国立国語研究所学術情報リポジトリ

日本語語構成情報データベースの構築

著者	淺尾 仁彦
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	120-125
発行年	2017
URL	http://doi.org/10.15084/00001465

日本語語構成情報データベースの構築

淺尾仁彦 (情報通信研究機構) *

Constructing a Database of Word Structures in Japanese

Yoshihiko Asao
(National Institute of Information and Communications Technology (NICT))

要旨

本研究では、形態素解析辞書『UniDic』への語構成情報の付与について紹介する。語構成情報とは、例えば名詞「招き猫」は、動詞「招く」と名詞「猫」の複合語であるといった情報を指す。日本語について語構成の情報が付与された公開データベースは、複合動詞など特定のカテゴリに限定されたものを別とすれば、管見のかぎり存在しない。このデータベースでは、『UniDic』に対して語構成情報をできるだけ網羅的に付与し、品詞・語種・アクセントなど『UniDic』に元々含まれている情報と組み合わせることにより、「名詞＋動詞の複合名詞」、「アクセントが無核の動詞の名詞化で、アクセントが有核のもの」といった複雑な条件での検索を行うことができ、語彙論・音韻論・形態論などの多様な分野で言語資源として活用可能である。合わせて、開発中の検索インタフェースの紹介を行う。

1. はじめに

近年、『UniDic』のような言語学的な観点の取り入れられた形態素解析辞書や、『日本語書き言葉均衡コーパス』(『BCCWJ』) をはじめとする形態素解析済みの大規模なコーパスが整備され、多様な分野の言語研究に活用できるようになった。

しかしながら、現在のところ、これらの言語資源からは形態論の研究で必要とされる情報を限定的にしか得ることができない。これはいわゆる「形態素解析」で解析される単位が実際には言語学的な意味での形態素（意味を担う最小単位）ではなく、それより大きい単位であるためである。一般に、形態素解析辞書、あるいはそれを用いた解析結果からは、屈折形態論に関する情報は得ることができるが、派生形態論に関する情報 (本研究では語構成情報と呼ぶ) は得ることができない。例えば、「出た」が動詞「出る」の連用形「出-」と助動詞「-た」から成るという情報は得られるが、「家出」は「家出」全体がそのまま辞書登録されており、これが名詞「家」と動詞「出る」の複合であるという情報は得られない。このため、例えば名詞と動詞が複合しているものを検索するという操作は、既存の言語資源では簡単に行うことができない。

形態素解析が言語学的な意味での形態素まで文を分割しないことには一定の合理性があると考えられる。例えば「持つ」という動詞の用法を調査する際に、「気持ち」という語の用例が全て動詞「持つ」の用例として扱われるのは通常、コーパス検索において期待される動作ではない (小椋ほか 2007)。一方で、語彙論、形態論、音韻論の研究では、しばしば語彙項目の内部構造が議論の対象となるため、既存のコーパスや辞書で語彙項目（としてその辞書で扱われているもの）の内部構造の情報に容易にアクセスできないことは、これらの分野における形態素解析やコーパスの有用性の限界となってしまう。

* asao@nict.go.jp

そこで、本研究では、形態素解析辞書である『UniDic』(伝ほか 2007) をベースとし、語構成情報を付与したデータベースを構築し、加工・再配布自由なデータとして順次公開する。また、合わせて、このデータに容易にアクセスできるよう、検索ツールを開発する。

本稿の構成は以下の通りである。2 節で、本研究で開発するデータの設計について述べる。3 節で、現在までのデータ構築状況について述べる。4 節では開発中の検索ツールについて紹介する。5 節でまとめと今後の課題について述べる。

1.1 関連研究

管見のかぎり、網羅的かつフリーで利用可能な日本語の語構成情報データベースは存在しないが、関連する言語資源として以下のものがある。『BCCWJ』の「短単位」は、ほぼ言語学的な意味での形態素に対応する「最小単位」に基づき、その組み合わせとして定義されており (小椋ほか 2011)、本研究で付与する語構成情報はこの「最小単位」のもつ情報と重なる部分がある（この最小単位自体は、公開されている形態素解析辞書では利用できない)。ただし、本研究で認定する語構成情報は、『BCCWJ』で定義されている「最小単位」と一致させることが目的ではない。また、後述のように単に形態素を認定するだけでなく、その範疇などについても、他の項目と関連づけることによって情報を付与することを意図している。

複合動詞については語構成情報を含むデータベース『複合動詞レキシコン』が公開されている (国立国語研究所 2015)。このデータベースは項構造や例文等の情報が充実する一方、収録されている項目は頻度の高いものに限定されているなど [(1)]、やや本研究とは目的が異なると考える。

英語、ドイツ語、オランダ語に関しては『CELEX2』という語彙データベースがあり (Baayen et al. 1996)、フリーではないが、各言語について網羅的な語構成情報が利用可能である。本研究はこのデータベースを 1 つのモデルとしている。

2. 設計

本研究では、語構成情報を形態素解析辞書『UniDic』をベースとして構築する。『UniDic』をベースとするメリットは、(i) ライセンス上、自由に加工・再配布を行うことができる言語資源であること、(ii) 『BCCWJ』などに付与された形態論情報と基本的に対応づけが可能なこと、(iii) 辞書への単語の収録基準が比較的明確であること、(iv) 語種やアクセントなど言語研究に有用な情報があらかじめ付与されていること、などが挙げられる。

本研究でそれぞれの語彙項目に対して付与される情報は以下の通りである。また、本研究で付与される情報のイメージ図を図 1 に示した。

- 形態素境界情報
- 語を構成する各形態素へのリンク
- 語形成に関わる付属情報（連濁、音便など）

例えば「飛び箱（とびばこ)」という項目に対しては以下のような情報が付与される。

- **飛び箱**
  - 境界情報: 飛び/箱, とび/ばこ
  - 形態素へのリンク: 「飛ぶ（とぶ)」、「箱（はこ)」へのリンク
  - 付属情報: 連濁

(1) 複合動詞レキシコンに収録されている複合動詞は 2,759 語だが、本研究では現在 7,842 語の複合動詞を認定している。

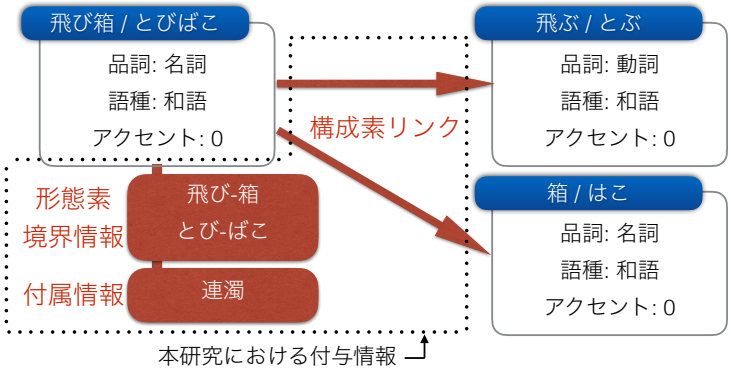

図 1　検索ツールの開発中の画面

リンク先の形態素が別途『UniDic』の見出し語として立項されている場合は、リンク先はその見出し語に紐付けられる。そのため、『UniDic』に付与されている情報を利用し、「飛び箱」の前部要素が動詞であり、アクセント型は 0 であるといった情報にアクセスすることも可能となっている。また、境界情報と形態素へのリンクを別々に付与することにより、例えばこの複合語における後部要素の形式が「ばこ」であることと、この形態素の単独での形が「はこ」であることの両方の情報にアクセス可能となっている。

合わせて、連濁など、形態音韻論に関する付属情報を付している。連濁の有無は、「ばこ」と「はこ」のような形を比較することによって機械的な判定を行うことも基本的に可能だが、検索の便宜のため直接ラベルを付与している。現状認めている付属情報には、連濁、半濁音化、音便 (促音便、撥音便)、音挿入 (促音挿入、撥音挿入、ノ挿入)、被覆形がある。

語を構成する動詞連用形および形容詞（イ形容詞）・形状詞（ナ形容詞など）語幹に関してはそれぞれ動詞・形容詞・形状詞へのリンクを付与する。

- **落ち込み**
  - 境界情報: 落ち/込み, おち/こみ
  - 形態素へのリンク: 「落ちる（おちる)」、「込む（こむ)」へのリンク
  - 付属情報: —
- **狭苦しい**
  - 境界情報: 狭/苦しい, せま/くるしい
  - 形態素へのリンク: 「狭い（せまい)」、「苦しい（くるしい)」へのリンク
  - 付属情報: —

動詞連用形や形容詞・形状詞語幹は、同じ形の名詞が立項されていても、動詞・形容詞・形状詞へのリンクを優先する（例えば形態素「落ち」は動詞の「落ちる」にリンクされ、名詞の「落ち」にはリンクされない)。そのため、名詞「落ち込み」から動詞「落ちる」、動詞「込む」へのリンクはあるが、動詞「落ち込む」や名詞「落ち」へのリンクはないことに注意が必要である。

3.　現在までの構築状況

本研究では、フリーなライセンスで提供されている『UniDic』の形態素解析用辞書 (unidic-mecab 2.1.2) に掲載されている 756,463 項目を、表記のゆれや活用の違いなどを吸収した

199,098 項目にまとめた [2]。この 199,098 項目について語構成情報を付与する。付与にあたっては、機械的な判定手法を援用しつつ、手作業によるチェックも行う。

現在までに、構成要素も『UniDic』に立項されている複合語を優先してデータの構築を行っている。原稿執筆時点では、複合動詞・複合形容詞については人手でのチェックを終えているが、複合名詞・複合形状詞については一部、人手でのチェックが残っている。現段階での暫定的な種類別の語数を、複合語を中心に表 1 にまとめた。表の数値は、今後のデータの修正によって変動する可能性がある。また、以下のようなものについては、語構成情報を整備中であり、表 1 では単純語と合わせて「その他/未処理」に含まれている。

- 派生接辞を含むもの　例：「小骨（こぼね）」「厚み（あつみ）」「羨ましい（うらやましい）」
- 漢字語根を含むもの　例：「出版（しゅっぱん）」「先手（せんて）」
- その他（3 つ以上の形態素を含むもの、複合語と考えられるが構成要素が立項されていないもの、略語、例外的な表記または音形をもつもの、語構成が不明のものなど）

なお、固有名詞、外来語、記号等に関してはその内部構造について情報を付与することは行わない予定である。表では固有名詞、外来語を名詞・形状詞には含めず、全て「その他の品詞」としている（「その他の品詞」の大部分は固有名詞と外来語である）。

表 1　現段階での暫定的な種類別の語数

語構成	名詞 (N)	動詞 (V)	形容詞 (A)	形状詞 (K)	その他の品詞
NN	7,088			28	
VN	3,279			4	
AN	1,050			43	
KN	34			5	
NV	4,489	232		29	
VV	2,225	7,842		16	
AV	340	24		5	
KV	12	1			
NA	198		153	52	
VA	26		23	2	
AA	16		28	5	
KA					
NK	15			12	
VK	3			1	
AK	3				
KK					
その他/未処理	60,108	3,179	621	1,340	106,697
合計	78,886	11,278	825	1,412	106,697

4. 検索ツール

現在、本研究で付与している情報および『UniDic』に元々付与されている情報を検索するためのウェブ UI を開発しており、現在、試験的に公開している [3]。図 2 は開発中の画面であり、動詞+動詞の複合名詞を検索した例を示している。

検索ツールを開発するのは以下のような理由による。本研究で整備するデータはそのまま

[2] 『UniDic』や『BCCWJ』で定義されている「語彙素」に近いものであるが、厳密には対応しない。
[3] http://asaokitan.net/jmorph/

テキストデータとしても公開する予定だが、直接そのデータを利用し、例えば形態素へのリンクをたどって前部要素の属性で絞り込むといった処理を行うにはある程度の知識が要求される。そのため、このような検索が簡単に行えるツールを提供することで、より広い分野の研究者にデータを利用してもらうことが可能になる。

この目的にウェブ UI を用いることには、データダウンロードなどの手間がなく、ユーザ側の環境を選ばないことや、また、データをウェブ上で公開することにライセンス上の問題がないことから、合理的であると考える。なお、ウェブページのソースコードも公開を予定している。

読み	表記	品詞	語種	前読み	前	後読み	後
サキオリ	裂き織り	名詞 普通名詞 一般 *	和	サク	裂く	オル	織る
サキワケ	咲き分け	名詞 普通名詞 一般 *	和	サク	咲く	ワケル	分ける
サグリウチ	探り撃ち	名詞 普通名詞 サ変可能 *	和	サグル	探る	ウツ	打つ
サグリツリ	探り釣り	名詞 普通名詞 一般 *	和	サグル	探る	ツル	吊る
サグリビキ	探り弾き	名詞 普通名詞 サ変可能 *	和	サグル	探る	ヒク	弾く
サゲシブリ	下げ渋り	名詞 普通名詞 一般 *	和	サゲル	下げる	シブル	渋る
サゲドマリ	下げ止まり	名詞 普通名詞 一般 *	和	サゲル	下げる	トマル	止まる
サゲフリ	下げ振り	名詞 普通名詞 一般 *	和	サゲル	下げる	フル	振る
サゲモドシ	下げ戻し	名詞 普通名詞 一般 *	和	サゲル	下げる	モドス	戻す
ササエアイ	支え合い	名詞 普通名詞 一般 *	和	ササエル	支える	アウ	合う
サシイデ	差し出で	名詞 普通名詞 一般 *	和	サス	差す-他動詞	イデル	出でる
サシオサエ	差し押さえ	名詞 普通名詞 サ変可能 *	和	サス	差す-他動詞	オサエル	押さえる

図 2　検索ツールの開発中の画面

原稿執筆時点の検索ツールでは、単語全体の表記・読み・品詞・語種、また構成素の表記・読み・品詞・語種、あるいはその組み合わせを指定して検索することができ、表記・読みについてはワイルドカードも使用可能である。ただし、アクセントや「連濁の有無」など付加情報を用いた検索、また同一の語彙素における表記や読みのバリエーションについては現段階では対応していない。音韻研究での有用性を考えると、正規表現検索、（仮名ではなく）ローマ字による検索、（字数ではなく）モーラ数を指定しての検索などが可能であればより有用なツールになると思われる。これらの点は今後の課題とする。

5.　まとめと課題

本研究では、『UniDic』への語構成情報の付加、およびその検索ツールの開発について紹介した。

本研究の主要な課題として、語構成の曖昧性が挙げられる。本研究は、語源についての情報を意図したものではなく、基本的には語構成に関する共時的な意識を反映させたものを意図している。しかしながら、語構成の意識には話者間の感覚の違いも大きく、合成語と見な

せるかどうかのグレーゾーンに位置する語も多いと考えられる。そのようなケースについての一貫した基準を現段階では持っていない。『BCCWJ』の短単位の定義の元となる最小単位の決定においては、さまざまなルールが用いられている (小椋ほか 2011)。例えば「えがく」や「まつりごと」など常用漢字表に掲げられた訓はそれ以上分割せず全体を最小単位として扱うことや、「いなずま」のように現代仮名遣いに関する内閣告示で「二語に分解しにくい」ために「ぢ」「づ」ではなく「じ」「ず」を用いると定められている語も、分割せず全体で最小単位として扱うという規定などである。これらのルールは、基準ごとに適用可能な項目が限られるうえ、(狭い意味での) 言語学的な基準とは言いにくいため、同様の基準を採用することが本研究の目的に即しているかどうかについては議論の余地がある。

本研究で構築するデータにはさまざまな拡張の可能性がある。例えば、構成要素間の関係についての情報 (項関係、付加詞関係、等位構造の区別など)、意味的情報、統語的情報 (項構造など)、頻度情報などを追加することが考えられる。現段階では、これらの情報を追加することは予定していないが、加工・再配布可能な形で公開することで、必要に応じてこれらの情報を自由に追加できるようにする。

謝　辞

本研究は JSPS 科研費 15H06258「構文形態論の形式モデルの構築に関する研究」の助成を受けたものである。

文　献

小椋秀樹・小木曽智信・小磯花絵 (2007). 「「現代日本語書き言葉均衡コーパス」の短単位解析について」　言語処理学会第 13 回年次大会発表論文集.

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, pp. 101–122.

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」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 (2015). 『複合動詞レキシコン』, http://vvlexicon.ninjal.ac.jp.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers (1996). *CELEX2*. Philadelphia: Linguistic Data Consortium.