『UniDic』を活用した語構造情報付与の試み：『日本語歴史コーパス』を対象に

著者	村山 実和子
雑誌名	言語資源活用ワークショップ発表論文集
巻	3
ページ	267-273
発行年	2018
URL	http://doi.org/10.15084/00001660

『UniDic』を活用した語構造情報付与の試み ―『日本語歴史コーパス』を対象に―

村山 実和子（国立国語研究所 言語変化研究領域）†

Annotation of Word Structures in Japanese using "UniDic" : A Study of the Corpus of Historical Japanese

Miwako Murayama (National Institute for Japanese Language and Linguistics)

要旨

本研究は『日本語歴史コーパス』に出現する合成語に対し，その内部構造に関する情報を新たに追加することで，日本語の語形成研究に使用可能なデータの構築をめざすものである。その方法として，各種コーパスに紐付いた解析用辞書「UniDic」の見出し語に対して，構成語情報を付与することを試みる。その設計方針と有用性を述べるとともに，現状の課題について報告する。

1. はじめに

国立国語研究所（以下，国語研）では，上代から近代までの日本語を通時的に研究するための基礎資料として，『日本語歴史コーパス（以下，CHJ）』（国立国語研究所 2016）の構築を進めている。すべてのテキストを斉一な単位で分割し，詳細な形態論情報を付与している点が特長である。ただし，その単位認定は基本的に現代語のコーパスに準拠するものであり，時代によっては「語」の認定に揺れが生じる場合もある。例えば，ある語の連続を複合語とみなすかどうか，現代語では内省可能なものでも，古代語ではその認定は難しい。そのため，CHJ では現状，奈良～鎌倉時代編までのコーパスでは複合動詞を認定しないが，室町時代編以降は現代語の規程に照らして複合動詞を認めるというように，時代ごとに異なる対処を行っている[1]。以下，用例中の「｜」は単位の境界を表しており，(1a)の［聞き入れ］が四段動詞「聞く」の連用形＋下二段動詞「入れる」の未然形，と分割されるのに対し，(1b)の［聞き入れ］は，下二段動詞「聞き入れる」の未然形として処理される。

(1) a. ｜御前｜近き｜人｜など｜の｜けしき｜ばみ｜言ふ｜を｜も｜聞き｜入れ｜ず｜
【出典】CHJ サンプル ID：20-枕草 1001_00120『枕草子』
b. ｜蛇｜情強｜に｜し｜て｜少し｜も｜聞き入れ（qiqiire）｜なんだ｜に｜因っ｜て｜
【出典】CHJ サンプル ID：40-天伊 1593_00067『天草版伊曾保物語』

また，単位認定の規程上（→3.2 節），前項または後項の独立性や語種によって，形態的に同一のものであっても，単独で扱う場合と，語の一部として扱う場合がある点にも注意が必要である。これは接頭・接尾語の類にしばしば見られるものである。例に挙げた「めく」は UniDic では単独の要素（品詞：接尾語-動詞的）として扱われている。しかし，上接語が象徴語など拘束的な要素である場合には，語の内部要素として処理される。

† m-murayama@ninjal.ac.jp

[1]『日本語歴史コーパス平安時代編』および『同　鎌倉時代編』形態論情報規程集の記述に拠る

(2) a. ｜鳥｜の｜声｜など｜も｜こと｜の｜外｜に｜春｜めき｜て｜
【出典】CHJ サンプル ID：30-徒然 1336_01019 『徒然草』
b. ｜濡れ｜たる｜やう｜なる｜葉｜の｜上｜に｜きらめき｜たる｜こそ｜
【出典】CHJ サンプル ID：30-徒然 1336_01137 『徒然草』

上代から近代まで幅広い時代の資料を検索対象とする CHJ においては，形態論情報付与のために一定の基準を設けることは運用上必要不可欠であるといえる。ただしそれは現代語の状況に立脚したものであるため，基本ルールから逸脱するものは時代ごと、あるいは語ごとに例外的な基準を設けることで対応している（それは各時代の規程集に明文化される）。そのような構築上の背景に留意することが，コーパス利用の前提となることは言うまでもない。しかしながら，このように場合によって単位が異なるケースをみると，語構成要素の情報（以下，構成語情報とする）を付与することで，データが扱いやすくなる場合もあるように思う。また，公開済のコーパスをベースとして構成語情報を追加することによって，時代・資料・品詞など様々な条件下で，前項・後項のバリエーションや出現状況の調査が可能になることが期待される。日本語の歴史的な語形成を考察する手がかりとして有用な情報であるといえよう。本発表では，CHJ に出現する語を対象に，構成語情報を付与する方法と，その有効性を検討し，現状の課題についても報告する。

２．関連研究

古代日本語の形容詞・形容動詞の語構造に着目し，計量的な分析を行うものとして，村田菜穂子氏の一連の研究が挙げられる。古代語に加え，中～近世期の資料に関しても調査が進められており，語構造に関する情報を付した語彙表が順次報告されている。合成語（複合語・派生語）の変遷を知るうえでも，歴史資料に見える語の分析方針を検討するうえでも重要な資料であるといえる。それらの語彙表には，出典，用例数，活用型，語構造についての情報が付されるが，あくまで一覧化したものであり，各語が用いられる構文の環境や，用法などをただちに参照できるものではない。公開中のコーパスのテキストに対して，各語の構成語情報を付与することができれば，データの汎用性，および再現性はより高くなることが期待される。

また動詞に関しては，オンラインデータベース「複合動詞レキシコン[2]」（国立国語研究所，2015）が公開されている。複合語自体の検索のみならず，前項・後項それぞれに検索でき，『現代日本語書き言葉均衡コーパス（BCCWJ）[3]』の例文と関連づけるなど，利便性の高いデータ提供を行っている。ただし，このデータベースに収録される語彙は「現在の日本語で一般的に使われている複合動詞」（2,790 語）が対象であり，「古語・古典語」は対象外であることが明記されている。

さて，本研究では，コーパスのテキストそのものではなく，コーパスと関連付けられた電子化辞書「UniDic」に，別途，語構造に関する情報を持たせることを計画している。この手法は，淺尾（2017）で提案されており，フリーライセンスで提供している UniDic の内容をもとに 199,098 項目に対して語構成情報を付与する内容が示されている。その研究結果を反

[2] http://vvlexicon.ninjal.ac.jp/

[3] http://pj.ninjal.ac.jp/corpus_center/bccwj/

映した検索ツールが試験公開されている[4]が，そのように品詞によらず，網羅的に日本語の語構成情報を付与したデータベースは他に類をみないものである。ただし，使用したデータのバージョンから，CHJ で追加された項目は含まれていないものと見られる。本研究では，以上の研究を参照しながら，歴史資料に出現する語に対する構成語情報の付与について検討を行う。

３．対象とするデータの概要

3.1 『日本語歴史コーパス（CHJ）』

本研究では，CHJ を対象として対象語の収集，分析，情報の付与を行う。表 1 に，CHJ に収録されているコーパスの一覧と延べ語数を示す（2018 年 7 月現在）[5]。

表 1　調査対象

サブコーパス	収録作品	延べ語数（短単位）
奈良時代編	Ⅰ万葉集	98,499
平安時代編	（仮名文学作品）	856,682
鎌倉時代編	Ⅰ説話・随筆／Ⅱ日記・紀行	821,010
室町時代編	Ⅰ狂言／Ⅱキリシタン資料	358,419
江戸時代編	Ⅰ洒落本	204,519
明治・大正編	Ⅰ雑誌	12523,750

CHJ は日本語の通時的研究のための基礎資料として整備が進められている。2018 年 3 月に，『室町時代編IIキリシタン資料』『江戸時代編I洒落本』のデータが加わり，部分的ではあるが，上代から近代まで一本化した資料を検索対象として扱えるようになった。これらのコーパスは，先述のとおり国語研の規定する斉一な単位（短単位）によってテキストを分割し，各単位に対して，形態論情報が付与されている。その形態素解析は，国語研が整備している電子化辞書 UniDic を利用して行われる。コーパス本文の短単位に付与された形態論情報と，UniDic に立項された見出し語の情報は，国語研の形態論情報データベースによって関連づけられている（小木曽・中村 2011）。したがって，UniDic の見出し語（語彙素）に対して，その構成語情報を付与することができれば，各語がどのコーパスにどのように出現するかについても容易に参照可能となる。

3.2 UniDic の見出し語について

UniDic の特長として、以下の 2 点が挙げられる（小椋ほか 2011）。

（ア）一語の認定基準がわかりやすく、判断の揺れが少ない「短単位」を見出し語として採用する。

（イ）表記や語形の違いに関わらず、同じ語であれば同一の見出しを与えるという方針をとり、語を階層化した形で登録している。階層構造の最上位を「語彙素」（辞書の見出しに相当）とし，語彙素＞語形（語形の違いを区別する層）＞書字形（表記の違いを区別する層）という階層を設ける。

「短単位」は，用例収集を目的とし，言語の形態論的側面に注目して規定された言語単位である。短単位の認定にあたっては，まず「最小単位」が規定される。その上で，文節

[4] http://asaokitan.net/jmorph/

[5]「語彙統計：バージョン 2017.3」（http://pj.ninjal.ac.jp/corpus_center/chj/201703.html）の数値に拠る。記号や空白はのぞく。万葉集，キリシタン資料，洒落本の語数は公表前のため，稿者の調査に拠る数値を示す。

の範囲内で短単位の規程に基づいて結合させる（又は結合させない）ことにより，短単位が認定される。この「最小単位」は、和語・漢語・外来語・記号・人名・地名の種類によって以下のように分類されている。その多くは、本研究で扱う語構成要素となりうることから、形態素境界や構成要素の判断をするにあたって、適宜参照する。

表 2　最小単位の分類（小椋ほか 2011、上 pp.7）

分類		例
一般		和語：豊か　大　雨 … 漢語：国語　研究　所 … 外来語：コール　センター　オレンジ …
数		一　二　十　百　千 …
その他	付属要素	接頭的要素：相　御　各 … 接尾的要素：兼ねる　がたい　的 …
	助詞・助動詞	だ　ます　か　から　て　の …
	人名・地名	星野　仙一　大阪　六甲 …
	記号	A　B　ω　イ　ロ　ア　JR …

4．作業方針と今後の課題

4.1　構成語情報の作成と UniDic との連携

本研究で目指す構成語情報付与のイメージを，図 1 に示した。構成語情報として，以下の情報を必須項目とする。

(ア) 語彙素 ID
(イ) 分類｛複合／派生／不明｝
(ウ) 連番（前項・後項等の位置を定めるもの）
(エ) 構成語_語彙素 ID（各構成要素の語彙素 ID）

UniDic に登録された語を一意に同定することのできる「語彙素 ID」の情報を利用することで，その語の内部構造に関する情報を入力する。図のように，各要素が UniDic に登録されている語であれば，その情報にリンクできるよう，語彙素 ID を入力する。語彙素 ID を利用して，UniDic の情報と対応づけることによって，UniDic に紐付いたコーパスの用例も適宜参照することが可能になる。

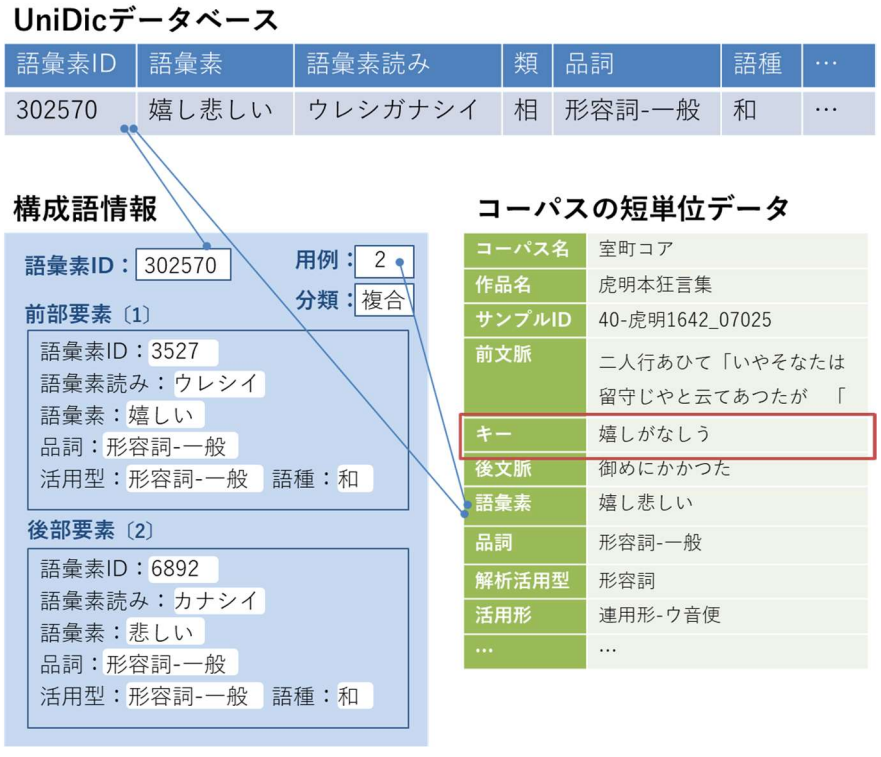

図 1　　UniDic・コーパスデータと対応付けた構成語情報のイメージ

なお，UniDic の編集を行うためのツールである「UniDic Explorer」には，語彙素に付与可能な情報の一つとして，「構成語情報」の入力欄が用意されている[6]（短単位で複合語となる語に対し，それを構成する複数の語彙素 ID が入力可能）。この機能を生かして UniDic に直接入力することも検討したが，その場合，各構成要素が UniDic の見出し語として立項されていることが条件となる。そのため，短単位以下の語を構成要素として付したい場合，UniDic の規定にそぐわない形式を登録するか，あるいは空欄とすることになる。そこで本研究では，CHJ に出現する語を抽出したうえで，必要な情報を入力した別表を作成し，対応づけることとした。本研究で作成したデータのうち，反映できるものについては，UniDic に還元することも視野に入れている[7]。なお、現在，試行として CHJ に出現する形容詞を対象としてデータの入力を進めている（表 3）。その作業内容の一部を、表 4 に示す（この別表と UniDic を対応付けたデータベースに関しては当日のポスターで紹介する）。

表 3　CHJ 室町時代編・江戸時代編に出現する動詞・形容詞の数

		キリシタン	狂言	洒落本
動詞	延べ	21,730	47,121	32,082
	異なり	1,959	1,868	2,167
形容詞	延べ	2,145	5,325	5,156
	異なり	221	238	347

表 4　UniDic と対応付ける構成語情報の入力例

コーパス	語彙素ID	語彙素	語彙素読み	類	語種	最小単位	分類	連番	語彙素ID（構成要素）	類	語彙素	語彙素読み
江戸	268920	甲斐無い	カイナイ	相	和	カイ/ナイ	複合	1	5615	体	甲斐	カイ
江戸	268920	甲斐無い	カイナイ	相	和	カイ/ナイ	複合	2	27442	相	無い	ナイ
江戸	290140	掛かりがましい	カカリガマシイ	相	和	カカリ/ガマシイ	派生	1	6016	用	掛かる	カカル
江戸	290140	掛かりがましい	カカリガマシイ	相	和	カカリ/ガマシイ	派生	2	8140	接尾	がましい	ガマシイ
江戸	93415	限り無い	カギリナイ	相	和	カギリ/ナイ	複合	1	6117	体	限り	カギリ
江戸	93415	限り無い	カギリナイ	相	和	カギリ/ナイ	複合	2	27442	相	無い	ナイ
江戸	6631	堅苦しい	カタクルシイ	相	和	カタ/クルシイ	派生	1	6603	固い	カタイ	カタイ
江戸	6631	堅苦しい	カタクルシイ	相	和	カタ/クルシイ	派生	2	保留	–	–	–
江戸	8361	聞き苦しい	キキグルシイ	相	和	キキ/グルシイ	複合	1	8399	用	聞く	キク
江戸	8361	聞き苦しい	キキグルシイ	相	和	キキ/グルシイ	複合	2	10523	相	苦しい	クルシイ

4.2　作業上の課題

現在，上記のように抽出した語彙に対して分類・情報付与の作業を進めているが，その過程で課題となった事例についていくつか紹介する。

語の認定に揺れが生じる例として，「モノ-」（名詞「物」から）、「ウス（ウソ）-」（形容詞「薄い」の語幹から）のように，一概にその品詞を確定できないものが存する。

(3)　a. もの おもはしき その人は嶋の内なる大見やに。名高き若づめ大角とて
【出典】CHJ サンプル ID：52-洒落 1826_01026『色深狭睡夢』

b. うそ 甘い（vfoamai）物を食らうた上なれば，何かは良からう
【出典】CHJ サンプル ID：40-天伊 1593_00002『天草版伊曽保物語』

また，個別の例ではあるが，語源俗解などにより，発生時とその後の時代とで，異なる語として認識されるものも存する。例として、現代語における「形容詞語幹＋クルシイ」は，中

[6] 複合動詞について試験的に入力されているものもあるが、基本的に未入力の状態であり、現時点では公開対象にない情報である。

[7] 語彙素 ID を使用し、連番によって要素の位置を定める考え方はこの「構成語情報」に拠っている。データを還元する可能性があることから、共通の仕様となるよう調整した。ただし、UniDic Explorer では複合語を対象としているため、複合・派生等の分類は行なわれない。

世～近世にかけては「クロシイ」という接尾辞による派生語であったが、近世後期以降、「～苦しい」と認識されるようになった形式である（村山 2012）。このような場合、共時的には接尾辞による派生語とみなしたいが、近現代の例では「苦しい」とするのが望ましい。

(4) a. なんじやいなかたくろしいその羽織もぬぎなんせんか
【出典】CHJ サンプル ID：52-洒落 1757_01005『聖遊廓』
b. 『まあ、そんな固苦しいことを言はないだつて、いゝでせう。』
【出典】CHJ サンプル ID：60M 婦倶 1925_12020『女人群像』

(5)の例はやや特殊なパターンではあるが、(4)の場合は、複合動詞の後項についても同様に、動詞と見るか接尾辞と見るか、判断が割れるものが現れうる。これらの形式については，先行研究も参照しながらなるべくその時代に即した分類を心がけたいものの，時代や語によって分類を変えるよりも，同形式として収集できるほうが結果として望ましいものと考える。したがって，まずは一定の基準にしたがって分類しておき，構成語情報には特記事項を加えるか，または注意が必要な形式としてリストアップすることで対処に代えたい。

５．おわりに

本稿では、『日本語歴史コーパス』に出現する語に対し、語構成要素に関する情報を付与する試みについて紹介した。淺尾（2017）でも述べられているが、やはり最大の課題は、語構造をどのように認定するかということになる。特に歴史的な資料を扱う上では，その語がどのように発生したかを明らかにすることには限界がある。一方で現代語の視点から，いかなる要素の結合であるかを分析することには，余計な解釈を付け足してしまう恐 れもあるといえよう[8]。とりわけ派生語については，何を接辞とみなすか，明確な基準が必要である。実作業を通して得られた課題について検討しつつ，一定の基準を策定し，複合・派生語の研究に資するデータとなるよう引き続き整備を進めていく。

謝　辞

本研究は JSPS 科研費 17K13471「派生・複合情報を付与した歴史コーパスによる語形成の歴史的研究」による成果の一部である。また，データベースの関連付けにあたって，中村壮範氏（国立国語研究所，コーパス開発センター）にご協力いただいた。記して感謝申し上げる。

文　献

淺尾仁彦（2017）.「日本語語構成情報データベースの構築」『言語資源活用ワークショップ 2016 発表論文集』，pp.120-125

池上尚（2016）.「中古語複合形容詞［名詞＋評価形容詞］の一語性」『国語語彙史の研究』35，pp.39-55

小木曽智信・中村壮範（2011）．『『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 改訂版』国立国語研究所（特定領域研究「日本語コーパス」平成 22 年度研究成果報告書）

小椋秀樹・小磯花絵・冨士池優美・宮内佐夜香・小西光・原裕 (2011). 『『現代日本語書き言葉均衡コーパス』形態論情報規程集第 4 版（下)』，特定領域研究「日本語コーパス」

[8] 語の認定には認識の差も関わる。前川・村田（2015）では、コーパスを使った語彙研究の有用性を示すとともに、品詞性や複合度については作成者や使用者の間で差が生じることが指摘されている。

平成 22 年度研究成果報告書 http://pj.ninjal.ac.jp/corpus center/bccwj/doc/report/JC-D-10-05-02.pdf.
国立国語研究所コーパス開発センター（池上尚）編（2016）.『日本語歴史コーパス　平安時代編』形態論情報規定集,　http://pj.ninjal.ac.jp/corpus_center/chj/doc/morph-heian-2016.pdf よりダウンロード可能
国立国語研究所コーパス開発センター（鴻野知暁）編（2017）.『日本語歴史コーパス　鎌倉時代編』短単位規定集 Ver.1.0, http://pj.ninjal.ac.jp/corpus_center/chj/doc/morph_kamakura_v1_0.pdf よりダウンロード可能
国立国語研究所(編) (2018).『日本語歴史コーパス』, (バージョン 2018.03,中納言バージョン 2.2.1) https://chunagon.ninjal.ac.jp/
前川武・村田菜穂子（2015）.「索引とコーパスを利用した形容詞語彙の採取について」『国語語彙史の研究』34 , pp.227-241
村田菜穂子（2005）.『形容詞・形容動詞の語彙論的研究』, 和泉書院.
村田菜穂子（2015）.「中古形容詞に見られる複合的方式についての一考察」『国語語彙史の研究』34 , pp.91-109
村山実和子（2012）.「接尾辞クロシイ考」『日本語の研究』8-4 , pp.16-30

関連 URL

コーパス検索アプリケーション『中納言』	https://chunagon.ninjal.ac.jp/
『日本語歴史コーパス』	http://pj.ninjal.ac.jp/corpus_center/chj/
『UniDic』	http://unidic.ninjal.ac.jp/