# 異表記同語情報を付与した辞書の整備

浅原 正幸　　高橋 由梨加　　松本 裕治

奈良先端科学技術大学院大学 情報科学研究科

{masayu-a,yurika,matsu}@is.naist.jp

**Abstract**

This paper reports on the progress of a dictionary maintenance project at NAIST. The project aims to augment the dictionary with orthographic variations of Japanese words. We describe the criteria for identifying orthographic variations. We also present the maintenance schema for the dictionary and the graphical user interface for the annotation task.

## 1　はじめに

自然言語処理のアプリケーションにおいて入力文中の語の同定が行われる。多くの場合形態素解析器などを用いて、入力文中の部分文字列から辞書中に含まれる語彙項目の同定を行なう。統計的手法や機械学習の導入により形態素解析器の精度が向上した。また、未知語処理モデル、未知語抽出モデルなどにより、大規模の生テキストから未知語の抽出が可能になった。

本発表では、奈良先端大で行なっている異表記同語情報を付与した辞書の整備の進捗について報告する。形態素解析は入力中に出現する語の同定を目的として用いられているアプリケーションではあるが、日本語では 1 つの語が複数の異表記を持つ場合があり、それらを別の語として同定してしまう。今回の辞書整備はこの問題を解決を目標としている。本稿では、まず、異表記だが同語と見なす判定基準を提示する。この中には既に関係付与が完了した基準と今後関係付与予定の基準の両方が含まれる。次に、どのように関係付与作業を進めたかについて述べる。最後に、実際に関係付与作業に用いた辞書管理用のデータベーススキーマと関係付与ツールの機能について述べる。判定基準および作業形態を提示することにより、辞書ユーザからの意見を広く求めたい。

## 2　異表記同語の整理作業

整理作業を行なう対象の辞書は ipadic-2.7.0 [1] である。まず、関係付与しようとしている異表記同語の判定基準を提示する。次に、作業の進め方を述べる。最後に、2004 年度の作業の進捗を示す。

### 2.1　判定基準

以下に我々が関係付与しようとしている異表記同語の判定基準を示す：

(A) ひらがな、カタカナ、漢字、英字など字種間の表記のゆれ
e.g.)「かえる」⇔「カエル」⇔「蛙」

(B) 漢字のゆれ
e.g.)「芸」⇔「藝」

(C) 送り仮名のゆれ
e.g.)「行なう」⇔「行う」

(D) 可能性に基づく品詞体系の整理
e.g.)「暇（名詞-一般）」、「暇（名詞-形容動詞語幹）」

(E) 品詞「名詞-非自立-∗」「名詞-接尾-∗」と対応する自立語とのリンク
e.g.)「こと（名詞-一般）」、「こと（名詞-非自立-一般）」

(F) カタカナ語の表記のゆれ
e.g.)「コンピューター」⇔「コンピュータ」

(G) 略語
e.g.)「特措法」⇔「特別 | 措置 | 法」

(H) 縮約形態
e.g.)「ちゃう」⇔「て | しまう」

(J) 固有名詞、専門用語
e.g.)「日ハム」⇔「日本ハム」

(A)〜(C) は表記のみが異なる同語である。「読み」の一致を前提としているために「読み」を手掛りとして候補集合の枚挙が可能である。(B) の漢字のゆれは、旧字と新字、異体字、同訓異字などが含まれる。同訓異字の認定は、複数の国語辞典を参照し、その中で 1 つでも認定されていれば、異表記同語とみなす。

(D) は伝ら [4] が解説している可能性に基づく品詞体系に基づく整理である。可能性に基づく品詞体系の考え方では用法による品詞の区別は行なわず、可能な用法を全て考慮した 1 つの品詞を付与を目標とする。例えば、「今度 | 行き | ます」の | 今度 | は副詞的な用法であるが、「今度 | が | 最後 | だ」の | 今度 | は名詞的な用法である。この | 今度 | という語に対し、いずれの用法についても「名詞-副詞可能」という品詞を付与する。ipadic では、この考え方が徹底しておらず、依然登録語の品詞のゆれが残っている。今回は以下の 5 つのゆれを対象に整理を行なった。

- {「名詞-一般」「名詞-サ変接続」} →「名詞-サ変接続」
- {「名詞-一般」「名詞-形容動詞語幹」} →「名詞-形容動詞語幹」
- {「名詞-一般」「名詞-副詞可能」} →「名詞-副詞可能」
- {「名詞-サ変接続」「名詞-形容動詞語幹」} →「名詞-形容動詞語幹」
- {「副詞-一般」「名詞-副詞可能」} →「名詞-副詞可能」

なお、これらの規則に適合する場合でも、異なる品詞間で語義が異なる場合には異表記同語とみなさない。例えば、「緑 | いっぱい | の | 自然」に出現する | 自然 (名詞-一般)| と「自然 | な | 動作」に出現する | 自然 (名詞-形容動詞語幹)| は語義が異なるために区別する。

(E) は、現在まで自立語との区別の基準が不明確であった品詞「名詞-非自立-*」「名詞-接尾-*」の整理である。明らかに不要な語彙項目 (e.g. 数えあげられない助数詞) を削除し、削除するか否か判断が困難な語彙項目については残しておいたまま、対応する自立語にリンクをはる。また連濁による有声化などの読みのゆれもここで関係付与する。

(F) はカタカナ語の長音などのゆれを対象とする。候補語を枚挙するために、増山ら [5] が提案した手法が利用できると考えている。

(G)(H) は、略語や話し言葉に頻出する縮約形態を対象とする。単位自体が変わるために、後に述べる複合語-構成語間の関係付与の枠組みで対処する。(K) は、固有表現の表記のゆれ。分野依存の専門的な知識が必要なため、現在は整理の対象としていない。また、(G)(H)(K) は候補集合を提示する手法が確立できていない。

## 2.2 作業の進め方

関係付与作業は以下に示す手順で行われる：

1. 何らかの方法で候補集合を枚挙する

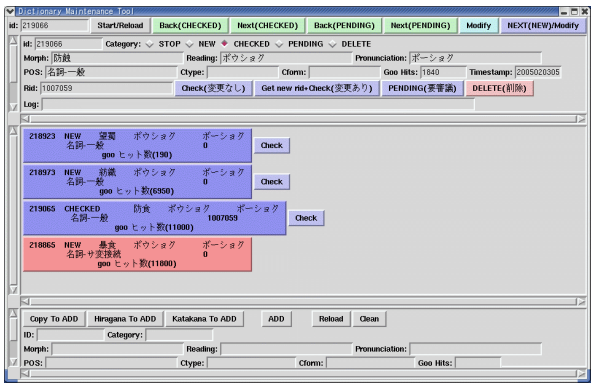


図 1: 作業者用関係付与ツール

2. 関係付与に必要な情報を提示する
3. 情報を基に作業者が異表記同語と認定した語の集合に同じ異表記同語 ID(rid) を付与する
4. 1. で枚挙されなかった候補集合を作業者が入力する

図 1 に実際に利用している関係付与ツールを示す。まず関係付与ツールは、上部に現在関係付与対象となっている単語の情報を示す。中央のフィールドには、対象となっている単語と異表記同語であろう候補集合を示す。同時に各単語の頻度や検索エンジンにおけるヒット数、各種国語辞典中での見出し語の立て方などを提示する。作業者は、提示された情報を基にして、関係付与対象となっている単語と中央のフィールドに示される異表記同語の両方に同じ異表記同語 ID (rid と呼ぶ) を付与する。枚挙されていない異表記同語については、下部のフィールドから入力を行ない候補集合に追加し rid を付与する。

関係付与作業をする際に、ある表層形について複数の異表記同語がある場合に、その表層形は全ての可能な rid を持つ。例えば、表層形「はし」は {「はし」,「橋」}、{「はし」,「端」}、{「はし」,「箸」} の 3 種類の異表記同語関係を持つ。一方、表層形「橋」は異表記同語関係 {「はし」,「橋」} 1 つのみを持つ（同様に「端」、「箸」も異表記同語関係を 1 つのみ持つ。）。同じ異表記同語関係を持つ単語に対して、いずれかの語が正規形であるというような情報の付与は、正規形の定義が困難であるために行なわない。また、異表記同語を付与する際に 2 つの単語の関係が (A)〜(F) のいずれかであるという情報の

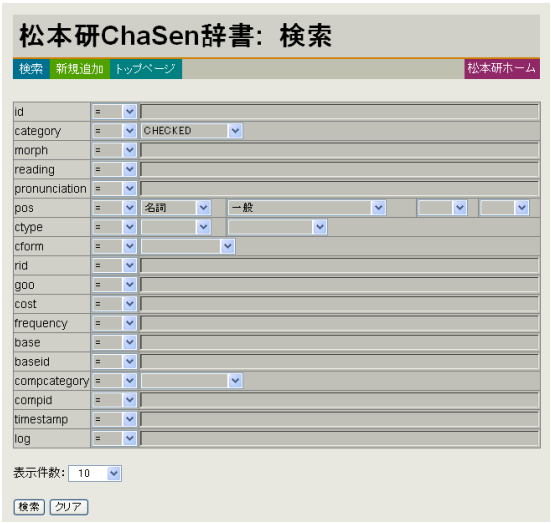


図 2: 作業監督者用閲覧ツール（検索要求画面）


図 3: 作業監督者用閲覧ツール（検索結果画面）

付与も行なわないが、(A)〜(F) と (G) と (H) と (J) 間は異なるカテゴリ名を導入し区別した。

### 2.3 作業の進捗

2004 年度は、ipadic 中の固有名詞以外の語に対して、前節の (A)〜(E) についての整備作業を作業者 4 人と作業監督者 2 人の合計 6 人で行なった。作業者が判断できなかった語彙項目のみについて、作業監督者が作業を行なう。2005 年 2 月 3 日現在で 109408 語に対し作業が完了し、作業の必要な残り語数は 15998 語である[1]。

[1]作業開始時は約 90000 語を対象としていた。可能な異表記同語を枚挙しているために作業完了語数は作業開始時の対象語数より多くなっている

## 3 辞書整備環境

本節では、構築した辞書整備環境について述べる。辞書データは SQL データベースサーバに格納し管理を行なう。管理されているデータは、表層形、読み、発音、品詞、活用型、活用形、現在付与している異表記同語を表現する関係 ID(rid)、異表記同語認定作業の進捗を表わすカテゴリ、検索エンジンのヒット数、形態素解析器で利用される単語生起コスト、原形、原形の単語に対するポインタ (baseid)、複合語関係付与作業の進捗を表わすカテゴリ、複合語情報、最終更新日時、その他作業ログなどのフィールドからなる。

異表記同語情報の整備作業は図 1 に示される GUI を用いて、2.2 節に示される手順で行なう。同じ GUI の利用により、作業者が判断できなかった単語について、作業監督者は関係付与作業を行なう。新語の登録作業以外の殆どがマウスのクリックにより行なうことができる。

関係付与作業用の GUI とは別に、web ベースの辞書閲覧ツール[2] を作成した。作業監督者は閲覧ツールを用いて作業を確認することができる。図 2 に検索要求画面を示す。上に示したデータベースに格納されている全ての項目に対して絞り込み検索を行ったり、各項目で昇順/降順ソートを行なうことができる。図 3 に検索結果画面の例を示す。各単語の左端の [詳細] ボタンを押すことにより、各単語の情報を表示することができる。表層形、読みの文字列をクリックすることにより、クリックした文字列を web 上の各種辞典で検索することができる。rid および baseid はクリックすることにより、既に関係付与された同じ rid もしくは baseid を持つ単語を表示することができる。

## 4 辞書整備の今後

2.2 節中の (A)〜(E) の作業が終了した時点で、一度辞書をリリースする。個々の事例について広く意見を得るために、web ベースの辞書閲覧ツールを公開し、外部の研究者からの意見を取り入れて基準を修正していく。

引き続き残りの辞書整備を続ける。まず (F) カタカナ語の表記のゆれの関係付与作業を行なう。次に単

[2]"cradle" (`http://cl.naist.jp/cradle/`)

語の単位の問題を解決するために、辞書中の複合語に対してその複合語を構成する要素へのリンクを付与する [2]。複合語情報の付与と同時に (G) 略語、(H) 縮約形態の関係付与作業についても行なう。前節で述べた web ベースの辞書閲覧ツールは、複合語情報を付与するツールでもある。長い単位の単語の先頭から、部分文字列検索を行ない、可能な構成語候補を枚挙することができる。作業者は、正しい構成語候補を単語の先頭から順に選んでいくことにより、複合語に含まれる構成語情報を付与することができる。

## 5 関連研究

佐藤 [3] は、異表記同語の認定基準を整理するとともに形態素解析器 JUMAN 用辞書を基準に基づき整備した。彼の予稿が発表される前から奈良先端大において、同様な異表記同語関係付与作業を開始していたが、彼の基準をもとにいくつかの基準を再構成した。基準に関していくつかの差異があるが、その差異は関係付与スキーマの違いに由来すると考える。彼のスキーマでは各語の「語義」（彼の論文では「抽象的な実体」）の 1 つ 1 つを認定し、その語義に対する異表記を枚挙する。これに対し我々のスキーマでは「表層形と読みの 2 つ組」（彼の論文では「表示」）に対し、全ての異表記を枚挙する。言い換えると、複数の「語義」について「表層形と読みの 2 つ組」と「その 2 つ組に対する可能な異表記」が全く同じ、若しくは片方の可能な異表記集合がもう片方の異表記集合に完全に含まれる場合には、その「語義」を区別しない。例えば、衣服などの合せ目を留める「ボタン」と機械を動作させるために押す「ボタン」は同じ漢字表記「釦」を持つ。この場合、我々の定義ではこの 2 つの語義を区別せず表層形「ボタン」は漢字表記「釦」を 1 つ持つとする。さらに「ボタン」は 3 つ目の語義として花の「ボタン」があり漢字表記「牡丹」を持つ。ゆえに 3 つの語義を持つ表層形「ボタン」に対して 2 つの可能な漢字表記「釦」と「牡丹」を定義する。本辞書は、ある「表層形」もしくは「表層形と読みの 2 つ組」に対し、可能な全ての異表記を枚挙することのみを想定しているため、厳密な語義の数を判定しない分だけ関係付与作業効率があがる。しかしながら、実際のアプリケーションによっては、本辞書により展開される異表記を、語義までを考慮して出現可能な異表記に制限するために、語義の曖昧性解消などの後処理が必要となる。

増山ら [5] は、異表記同語情報の整備における部分問題であるカタカナ語の表記のゆれを半自動的に獲得する手法を提案している。現在はカタカナ語の整備を開始していないが、彼らの提案手法を利用する予定である。

浅原ら [2] は、単位の問題を緩和するために複合語辞書の構成スキーマについて提示している。複合語、略語、縮約形態についての関係付与は彼らのスキーマにならう予定である。

## 6 おわりに

本稿では、奈良先端大で行なっている ipadic に対する異表記同語情報を付与作業の進捗について報告した。2004 年度は字種間の表記ゆれなど、意見一致しやすい基準に関し関係付与を行なってきた。2005 年度は引き続きカタカナ語の表記ゆれと略語や縮約形態を含めた複合語の関係付与を行なっていく予定である。

## 参考文献


[1] 浅原正幸, 松本裕治. IPADIC ユーザーズマニュアル. 奈良先端科学技術大学院大学, 2002.

[2] 浅原正幸, 米田隆一, 山下亜希子, 伝康晴, 松本裕治. 語長変換を考慮したコーパス管理システム. 情報処理学会論文誌, Vol. 43, No. 07, pp. 2091–2097, 2002.

[3] 佐藤理史. 異表記同語認定のための辞書編纂. 情報処理学会研究会報告 (自然言語処理研究会), 2004-NL-161, pp. 97–104, 2004.

[4] 伝康晴, 宇津呂武仁, 山田篤, 浅原正幸, 松本裕治. 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, 2002.

[5] 増山毅司, 関根聡. 大規模コーパスからのカタカナ語の表記の揺れリストの自動構築. 言語処理学会第１０回年次大会発表論文集, pp. 29–32, 2004.