第９章　『表記統合辞書』の構築

山口 昌也

１　はじめに

『表記統合辞書』は，言語研究・自然言語処理用に開発された，同語判別のための基礎データである。

本辞書を作成した背景には，各種の自然言語処理や言語分析を行う際に問題となる，異表記の扱いが挙げられる。例えば，形態素解析システムを使って，形態素の生起確率を測定することを考える。このとき，「表す」「表わす」，「りんご」「リンゴ」などのように，送りがなや字種の違いにより表記が異なる形態素については，それぞれ別の形態素として認定される。したがって，語の生起確率を測定するためには，同語の判定を行い，出現頻度を統合する必要がある。

そこで，『表記統合辞書』では，奈良先端科学技術大学院大学（松本研究室）で開発された形態素解析システム『茶筌』[1]の解析結果に同語判別のための情報を付与することを前提に設計する。具体的には，『茶筌』の形態素辞書IPADIC(ver. 2. 4. 4)[2]の各辞書項目に対して，同語判別のための情報を付与する。これにより，表記上の違いにより，『茶筌』が異なった形態素として認定された形態素を統一的に扱うことができるようになる。

『表記統合辞書』における同語と判断するための基準は，可能な限り，表記上の知識のみに基づくものとし，意味的な知識が必要な場合は，同語と判断しないことにした。このような基準を設けたのは，形態素解析結果の利用方法はさまざまであり，汎用的な側面からは，意味的な基準を規定することが困難であると判断したからである。

本稿では，次節で『表記統合辞書』の内容について解説した後，３節で利用方法と利用例を示す。さらに，４節では，『表記統合辞書』の一般公開の概要と利用状況を報告し，５節でまとめを行う。

２　『表記統合辞書』の内容

2.1　概要

すでに述べたように，『表記統合辞書』の各辞書項目には，同語判別のための情報が付与されている。また，形態素解析システム『茶筌』用の形態素辞書IPADIC(ver. 2. 4. 4)と対応付けることができるように設計されており，『茶筌』の解析結果に対して同語判別をするための基礎情報として利用することができる。

『表記統合辞書』の辞書項目数は，33428である。品詞ごと辞書項目数を表１に示す。

表１　品詞別の辞書項目数

名詞	17067	接続詞	58
動詞	13300	接頭詞	18
形容詞	1377	連体詞	41
副詞	1460	助詞	43
感動詞	58	助動詞	6

2.2　辞書の構造

『表記統合辞書』は，次の構造の辞書項目からなる。実際の辞書項目の例を図１に示す。なお，辞書ファイルの物理的な形式については，3.1節で詳説する。

見出し部　＋　同語判別情報

見出し部は，IPADICの「見出し語」「読み」「品詞名」「活用型」の組からなる。これらは，本辞書の辞書項目を特定するだけでなく，IPADICの辞書項目を一意に同定するための情報としても利用される。例えば，図１（４）の例では，見出し部は「すばやい」「スバヤイ」「形容詞-自立」「形容詞・アウオ段」とな

る。非活用語は，図１（１）のように「活用型」の欄が空欄となる。

同語判別情報は，『表記統合辞書』中の他の辞書項目へのリンク情報として表現される。具体的には，同語と判断される辞書項目の見出し部の「見出し語」を列挙したものである。

図１（１）を見てみよう。この例の場合，同語判別情報は，「編みもの/編み物/編物」となる。これは，見出しが「編み物」と「編物」の辞書項目，つまり，辞書項目（２）（３）と同語であることを意味する。同語判別情報に，当該辞書項目の見出し部である「編み物」が含まれているのは，見出し語以外の情報がまったく同一の辞書項目を容易に判別するためである。辞書項目（１）（２）（３），および，（４）（５）はそれぞれ同語判別情報が同一であり，これらの辞書項目が，表記以外の情報がまったく同一であることがわかる（当該の辞書項目の見出し語が含まれないと，同語判別情報が同一にならない）。それに対して，辞書項目（６）（７）（８）は，同語判別情報がそれぞれ異なるので，完全には同語でないということがすぐに判定できる。

（１）	編みもの	アミモノ	名詞-一般		編みもの/編み物/編物
（２）	編み物	アミモノ	名詞-一般		編みもの/編み物/編物
（３）	編物	アミモノ	名詞-一般		編みもの/編み物/編物
（４）	すばやい	スバヤイ	形容詞-自立	形容詞・アウオ段	すばやい/素早い
（５）	素早い	スバヤイ	形容詞-自立	形容詞・アウオ段	すばやい/素早い
（６）	厩舎	キュウシャ	名詞-一般		きゅう舎/厩舎
（７）	鳩舎	キュウシャ	名詞-一般		きゅう舎/鳩舎
（８）	きゅう舎	キュウシャ	名詞-一般		きゅう舎/厩舎/鳩舎

図１　辞書項目の例

2.3　同語と判別するための規則

2.2節で示した同語判別情報を付与するための基準，つまり，IPADICの二つの辞書項目を同語であると判断するための規則を示すことにする。

まず，IPADICの二つの辞書項目を同語と判断するための前提条件として，（１）同一品詞であること，（２）読みが同一であることを挙げる。これは，IPADIC上の情報においては，次のことを意味する。ただし，固有名詞（IPADICでは，「品詞」が「名詞-固有名詞」）の辞書項目については，表記による同語／別語の判別が困難であるため，規則適用の対象外とした。

・IPADICの品詞，活用型が同一
・IPADICの読みが同一

以上の前提条件を満たした上で，次の規則に合致する辞書項目同士を同語とする。規則の作成にあたっては，可能な限り，表記上の知識のみに基づくものとし，意味的な知識が必要な場合は，同語と判断しないことにした。なお，以下の規則中に示した実例は，左からIPADICの「読み」「品詞」「活用型」（見出し部），「同語判別情報」である。

1)　送り仮名による違い

a）送り仮名の有無

ワリツケ	名詞-一般	割り付け/割付/割付け

b）促音，撥音の有無

スットンキョウ	名詞-形容動詞語幹	素っ頓狂/素頓狂
マンマル	名詞-形容動詞語幹	真ん丸/真丸

2)　字種による違い

a）ひらがな，カタカナ，漢字

イス	名詞-一般	いす/イス/椅子
イキイキ	副詞-助詞類接続	いきいき/生き生き
ヨビオコス	動詞-自立	よびおこす/呼び起こす

b）一般名詞，および，数詞における漢数字，アラビア数字，ローマ数字

ハチミリ	名詞-一般	８ミリ/八ミリ
レイ	名詞-数	○/〇/零

c）アルファベット表記とカタカナ表記

エヌジー	名詞-一般	ＮＧ/エヌジー
ヘルツ	名詞-接尾-助数詞	ｈｚ/ヘルツ

d）「ヵ」，「か」，「カ」，「ヶ」，「ケ」，「箇」，「個」

カソン	名詞-接尾-助数詞	か村/カ村/ケ村/ヵ村/ヶ村

e）アルファベットの大文字・小文字（２文字以上の形態素の場合）

ペーハー	名詞-一般	ＰＨ/ｐＨ

f）名詞-接尾-助数詞で同一の単位における字種

トン	名詞-接尾-助数詞	ｔ/トン/噸

3）記号類による違い

a）読点・中黒の違い，読点・中黒の有無

ショウチュウガクセイ	名詞-一般	小、中学生/小・中学生/小中学生

b）「々」，「ゝ」などの踊り字の種類，有無

アラアラシイ	形容詞-自立	荒々しい/荒荒しい
イヨイヨ	副詞-一般	愈/愈々
シバシバ	副詞-一般	屡/屡々

2.4　未解決の問題

2.3節で規定した規則は，できる限り，表記上の知識のみに基づいて規定したため，同語判別を行う上で，問題となる場合がある。この節では，既知の問題点について述べる。

問題点は，大きく分けて二つある。一つは，規則の不足により，同語と判別できない問題である。具体的には，カタカナ語の異表記，漢字の異体字によるものである。主な例を次に示す。

●カタカナ語の異表記
- 長音の有無（例：「コンピュータ」「コンピューター」）
- 発音表記（例：「バイオリン」，「ヴァイオリン」）

●漢字の異体字（例：暁/曉，虱/蝨）

もう一つの問題は，規則の記述能力の低さから，過度に同語と判断してしまう問題である。この問題は，特に，規則２）のａ）により，漢字表記とひらがな・カタカナ表記を同一に扱う場合に発生する。例えば，「度々」「たびたび」，「鮭」「サケ」「さけ」など正しく機能する例も多い。その一方，次の例のように，「印」「イン」「いん」の場合のように，（文脈によっては使われる場合があるかもしれないが）ほとんどの場合，同語とはならない例も発生してしまう。

３　『表記統合辞書』の利用方法

3.1　辞書ファイルの形式

『表記統合辞書』は，電子的なファイルで配布される。実際の利用方法を説明する前に，『表記統合辞書』ファイルの形式について説明しておく。

『表記統合辞書』ファイルは，タブ区切りのテキストファイルである。詳細な形式は，次のとおりである。

●文字コード：　Shift_JIS
●改行文字：　CR/LF
●各フィールドの内容
- 第１フィールド：IPADICの「見出し語」
- 第２フィールド：IPADICの「読み」

・第３フィールド：IPADICの「品詞名」
・第４フィールド：IPADICの「活用型」
・第５フィールド：同語判別情報

3.2　『茶筌』の解析結果への適用

適用方法

『茶筌』と『表記統合辞書』との連携は，『茶筌』の出力結果と『表記統合辞書』の「見出し部」とを対応付けることにより行う。ただし，『表記統合辞書』は，IPADIC（ver. 2. 4. 4）の各辞書項目に，同語判別のための情報を付与したものである。したがって，『茶筌』を利用する場合は，IPADIC（ver. 2. 4. 4）と組み合わせて使用する必要がある[3]。

次に，実際の例を用いて，『茶筌』の結果に同語判別情報を付与する手順を説明する。

（1）『茶筌』による解析

次の例は，『茶筌』で「赤いりんごをたべた」を解析した結果である[4]。左から，「見出し」（出現形），「見出し」（基本形），「読み」（基本形），「品詞」，「活用型」，「活用形」である。

赤い	赤い	アカイ	形容詞-自立	形容詞・アウオ形容詞	基本形
りんご	りんご	リンゴ	名詞-一般		
を	を	ヲ	助詞-格助詞-一般		
たべ	たべる	タベル	動詞-自立	一段	連用形
た	た	タ	助動詞	特殊・タ	基本形

（2）『表記統合辞書』の検索

『茶筌』の解析結果の「見出し」（基本形），「読み」（基本形），「品詞」，「活用型」と『表記統合辞書』の見出し部を比較し，一致するものを列挙する。結果は，次のようになる。左の４列が見出し部（「見出し」「読み」「品詞」「活用型」），５列目が同語判別情報である。

りんご	リンゴ	名詞-一般		りんご/リンゴ/林檎/苹果
食べる	タベル	動詞-自立	一段	たべる/食べる

（3）同語判別情報の対応付け

（2）で検索された結果を（1）の結果と対応付ければ，『茶筌』の解析結果への同語情報付与は完了である。（1）～（3）の処理を機械的に行うには，次のような方法がある。

a)　Perlなどのスクリプトで行う方法
b)　『茶筌』の解析結果と『表記統合辞書』をリレーショナルデータベースに表として格納し，両者を結合する方法

なお，a)の方法については，語種辞書『かたりぐさ』（本報告書第３部第８章の「語種辞書『かたりぐさ』の開発」（茂木俊伸）を参照のこと）の公開用Webページで公開されているPerlスクリプトが参考になる。

（4）同語判別

形態素解析結果に付与された同語判別情報を使って，同語判別を実際に行う際，付与された同語判別情報をどのように利用するかは，利用者によって異なると思われる。例えば，2.4節で指摘した問題を避けるため，カタカナ・ひらがなだけからなる辞書項目は他の辞書項目と同語とはしない，などである。ここでは，次の条件の下で，ある言語資料中に出現する形態素の出現頻度を測定する状況を考えてみる。

a)　語判別情報はすべて採用
b)　言語資料中に出現した範囲内で同語判別を実施

そして，（3）までの処理を行い，個々の形態素の出現頻度が次のようになったとしよう（右端の数値は

出現頻度)。このとき，同語判別情報を見ると，「りんご」と「リンゴ」は同一である。したがって，上記の条件a)から，同語であると判断する。さらに，b)の条件を適用すると，測定対象の言語資料に出現しなかった「苹果」は，同語判別情報から削除してもよいことになる。これにより，「りんご」「リンゴ」「林檎」すべての同語判別情報が同一になるので，この３種の形態素を同語と判断する。この結果，出現頻度は，３種の形態素を合算したものとなる。

りんご	リンゴ	名詞-一般	りんご/リンゴ/林檎/苹果	20
リンゴ	リンゴ	名詞-一般	りんご/リンゴ/林檎/苹果	5
林檎	リンゴ	名詞-一般	りんご/リンゴ/林檎	3

適用例

次に，形態素解析する際，どの程度，『表記統合辞書』の情報が利用されるかを検証するために，実際のデータに対して，（１）～（４）に示した方法で，形態素の出現頻度を計測してみた。実験条件，および，結果は次のとおりである。

実験条件
　解析対象： 毎日新聞 2002年（日外アソシエーツ）
　解析結果：
　　延べ形態素： 29988722
　　異なり形態素： 61372

ただし，「名詞-固有名詞」「未知語」「名詞-数」は計測の範囲外とする。

適用結果
　同語判別情報が付与された形態素数
　　延べ： 5090886（全体の17.0%）
　　異なり： 17873（全体の29.1%）
　適用後の異なり形態素数： 50585

この結果のとおり，適用前の異なり形態素数61372から適用後は50585に統合され，適用前の異なり形態素のうち，約8.2%が別の形態素と同語であると判別されていることがわかる。

４　データの公開と利用状況

『表記統合辞書』は，2005年7月より国立国語研究所のWebページ「言語データベースとソフトウェア」(http://www.kokken.go.jp/lrc)で無償公開している。使用条件は広範な利用と発展的な利用を考慮し，(IPADICと同様）一定の条件の下，複製，改変，再配布を許可している。

本節では，データの利用状況を分析する。分析には，ダウンロードする際に実施しているアンケートの結果を利用する。アンケートの実施条件は，次のとおりである。

　期間： 2005年7月11日～2006年12月23日
　設問： 所属，身分，専門分野，年齢，使用OS，コンピュータ使用歴
　　（各設問に対する回答は，すべて複数回答可）
　回答総数： 71件

なお，アンケートへの回答は，自由であり，回答しなくてもダウンロードすることは可能である。したがって，全利用者へのアンケート結果でないことに注意されたい。

アンケートの集計結果を図２に示す。

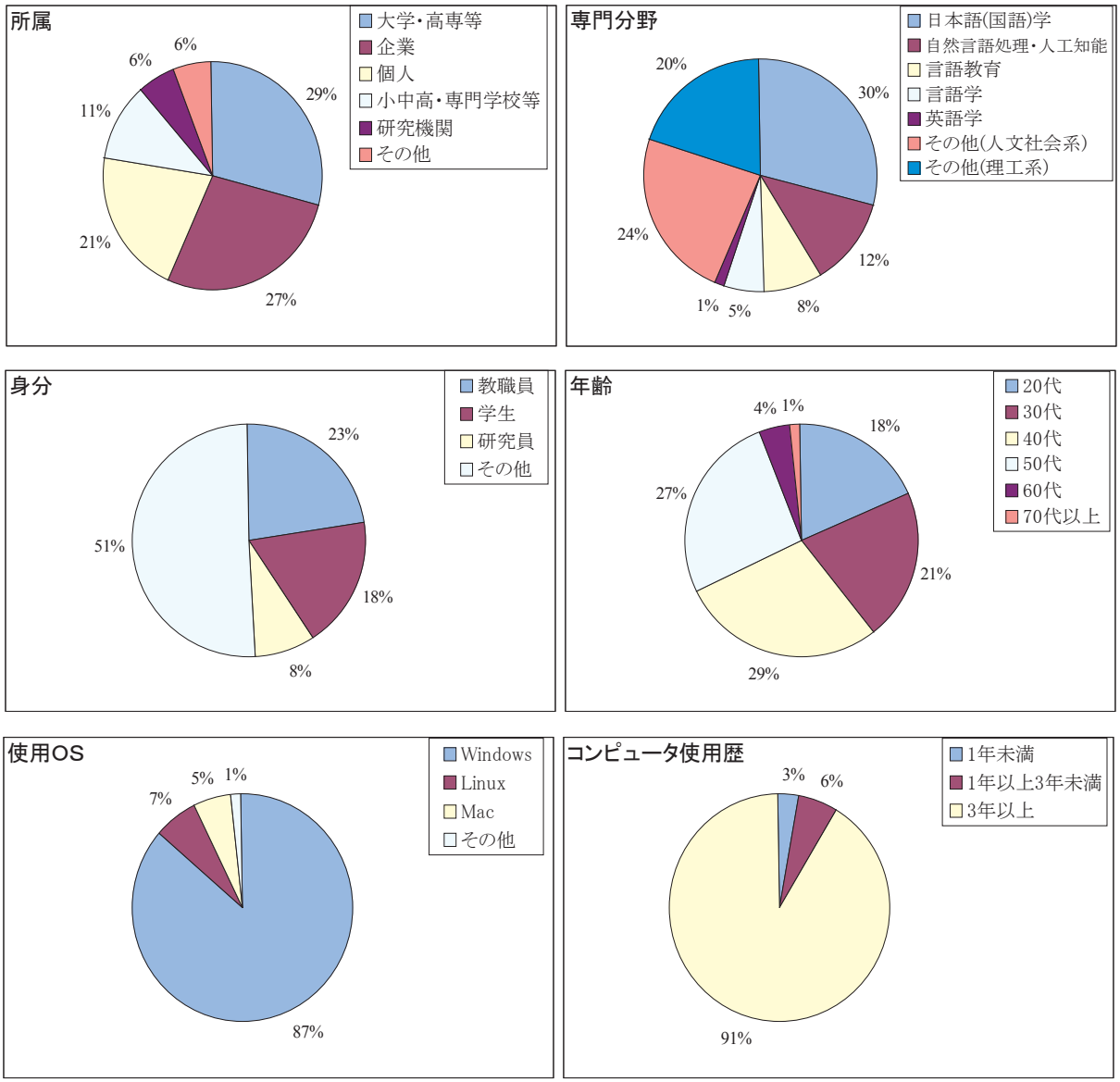

図２　アンケート結果

図２から利用者の特徴を分析してみる。まず，利用者の専門分野（図２「専門分野」）を見てみる。すると，日本語（国語）学・その他（人文社会系）が全体の54%，自然言語処理・人工知能・その他（理工系）は34%である。言語教育・言語学・英語学をあわせると，文科系の利用者の割合は66%となり，理科系の利用者の割合よりも大きいことがわかる。この結果は，『表記統合辞書』を自然言語処理システムの一部として利用しようというより，言語学的な分析に利用しようとしている利用者が多いことを示していると予想できる。

次に，「年齢」「所属」「身分」の面から分析してみると，利用者が実際の業務で利用しようとしていることが予想される。なぜならば，（１）「年齢」を見てみると，20代～50代が全体の95%をしめており，おおむね大学生以上の利用者であること[5]，（２）さらに，「所属」を見ると，「大学・高専等」「企業」「小中高・専門学校等」「研究機関」の合計が全体の73%を占めていること，この両者を勘案すると，実際の業務に携わっていることが読み取れるからである。

最後に「使用OS」と「コンピュータ歴」を見てみると，「３年以上」のコンピュータ歴の利用者が全体の91%を占めており，コンピュータの利用には慣れている利用者が大部分であることがわかる。その一方で，使用ＯＳは Windows が全体の87%を占めること，（前述のとおり）文科系の利用者が多いことを考えると，『表記統合辞書』を容易に利用できるような環境を提供することは有用であると考えられる。

５　おわりに

本稿では，『表記統合辞書』の構造や同語判別のための規則，および，利用方法について解説した。さらに，『表記統合辞書』の一般公開の状況を示すとともに，利用者の分析結果を示した。

注

1　http://chasen-legacy.sourceforge.jp/ で配布されている。詳しくは，左記 Web ページ，もしくは，松本ら（2001）を参照のこと。
2　詳細は，松本ら（2002）を参照のこと。
3　IPADIC(ver.2.4.4)は，『茶筌』のWebページ上からは削除されているが，次のURLから直接ダウンロードできる。
http://chasen.naist.jp/stable/ipadic/ipadic-2.4.4.tar.gz
また，Windows版の『茶筌』であるWinChaを使用する場合は，IPADIC(ver.2.4.4)を同梱しているcha21244sp5.exe を利用のこと。
4　オプションの設定によっては異なった結果が出る。設定の詳細は，『茶筌』のマニュアル（松本ら（2002））を参照のこと。
5　アンケートの項目には，「10代未満」「10代」もあったが，該当者はいなかった。

参考文献

日外アソシエーツ（2003）「CD－毎日新聞2002データ集」日外アソシエーツ
松本裕治，浅原正幸(2001)『IPADICユーザーズマニュアルver.2.4.4』
松本裕治ら（2002)『形態素解析システム『茶筌』version 2.2.9 使用説明書』

付記

『表記統合辞書』の企画・開発は，本稿の筆者である山口のほか，桐生りか，茂木俊伸（現鳴門教育大学），田中牧郎が担当した。
『表記統合辞書』の基礎データを構築されたICOT，および，形態素解析システム『茶筌』の開発者の方々に感謝する。