

# 国立国語研究所学術情報リポジトリ

UniDic :

短単位辞書データベースと形態素解析用辞書

メタデータ	言語: Japanese 出版者: 公開日: 2021-06-25 キーワード (Ja): キーワード (En): 作成者: 岡, 照晃 メールアドレス: 所属:
URL	<a href="https://doi.org/10.15084/00003419">https://doi.org/10.15084/00003419</a>



# UniDic – 短単位辞書データベースと形態素解析用辞書 –

C4

岡 照晃（国立国語研究所コーパス開発センター・特任助教）

## 『UniDic』とは

国語研の規定した齊一な言語単位（短単位）と階層的見出し構造に基づく電子化辞書の

### ■ 設計方針

およびその実装としてのリレーションナルデータベース

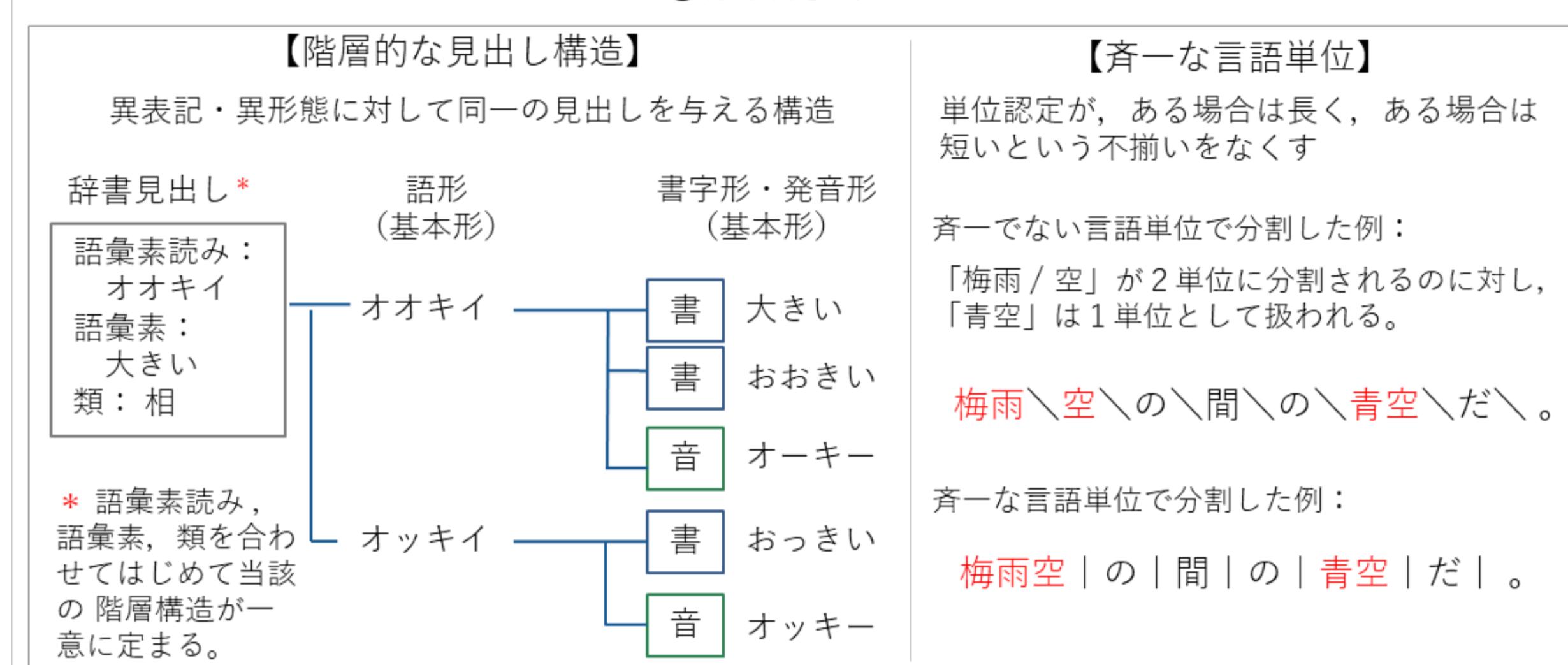
### ■ UniDicデータベース

と、そのデータベースからエクスポートし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

### ■ 解析用UniDic の総称

<http://taku910.github.io/mecab/>

#### ① 設計方針



実装

#### 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	…

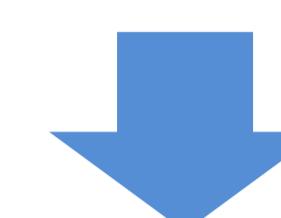
## 『UniDic』の目的

国語研で構築しているコーパスアノテーションの支援

⇒生文を短単位に切り出し、形態論情報を付ける作業の支援

### アノテーション例:

青空におおきい雲が浮かんでいる。

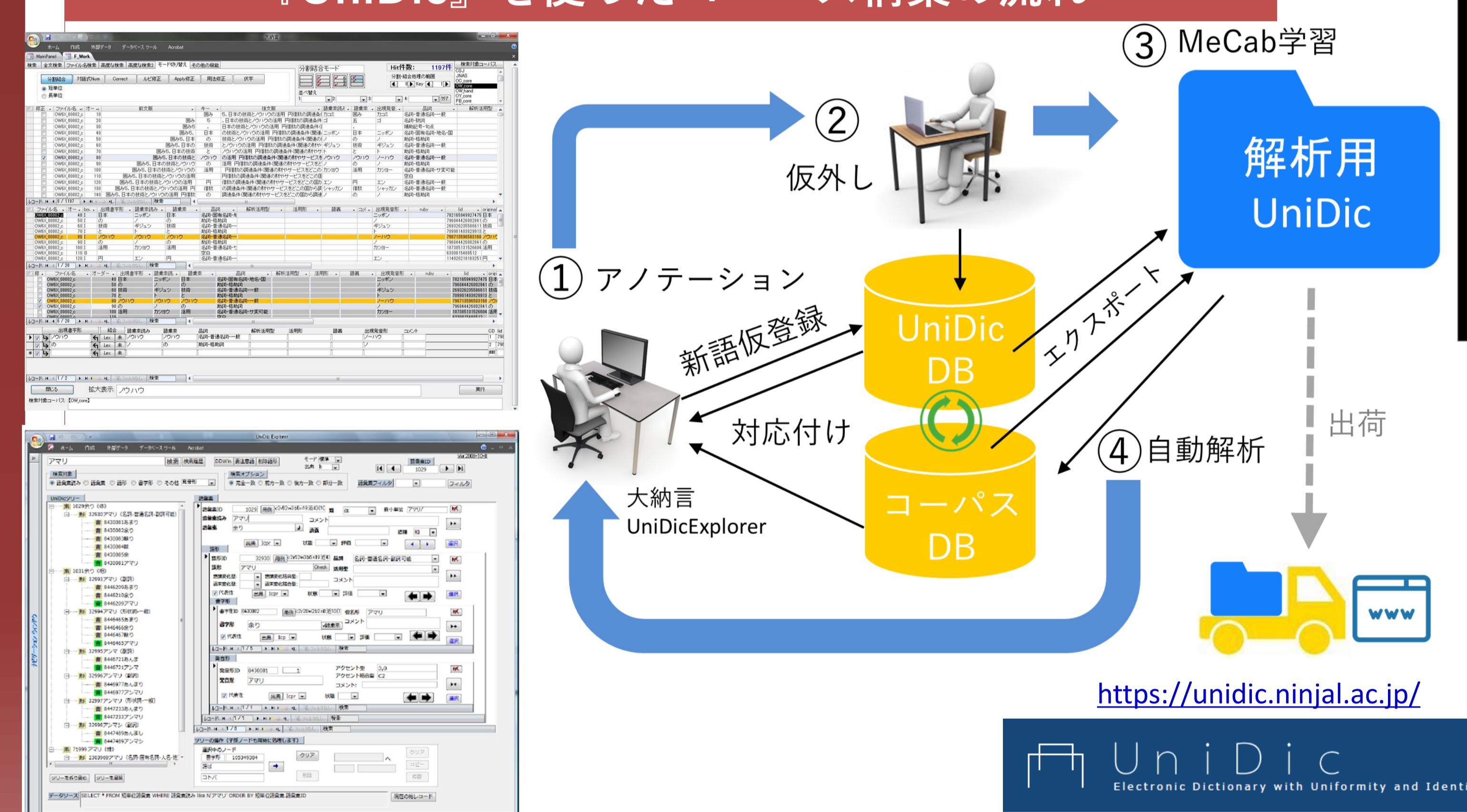


書字形(出現形)	発音形(出現形)	語形(出現形)	品詞	語彙素	語彙素読み	語彙素類	語種
青空	アオゾラ	アオゾラ	名詞-普通名詞-一般	青空	アオゾラ	体	和
に	ニ	ニ	助詞-格助詞	に	ニ	格助	和
おおきい	オーキー	オオキイ	形容詞-一般	大きい	オオキイ	相	和
雲	クモ	クモ	名詞-普通名詞-一般	雲	クモ	体	和
が	ガ	ガ	助詞-格助詞	が	ガ	格助	和
浮かん	ウカン	ウカン	動詞-一般	浮かぶ	ウカブ	用	和
で	デ	デ	助詞-接続助詞	て	て	接助	和
いる	イル	イル	動詞-非自立可能	居る	イル	用	和
。			補助記号-句点	。		補助	記号

## 『UniDic』の持つリッチな形態論情報

品詞、活用型、活用形、語彙素読み、語彙素（+語彙素細分類）、書字形出現形、書字形基本形、発音形出現形、発音形基本形、仮名形出現形、仮名形基本形、語種、語彙素類、語形出現形、語形基本形、語頭変化化型、語頭変化形、語頭変化結合型、語末変化化型、語末変化形、語末変化結合型、アクセント型、アクセント結合型、アクセント修飾型…

## 『UniDic』を使ったコーパス構築の流れ



```
$ echo "青空におおきい雲が浮かんでいる。" | mecab
名詞,普通名詞,一般,,,アオゾラ,青空,アオゾラ
助詞,格助詞,,,ニ,に,ニ
形容詞,一般,,形容詞,連体形-一般,オオキイ,大きい,オーキー
名詞,普通名詞,一般,,,クモ,雲,クモ
助詞,格助詞,,,ガ,が,ガ
動詞,非自立可能,,五段-バ行,連用形-撥音便,ウカブ,浮かぶ,ウカン
助詞,接続助詞,,,デ,で,デ
動詞,非自立可能,,上一段-ア行,終止形-一般,イル,居る,イル
補助動詞,句点,,,。
EOS
```

## Point !

コーパス構築の際は、

- 解析用UniDicを使った自動解析結果（精度は100%ではない）を基に①のアノテーション（形態論情報修正）を行う。
- アノテーションの際、UniDic DBに未登録の短単位が見つかれば、新たにUniDic DBに登録する。
- 解析用辞書は整備中のコーパスとUniDic DBを基に作成されるので、コーパス構築が進むにつれて、解析用UniDicの解析精度は向上していく。

UniDicデータベースは、コーパスのデータベースと参照関係にある。

### コーパスが完成した際：

コーパスデータベース中の短単位は、

- UniDicデータベースに登録されており、
- UniDicデータベース中の一意のエントリを参照する（リンク付けられた）状態になっている。

## 『UniDic データベース』と コーパスデータベースの関係

### 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	語彙素類	語種	…	order	文字
すもも	スモモ	スモモ	名詞 -普通名詞 -一般	李	スモモ	体	和語		10	す
もも	モモ	モモ	名詞 -普通名詞 -一般	桃	モモ	体	和語		20	も
も	モ	モ	助詞 -係助詞	も	モ	係助	和語		30	も

### コーパスデータベース (文字列テーブル)

すもも	スモモ	スモモ	名詞 -普通名詞 -一般	李	スモモ	体	和語		40	も
もも	モモ	モモ	名詞 -普通名詞 -一般	桃	モモ	体	和語		50	も
も	モ	モ	助詞 -係助詞	も	モ	係助	和語		60	も

## 参考文献

- 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).
- 伝 康晴. 多様な目的に適した形態素解析システム用電子化辞書, 人工知能学会誌, Vol.24, No.5, pp.640-646 (2009).
- 伝 康晴, 浅原 正幸: 「リレーションナル・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, pp.77-84 (2001).