様式 C-19

科学研究費補助金研究成果報告書

平成２３年６月１６日現在

機関番号：62618
研究種目：基盤研究(C)
研究期間：2008～2010
課題番号：20520428
研究課題名（和文）辞書用例の記述仕様標準化のための実証研究

研究課題名（英文）Practical Research for Lexical Illustration of the Actual Use of Words and Specification Standardization for their Lexical Description

研究代表者
柏野　和佳子（KASHINO　WAKAKO）
大学共同利用機関法人　人間文化研究機構　国立国語研究所・言語資源研究系・准教授
研究者番号：50311147

研究成果の概要（和文）：現在，コンピュータによるコーパス（言語のデータベース）からの用例抽出の自動化が進んでいる。自動的に抽出した大量の用例を有効に利用できる情報として蓄積するためには，人手による分析と編集を欠かすことができない。そこで，用例抽出後の，用例分析と用例記述過程の手続き化を行った。また，用例記述の仕様を策定した。加えて，コーパスから得られる辞書情報分析を 2 つ行った。1 つは，和語や漢語のカタカナ表記の使用実態を明らかにした。もう 1 つは，古い語の現代語としての使用実態を明らかにした。

研究成果の概要（英文）：As computational extraction of word uses from large language databases is becoming popular, manual analysis is becoming even more important in order to accumulate useful pieces of information from huge volumes of automatically stored text data. This study therefore addressed standardization of lexical description, and determined a set of standard processes for the use-case analysis and lexical description, clarifying specifications required for lexical description of word use. Furthermore, based on a newly-compiled large-scale Japanese corpus (BCCWJ), statistical analysis for Katakana representation for Japanese native words and ancient words was performed.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,500,000	450,000	1,950,000
２００９年度	1,200,000	360,000	1,560,000
２０１０年度	700,000	210,000	910,000
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：　人文学
科研費の分科・細目：日本語学，語彙・意味
キーワード：用例，辞書，国語辞典，コーパス

１．研究開始当初の背景

辞書の用例を充実させるためには，十分に収集できるための言語資源の整備，用例の抽出方法及び，記述仕様の確立が必要である。さらに，大規模な用例記述を実現させるためには，それら方法論の標準化，マニュアル化が必要になる。

現在，1 億語を超える規模の『現代日本語書き言葉均衡コーパス』が構築されつつある。また，用例の抽出に関しては，言語処理研究において格パターンやコロケーションの抽出といったツールの開発，公開が進められている。

自動的に抽出した大量の用例をより有効に利用できる情報として蓄積するためには，人手による分析と編集を欠かすことができない。そして，大規模な辞書用例集の構築に向けて，用例記述の仕様を策定する必要がある。

２．研究の目的

大規模な辞書用例集の構築に向けて，本研究では具体的に次の目的を立てる。

(1) 大規模コーパスを利用した用例抽出から，分析，記述に至る一連の過程を手続き化する。
(2) 用例記述の仕様の標準化，マニュアル化を計る。
(3) 小規模な用例集を構築することで，方法論を実証し，大規模構築への見通しを立てる。
(4) コーパス分析を辞書記述に活かす具体例を明らかにする。

３．研究の方法

(1)現行の国語辞典の用例分析

一般の国語辞典に必要な用例を明らかにするため，現行の国語辞典の多義語の意味記述と，用例の掲載状況を比較，分析する。対象とする辞書は,『岩波国語辞典』第六版（岩波書店),『新明解国語辞典』第六版（三省堂),『新選国語辞典』第八版（小学館),『明鏡国語辞典』(大修館書店)，以上 4 冊である。

(2)コーパスから得られる用例の分析

格フレーム検索（京都大学黒橋研究室）を用いて，京都大学黒橋研究室にて公開されている Web5 億文から自動構築した格フレームから得られる用例と,『現代日本語書き言葉均衡コーパス』の領域内公開データから得られる用例とを比較分析する。
また，格フレームから得られる用例と現行の国語辞典の記載用例との比較分析を行う。

(3)小規模用例集の構築

(2)と(3)より得られた用例を分析，編集し，辞書に載せるべき用例の記述を行う。

(4)仕様の策定

コーパスを利用した用例抽出から，分析，記述に至る一連の過程の手続き化,及び，用例記述の枠組みと仕様の策定を行う。

(5)コーパスから得られる辞書情報の分析

① コーパス分析を辞書記述に活かすという観点から,『現代日本語書き言葉均衡コーパス』を用いて，和語や漢語のカタカナ表記の実態を調査，分析する。
② 現代語の辞書記述では，古い語の見出しとしての採否や用法記述が問題になる点に着目し，国語辞典に「古語的」「古風」と注記される語について,『現代日本語書き言葉均衡コーパス』を用いて，その使用実態を調査，分析する。

４．研究成果

(1)用例提示が貢献すべき用途

現行の国語辞典の用例分析より，用例提示が貢献すべき用途を次のとおり，明らかにした。

(ア)意味・用法理解
(イ)文型理解
(ウ)慣用例の理解
(エ)典型例の理解
(オ)複合語・派生語・関連語の理解
(カ)位相・文体情報の理解
(キ)誤用の理解
(ク)実使用の証拠の理解
(ケ)例示の網羅性の理解
(コ)初出の提示

(2)格フレーム検索結果と国語辞典の記載用例との比較分析

辞書に記載すべき用例を明らかにするため，大規模コーパスで自動抽出される用例と，現行の国語辞典の記載用例との比較分析を行った。
25 の動詞を対象に，京大黒橋研究室で開発，公開されている，「格フレーム」(Web 上の約 16 億文の日本語テキストから自動的に構築したもの）の検索結果から，高頻度用例として抽出された用例と，国語辞典（岩波国語辞典，新選国語辞典，明鏡国語辞典）に記載されている用例とを比較した。
比較は次のとおり行った。

a.「格フレーム」の名詞（主にヲ格）が一致する用例が国語辞典にある場合
b.「格フレーム」の名詞は一致せずとも似た用例が国語辞典にある場合
c.「格フレーム」と似た用例がなく，その用例が分類できそうな語義が国語辞典にはない場合
d.国語辞典にある用例が「格フレーム」の用例として抽出できていない場合

調査した範囲内では上記 a.に該当するものは予想以上に少なく，多くは b.であった。c.の該当例はほとんどなかったが，d.の該当例はいくつか見つかった。

たとえば，動詞「使う」の場合，a.は「金／筋肉／気／神経／頭／英語／魔法を使う」の 7 例。また，d.に該当したものは「人を使う（のは難しい）」や「弁当を使う（＝食べる）」であった。

(3)小規模用例集の構築と仕様の策定

国語辞典の記載用例，格フレームの検索結果，『現代日本語書き言葉均衡コーパス』からの抽出用例を選別，分析し，辞書に記載すべき用例記述の分析を行い，動詞 15 語の用例記述を行った。また，コーパスからの用例抽出，分析，記述に至る一連の過程の手続き化,及び，用例記述方法のマニュアル化を行った。

これにより，コンピュータによって自動的に抽出される大量の用例の有効利用方法を示すことができた。この成果は，今後，大規模な用例集構築へつなげることができるものである。

(4) コーパスから得られる辞書情報の分析―カタカナ表記に関して―

先行研究では，新聞や若者雑誌に目立つカタタカナ表記語のタイプは，大きく分けて，①「表外漢字」「表外音訓」「表外熟字訓」を含む語,②動植物名の語,③擬音語・擬態語,④そのほか（①～③以外）の 4 タイプあると報告されている。新聞，雑誌と書籍との比較を目的に，これら 4 タイプの語が揃うよう，新聞や雑誌の先行調査の対象語より，本調査の対象語として 55 語を選定した。

今回の調査対象語のうち，カタカナ表記の使用頻度の多い，上位 12 語（パチンコ，モテる，ネタ，コツ，メリハリ，サヨウナラ，ゴミ，リンゴ，エビ，バラバラ，ニキビ，カッコ）については，明らかにカタカナ表記が優勢であることが確認できた。

しかしながら，残り 43 語は，必ずしも漢字表記やひらがな表記よりもカタカナ表記が優勢ではなかった。ひらがなの表記比率が 50%を超えるものもあれば（がっかり，びっくり，はまる，おしゃれなど），漢字の表記比率が 50%を超えるもの（謎，金，闇，米，虎，無駄など）も多くあった。

この実態調査によって，雑誌や新聞とは異なる書籍の使用傾向が明らかにできた。具体的には次のとおりである。

① 「表外漢字」「表外音訓」「表外熟字訓」などを含む語：漢字表記が避けられ，カタカナ表記される傾向が確かに見られる。しかし，その一方，本調査では，ひらがな表記の傾向が強いもの（例：「クセ」「キレイ」）や，漢字表記を避けない傾向があるもの(例:「エサ」「カッコウ」)，それでも漢字表記の傾向が強いもの（例：「ヤミ」「ナゾ」）も見られる。

② 動植物名の語：新聞や若者雑誌同様に，カタカナ表記の傾向がある。「表外漢字」であればその傾向はさらに強い。しかし，「リンゴ（林檎）」の漢字表記率の低さに対し，「ブドウ（葡萄）」の漢字表記率はさほど低くない，など，個別の傾向は若干異なる。

③ 擬音語・擬態語：若者雑誌ほど顕著ではないが，擬音語はもちろん，擬態語もカタカナ表記の傾向はある。しかし，「バラバラ」はカタカナ比率が高いが，「ガッカリ」「ビックリ」はひらがな表記率の方が高いなど，これも個別の傾向は異なる。

④ 第一義でない用法・特別な意味を加味する用法の語：本調査では「コツ」がその典型であることを示し，カタカナ比率が高いが,「カネ」は漢字表記率が高く,「ズレ」はひらがな表記率が高い。

⑤ 強調用法のある語：単なる強調によりカタカナ表記される傾向があると言われていた語は「ナゾ」「ニセ」「ハズレ」であるが，いずれも本調査ではカタカナ表記率が目立って高くはなく，カタカナ表記される傾向はややある，という程度となった。

⑥ 音を明示する用法のある語：音を明示したい場合にカタカナ表記される傾向のある語に該当する例が，本調査中の「ダメ」や「イヤ」などであると考えられる。これらはひらがな表記や漢字表記も少なくはないが，カタカナ表記されているという傾向は確かに読み取ることができる。

(5)コーパスから得られる辞書情報の分析―古い語に関して―

現代語の辞書で「古い」語をどう扱うべきかを明らかにするために，岩波国語辞典で「古語的」「古風」と注記される語を対象に，その使用実態を『現代日本語書き言葉均衡コーパス』に収録されている約 3,000 万語分の「書籍」テキストを用いて調査した。

その結果，現代語としての古い語を記述するためには，得られた用例が「現代語」の用例であるのか，そうではなく「非現代語」の用例であるのかを区別する必要があることを明らかにした。

コーパスより得られる用例は，大きく次の

4つに分類することができた。

① 古典の引用で使用されるもの 【古典の非現代語】
② 時代小説や歴史小説で使用されるもの 【享受と創造の非現代語】
③ 現代語であるが古い文体の中で使用されるもの 【古い文体の現代語】
④ 古風であるが現代語として使用されているもの 【現代語】

たとえば,「ものども【者共】」には,各分類に該当する次のような用例がある(太字,下線は本稿著者による)。

・『丹州三家物語』のいう「降参の者共は念比に扶助して皆家臣となる」状態であったろう。(吉村豊雄,1940年代生まれ,オメガ社|編『地方別日本の名族』新人物往来社,1989年) →①「古典の引用」(※下線が古典の引用部分。)

・「御前が鎌倉を不在にすれば、必ずや関東、東北、北陸の叛意を含む者どもが鎌倉をうかがうでしょう」(森村誠一,1930年代生まれ,『太平記 上』角川書店,2002年) →②「時代小説や歴史小説」(※下線部分,つまり,テキスト全体が時代小説の本文。この例は鎌倉時代末期から南北朝時代の時代設定。)

・王は旅行で疲れてゐたから、早く床に就き、寝殿には二人の侍従が(当時の習慣に依つて)すぐ側に眠りました。彼はマクベスの歓待を非常に悦び、寝所に退くまへに金を役役の者どもに贈りました。(野上弥生子訳,1880年代生まれ,『野上弥生子全集』第2期第20巻,岩波書店,1987年) →③「古い文体」(※下線箇所はいずれも旧仮名遣い。テキスト全体が旧仮名遣いである。)

・雲になってしまったあわれな者どもは、ゆく先がわからなくて、さまよいつづけた手紙たちです。あて名がちがっていたり、町名を書かなかったり、届きようのない手紙たちの、なれの果てですよ。(福永令三,1920年代生まれ,『クレヨン王国新十二か月の旅』講談社,1988年) →④「現代語」

古典の引用使用であるようなものは,現代語の辞書記述の対象とする必要はないものである。しかし,それ以外の用法は,いずれも現代語の辞書記述の対象として十分に考慮すべきものであろう。

本調査により,「古い」とされる語を現代語に位置づけて記述する重要性と,たとえば「古風」とただひとくくりにはせず,使用傾向に即した用例,用法の詳細記述を行うことの有用性を明らかにした。

(5)岩波国語辞典改版のための用例選定

一連の分析,成果を踏まえ,岩波国語辞典第七版改訂の際の用例選定を行った。

５．主な発表論文等
(研究代表者、研究分担者及び連携研究者には下線)

〔学会発表〕(計3件)

①柏野和佳子,奥村学,国語辞典に「古風」と注記される語の使用実態調査 ―『現代日本語書き言葉均衡コーパス』を用いて―,言語処理学会 第17回年次大会,2011年3月9日,豊橋技術科学大学

②柏野和佳子,奥村学,国語辞典に「古い」と注記される語の現代書き言葉における使用傾向の調査,情報処理学会 人文科学とコンピュータ研究会,2010年10月30日,国立国語研究所

③柏野和佳子,奥村学,和語や漢語のカタカナ表記―『現代日本語書き言葉均衡コーパス』における使用実態―,計量国語学会第53回大会,2009年9月12日,大正大学

〔図書〕(計1件)

西尾実・岩淵悦太郎・水谷静夫編,岩波書店,岩波国語辞典 第七版(分担執筆),2009,1722

６．研究組織

(1)研究代表者

柏野 和佳子(KASHINO WAKAKO)大学共同利用機関法人 人間文化研究機構 国立国語研究所・言語資源研究系・准教授
研究者番号:50311147

(2)研究分担者
(0)

研究者番号:

(3)連携研究者
(0)

研究者番号: