多重の読みを持つテキストのコーパス化

多重の読みを持つテキストのコーパス化

小木曽 智信（国立国語研究所言語変化研究領域）

Making corpus of Japanese text including multiple readings

Toshinobu Ogiso (NINJAL)

要旨　日本語のテキストには，本文漢字の通常の読みを示すのではない特殊な読みをもつ振り仮名（たとえば「強敵」と書いて「とも」とふりがなを振る類）や，掛詞（「ながめ」を「眺め」「長雨」の両用に読む類から，語形の一部から別の語を連想させる類まで），各種の洒落など，意図的に多重の読みを持たされたテキストが少なくない。従来のコーパスではこのような多重の読みは切り捨てられ，選択されたただ一つの読みを配置することが多かった。本発表では，このような多重の読みを持つテキストについて，主として『日本語歴史コーパス』の事例を整理して示すとともに，そのあるべきコーパスアノテーションの方法について論じる。

１．はじめに

テキストが多重の読みを持つと言うとき、まず想像されるのはその解釈の曖昧性かもしれない。たとえば、古典読解における文学的・文献学的なテキスト解釈の曖昧性の問題、自然言語処理における形態素解析や統語解析、先行詞の同定の曖昧性などがあげられようか。ここで言う読みは、発音形の表示というレベルから解釈のありかたまで多様である。これら多重の読みは、一種の「謎」として書き手によって残されることもありうるが、通常は唯一の解が定まっているものであって、多重の読みがあるとしても意図的に仕掛けられたものではない。これに対して、洒落や掛詞のように、意図的に多重の読みが仕込まれたテキストがある。複数の読みは形式上必ずしも明示的ではないが、複数の読みを持つこと自体に一定の価値を置くテキストである。また、ルビによって本文とはちがう別の読みが明示されている場合もある。

これまでに構築された日本語コーパスにおいては、以上のような各種の多様な読みを持つテキストであっても、原則として一つの読みだけが選択され、他の可能性は捨象されてきた。形態論情報などの言語情報アノテーションは一つの読みについてのみ行われている。しかし本来であれば、多重の読みを持つこと自体に価値があったり、はっきりと多重の読みが示されたりするテキストについては、コーパス化においても、その点への配慮が欠かせないはずである。本稿ではこのような問題意識の下、多重の読みを持つテキストとして、まずは書き手によって意図的に残されたもので、かつ、語や句を単位としたものを取り扱う。具体的には、漢字の固定的な音訓以外の読みを表示するルビと、掛詞・洒落の例である。多重の読みをどうしても扱わなければ済まないこのような場合を例として、コーパスにおいて多重の読みを取り扱う方法について検討したい。

２．自由ルビ

ルビは通常、本文のフリガナとして用いられる。フリガナは、本文の読みを一意に示すことに主眼があるのであって、本来は読みの曖昧性（＝多重の読みの可能性）を抑制するものである。それは漢字の音訓の表示に留まらず、「時雨」のような熟字訓であっても変わらな

い。しかし、この用法を逸脱して、本文の場面・文脈に即した説明的な読みを表示したり、逆にルビの読みの意味説明を親文字が行ったりするタイプのものがある。親文字とルビとの関係が一般的な慣習を離れて、ルビが自由に付与されているという観点から、本稿ではこうしたものを「自由ルビ」と呼ぶことにする。自由ルビと親文字との関係は多種多様のものがあり、混質的である。こうしたものは『現代日本語書き言葉均衡コーパス』にも次のように数多く見られる。

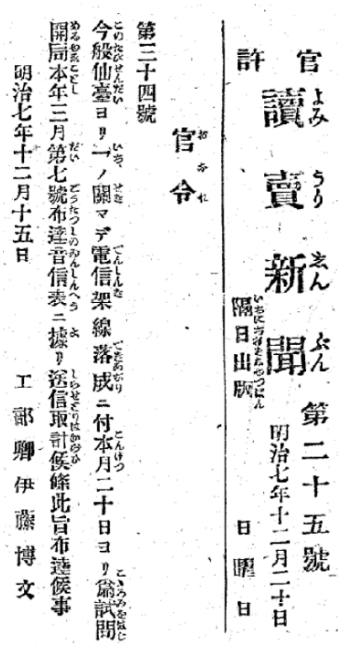
公権力　所有物　連帯責任　超魔　食べたい　ソウルフード　女主人　仮面の男
冥界の神　学際的研究　共体験　建国の父　女に磨きをかけたい症候群
オクスフォード英語辞典　お入り　海蛮人　海蛮人　海蛮人

しかしこれらは臨時的であり、頻度も必ずしも高くはない。

一方、近世・近代の資料にはシステムティックに自由ルビが使われる資料がある。一つのタイプは、図1に示す讀賣新聞のように、「隔日」に「いちにちおき」、「官令」に「おふれ」、「今般」に「このたび」、「落成」に「できあがり」とルビを振るように、難しい漢語に平易な日常語の読みを付けて理解を助けるものである。固い文語文と、日常語のルビによる平易な読みの二重のテキストとなっている。

もう一つのタイプは大部分が話し言葉の台詞からなり、ルビでその台詞の音形を示しつつ意味を本文の漢字列によって示すものである。図2に『太陽』所収の近代の作品の例を挙げる。「誤魔化す」「鳥渡」などは当時通行のフリガナともいえるが、「騙取し」「隔意」などは「騙取」「隔意」の漢語を本文にあて漢字によって意味を表示したものと思われる。このタイプのテキストは近世からあり、洒落本（市村・村山 2017）、人情本（藤本ほか 2017）のコーパス化において問題となったルビがそれぞれ例示されている。特に人情本においてはこの種のルビが多い。

二つのタイプの前者は固い書き言葉の本文に対しルビによって日常語の読みを説明的に示したものであり、後者は、平易な話し言葉のテキストをルビで表し、本文の漢字で意味を説明的に示したものであって、両者は対蹠的な位置にある。


官許 讀賣新聞 第二十五號
明治七年十二月二十日
隔日出版
日曜日

官令

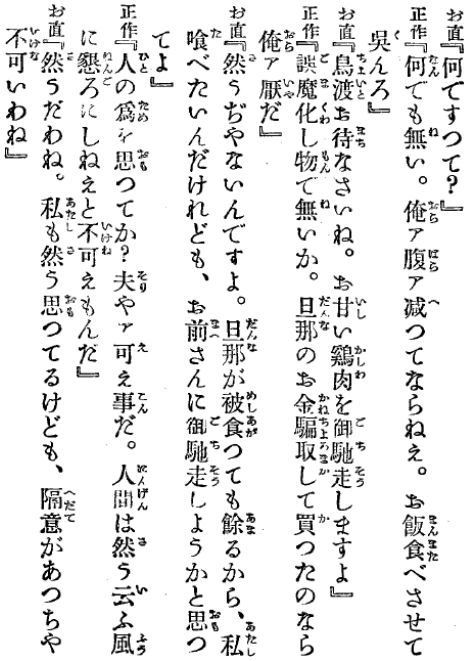
第二十四號

今般仙臺ヨリ一ノ關マデ電信架線落成ニ付本月二十日ヨリ爲試開
開局本年三月第七號布達音信表ニ據リ逓信取扱候條此旨布達候事

明治七年十二月十五日

工部卿伊藤博文

図 1　讀賣新聞 明治 7 年 12 月 20 日より


だが嘘突きだ。不可えなア』

お直『何ですつて？』

正作『何でも無い。俺ア腹ア減つてならねえ。お飯食べさせて呉んろ』

お直『鳥渡お待なさいね。お甘い鷄肉を御馳走しますよ』

正作『誤魔化し物で無いか。旦那のお金騙取して買つたのなら俺ア厭だ』

お直『然うぢやないんですよ。旦那が被食つても餘るから、私喰べたいんだけれども、お前さんに御馳走しようかと思つてよ』

正作『人の爲を思つてか？夫やア可え事だ。人間は然う云ふ風に懇ろにしねえと不可えもんだ』

お直『然うだわね。私も然う思つてるけども、隔意があつちや不可いわね』

図 2　田口掬汀「喜劇　嘘の世界」『太陽』明治 42（1909）年 12 号,p.108 より

コーパス化にあたり、讀賣新聞のような二重の本文はどちらか一方の読みを取っただけでは不足するし、人情本のようなタイプでも本文の漢語は用例として検索可能にするべきである。やはりルビと本文の二重の読みをコーパスで適切に扱う必要性が理解されよう。なお、近世・近代のテキストでは右だけでなく左側にルビが付されることがあり、その場合には三重の読みが重なる可能性もある。

3．掛詞・洒落とシンタグム

掛詞や洒落は、いずれも語形の一致や類似をもととしてもう一つの語としての読みをイメージさせるもので、その点では同じ構造を持っている。ここでは、韻文の掛詞を中心に見ていくことにする。掛詞や洒落の多重の読みは本文に内在するものといえ明示はされないが、音形の類似から多重の読みを可能にするために仮名書きされるなど、別の読みを喚起するための工夫が行われる場合もある。有名な和歌をもとに、掛詞のタイプについて確認しておきたい。次の歌は「ふる」が「古る/振る」、「ながめ」が「眺め/長雨」の二重に読まれる例である。二つの掛詞は関連するが統語関係までは持たず、個々の語が 2 重の意味を持つものとして扱える。

a. 花の色は移りにけりないたづらにわが身世に[ふる] [ながめ]せしまに（古今 113）

次は、「いなば」が「去なば/稲羽」、「まつ」が「待つ/松」 の二重に読まれる例であるが、二つの読みを介在して別の統語的なつながりが成り立っている。「立ち別れ去なば」と「稲羽の山の峰におふる松」は別の文であり、掛詞「いなば」が両者を仲介しているのである。

b. 立ち別れ[いなば]の山の峰におふる[松]とし聞かばいま帰り来む（古今 365）

掛詞は和歌に限られない。次に示すのは、現在コーパス化が進む近松の浄瑠璃（上野 2016）における例である。「なつ」が「夏」であると同時にその一部が「無」の掛詞となっている。さらに、「をりは」は「折羽(双六)」と「降り端」、「こひ目」は「乞い目」と「恋目」の二重になる。ここでも、「な」を介在にして別の文に連なっていく。

c. めぐれば。罪も[な][つ]の雲、あつくろしとて、駕籠をはや。[をりは]の[こひ目]、
（曽根崎心中 p.15）

最後は、洒落本に見られた洒落の例である。「良し」と「吉野」の二重の読みとなっている。

d. 何さ、こゝが[よし]の葛さ（傾城買四十八手 p.109）

以上の例からわかるとおり、多重の読みを持つテキストをシンタグマティックな関係として見たとき、単純な直線的な配列ではあらわせない。a.のように別の語をイメージ喚起するだけであれば、当該部分に二重の形態論情報を付与すれば済む。しかし、b. c.のような例では、イメージされたもう一つの読みを契機に別の文に乗り換える（場合によってはその語また元の文の続きに戻る）ことになるため、二つの意味の主従関係が、前文と後文で入れ替わることになる。d.の洒落は後文が続かないが、b. c.の前半と同じ構造である。

自由ルビの場合、多重化しても基本的には文の範囲は同じであるのに対し、係り結びや洒落では、このような複線的な関係が現れるため、単純に形態論情報を二重化が必要なだけでは十分でなく、文境界や統語関係のアノテーションにおいて問題を生じることになる。

４.多重の読みとコーパス化

従来の国語研究所のコーパスでは多重の読みのうちのただ一つの読みが選択されてきた。たとえば、『現代日本語書き言葉均衡コーパス』では、形態論情報が付されるのは本文文字列に対してであり、ルビはタグとしては付与されるものの形態論情報付与の対象とされていない。その反対に、『日本語歴史コーパス』の試作版として公開されている洒落本のコーパスでは自由ルビについては本文とルビを入替え、元のルビを形態論情報付与の対象とする一方で、元の本文には形態論情報が付与されていない。このような取り扱いは故あってのことではあるが、テキストが持つ重要な情報をすくい上げられていない点でやはり不十分と言わざるを得ない。

現在構築が進む『日本語歴史コーパス』では、本稿で取り上げたような近世・近代資料のコーパス化に取り組んでいるため、多重の読みを適切に扱うことが欠くことのできないこととなってきた。そこで当面の対応として、形態論情報データベース（小木曽・中村 2014）を拡張し、本文文字列に対して多重の形態論情報を付与できるようにした。文字単位で、異なる範囲にまたがる形で多重に情報を付けることが可能である。これにより、自由ルビや掛詞・洒落について、形態論情報のレベルでは対応が可能になった。今後、これを活用して和歌や近世・近代資料のコーパス構築を進める予定である。それでも掛詞の文の二重性については十分に扱えておらず、今後の課題である。

５．おわりに

本稿では、多重の読みを持つテキストの一部としてルビや掛詞・洒落を例示し、コーパス化する場合の課題について見た。その上で、限定的ながら一つの本文に多重の形態論情報を付与することでこの問題に対処した。

将来的には、冒頭に述べたような解釈の曖昧性も含めて、解が一つに定められない場合には複数の読みをそのままコーパスに格納できるようにすることも求められるだろう。コーパスでテキストの多重の読みを扱おうとする試みは緒に就いたばかりである。

謝　辞

本研究は国立国語研究所の共同研究プロジェクト「通時コーパスの構築と日本語史研究の進展開」および科研費基盤(A)「日本語歴史コーパスの多層的拡張による精密化とその活用」による成果の一部である。

文　献

小木曽智信・中村壮範（2014）．『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用，自然言語処理, 21(2), pp.301-332.

上野左絵（2016）．近松浄瑠璃本のコーパス化―「語り」のテキストをどう扱うか,人文科学とコンピュータシンポジウム論文集 2016, pp. 25-30.

小木曽智信（2016）.『日本語歴史コーパス』の現状と展望，國語と國文學, 93(5), pp.72-85.

藤本灯，北﨑勇帆，市村太郎，岡部嘉幸，小木曽智信，高田智和（2017）.「人情本コーパス」の設計と構築，国立国語研究所論集, 12, pp.1-12.

市村太郎，村山実和子（2017）. 洒落本コーパス構築の試行，国立国語研究所論集, 12, pp.29-45.

関連 URL

『日本語歴史コーパス』　http://pj.ninjal.ac.jp/corpus_center/chj/