



『UniDic』の拡張計画

国立国語研究所 コーパス開発センター
特任助教 岡 照晃

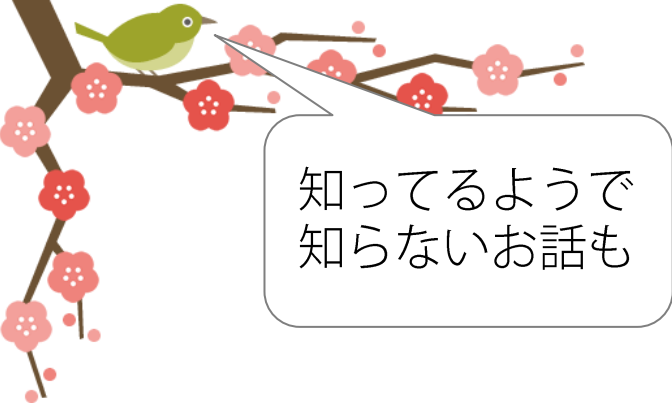
『UniDic』 拡張計画とは？

- コーパス開発センターのプロジェクト
「日本語言語資源の包括的高度共同利用環境の整備」
の項目の一つ
 - UniDicの量的拡張と質的拡張
 - 量的拡張： 【目標】 年間1,000語の追加
 - 参考： 広辞苑の改版1998→2008年 10,000語の追加
 - 質的拡張：
 - 意味情報（分類語彙表番号）の付与



補足 1 : 現状のUniDicデータベース登録短単位数 (2017/3/31 現在)

- 展開した書字形出現形数 :
 - 総数 (含 歴史コーパス用) : 1,734,897
 - 現代語 : 894,992
- 語彙素数 :
 - 総数 (含 歴史コーパス用) : 258,277
 - 現代語 : 229,527



広報活動

『UniDic』の拡張計画

国立国語研究所 コーパス開発センター
特任助教 岡 照晃

UniDic クイズ

- UniDicの説明として、自分の直観的な理解に一番近いものに手を上げてください。
 - ① 短単位が収録された辞書である
 - (形式は問わない)
 - ② 短単位が格納されたデータベースである
 - ③ 短単位でエントリされたMeCab用の辞書である



UniDic クイズ

- UniDicの説明として、自分の直観的な理解に一番近いものに手を上げてください。

① 短単位が収録された辞書である
－（形式は問わない）

② 短単位が格納されたデータベースである

③ 短単位でエントリされたMeCab用の辞書である



UniDic クイズ

- UniDicの説明として、自分の直観的な理解に一番近いものに手を上げてください。

① 短単位が収録された辞書である

– (形式は問わない) ←対外的な簡単な説明

② 短単位が格納されたデータベースである

↑ 所内のアノテーション作業者的理解

③ 短単位でエントリされたMeCab用の辞書である

↑ UniDicをMeCabの辞書として使っている方々の理解



『UniDic』とは（2017年現在）

- UniDicとは，国語研の規定した斉一な言語単位（短単位）と，階層的見出し構造に基づく電子化辞書の
 - － ① 設計方針
- および，その実装としてのリレーショナルデータベース
 - － ② UniDicデータベース
- と，そのデータベースからエクスポートされた短単位をエントリとする形態素解析器MeCab用の解析用辞書
 - － ③ 解析用UniDic

の総称です。



①設計方針

【階層的な見出し構造】

異表記・異形態に対して同一の見出しを与える構造

辞書見出し*

語彙素読み：
オオキイ
語彙素：
大きい

語形
(基本形)

オオキイ

書字形・発音形
(基本形)

書 大きい

書 おおきい

音 オーキー

オッキイ

書 おっきい

音 オッキー

* 語彙素読みと
語彙素を合わせて
はじめて当該の
階層構造が一意に
定まる。

実装

【斉一な言語単位】

単位認定が、ある場合は長く、ある場合は短いという不揃いをなくす

斉一でない言語単位で分割した例：

「幾何学」が1単位なのに対し、「心理学」は2単位に分割される。

幾何学 / と / 心理 / 学 / を / 学ぶ / 。

斉一な言語単位で分割した例：

幾何 / 学 / と / 心理 / 学 / を / 学ぶ / 。

②UniDicデータベース



エクスポート

③解析用UniDic

書字形 (基本形)	発音形 (基本形)	語形 (基本形)	語彙素 読み	語彙素	...
大きい	オーキー	オオキイ	オオキイ	大きい	...
おおきい	オーキー	オオキイ	オオキイ	大きい	...
おっきい	オッキイ	オッキイ	オオキイ	大きい	...

セッション3 14:15－14:45

「『日本語歴史コーパス』に出現した新規語の『**UniDic**』への登録について」

鴻野 知暁（国立国語研究所）

さて、この『**UniDic**』はどこを指しているのでしょうか？

①設計方針

【階層的な見出し構造】

異表記・異形態に対して同一の見出しを与える構造

辞書見出し*

語彙素読み：
オオキイ
語彙素：
大きい

語形
(基本形)

オオキイ

書字形・発音形
(基本形)

書 大きい
書 おおきい
音 オーキー
書 おつきい
音 オッキー

* 語彙素読みと
語彙素を合わせて
はじめて当該の
階層構造が一意に
定まる。

実装

【斉一な言語単位】

単位認定が、ある場合は長く、ある場合は
短いという不揃いをなくす

斉一でない言語単位で分割した例：

「幾何学」が1単位なのに対し、「心理学」は
2単位に分割される。

幾何学 / と / 心理 / 学 / を / 学ぶ / 。

斉一な言語単位で分割した例：

幾何 / 学 / と / 心理 / 学 / を / 学ぶ / 。

②UniDicデータベース



エクスポート

③解析用UniDic

書字形 (基本形)	発音形 (基本形)	語形 (基本形)	語彙素 読み	語彙素	...
大きい	オーキー	オオキイ	オオキイ	大きい	...
おおきい	オーキー	オオキイ	オオキイ	大きい	...
おつきい	オッキー	オッキー	オオキイ	大きい	...

①設計方針

【出し構造】

一の見出しを与える構造

書字形・発音形
(基本形)

- 書 大きい
- 書 おおきい
- 音 オーキー
- 書 おつきい
- 音 オッキー

【斉一な言語単位】

単位認定が、ある場合は長く、
短いという不揃いをなくす

斉一でない言語単位で分割した例

「幾何学」が1単位なのに対し、
2単位に分割される。

幾何学 / と / 心理 / 学 / を

斉一な言語単位で分割した例：

幾何 / 学 / と / 心理 / 学 / を

③解析用UniDic

書字形 (基本形)	発音形 (基本形)	語形 (基本形)	語彙素 読み	語
大きい	オーキー	オオキイ	オオキイ	大
おおきい	オーキー	オオキイ	オオキイ	大
おつきい	オッキー	オッキイ	オオキイ	大

UniDic Explorer :
DB操作のインターフェース

②UniDicデータベース

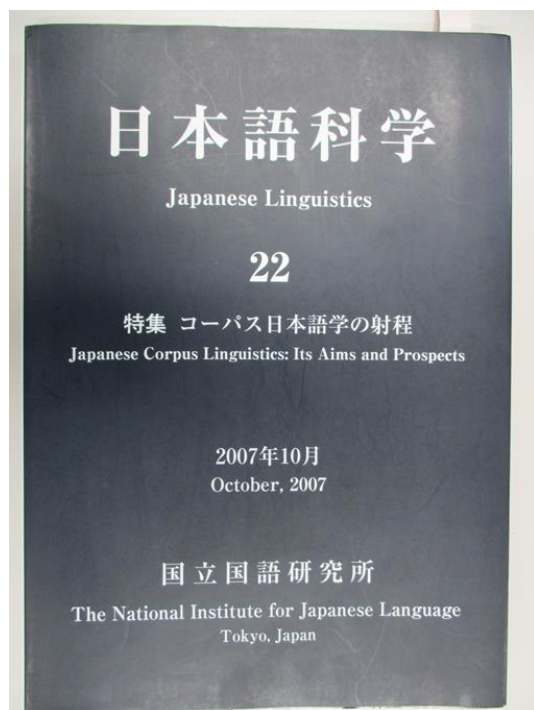


エクスポート



実は、このあたりのお話は この論文にしっかり載っています

- 伝康晴，小木曾智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵：
「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，Vol.22，pp. 101-123 (2007).



章立て

1. はじめに
2. 電子化辞書の設計
- ...
3. 実装
3. 1 辞書データベースシステム
3. 2 形態素解析システムでの利用
- ...

『UniDic』関連の論文でよく引用されている論文ですが，権利関係の都合，ネット上で電子的に公開されていません。

補足 2.1：UniDicの目的

- UniDicの第一の目的は、国語研のコーパス構築促進
 - UniDicデータベースはコーパスデータベースと参照関係にあり、コーパス中に出現する短単位は、
 - すべてUniDicデータベースに登録されている
 - すべてUniDicデータベース中の一意のエントリを参照する状態になっている

 コーパスと辞書の統合的運営

補足2.2：コーパスと辞書の統合的運営

UniDicデータベース

書字形 (出現形)	発音形 (出現形)	語形 (出現形)	品詞	語彙素	語彙素読み	類	...
すもも	スモモ	スモモ	名詞 -普通名詞 一般	李	スモモ	体	
もも	モモ	モモ	名詞 -普通名詞 一般	桃	モモ	体	
も	モ	モ	助詞 -係助詞	も	モ	係助	

コーパスデータベース
文字列テーブル

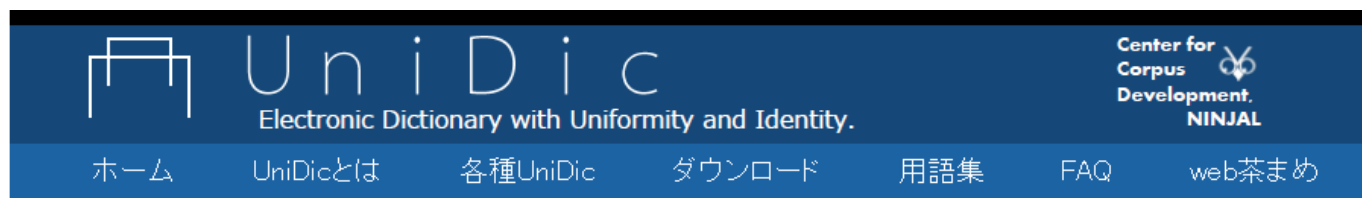
order	文字
10	す
20	も
30	も
40	も
50	も
60	も
70	も
80	も
90	も



補足2.3：コーパスと辞書を統合的に運営する利点

- 短単位情報アノテーションの際、コーパスには、出現した各短単位がUniDicデータベース中のどのエントリに対応しているかを選択していけばいい
 - コーパス中の異なる場所に出現する同一の短単位に対し、一部の属性を異なって与えてしまうというミス在未然に防ぎ、コーパス中に不整合が生じる可能性が排除できる。
- 現時点のUniDicデータベースに存在しない情報が、新たにUniDicデータベースへ追加された場合でも、データベース間のリンクでコーパスへの反映は瞬時に行える

そこで現在、『UniDic』公式サイトを 新しく整備しています (近日公開予定)



UniDicとは？ (2017年現在)

UniDicとは、国語研の規定した斉一な単位 ([短単位](#)) と、階層的見出し構造に基づく電子化辞書の ① 設計方針および、その実装としてのリレーショナルデータベース (② UniDicデータベース) と、そのデータベースからエクスポートされた短単位をトークンとする形態素解析器MeCab用の解析用辞書 (③ 解析用UniDic) の総称です。

本サイトでは③の 解析用UniDic を公開・配布しています。

また、解析用UniDicは[短単位](#)をMeCab辞書のトークンとしているため、UniDicを使った形態素解析は「短単位 (自動) 解析」と呼びます。

① 設計方針

【階層的な見出し構造】

異表記・異形態に対して同一の見出しを与える構造

辞書見出し*

語彙素読み：
オオキイ
語彙素：
大きい

語形
(基本形)

オオキイ

書字形・発音形
(基本形)

書 大きい
書 おおきい
音 オーキー

【斉一な言語単位】

単位認定が、ある場合は長く、ある場合は短いという不揃いをなくす

斉一でない言語単位で分割した例：

「幾何学」が1単位なのに対し、「心理学」は2単位に分割される。

幾何学 / と / 心理 / 学 / を / 学ぶ / 。

UniDic公開サイトの現状…

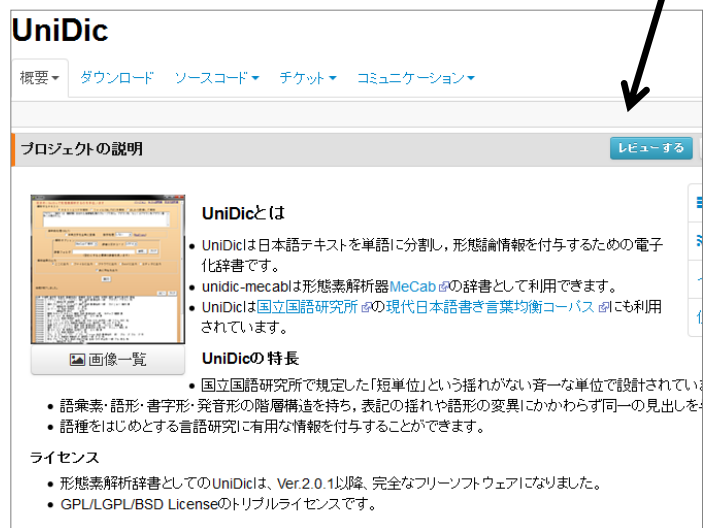
http://pj.ninjal.ac.jp/corpus_center/unidic/



http://chamame.ninjal.ac.jp/chamame_unidic_download.html



<https://ja.osdn.net/projects/unidic/>



<http://www2.ninjal.ac.jp/lrc/index.php?UniDic%2F%B6%E1%C2%E5%CA%B8%B8%ECUniDic>



ちなみに、このサイトは 既に活動を停止しています…

UniDic

<http://download.unidic.org/>

UniDicウェブサイト



ホーム 旧バージョンのUniDicの利用 旧バージョンの情報

UniDicトップページ

現在、新しいUniDicの公開のため準備中です。

- 形態素解析辞書としてのUniDic(unidic-mecab)は、Ver.2.0.1以降、オープンソースのソフトウェアとしてsourceforge.jpから公開を行っています。MeCab用のUniDicやWindows用のパッケージが必要な方は[こちらからダウンロード](#)してください。
 - <http://sourceforge.jp/projects/unidic/>
- 今後このサイトでは、XML版などの新しい形式でUniDicを公開する予定です。もうしばらくお待ちください。新しい形式のUniDic2については下記をご覧ください。
 - http://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/A6-3.pdf

検索

最近のコメント

アーカイブ

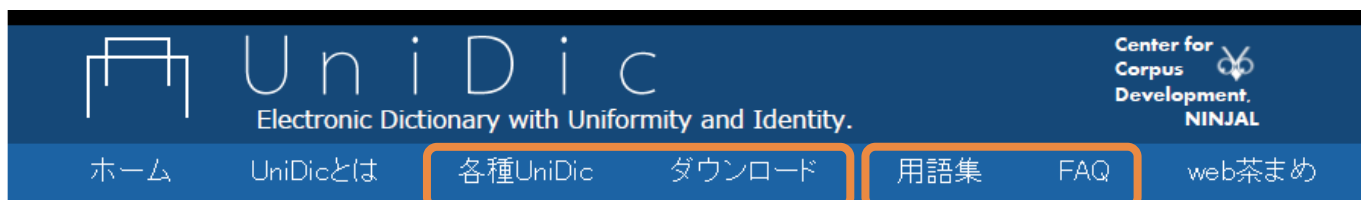
カテゴリ

■ カテゴリーなし

メタ情報

- ログイン
- 投稿の RSS
- コメントの RSS
- WordPress.org

分散しているサイトを統合し，かつ用語集や，FAQなどを完備したサイトを建設中です

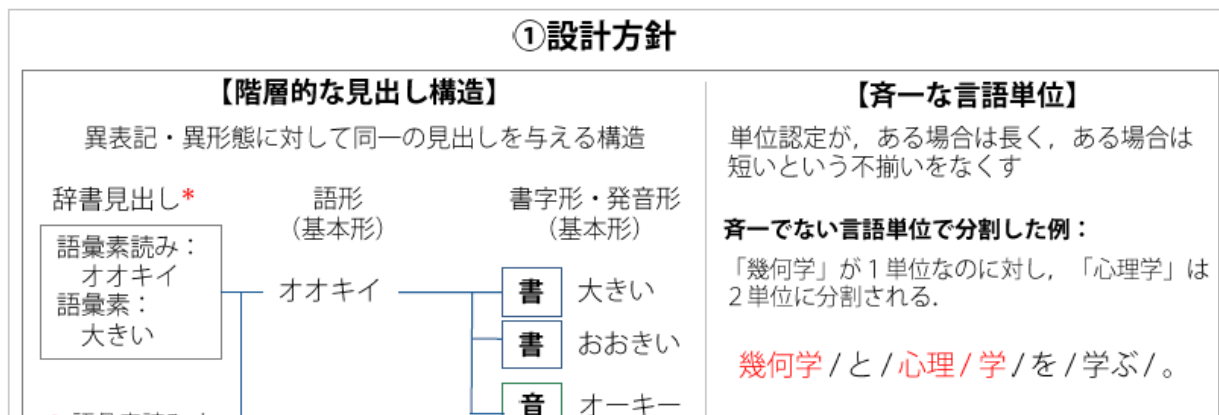


UniDicとは？（2017年現在）

UniDicとは，国語研の規定した斉一な単位（[短単位](#)）と，階層的見出し構造に基づく電子化辞書の ① 設計方針および，その実装としてのリレーショナルデータベース（② UniDicデータベース）と，そのデータベースからエクスポートされた短単位をトークンとする形態素解析器MeCab用の解析用辞書（③ 解析用UniDic）の総称です。

本サイトでは③の 解析用UniDic を公開・配布しています。

また，解析用UniDicは[短単位](#)をMeCab辞書のトークンとしているため，UniDicを使った形態素解析は「短単位（自動）解析」と呼びます。



「短単位」とは

- これまでの対外的な，簡略化した説明：
 - （形態素に相当する？） 斉一な言語単位
 - 文字列検索よりPrec.高く， Rec.重視で用例の検索できる
-
- [小椋ら, 2004] より
 - CSJの単位の設計に当たっては，語彙調査と同様に，まず目的を設定した上で，その目的に適した単位を設計することとした（操作主義的な立場）
 - 目的：CSJを使ってどのような国語研究を行なうのか？
 - （１）用例を収集し，語彙・語法の研究を行う→**短単位**
 - （２）品詞の分布などの計量分析によって言語の構文的な機能の研究を行う→**長単位**

補足 3：短単位と長単位の違い

- 短単位：
 - 用例収集を目的とした単位
 - 言語の形態論的側面に着目して規定した言語単位
 - ≡ 単語の内部構造を扱う領域の単位
- 長単位：
 - 言語的特徴の解明を目的とした単位
 - 言語の構文的な機能に着目して規定した言語単位
 - ≡ 単語を組み合わせて文を作る領域の単位
- 短単位は、検索用途の単位です！
- 構文解析には、専用の長単位を使いましょう！

しかし「短単位」といいつつ、 短くない書字形が入ってますよね？

lex.csv

あいかぎ,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,あいかぎ,アイカギ,あいかぎ,あいかぎ,
和,,,,,,,,,29695538242048
合いかぎ,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合いかぎ,アイカギ,合いかぎ,合いかぎ,
和,,,,,,,,,29695437578752
合い鍵,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合い鍵,アイカギ,合い鍵,合い鍵,
和,,,,,,,,,29695504687616
合鍵,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合鍵,アイカギ,合鍵,合鍵,和,,,,,,,,, 29695471133184
会い掛けれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,会い掛けれ,アイカケ
レ,会い掛ける,会い掛ける,和,,,,,,,,, 24146933304140513
遭いかけれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,遭いかけれ,アイカケ
レ,遭いかける,遭いかける,和,,,,,,,,, 24146933237031649
遭い掛けれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,遭い掛けれ,アイカケ
レ,遭い掛ける,遭い掛ける,和,,,,,,,,, 24146933270586081
会い掛けりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,会い掛けりゃ,アイ
カケリゃ,会い掛ける,会い掛ける,和,,,,,,,,,24146933304140514
遭いかけりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,遭いかけりゃ,アイ
カケリゃ,遭いかける,遭いかける,和,,,,,,,,, 24146933237031650
会い掛ける,会い掛ける,和,,,,,,,,*,*,24146933304140514
遭いかけりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,遭いかけりゃ,アイ
カケリゃ,遭いかける,遭いかける,和,,,,,,,,, 24146933237031650

実際の「短単位」の規定

- 原則として「一般」に分類した和語・漢語の**最小単位** 2つの1次結合を1短単位とする
- **最小単位**：
 - 言語学でいう形態素に概ね相当するが、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。
 - [伝ら, 2007]
 - 最小単位は、短単位の認定を行うために規定するものであり、短単位の認定のために持っておくという性質のものである。したがって、この最小単位の境界はCSJのデータには示されていない。
 - [小椋ら, 2004]

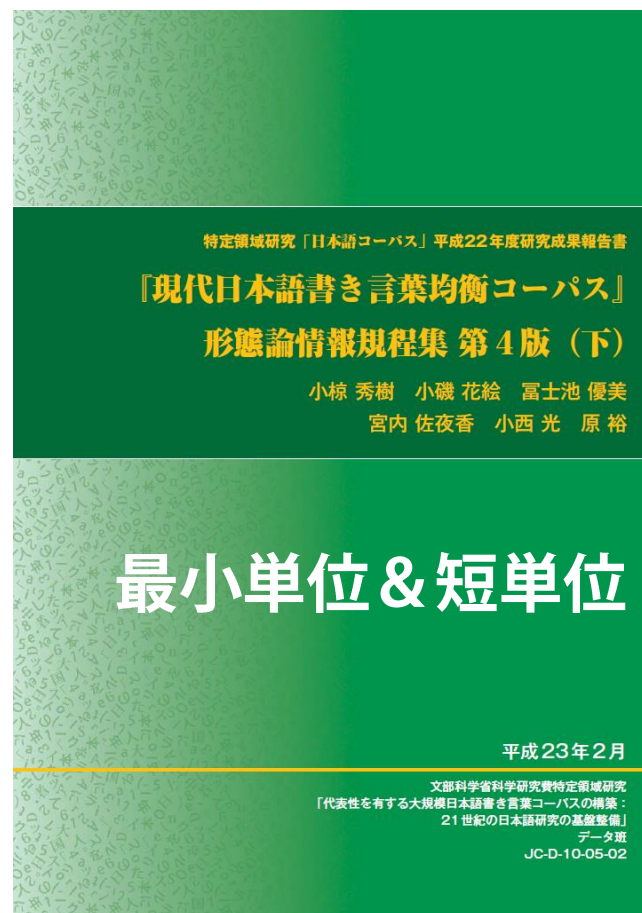
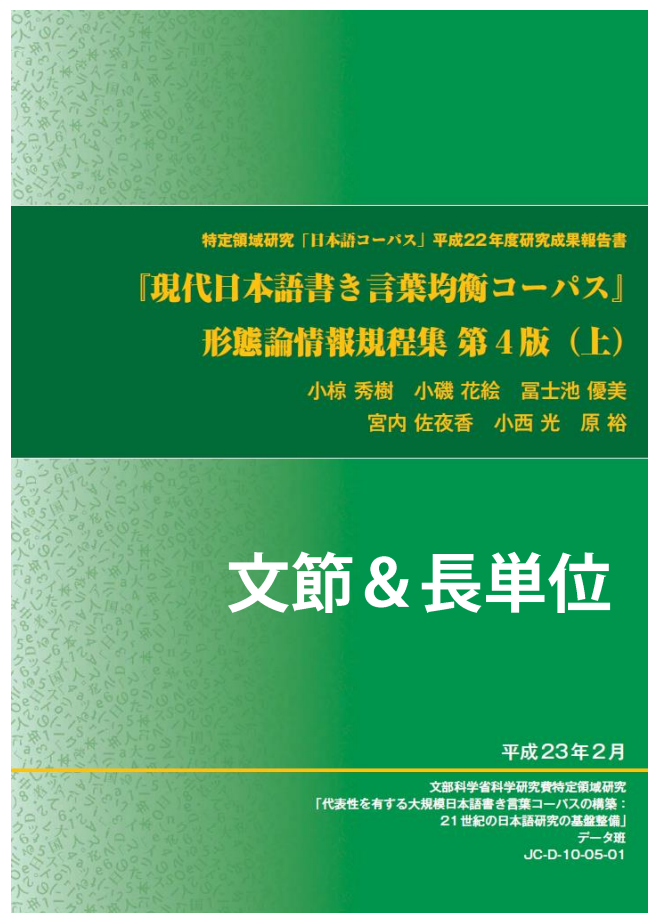
最小単位の分類と結合方法

分類		例	結合方法
一般	和語	豊か 大 雨...	2つまで結合可
	漢語	国 語 研 究 所...	2つまで結合可
	外来語	コール センター...	結合不可
付属要素	接頭的要素	御 各...	結合不可
	接尾的要素	兼ねる がたい 的...	結合不可
その他	記号	、 ・ 。 「 」...	結合不可
	数	一 二 十 百 千...	「数」同士で結合可
	人名・地名	星野 仙一 大阪 六甲...	結合不可
	助詞・助動詞	だ ます か から て の...	結合不可



BCCWJの形態論情報規程集

http://pj.ninjal.ac.jp/corpus_center/bccwj/doc.html#01



この2冊はコーパス開発センターのサイトにおいてあります。

実は，短単位の語彙素には， 構成語情報がついてる

- サキチル-咲き散る
 - サク-咲く
 - チル-散る
- ミダレサク-乱れ咲く
 - ミダレル--乱れる
 - サク-咲く
- サキニオウ-咲き匂う
 - サク-咲く
 - ニオウ 匂う
- サキソメル-咲き初める
 - サク-咲く
 - ソメル-初める
- サキイデル-咲き出でる
 - サク-咲く
 - イデル-出でる

モード: 仮登録 出典:

検索オプション
☒ 完全一致 ☐ 前方一致 ☐ 後方一致 ☐ 部分一致

DDWin goo辞書 要注意語 削除語形

語彙素ID: 28007

語彙素フィルタ: フィルタOFF

語彙素

語彙素ID: 28007 用例 万3:古141:漢44:徳 類: 用 最小単位: ナヤム/ 語彙素: 1/1 自至情報全列ア Lex

語彙素読み: ナヤム 語義: 語種: 和 構成語情報

語彙素: 悩む コメント:

分類語彙素番号 出典: Icpv 状態: 評価: 目: 至:

新規 選択 削除

語形

語形ID: 896228 用例 漢1:*(1) 品詞: 動詞-一般 語形: 4/4 Lex

語形: ナヤブ Check 活用型: 文語四段-バ行

語頭変化型: 語頭変化結合型: コメント:

語末変化型: 語末変化結合型:

代表性: 出典: 古 状態: 評価: 目: 中古 至: 中世

新規 選択 削除

書字形

書字形ID: 22943 書字形: 悩ぶ 仮名形: ナヤブ 代表性: 代

発音形

発音形ID: 229434369 サブID: 1 アクセント型: 発音形: 1/1 Lex

発音形: ナヤブ アクセント結合型:

代表性: 代 出典: 古 状態: コメント:

新規 選択 削除

語彙素数258,277中,
2,344語彙素に付与

UniDicで一番多い F A Q

- Q. UniDicの列が何を出力しているかわからない
- A.
 - rewrite.defをみれば、ソースファイルのcsvの各列が何か分かります.
 - 出力で列が入れ替えられている場合、dicrcを見ればわかる
 - 「こおりつけ！」を解析すると、大体わかります.



rerwrite.defの先頭

```
# node:
# $1: pos1
# $2: pos2
# $3: pos3
# $4: pos4
# $5: cType
# $6: cForm
# $7: lForm
# $8: lemma
# $9: orth
# $10: pron
# $11: orthBase
# $12: pronBase
# $13: goshu
# $14: iType
# $15: iForm
# $16: fType
# $17: fForm
```



dicrcの先頭

```
; List of features  
; f[0]: pos1  
; f[1]: pos2  
; f[2]: pos3  
; f[3]: pos4  
; f[4]: cType  
; f[5]: cForm  
; f[6]: lForm  
; f[7]: lemma  
; f[8]: orth  
; f[9]: pron  
; f[10]: orthBase  
; f[11]: pronBase  
; f[12]: goshu  
; f[13]: iType  
; f[14]: iForm  
; f[15]: fType  
; f[16]: fForm
```

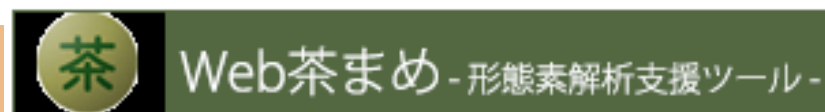


「こおりつけ！」の解析結果

選択辞書:現代語

入力文字:

こおりつけ！



結果

辞書	文境界	書字形 (=表層形)	語彙 素	語彙 読み	品詞	大分類	中分類	小分類	活用型	活用形	発音 形出現形	仮名 形出現形	語種	書字形(基本形)	発音形 (基本形)	仮名 形(基本形)	語形 (基本形)
現代語	B	こおりつけ	凍り付く	コオリツク	動詞-一般	動詞	一般		五段-力行	命令形	コーリツケ	コオリツケ	和	こおりつく	コーリツク	コオリツク	コオリツク
現代語	I	！	！		補助記号-句点	補助記号	句点						記号	！			

「こおりつけ」は、書字形、発音形、仮名形がすべて異なるので見分けるのに便利.



UniDicで一番多い F A Q

- Q. UniDicの列が何を出力しているかわからない
- A.
 - rerite.defをみれば、ソースファイルのcsvの各列が何か分かります。
 - 出力で列が入れ替えられている場合、dicrcを見ればわかる
 - 「こおりつけ！」を解析すると、大体わかります。
 - 解析用UniDic 3.0では、「-Overbose」オプションで列名を併記して出力するようにします。

最後の列に見慣れない数字列がありました…

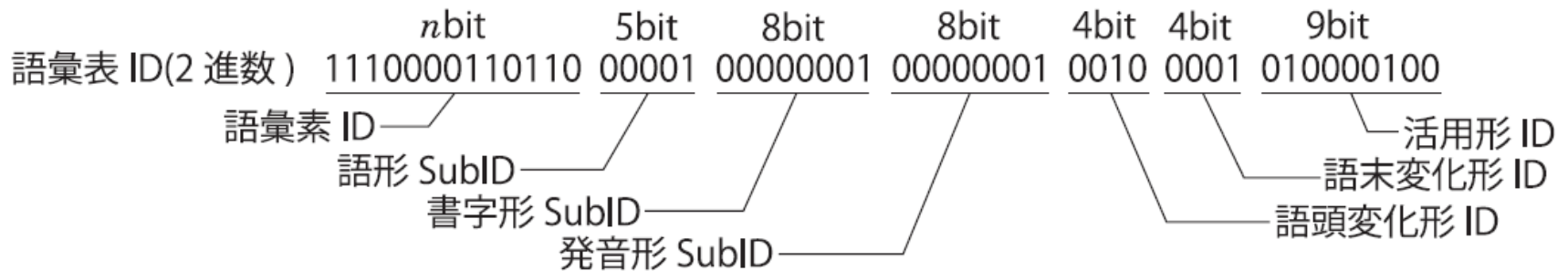
lex.csv

あいかぎ,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,あいかぎ,アイカギ,あいかぎ,あいかぎ,
和,,,,,29695538242048
合いかぎ,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合いかぎ,アイカギ,合いかぎ,合いかぎ,
和,,,,,29695437578752
合い鍵,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合い鍵,アイカギ,合い鍵,合い鍵,
和,,,,,29695504687616
合鍵,0,0,0,名詞,普通名詞,一般,,,,アイカギ,合鍵,合鍵,アイカギ,合鍵,合鍵,和,,,,, 29695471133184
会い掛けれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,会い掛けれ,アイカケ
レ,会い掛ける,会い掛ける,和,,,,, 24146933304140513
遭いかけれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,遭いかけれ,アイカケ
レ,遭いかける,遭いかける,和,,,,, 24146933237031649
遭い掛けれ,0,0,0,動詞,一般,,,下一段-力行,仮定形-一般,アイカケル,会い掛ける,遭い掛けれ,アイカケ
レ,遭い掛ける,遭い掛ける,和,,,,, 24146933270586081
会い掛けりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,会い掛けりゃ,アイ
カケリゃ,会い掛ける,会い掛ける,和,,,,,24146933304140514
遭いかけりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,遭いかけりゃ,アイ
カケリゃ,遭いかける,遭いかける,和,,,,, 24146933237031650
会い掛ける,会い掛ける,和,,,,,*,*,24146933304140514
遭いかけりゃ,0,0,0,動詞,一般,,,下一段-力行,仮定形-融合,アイカケル,会い掛ける,遭いかけりゃ,アイ
カケリゃ,遭いかける,遭いかける,和,,,,, 24146933237031650

語彙表 I D

- UniDic DB中の短単位を一意識別するための I D
 - 3 8 bitシフトすれば，語彙素IDになる

語彙表 ID(10 進数) 1985176867586692



商用利用に料金が発生する「分類語彙表番号」や，
構成語情報などデフォルトで提供していない列項目は，このIDを介し
て別途提供していく予定

こういった項目が秘匿されているかは，新公式サイトで適宜公開していきます。

解析用 UniDic 3.0 公開計画

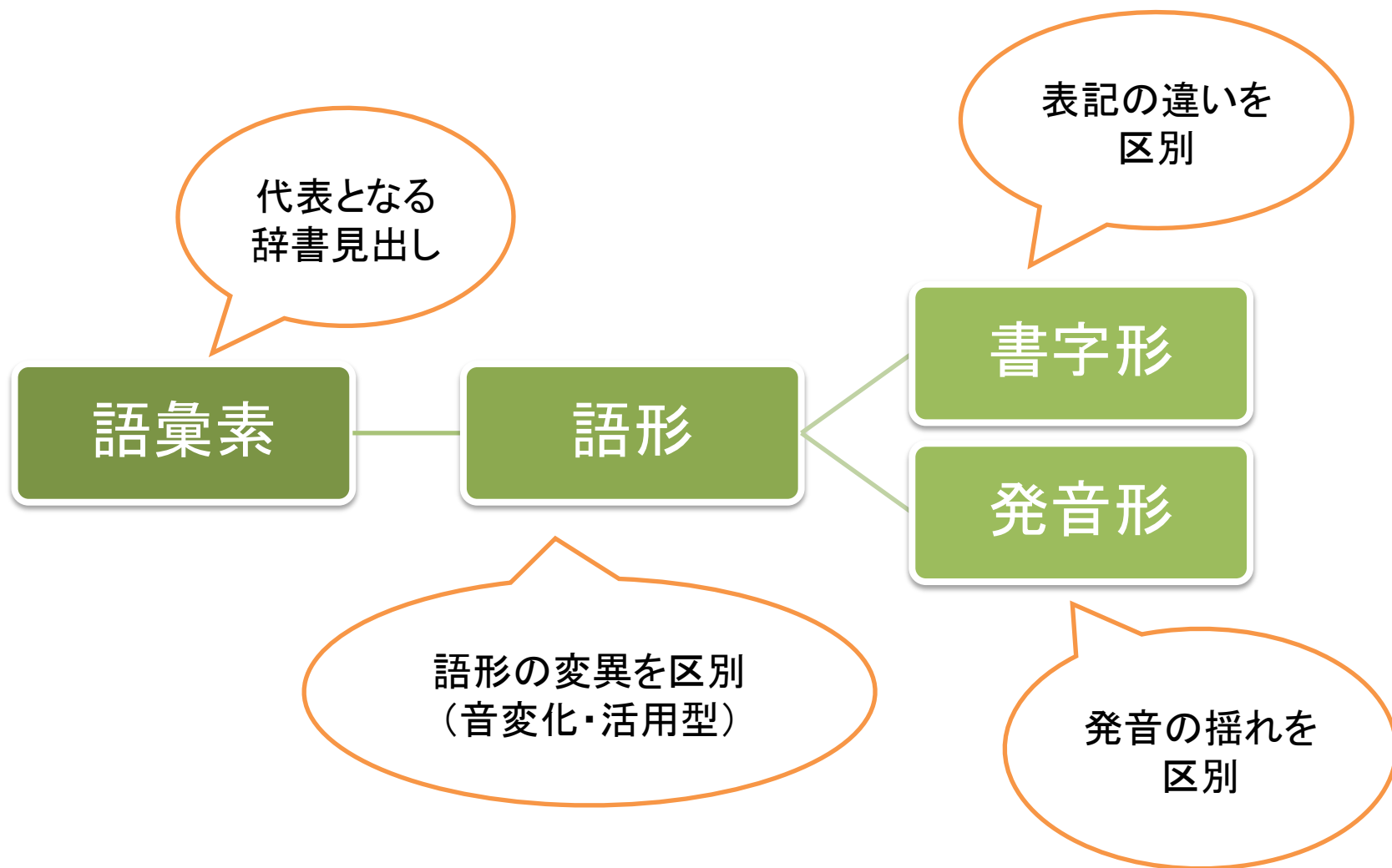
- 新しい公式サイトで公開
 - 用語, FAQの整備
 - 「-Overbose」オプションで列名を表示可能
 - 語彙表IDを介したデフォルト配布項目以外の別途提供
 - 発音形認定性能の強化
 - MeCab用素性テンプレートの再検討
 - α

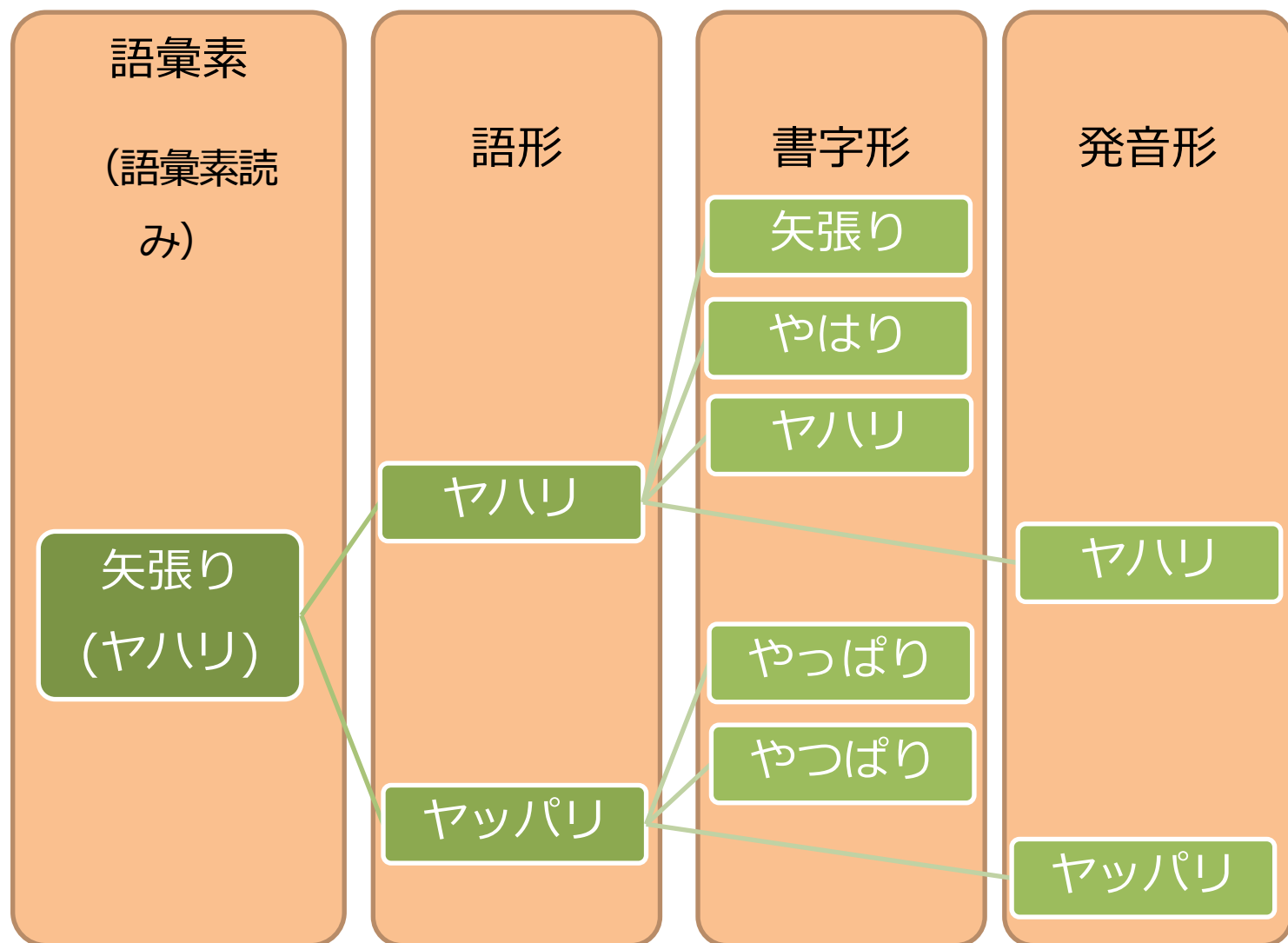
UniDic's Features

@ NAIST CL lab Autumn Camp 2014

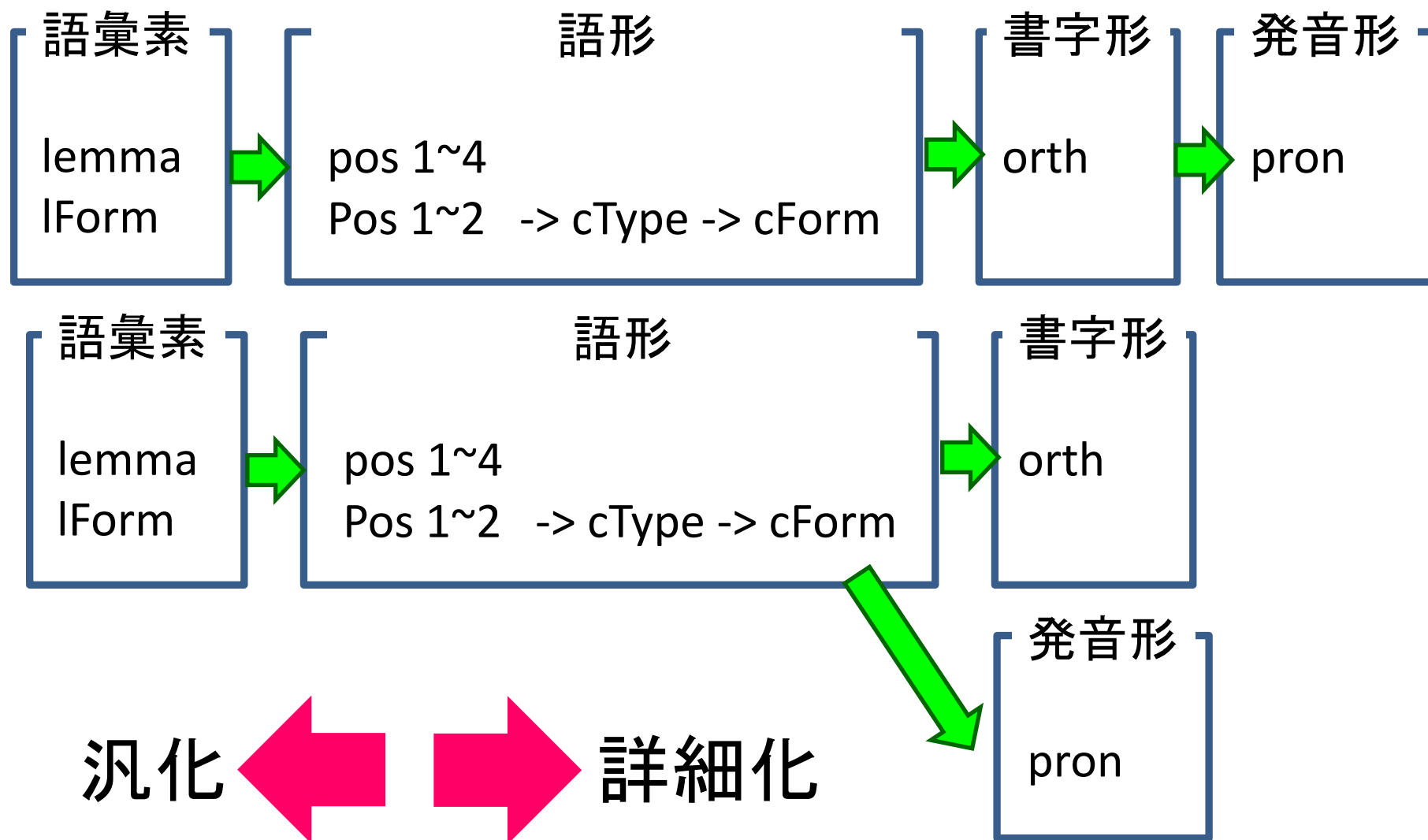
実は、3年前に既に解析用UniDicの素性の改良を行っていました。

見出し語の階層構造

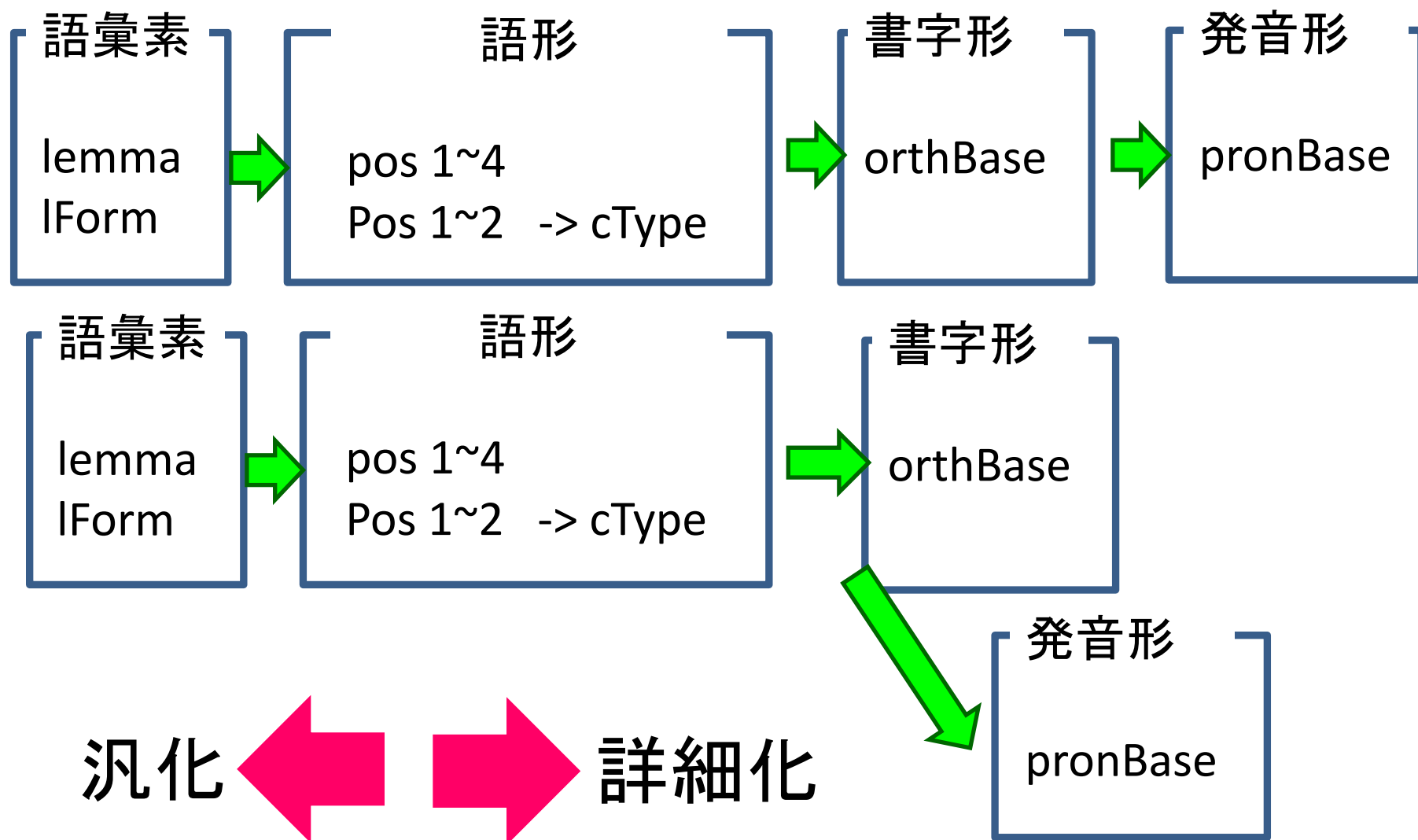




UniDicの階層構造を反映したUnigram素性 (1)



UniDicの階層構造を反映したUnigram素性（2）



連接素性

- 以下の2軸を基礎に考える
- 統語的連節
 - 品詞（もしくはそれに準ずるもの）同士の連接
 - 活用形と品詞の連接
- 語彙的連接
 - 各語ごとの連接しやすさ（品詞+活用(leaf)で代用）

素性数・モデルサイズの削減

- 学習用コーパス：約 1 2 万文
- 評価用コーパス：約 1 万文
 - BCCWJコア+人手修正
 - CSJ
 - RWCP

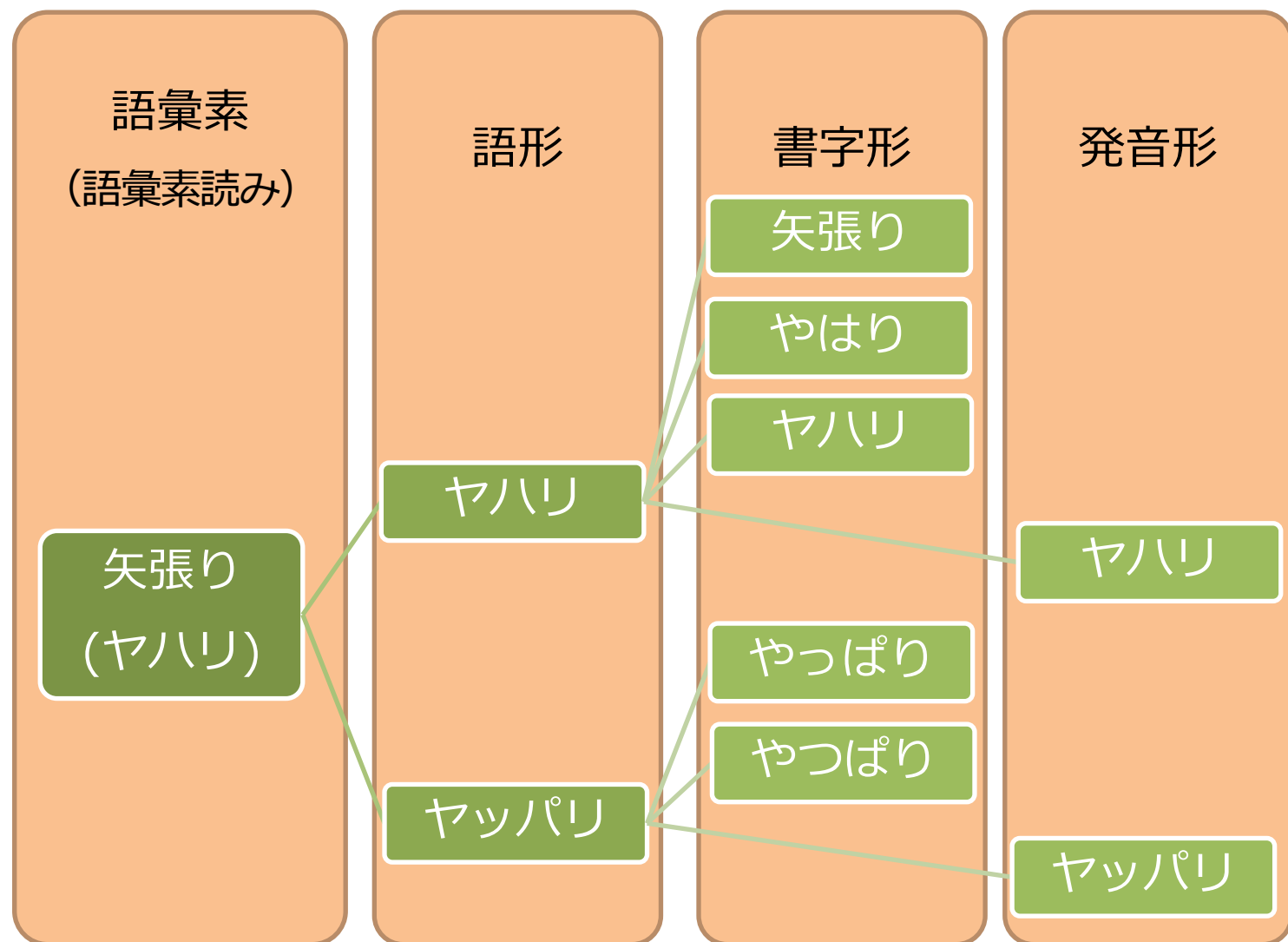
	素性テンプレート 行数 (U)	素性テンプレート 行数 (B)	素性数	モデルサイズ
解析用UniDic 2.1.2	62	201	6,506,081	535M
新素性	93	81	5,743,431	401M



解析性能の向上

- F1値の評価

		境界認定	品詞認定	語彙素認定	発音認定
未知語なし	従来素性	99.46	98.43	98.07	96.68
	新素性	99.47	98.40	98.09	97.52
書字形レベルの未知語あり	従来素性	98.41	97.13	96.74	95.35
	新素性	98.45	97.13	96.80	96.22
語彙素レベルの未知語あり	従来素性	98.75	97.62	97.25	95.86
	新素性	98.76	97.60	97.29	96.72



『UniDic』拡張計画 まとめ

- UniDicの量的拡張と質的拡張
 - 量的拡張： 【目標】年間1,000語の追加
 - 参考： 広辞苑の改版1998→2008年 10,000語の追加
 - 質的拡張：
 - 意味情報（分類語彙表番号）の付与
- 新しい公式サイトで 解析用UniDic3.0 公開
 - 用語, FAQの整備
 - 「-Overbose」オプションで列名を表示可能
 - 語彙表IDを介したデフォルト配布項目以外の別途提供
 - 発音形認定性能の強化
 - MeCab用素性テンプレートの再検討
 - $+a$

