駄洒落データベースの拡張及び分析

Extension and Analysis of Pun Database in Japanese

荒木 健治[1*] 佐山 公一[2] 内田 ゆず[3] 谷津 元樹[4]
Kenji Araki[1] Kohichi Sayama[2] Yuzu Uchida[3] Motoki Yatsu[4]

[1] 北海道大学 [2] 小樽商科大学 [3] 北海学園大学 [4] 青山学院大学
[1] Hokkaido University [2] Otaru University of Commerce [3] Hakkai-Gakuen University
[4] Aoyama Gakuin University

Abstract: スマートフォン上の対話エイジェントや AI スピーカーの普及に見られるように人間らしい振る舞いをする対話システム実現への欲求は年々急激に高まりつつある．しかし，現状で普及している対話システムは基本的にタスク指向型システムであり，話していて楽しいシステムとはなっていない．これは，非タスク指向型システム（雑談システム）の完成度が低いためである．我々は対話システムへ高度化したユーモア処理を導入することによりこの現状を打破することを目指している．そこで，ユーモア処理の高度化のための第一段階として 51,000 件の駄洒落を収録した駄洒落データベースの開発を行った．本稿では，この駄洒落データベースにさらに約２万件追加し，約７万件まで拡張した際の方法や問題点，及び拡張された駄洒落データベースを用いて行なった分析結果について述べる．

1 はじめに

少子高齢化社会の急激な進展に伴い，高齢者単身世帯の急激な増加が予想される．この際に最も重要となるのが，心のケアである．心のケアを行うためには，介護者などを雇用することが最善の方法であるが，昨今の政府の緊縮予算では困難である．そこで，この問題を解決するために人間と同等の対話能力を持つシステムの実現が喫緊の課題となっている．

このような背景のもと，現在スマートフォン上の対話エイジェントや AI スピーカーの急激な普及が進んでいる．しかし，現状で普及しているこれらの対話システムは基本的にタスク指向型システムであり，話していて楽しいシステムとはなっていない．これは，非タスク指向型システム（雑談システム）の完成度が低いためである．そこで，我々はこの問題を解決するために，これまで非タスク指向型の対話システムの研究を行ってきた [1, 2]．これらの研究では，一定の成果は得られているが，応答精度の低さや会話が盛り上がらない等，依然として多くの問題が残されている．

これらの問題を解決するために，対話システムが意味的に整合性のある応答を行うという面での精度の向上はもちろんであるが，仮に意味的に整合性のある精度の高い応答を行うことができたとしてもユーザが対話システムと楽しく対話ができ，会話が盛り上がるということにはならない．対話システムと楽しく対話を行えるようになるためには，人間並みにユーモアを理解し，生成できるシステムの実現が必須である．しかし，ユーモア処理の研究は少なく学問としてしっかりと確立されているとは言い難い．これは，日本文化におけるエンターテインメントに対する価値観の低さも一つの要因であると考えられる．欧米文化圏では，フォーマルな場でもユーモアを織り交ぜながら聴く人を飽きさせることなくスピーチすることが知的レベルの高さを示す一つの指標とされているのに対し，日本では内容を正確に伝えることに重きが置かれている．

このようなユーモア処理の研究の遅れを取り戻すために我々のユーモア処理のプロジェクトでは，駄洒落，ストーリージョーク，アメリカンジョーク，ワンライナー（１行ジョーク），皮肉（アイロニー），なぞなぞなどにおいて人間を超える面白いユーモアを工学的に実現することを最終目的としている．

これまでのユーモアに関する研究としては，なぞなぞの答えを駄洒落として生成するキムらの研究 [3] がある．キムらの手法では，手作業で小規模な辞書を作成し駄洒落を自動生成する実験を行い，既存の一般的な辞書と比較している．その結果，使用する辞書に性能が大きく依存することが確認され，大規模な駄洒落データベースを構築することの有用性が明らかにされている．

また，ことわざの文末を変更して意外性による笑い

*連絡先： 北海道大学大学院情報科学研究科
札幌市北区北１４条西９丁目
E-mail: araki@ist.hokudai.ac.jp

を狙った山根らの研究 [4] がある．山根らの研究では，ことわざを「すかし」と呼ばれるオチを利用することにより面白い文章を生成するシステムを提案している．「すかし」を用いたシステムが，一定のレベルの面白い文章を生成することは実証されているが，その他の面白さを生み出す手法に関する言及はなく，やはりユーモアを広範囲かつ大規模に収集し，データベースを構築する必要があるものと考えられる．

また，画像からユーモア発話を自動生成し，そのことにより対話継続欲求を向上させようとした二又らの研究 [5] がある．二又らの研究では対話システムにおいて画像からユーモアを生成することにより，対話が途切れた場合でもユーモアにより対話を継続したいという欲求が向上することを確認している．しかし，文献 [5] で実験に用いられているユーモア発話は 50 文のみであるため大規模なユーモア発話を用いて実用的な性能評価を行う必要がある．このためにもユーモアに関する大規模なデータベースの構築は必須のものとなる．

我々もこれまで駄洒落の自動生成に関する研究 [6]，対話システムに駄洒落生成システムを組み合わせたもの [7, 8]，物語性のあるテキストをユーモアとノンユーモアに自動的に分類するもの [9]，ストーリージョークに出現する単語間の意味的な類似度の評価により，話しの「落ち」を検出しユーモアの自動認識を行う手法 [10]，駄洒落における子音の音韻類似性及び SVM を用いた駄洒落検出手法の研究を行ってきた [11]．これらの研究を行う際に問題となったのが，評価を行うための標準的なデータセットが存在せず，他の手法と実験結果を正確に比較し評価することができないということである．また，認識・生成を行う際にユーモアの面白さに対する評価手法も確立されていない．

そこで，ユーモア処理の高度化のための第一段階として 51,000 件の駄洒落を収録した大規模な駄洒落データベースの開発を行った [12]．また，開発された駄洒落データベースを用いて駄洒落の面白さを認識する理解の手続きについての考察を行った [13]．さらに，駄洒落データベースを用いた駄洒落生成システムの開発を行い，どのようなお題に対しても約 80 %の精度で駄洒落を生成できることを確認した [14]．また，駄洒落データベースを用いて駄洒落中に含まれるオノマトペの分析を行った [15].

このような結果から大規模な駄洒落データベースを開発し，種々の研究を進めることはユーモア処理の研究を進める上で非常に有効であることが確認されている．しかし，文献 [15] で述べているようにオノマトペの分析を行うには駄洒落データベース中に含まれるオノマトペの量が少なく不十分である．また，文献 [14] で述べているように大規模な駄洒落データベースを駄洒落の自動生成に利用することの有効性は確認されたが，日常的に使用する語でも駄洒落データベースにお題として含まれていないものがある．そこで次の段階として，昨年度に開発した 51,000 件の駄洒落データベースを拡張することを試みた．駄洒落データベースは最終的には現在の約５万件の倍の量の１０万件まで拡張する予定であるが，本稿では，その途中段階として約２万件を追加し約７万件とした際の駄洒落データベースの拡張方法及び拡張された駄洒落データベースを用いて行なった分析結果について述べる．

2 駄洒落データベース拡張のための駄洒落収集方法

2.1 既存の駄洒落データベースにおける同義駄洒落の削除

前回駄洒落データベースの構築を行った際には，表 1 に示すように９つのサイトより合計 52,995 件収集した [12]．それらの駄洒落より字面上完全に一致している重複したものの削除を行った結果，51,872 件となった．ここから 51,000 件をランダムに選択することにより駄洒落データベースに収録する駄洒落を決定した．さらに，駄洒落データベースを詳細に調査した結果，駄洒落の種類，種表現，変形表現が同じで，句読点が一字入ったり，助詞が一語変わったりしただけで意味的にほとんど同じものがいくつか存在することが明らかとなった．そこで，駄洒落の種類，種表現，変形表現が同じで，その他の部分が一語だけ異なるものを 328 組抽出した．この例を以下に示す．この例のように助詞の「は」と「に」の違いがあるだけで意味的に変わりはないので，一方を削除する必要がある．

- 11173 (秋田) は [飽き た] 1[1]
- 32729 (秋田) に [飽き た] 1

しかし，以下の例のように一語異なる場合でも意味的に異なる場合が存在する．この場合には，異なる部分が「盗難」と「ずれ」で意味的に異なっているので，削除する必要はない．

- 15736 (帽子) の 盗難 [防止] 1
- 16046 (帽子) の ずれ [防止] 1

一語異なる場合に，意味的に異なっているかどうかは人手により判断する必要がある．そこで，第一著者が抽出された駄洒落の種類，種表現，変形表現が同じでその他の部分が一語だけ異なる駄洒落 328 組みに対して，意味的に異なるかどうかを判断し，意味的に同

[1] ( ) [ ] や駄洒落のタイプを表す文末の数字や文頭の通し番号については 3 章で説明する．

じもの（これを同義駄洒落と呼ぶ）の削除を行った．この結果，同義駄洒落のペアは 198 組となった．このようにして得られた 198 組の同義駄洒落のペアを用いて，同義駄洒落のペアのうち通し番号の大きいものの削除を行った．この結果 161 件の駄洒落が同義駄洒落として削除された．これは，198 組の中に同じ駄洒落とペアとなっていたものが 37 件あったためである．この結果，51,000 件の駄洒落データベースは 50,839 件に減少した．なお，同義駄洒落を削除した際に削除された同義駄洒落の通し番号は欠番とし，他の駄洒落の通し番号の変更は行わないものとした．

2.2 新たな駄洒落のスクレイピング方法および収集結果

本節では駄洒落データベースを拡張するために行なった新たな駄洒落収集方法について述べる．文献 [12] で駄洒落データベース構築のために収集した駄洒落はインターネット上に存在するものをスクレイピングすることにより行った．この際に収集を行ったサイトを表 1 に示す．この作業を行なったのは 2013 年 10 月であったので，前回収集を行なった時点より約 5 年間経過している．このため，まずこれらのサイトに最近 5 年間で新たに登録された駄洒落をスクレイピングすることにより新たな駄洒落の収集を試みた．新規に収集された駄洒落の件数（これを新収集件数と呼ぶ）を表 1 に示す．ここで新収集件数は，すでに駄洒落データベースに登録されている駄洒落と字面上で完全一致するものを除いたものである．また，一つのサイトで重複して字面上で完全一致する駄洒落が収集された場合には，同じ駄洒落が重複して登録されることを防ぐため収集を行わない．

2.1 で述べた同義駄洒落を除去した既存の駄洒落データベース 50,839 件に収録されていないものという条件で新たな駄洒落の収集を行った．表 1 に示すように，各サイトより前回駄洒落のスクレイピングを行なった 2013 年 10 月以降の約 5 年間に新たに登録された駄洒落の収集を行い，合計 53,099 件の駄洒落が得られた．

3 駄洒落データベースの拡張方法

本章では駄洒落データベースの拡張方法について述べる．駄洒落データベースの拡張におけるタグ付け作業は，クラウドソーシングを用いて行った．この際のタグ付け基準の統一方法などについて述べる．

3.1 駄洒落の種類について

駄洒落は音韻的に似ている二つの語で意味的に離れているものを 1 文中に存在させることによる意外性を用いてユーモアを表現するものである．したがって，駄洒落には音韻的に似ている 2 つの区間が存在し基となるフレーズを種表現，種表現より作成される音韻的に類似した区間を変形表現と呼ぶ．この種表現，変形表現の状態，有無により収集された駄洒落は表 2 に示す 4 種類に分類される [16]．

表 2 に示すように併置型駄洒落は，種表現と変形表現が明示的に文内に存在するものである．併置型のうち種表現と変形表現がひらがな表記で字面上完全に一致するものを Perfect と呼び，種表現と変形表現がひらがな表記で字面上完全には一致しないものを Imperfect と呼ぶ．重畳型は，種表現が文内に明示的に表現されていないもので，背景知識や文脈上に存在するものである．このため，重畳型では駄洒落中には変形表現しか存在しない．また，不明とは，駄洒落として解釈できないものである．

併置型 Perfect は，本来音的に同じものが存在する場合であるが，今回のタグ付け作業においては，ひらがな表記で字面上完全に一致するものとした．この理由は，音的に同じものでもひらがな表記にすると様々なものが存在し，タグ付けする際にこの部分で判断が分かれる可能性があるためである．そこで，ひらがな表記を近似的に音の表記ということにして同じかどうかの判断を行うこととした．以下に例を示す．

- 51085 (生姜 加え た) ので [少額 は 得 た] 2

この例の場合，種表現の「生姜 加え た」と「少額 は 得た」は音的には同じであるが，種表現のひらがな表記は「しょうが くわえ　た」であり，変形表現のひらがな表記は「しょうがく　は　えた」と異なるので，駄洒落種類は併置型 Imperfect を表す 2 となる．

3.2 駄洒落データベースのタグ付けについて

タグ付けは以下の 4 つの項目について行った．

- 種表現
- 変形表現
- 種表現，変形表現の対応付け
- 駄洒落種類

駄洒落データベースにおけるフォーマットを図 1 に示す．図 1 で駄洒落の種類については，表 2 に示す番号を駄洒落の最後部にスペースを一コマ入れて表示す

表 1: 駄洒落収集元一覧および新たに収集された駄洒落の件数

Web サイト名	URL	既存収集件数	新収集件数
ダジャレナビ	http://www.dajarenavi.net/pc/i_today_index.htm	39,120	44,605
Dajare Station	http://dajare.jp	8,795	6,308
ダジャレネット	http://www.dajare.net	1,621	345
ひとくちダジャレ大集合	http://www.biwa.ne.jp/~aki-ina/gyagu.htm	1,067	383
ダジャレ集ダジャレ事典	http://dajareshuu.web.fc2.com	982	239
ダジャレの缶詰	http://www.geocities.jp/pikumin_hiroba/dajare.html	572	63
駄洒落倶楽部	http://with2.net/dajakura	428	1,146
ダジャレ広場	http://www1.ocn.ne.jp/~origo/dazyare	303	0
駄洒落を言ったのは誰じゃ？	http://wtpage.info/dajare	107	10
合計	字面上で完全一致した重複除去前	52,995	
	字面上で完全一致した重複除去後	51,872	
	ランダムに選択	51,000	
	同義駄洒落除去後	50,839	
	新収集件数		53,099
	タグ付け件数		19,308
	拡張された駄洒落データベース		70,147
	同義駄洒落除去後		68,648

注）ダジャレ広場の新収集件数が 0 なのは，すでに HP が存在しなかったためである．

表 2: 駄洒落の種類と例

種類		説明	例
併置型	1.Perfect	種表現と変形表現がひらがな表記で字面上完全に一致しているもの	(大将) が [大賞] を 獲得
	2.Imperfect	種表現と変形表現がひらがな表記で字面上一致していないもの	(きちんと) 整理 さ れ た [キッチン]
3. 重畳型		種表現が背景知識，文脈上に存在し明示的には存在しないもの	[すい ま 千羽鶴]
4. 不明		駄洒落として解釈できないもの	「 あ 、 あれ 山 だ ！ 」

注） ( ) は種表現， [ ]:は変形表現を表す．

通し番号	S駄洒落本体 (種表現)N1 [変形表現]N2	S種類

注）S:空白1コマ, N1, N2:対応関係を示す数字

図 1: タグ付けのフォーマット

ることとした．また，先頭に通し番号を付与し，その後ろにスペースを 1 コマ入れて，駄洒落本体を表記する．この際，種表現は ( )，変形表現は [ ] で示した．また，複数の種表現，変形表現が存在する場合には，各記号の後ろに対応する駄洒落のペアに対して同じ数字を表記することにより対応関係を表現した．なお，元々駄洒落の中に存在した ( ) [ ] は，予め全て< >に変換を行うことにより識別している．このフォーマットを用いて，タグ付けされた駄洒落の例を以下に示す．

- 27227 議題 ！「 (カウボーイ)1 & (エイリアン)2 」 を [買う 暴威]1 & [営利 案]2 11

27227 番には駄洒落が 2 つ含まれている．一つは「カウボーイ，買う 暴威」であり，もう一つは「エイリアン，営利案」である．これらの対応関係を示すために，前者の駄洒落には，種表現，変形表現の後ろに空白を入れずに 1 を付与し，後者の駄洒落には，2 を付与する．また，最後に空白を入れて 2 つの駄洒落の種類を表す 11 を付与する．これは併置型 Perfect の駄洒落が 2 つ存在していることを示している．

また，駄洒落は事前に形態素解析ツール MeCab[17] を用いて，形態素解析を行なっている．変形表現は，通常の単語表現が変形していることが多く新出語となるため形態素解析として正しい分割を定めることが困難な場合が多い．このため，今回はタグ付けに重点を置き，形態素解析結果については，変更を行なっていない．駄洒落の形態素解析については，正しい分割方法を定めることも含めて，今後の課題である．

3.3 拡張後の同義駄洒落の削除について

新たに収集した約 5 万件の駄洒落のうち約 2 万件（19,308 件）の駄洒落に対してクラウドソーシングを利用してタグ付けを行い，駄洒落データベースへの追加登録を行った．この結果，拡張された駄洒落データベースの登録件数は 70,147 件となった．この際，字面上完全に一致する駄洒落は新規の駄洒落とはならないので登録されていないが，2.1 で述べたようにこの中には，駄洒落の種類，種表現，変形表現が同じでその他の部分が一語だけ異なるが意味的には同じである同義駄洒落が含まれていた．

そこで，拡張された駄洒落データベースに対しても 2.1 で述べた方法と同様の方法で同義駄洒落の削除を行った．その結果，駄洒落の種類，種表現，変形表現が同じで，その他の部分が一語だけ異なるものが 1,752 組存在した．次に，第一著者がこの 1,752 組を意味的に異なるものと同じものに分類した．この結果，意味的に異なるものが 224 組，同じものが 1,528 組となった．この 1,528 組の情報を用いて同義駄洒落のペアのうち登録番号が大きいものの削除を行った．この結果，表 1 に示すように当初 70,147 件登録されていたものが，1,499 件減少し，最終的な登録件数は，68,648 件となった．

4 クラウドソーシングによるタグ付け作業について

表 1 に示すように今回駄洒落データベースの拡張のために新たに収集した駄洒落は 53,099 件であったが，現在タグ付け作業が終了しているのは，19,308 件である．この 19,308 件の内訳は，表 1 の Dajare Station の全て 6,308 件，ダジャレナビの一部 13,000 件である．本章では，駄洒落データベースの拡張における約 2 万件の駄洒落を追加するためのクラウドソーシングによるタグ付け作業の詳細について述べる．

4.1 作業方法

タグ付け作業の手順を以下に示す．データは 1,000 件ずつに分割されているので，1000 件ごとに以下の手順で作業を進めた．

1. 2 名の作業者が 4.2 で述べるタグ付けの基準をもとに同一のデータに対しタグ付けを行う．
2. 2 名の加工結果をプログラムを用いて機械的に比較することにより相違点を両者に提示し相違点が 10 %以下になるまで双方が独立して修正を行う．
3. 相違点が 10 %以下になった段階で相違点に対して第一著者が正誤の判定を行う．
4. 2 名の作業者に相違点の正誤の判定結果とタグ付の完成版を渡す．

タグ付けの精度を高めるために同一データを 2 名で行い，その相違点を比較することにより確認を行うこととしたのは，以下に述べる理由によるものである．2 名の作業者はお互いに誰かを知らず，相違点を提示され同時に修正作業を進める．このため相手に相談することもできず，相手の修正方法も知らない．このことにより仮にお互いに相手のタグ付け方法に納得して変更すると相違点が減少しないということになってしまう．このような状況からタグ付け方法の基準に基づく

正確な判断が求められることになる．また，1,000 件の作業が終了するごとに残り約 10 %の相違点の正誤の判定結果とタグ付けされた完成版を渡し，次回以降の作業の参考にしていただいた．このことにより誤りの原因がフィードバックされ精度が次第に向上することが期待される．

しかし，この方法は 2 名の作業者のタグ付け結果が異なっている部分のみを見て確認を行っているため，2 名が同時に同じ間違い方をすると誤りが見逃されてしまうという問題がある．このような場合は，非常に少ないという仮定のもとに行った．この同じ間違い方をしたために見逃された誤りは 5 章で統計的な処理をする際にプログラムがエラーにより停止することにより発見された．最終的な結果を見てみると約 2 万件のタグ付けのうち 2 名が同時に同じ間違い方をしたために誤りが見逃された場合は 27 件と非常に少なく 0.1 %であった．このことは，この仮定が正しかったことを示している．

作業者の選定方法であるが，最初の 2 名は 4.2 で述べるタグ付けの基準の説明書と 50 件の駄洒落を渡し，それらを用いてタグ付けをしてもらい，その結果を見て精度の高いタグ付けを行った人にお願いするという方法で行ったが，3 名以降は作業者のプロフィールを見て適切と思われる人に直接依頼するという形で行った．2 名以上の作業者にお願いしたのは作業スピードを早めるために 3 組の作業者で並行して作業を進めたためである．なお，1,000 件のデータを一週間で行うように締め切りを設定した．

タグ付けを行った作業者の一覧を表 3 に示す．表 3 に示すように作業者の年代は 20 代から 60 代まで様々であり，性別は男性 3 名，女性 6 名である．9 名の作業者のうち 3 名は，クラウドソーシングの運営会社から優秀な作業者として認定されている．また，1,000 件を一度作業しただけで終了したのは，D と H の 2 名のみであった．これは，作業者の継続性の高さを示していると考えられる．

表 3: タグ付け作業者一覧

作業者名	年代	経歴	認定
A	60 代前半女性	会計・財務・経理	○
B	20 代前半女性	イラストレーター	
C	20 代前半男性	その他職種	
D	20 代後半男性	その他職種	○
E	20 代後半女性	その他専門職	
F	30 代後半男性	クリエイティブディレクター	
G	30 代後半女性	マーケティング	○
H	20 代後半女性	その他職種	
I	40 代後半女性	その他職種	

相違点数の推移を表 4 に示す．1,000 件ごとに最終的な相違点の正誤の判定結果と完成版のタグ付け結果を渡しフィードバックを行っているため，同一ペアでタグ付け作業を行うと相違点数が次第に減少すると予想していたが，実際の相違点数の推移を見ると可能な限り同一ペアでタグ付け作業を行っているにも関わらず，相違点数が減少するということはなくデータごとにバラツキが見られる．これは，同一ペアで行う作業量がまだ少ないためと考えられ，今後同一ペアで行うデータ量が増えるにつれて減少するものと考えられる．

4.2 タグ付けの基準について

次にタグ付けの基準について述べる．以下で述べる内容を箇条書きにして作業者に渡している．作業内容としては以下の 4 点である．駄洒落の種類については，表 2 に示す 4 種類に分類してもらうこととした．

1. 種表現の範囲を記入
2. 変形表現の範囲を記入
3. 種表現，変形表現の対応関係の記入
4. 駄洒落の種類を記入

次に作業上の注意点について述べる．駄洒落は予め単語ごとにスペースで分割されているが，この分割は変更しないものとした．これは，3.1 で述べたように形態素解析は MeCab を用いて行っているため誤りが存在する．特に駄洒落における変形表現は，未知語が多くその形態素解析を正しく行うことは，非常に困難である．このため，今回は誤りを含むことを前提とした上でこのまま変更せず，形態素の区切り部分で種表現．変形表現の範囲を決定することとした．駄洒落の正しい形態素解析については，正しい形態素の基準を作成することも含めて今後の課題である．

一方，そもそも形態素解析を行わずにタグ付けを行うべきであるという考えもあるが，そうなるとタグ付けする箇所が文字単位となり，その候補が非常に多くなってしまい，相違点が多くなるという問題がある．このため今回はこのような方法で行った．また，番号で駄洒落の種類を書く際には，その前に半角スペースを１コマ入れることや記号，数字，スペースの入力はすべて半角で行うととした．

3.2 で述べたように１件の駄洒落中に複数の種表現と変形表現のペア（これを種・変形表現ペアと呼ぶ）が複数存在する場合には，対応する種・変形表現ペアの後ろに番号を付与することとし，各々の駄洒落の種類を文末にスペースを一コマ入れて，種表現の出現順に書くこととした．以下に例を示す．

表 4: タグ付けの相違点数の推移

データ名		作業者名		相違点数			誤り数		
		作業者１	作業者２	１回目	２回目	３回目	作業者１	作業者２	両者
Dajare Station	ds1	A	B	298	62		22	39	1
	ds2	A	B	356	169	26	14	6	1
	ds3	A	B	252	91		22	61	8
	ds4	A	B	253	107		46	58	3
	ds5	A	B	337	159	91	60	28	3
	ds6	A	B	275	147	71	31	36	4
	ds7[注]	A	B	192	78		36	39	3
ダジャレナビ	dv1	C	D	536	170	87	78	7	2
	dv2	C	E	211	72		35	39	1
	dv3	C	F	536	36		13	16	7
	dv4	C	E	285	180	128	51	71	6
	dv5	C	E	224	116		43	73	0
	dv6	C	E	168	66		31	33	2
	dv7	G	H	189	98		31	66	1
	dv8	G	B	187	96		25	70	1
	dv9	G	B	231	123		19	102	2
	dv10	G	B	161	66		33	31	2
	dv11	G	B	152	42		17	23	2
	dv12	F	I	255	123		78	45	0
	dv13	F	I	198	67		36	31	0
合計				5,296	2,068	403	721	874	47
				7,767			1,642		

注）ds7 はデータ数が 308 件であったため 1,000 件相当に換算して表示している．

- 48445 (格闘 家)1 の (帽子)2 で [核 投下]1 を [防止]2 11

種表現を取り出す範囲についてであるが，単独で意味がわかる範囲でできるだけ短く抽出することとした．これは，種表現を短くすることにより同じ種表現から生成される駄洒落を一括して捉えるためである．また，変形表現との対応関係を考え，変形表現の方で対応するものが単語単位に分割されていない場合には，種表現の方も分割せず一語にするものとした．これは，種表現の変化により変形表現がどのように変化するかということを捉える必要があるためである．以下に例を示す．

- 46989 (妻子 用)1 と (父子 用)2 、 [再 使用]1 と [不 使用]2 11

変形表現を取り出す範囲についても，単独で意味がわかる範囲でできるだけ短く抽出するものとした．以下に例を示す．

- 4 (高菜) 、 [あっ た か な] ？ 1

この場合，種表現は「高菜」なので，種表現のひらがな表現に対応する部分は「た　か　な」になる．しかし,「た　か　な」では意味がわからないので,「あった か な」を変形表現として抽出する．

また，種表現，変形表現が重複している場合については，以下に挙げる例のように種表現，変形表現に複数の対応する番号を付与することや種表現，変形表現の範囲を重複させるという処理により対応することとした．

- 42579 (姉妹)12 の [獅子舞]1 は もう [おしまい]2 11
- 24223 ((猫)1 カフェ)2 で 、 [ニャンと]1 [猫 が 屁]2 22
- 21624 (タラ)1 が [[足ら ん]1 ち ゅら]2 13

駄洒落の種類について，併置型の Perfect については，ひらがな表記で同じものとしたが，以下のような場合，音的には同じだがひらがな表記では異なってしまう．3.1 で述べたように音的に同じということで判断が分かれ統一できない可能性があるので，ひらがな表

記を近似的に音の読みということにして判断するものとした．以下に例を示す．

- 51087 (妹 さん) 、 [芋 落とさ ん] ？ 2
- 13413 (シュー) が しぼん で く ！ ！ [プシュ 〜〜〜 !] 2

また，長音の「ー」の処理については，ひらがな表記としては異なるが音としては同じものの代表的なものとして，例外的に同じものとした．以下に例を示す．

- 19813 (オーボエ) の [応募へ] 行く 1

種表現，変形表現中に”？ 、「 」 。・”などが存在する場合については，ひらがな表記にした場合に完全一致しないので併置型 Imperfect となる．したがって，この場合は駄洒落の種類は 2 となる．なお，＜ ＞の場合には，読みを表すので駄洒落の種類は 1 となる．この例を以下に示す．

- 7 (コートジボアール) に は 、 [コート 地 、 倍 ある] 2
- 36266 (アンデス) メロン の 中 は [餡 < あん > です] ！ 1

種表現や変形表現を切り出す際に助詞から始まるフレーズを認めるかどうかということについては，そのように切り出しても意味がわかるかどうかということに依存する．これは，意味がわかるという基準の他にできるだけ短くするという基準もあるためである．以下に挙げる例のように「と ランプ」と言う表現があれば，「○○とランプ」と言うように容易に推測できるということからこのような表現でも抽出できるものとした．

- 51192 (トランプ) [と ランプ] 1

種表現と変形表現の位置関係については，種表現は変形表現より前にあるものとしている．このようにした理由は駄洒落は音で聞くものなので，先に聞いた単語から変形表現を連想するからである．以下に例を示す．

- 41786 (度重なる) [旅] ！ 2

語尾を含めるかどうかについてであるが，語尾については一連のフレーズである場合には，含めるものとした．一連のフレーズであるかどうかは判断が分かれる部分であり，駄洒落の場合，駄洒落を面白くさせるための語調，言い方という問題もある．このようなことを総合的に考えて語尾を含めるかどうかを決定するのであるが，この部分がタグ付けの不一致に繋がった可能性があるので今後検討する必要がある．以下にこの例を示す．

- 53029 (痛く し ない) から 、 [委託 し ない で ね] ！ 1
- 1541 [太ら ぬ タヌキ の 母さん よー] ！ 3

この例で「委託 し ない で ね !」という表現の種表現に対応する最小単位は「委託 し ない」であるが，一連のフレーズを考えると「委託 し ない で ね」ということになる．「！」を含めない理由は，意味がわかる範囲でできるだけ短く取るという基準からである．

固有名詞については，原則として各単語ごとに分割せず一つの種表現，変形表現とするものとした．これは固有名詞は，一つの単位として意味を持つものでそれを分割すると意味が変わるかわからなくなってしまうためである．以下に例を示す．

- 56037 (菅 直人) は [敏感 な 男 だ] 。 1

接頭辞についてであるが，接頭辞は無くても意味がわかるので，できるだけ短くという観点から含めないものとした．以下に例を挙げる．

- 56814 お (兄 が 立っ た) 場所 は [新潟] 。 2

このようにタグ付けの基準については，詳細な点について説明し，作業者に渡しているが，それでも不明な点がある場合には，第一著者に質問をしてもらうこととした．不明なものの番号を書いておき，作業結果を報告する時にまとめて質問するということでも良いということを伝え作業を開始していただいた．

4.3 タグ付けの誤りについて

4.3.1 作業者別の誤り数について

表 5: 作業者別の誤り率

作業者	作業件数	誤り数 [個]	誤り率 [%]
A	6,308	252	4.0
B	10,308	521	5.1
C	6,000	269	4.5
D	1,000	9	0.9
E	4,000	225	5.6
F	3,000	137	4.6
G	5,000	133	2.7
H	1,000	67	6.7
I	2,000	76	3.8

表 4 にタグ付けの相違点数の推移を示す．表 4 で同一ペアでも相違点数が次第に減少しなかった理由は，4.2 で述べたタグ付けの基準の中で種表現，変形表現を「意味がわかる最小の単位」としたために「意味がわかる」

という部分で主観が入り込む余地が多かったためと考えられる．このような誤りは，データ数が増えると学習効果により次第に減少するものと思われるが，今後は基準を作成する際に可能な限り主観が入り込むことの無いような基準を作成する必要がある．

次に，作業者別の誤り率を表 5 に示す．表 5 で誤り数とは，表 4 に示す最終的な約 10 %の相違点に対して第一著者が正誤の判定結果で各作業者の誤り件数に両者の誤り件数を加算したものである．作業件数については，1,000 件から 10,308 件と作業件数に差があるものの誤り率については，おおよそ 4 %から 6 %となっていてヒューマンエラーの範囲内に収まっているものと考えられる．D については，0.9 %と非常に低い値になっているが，作業件数が 1,000 件のみだったため作業件数が数千件になった場合にどのようになるかは不明である．

4.3.2 駄洒落の種類の誤りについて

本稿では，文献 [12] で述べられている 51,000 件の駄洒落データベースを「基本」，今回約 7 万件まで拡張した駄洒落データベースを「拡張」，駄洒落データベースを拡張するために使用した約 2 万件の駄洒落を「拡張部」と呼ぶ．

表 6 に駄洒落の種類の誤り数を示す．表 6 で「１ー＞２」とは種類１を種類２にした誤りであり,「２ー＞１」は種類２を種類１にした誤りである.「基本」と「拡張部」で比較すると 91.1 %は「基本」での誤りであった．これは,「基本」を作成する際には 3 名の作業者が相談しながら行ったため駄洒落の種類１（併置型 Perfect）と種類２（併置型 Imperfect）を区別する規則を明確に作らなくても問題となる駄洒落が出現するごとに相談して決めて行くことができたが，逆にこのことがタグ付けに揺れが生じる原因になったものと考えられる．これに対して,「拡張部」を処理する際にはクラウドソーシングで 9 名の作業者が行ったため明確に文章で規則を記述する必要が生じ，タグ付けに曖昧さが生じなくなったものと考えられる．

また,「基本」を作成した際には，クロスチェックを行い，さらに意見が別れる場合には 3 人目の作業者が決定することとしたが，この作業は全て手作業であったため誤りの見落としが発生したものと考えられる．しかし,「拡張部」のタグ付けは 2 名の作業者が同時に同じデータをタグ付けし，1,000 件終了するたびに機械的に異なるものを検出し，それを両者に提示し修正作業を行うという形にしたために双方のタグ付けの相違点が両者にフィードバックされ，統一が図られたと考えられる．また，1,000 件の 10 %である 100 件以下の相違点については，第一著者が一人で正誤を決定したこともタグ付けの統一に有利に働いたものと考えられる．

次に誤りのパターンを見ると 66.9 %が種類２を種類１と間違えたものであった．これは 4.2 に述べたように駄洒落の種類を決定する詳細な規則が明示的に整理されておらず，作業者間で相談しながら進めたことにより「基本」おいてこの部分の意思統一が作業者間で不十分であったためである．

このように「基本」において行った 3 名の作業者によるタグ付け作業よりクラウドソーシングを用いて行った 9 名の作業者によるタグ付け作業の方が，精度の高いタグ付けを行う上で有効であることが確認されたので今後の作業もクラウドソーシングを用いて行う予定である．

5 拡張された駄洒落データベースの分析

本章では，拡張された駄洒落データベースの総語彙数，種表現語彙数などの統計的な分析結果について文献 [12] で述べた「基本」での分析結果との比較を行う．また，拡張された駄洒落データベースにおいて種表現，変形表現で頻出するものについても「基本」と比較しながら考察する．

5.1 統計的な分析結果の比較

表 7 に示すように併置型の駄洒落が占める割合が 96.3 %から 98.8 %に増加し，さらに併置型が多くなっていることが確認された．また，併置型の Perfect と Imperfect の割合は 47.8 %，48.5 %から 48.2 %，48.5 %となり，Perfect の割合が少々上昇し，Imperfect と同じ割合になっている．その分，不明の割合が少なくなっている．これはスクレイピングの精度が向上したためと考えられる．

また，2.1 でも述べたように「基本」の駄洒落件数は 51,000 件であるが，1 件の駄洒落の中に複数の種・変形表現ペアが含まれている場合があるので，総分類数は 51,685 個となっている．51,000 件の駄洒落の中には人間が読んでも理解できない駄洒落ではないもの (種類４の不明）も 1,102 件存在するので，これを引くと実際に駄洒落の含まれる件数は，49,898 件となる．また，総分類数は 51,685 個の中には 1,102 件の駄洒落ではないものも含まれるので，実際の種・変形表現ペアの総数は 50,583 個となる．したがって，1 件あたりの平均種・変形表現ペア数は，50,583/49,898=1.01 個となる．これは，1 件の駄洒落中に約 1 件の種・変形表現ペアが存在するということになる．これに対して「拡張」の駄洒落件数は 68,648 件であるが，総種・変形表現ペア数は 70,168 個となっている．68,648 件の駄洒落の中には種

表 6: 駄洒落種類の誤り数

タイプ	基本		拡張部		タイプ別合計	
	個数	割合 [%]	個数	割合 [%]	個数	割合 [%]
1ー＞2	135	31.4	21	50.0	156	33.1
2ー＞1	295	68.6	21	50.0	316	66.9
基本・拡張部別合計	430	91.1	42	8.9	472	100.0

注）種類 1 は併置型 Perfect, 種類 2 は併置型 Imperfect を示す.

表 7: 駄洒落の収録数及び割合の比較

種類		基本		拡張	
		数 [個]	割合 [%]	数 [個]	割合 [%]
併置型	1.Perfect	24,718	47.8	33,786	48.2
	2.Imperfect	25,081	48.5	34,049	48.5
	合計	49,799	96.3	67,835	98.8
3. 重畳型		784	1.5	1,103	1.6
4. 不明		1,102	2.1	1,230	1.8
総分類数 [個]		51,685		70,168	
総種・変形表現ペア数 [個]		50,583		68,938	
総駄洒落件数 [件]		51,000		68,648	
不明を除く総駄洒落件数 [件]		49,898		67,418	
総単語数 [語]		455,276		630,303	
総文字数 [文字]		1,499,160		2,066,309	
平均種・変形表現ペア数 [個]		1.01		1.02	
平均単語数 [語]		8.93		9.18	
平均文字数 [文字]		29.40		30.10	
平均単語長 [文字]		3.29		3.28	

表 8: 種・変形表現ペアの出現回数別頻度の比較

出現回数	基本			拡張		
	頻度	総出現数	占有率 [%]	頻度	総出現数	占有率 [%]
1	38,904	38,904	76.91	37,933	37,933	55.02
2	3,713	7,426	14.68	11,196	22,392	32.48
3	768	2,304	4.55	1,659	4,977	7.22
4	227	908	1.80	482	1,928	2.80
5	85	425	0.84	153	765	1.11
6	48	288	0.57	67	402	0.58
7	20	140	0.28	28	196	0.28
8	14	112	0.22	14	112	0.16
9	5	45	0.09	9	81	0.12
10	2	20	0.04	7	70	0.10
11	1	11	0.02	3	33	0.05
12	0	0	0.00	3	36	0.05
13	0	0	0.00	1	13	0.02
総種・変形表現ペア数		50,583	100.00		68,938	100.00
種表現語彙数	43,010			50,546		
変形表現語彙数	43,787			51,555		

類４（不明）が 1,230 件存在するので，これを引くと実際に駄洒落の含まれる件数は，67,418 件となる．また，総分類数は 70,168 個であるが，この中には 1,230 件の種類４（不明）も含まれるので，実際の種・変形表現ペアの総数は 68,938 個となる．したがって，1 件あたりの平均種・変形表現ペア数は，68,938/67,418=1.02 個となり,「基本」に比べて少々増加し，複雑な駄洒落が増えていることがわかる．

また,「基本」の総単語数は 455,276 語，総文字数は 1,499,160 文字である．平均単語数は 8.93 文字，平均文字数が 29.4 文字であることから，1 件の駄洒落は約 9 単語から構成され，平均単語長が 3.29 文字であることから平均 3 文字程度の単語から構成されていることがわかる．これに対して,「拡張」の総単語数は 630,303 語，総文字数は 2,066,309 文字である．平均単語数は 9.18 文字，平均文字数が 30.10 文字であることから，1 件の駄洒落を構成する単語数が 0.25 ポイント増加し，長い駄洒落が多くなっていることがわかる．また，平均単語長が 3.28 文字と「基本」の場合とほぼ同じである．このことは，駄洒落１件あたりに含まれる種・変形表現ペア数が増加し，長い駄洒落が多くなっていることとも合致するが，駄洒落を構成する単語の長さは変わっていないので，長い単語を使うことが増えたのではなく，これまでと同じ 3 文字程度の単語を使い，複数の駄洒落を含むことにより文長が長くなっていることがわかる．

5.2 種・変形表現ペアの出現回数および頻出表現の比較

表 8 に「基本」と「拡張」の種・変形表現ペアの出現回数別の頻度の比較を示す．また，図 2 にそのグラフを示す．表 8 で占有率は，出現回数を総種・変形表現ペア数で割ったものである．総種・変形表現ペア数は，表 7 に示す 50,583 個，68,938 個と一致している．これは，種類別に合計したものと種表現，変形表現ごとに集計したものが一致したことを示し，少なくともすべての駄洒落について，その記号による種表現，変形表現の指定と駄洒落の種類の指定が行われたことを示している．

また，表 8 から「基本」の場合には，１回から４回出現するもので，97.94 %を占めていて，ほぼ全てのものが４回以下しか出現していないことがわかる．また,「拡張」も 1 回から 4 回出現するもので 97.52 %を占めていて同様の傾向を示している．しかし，異なる部分として１回出現するものが 76.91 %から 55.02 %と 21.89 ポイントも低下し，この分が２回，3 回出現するものの増加分となっている．

また，種表現の語彙数が 43,010 語から 50,546 語に増加している．駄洒落 19,308 件を加えても種表現としての語彙数は，7,536 語しか増加していない．これは，39.0 %の増加に留まっている．すなわち，残りの約 6 割の約 12,500 件はすでに使用されていた種表現（お題）であるということである．このことが，2 回，3 回出現するものの増加に繋がったと考えられる．このことは，インターネット上に存在する駄洒落の種表現のバリエーションが次第に収束しつつあることを示している．

この傾向は駄洒落データベースの網羅率を上げ，どのようなお題に対しても駄洒落を生成できるシステムを開発するという観点からは良いことではない．この傾向は駄洒落データベースをさらに拡張するにつれ強まるものと考えられ，今後拡張する際に種表現の語彙数を増やす新たな方法を考案する必要がある．また，変形表現の語彙数は 43,787 語から 51,555 語に増加し，7,768 語増加している．変形表現についても今回新規に追加した駄洒落 19,308 件の追加に対して 39.6 %の増加であり，種表現と同様に変形表現の語彙数を増やす新たな方法を考案する必要がある．

また,「基本」では種表現の存在しない重畳型の変形表現の種類は 777 件であるが，表 7 で重畳型の駄洒落が 784 個あったことから，重畳型で複数回出現したものは 7 件であることがわかる．これに対して「拡張」では，重畳型の変形表現の種類は 1,009 件であるが表 7 で重畳型の駄洒落が 1,103 個あったことから，複数回出現したものが 94 件と急増している．これは,「基本」において重畳型が，文脈や背景知識に種表現が存在することから，明らかにわかるものでなくてはならず，そのことが複数回の出現を妨げているものと考えていたが,「拡張」において長い駄洒落が増えたことから重畳型でも以下のような一部分を利用したものが増え，複数箇所で利用できるようになったことが原因と考えられる．

- 64328 今夜 の お供 は 、 [天才 バー ボン] ！ 3

表 9 に種・変形表現ペアの出現頻度ランキングの比較として「基本」,「拡張」それぞれの上位１０組を示す.「基本」,「拡張」共に最大頻度のペアは「ドイツーどいつ」である．２位以下は順位の変更はあるが，どれも短い単語である．これは短い単語である程，駄洒落を思い付きやすいためであると考えられる．この傾向は「基本」と「拡張」で変わりがない結果となった．

次に，一つの種表現からどれだけ多様な変形表現を作成できるのかについて「基本」と「拡張」での変化を調査するために種表現から作成される変形表現の種類数のランキングの比較を行った．表 10 に上位 10 位までを示す．また，占有率は，駄洒落の種類数を変形表現語彙数で割ったものである．変形表現語彙数は表 8 に示すように「基本」は 43,787 種,「拡張」は 51,555 種であった．

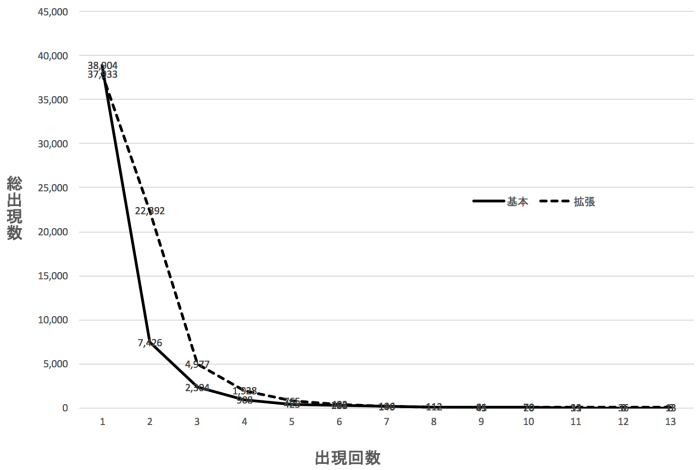

図 2: 種・変形表現ペアの出現回数別頻度の推移の比較

表 9: 種・変形表現ペアの出現頻度ランキングの比較

順位	基本			拡張		
	種表現	変形表現	頻度	種表現	変形表現	頻度
1	ドイツ	どいつ	11	ドイツ	どいつ	13
2	帽子	防止	10	韓国	勧告	12
3	理科	理解	10	仏陀	破った	12
4	秋田	飽き た	9	タクシー	わたくしー	12
5	ケーキ	景気	9	ケーキ	景気	11
6	医院	いい ー ん	9	帽子	防止	11
7	鮭	酒	9	彼	カレー	11
8	朝食	超 ショック	9	秋田	飽きた	10
9	胃	いい	8	カバ	カバー	10
10	韓国	勧告	8	会長	快調	10

注) 基本については頻度 8 回が 14 組あるのでそのうち 2 組を示す.
拡張については頻度 10 回が 7 組あるのでそのうち 3 組を示す.

表 10: 一つの種表現から生成される変形表現種類数のランキングの比較

順位	基本			拡張		
	種表現	変形表現種類数	占有率 [%]	種表現	変形表現種類数	占有率 [%]
1	ブドウ	35	0.080	虎	86	0.170
2	サメ	34	0.078	患者	80	0.158
3	手紙	32	0.073	貝	79	0.156
4	魚	31	0.071	甲斐	75	0.148
5	レタス	30	0.069	ブリ	75	0.148
6	イカ	30	0.069	手紙	68	0.138
7	猫	30	0.069	タイ	66	0.130
8	メジャー	30	0.069	バッタ	64	0.127
9	仙台	29	0.066	バンダナ	64	0.127
10	内容	29	0.066	ブドウ	62	0.123

表 11: 一つの種表現から生成される変形表現種類数上位３位までの変形表現の比較

基本
順位：1，種表現：ブドウ，変形表現種類数：35, 占有率：0.069 %
運ぶ どー 2：内部 どう 1：忍ぶ どー 1：浮かぶ どー 1：呼ぶ どー 1：飛ぶ どー 1：思う存分 どう 1： き ょほぉおぉ 〜〜〜 1：半分 どう 1：武道館 1：１ 粒 ー どう 1：区分 どう 1：危 どう ござい ます 1： 叫ぶ どー 1： 結ぶ どー 1：一 つぶ どう 1：取り分 どう 1：遊ぶ どー 1：論文 どう 1：株 どう 1：粒 どう 1： 当分 どう 1：並ぶ どー 1：気分 どう 1：呼ぶ どー 1：今時分 どう 1：文 どう 1：処分 どう 1：渋 どう 1： 全部 どう 1：学ぶ どー 1：成分 どう 1：選ぶ どー 1：説明 文 どう 1：一 粒 どう 1
順位：2，種表現：サメ，変形表現種類数：34, 占有率：0.067 %
覚め た 3：ジョーズ 2：サメザメ 2：慰める 2：冷め た 2：氷雨 1：は しゃ ー ぐ 1：３ 名 1：６ シャーク 1： サメー 1：ジャーズ 1：ジョーンズ 1：慰め 1：霧雨 1：冷める 1：秋雨 1：シャーックリ 1：杓子定規 1： 冷め た 顔 し てる ね 1：お 酌 1：さめ て いる 1：サメー <寒い> 1：余裕 しゃ ー く しゃく 1：小雨 1： さぁ 〜 めった に 見 ない 1：シャーク に さわる 1：納め 1：興ざめ 1：浅 め 1：泳ぎ が ジョーズ 1： こしゃ ー く な 1：見納め 1：青ざめ た 1：寝覚め 1
順位：3，種表現：手紙，変形表現種類数：32, 占有率：0.063 %
入れ た ー 5：言いそびれ た ー 1：あふれ た ー 1：遅れ た ー ！ <遅 レター> 1：忘れ た ー <レター> 1： イラストレーター 1：折れ た ー 1：破れ た ー 1：しら ばっ くれ た ー 1：埋もれ た ー 1： 汚れ た ー 1： 疲れ た ー 1：たてがみ 1：やぶれ た ー 1：手 が 見える 1：触れ た ー 1：読ま れ た ー 1：盗ま れ た ー 1： 慣れ た ー 1：れ た - 1：渡さ れ た ー 1：別れ た ー 1：くれ た ー 1：抜か れ た ー 1：あきれ た ー 1： 遅れ た ー 1：はがれ た ー 1：潰れ た ー 1：忘れ た ー 1：ナレーター 1：流れ た ー 1：濡れ た ー 1
拡張
順位：1，種表現：虎，変形表現種類数：86, 占有率：0.170 %
トラップ 6:トラウマ 5:トライ 3:見 たい が ー 3:大河 3:通ら ん 3:捕らえる 2:捕らえ た 2:捕ら ぬ 2:鳴い とら 2: 捕らえろ 2:取ら 2:とらわれる 2:渡来 2:ガ タイ が ー 2:捕らわれ た 2:取ら れる 2:なり たい が ー 2:たいがい 2: トラブル 2:トランク 2:トラベル 2:ドラマ 1:退学 1:鯛 が 好き 1:タイ が 好き 1:い tiger 1:とら れる 1: Try 1:タイ が 1:痛い が ー 1:対外 1:捕らえ たい が ー 1:冷たい が ー 1:逃げ たい が ー 1:反撃 し たい が ー 1: 鯛 が 1:対岸 1:取ら れ た 1:気 を とら れる 1:トラック 1:トラブル メーカー 1:トランクス 1:トランプ 1: トラフィック 1:トラウト 1:吠え とら 1:トラディショナル だ 1:怒っ とら 1:捕まっ とら 1:みとら ん 1 見 とら 1:酔っ とら 1
順位：2，種表現：患者，変形表現種類数：80, 占有率：0.158 %
直感 じゃ 3:時間 じゃ 3:お かん じゃ 3:鈍感 じゃ 2:送還 じゃ 2:アメリカン じゃ 2:噛ん じゃ ダメ 2:勘 じゃ 2: かんかん じゃ 2:噛ん じゃっ た 2:観 じゃ 2:感じ や 2:快感 じゃ 2:缶 じゃ 2:血管 じゃ2:予感 じゃ 2:間 じゃ 2: 敏感 じゃ 2:悪寒 じゃ 2: 器官 じゃ 2:違和感 じゃ 2:共感 じゃ 2:反感 じゃ 2:期間 じゃ 2:ご 帰還 じゃ 1: あっけらかん じゃ 1:痴漢 じゃ 1:噛ん じゃ お 1:帰還 じゃ 1:交換 じゃ 1:脂肪 肝 じゃ 1:気管 じゃ 1: 危機 感 じゃ 1:厭世 観 じゃ 1:価値 観 じゃ 1:創刊 じゃ 1:週刊 じゃ 1:月刊 じゃ 1:朝刊 じゃ 1:夕刊 じゃ 1: 日刊 じゃ 1:好 かん じゃ 1:一環 じゃ 1:盛ん じゃ 1:阿鼻叫喚 じゃ 1:習慣 じゃ 1:実感 じゃ 1:空間 じゃ 1: 圧巻 じゃ 1:感謝 1:玄関 じゃ 2:一 週間 じゃ 1
順位：3，種表現：貝，変形表現種類数：79, 占有率：0.156 %
絵画 4:買い ます 3:かい 3:買い 3:会 3:かい ー 3:買い付け 2:食べ がい 2:下位 2:奇奇怪怪 2:買い控え 2:海岸 2: 甲斐 2:可愛い 2:開会 式 2:かいがいしい 2:回 2:海底 2:食 べ たかい 1:か いっ 1:買い取る 1:会話 1:３ 回 1: もう １ 回 1:です かい 1:買い物 1:かい ーーー っ 1:描い た 1:かい ぃ 〜 の 1:概念 1: かい かい 1:買い たい 1:買占め 1:買いあさる 1:買いかぶる 1:買い 損 1:買い まし た 1:飼い なさい 1:飼い犬 1: 書い て 1:描い て 1:欠い て 1:掻い て 1:快挙 1:開眼 1:快感 1:快調 1:解釈 1:会議 1:回収 1:解決 1: 怪死 1:解凍 1:回転 1

注) 変形表現の後ろの数字は出現頻度を表す.

表 12: 一つの種表現から生成される種・変形表現ペア数のランキング

順位	基本			拡張		
	種表現	種・変形表現ペア数	占有率 [%]	種表現	種・変形表現ペア数	占有率 [%]
1	メジャー	51	0.101	貝	54	0.105
2	マスター	46	0.091	虎	53	0.103
3	魚	45	0.089	患者	52	0.101
4	サメ	40	0.079	甲斐	51	0.099
5	イカ	40	0.079	ブドウ	43	0.083
6	時計	39	0.077	ブリ	43	0.083
7	カッター	38	0.075	鳥	42	0.081
8	火星	37	0.073	タイ	39	0.076
9	内容	37	0.073	トロ	37	0.072
10	ブドウ	36	0.071	栗	37	0.072

注)「拡張」については種・変形表現ペア数 37 個のものが 3 つあるのでそのうちの 2 つを示す.

表 10 で「基本」で最も変形表現の種類数が多かった種表現「ブドウ」は表 11 から 35 種であるが,「拡張」で最も変形表現の種類数が多かった種表現「虎」は 86 種であり，約 2.5 倍増加している．また，占有率も 0.069 %から 0.170 %に約 2.5 倍増加している．このように「拡張」においては，上位の種表現で変形表現のバリエーションが大幅に増加している．この傾向は，２位，３位の「患者」,「貝」についても同様の傾向となっている.

また，表 11 に変形表現の種類数が多い種表現上位３位までの具体的な変形表現の比較を示す.「基本」と「変形」で変形表現種類数が最大の「ブドウ」と「虎」を比較すると,「ブドウ」が語尾の「どう，どー」などで駄洒落を作成しているのに対し,「虎」では「トラップ,トラウマ」などの自立語の一部として使用する一方で「見 たい が ー」のように英訳の「タイガー」を語尾に使うなど多様な方法で使用されている．このことが約 2.5 倍の増加に繋がったものと考えられる.

次に，一つの種表現がどれだけ多くの種・変形表現のペアを作成できるかを調査するために一つの種表現から作成される種・変形表現ペア数のランキングを行った．表 12 に上位 10 位までを示す．また，占有率は，種・変形表現ペア数を総種・変形表現ペア数で割ったものである．種・変形表現ペアの総数は表 8 に示すように「基本」は 50,585「拡張」は 68,938 であった．表 12 で,「基本」は上位５位までのうち 4 語がカタカナ語であるのに対し「拡張」ではカタカナ語が１語のみに留まっている.「拡張」ではカタカナ語に変わって「貝，虎」などの漢字語で短い読みの語が多くなっている．これは，表 10 に示す種類数でも同様の傾向がある．カタカナ語でも漢字語でも比較的読みが短い語は，駄洒落の変形表現を作成する際の自由度が増し，比較的容易に駄洒落を思い付くことができるためと考えられる.

6 おわりに

本稿では，まず始めに現状で普及しているスマートフォン上の対話エイジェントや AI スピーカーなどの対話システムが非タスク指向型システムとしては不十分であることを述べ，その性能を人間並みにするためにはユーモア処理の高度化が重要であることを述べた．その第一段階として昨年 51,000 件という大規模な駄洒落データベースの構築を行ったが，駄洒落中に含まれるオノマトペの分析や駄洒落の自動生成に関する研究では，まだ規模が不十分であることが明らかとなった.

そこで，駄洒落データベースにさらに約２万件追加し約７万件まで拡張を行った．本稿では，拡張を行う際に用いたクラウドソーシングによるタグ付け作業が,昨年行った３名の作業者を雇用して行った方法より精度の点において高くなることを述べた．また，クラウドソーシングでタグ付けを行う際の基準についても詳細に述べ，作業者間での相違点数や誤り率についても述べた．また，駄洒落の種類の誤りの分析結果についても述べた.

次に，構築した駄洒落データベースの収録語数，種表現の語彙数などの統計的な分析を昨年度構築した約５万件の「基本」と約７万件の「拡張」を比較しながら行った．この結果，約６割がすでに使用されていた種表現（お題）であり，新たな駄洒落収集方法の導入が必要であることが明らかとなった.

また，拡張された駄洒落データベースに頻出する種表現と変形表現の分析を行った結果,「基本」の際にカタカナ語が上位を占めていたのに対して,「拡張」では漢字語が上位を占めていたが，いずれも読みとしては短い語であり，表記上の相違に関係なく短い読みの語が駄洒落を思い付きやすいことが確認された.

今後の課題としては，以下のような点が挙げられる．現在，駄洒落種類，種表現，変形表現が同じで一語だ

け異なり意味的に同じものは同義駄洒落として削除しているが，そもそも非常に類似した駄洒落でも種表現，変形表現を取り出す範囲が微妙に異なっている場合には，同義駄洒落として検出されないので非常に類似した駄洒落が十分に除去ができていない可能性がある．この点については，今後詳細に調査する予定である．また，拡張された駄洒落データベースの分析結果からネット上をスクレイピングすることにより駄洒落を収集するという方法では，新たな種表現の駄洒落を収集することが困難になってくるものと考えられる．そのため，駄洒落データベースに登録されていない種表現（お題）に対して新たな駄洒落の収集方法を考える必要がある．

謝辞

本研究は科研費（基盤研究 (C)17K00294）の助成を受けたものである．

参考文献

[1] Rafal Rzepka, Shinsuke Higuchi, Michal Ptaszynski, Pawel Dybala and Kenji Araki: When Your Users Are Not Serious, 人工知能学会論文誌, Vol. 25, No. 1, pp.114-121, 2010.

[2] Arnaud JORDAN and Kenji ARAKI: Real-time Language-Independent Algorithm for Dialogue Agents, 知能と情報（日本知能情報ファジィ学会誌）， Vol.28, No.1，pp.535-555, 2016.

[3] キム・ビンステッド，滝澤修：日本語駄洒落なぞなぞ生成システム「BOKE」, 人工知能学会誌， Vol.13, No.6, pp.920-927, 1998.

[4] 山根宏彰，萩原将文：笑いを生むことわざすかしの自動生成システム，知能と情報 (日本知能情報ファジィ学会誌)， Vol.24, No.2, pp.671-679, 2012.

[5] 二又 航介，藤倉 将平，菊池 英明：対話システムにおける画像からのユーモア発話の自動生成とそれによる対話継続欲求の向上，言語処理学会 第 24 回年次大会 発表論文集, pp.520-523, 2018.

[6] Jonas Sjobergh and Kenji Araki: Robots Make Things Funnier, New Frontiers in Artificial Intelligence, Lecture Notes in Artificial Intelligence, Vol.5447, pp.306-313, 2009.

[7] Pawel Dybala, Rafal Rzepka and Kenji Araki: Humor Prevails! – Implementing a Joke Generator into a Conversational System, Lecture Notes in Artificial Intelligence, Vol. 5360, pp.214-225, 2008.

[8] 谷津　元樹，荒木 健治: 話題遷移に適応した駄洒落ユーモア統合型対話システムの性能評価, 人工知能学会第 2 種研究会 ことば工学研究会資料,SIG-LSE-B601-3, pp.23-27, 2016.

[9] 天谷祐介，荒木健治，ジェプカ・ラファウ: テキストの面白さの評価によるユーモアの認識, 第２８回ファジィシステムシンポジウム (FSS2012) 講演論文集，pp.233-238, 2012.

[10] 天谷 祐介, ジェプカ ラファウ, 荒木 健治: 単語間類似度を用いた物語ユーモア認識手法の性能評価, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B301-10, pp.63-69, 2013.

[11] 谷津元樹，荒木健治：子音の音韻類似性及び SVM を用いた駄洒落検出手法, 知能と情報 (日本知能情報ファジィ学会誌), Vol.28, No.5, pp.833-844, 2016.

[12] 荒木健治，内田ゆず，佐山公一，谷津元樹：駄洒落データベースの構築及び分析, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B702-3, pp.13-24, 2017.

[13] 佐山公一，荒木健治：コンピュータが駄洒落で笑わせる？—駄洒落の面白さを認識する理解の手続きー, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B702-4, pp.25-32, 2017.

[14] 荒木健治：駄洒落データベースを用いた駄洒落生成システムの性能評価, 人工知能学会第 2 種研究会 ことば工学研究会資料, SIG-LSE-B703-8, pp.39-48, 2018.

[15] 内田ゆず，荒木健治：駄洒落に含まれるオノマトペの特徴分析, 言語処理学会第２３回年次大会発表論文集， pp.741-744, 2017.

[16] 滝澤修：記述された「併置型駄洒落」の音素上の性質，自然言語処理，Vol.2, No.2, pp.3-22, 1995.

[17] Taku Kudo, Kaoru Yamamoto, Yuji Matsumoto: Applying Conditional Random Fields to Japanese Morphological Analysis, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004), pp.230-237, 2004.