近世口語資料の形態素解析の試み

小木曽智信（国立国語研究所言語資源研究系）†
市村太郎（国立国語研究所コーパス開発センター）
鴻野知暁（国立国語研究所コーパス開発センター）

Preliminary Study of Morphological Analysis of Early Modern Japanese

Toshinobu Ogiso (National Institute for Japanese Language and Linguistics)
Taro Ichimura (National Institute for Japanese Language and Linguistics)
Tomoaki Kono (National Institute for Japanese Language and Linguistics)

１．はじめに

国立国語研究所では「日本語歴史コーパス」の一部として、近世の口語を反映した資料群のコーパス化計画を進めている（近藤 2012）。このうち、虎明本狂言集と洒落本の一部については、すでに電子化・形態論情報の付与に着手している。形態論情報の付与にあたっては、高精度で均質なタグ付けのために形態素解析が欠かせないが、従来の形態素解析辞書では近世の口語文を十分な精度で改正することができなかった。そのため、発表者らは UniDic（中古和文 UniDic・近代文語 UniDic）の見出し語データと整備済みの狂言・洒落本等のコーパスを用いて、近世口語文を解析するための辞書の作成に取り組んでいる。本発表では、この近世口語文を対象とした形態素解析の方法、現在達成している解析精度、今後の見通しについて論ずる。

２．狂言・洒落本のコーパス化

狂言は、中世から近世にかけての言語資料として重要な位置を占めている。登場人物が多彩で身分関係が明確であること、対話劇の形で進行し場面・状況が明確であることから、口語資料としての価値は極めて高い。狂言資料の中でも『虎明本』は、寛永 19 年（1642）大蔵流十三世宗家大蔵弥太郎虎明の手による大蔵流の祖本である。本狂言 237 曲を収めており、狂言の類別や詞章の整備された台本として、質・量とも第一級の資料である。その詞章には、中世、室町時代の言葉を伝承していると見られる点、書写当時である近世初期の日常語の影響を受けたと思われる点、舞台言語として整理され固定化・類型化する兆候が見られる点がある。狂言史上の位置を踏まえ、他の台本との比較ということが不可欠であるが、注釈書や総索引が整備され、中世から近世の言語資料として広く利用されてきた（小林・市村 2013）。虎明本では、狂言台本としての性格上、通常であれば漢字表記される語に仮名書きが多い特徴があり辞書未登録の表記が発生しやすく、形態素解析を難しくしている。一方、同一人による写本であるため、全体として均質性も持つ。以下に虎明本「あさう」の一部を掲げる。

例：（あさう）

「〽信濃の国の住人、あさうのなにがしです、そせうの子細あつて、在京仕る処に、安堵の御教書を下され、新地を拝領いたし、あまつさへおいとまを下された、のさ者をよび出し、よろこばせうとぞんずる、藤六あるかやい　〽お前に　〽下六もよべ　〽やい下六めすはやい　〽何とめすといふか　〽あふ　〽とういふてくれひで、お前に〽両人ながらはやかつた、やいなんぢらがよろこぶ事があるは、　〽そはまづめでたひ事で御ざあるが、何事で御ざあるぞ　〽永々在京いたす程にと有て、あんどの御教書を下され、新地を拝領して、おいとままで下されたが、かたじけなひ事ではなひか

† toigso@ninjal.ac.jp

洒落本は、登場人物の会話部分に当時の話し言葉が反映されているとされ、日本語史研究上、近世後期の口語の実態を探る上での重要資料である。大きく分けて江戸版と上方版があり、その口語体の会話部分はそれぞれの地域の言葉を反映する場合も多い。また年代も 18C 後半から 19C 前半までと幅広く、近・現代語への過渡的状況を伺うのに適している。方言や中央語の形成を知る上でも、不可欠な資料である（市村ほか 2013）。洒落本は、作品ごとに内容が異なるだけでなく、江戸・上方で言語そのものが大幅に異なっており、作者・形式も多様で均質性は低い。さらに言葉遊び的な要素をしばしば含むため、形態素解析やコーパス化には課題が多い。以下に洒落本の一つ『聖遊廓』の一部を掲げる。

例：(聖遊廓)

爰に聖人のかよひたまへる郭(くるわ)あり　揚屋(あげや)の亭主(ていす)は李白(りはく)とかや　中にも孔子はくるわにてすいといはれて端手(はで)ならず　ゑちご縮(ちゞみ)のかたびらにもんろの羽織(はおり)すそながく深(ふか)あみがさにあわざうり古金買(ふるかねかい)の目利(めきゝ)にも太夫かいとは見へざりし　李白がかたへ御入りあれば　▲亭主李白 是は仁さまおめづらしい さあ／＼おくへ ともてはやす　▲孔子 なんと李す此中は久しいの 無事で珍重／＼　と座敷へ行　▲李白女房滝(たき)　是はおめづらしいおかほ。 おうはさばつかり申ておりました　▲中居なつ もし仁さま此中横堀でお見うけ申ましたゆへ大かたおよりなさるであろふとぞんじましたに。よふまたせなさつたの　▲孔子 ヲヽよりたかつたけれども行時(ゆくとき)に径(こみち)によらず。

３．学習・評価用コーパス

狂言と洒落本のテキストのうち、現在、表１に示すものが単語情報付きのコーパスとして整備済みである。文書構造のアノテーション、濁点付与・文境界付与等の本文整備を施した後、既存の形態素解析辞書を用いて形態素解析を行って形態論情報データベースに格納し、その後人手によって解析の誤りを修正したものである。

表１　近世口語資料の人手修正済みコーパス

ジャンル	狂言	洒落本	人情本・滑稽本	*合計*
語数（短単位）	9515	25594	14361	*49470*

狂言の内訳は、次の 8 曲である。

「ゑびす大黒」「連歌毗沙門」「入間川」「あさう」「ゑびす毗沙門」「あさいな」「わかな」「腹不立」

洒落本の内訳は次の 5 作品である。本文は、『陽台遺編・妣閣秘言』『風流裸人形』『興斗月』は「洒落本大成」、『遊子方言』『跕婦人伝』は小学館「新編日本古典文学全集」によっている。

『遊子方言』　1770（明和 7）年，江戸
『跕婦人伝』　1749（寛延 2）年，江戸
『陽台遺編・妣閣秘言』　1758（宝暦 8）年，大阪
『風流裸人形』　1779（安永 8）年，京都
『興斗月』　1836（天保 7）年，京都

なお、人情本・滑稽本は、滑稽本が『浮世床』初編の一部、人情本が『春告鳥』初編の一部でいずれも本文は「新編日本古典文学全集」によっている。

表 1 のコーパスのうち、狂言・洒落本のそれぞれ約１割を文単位でランダムサンプリングして精度評価用コーパスを作成した。語数を表 2 に示す。

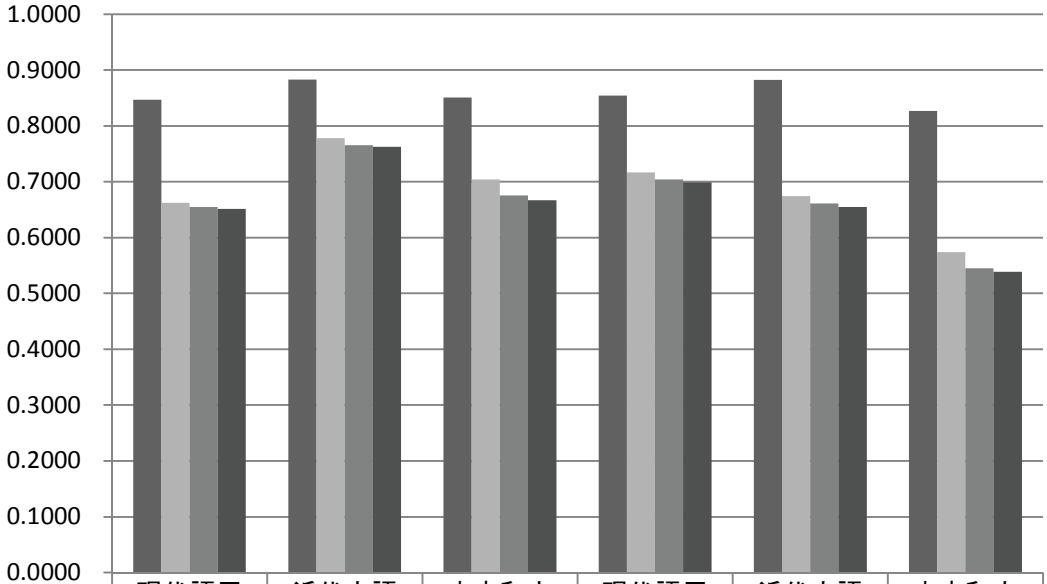
表２　狂言・洒落本の評価用コーパス

ジャンル	狂言	洒落本
語数（短単位）	1030	2448

４．既存の UniDic による解析精度

狂言・洒落本のテキストは、いずれも現代語とは大幅に異なる上に、典型的な古文である平安和文などとも大きく異なる文体で書かれている。発表者らは、これまでに歴史的な資料を対象とした形態素解析辞書として現代語用の UniDic をもとに「中古和文 UniDic」「近代文語 UniDic」を開発・公開してきたが、このいずれも狂言・洒落本の解析には適していない。

形態素解析器に MeCab（Kudo et al. 2004）[1]を用い、現代語用の UniDic と中古和文 UniDic、近代文語 UniDic のそれぞれで狂言・洒落本の評価用コーパスを解析し、精度を評価した。結果を図 1 に示す。数値は F 値（再現率と適合率の調和平均）である。



	現代語用	近代文語	中古和文	現代語用	近代文語	中古和文
	― 狂言 ―			― 洒落本 ―		
lv.1	0.8471	0.8829	0.8506	0.8543	0.8822	0.8269
lv.2	0.6626	0.7778	0.7040	0.7168	0.6744	0.5740
lv.3	0.6551	0.7653	0.6753	0.7044	0.6610	0.5450
lv.4	0.6513	0.7624	0.6667	0.6988	0.6549	0.5385

図 1　既存の UniDic による狂言・洒落本テキストの解析精度

図 1 の lv.1 は、解析結果において単語の境界が正しかったかどうか、lv.2 は境界が正しいことに加えて単語の品詞・活用型・活用形も正しく認定されていたかどうかを見るものである。lv.3 は lv.1 と lv.2 に加えて語彙素の認定も正しかったかどうかを見る。たとえば「金」が「キン」でなく「カネ」と解析されているかどうかという違いに相当する。lv.4 は、発音（読み）の違いが正しく認定されているかどうかを評価するもので、lv.1・lv.2・lv.3 が正しいことに加え、さらに読み方が正しいかどうかを見る。たとえば「言語」が文脈にあわせて「ゲンゴ」ではなく「ゴンゴ」と解析されているかどうかに相当する。

[1] 学習・解析ともに使用した MeCab のバージョンは 0.993 である。

現代語のコーパス構築において必要とされた解析精度が、語彙素認定（lv.3）において 98%であった。歴史コーパスの構築でも、おおむね 96%以上の解析精度が必要とされ、「中古和文 UniDic」「近代文語 UniDic」はほぼその解析精度を実現していた。しかし、図 2 にみるように、既存の解析辞書では近世口語資料の解析は難しく、狂言のテキストが最高で 76.5%、洒落本のテキストが最高で 70%程度の解析精度に留まっている。本格的なコーパス構築に用いるには大幅に精度が不足しており、新たな解析辞書を作成する必要がある。

５．近世口語共通辞書の解析精度

表 1 のコーパスのうち、表 2 の評価用を除く全てを利用して学習を行い、近世口語用の形態素解析辞書を作成した。見出し語は、従来の中古和文 UniDic・近代文語 UniDic で用いていたものに、表 1 のコーパスで出現した語を追加したものを利用した。見出し語は、活用形展開後で総計 134 万に上る。近世口語では利用されない語彙を含むが、①どの語が不要であるかを事前に判断することは必ずしも容易ではないこと、②古文の形態素解析辞書にとって見出し語の肥大化は大きな問題ではないこと、③不要語があることによる解析精度への悪影響は特に認められなかったこと、によりそのまま利用している。

表 3 にこの近世口語共通辞書の解析精度を示す。各レベルの意味は図 1 と同様、数値は F 値である。

表 3　近世口語共通辞書の解析精度

	狂言	洒落本
lv.1	0.9639	0.9681
lv.2	0.8652	0.8635
lv.3	0.8613	0.8545
lv.4	0.8594	0.8491

既存の辞書と比較すると精度は向上しているが、コーパス構築に十分な性能とはいえない。原因として学習用コーパスの絶対的な不足が考えられる。現状の学習用コーパスの量は約 4.5 万語だが、近代文語 UniDic では約 64 万語、中古和文 UniDic では約 82 万語を用いており、現在の学習用コーパスの量では不十分である。

しかし、量の問題とは別に、狂言と洒落本という質的にかなり異なるテキストを近世口語として一括していることにも問題があると考えられる。

６．狂言・洒落本専用辞書の解析精度

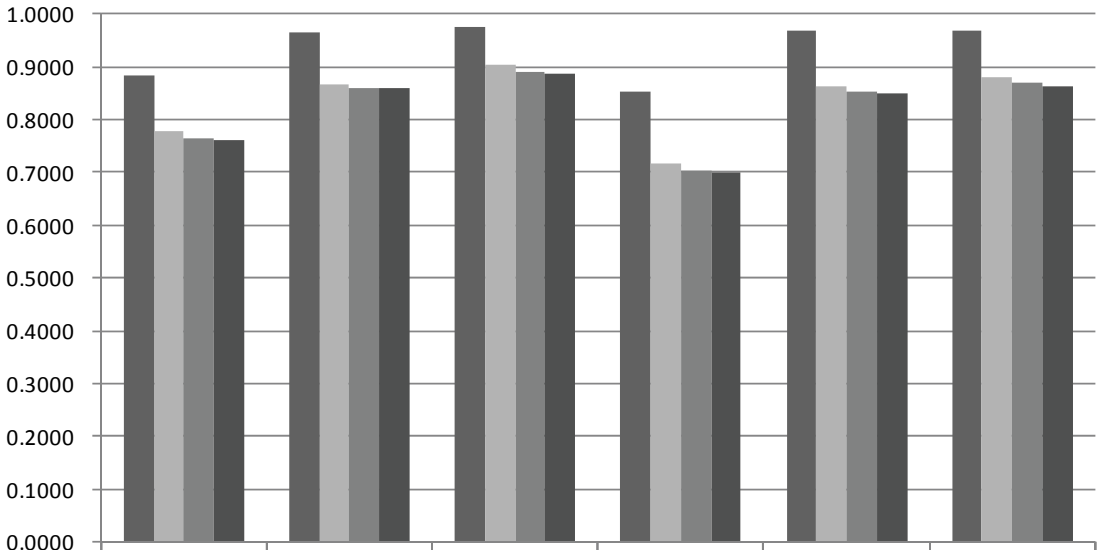
予備的な実験により、歴史的資料の形態素解析を行う際には、異分野のテキストによる学習結果を流用するより、少量であっても専用コーパスによる学習が効果的であることが分かっている。そこで、狂言と洒落本を分割し、それぞれの専用辞書を作成して近世口語共通の辞書と解析精度を比較することにする。分割により、もともと不十分であった学習用コーパスの量はほぼ半減することになるが、専用の学習用コーパスのみを利用することによるメリットがそれを上回る可能性がある。

狂言の学習は狂言のコーパスだけを学習に利用し、洒落本は、滑稽本・人情本に時代的にも内容的にも比較的近いため、これらを利用するもの（A）と純粋に洒落本だけを利用するもの（B）の 2 通りを作成した。それぞれの辞書による解析精度を表 4 に示す。見方は表 3 と同様であるが、表 4 では狂言と洒落本が、それぞれ別の辞書による評価結果となっていることに注意されたい。表 4 の数値は、学習用コーパスの量を大幅に減らしたものであるにもかかわらず、表 3 の共通辞書による解析精度よりも向上している。したがって、狂言と洒落本とでは別の解析辞書を用意すべきであることが分かる。洒落本の（A）（B）についても、滑稽本や人情本を交えない（B）のほうがよりよい精度となっている。

表 4　狂言・洒落本専用辞書の解析精度

	狂言専用	洒落本用（A）	洒落本専用（B）
lv.1	0.9747	0.9695	0.9699
lv.2	0.9026	0.8738	0.8791
lv.3	0.8900	0.8636	0.8688
lv.4	0.8870	0.8583	0.8627

図 2 は、これまでに精度を確認してきた辞書による解析精度をグラフにまとめたものである。既存の辞書による解析結果のうち最良のもの（狂言は近代文語 UniDic、洒落本は現代語用 UniDic）と、近世口語共通辞書、専用辞書の解析結果を比較した。



	既存最良（近代文語）	共通辞書	専用辞書	既存最良（現代語）	共通辞書	専用辞書
	― 狂言 ―			― 洒落本 ―		
lv.1	0.8829	0.9639	0.9747	0.8543	0.9681	0.969919
lv.2	0.7778	0.8652	0.9026	0.7168	0.8635	0.879065
lv.3	0.7653	0.8613	0.8900	0.7044	0.8545	0.868834
lv.4	0.7624	0.8594	0.8870	0.6988	0.8491	0.862696

図 2　各辞書による狂言・洒落本テキストの解析精度比較

このように提案手法により精度はかなり向上したものの、現状の解析精度は狂言・洒落本ともに lv.3（語彙素認定）で 9 割を切っており、コーパス構築のために十分な精度には達していない。これは、学習用コーパスの量の絶対的な不足に起因するものと考えられる。

7．おわりに

新たにコーパスからの学習を行って狂言用と洒落本用の形態素解析辞書を作成し、既存の辞書を上回る精度で解析を行うことが可能になった。また、近世口語を一括した共通辞書を作成する場合と、対象分野を分割した専用辞書を作成する場合とで精度の比較を行うことにより、狂言と洒落本とは別に扱うことが良いことが確認された。

現状では、狂言と洒落本を分割すると学習用のコーパスが大きく不足するため、解析精度は語彙素認定で 90%に達していない。今後、それぞれの学習用コーパスの量を増やし、見出し語の増補を行うことで専用辞書を充実させ、狂言・洒落本のコーパス構築に資するものにしていきたい。

謝　辞

本研究は JSPS 科研費 24520522 の助成を受けたものである。また、本研究の一部は国立国語研究所の共同研究プロジェクト「通時コーパスの設計」および「統計と機械学習による日本語史研究」による研究成果を含む。

文　献

市村 太郎，河瀬 彰宏，小木曽 智信(2012)「近世口語テキストの構造化とその課題」情報処理学会研究報告 人文科学とコンピュータ研究会報告(CH96) pp.1-8

市村 太郎，河瀬 彰宏，小木曽 智信(2013)「洒落本コーパスの構造化 ―仕様と事例の検討―」『第 3 回コーパス日本語学ワークショップ予稿集』pp.249-258

小林 正行，市村 太郎（2013）「『虎明本狂言集』コーパスの構造化 －仕様と事例の検討－」『第 3 回コーパス日本語学ワークショップ予稿集』pp.323-332

近藤泰弘(2012)「日本語通時コーパスの設計について」『国語研プロジェクトレビュー』3 pp.84-92

伝康晴，小木曽智信，小椋秀樹，山田篤，峯松信明，内元清貴，小磯花絵(2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』22 号 pp.101-122.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. (2004). Applying conditional random fields to Japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing* , Barcelona, Spain, pp.230–237.

関連 URL

MeCab: Yet Another Part-of-Speech and Morphological Analyzer　https://code.google.com/p/mecab/

NINJAL「通時コーパス」プロジェクト　http://historicalcorpus.jp

UniDic　http://sourceforge.jp/projects/unidic/

近代文語 UniDic，中古和文 UniDic　http://www2.ninjal.ac.jp/lrc/index.php?UniDic