第８章　　語種辞書『かたりぐさ』の開発

茂木 俊伸

１　はじめに

語種辞書『かたりぐさ』は，現代日本語の語種の調査・研究に資するために開発された基礎データである。

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（西尾(2002)）。

日本語の語種は，一般的に，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。また，これらの複数から構成される語は「混種語」と呼ばれる。この語種は，日本語の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。

語種に関するこれまでの量的研究としては，国立国語研究所による一連の成果（最近の調査としては，国立国語研究所(2005)を参照）などがあるが，資料に現れた文を語に分解し，語種を付けるという一連の作業を人手によって行う形の調査が主流であった。

一方，現在，大量のテキストデータ（電子化テキスト）を用いた日本語研究が可能になりつつある。語彙論の分野においても，これまでにない大規模なデータを対象とした調査による新たな成果が期待されているものの，その反面，従来のような人手によるデータの処理という方法には量的な限界が近づいている。

以上のような背景の下，国立国語研究所研究開発部門では，大量データを用いた語種の調査・研究のためのツールとして，語種辞書『かたりぐさ』を開発・公開した。本稿では，この『かたりぐさ』の内容に関する報告を行う。

２　『かたりぐさ』の内容

2.1　概要

語種辞書『かたりぐさ』[1]は，電子化辞書『IPADIC』（IPADIC-2.4.4，2001年3月）をベースとして作成された。

IPADIC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（ChaSen)』に付属しているデータである[2]。

『かたりぐさ』は，IPADICに登録されている項目（語や形態素）のうち，固有名詞（142,155語）と記号（150語）および一部の項目の重複を除いた，計91,319語から成っている。

これらの項目には，人手で確認した上で，和語，漢語，外来語，混種語の４分類に対応する記号（それぞれ「和」「漢」「外」「混」）の形で，語種情報が付与されている。語種別の項目数は，次の表に示すとおりである（この表は，『かたりぐさ』バージョン1.0.1（2005年9月）に基づく。表中の「その他」については，2.2節で後述する)。

表　『かたりぐさ』の語種別項目数

語種	項目数
和語	37,432
漢語	37,680
外来語	7,180
混種語	7,381
その他	1,646
計	91,319

なお，『かたりぐさ』（バージョン1.0，2004年12月）から1,000項目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ている。

2.2　構造

IPADICは，各項目の「見出し語」（表記），「読み」，「発音」，「品詞名」といった情報から構成されている。『かたりぐさ』では，次のように，これらのうちの「見出し語」，「読み」，「品詞名」，「活用型」を利用し，新たに「語種」の情報を加えた。この語種情報は，先に触れたように，「和」「漢」「外」「混」という記号として付与されている。

(1)

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
（見出し語）	**（読み）**	**（品詞）**	**（活用型）**	**（語種）**

ただし，「語種」欄は，上の表で「その他」とした項目に関して，複数の語種を付与した場合と，語種を付与していない場合がある。

まず，次のような項目には，複数の語種が与えられている。

(2)　a. IPADICで複数の読みが認められている語（1,394語）

例：「月」には「{ツキ/ゲツ/ガツ}」のように複数の読みが認められている。この場合，「和/漢/漢」のように，スラッシュ区切りでそれぞれの読みに対応する語種を示す。

b. 一つの読みで複数の語種が想定される語（112語）

例：カタカナ表記の「カバ」は，語として「樺」（和語）と「河馬」（漢語）が想定できる。この場合，「和，漢」のように，カンマ区切りで語種を示す。

また，次のような項目（140語）には語種が与えられていない。

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

語種情報の付与作業の際に定めた原則（使用した資料と付与の方法）については次の2.3節で具体的に示すが，(3b)は，IPADICに項目として立てられているものの，資料で実在が確認できなかった語である（例：「スボールクラブ」（名詞），「うさんい」（形容詞））。また，(3c)は，国語辞典では語種が判明しなかった語である（例：「胡蝶花（シャガ）」）。

2.3　語種情報の付与規則

IPADICの各項目に語種を付与するための調査には，見出し語の表記によって語種が示されている『新潮現代国語辞典（第２版）』（新潮社，2000年，以下『新潮』とする）を用いた。『かたりぐさ』の語種情報は，原則としてこの『新潮』に従っている。

一方で，IPADICの辞書項目には，きわめて多様な語が含まれている。『新潮』に収録されていない語については，必要に応じて以下の補助的資料を用いた。

(4)　a.『現代語・古語 新潮国語辞典（第２版）』（新潮社，1995年）
b.『日本国語大辞典（第２版）』（小学館，2000-2001年）
c.『広辞苑（第５版）』（岩波書店，1998年）

また，漢字と読みの対応を確認するために，『今昔文字鏡（単漢字10万字版）』（エーアイ・ネット／紀伊國屋書店，2003年）を使用した。

まず，当該の語が(4)の資料で確認できた場合は，読みに従って（必要があればより小さな単位に分解し

た上で)，語種を判断した。次に挙げるのは，『日本国語大辞典（第２版)』で確認できた例である。

(5)　a.「秋高（アキダカ)」：
　　米相場関連の用語　…　「秋」(和語）＋「高」(和語）→「和」
　　b.「外販（ガイハン)」：
　　「外交販売」の略　…　「外」(漢語）＋「販」(漢語）→「漢」

さらに，上記の辞典類にも記載のない場合（特に専門用語や新語等）は，新聞記事データベース（毎日，読売，日経の３紙）および全文検索型サーチエンジン（Google）を用いて実例の確認を行い，辞典類を利用した場合と同じ手順で語種を付与した。

(6)　a.「歌伴（ウタバン)」：
　　歌の伴奏，バックバンド …　「歌」(和語）＋「伴」(漢語）→「混」
　　b.「益税（エキゼイ)」：
　　税金関連の用語 …　「益」(漢語）＋「税」(漢語）→「漢」
　　c.「オーヴァードーズ」：
　　薬物の過剰摂取（原語は英語“overdose”）　→「外」

なお，語源が諸説ある語や語源未詳の語の場合，辞典によって（あるいは同一の辞典でも版によって）掲げられた語種が異なることがある。このような場合は『新潮』に従った（例：「背広」には外来語説があるが，『新潮』の見出しの表記に従い，和語とした)。

2.3.1　複合語・連語の処理（付則１）

IPADICの辞書項目には，複合語や連語が含まれている場合がある。このように『新潮』の見出し語よりも長い語（句）は，見出し語レベルに分解した上で語種を判断した。

(7)　a.「推して知るべし」：
　　…「推す」(和語）＋「て」(和語）＋「知る」(和語）＋「べし」(和語）→「和」
　　b.「こんなふうに」：
　　…「こんな」(和語）＋「風」(漢語）＋「に」(和語）→「混」

2.3.2　略語の処理（付則２）

略語は，表記にかかわらず，元の語に基づいて語種を判断した。

(8)　a.「断トツ」：
　　…「断然」(漢語）＋「トップ」(外来語）→「混」
　　b.「駐禁」：
　　…「駐車」(漢語）＋「禁止」(漢語）→「漢」

また，略号は，読みに従って判断した（例：「(株)（カブシキガイシャ)」→「漢」)。

2.3.3　表記ゆれの処理（付則３）

IPADICにおいて同一の語がさまざまな表記形で登録されている場合，語種の判断に支障がないかぎりにおいて，表記や読みのゆれ（まぜ書き，送りがなのゆれ，長音表記の有無，音の清濁のゆれ等）は許容した。

(9)　a.「でんぷん／でん粉／澱粉」：
　　…『新潮』では「デンプン【澱粉】」→ すべて「漢」
　　b.「堂々めぐり／堂々回り／堂々巡り／堂堂回り／堂堂巡り」：
　　… 同「ドウドウめぐり【堂堂巡(り)・堂堂回(り)】」→ すべて「混」

また，当て字の類は，読みに従って判断した（例：「倶楽部（クラブ)」「遊ゴロ（ショートゴロ)」→「外」)。

2.3.4　語種分類に関する補足

『かたりぐさ』の語種の４分類は，第１節で概略を述べたとおりであるが，漢語・外来語・混種語の認定に当たっては，以下のような基準を設けた。

一般的に漢語とされるものには，各時代に中国語から入った語のほか，和製漢語（例：「火事」「大根」）がある。『かたりぐさ』では，これら以外に，他の言語から中国語の音訳を経て入った語（例：「刹那（セツナ）」「盂蘭盆（ウラボン）」），および朝鮮漢字音で読む語（例：「諺文（オンモン）」「温突（オンドル）」）についても，漢語とした（『新潮』には，それぞれ「～の音訳」「～の朝鮮字音」という注記がある）。

また，外来語は，一般的な分類に沿って，欧米系の諸言語から取り入れられた語のほか，和製英語（例：「マイカー」「ナイター」），アイヌ語（例：「オットセイ」「シシャモ」），現代中国語（例：「ラーメン」「高粱（コーリャン）」）等も含めている。

混種語は，和語・漢語・外来語の異種の組み合わせ，すなわち，「外来語＋和語」（例：「ガラス張り」「生ビール」），「外来語＋漢語」（例：「テレビ局」「豚カツ」），「和語＋漢語」（例：「切符」「牛鍋」，いわゆる重箱読み・湯桶読みの語）という構成を持つ語である。原語の異なる外来語が結合しているもの（例：「テーマパーク」（ドイツ語“Thema”＋英語“park”））は，混種語ではなく外来語とした。また，IPADICには，「愛する」のようないわゆるサ変活用動詞，「徐々に」のような語尾の付いた副詞が登録されているが，これらも混種語に含めている。

2.4 問題点

以上，『かたりぐさ』の概要について述べてきたが，ここで，利用の際に生じうる問題にも触れておきたい。

次に例として挙げる問題（本報告書第３部第４章「新聞記事における語彙の時間的変化分析―語種との関係を中心に―」（山口昌也）も参照）は，『茶筌』やIPADICに依存する『かたりぐさ』の仕様上，避けられないものであり，それゆえに特に留意が必要な点である。

(1) 形態素解析精度

『茶筌』による形態素解析は，ケアレスミスがない一貫した処理を行ってくれる一方で，人の目から見て明らかに「誤り」であるという結果を出力することがある。第３節で後述するように，『かたりぐさ』は，『茶筌』の形態素解析の結果に語種情報を加える形で利用するため，形態素解析の精度が語種の分析に影響を与えることになる。

例えば，実際の語種の分析において，形態素解析の結果をランダムサンプリングし，人手で確認することによって見積もられた精度は，新聞データでは約98％（山口ほか(2004)），雑誌データでは約96％（茂木ほか(2005)）であった。すなわち，『かたりぐさ』の適用以前に，ある程度の処理上の「誤り」が生じ，データが大量になればなるほど，それを確認することも困難になるわけである。

この点をどのように評価するかは，分析者の立場や分析の目的によって分かれると思われるが，いずれにせよ，この種の誤りが生じることを前提として分析に利用していく必要がある。

(2) 調査単位

語彙調査の際に，「語」をどのように認定するのかは大きな問題となる。国立国語研究所の調査においても，形態素レベルや文節レベルなど，さまざまな調査単位が用いられてきた（中野(1998)）。

どの調査単位を用いるにせよ，その区切り方は一定の基準に基づいて統一されている必要があるが，IPADICの辞書項目には，句から形態素まで，さまざまな単位が混在している。したがって，『かたりぐさ』によって得られた語種構成（各語種の比率）は，必ずしも均質な単位に基づくものではない。

(3) 未知語

IPADICの辞書項目として登録されていない語がテキストに出現する場合，『茶筌』の形態素解析の段階で，文字列として切り出されて品詞名「未知語」が与えられるか，もしくは複数の語に（過）分割される。『かたりぐさ』は，未知語に語種を付与できず，また誤った解析結果には誤った語種を与える可能性があるため，IPADICに未登録の語の処理は，語種の調査にも影響を及ぼす。

例えば，『かたりぐさ』のベースとなっているIPADIC-2.4.4は2001年に公開されたものであり，それ以降の新語は特に問題となりうる。茂木ほか(2005)の雑誌データの分析からは，頻度11以上の語（2,102語）における未知語の割合は0.4%であるのに対し，頻度１の語（11,840語）では17.3%を占め，特に異なり語数の集計において，未知語の扱いが与える影響が大きいことが分かっている。

以上，３つの大きな問題を挙げたが，このような事情から，『かたりぐさ』を利用して集計された語種構成と，例えば国立国語研究所の一連の語彙調査で明らかにされている語種構成とを単純に比較することは，たとえ同種の資料の調査結果であっても，困難である。

『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詳しく検討していく必要があるが，この点は今後の課題となっている。

３　『かたりぐさ』の利用方法

3.1　データの形式

『かたりぐさ』は，それ自体はタブ区切りテキスト形式（文字コードはShift_JIS）のファイルであり，単体では動作しない。

このため，あるテキストの語種構成を調べる際には，次の図のような，①形態素解析，②語種情報の付与，という２段階の手順を踏むことになる。

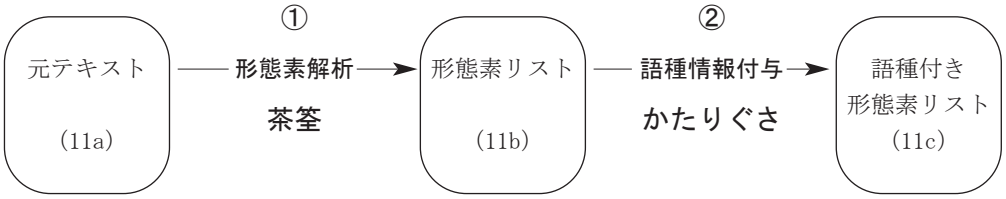

図　『かたりぐさ』を用いた語種調査の手順

このうち①の過程では，対象となるテキストを『茶筌』を使って形態素（語）に分割する。次に，②の過程では，『茶筌』から出力された解析結果（形態素のリスト）と『かたりぐさ』の情報とを照合し，解析結果に語種情報を付与する。このような手順により，語種情報が付与された形態素のリストが得られることになる。

次に，Windows環境においてこれらの手順を実行した例を，具体的に見ていく。

3.2　Windows環境での利用方法

まず準備段階として，Windows上で『かたりぐさ』を利用するためのファイルを揃える必要がある[3]。『かたりぐさ』の利用には，語種構成の調査対象となるテキストファイルのほかに，以下のものが必要である。

(10)　a. 形態素解析システム『茶筌』Windows版（WinCha）
　　b. プログラミング言語Perl（バージョン5.8.6 以上）
　　c. 語種辞書『かたりぐさ』
　　d. 『かたりぐさ』適用のためのパッケージ

(10a)は，『茶筌』の公式ページ（注2参照）で配布されている，IPADIC-2.4.4を同梱したバージョン（cha21244sp5.exe）が必要である。(10d)は，『かたりぐさ』用の処理ファイルをまとめたものであり，『かたりぐさ』の配布ページで入手できる。

以上の準備が整ったら，先の図の①の段階（『茶筌』による形態素解析）に入る。

例えば，(11a)のような文が含まれるテキストファイルを『茶筌』で形態素解析すると，(11b)のような解析結果のファイルが出力される（左から「表層語（出現形）」，「基本形」，「読み」，「品詞」，「活用型」，「活用形」）。

(11)　a. **元テキスト**：

この例文でテストしてみます

b. **形態素解析結果**：

この	この	コノ	連体詞		
例文	例文	レイブン	名詞-一般		
で	で	デ	助詞-格助詞-一般		
テスト	テスト	テスト	名詞-サ変接続		
し	する	シ	動詞-自立	サ変・スル	連用形
て	て	テ	助詞-接続助詞		
み	みる	ミ	動詞-非自立	一段	連用形
ます	ます	マス	助動詞	特殊・マス	基本形

次に，3.1節の図の②の段階（語種情報の付与）として，この解析結果に『かたりぐさ』を適用する。

『かたりぐさ』適用パッケージに含まれる実行用のファイルをダブルクリックすると，『かたりぐさ』の情報（(1)）と形態素解析結果のテキストファイル（(11b)）との照合が行われ，次の(11c)のような語種情報の付いた形態素のリストが得られる。(11c)の右２列（点線で囲んだ部分）が，語種と，その判定の基準となった読みの情報である。

(11)　c. **語種付与結果**：

この	この	コノ	連体詞			和　コノ
例文	例文	レイブン	名詞-一般			漢　レイブン
で	で	デ	助詞-格助詞-一般			和　デ
テスト	テスト	テスト	名詞-サ変接続			外　テスト
し	する	シ	動詞-自立	サ変・スル	連用形	和　スル
て	て	テ	助詞-接続助詞			和　テ
み	みる	ミ	動詞-非自立	一段	連用形	和　ミル
ます	ます	マス	助動詞	特殊・マス	基本形	和　マス

調査対象のテキストの語種構成は，最終的に得られたテキストファイル（(11c)）を表計算ソフトなどで読み込んで集計することができる。

『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によって，テキストの量を問わず，自動的な語種情報の付与を可能にすることにある。先に2.4節で述べた問題はあるものの，従来は人手による作業に長い期間がかかっていた量のテキストであっても，きわめて短時間で語種に関する情報が得られるのである。

４　『かたりぐさ』の公開

『かたりぐさ』は，2004年12月に，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(http://www.kokken.go.jp/lrc/)において，バージョン1.0の無償配布を開始した。2005年9月には，その後確認された10語の語種付与の誤りを修正した最新版（バージョン1.0.1）を公開している。

また，『かたりぐさ』は単体で動作しない（第３節参照）ことから，利用方法に関する情報を充実させる必要があると考え，2005年3月に「Windows環境での利用方法」というサポートページを作成した。

『かたりぐさ』の配布ページにおいて任意で回答を依頼しているアンケート（2007年2月20日現在の回答数190）によると，『かたりぐさ』をダウンロードした利用者の専門分野は，国語（日本語）学，言語学，言語教育といったいわゆる文系分野が多く（68.4%），自然言語処理などの理系分野（22.1%）はむしろ少数派である（この他に，文系・理系両方の分野2.6%，無回答6.8%）。また，利用者の使用OSはWindowsのみという回答が圧倒的に多い（86.8%）。これらのことから，日本語研究者への情報提供を念頭に置いた上記のサポートページには，一定の意義があったと考えられる。

なお，『かたりぐさ』は，新聞（山口ほか(2004)）や雑誌（茂木ほか(2005)）の語種構成の分析に実際に使用されているほか，国立国語研究所の「「外来語」言い換え提案」に関する基礎データ作成（国立国語研究所「外来語」委員会(2006)，田中(2006)）にも利用されている。また，一般公開後には，宋(2005)，橋本(2006)といった活用事例が現れている。

５　おわりに

語種辞書『かたりぐさ』は，比較的手軽に大量のテキストデータを処理し，その調査結果を速報性をもって提供できるという，これまでにない形の語彙研究を可能にするツールである。また，同じ調査対象のテキストデータが入手できれば，研究者間で分析の再現が容易にできるという点（検証可能性）も，利点として挙げることができる。

その一方で抱えるさまざまな制約の中で，『かたりぐさ』をどのように調査・研究，あるいは今後のツール開発に活用できるのか，今後も検討を行っていきたい。

注

1　この名称は，「語種」の訓読みに由来している。

2　『茶筌』およびIPADICは，公式ページ（http://chasen-legacy.sourceforge.jp/）で入手できる。IPADICの内容および形式に関しては，添付の「ユーザーズマニュアル」を参照されたい。『茶筌』の概要に関しては，松本（2000）を参照。

3　以下の手順やファイルの入手方法の詳細は，『かたりぐさ』配布ページ（第４節参照）の「Windows環境での利用方法」で公開している。

参考文献

国立国語研究所（2005）『現代雑誌の語彙調査－1994年発行70誌－』国立国語研究所

国立国語研究所「外来語」委員会(編)（2006）『分かりやすく伝える　外来語言い換え手引き』ぎょうせい

宋　浩（2005）「日本語における基本語彙選定に関する研究－動詞格フレームを中心的に利用して－」慶尚大学校大学院日本学科博士学位論文

田中牧郎（2006）「現代社会における外来語の実態」国立国語研究所(編)『新「ことば」シリーズ19　外来語と現代社会』38-46，国立印刷局

中野　洋（1998）「第４章　言語の統計」長尾真ほか(編)『岩波講座言語の科学９　言語情報処理』149-199，岩波書店

西尾寅弥（2002）「第４章　語種」斎藤倫明(編)『朝倉日本語講座４　語彙・意味』79-109，朝倉書店

橋本和佳（2006）「『分類語彙表』の増補改訂と外来語の増加－「1.4　生産物および用具」について－」倉島節尚(編)『日本語辞書学の構築』105-117，おうふう

松本裕治（2000）「形態素解析システム「茶筌」」『情報処理』41(11)，1208-1214，情報処理学会

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．（ユーザーズマニュアルを含むIPADIC-2.4.4は，http://chasen.naist.jp/stable/ipadic/で入手可（2007/2/20現在））

茂木俊伸・山口昌也・丸山岳彦・田中牧郎（2005）「語種辞書『かたりぐさ』の開発と月刊雑誌の語種構成分析」『言語処理学会第11回年次大会発表論文集』341-344

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』113-116，山口昌也「新聞記事における語彙の時間的変化分析―語種の観点から―」として本書第３部第４章に収録

付記

語種辞書『かたりぐさ』の企画・開発は，本稿の筆者である茂木（2005年3月まで国立国語研究所に在籍）のほか，山口昌也，桐生りか，田中牧郎が担当した。本稿は，『かたりぐさ』添付の「利用マニュアル」に加筆したものである。

本データの基礎となる辞書の構築に携わられた方々，また『かたりぐさ』の活用事例をご報告くださった方々に，心より感謝申し上げる。