The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics

コーパス日本語学のための言語資源
―― 形態素解析用電子化辞書の開発とその応用 ――

伝　康晴（千葉大学）　小木曽　智信（国立国語研究所）　小椋　秀樹（国立国語研究所）　山田　篤（京都高度技術研究所）
峯松　信明（東京大学）　内元　清貴（情報通信研究機構）　小磯　花絵（国立国語研究所）

キーワード
電子化辞書，形態素解析，データベース，単位の斉一性，見出しの同一性

要　旨
コーパス日本語学への応用を指向した形態素解析用電子化辞書 UniDic を開発した。大規模コーパスに対する形態論情報付与作業には，計算機を用いた形態素解析システムの利用が不可欠であるが，既存の形態素解析システム用辞書には，コーパス日本語学への応用を考える上でさまざまな不都合がある。1つは，単位の認定がある場合には長く，ある場合には短いといった不揃いがあることであり，もう1つは，異表記や異形態に対して同一の見出しが与えられないということである。言語研究で重要な要件となる，このような単位の斉一性や見出しの同一性への対処といったことを中心に，本電子化辞書の設計方針とそれを実装した辞書データベースシステムについて述べる。さらに，この設計の有用性を示すため，表記や語形の変異に関するコーパス分析の事例を紹介する。

1. はじめに

本稿では，コーパス日本語学への応用を指向した形態素解析用電子化辞書 UniDic の開発について述べる。コーパス中のテキストを単語に分割し，それぞれに対して見出し語・品詞・語種などの形態論情報を与えることは，語彙・語法の研究や品詞の分布などから言語資料の特徴を明らかにしようという研究にとって欠かせない。従来，この種のデータ作成においては，人手による作業が行なわれてきた（たとえば国立国語研究所 1962, 1987, 1995）。しかし，大規模コーパス（たとえば100万語程度）に対して人手で形態論情報を与えることは多大な時間と多数の作業者を要し，その規模が500万語～1000万語以上ともなると人手による作業は実際上不可能となる。

これに対して，日本語の自然言語処理分野では，計算機を用いた自動形態素解析システムの研究が早くからなされており，JUMAN [1] や ChaSen [2] といったフリーソフトウェアが広く普及している。これらの形態素解析システムをコーパスへの形態論情報付与作業に利用することは，対象データの巨大化に伴って不可欠になってきている。実際，国立国語研究所が中心となって構築した『日本語話し言葉コーパス(CSJ)』（前川 2004）では，100万語相当のテキストに対して手作

業で情報付与した後，それらを学習データとして開発された形態素解析システム（Uchimoto et al. 2004）によって残りの約650万語に対する情報付与を行なっている。また，筆者たちも構築に関わっている『現代日本語書き言葉均衡コーパス』（山崎ほか 2006）では，その規模が1億語を超えることから，形態素解析システムの全面的な利用を当初から想定している。

その一方で，形態論情報付きコーパスを日本語研究に利用することを考えた場合，JUMANやChaSenなどの既存の形態素解析システムの出力にはさまざまな不都合がある。第一に，単位の斉一性の問題がある。中野(1998：156)は，言語の計量的研究における調査単位が備えているべき条件の1つとして以下のことを挙げている。

> ある言語現象に対して，あいまいさや矛盾がなく，一義的にその単位を切り取ることができること。切り取られた単位は等質であること。

単位が等質であるとは，ある同一の単位設計のもとで，ある場合には長く，ある場合には短く単位を認定するといったことがないということである。既存の形態素解析システムは，この問題にほとんど対処できていない。たとえば，JUMAN 5.1では，「幾何学」は1語として解析されるが，「心理｜学」は2語として解析され，「不心得者」と「無骨｜者」でも同様の不揃いが生じる。ChaSenの標準辞書であるipadic 2.7.0でも同様の問題が生じる。

第二に，見出しの同一性の問題がある。中野(1998：156)は，調査単位が備えているべきもう1つの条件として以下のことを挙げている。

> 切り取られた単位に対して，その見出しが決まること。ある単位と別の単位とが同じ見出しを持つか否かが見分けられること。

この問題は，異表記や異形態の扱いに関わる。計量的な語彙研究においては，送り仮名の違い（「表わす」と「表す」など）や新旧字体の違い（「撹乱」と「攪乱」など）を無視して同じ語とみなしたい場合が多い（あるいは，いったん同一の語とみなした上で，その変異を調べたいことが多い）。JUMANは5.0以降この問題に対処しているが，ChaSenではこの問題を解決できない[3]。また，とくに話し言葉を対象とするような場合，「大きい」と「おっきい」や「やはり」と「やっぱり」「やっぱし」のような語形の変異についても，一方は他方が変化した形であるという情報が欲しいことが多い。JUMANもChaSenもこの問題に対処していない[4]。

このような問題を解決すべく，筆者らはコーパス日本語学への応用を指向した形態素解析用電子化辞書UniDicの開発を数年前から進めてきた（伝ほか 2002）。本辞書の開発にあたっては，国立国語研究所のCSJ開発担当者・語彙調査担当者と当初から議論を重ね続けており，上述したような問題に対して一通りの解決策を提供している。また，平成18年度以降は，国立国語研究所研究開発部門言語資源グループと共同開発に着手し，辞書データベースシステムの開発や語彙の拡充などで飛躍的な進歩を遂げている。

本稿は，単位の斉一性や見出しの同一性への対処といったことを中心に，本電子化辞書の設計方針とそれを実装した辞書データベースシステムについて述べる。さらに，この設計の有用性を示すため，表記や語形の変異に関するコーパス分析の事例を紹介する。なお，本電子化辞書は，形態素解析システムChaSenで利用可能な形態で一般に公開されている[5]。

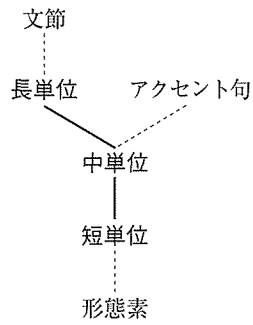

図1　単位の設計

2. 電子化辞書 UniDic の設計

2.1. 問題点の解決策

1節で述べた，単位の斉一性と見出しの同一性という2つの問題に対して，本電子化辞書 UniDic では以下のような解決策を与えた。

2.1.1. 単位の斉一性

何をもって基本の単位とするのがよいかということは，そのデータを用いてどのような研究を行なうかという目的に依存してしか決まらない。国立国語研究所の語彙調査（国立国語研究所 1962, 1987, 1995）においても，調査対象・目的に応じてその都度異なる単位が用いられてきている。これらの単位は大別して，長い単位の系列（α 単位・W 単位・長い単位・長単位）と短い単位の系列（β 単位・M 単位・短単位）に分けることができる（小椋ほか 2004)。これらの単位は，ある1つの単位をとったときには，「あいまいさや矛盾がなく，一義的にその単位を切り取ることができる（中野 1998)」ように単位認定手続きが厳格に与えられている。

本研究では，語彙調査におけるこのような単位設計方針に習って，複数の粒度の単位を設けることにし，『日本語話し言葉コーパス(CSJ)』で採用された短単位と長単位（小椋ほか 2004）を採用した（次節で述べるように，CSJ のものから一部変更されている)。数ある単位のうち CSJ の短単位・長単位を採用したのは，これらがもっとも最近になって策定されたものであり，もっとも洗練されていること，話し言葉に適用できることがおもな理由である。

短単位は，現代語で意味を持つ最小の単位（最小単位[6]）2個を1回結合したものであり，長単位は，文節を自立語と付属語（複合辞を含む）に分けたものである。これに対して，本研究では，音声研究への応用を念頭に置き，短単位と長単位の中間的な長さの単位として中単位も設けた。中単位はアクセント句の構成単位となることを想定している。3つの単位の関係を図1に示す。

これら3つの単位はいずれも手続き的に定義されており，あいまいさや矛盾のない単位認定が可能である。また，それぞれ，形態素・アクセント句・文節との関わりのもと，ほぼ等質に設計

されている。これらの単位を目的に応じて使い分けることにより，単位の斉一性を保持しつつ，幅広い応用研究に供することができる。

2.1.2. 見出しの同一性

何をもって同一の見出しとするかということは，簡単な問題ではない。たとえば，「熱い」「暑い」「厚い」「篤い」のうち，どれとどれが同じ見出しであるかという判断は国語辞典によってさまざまに異なる（中野 1998）。その一方で，送り仮名や新旧字体の違い，あるいは，活用による語形の変化など，明確に同語と判断できるものもある。

本研究では，見出しの同一性問題を以下のように細分化して考えた。

1. 語形の変異
   (a) 活用語の語尾変化（「書く」と「書か」「書き」「書け」など）
   (b)（語の複合に伴う）語頭音の変化（「ハタケ（畑）」と「バタケ」など）
   (c)（語の複合に伴う）語末音の変化（「サンカク（三角）」と「サンカッ」など）
   (d) 口語活用と文語活用の違い（「潔い」と「潔し」など）
   (e) サ行変格活用の五段化・上一段化（「愛する」と「愛す」，「信ずる」と「信じる」など）
   (f) 外来語の語形の違い（「アイデア」と「アイディア」など）
   (g) 慣用読みによる変化（「チョウフク（重複）」と「ジュウフク」など）
   (h) その他の音の転化（「大きい」と「おっきい」，「あなた」と「あんた」など）
2. 表記の変異
   (a) 送り仮名の違い（「表わす」と「表す」など）
   (b) 新旧字体の違い（「撹乱」と「攪乱」など）
   (c) 漢字と仮名の違い（「表わす」と「あらわす」，「猫」と「ネコ」など）
   (d) 漢字の違い（「愛敬」と「愛嬌」など）
   (e) 外来語の表記の違い（「データ」と「データー」など）
3. 発音の変異
   (a) 外来語の発音の違い（「データ」と「データー」など）

これらの変異をとらえるために，以下の4つのレベルで見出しを定義した（図2左）。

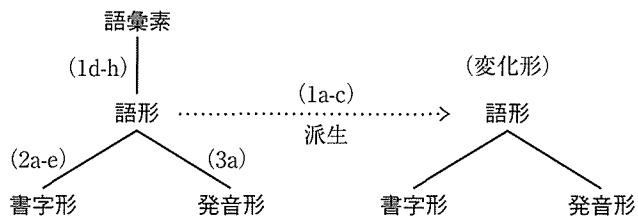

図2　見出しの設計

語彙素　変異を考慮せず，元来同一と見なしうる語に対して同一の見出しを与えたもの

語形　同じ語彙素に所属するものに対して，語形の変異を区別したもの

書字形　同じ語形に所属するものに対して，表記の変異を区別したもの

発音形　同じ語形に所属するものに対して，発音の変異を区別したもの

語形の変異は，活用や音の変化などにより，元来同一の語であったものの形態が変化した場合である。このうち，(1a)活用語の語尾変化・(1b)語頭音の変化（連濁など）・(1c)語末音の変化（促音化など）については，容易に同一性を認めることができる。実際，国語辞典では，基本的な形（活用語の終止形や語頭が清音・語末が直音のもの）のみを見出しとして挙げ，変化した形は掲載していないのが普通である。これは，変化した形を別見出しとはとらえていないということである。本研究でも，基本形のみを見出しとして登録し，変化形は別途記述した変化表を用いて派生する方式をとった（図２右）。

一方，(1d)口語活用・文語活用については，国語辞典では，口語活用の見出しのもとに，文語活用を併記するという形で，見出しの同一性を表現している。本研究では，これを語形の変異ととらえ，同一の語彙素に所属する異なる語形として登録した。(1f)外来語の語形の違いについても，同じ見出しのもとに括弧書きによる併記や定義文中で別語形を挙げている国語辞典が多く，本研究では語形の変異として扱った。ただし，「データ」と「データー」など，長音の有無をはじめ，発音にかかわる違いは，語形の変異ではなく，表記・発音の変異として扱った。さらに，(1e)サ行変格活用の五段化・上一段化・(1g)慣用読みによる変化・(1h)その他の音の転化については，国語辞典では，別見出しとしてあげつつ，一方を他方への参照見出しとしている場合が多い。本研究ではこれらも語形の変異とし，同一の語彙素に所属する異なる語形として登録した。

表記の変異は，同じ音を持つ語が異なる表記で記される場合である。書き言葉に固有の変異であることから，このレベルの変異を区別した見出しを書字形と名づけた。この種の変異のうち，(2a)送り仮名の違いと(2b)新旧字体の違いについては，容易に同一性を認めることができる。(2c)漢字と仮名の違いは多少問題である。たとえば，「ネコ」「ヒマワリ」などの動植物名は仮名書き（平仮名・片仮名とも）も広く用いられるが，「単位」「研究」などの漢語名詞が仮名書きされることはほとんどない。そこで，漢字と仮名による表記の変異はどこまで対象とするかを選択的にとらえた。(2d)漢字の違いはさらに問題である。本節冒頭に述べたように，異なる漢字で表記された語のどれとどれを同じ見出しとするかの判断は国語辞典によっても揺れがある。本研究では，『岩波国語辞典第六版』をおもに参考にしながら，国語辞典中で同一見出しになっているものでも，表記によって固有の語義が立てられている場合は，同一語の表記の変異とは見なさず，「元来同一と見なすことのできない語」として異なる語彙素に所属させた。同一見出しで，表記による語義の区別もない場合にのみ，同一の語形に所属する異なる書字形として登録した。

(2e) 外来語の表記の違いについては，(3a)外来語の発音の違いと対にして考えたほうがわかりやすい。発音の変異を区別した見出しとして，発音形というレベルを設けた。書字形が書き言葉に固有のものであるのに対して，発音形は話し言葉に固有のものである。たとえば，“data”と

いう語を考えると，これは書き言葉で「データ」と書く場合もあれば，「データー」と長音を付けて書く場合もある。同様に，話し言葉で「データ」と発声する場合もあれば，「データー」と延ばして発声する場合もある。これら表記と発音の変異の間には，必ずしも対応関係がないように見える。すなわち，「データ」と書いて，「データ」と読むことも「データー」と読むこともある。逆に，「データー」と書いて，「データー」と読むことも「データ」と読むこともある。ここでの発音の変異は，多分に個人差や状況によるものであり，表記に拘束されてのものではない。これに対して，“news”を「ニュース」と書いた場合には「ニュース」とのみ読み，「ニューズ」と書いた場合には「ニューズ」とのみ読むだろう。つまり，ここでは表記と発音の間に対応関係がある。

本研究では，“data”のようなケースでは，語形は1種類だけ設け，書字形と発音形にそれぞれ2種類の変異を認めた。書字形と発音形は互いに他方から独立して定義されており（図2参照），表記と発音の変異を組み合わせると上記のように2 × 2 = 4通りが可能である。これに対して，“news”のようなケースでは，2種類の語形を設け，それぞれ書字形と発音形には変異を認めなかった。この場合，表記と発音の変異を組み合わせても2通りしかない。このような外来語の表記・発音の変異については，『日本語話し言葉コーパス』の転記テキスト表記法（小磯ほか2006）の策定において利用された片仮名表記の揺れに関する資料を参考にした。

2.2. 階層的単位

2.1で述べたように，電子化辞書UniDicでは，単位の斉一性問題に対処するために，短・中・長単位からなる階層的単位設計を採用した。本節では，階層的単位についてより詳しく述べる。

階層的単位の例を図3に示す。「外来語仮名表記を調査した」という文に対して，短・中・長単位でそれぞれ8・6・4単位が認定される。もっとも大きな違いは「外来語仮名表記」の部分に見られる。この部分は，長単位では1単位だが，短単位では「外来｜語｜仮名｜表記」と4単位に分割され，中単位では「外来語｜仮名表記」の2単位に分割される。

	外来	語	仮名	表記	を	調査	し	た
短単位	名詞	名詞	名詞	名詞	助詞	名詞	動詞	助動詞
中単位	名詞		名詞		助詞	名詞	動詞	助動詞
長単位	名詞				助詞	動詞		助動詞

図3　階層的単位の例

■短単位　短単位は，原則として，現代語で意味を持つ最小の単位（最小単位）2個が1回結合したものである。たとえば，「外（ガイ）」と「来（ライ）」という2つの最小単位が結合して，「外来」という短単位ができる。ただし，最小単位2個の1回結合を1短単位とするのは原則であって，1最小単位を1短単位とする場合や3最小単位以上の結合を1短単位とする場合など，

いくつか例外規定がある。詳細は小椋(2006)を参照のこと。

本研究の短単位では，CSJ 短単位から以下の点を変更した。

- 外来語は1回結合せず，1最小単位で1短単位とした[7]。
- 補助記号（「・」「,」「。」「(」「)」など）を独立の最小単位として認定し，1最小単位で1短単位とした。
- 意志・推量の助動詞「う」「よう」を独立の最小単位とせず，活用語尾として活用語の最小単位に含めた[8]。

本研究では，市販の国語辞典や既存の形態素解析用辞書・コーパスから短単位に当てはまるものを人手で選定し，辞書登録した。この辞書を形態素解析システムで利用することによって，短単位の自動解析を行なう。

■長単位　長単位は，概ね，文節を自立語と付属語（複合辞を含む）に分けたものである。たとえば，「外来語仮名表記を」という文節は，自立語「外来語仮名表記」と付属語「を」に分けられる。このそれぞれが1長単位である。長単位認定規則の詳細は小椋(2006)を参照のこと。

長単位は，1短単位からなるか，あるいは，2つ以上の短単位を複合したものからなる。その複合パターンはさまざまであり，すべての長単位を前もって辞書登録しておくのは非現実的である。そこで，本研究では，長単位を辞書という形で記述するのではなく，短単位から長単位を自動構成するソフトウェアを作成するという方針を採ることにした。このような試みは，CSJ 短単位から長単位の自動構成としてすでにある程度の成功を収めている（Uchimoto et al. 2004)。

長単位では，「に関して」「という」「かもしれない」などの複合辞を1単位として認めている。複合辞の一覧は，小椋(2006)に挙げられたものを，ipadic 2.7.0などを参考にしながら，拡張しているところである。Uchimoto et al. (2004)のシステムでは，このような複合辞についても，コーパスからの学習によって，概ね正しく認定できる。

現時点ではまだ，本研究における長単位の自動解析精度を評価する段階にはいたっていない。長単位の扱いに関しては，今後，稿を改めて詳しく報告する。

■中単位　中単位は，短単位と長単位の中間に位置する単位である。これはアクセント句の構成単位となることを想定している。UniDic の開発当初から想定されていた応用分野の1つに，テキスト音声合成がある。テキスト音声合成とは，漢字仮名混じりテキストに対して，読みとアクセントを付与し，自然な音声として読み上げるソフトウェアのことである。そのためには，「段々」と「畑」が複合すると，「ハタケ」が連濁して「バタケ」となることや，アクセント型が頭高型の「ダ⌝ンダン」と平板型の「ハタケ」から中高型の「ダンダンバ⌝タケ」が生じることなどが導出できなければならない。

中単位はこのような処理に役立つ。語の複合による語頭・語末の音変化やアクセントの移動は，右分岐構造によって阻害されることが多い（佐藤 1990，窪園 1995)。たとえば，「外来語仮名表記」は［[外来 語］［仮名 表記]］のような内部構造を持ち，「仮名表記」の部分は右分岐に

なっている。このため，ここでの「仮名」は，「万葉仮名（マンヨウガナ）」にみられるような連濁は起こさない。中単位はこのような語の内部構造に従った単位であり，長単位を超えない範囲で，直接的な係り受け関係を持つ，隣接する短単位同士を結合したものとして定義できる。ただし，この定義の妥当性については，今後さらに検討する必要がある。

長単位同様，中単位も前もって辞書登録しておくことは非現実的であり，短・長単位をもとにした自動構成ソフトウェアの開発を計画している。

2.3. 階層的見出し

2.1で述べたように，電子化辞書 UniDic では，見出しの同一性問題に対処するために，語彙素・語形・書字形・発音形からなる階層的見出しを採用した。本節では，階層的見出しについてより詳しく述べる。本節で説明する階層的見出しの設計は，短・中・長単位のいずれのレベルにも適用できる。しかし，中・長単位に対してすべての語をあらかじめ辞書登録するというのは現実的ではなく，以下に述べる見出し語の登録は，実際には短単位に対してのみ行なっている。

階層的見出しの例を図4に示す。ここには，異なる8つの語彙素が掲載されている。このうち，「大きい」「貴方」「ニュース」はそれぞれ2つずつ語形を持っている。他の5つの語彙素はいずれも単一の語形しか持たない。語形のレベルには11個の項目がある。「アラワス」という同じ語形が2箇所に現れるが，これらは上位の語彙素のレベルで別項目（「表わす」と「著わす」）として区別されている。11個の語形のうち，6項目は送り仮名や漢字と仮名の違いによって2つ以上の書字形を持つ。11個の語形のほとんどは，単一の発音形しか持たないが，「データ」のみ2つの発音形を持つ。「データ」は書字形も2つ持つため，可能な書字形と発音形の組み合わせは4通りある。

■語彙素　語彙素は，語形・表記・発音の変異を考慮せず，意味・文法機能が同一であると見なしうるものに同一の見出しを与えたものである。見出しは，語彙素読み・語彙素表記・語彙素細分類・類の4つの属性によって定義される（図4では語彙素読み・語彙素表記のみを記してある）。

語彙素読み　その項目の読みを片仮名で記したもの
語彙素表記　同一の読みを持つ項目を区別するために漢字仮名混じりで表記したもの
語彙素細分類　語彙素表記によっても区別できない項目を区別するための細分類
類　その項目の意味・文法範疇を記したもの

語彙素読み・語彙素表記は，通常の国語辞典の見出しに相当するものである。語形や表記の変異を持ちうるものに関しては，CSJ 短単位の代表形や『岩波国語辞典』の見出しなどを参考にして，代表的と考えられる読み・表記を語彙素読み・語彙素表記として与えた。語彙素細分類は，これらで区別できない項目を区別するためのものである。おもに外来語（“light/right” の意味の「ライト」など）や省略語（「ブルジョア / ブルドッグ」を略した「ブル」など）の区別に用い，原語や完全形を記した。類は，意味・文法範疇を分類するものである。語彙素読み・語彙素表記

語彙素	語形	書字形	発音形
オオキイ【大きい】	オオキイ	大きい	オーキー
		おおきい	
	オッキイ	おっきい	オッキー
アナタ【貴方】	アナタ	貴方	アナタ
		あなた	
	アンタ	あんた	アンタ
アラワス【表わす】	アラワス	表わす	アラワス
		表す	
		あらわす	
アラワス【著わす】	アラワス	著わす	アラワス
		著す	
		あらわす	
カナ【仮名】	カナ	仮名	カナ
		かな	
カメイ【仮名】	カメイ	仮名	カメー
データ【データ】	データ	データ	データ
			データー
		データー	
			データ
ニュース【ニュース】	ニュース	ニュース	ニュース
	ニューズ	ニューズ	ニューズ

図4　階層的見出しの例

が同一で，意味・文法機能が異なるものは類で区別される。『分類語彙表（増補改訂版）』（国立国語研究所 2004）にある体・用・相の類に加え，姓・名・地名などの固有名詞や格助詞・接続助詞・助動詞などの付属語類についても，異なる項目が互いに区別できるよう分類を設けた。

語彙素はあくまでも辞書見出しを立てる目的から設定したものであり，国語辞典に見られる語義の分類や『分類語彙表』の部門以下に相当する細かい分類は行なわなかった。

■語形　語形は，同一の語彙素に対して，形態の違いを区別したものである。このレベルの見出しは，語形基本形・品詞・活用型の3つの属性によって定義される（図4では語形基本形のみを記してある）。

語形基本形　その項目の語形を片仮名で記したもの

品詞　その項目の品詞を記したもの

活用型　その項目が活用語である場合に，その活用の型を記したもの

語形基本形は，活用や複合による語頭・語末音の変化を無視し，ある語の基本的な形を片仮名で記したものである。「オオキイ」と「オッキイ」，「アナタ」と「アンタ」など，同一の語彙素

に所属する語形の異なる語は，この属性によって区別される。同一の語形を持つ語に対して，しばしば，品詞や活用型の差異を区別した。たとえば，一般名詞の「同期」(「同じ時期」の意）とサ変名詞の「同期」(「２つ以上の信号や処理のタイミングが合うこと」の意）や「死ぬ」の口語五段活用と文語ナ行変格活用などである。

品詞は，学校文法に概ね基づきつつ，ipadic 2.7.0や『岩波国語辞典』を参考にしながら，自動形態素解析に必要なレベルにまで詳細化した。「全て」のように，文脈に応じて複数の品詞（この場合，名詞と副詞）で用いられる語については，「名詞-普通名詞-副詞可能」のような複合的な品詞を与えた。品詞の一覧については伝ほか(2007)を参照のこと。

活用型も学校文法と ipadic 2.7.0を参考にして設定した。UniDic では，活用語については終止形のみを辞書登録し，その他の活用形は活用表を用いて派生している（2.4参照)。そのため，異なる活用パターンはすべて，異なる活用型として分類する必要がある。とくに話し言葉を考えた場合，通常よりもずっと詳細な分類が必要になる。たとえば，口語形容詞は，学校文法では単一の活用型しか持たず，「カロ・ク／カッ・イ・イ・ケレ」と活用するとされている。しかし，実際の話し言葉では，「高い（タカイ)」の連用形として「タコウ」の形や終止形・連体形として「タケエ」の形がある。つまり，語幹部分の末尾も変化を生じる。このことを踏まえて活用パターンを記述すると，「カカロ・カク／カカッ／コウ・カイ／ケエ・カイ／ケエ・カケレ」のようになる。これは，語幹末の音に応じて異なる活用型を設定する必要があることを意味する。

■書字形　書字形は，同一の語形に対して，表記の違いを区別したものである。このレベルの見出しは，書字形基本形によって定義される。

書字形基本形　その項目の表記を漢字仮名混じりで記したもの

書字形基本形は，ある語の可能な表記を漢字仮名混じりで記したものである。たとえば，「表わす」「表す」「あらわす」などの表記の違いは，この属性によって区別される。ただし，「あらわす」のような仮名による表記については，選択的に辞書登録している。選択に際しては ipadic 2.7.0を参考にした。

■発音形　発音形は，同一の語形に対して，発音の違いを区別したものである。このレベルの見出しは，発音形基本形によって定義される。

発音形基本形　その項目の発音を片仮名で記したもの

発音形基本形は，ある語の可能な発音を片仮名で記したものである。たとえば，「データ」「データー」などの発音の違いは，この属性によって区別される。ただし，標準的でない変異（「国語」を「コッゴ」と発音など）は対象としない。話し言葉の実情にあわせ，同一母音の連続や[ei]，[ou]連続の２モーラ目を長音で表わした（最小単位境界をまたぐ場合を除く)。

以上述べた各階層の見出しを定義する属性をまとめると，表１の見出し属性のようになる。

表1　UniDicで用いられる属性

階層	見出し属性	その他の属性
語彙素	語彙素読み，語彙素表記	語種，意味分類
	語彙素細分類，類	
語形	語形基本形	語頭変化型,語末変化型
	品詞，活用型	簡略活用型,語頭変化結合型，語末変化結合型
書字形	書字形基本形	活用型書字形分類，語頭変化型書字形分類，語末変化型書字形分類
発音形	発音形基本形	活用型発音形分類，アクセント型，アクセント結合型

表2　活用表

活用型：形容詞-ア段-カイ				
活用形	語形	書字形		発音形
		漢字	仮名	
意志推量形	カカロウ	かろう	かかろう	カカロー
連用形-一般	カク	く	かく	カク
連用形-促音便	カカッ	かっ	かかっ	カカッ
連用形-ウ音便	コウ	う	こう	コー
終止形-一般	カイ ケエ	い え	かい けえ	カイ ケー
連体形-一般	カイ ケエ	い え	かい けえ	カイ ケー
仮定形-一般	カケレ	けれ	かけれ	カケレ

2.4. 変化表

2.1で述べたように，活用語の語尾変化や複合に伴う語頭・語末音の変化（連濁や促音化など）は，変化表を用いた派生として扱う。

2.4.1. 活用による変化

辞書登録された活用語の終止形からすべての活用形を得るために，表2のような活用表を用いる。この表は，語幹末が「カ」で終わる形容詞の活用パターンを定義している（説明の都合上簡略化してある）[9]。たとえば，「タカイ（高い）」の活用型は「形容詞-ア段-カイ」と定義されており，この活用表から「タカカロウ・タカク／タカカッ／タコウ・タカイ／タケエ・タカイ／タケエ・タカケレ」の変化形が得られる（UniDicでは意志推量形を立てていることに注意）。

この表はさらに，書字形や発音形がどのように変化するかも定義している。この例では，書字形の変化には2通りある。「漢字」と書かれた列は語幹が漢字の場合の変化であり，「高かろう・高く／高かっ／高う・高い／高え・高い／高え・高けれ」のように語尾のみが変化する。一方，「仮名」と書かれた列は語幹末が仮名の場合の変化であり，「暖かかろう・暖かく／暖かかっ／暖

こう・暖かい／暖けえ・暖かい／暖けえ・暖かけれ」のように語幹末も含めて変化する。このため，同じ「形容詞-ア段-カイ」という活用型であっても，いずれの変化パターンをとるかが書字形ごとに別個に指定されている。「漢字」「仮名」のように，書字形ごとの活用変化の違いを分類したものを活用型書字形分類とよぶ。同様の分類が発音形に関してもなされている[10]。

このような事情から，UniDic における活用型の分類は非常に詳細になっている。しかし，活用表を用いてすべての活用形がひとたび派生されたなら，もはやこのような詳細な分類は必要ない。コーパスに情報付与を行なったり，語彙調査に用いたりする場合にはむしろ煩雑すぎる。そこで，学校文法と同程度の詳細さ（たとえば口語形容詞はすべて「形容詞型」）の「簡略活用型」を別途記述し，形態素解析システムの出力にはそちらを利用した（一覧は伝ほか(2007)を参照）。

2.4.2. 語頭音の変化

活用による変化以外で，基本形からの派生によって扱うものに，語の複合に伴う語頭音の変化（連濁）がある。たとえば，「段々畑」の「バタケ」や「株式会社」の「ガイシャ」は，語頭音が濁音化しているという点を除けば，「ハタケ」や「カイシャ」と同一の語と見なすことができる。そこで，これらの濁音化した形態を，活用表と同様な表を使って，基本形から派生しようというのである。まず，語頭音の変化を生じうる語に対して，変化パターンの違いを分類し，変化の型（活用語の活用型に相当する）を与えた。これを語頭変化型とよぶ。次に，語頭変化型ごとに変化表によって，語形・書字形・発音形の変化パターンを定義した。

表 3 に語頭変化表の例を挙げる。たとえば，「ハタケ（畑）」の語頭変化型は，「ハ濁音」と定義されており，この変化表から「ハタケ・バタケ」の変化形が得られる。書字形の変化には，語頭変化型書字形分類に応じて 2 通りある。書字形が漢字で書かれている場合は「畑」で変化しない。一方，仮名で書かれている場合は，「はたけ・ばたけ」のように変化する。同様に，接尾辞「ハイ（杯）」の語頭変化型は，「ハ混合」と定義されており，変化表から，「ハイ・バイ・パイ」の 3 つの変化形が得られる（それぞれ数詞の「二」「三」「一」と結合するときの形）。

このように，活用語の活用形を派生するのとまったく同じ方式で，連濁した形を派生できる。ただし，これはあくまでも連濁した形を辞書として提供する方式について述べたものであり，ある文脈において基本形と濁音形のどちらを選択するかというのはまた別の問題である。後者は形態素解析の処理系の問題となる。

表 3　語頭変化表

語頭変化型：ハ濁音				
語頭変化形	語形	書字形		発音形
		漢字	仮名	
基本形	ハ		は	ハ
濁音形	バ		ば	バ

語頭変化型：ハ混合				
語頭変化形	語形	書字形		発音形
		漢字	仮名	
基本形	ハ		は	ハ
濁音形	バ		ば	ハ
半濁音形	パ		ぱ	ハ

2.4.3. 語末音の変化

語の複合に伴う語末音の変化も同じ方法で扱える。たとえば，「サンカク（三角）」が接尾辞の「ケイ（形）」と複合すると，「サンカッ」のように語末音が促音化する。語頭変化の場合と同様に，語末変化型と語末変化表を定義することで，この種の変化形を派生できる。

語末音の変化は数詞でよく見られる。現状では以下のものを扱っている。

- 促音化：「イチ（一）」→「イッ」（「回」などにつながるとき）
- 促音添加：「ヨ（四）」→「ヨッ」（「つ」などにつながるとき）
- 撥音添加：「ヨ（四）」→「ヨン」（「回」などにつながるとき）
- 長音添加：「ヨ（四）」→「ヨウ」（ヒイ，フウ，ミイ…と数えるとき）

2.5. その他の属性

以上の属性以外にもいくつかの属性を記述（計画中を含む）している（表1参照）。

語彙素　語種，意味分類

語形　語頭変化結合型，語末変化結合型

発音形　アクセント型，アクセント結合型

語彙素レベルでは，漢語・和語・外来語・混成語などの語種の記述を計画している。また，『分類語彙表』（国立国語研究所 2004）の意味分類を記述することも計画している。ただし，意味分類に応じて語彙素を細分化するのではなく，複数の分類が当てはまる項目については併記する形をとる。『分類語彙表』に記載がない項目についてどうするかなど，詳細は今後の課題である。

語形レベルでは，語頭・語末音の変化に関連して，語頭・語末変化結合型を記述している。たとえば，「イッポン（一本）」という複合語を考えると，「ホン」の語頭が半濁音化するのは，先行要素が数詞「一」であることに依存している。そこで，「一」の辞書項目に「「ホン」の語頭を半濁音化させる」という情報を記載しておけば，形態素解析時の語頭変化形の選択に役立てることができる。この種の情報を分類したものを語頭変化結合型とよぶ。同様に，助数詞「本」の辞書項目に「「イチ」の語末を促音化させる」という内容を持った語末変化結合型を記載することで，「イチ」の語末変化形の選択に役立てることができる。数詞と助数詞を中心としてこのような情報を記述している。詳細は伝ほか（2002，2007）を参照のこと。

発音形レベルでは，アクセント情報を記述している。各語が単独発声されたときのアクセント位置を語頭からのモーラ数によって記した。記述には，『NHK 日本語発音アクセント辞典』『三省堂大辞林』などを参考にした。複数のアクセント型が可能な場合は併記した。さらに，頭高型の「ダ[1]ンダン」と平板型の「ハタケ」から中高型の「ダンダンバ[1]タケ」が生じるといったことを導出するために，アクセントの移動に関わる情報（アクセント結合型）を記述した。たとえば，「ハタケ」のアクセント結合型はC2型と記述されており，これは先行要素との結合点の次のモーラにアクセントが置かれることを意味する。アクセント結合型の記述は，匂坂・佐藤(1983)の定式化を参考にした。詳細は伝ほか(2002，2007)，Minematsu et al. (2003)を参照のこと。

3. 実装

本節では，電子化辞書 UniDic の計算機上での実装について簡単に述べる。システム全体の構成を図5に示す。

3.1. 辞書データベースシステム

辞書データベースシステムは，電子化辞書 UniDic をリレーショナルデータベースとして実現したものである。実装には Microsoft SQL Server 2005を用いた。

リレーショナルデータベースは，表形式のデータ（テーブル）を複数個関連付け，データベース全体を有機的に構成したものである。本システムでは，表1の4階層の見出しをそれぞれテーブルとして記述し，それらの間の階層関係を ID を介した参照関係によって表現した。同一の上位項目に複数の下位項目が所属する場合は，表1の見出し属性によって互いに識別される。見出し属性に加えて，その他の属性も階層に従って各テーブルに記述した。さらに，活用語尾や語頭・語末音の変化に関する変化表もそれぞれテーブルとして記述した（図5「辞書データベース」）。

データベース理論の観点から見ると，この設計はデータベースの正規化を実現していることになる。正規化とは，同一の情報を複数の箇所に重複して記述することなく，「1事実1箇所」の形にすることをいう。たとえば，「表わす」の語種が和語であることは，この語が「表わす」「表す」「あらわす」のいずれで表記されるかということに関わらない。そこで，語種情報は，書字形ごとに記述するのではなく，より上位の語彙素のレベルで記述すれば十分である。むしろ，このようにすることで，書字形によって異なる語種を与えてしまうというミスが生じる危険性を回避できる。同様に，ある語の品詞や語種は活用形の違いによらず一定であるから，これらの情報は終止形に対してのみ与え，その他の活用形については語尾変化に関する情報だけを記述すればよい。このような「1事実1箇所」の原則を守ることは巨大なデータベースの一貫性を維持する上で必要不可欠であり，前節で述べた UniDic の設計はそれを自然に実現している（データベースの設計や正規化に関する平易な入門書として高橋・飯室(2002)などを参照）。

一方で，このような階層関係を意識しながら，辞書登録作業を行なっていくことは，作業者に多大な負担を強いることになる。そこで，国立国語研究所において，辞書登録作業を支援するためのユーザインターフェースを Microsoft Access のフォーム機能を用いて作成した（図6）。このインターフェースを用いると，語彙素・語形・書字形・発音形のさまざまなレベルで既存の登録内容を検索でき，検索結果を階層関係とともに表示することができる（図6の左側のツリー表示）。また，階層構造を直接編集することもできる。たとえば，ある書字形を別の語彙素として独立させたり，既存の別の語形の下に移動したりといったことができる。

このようなインターフェースを用いて，国立国語研究所研究開発部門言語資源グループにおいて辞書登録作業を進めている。2007年4月に公開した辞書では，語彙素10万6千・語形11万・書字形13万6千・発音形11万の登録数（いずれも概数）を数え，その後も随時拡張している。

リレーショナルデータベースでは，一貫性の保持のため，さまざまなテーブルに分散して情報

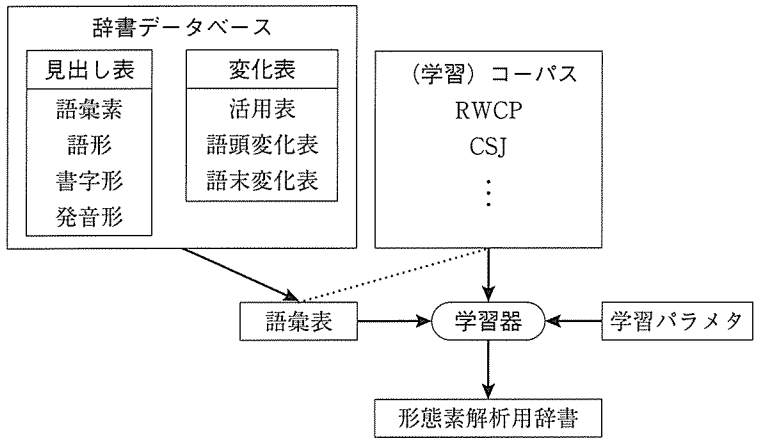

図5　システムの構成

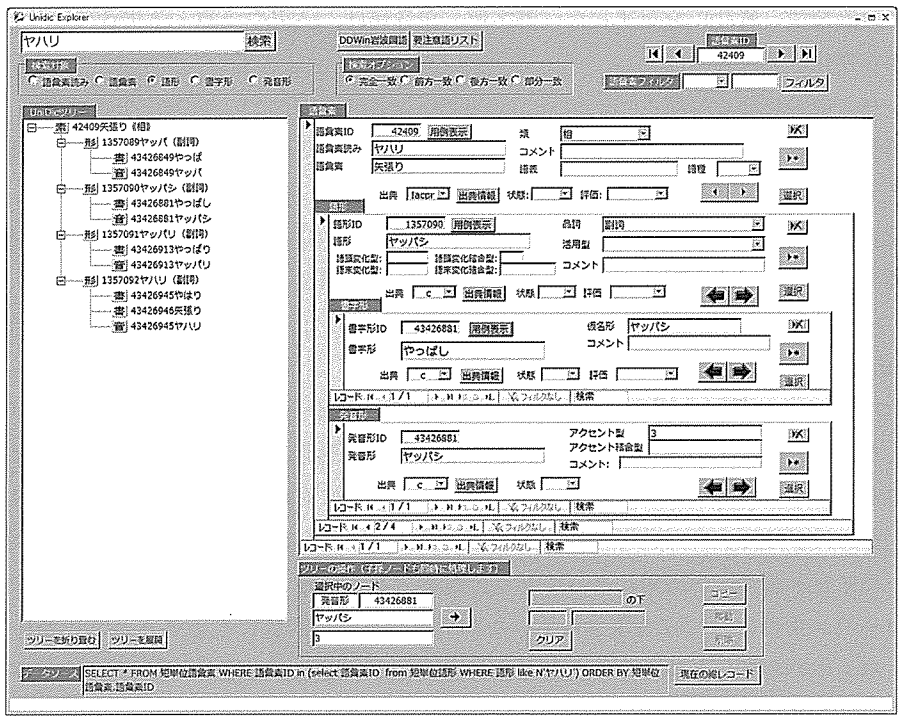

図6　辞書登録フォーム

が記述されている。これらを1つの表にまとめて利用したいことがある。リレーショナルデータベースでは，そのためにテーブルの結合という機能を提供している。これによって，「表わす」「表す」「あらわす」の各書字形に，上位階層に記述された「和語」という語種情報が複写される。同様に，終止形「表わす」に記述された品詞や語種情報が「表わさ」「表わし」「表わせ」などの活用形に複写される。このようにして，すべての語彙素・語形・書字形・発音形・活用形および語頭・語末変化形を1枚の表に展開したものを語彙表とよぶ。語彙表は約35万7千項目からなり，コーパスとの整合性維持に利用するとともに，形態素解析用辞書を生成するもととなる。

3.2. コーパスとの統合管理

ChaSenなどの形態素解析システムでは，システムの動作を制御するために，コーパス中の語の分布の統計情報を利用している。たとえば，各語の生じやすさや品詞間の結合しやすさなどを学習コーパスから取得している。本研究でもこの方式を採用した。

そのために，学習コーパスを短単位に分割・情報付与し，リレーショナルデータベースに登録した。学習コーパスとして，書き言葉の『RWCPテキストコーパス』(RWCP 1998)と話し言葉の『日本語話し言葉コーパス』（前川 2004）をおもに用いた[11]。前者は約95万単位，後者は約100万単位からなる。データベース中では，各語のテキスト中での位置情報・文境界情報および以下の形態論情報（属性）を記述した（書字形・発音形は，活用や語頭・語末音の変化に応じて基本形から変化させたもの)。

コーパス中の形態論情報：

語彙素読み・語彙素表記・語彙素細分類・品詞・簡略活用型・活用形・書字形・発音形

本システムでは，コーパス中に出現する語はすべて語彙表に掲載されているものとし，コーパスと語彙表との間に参照関係を設定した（図5の点線)。このように，コーパスと辞書とを統合的に管理することには，以下のような利点がある（伝・浅原 2001，浅原ほか 2002)。

1. コーパス中の異なる位置に出現する同一の語に対して，一部の属性を異なって与えてしまうというミスを未然に防止する。
2. コーパス中に陽に表現されていない情報を辞書記述から派生的に取得することができる。

1は，本質的には「1事実1箇所」と同じことである。形態論情報は辞書中に記述されているのであるから，コーパス中には各単位が辞書中のどの項目に対応するかだけを記載すれば十分であるし，それによってコーパス中の形態論情報の記述に不整合が生じる危険性を回避できる。2は，コーパス日本語学への応用にとってとくに重要である。たとえば，意味分類や語種といった情報はコーパス中には陽には表現されていない（現時点では辞書にも記述されていない)。しかし，コーパス中の各単位と辞書中の項目との対応関係がひとたび設定されたなら，これらの情報は（辞書中に与えられれば）いつ何時でもコーパスに複写することができる。このことは，本電子化辞書のコーパス日本語学における位置付けの重要性を示している。つまり，本電子化辞書は，たんに形態素解析用辞書というだけでなく，コーパス中に記述される情報を漸進的に豊富にしていく貴重な資源となっている。

3.3. 形態素解析システムでの利用

電子化辞書 UniDic を自動形態素解析に利用するため，語彙表と学習コーパスから形態素解析用辞書を生成した（図 5 下部）。ChaSen や MeCab[12]などの形態素解析システムでは，処理エンジンと辞書とが独立しており，標準辞書とは別の辞書を用いてシステムを構成することができる。本研究では，人文系分野も含めて広く普及していることから，ChaSen を採用した。

辞書データベースから ChaSen 用辞書を生成するための学習器を作成した。学習モデルには，拡張隠れマルコフモデル（浅原・松本 2002）を用いた。モデルの調整のための学習パラメタを適宜設定した。学習には，前項で述べた『RWCP テキストコーパス』（新聞記事），『日本語話し言葉コーパス』（学会講演や模擬講演）を中心とする学習コーパス約200万単位を用いた。

得られた形態素解析用辞書の解析精度を評価した。評価は，インサイド評価とアウトサイド評価の 2 通りを行なった。インサイド評価とは，統計学習に用いたデータそのものを解析し，精度を測ることをいう。これに対して，アウトサイド評価とは，コーパスを学習データとテストデータに分け，前者で学習したモデルを用いて後者を解析し，精度を測ることをいう。ここでは，9 割（約180万単位）を学習データ，1 割（約20万単位）をテストデータにランダムに割り当てた。ただし，語彙は全データから取得した（つまり未知語はない）。解析精度（F 値：再現率と適合率の調和平均）は，インサイド/アウトサイド評価でそれぞれ，単位認定99.5%/99.3%，品詞（活用型・活用形）認定98.0%/97.7%，語彙素（読み・表記）認定97.6%/97.2%であった。なお，現在公開されている辞書は，ここでのインサイド評価用に作成したもの（学習コーパス全体から学習したもの）である。

4. コーパス日本語学への応用

本電子化辞書 UniDic の設計の有用性，とくに見出しの同一性を表現したことの利点を示すため，表記や語形の変異に関するコーパス分析の事例を紹介する。なお，ここでの分析事例は，あくまでも UniDic のコーパス日本語学分野での有用性を示すためのものであり，特定の調査目的を持つものではないことを断っておく。

4.1. 話し言葉に見られる語形の変異

話し言葉では，「大きい」に対する「おっきい」，「あなた」に対する「あんた」など，さまざまな語形の変異が見られる。これを『日本語話し言葉コーパス』（学会講演と模擬講演）を使って分析してみよう。ここでは，2.1.2で挙げた語形の変異のうち，(1a)活用語の語尾変化・(1b)語頭音の変化（連濁など）・(1c)語末音の変化（促音化など）は除いて考えたい。もとのコーパスには，UniDic の語彙素読み・語彙素表記に相当する代表形・代表表記が与えられているが，これは活用や連濁・促音化による変化など，あらゆる変異を同一化したものであり，ここでの目的に使うには情報が足りない。UniDic を用いると，辞書から語形基本形などの情報を取得でき，より精密な集計が簡単にできる。

表 4 は，出現頻度50以上の語彙素のうち，2 つ以上の語形を持つものについて，語形の変異の

表4　CSJにおける語形の変異（内訳の括弧内は語形ごとの比率）

語彙素	品詞	頻度	語形の内訳
エート【えーと】	フィラー	4432	エート（42.4%），エーット（18.6%），エートー（15.6%），エット（8.8%），エト（5.2%），エットー（4.5%），エーットー（3.7%），エトー（1.3%）
ント【んと】	フィラー	119	ンート（43.7%），ント（20.2%），ンートー（10.9%），ンーット（6.7%），ントー（5.9%），ンット（5.9%），ンットー（3.4%），ンーットー（3.4%）
ヤハリ【矢張り】	副詞	1291	ヤッパリ（50.0%），ヤハリ（43.5%），ヤッパ（4.4%），ヤッパシ（2.0%）
ケレド【けれど】	接続助詞	6385	ケレド（50.4%），ケド（49.6%）
イー【いー】	フィラー	468	イ（51.3%），イー（48.7%）
アー【あー】	フィラー	3698	ア（52.2%），アー（44.7%），アッ（3.1%），アア（0.0%），アーア（0.0%）
アノ【あの】	フィラー	11995	アノー（52.3%），アノ（44.1%），アーノー（1.9%），アーノ（1.7%），アンノ（0.0%）
ノ【の】	準体助詞	20112	ン（52.5%），ノ（47.5%）
ツウ【つう】	助動詞	139	ッツウ（52.5%），ッチュウ（30.2%），ツウ（14.4%），チュウ（2.9%）
アマリ【余り】	副詞	468	アマリ（53.4%），アンマリ（45.0%），アンマ（1.6%）

大きなもの（最頻語形の比率の小さな語）を上位10項目まで示したものである（品詞は簡略化してある）。ただし，学会講演では，文語の文を引用したり，語のメタ的な引用を用いたりすることがあるので，それらは分析から除外した。フィラーに関する変異が多いが，それ以外にも「やはり」「けれど」「の」「あまり」などで話し言葉に特有の変異が見られる。とくに，副詞「やはり」や準体助詞「の」では，異形態である「ヤッパリ」「ン」のほうが多用されている。

4.2. 書き言葉に見られる表記の変異

新聞記事や雑誌などでは，「表わす」「表す」「あらわす」など，同一の語に対する表記の揺れが多く見られる。ここでは，『RWCPテキストコーパス』（新聞記事）を使ってこのような表記の変異を分析してみよう。ただし，2.1.2で挙げた(2a)-(2e)の変異を区別なく扱うものとする。

表5は，出現頻度50以上の語形のうち，2つ以上の書字形を持つものについて，表記の変異の大きなもの（最頻書字形の比率の小さな語）を上位10項目まで示したものである[13]。ただし，UniDicでは，位取りされる算用数字をすべて漢数字に変換して扱っているため[14]，数詞は分析から除外した。動詞・動詞性接尾辞の漢字・仮名の表記違いが多いことがわかる。とくに，動詞では「見る」「付ける（ヅケル）」など補助動詞として使われる語が多い（後者は濁音化していることから補助動詞用法であることがわかる）。名詞では，「今日」「始め」など，副詞的に用いたり，連用修飾句を導いたりする語が目立つ。

表5　RWCP コーパスにおける表記の変異（内訳の括弧内は書字形ごとの比率）

語彙素	語形	品詞	頻度	書字形の内訳
イエル【言える】	イエル	動詞	114	言える (50.9%), いえる (49.1%)
カップ【カップ】	カップ	普通名詞	71	カップ (53.5%), 杯 (46.5%)
ミル【見る】	ミル	動詞	1014	見る (53.9%), みる (46.1%)
オモシロイ【面白い】	オモシロイ	形容詞	63	面白い (54.0%), おもしろい (46.0%)
キョウ【今日】	キョウ	普通名詞	115	きょう (54.8%), 今日 (45.2%)
サマ【様】	サマ	名詞性接尾辞	111	様 (55.0%), さま (44.1%), サマ (0.9%)
ハジメ【始め】	ハジメ	普通名詞	123	はじめ (55.3%), 初め (41.5%), 始め (3.3%)
ツケル【付ける】	ヅケル	動詞	56	づける (57.1%), 付ける (42.9%)
スギ【過ぎ】	スギ	動詞性接尾辞	52	過ぎ (59.6%), すぎ (40.4%)
トル【取る】	トル	動詞	354	とる (60.2%), 取る (39.8%)

5. おわりに

本稿では，コーパス日本語学への応用を指向した形態素解析用電子化辞書 UniDic の開発について述べた。言語研究で重要な要件となる単位の斉一性や見出しの同一性への対処を中心に，本電子化辞書の設計方針とその実装について述べ，コーパス日本語学への応用事例を紹介した。作成した形態素解析用辞書は，形態素解析システム ChaSen で利用可能な形態で一般に公開されている。

今後の課題としてはおもに 3 つの問題が挙げられるだろう。第一に，語彙の拡充である。UniDic の語彙素数はすでに10万項目を超えているが，これらは市販の国語辞典や『RWCP テキストコーパス』（新聞記事），『日本語話し言葉コーパス』（学会講演や模擬講演）からおもに取得したものである。雑誌やインターネット上の文書を対象とする場合には，まだまだ十分とはいえない。また，すべての語を登録し尽くすことは原理的に不可能であるから，未知語に対処する技術も並行して考える必要がある。

第二に，記載する情報の拡大である。現状の UniDic に記載されている情報は，基本的な形態論情報とアクセント情報に限られる。辞書とコーパスの統合によって，コーパスの情報を漸進的に豊富にしていく，というアイデアを実のあるものにするには，まずもって辞書中に豊富な情報を記載していかなければならない。現在，語種と意味分類を記載する計画があるが，とくに後者については，参考資料中に記載がないものはどうするかなど課題も多い。

第三に，形態素解析システムの高精度化である。現状の解析精度97%以上という数字は決して低いものではないが，コーパスへの自動情報付与に利用することを考えると十分とはいえない。とくに，既存の形態素解析システムでは，単位・品詞認定のみが重視され，語彙素認定は考慮されていない。実際，本研究で採用した ChaSen では，語彙素認定に関して極めて貧弱なモデルしか提供できない。語彙素レベルの見出しを与えることは，UniDic の主たる目的の 1 つであるから，このレベルでの精度向上は欠かせない。今後，MeCab など最新の統計モデルを備えたシステムを採用するなどして，対処するつもりである。

これらの課題に対して，開発者として拡張・改善に取り組んでいくとともに，随時改訂版を公開し，利用者からのフィードバック（未登録語の報告や新たな記載情報に関する提案・データ提供）を得つつ，コーパス日本語学のための言語資源としてより価値を高めていきたい。

注

1 http://nlp.kuee.kyoto-u.ac.jp/nl-resource/juman.html

2 http://chasen-legacy.sourceforge.jp/

3 ChaSen 開発チームは，この問題に対する対処を現在進めている（浅原ほか 2005）。ただし，異形態の扱いについては対処していない。

4 JUMAN 5.1はこの問題に部分的には対処しているようである。たとえば，「温かい」と「あったかい」に対しては同一の“代表表記”が与えられている。一方，「やはり」と「やっぱり」の同一性は表現されていない。JUMAN 辞書ではもともと話し言葉の語彙が少なく，語形の変異に関して体系的に対処しようとは考えていないように思われる。

5 http://download.unidic.org/

6 言語学でいう形態素に概ね対応するが，活用語の活用語尾を独立した単位としないなどの違いがある。

7 『現代日本語書き言葉均衡コーパス』の形態論情報の記述でも，同様の変更がなされている。

8 「う」の発音が長音「ー」となるため，テキスト音声合成などへの応用で不都合があること，話し言葉で「行こ」などの「う」を省略した形があり，記述が困難であることなどが理由である。

9 表2に挙げた以外にも，話し言葉に現れるさまざまな活用形（意志推量形の末尾を縮めた「高かろ」，終止形の末尾を促音化した「高っ」，仮定形の融合形「高けりゃ」「高きゃ」など）に対応している。

10 発音形で変化パターンの差異が生じるのは，たとえばカ行・ガ行五段動詞の場合である。「キク（聞く）」のように，語幹末がイ段またはエ段で終わる場合，連用形イ音便の発音が「キー（タ）」のように長音になる。これに対して，「サク（咲く）」のように，語幹末がイ段・エ段以外で終わる場合，連用形イ音便は「サイ（タ）」のように「イ」である。

11 RWCP コーパスは，オリジナルのものに ChaSen 開発チームが修正を加えたものをさらに短単位に再分析した。CSJ も，CSJ 短単位と本研究の短単位の違いに応じて，適宜再分析した。

12 http://mecab.sourceforge.jp/

13 「カップ」の書字形「杯」は，RWCP コーパスで「W 杯」を「ワールド｜カップ」と認定していることから。この認定自体は再考の余地がある。

14 短単位では，数詞は千・百・十の桁ごとに単位分割する（「二千｜七」など）。このため，算用数字のままでは扱いにくく，形態素解析時には前処理で漢数字に変換している。コーパス中でも同様に漢数字で表現している。

参考文献

浅原正幸・高橋由梨加・松本裕治(2005)「異表記同語情報を付与した辞書の整備」『言語処理学会第11回年次大会発表論文集』，604-607

浅原正幸・松本裕治(2002)「形態素解析のための拡張統計モデル」『情報処理学会論文誌』43，685-

695
浅原正幸・米田隆一・山下亜希子・伝康晴・松本裕治(2002)「語長変換を考慮したコーパス管理システム」『情報処理学会論文誌』43, 2091-2097
小椋秀樹(2006)「形態論情報」『国立国語研究所報告124 日本語話し言葉コーパスの構築法』, 133-186, 国立国語研究所
小椋秀樹・山口昌也・西川賢哉・石塚京子・木村睦子(2004)「『日本語話し言葉コーパス』における単位認定基準について」『日本語科学』16, 93-113, 国書刊行会
窪園晴夫(1995)『語形成と音韻構造』くろしお出版
小磯花絵・西川賢哉・間淵洋子(2006)「転記テキスト」『国立国語研究所報告124 日本語話し言葉コーパスの構築法』, 23-132, 国立国語研究所
国立国語研究所(1962)『国立国語研究所報告21 現代雑誌九十種の用語用字(1)』秀英出版
国立国語研究所(1987)『国立国語研究所報告89 雑誌用語の変遷』秀英出版
国立国語研究所(1995)『国立国語研究所報告112 テレビ放送の語彙調査I』秀英出版
国立国語研究所 (編)(2004)『分類語彙表 増補改訂版』大日本図書
匂坂芳典・佐藤大和(1983)「日本語単語連鎖のアクセント規則」『電子情報通信学会論文誌』J66-D, 849-856
佐藤大和(1990)「複合語におけるアクセント規則と連濁規則」杉藤美代子 (編)『日本語と日本語教育第2巻：日本語の音声・音韻 (上)』, 233-265, 明治書院
新情報処理開発機構(RWCP)テキスト・サブ・ワーキンググループ(1998)『研究開発用知的資源：タグ付きテキストコーパス報告書』
高橋栄司・飯室美紀(2002)『基礎からのデータベース設計』ソフトバンクパブリッシング
伝康晴・浅原正幸(2001)「リレーショナル・データベースによる統合的言語資源管理環境」『第1回「話し言葉の科学と工学」ワークショップ講演予稿集』, 77-84
伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治(2002)「話し言葉研究に適した電子化辞書の設計」『第2回「話し言葉の科学と工学」ワークショップ講演予稿集』, 39-46
伝康晴・山田篤・小椋秀樹・小磯花絵・小木曽智信(2007)『unidic version 1.3.0ユーザーズマニュアル』, http://download.unidic.org/ よりダウンロード可能
中野洋(1998)「言語の統計」『岩波講座 言語の科学9 ：言語情報処理』, 149-199, 岩波書店
前川喜久雄(2004)「『日本語話し言葉コーパス』の概要」『日本語科学』15, 111-133, 国書刊行会
山崎誠・前川喜久雄・田中牧郎・小椋秀樹・柏野和佳子・小磯花絵・間淵洋子・丸山岳彦・山口昌也・秋元祐哉・稲益佐知子・吉田谷幸宏(2006)「代表性を有する現代日本語書き言葉コーパスの設計」『言語処理学会第12回年次大会発表論文集』, 440-443
Minematsu, N., Kita, R., & Hirose, K. (2003) Automatic Estimation of Accentual Attribute Values of Words for Accent Sandhi Rules of Japanese Text-to-Speech Conversion, *IEICE Transactions on Information and Systems*, E86-D, 550-557.
Uchimoto, K., Takaoka, K., Nobata, C., Yamada, A., Sekine, S., & Isahara, H. (2004) Morphological Analysis of the Corpus of Spontaneous Japanese, *IEEE Transactions on Speech and Audio Processing*, 12, 382-390.

謝 辞

本電子化辞書の開発には，情報処理振興事業協会「擬人化音声対話エージェント基本ソフトウェ

アの開発」プロジェクト（代表：東京大学・嵯峨山茂樹教授），情報処理学会「音声対話技術コンソーシアム」(ISTC)（代表：豊橋技術科学大学・新田恒雄教授），文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築」（領域代表：国立国語研究所・前川喜久雄グループ長）からの助成を得ています。また平成18年度からは，国立国語研究所の研究課題「大規模汎用日本語データベースの構築とその活用に関する調査研究」の下，同研究所研究開発部門言語資源グループと共同開発を行なっています。

（投稿受理日：2007年 1 月31日）
（最終原稿受理日：2007年 5 月21日）

伝　康晴（でん　やすはる）
千葉大学文学部
263-8522　千葉市稲毛区弥生町1-33
den@cogsci.L.chiba-u.ac.jp
小木曽　智信（おぎそ　としのぶ）
国立国語研究所研究開発部門
togiso@kokken.go.jp
小椋　秀樹（おぐら　ひでき）
国立国語研究所研究開発部門
ogura@kokken.go.jp
山田　篤（やまだ　あつし）
京都高度技術研究所研究開発部
yamada@astem.or.jp
峯松　信明（みねまつ　のぶあき）
東京大学大学院新領域創成科学研究科
mine@gavo.t.u-tokyo.ac.jp
内元　清貴（うちもと　きよたか）
情報通信研究機構情報通信部門
uchimoto@nict.go.jp
小磯　花絵（こいそ　はなえ）
国立国語研究所研究開発部門
koiso@kokken.go.jp

The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics

DEN Yasuharu
Chiba University

OGISO Toshinobu
The National Institute for Japanese Language

OGURA Hideki
The National Institute for Japanese Language

YAMADA Atsushi
ASTEM

MINEMATSU Nobuaki
University of Tokyo

UCHIMOTO Kiyotaka
National Institute of Information and Communications Technology

KOISO Hanae
The National Institute for Japanese Language

Keywords

electronic dictionary, morphological analysis, database system, uniformity of units, identity of indexes

Abstract

In this paper, we describe the design and the implementation of an electronic dictionary for morphological analysis, UniDic, which aims particularly at application to Japanese corpus linguistics. It has been indispensable for the development of a large-scale corpus to utilize an automatic morphological analyzer on computer. The existing dictionaries for morphological analyzers, however, reveal lots of problems when used in corpus linguistics, such as unevenness in defining a unit and failure in handling allomorphs and orthographic variants. Our dictionary, in contrast, deals with the uniformity of units and the identity of indexes, which are important requirements for linguistic analysis of corpora. We adopt multi-level definition of word units, consisting of short-, middle-, and long-unit words, and structured representation of indexes, composed of lemma, word form, orthography, and pronunciation. We develop a database system that straight-forwardly implements this design of the dictionary and a friendly user-interface for dictionary builders to be capable of searching and registering entries with grasping the complex structure of the indexes. We also show how this structured representation benefits us in analyzing morphologically annotated corpora, presenting case studies that investigate the variation of word form in spoken language corpus and the variation of orthography in written language corpus.