

多様な目的に適した形態素解析システム用電子化辞書の開発

The development of a multi-purpose electronic dictionary for morphological analyzers

伝 康晴	千葉大学文学部
Yasuharu Den	Faculty of Letters, Chiba University
山田 篤	(財) 京都高度技術研究所
Atsushi Yamada	ASTEM
峯松 信明	東京大学大学院新領域創成科学研究科
Nobuaki Minematsu	Graduate School of Frontier Sciences, The University of Tokyo
内元 清貴	(独) 情報通信研究機構
Kiyotaka Uchimoto	National Institute of Information and Communications Technology
小磯 花絵	(独) 国立国語研究所
Hanae Koiso	The National Institute for Japanese Language
小木曾 智信	(独) 国立国語研究所
Toshinobu Ogiso	The National Institute for Japanese Language

1. 電子化辞書班の研究目的

本計画研究の目的は、研究代表者の伝らが従来開発を進めてきた形態素解析システム用電子化辞書 UniDic (伝ほか 2002)を整備・拡充・改良することにより、(1) 本研究領域が目指す大規模書き言葉コーパスの構築を支援するとともに、(2) 日本語学・日本語教育学における語彙・文法調査研究、自然言語処理における構文・意味解析研究、音声情報処理におけるテキスト音声合成研究など、多様な目的に適した統合的な電子化辞書およびその利用システムを提供することにある。

具体的には以下のことを行なう。

- (1) 本研究領域のコーパスで使用する形態論体系と整合的になるよう UniDic を整備し、データ班に提供するとともに、データ班からのフィードバックにより拡充・改良を行なって、短単位で10万語以上の電子化辞書を開発する。
- (2) 発音・音変化やアクセント型・アクセント変化に関する情報や意味分類などを記述し、日本語学・日本語教育学・自然言語処理・音声情報処理など多様な目的に利用できるようにする。
- (3) 語彙形態論研究に適した短単位、音声研究に適した中単位、構文・意味研究に適した長単位という複数粒度の「語」を高精度(98%以上)で自動構成するシステムを提供する。

以上により、多彩な情報を複数の粒度の「語」に対して与える電子化辞書とその利用システムを開発する。

日本語の電子化辞書は数多く存在する(たとえば、EDR 辞書・IPAL 辞書や各出版社の電子化辞書など)が、これらは基本的に検索目的であり、形態素解析システムなどオンライン処理での利用を想定したものではない。一方、形態素解析システム(茶釜や Juman など)付属の辞書は、自動解析に必要な範囲の情報の記述にとどまっており、計量言語学的な研究に必要な語の同一性(異表記や異形態の扱い)や単位の斉一性(語を認定する一定の基準)の問題を解決していない。UniDic は、開発当初から本研究領域の主要メンバーである国立国語研究所グループと議論を重ね続けており、日本語コーパス言語学分野において、比類のない重要性をもった研究リソースとなっている。

2. 従来の研究経過

伝・山田・峯松は、情報処理振興事業協会「擬人化音声対話エージェント基本ソフトウェアの開発」プロジェクト(平成12~14年度・代表:嵯峨山茂樹)において、UniDicの開発に着手した。プロジェクト終了後も共同研究を続け、平成16年度に3万5千語からなる形態素解析システム用辞書を作成・公開した。本辞書は、テキスト音声合成システムにおいて重視される、数詞・助数詞の音変化(「一(イ

チ)」「本(ホン)」が結合して「イッポン」になるなど)やアクセント変化(「言語(ゲンゴ)」「学(ガク)」が結合して「ゲンゴ・ガク」になるなど)といった音韻論的現象の扱いを重視しており、同様の主旨の電子化辞書が皆無の中、音声言語処理分野に重要な貢献をもたらした。

また、伝・山田・内元・小磯は、本研究領域代表者である前川喜久雄・国立国語研究所グループ長が中心となって進めた文科省科学技術振興調整費開放的融合研究「話し言葉の言語的・パラ言語的構造の解析に基づく『話し言葉工学』の構築」プロジェクト(平成11~15年度)に参加、伝・小磯は、その後の科学研究費補助金(基盤研究B)「話し言葉コーパスに基づく言語変異現象の定量的分析」(平成16~18年度・代表:前川喜久雄)にも参加し、本研究領域内の他班のメンバーとの協力関係を十分整えている。さらに、伝・小磯・小木曾は、国立国語研究所 KOTONOHA プロジェクトの一環として、平成18年度初頭から UniDic の開発環境整備・語彙拡充作業にすでに着手している。

本計画研究は、これらの流れを統合し、より完成度の高い電子化辞書の開発を目指して、従来からの共同研究をさらに推し進めるものである。

3. 研究計画

UniDic の特色は以下の3つにまとめられる(詳細は付録参照)。

- 見出し・表記・品詞・活用型などの基本的な辞書情報に加えて、代表形・発音・アクセント型・語種・意味分類および音変化やアクセント変化に関わる情報など、多彩な情報を記述し、多様な目的に供する。
- 語彙形態論研究、音声研究、構文・意味研究といった異なるレベルの研究要請に適した3つの異なる粒度の「語」(短・中・長単位)を認定し、多様な応用研究に供する。
- 異表記(「表わす」と「表す」など)や異形態(「小さい」と「ちっさい」「ちっちゃい」など)に関する情報を反映した、階層化されたデータ構造を持ち、実データの多様性に対処するとともに、計量言語学的な研究に必要な正規化の問題に対処する。

これらの特色を実現するための具体的な研究課題は、以下のようにまとめられる。

- (1) 辞書データベースの開発(担当:伝・小木曾)
- (2) 辞書データベースへの情報登録(担当:伝・小木曾・小磯)
- (3) 形態素解析システムでの運用(担当:伝・小木曾)
- (4) 音変化・アクセント変化の調査とシステム開発(担当:山田・峯松・小磯)
- (5) 中・長単位構成の調査とシステム開発(担当:内元)

以下、今年度と来年度以降に分けて、詳細を述べる。

3.1 今年度

研究課題ごとに計画を述べる。

- (1) 伝・小木曾は、短単位に関する階層化されたデータ構造を関係データベーススキーマとして定式化し、SQL サーバと Microsoft Access のフォーム機能を用いた辞書データベースシステムを実装する(すでに進行中)。サーバは国立国語研究所のネットワーク内で運用する。また、国立国語研究所外の研究分担者とのデータ共有の方式についても検討する。
- (2) 伝・小木曾・小磯は、国立国語研究所内にアルバイト数名を雇用し、研究課題(1)の辞書データベースシステムを用いて、データ班から提供された未登録語(見出し・表記・品詞・活用型)の登録を行なう。さらに、既登録語も含めて、研究課題(4)で作成される発音・アクセント型・音変化制約・アクセント変化制約、および、データ班から提供される語種・語構成情報の登録も行なう。
- (3) 伝・小木曾は、国立国語研究所内に研究補佐員1名を雇用し、形態素解析用統計モデルの作成

と形態素解析システムでの運用を行なう。現在、学習コーパスとして RWCP コーパス(RWCP 1998)と『日本語話し言葉コーパス』(前川 2004)を中心に用いており、統計モデルとして拡張HMM(浅原・松本 2003)を用いている。実行エンジンとしては茶筌(松本 2000)を用いている。データ班提供のコーパスを対象に性能評価実験を行ない、モデルの改良を行なう。さらに、語形選択精度の向上のため、CRF(条件付確率場)モデル(Lafferty et al. 2001)の導入を検討する。

- (4) 山田・峯松・小磯は、国立国語研究所内に研究補佐員1名を雇用し、『日本語話し言葉コーパス』中の長単位の発音・アクセント情報を調査し、音変化・アクセント変化の記述に有効な素性を抽出する。これにより、UniDicの既存の記述体系(伝ほか 2002; 黒岩ほか 2005)を見直し、また、数詞・助数詞以外の音変化(連濁など)を扱えるように拡張する。峯松は被験者実験によるアクセント調査も行なう。得られた結果は随時、研究課題(2)で辞書登録する。
- (5) 内元は、短単位から長単位を自動構成するシステムのプロトタイプを作成する。システムには統計的チャンキングモデルを採用し、長単位情報が付与された『日本語話し言葉コーパス』を学習データとして統計モデルを作成する。

年度末には、語彙数7万語以上、短単位解析精度(語形選択含む)97%以上の形態素解析システム(茶筌)用辞書を公開する。これら作業の進行調整のために9月・11月・1月・3月に会合を行なう。

3.2 来年度以降

研究課題ごとに計画を述べる。

- (1) 伝・小木曾は、ツール班が提供しデータ班が利用するコーパス管理システムと、本班が利用する短単位辞書データベースシステムとの統合的運用の仕組みを作る(平成19年度中)。平成20年度以降は、中・長単位に関する記述枠組みの設計とシステム開発を手がける。
- (2) 伝・小木曾・小磯は、アルバイト数名を引き続き雇用し、データ班からのフィードバックを受けて、語彙の拡充を行なう。平成20年度からは、意味情報として、国立国語研究所が従来開発を進めてきた『分類語彙表』の分類コードを適宜拡充しつつ辞書登録する。
- (3) 伝・小木曾は、研究補佐員1名を引き続き雇用し、形態素解析用統計モデルの改良を行なう。とくに、実行エンジンとしてCRFモデルを採用したもの(茶筌の次期版?)を利用し、語形選択精度の向上を目指す。
- (4) 山田・峯松・小磯は、研究補佐員1名を引き続き雇用し、音変化・アクセント変化記述の改良を行なう。また、研究課題(5)の中単位構成システム(音変化・アクセント変化が及ぶ語列の範囲を特定する処理)を利用し、音変化・アクセント変化処理の高精度化を目指す。
- (5) 内元は、長単位構成システムを完成させ、データ班に提供し、長単位情報付与作業に利用してもらう(平成19~20年度中)。さらに、研究課題(4)で利用する中単位構成システムの開発を行なう。

各年度末に形態素解析システム用辞書の改訂版を公開する。最終的に、以下の成果物を公開する。

- 語彙数10万語以上、短単位解析精度(語形選択含む)98%以上の形態素解析システム用辞書
- 精度98%以上の長単位構成システム
- 発音・アクセント精度95%以上の音変化・アクセント変化処理システム

これら作業の進行調整のために、毎年度、隔月の頻度で会合を行なう。

4. その他

データ班とは、研究課題(2)(3)を実質共同で行なうことになる。ツール班には、研究課題(3)に関して、本辞書の潜在能力を最大限引き出すための実行エンジンの改良(CRFモデルなどによる任

意素性の扱いなど)を期待したい。研究項目 B01 の各班についても、辞書に記載してほしい情報の要望をいただくなど、積極的な交流を行ないたい。

文献

- 浅原正幸・松本裕治 (2002). 「形態素解析のための拡張統計モデル」 情報処理学会論文誌, 43, 685-695.
- 伝康晴・宇津呂武仁・山田篤・浅原正幸・松本裕治 (2002). 「話し言葉研究に適した電子化辞書の設計」 第2回「話し言葉の科学と工学」ワークショップ講演予稿集, pp. 39-46.
- 黒岩龍・峯松信明・広瀬啓吉 (2005). 「日本語テキスト音声合成用アクセント結合規則の改良」 日本音響学会 2005 年度秋季講演論文集, pp. 427-428.
- Lafferty, J. D., McCallum, A., & Pereira, F. C. N. (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data”, *Proceedings of the 18th International Conference on Machine Learning*, pp. 282-289.
- 前川喜久雄 (2004). 「『日本語話し言葉コーパス』の概要」 日本語科学, 15, 111-133.
- 松本裕治 (2000). 「形態素解析システム『茶釜』」 情報処理, 41, 1208-1214.
- 新情報処理開発機構 (RWCP) テキスト・サブ・ワーキンググループ (1998). 研究開発用知的資源: タグ付きテキストコーパス報告書.

付録

A. UniDic の概要

A.1 UniDic の基本理念

- 言語学・国語学など人文系の言語研究に違和感なく使える。
 - 齊一な単位設定
 - 学校文法に準じた品詞体系の採用
 - 語形を中心にした語の認定
- 音声言語研究に使える。
 - 実情に即した発音情報の記述
 - アクセント型情報の記述
 - 音変化・アクセント変化に関わる制約の記述

A.2 複数粒度の単位認定

UniDic では、研究目的に応じて複数粒度の単位を認定する (図 1 参照)。

- 短単位は、形態素 (最小単位) の結合体であり、語彙形態論研究に適する。
- 中単位は、語の意味構造と関係した、音変化・アクセント変化が及ぶ範囲の複合語であり、音声研究に適する。
- 長単位は、文節を自立語と付属語に分けたものであり、構文・意味研究に適する。

UniDic はこのうち短単位を辞書記述によって、中・長単位は自動構成システムによって認定する。

	外来	語	仮名	表記	を	調査	し	た
短単位	名詞	名詞	名詞	名詞	助詞	名詞	動詞	助動詞
中単位	名詞		名詞		助詞	名詞	動詞	助動詞
長単位	名詞				助詞	動詞		助動詞

図 1 複数粒度の単位認定の例

A.3 「語」の階層的定義

UniDic では、「語」をさまざまな属性によって定義する。これらの属性の継承関係によって、以下の4つの階層での「語」の定義が得られる（図2参照）。

- 語彙素：意味・文法機能を同じくする語の集合に対応し、語彙素読み・語彙素・類の3属性によって定義される。
- 語形：国語辞典の見出しに相当し、同じ語彙素に属する語群を語形の違いによって細分化したものである。たとえば、「アナタ」「アンタ」は同じ語彙素に属する異なる語形である。
- 書字形：語形を表記の違いによって細分化したものである。たとえば、「表わす」「表す」は同じ語形に属する異なる書字形である。
- 発音形：語形を発音の違いによって細分化したものである。たとえば、「ウインドーズ」「ウィンドーズ」は同じ語形に属する異なる発音形である。

なお、書字形と発音形の間には階層関係はなく、階層のレベルは、

語彙素 > 語形 > 書字形, 発音形

の3段階である。

語彙素	語形	書字形	発音形
アナタ【貴方】《体》	アナタ	貴方	アナタ
		あなた	アナタ
	アンタ	あんた	アンタ
アラワス【表す】《用》	アラワス	表わす	アラワス
		表す	アラワス
		あらわす	アラワス
アラワス【著わす】《用》	アラワス	著わす	アラワス
		著す	アラワス
		あらわす	アラワス
カナ【仮名】《体》	カナ	仮名	カナ
		かな	カナ
カメイ【仮名】《体》	カメイ	仮名	カメイ
ウインドウズ《体》	ウインドウズ	ウインドウズ	ウインドーズ
			ウィンドーズ
		ウィンドウズ	ウインドーズ
			ウィンドーズ

図2 「語」の階層的定義の例

A.4 多彩な辞書情報

UniDic では、「語」を定義する属性に加え、さまざまな辞書情報を記載する。これらは、各語の語彙・形態・音韻的性質を記述したものであり、各階層に分けて記述される（表1）。

表 1 多彩な辞書情報

階層	定義属性	辞書情報
語彙素	語彙素読み・語彙素・類	意味分類
語形	語形	品詞・活用型・語種・語構成 語頭変化型・語頭変化形・語頭変化結合型 語末変化型・語末変化形・語末変化結合型
書字形	書字形	活用型書字形分類 仮名書字形
発音形	発音形	活用型発音形分類 アクセント型・アクセント結合型

A.5 品詞体系

A.5.1 品詞

品詞は4階層で定義され、大分類は15種類、細分類まで展開すると53種類である。大分類の15種類は以下のとおり。

- ①名詞 ②代名詞 ③形状詞 ④連体詞 ⑤副詞 ⑥接続詞 ⑦感動詞 ⑧動詞 ⑨形容詞
⑩助動詞 ⑪助詞 ⑫接頭辞 ⑬接尾辞 ⑭記号 ⑮補助記号

A.5.2 活用型

活用型は6階層で定義され、大分類は17種類、書字形・発音形分類まで展開すると259種類である。大分類の17種類は以下のとおり。

- ①五段 ②上一段 ③下一段 ④カ行変格 ⑤サ行変格 ⑥文語四段 ⑦文語下一段
⑧文語上二段 ⑨文語下二段 ⑩文語カ行変格 ⑪文語サ行変格 ⑫文語ナ行変格
⑬文語ラ行変格 ⑭形容詞 ⑮文語形容詞 ⑯助動詞 ⑰文語助動詞

A.5.3 活用形

活用形は2階層で定義され、大分類は9種類、小分類まで展開すると48種類である。語彙定義辞書には基本形のみ記述され、活用形は別途、活用辞書に記述される。大分類の9種類は以下のとおり。

- ①語幹 ②未然形 ③意志推量形 ④連用形 ⑤基本形 ⑥連体形 ⑦仮定形 ⑧已然形
⑨命令形

B. UniDic と形態素解析

上述のように、UniDic は語彙素・語形・書字形・発音形という階層で定義されるが、形態素解析システム用辞書としては、書字形をキーとし、統計情報を付加した形式で提供される。UniDic による「語」の認定とは、表1中の7つの定義属性（と活用形）を同定することであり、これは従来の形態素解析システムが出現形の分ち書きと品詞（と活用形）の同定のみですませているのに比べると、問題をより困難にしている。人文系言語研究においては、「語」を同定することが形態素解析の主要な目的であり、高精度な語形選択（同形異音語の識別）が必要となる。既存の形態素解析システムの統計モデルは、この課題に対して極めて貧弱であり、その改良が UniDic を用いた形態素解析の重要課題の一つとなる。