近代文語 UniDic ユーザーズマニュアル

Ver. 1.2　2012/04/23　小木曽智信

1. 近代文語 UniDic とは

近代文語 UniDic は、現代語向けの UniDic をもとにして近代文語文を解析できるようにした形態素解析辞書です。現代語版と同様、次のような言語研究に適した特長を持っています。

- 「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。

ただし、UniDic の特長である発音形・アクセント等の音声情報は、現代の読み方による限定的なものです。

近代文語 UniDic が主たる対象とするテキストは、近代の文語論説文（明治普通文）です。一般的な文語論説文のテキストであれば、概ね 95%～98%程度の精度で解析することができます。

近代文語 UniDic は現代語の UniDic をもとにして作成されており、見出しに互換性があります（見出し語についてみると、近代文語 UniDic はほぼ現代語 UniDic のスーパーセットになっています[1]）。このため、現代語版の UniDic で解析した結果と近代文語 UniDic で解析した結果を比較することができます。

利用条件

近代文語 UniDic の利用条件は次の通りです。現代語版 UniDic と同等の内容です。

【近代文語 UniDic 利用条件】

1. 近代文語 UniDic の著作権は，小木曽智信，小椋秀樹，近藤明日子 および The UniDic consortium（代表：伝康晴）が保持する。
2. 近代文語 UniDic を複製又は改変することは，個人的な利用に限り認める。
3. 近代文語 UniDic 及びこれを改変したものを再配布してはならない。

[1]顔文字など近代文語文には現れない一部の見出し語は削除されています。

4. 近代文語 UniDic を利用して行った研究等の成果を公表する場合は，近代文語 UniDic を利用したことを明記すること。
5. 営利を目的として，近代文語 UniDic を利用する場合は，事前に著作権者と協議すること。
6. 近代文語 UniDic を利用することによって，直接的・間接的に生じたいかなる損害についても，著作権者は賠償する責任を負わない。
7. 本文書に定めのない事項については，著作権者と協議すること。

ダウンロードサイト

近代文語 UniDic は次の Web ページからダウンロードすることができます。

http://www2.ninjal.ac.jp/lrc/index.php?UniDic

2. 利用方法

近代文語 UniDic は Windows 版パッケージとバイナリ辞書の二つの形態で配布されています。一般の方は前者をお使いください。MeCab 等のツールに詳しい方、Windows 以外の OS を利用している方はバイナリ辞書をお使いください。

Windows 版パッケージ

（Windows 上で簡単に利用できるようにした配布形態です。）

Windows パッケージは簡単にインストールして利用することができるようになっています。ダウンロードしたファイルを実行するとインストーラが起動しますので、あとは画面上の指示に従ってください。ただし、近代文語 UniDic の他に，形態素解析器として ChaSen または MeCab をインストールする必要があります。詳しくは「インストールマニュアル」を御覧ください。

形態素解析は、同梱されている解析用のプログラム「近代茶まめ」を起動することで手軽に利用できます。インストール後に現れる「近代茶まめ」アイコンから起動してください。詳しい使い方は「茶まめマニュアル」を御覧ください。

バイナリ辞書

（詳しい方向けの配布形態です。辞書本体以外のツールを同梱していませんので、現代語版 UniDic のツールをご利用ください。）

ダウンロードしたファイルを展開し、ディスク上の任意の場所にコピーしてください。辞書は UTF-8 版のみ公開しています（エンディアンが同じであれば 64bit 環境でも使えます）。解析のためには形態素解析器が必要ですので、ダウンロードした辞書にあわせ、ChaSen 2.4.3 以降または MeCab0.96 以降をインストールしてください。

形態素解析の実行は次のように行います。

MeCab の場合

```
mecab -d /usr/local/unidic/dic/unidic-mecab FILE
```

ChaSen の場合　（UniDic 標準の XML タグ形式で出力する場合）

```
chasen -i w -r /usr/local/unidic/dic/unidic-chasen FILE
```

ChaSen で表形式テキストの出力を利用したい場合には、現代語版 UniDic パッケージ付属の chauni を利用してください。

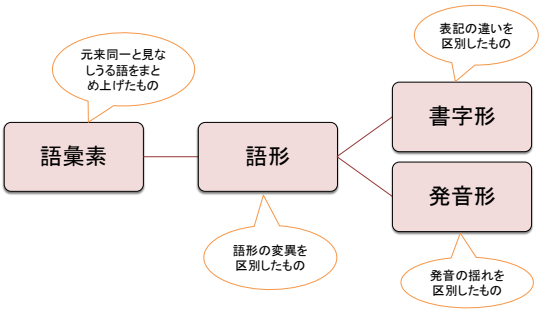
出力結果の列名は、現代語版 UniDic のユーザーズマニュアルを御覧ください。また、辞書に同梱されている chasenrc, dicrc を参考にしてください。

ソース辞書

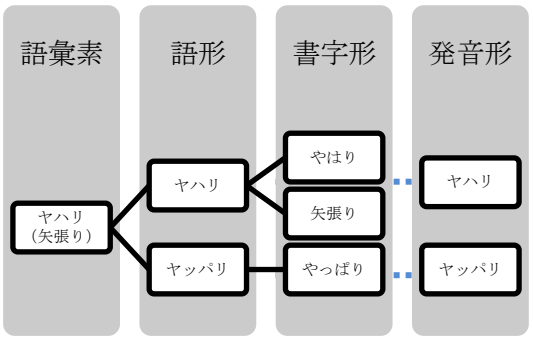
近代文語 UniDic のソース辞書は現在一般公開を行っていません。研究目的で利用される場合にはデータをお送りしますので、必要な場合には連絡先までご相談ください。

3. UniDic の出力

UniDic は次に示すような階層構造を持っています。



UniDic の階層構造



見出し語の例

「語彙素」は国語辞典の見出し語に相当するレベルで、語種などの情報はここに記述されます。「語形」は異語形を区別するレベルで、たとえば「アマリ（余り）」に対する「アンマリ」「アンマシ」「アンマ」といった異語形、上一段活用と文語上二段活用といった活用の違いのほか、可能動詞形もここで区別されます。「書字形」は異表記を区別するレベルで、漢字を使うか仮名書きするかといった違いのほか、送り仮名の揺れもここに記述されます。「発音形」は発音やアクセントなどの情報が記述されます。

UniDic の解析結果には次のような情報が出力されます。

項目	**説明**	**階層**
書字形	（活用後の）書字形＝テキストに現れた出現形	書字形
発音形	（活用後の）発音形（長音は「ー」）	発音形
語彙素読み	辞書のカナ見出しに相当	語彙素
語彙素（-語彙素細分類）	辞書の代表表記に相当	語彙素
品詞	詳しくはこのマニュアルの 4.品詞体系 を参照	語形
活用型		
活用形		
語形	異語形の場合、語彙素読みと異なる	語形
語種	和語・漢語等の語種	語彙素

※ 「茶まめ」ではこれ以外に出典・文境界の情報が出力されます。詳しくは茶まめの説明書を御覧ください。

※ ChaSen 版ではこれ以外にアクセントや語頭・語末変化などの情報も出力されます。詳しくは現代語版 UniDic のマニュアルを御覧ください。

4. 品詞体系

以下に近代文語 UniDic の品詞・活用型・活用形のリストを掲げる。近代文語 UniDic の見出し語は現代語版 UniDic の見出し語を含んでいるため、口語の活用型についても掲げる。

品詞

品詞	大分類	中分類	小分類	細分類	類
名詞-普通名詞-一般	名詞	普通名詞	一般		体
名詞-普通名詞-サ変可能			サ変可能		体
名詞-普通名詞-形状詞可能			形状詞可能		体
名詞-普通名詞-サ変形状詞可能			サ変形状詞可能		体
名詞-普通名詞-副詞可能			副詞可能		体
名詞-普通名詞-連用可能			連用可能		体
名詞-固有名詞-一般		固有名詞	一般		固有名
名詞-固有名詞-人名-一般			人名	一般	人名
名詞-固有名詞-人名-姓			人名	姓	姓
名詞-固有名詞-人名-名			人名	名	名
名詞-固有名詞-地名-一般			地名	一般	地名
名詞-固有名詞-地名-国			地名	国	国
名詞-数詞		数詞			数
名詞-助動詞語幹		助動詞語幹			体
代名詞	代名詞				体
形状詞-一般	形状詞	一般			相
形状詞-タリ		タリ			相
形状詞-助動詞語幹		助動詞語幹			助動
連体詞	連体詞				相
副詞	副詞				相
接続詞	接続詞				他
感動詞-一般	感動詞	一般			他
感動詞-フィラー		フィラー			他
動詞-一般	動詞	一般			用
動詞-非自立可能		非自立可能			用
形容詞-一般	形容詞	一般			相
形容詞-非自立可能		非自立可能			相

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

※ＡＡはアスキーアートの略

活用型

動詞（文語）

活用型（コーパス）	補足説明
文語力行変格	

活用型（コーパス）	補足説明
文語サ行変格	
文語ザ行変格	「-ず」型の一字漢語サ変動詞
文語ナ行変格	
文語ラ行変格	
文語上一段-カ行	
文語上一段-ナ行	
文語上一段-マ行	
文語上一段-ヤ行	
文語上一段-ワ行	
文語上二段-タ行	
文語上二段-ダ行	
文語上二段-ハ行	
文語上二段-バ行	
文語上二段-ヤ行	
文語下二段-ア行	
文語下二段-カ行	
文語下二段-ガ行	
文語下二段-サ行	
文語下二段-ザ行	
文語下二段-タ行	
文語下二段-ダ行	
文語下二段-ナ行	
文語下二段-ハ行	
文語下二段-バ行	
文語下二段-マ行	
文語下二段-ヤ行	
文語下二段-ラ行	
文語四段-カ行	
文語四段-ガ行	
文語四段-サ行	
文語四段-タ行	

活用型（コーパス）	補足説明
文語四段-ハ行	
文語四段-バ行	
文語四段-マ行	
文語四段-ラ行	

助動詞（文語）

活用型（コーパス）	補足説明
文語助動詞-キ	
文語助動詞-ケム	
文語助動詞-ケリ	
文語助動詞-ゴトシ	
文語助動詞-ザマス	
文語助動詞-ザンス	
文語助動詞-ズ	
文語助動詞-タリ-完了	
文語助動詞-タリ-断定	
文語助動詞-ツ	
文語助動詞-ナリ-伝聞	
文語助動詞-ナリ-断定	
文語助動詞-ヌ	
文語助動詞-ベシ	
文語助動詞-マジ	
文語助動詞-ム	
文語助動詞-ラシ	
文語助動詞-ラム	
文語助動詞-リ	
文語助動詞-ンス	近世上方語
無変化型	

形容詞（文語）

活用型（コーパス）	補足説明
文語形容詞-ク	
文語形容詞-シク	

文語形容詞-ジク	「いみじ」など
文語形容詞-多シ	「多し」（終止「多かり」）

動詞（口語）

活用型（コーパス）	**補足説明**
カ行変格	
サ行変格	
ザ行変格	「-ずる」型の一字漢語サ変動詞
上一段-ア行	
上一段-カ行	
上一段-ガ行	
上一段-ザ行	
上一段-タ行	
上一段-ナ行	
上一段-ハ行	
上一段-バ行	
上一段-マ行	
上一段-ラ行	
下一段-ア行	
下一段-カ行	
下一段-ガ行	
下一段-サ行	
下一段-ザ行	
下一段-タ行	
下一段-ダ行	
下一段-ナ行	
下一段-ハ行	
下一段-バ行	
下一段-マ行	
下一段-ラ行-一般	
下一段-ラ行-呉レル	「呉れる」（命令形「くれ」）
五段-カ行-イク	「行く（イク）」（連用形促音便あり）
五段-カ行-ユク	「行く（ユク）」（連用形に音便なし）

活用型（コーパス）	補足説明
五段-カ行-一般	
五段-ガ行	
五段-サ行	
五段-タ行	
五段-ナ行	
五段-バ行	
五段-マ行	
五段-ラ行-アル	
五段-ラ行-一般	
五段-ワア行-イウ	「言う」（イーマス／ユー）
五段-ワア行-一般	

助動詞（口語）

活用型（コーパス）	補足説明
助動詞-ジャ	
助動詞-タ	
助動詞-タイ	
助動詞-ダ	
助動詞-デス	
助動詞-ドス	関西（京都）方言
助動詞-ナイ	
助動詞-ヌ	
助動詞-ヘン	関西方言
助動詞-マス	
助動詞-ヤ	
助動詞-ヤス	「～でやす」
助動詞-ラシイ	

形容詞（口語）

活用型（コーパス）	補足説明
形容詞	

活用形

大分類	活用形	補足説明
語幹	語幹-サ	形容詞「無い」「良い」に、様態の助動詞「そうだ」が接続するときの形（「無さ-そうだ」「良さ-そうだ」）
	語幹-一般	
未然形	未然形-サ	サ変（ザ変）に、助動詞「せる」「れる」が接続するときの形（「さ-せる」「さ-れる」）
	未然形-セ	サ変（ザ変）に、助動詞「ず」が接続するときの形（せ-ず）
	未然形-一般	
	未然形-撥音便	ラ行五段活用動詞の一部で起こる撥音便（「知ん-ない」）
	未然形-補助	形容詞カリ活用未然形（「少なから-ず」）
意志推量形	意志推量形	意志・推量の助動詞「う」「よう」が接続した形全体（「行こう」「見よう」）
連用形	連用形-イ音便	
	連用形-ウ音便	
	連用形-キ接続	文語サ行四段活用の動詞に過去の文語助動詞「き」が接続する場合に「～せ」の形を取ることがある（「許せ-し」）cf.「文法上許容スベキ事項」
	連用形-ト	断定の文語助動詞「たり」の連用形「と」
	連用形-ニ	断定の助動詞「だ」・文語助動詞「なり」の連用形「に」
	連用形-一般	
	連用形-促音便	
	連用形-撥音便	
	連用形-省略	関西方言などで形容詞連用形が省略された形をとることがある（「欲し-ない」）
	連用形-融合	断定の助動詞「だ」の連用形に後続する係助詞「は」が融合した形（「じゃ」）
	連用形-補助	文語形容詞・文語助動詞「ず」のカリ活用連用形（「無かり」「ざり」）
終止形	終止形-ウ音便	文語ハ行四段活用動詞の終止形がウ音便化することがある（「給う［タモー］」「候［ソーロー］」）
	終止形-一般	

大分類	活用形	補足説明
	終止形-促音便	形容詞の「高っ」「痛っ」などの形
	終止形-撥音便	助動詞「ず」の終止形に撥音便形がある（「（しませ）ん」） また関西方言などで撥音便形になることがある(「てん(な)」)
	終止形-融合	断定の助動詞「だ」の終止形に前接する「と」の音と融合した形（「（何のこっ）ちゃ」）
	終止形-補助	文語形容詞「多し」のカリ活用終止形（「多かり」）
連体形	連体形-ウ音便	文語ハ行四段活用動詞の連体形がウ音便化することがある （「給う［タモー］」「候［ソーロー］」）
	連体形-一般	
	連体形-撥音便	助動詞「ず」の連体形がしばしば「ん」となるほか、動詞でも「すん（の）」のように準体助詞「の」の前で撥音になる また文語助動詞「む」「けむ」の連体形が「ん」となる
	連体形-省略	
	連体形-補助	
已然形	已然形	
	已然形-一般	
	已然形-補助	
仮定形	仮定形-一般	
	仮定形-融合	
命令形	命令形	
	命令形-一般	
ク語法	ク語法	文語専用

近代文語 UniDic における例外的処理

近代文語 UniDic では、やむを得ない要因により、現代語の短単位規定とは異なった単位分割を行った場合があります。

- 表記上の問題から短単位を結合した例（括弧内は全体に与えた品詞）

  例：しかのみならず　加之　（接続詞）

  例：といえども　と雖も　（接続助詞）

- ク語法（次のパターンにわかれている）

  曰く・：名詞-普通名詞-副詞可能（現代語 UniDic と共通）

  らく：助動詞「り」の活用形扱い（活用形＝ク語法）

  まく：助動詞「む」の活用形扱い（活用形＝ク語法）

  一般の四段活用動詞のク語法：当該動詞の活用形扱い（活用形＝ク語法）

  四段活用動詞に接続する「らく」：接尾辞-名詞的-一般

近代文語 UniDic の更新履歴

- 2012/04/23　近代文語 UniDic Ver.1.2 Windows 版パッケージ,バイナリ辞書を公開。
- 2009/08/12　近代文語 UniDic Ver.1.1 Windows 版パッケージ,バイナリ辞書を公開。（現代語版 UniDic 1.3.12 と同期）
- 2009/03/31　近代文語 UniDic Ver.1.0 Windows 版パッケージ,バイナリ辞書を公開。
- 2008/11/30　近代文語 UniDic Ver.0.9 Windows 版パッケージ,バイナリ辞書を公開。
- 2008/08/01　近代文語 UniDic Ver.0.8 Windows 版パッケージ,バイナリ辞書を公開。
- 2008/05/19　近代文語 UniDic Ver.0.7a Windows 版パッケージ（修正版）を公開。
  - 複数ファイルの解析に失敗するエラーに対応。
  - ファイル末に EOF コードが付いているテキストファイルに対応。
- 2008/04/11　近代文語 UniDic Ver.0.7 Windows 版パッケージを公開。
- 2008/03/28　近代文語 UniDic Ver.0.7 バイナリ辞書を公開。

参考文献

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「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 号 pp.101-122.

小椋秀樹・小磯花絵・冨士池優美・原裕（2009）『『現代日本語書き言葉均衡コーパス』形態論情報規程集 改定版』国立国語研究所内部報告書　LR-CCG-08-03
小木曽智信・中村壮範（2009）『『現代日本語書き言葉均衡コーパス』形態論情報データベースの設計と実装』国立国語研究所内部報告書　LR-CCG-08-04

小木曽智信・伝康晴・渡部涼子・近藤明日子（2009）「現代語コーパスの利用による近代語形態素解析の精度向上」『言語処理学会第 15 回年次大会発表論文集』 pp.801-804
小木曽智信・小椋秀樹・近藤明日子（2008a）「近代文語文を対象とした形態素解析辞書の開発」『言語処理学会第 14 回年次大会発表論文集』 pp.225-228
小木曽智信・小椋秀樹・近藤明日子（2008b）「近代文語文を対象とした形態素解析辞書・近代文語 UniDic」『日本語学会 2008 年度春季大会予稿集』 pp.211-218