文書構造を利用した近世期洒落本の形態素解析

市村太郎　小木曽智信
人間文化研究機構　国立国語研究所
{tichimura, togiso}@ninjal.ac.jp

1　はじめに

日本語の歴史を研究するに当たっては，対象となる言葉の用例を収集し分析するのが何よりも重要である．だが，これまでは資料を目視して収集するのが一般的であり，多大な労力を必要とした．一部電子テキストが公開されている資料についてはそれが利用されることもあったが，ある言葉の網羅的な用例の収集に際しては，日本語の表記の特性が検索の障壁となった．例えば「ちょっと」という語には，「ちよつと」は勿論，「チヨツと」「一寸」「一寸と」「鳥渡」「些と」…などと様々な表記が存在するように，日本語の単語には，平仮名・片仮名・漢字およびこれらが混在したものなど，何通りもの表記パターンが存在する．言語感覚の異なる歴史的資料においては，これらを網羅的に検索することは現代語の場合以上に困難である．動詞や形容詞などの用言では，これに語尾の活用変化も加わる．

このような表記や語形変化の問題を克服し，網羅的かつ大規模な日本語史研究を行うことを目的として開発されているのが，『日本語歴史コーパス』[1]である．『日本語歴史コーパス』は，古代から現代に至る日本語の歴史をたどる上での重要資料を集めたコーパスである．XML 形式で文書構造情報が付与されるとともに，古文に対応した形態素解析辞書 UniDic を作成し，形態素解析と解析結果の人手修正によって形態論情報(短単位)が付与されるのが特徴である．これにより，検索の際，多様な表記バリエーションがカバーされる．

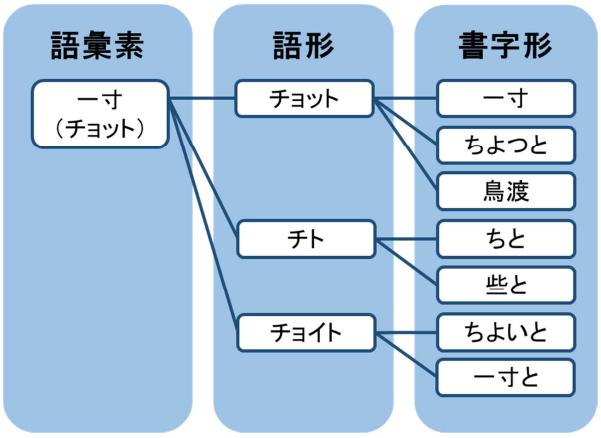
「洒落本コーパス」[2]は，『洒落本大成』[3]を底本とし，『日本語歴史コーパス』中の「江戸時代編」の中核となることが期待される資料体であるが，表記等においてとりわけ多様なテキストであり，形態論情報付与上の大きな課題を孕んでいるため，その克服が求められる．本稿では，この洒落本に対応した UniDic を構築し，形態素解析を行う際，いかなる点が課題となるかを挙げ，それに対処する一手法を提案する．

2　形態論情報の特徴

UniDic では，「表記のゆれや語形の変異にかかわらず同一の見出しを与えるため，見出しとして語彙素・語形・書字形・発音形のレベル(p.100)」(小木曽[4])が設定されている．以下に見出しの内容を挙げるとともに，後の図 1 に階層的見出しの例を挙げる．

語彙素：国語辞典の見出し語に相当するレベル
語形　：異語形を区別するレベル
書字形：異表記を区別するレベル
発音形：発音を区別するレベル

図 1　UniDic の階層的見出しの例(発音形は略)



コーパス全体にこのような階層的な形態論情報を付

与することで, 例えば親見出しとなる語彙素「一寸」(または語彙素読み「チョット」)で検索すれば,「鳥渡」「ちと」「一寸と」など，あらかじめ想定しがたい異語形・異表記の用例を同時に網羅的に取得することができる.

3　洒落本資料の特徴と問題の所在

現在公開されている『日本語歴史コーパス』中の「平安時代編」・「室町時代編 I 狂言」では，コーパスの充実に伴い, UniDic の解析精度の着実な向上が見られた.

例えば，時代的にも文書構造的にも近い「室町時代編 I 狂言」を対象とした狂言専用 UniDic は，2013 年 8 月の時点では，解析精度約 89%(語彙素認定レベル)であったものが, 2015 年 3 月時点では約 96%と，十分に高い精度での解析が行えるレベルに到達した.

一方，洒落本においては，狂言と並行してコーパスと UniDic の整備を進めていたものの，2015 年 3 月時点では, 解析精度約 86%と, 依然低いままであった[4].

表 1　2015 年 3 月時点での狂言用・洒落本用 UniDic の解析精度(F 値・小木曽，鴻野，市村[5]より)

	Level 1 単語境界	Level 2 品詞認定	Level 3 語彙素認定	Level 4 発音形認定
狂言	0. 9887	0. 9700	0. 9613	0. 9595
洒落本	0. 9650	0. 8710	0. 8561	0. 8535

このような状況が生じた背景には,「室町時代編 I 狂言」の構築を優先して進めていたことによるコーパス整備の相対的な遅れももちろん関係するのだが，以下に挙げるような洒落本のテキストの持つ性質が，大きく関わっているのではないかと疑われた.

3.1　表記と語形の多様性

洒落本などの近世口語資料には，コーパスを構築するにあたって乗り越えなければならない表記上の課題が存在することは，市村[6]等で述べたところである.

第一に仮名遣いや送り仮名等の多様性の問題がある. 例えば，口語形容詞の終止形・連体形における語尾のイの表記には，次のようなバリエーションがある.

(1)　かたくろしい(『聖遊郭』)

(2)　きぶひ酢じや(『聖遊郭』)

(3)　おかしゐ。もんだね(『甲駅新話』)

また動詞の「言う」に接続助詞の「て」が付いた形には「いつて」「いゝて」もあれば「言つて」「いひて」「いうて」「言うて」「言ふて」「云ひて」「言て」「云て」「ゆふて」…などが考えられる.

これに加えて, 片仮名平仮名の混在, 踊り字の使用, 濁点無表記箇所の存在などが，表記パターンを増大させることとなる(なおこれらについてはタグ付きで想定される平仮名に置き換えて対処している).

これらの表記上の問題は，おおむね「室町時代編 I 狂言」にも見られるものであるが, 狂言と洒落本では, テキストの多様性が全く異なる.『虎明本狂言集』は大蔵虎明一人の手になるものであり，また伝承が重んじられる舞台台本資料であって，定型表現なども多い. 当然仮名遣いなども，一定の範囲での使用となる. 対して「洒落本コーパス」が対象とする洒落本は，作品ごとに作者が異なり，舞台や出版地が江戸・上方のものが混在していて，言語そのものが大幅に異なっている. そのため，テキストの均質性が低い. このような多様性が解析精度に影響しているとみられる.

ただし，このような多様性の問題は，例えば現代語のコーパスでも多かれ少なかれ生じたものであり，学習用コーパスの分量を増やし，表記パターンを集積していくことで, ある程度対応可能であると推測される.

3.2　文体・文法体系の混在

上述の表記の問題のような, データの蓄積に関わる, 比較的個別的な問題の他に，より体系的な，形態素解析上の課題が存在する. それは,「会話文」と「地の文」における文体および文法体系の異なりである. 次に挙げるのは，1822 年京都版の洒落本『箱まくら』の登場人物の会話(4), 登場人物を紹介する割書き(5), 序文(6)の例である(下線は筆者).

(4)　中ゐお七あの子はやかたもよしおもしろい気だてのあるげいこさんじやが。誰が目もちがはんもので。これまでおよびなさつた子ども衆のうち

では。いつちあの子がよろしい中ゐお八才さんのちやりはのけて。旦那さん。はるさんにあたりがつきましたか旦みながそうほめればいやでもすきでもさうせざなるまい中ゐお十まあそうだんはあとにして。皆さんをおしらせいな里それがよい

(5) 九兵へ国よりつれてのぼりし手代と見ゆれど。これは。国にてかねもちの家へ入こんでおひげのちりをとり。世わたりするものゆゑ京都へも両三度のぼつたこともあるゆゑ。どこともなふ様子よし才兵へ一目見るからしれた宿の亭主なり。もつとも四人一座三四度めのあそび(地の文・割書き)

(6) 傀儡師の頭掛。無量の仏像を現し。藪医者の手提。万国の品類を蔵む。其要いづれも箱にあり。凡多かる箱の中に。箱御祓の神祇あれば。賽銭箱の釈教あり。はこ入娘の恋あれば。棺に蓋する無常の世中。食て圄して寐て起て。つら/＼惟る所。方今や絃妓の齎せる三絃ばこほど奇しきはなし。(序文)

それぞれの特徴的な言語状況を下記に整理した.

会話文—用例(4)

動詞：ほめる（下一段活用)

形容詞：おもしろい・よろしい(語尾イ)

副詞：いつち　感動詞：まあ　助詞：いな

助動詞：じや・ん・まい　融合形：せざ(せずは)

敬語：さん・お～なさる・ます・お＋命令

地の文・割書き—用例(5)

動詞：見ゆ(下二段活用)　形容詞：よし(ク活用)

助詞：ど・にて・ゆゑ　助動詞：き・断定なり

序文—用例(6)

動詞：蔵む(下二段活用)・あり(ラ変活用)・惟る

形容詞：多し(カリ活用)・奇し(シク活用)・

なし(ク活用)　副詞：凡・つら/＼　助動詞：り

これに見るように，会話文では，動詞の一段化が進み，形容詞も口語活用が用いられ，口語助動詞が用いられるなど，比較的現代の話し言葉に近い状況である．対して地の文・割書きや序文では，二段活用動詞やク活用形容詞，文語の助詞・助動詞が見られ，文語体書き言葉の特徴を示している．つまり，一作品のなかで，会話文では口語，他では文語という文体や文法体系の混在が見られ，両者は大きく乖離しているのである．

現在 UniDic の活用体系は，文語(古典文法)と口語(現代語文法)の 2 分法を採用しており，文体や文法体系が混在したままのテキストに対して自動形態素解析を行えば，当然それによる誤りが頻出することとなる．

4　文書構造を利用した形態素解析

洒落本の形態論情報付与においては，会話文は口語，その他では文語を原則として処理する必要がある．そこで，形態素解析前のコーパスに付与された XML タグを利用し，文書構造別の解析を行うこととした．

「洒落本コーパス」では，図 2 のように，会話文は speech 要素によってマークされている．

図 2　『南閨雑話』の XML データ(主要なタグのみ)

```
<s>程なくてうし。硯ぶた。鉢肴。たばこぼん。ども持きたる</s>
<speech><s><speaker>若</speaker></s><s>さあ皆様へお出なさりまし</s></speech>
<speech><s><speaker>里</speaker></s><s>こつちへ。おはいりねんし</s></speech>
<warigaki><s>女郎共すらりとならび忠治を見てにつこり</s></warigaki>
<speech><s><speaker>忠</speaker></s><s>やあ是は。いづれも様方。おそろい</s></speech>
<s>里</s><warigaki><s>盃をはじめて大じんへさす</s><s>それより段/＼盃もすみて</s></warigaki>
```

これを利用し，人手修正済みのコーパスから，speech 要素内のテキスト，speech 要素外のテキスト，両方を含むテキスト全体（＝混合テキスト）の 3 種の学習用コーパスをそれぞれ作成し，3 種の辞書を作成した．3 種の辞書とも，見出し語のデータは洒落本用の語彙や表記を補った同一のもの(約 185800 語彙素，約 345500 書字形）を用いた．自動形態素解析を行う際，解析対

象 XML データ中の speech 要素でマークされた箇所には speech 要素内の言語による辞書を，それ以外の箇所には speech 要素タグ外の言語による辞書を使用し，専用の辞書で解析し分けた．比較用のベースラインとして混合テキストで学習した辞書でも解析を行った．その結果を表 2・表 3 に示す．なお，解析精度の評価は，小木曽[7]と同様に，UniDic の階層に合わせて「境界認定」「品詞認定」「語彙素認定」「発音形認定」の 4 段階で行い，精度は F 値で示した．

この結果，どの場合においても提案手法によって明らかな精度向上が認められた．また表 4 のとおり，会話文と地の文をあわせたテキスト全体でも，語彙素認定レベルでベースラインでは 0.8877 であったところが専用辞書の組み合わせでは 0.9085 と，大きく向上している．

表 2　洒落本・会話文の形態素解析精度

	ベースライン	提案手法
Lv.1 境界	0.9778	0.9799
Lv.2 品詞	0.9076	0.9250
Lv.3 語彙素	0.8980	0.9157
Lv.4 発音形	0.8928	0.9099

表 3　洒落本・地の文等の形態素解析精度

	ベースライン	提案手法
Lv.1 境界	0.9775	0.9771
Lv.2 品詞	0.8885	0.9151
Lv.3 語彙素	0.8725	0.8987
Lv.4 発音形	0.8663	0.8914

表 4 洒落本･全体での形態素解析精度

	ベースライン	提案手法
Lv.1 境界	0.9777	0.9788
Lv.2 品詞	0.8999	0.9211
Lv.3 語彙素	0.8877	0.9085
Lv.4 発音形	0.8821	0.9024

また，会話文と地の文では，会話文を対象としたほうがやや精度が良い状況であり，語彙素認定レベルで約 92%の精度を得ることができた．地の文等では，先の序文の例(6)に見られる当て字など標準的でない漢字表記が比較的多く出現することや，会話文の引用などが混在することが影響するとみられる．

まとめ

会話文・地の文で明確に文体・文法体系が分かれる洒落本では，形態素解析の際に文書構造の情報を利用することが有効であることが確認された．今後，コーパスの充実によりデータの蓄積を行うことで，一層の精度向上を目指すとともに，同様の文書構造を持つ人情本など，他の近世資料にも応用したい．

参考文献

[1] 国立国語研究所コーパス開発センター．日本語歴史コーパス バージョン 2015.3, 中納言バージョン 2.0.1, https://chunagon.ninjal.ac.jp/ (2016 年 1 月 7 日確認)，2015.

[2] 国立国語研究所コーパス開発センター(市村太郎ほか)(編)．ひまわり版「洒落本コーパス」(日本語歴史コーパス江戸時代編)バージョン 0.5, http://pj.ninjal.ac.jp/corpus_center/chj/edo.html#share (2016 年 1 月 7 日確認)，2015.

[3] 洒落本大成編集委員会(編). 洒落本大成(全 30 冊), 中央公論社，1978-88.

[4] 小木曽智信．形態素解析．前川喜久雄(監修)，山崎誠(編), 講座日本語コーパス 2 書き言葉コーパス―設計と構築―, 第 5 章, pp.89-115, 朝倉書店, 2014.

[5] 小木曽智信，鴻野知暁，市村太郎．狂言台本の形態素解析．日本語学会 2015 年度春季大会予稿集，pp.161-166，2015.

[6] 市村太郎．近世口語資料のコーパス化―狂言・洒落本のコーパス化の過程と課題―. 日本語学 11 月臨時増刊号日本語史研究と歴史コーパス 33(14)，pp.96109，2014

[7] 小木曽 智信，小町守，松本裕治．歴史的日本語資料を対象とした形態素解析, 自然言語処理 20(5), pp. 727-748，2013.