日本語歌謡曲の歌詞における性差と時代変化のテキスト分類を用いた検討

木田菜月　久野雅樹
電気通信大学大学院情報理工学研究科
k2230049@edu.cc.uec.ac.jp, hisano@uec.ac.jp

概要

本研究では，作詞家や歌手の性差による歌詞の違いに注目し，その違いを明らかにするためのテキスト分類を行い，その性能から差異の実態を明らかにした．性差によって言葉の使い方にどのような違いがあるかを見るために文末詞に着目し，使用割合の違いから性差による差異が認められた．この性差による違いを小説の登場人物を対象とした先行研究と比較すると，歌詞と小説には違いが見られ，さらに，作詞家の性差によって時代変化に違いがあるのかを明らかにするため，男女別に年代ごとのテキスト分類を行ったところ，男女それぞれある程度の年代予測力が見られた．

1　はじめに

日本語コーパスを対象とした特徴分析，テキストの分類や生成の研究は，言語資源の整備と自然言語処理技法の発展にともない盛んに行われている．本研究では日本語歌詞を対象に研究を行う．音楽における歌詞には，時代ごとの言葉の使われ方，作詞家や楽曲のジャンルの違いによる構造の違いが大きく反映されていると言える．

歌詞を対象とした計量的な研究の例を挙げると大出らは，1978 年から 2012 年に日本レコード大賞および優秀作品賞を受賞した楽曲の歌詞について計量テキスト分析を行い，ネガティブな内容からポジティブな内容への変化を明らかにした [1]．細谷らは，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 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 曲対して，分類実験を行った．結果としてカーネル主成分分析では，歌手ごとの一定の集中傾向が観察されたが，発行年の影響は明確には観察されなかった [2]．

このように歌詞研究は様々に行われているが，流行歌や数名の歌手の一部の楽曲など特定の楽曲のみを対象としたものがほとんどである．そこで，本研究では大規模な歌詞コーパスである歌詞コンテンツデータ集に含まれる日本語歌詞の楽曲を対象にテキスト分類を行うことで，より全般的な特徴を明らかにすることを目的とし，作詞家や歌手の性差によるテキスト分類と男女別年代ごとのテキスト分類を行った．

2　性差によるテキスト分類

2.1　使用したデータセット

分析対象として，株式会社シンクパワーが提供している歌詞コンテンツデータ集 [3] を利用した．この歌詞コンテンツデータ集には，約 43 万曲分の曲名，アーティスト名，アルバム名，作詞者，作曲者，発売日，歌詞などの情報が収録されている．本研究では，日本語歌詞を対象とするため，このコーパスから歌詞に少なくとも平仮名またはカタカナが含まれる楽曲のみを抽出した．

そして，1985 年から 2020 年に発売されたこのコーパスに含まれる曲数の多い作詞家上位 10 人の楽曲 15722 曲と歌手上位 10 組の楽曲 7116 曲を対象にそれぞれ実験を行った．楽曲を用いた作詞家および歌手は表 1 の通り．

2.2　前処理

歌詞データに含まれる鍵括弧と二重鍵括弧以外の括弧は，ルビであったり歌っている人の名前であったりする場合が多いため，括弧とその中身の除去を行った．また，言葉により着目するため，歌詞データに含まれる改行コードの除去，および歌詞の鍵括弧，二重鍵括弧，句読点の除去を行った．

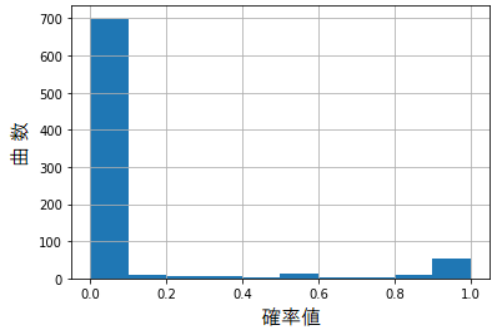

図 1　女性作詞家楽曲の確率値の分布

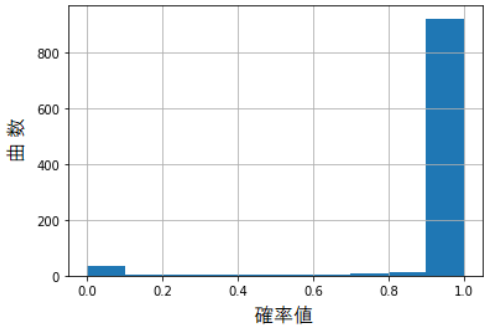

図 2　男性作詞家楽曲の確率値の分布

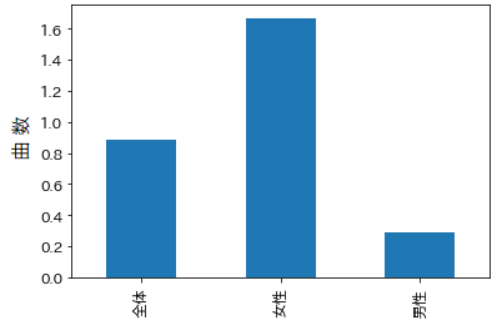

図 3　ですねの使用割合 (作詞家)

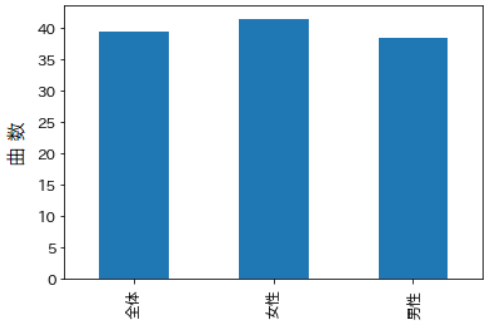

図 4　かなの使用割合 (歌手)

表 1　実験に楽曲を用いた作詞家と歌手

作詞家	歌手
秋元康	浜崎あゆみ
つんく	嵐
松井五郎	槇原敬之
畑亜貴	TUBU
松本隆	松任谷由実
阿久悠	B’z
こだまさおり	DREAMS COME TRUE
吉田美和	SMAP
桑田佳祐	氷川きよし
中島みゆき	ゆず

2.3　実験手順

テキスト分類には GRU モデルを用いた．GRU は長期記憶と短期記憶のバランスも学習でき，LSTM よりもゲートが少なくセルも必要ないため，状態変数の数が同じであれば LSTM よりも少ない計算量・使用空間量で済ませることができるためである [4]．

まず，データセットに女性を 0，男性を 1 としてクラス情報を付加し，学習用が 80 %，テスト用が 20 %になるように分割した．分割したデータセットからボキャブラリを構築，単語の ID 化とパディングを行い GRU に学習させた．そして，テスト用データを用いて学習したモデルの性能評価を行ったのち予測した確率値の分布を出した．最後に小説の登場人物を対象とした研究 [5] と比較するため，文末詞 11 語に関して，全体の楽曲，女性と分類した楽曲，男性と分類した楽曲それぞれの 100 曲あたりに含まれる曲数を出した．同様の手順を歌手のデータセットに対しても行った．歌手の性別に関してはボーカルの性別を歌手の性別としている．

2.4　性能評価と確率値

性能評価には予測結果の適合率と再現率，また，これらから計算される F1 値を用いた．また，予測結果は，それぞれのクラスに対してすべてのクラスの確率値の合計が 1 となるように確率値を出したのち，確率値の値が一番高いクラスを予測結果としている．本実験は 2 値分類であるため，クラス 1 すなわち男性のクラスの確率値が 0 に近ければ近いほど女性的であり，1 に近ければ近いほど男性的な歌詞と言える．

2.5　結果と考察

分類性能は表 2 のようになった．

また，作詞家の性別予測の確率値の分布は図 1，図 2 の通り．さらに，文末詞 11 語それぞれの語が

表 2　それぞれの性能評価

	作詞家	歌手
適合率	0.9206	0.9607
再現率	0.9463	0.9617
F1 値	0.9333	0.9612

100 曲あたりに含まれる曲数を全体の楽曲，女性の楽曲，男性の楽曲で比較した結果の一部を図 3，図 4 に示す．

作詞家の確率値の分布を見てみると，女性作詞家楽曲に対する確率値はほとんどが 0.05 以下，男性作詞家楽曲に対する確率値はほとんどが 0.95 以上であることが分かる．また，歌手の性差による分類でも同様の結果が得られた．このことから，性差による歌詞の違いが明確にあるということが分かる．

作詞家　分類された楽曲の特徴として文末詞に着目したが，男性作詞家が比較的よく使う文末表現は「だもの・なのよ・だよね」であり，女性が比較的よく使う文末表現は「ですか・ますか・ですね」であることが分かった．特に女性作詞家に関しては，「です・ます」の使用が特徴である考えられる．また，「ですか・ますか・ですね」は小説の登場人物を対象とした研究では，中性的または男性的な文末詞という結果が出ている [5]．このことから，小説に見られる性差の特徴と歌詞に見られる性差の特徴には違いがあることが分かった．

歌手　「かしら」，「だわ」や「ですね」，「ますね」が男性に比べて極端に多いところは小説を対象としたものと一致していた．特に「かしら」，「だわ」は非常に女性的な語とされていた [5]．しかし，先行研究では非常に男性的な語とされていた「かな」は男女ともに使われていることが分かった．他にも「だよね」や「だもの」，「なのよ」は逆の傾向が見られた．このことから，作詞家のみならず歌手においても小説に見られる性差の特徴と歌詞に見られる性差の特徴には違いがあることが分かる．

3　男女別年代ごとのテキスト分類

3.1　使用したデータセット

これまでの実験と同様のコーパスに含まれる曲数の多い男性作詞家上位 14 人の楽曲を曲数がある程度均等になるように発売年で 5 区分に区切ったデータセット、および同様にして発売年で 5 区分に区切った曲数の多い女性作詞家上位 16 人の楽曲のデータセットの 2 つのデータセットで実験を行った．楽曲を用いた作詞家は表 3 に示す．また，男性，女性それぞれのデータセットの区分は表 4 の通り．

表 3　楽曲を用いた作詞家

男性作詞家		女性作詞家	
名前	曲数	名前	曲数
秋元康	3166	畑亜貴	1051
つんく	1949	こだまさおり	781
松井五郎	1313	吉田美和	690
松本隆	1050	中島みゆき	611
阿久悠	615	岩里祐穂	561
桑田佳祐	688	藤林聖子	587
森雪之丞	534	三浦徳子	369
売野雅勇	441	松任谷由実	419
奥田民生	578	岡村孝子	280
前田亘輝	638	竹内まりや	389
小田和正	488	YUKI	437
草野正宗	532	吉元由美	308
福山雅治	474	ayu hamasaki	401
さだまさし	426	及川眠子	268
		椎名林檎	361
		AIKO	376

表 4　それぞれの年代区分

区分	男性	女性
0	1985-1995	1985-2000
1	1996-2003	2001-2008
2	2004-2009	2009-2012
3	2010-2013	2013-2015
4	2014-2020	2016-2020

3.2　実験手順

前処理としてこれまでの実験と同様の処理を行った．その後，男性作詞家のデータに区分ごとのクラス情報を付加し，学習用が 80 %，テスト用が 20 % になるように分割した．分割したデータセットからボキャブラリを構築，単語の ID 化とパディングを行い GRU に学習させた．テスト用データを用いて学習したモデルの性能評価を行い，予測した確率値の分布を出した．女性作詞家のデータに対しても同様の手順を行った．

3.3　性能評価

性能評価には予測結果の適合率と再現率から計算される F1 値のマクロ平均とマイクロ平均を用いた．

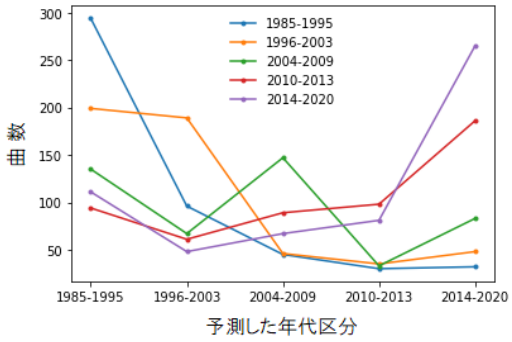

図 5　男性作詞家楽曲の年代区分ごとの予測値の分布

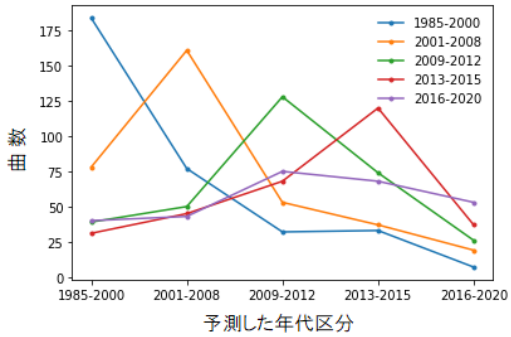

図 6　女性作詞家楽曲の年代区分ごとの予測値の分布

また，予測結果は 2 値分類と同様に，それぞれのクラスに対してすべてのクラスの確率値の合計が 1 となるように確率値を出したのち，確率値の値が一番高いクラスを予測結果としている．

3.4　結果と考察

分類性能は表 5 のようになった．

表 5　それぞれの性能評価

	男性	女性
マクロ平均	0.3722	0.3956
マイクロ平均	0.3850	0.4094

また，それぞれの年代区分の楽曲をどの年代区分の楽曲と予測しているかの予測値の分布は図 5，図 6 の通り．

男性作詞家　分類性能だけを見るとかなり低い．しかし，正解と予測の関係を見てみるとほとんどの年代区分で正しい年代区分またはその近い区分の予測値が高くなっている．例えば，正解が区分 0(1985-1995) と区分 1(1996-2003) のグラフを見ると，区分 0(1985-1995) の楽曲と予測している曲が多い．また，区分 2(2004-2009) では依然として区分 0(1985-1995) も多いが区分 2(2004-2009) の方が多くなっており，区分 3(2010-2013)，区分 4(2014-2020) では区分 4(2014-2020) と予測する曲が圧倒的に多くなっていることが分かる．このことから，区分 0(1985-1995) と区分 1(1996-2003) は近い傾向があり、年を追うごとに変化，区分 3(2010-2013) 以降は，区分 0(1985-1995)，区分 1(1996-2003) と比べて大きく変化していると考えられる．

女性作詞家　分類性能だけを見るとかなり低く，男性作詞家との差もあまりない．しかし，正解と予測の関係を見てみると，ほとんどの年代区分で正しい年代区分の予測値が高くなっている．この結果から，男性作詞家の結果よりも顕著になっていると言える．しかし，区分 4(2016-2020) の予測だけは正しく予測されていない曲が多い．これは，近年の楽曲は過去の楽曲の特徴を反映したものが増えているとも考えられる．しかし，カバー曲や再録曲の影響，また，女性作詞家のデータセットは男性作詞家のものに比べて少ないので，区分ごとの曲の偏りによって見られる特徴の可能性も高いとも言える．そのため，今後その影響について検討する必要がある．

4　おわりに

歌詞には作詞家や歌手の性差が反映されていることが，GRU を用いた性差によるテキスト分類から示された．また，男女別での年代ごとのテキスト分類実験において，歌詞は時代によって変化していることが示された．このように歌詞に反映された性差や時代ごとによる特徴の実態が得られたことで，今後歌詞生成研究などへの応用が考えられる．

しかし，本研究で用いた歌詞コーパスには，カバー曲や再録曲の歌詞も多く含まれていたため，時代変化を見るという点で問題が残っている．今後，このようなデータをどのように扱っていくかが課題である．

参考文献

[1] 大出彩, 松本文子, 金子貴昭. 「流行歌から見る歌詞の年代別変化」. 情報処理学会じんもんこん 2013 論文集 (4), pp. 103–110, 2013.

[2] 細谷舞, 鈴木崇史. 「女性シンガーソングライターの歌詞の探索的分析」. 情報処理学会じんもんこん 2010 論文集 (15), pp. 195–202, 2010.

[3] 株式会社シンクパワー.「歌詞コンテンツデータ集」. 2020.

[4] K. Cho, B. Merrienboer, C. Gulcehere, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. **EMNLP**, pp. 1724–1734, 2014.

[5] 黒須理紗子. 「女ことば・男ことばの研究：差異と変遷」. 日本文學 vol.104, pp. 187–207, 2008.