歌詞における聞き慣れない表現と誤りとの関連性の分析

松本 和幸*　篠山 学**　宮内 弘輔*　吉田 稔*　北 研二*
*徳島大学, **香川高等専門学校
{matumoto, mino, kita}@is.tokushima-u.ac.jp, sasayama@di.kagawa-nct.ac.jp

1　はじめに

ある楽曲のタイトルやアーティスト名を知りたいとき，その検索クエリとして歌詞を用いる場合がある．歌詞はタイトルやアーティスト名よりも情報量が多い分，一言一句を正しく記憶することが難しいため，誤った歌詞を歌詞検索システムに入力してしまうことも多い．

Google などの Web 検索エンジンでは，検索クエリに誤りが含まれる可能性が高い場合に，正しいと思われるクエリによる検索結果を提示する「もしかして」機能が利用できる．この機能により Web 検索エンジンを用いて間違いを含む歌詞により目的の楽曲を探せる場合もあるが，あまり有名でない楽曲の場合は，検索結果として提示されにくいという問題がある．

一般に，耳で聞いて記憶した情報に基づき書き起した歌詞片を用いて楽曲検索をおこなうことが多いことから，ユーザがあまり知らない単語やフレーズを用いて検索することは少ないと思われる．このことから，入力されたクエリに対し，単語やフレーズの認知度を手がかりに，誤りと思われる箇所を特定したり，正しい表現に訂正したりすることが可能ではないかと考える．

本稿では，歌詞において特徴的な表現について，歌詞コーパスと Web コーパスにおける出現頻度を比較することで分析する．また，聞き慣れない表現と誤りが起こるフレーズとの関連を，実際に楽曲の歌詞の書き起こしをおこなうことで得た歌詞片を用いることにより調査し，歌詞に特徴的な表現との比較もおこなう．

2　従来研究

楽曲の歌詞について分析した研究はこれまでに幾つか存在するが，歌詞検索における誤りについて研究したものはあまり無い．篠山ら [1] は，歌詞の間違い方のタイプを調査するため，Yahoo!知恵袋に投稿された，うろ覚えの歌詞を記述して曲名を尋ねている質問と，その回答の収集をおこなった．このデータに基づき，歌詞の誤りを「表記の揺れ」,「単語の欠落」,「単語の挿入」,「単語の置換」,「イメージ」の 5 種類に分類している．これらの誤りの種類ごとに統計をとったところ，表記の揺れが 43.2%ともっとも多く，次いで単語の置換が 38.2%であったと報告している．

表記の揺れも単語の置換も，意味的または音的に似ている語に置き換わってしまうことが多いことから，意味または音が類似する語をクエリに追加することが有効であると考えられる．しかし，クエリ中に含まれる語に類似する語を概念辞書などから得て検索をおこなう場合，誤っていない語の類似語を検索クエリとして追加してしまうと，検索結果が改悪されるという問題が起きる．このことからも，誤り箇所を特定することが重要であることがわかる．

3　書き起こしデータの誤り分析

本節では，歌詞の書き起こしにより取得したデータから，誤った単語の分析をおこなう．歌詞の書き起こしに用いた楽曲（歌詞）数は，合計 40 曲であり，のべ 20 名の被験者に楽曲を実際に聞いてもらった後，テキストに書き起こしてもらうことでデータを取得した．被験者のうち数名は，楽器演奏などの音楽経験者であり，また，ほぼ半数以上が日常的に音楽鑑賞している．

得られたデータのうち，単語の表記揺れや置換に該当する誤りが含まれるフレーズの種類は全部で 80 種類あった．1 つのフレーズに 1 箇所のみの誤りを含むものもあれば，2 箇所以上含まれるものもある．

まず，誤りの出現傾向について見ることにする．誤りがフレーズ内のどの位置に現れるかを見ることで，誤り位置を予測する手がかりとなるかどうかを検討する．

表 1 に，誤りが出現する位置を正解フレーズ内の先頭からの文字数で表した値を，文字数で割った値の頻度分布を示す．この表から，誤りが先頭のほうに含まれる割合が比較的高いことがわかるが，フレーズの長さが全体的に短く（平均約 14 文字），データ数が不十分なため，これだけでは誤り位置の特定は困難である．

つぎに，誤りが含まれる単語について，その読みを分析する．聞き誤りにより，読みが類似する語に置換される可能性が高い．誤った語と元の語とを形態素解析により読みに変換し，その読みの間のレーベンシュタイン距離の分布を求めたところ，図 1 のようになった．

この分布を見ると，含まれる誤りの多くが，元の語との距離が 0 の語，すなわち同音異義語や文字種の表記違いに置換されてしまっている．実際の歌詞における表記と，被験者による書き起こしの表記に食い違いが出たた

表 1: 誤り箇所の位置の分析

位置	頻度
0.0〜0.1	42
0.1〜0.2	10
0.2〜0.3	12
0.3〜0.4	9
0.4〜0.5	15
0.5〜0.6	24
0.6〜0.7	11
0.7〜0.8	12
0.8〜0.9	8

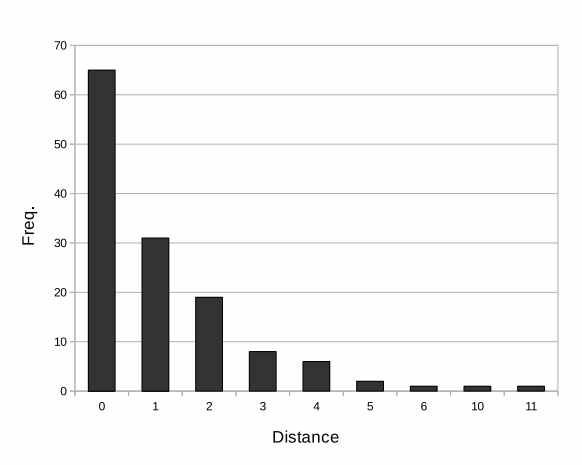

図 1: 誤った語と元の語の読みの距離の分布

めである．このことから，誤りが起きた歌詞に原因があるとは考えにくい．楽曲の聞き取りやすさや，被験者のリスニング力によって違いが出ると考えられる．

4 歌詞に特徴的なフレーズとの比較

本節では，歌詞において頻出し，一般の文章ではあまり用いられないフレーズを歌詞に特徴的な表現とし，誤りが起きたフレーズと，歌詞に特徴的な表現との比較を通して，誤りが起きやすい原因が何であるかを分析する．

歌詞検索システムに入力されたクエリは，ユーザが耳で聞いて認知したものであるため，意味的に理解できないフレーズでは記憶されにくいと考える．また，歌詞は限られた長さで，曲に合わせるため，慣用的な表現もよく用いられる．慣用的な表現は，普段聞き慣れているものが多い．たとえば,「世界が終わる」というフレーズは歌詞においてよく出現するものであるが，聞き慣れた表現といえる．しかし,「終わる」という語を，意味的に似ている「停止する」という語に置き換えた「世界が停止する」というフレーズは，歌詞にはあまり出現せず，通常の話し言葉や書き言葉においても一般的な表現ではない．つまり，歌詞に特徴的な表現でも，一般的な表現と共通するものがあると考えられる．

フレーズが一般的な文章で使用されるかどうかの判定のために，日本語 Web コーパス[1]（以下 n-gram コーパス）を用いる．このコーパスには，Web 上の大規模な文書から一定期間収集した文書から得た文字 n-gram および形態素 n-gram が出現頻度とともに登録されている．

ユーザがうろ覚えの歌詞について検索する際，ただ単に一般の文章で頻出するフレーズではなく，歌詞として聞き慣れたフレーズや単語を検索クエリとして入力する可能性がある．このような，歌詞に特徴的なフレーズを分析するため，篠山らが Web から自動収集した約 10 万曲分の歌詞が登録された歌詞コーパス [1] を用いて，形態素 n-gram の出現頻度の分析をおこない，n-gram コーパスにおける出現頻度との比較をおこなう[2]．

歌詞コーパス中の単語数と n-gram コーパスの単語数の差が大きいため，各 n-gram の n の値ごとに，最大出現頻度の値で出現頻度を割った相対頻度値を用いて，その相対頻度値の差に歌詞コーパスにおける相対頻度をかけた値をスコア（歌詞特徴スコア）として，歌詞と一般文章とで出現頻度に違いがあるフレーズを調べた．式 1 により，歌詞特徴スコアを計算する．式中の $freq_{r,l}(p_i), freq_{r,w}(p_i)$ は，それぞれ，フレーズ（n-gram）p_i の歌詞コーパス，n-gram コーパス中の相対頻度値を表す．

$$Score(p_i) = (freq_{r,l}(p_i) - freq_{r,w}(p_i)) \times freq_{r,l}(p_i) \quad (1)$$

表 2 に，歌詞特徴スコアの上位 20 件の，フレーズのリストを示す．

表 2: 歌詞に特徴的なフレーズ（上位 20 件）

を 見 て い た / 風 に 吹か れ て
離れ て い て も / いつ の 日 か
が 教え て くれ た / どこ に い て も
何 度 も 何 度 / な 気 が し て
いつ の 日 に か / 目 を 閉じ て
何 が あっ て も / 度 も 何 度 も
も い つ まで も / よう な 気 が し
た 気 が し た / て き た ん だ
の せい に し て / 気 に し ない で
誰 も 知ら ない / まで も い つ まで

このリストを見ると，歌詞において頻出する表現は，単語単位で見ると一般的なものが多いことがわかる．しかし，たとえば,「いつの日にか」という表現が日常会話で使用されることは稀であるため，歌詞に特徴的な表現をリストアップできている．

[1] http://s-yata.jp/corpus/nwc2010/ngrams/

[2] 本研究では，形態素 n-gram の出現頻度 500 以上のものを対象とした．

ここで，誤りが含まれるフレーズは，普段聞き慣れない単語が含まれるものであると推測する．普段聞き慣れない単語が含まれることで，書き起こしの際に誤りが生じやすくなると考える．普段聞き慣れないかどうかの指標として，n-gram コーパスにおける出現頻度だけでなく，単語のなじみ度合いを表す単語親密度も考慮することにした．単語親密度は,「日本語の語彙特性」[2] に収録されている，各単語の認知率に基づく，なじみの程度を数値化したものである．

歌詞データベースにおける出現頻度上位 80 件 (phrase-1) および，書き起こしにおいて誤りが起きたフレーズ 80 件 (phrase-2) に対し，それぞれのフレーズごとに形態素 3-gram を抽出し，n-gram コーパスにおける 3-gram の出現頻度を得て平均値を求める．また，各フレーズ内の内容語の単語親密度の平均値を計算した．

フレーズ内の形態素 3-gram の出現頻度の平均値および内容語の単語親密度の平均値の比較結果を，表 3 に示す．また，図 2 と図 3 は，それぞれのフレーズごとに形態素 3-gram の出現頻度の平均値を，値の大きい順にソートしてグラフで表したものである．

表 3: 3-gram 出現頻度の平均値および内容語の単語親密度平均値の比較

	3-gram 出現頻度	単語親密度
phrase-1	31,180,815	6.18
phrase-2	876,205	6.04

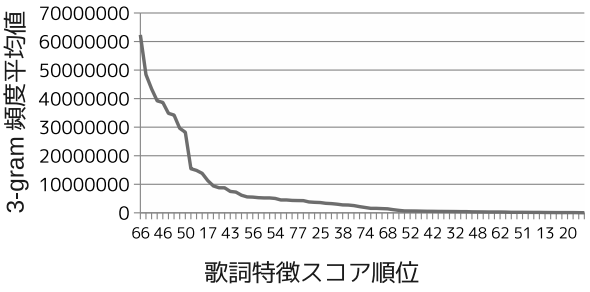

図 2: 歌詞コーパスにおける歌詞特徴スコア上位 80 件についての 3-gram 出現頻度平均値

この結果から，フレーズから抽出した 3-gram の出現頻度の平均値に大きな差があることがわかった．誤りが含まれやすいフレーズは，一般的な文章での使用頻度が少なく，結果として「聞き慣れない」表現となり，書き起こしにおいて誤りが含まれやすくなると考えられる．

極端なものでは，3-gram の出現頻度の平均値が 0 になるフレーズが，80 件中 8 件あった．そのフレーズの一部を，表 4 に示す．

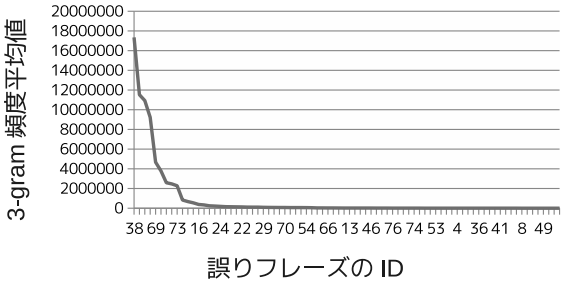

図 3: 誤りフレーズ 80 件についての 3-gram 出現頻度平均値

この例を見ればわかるように，形態素解析により正しく分割できなかったものや,「書（ふみ） よむ」のように特殊な表現や，ひらがなのみで表記された表現が含まれる．ここで,「書（ふみ） よむ」という表現は，唱歌としてよく知られている楽曲「蛍の光」の歌詞に出現するものである．一般文章ではあまり用いられないが，聞き慣れない表現とはいえない．検索対象とする n-gram コーパスの出現頻度のしきい値を 500 からさらに低くした場合には，これらのフレーズも 3-gram の出現頻度平均値が 0 とならない場合もあると考えられるため，しきい値の決定方法を検討する必要がある．

表 4: 3-gram 出現頻度の平均値が 0 となったフレーズ例

もう一度 明日 へ try
あの 香り とともに 花火 が ぱっと 開く
溜め息 一つ 堕ち た 花びら
チクッ っと さ さる トゲ が イタイ
ここ に 未だ 還ら ない
書 よむ 月日 重ね つつ
つ の だせ やり だせ あ たま だせ

一方で，歌詞に特徴的なフレーズの上位 80 件には部分的な重複が含まれることもあり，わずかであるが単語親密度が高くなるという結果が得られた．この結果から，単語単位でのなじみ度合いが，あるフレーズを聞き慣れているか否かに直接関連するわけではないが，よく出てくるフレーズには比較的，単語親密度の高い単語が多く含まれることが分かる．歌詞に特徴的なフレーズで用いられる単語の例を表 5 に示す．この表に示されるような単語は，単語親密度が高く，ユーザが歌詞検索クエリとして入力しやすい語ではないかと考える．

5 おわりに

本稿では，歌詞検索における誤りは，普段聞き慣れない表現の場合に起こるのではないかと考え，楽曲の歌詞

表 5: 歌詞に特徴的なフレーズにおける単語親密度が得られた内容語の例

誰，気，見る，何，君，手，中，人，目，日，夢，教える，止める，言う，前，思う，心，恋，信じる，待つ，吹く，忘れる，声，空，風

の書き起こしにおいて誤りの起きたフレーズについて誤りの出現位置や，誤った語の読みに関して分析をおこなった．その結果，誤りの出現位置には顕著な傾向はみられなかったが，読みに関しては，同音異義語や異表記への間違いが大多数であることがわかった．

また，歌詞コーパスにおける出現頻度と，n-gram コーパスにおける出現頻度に基づいて歌詞に特徴的なフレーズを抽出し，誤りの起きたフレーズとの「聞き慣れている」かどうかの観点から比較をおこなった．この結果，歌詞に特徴的な表現であっても，3-gram の出現頻度の平均値をみると，大きな値が得られたため，普段聞き慣れている単語列が多く用いられていることがわかった．一方で，誤りの起きたフレーズは，出現頻度からみると歌詞に特徴的なフレーズとはいえなかった．また，聞き慣れていない単語も多く含まれていた．

今後は，歌詞検索において起きる誤り箇所の特定の手がかりについて，さらに詳しく分析を進めていきたい．今回は歌詞に特徴的なフレーズかどうかを，単に出現頻度のみで判定したため，低頻度であるが「歌詞らしさ」が出ているような表現を考慮できなかった．こうした表現の抽出には,「歌詞らしさ」を判定する，別の指標を見つける必要がある．また，より多くの，様々な種類の誤りデータが必要なため，書き起こし以外の収集方法を検討したい．

参考文献

[1] 篠山学, 松本和幸. 歌詞検索のための意味情報を用いたクエリの拡張. HCG シンポジウム 2013, HCG2013-B-1-3, Vol. HCG2013, No. B-1-3, pp. 38–42, 2013.

[2] 天野成昭, 近藤公久. NTT データベースシリーズ 日本語の語彙特性. 三省堂, 2008.