語種辞書『かたりぐさ』の開発と月刊雑誌の語種構成分析

茂木俊伸，山口昌也，丸山岳彦，田中牧郎
（国立国語研究所）

はじめに

国立国語研究所では，大量データを用いた現代日本語の語彙調査・研究に資するため，語種辞書『かたりぐさ』を開発した[1]。

本発表では，この『かたりぐさ』の概要の紹介とともに，その利用例として，2003 年発行の月刊雑誌の語種構成の分析の報告を行う。

1. 背景

現在，大量のテキストデータを用いた日本語研究が可能になりつつあり，語彙論に関してもこれまでにない大規模な調査による新たな成果が期待される。その反面，人手による形態素解析等の処理は，量的な限界が近づいている。

語種は，語彙の構造を捉えるための観点の中でも基本的なものの一つとされてきた。これまでの関連する研究では，国立国語研究所による一連の調査のような人手処理や，文字種による文字列の抽出といった方法が取られている。

語種辞書『かたりぐさ』は，奈良先端科学技術大学院大学 松本研究室によって開発，公開されている形態素解析システム『茶筌』付属の電子化辞書『IPADIC』（IPADIC-2.4.4）をベースとして，IPADIC の各辞書項目の「見出し語」（表記），「読み」，「品詞名」，「活用型」の情報に，新たに「語種」の情報を加えたものである。

例）	冬休み	フユヤスミ	名詞-一般		和
	冬至	トウジ	名詞-一般		漢
	凍る	コオル	動詞-自立	五段・ラ行	和
					語種

『かたりぐさ』の最大の利点は，『茶筌』の形態素解析結果を利用することで，テキストの量を問わず，基本的に語彙のレベルで，自動的な語種情報の付与を可能にする点にある。

その利用においては，データの扱いに関するいくつかの重要な問題（形態素解析の精度，形態素の単位認定や同語異語判別，未知語の問題等）への注意が必要になるが，分析の再現，検証が可能であることや，調査結果を速報性をもって提供できるという点は，語彙研究にとって大きな魅力であると言える。

2. 『かたりぐさ』の概要

2.1 語種とは

日本語の語彙は，その出自（どのようにしてその語が日本語の中で使われるようになったか）によって分類することができる。この分類を，「語種」という。

日本語の語種は，日本語に固有の「和語」，中国語からの借用語である「漢語」，それ以外の言語（主に西洋の諸言語）からの借用語である「外来語」に大別される。これらの複数から構成される語は，「混種語」と呼ばれる。

以上の，和語，漢語，外来語，混種語の 4 種が，一般的な語種の分類である。

2.2 『かたりぐさ』の語種情報

『かたりぐさ』は，IPADIC に登録されている語（形態素）のうち，固有名詞（142,155 語）と記号（150 語）を除いた，91,319 語から成っている。

『かたりぐさ』の語種情報は，IPADIC の各項目を人手で確認し，原則として『新潮現代国語辞典（第 2 版）』（新潮社，2000 年）に従って付与されている（詳細については，『かたりぐさ』添付のマニュアルを参照）。

[1] 2004 年 12 月から，国立国語研究所「言語データベースとソフトウェア」ウェブページにて無償で配布を開始している（http://www.kokken.go.jp/lrc/）。

語種情報は，2.1 節の 4 分類に基づき，「和」「漢」「外」「混」という記号で表されている。それぞれの語数は，「和」37,435 語，「漢」37,679 語，「外」7,176 語，「混」7,381 語である。

ただし，次のような項目には複数の語種が与えられている（かぎカッコが見出し語（表記），その中の丸カッコが読みを表す）。

- IPADIC で複数の読みが認められている語
  → 「月（{ツキ/ゲツ/ガツ}）」ならば「和/漢/漢」のように，スラッシュ区切りで対応する語種を示す。（1,394 語）

- 一つの読みで複数の語種が想定される語
  → 「カバ（カバ）」（「樺」と「河馬」が想定できる）ならば「和,漢」のように，カンマ区切りで語種を示す。（111 語）

また，次のような項目（143 語）には語種が与えられていない。

- IPADIC において，読みが記号になっている「、」「,」「.」「・」の 4 語
- 単語として特定できなかった語
- 語種が不明の語

なお，『かたりぐさ』から 1,000 語をランダムサンプリングし，人手でチェックしたところ，約 99.9%の精度で語種情報が付与されていることが分かった。

3. 月刊雑誌の分析

多様なジャンル，形式の文章を含む雑誌は，語彙の使われ方も多様であり，語種という観点から検討する資料として有効であると考えられる。ここでは，現代の書きことばの実態を捉えるために，2003 年発行の月刊誌（8 月号～12 月号）50 誌を資料として用いた。

3.1 調査対象

調査対象とする 50 誌の選定においては，以下の基準を原則とした。

a) 月刊誌であること。
b) 発行部数が公称 24 万部もしくは ABC 発行部数 16 万部以上であること。
c) 使用言語が日本語であること。
d) 書店で販売されていること。

さらに，これらに該当するものを，『雑誌新聞総かたろぐ 2003 年版』（メディア・リサーチ・センター），『月刊メディア・データ 一般雑誌レート&データ版』2003 年 6 月特大号（同），『雑誌のもくろく 2003 年版』（雑誌目録刊行会）の分類・分野をなるべく広く取る形で，発行部数等を考慮に入れながら絞り込んだ。選定された雑誌は，以下のとおりである（ジャンルは『雑誌のもくろく 2003 年版』による）。

「児童・学生」（3 誌）
Animage，My Birthday，蛍雪時代

「女性」（4 誌）
mini, MORE, 家庭画報, ポップティーン

「家庭」（6 誌）
dancyu，ESSE，QUANTO，新しい住まいの設計，壮快，ひよこクラブ

「大衆」（2 誌）
L magazine，Myojo

「総合・文芸」（7 誌）
一個人，財界人，小説宝石，短歌研究，俳句，文學界，文藝春秋

「趣味」（22 誌）
BACKSTAGE PASS, BE-PAL, GOLF DIGEST, HOBBY JAPAN，Lure magazine，Option，SCREEN, Swing Journal, Tennis Classic Break, カメラマン, 月刊碁ワールド, サッカーズ, 月刊ザテレビジョン（首都圏版）, 月刊ジャイアンツ, 趣味の園芸, 鉄道ファン, 月刊バスケットボール, パチスロ必勝ガイド, ヤングマシン, 優駿, ラジオライフ, 旅行読売

「専門」（6 誌）
MONEY japan，Newton，経済セミナー，日経 PC21，日経 TRENDY，法学教室

このうち「趣味」に関しては，含まれる分野が多岐にわたるため，選定された雑誌の数が多くなっている。

3.2 分析データ

各雑誌の本文（広告や付録等を除く）部分から，文末の句点もしくはそれに相当する記号（「。」「.」「？」等）を含む文をランダムに 200

文ずつ，50 誌で計 10,000 文，抽出した[2]。

さらに，Windows 版『茶筌』（WinCha 2000 R2）で形態素解析を行い，『かたりぐさ』を用いて解析結果に語種情報を付与した。形態素解析の精度は，全 224,597 形態素から 1,000 形態素をランダムサンプリングし，人手でチェックしたところ、約 96.0%であった。

3.3 月刊雑誌の語種構成

今回の調査の基礎的なデータは，次のとおりである（調査単位は，基本的に IPADIC に依存する）。

延べ語数：	224,597 語	
うち，	固有名詞	5,187 語
	記号	34,287 語
	未知語[3]	3,802 語
	助詞・助動詞	71,630 語
異なり語数：	23,008 語	

ここでは，固有名詞，記号，未知語，助詞・助動詞，数詞，および，複数の語種が付与された語，語種が付与されなかった語（2.2 節参照）を除いた範囲（延べ 102,252 語，異なり 16,323 語）を，語種の集計対象とした。

【表 1】月刊雑誌の語種構成（延べ語）

語種	語数	%
和語	52,971	51.8
漢語	38,321	37.5
外来語	9,011	8.8
混種語	1,949	1.9

【表 2】月刊雑誌の語種構成（異なり語）

語種	語数	%
和語	6,252	38.3
漢語	7,334	44.9
外来語	2,091	12.8
混種語	646	4.0

[2] 文頭から最初の「文末」までを一文としたため，例えば，“ 太郎「うん。 ”のように，引用記号やカッコ等が変則的に含まれている場合がある。

[3] 形態素解析の際，IPADIC に登録されていない文字列が形態素として切り出された場合，「未知語」という品詞が付けられる。

ただし，実態をより正確に捉えるために人手による補正を行った場合，これらの値は変動する可能性が高い。例えば，頻度 11 以上の語（異なり 2,102 語）における未知語の割合は 0.4%であるが，頻度 1 の語（11,840 語）では 17.3%を占め，特に異なり語の集計では未知語の扱いが与える影響が大きいと考えられる。

3.4 従来の調査との比較

山口ほか（2004）は，今回の雑誌の調査と同じ条件で，新聞（毎日新聞 CD-ROM9 年分）の調査を行っている。【表 3】にその結果を示す。

【表 3】新聞の語種構成（毎日新聞 9 年分）

語種	延べ語%	異なり語%
和語	39.37	39.81
漢語	54.09	44.46
外来語	5.03	8.55
混種語	1.51	7.19

【表 1～3】を単純に比較すると，延べ語では，雑誌は新聞に比べ和語・外来語の比率が高く，漢語の比率が低い。異なり語では，雑誌の外来語の比率がやや高くなっている。

なお，最近の雑誌の語種構成を扱った調査として，1994 年発行の月刊雑誌 70 誌を対象とした山崎・小沼（2004）がある。今回の調査とは，データの規模，調査単位，処理方法や語種の集計範囲等，条件が大きく異なるため，結果の比較そのものが困難であるが，『かたりぐさ』をより使いやすいものにするためには，どのような要因がどのように調査結果に作用するのかを今後詳しく検討していく必要がある。

3.5 ジャンルによる差

3.5.1 全体的な傾向

これまでの研究でも，雑誌のジャンルによって，語種構成に差があることが指摘されてきた。

個々の雑誌の題材や内容によって異なりがあるが，大まかな傾向を見るために，3.1 節の各ジャンルごとにまとめて語種構成（延べ語）を集計したものが【表 4】である。

【表 4】ジャンル別の語種構成（延べ語%）

	和	漢	外	混
児童･学生	54.6	36.9	6.3	2.2
女性	57.4	26.7	14.0	2.0
家庭	55.9	32.8	9.1	2.2
大衆	55.3	31.4	10.8	2.5
総合･文芸	58.7	36.8	2.6	1.9
趣味	51.9	35.0	11.3	1.9
専門	39.1	53.1	6.2	1.6
全体	51.8	37.5	8.8	1.9

語種別に見ると，和語は「専門」が低いほかは，それほど顕著な差は見られない。漢語は，同じく「専門」が突出して高く「女性」が低い。外来語は，最も低い「総合・文芸」から最も高い「女性」まで，ばらつきが見られる。

先の【表 3】の新聞の語種構成は，ここでのジャンルでは「専門」に近いと言える。

3.5.2 雑誌ごとの特徴

次に，各語種の比率（延べ語）の上位・下位 10 位までに来る雑誌を，具体的に見ていく。

[1] 和語

上位では「総合・文芸」の雑誌が目立ち，下位では「専門」の 6 誌が 10 位以内にすべて含まれる（カッコ内に比率を示す)。

上位　小説宝石(66.2)，My Birthday(65.3)，文學界(64.7)，月刊碁ワールド(64.2)，BACKSTAGE PASS(63.6)，ひよこクラブ(63.4)，俳句(63.3)，Myojo(63.2)，短歌研究(63.2)，趣味の園芸(63.0)

下位　法学教室(33.9)，経済セミナー(37.3)，日経PC21(38.9)，HOBBY JAPAN(39.6)，鉄道ファン(40.7)，QUANTO(41.3)，日経 TRENDY(42.4)，MONEY japan(42.8)，Newton(44.5)，蛍雪時代(44.7)

[2] 漢語

上位・下位ともに 1 位が突出している。上位では，和語の下位と同様，「専門」の 6 誌すべてが 10 位以内に入る。下位では，「女性」「趣味」の雑誌が目立つ。

上位　法学教室(63.1)，経済セミナー(56.5)，鉄道ファン(52.4)，Newton(49.9)，財界人(49.8)，蛍雪時代(49.4)，MONEY japan(47.7)，日経 PC21(45.5)，日経 TRENDY(43.7)，旅行読売(43.1)

下位　mini(17.4)，BACKSTAGE PASS(22.7)，ポップティーン(23.8)，GOLF DIGEST(24.8)，My Birthday(24.8)，Myojo(25.4)，MORE(26.6)，Lure magazine(28.0)，Tennis Classic Break(28.6)，ESSE(28.7)

[3] 外来語

個々の雑誌を見た場合，ジャンル間の比較よりも上位と下位の差が大きくなる。上位では，「女性」の 1 位以下で「趣味」の雑誌が目立つ。下位では，10 位以内に「総合・文芸」の 7 誌すべてが含まれている。

上位　mini(29.9)，ヤングマシン(19.0)，QUANTO(18.2)，Swing Journal(18.1)，Option(17.8)，Lure magazine(17.6)，HOBBY JAPAN(17.1)，Tennis Classic Break(16.7)，GOLF DIGEST(14.9)，日経 PC21(14.0)

下位　俳句(1.5)，法学教室(1.6)，小説宝石(1.9)，短歌研究(2.0)，月刊碁ワールド(2.3)，文學界(2.4)，一個人(2.5)，文藝春秋(3.5)，財界人(3.9)，Newton(4.2)

おわりに

本発表では，語種辞書『かたりぐさ』の概要と月刊雑誌 50 誌の語種構成について報告した。

『かたりぐさ』は，比較的手軽に大量のテキストを処理できるという点で，大きな利点を持つ。反面，同時にさまざまな制約を抱えてもおり，日本語の姿を適切に捉えるためにはさらなる改善が必要である。今後も実際の資料で調査を行いながら，検討を行っていきたい。

参考文献

松本裕治・浅原正幸（2001）「IPADIC ユーザーズマニュアル（version 2.4.4)」奈良先端科学技術大学院大学 情報科学研究科 松本研究室.

山口昌也・茂木俊伸・桐生りか・田中牧郎（2004）「語種との関係に基づいた新聞記事における語彙の時間的変化分析」『社会言語科学会第 13 回大会発表論文集』pp.113-116.

山崎誠・小沼悦（2004）「現代雑誌における語種構成」『言語処理学会第 10 回年次大会発表論文集』pp.670-673.