女性シンガーソングライターの歌詞の探索的分析

細谷 舞
東洋大学 社会学部

鈴木 崇史
東洋大学 社会学部

本研究では，現代日本の女性シンガーソングライターの楽曲の歌詞を探索的に分析する．歌詞は，従来から，その時代の社会や文化をあらわす表象として重要なものと考えられてきた．とりわけ，女性シンガーソングライターは，現代日本のヒットチャートの 1 ジャンルを構成するものであり，新たな日本語やポピュラーカルチャーを理解するために重要な役割をもっている．本研究では，計算文体論の最新の手法を用いて，10 人の女性シンガーソングライターの楽曲の歌詞を探索的に分析した．結果，それぞれの歌手の歌詞の特徴が明確に示された．本研究は，現代日本の社会や文化を理解するための重要な知見を提出するものである．

Exploratory analysis of popular songs made by Japanese female singer-songwriters

Mai Hosoya
Faculty of Sociology
Toyo University

Takafumi Suzuki
Faculty of Sociology
Toyo University

In this paper, we analyzed popular songs made by Japanese female singer-songwriters. Popular songs are a good representation of the modern culture and society they are created in and spread throughout. The songs by female singer-songwriters account for a large part of the current Japanese hit charts, and have especially important roles for understanding Japanese popular culture. In this paper, we applied new methods in computational stylistics to texts of the songs. The results clearly show the respective character of the ten female singer-songwriters. Our findings will be very important for understanding modern Japanese popular culture and society.

1．はじめに

大量・多種データの蓄積と自然言語処理技術の進展により，計算機ベースのテキスト研究は，大いに進展している．著者推定やジャンル判定など計算文体論の技術は，人文科学研究にとって欠かせないものになりつつある[3; 7; 19].

一方，流行歌は，その時代や社会の表象として，重要な研究対象となってきた[17]．とりわけ，女性シンガーソングライターは，現在，日本全体の楽曲売り上げにおいて，大きな部分を占め[22; 23]，これを分析することは，新たな日本語やポピュラーカルチャーを理解するために，重要な意義をもつ[9-15].

このような背景のもと，本研究では，過去 30 年間にわたる，現代日本の女性シンガーソングライターの楽曲の歌詞を，計算文体論の手法を適用することで，探索的に分析する．

計算文体論の方法論的観点からは，本研究で扱う問題は，(a)延べ語数が極めて少ない，(b)複数の要因（歌手，発行年等）が影響し得る，(c)文体のみならず，内容も関連する，(d)テキスト分類の性能のみならず，分類結果の言語学的，社会学的解釈に関心がある，以上の特徴をもつ．これらの点に留意して，特徴量と統計手法を選択することで，方法論的にも有意義なケーススタディを提供する．本研究は，現代日本の文化と社会の理解に実質的な貢献をもたらすとともに，計算文体論に方法論的な貢献をもたらすものである．

本研究，次節以降の構成は以下の通りである．まず，第 2 節で関連研究を整理し，第 3 節で本研究に用いたデータについて，第 4 節で統計手法について記述する．第 5 節で本研究の結果を議論し，第 6 節で本研究全体を総括する．

2．関連研究

計算文体論の領域では，すでに従来から，計算機を用いた計量的文体分析が多くなされてきた[3; 7; 19]．その技術は，従来からある，著者推定，ジャンル判定，テキストの時代測定などの他，近年では，スパムフィルタリン

グや犯罪捜査など，様々な応用に利用されている[1]．また，従来，この分野では，テキスト分類の問題として研究が行われてきたのに対し，近年，分類に有効な特徴量に注目し，これを解釈に利用することに主眼を置く研究が注目を集めている[1; 4]．本研究でも，分類問題を適用するのみならず，分類に有効な特徴量に注目し，これを分析することで，言語学的，社会学的解釈を導き出すことをめざす．

一方，歌謡曲，流行歌を題材として，その時代背景，文化，社会等について，社会学的分析を行った研究は多くある[5; 16; 19-21]．これらは，一部，歌詞の分析も含んでいるものの，その中心は，背景となる資料やインタビューの質的な分析であり，個々の歌手に関する議論も，これらをもとに，社会的要因から，説明しようとするものが多い．これに対して，本研究は，歌詞自体をより直接的，積極的に分析対象とし，そこから歌手，さらには，その時代の文化や社会の特徴を明らかにしようとするものである．

歌詞の言語学的分析は，古くからあり[17]，とりわけ，伊藤の一連の研究[9-15]は，同時期の歌謡曲を分析対象としているという点で，本研究と関心を同じくするものである．しかし，これら従来研究は，特定の歌手に分析対象を絞り，また，個々の語や，外来語・日本語比に注目して研究を行っている．これに対し，本研究は，より体系的に時代を代表するシンガーソングライターを抽出し，また，計算文体論の最新の手法を適用する点に特色がある．

3. データ

1979 年から 2009 年までの，オリコン年間シングルランキング TOP100 [22; 23]より，調査対象全期間で 5 曲以上ランクインした女性シンガーソングライター10 名 116 曲を選択した．「歌ネット」[1]，「うたまっぷ」[2]を参照し，歌詞のテキストファイルを作成[3]，タイトル，スペース，説明個所を削除した後，MeCab[4] による形態素解析を適用，全ての語の相対頻度を計算し，テキスト - 特徴量行列を作成，これを特徴量として利用することとした．[5] 全テキストにおける，延べ語数は 27,663，異なり語数は 3,971 である．

既存の，著者推定，ジャンル判定では，機能語や文字 n-gram が頑強な分類結果をもたらすことが知られているが，上述(d)の理由から，文字ベースよりも形態素ベースの特徴量が適切であると判断し，さらに，(a)，(c)の理由から，機能語のみならず，内容語も含めることが適当であると判断した．

4. 統計手法

4.1. カーネル主成分分析

上述，(b)の理由から，まず，テキスト - 特徴量行列に対し，カーネル主成分分析を適用することとした．これによって，特徴量に影響を与える主要因を抽出し，また，テキスト間の距離を可視化できる．カーネル関数として，ガウシアンカーネルを用い，パラメータは，$\sigma=0.1$ を与えた．従来研究では，主成分分析，因子分析，対応分析等が利用されてきたが[6]，カーネル主成分分析は，これらより，柔軟なパラメータ設定が可能であり，探索的分析のためには，より適切であると考え，本研究ではこれを利用することとした．

4.2. ランダムフォレスト機械学習法

次に，以下の要領で，ランダムフォレスト機械学習法[2]を適用し，歌手ごとに，テキスト分類を行った．まず，i 行，j 列のテキスト - 特徴量行列から，重複を許して列（特徴量）を複製し，1,000 個のブートストラップサンプルを作成，それぞれのブートストラップサンプルから$\sqrt{j}$ 列をランダムサンプリングで抽出した．それぞれのランダムサンプリングから，ジニ計数を用いて決定木を作成，ブートストラップサンプルの 2/3 の多数決から新しい決定木を構築し，残りの 1/3 を評価に用いた．[6] 実験の評価には，エラー率を用いた[6]．分類結果を観察することで，データセットに対して，特別な特徴をもつ（分類性能の高い）歌手，特徴の少ない（分類性能の低い）歌手を明らかにすることができる．

---

[1] www.uta-net.com

[2] www.utamap.com

[3] シンガーソングライターは，90%以上の楽曲で，自ら作詞作曲を行った歌手と定義した．

[4] mecab.sourceforge.net

[5] いわゆる，Bag-of-words モデルである．

[6] これを，Out-of-Bag テストと呼び，評価用データを OOB データと呼ぶ．

表１　基本データ

	曲数	延べ語数		
		平均	標準偏差	変動係数
aiko	13	217.08	49.59	22.84
広瀬香美	7	273.00	31.37	14.45
松任谷由実	13	155.31	40.51	18.66
中島みゆき	20	211.30	35.82	16.50
大黒摩季	13	231.38	48.84	22.50
大塚愛	15	209.33	49.99	23.03
椎名林檎	9	199.78	46.19	21.28
竹内まりや	8	199.25	48.22	22.21
宇多田ヒカル	20	235.70	51.69	23.81
YUI	10	243.10	68.61	31.61

さらに，以下で示される変数の重要度（VI_{acu}）を計算する．

$$VI_{acu} = \frac{mean(C_{oob} - C_{per})}{s.e.}$$

ただし，ここで，C_{oob}は，OOBデータにおける正解数，C_{per}は，OOBデータのm変数をランダムに置換した場合の正解数，$s.e.$は，標準誤差をあらわす．VI_{acu}によって，分類に有効な変数，すなわち，分類に寄与の大きい重要語を明らかにすることができる．本研究では，分類実験全体および，それぞれの歌手の分類について，有効な変数の抽出を行った．

ランダムフォレストは，日本語の著者推定に関して，もっとも高い性能が確認されており[8]，マルチクラスへも対応している．さらに分類に有効な特徴量を分類実験と同時に計算することができることから，本研究の目的に対して最適である[4]．

５．結果と考察

5.1.　基礎的な観察結果

表１は，それぞれの歌手について，分析期間にランクインした楽曲数ならびに１曲あたりの延べ語数を示したものである．楽曲数は7曲から20曲の範囲にあり，１曲あたり延べ語数は，155語から273語の範囲にある．宇多田ヒカル，中島みゆきなどは，ランクインする曲が多く，広瀬香美，竹内まりや，椎名林檎などは，相対的にランクインする曲が少ない．広瀬香美は相対的に曲が長く，松任谷由実は，相対的に曲が短い．また，YUIや宇多田ヒカルは，標準偏差，変動係数が高いことが注目される．[7]

5.2.　カーネル主成分分析の結果

図１は，カーネル主成分分析の結果を示したものである．歌手名をラベルとして，第三主成分までをプロットしてある．宇多田ヒカルやYUIが右手に，中島みゆきが左手に，aikoが上手にといった，歌手ごとの一定の集中傾向が観察される．一方，発行年の影響は，自明には観察されない．[8] 第一主成分に外来語の影響（宇多田ヒカル，YUI），第三主成分に代名詞の影響（aiko）が他よりもやや強いと示唆されるが，三軸とも，様々な語から構成されており，軸の解釈は自明ではない．

[7] この点は，外来語の影響ではないかと推察される．

[8] 発行年をラベルとした散布図も作成したが，結果は同様であった．

図１　カーネル主成分分析の結果

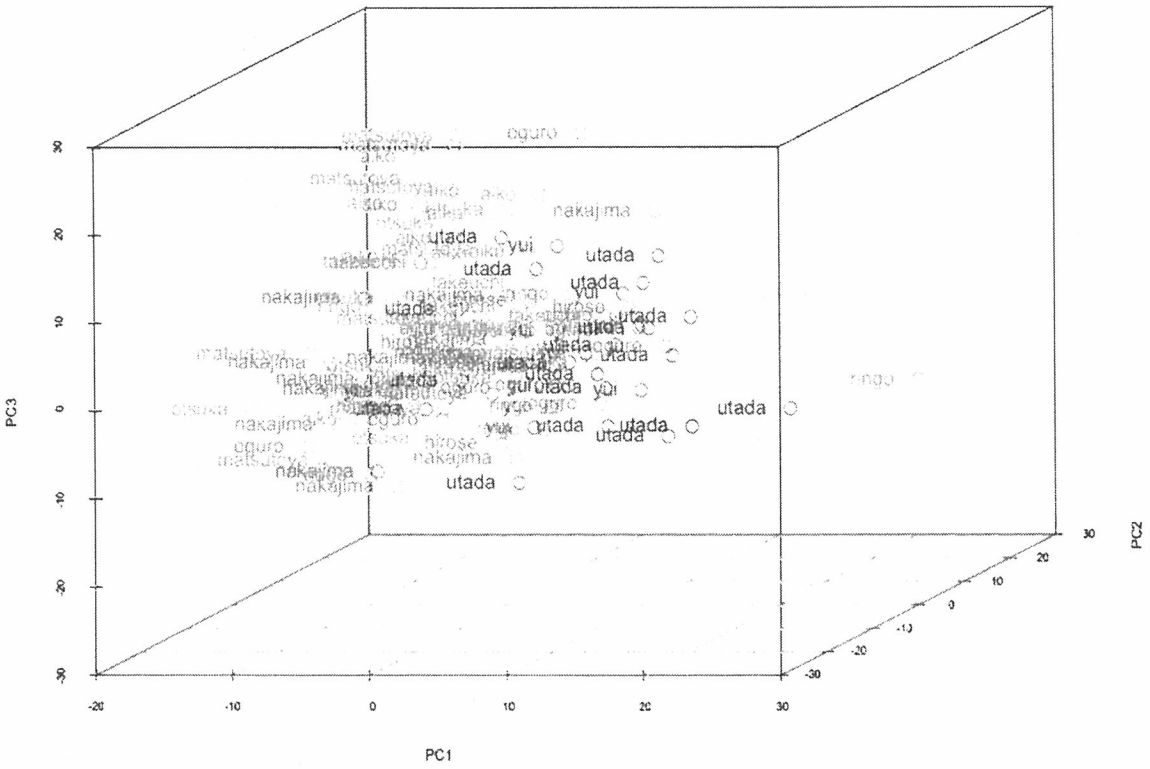

表２　ランダムフォレストの結果

	aiko	広瀬香美	松任谷由実	中島みゆき	大黒摩季	大塚愛	椎名林檎	竹内まりや	宇多田ヒカル	YUI	エラー率
aiko	13	0	0	0	0	0	0	0	0	0	0.00
広瀬香美	0	0	0	4	0	0	0	0	3	0	1.00
松任谷由実	0	0	2	10	0	0	0	0	1	0	0.85
中島みゆき	0	0	0	19	0	0	0	0	1	0	0.05
大黒摩季	0	0	1	1	1	2	0	0	8	0	0.92
大塚愛	1	0	0	2	0	10	0	0	2	0	0.33
椎名林檎	1	0	1	1	0	0	2	0	4	0	0.78
竹内まりや	0	0	0	1	1	0	0	2	3	1	0.75
宇多田ヒカル	0	0	0	3	0	0	0	0	17	0	0.15
YUI	0	0	0	0	0	0	0	0	5	5	0.50

表３　分類実験全体に帰与の大きい特徴量

	語	VI_{acu}
1	あたし	0.0125
2	し	0.0078
3	1	0.0067
4	私	0.0059
5	様	0.0053
6	だ	0.0052
7	を	0.0050
8	幸せ	0.0047
9	た	0.0045
10	君	0.0045
11	なんて	0.0039
12	いい	0.0039
13	へ	0.0038
14	よ	0.0036
15	愛	0.0032
16	ん	0.0032
17	たい	0.0032
18	ない	0.0032
19	から	0.0031
20	事	0.0031

5.3.　ランダムフォレストによる分類結果

表２は，ランダムフォレストによる分類結果を示したものである．行は，被判別テキスト，列が判別結果をあらわす．エラー率からは，aiko，中島みゆき，宇多田ヒカルらは，分類性能が高く，広瀬香美，大黒摩季，松任谷由実らは，分類性能が低いことが示されている．これは，データセット中で，前者は特別な特徴をもち，後者は，特徴が少ないことを示している．分類性能には，テキスト数，延べ語数も影響するものの，aiko などは，両者ともに極端に大きいわけではなく，とりわけ個性的な歌詞の特徴をもつことが示唆される．

5.4.　重要な特徴量

表３は，ランダムフォレストを用いた分類実験全体において，寄与の大きい特徴量上位20変数を示している．「あたし」，「私」などの代名詞，「し」，「だ」，「を」などの助詞・助動詞が多く含まれており，これらの語が，歌手の判別に有用であることを示している．

また，「なんて」，「いい」，「よ」などに見られる通り，日常的な口語体に用いる語が多く含まれており，これは，女性シンガーソングライターの歌詞，あるいは、現代日本の歌謡曲一般の歌詞の特徴を示すものと推察される．

表４は，ランダムフォレストを用いた分類実験において，シンガーソングライター別に寄与の大きい特徴量上位 20 変数を示している．全体的な特徴としては，(a)機能語が多く含まれるものの，一定程度内容語も含まれる，(b)代名詞，とりわけ人称代名詞が多く含まれる，(c)特定の歌手においては，外来語の影響がみられる，(d)特定の歌手において，口語体が多く含まれる，(e)「コト」と「事」など，表記レベルの差異もみられる，以上があげられる．

以下，エラー率が高かった，広瀬香美，松任谷由美，大黒摩季を除く７名の歌手の特徴について，結果の一部を議論する．

表４　個別シンガーソングライターの分類に帰与の大きい特徴量

	aiko		広瀬香美		松任谷由実		中島みゆき		大黒摩季	
	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}
1	あたし	0.0558	Love	0.0124	を	0.0193	し	0.0261	し	0.0058
2	様	0.0379	人	0.0111	愛	0.0092	だ	0.0202	から	0.0043
3	事	0.0227	永遠	0.0073	あたし	0.0076	へ	0.0178	あたし	0.0040
4	程	0.0077	私	0.0072	だ	0.0073	よ	0.0151	よう	0.0039
5	私	0.0065	君	0.0049	じゃ	0.0072	あたし	0.0133	過去	0.0034
6	光	0.0063	二	0.0043	なんて	0.0059	たい	0.0127	に	0.0027
7	目	0.0061	を	0.0036	見	0.0055	たち	0.0116	する	0.0024
8	頬	0.0055	は	0.0035	私	0.0052	に	0.0098	を	0.0023
9	も	0.0051	ちゃ	0.0035	君	0.0050	を	0.0080	1	0.0020
10	想い	0.0049	プレゼント	0.0034	ない	0.0044	私	0.0078	君	0.0018
11	触れ	0.0042	きっと	0.0030	ば	0.0038	た	0.0076	う	0.0017
12	し	0.0039	し	0.0030	いい	0.0035	いい	0.0072	見	0.0016
13	た	0.0038	あたし	0.0028	想い	0.0033	は	0.0068	だって	0.0015
14	唇	0.0037	くれ	0.0027	とき	0.0032	から	0.0063	ん	0.0014
15	愛	0.0032	た	0.0026	こと	0.0032	人	0.0061	また	0.0013
16	...	0.0028	なんて	0.0024	する	0.0031	?	0.0060	が	0.0013
17	あなた	0.0027	ちゃっ	0.0023	時	0.0029	も	0.0057	少し	0.0013
18	いい	0.0025	が	0.0022	へ	0.0028	ん	0.0054	愛	0.0012
19	だって	0.0024	幸せ	0.0021	よ	0.0028	ない	0.0051	ゆく	0.0012
20	よう	0.0024	愛	0.0020	し	0.0026	あなた	0.0050	幸せ	0.0011

	大塚愛		椎名林檎		竹内まりや		宇多田ヒカル		YUI	
	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}	語	VI_{acu}
1	1	0.0439	あたし	0.0113	私	0.0143	君	0.0141	なんて	0.0213
2	幸せ	0.0280	居	0.0106	街	0.0123	I	0.0097	でしょ	0.0165
3	あたし	0.0144	だ	0.0103	し	0.0117	you	0.0091	ん	0.0109
4	2	0.0105	た	0.0103	恋	0.0104	いい	0.0091	?	0.0075
5	私	0.0099	人	0.0086	ふたり	0.0097	あたし	0.0079	の	0.0074
6	も	0.0073	私	0.0078	想い	0.0078	'	0.0069	た	0.0072
7	コト	0.0067	其の	0.0069	だ	0.0074	baby	0.0062	って	0.0065
8	ない	0.0066	いい	0.0062	た	0.0072	どこ	0.0058	私	0.0053
9	を	0.0063	でしょ	0.0052	いい	0.0068	ん	0.0053	ない	0.0050
10	だ	0.0063	なんて	0.0051	を	0.0065	し	0.0047	わかっ	0.0044
11	する	0.0063	様	0.0051	あたし	0.0063	あなた	0.0041	人	0.0040
12	し	0.0062	生命	0.0048	出会っ	0.0061	て	0.0041	出来	0.0039
13	君	0.0059	れ	0.0045	感じ	0.0058	無い	0.0036	だ	0.0037
14	なぁ	0.0055	を	0.0043	'	0.0053	心	0.0033	あたし	0.0037
15	なんて	0.0054	君	0.0043	は	0.0053	愛	0.0033	を	0.0036
16	と	0.0052	たい	0.0040	本当	0.0052	can	0.0030	も	0.0030
17	たい	0.0052	心	0.0040	たい	0.0049	幸せ	0.0028	夜	0.0029
18	愛	0.0050	など	0.0038	ない	0.0046	時	0.0026	じゃ	0.0028
19	見	0.0047	て	0.0037	選ん	0.0044	少し	0.0026	たい	0.0027
20	な	0.0044	よ	0.0036	電話	0.0043	だって	0.0025	あなた	0.0024

- aiko：代名詞「あたし」が特徴的なほか，一文字の漢字が多く含まれている．また，「目」，「頬」，「唇」など身体の一部を多く含んでおり，これは他の歌手には見られない特徴である．
- 中島みゆき：ひらがなが多く含まれている．他の歌手に比べて内容語が少なく，機能語，とりわけ，助詞・助動詞が多く含まれている．
- 大塚愛：「1」や「2」など，英数字が上位に出現する．また，「コト」や「なぁ」など，よりくだけた語が出現する．aiko では，漢字表記の「事」が出現する一方，大塚愛では，カタカナ表記の「コト」が出現している点も注目に値する．漢数字や漢字を使わないことは，歌詞の見た目を柔らかい印象にするものである．
- 椎名林檎：aiko と同様「あたし」が使われるほか，「其の」と漢字表記が使われていることが注目に値する．このような漢字表記は，大塚愛とは対比的であり，歌詞の見た目を堅い印象にするものである．
- 竹内まりや：比較的，内容語が多いことが特徴的である．竹内まりやは中島みゆきとともに，これらの歌手の中では比較的時代が離れており，「街」，「恋」，「ふたり」，「想い」，「電話」などは，70 年代から 80 年代の，とりわけ，男女間のコミュニケーションのあり方をあらわす要素を含むものである．
- 宇多田ヒカル：外来語が多く含まれている．また，「君」という語が分類にもっとも寄与が大きい．他の歌手は二人称を表す「君」や「あなた」などの語よりも「あたし」や「私」といった一人称の語が上位に出現するのに対して，宇多田ヒカルだけが一人称よりも二人称が上位にきている．
- YUI：「なんて」，「でしょ」など，口語的な表現が上位に出現する．また，「?」も上位に含まれており，歌詞の中で，問いかけが多く含まれることが推察される．

6．おわりに

以上，本研究では，過去 30 年にわたる，現代日本の，女性シンガーソングライターの楽曲の歌詞を探索的に分析した．計算文体論の最新の手法を利用することで，それぞれの歌手を区別する特徴を明らかにすることができた．

本研究は，探索的分析を中心とし，特徴語に関しても，ランダムフォレストが返す重要変数を分析するにとどまっている．今後，本研究の結果をもとに，各シンガーソングライターに関する仮説を構築し，また，それぞれの語が実際に利用される文脈を，より詳細に検討することで，より厳密な知見を提出することを目指したい．

また，本研究では，結果が，どの程度，時代背景によるものであり，また，どの程度，歌手の個性や思想によるものかについても，十分な分析を行っていない．この点については，歌手のインタビューや当時の雑誌，新聞を含む，一次資料を分析することで，今後さらに検討していきたい．

謝辞

本研究は，国立情報学研究所公募型共同研究「多種テキストからのコミュニケーション・スタイルの抽出ならびにその分析と応用（代表者：鈴木崇史）」より，一部支援を受けています．ここに記して謝意を表します．

参考文献

[1] Shlomo Argamon, Casey Whitelaw, Paul Chase, Sodhan Raj Hota, Navendu Garg, and Shlomo Levitan. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology,* Vol.58, No.6, pp.802-822, 2007.

[2] Leo Breiman. Random forests. *Machine Learning,* Vol.45, pp.5-23, 2001.

[3] Anthony Kenny. *The Computation of Style: An Introduction to Statistics for Students of Literature and Humanities.* Pergamon Press, Oxford, 1982.

[4] Takafumi Suzuki. Extracting speaker-specific functional expressions from political speeches using random forests in order to investigate speakers' political styles. *Journal of the American Society for Information Science and Technology*, Vol.60, No.8, pp.1596-1606, 2009.

[5] 菊池清麿．日本流行歌変遷史: 歌謡曲の誕生から J・ポップの時代へ．論創社，東京，2008.
[6] 金明哲．R によるデータサイエンス．森北出版，東京，2007.
[7] 金明哲，村上征勝．文章の統計分析とは．甘利俊一，竹内啓，竹村彰通，伊庭幸人（編），言語と心理の統計: ことばと行動の確率モデルによる分析．岩波書店，東京，pp.3-57，2003.
[8] 金明哲，村上征勝．ランダムフォレスト法による文章の書き手の同定．統計数理，Vol.55，No.2，pp.255-268，2007.
[9] 伊藤雅光．表記からみた松任谷由実の歌詞(3): 曲名の字種比率の変遷．日本語学，pp.79-87，1997.
[10] 伊藤雅光．表記からみた松任谷由実の歌詞(4): 歌詞とはなにか．日本語学，pp.78-87，1997.
[11] 伊藤雅光．表記からみた松任谷由美の歌詞(4): 歌詞とはなにか．日本語学，Vol.16，No.3，pp.78-87，1997.
[12] 伊藤雅光．ユーミンの言語学(41): ユーミンソングの基本語彙．日本語学，Vol.20，No.2，pp.74-82，2001.
[13] 伊藤雅光．ポップス系流行歌の語彙調査における外来語と外国語の判定基準．計量国語学，Vol.23，No.2，pp.110-130，2001.
[14] 伊藤雅光．計量言語学入門，大修館書店，東京，2002.
[15] 伊藤雅光．歌謡曲の中の外来語・外国語．日本語学，Vol.22，No.8，pp.40-49，2003.
[16] 見崎鉄．J ポップの日本語: 歌詞論．彩流社，東京，2002.
[17] 水谷静夫．数理言語学．培風館，東京，1982.
[18] 村上征勝．シェークスピアは誰ですか?: 計量文献学の世界．文春新書，2004.
[19] 田家秀樹．読む J-POP: 1945-2004．朝日文庫，東京，2004.
[20] 烏賀陽弘道．J ポップの心象風景．文春新書，東京，2005.
[21] 烏賀陽弘道．J ポップとは何か: 巨大化する音楽産業．岩波新書，東京，2005.
[22] コンフィデンス年鑑（1970-1979）．オリコン・エンタテインメント，1970-1979.
[23] オリコン年鑑（1980-2009）．オリコン・エンタテインメント，1980-2009.