UniDic version 1.3.9 ユーザーズマニュアル

伝 康晴　　山田 篤　　小椋 秀樹　　小磯 花絵　　小木曽 智信

2008 年 7 月

UniDic version 1.3.9 Users Manual

Yasuharu Den, Atsushi Yamada, Hideki Ogura, Hanae Koiso, and Toshinobu Ogiso

version 1.3.0 2 April 2007
version 1.3.5 12 October 2007
version 1.3.8 25 April 2008
version 1.3.9 15 July 2008

目次

はじめに

UniDic は、形態素解析システム用の日本語辞書です。ChaSen 用（UniDic-chasen）と MeCab 用（UniDic-mecab）があります。ChaSen も MeCab も、言語処理のためのソフトウェアとしてフリーで公開され、広く用いられています。

ChaSen 用辞書としては、奈良先端科学技術大学院大学より公開されている IPA 体系にもとづく日本語辞書 ipadic がありますが、本辞書は、言語学・国語学や音声情報処理など、より多様な目的に適した体系にもとづくものです。具体的には、以下の特徴を持ちます。

- 国立国語研究所で規定した「短単位」という揺れがない斉一な単位で設計されています。
- 語彙素・語形・書字形・発音形の階層構造を持ち，表記の揺れや語形の変異にかかわらず同一の見出しを与えることができます。
- アクセントや音変化の情報を付与することができ、テキスト音声合成などに利用することができます。

UniDic 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）、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）、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：21 世紀の日本語研究の基盤整備」（平成 18〜22 年度、領域代表者：国立国語研究所・前川喜久雄）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」のもと、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。

本辞書に関するお問い合わせは以下にお願いします。

Tel: 042-540-4300（国立国語研究所代表）
E-mail: unidic@kokken.go.jp

第 I 部
実践編

1　インストール

UniDic のインストールおよび実行には、ChaSen（ver. 2.4.0 以降）と MeCab（ver. 0.96 以降）の一方もしくは両方が必要です。あらかじめインストールしておいてください。

1.1　パッケージ版のインストール

パッケージ版（Windows 用・Linux/Cygwin 用）をお使いの方は、パッケージをインストールすればそのまま『茶まめ』（Windows 用）・chauni（Linux/Cygwin 用）からお使いいただけます。

Linux/Cygwin 用パッケージでは、configure 時のオプションによって、ChaSen/MeCab 用辞書のいずれをインストールするかを選択できます。デフォルトでは両者ともインストールします。

```
./configure --with-use-mecab=0   # ChaSen 用辞書のみをインストール
./configure --with-use-chasen=0  # MeCab 用辞書のみをインストール
```

注意　Cygwin でインストールする際は、configure 時に Cygwin のトップディレクトリを指定してください。

```
./configure --with-systemtop=D:/Cygwin
```

注意　パッケージ版の辞書は文字コードが utf8 になっています。ChaSen を直接実行する場合は、-i オプションで文字コード w を指定してください。

```
chasen -i w < 入力ファイル
```

1.2　バイナリ辞書の個別インストール

バイナリ辞書を個別インストールする場合は以下の手順に従ってください。

1.2.1　UniDic-chasen のインストール

■Windows の場合

1. 圧縮ファイル unidic-chasen139_XXXX.zip を解凍する。unidic-chasen139_XXXX フォルダができる（XXXX は utf8, sjis, eucj のいずれか）。
2. ChaSen がインストールされているフォルダ（標準では C:\Program Files\ChaSen）にすでに dic フォルダがある場合は、あらかじめ削除（あるいは名前を変更）し、新たに空の dic フォルダを作成する。
3. 1 でできたフォルダの**中身**を 2 の dic フォルダにコピーする。

■Linux/Cygwin の場合

1. 圧縮ファイル unidic-chasen139_XXXX.tar.gz を解凍する。unidic-chasen139_XXXX ディレクト

リができる（XXXX は utf8, eucj, sjis のいずれか）。

2. ChaSen 辞書ディレクトリ（標準では/usr/local/lib/chasen/dic）[*1]にすでに unidic ディレクトリがある場合は、あらかじめ削除（あるいは名前を変更）し、新たに空の unidic ディレクトリを作成する。
3. 1 でできたディレクトリの**中身**を 2 の unidic ディレクトリにコピーする。
4. 3 でコピーした unidic ディレクトリ内の chasenrc ファイルを$HOME/.chasenrc にコピーし、冒頭の GRAMMAR ディレクトリの指定を適宜書き換える。

注意　Cygwin では、GRAMMAR ディレクトリの指定は Windows のパス名にする必要があります。

```
(GRAMMAR D:/Cygwin/usr/local/lib/chasen/dic/unidic)
```

1.2.2　UniDic-mecab **のインストール**

■Windows **の場合**

1. 圧縮ファイル unidic-mecab139_XXXX.zip を解凍する。unidic-mecab139_XXXX フォルダができる（XXXX は utf8, sjis, eucj のいずれか）。
2. MeCab がインストールされているフォルダ（標準では C:\Program Files\MeCab）の辞書フォルダ dic にすでに unidic フォルダがある場合は、あらかじめ削除（あるいは名前を変更）し、新たに空の unidic フォルダを作成する。
3. 1 でできたフォルダの**中身**を 2 の unidic フォルダにコピーする。

注意　MeCab の実行時に-d オプションによって辞書ディレクトリの位置を指定してください。

```
mecab -d "C:\Program Files\MeCab\dic\unidic" 入力ファイル
```

■Linux/Cygwin **の場合**

1. 圧縮ファイル unidic-mecab139_XXXX.tar.gz を解凍する。unidic-mecab139_XXXX ディレクトリができる（XXXX は utf8, eucj, sjis のいずれか）。
2. MeCab 辞書ディレクトリ（標準では/usr/local/lib/mecab/dic）[*2]にすでに unidic ディレクトリがある場合は、あらかじめ削除（あるいは名前を変更）し、新たに空の unidic ディレクトリを作成する。
3. 1 でできたディレクトリの**中身**を 2 の unidic ディレクトリにコピーする。

注意　MeCab の実行時に-d オプションによって辞書ディレクトリの位置を指定してください。

```
mecab -d /usr/local/lib/mecab/dic/unidic 入力ファイル
```

1.3　ソース辞書のインストール

ソース辞書をインストールする場合は以下の手順に従ってください。

[*1] ChaSen 辞書ディレクトリは chasen-config --dicdir コマンドで知ることができる。
[*2] MeCab 辞書ディレクトリは mecab-config --dicdir コマンドで知ることができる。

1.3.1 UniDic-chasen のインストール

■Windows の場合

1. 圧縮ファイル unidic-chasen139src.zip を解凍する。unidic-chasen139src フォルダができる。
2. ChaSen がインストールされているフォルダ（標準では C:\Program Files\ChaSen）にすでに dic フォルダがある場合は、あらかじめ削除（あるいは名前を変更）する。
3. 1 でできた unidic-chasen139src フォルダを ChaSen フォルダにコピーする。
4. 3 でコピーした unidic-chasen139src フォルダ内の Makefile.bat ファイルをダブルクリックし、実行する。辞書は、2 のフォルダの dic フォルダ内にインストールされる。

注意　ある品詞（たとえばフィラー）を辞書に含めたくない場合は、ステップ 4 を実行する前に、該当する辞書ファイル（たとえば Filler.dic）を.dic 以外の拡張子に改名してください。

注意　ソース辞書は文字コードが utf8 になっています。他の文字コードで辞書を作成したい場合は、すべてのソース辞書ファイル（2 節参照）の文字コードを変換し、Makefile_sjis.bat または Makefile_eucj.bat を実行してください。ただし、一部の文字は utf8 から変換できませんので、変換時にエラーになる語は適宜辞書ファイルから削除してください。

■Linux/Cygwin の場合

1. 圧縮ファイル unidic-chasen139src.tar.gz を解凍する。unidic-chasen139src ディレクトリができる。
2. 1 でできた unidic-chasen139src ディレクトリに移動し、./configure && make を実行する。
3. make install を実行する。辞書は、標準では/usr/local/lib/chasen/dic/unidic にインストールされる。
4. 3 でできた unidic ディレクトリ内の chasenrc ファイルを$HOME/.chasenrc にコピーし、冒頭の GRAMMAR ディレクトリの指定を適宜書き換える。

注意　ある品詞（たとえばフィラー）を辞書に含めたくない場合は、ステップ 2 で configure を実行する際に、該当する辞書ファイルを--with-exclude-dic オプションで指定してください（複数ある場合はコンマ（,）で区切って並べる）。

```
./configure --with-exclude-dic=Filler.dic
```

注意　Cygwin でインストールする際は、configure 時に Cygwin のトップディレクトリを指定してください。

```
./configure --with-systemtop=D:/Cygwin
```

また、GRAMMAR ディレクトリの指定は Windows のパス名にする必要があります。

```
(GRAMMAR D:/Cygwin/usr/local/lib/chasen/dic/unidic)
```

注意　ソース辞書は文字コードが utf8 になっています。他の文字コードで辞書を作成したい場合は、すべてのソース辞書ファイル（2 節参照）の文字コードを変換し、configure の際に、with-encoding=s（Shift-JIS の場合）または e（EUC-JP の場合））を指定し、make install してください。ただし、一部の文字は utf8

から変換できませんので、変換時にエラーになる語は適宜辞書ファイルから削除してください。

1.3.2 UniDic-mecab のインストール

■Windows の場合

1. 圧縮ファイル unidic-mecab139src.zip を解凍する。unidic-mecab139src フォルダができる。
2. MeCab がインストールされているフォルダ（標準では C:\Program Files\MeCab）の辞書フォルダ dic にすでに unidic フォルダがある場合は、あらかじめ削除（あるいは名前を変更）する。
3. 1 でできた unidic-mecab139src フォルダを MeCab フォルダにコピーする。
4. 3 でコピーした unidic-mecab139src フォルダ内の Makefile.bat ファイルをダブルクリックし、実行する。辞書は、2 のフォルダの unidic フォルダ内にインストールされる。

注意　ある品詞（たとえばフィラー）を辞書に含めたくない場合は、ステップ 4 を実行する前に、該当する辞書ファイル（たとえば Filler.csv）を.csv 以外の拡張子に改名してください。
注意　デフォルトでは文字コードが utf8 でバイナリ辞書が作成されます。他の文字コードで辞書を作成したい場合は、Makefile_sjis.bat または Makefile_eucj.bat を実行してください。ただし、一部の文字は utf8 から変換できず、バイナリ辞書からは削除されます。
注意　MeCab の実行時に-d オプションによって辞書ディレクトリの位置を指定してください。

```
mecab -d "C:\Program Files\MeCab\dic\unidic" 入力ファイル
```

■Linux/Cygwin の場合

1. 圧縮ファイル unidic-mecab139src.tar.gz を解凍する。unidic-mecab139src ディレクトリができる。
2. 1 でできた unidic-mecab139src ディレクトリに移動し、./configure && make を実行する。
3. make install を実行する。辞書は、標準では/usr/local/lib/mecab/dic/unidic にインストールされる。

注意　ある品詞（たとえばフィラー）を辞書に含めたくない場合は、ステップ 2 で configure を実行する際に、該当する辞書ファイルを--with-exclude-dic オプションで指定してください（複数ある場合はコンマ（,）で区切って並べる）。

```
./configure --with-exclude-dic=Filler.csv
```

注意　デフォルトでは文字コードが utf8 でバイナリ辞書が作成されます。他の文字コードで辞書を作成したい場合は、configure の際に、with-charset=sjis（Shift-JIS の場合）または euc-jp（EUC-JP の場合））を指定し、make install してください。ただし、一部の文字は utf8 から変換できず、バイナリ辞書からは削除されます。
注意　MeCab の実行時に-d オプションによって辞書ディレクトリの位置を指定してください。

```
mecab -d /usr/local/lib/mecab/dic/unidic 入力ファイル
```

2 UniDic-chasen のファイル群

2.1 品詞定義ファイル

grammar.cha には、品詞のリストが記述されている。活用のある品詞の場合、品詞名の末尾に % をつける。活用のある品詞では、ctypes.cha に可能な活用型を、cforms.cha に可能な活用形を記述する必要がある。

```
(助詞
        (係助詞)
        (副助詞)
        (接続助詞)
        (格助詞)
        (準体助詞)
        (終助詞))
(動詞
        (一般 %)
        (非自立可能 %))
```

2.2 活用型定義ファイル

ctypes.cha には、活用のある品詞がどのような活用型を取るかが記述されている。

```
((動詞 一般)
        (カ行変格
         サ行変格
         ザ行変格
         上一段-ア行
         ...
         文語四段-ラ行))
```

2.3 活用形定義ファイル

cforms.cha には、各活用型がどのような活用形を取るかが記述されている。

```
(上一段-ア行
        ((仮定形-一般    *)
         (仮定形-融合    *)
         (命令形         *)
         (意志推量形     *)
         (未然形-一般    *)
         (終止形-一般    *)
         (終止形-撥音便  *)
         (連体形-一般    *)
         (連体形-撥音便  *)
         (連体形-省略    *)
         (連用形-一般    *)))
```

ChaSen 標準の ipadic とは異なり、UniDic-chasen では、cforms.cha には活用形の名称のみが定義されており、活用語尾に関する記述はない。すべての語の活用形は語彙定義ファイル中に明示的に記載されている。

2.4 語彙定義ファイル

.dic の拡張子を持つファイルには、単語の一覧が記述されている。品詞ごとにファイルに分かれている。

```
(POS (助詞 係助詞))
  ((LEX (は 0)) (READING ハ) (PRON ワ)
   (INFO 'orthBase="は" kanaBase="ハ" pronBase="ワ"
         lForm="ハ" lemma="は" form="ハ"
         aConType="動詞%F2@0,名詞%F1,形容詞%F2@-1" goshu="和"'))

(POS (名詞 普通名詞 一般))
  ((LEX (ねこ 4000)) (READING ネコ) (PRON ネコ)
   (INFO 'orthBase="ねこ" kanaBase="ネコ" pronBase="ネコ"
         lForm="ネコ" lemma="猫" form="ネコ"
         aType="1" aConType="C3" goshu="和"'))
```

活用のある語では、すべての活用形が記載されている。これにより、活用形ごとに異なる属性を与えることができる（下例の aModType 参照）。

```
(POS (動詞 一般))
  ((LEX (起きれ 261)) (READING オキレ) (PRON オキレ)
   (CTYPE 上一段-カ行) (CFORM 仮定形-一般) (BASE 起きる)
   (INFO 'orthBase="起きる" kanaBase="オキル" pronBase="オキル"
         lForm="オキル" lemma="起きる" form="オキル"
         aType="2" aConType="C1" goshu="和"'))
(POS (動詞 一般))
  ((LEX (起きりゃ 261)) (READING オキリャ) (PRON オキリャ)
   (CTYPE 上一段-カ行) (CFORM 仮定形-融合) (BASE 起きる)
   (INFO 'orthBase="起きる" kanaBase="オキル" pronBase="オキル"
         lForm="オキル" lemma="起きる" form="オキル"
         aType="2" aConType="C1" goshu="和"'))
(POS (動詞 一般))
  ((LEX (起きよ 261)) (READING オキヨ) (PRON オキヨ)
   (CTYPE 上一段-カ行) (CFORM 命令形) (BASE 起きる)
   (INFO 'orthBase="起きる" kanaBase="オキル" pronBase="オキル"
         lForm="オキル" lemma="起きる" form="オキル"
         aType="2" aConType="C1" aModType="M2@1" goshu="和"'))
...
(POS (動詞 一般))
  ((LEX (起き 261)) (READING オキ) (PRON オキ)
   (CTYPE 上一段-カ行) (CFORM 連用形-一般) (BASE 起きる)
   (INFO 'orthBase="起きる" kanaBase="オキル" pronBase="オキル"
         lForm="オキル" lemma="起きる" form="オキル"
         aType="2" aConType="C1" aModType="M4@1" goshu="和"'))
```

UniDic-chasen の語彙定義にはさまざまな属性が含まれ、標準の ChaSen で用意されている属性スロットでは記述しきれない。そのため、標準的な属性以外は、属性・値対の形式で INFO スロットにまとめて記載している。各属性の意味は、4.4 節を参照。

2.5 連接規則ファイル

connect.cha には、連接規則が記述されている。連接規則とは、ある要素とある要素がどれくらいつながりやすいかを規定したものである。

```
(( (((名詞 普通名詞 一般)))
   (((接尾辞 名詞的 一般))) )
 814)

(( (((動詞 一般) 上一段-ア行 未然形-一般))
   (((助動詞) 助動詞-ナイ 終止形-一般 ない))) )
 147)

(( (((*)))
   (((助動詞) 助動詞-ダ 連用形-一般 で))) )
 8000)
(( (((助詞 準体助詞) * * の))
   (((助動詞) 助動詞-ダ 連用形-一般 で))) )
 425)
```

2.6 chasenrc ファイル

chasenrc には、ChaSen の実行に必要なさまざまなオプションが定義されている。

```
(GRAMMAR /usr/local/lib/chasen/dic)
(DADIC chadic)

(UNKNOWN_POS (名詞 普通名詞 一般))

(OUTPUT_FORMAT  ; 以下、実際には 1 行
  "<cha:W1 orth=\"%m\" kana=\"%?U/%m/%y/\" pron=\"%?U/%m/%a/\"
           pos=\"%U(%P-)\"%?T/ cType=\"%T \"//%?F/ cForm=\"%F \"//
           %?I/ %i//>%m</cha:W1>\n")

(OUTPUT_COMPOUND "SEG")

(EOS_STRING "")

(DEF_CONN_COST 10000)
(POS_COST
  ((*)       1)
  ((UNKNOWN) 30000) )

(CONN_WEIGHT  1)
(MORPH_WEIGHT 1)
(COST_WIDTH   0)

(ANNOTATION (("<" ">") "%m\n"))
```

おもなオプションを以下に示す。

GRAMMAR　辞書がインストールされているディレクトリを指定する。
UNKNOWN_POS　未知語をどのような品詞として扱って連接規則を適用するかを指定する。
OUTPUT_FORMAT　出力フォーマットを指定することにより、解析結果の出力形式を変えることができる。
EOS_STRING　解析結果の文末に表示する文字列を指定する。
ANNOTATION　ある文字列で始まり、ある文字列で終わる一連の文字列を注釈として扱い、その文字列の部分を解析時に無視させることができる。上例では、xml タグ（‘<’ と ‘>’ で囲まれた部分）をそのまま出力するよう指定している。

UniDic-chasen では、標準の ChaSen で用意されているよりもはるかに多くの属性が定義されている。そこで、出力には xml 形式を用い、これらの属性を属性・値対の形式で出力することを推奨している。OUTPUT_FORMAT を上例のように指定することで、以下のような出力が得られる。

```
;以下、実際の <cha:W1>...</cha:W1>は 1 行
<cha:W1 orth="解析" kana="カイセキ" pron="カイセキ"
        pos="名詞-普通名詞-サ変可能"
        orthBase="解析" kanaBase="カイセキ" pronBase="カイセキ"
        lForm="カイセキ" lemma="解析" form="カイセキ"
        aType="0" aConType="C2" goshu="漢">解析</cha:W1>
<cha:W1 orth="結果" kana="ケッカ" pron="ケッカ"
        pos="名詞-普通名詞-副詞可能"
        orthBase="結果" kanaBase="ケッカ" pronBase="ケッカ"
        lForm="ケッカ" lemma="結果" form="ケッカ"
        aType="0,1" aConType="C2" goshu="漢">結果</cha:W1>
<cha:W1 orth="を" kana="ヲ" pron="オ"
        pos="助詞-格助詞"
        orthBase="を" kanaBase="ヲ" pronBase="オ"
        lForm="ヲ" lemma="を" form="ヲ"
        aConType="動詞%F2@0,名詞%F1,形容詞%F2@-1" goshu="和">を</cha:W1>
<cha:W1 orth="表示" kana="ヒョウジ" pron="ヒョージ"
        pos="名詞-普通名詞-サ変可能"
        orthBase="表示" kanaBase="ヒョウジ" pronBase="ヒョージ"
        lForm="ヒョウジ" lemma="表示" form="ヒョウジ"
        aType="0,1" aConType="C2" goshu="漢">表示</cha:W1>
<cha:W1 orth="し" kana="シ" pron="シ"
        pos="動詞-非自立可能" cType="サ行変格" cForm="連用形-一般"
        orthBase="する" kanaBase="スル" pronBase="スル"
        lForm="スル" lemma="為る" form="スル"
        aType="0" aConType="C3" goshu="和">し</cha:W1>
<cha:W1 orth="ます" kana="マス" pron="マス"
        pos="助動詞" cType="助動詞-マス" cForm="終止形-一般"
        orthBase="ます" kanaBase="マス" pronBase="マス"
        lForm="マス" lemma="ます" form="マス"
        aConType="動詞%F4@1" goshu="和">ます</cha:W1>
```

ここから、xslt スクリプトなどによって必要な属性のみを取り出せばよい。uniutils モジュールの xml2txt.xsl にサンプルがあるので、参照のこと。

3 UniDic-mecab のファイル群

3.1 語彙定義ファイル

.csv の拡張子を持つファイルには、単語の一覧が記述されている。品詞ごとにファイルが分かれている。これらは、UniDic-chasen の.dic ファイルに相当する。

3.2 その他の定義ファイル

.def の拡張子を持つファイルには、連接規則など、各種のモデル定義情報が記述されている。詳しくは、http://mecab.sourceforge.net/の解説を参照。

3.3 dicrc ファイル

dicrc ファイルには、MeCab の実行に必要なさまざまなオプションが定義されている。

```
cost-factor = 700
bos-feature = BOS/EOS,*,*,*,*,*,*,*,*,*,*,*,*
eval-size = 9
unk-eval-size = 4
max-grouping-size = 10

output-format-type = unidic

node-format-unidic = %m\t%f[10]\t%f[6]\t%f[7]\t%F-[0,1,2,3]\t%f[4]\t%f[5]\t%f[12]\n
unk-format-unidic  = %m\t%m\t%m\t%m\t%F-[0,1,2,3]\t%f[4]\t%f[5]\t%f[12]\n
eos-format-unidic  = EOS\n
```

出力フォーマットを変更する場合は、output-format-type に任意のフォーマット名を記載し、node-format-XXX, unk-format-XXX, bos-format-XXX, eos-format-XXX でそれぞれ単語・未知語・文頭・文末の出力形式を指定する（XXX はフォーマット名）。なお、f[0]〜f[12] の各属性の意味については、MeCab 辞書ディレクトリ内の rewrite.def ファイルの冒頭にあるコメントを参照のこと。

dicrc のその他のオプションについては、http://mecab.sourceforge.net/の解説を参照。

第 II 部
解説編

4　UniDic の概要

UniDic は、既存の形態素解析辞書と比べて、言語学・国語学や音声情報処理など、より多様な目的に適した体系にもとづいている。以下、単位設計・見出し設計・音韻論情報という 3 点から説明する。詳細は、伝ほか「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」(『日本語科学』 22, 2007 年 10 月）を参照のこと。

4.1　単位設計

既存の形態素解析辞書では、何をもって 1 語とするかの基準が曖昧であり、長い語や短い語が混在していた。このような単位の不ぞろいは、とくに言語学・国語学への応用において不都合である。

この問題を解決するため、UniDic では、国立国語研究所で規定した「短単位」という揺れがない斉一な単位を採用している。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 個を 1 回結合したものである。たとえば、「母親」「食べ歩く」「音声」「無口」などが該当する。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。短単位の概要は、小椋ほか「『現代日本語書き言葉均衡コーパス』における短単位の概要」(『特定領域「日本語コーパス」平成 18 年度公開ワークショップ（研究成果報告会）予稿集』, pp. 101–108, 2007 年 3 月）を参照のこと。

なお、UniDic で採用している短単位と国立国語研究所の短単位には、ごく微細な違いがある。国立国語研究所の短単位では、意志・推量の助動詞「う」「よう」を独立した語として扱っているが、UniDic では、これらを活用語尾として扱っている。そのため、「意志推量形」という特別な活用形を設けている（5.2 節参照）。これは、「う」の発音形が常に長音であること、話し言葉で「う」が省略される場合があることなどを考慮し、より扱いやすくするためである。

4.2　見出し設計

既存の形態素解析辞書では、テキストに出現する形（活用語ではその終止形）をもって見出しとしていた。すなわち、「あらわす」は「表わす」の意味でも「著わす」の意味でも同じ見出しであり、その一方で、「表わす」と「表す」は別の見出しであった。さらに、「カナ（仮名）」もしくは「カメイ（仮名）」のように明らかに別語の場合でも、漢字表記が同じであれば同じ見出しが与えられていた。このような表記にもとづく見出しは、言語学・国語学への応用において不都合であると同時に、情報検索などへの応用においても不都合である。

この問題を解決するため、UniDic では、語彙素・語形・書字形・発音形という 4 階層からなる階層的見出しを採用している。図 1 の例によって説明する。語彙素は、「オオキイ」「オッキイ」のような語形の変異や「表わす」「表す」「あらわす」のような表記の揺れを考慮せず、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である。国語辞典の見出しに相当する。語形は、同じ語彙素に所属するものに対して、「オオキイ」「オッキイ」のような語形の変異を区別したものである。少し大きな国語辞典では、このレベルの見出しを掲げ、一方を他方への参照見出しとしている場合が多い。書字形は、同じ語形に所属するものに対して、「表わ

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 1　階層的見出しの例

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

図 2　書字形を左端に配置して図 1 を書き直したもの

す」「表す」「あらわす」のような表記の揺れを区別したものである。これらはテキストに出現する形である。発音形は、同じ語形に所属するものに対して、「データ」「データー」のような発音の揺れを区別したものである。これらは音声データに出現する形である。

図 1 は、辞書の設計という観点から階層的見出しを例示したものである。これに対して、形態素解析では、テキストに出現する形（書字形）から見出しを同定しなければならない。この点を理解するために、書字形を左端に配置して図 1 を書き直したものが図 2 である。「大きい」「おおきい」「おっきい」に対して同一の語彙素が得られること、「大きい」「おおきい」に対してはさらに同一の語形も得られることが見て取れる。また、「表わす」「表す」に対しても同一の語彙素・語形が得られる。一方、「あらわす」や「仮名」では、語彙素の曖昧性が生じる。UniDic を用いた形態素解析では、このような曖昧性に対しても対処しなければならない。これは従来の日本語形態素解析が対象としていなかった新しい問題であり、現状ではまだ高い精度での語彙素同定は実現できていない。今後この処理の高精度化が課題である。

4.3 音韻論情報

従来の日本語形態素解析で扱われていなかったもう 1 つの問題として、複合語に対して正しい発音やアクセントを与えるという問題がある。たとえば、「イチ（一）」と「ホン（本）」が複合すると、「イッポン」になり、「サン（三）」と「ホン（本）」が複合すると、「サンボン」になるといった音の変化がある。また、アクセントに関しても、頭高型の「シャ⌉カイ（社会）」と平板型の「セーカツ（生活）」が複合して「社会生活」になると、4 モーラ目に核を持つ中高型（「シャカイセ⌉ーカツ」）になるといった変化がある。これらは、テキスト音声合成などへの応用においてとくに重要な問題である。

UniDic では、これらの問題に対処するために、さまざまな音韻論情報を記述している。まず、「ホン（本）」の語頭音が「ボ」や「ポ」に変わる現象を扱うため、語頭変化型・語頭変化形という属性を設けている。語頭変化型は語頭音の変化のパターンを記したもの（たとえば「ホ混合」型）であり、語頭変化形は特定の語における変化の形を指定したもの（たとえば「濁音形」）である。同様に、「イチ（一）」の語末音が「ッ」に変わる現象を扱うため、語末変化型・語末変化形という属性を設けている。語末変化型は語末音の変化のパターンを記したもの（たとえば「チ促」型）であり、語末変化形は特定の語における変化の形を指定したもの（たとえば「促音形」）である。さらに、語頭変化形の決定に際しては前接要素が、語末変化形の決定に際しては後続要素が何であるかが影響を与える。たとえば、「ホン（本）」が濁音形を取るのは、前接要素が「三」の場合であり、「イチ（一）」が促音形を取るのは、後続要素が「本」「階」「杯」などの場合である。このような隣接要素に対する影響力を記すために、語頭変化結合型・語末変化結合型という属性を設けている。これらの音変化に関わる属性をもとにして、複合語に対する正しい発音が chaone モジュールによって選択される。

アクセントの扱いもこれと類似である。まず、各語を単独で発声した場合のアクセント型を記述している。この単独発声のアクセント型が変化する状況は 3 通りある。1 つめは活用のある語が特定の活用形を取る場合の変化（たとえば「オキ⌉ル」→「オ⌉キ（タ）」）であり、2 つめは複合語を作る際の変化（たとえば「シャ⌉カイ + セーカツ」→「シャカイセ⌉ーカツ」）であり、3 つめは自立語に助詞・助動詞が結合する際の変化（たとえば「ハナ⌉シ」→「マス」→「ハナシマ⌉ス」）である。最初の変化を扱うために、アクセント修飾型という属性を設け、アクセント変化を生じる活用形に対してその変化パターンを記す。残りの 2 種類の変化を扱うために、アクセント結合型という属性を設け、後部要素（接頭辞においては前部要素である接頭辞）が全体のアクセント型にどのような変化をもたらすかのパターンを記す。これらのアクセント変化に関わる属性をもとにして、活用形・複合語・アクセント句に対する正しいアクセントが chaone モジュールによって付与される。

表 1　UniDic で用いられる属性

階層	属性の名称	属性ラベル	説明
語彙素	語彙素読み	lForm	語彙素見出し（カタカナ表記）
	語彙素表記	lemma	語彙素見出し（漢字仮名混じり表記）
	語種	goshu	語種の名称
語形	語形基本形	form	語形見出し
	品詞	pos	品詞の名称
	活用型	cType	活用の種類（型）
	活用形	cForm	活用の形
	語頭変化型	iType	語頭音変化の種類（型）
	語頭変化形	iForm	語頭音変化の形
	語頭変化結合型	iConType	後続要素の語頭変化形への制約の種類（型）
	語末変化型	fType	語末音変化の種類（型）
	語末変化形	fForm	語末音変化の形
	語末変化結合型	fConType	前接要素の語末変化形への制約の種類（型）
書字形	書字形基本形	orthBase	書字形見出し
	書字形出現形	orth	書字形基本形が活用変化を受けたもの
	仮名形基本形	kanaBase	書字形基本形をカタカナ表記にしたもの
	仮名形出現形	kana	書字形出現形をカタカナ表記にしたもの
発音形	発音形基本形	pronBase	発音形見出し
	発音形出現形	pron	発音形基本形が活用変化を受けたもの
	アクセント型	aType	アクセント核の位置
	アクセント修飾型	aModType	活用によるアクセント変化の種類（型）
	アクセント結合型	aConType	前接（後続）要素との結合時のアクセント変化の種類（型）

4.4　属性一覧

以上、さまざまな属性をまとめると表 1 のようになる。「属性の名称」は、これまで本文中で説明してきた属性名であり、「属性ラベル」は、ChaSen 版で、添付の chasenrc を用いて形態素解析した結果を xml 形式で出力した場合の属性ラベルである。これらのうち、「書字形出現形」「仮名形出現形」「発音形出現形」「品詞」「活用型」「活用形」は通常の ChaSen 辞書の属性スロット（それぞれ LEX, READING, PRON, POS, CTYPE, CFORM）に記述され、残りは属性・値対の形式で INFO スロットに記述されている（2.4 節の例を参照）。MeCab 版では、属性名なしでコンマ区切りで出力される。

5　品詞体系

UniDic の品詞・活用型・活用形は概ね学校文法に準拠している。これは、多様な目的に供するために、なるべく多くの者に受け入れられる標準的な品詞体系を採用するのが望ましいと考えたからである。しかし、学校文法の品詞・活用型・活用形は、とくに自然言語処理などの工学的応用の上では分類が粗すぎることがある。そこで、UniDic では、ipadic で採用されている IPA 体系を参考にしつつ、学校文法の品詞・活用型・活用形を細分化して階層的な品詞体系を設計した。階層の上部のみを取り出すと、「動詞」「下一段-ア行」「連用形」といった学校文法的な分類が得られる。

表 2　品詞分類

大分類	中分類	小分類	細分類
名詞	普通名詞	一般 サ変可能 形状詞可能 サ変形状詞可能 副詞可能	
	固有名詞	一般	
		人名	一般 姓 名
		地名	一般 国
		組織名	
	数詞		
	助動詞語幹		
代名詞			
形状詞	一般 タリ 助動詞語幹		
連体詞			
副詞			
接続詞			
感動詞	一般 フィラー		
動詞	一般 非自立可能		
形容詞	一般 非自立可能		

大分類	中分類	小分類	細分類
助動詞			
助詞	格助詞 副助詞 係助詞 接続助詞 終助詞 準体助詞		
接頭辞			
接尾辞	名詞的	一般 サ変可能 形状詞可能 副詞可能 助数詞	
	形状詞的		
	動詞的		
	形容詞的		
記号	一般 文字		
補助記号	一般 句点 読点 括弧開 括弧閉		
空白			

5.1　品詞

品詞の一覧を表 2 に示す。以下、注意が必要なものを説明する。

■名詞-普通名詞-{ サ変可能, 形状詞可能, サ変形状詞可能 }　普通名詞のうち、「運動（する）」のように形式的な意味の「する」「できる」などが直接続き、動詞として用いられることがあるもの、「安全（な）」のように「な」（助動詞「だ」の連体形）が直接続き、形容動詞として用いられることがあるもの、「心配（する・な）」のように両者が可能なものをそれぞれ、「名詞-普通名詞-サ変可能」「名詞-普通名詞-形状詞可能」「名詞-普通名詞-サ変形状詞可能」に分類する。これらは、あくまでも「用いられる可能性がある」ことを示すものであり、特定の文脈で実際にサ変動詞や形容動詞として用いられているか否かには関わらない。

■名詞-普通名詞-副詞可能　普通名詞のうち、「大体」「近々」のように助詞を伴なわずに連用修飾語になるものを「名詞-普通名詞-副詞可能」に分類する。これには、「頃」「時」のように句や節による連体修飾を受けて

連用修飾節になるものも含む[*3]。これも可能性を示すものであって、実際に助詞を伴なわずに連用修飾しているか否かに関わらない。

■名詞-固有名詞-一般　人名・地名・組織名以外の固有名詞であり、「平成」のような元号や「ウィンドウズ」のような商品名などが該当する。

■名詞-助動詞語幹　いわゆる伝聞の助動詞「そうだ」の語幹部分「そう」。

■形状詞-一般　「静か」「健やか」など、いわゆる形容動詞の語幹部分。ただし、名詞としての用法があるものは、「名詞-普通名詞-形状詞可能」に分類する。

■形状詞-タリ　「釈然」「錚々」など、いわゆるタリ活用の形容動詞の語幹部分。

■形状詞-助動詞語幹　一般に助動詞とされる「そうだ（様態）」「ようだ」「みたいだ」の語幹部分。

■動詞-非自立可能　動詞のうち、「する」「できる」のように「名詞-普通名詞-サ変可能」に直接続くことがあるものや「始める」「くる」のように動詞連用形（+ 接続助詞「て」）に接続して補助動詞として用いられることがあるものを、「動詞-非自立可能」に分類する。可能性を示すものであって、実際にサ変動詞や補助動詞として使われているか否かに関わらない。

■形容詞-非自立可能　形容詞のうち、「ない」「よい」のように形容詞・形容詞活用型助動詞の連用形や「欲しい」のように動詞・動詞活用型助動詞の連用形 + 接続助詞「て」に接続して補助形容詞として用いられることがあるものを、「形容詞-非自立可能」に分類する。可能性を示すものであって、実際に補助形容詞として使われているか否かに関わらない。

■接尾辞-名詞的-{ サ変可能, 形状詞可能, 副詞可能 }　名詞に接続する接尾辞のうち、「（活性）化」「（東洋）風」「（仕事）中」のように作られた複合名詞がサ変可能・形状詞可能・副詞可能であるものをそれぞれ、「接尾辞-名詞的-サ変可能」「接尾辞-名詞的-形状詞可能」「接尾辞-名詞的-副詞可能」に分類する。

■接尾辞-形状詞的　名詞に接続する接尾辞のうち、「（健康）的」「（自慢）気」のように形状詞を作るもの。

■接尾辞-動詞的　名詞に接続する接尾辞のうち、「（汗）ばむ」「（大人）ぶる」のように動詞を作るもの。

■接尾辞-形容詞的　名詞に接続する接尾辞のうち、「（安）っぽい」「（書き）易い」のように形容詞を作るもの。

■記号-{ 文字, 一般 }　記号のうち、「A」「α」「Σ」のようにアルファベットやギリシャ文字は「記号-文字」に、音階を表わす「ド」「ミ」「ソ」などは「記号-一般」に分類する。

■補助記号-{ 句点, 読点, 括弧開, 括弧閉, 一般 }　補助記号のうち、「。」「.」「！」などは「補助記号-句点」に、「、」「,」などは「補助記号-読点」に、「(」「《」「「」などは「補助記号-括弧開」に、「)」「》」「」」などは「補助記号-括弧閉」に、「・」「※」「’」などは「補助記号-一般」に分類する。

■空白　行頭の字下げなどの全角空白「　」。

[*3] 後者は前者とは意味合いが異なるため、将来的には別の品詞にする予定である。

5.2 活用型

5.2.1 動詞の活用型

動詞の活用型の一覧を表 3 に示す。以下、注意が必要なものを説明する。

■五段-カ行-イク　カ行五段活用動詞のうち、「イク（行く）」の活用型。助動詞「た」・接続助詞「て」が接続する場合、促音便になる。

■五段-カ行-ユク　カ行五段活用動詞のうち、「ユク（行く）」の活用型。連用形に音便形がない。

■五段-ラ行-アル　ラ行五段活用動詞のうち、「いらっしゃる」「おっしゃる」「くださる」「ござる」「なさる」の活用型。助動詞「ます」が接続する場合、イ音便になる。また、命令形の活用語尾が「い」となる。

■五段-ワア行-イウ　ワア行五段活用動詞のうち、「イウ（言う）」の活用型。連用形・終止形・連体形の発音形が「イー（マス）」「ユー」のように長音化する。

■五段-ワア行-ユウ　ワア行五段活用動詞のうち、「ユウ（言う）」の活用型。終止形・連体形の発音形が「ユー」のように長音化する。

■下一段-ラ行-呉レル　ラ行下一段活用動詞のうち、「呉れる」の活用型。命令形に「ーれろ」「ーれよ」のほか「ーれ」の形がある。

■ザ行変格　サ行変格活用動詞のうち、終止形が「ーずる」の形のもの。

■文語四段-ハ行+{ う, ふ }　文語ハ行四段活用動詞のうち、書字形が「買う」のように新仮名遣いで記されたものを「文語四段-ハ行+う」に、「買ふ」のように旧仮名遣いで記されたものを「文語四段-ハ行+ふ」に分類する（「う」「ふ」の前の「+」に注意）。

■文語下二段-ダ行+{ ず, づ }　文語ダ行下二段活用動詞のうち、書字形が「出ず」のように新仮名遣いで記されたものを「文語下二段-ダ行+ず」に、「出づ」のように旧仮名遣いで記されたものを「文語下二段-ダ行+づ」に分類する（「ず」「づ」の前の「+」に注意）。

■文語ザ行変格　文語サ行変格活用動詞のうち、終止形が「ーず」の形のもの。

5.2.2 形容詞の活用型

形容詞の活用型の一覧を表 4 に示す。以下、注意が必要なものを説明する。

■文語形容詞-多シ　文語形容詞「多し」の活用型。終止形に「多し」「多かり」の 2 つの語形がある。

■無変化型　「良か」など、「ーい」の形を取らない、おもに標準語以外の方言の形容詞は、テキスト中には通常終止形しか出現しないので、活用型は「無変化型」とする。

5.2.3 助動詞の活用型

接続助詞「て」+ 補助動詞「おく」が融合した「とく」など、動詞・形容詞由来の助動詞については、動詞・形容詞型の活用型に分類する。その他の助動詞については、語形ごとに個別に活用型を分類する。

表 3　活用型分類（動詞）

大分類	行分類	小分類
五段	カ行	一般 イク ユク
	ガ行	
	サ行	
	タ行	
	ナ行	
	バ行	
	マ行	
	ラ行	一般 アル
	ワア行	一般 イウ ユウ
上一段	ア行 カ行 ガ行 ザ行 タ行 ナ行 ハ行 バ行 マ行 ラ行	
下一段	ア行	
	カ行	
	ガ行	
	サ行	
	ザ行	
	タ行	
	ダ行	
	ナ行	
	ハ行	
	バ行	
	マ行	
	ラ行	一般 呉レル
カ行変格		
サ行変格		
ザ行変格		

大分類	行分類	小分類	書字形分類
文語四段	カ行		
	ガ行		
	サ行		
	タ行		
	ハ行		う ふ
	バ行		
	マ行		
	ラ行		
文語上二段	カ行 ガ行 タ行 ダ行 ハ行 バ行 マ行 ヤ行 ラ行		
文語下二段	ア行		
	カ行		
	ガ行		
	サ行		
	ザ行		
	タ行		
	ダ行		ず づ
	ナ行		
	ハ行		
	バ行		
	マ行		
	ヤ行		
	ラ行		
	ワ行		
文語カ行変格			
文語サ行変格			
文語ザ行変格			
文語ナ行変格			
文語ラ行変格			

表 4　活用型分類（形容詞）

大分類	小分類
形容詞	
文語形容詞	ク シク 多シ
無変化型	

5.2.4　接尾辞の活用型

「接尾辞-動詞的」「接尾辞-形容詞的」はそれぞれ、動詞型・形容詞型の活用型に分類する。

5.3　活用形

活用形の一覧を表 5 に示す。以下、注意が必要なものを説明する。

■語幹-サ　形容詞のうち、「無い」「良い」の語幹。いわゆる様態の助動詞「そうだ」が接続する場合、「無さ（そうだ）」「良さ（そうだ）」のように語幹に「さ」が接続する。

■未然形-{ サ, セ }　サ行・ザ行変格活用動詞の未然形のうち、助動詞「せる」「れる」に続く形（「さ（せる）」）を「未然形-サ」、助動詞「ず」に続く形（「せ（ず）」）を「未然形-セ」に分類する。

■未然形-撥音便　「分かん（ない）」など、ラ行五段活用動詞の一部に、未然形語尾が撥音便になるものがある。

■意志推量形　UniDic では、いわゆる意志・推量の助動詞「う」「よう」を立てず、動詞・形容詞・助動詞の活用形として意志推量形を設けている。「行こう」「食べよう」「高かろう」「でしょう」など。なお、語末が促音化したり（「行こっ」）、省略されたり（「行こ」）したものもこの活用形に分類する。

■連用形-ト　文語助動詞「たり」の連用形に「と」の形がある。

■連用形-ニ　助動詞「だ」の連用形に「に」の形がある。

表 5　活用形分類

大分類	小分類
語幹	一般 サ
未然形	一般 サ セ 撥音便 補助
意志推量形	

大分類	小分類
連用形	一般 ト ニ イ音便 ウ音便 促音便 撥音便 融合 省略 補助

大分類	小分類
終止形	一般 ウ音便 促音便 撥音便 融合 補助

大分類	小分類
連体形	一般 ウ音便 撥音便 省略 補助
仮定形	一般 融合
已然形	
命令形	

■連用形-融合　助動詞「だ」の連用形と後続する係助詞「は」が融合した「(それ）じゃ（ない)」の形がある。

■終止形-ウ音便　文語ハ行四段活用動詞「給う」に、終止形ウ音便（「たもう」）がある。

■終止形-促音便　「高っ」「安っ」など、形容詞の終止形で末尾の「い」が促音化することがある。

■終止形-撥音便　助動詞「ず」などに、終止形撥音便（「(ありませ）ん」）がある。また、一段動詞などの終止形で、終助詞「な」や助動詞「ねん」に連接する際に、撥音便化することがある（「(来）てん（な)」)。

■終止形-融合　助動詞「だ」の終止形に、前接する「と」の音と融合した「(何のこっ）ちゃ」の形がある。

■終止形-補助　文語形容詞「多し」の終止形に「多かり」の形がある。

■連体形-ウ音便　文語ハ行四段活用動詞「給う」に、連体形ウ音便（「たもう」）がある。

■連体形-撥音便　「食べん（ので)」など、動詞連体形が準体助詞や終助詞の「の」に連接する際に、撥音便化することがある。

■連体形-省略　「落ち（んです)」など、動詞連体形が準体助詞「ん」に連接する際に、活用語尾が省略されることがある。

■仮定形-融合　形容詞や助動詞「たい」「ない」「らしい」「ぬ」の仮定形が接続助詞「ば」と連接するときに、「ーけりゃ」「ーきゃ」や「にゃ」の形で融合することがある。

6　音韻論情報

UniDic では、音韻論情報として、音変化やアクセント変化に関わる属性を記述している。以下、これらを簡単に説明する。

6.1　語頭変化型・語頭変化形

語頭変化型と対応する語頭変化形の一覧を例とともに表 6 に示す。

6.2　語末変化型・語末変化形

語末変化型と対応する語末変化形の一覧を例とともに表 7 に示す。

6.3　語頭変化結合型

語の複合に際して、後部要素の語頭音がどの形を取るかは、前部要素の種類によって制約される。このような制約を記述した語頭変化結合型の例を表 8 に示す。各語頭変化結合型と各語頭変化型とが交差するセルには、その組み合わせのときに、どの語頭変化形を取るかが記載されている。

表 6　語頭変化型・変化形分類

語頭変化型	語頭変化形	例
カ濁	基本形（カ） 濁音形（ガ）	カイ（階）
キ濁	基本形（キ） 濁音形（ギ）	キライ（嫌い）
ク濁	基本形（ク） 濁音形（グ）	クルシイ（苦しい）
ケ濁	基本形（ケ） 濁音形（ゲ）	ケン（軒）
コ濁	基本形（コ） 濁音形（ゴ）	コク（国）
サ濁	基本形（サ） 濁音形（ザ）	サン（山）
シ濁	基本形（シ） 濁音形（ジ）	ショ（所）
ス濁	基本形（ス） 濁音形（ズ）	スキ（好き）
セ濁	基本形（セ） 濁音形（ゼ）	セメ（攻め）
ソ濁	基本形（ソ） 濁音形（ゾ）	ソク（足）
タ濁	基本形（タ） 濁音形（ダ）	タメシ（試し）
チ濁	基本形（チ） 濁音形（ヂ）	チョウシ（調子）
ツ濁	基本形（ツ） 濁音形（ヅ）	ツキアイ（付き合い）
テ濁	基本形（テ） 濁音形（デ）	テラ（寺）
ト濁	基本形（ト） 濁音形（ド）	トマリ（止まり）

語頭変化型	語頭変化形	例
ハ濁	基本形（ハ） 濁音形（バ）	ハシ（橋）
ハ半濁	基本形（ハ） 半濁音形（パ）	ハク（泊）
ハ混合	基本形（ハ） 濁音形（バ） 半濁音形（パ）	ハイ（杯）
ヒ濁	基本形（ヒ） 濁音形（ビ）	ヒョウシ（拍子）
ヒ半濁	基本形（ヒ） 半濁音形（ピ）	ヒン（品）
ヒ混合	基本形（ヒ） 濁音形（ビ） 半濁音形（ピ）	ヒキ（匹）
フ濁	基本形（フ） 濁音形（ブ）	フソク（不足）
フ半濁	基本形（フ） 半濁音形（プ）	フン（分）
フ混合	基本形（フ） 濁音形（ブ） 半濁音形（プ）	フリ（振り）
ヘ濁	基本形（ヘ） 濁音形（ベ）	ヘタ（下手）
ヘ半濁	基本形（ヘ） 半濁音形（ペ）	編（ヘン）
ヘ混合	基本形（ヘ） 濁音形（ベ） 半濁音形（ペ）	遍（ヘン）
ホ濁	基本形（ホ） 半濁音形（ボ）	ホレ（惚れ）
ホ半濁	基本形（ホ） 半濁音形（ポ）	ホ（歩）
ホ混合	基本形（ホ） 濁音形（ボ） 半濁音形（ポ）	本（ホン）
ワ混合	基本形（ワ） 濁音形（バ） 半濁音形（パ）	羽（ワ）

表 7　語末変化型・変化形分類

語末変化型	語末変化形	例
キ促	基本形（キ） 促音形（ッ）	セキ（赤）
ク促	基本形（ク） 促音形（ッ）	ロク（六）
チ促	基本形（チ） 促音形（ッ）	イチ（一）
ツ促	基本形（ツ） 促音形（ッ）	ベツ（別）
十促	基本形（ジュウ） 促音形（ジュッ/ジッ）	ジュウ（十）

語末変化型	語末変化形	例
ア長促添	基本形（） 長音添加形（ア） 促音添加形（ッ）	ヤ（八）
イ長添	基本形（） 長音添加形（イ）	ニ（二）
イ長促添	基本形（） 長音添加形（イ） 促音添加形（ッ）	ミ（三）
ウ長促添	基本形（） 長音添加形（ウ） 促音添加形（ッ）	ム（六）
ウ長促撥添	基本形（） 長音添加形（ウ） 促音添加形（ッ） 撥音添加形（ン）	ヨ（四）
オ長添	基本形（） 長音添加形（オ）	ゴ（五）

表 8　語頭変化結合型分類

	後部要素													
語頭変化結合型	カ濁	ケ濁	ソ濁	ハ半濁	ハ混合	ヒ半濁	ヒ混合	フ半濁	ヘ半濁	ヘ混合	ホ半濁	ホ混合	ワ混合	例
N1	基	基	基	半	半	半	半	半	半	半	半	半	基	イチ（一）
N3	濁	濁	濁	半	濁	半	濁	半	半	半/濁	半	濁	基/濁	サン（三）
N4	基	基	基	半	基	基	基	半	半	基	基	基	基	ヨン（四）
N6	基	基	基	半	半	半	半	半	半	半	半	半	基/半	ロク（六）
N8	基	基	基	半	半	半	半	半	基/半	基/半	基/半	基	基	ハチ（八）
Nj	基	基	基	半	半	半	半	半	半	半	半	半	半	ジュウ（十）
Nh	基	基	基	半	半	半	半	半	半	半	半	半	半	ヒャク（百）
Ns	基	基	基	半	基	半	濁	半	半	半	半	濁	濁	セン（千）
Nm	基	基	基	基	基	基	濁	基	半	半	半	濁	濁	マン（万）
Nn	基	基	基	半	濁	半	濁	半	半	半	半	濁	濁	ナン（何）

基: 基本形，　濁: 濁音形，　半: 半濁音形

6.4　語末変化結合型

語の複合に際して、前部要素の語末音がどの形を取るかは、後部要素の種類によって制約される。ただし、この制約は数詞の種類ごとに多分に異なり、かなり煩雑なので、ここでは省略する。

表 9　アクセント修飾型分類

アクセント修飾型	活用形のアクセント型			例
	基本形のアクセント型が			
	0 型	1 型	それ以外	
M1@M	$N_0 - M$			意志推量形
M2@M	$N_0 - M$	M_0		一段・サ変活用の命令形
M4@M	M_0		$M_0 - M$	一段活用の未然形

N_0: 当該活用形のモーラ数,　M_0: 基本形のアクセント型,　M: 当該活用形のアクセント修飾価

表 10　アクセント結合型分類（普通名詞・接尾辞）

アクセント結合型	複合語のアクセント型	例
C1	$N_1 + M_2$	テツヅキ（手続き）, ニチカン（日間）
C2	$N_1 + 1$	セイカツ（生活）, ジカン（時間）
C3	N_1	ワン（湾）, ガク（学）
C4	0	シマ（島）, ケイ（系）
C5	M_1	ドノ（殿）

N_1: 前部要素のモーラ数,　M_1: 前部要素のアクセント型,　M_2: 後部要素のアクセント型

6.5　アクセント型

アクセント型は、アクセントの位置を先頭からのモーラ数で数えることによって表わす。たとえば、「シャ]カイ（社会）」のアクセント型は 1、「カタカ]ナ（片仮名）」のアクセント型は 3 である。0 は平板型を示す。複数のアクセント型が可能な場合は、コンマで区切って併記する。並びの順が優先順序を表わす。

6.6　アクセント修飾型

活用のある語が特定の活用形を取る場合に、基本形のアクセント型が変化することがある。この変化の種類をアクセント修飾型で表わす。アクセント修飾型の一覧を表 9 に示す。

6.7　アクセント結合型

複合語を作ったり、自立語に助詞・助動詞が結合したりする際に、アクセントの位置が変化することがある。この変化は、後部要素（接頭辞との結合では前部要素である接頭辞）の種類によって定まり、アクセント結合型によって表わす。後部（前部）要素の品詞ごとに、アクセント結合型の一覧を表 10（普通名詞・接尾辞）、表 11（接頭辞）、表 12（助詞・助動詞）にそれぞれ示す[*4]。 これらの表の内容は、chaone モジュールに規則として実装されている。

一般に、3 モーラ以上の普通名詞・接尾辞については、平板型と尾高型の語は C2 型に、それ以外は C1 型に分類される。ただし、例外も多くある。2 モーラ以下の普通名詞・接尾辞や接頭辞・助詞・助動詞のアクセン

[*4] 助詞・助動詞のアクセント結合型は、実際には、前部要素の品詞（名詞・動詞・形容詞）によって異なる場合がある。表 12 の記述は代表的なものである。

表 11　アクセント結合型分類（接頭辞）

アクセント結合型	複合語のアクセント型		例
	後部要素のアクセント型が		
	0 型・N_2 型	それ以外	
P1	0	$N_1 + M_2$	御（ゴ）
P2	$N_1 + 1$	$N_1 + M_2$	総（ソウ）
P4	$N_1 + 1$	M_1	両（リョウ）
P6	0		再来（サライ）
P13	M_1		現（ゲン）
P14	M_1	$N_1 + M_2$	要（ヨウ）

N_1: 前部要素のモーラ数，　M_1: 前部要素のアクセント型，　N_2: 後部要素のモーラ数，　M_2: 後部要素のアクセント型

表 12　アクセント結合型分類（助詞・助動詞）

アクセント結合型	文節のアクセント型		例
	前部要素のアクセント型が		
	0 型	それ以外	
F1	M_1		が（格助詞）, た（助動詞）
F2@M	$N_1 + M$	M_1	か（終助詞）, です（助動詞）
F3@M	M_1	$N_1 + M$	せる（助動詞）
F4@M	$N_1 + M$		ます（助動詞）
F5	0		だけ（副助詞）
F6@M, L	$N_1 + M$	$N_1 + L$	たり（副助詞）

N_1: 前部要素のモーラ数，　M_1: 前部要素のアクセント型，　M, L: 後部要素のアクセント結合型

ト結合型を一般的に定めることはできない。

7　その他の情報

7.1　語種

語の出自を分類した語種情報として、表 13 のものを記述している。

表 13　語種分類

語種ラベル	説明
和	和語
漢	漢語
外	外来語
混	混種語
固	固有名
記	記号
他	その他

付録 A　変更履歴

A.1　Version 1.3.0 からの変更点

- 活用のある語の「基本形」を「終止形」と「連体形」に区別した。

A.2　Version 1.3.5 からの変更点

- MeCab 版を作成した。
- 語種情報を追加した。

A.3　Version 1.3.8 からの変更点

- MeCab 版の辞書の並び順を変更した。