.1834–1842 (2002).
5) 浅原正幸，松本裕治：ipadic version 2.6.1 ユーザーズマニュアル (2003). http://chasen.aist-nara.ac.jp/chasen/doc/ipadic-2.6.1-j.pdf
6) 黒橋禎夫，長尾 眞：京都大学テキストコーパス・プロジェクト，言語処理学会第 3 回年次大会発表論文集，pp.115–118 (1997).
7) 国立国語研究所：現代語複合辞用例集 (2001).
8) 森田良行，松木正恵：日本語表現文型，NAFL 選書 5，アルク (1989).
9) グループ・ジャマシイ：日本語文型辞典，くろしお出版 (1998).
10) 工藤 拓：形態素解析器 MeCab. http://chasen.org/~taku/software/mecab/
11) Chklovski, T. and Mihalcea, R.: Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation, *Proc. Conference on Recent Advances in Natural Language Processing* (*RANLP2003*) (2003).
12) Ng, H.T., Lim, C.Y. and Foo, S.K.: A Case Study on Inter-Annotator Agreement for Word Sense Disambiguation, *Proc. ACL SIGLEX Workshop on Standadizing Lexical Resource* (*SIGLEX99*), pp.9–13 (1999).
13) Carletta, J.: Assessing Agreement on Classification Tasks: The Kappa Statistic, *Computational Linguistics*, Vol.22, No.2, pp.249–254 (1996).
14) 前川喜久雄：『日本語話し言葉コーパス』の概観 ver.1.0 (2004). http://www2.kokken.go.jp/~csj/public/members_only/manuals/overview10.pdf
15) Hasida, K., Isahara, H., Tokunaga, T., Hashimoto, M., Ogino, S., Kashino, W., Toyoura, J. and Takahashi, H.: The RWC text databases, *Proc. 5th International Conference on Language Resources and Evaluation*, pp.457–652 (1998).
16) 小椋秀樹，山口昌也，西川賢哉，石塚京子，木村睦子：『日本語話し言葉コーパス』の形態論情報の概要 ver.1.0 (2004). http://www2.kokken.go.jp/~csj/public/members_only/manuals/pos_20040320.pdf
17) 河原大輔，黒橋禎夫，橋田浩一：「関係」タグ付きコーパスの作成，言語処理学会第 8 回年次大会発表論文集，pp.495–498 (2002).
18) 日本電子化辞書研究所：EDR 電子化辞書仕様説明書 (1993). http://www2.nict.go.jp/kk/e416/EDR/J_index.html
19) Shudo, K., Narahara, T. and Yoshida, S.: Morphological Aspect of Japanese Language Processing, *Proc. 8th International Conference on Computational Linguistics* (*COLING'80*), pp.1–8 (1980).
20) 首藤公昭，吉村賢治，武内美津乃，津田健蔵：日本語の慣用的表現について—語の非標準的用法からのアプローチ，情報処理学会研究報告，Vol.1988-NL-66, pp.1–7 (1988).
21) 首藤公昭，小山泰男，高橋雅仁，吉村賢治：依存構造に基づく言語表現の意味的類似度，電子情報通信学会研究報告，Vol.NLC98-30, pp.33–40 (1998).
22) Shudo, K., Tanabe, T., Takahashi, M. and Yoshimura, K.: MWEs as Non-propositional Content Indicators, *Proc. 2nd ACL Workshop on Multiword Expressions: Integrating Processing* (*MWE-2004*), pp.32–39 (2004).
23) 兵藤安昭，若田光敏，池田尚志：文節ブロック間規則による浅い係り受け解析と精度評価，電子情報通信学会研究報告，Vol.NLC98-30 (1998).
24) 兵藤安昭，池田尚志：文節単位のコストに基づく日本語文節解析システム，言語処理学会第 5 回年次大会発表論文集，pp.502–504 (1999).
25) 兵藤安昭，村上 裕，池田尚志：文節解析のための長単位機能語辞書，言語処理学会第 6 回年次大会発表論文集，pp.407–410 (2000).
26) 伊佐治和哉，山田将之，池田尚志：長単位の機能語を辞書に持たせた文節構造解析システム ibukiC，言語処理学会第 10 回年次大会発表論文集，pp.636–639 (2004).
27) 土屋雅稔，宇津呂武仁，佐藤理史，中川聖一：形

態素情報を用いた日本語機能表現の検出，言語処理学会第 11 回年次大会発表論文集，pp.584–587 (2005).
28) 注連隆夫，内元清貴，土屋雅稔，高木俊宏，宇津呂武仁，佐藤理史，井佐原均：機械学習を用いた日本語複合辞のチャンキング，情報処理学会研究報告，Vol.2005-NL-170 (2005).

(平成 17 年 10 月 21 日受付)
(平成 18 年 4 月 4 日採録)



**土屋 雅稔**（正会員）

1998 年京都大学工学部電気工学科第二学科卒業．2004 年同大学大学院情報学研究科知能情報学専攻博士課程単位認定退学．京都大学修士（情報学）．2004 年より豊橋技術科学大学情報メディア基盤センター助手．自然言語処理に関する研究に従事．

**宇津呂武仁**（正会員）

1989 年京都大学工学部電気工学第二学科卒業．1994 年同大学大学院工学研究科博士課程電気工学第二専攻修了．京都大学博士（工学）．奈良先端科学技術大学院大学情報科学研究科助手，豊橋技術科学大学工学部情報工学系講師，京都大学情報学研究科知能情報学専攻講師を経て，2006 年より筑波大学大学院システム情報工学研究科知能機能システム専攻助教授．自然言語処理の研究に従事．

**松吉 俊**（学生会員）

2003 年京都大学理学部卒業．2005 年同大学大学院情報学研究科修士課程修了．現在，同大学院情報学研究科博士後期課程在学中．自然言語処理の研究に従事．

**佐藤 理史**（正会員）

1983 年京都大学工学部電気工学第二学科卒業．1988 年同大学大学院工学研究科博士後期課程電気工学第二専攻研究指導認定退学．京都大学工学部助手，北陸先端科学技術大学院大学情報科学研究科助教授，京都大学大学院情報学研究科助教授を経て，2005 年より名古屋大学大学院工学研究科電子情報システム専攻教授．工学博士．自然言語処理，情報の自動編集等の研究に従事．

**中川 聖一**（正会員）

1976 年京都大学大学院博士課程修了．同年京都大学情報工学科助手．1980 年豊橋技術科学大学情報工学系講師．1990 年教授．1985〜1986 年カーネギメロン大学客員研究員．音声情報処理，自然言語処理，人工知能の研究に従事．工学博士．1977 年電子通信学会論文賞，1988 年 IETE 最優秀論文賞，2001 年電子情報通信学会論文賞各受賞．電子情報通信学会フェロー．著書『確率モデルによる音声認識』（電子情報通信学会編），『音声聴覚と神経回路網モデル』（共著，オーム社），『情報理論の基礎と応用』（近代科学社），『パターン情報処理』（丸善），『Spoken Language Systems』（編著，IOS Press）等．