歌詞の自動翻訳のための発音を考慮した訳語選択に関する研究

池田昂太郎[†]　松平　茅隼[†]　加藤　大貴[†]　平山　高嗣[††,†]　駒水　孝裕[†]

井手　一郎[†]

† 名古屋大学　〒464-8601 愛知県名古屋市千種区不老町

†† 人間環境大学　〒444-3505 愛知県岡崎市本宿町上三本松 6-2

E-mail: † fikedak, matsuhirac, katohg@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp,

†† t-hirayama@uhe.ac.jp

あらまし　一般的な翻訳とは異なり，歌詞の翻訳では，意味以外にも音数，リズム，発音などの考慮すべき要素がある．これらの要素のうち，発音，特にその類似性については，従来あまり注目されてこなかった．そこで本研究では，自動翻訳における訳語選択アルゴリズムにおいて，音声学の知見に基づいて定義した発音の類似度を導入し，意味と発音のバランスを調整するための重みを設定することで，意味だけでなく発音の類似性も考慮した歌詞の自動翻訳手法を提案する．定量評価実験では，まず歌詞全体における意味と発音の類似度に対する評価指標をそれぞれ定義した．そして，専門家が翻訳した既存の訳詞を基準として，意味と発音の両方の類似性を保つと考えられる適切な重みを決定した．定性評価実験では，その重みを用いて出力された訳詞を人間に提示し，その訳詞が原詞の意味および発音と類似して感じられるかどうか評価した．その結果，両観点において原詞との類似性が確認されたことから，意味と発音の両方の類似性を考慮した訳詞が得られたことが示唆された．

キーワード　自動歌詞翻訳，発音，訳語選択

A Study on Word Selection Considering Pronunciation for Automatic Lyric Translation

Kotaro IKEDA[†], Chihaya MATSUHIRA[†], Hirotaka KATO[†], Takatsugu HIRAYAMA[††,†],

Takahiro KOMAMIZU[†], and Ichiro IDE[†]

† Nagoya University　Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8601 Japan

†† University of Human Environments　6-2 Kami Sanbonmatsu, Motojuku-cho, Okazaki, Aichi, 444-3505 Japan

E-mail: † fikedak, matsuhirac, katohg@cs.is.i.nagoya-u.ac.jp, taka-coma@acm.org, ide@i.nagoya-u.ac.jp,

†† t-hirayama@uhe.ac.jp

Abstract　Unlike general translation, lyric translation requires considering various factors other than meaning, such as number of sounds, rhythm, and pronunciation. Among these factors, taking pronunciation into account for automatic lyric translation is underexplored. Therefore, this study proposes an automatic lyric translation method that considers both the meaning and pronunciation similarities. For this, we introduce the pronunciation similarity defined based on phonetics into the word selection algorithm along with a weight to balance the importance between the meaning and pronunciation. In the quantitative evaluation, we first defined metrics for each of the meaning and pronunciation similarities between the entire lyrics before and after translation. Next, we explored a suitable weight that would guarantee both similarities. A qualitative evaluation assessed the agreement between the translated lyrics generated under this weight and human perception with regard to the meaning and pronunciation similarities. The result indicated that the translated lyric retains both the meaning and pronunciation properties of the original lyric.

Key words　Automatic lyric translation, pronunciation, word selection

語 $w^t_{i,j}$ から $n_{i,j}$ 個の母音 $v^t_{i,j,k}$ からなる母音列 $a^t_{i,j}$ を得られるとする．すなわち，$a^t_{i,j} = (v^t_{i,j,1}, v^t_{i,j,2}, ..., v^t_{i,j,k}, ..., v^t_{i,j,n_{i,j}})$ である．同様に，原詞のフレーズ p^s_i は m_i 個の母音列 $a^s_{i,j}$ からなり，$a^s_{i,j}$ は $n_{i,j}$ 個の母音を含む．すなわち $a^s_{i,j} = (v^s_{i,j,1}, v^s_{i,j,2}, ..., v^s_{i,j,k}, ..., v^s_{i,j,n_{i,j}})$ である．翻訳の際には，まずフレーズ p^s_i を訳語選択アルゴリズムに入力として与える．従来の Beam Search では，j 回目の訳語選択の際に，原詞のフレーズの意味および $j-1$ 回目までに生成した訳語 $w^t_{i,1}, w^t_{i,2}, ..., w^t_{i,j-1}$ に応じて翻訳モデル M の語彙の各訳語候補に対して，翻訳モデル M が意味に関する条件付き尤度 $M(w^t_{i,j} \mid w^t_{i,1}, ..., w^t_{i,j-1}, p^s_i)$ を推定する．ここで，翻訳モデル M は共起に基づいて 2 言語間の各単語間の関係性を学習していることから，本報告ではこれを意味に関する尤度とみなす．提案手法では，この意味に関する尤度に対して，さらに発音に関する尤度 $\phi_a(a^s_{i,j}, a^t_{i,j})$ を加算する．

このようにして求められた尤度を Softmax 関数により確率化し，Beam Search による訳語選択に用いる．すなわち，$w^t_{i,j}$ に対する Beam Search の枝刈りに用いる確率は，

$$\mathrm{Softmax}\Big(M(w^t_{i,j} \mid w^t_{i,1}, ..., w^t_{i,j-1}, p^s_i) + \theta\phi_a(a^s_{i,j}, a^t_{i,j})\Big) \quad (4)$$

である．ここで，θ は訳語選択の際に母音の類似度を考慮する度合いを指定する重みである．

原詞のフレーズ p^s_i を入力とした訳語選択アルゴリズムにおいて，j 回目までに選択された語 $w^t_{i,1}, ..., w^t_{i,j}$ が含む母音の個数の総和が，p^s_i が含む母音の個数と等しくなるまで訳語の生成を繰り返すことで訳詞のフレーズ p^t_i を生成する．

3. 評価実験

提案手法では発音に関する尤度に対する重み θ を設定しているが，これが小さすぎると発音を考慮しない音数を合わせただけの翻訳に，大きすぎると意味を無視した翻訳になってしまう．そのため，発音の類似性を考慮した翻訳として適切な値を決定する必要がある．その上で，実際に翻訳された訳詞に関して，原詞と発音が類似した翻訳が行なわれていることおよび原詞の意味が大きく損なわれていないことを確認する必要がある．

3.1 実験設定

本項では，2. で提案した手法に対して，具体的な実験設定を与える．提案手法は原理上，任意の言語間において適用可能であるが，本実験では原詞の言語を英語，訳詞の言語を日本語とする．

英語で記述された原詞のフレーズを IPA へ変換する際には，Python パッケージ eng-to-ipa(注1)を用いて，そのフレーズに含まれる英単語をまずそれぞれ IPA へ変換する．その後，それらをさらに音節ごとに分割することで母音列を得る．一方，日本語の訳語候補 $w^t_{i,j}$ を IPA に変換する際は，まず日本語形態素解析エンジン MeCab（mecab-ipadic-neologd）[14] を用いて読み仮名を付与し，その後あらかじめ人手で作成した対応表に基づいて IPA への変換を行なう．

本実験では訳語候補を生成し，意味に関する尤度を計算するための翻訳モデルとして，ニューラルネットワークによる機械翻訳のフレームワークである MarianNMT [15] を用いた翻訳モデル staka/fugumt-en-ja(注2)を利用する．また，生成される訳詞の多様性を高めるため，本報告で改良する訳語選択アルゴリズムの実装として Diverse Beam Search [16] を用いる．提案手法では，選択された訳語からなる訳詞のフレーズの音数の合計が原詞のフレーズの音数と等しくなった状態で，次の訳語候補として音数が 0 であるもの（例えば句読点などの記号）が上位に現れるものをより優先的に生成する処理を行なう．これは，音数が 0 である訳語候補の尤度が高くなるときが文の意味の切れ目であるという仮定に基づいて，音数が一致した時点で文として意味が切れているフレーズが生成されやすくするねらいがある．

本実験では原詞として，口の動きの類似性，すなわち母音の調音の仕組みの類似性を考慮して翻訳された訳詞が既に存在する Disney 映画 “Frozen” の主題歌 “Let It Go”（以降では原曲と表記する）の一部を用いた． 2. で定義した発音の類似度を用いてこの原詞 L^s と訳詞 L^t 全体にわたって母音の相違度および類似度を計算するためには，原詞と訳詞に含まれる母音の個数が一致していなければならない．翻訳の都合上，原詞と訳詞に含まれるフレーズは同数であるとする．原詞および訳詞がそれぞれ l 個のフレーズ p^s_i，p^t_i から構成されるとき，$L^s = (p^s_1, p^s_2, ..., p^s_i, ..., p^s_l)$，$L^t = (p^t_1, p^t_2, ..., p^t_i, ..., p^t_l)$ と表す．提案する訳語選択では原詞のフレーズ p^s_i を入力として訳語選択を繰り返し，p^s_i と同じ個数の母音を含む訳詞のフレーズ p^t_i を生成することで，原詞と訳詞に含まれている母音の個数を一致させる．本報告における翻訳システムは，この訳語選択アルゴリズムを原詞 L^s の全てのフレーズに対して適用することで，原詞全体に対する翻訳，すなわち訳詞 L^t を得る．

このようにして得られる原詞と訳詞全体における発音と意味の類似度に関する評価指標を以下で定義する． 2. で定義したように，原詞および訳詞のフレーズ p^s_i，p^t_i はともに m_i 個の母音列 $a^s_{i,j}$，$a^t_{i,j}$ からなり，$a^s_{i,j}$，$a^t_{i,j}$ はともに $n_{i,j}$ 個の母音をもつ．すなわち，原詞および訳詞のフレーズが含む母音の個数を $|p^s_i|$，$|p^t_i|$ とすると，$|p^s_i| = |p^t_i| = \sum_{j=1}^{m_i} n_{i,j}$ である．まず， 2. で定義した母音間の Euclidean 距離を用いて，フレーズ p^s_i，p^t_i の間の発音の相違度を

$$d_p(p^s_i, p^t_i) = \frac{1}{|p^s_i|} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{i,j}} d_v(v^s_{i,j,k}, v^t_{i,j,k}) \quad (5)$$

と定義すると，$d_L(L^s, L^t)$ は $[0, d_{\max}]$ の範囲内にある．次に，

$$\phi_p(p^s_i, p^t_i) = 1 - \frac{d_p(p^s_i, p^t_i)}{d_{\max}} \quad (6)$$

と定義し，これを用いて歌詞間の発音の類似度を

$$\phi_L(L^s, L^t) = \frac{1}{l} \sum_{i=1}^{l} \phi_p(p^s_i, p^t_i) \quad (7)$$

(注1)：mphilli, Mitchellpkt, CanadianCommander, and timvancann, “eng-to-ipa”, https://pypi.org/project/eng-to-ipa/（2024/2/19 閲覧）

(注2)：Satoshi Takahashi, “staka/fugumt-en-ja”, https://huggingface.co/staka/fugumt-en-ja/（2024/2/19 閲覧）

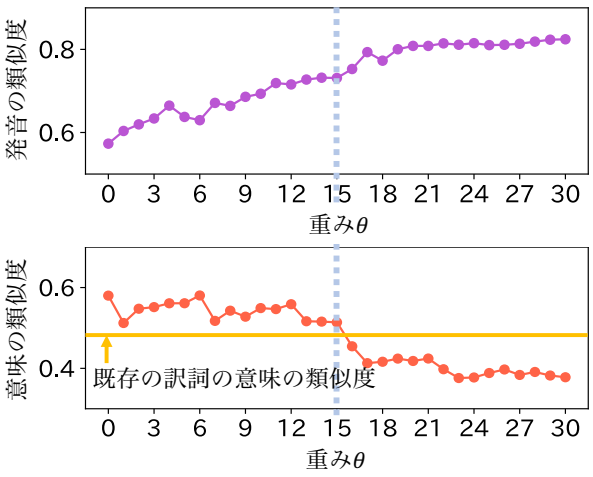

図 4: 重み θ に応じた発音および意味の類似度の変化.

と定義する．ϕ がとりうる値の範囲は [0,1] であり，1 に近いほど原詞と訳詞の発音の類似性が高く，0 に近いほど低いといえる．

また，フレーズ間の意味の類似度を

$$\sigma_p(p_i^s, p_i^t) = s\Big(\mathrm{emb}(p_i^s), \mathrm{emb}(p_i^t)\Big) \tag{8}$$

と定義し，歌詞間の意味の類似度を

$$\sigma_L(L^s, L^t) = \frac{1}{l}\sum_{i=1}^{l} \sigma_p(p_i^s, p_i^t) \tag{9}$$

と定義する．ここで，emb(·) は文を埋め込みへ変換する関数，s は 2 つのベクトル間の余弦類似度である．本実験では emb(·) として Sentence Transformer モデルの 1 つである paraphrase-multilingual-mpnet-base-v2 [17] を用いる．σ がとりうる範囲は [0,1] であり，1 に近いほど原詞と訳詞の意味の類似性が高く，0 に近いほど低いといえる．

3.2 定量評価実験

本項ではまず，提案手法に対して，意味と発音の類似性を考慮した翻訳として適切な重みを決定する定量評価実験を行なう．具体的には，3.1 項で定義した評価指標 $\phi_L(L^s, L^t)$ および $\sigma_L(L^s, L^t)$ を用いて，提案手法により出力された翻訳結果を評価する．

原詞と訳詞の間における発音の類似性と意味の類似性は，意味を重視しすぎると発音を考慮しない音数を合わせただけの翻訳に，発音を重視しすぎると意味を無視した翻訳になるというトレードオフの関係にあると考えられる．そのため，式 (4) において，意味および発音の両方の類似性を保った翻訳結果を出力するパラメータとして適切な重み θ を見つける必要がある．そのために，本実験では原詞 “Let It Go” に対する既存の訳詞「レット・イット・ゴー〜ありのままで〜」 [18] について，意味を最低限担保したうえで限界まで発音を近づけた訳詞とみなす．そして，この訳詞と原詞との間の意味の類似度を下回らない最大の θ を適切な重みとして選択し，その値を用いて生成した訳詞を提案手法の出力とする．

重み θ の変化に対する発音および意味の類似度の変化を図 4 に示す．原詞 L^s と既存の訳詞 L_e^t の間の意味の類似度 $\sigma_L(L^s, L_e^t)$ は 0.473 となった．θ を 1 ずつ大きくした結果，θ が 16 のとき初めて提案手法による訳詞の意味の類似度が $\sigma_L(L^s, L_e^t)$ を下回った．このことから，本報告では $\theta = 15$ のときの訳詞を意味と発音をともに考慮した適切な訳詞 L_p^t と定める．この際の原詞と訳詞の間の発音の類似度 $\phi_L(L^s, L_p^t)$ は 0.731，意味の類似度 $\sigma_L(L^s, L_p^t)$ は 0.514 であった．

また，理論上の発音の類似度の上限かつ意味の類似度の下限であると考えられる，原詞の発音を日本語に無理矢理当てはめた訳詞 L_∞^s を生成する．具体的には，原詞の各母音に対して発音平面上で最も Euclidean 距離が小さい訳詞の母音を選択することで L_∞^t を作成する．ただし，本実験では意図せず意味があるフレーズが生成される可能性を極力小さくするために，子音についても原詞に近いと思われるものを人手で選択して作成する．このときの原詞と訳詞の間の発音の類似度 $\phi_L(L^s, L_\infty^t)$ は 0.855，意味の類似度 $\sigma_L(L^s, L_\infty^t)$ は 0.379 となった．重み θ を大きくすると，発音および意味の類似度がそれぞれ上限および下限に漸近していく様子が見られた．この上限および下限は実験によらずに設定した理論上の限界であることから，提案手法が発音の類似度を考慮する手法として妥当であることが示唆される．

また，既存の訳詞 L_e^t の発音の類似度 $\phi_L(L^s, L_e^t)$ は 0.600 であり，$\phi_L(L^s, L_p^t) > \phi_L(L^s, L_e^t)$ がなりたつため，既存の訳詞が発音を類似させることを直接の目的として翻訳されたものではないとはいえ，提案手法の妥当性が裏付けられたと言える．

3.3 定性評価実験

次に，意味と発音のそれぞれにおいて，発音を考慮しない訳詞 L_0^t と考慮した訳詞 L_p^t のどちらが人間にとって原詞 L^s に類似していると感じられるか評価する定性評価実験を行なう．ここで，訳詞 L_0^s は，提案手法において $\theta = 0$ とした際に出力される，発音の類似度を全く考慮しない，音数を合わせただけの訳詞である．

意味と発音のそれぞれについてアンケートを作成し，発音の類似性に関する実験の後に，意味の類似性に関する実験を行なう．発音の類似性は歌唱音声のみ，意味の類似性はテキストのみによって判断させる．質問の形式として，設問ごとに原詞と 2 つの訳詞 A，B による歌唱音声および原詞と 2 つの訳詞 A，B のテキストを順に提示し，原詞と 2 つの訳詞との類似性を「A の方が近い」・「A の方がやや近い」・「差がない」・「B の方がやや近い」・「B の方が近い」の 5 段階の中から選択させる．また，実験協力者が何度も繰り返し聴き直せるように，HTML 形式で歌唱データを提示する．この際，データの提示順に関するバイアスを排除するため，A と B のどちらを L_0^t とし他方を L_p^t とするかを，設問および実験協力者ごとに無作為に入れ替える．実験協力者の英語および日本語能力の差異による影響を極力排除するため，日本語を母語とする大学生 8 名を対象者とする．なお，実験状況を揃えるために，イヤホンの使用を指示する．

発音の類似度に関する実験では，実際の歌唱音声データを実験協力者に提示し発音の類似性を評価させる．本実験で用いる原詞の歌唱データとして原曲を，また訳詞の歌唱音声データと

して，音声合成ソフトウェア VoiSona(注3)を利用して生成された歌唱音声データを使用する．この際，聴き取りの妨げにならないよう，原曲に対して Vocal Remover and Isolation(注4)による音源分離を施し，歌唱音声のみを抽出したものを提示する．訳詞の歌唱データは，原曲の歌唱音声の旋律を筆者が採譜して MIDI 形式にして VoiSona に取り込み，訳詞を入力することで作成する．

実験の実施後，L_0^t の方が原詞に近いと評価した回答を −2，L_0^t の方が原詞にやや近いと評価した回答を −1，差がないと評価した回答を 0，L_p^t の方が原詞にやや近いと評価した回答を 1，L_p^t の方が原詞に近いと評価した回答を 2 として，全ての実験協力者および全ての設問に対して発音および意味の類似性に関する評価値の平均をとった．その結果，発音の平均は 0.17（= 13/80），意味の平均は 0.04（= 3/80）となり，いずれも正の値となった．このことから，適切な重みでの提案手法による翻訳結果が，実際に人間にとって原詞の発音と近く感じられ，なおかつ意味が大きく損なわれていないと感じられたことが示唆される．

4. ま と め

本研究では，原詞と訳詞の間の発音の類似性を考慮することにより，歌詞をより良く翻訳することを目的とした．本報告では，そのために発音のうち特に母音に着目して，原詞と訳詞の間の発音の類似度を定義し，それを用いた訳語選択アルゴリズムを提案した．原詞と訳詞の間の発音の類似度に関しては，母音の調音の仕組みを表す IPA-Chart [11] を Euclidean 平面上で再現し，この平面における各母音間の Euclidean 距離を用いて類似度を表現した．訳語選択に関しては，既存の翻訳システムにおいて一般的に用いられる訳語選択手法の 1 つである Beam Search [13] の尤度計算に発音の類似度を統合することで，発音の類似性を考慮した訳語選択を実現した．原詞のフレーズに対してこの訳語選択アルゴリズムを繰り返し適用することで，発音が類似し，かつ音数が合う「歌える」訳詞に翻訳できるようにした．また，発音の類似度に対する重みをパラメータとして設定し，意味と発音のバランスを調整できるようにした．評価実験では，まず原詞と訳詞の間の意味と発音の類似度に関する評価指標をそれぞれ定義した．次に，その評価指標を用いて，既存の訳詞に基づいて意味および発音の両方の類似性を考慮した翻訳として適切な重みを決定した．そして，人間の主観による評価実験を行ない，実際に人間にとって原詞の発音と類似して感じられることを，また原詞がもつ意味が大きく損なわれていないことを確認した．

今後の課題として，子音の類似性の考慮，発音の旋律中の位置に応じた重みづけ，実際の歌詞データによる翻訳モデルの Fine-tuning，音声や映像を用いたマルチモーダルな翻訳などが考えられる．

謝辞　本報告の内容について御助言を頂きました名古屋大学博士前期課程 宮川由衣さんに深く感謝致します．

（注3）：（株）テクノスピーチ, “Voisona”, https://voisona.com/（2024/2/19 閲覧）

（注4）：Vocal Remover and Isolation, https://vocalremover.org/ja/（2024/2/19 閲覧）

文　献

[1] Peter Low, “Singable translations of songs,” Perspectives: Studies in Translation Theory and Practice, vol.11, no.2, pp.87–103, Jan. 2003.

[2] Kyron Basu, “Phonetic Journey: Sound in singable translations,” Master’s thesis, University of Victoria, Greater Victoria, BC, Canada, Aug. 2020.

[3] Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang, “Songs across borders: Singable and controllable neural lyric translation,” Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol.1, pp.447–467, Toronto, ON, Canada, July 2023.

[4] Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Tagyoung Chung, Jing Huang, and Nanyun Peng, “Unsupervised melody-to-lyrics generation,” Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol.1, pp.9235–9254, Toronto, ON, Canada, July 2023.

[5] 山本金雄，“歌唱における日本語の発音に関する一考察，” 千葉大学教育学部研究紀要 第 2 部，vol.22，pp.233–249，Aug. 1973．

[6] 金田一晴彦，日本語 新版（上），岩波書店，東京，1988．

[7] 斎藤純男，日本語音声学入門［改訂版］，三省堂，東京，2005．

[8] 木石岳，歌詞のサウンドテクスチャー うたをめぐる音声詞学論考，白水社，東京，2023．

[9] Kristen Anderson-Lopez, “Let It Go,” 2013.

[10] Mónica Martín Castaño, “Translating Disney songs from The Little Mermaid (1989) to Tarzan (1999): An analysis of translation strategies used to dub and subtitle songs into Spanish,” PhD thesis, University of Warwick, Conventry, England, UK, Jan. 2017.

[11] International Phonetic Association, Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet, Cambridge University Press, Cambridge, England, UK, 1999.

[12] Benjamin Bay, Paul M Bodily, and Dan Ventura, “Text transformation via constraints and word embedding,” Proceedings of the 8th International Conference on Computational Creativity, pp.49–56, June 2017.

[13] Raj Reddy, “Speech understanding systems: Summary of results of the five-year research effort at Carnegie-Mellon University,” Technical report, Carnegie-Mellon University, Department of Computer Science, Pittsburgh, PA, USA, Aug. 1977.

[14] 佐藤敏紀，橋本泰一，奥村学，“単語分かち書き辞書 mecab-ipadic-NEologd の実装と情報検索における効果的な使用方法の検討，” 言語処理学会第 23 回年次大会，no.NLP2017-B6-1，Mar. 2017.

[15] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch, “Marian: Fast neural machine translation in C++,” Proceedings of ACL 2018, System Demonstrations, pp.116–121, Melbourne, VIC, Australia, July 2018.

[16] Ashwin Vijayakumar, Michael Cogswell, Ramprasaath Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra, “Diverse beam search for improved description of complex scenes,” Proceedings of the AAAI Conference on Artificial Intelligence, vol.32, no.1, pp.7371–7379, April 2018.

[17] Nils Reimers and Iryna Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp.3982–3992, Nov. 2019.

[18] 高橋知伽江，“レット・イット・ゴー〜ありのままで〜，” 2013．