『虎明本狂言集』における会話文の計量分析

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

1　はじめに

文芸―物語，評論，随筆，戯曲などの言語を媒介とする芸術―の理解は，人間の高次感性と高度な知識操作（情報処理）を伴うものであり，そのプロセスの解明は，おもに心理学と認知科学の分野において進められてきた．

とくに娯楽の対象である物語や戯曲は，虚構の世界が描写されており，鑑賞者は，背景知識とテキスト―登場人物の言説・行為―から得られる情報を統合することで人物・出来事・場面を表象（理解）する [1]．

本研究では，これまで自然言語処理の研究対象とされてこなかった日本伝統芸能の笑劇である狂言を分析対象とし，狂言台本のテキストに計量分析を施すことで，登場人物の身分・役割がどのように規定されるのか，語彙の側面から検証する．具体的には，狂言の登場人物の会話文を身分・役割ごとに分類したテキストコーパスを構築し，頻出語彙の共起ネットワークと統計指標を用いて特徴を可視化することで，各々の身分・役割に特有の概念を比較する．

2　狂言テキストの特徴とその意義

狂言は，中世から近世にかけて重要な位置を占める言語資料である．とくに次の 3 点の特徴を有することから口語資料としての価値は極めて高い：(1) 物語がダイアログ形式で進行すること；(2) 会話部分と場面・状況描写が明確であること；(3) 登場人物が多彩であり，身分・役割が分かれていること．

したがって，狂言台本のテキストを機械可読な形式に整備し，計量分析を施すことは，近世口語の実態解明に寄与するだけでなく，日本語史や書誌学などの人文学的研究を促進する意義がある．

これまで日本語学において狂言を扱った研究には，例えば特定の助動詞に着目した用例分析 [2] や語彙の使用傾向をめぐる特定の作品の台本比較 [3] などが伝統的に進められているものの，狂言台本に含まれる作品群を対象とした巨視的な研究は行われていなかった．とくに，狂言の登場人物の言説を網羅的に分析し，その人物が担う役割を読み解いた研究は存在しない．本研究では，コーパス言語学の観点からこれを実施する．

3　分析対象

3.1　『虎明本狂言集』のテキストコーパス

国立国語研究所（以下，国語研）では，「日本語歴史コーパス設計」プロジェクトの一環として近世口語資料などの古典資料の電子化・構造化・形態素解析を実施している [4]．

本研究では，国語研が電子化・構造化を進めている『虎明本狂言集』（以下，『虎明本』）のテキストコーパス [5] を用いる．

狂言テキストそのものは，台詞とト書きから成る台本本文に，舞台外に関する注釈を付与した演劇資料である．複数の狂言資料の中でも『虎明本』は，寛永 19 年（1642）に大蔵流十三世宗家大蔵弥太郎が手掛けた大蔵流の祖本であり，計 237 曲の作品が収められている．『虎明本』の各作品は，テーマごとに 8 つに類別され，詞章を整備した質・量ともに第一級の狂言台本である．

本研究では．全 8 類のうち最初の 4 類―脇狂言之類，大名狂言類，聟類・山伏類，鬼類・小名類―に含まれる全作品を対象に，登場人物の会話文を身分・役割に応じて分類し，新たにテキストコーパスを準備する．使用するテキスト数は 124，サイズは 2.29MB である．

3.2　狂言の形態素解析

『虎明本』では，狂言台本の性格上，通常であれば漢字表記されるべき語に仮名書きが多く使用されている．必然的に，このようなテキストには辞書未登録の表記が多く含まれるため，形態素解析を困難にしてし

まう．この問題に対して，小木曽ら [6] は，新たにコーパスから学習を行い，既存の辞書の解析精度を上回る狂言専用の形態素解析辞書を作成した．

表 1 に，狂言テキストの最新の解析精度を示す．解析精度は Lv.1 から Lv.4 まで 4 つの段階に分けてあり，それぞれ次の点を評価している：(Lv.1) 単語の境界判定が正しいかどうか；(Lv.2) Lv.1 に加えて，単語の品詞・活用型・活用形の認定が正しいかどうか；(Lv.3) Lv.2 に加えて，語彙素の認定が正しいかどうか；(Lv.4) Lv.3 に加えて，発音（読み）の違いの認定が正しいかどうか．ただし，数値は F 値（再現率と適合率の調和平均）であり，学習・解析には形態素解析器 MeCab のバージョン 0.993[7] を用いている．本研究では，この狂言専用の形態素解析辞書に対してさらに人手修正を加えて改良した最新の辞書を用いて，登場人物の会話文を形態素解析する．

表 1: 狂言の解析精度（%）

Lv.1	Lv.2	Lv.3	Lv.4
0.9859	0.9583	0.9512	0.9486

3.3 登場人物の身分・役割と発話回数

4 類のうち発話回数が多い登場人物の上位 10 位を，語彙の総数とともに表 2 に示す．

表 2 より，(1 位) 冠者・下人と (2 位) 主は，発話回数・語彙の総数が他の登場人物よりも圧倒的に多いことがわかる．続いて (3 位) 商人，(5 位) 百姓・漁師，(7 位) すっぱ，(8 位) 山伏，(9 位) 出家などの職種，(4 位) 聟，(6 位) 舅，(10 位) 妻・女などの親類関係が多く登場する．これらのうち，冠者・下人と主は主従関係，聟と舅は家族関係にあり，身分の差から使用される言語が異なることが報告されている [8]．

また，会話 1 回あたりに使用する平均語彙数（語彙の総数 ÷ 発話回数）を求めると，上位 10 位以内の登場人物では，山伏が最も多く 27.49 語使用し，出家が最も少なく 4.20 語，続いて，すっぱが 7.51 語使用する．それ以外の身分・役割はおおよそ 20 – 24 語の範囲内で語彙を使用することがわかった．

以降では，表 2 に列挙した上位 10 位までの登場人物を対象に身分・役割の違いを比較する．

4 ネットワーク分析

4.1 頻出語彙の共起ネットワーク

ネットワーク分析とは，内容の概念構造をネットワークを用いて計量的に可視化する分析方法である．テキストからネットワークを作成することは，概念同士の関連を視覚的に捉え易くするだけでなく，グラフ理論・情報理論で培われた様々な解析手法をテキスト分析に適用できる利点がある．

表 2: 登場人物の発話回数（上位 10 位）

順位	登場人物	発話回数	語彙の総数
1	冠者・下人	1,565	31,884
2	主	1,205	26,262
3	商人	271	6,335
4	聟	259	5,819
5	百姓・漁師	206	4,400
6	舅	130	3,216
7	すっぱ	120	901
8	山伏	107	2,941
9	出家	104	437
10	妻・女	101	1,973

表 2 に列挙した上位 10 位の登場人物の会話文を形態素解析し，登場人物ごとに名詞の頻出語彙の共起ネットワークを作成する．紙面の制約上，例として図 1 に (3 位) 商人，図 2 に (5 位) 百姓・漁師の共起ネットワークのみ掲載する．ただし，ここでは頻出語彙上位 50 語に着目し，頻出語彙の影響力をネットワークに反映させるために，ノードの大きさは語彙の出現頻度の 2 乗根に比例させ，エッジの距離は共起頻度に反比例させている．

4.2 ネットワーク中心性

各ネットワークにおける中心概念を捉えるために，複数のネットワーク中心性を計算し，中心性の値が上位にある語彙を求めた．表 3 にその結果の一部を示す．

計算の結果 (表 3) より，今回分析した全ての身分・役割において「事」「物」「者」「程」「人」「所」の語彙が共通して上位に出現した．寿岳 [9] によれば，「程」を除くこれらの語彙は，狂言の基本語彙の上位 50 位に含まれており，これらの語彙が狂言作品の会話文一般における言及対象の存在であることがわかる．

その一方で，登場人物ごとに特徴的な語彙も上位に出現する．例えば，(3 位) 商人では「商売」「羯鼓」「棚」，(5 位) 百姓・漁師では「年貢」「国」「上頭」，(8 位) 山伏では「苦行」「難行」「数珠」などが含まれる．

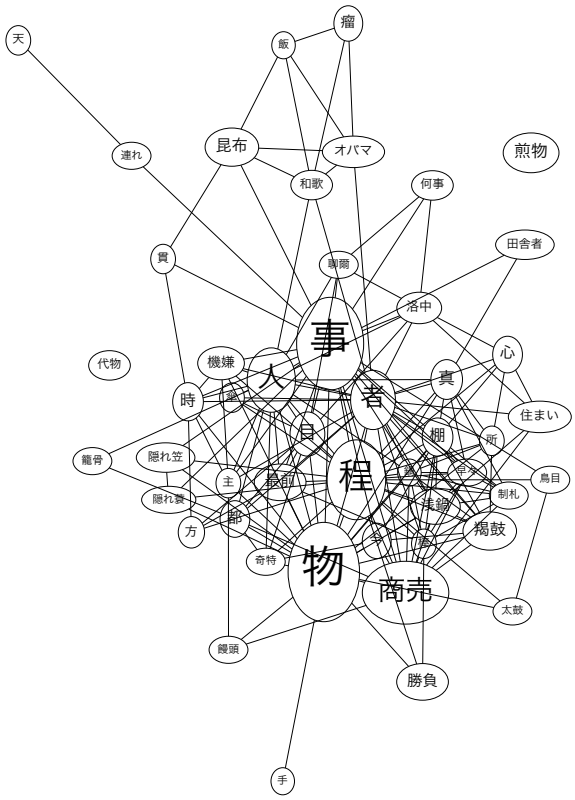

図 1: 商人の共起ネットワーク

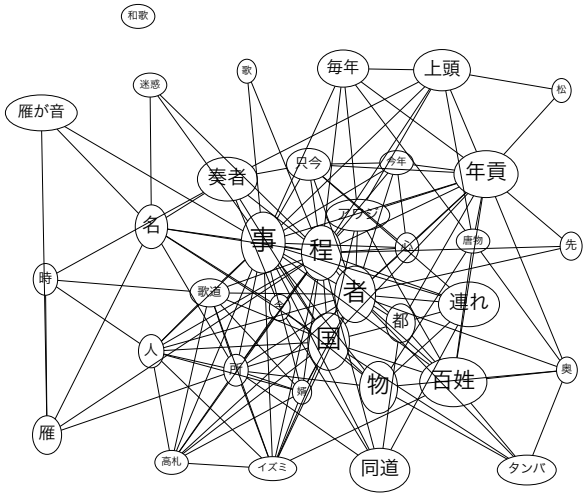

図 2: 百姓・漁師の共起ネットワーク

表 3: ネットワーク中心性（上位 5 位）

(i) 冠者・下人

近接		媒介	
程	4.43	程	28.42
事	4.42	事	23.58
人	4.39	物	9.13
物	4.38	人	8.60
者	4.37	者	4.79

(iii) 商人

近接		媒介	
程	5.60	程	24.86
物	5.57	事	19.79
事	5.56	人	17.98
人	5.55	物	16.70
者	5.54	者	13.68

(v) 百姓・漁師

近接		媒介	
事	4.17	事	18.29
程	4.17	程	17.23
者	4.15	物	11.00
年貢	4.15	年貢	10.66
国	4.14	雁	10.24

(viii) 山伏

近接		媒介	
山伏	5.62	程	21.93
程	5.61	山伏	16.35
行	5.58	事	11.48
難行	5.57	行	7.67
奇特	5.56	駆け出	7.54

5 身分・役割に特徴的な語彙

各登場人物の会話文から作成した頻出語彙の共起ネットワークは，共通する語彙が上位に出現してしまうため，身分・役割に特有の語彙が見え難いという問題がある．ここでは村井・徃住 [10] で用いられている $CIDF$（Comparative IDF）を狂言のテキストコーパスに適用し，登場人物の身分・役割に特有の語彙を抽出する．

$CIDF_i$ は，任意の単語 t_i について，idf_{a}（全テキスト D_{a} において，その単語が出現するテキスト数 d_{a} から求まる IDF 値）と，idf_{g}（特定ジャンル―ここでは登場人物の身分・役割―の全テキスト D_{g} において，その単語が出現するテキスト数 d_{g} から求まる IDF 値）の比として定義される指標である：

$$CIDF_i = \frac{idf_{\mathrm{a}}}{idf_{\mathrm{g}}} = \frac{\left(\log \dfrac{|D_{\mathrm{a}}|}{|\{d_{\mathrm{a}} : t_i \in d_{\mathrm{a}}\}|}\right)}{\left(\log \dfrac{|D_{\mathrm{g}}|}{|\{d_{\mathrm{g}} : t_i \in d_{\mathrm{g}}\}|}\right)}.$$

ただし，出現語彙数が少ない場合に $CIDF_i$ 値が極端に高くなるため，身分・役割に特徴的な語彙を厳密に決定するためには，より多くのテキストデータを分析対象に用いる必要がある．

表 4 は，各登場人物の会話文中に出現する名詞について計算した $CIDF_i$ 値上位 10 語である．この結果では，ネットワークの中心性について共通して出現した「事」「物」「者」「程」といった語彙は，いずれの登場人物についても順位が低く，各々の身分・役割に特有の語彙が上位に抽出されていることがわかる．

狂言の登場人物はこれらの特徴的な語彙（言及対象・概念）を使い分けることで場面における身分・役割を規定していると考えることができる．

表 4: 身分・役割に特徴的な語彙（上位 5 位）

（i）冠者・下人

語彙	$CIDF_i$
御意	1.678
由	1.422
一所	1.419
供	1.412
普請	1.386

（ii）主

語彙	$CIDF_i$
辺り	2.824
住まい	2.292
冠者	2.128
太郎	2.007
奴	1.884

（iii）商人

語彙	$CIDF_i$
田舎者	2.827
買い手	2.749
代物	2.616
勝負	2.597
洛中	2.597

（iv）聟

語彙	$CIDF_i$
婿入り	6.642
婿	6.166
無音	3.616
花婿	3.150
礼	3.045

（v）百姓・漁師

語彙	$CIDF_i$
国	∞
百姓	∞
年貢	24.363
上頭	23.481
毎年	9.621

（vi）舅

語彙	$CIDF_i$
婿	5.117
御出で	3.799
内々	3.205
娘	2.758
高札	2.461

（vii）すっぱ

語彙	$CIDF_i$
心	6.693
カミギョウ	3.847
サンジョウ	3.847
祇候	3.847
後日	3.847

（viii）山伏

語彙	$CIDF_i$
祈り	24.363
ぼろおん	12.309
山伏	10.899
駆け出	7.922
カズラキ	7.922

（ix）出家

語彙	$CIDF_i$
法	∞
タテヤマ	7.693
経	7.693
住持	7.693
本寺	7.693

（x）妻・女

語彙	$CIDF_i$
父	4.867
上	3.592
物見	3.085
髪	2.475
夫	2.427

6 おわりに

本研究では，狂言台本のテキストに対して，ネットワーク中心性と $CIDF$ を指標として利用し，登場人物の身分・役割を規定する特徴的な語彙（言及対象・概念）を抽出した．今後は，狂言台本に含まれる全 8 類・計 237 曲を電子化し，テキストコーパスを整備することで，登場人物の身分・役割を規定する概念の抽出だけでなく，これまで実施されて来なかった 8 類の構造比較に着手する．

また，本研究では事物の概念を表す名詞にのみ着目したが，名詞以外の品詞―登場人物の動作を表す動詞，性質を表す形容詞，従来日本語学で調査されてきた助動詞など―の使用傾向を分析することで，狂言作品の世界をより立体的に捉えることを目指す．とくに，狂言テキストの分析結果を同時代の洒落本コーパス [11] や，他の時代の和文コーパスの分析結果と比較し，通時的に日本語の歴史的変化を探ることで，近世口語の実態解明に寄与していきたい．

参考文献

[1] 良峯徳和, 徃住彰文. 虚構言説理解過程の制御システムモデル. *Cognitive Studies*, Vol. 8, No. 4, pp. 384–391, 2001.

[2] 蜂谷清人.「おりゃらします」考. 狂言台本の国語学的研究, 第 8 章, pp. 149–164. 笠間叢書, 1977.

[3] 張元哉. 狂言台本の「鬼の継子」における語彙の計量的考察. 日本研究, Vol. 21, pp. 377–390, 2003.

[4] 市村太郎, 河瀬彰宏, 小木曽智信. 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報告, Vol. CH96, pp. 1–8, 2012.

[5] 小林正行, 市村太郎.『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, 2013.

[6] 小木曽智信, 市村太郎, 鴻野知暁. 近世口語資料の形態素解析の試み. 第 4 回コーパス日本語学ワークショップ予稿集, pp. 145–150, 2013.

[7] 工藤拓. Mecab: Yet another japanese dependency structure analyzer. `https://code.google.com/p/mecab/`（2014.01.20 参照）.

[8] 伊藤達弘, 笹川祥生, 藪田夏雄. 狂言における人間関係と言語 特に主従関係について. 京都府立大学国語国文学会誌, Vol. 3, pp. 41–52, 1962.

[9] 寿岳章子. 近代の文体. 佐藤喜代治（編）, 文体史・言語生活史, 第 3 章, pp. 115–167. 大修館書店, 1972.

[10] 村井源, 徃住彰文. テキスト批評の計量化に向けて―書評の計量分析―. 情報知識学会誌, Vol. 19, No. 2, pp. 120–125, 2009.

[11] 河瀬彰宏, 市村太郎, 小木曽智信. TEI:P5 に基づく近世口語資料の構造化とその問題点. じんもんこん 2013 論文集, pp. 7–12, 2013.