RWCテキストデータベースにおける口語・古語等の扱い

荻野 紫穂* 橋本 三奈子† 木村 朗子‡ 阿部圭志§
shiho@trl.ibm.co.jp hasimoto@ipastc.stc.ipa.go.jp kimura@stc.ipa.go.jp abe@rainbow.net.au

1 はじめに

新情報処理開発機構 (RWCP) テキストデータベースグループでは、1994年度から1996年度にかけて、単語分割、読み付け、品詞付けなどを施したテキストデータベースを作成してきた [4,8]。対象データは通商白書 (1994年度対象データ) 及び毎日新聞 (1995年度対象データ) などで、1994年度分については機械で形態素解析を施した結果を人手修正したものを、1995年度分については、機械処理出力と、その一部を人手修正したものとの双方を、公開している。

このうち、新聞記事には特に、文語体や旧仮名遣いの表現や、口語体特有の表現（以下、口語表現と記す）が多く現れた。近年は口語表現の解析がさかんになり、その特有な表現をどう扱っていくかに関する研究もかなり見られる [2,3,6]。これらは、大量のデータを解析する際に現れたそれぞれの現象への対処法を具体的に示している。しかし、口語表現は非常に多様なため、RWCテキストデータベースの処理に当たって、これらに記述のない表現をも処理する必要が生じた。文語体表現そのものについては各辞書 [1] や文法書に活用表や助詞一覧表が載っているが、現代文にも多発する文語的な表現で、[1] の活用表では扱い切れないものを、現代語を扱う品詞体系の中でどう位置付けていけばよいかは明らかでなく、現代語と文語との双方を共存させる品詞体系が必要となった。

更に人手修正作業では、複数の作業者が時間をかけずに判断でき、かつ、コーパス内での単語構成規則と各単語の構成状態との双方が統一されているような修正基準を、一つ一つ確認する手順が要求されている。

本稿では、RWCテキストデータベース構築に当たって、機械処理と人手修正双方で起こった問題点と、形態素タグ（以下、品詞と記す）を付与する際の大まかな方針を述べる。それを踏まえて、3章では、現代語の中の文語的表現をどう扱うか、4章では、対象データに現れた口語表現と、その扱いについて述べ、大規模テキストに現れる多様な表現を、RWCテキストデータベースの品詞体系ではどう扱っているかについて概観する。品詞体系 (THiMCO) そのものや、細かい品詞の説明については、[7,8] を参照されたい。

表現	辞典 [1]	機械	人手	
			可能性	用法
A 勉強 する	b	b	b	b
B あいさつ する	b	x	b	b
C あいさつ を	b	x	a, b	a
D タバコ する	a	a	a, b	a, b

a 名詞
b 名詞 サ変接続
x 未知語

表 1: 機械 vs 人手, 可能性 vs 用法

2 複数の観点から見たタグ付け方針

2.1 機械処理と人手修正

RWCテキストデータベースの品詞付け・読み付け作業は、まず機械で形態素解析をした後、その出力を複数の人手修正者が修正するという形態を採っている。データ公開は機械出力と人手修正との双方について行なっているが、機械処理の結果をそのまま出した時に便利な単語切り方針・品詞体系と、人手修正をしやすい単語切り方針や各修正作業者によって揺れが生じにくい品詞体系とは微妙にずれている。本項と次項とで、このずれに対する現在の対策を述べる。

まず、人手修正作業者間の揺れと品詞付与方針についての例を、表1に例を挙げる。

THiMCOには、[名詞] と [名詞 サ変接続] の双方の品詞があり、概念的には、前者は例えば「机」、後者は例えば「勉強 (する)」などに付与される。実際の品詞付与作業において、辞書への二重登録をできるだけ避けたいなどの理由から、サ変用法のある名詞は、名詞として使われていても (表1C) 動詞的に使われていても (表1A,B)、全て [名詞 サ変接続] を付与しようとする (表1人手/可能性) と、機械処理出力が未知語だった場合な

*日本ＩＢＭ(株)東京基礎研究所
†情報処理振興事業協会技術センター
‡東京大学大学院人文社会系研究科
§St. Finbar's Primary School

どには、人手修正作業において、サ変用法があるとした作業者と、あるとしなかった作業者との間で非常に結果が揺れやすい。またこういった付与方針は、ほぼ全ての名詞についてサ変用法があるかどうか判断することを、修正作業者に強いることになり、チェックに時間がかかる可能性が大きい。このため、このような場合には、実際に使われている用法から判断して品詞を付与する方針を採った(表1人手/用法)。こうすると、人手修正の結果は、作業者によって揺れずに済み、作業者への負担も多少軽減される。ただし、依然として、サ変用法があるかないかの判断に揺れは生ずる(表1D)。更にこの問題は、[名詞]/[名詞 副詞可能]/[名詞 副詞可能 副詞的]や[副詞]/[接続詞]などでも見られる。

このような経験から、機械処理では、主に人手修正のしやすさと、機械出力情報の使いやすさを考えて品詞を付与することを第一目的とした。機械出力では余分な情報も出力するが、それは人手修正後には削除し、揺れの削減を図る。コーパス内の情報付与規則の正確さがそれほど期待されない機械出力においては、もし使用者の参考になるならば、公開時に削除せずそのまま残してもよい、といった考えである。この方針の結果として、機械出力にしか出現しない品詞・人手修正にしか出現しない品詞項目を立てることも検討している。

2.2 単語構成規則の統一性

機械処理では辞書登録された単語が一語という単位を持つ。処理の効率や、アプリケーションの便宜に応じて、非常に長い単位の単語が登録されることもあるが、そのアプリケーションにとって不自由がなければ、辞書を通して一貫した単語構成規則が要求されることは少ない。例えば「白っぽい」という単語が登録されいる一方で、「青っぽい」が一語ではなく「青」+「っぽい」と解析されるとしても、機械翻訳にとって非常に本質的な問題とはならない。

しかし、人手修正作業においては、処理前に事前に定められた単語構成規則が不確定であると、修正作業者に負担を与えることが多い。仮に、「機械出力結果をできるだけ優先する」という方針を採ったとしても、未知語に対して人手修正を施す際には、単語構成規則が不確定であることが、各作業者による修正結果の揺れが生じる原因ともなる。

人手修正したコーパスを作成するという面では、単語構成規則がコーパス内でできれば閉じているほうが、変換処理などをかけやすく、機械処理出力に対して付加価値が高く、より応用分野の広いものになる。本稿のコーパス作成では、単語はできるだけ細かく切る・単語構成規則はできるだけ形から分かるように定める、という方針をとり、特に人手修正コーパスの単語構成規則の一貫性と、作業者による揺れの削減を図った。また、口語処理などから生じる新たな接尾語などを導入する際も、上の方針に従って、できるだけ形から分かる接続規則を判断基準とできるように考えた。

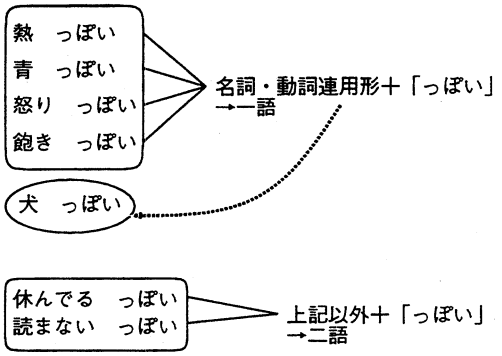

図 1: 新しい接尾語の導入

例えば、図1の「っぽい」という接尾語を導入する際、一語の形容詞として定着しやすい名詞/動詞連用形+「っぽい」という形は一語として扱い、口語体に現れる活用語終止形+「っぽい」は二語に分けて扱う、という規則をたてた。この結果、あまり一語の形容詞として定着していない「犬っぽい」という表現も一語として扱われることになるが、作業者負担の軽減と単語構成規則の一貫性から、現在の段階でこれが一語として扱われることを認めている。

しかしなお、単語構成規則の一貫性も、作業者による揺れの削減も、必ずしも十分に達成はできていない。そこで現在では、形から判断できないものについては、一覧表を作成し、それを各作業者および機械処理の判断基準にすることを検討している。

3 文語

3.1 文語残存形と現代活用形の共存

現代語の活用は、例えば[1]の巻末付録に代表される活用表にあてはまるものがほとんどであるため、THiMCOの活用形の主要部分も、それらの一般的な品詞体系とほぼ同じ形にになっている。しかし、データの中には文語活用が残存しているものもある。図2に例を挙げる。

図2に現れる動詞「うる」は、現代語でもよく使われるが、現代語にはあまり存在しない下二段活用の残存形

(1-a) あり う べからざること
(1-b) あり うる べからざること
(1-c) あり える べからざること
(1-d) あり え ない

図 2: 文語残存形

文語終止形	現代語終止形	文語連体形
寒し	寒い	寒き (体言接続)
悪し	-	悪しき (文語体言接続)

表 2: 形容詞の扱い

である。現代語にはこれとは別に、現代語下一段活用である「える」(1-c, 1-d) も存在するため、「うる」はいわゆる終止・連体・仮定形の活用形だけを持つ。現代語では「うる」が一般に終止形として使用される (1-b) が、「べし」に接続する際には文語終止形 (1-a) も依然として使用されている。現代語において、「べし」に接続する形を特別に持っているのは、ほとんどサ変動詞だけであり、「うる」は特殊な部類に入る。

また形容詞の場合、かなり多くの語について文語連体形が現代語でも使用されるが、逆に文語連体形はよく使われても、その終止形は現代語活用には直接的には残っていないものもある (表 2)。

このように、文語残存形が現代語での形と共存しており、他の現代語単語と比べて特殊な活用形を残している場合には、できるだけ活用型を現代語表記のものと合わせ、現代語と共通した活用形には、現代語の品詞を付与して、現代語表記の活用形では扱えないものだけを文語と扱うとした (表 3)。文語的な活用であっても、多くの現代語でその形が存在する場合はその活用を現代語の活用に加え、逆に文語活用形は存在するが、現代語としての活用形が残存していないものには文語の品詞を付与する方針を採った (表 2)。同様に、ある活用形が現代語においてあまり使用されず (例 「ゆく」 連用

活用形	終止形
うる, うれ	うる (動詞 一段)
え, える, えれ, えよ	える (動詞 一段)
う	う (動詞 文語)

表 3: 「うる」/「える」の扱い

文句をいつてゐる
天は自ら助くるものを助く

図 3: 文語扱いの語例

形)、その活用形を敢えて作ると文語のそれと等しくなる場合 (例 「ゆきて」) には、現代語において欠けている活用形は文語とし、他の活用形には現代語の品詞を付与するとした。例えば動詞では [動詞 文語]、形容詞では [形容詞 文語見出し形/文語体言接続] などが文語扱いの品詞となる。

一般に、文語と現代語との活用形が共通している場合、表記から明らかに文語と思われる場合には文語として扱い、それ以外は現代語として扱うものとする。これは、文語・現代語の分類基準がはっきりするまで、人手作業者による修正結果の揺れを生じさせにくくするための暫定処置でもある。

3.2 旧仮名遣いと文語活用

現在、[動詞 文語] の分類には、上二段活用・ラ行変格活用など、現代語にはほぼ残っていない動詞と、現代語動詞旧仮名使いによる表記が混在している (図 3)。

本来ならば、少なくとも、現代語動詞旧仮名遣い表記と、現代語にはほぼ残っていない活用とは、分けるべきであるかもしれないが、現代語動詞旧仮名遣い表記と、文語動詞そのものが同形である場合、人手修正作業において判断が揺らぐ可能性もあるので、現在のところ、これら全てをまとめて文語として扱っている。

3.3 文語助動詞・文語助詞

助動詞は、文語的活用をするものや、ほぼ文語体にだけ出現するものを [助動詞 文語] として扱った。ここには「ごとし」「べし」のように現代文にも頻出するものも含めている。文語扱いするものとしないものとに関する判断が、各修正作業者によって揺れにくくするとともに、修正者が余分な時間を文語助動詞の判断に費やさず、文語以外の処理にできるだけ力を注ぐことができるようにする、という理由である。

文語助詞に関しては、特に文語特有の品詞は立てていない。助詞は特別な活用などもなく、文語用法と現代語用法とが判別しにくいからである。このため、文語用法と見られる助詞でも、現存する現代語用助詞の品詞の枠組に入れられる場合はそれを流用し、入れられないものは、[助詞 特殊] に分類するか、解析せずに [名詞 引用文字列] に分類するなどの処置をとった。

4 口語

4.1 一体化と省略形

「読みゃ」「赤きゃ」「赤けりゃ」のように、活用語尾と助詞とが一体化したものは、活用語の縮約形と認めた。また、「(読んで)る」「(見て)る」のように語幹が省略されたものも、存在する活用形(空でない活用形)については、語幹が省略されていない元の単語と同じ品詞を付与した[2,3,6]。

「(学生)じゃ(ない)」のように助動詞と助詞が一体化したものは、助動詞と認め、他の活用形が特別に存在しない場合は、[助動詞 不変化型]とした。名詞と助詞・助動詞が一体化したものについては、「ありゃ」「こりゃ」などは代名詞縮約形と認めた。これに対し、「(言わん)こっちゃ(ない)」などは、現在は[引用文字列]として扱う方針である。

4.2 語尾変化

終止/連体形が「る」で終わるラ行五段活用、一段活用、カ行変格活用、サ行変格活用の動詞型活用語は、「の」に連なる形が「帰ん(の)」「見ん(の)」「来ん(の)」「すん(の)」のように「ん」に変わることがある。この形を、それぞれの活用型の[体言接続特殊]という活用形として定義した。また、助動詞ナイに連なる形が「られ」となるラ行五段活用、一段活用の動詞型活用語は、「(見)らん(ない)」のように、「れ」が「ん」に変わることがある。この形を、それぞれの活用型の[未然ナイ接続特殊]として定義した。

4.3 長音等による揺れ

口語表現では、「するぅ」「赤いー」「(読む)ねえ」「(し)ますぅっ」「げ(、元気?)」などのように、長音や長音を表す母音(以下、長音表現と呼ぶ)、促音、言い淀みなどが表記に現れる。

「するぅ」「赤いー」などのように自立語につく長音表現や促音については、長音表現や促音を機能語から分離し、[その他 間投辞]とした。付属語につく長音表現や促音については、「ねえ」「なあ」のように高頻度で出現するものや、特に終助詞として扱いやすいものは、付属語と一体化させて、その付属語と同じ品詞を付与した。それ以外のものは、自立語につくものと同様、[その他 間投辞]とした。また、今回は現れなかったが、自立語のなかに長音表現が含まれる「すーるー」などの表現については、現在のところ、[名詞 引用文字列]とする予定である。

言い淀みについては、[その他 間投辞]として処理した。

基本方針は以上のものだが、これらの表現は非常に多様で、最初に一貫した方針を作りづらかったため、こうした語の処理の一貫性は、現在のところ、あまり達成されていず、統一基準が課題となっている。

4.4 方言

一体化口語に現れる助動詞類で、「(読ま)ねえ」などのように、終止形以外の活用形を、元の助動詞とほぼ共有しているものや、「(学生)っす」「読む(っす)」のように、元の助動詞を推測することが困難で、終止形以外の活用形が出現しない助動詞については、[助動詞 不変化型]を付与した。また、方言などで、標準語に対応する助動詞を推測できるもので、終止形以外の活用形があまり出現しないものについても、同様に[助動詞 不変化型]を付与した。一方で、あまりに標準語の対応表現を推測しにくいものなどについては解析をせず、[名詞 引用表現]とした。

5 おわりに

RWCテキストデータベースの品詞付与作業における基本方針と、文語・口語の扱いを概観した。今後は、特に人手修正コーパスの一貫性を向上させるとともに、より深い情報を付与することを検討している。

謝辞

本稿で述べたことは、井佐原均氏、徳永健伸氏、元吉文男氏(以上RWCデータベースワークショップテキストグループ委員)、桑畑和佳子氏(IPA技術センター)らとの議論を通じて得られた。また、本稿の元となった人手修正作業には、斉藤初江氏、佐藤幸子氏、鈴木悟子氏、谷崎淳哉氏、玉井陽子氏、中川建司氏、山下智弥氏らが携わってくださった。上記の方々及び関係者の皆様に感謝の意を表する。

文献

[1] 西尾, 岩淵, 水谷編 (1994) 岩波国語辞典第五版.
[2] 竹内, 福永 (1991) 話し言葉に対する形態素解析, 情報処理学会第42回全国大会講演論文集 vol. 3, pp. 5 - 6.
[3] 竹元, 福島 (1993) 口語的表現を含む日本語文の形態素解析, 情報処理学会第46回全国大会講演論文集 vol. 3, pp. 109 - 110.
[4] 井佐原他 (1995) RWCにおける品詞情報付きテキストデータベースの作成, 言語処理学会第1回年次大会発表論文集, pp. 181 - 184.
[5] 渕, 米澤 (1995) 日本語形態素解析システムのための形態素文法, 自然言語処理 vol. 2, no. 4, pp. 37 - 65.
[6] 黒橋, 坂口, 長尾 (1996) 京都大学におけるテキストコーパスの作成, 「大規模テキストコーパスの作成及び共有の問題」シンポジウム論文集, pp. 19 - 26.
[7] 技術研究組合 新情報処理開発機構 (1996) 平成7年度RWCテキストデータベース報告書.
[8] 技術研究組合 新情報処理開発機構 (1997発行予定) 平成8年度RWCテキストデータベース報告書.