# Computational stylistic analysis of popular songs of Japanese female singer-songwriters

anonymous
anonymous

## Abstract

This study analyzes popular songs composed by Japanese female singer-songwriters. Popular songs are a good representation of modern culture and society. Songs by female singer-songwriters account for a large portion of the current Japanese hit charts and particularly play an important role in understanding the Japanese language and communication style. In this study, we applied new methods of computational stylistics to the lyrics of the songs. The results clearly show differences in the characteristics of 10 female singer-songwriters, and we found that the 'visualization of the lyrics' is a typical characteristics of current singer-songwriters. Our findings provide an important case study for computational stylistics and can also be useful for understanding Japanese cultural trends.

## Introduction

When listening to music people often think "that sounds like so-and-so' s song" or "that song sounds like so-and-so." What is it that triggers people to make these judgments? Maybe it is the chord progression, the type of instrument accompaniment, the rhythm, or perhaps elements that cannot be specifically explained, such as the impression that the music makes, all of these can have an effect. However, of the factors that greatly influence these judgments, the lyrics have the most significant bearing. There are many factors associated with lyrics that influence these judgments, such as unique expressions and the visual effect of the lyrics on a lyric sheet as well as the repetition of certain words. Of course, it is said that a song can be attributed to an artist because certain lyrics are applied to a certain chord progression, but if these are then separated, there will be cases when it is impossible to speculate whose song it is. However, in many cases, which artist a song belongs to can be deduced just by reading the lyrics. If the individual characteristics of an artist really are clearly expressed in the lyrics and people do not notice, then what are these characteristics? Moreover, what specific factors trigger the intuition people feel which makes them say "this sounds like so-and-so?"

If we turn our eyes to the situation in modern Japan, there are in fact not only communities on social networks relating to "singers" and "music", but also communities relating to the lyrics of certain singers (for example, consider Mixi[1]). On one of Japan's foremost music information sites, Barks Global Music Explorer[2], a feature article called "Ranking of artists with heart-stopping lyrics" is regularly published. Whether it is online or offline communities, lyrics are a subject often touched on by the young generation. The way people communicate is changing owing to the development of communication equipment such as cell phones and PCs, and the resulting spread of systems such as Mixi and Twitter. People who are complete strangers and who cannot even see each other's faces can now communicate just because they have something in common. Many people have a shared knowledge of popular song lyrics and are also able to relate to them. It is thought that when people are receptive to popular music, one of the most important factors is the lyrics. Popular lyrics and lyrics people relate with are a useful medium for analyzing not only the communication between singers and fans, but also between fans and other fans or, in other words, the way in which ordinary people communicate with each other daily.

Also in a sociological theory, the lyrics of popular songs are said to be a good representation of the tastes and linguistic sensitivity of the people of that generation. According to Mita [1], popular songs

[1] mixi.jp
[2] www.barks.jp

are not just things we listen to, but also things we hum along and sing along to. In particular, karaoke has established as a popular pastime in Japan, especially for the younger generation, and popular songs reflect people's feelings, fashions, and their manner of communication. Many popular arts, such as novels, movies, and television programs, are used as factors for analyzing culture and society. However, people connect with these passively, whereas, even though they are passively imparted, popular songs are a medium in which people can actively participate by humming and singing [1]. An individual's writing, such as notes, diaries, and journals, are also used to analyze a person's lifestyle, but these are documents that record the specific independent opinions of an individual person and unlike popular songs, which are on trend; in other words, they are songs that are accepted by the majority of the public and contain lyrics that are also accepted [1]. In this regard, songs represent the tastes and linguistic sensitivity of the people of that generation, and are a more appropriate medium for observing the macroscopic shifts and changes in Japanese culture.

Singer-songwriters are singers that compose their own lyrics and music and they form one of the major genres in the hit charts in modern Japan [2]. They have a long tradition in Japanese music, but they began to flourish in the 1970s in Japan as a small genre, and since then, female singer-songwriters have increased the genre's importance [3]. The message of their lyrics is very powerful and their impact is significant, thus singer-songwriters are an extremely interesting subject matter for lyrical analysis, in particular lyrical analysis using computational stylistics methods as conducted in this study, because they are important for investigating the way in which people in modern Japan communicate.

There have been some studies that tried to investigate the characteristics of language and communication styles from popular songs, although their methods mainly involved qualitative analyses of source documents or interviews with artists and eminent personalities in these fields. Some studies stressed the importance of investigating the lyrical characteristics of popular songs, but in fact they focused only on specific songs or singers for their analyses [4]. Turning our eyes to the field of computational stylistics, the current scope of research has expanded considerably with an increasing variety of available data and the development of natural language processing techniques. In addition to the conventional applications such as authorship attribution and genre-based text classification, computational stylistics methods are used for various new applications such as authorship profiling, computational sociolinguistics, and plagiarism detection [5–7]. These methods must be very useful for analyzing the textual characteristics of lyrics.

Considering this backdrop, this study uses computational stylistics methods to analyze popular Japanese songs composed by Japanese female singer-songwriters over the past 30 years. From the viewpoint of computational stylistics, the purpose of this study has the following four characteristics: (a) the number of tokens is considerably small, (b) several factors (e.g., authors and eras) affect the textual characteristics,(c) the content as well as the style can affect the textual characteristics, and (d) we wanted to obtain linguistically and sociologically meaningful findings instead of just enhancing the classification method. We deliberately selected the features and methods appropriate for our purposes to obtain an important case study for computational stylistics. We also attempted to provide useful knowledge for understanding current Japanese language and communication styles.

# Data and Methods

## Data

We selected 116 songs composed by 10 singer-songwriters (those who wrote more than 90 % of their songs) that appeared on the Oricon annual top 100 single hit chart [8,9] more than five times in the past 30 years. We collected the text data from Uta Net (www.uta-net.com) and Uta Map (www.utamap.com), which are databases of Japanese popular songs, and excluded the titles, spaces, and explanatory phrases. We applied morphological analysis by MeCab (mecab.sourceforge.net), a Japanese morphological analysis system. We calculated the relative frequencies for each text file, and constructed a text-feature matrix.

Previous authorship attribution studies showed that function words or character based n-grams provide a more robust classification performance [10]; however, we selected bag-of-words as features due to reason (d), as mentioned in Introduction, that states that our purpose is to obtain linguistically and sociologically meaningful findings instead of enhancing the classification performance. We used content words as well as function words because of the aforementioned reason (a) the number of tokens is considerably small and (c) the content as well as the style can affect the textual characteristics. It is possible to analyze more specific aspects with analyses that use only function words and content words. We intend to study these subjects in the future.

## Statistical methods

When we considered that several factors affect the textual characteristics, we first applied the kernel PCA to the matrix in order to examine the factors affecting the lyrical characteristics of the songs. We selected the Gaussian kernel and parameter sigma = 0.1, which gave the best possible results for interpretation after attempting several kernels and parameters. Previous studies used a principal component analysis, a correspondent analysis, or a factor analysis for these tasks; however, the kernel PCA sets up kernels and parameters more flexibly; thus, it is more suitable for use in preliminary and exploratory research, such as this paper.

Second, we applied random forests machine learning methods [11] in order to perform the classification experiments for the 10 selected singers. This method has proved the best performance for authorship attribution in Japanese [12], and has also been the best method for similar types of feature extraction tasks [6].

We analyzed the results of the classification to find the singers who had special individual lyrical characteristics (high classification performance) and those who had common lyrical characteristics (low classification performance). In addition, we extracted the important features for classification in order to find the special distinguishing lyrical characteristic of every singer-songwriter.

Random forests are an improved method of bagging [13], which is an ensemble learning method. The purpose of ensemble learning is to improve the classification performance of previous statistical methods, such as decision trees, by repeatedly performing the experiments and calculating the mean or majority votes of the results. However, the results will always be identical when the same data is used to perform these experiments; thus, ensemble learning methods including bagging usually use bootstrap samplings from the original data to repeat the experiments. The major advantage of using random forests instead of bagging is the extraction of a random subset from each bootstrapping sample, which enlarges the variances in bootstrapping samples [11, 14].

We first sampled from $i$ cases at random from the original text-feature matrix $M_{i,j}$ with replacements to make a bootstrap sample, and we extracted random subsets of $\sqrt{j}$ variables from a bootstrap sample to make a sample for constructing an unpruned decision tree. To split the nodes, we used the Gini index formalized as follows:

$$Q_\tau(T) = \sum_{k=1}^{K} p_{\tau k}(1 - p_{\tau k}),$$

where $p_{\tau k}$ represents the proportion of data points in region $R_\tau$ assigned to class $k$ ($k = 1, \ldots, K$), which vanishes for $p_{\tau k} = 0$ and $p_{\tau k} = 1$ and has a maximum at $p_{\tau k} = 0.5$ [15]. These sampling, extraction, and tree-constructing processes were repeated 1000 times, and we constructed a new classifier by a majority vote of the set of trees. Two-thirds of the extracted bootstrap samples were used for constructing the model and the other one-third were left out for testing the model.[3]

An important characteristic of random forests is that it returns variable importance ($VI_{acu}$) for the classification experiments. To calculate variable importance, we first put down the out-of-bag cases and

[3]This is called as Out-of-bag tests.

counted the number of votes cast for the correct class, and then randomly permuted the values of variable $\sqrt{j}$ in the out-of-bag cases, and put these cases down the tree. We subtracted the number of votes for the correct class in the variable-$\sqrt{j}$-permuted out-of-bag data from the number of votes for the correct class in the original untouched out-of-bag data. We calculated the average of this number over all trees in the forest, and that number was the raw importance score for variable $\sqrt{j}$. Finally, we divided the raw score by the standard error of the variable in the calculation over all trees, and that number was $VI_{acu}$ for variable $m$ [11, 16]. $VI_{acu}$ represents the importance of variables for classification. This method is advantageous because this study aims to compare the contributions of the morphemes rather than achieving the best performance. This method calculates important variables directly contributing to the classification in the experiment; thus, it is most suitable for this study.

# Results and discussion

# Basic data for selected singer-songwriters

Table 1 shows the year the selected female singer-songwriters made their debuts and their ages at those times. The debut years and debut ages were taken from the official websites of the respective singer-songwriters. First, we found that the debut age ranged from the late teens to early 20s. The average debut age is 20.7. In general, because of the high profile of Hikaru Utada, the impression is that the debut age has fallen in recent years. However, these results indicate that the debut age has thus far barely changed.

Next, if we look at the data by decade, there were three singer-songwriters in the 1970s who regularly had hit songs, in the 1990s there were five, and in the 2000s there were two. In contrast, there were no singer-songwriters in the 80s that matched our selection criteria. In the 1980s, instead of singer-songwriters, female singers referred to as idols frequented the popular chart. Around that time, popular music programs were established on TV and were beginning to take off and, rather than singer-songwriters who pour forth their beliefs and feelings in their lyrics, it was the attractive, generally well-received idols who attracted sponsors that were popular [17]. In contrast to this, in the 1990s, there was a gathering momentum towards female social equality and independence, and the number of women and young people buying music was increasing [17].As a result, the genre of the female singer-songwriter seemed to be energized once more in place of the spectacle of the idol. It can be said that the results of this study clearly reflect the changes in Japanese society and the changes in the music industry.

## Basic characteristics for selected songs

Table 2 represents the number of texts, and the mean and standard deviations with coefficients of variances of the number of tokens for the 10 singer-songwriters. The number of texts ranges from 7 to 20, and the mean number of tokens ranges from 155 to 273. The table shows that Utada and Nakajima have a large number of songs, whereas Hirose, Takeuchi and Shina have a small number of songs in our dataset. It also shows that Hirose generally prefers short songs, and YUI and Utada have large variances in the length of their songs. As explained later, this may be because they preferred English words.

## Results on kernel principal component analysis (kernel PCA)

Figure 1 and 2 show three dimensional scatter plots of the first three principal components by kernel PCA. The results show that the songs by the same singer are grouped, for example, aiko's songs are grouped in the upper side, Utada and YUI's songs toward the right, and Nakajima's songs toward the left. In the kernel PCA, compared with the factors concerning the singer, the factors concerning the era are not clearly observed, which was also true when the release years were used as labels.

The grouping results and a qualitative analysis of the songs shows that the first principal component includes the factor of loanwords because Utada and YUI's songs have many of such words, and the third principal component includes the factor of pronouns, because aiko's songs have many special first person pronouns. However, these three axes are composed of many words, thus their labeling becomes difficult.

## Results on random forests

Table 3 represents the confusion matrix constructed by random forests. The results show that the songs by aiko, Nakajima, and Utada have high classification performance, whereas those by Hirose, Oguro, and Matsutoya have low classification performance. These results also show that the songs by the first three singer-songwriters have a special individual characteristics among Japanese female singer-songwriters, while those by the last three singer-songwriters have common characteristics. It is certain that the number of texts and the number of tokens can affect classification performance. However, aiko, for example, has no special lyrical characteristics within these textual characteristics; thus, these results also show the special lyrical characteristics of the singers.

Table 4 represents the results of the top 20 important features obtained from the classification experiment calculated by random forests. The results show that the top 20 important features include many pronouns, final particles, and auxiliary verbs. The results also show that these words are particularly important for discriminating the songs by 10 Japanese female singer-songwriters.

Next, we conducted a classification experiment using the lyrics of the three singer-songwriters who made their debut in the 1970s and the lyrics of the seven singer-songwriters who have made their debut since the 1990s. According to the kernel PCA, differences based on singer are more significant than those based on the simple era labels. Therefore, era classification was conducted according to the debut year of the singer rather than by decade. For example, in the 1990s there are of course compositions written by Miyuki Nakajima, but these are treated as Miyuki Nakajima's compositions. Table 5 shows the confusion matrix and Table 6 shows the top 20 important features in the experiment. We found that a number of trends can be identified in Table 5 because there are many songs from the new era, but a moderate classification was obtained from the confusion matrix results. Many function words, such as "wa" and "wo" and many personal pronouns, such as "watashi (I)," are included in Table 6 as well as in Table 4. As we shall see later, the other words of high importance in the singer-songwriter classification, those that contribute significantly to the classification, also achieve high values in the classification according to era.

Table 7 represents the top 20 important features for discriminating the 10 singer-song-writers. Next, we discuss the results in the last table together with the qualitative analysis results. In particular, the analysis for Komi Hirose, Yumi Matsutoya, and Maki Oguro, for whom the classification experiment performance was poor, was supplemented by a mainly qualitative study of their lyrics.

### aiko

The first person pronoun 'atashi (I)' was the most important feature, although 'watashi (I)' is usually used in spoken Japanese. Pronouns have strong discriminant power in authorship attribution [6, 10]; thus, we inferred that this lyrical characteristic of her songs led to the lowest error rates in Table 2. In addition, words representing parts of the body, such as 'me (eyes)', 'hoho (cheeks)', and 'kuchibiru (lips)' appeared at 7, 8 and 14 ranks, and we inferred that they show her special lyrical characteristic. This is apparent in her use of the following lyrics for example, "夏髪が頬を切る (natsukami ga hoho wo kiru; summer hair brushes against my cheeks)" ("KissHug"), "赤く染まる指先や頬を (akaku somaru yubisaki ya hoho wo; your blushing fingertips and cheeks)" ("スター [star]"), "唇かんで指で触ってあなたとのキス確かめてたら (kuchibiru kande yubi de sawatte anata tono kisu tashikame te tara; bite my lip, touch it with my finger, when I go over kissing you in my mind)" ("ボーイフレンド [Boyfriend]"), "頬は熱くなって　たまに悲しくもなった (hoho wa atsuku natte tamani kanashiku mo natta; my cheeks burned hot and

now and again I felt sad" ("アンドロメダ [Andromeda]") and "三角の耳した天使は恋のため息聞いて目を丸くしたあたしを指さし (sankaku no mimishita tenshi wa koi no tameiki kiite me wo maruku shita atashi wo yubisashi; the angel with the triangular ears heard a sigh of love and pointed at me, my eyes wide" ("花火 [Hanabi; Firework]"). Many other words representing parts of the body that are not in the ranking of important words, such as "tsume (nails)," "kami (hair)," "yubi (fingers)," "mune (chest)," "mimi (ears)," and "te (hands)," are apparent in such lyrics as, "あたしの髪が揺れる距離の息づかいやきつく握り返してくれた手はさらに　消えなくなるのにね (atashi no kami ga yureru kyori no ikidsukai ya kitsuku nigirikaeshite kureta te wa sarani kienaku naru nonine; your breath and the hand that gripped me tightly does not fade with a shake of my hair" ("アンドロメダ [Andromeda]") and "深爪したことも (fukadsume shita koto mo; and a nail cut too close" ("キラキラ [Kirakira; sparkle]").

**Komi Hirose**

Many content words is ranked in the top ranked words. Contrary to function words, content words were known to serve as noises for better classification between authors [6]; therefore, we inferred that this lyrical characteristic of her songs led to high error rates shown in Table 2. It can be inferred that many of Hirose' s song lyrics concern the theme of love because the words "love" and "ai (love)" appear as one of the important words. It is thought that the word "love" ranked top because there are many phrases that contain "love" applied to sections where the lyrics are repeated, such as "Fall in Love ロマンスの神様　この人でしょうか (Fall in Love romansu no kamisama konohito de shou ka?; fall in love, is this person the God of romance?" ("ロマンスの神様 [Romance no kamisama; God of Romance]"), "私だけに White Love Song 歌ってほしいの (watashidakeni White Love Song utatte hoshii no; I want you to sing a white love song just for me" ("ゲレンデがとけるほど恋したい [Garendega tokeru hodo koi shitai; I want to love you in a way that will melt the ski slopes])") and "ずっと Eternal Love (zutto Eternal Love; forever, eternal love)" ("Promise"). In contrast to the other artists, Komi Hirose was the only artist to have the word love in romanized writing in the top ranked words. Furthermore, Komi Hirose conjures a powerful image of "winter" in Japan, but not a single word relating to winter featured in the top ranking words, an unexpected result.

**Yumi Matsutoya**

The particle "wo" that represents the object is ranked in the top ranked words. Examples of lyrics that use "wo" include, "初めて言葉を 交わした日の その瞳を 忘れないで (Hajimete kotoba wo kawashita hi no sono hitomi wo wasure nai de; don't forget the way your eyes were when we spoke for the first time)" ("守ってあげたい (Mamotte agetai; I want to protect you)") and "夢をくれし君の 眼差しが肩を抱く (yumewo kureshi kimi no manazashi ga kata wo daku; the protective gaze of the one who dreams of me wraps around my shoulders)" ("春よ、来い [Haruyo, koi; spring comes]"). In addition it can be inferred that many of Yumi Matsutoya's lyrics concern love because the word "ai (love)" appears in the top rankings. Picking out several lyrics as they appear reveals lyrical themes concerning love. For example, "愛をくれし君の なつかしき声がする (ai wo kureshi kimi no natsukashiki koe ga suru; I hear the longed-for voice of the one who loves me)" ("春よ、来い [Haruyo, koi; spring comes]") and "私を愛したことを後悔はしていないかしら (watashi wo aishita koto wo koukai ha shite inai kashira; I wonder if you regret loving me)" ("輪舞曲 (Rondo)").

**Miyuki Nakajima**

Hiragana appeared at all except 10, 15 and 16 ranks, and many of Hiragana words were function words, though other singer-songwriters included more content words. We inferred that she had a special lyrical characteristic on the basis of the function words rather than the content words. Function words have strong discriminant power in authorship attribution [6,10]; thus, we inferred that this lyrical characteristic of her songs led to low error rates shown in Table 2.

**Maki Oguro**

Compared to the other artists, there are few characteristic words and also the classification performance is poor. We conclude that either Maki Oguro is an artist with few characteristic words or she is an artist with an extensive vocabulary who does not use similar expressions. However, it should be noted is that the word "kako (past)" is included only in Oguro's important words. Words that represent time, such as "kako (past)", "mirai (future)" and "genzai (present)", are not commonly heard from the other artists. Specifically, phrases such as, "泣きながらあなたを諦めようとした過去も (nakinagara anata wo akirame you to shita kakomo; the past when I cried and tried to give you up" ("熱くなれ [Atsukunare; heat up]") and "過去を責めても あなたは帰らない (kako wo semetemo anata wa nera nai; even if I blame the past, you do not return)" ("チョット [Chotto; a moment]"), appear in her lyrics.

**Ai Otsuka**

Mathematical numbers such as 1 and 2 appeared in the 1 and 4 ranks, while we can use kanji to indicate the same meaning in Japanese. In addition, katakana in Japanese ('koto (thing)') appeared at 7 rank and chatter expressions such as 'Naa' and 'nante' appeared at 14 and 15 ranks. We inferred that these expressions make her songs moderate and accessible, and can represent her special lyrical characteristics. Examples of her lyrics include "恋してれば 全てが 2 倍 Power (aishitereba subetega 2bai Power; When we're in love, everything grows to twice, the power)", "チューすれば 1 話始まる Story (chusureba 1wa hajimaru Story; When we kiss one story begins)" ("CHU-LIP"), "もう 1 つ食べたいわ (mou hitotsu tabetai wa; I want to eat one more)," "もう 1 杯飲みたいわ (mou ippai nomi tai wa; I want to drink one more)", "1 人はとてもめんどうだから (hittori wa totemo mendou dakara; It is tedious being alone)" ("フレンジャー [Frienger]"), "夏の終わりに 2 人で抜けだした この公園で見つけた (natsu no owarini futari de nukedashita kono koen de mitsuketa; we snuck out at the end of summer and met at this park)", and "1 番に君が好きだよ 強くいられる (ichiban ni kimi ga suki dayo tsuyoku irareru; I love you more than anything and it makes me strong)" ("Planetarium"). The fact the word "koto (thing)" written in kanji appeared in Aiko's ranking, but the same word in katakana appears in Otsuka's ranking and is also noteworthy. Using hiragana and katakana instead of Chinese numerals and kanji softens the appearance of her lyrics. It does not appear in the top important words, but the use of "toki (when)" in katakana instead of kanji in "好きなトキ出かけて 好きなトキ甘えて (sukina toki dekakete sukina toki amaete; going out when I want to, spoiling myself when I want to)" ("ネコに風船 [Nekoni fuusen; cat with a balloon]"), is an example of this.

**Ringo Shina**

Kanji appeared at 2, 5, 6, 7, 11, 12, 15, 17 ranks. Even function words such as 'sono (its)' and 'made (till)' were used in kanji, although these words are usually written in hiragana in modern Japanese. We inferred that, contrary to Otsuka, these expressions make her songs solid and aggressive, and can represent her special lyrical characteristic.

The lyrics "グラスよりも其の御口に注いで戴きたいのだもの (gurasu yorimo sono okuchi ni sosoide itadakitai no da mono; pour into not my glass but my mouth)" (真夜中は純潔 [Moyonaka wa junketsu; midnight is chaste]) are a specific example of this. In addition, Ai Otsuka and Ringo Shina use "居 (i; presence)" as representing the same meaning, but Shina often ventures to use the kanji. For example, Otsuka writes, "いつだって そこにいてあげるんだ (itsu datte soko ni ite agerunda; I'll always be there for you)" ("フレンジャー [Frienger]"), but Shina writes "一番愛しいあなたの声迄 掠れさせて居たのだろう (ichiban itoshii anata no koe made kasuresasete ita no darou; turned even your voice, the thing I love most, hoarse)" ("罪と罰 [Tsumi to batsu; crime and punishment]") and "ずっと繋がれて 居たいわ (zutto tsunagarete itai wa; I want to stay connected to you forever)" ("本能 [honnou; instinct]"). Other examples of her lyrics that do not appear in the top ranking in which she also uses kanji are "返して貰うまでもない筈 (kaeshite morau mademo naihazu; you shouldn't even go so far as to get it back)" ("あ

りあまる富 [Ariamaru tomi; excessive wealth]”), “もっと中迄入って (motto nakamade haite; get deeper inside)” (“本能 [honnou; instinct]”) and “此処に居て (kokoni ite; stay here)” (“ギブス [Gibusu; cast]”).

**Mariya Takeuchi**

Content words appeared at 2, 4, 5, 6, 12, 13, 16, 19, 20 ranks, ; therefore, we inferred that this lyrical characteristic of her songs led to high error rates shown in Table 2 as well as Komi Hirose. Among the 10 singer-songwriters, Takeuchi and Nakajima belonged to the 1970 and 1990 era, their lyrical characteristics were considerably different. In Takeuchi’s songs, ‘Denwa (telephone)’ that represent the communication styles in that era appeared at 20 rank, while e-mail or cell phones were not famous during that era. We inferred that these words represent her special lyrical characteristic, as well as people’s communication styles during that era. It can also be inferred that the subject of many of her songs is love. Specific examples include, “私だって命がけの恋に憧れることはある (watashi datte inochigake no koini akogareru kotowa aru; I yearn for a desperate love)” (“純愛ラプソディ [Junai Rhapshody]”), “手放した恋を今 あなたも悔やんでるなら (tebanashita koi wo ima anatamo kuyanderu nara; if you also regret throwing our love away)” and “電話ぐらいくれてもいいのに (denwa gurai kuretemo iinoni; just a phone call would do)” (“シングル・アゲイン [single again]”) and “にぎわう街の音がかすかに聞こえる (nigiwau machi no oto ga kasukani kikoeru; I hear the faint sound of the bustling city)” (“カムフラージュ [camouflage]”).

**Hikaru Utada**

The second person pronoun ‘Kimi (you)’ was the most important distinguishing characteristic of Utada’s songs, while first person pronouns were more important in the other nine singers.

In the case of other artists, words that represent the first person, such as “atashi (I)” and “watashi (I)” appear in the top rankings more often than words that represent the second person, such as “kimi (you)” and “anata (you),” and it is only in Utada’s lyrics that second person pronouns out rank first person pronouns. Specific examples of her lyrics with “kimi (you)” include, “ありがとう、と君に言われると なんだかせつない (arigatou, to kimi ni iwareru to nandaka setsunai; when you tell me thank you, it is somehow bittersweet)” (“Flavor of Life”), “いいじゃないか キャンバスは君のもの (iija nai ka kyanbasu wa kimi no mono; that’ll do, right? The canvas is yours)” (“Colors”), “君という光が私をみつける (kimi to iu hikari ga watashi wo mitsukeru; a light called “you” has found me)” (“光 [hikari; light]”), “もっと君に近づきたいよ (motto kimini chikazuki taiyo; I want to get closer to you)” and “そっと君に手を伸ばすよ (sotto kimi ni te wo nobasuyo; reaching out for you)” (“SAKURA ドロップス [sakura drops]”) and “近づきたいよ 君の理想に (chikazuki taiyo kimi no risouni; I want to get closer to your ideal)” (“Can You Keep a Secret”). In addition, many English words such as ‘I’ and ‘baby’ appeared at 2, 3, 6, 7, 16 ranks, and we inferred that they represent her special lyrical characteristics. It is inferred that she uses a lot of imported words because she was born in New York in America.

**YUI**

Chatter expressions such as ‘nante’, ‘desyo’, ‘tte’, ‘ja’, as well as question marks, appeared at 1, 2, 4, 7, 18 ranks.

YUI moved to Tokyo from Fukuoka at the age of 17 and is the youngest artist in this analysis. It is inferred that her lyrics contain Japanese expression of the new generation. Although YUI is commonly known to prefer English words in her lyrics [3],[4] but this characteristics of her lyrics did not clearly appear in Table 2, because she used various types of English words, and the number of tokens of these words was rather small; thus, it can not be used well for her classification. In addition, the word “yoru (night)” appeared in YUI’s important words, but not those of any of the other artists. The appearance of words that represent periods of time within a day, such as “asa (morning)” and “hiru (afternoon)” and

[4]For example, all the titles of her songs are in English or written in alphabets [3].

“gozen (morning)” and “gogo (afternoon),” clearly demonstrates a world view in terms of time in the lyrics. Specifically, in lyrics such as “星の夜 願い込めて CHE.R.RY (hoshi no yoru negai komete che.r.ry; I make a wish in the starry night, cherry)” (“Che.r.ry”), “なみだいろ 声が聞こえない夜は (Namida iro koe ga kikoe nai yoru wa; the color of tears on nights when I can’t hear your voice)” (“Namidairo (the color of tears)”) and “わすれちゃいそうな 夜の真ん中 (wasure chai souna yoru no mannaka; in the middle of the night I begin to forget that)” (“Again”), it is possible to understand the specific time of day and imagine the situation as you listen and become more emotionally involved with the lyrics.

Furthermore, it was not part of the lyrical analysis this time, but all of the names of YUI’s songs are written using a foreign language. YUI was born in Fukuoka Prefecture and does not have the same unusual background as Hikaru Utada, and therefore it could be said that this is an idiosyncrasy of YUI’s. YUI often romanizes expressions that were not originally in English, such as, “CHE.R.RY” and “Namidairo,” and it is possible that she has deliberately used an English title to address any preconceptions that might arise from using a Japanese title. These points require further study.

## Further discussion

In the section above, we analyzed the characteristics of individual singer-songwriters. The common point that emerged from this analysis is that love is invariably an important subject for women regardless of the era. Many of the listeners to songs written and sung by women such as Yumi Matsutoya and Mariya Takeuchi in earlier decades and aiko and Ai Otsuka in recent decades are young women, and love is a common theme and an important feature of communication for this age group.

However, the style used to talk about love differs by era and singer. For example, singers like Mariya Takeuchi and Yumi Matsutoya use many words related less to linguistic expression and more to lyrical content, imagery, and feelings, expressing their individuality through such content and imagery. In contrast, modern singers assert a stronger linguistic and stylistic individuality: aiko with her concrete representations of the body, Utada with imported words, YUI with colloquialisms, Otsuka with numerals, hiragana, katakana, and so on, and Shiina with formal words and words of Chinese origin. This is a main feature of the lyrics of Japanese female singer-songwriters from the 1990s onwards, following the idol period of the 1980s.

Naturally, with this change there is also a corresponding change in the listeners. Since 2000, communication on the Internet and by cell phone has increased and people’s conditions for communication and the style of language people use when communicating continues to change significantly. The arrival of the Internet is producing a new Japanese language that is neither spoken nor written language [18], and the shape of the Japanese language and, in particular, the way in which young people communicate, is changing. As a new style emerges from the past and traditional styles of writing narratives continue to change, people are also choosing different styles of, using different expressions for talking about the same emotions. We conclude that this change in the conditions for communication among such people corresponds to the change in the lyrics of singer-songwriters.

According to the findings of our research, the most significant change in the lyrics of current singer-songwriters can be conceptualized as ’the visual effects of the lyrics.’ This partially corresponds with the general change in Japanese characters and communication styles among Japanese youngsters. However, this concept includes aspects that are specific to the stylistic characters of their lyrics.

As we mention above, a special characteristic of current singer-songwriters is their varied characters. For example, a singer prefers “koto (thing)” in katakana, while another prefers it in kanji. Simultaneously, another singer favors “sono (that)” in hiragana, whereas another prefers it in kanji. These examples correspond with general changes in Japanese during the 1990s and 2000s as well as the utilization of the Interne, mobile phones, and e-mail as popular communication tools. More specifically, among the young females, the emotions or feelings that the ’characters’ transmit have become particularly important. This has led to the overall popularity of emojis and emoticons. However, the styles of their lyrics still include an independent character that differs from the general changes in Japanese.

With regard to emojis and emoticons, these general changes in Japanese are written ones. However, written lyrics also have an important role: they must be sung. Thus, we assume that in the dynamic era of Japanese and its communication styles, singer-songwriters must be sensitive in their overall approach to writing lyrics, in which the meanings of the words and their characters must effectively transmit personal feelings or emotions. As mentioned in the introduction, karaoke is a popular activity among young people, in which 'singing' the displayed lyrics is equally important as 'being heard'. Current female singer-songwriters select personal lyrical writing styles that correspond to the latest communication styles. They must naturally consider that their lyrics will be 'sung'. However, we must assume that they are sensitive to the fact that their lyrics will also be 'seen'.

Compare to this aforementioned approach, the use of loan words is not a unique characteristic that reflects the current communication styles of youngsters. Our results indicate that some singers, such as Utada and YUI, prefer using English words in their lyrics. This was also seen in the lyrics of singer-songwriters during the 1980s, including Nakajima, Takeuchi, and Matsutoya. In this case, English words could provide an alienation effect when heard, but visually, they are still written in the standerd manner. Therefore, we can conclude that this is a secondary point of discussion regarding the lyrics of current female singer-songwriters.[5]

## Conclusion

In summary, this study analyzed popular Japanese songs composed by Japanese female singer-songwriters over the past 30 years, by using methods of computational stylistics. Our texts contained various characteristics as mentioned in Introduction; thus our study using kernel PCA and random forests provided an important case study for computational stylistics. We also provided empirical knowledge for understanding the Japanese language, and communication styles.

The results of this study showed that although love is a common theme in the lyrics of female singer-songwriters, the manner of expressing love varies significantly depending on the era and the individual singer. The subtle topic of love certainly differs for every singer-songwriter as well as every song. However, we found many content-independent characteristics, such as the selection of kanji and kana, in singer-songwriters after the 2000s. These kinds of individuality are what make ordinary people make comments like "This sounds like such and such a singer," and we were able to clarify these differences by applying machine learning. The results of this study suggest that the variety of styles is an important characteristics of current singer-songwriters. Thus, we can conclude that they emphasize linguistic and stylistic factors, including word play.

In this study, we took the theoretical premise that lyrics are important for understanding modern society and submitted it to empirical analysis. According to Mita [1], popular songs are first 'sung', and 'The action of 'Singing' is primary and open to a wider range of people than is the action of 'Writing', thus it is closer to the ordinary'. However, in contemporary Japanese society, in which karaoke is popular and communication styles have changed, the popular songs are 'seen' as well as 'sung'. This 'visualization of the lyrics' is a typical characteristic of current singer-songwriters. Therefore, this study adds this point to Mita's theory for identifying the application of popular songs in current Japanese popular culture and society.

The shifts in modern Japanese culture can be investigated further by analyzing each respective genre, lyrics in conjunction with the other non-lyric musical elements of popular music. It is no doubt possible to enhance the results of this study by applying content analysis to the lyrics and by using a larger data set. We intend to work on these points in the future.

[5] YUI's CHE.R.RY is an exception that uses 'the visual effects of the lyrics' even in English.

# References


1. Mita M (1975) Kindai Nihon no Shinjo no Rekishi: Ryukoka no Syakaishinri-shi. Tokyo: Kodansya.
2. Kikuchi K (2008) Nihon Ryuko-ka Hensen Shi: Kayo Kyoku no Tanjo kara J-Pop no Jidai e. Tokyo: Ronsosya.
3. Hosoya M (2010) Analyses of female singer-songwriters. Undergraduate Thesis, Department of Media and Communications, Faculity of Sociology, Toyo University.
4. Ito M (2001) The judgement standard for distinguishing loan words of western origin from western words in Japanese pop songs. Mathematical Linguistics 23: 110-130.
5. Argamon S, Whitelaw C, Chase P, Raj Hota S, Garg N, et al. (2007) Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology 58: 802-822.
6. Suzuki T (2009) 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 60: 1596-1606.
7. Potthast M, Stein B, Barrón-Cedeño A, Rosso P (2010) An evaluation framework for plagiarism detection. In: Proceedings of COLING2010: the 23rd International Conference on Computational Linguistics. pp. 997-1005.
8. (1970-1979) Confidence Annual Report. Original Confidence.
9. (1980-2009) Oricon Annual Report (Oricon Year Book). Original Confidence (Oricon).
10. Stamatatos E (2009) A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60: 538-556.
11. Breiman L (2001) Random forests. Machine Learning 45: 5-23.
12. Jin M, Murakami M (2007) Authorship identification using random forests. Proceedings of the Institute of Statistical Mathematics 55: 255-268.
13. Breiman L (1996) Bagging predictors. Machine Learning 24: 123-140.
14. Jin M (2007) R ni yoru Deta Saiensu. Tokyo: Morikita Publishing Co., Ltd.
15. Bishop CM (2006) Pattern Recognition And Machine Learning. New York: Springer Science + Business Media, LLC.
16. Breiman L, Cutler A (2004). Random forests. www.stat.berkeley.edu/ ˜breiman/ RandomForests (accessed Jan. 6, 2011).
17. Ugaya H (2005) Jpopu towa Nani ka: Kyodaika-suru Ongaku Sangyo. Tokyo: Iwanami Syoten.
18. Suzuki T, Kawamura S, Aizawa A (2012) Stylistic analysis of text submission to Japanese Q&A communities. Journal of Quantitative Linguistics (to appear).

# Figures

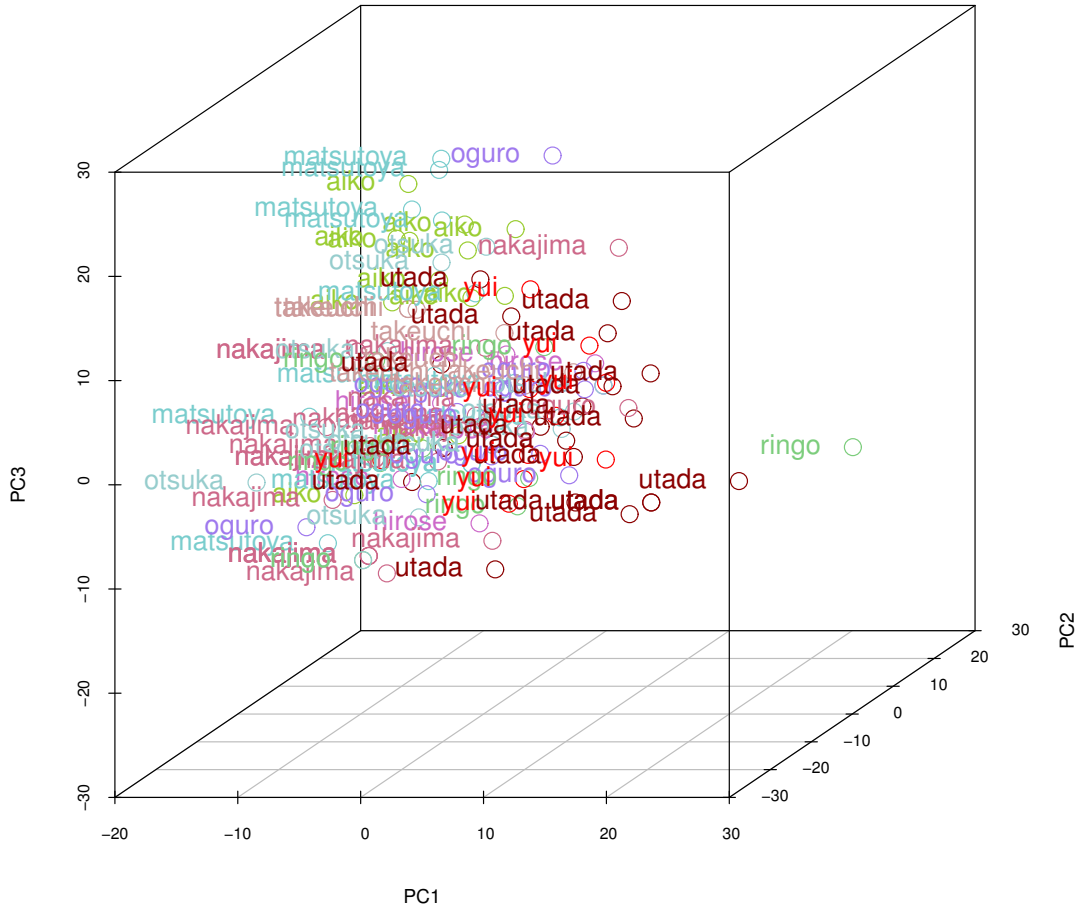


**Figure 1. Three dimensional scatter plots of the first three principal components by kernel PCA.** The labels represent the names of the singer-songwriters.

# Figures

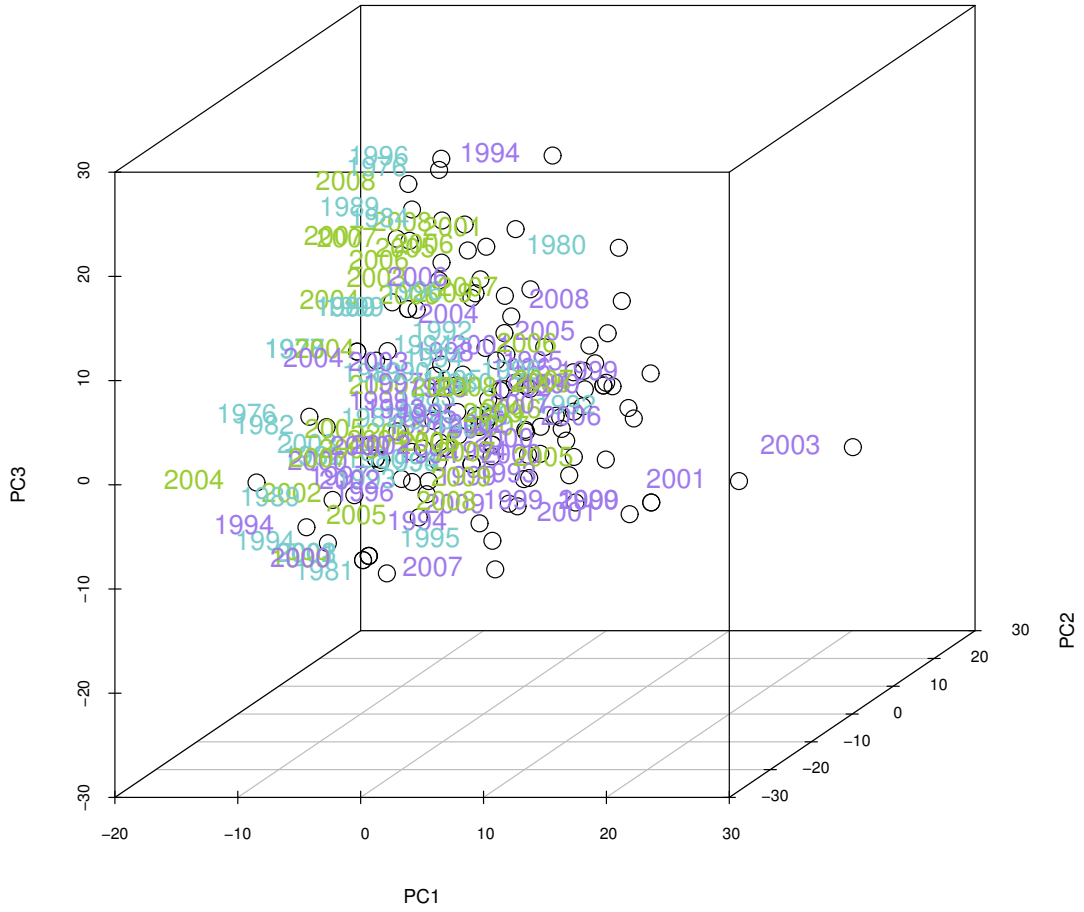


**Figure 2. Three dimensional scatter plots of the first three principal components by kernel PCA (2).** The labels represent the years of release.

# Tables

**Table 1. Female singer-songwriters**

| | Debut year | Debut age |
|---|---|---|
| aiko | 1998 | 23 |
| Komi Hirose | 1992 | 26 |
| Yumi Matsutoya | 1973 | 18 |
| Miyuki Nakajima | 1975 | 23 |
| Maki Oguro | 1992 | 22 |
| Ai Otsuka | 2003 | 21 |
| Ringo Shina | 1998 | 19 |
| Mariya Takeuchi | 1978 | 23 |
| Hikaru Utada | 1998 | 15 |
| YUI | 2005 | 17 |

**Table 2. Results of basic characteristics**

| | | number of tokens | | |
|---|---|---|---|---|
| | number of texts | mean | s.d. | c.v. |
| aiko | 13 | 217.08 | 49.59 | 22.84 |
| Komi Hirose | 7 | 273.00 | 31.37 | 14.45 |
| Yumi Matsutoya | 13 | 155.31 | 40.51 | 18.66 |
| Miyuki Nakajima | 20 | 211.30 | 35.82 | 16.50 |
| Maki Oguro | 13 | 231.38 | 48.84 | 22.50 |
| Ai Otsuka | 15 | 209.33 | 49.99 | 23.03 |
| Ringo Shina | 9 | 199.78 | 46.19 | 21.28 |
| Mariya Takeuchi | 8 | 199.25 | 48.22 | 22.21 |
| Hikaru Utada | 20 | 235.70 | 51.69 | 23.81 |
| YUI | 10 | 243.10 | 68.61 | 31.61 |

The number of texts, the mean, standard deviations and coefficients of variation of the number of tokens for the 10 singer-songwriters are shown.

**Table 3. Results of the classification experiment by using random forests**

| | aiko | Komi Hirose | Yumi Matsutoya | Miyuki Nakajima | Maki Oguro | Ai Otsuka | Ringo Shina | Mariya Takeuchi | Hikaru Utada | YUI | error rates |
|---|---|---|---|---|---|---|---|---|---|---|---|
| aiko | 13 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 |
| Komi Hirose | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 3 | 0 | 1.00 |
| Yumi Matsutoya | 0 | 0 | 2 | 10 | 0 | 0 | 0 | 0 | 1 | 0 | 0.85 |
| Miyuki Nakajima | 0 | 0 | 0 | 19 | 0 | 0 | 0 | 0 | 1 | 0 | 0.05 |
| Maki Oguro | 0 | 0 | 1 | 1 | 1 | 2 | 0 | 0 | 8 | 0 | 0.92 |
| Ai Otsuka | 1 | 0 | 0 | 2 | 0 | 10 | 0 | 0 | 2 | 0 | 0.33 |
| Ringo Shina | 1 | 0 | 1 | 1 | 0 | 0 | 2 | 0 | 4 | 0 | 0.78 |
| Mariya Takeuchi | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 2 | 3 | 1 | 0.75 |
| Hikaru Utada | 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 |

Confusion matrix and error rates are shown.The rows represent the original data and the columns represent the results from the classification experiments; thus, it represents that all the aiko's songs were recognized from her texts, whereas none of Hirose's songs were identified from her lyrics.

**Table 4. Top 20 important features for the classification experiment**

| rank | morphemes | $VI_{acu}$ |
|---|---|---|
| 1 | atashi | 0.0125 |
| 2 | shi | 0.0078 |
| 3 | 1 | 0.0067 |
| 4 | watashi | 0.0059 |
| 5 | you | 0.0053 |
| 6 | da | 0.0052 |
| 7 | wo | 0.0050 |
| 8 | shiawase | 0.0047 |
| 9 | ta | 0.0045 |
| 10 | kimi | 0.0045 |
| 11 | nante | 0.0039 |
| 12 | ii | 0.0039 |
| 13 | e | 0.0038 |
| 14 | yo | 0.0036 |
| 15 | ai | 0.0032 |
| 16 | n | 0.0032 |
| 17 | tai | 0.0032 |
| 18 | nai | 0.0032 |
| 19 | kara | 0.0031 |
| 20 | koto | 0.0031 |

The top 20 important features calculated by random forests. The definition of $VI_{acu}$ is described in the Data and Methods section.

**Table 5. Results of the classification experiment by using random forests**

| | old | new | error rates |
|---|---|---|---|
| old | 11 | 30 | 0.73 |
| new | 0 | 87 | 0 |

Confusion matrix and error rates are shown.The rows represent the original data and the columns represent the results from the classification experiments.

**Table 6. Top 20 important features for the two-class (old/new singer-songwriters) classification experiment**

| rank | morphemes | $VI_{acu}$ |
|---|---|---|
| 1 | wo | 0.00824 |
| 2 | watashi | 0.00819 |
| 3 | da | 0.00545 |
| 4 | ai | 0.00501 |
| 5 | hitori | 0.00478 |
| 6 | ha | 0.00420 |
| 7 | nagare | 0.00375 |
| 8 | shi | 0.00306 |
| 9 | tai | 0.00302 |
| 10 | atashi | 0.00290 |
| 11 | ? | 0.00287 |
| 12 | koto | 0.00232 |
| 13 | tachi | 0.00188 |
| 14 | hitotsu | 0.00187 |
| 15 | machikado | 0.00185 |
| 16 | hitomi | 0.00174 |
| 17 | n | 0.00162 |
| 18 | no | 0.00154 |
| 19 | chigai | 0.00138 |
| 20 | yuku | 0.00136 |

The top 20 important features calculated by random forests. The definition of $VI_{acu}$ is described in the Data and Methods section.

**Table 7. Top 20 important features for distinguishing the 10 singer-songwriters**

| | aiko | | Komi Hirose | | Yumi Matsutoya | | Miyuki Nakajima | | Maki Oguro | |
|---|---|---|---|---|---|---|---|---|---|---|
| rank | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ |
| 1 | atashi | 0.0558 | Love | 0.0124 | wo | 0.0193 | shi | 0.0261 | shi | 0.0058 |
| 2 | you | 0.0379 | hito | 0.0111 | ai | 0.0092 | da | 0.0202 | kara | 0.0043 |
| 3 | koto(kanji) | 0.0227 | eien | 0.0073 | atashi | 0.0076 | e | 0.0178 | atashi | 0.0040 |
| 4 | hodo | 0.0077 | watashi | 0.0072 | da | 0.0073 | yo | 0.0151 | you | 0.0039 |
| 5 | watashi | 0.0065 | kimi | 0.0049 | ja | 0.0072 | atashi | 0.0133 | kako | 0.0034 |
| 6 | hikari | 0.0063 | ni | 0.0043 | nante | 0.0059 | tai | 0.0127 | ni | 0.0027 |
| 7 | me | 0.0061 | wo | 0.0036 | mi | 0.0055 | tachi | 0.0116 | suru | 0.0024 |
| 8 | hoho | 0.0055 | wa | 0.0035 | watashi | 0.0052 | ni | 0.0098 | wo | 0.0023 |
| 9 | mo | 0.0051 | cha | 0.0035 | kimi | 0.0050 | wo | 0.0080 | 1 | 0.0020 |
| 10 | omoi | 0.0049 | purezento | 0.0034 | nai | 0.0044 | watashi | 0.0078 | kimi | 0.0018 |
| 11 | hure | 0.0042 | kitto | 0.0030 | ba | 0.0038 | ta | 0.0076 | u | 0.0017 |
| 12 | shi | 0.0039 | shi | 0.0030 | ii | 0.0035 | ii | 0.0072 | mi | 0.0016 |
| 13 | ta | 0.0038 | atashi | 0.0028 | omoi | 0.0033 | wa | 0.0068 | datte | 0.0015 |
| 14 | kuchibiru | 0.0037 | kure | 0.0027 | toki | 0.0032 | kara | 0.0063 | n | 0.0014 |
| 15 | ai | 0.0032 | ta | 0.0026 | koto (hiragana) | 0.0032 | hito | 0.0061 | mata | 0.0013 |
| 16 | ... | 0.0028 | nante | 0.0024 | suru | 0.0031 | ? | 0.0060 | ga | 0.0013 |
| 17 | anata | 0.0027 | chaa | 0.0023 | toki | 0.0029 | mo | 0.0057 | sukoshi | 0.0013 |
| 18 | ii | 0.0025 | ga | 0.0022 | e | 0.0028 | n | 0.0054 | ai | 0.0012 |
| 19 | datte | 0.0024 | shiawase | 0.0021 | yo | 0.0028 | nai | 0.0051 | yuku | 0.0012 |
| 20 | you | 0.0024 | ai | 0.0020 | shi | 0.0026 | anata | 0.0050 | shiawase | 0.0011 |
| | Ai Otsuka | | Ringo Shina | | Mariya Takeuchi | | Hikaru Utada | | aiko | |
| rank | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ | morphemes | $VI_{acu}$ |
| 1 | 1 | 0.0439 | atashi | 0.0113 | watashi | 0.0143 | kimi | 0.0141 | nante | 0.0213 |
| 2 | shiawase | 0.0280 | i | 0.0106 | machi | 0.0123 | I | 0.0097 | desyo | 0.0165 |
| 3 | atashi | 0.0144 | da | 0.0103 | shi | 0.0117 | you | 0.0091 | n | 0.0109 |
| 4 | 2 | 0.0105 | ta | 0.0103 | koi | 0.0104 | ii | 0.0091 | ? | 0.0075 |
| 5 | watashi | 0.0099 | hito | 0.0086 | futari | 0.0097 | atashi | 0.0079 | no | 0.0074 |
| 6 | mo | 0.0073 | watashi | 0.0078 | omoi | 0.0078 | ’ | 0.0069 | ta | 0.0072 |
| 7 | koto (katakana) | 0.0067 | sono | 0.0069 | da | 0.0074 | baby | 0.0062 | tte | 0.0065 |
| 8 | nai | 0.0066 | ii | 0.0062 | ta | 0.0072 | doko | 0.0058 | watashi | 0.0053 |
| 9 | wo | 0.0063 | desyo | 0.0052 | ii | 0.0068 | n | 0.0053 | nai | 0.0050 |
| 10 | da | 0.0063 | nante | 0.0051 | wo | 0.0065 | shi | 0.0047 | waka | 0.0044 |
| 11 | suru | 0.0063 | you | 0.0051 | atashi | 0.0063 | anata | 0.0041 | hito | 0.0040 |
| 12 | shi | 0.0062 | seimei | 0.0048 | deatt | 0.0061 | te | 0.0041 | deki | 0.0039 |
| 13 | kimi | 0.0059 | re | 0.0045 | kanji | 0.0058 | nai | 0.0036 | da | 0.0037 |
| 14 | naa | 0.0055 | wo | 0.0043 | ’ | 0.0053 | kokoro | 0.0033 | atashi | 0.0037 |
| 15 | nante | 0.0054 | kimi | 0.0043 | wa | 0.0053 | ai | 0.0033 | wo | 0.0036 |
| 16 | to | 0.0052 | tai | 0.0040 | hontou | 0.0052 | can | 0.0030 | mo | 0.0030 |
| 17 | tai | 0.0052 | kokoro | 0.0040 | tai | 0.0049 | shiawase | 0.0028 | yoru | 0.0029 |
| 18 | ai | 0.0050 | nado | 0.0038 | nai | 0.0046 | toki | 0.0026 | ja | 0.0028 |
| 19 | mi | 0.0047 | te | 0.0037 | eran | 0.0044 | sukoshi | 0.0026 | tai | 0.0027 |
| 20 | na | 0.0044 | yo | 0.0036 | denwa | 0.0043 | datte | 0.0025 | anata | 0.0024 |

The top 20 important features calculated by random forests. The definition of $VI_{acu}$ is described in the Data and Methods section.