The University of Osaka
Institutional Knowledge Archive

Title	『歌詞コンテンツデータ集』の概要と研究利用の可能性
Author(s)	東条, 佳奈
Citation	現代日本語研究. 2021, 13, p. 47-64
Version Type	VoR
URL	https://doi.org/10.18910/88321
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

『歌詞コンテンツデータ集』の概要と研究利用の可能性

Overview of “Lyrics Data” and Its Potential for Research Use

東条　佳奈
TOJO Kana

キーワード：歌詞コーパス，Ｊポップ，言語資源

要　旨

本稿は，株式会社シンクパワー制作の資料『歌詞コンテンツデータ集』の概要について，特徴，仕様，注意点を述べた上で，本データ集を利用した調査の一例を示しながら，研究利用の可能性について検討を行ったものである。データの選定・収録方法が不明瞭であることなど，いくつか扱い方に留意すべき点はあるものの，個人の研究では資料の収集手段が限られる歌詞情報を豊富に収録した本データ集の利便性は高いと思われる。

１．はじめに

本稿は，資料『歌詞コンテンツデータ集』の概要について紹介し，その研究利用の可能性について検討するものである。これまで，歌詞という言語資源を対象にした調査では，特定の歌手の曲を対象とするか，オリコンランキング等を利用したヒットソングを分析するか，という手法が主に行われてきたといえる。いずれにしても，調査者自身がデータを一から収集する必要があった。しかし，インターネット上の歌詞掲載サイトの多くは，著作権保護の都合上，コピー＆ペーストができない仕様となっており,テキストデータ化する場合には，手で打ち込む必要があった。そのため，大量に収集する場合，作業負担が大きくなるという状況にあったといえる。

株式会社シンクパワー制作の『歌詞コンテンツデータ集』(Version1.0）は，歌詞とその周辺情報が収録されたデータ集である。本稿では，本データ集がど

のようなものであるか,その概要についてまとめ,利用法について検討を行う。本データ集の利便性が高ければ，今後の歌詞研究に役立つものになると思われる。

２．利用規約について

『歌詞コンテンツデータ集』は,「曲名・歌手名・作詞者名・作曲者名・歌詞本体・発売年のデータ集」である。制作したシンクパワーより販売委託を受けている一般社団法人日本からだ・ぶんかストリートを通じて，本製品 CD-ROM を購入するという流れとなる。利用規約によれば，利用者は学術研究目的でのみ本データを利用することができ，利用の範囲は，申込者個人または申込者の属する課，もしくは研究室に限定されるものとされている。

また，利用者は本データを使用して得られた知見に関する研究発表あるいは成果の公表を行なうことができ，その際，発表論文には,「歌詞コンテンツデータ集」を使用したことを明記することと規定されている。また，成果の公開と同時に内容を書面により報告することが義務づけられている。使用許諾の範囲内において，データや処理プログラムを公開できることも明記されている。

なお，利用期間は 1 年間で，継続して使用するには毎年更新が必要である。

３．データ集の仕様

「『歌詞コンテンツデータ集』仕様書」では，文字コード，ファイル形式，データ項目についての記載がある。

文字コードは UTF-8，ファイル形式は csv である。データ項目は，lyricsId（歌詞 ID)，title（曲名)，artist（アーティスト名)，album（アルバム名)，writer（作詞者)，composer（作曲者)，uploadDate（投稿日)，releaseDate（発売日)，txt（歌詞情報）の 9 項目である。このうち，アルバム名，作詞者，作曲者，発売日情報は，Null（空白）のデータも含むものとなっている。これは，同じアルバムに収録されている曲は省略されるためであると思われる。なお，投稿日は,「当該データの本データ登録日」となっており，データを確認すると「2010/6/7」から「2020/12/31」の間に登録されたものとなっている。そのほか，歌詞情報には改行コードとして¥n が入力されている。

また注意書きとして,「CD 単位でデータを登録するため，同曲でも複数アル

バムに収録されている曲はアルバム違いとして重複登録されている場合がある」ことが記載されている。

４．収録されている歌詞コンテンツの詳細

４．１．収録曲について

データ項目をテキストエディタで確認すると，それぞれカンマで区切られており，歌詞 ID から歌詞情報までが一つの行の中に含まれている（図１）。本データは文字コードが UTF-8 のため，Excel で作業する際は，文字化けを防ぐための前処置が必要となる。筆者は Excel のデータタブ内の「テキストまたは csv から」のメニューからファイルを読み込むことで，タブ形式で表示できるようにした（図２）。

```
1  lyricsId,title,artist,album,writer,composer,uploadDate,releaseDate,txt↓
2  1,Amefrica,flex life,Japonica,青木 里枝,大倉 健,2010-06-07,2004-04-28,潤しておく
3  れ乾いた心を¥n愛に飢えてどしゃぶりさ僕は¥n¥n情けない僕のメランコリーに君はうなだ
4  れて¥n重なりあえないもどかしさにまいるわ¥n¥n言葉はいらないと昔から言うけれど¥nそ
5  んなの嫌いだわ 少しだけ苦手だわ¥nでも叫んでる¥n¥n潤しておくれ乾いた心を¥n愛に飢
6  えてどしゃぶりさ僕は¥n¥n情けない僕のメランコリーは きっとうまれた朝から¥n君に多
7  くを求め過ぎてまいるわ¥n¥n独りでも生きていく。偉そうに言うけれど¥nそんなの嫌いだ
8  わ 少しだけ勝手だわ¥nでも叫んでる¥n¥n潤しておくれ乾いた心を¥n夕立みたいにどしゃ
9  ぶりの愛で¥n¥nあぁ 音も立てず僕は泣けないよ¥n荒ぶるこの胸をお願いなだめて¥n¥n潤
10 しておくれ乾いた心を¥n夕立みたいにどしゃぶりの愛で¥n¥nゆるしておくれ身勝手な僕を
11 ¥n通いあうまで伝える愛を君に¥n↓
```

図１：テキストファイルでのデータ表示例

	A	B	C	D	E	F	G	H	I	J
1	lyricsId	title	artist	album	writer	composer	uploadDate	releaseDate	txt	
2	1	Amefrica	flex life	Japonica	青木 里枝	大倉 健	2010/6/7	2004/4/28	潤しておくれ乾いた心	
3	2	A LIE	EGO-WRAPPIN'	ベストラッ	Yoshie Na	Masaki M	2010/6/7	2008/10/15	You tried to appease	
4	3	BxKxRxxx	リア・ディゾン	Communic	白井 裕紀	リア・ディ	2010/6/7	2008/8/20	意味ないなんて 言わ	
5	4	B.I.O	BoA	VALENTI	Kenn Kato	Kosuke M	2010/6/7	2003/1/29	未来へと はやる心が	
6	5	Beladon'	Diggy-MO'	Diggyism	Diggy-MO	Diggy-MO	2010/6/7	2009/3/25	Since the 50's, nah e	
7	6	Bangalicious feat.土屋アンナ	ravex	TRAX	LISA	ravex	2010/6/7	2009/4/29	Jack yo body¥nJack J	
8	7	If I Never See Your Face Again	Maroon 5	It Won't Be	ジェイムス	ジェイムス	2010/6/7	2009/3/4	Now as the summer	
9	8	Der Rhein	DER ZIBET	アリ	ISSAY/三	DER ZIBET	2010/6/7	1996/3/23	酒の味 苦いだけ¥n苦	
10	9	EWIG...(warum?)	EXILE	HEART of	SOHJIN/D	SOHJIN/D	2010/6/7	2004/9/29	人は何故生きている	
11	10	FF	369	369.2	369/RYO/	369/RYO/	2010/6/7	2009/3/18	気持ちはいつでも前の	
12	11	I・N・G	MEGARYU	我流旋風	MEGARYU	MEGARYU	2010/6/7	2006/7/12	情熱を奮い起こせ 今	
13	12	JUVES	Diggy-MO'	Diggyism	Diggy-MO	Diggy-MO	2010/6/7	2009/3/25	Speculation 遊ぶ Lov	
14	13	JAP	abingdon boys s	JAP	Takanori N	Shibasaki	2010/6/7	2009/5/20	Inside Out ぶった斬れ	

図２：エクセルでのデータ表示例

図２を見ると，歌詞 ID は１番から連番になっており，連番の基準は概ね曲名

のアルファベット順（昇順）に並んでいるということがわかる。しかし，歌詞IDNo.7「If I Never See Your Face Again」の次が「Der Rhein」（IDNo.8）になっているなど，必ずしもそのアルファベット順には並んでいないものも散見される。英題ではない日本語の曲名はIDNo.42から始まり，五十音順にID番号が振られているが，「ありがとう!おともだち。」（IDNo.52），「愛する人の名前を日記に」（IDNo.53），「OUTGROW ～Ready butterfly～」（IDNo.54）と，「あり」→「あい」のように五十音順と一致しないものや，アルファベット順に並べるはずの英題が五十音順の方に含まれているなど，どのような順序で番号が付けられているのかが不明瞭な点もある。また，本データの登録日であるuploadDateを参照すると，2011年以降は，曲名の五十音順ではなく，アーティストごとにまとめて登録されているようであった。

歌詞IDのみを見ると，No.1からNo.2935397まであるが，途中連番が飛んでいるものも多くあり，収録曲数は431,544曲である。このなかには，英語曲など，日本語を全く歌詞に含まない曲も含まれる。英語歌詞の曲は，洋楽のカバーであるものもあれば，邦楽であるが英語歌詞というものもある。

また，前述のとおり，本データはCD単位でデータが登録されており，アルバム違いの重複登録の曲がある。歌詞（txt）欄の冒頭20字を対象に，重複確認を行ったところ，78,436曲に重複が見られた[1)]。

４．２．楽曲のリリース年について

本データ内の楽曲を収録したCDがいつ発売されたかは，releaseDate（発売日）の項目を参照する。releaseDate の内容を元に，発売年ごとの楽曲数をカウントしたものが表１である。表１を見ると，年代ごとにばらつきがあることが見て取れる。1965年の1曲はBob Dylanの“Like a Rolling Stone”である。また，曲数が微増する1985年の89曲のうち43曲はベストアルバムであり，同年に新規発表された楽曲となるともっと少ない。

毎年10,000曲以上,安定して収録されるようになるのは2006年以降である。一般社団法人日本レコード協会が掲載している統計情報「新譜数推移」[2)]によれば，2005年と2006年の邦盤（歌謡曲，軽音楽，アニメーション，その他）の発売状況に大きな開きはないため（2005年が9,427曲，2006年が9,774曲

である)，新譜数とは関係ないところで収録方法の基準が変わっているものと思われる。

表１：発売年ごとの曲数

発売年	曲数	発売年	曲数	発売年	曲数	発売年	曲数	発売年	曲数
1965	1	1990	2451	2000	5690	2010	13058	2020	12393
1983	1	1991	3554	2001	6186	2011	13057		
1984	2	1992	3092	2002	6278	2012	15490		
1985	89	1993	3511	2003	6346	2013	14605		
1986	309	1994	3946	2004	6305	2014	15367		
1987	467	1995	4471	2005	7880	2015	13326		
1988	795	1996	4043	2006	10859	2016	11805		
1989	1338	1997	5581	2007	17764	2017	12493		
		1998	6142	2008	10041	2018	11530		
		1999	5696	2009	11240	2019	12263		
～80年代計	3,002	90年代計	42,487	00年代計	88,589	10年代計	132,994	20年代計	12,393
発売年月日記載なし					152,079				
総計					**431,545**				

表１で挙げている発売年は，必ずしも新規に発表された楽曲のものだけではなく，一度シングルで発売された曲を収録したアルバムが新たに発売された場合には，アルバムの発売年に重複して曲数がカウントされる。そのため，発表年を重視する調査の場合には，releaseDate に掲載されている情報がどの CD の発売日のことなのかに注意しなければならない。

また，表１で示したように，発売日の記載がないものが 15 万曲に上る。発売日記載なしのものには，同じアルバムに収録されている曲のほか，ベストアルバムなど,楽曲単体の発売日とアルバムの発売日が異なるものなどが含まれる。ただし,前述の通り,同じアルバムに収録されている曲であっても,releaseDate を記載している曲もあるため，アルバム内の楽曲と対照させながら，精査していく必要がある。

なお，楽曲によっては，uploadDate（投稿日）の方が releaseDate（発売日）よりも早いものがある。データ内に既に登録済の楽曲が新譜のアルバムとして再発売された際にこのようなことが生じるようである。

５．歌詞データを用いた分析例

５．１．2020 年歌詞タイトルの調査

５節では，本データの利用法の検討として，試みにいくつか簡単な調査を行ってみたい。

伊藤（2017，2019）では，シンガーソングライターの松任谷由実と中島みゆきの楽曲の歌詞およびヒットランキングに上がる流行歌・Ｊポップに，日本語回帰現象が見られることが指摘されている。日本語回帰現象とは，外国語の使用が 1980 年頃から増加するものの，2000 年頃からは減少する傾向があるというものである。伊藤は，外国語が減少する一方で，和語が増加する傾向にあることを指摘している。

伊藤（2017）が行った日本語回帰現象を示す調査の一つに，タイトル語種構成比率の変遷というものがある。タイトル語種とは,「タイトル全体を 1 語のように見なして，語種判定の基準で判定したタイトルの語種構成」（伊藤 2017:6-7）であり，具体的には以下のように分類されるという。

(1) 和語タイトル :「また君に恋してる」（坂本冬美）
漢語タイトル :「少女飛行」（ぱすぽ☆）
外来語タイトル :「ヘビーローテーション」（AKB48）
混種語タイトル :「不自然なガール」（Perfume）
外国語タイトル :「I Wish For You」（EXILE）

（伊藤 2017:6-7 より抜粋）

加えて，外来語と外国語とを区別する基準として以下のものを挙げている。

(2) 外来語基準 : かなかローマ字で表記された外国出自の単語は外来語
外国語基準 : アルファベット表記された外国出自の単語は外国語

（伊藤 2017:7 より番号を変え抜粋）

このような基準により判定されたタイトルの語種構成は，混種語（1960 年）→和語（1970 年）→和語・混種語（1980 年）→和語・混種語・外国語（1990 年）→外国語（2000 年）→和語・外国語（2010 年）のように中心傾向が変化しているという。

そこで，2020 年のタイトルではどのような様相であるのかについて，『歌詞コンテンツデータ集』を用いて調査を行った。

伊藤（2017）では，ヒットランキングに載っている曲という指定を行っているが，『歌詞コンテンツデータ集』では売り上げ枚数の情報がないため，「日本語を全く含まない歌詞の曲を除く」，「2020 年以前に発表された曲を除く」という制限を加えた中から，random 関数を用いて 100 曲を選定するという方法で行った[3)]。

対象としたタイトルは 100 種のため，割合と実数は同じ内容となるが，以下のような結果となった。

表２：2020 年歌詞タイトル 100 種の語種構成

外国語タイトル	52	例）Trust In You（JUJU）
和　語タイトル	17	例）イノチノアカシ（アーバンギャルド）
混種語タイトル	15	例）白鍵と黒鍵のあいだで（TK from 凛として時雨）
漢　語タイトル	8	例）夜間飛行少年（クジラ夜の街）
外来語タイトル	6	例）ヘルレイザー（Creepy Nuts）
他	2	40000000%（なみへぇ），777（Ryohu）
計	100	

表２のように，大半を占めるのは「外国語タイトル」，次いで「和語タイトル」，拮抗して「混種語タイトル」という結果となった。検索範囲をヒットソングに限らなかったために，日本語回帰現象に沿わない結果になったとも考えられる。ヒットソングとランキング上位から漏れた曲とでの比較を行うと，また異なる結果が見えるかもしれない。また，その他として，数字（＋記号）のみのタイトルも見られた。表 2 に該当する 100 タイトルの中には入らなかったが，記号との組み合わせとなるタイトルも少なくない。

特に，曲名における「#」記号の使われ方は，近年で異なっているようである。#記号が用いられているタイトルは，重複を除くと 273 あるが，2010 年以前は「デッサン#1」(ポルノグラフィティ)，「自転車ラプソディ#1」(Something else)のように，数詞の前に用いる記号として専ら使われていたが，特に 2014 年以降は，「#拡散希望」((feat. 狐火）MC 小法師)「#インスタのストーリーで男とシーシャを吸ってる女」(TOFU)，「#スグ消シマス」(GOTCHAROCKA)，「#CD が売れ

ないこんな世の中じゃ（偽）」（ゴールデンボンバー）「#いいね!」（板野友美）など，ソーシャルメディアにおけるハッシュタグを意識した曲名が散見される。検索語を定めずとも実例の一覧を確認できるデータ集だからこそ，目に入る変化であると思われる。

５．２．句，文レベルの名称となるアーティスト名

前項の表２の例では，「(TK from) 凛として時雨」，「クジラ夜の街」など，句からなるアーティスト名を示した。バンド・グループの名称ではほかにも，「ぜんぶ君のせいだ。」「神様，僕は気づいてしまった」「ずっと真夜中でいいのに。」など，句レベルを超え，文レベルになるようなものもある。

こうしたアーティスト名にはどのようなものがあるのかについて，『歌詞コンテンツデータ集』を用いて調査を行った。本調査では，文レベルでの名づけがなされているアーティスト名を収集するため，artist の項目を対象に，助詞「は」，「が」を含むものについて検索を行った。その結果該当した 22 組のアーティスト名を，楽曲データの初収録年と共に表３に示した。

表３：文レベルの名称となるアーティスト名

それでも世界が続くなら（2013）／森は生きている（2013）／愛はズボーン（2015）／神様、僕は気づいてしまった（2017）／そして僕達は途方に暮れる（2018）／家の前でゴリラが死んでる（2018）／如何様詐欺師は夜うごく（2019）／原因は自分にある。（2019）／神はサイコロを振らない（2019）／ゆとりがない（2019）／男はくさいよ（2020）／斉藤正法は架空の人物です。（2020）／バカがミタカッタ世界。（2020）／パピプペポは難しい（2020）／月には兎がいる（2020）／どうして友達がいないのか。（2020）／それは、まさに肉球（2020）／生活は忘れて（2020）／オリーブがある（2020）／。明日はきっと雨（2020）／君の声を僕は知らない（2020）／テスラは泣かない。（2020）

もっとも早くて 2013 年，半数は 2020 年であり，従来はなかった長い名称のアーティスト名が出現してきているということがわかる。そしてこれらの語種構成を 5.1 節と同様の手順で見ると，和語名が 10，漢語と和語の混種語名が 8，

外来語と和語の混種語名が 3，その他が 1（「愛はズボーン」のズボーンが不明瞭なため[4]）となる。文単位となっているため，和語が増えるのは当然ともいえる。新たにデビューするアーティストが，和語を用いた文をグループ名として選定するのは，他のグループとの差別化ということもあるだろうが，もし，伊藤（2017，2019）が指摘している日本語回帰現象，J ポップの和風化の流れの新たな兆候だとすると興味深い。今回は文となっている名称のみに限ったが，今後は句や長大な複合語となっている名称と合わせて，アーティスト名の変化を観察してみたい。

５．３．個別のアーティストの歌詞を対象とした調査

５．３．１．Official 髭男 dism と米津玄師の比較

本節では，個別のアーティストの楽曲の歌詞を対象にした調査を行う。対象としたのは，レコチョクの「年間ランキング 2020 年シングル」の上位 50 位までのランキング[5]に最もランクインしていた「Official 髭男 dism」の 7 曲と，作詞作曲した楽曲提供を含む 5 種 6 曲がランクインしていた「米津玄師」である。

いずれも男性アーティストである。Official 髭男 dism（以下，ヒゲダン）は日本の 4 人組バンドであり，今回対象とした曲の作詞作曲はすべてボーカル・キーボードを担当する藤原聡が担当している。いっぽう，米津玄師（以下，米津）は，2009 年に「ハチ」名義でニコニコ動画へオリジナル曲を投稿しはじめ，2012 年以降は本名の米津玄師として自身がボーカルを務めながら，シンガーソングライターとして活動している（米津玄師 official site より）。

両者とも数々の賞を受賞しており，ヒゲダンは 2020 年には各音楽配信サイトにて計 22 冠を達成し[6]，米津は 2020 年にリリースしたアルバム「STRAY SHEEP」などが，2020 年の年間ランキング 50 冠を達成しており[7]，いずれも近年を代表する人気アーティストといえる。

本調査では，試みに，ヒゲダン・藤原の作詞作曲したヒットソング 7 曲「115 万キロのフィルム」（2018 年），「Pretender」，「イエスタデイ」，「宿命」（いずれも 2019 年），「I LOVE...」「HELLO」「Laughter」（いずれも 2020 年，先行リリースを含む）と，米津が作詞作曲したヒットソング 5 曲「Lemon」（2018 年），「パプリカ」（2018 年楽曲提供，本人歌唱版 2020 年），「馬と鹿」（2019 年），「まち

がいさがし」（2019 年楽曲提供，本人歌唱版 2020 年），「感電」（2020 年）の計 12 曲について，主に形態素解析した結果について観察を行った。本データの特性を生かすには大量の曲を調査する方がよいが，解析結果を目視で確認するため，今回は少数に留めた。

５．３．２．品詞別集計結果と誤解析について

本調査ではまず，「Web 茶まめ」の「現代語辞書」を用いて短単位での形態素解析を行い，結果をアーティスト別に集計した。品詞別の延べ語数集計結果が表４である。対象とした曲数が異なるため，割合の情報も併記した。なお，品詞名は Unidic の辞書のものをそのまま利用している。

表４：品詞別集計結果と割合

Official髭男dism（7曲）			米津玄師（5曲）		
品詞	集計	割合	品詞	集計	割合
助詞	776	30.6	助詞	426	30.6
名詞	571	22.5	名詞	294	21.2
動詞	386	15.2	動詞	257	18.5
助動詞	295	11.6	助動詞	187	13.5
記号	102	4	代名詞	78	5.6
代名詞	94	3.7	副詞	38	2.7
形容詞	91	3.6	形容詞	39	2.8
副詞	48	1.9	形状詞	25	1.8
接尾辞	46	1.8	接尾辞	21	1.5
形状詞	40	1.6	連体詞	15	1.1
補助記号	35	1.4	補助記号	4	0.3
連体詞	34	1.3	感動詞	3	0.2
感動詞	13	0.5	接続詞	2	0.1
接頭辞	7	0.3	接頭辞	1	0.1
接続詞	1	0	総計	1390	100
総計	2539	100			

両者共通して助詞が最も多く，名詞・動詞・助動詞と続き，またそれらの割合も近いという結果となった。ヒゲダンの方のみ「記号」があるが，これは歌詞中でアルファベット表記される外国語（I Love など）が解析できずに，それぞれ 1 字ずつ記号としてカウントされてしまったためである。米津の歌詞にはアルファベット表記の語がなかったため，こうした差が生じたと思われる。

伊藤（2019）は，流行歌・J ポップの歌詞には話し言葉性の高い語が多用されることを指摘している。Web 茶まめでは「現代語話し言葉辞書」も選択可能であるが,「現代語辞書」で解析した結果と比較したところ，結果はほとんど変わらず，むしろ「現代語話し言葉辞書」の語分割の方が多く誤解析が見られたため，本稿では「現代語辞書」を選択している。

新聞などの硬い文章に対して，歌詞は整然としていない文章となるが，今回対象とした楽曲の歌詞においては，日本語の語分割における明らかな誤解析は以下に挙げる数例のみであった。

(3) 自分にとっての正しさを創造してみるよ 大事にするよ（「Laughter」）

（誤）取っ手（名詞）

（正）とって（動詞-一般「取る」連用形＋助詞-接続助詞「て」）

(4) エンドロールなんてもん作りたくもないから（「115 万キロのフィルム」）

（誤）も（助詞-係助詞「も」）＋ん（感動詞-フィラー「んー」）

（正）もん（名詞-普通名詞-一般「物」）

(5) 夢ならばどれほどよかったでしょう（Lemon）

（誤）ほどよい（形容詞-一般「程良い」）

（正）ほど（助詞-副助詞「程」）＋よい（形容詞-非自立可能「良い」）

(6) かかと弾ませこの指とまれ

（誤）ともあれ（副詞）

（正）とまれ（動詞-一般「止まる」命令形）

本データ集では歌詞の行区切りの箇所に”¥n”が挿入されている。（3）については,「自分に」と「とっての」の間に改行が挟まれていたために誤解析が生じたのだと思われる。語分割のミスの場合，データ上でどのような表記になっていたかについても確認をしたほうがよいだろう。

なお，数詞とアルファベット表記の語は，1 文字ずつの過分割になってしまったり，単語の途中で分割されたり（「Beautiful」を「B」「eautiful」と分割する),一単語にまとめられていても,品詞情報が名詞になる(「Cry」「Forever」)など，ほとんど得たい結果での解析ができておらず，課題が残った。

伊藤（2017，2019）では，歌詞の語彙表を作成する際にはいわゆる「長い単位」での語分割が推奨されているため,今後,長単位での解析も試みてみたい。また，伊藤（2017）では，歌詞には，擬態語や擬声語が一語文として使用されることから,「擬音詞」「擬態詞」というべき品詞が必要であることが提唱されている。本調査においてはオノマトペはほとんど見られなかったが，使用されていた擬声語（「にゃんにゃんにゃん」「わんわんわん」(いずれも「感電」)）は解析がうまくできていなかったため，解析時の検討事項としておきたい。

５．３．３．語種の比較

前項では,ヒゲダンの歌詞にのみ英語が含まれていることを示した。そこで，解析結果の語種情報をもとに，集計を行った。その結果が表５である。集計欄の（ ）内の数字は異なり語数を示している。

表５：歌詞における語種と割合

Official髭男dism			米津玄師		
語種	集計	割合	品詞	集計	割合
和語	2036（502）	80.2	和語	1257（363）	90.4
漢語	268（170）	10.6	漢語	94（59）	6.8
記号	137（23）	5.4	外来語	22（13）	1.6
外来語	66（43）	2.6	混種語	12（9）	0.9
混種語	16（13）	0.6	記号	4（2）	0.3
固有名	5（1）	0.2	固有名	1（1）	0.1
語種情報なし	11（5）	0.4	総計	1390（447）	100
計	2539（757）	100			

表５のヒゲダン項目内の「語種情報なし」は，前述の「HELLO」などの英単語である。また，本来一単語となるはずの英語がアルファベット１文字ずつに分

割され，記号と判定されている。それらを除くと，？や鍵括弧，句読点などの補助記号 7 種となる。

両者を割合で比較すると，米津の方が漢語・外来語でわずかに少なく，その分，和語の割合が高いことがわかる。

今回は曲数・語数を揃えていないため，より差を明確にするためには，データを増やし，精査した上で検討する必要があるだろう。曲数を増やしていくことで，先行研究との比較も可能になると思われる。

５．３．４．頻度上位の自立語

本節では，具体的にどのような語が歌詞中に出現しているのかを述べる。形態素解析結果から，自立語，特に動詞，形容詞（イ形容詞），名詞，副詞，代名詞について，高頻度順に 10 種程度の語を抽出し，まとめたものが表 6 である。

総語数が少ないためそれぞれの語の頻度も少なく，このまま傾向を取り出すことは難しいが，一方には高頻度に出現するがもう一方には出現していない（低頻度の語）というものとして，ヒゲダンの歌詞における名詞「中（なか）」が挙げられる。また，人称代名詞においても両者には違いが見られた。

表６：自立語における頻度上位語の比較

	動詞	頻度	イ形容詞	頻度	名詞	頻度	副詞	頻度	代名詞	頻度
ヒゲダン	スル【する】	24	ナイ【ない】	37	ナカ【中】	12	モット【もっと】	6	ボク【僕】	27
	アル【ある】	8	イイ【いい】	9	ヒト【ひと（一）】	10	モウ【もう】	5	キミ【君】	27
	イル【いる】	8	ツヨイ【強い】	3	イマ【今】	9	ハルカ【遥か】	3	ソレ【それ】	11
	チガウ【違う】	7	アマイ【甘い】	3	ド【度】	8	トテモ【とても】	3	ナン【何】	8
	ワスレル【忘れる】	6	ウツクシイ【美しい】	3	ジブン【自分】	7	ドウ【どう】	3	イツ【いつ】	6
	シマウ【しまう】	5	ツライ【辛い】	2	コト【こと】	7	キット【きっと】	3	ココ【ここ】	4
	ナル【なる】	5	ワルイ【悪い】	2	セカイ【世界】	7	タッタ【たった】	2	ドレ【どれ】	4
	ウタウ【歌う】	5	ヨイ【よい】	2	ユメ【夢】	6	タダ【ただ】	2	コレ【これ】	3
	ススメル【進める】	5	アツイ【熱い】	2	ヤツ【やつ】	6	ソウ【そう】	2	ダレ【誰】	3
			イタイ【痛い】	2	ナン【何】	6	スコシ【少し】	2	ソリャ【そりゃ】	1
			クライ【暗い】	2	テ【手】	6				
			タダシイ【正しい】	2						
米津	イル【いる】	18	ナイ【ない】	10	コト【こと】	8	マダ【まだ】	7	アナタ【あなた】	16
	スル【する】	12	ヨイ【よい】	4	ハナ【花】	7	ドウ【どう】	5	ダレ【誰】	10
	オモウ【思う】	5	ツヨイ【強い】	3	ママ【まま】	6	キット【きっと】	4	ナン【何】	11
	キエル【消える】	5	フカイ【深い】	2	ヒ【日】	6	タダ【ただ】	4	ボク【僕】	8
	ノコル【残る】	4	ニガイ【苦い】	2	ユメ【夢】	5	モウ【もう】	3	キミ【君】	8
	サク【咲く】	4	クライ【昏い】	2	ココロ【心】	5	イマダ【未だ】	3	ワタシ【わたし】	6
	イク【行く】	4	ワルイ【悪い】	2	ムネ【胸】	5	マッスグ【真っ直ぐ】	2	ソレ【それ】	4
	アル【ある】	4	イイ【いい】	2	マチガイ【間違い】	4	タッタ【たった】	2	オマエ【お前】	3
	シル【知る】	4	アワイ【淡い】	2	テ【手】	4			コレ【これ】	3
	ハレル【晴れる】	4			モノ【もの】	4			ドッ【どっ（どこ）】	3
	フレル【触れる】	4			ヒト【一（ひと）】	4				
	ヨブ【呼ぶ】	4			イマ【今】	4				

「中」は，「I LOVE...」に7例，「イエスタデイ」に2例，「Laughter」に2例，「115万キロのフィルム」に1例と，4曲で用いられていた。突出して多い「I LOVE...」では，「僕が見つめる景色のその中に」「高まる愛の中」「変わる心情の中」「まるで水槽の中」「鼠色の街の中で」「重なる愛の中」「濁った感情の中」のように使われていた。「高まる愛の中」と「変わる心情の中」は連続したメロディで使われており，同じメロディラインの箇所で「重なる愛の中」と「濁った感情の中」が出てきているところから，意図的に「中」という語を使用しているものと思われる。同じメロディラインで，「中」の前後の語のみ変えているという点では，「115万キロのフィルム」も同様である。ここでは「ポケットの中で怯えたこの手はまだ忘れられないまま」という前半に出てくるフレーズが，後半で「ポケットの中で震えたこの手で今君を連れ出して」と，曲の中のストーリーに合わせて変化をしている。米津の歌詞では，「中」は3例，2曲で使用されていた。

人称代名詞については，米津の歌詞では「わたし」「僕」（一人称），「あなた」「君」「お前」（二人称）と5種類使用されているのに対し，ヒゲダンの歌詞では「僕」と「君」の2種類である。

ヒゲダンの歌詞における「僕」と「君」はいずれも27例であったが，「Laughter」を除く6曲すべてにこれらの人称代名詞の使用が見られた。やや偏った「君」の使用が見られた「Pretender」（9例）と「イエスタデイ」（7例）はいずれも恋愛をテーマとした曲のため，「君」への言及が多くなったものと思われる。

米津の歌詞中の人称代名詞で最も多かった「あなた」は，「Lemon」（11例）と「パプリカ」（5例）の2曲で用いられているという偏った結果となった。「パプリカ」では，「夏が来る 影が立つ あなたに会いたい」と「喜びを数えたら あなたでいっぱい」という同じメロディラインの箇所と，「心遊ばせあなたにとどけ」というフレーズの繰り返しが3回という内訳であった。また「パプリカ」には，一人称の使用はなかった。

「Lemon」では，今はそばにいなくなってしまった「あなた」に向かって語りかける内容の歌詞となっている。そのため，全編にわたり，「あなた」という二人称が用いられていると考えられる。なお，「Lemon」では，「あなた」と対になる一人称として「わたし」が6例用いられている。こうした，目の前に語りか

ける相手がいない状態で語りかける歌詞を，伊藤（2017:45）は「対他独白型」と呼び，歌詞の「語り」の文体の体系と構造についての分析を行っている。伊藤は主に「中島みゆき」と「松任谷由実」を分析の対象に取り上げているが，本調査での 12 曲というわずかな曲数でも人称の使用に差異が見られたため，アーティストごとにどのような傾向を取り出せるのかということも，分析の観点になりうると思われる。

以上，いくつかの観点から個別のアーティストの楽曲を取り上げたが，今回は調査対象を少数に絞ったため，いずれも試験的な調査となった。

本データ集には，最新に近い歌詞データが多く収録されている。そのため，本データ集を利用して再検討することで，従来先行研究で行われてきたこととの比較が可能になると思われる。

６．本データ集の使用における利点と注意すべき点

本稿では,『歌詞コンテンツデータ集』の特徴と仕様について述べた上で，本データ集を利用した調査例をいくつか示しながら，研究利用の可能性について検討を行った。最後にまとめとして，本データ集の利点と欠点（注意すべき点）について述べる。

まず，利点として，①収集・使用が容易になること，②豊富なデータを扱えること，③検索や分析の自由度が高いこと，などが挙げられる。

①の収集の容易さであるが，歌詞情報は著作権の関係で，コピー＆ペーストが出来ず，従来収集が難しかった。本データ集を用いることで，一から収集する手間が省けると思われる。また，データ集の規約の範囲で使用すれば，歌詞情報の利用許諾についても，個人で取る必要がないという点もメリットであると思われる。

②については，43 万件以上という豊富な歌詞データを対象とした研究が可能であるという点である。大量のデータを扱うことで，それだけ充実した分析が可能になると思われる。また，用例数のばらつきはあるものの，1980 年代からの歌詞データが収録されているため，変化を見ることも可能である。一気に大量の曲を対象に全文検索することが可能なウェブサイトもあるが，検索結果では歌詞の一部分しか閲覧することができず，また，結果を取り込む際は一つ一

つ手作業で保存しなければならない。そうした点で，本データ集は有用であると思われる。

③として，テキストデータとして読み込めば，正規表現などを用いた多様な検索も可能であり，その分，自由度が上がる。また，歌詞本文やタイトルを実例として一覧で確認できるため，検索語を定めずとも，分析が可能となる。データ量が多いため，統計を用いた分析も可能だろう。

一方で，注意すべき点や欠点もある。まず，それぞれの年で収録されている歌詞データの量にばらつきがあることや，選定基準が不明瞭であることが挙げられる。本データ集には邦楽だけでなく洋楽も多く含まれているが，どのような基準で収録が選定されているのかが分からなければ，データの代表性という点にも疑問が生じる。次に，再録曲も新曲と同じように収録されていること，発売年情報がない曲が大量にあることも注意すべき点である。重複の数も多いため，分析に応じた精査が必要になる。

最後に，データの量が大きすぎるという点がデメリットとして挙げられる。Biber, et al.（1998＝2005）では，コーパスが大規模なほど分析が充実するという一方で，データが大きくなればその分，作業負担が増え，手に負えなくなることが述べられている。こうしたジレンマを石井（2014）は，BCCWJ をはじめとしたコーパスの巨大化が抱える問題であるとして，「コーパスのパラドクス」というべき逆説的な事態であると指摘している。本データ集は，BCCWJ の「中納言」や「KOTONOHA」のように，ウェブ上に検索アプリケーションがあるものではなく，分析者のパソコン上での作業が必要である。そもそも，用例の閲覧や整理以前に，43 万行もの Excel ファイルは，開く・保存するといった作業だけでも処理に時間を要する。形態素解析にかけるにも，細かくファイル自体を分割しなければならず，分割作業中も処理落ちやスムーズに進まないなどの問題があった。こうしたデータでは当然，得られる用例も大量となるため，分析にかかる作業者への負担は大きくなる。そのため，巨大なデータを容易に扱うための工夫や手法が必要であると思われる。

以上のようなデメリットがあるものの，これまでまとまったデータ集がなかった歌詞情報を研究利用できるという点で，『歌詞コンテンツデータ集』は価値と可能性があると思われる。本稿ではごく一部を対象とした試験的調査しか行

わなかったが，今後，より対象を広げ，本データ集を利活用できるように努めたい。

注

１）本調査はExcelのCOUNTIF関数を用いて行った。完全一致である場合に重複と判定されるため，スペースの有無といった表記ゆれがあると，同じ曲であっても別の曲と判定されている可能性がある。また，冒頭 20 字が挨拶（「Hello, Hello……」）やハミング（「la la la la la……」）などが連続する曲の場合，異なる楽曲であっても重複した楽曲として判定されていることも考えられる。

２）2005年と2006年のものを参照。
https://www.riaj.or.jp/f/data/others/sp.html

３）本来であれば，ランキング情報を確認し，そこから調査を行うべきであるが，本稿では『歌詞コンテンツデータ集』を使用することを念頭においたため，このような手法を取った

４）“I was born”のもじりであると思われる。

５）集計期間は2019年12月1日～2020年11月30日，ダウンロード販売数を集計対象とした結果である（https://recochoku.jp/special/100891/）。

６）PONY CANYON NEWS「Official髭男dism，各音楽配信サイトなどにて2020年年間ランキング22冠を獲得！さらに，Yahoo!検索大賞2020ミュージシャン部門賞受賞！」
（2020年12月10日，https://news.ponycanyon.co.jp/2020/12/45043）。

７）米津玄師official site REISSUE RECORDS「「STRAY SHEEP」2020年 年間ランキング/受賞一覧」
（2021年5月14日,https://reissuerecords.net/2021/05/14/「stray-sheep」2020年-受賞年間ランキング一覧/）

参考文献

石井正彦（2014）「第1章　コーパスを用いた日本語研究の特徴―語彙・文法を中心に―　1.2 コーパスを用いた語彙研究の特徴」田野村忠温編（2014）

『講座日本語コーパス6　コーパスと日本語学』朝倉書店，pp.3-13
伊藤雅光（2017）『Jポップの日本語研究―創作型人工知能のために―』朝倉書店.
伊藤雅光（2019）「第7章　流行歌・Jポップの言葉―自己組織化現象としての日本語回帰―」田中牧郎編『シリーズ〈日本語の語彙〉7　現代の語彙―男女平等の時代―』朝倉書店，pp.84-97.
Biber, D., Conrad, S. & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use,* Cambridge University Press. （齋藤俊雄ほか訳『コーパス言語学―言語構造と用法の研究―』，南雲堂，2003）

参考URL

一般社団法人日本レコード協会「統計情報　新譜数推移」2005，2006.
〈https://www.riaj.or.jp/f/data/others/sp.html〉（2021年12月1日最終アクセス）
米津玄師 official site REISSUE RECORDS
〈https://reissuerecords.net/〉（2021年12月1日最終アクセス）
レコチョク「年間ランキング2020 シングル」
〈https://recochoku.jp/special/100891/〉(2021年12月1日最終アクセス)
Official髭男dism オフィシャルホームページ
〈https://higedan.com/〉（2021年12月1日最終アクセス）
Web茶まめ
〈https://chamame.ninjal.ac.jp/〉（2021年12月1日最終アクセス）

付記

本稿で用いた『歌詞コンテンツデータ集』は，大阪大学大学院文学研究科日本語学講座が株式会社シンクパワーおよび一般社団法人日本からだ・ぶんかストリートと交わした利用規定書に基づき使用したものである。また，本稿では，データ本体のほか，同梱の「仕様書」も資料として使用している。

（文学研究科助教）