

異表記同語認定のための辞書編纂

佐 藤 理 史

京都大学大学院情報学研究科知能情報学専攻

日本語の語の異表記（表記のゆれ）をどのように扱うかについてのガイドラインを作成し、そのガイドラインに基づいて JUMAN4.0 付属の辞書を整理した。この辞書を用いることにより、異表記同語を機械的に認定することが可能になった。この辞書を用いて、新聞記事における表記のゆれを調査したところ、約 10% の表記がゆれていることが明らかになった。

Identifying Spelling Variations of Japanese Words

SATOSHI SATO

Department of Intelligence Science and Technology
Graduate School of Informatics, Kyoto University

Almost every word in Japanese has two or more spellings. This demands identification of spelling variations. This paper presents a guideline for handling spelling variations of Japanese words, and a word dictionary that is compiled by using the guideline. This paper also describes existing spelling variations on newspapers, which were detected by using the dictionary.

1. はじめに

形態素解析システムは、現在、日本語処理の各種応用において、広く用いられている。誰もが利用できる JUMAN や ChaSen の出現は、日本語処理のハードルを低くするのに、大きく寄与してきた。

しかしながら、筆者は、現在の形態素解析システムは次の 2 つの問題を抱えていると考える。

- JUMAN や ChaSen は「文を形態素の列に変換する」システムであると称しているにもかかわらず、その結果を語の列とすり替えることが事実上広く行なわれている。「形態素 = 語」と見なすことは近似的には問題ないが、あくまでも近似である。
- いわゆる異表記の問題がほとんど考慮されていない。現在のシステムでは、表記が異なれば別形態素と見なされる。たとえば、「等しい」と「ひとしい」は、まったくの別語とみなされる。

これらの問題の根底には、「語とは何か」という定義の欠如、あるいは、「何を語として認定するか」という原則の欠如がある。これらの欠如の結果、形態素と語の区別、表記と語の区別があいまいとなり、上記の問題が生じている。

我々が最終的にほしいものは、「文を形態素の列に変換する」システムではなく、「文を語の列に変換する」システムである。これを実現するためには、システムの出力の単位となる「語」を明確に定義することが必要である。なお、「文を語の列に変換する」処理は、形態素解析と呼ぶのは不適切であり、語認定、あるいは、セグメンテーションと呼ぶのが適切であると考ええる。

日本語の「語」を明確に定義するためには、語の単位（長さ）と同語異語の判定を、明確に規定する必要がある。日本語は、表記法において語の区切り記号を持たず、語境界があいまいである。また、日本語の語のほとんどが、複数の表記（漢字表記、ひらがな表記）を持つ。このため、表記だけをよりどころに語を認定するという方法は、実質的に機能しない。これらは、日本語固有の問題である。さらに、複合語や慣用句の場合には、文を構成する単位としての「語」が、意味的处理の際の単位（意味単位としての「語」）と一致しない（ずれがある）という問題がある。これらの問題がからみあって、日本語の語の認定を非常に難しいものとしている。

語認定システム（より正確には、語認定用辞書）が、ある特定の応用システムに依存した専用システムであるならば、語の認定は、その後の処理の都合に合わせてよ

い。実際、機械翻訳システムなどでは、そのような都合主義による語認定がなされてきた。しかしながら、多くの応用システムで使用を想定した汎用システムを目指すのであれば、語の認定に対する原則を明らかにし、その原則にのっとった語認定を実現する必要がある[☆]。このためには、語認定用辞書の編纂のガイドラインが必要である。

このような考えに基づき、本研究では、次の3つのことを行なった。

- いわゆる異表記の問題、すなわち、先に上げた2つの問題のうち、後者の問題をどのように扱うべきかについてのガイドラインを定めた。
- そのガイドラインにしたがって、JUMAN4.0[a]の辞書を整理した。この辞書を用いることにより、表記が異なる同語を認定することが可能となった。
- 整理した辞書を用いて、新聞における表記のゆれを調査した。

2. 異表記同語問題と語認定

2.1 同語異語の判定問題

日本語の語の認定問題は、大きく、単位の問題と、同語異語判定の問題に分けることができる。

- 単位の問題：どの長さを1語と考えるかという問題である。これは、テキストの表記法において、語の区切り記号が存在せず、語境界があいまいであることに起因する。国立国語研究所が行なった各種語彙調査では、複数の単位が採用されている^{1),2)}。
- 同語異語判定の問題：どの範囲を1語と考えるか、つまり、語の広がりに関わる問題である。これは、主に、語が複数の表記を持つことに起因する。

この2つの問題を同時に扱うのは無謀であるので、本研究では、後者のみを扱う。

後者の問題、すなわち、同語異語の判定が難しいのは、表記と語との間に複雑な対応関係があるからである。表記と語がほぼ1対1に対応する言語(例えば、英語)では、表記が等しいものを語とする(語の同一性の判定を、表記の同一性で判定する)ことにしても、特に問題は生じない。しかし、日本語では、(1)ほとんどの語が複数の表記を持つ、(2)ひらがな表記やカタカナ表記においては、明らかな別語が同一表記となる、という事実があり、語の同一性を表記の同一性で判定することは、あまりに粗い近似となってしまう。

同語異語の判定問題には、比較的コンセンサスが得やすい問題と、コンセンサスが得にくい問題がある。比較的コンセンサスが得やすい問題は、いわゆる「表記のゆれ」と呼ばれるものである。たとえば、「等しい」と「ひとしい」は、漢字を用いた表記、ひらがなのみの表記と

いう違いはあるが、どちらも同じ語を表していることにほとんどの人が同意するであろう。一方、コンセンサスが得にくい問題は、いわゆる「同訓異字」と呼ばれるもので、たとえば、「熱い」、「暑い」、「厚い」、「篤い」などがその典型的な例である。これらの語の同語異語判定(見出し語のたて方)は、市販の国語辞典でもまちまちであり、容易に結論がでるような問題ではない。

このような認識より、本研究では、「表記のゆれ」、すなわち、コンセンサスが得られやすい範囲の異表記同語を扱い、そのような異表記を同語と認識することができる語認定システムを実現することを目標とする。なお、上記にあげたような同訓異字の和語は、原則として別語として扱う。

2.2 語、表記、読み、表示

以降の議論のために、ここでは、いくつかの用語を定義する。

- (1) 語を表記と切り離し、抽象的な実体と考える。それぞれの語にはIDを割り当てる。語の同一性は、このIDの同一性により判定する。このIDを語IDと呼ぶ。
- (2) 語の表記、読み、表示を以下のように定義する。
 - 表記(H): 語をテキストとして書くときに使用する文字列。
 - 読み(Y): 発音の目安を記述したひらがな(あるいは、カタカナ)文字列。
 - 表示(R): 表記と読みの組。 $R = \langle H, Y \rangle$
- (3) 語は、一般に複数の表示を持ち、そのいずれかによって示される。

語を表記や読みと切り離し、抽象的な実体と考えることが、ここでの出発点となる。抽象的な実体を一意に指し示す記号として、語IDを導入する。このようなIDを導入すれば、語の同一性を判定することはたやすい。

次に、我々が表記と呼んでいるものをすこしばかり整理する。語を指し示すことができるメディアは、音声とテキストがある。音声の場合は「音」であるが、日本語の場合は、これを適度に抽象化した「読み」が存在するので、これを音の代用として用いる。いわゆる「表記」は、テキストにおいて語を表示したものである。

日本語には、「人気(にんき)」と「人気(ひとけ)」のように、漢字表記はまったく同じで読みが異なる別語が存在する。これらを明確に区別する手段を提供するために、表記と読みの組を導入する。この1組を語の表示形式と考え、これを表示と呼ぶことにする。以上が、上記の定義のここである。

なお、表記と読みのデータ形式は、いずれも文字列であるので、それらの同一性は、文字列の同一性として判定できる。表示は、表記と読みの組であるので、これも文字列の同一性を基盤として、同一性を判定できる。

2.3 語認定とのための辞書

これまで異語同語問題を最も真剣に考えてきたのは、語

[☆] そのような原則がなければ、それぞれの応用システムで、文の構成単位としての「語」を、意味単位としての「語」にまとめ直す処理を設計できないことになる。

集調査の研究においてであろう。語彙調査は、あるコーパスに語彙がどのように分布しているか調べるものであるが、語の数を数えるためには、その単位、つまり語を明確に定める必要がある。また、語の頻度を数えるためには、2つのコンテキストに出現する表記が同語なのか別語なのかを判定できなければならない。

一方、今、我々が考えるのは、語認定のための辞書をどのように作るかという問題である。実際にコーパス中の語を認定するのはプログラムが行なう作業であり、語認定のための辞書は、その基準を辞書という形でプログラムに伝えるものとなる。

ここで、語認定という処理を、もういちど考え直してみよう。今までの議論を整理すると、次のようになる。

- (1) 概念レベルの定義：この処理は、「文を語の列に変換する」処理である。
- (2) 入力「文」は、「文字列」として与えられる。
- (3) 出力「語の列」は、「語 ID の列」として表現する[☆]。

さて、文（文字列）は、表記を連結したものであるから、概念的に、語認定は、次のような3段の処理として整理できる。

- (1) 区切り記号の挿入：文字列としての文に、表記の区切りを挿入して、文を表記の列に変換する。
- (2) 読みの決定：それぞれの表記に対する読みを決定する。この処理により、表記の列は、表示の列となる。
- (3) 語の同定：それぞれの表示が指し示す語を同定して、語の列に変換する。

上記の処理をどのように実現するかは一旦棚上げして、語に関してどのようなデータが最低限必要であるかを考えよう。まず、区切り記号の挿入のためには、語を表示するために用いられる表記のリストが必要である。次に、読みの決定のためには、それぞれの表記に対して、許される読みのリストが必要である。最後に、語の同定のためには、それぞれの表示が、どの語を表しうるかの対応表が必要である。すなわち、この3段の処理を実行するためには、語に関するデータとして、次のような形式のデータ（表示と語 ID との対応関係）があればよい。

$$\langle R, ID \rangle = \langle \langle H, Y \rangle, ID \rangle$$

言い換えるならば、語認定用の辞書では、少なくとも、上記のような形式で、表示と語 ID との対応関係を定義する必要があるということである。

形態素解析システムとして広く用いられている JUMAN や ChaSen の辞書は、表示のリストと見なすことができる。しかしながら、それらの表示がどのような語を表しうるかを明確に定義していないため、我々の用途には不十分である。我々が求める辞書を作成するためには、こ

れらの辞書に含まれているそれぞれの表示に対して、その表示が表しうる語を明確に定義する必要がある。これは、それぞれの表示に語 ID を割り当てるということである。

2.4 表示と語の対応関係

表示と語の対応関係を、次の3種類に分けて整理しよう。

- (1) 表示と語が1対1で対応する。このことは、次のことを意味する。
 - もし2つの表示が同一であれば、それらは同語を表している。
 - もし2つの表示が異なれば、それらは別語を表している。これが大原則となる。
- (2) 2つ以上の表示が1つの語に対応する。これは、ある語が複数の表示を持つということに対応する。次の3つのケースに細分できる。
 - 別表記同語：読みは同一で、表記だけが異なる。
 - 別読み同語：表記は同一で、読みだけが異なる。
 - 別表示同語：表記も読みも異なる。
- (3) 1つの表示が、2つ以上の語に対応する。これは、ある表示が（文脈に応じて）複数の語を表しうることに対応する。これを同一表示別語と呼ぶ。

後の2つが例外であり、これをどのような場合に適用するかを定めることにより、表示に対する語 ID の割り当て方法が定まる。

3. 異語同語判定のガイドライン

3.1 語に関する基本的な考え方

語は、形と意味を持った実体だと考える。形とは、表示を抽象化した概念である。

- 大原則1：意味が明らかに異なるものは同語と考える。

語は、通常、複数の語義（意味）を持つ。しかしながら、それらの間には、語としてのまとまりを示すような関連が観察されるのが普通である。ここでの「意味」は、複数の語義の総体と考える。

たとえば、「あおい（青い）」と「あおい（葵）」を考えよう。これらは、明らかに意味が異なるので同語とは考えない。

- 大原則2：形が明らかに異なるものは同語とは考えない。

意味が同じものをすべて同語だと考えるならば、いわゆる同義語は、すべて同じ語とすることになってしまう。そこで、やはり、形というものを考慮せざるを得ない。

たとえば、「外海（そとうみ）」と「外海（がいかい）」は、意味はほとんど同じであるが、和語と漢語という明らかな形の違いがあるので、同語とは考えない。

逆にいうならば、2つの表示を同語と見なす場合は、同じ意味を表し、かつ、形に何らかの強い関連が見られる場合ということになる。一方、1つの表記が2つの語を

[☆] 実際には、それ以外の情報（たとえば、活用形）を出力することが必要であるが、ここでは、議論を簡単にするため、「語 ID の列」を出力するものとする。

表すと考える場合は、あきらかに意味の異なるものを指し示す場合となる。

なお、以下に示すガイドラインは、2つの表示 R_1, R_2 が与えられた時、これらに同語とみなす（同じ語 ID を割り当てる）か、別語とみなす（異なる語 ID を割り当てる）かを定める指針を与えるものである。具体例の記述では、＝ は同語と見なすことを、≠ は同語と見なさないことを表す。

3.2 別表記同語

別表記同語は、表記のみが異なる同語である。数は一番多い。読みが同一の表示が、その候補となる。

次の場合を同語と見なす。

- 文字種の違い（漢字、ひらがな、カタカナ表記の違い）
例：等しい＝ひとしい、曖昧＝あいまい、癌＝ガン
- 送り仮名のゆれ
例：合い言葉＝合言葉、表す＝表わす
- 漢字のゆれ
異なる漢字が用いられることがあるが、その使い分けが意味的な違いに関係しないと考えられるものは、同語とみなす。具体的には、次の2種類。
 - － 旧字と新字
例：愛甞＝愛玩
 - － 使い分けられない漢字
例：合榧＝相榧、目差し＝眼差し

次のような場合は、同語とはみなさず、別語と見なす。

- 明らかに意味が異なるもの
例：青い（あおい）≠ 葵（あおい）
- 異なる漢字で表記される字音語
例：間（かん）≠ 館（かん）≠ 管（かん）
- 同訓異字をもつ和語
例：青い（あおい）≠ 蒼い（あおい）

3.3 別読み同語

別読み同語は、読みだけが異なる同語である。表記は同じで、読みが異なるものが候補となる。数は少ないが、判定は難しい。

次のような場合に、同語とみなす。

- 発音のゆれと思われるもの。
例：行き違い（いきちがい）＝行き違い（ゆきちがい）、旅客（りょきやく）＝旅客（りょかく）
- どちらの読みも日常的に使われ、かつ、意味に差がない場合。
例：世論（せろん）＝世論（よろん）

次のような場合は、別語とみなす。

- 和語と漢語
例：外海（そとうみ）≠ 外海（がいかい）
- 使用にあきらかな差がみられるもの
例：私（わたし）≠ 私（わたくし）

3.4 別表示同語

表記も読みも異なる同語である。同語の基準を明確に

表 1 JUMAN4.0 付属の辞書の構成

ファイル名	エントリ数	品詞等
ContentW.dic	35770	内容語
Noun.hukusi.dic	40	副詞的な名詞
Noun.keishiki.dic	9	形式名詞
Noun.suusi.dic	46	数詞
Demonstrative.dic	76	指示詞
Assert.dic	1	判定詞
AuxV.dic	20	助動詞
Postp.dic	94	助詞
Prefix.dic	55	接頭辞
Suffix.dic	387	接尾辞
Special.dic	123	特殊
Rengo.dic	238	連語
Noun.koyuu.dic	29811	固有名詞

設定することはかなり難しい。

以下のものを同語とみなす。

- 外国語を表記する際のゆれと思われるもの。
例：コンピュータ＝コンピューター
ただし、日本語において使い分けが定着しているものは、別語と見なす。
例：カップ≠コップ

3.5 同一表示別語

表示が同一であるにもかかわらず、2つ以上の語に対応させるもの。つまり、ある表示 R に対して、 $\langle R, ID_1 \rangle, \dots, \langle R, ID_n \rangle$ の N つの対応を定義するもの。以下の3つの場合がある。

- ひらがな表記
- カタカナ表記
- 2つ以上の意味がある外来語

4. JUMAN 辞書の整理

前節で述べたガイドラインに従って、JUMAN 辞書の整理を行なった。

4.1 JUMAN4.0 付属の辞書

JUMAN のマニュアル³⁾では、形態素を次のように説明している。

形態素 m の形態品詞、品詞細分類、活用型、活用形、表層形がそれぞれ $H1, H2, K1, K2, M$ であるとき、これを5項組 $(H1 H2 K1 K2 M)$ によって表わし、「形態素 m の形態素構造」と呼ぶ。

マニュアルには明示的に書かれていないが、形態素構造がまったく同一ものを同一形態素とみなすものと思われる。このことから、JUMAN の形態素は、2.2 節で導入した表示に対応する。

JUMAN4.0 付属の形態素辞書の構成を表 1 に示す。いわゆる内容語は、ContentW.dic というファイルにまとめられており、その数は 35,770 エントリである。

このエントリ数は、そのまま、形態素数を意味するわけではない。ContentW.dic の一部を図 1 に示す。この図に示すように、JUMAN 辞書のエントリは、1つの形

(名詞 (普通名詞 ((読み あいこく)(見出し語 愛国 (あいこく 1.6) (あいこく 1.6))))))
 (名詞 (普通名詞 ((読み あいこくしん)(見出し語 愛国心 (あいこくしん 1.6))))))
 (名詞 (普通名詞 ((読み あいことば)(見出し語 合い言葉 (あいことば 1.6))))))
 (名詞 (普通名詞 ((読み あいことば)(見出し語 合言葉 (合言葉 1.6) (合ことば 1.6) (合ことば 1.6) (あいことば 1.6))))))
 (名詞 (サ変名詞 ((読み あいご)(見出し語 愛護 (あいご 2.0) (あいご 1.6) (あいご 1.6))))))

図 1 JUMAN4.0 付属の辞書の記述形式

態素を定義するものではなく、表記のみが異なる複数の形態素を定義するものとなっており、異表記を扱うことを指向した記述形式となっている。

- しかしながら、この扱いは中途半端である。すなわち、
- (1) 形態素解析結果において、出力されたそれぞれの形態素がどのエントリで定義された形態素であるかという情報は得られない。つまり、(a) 同じエントリで定義された形態素であっても、形態素構造が異なれば、同じエントリで定義されているものであるかどうかかわからない。(b) 異なるエントリで定義された形態素であっても、形態素構造が同一であれば、それらの区別がつかない。
 - (2) 同語と考えるのが妥当であるものが、別エントリとして定義されているものがある。例えば、図 1 の 3 番目のエントリ (合言葉) と 4 番目のエントリ (合い言葉) は、同語と考えるのが妥当である。

前者の問題は、JUMAN の辞書の記述項目として使用できる「意味記述」に、語 ID を記述することによって解決できる。後者の問題を解決するためには、同語とみなすべきエントリをまとめあげる作業が必要である。

4.2 整理作業とその結果

異語同語判定の作業、すなわち、JUMAN 辞書のエントリのうち、同語とみなすべきものをまとめる作業は、オリジナルの辞書に 137 エントリを追加した後の 35,907 エントリの状態から開始した。

まず、すべてのエントリにそれぞれ別の語 ID を付与した。次に、同語の可能性のあるエントリの組をプログラムでリストアップし、それを作業者がチェックし、同語とすべき組に同一の語 ID を付与しなおすということを行った。作業者は 2 名であり、まず、1 名 (言語学専攻の博士課程の学生) が原案を作成し、もう 1 名 (筆者) がそれをチェックし、最終的な判断を下した。

以下に示す順で作業を行なった。

(1) 送り仮名のゆれ

送り仮名のゆれと思われるもののリストをプログラムで作成した。具体的には、品詞、活用型、読みが同じで、第 1 表記 (各エントリで最初に定義されている表記) が異なるものを取り出し、ダイナミックプログラミング法で差分を取り出し、その差が、漢字とひらがなの差であるものをリストアップした。こうして作成したリストを作業者がチェック

ガン (癌, 雁)、シミ (衣魚, 染)、スキ (隙, 鋤)
 コマ (駒, 独楽)、マス (鱒, 枺)、カラ (空, 殻)
 クマ (熊, 隈)、ガマ (蝦蟇, 蒲)

図 2 複数の語の表記として採用したカタカナ表記

表 2 作業の概要

作業名	チェックした量		まとめあげ	
	組数	エントリ数	組数	エントリ数
送り仮名のゆれ	924	1903	924	1903
カタカナ表記	374	911	316	652
別読み同語	247	506	36	72
ナ形容詞とサ変名詞	65	140	40	80
別表示同語	-	-	12	24
別表記同語	3073	6416	472	996

- し、最終的に異語同語を判定した。この作業により、1903 エントリを 924 エントリにまとめた。
- (2) カタカナ表記
 まず、第 1 表記がカタカナ表記であるエントリを取り出し、それと同じ読みのエントリがある場合に、それらを組としてリストアップした。これを作業者がチェックし、同語と認められる場合は、同語としてまとめた。なお、カタカナ表記が複数の語の表記として認められる場合は、複数の語に重複して登録した。このような語は全部で 8 エントリあった (表 2)。この作業により、652 エントリを 316 エントリにまとめた。
- (3) 別読み同語
 表記、品詞、活用型が同じで、読みが異なるものをリストアップした。このリストを作業者がチェックし、同語と認められるものをまとめた。72 エントリを 36 エントリにまとめた。
- (4) ナ形容詞とサ変名詞のまとめあげ
 ナ形容詞の語幹とサ変名詞で、表示が一致するものをリストアップし、同語と思われるものをまとめた。80 エントリを 40 エントリにまとめた。
- (5) 別表示同語
 外来語の表記のゆれをまとめた。24 エントリを 12 エントリにまとめた。
- (6) 別表記同語
 再度、品詞、活用型、読みが同じで、第 1 表記が異なるものをリストアップし、これらをチェックした。996 エントリを 472 エントリにまとめた。

これらの作業の概要を表 2 に示す。

作業終了後のエントリを語として認定した。作業全体では、35,907 エントリが 34,000 語にまとまったことになる。なお、別の作業も並行して行なったため、この数には、新たに追加した 13 語が含まれている。作業前と作業後のエントリ数、語数を表 3 に示す。この表から、全エントリ (35,907) の約 1 割のエントリ (3,560) がまとめ上げの対象となったということがわかる。なお、この作業を通して、もっとも多くのエントリが 1 語にまとまったのは、「まなざし」であった (図 3)。

表 3 語の整理作業前と作業後

	語	エントリ
作業開始時点	—	35,907
作業終了時点	34,000	35,915
1 エントリ ⇒ 1 語	32,355	32,355
2 エントリ ⇒ 1 語	1,443	2,886
3 エントリ ⇒ 1 語	117	351
4 エントリ ⇒ 1 語	53	212
5 エントリ ⇒ 1 語	5	25
6 エントリ ⇒ 1 語	13	78
8 エントリ ⇒ 1 語	1	8
小計 (2~8)	1,632	3,560
(新たに追加したもの)	13	—

((品詞 名詞 普通名詞) (読み まなざし) (表記 眼差 (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 眼差し (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 眼指 (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 眼指し (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 目差 (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 目差し (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 目指 (まなざし 1.6)))
 ((品詞 名詞 普通名詞) (読み まなざし) (表記 目指し (まなざし 1.6)))

図 3 8 エントリを 1 語にまとめたもの

4.3 保留事項

今回の作業では、以下に示す事項を保留事項として決定を先送りした。

- 別品詞同語

語は、唯一の品詞を持つとするのか、複数の品詞を持つもの (多品詞語) を許すのか。これは、品詞を、語に固有の性質と考えるのか、あるいは、文内の役割と考えるのか、に大きく依存する。今回の作業では、ナ形容詞の語幹とサ変名詞の両方で登録されていたもの (たとえば、「安心」) を同語としてまとめ上げる作業を行なったが、それ以外の品詞のペアに関しては、作業を行なわなかった (つまり、別語となっている)。たとえば、「ゆっくり (副詞)」と「ゆっくりする (サ変動詞)」など。この問題は、派生をどう考えるかにも深く関連している。同様の問題に、動詞の可能形を活用形 (原形と同語) と考えるのか派生語 (別語) と考えるかという問題がある。

- 副詞の異形

「ゆっくり」、「ゆっくりと」のような副詞の異形に関しても、判断を保留し、まとめあげを行なわなかった。

5. 日本語中核語彙集 JCore

前章の作業は、日本語中核語彙集 JCore の作成の一環として行なった。JCore は、醍醐プロジェクト⁴⁾の語彙部門の中核となる辞書として作成中のもので、その最初の版 (JC1) は、次の特徴を持っている。

- 語 ID の導入：前述
- 表示 (表記) の整理：前述
- 語の難易度情報を付与：
内容語に対して、語の難易度 (重要度) を付与した。難易度は、やさしい方から、A1(2500 語), A2(2500 語), B(5000 語), C(10000 語), F の 5 段階である。

- 表示の難易度情報を付与：

表示 (表記と読み) の組に対して、表示 (漢字) の難易度を与付した。難易度は、やさしい方から、A0(ひらがな等)、A1(300 字), A2(1000 字), B(常用漢字), C(JIS 第一水準), F の 6 段階である。

- JUMAN4.0 システムとの連携：

JC1 は、プログラムによって JUMAN4.0 システム用の辞書にコンバートすることができる。

JC1 を組み込んだ JUMAN4.0 による語認定結果を図 4 に示す。それぞれの表示に対して、可能な語に関する情報が、JUMAN4.0 の「意味記述」のスロットに出力される。たとえば、「危険だ」に対する情報 JC1:cA105680K!A2 は、!の前の JC1:cA105680K が語 ID を表し、!の後の A2 が、表示の難易度を表している。語の ID は、JC1:xXnnnnnnT の形式となっており、X、または、Xn が語の難易度を表している。すなわち、「危険だ」の難易度は A1 である。また、T は、和語 (W)、漢語 (K)、外来語 (G)、混種語 (H) の区別を表している。表記と品詞が同一で、複数の語の可能性がある場合は、語 ID と表示難易度の組が複数出力される。また、品詞が異なるものは、JUMAN の-m オプションを指定した場合に出力される。たとえば、「いった」は、「いう (言う)」、「いく (行く、逝く)」、「いる (要る、入る)」の可能性がある。

6. 新聞における表記のゆれ

JC1 を用いることにより、テキストにどの程度の表記のゆれが存在するかを定量的に明らかにすることができる。ここでは、京都大学テキストコーパス version 3.0 (京大コーパス [b])^{*}を対象に、次の手順で表記のゆれの程度を調べた。

- (1) 京大コーパスから、固有名詞以外の内容語 (形態素) を抽出した。ここで抽出した形態素は、機能語 (品詞の大分類が、判定詞、助詞、助動詞、接頭辞、接尾辞、特殊のいずれか) でなく、かつ、固有名詞 (品詞の小分類が、数詞、人名、組織名、地名のいずれか) でもない形態素である。その後、活用する形態素の活用形を基本形に変換した。この結果、総数で 426,650 個、異なりで 28,821 種類の形態素が得られた。京大コーパスの形態素の数を内容語、機能語、固有名詞別に集計した表を、表 4 に示す。なお、異なりの認定では、品詞小分類は無視した。先に述べたように、ここでの形態素は表示に相当する。
- (2) 得られた表示 (内容語) が、JC1 に存在するかどうかを調べた。京大コーパスの作成で用いられた JUMAN は、おそらく version 3.61 であり、JC1 のベースとなった JUMAN4.0 とは、辞書が全く異

^{*} CD 毎日新聞 1995 年版の約 4 万文に対して、JUMAN および KNP で自動解析を行い、その結果を人手修正したコーパス。

危険 きけん 危険だ 形容詞 * ナ形容詞 語幹 (JC1:cA105680K!A2)
 が が が 助詞 格助詞 * * (JC1:jX000010W!A0)
 切迫 せっぱく 切迫 名詞 サ変名詞 * * (JC1:cC441370K!B)
 して して する 動詞 * サ変動詞 タ系連用テ形 (JC1:cA111400W!A0)
 生じる しょうじる 生じる 動詞 * 母音動詞 基本形 (JC1:cB370110H!A1)

きのう きのう きのう 名詞 時相名詞 * * (JC1:cA105900W!A0)
 @きのう きのう きのう 名詞 サ変名詞 * * (JC1:cA229350K!A0)
 がっこう がっこう がっこう 名詞 普通名詞 * * (JC1:cA104760K!A0)
 に に に 助詞 格助詞 * * (JC1:jX000030W!A0)
 いった いっぱい いう 動詞 * 子音動詞ワ行 タ形 (JC1:cA101220W!A0)
 @いった いっぱい いく 動詞 * 子音動詞カ行促音便形 タ形 (JC1:cA101320W!A0 JC1:cF505000W!A0)
 @いった いっぱい いる 動詞 * 子音動詞ラ行 タ形 (JC1:cA101940W!A0 JC1:cF508810X!A0)

図 4 JUMAN/JC1 による語認定結果の例

表 4 京大コーパスの形態素数

	総数	異なり
内容語	426650	28821
固有名詞	63609	8969
機能語	456807	616
総計	947066	38406

表 6 表示の種類

表示数	語数
2	1030
3	61
4	2

なる。そのため、表記、読み、品詞大分類のいずれもが等しいことを条件に、京大コーパスの形態素(表示)と JC1 の表示の対応付けを行なった。この結果を表 5 に示す。総数で 403,045 個、異なりで 23,065 種類の表示が JC1 に存在した。

- (3) JC1 に含まれる表示のみを対象として、各表示に対して、語が一意に定まるか、あるいは、一意に定まらない(複数の語に対応する)かを調べた。その結果を表 5 に示す。語が一意に定まらない表示は、総数で 60,339 個(15.0%)、異なりで 612 種類(2.7%)であった。
- (4) 語が一意に定まる表示(22,453 種類)を対象として、表記のゆれ、すなわち、複数の表示が一つの語に対応する現象の存在を調べた。その結果を表 5 に示す。表記のゆれを持つ語は 1,093 語存在し、それを表すのに 2,251 種類の表示が使用されていた。
- (5) こうして得られた、表記のゆれが観察された 1093 語について、それぞれの語に対する表示の種類、最多表示の出現割合、表記のゆれの種類について調べた。これらの結果を、表 6、表 7、表 8 に示す。なお、表 8 には例も示した。

京大コーパスは、機械で解析した後、その結果を手で修正した「正解」コーパスである。本調査では、この解析結果を信用することとし、テキストを JUMAN/JC1 で解析するのではなく、解析済コーパスの形態素と JC1

表 7 最多表示の全体に占める割合

割合	語数
30%以上 40%未満	6
40%以上 50%未満	6
50%以上 60%未満	237
60%以上 70%未満	242
70%以上 80%未満	132
80%以上 90%未満	200
90%以上	270

の語を対応付ける方法を採用した*。

本調査の結果は、次のようにまとめることができる。

- (1) 新聞は、統制が効いており、表記のゆれはほとんど存在しないのではないかと考えていたが、その予想は覆えされた。表記が JC1 に含まれており、かつ、語が一意に定まるもの限定して考えるならば、語を分母にすると、内容語の約 5%(1093/21295)において、表記のゆれ(複数表示の使用)が観察された。また、表示を分母とすると、異なりで約 10%(2251/22453)が、総数でも約 10%(36178/342706)が、ゆれに含まれる表示であった。これは、我々の予想をはるかに越える量であった。表記のゆれがもたらす影響は個々の応用に依存するが、この数字から考えると、ほとんどの応用において、何らかの方法で表記のゆれに対処することは必要であろう。
- (2) 表記のゆれの大半は、字種の違いであった。
- (3) 送り仮名、および、漢字の使用は、かなり統制が効いており、ゆれはあまり観察されなかった。
- (4) 複数の表示が使用される場合、ある表示が多数(90%以上)を示す場合も多い(270 件)が、一番使われる表示が 50%-60%と、複数の表示がほぼ均等に使われる場合も少なくない(237 件)。

京大コーパスの品質、辞書(JC1)の品質、調査方法の不十分さ(手でチェックしていない)等の問題により、本調査の結果の信頼性は、人手ですべてを行なった過去の語彙調査のように高くはない。しかしながら、新聞記

* しかしながら、コーパスを眺めてみると、あきらかな解析誤りが観察される。本調査の結果(表記のゆれのデータ)にも、その誤りが含まれている。

表 5 京大コーパスにおける表記のゆれ

	表示		語	
	総数	異なり	(異なり)	
1. JC に含まれる表示	403045	94.5%	23065	80.0%
a. 語が一意に定まる	342706	(85.0%)	22453	(97.3%)
i. 一つの表示のみ	306528	[89.4%]	20202	[90.0%]
ii. 複数の表示	36178	[10.6%]	2251	[10.0%]
b. 語が一意に定まらない	60339	(15.0%)	612	(2.7%)
2. JC に含まれない表示	23605	5.5%	5756	20.0%
計	426650		28821	

注：丸括弧および角括弧内のパーセントは、それぞれの内訳の割合を示す。

表 8 表記のゆれの種類と例

ゆれの種類	語数	例
品詞・活用型が異なる	21	安心 (名)/20=安心だ (形)/3, 迷惑 (名)/20=迷惑だ (形)/3, それ (指)/1045=それ (名)/2
表示が異なる	4	こまぬく/4=こまぬく/1, つぶる/4=つむる/3, ロマンチックだ/1=ロマンティックだ/14
送り仮名のゆれ	79	扱い/53=扱/2, 入り口/26=入口/1, 取り締まり/8=取締/12
異漢字の使用	9	足下/1=足元/20, 火炎/1=火焰/1, 退廃/1=頹廢/1
字種の違い	917	
ひらがな vs 漢字	790	あした/3=明日/21, あっせん/19=幹旋/2, かたち/11=形/201
カタカナ vs 漢字	48	アワ/1=粟/1, イノシン/10=猪/2, カネ/86=金/162
ひらがな vs カタカナ	71	あご/2=アゴ/1, いす/2=イス/4, うさぎ/1=ウサギ/5
漢字/ひらがな vs カタカナ	6	大なた/2=大ナタ/1, 落ち/1=オチ/1, けた違いだ/2=ケタ違いだ/1
漢字 vs 数字	1	二人乗り/1=2人乗り/1
カタカナ vs 数字	1	トップテン/1=トップ10/4

名:名詞、形:形容詞、指:指示詞。“/” の後の数字は、出現数を示す。

事にどの程度の表記のゆれが存在するかについての、ひとつの目安にはなるだろう。

7. おわりに

本稿では、日本語の語の表記のゆれをどのように扱うかについてのガイドラインを示し、それに基づいて JUMAN4.0 付属の辞書を整理したことを報告した。さらに、整理した辞書を用いて、新聞記事 (京大コーパス) に見られる表記のゆれを調査した結果について報告した。これまで、新聞は十分に編集が行き届いており、表記のゆれはほとんど存在しないと考えられてきたが、実際には、かなりの量の表記のゆれが存在することが判明した。

今回整理した辞書では、語の単位の問題は未解決である。また、事実上使われない混ぜ書き表記も、十分に駆逐されていない。これらの問題を解決しなければ、表記のゆれに対する最終的な結論を下すことはできない。しかしながら、本研究の結果は、一つの予想として、「表記のゆれは、我々が思っているよりも数多く存在し、量的に無視できる問題ではない」ことを示唆する。

謝辞 本研究は、科学研究費補助金基盤研究 (B)(2) 「言い換えを中心としたテキスト自動編集技術とその機械翻訳への応用」 (課題番号 13480097)、および、21 世紀 COE プログラム「知識社会基盤構築のための情報学拠点形成」の支援を受けて行なった。

使用した言語ツール・資源

- [a] 日本語形態素解析システム JUMAN4.0
(www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html)
- [b] 京都大学テキストコーパス/CD 毎日新聞 1995 年版
(www.kc.t.u-tokyo.ac.jp/nl-resource/corpus.html)

参考文献

- 1) 林大監修: 図説日本語, 角川書店 (1982).
- 2) 長尾真, 黒橋禎夫, 佐藤理史, 池原悟, 中野洋: 言語情報処理, 岩波書店 (1998).
- 3) 黒橋禎夫, 河原大輔: 日本語形態素解析システム JUMAN version 4.0, 東京大学大学院情報理工学系研究科 (2003).
- 4) 佐藤理史, 土屋雅稔, 村山賢洋, 麻岡正洋, 王晴晴: 日本語文の規格化, 情報処理学会自然言語処理研究会, 2003-NL-153, pp. 133-140 (2003).