駄洒落データベースの構築及び分析

Construction and Analysis of Pun Database in Japanese

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

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

Abstract: 高齢者単身世帯の急増に伴う話し相手の需要の急増，ユーモアの医療への応用を考える場合，対話システムの精度の向上により人間らしい振る舞いをするシステムの開発が必須となる．このためユーモア処理の需要は今後急速に高まると考えられるが，ユーモアの面白さを評価する手法及び標準的なデータセットが確立されていないのが現状である．そこで，標準的なデータセット開発の第一段階として駄洒落データベースの構築を行った．本稿では，駄洒落データベース構築方法，構築過程での種々の問題とその解決方法，分析結果等について述べる．

1 はじめに

日本社会は，少子高齢化と独身若年層の急増により今後これまで人類が経験したことのないような高齢者単身世帯が急増する事態に直面する．高齢者単身世帯で重要となるのが，高齢者の孤独感を緩和することである．すなわち，心のケアが最重要課題となる．心のケアのためには高齢者単身世帯を訪問する人を多く雇用する必要がある．また，ユーモアにより発する笑いは医療的な効果が科学的に実証されている．しかし，このためには膨大な経費が必要となり，昨今の医療，福祉予算の急増と政府の緊縮財政を考えると現実的ではない．

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 [1, 2]．しかし，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，対話システムのさらなる精度の向上はもちろんであるが，より人間らしい振る舞いのできる対話システムを開発することが必須である．人間の対話と対話システムの対話との違いは種々あるが，ユーモアを認識したり，生成したりすることが人間並にできないことが大きな要因の一つである．しかし，ユーモアを自動的に認識・生成するという工学的な面での研究は少なく，依然として学問としてしっかりと確立されているとは言い難い．特に，その大きな要因は，実験結果を定量的に評価するための手法が確立されていないということである．

本研究の最終的な目的は，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），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である．そのためには評価用の標準データセットの開発およびユーモアの面白さを評価するための評価手法の開発が必要となる．

これまでのユーモアに関する研究としては，日本語を対象としたものとして，なぞなぞの答えを駄洒落として生成するもの [3]，ことわざの文末を変更して意外性による笑いを狙ったもの [4] 等がある．また，英語を対象とした研究としては，One-liner というジョークを対象とした研究 [5, 6] がある．

我々はこれらの研究に対して，これまで駄洒落の自動生成に関する研究 [7]，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 [8, 9]，物語性のあるテキストをユーモアかノンユーモアかに分類するもの [10] や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の開発を行ってきた [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存在しない．

現状では，各々の実験の都度，複数の被験者が主観的な評価を行っているが，評価者は少数であり，他研究との性能面での正確な比較は難しい．このような現状はユーモアの研究を進める際に大きな障害となる．そこで，本稿では，実験の際に用いる標準的なデータセッ

*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

トを定め，ユーモア全般の面白さに関する評価手法を確立し，ユーモアの研究を行う基盤を整備するための第一歩として，標準的な駄洒落データベースの構築および分析を行った結果について述べる．

2 駄洒落の収集方法

本章では駄洒落の収集方法について述べる．駄洒落データベースに収録する駄洒落はインターネット上に存在するものをクローリングすることにより行った．収集を行なったサイトを表１に示す [12].

表 1 に示すように９つのサイトより合計 52,995 件の駄洒落の収集を行った．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 件の駄洒落を収集した．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．なお，文献 [12] でも同様のデータを用いているが，文献 [12] の表 2 では重複文除去後の件数が 45,970 件となっている．これは，字面上完全に一致する駄洒落以外に表層上の表現が極めて近い駄洒落も重複として削除したためである．今回は，文献 [12] で用いた表層表現が極めて近い駄洒落という基準が曖昧であったため字面上完全に一致したもののみを削除することとした．

3 駄洒落データベースの構築方法

本章では駄洒落データベースの構築方法について述べる．駄洒落データベースの構築は，3 名の作業者に報酬を支払うことにより行った．作業者は，2 名が 20 代の理系男子大学生であり，1 名が 20 代の理系男子大学院生である．

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り意外性によるユーモアを表現するものである．

駄洒落には，音韻的に似ている２つの区間が存在する．駄洒落の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 [13].

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．

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 3 つの項目について行った.

- 種表現
- 変形表現
- 駄洒落種別

これらの 3 つの項目について前述した作業者 3 名が，各自 17,000 件の駄洒落についてタグ付けを行う．次に，作業者と異なる作業者が，クロスチェックを行う．クロスチェックで意見が分かれたものについては，3 人目の作業者がどちらかを判断する．さらに 3 人目の作業者も意見が異なる場合には，作業者全員で討論を行い，決定する．また，この作業とは別に，週に一度第一著者を含めて 2 時間程度のミーティングを行い，タグ付けで意見が分かれているもの，難しい例などを報告することによりタグ付け作業の統一性を図った．この作業は約 2ヵ月におよび，各自平均 110 時間，合計約 330 時間の時間を要した.

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ることとした．また，通し番号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 [ ] で囲んだ．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，重畳型の場合には，変形表現しかないため，変形表現のみに数字を付与した．なお，元々駄洒落の中に存在した ( ) [ ] は，全て< >に変換を行うことにより識別している.

番号	S駄洒落本体 (種表現)N1 [変形表現]N2	S種別

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

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

表 1: 駄洒落収集元一覧

Web サイト名	URL	収集件数
ダジャレナビ	http://www.dajarenavi.net/pc/i_today_index.htm	39,120
Dajare Station	http://dajare.jp	8,795
ダジャレネット	http://www.dajare.net	1,621
ひとくちダジャレ大集合	http://www.biwa.ne.jp/~aki-ina/gyagu.htm	1,067
ダジャレ集ダジャレ事典	http://dajareshuu.web.fc2.com	982
ダジャレの缶詰	http://www.geocities.jp/pikumin_hiroba/dajare.html	572
駄洒落倶楽部	http://with2.net/dajakura	428
ダジャレ広場	http://www1.ocn.ne.jp/~origo/dazyare	303
駄洒落を言ったのは誰じゃ？	http://wtpage.info/dajare	107
合計	字面上で一致した重複除去前	52,995
合計	字面上で一致した重複除去後	51,872

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.

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,「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 を書き，後者の駄洒落には，2 を書いている．また，最後に空白を入れて 2 つの駄洒落の種別を表す 11 を書いている．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[14] 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いため，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．

また，変形表現として取り出す範囲であるが，変形表現としての長さをできるだけ短くとるが，種表現が対応しているからといって変形表現だけで意味をなさないものは，意味がわかる範囲まで長く取ることにした．例えば，以下のような例である．

- 4 (高菜) 、 [あっ た か な] ？ 1

この例では，変形表現を「た　か　な」とすることも考えられるが,「た　か　な」は，単独では意味がわからず,「あっ　た　か　な」　の一部であるので，この場合は,「あっ　た　か　な」を変形表現とした. このような方針で変形表現のタグ付けを行なった．

4　特徴的な駄洒落について

本章では，駄洒落データベースに含まれる特徴的な駄洒落をいくつか取り上げ，駄洒落の構成や作成方法について考察する．特徴的な駄洒落を表 3 に示す．以下では，表 3 の例についてそれぞれ説明する．

- 1　(坊っちゃん) が [ぼっ ちゃんと] 水 に 飛び 込む 1

1 番は，通常の例で種表現が「坊ちゃん」で変形表現が「ぼっちゃんと」である．この場合，種表現と同じ読みが変形表現に含まれているので，駄洒落の種類としては併置型の Perfect である 1 を末尾に付与している．

- 539 (老化)12 し てる 人 と [廊下]1 を [走ろ う か]2 11

539 番は，種表現「老化」に対して，変形表現が「廊下」と「走ろうか」の 2 つある．このため対応関係を明らかにするために種表現，変形表現それぞれの後ろに対応する数字を付与している．つまり，この例は以下のような二つの駄洒落が存在し，二組の種表現，変形表現を含んでいるということになる．

- 種表現：老化，変形表現：廊下
- 種表現：老化，変形表現：走ろうか
- 1634　「 (ふふふ ふふふ ふ ふふふ) 」 で [豆腐 <ふ が 10 こ>] 2

1634 番は，種表現「ふふふ　ふふふ　ふ　ふふふ」に対して，変形表現が「豆腐」であるが，これは「ふ」が 10 個あるためである．この説明が「豆腐」の後ろに「ふが 10 こ」と書いてある．駄洒落は本来音で聞くものであるが，1634 番の例は音だけでは駄洒落を理解することが困難で文字で書いて，さらに説明を加えることにより駄洒落と理解される例である．

次に 3 つの駄洒落が含まれている例について述べる．

- 4360 (朝食)1 抜き で [超 ショック]1 、 (昼食)2 抜き で [中 ショック]2 、 (晩食)3 抜き で [一番 ショック]3 222

4360 番は，以下に示す 3 つの駄洒落が含まれている例である．

- 種表現：朝食，変形表現：超ショック
- 種表現：昼食，変形表現：中ショック
- 種表現：晩食，変形表現：一番ショック

次に二つの異なる種類の駄洒落が含まれている例について述べる．

- 6289 「 邪魔 な (ドイツ)12 人 は [どい つ]1 だ ？ 」 「 全員 [邪魔 <ん>]2 です 。 」 12

6289 番には，2 つの駄洒落が含まれている．種表現が「ドイツ」であり，変形表現が「どいつ」と「邪魔<ん>」なので，前者は併置型 Perfect であり，後者は併置型 Imperfect である．そこで，駄洒落の種類として，変形表現の出現順に駄洒落の種類を表す 1,2 というタグ付けを行っている．

- 7180 (刑事 [だけ) 遺児 抱け] 。 意地 で も 1

7180 番は，種表現と変形表現が重複して出現している例である.「刑事だけ遺児抱け」という文字列は,「刑事だけ」と「け遺児抱け」の駄洒落である.「刑事，遺児，意地」なども考えられるが，最長のものを取ると前者の解釈となる．これに形態素区切りを考慮して,「(刑事 [だけ) 遺児 抱け]」となる．

表 3: 特徴的な駄洒落一覧

番号	駄洒落
1	(坊っちゃん) が [ぼっ ちゃんと] 水 に 飛び込む 1
539	(老化)12 し てる 人 と [廊下]1 を [走ろ う か]2 11
1634	「 (ふふふ ふふふ ふ ふふふ) 」 で [豆腐 <ふ が 10 こ>] 2
4360	(朝食)1 抜き で [超 ショック]1 、 (昼食)2 抜き で [中 ショック]2 、 (晩食)3 抜き で [一番ショック]3 222
6289	「 邪魔 な (ドイツ)12 人 は [どい つ]1 だ ？ 」 「 全員 [邪魔 <ん>]2 です 。 」 12
7180	(刑事 [だけ) 遺児 抱け] 。 意地 で も 1
8184	(くも)1 を [食らう ど]1 。 え ？ (くも)2 は [酸っぱい だ]2 22
10021	(周 富 徳)1 が (周 富輝)2 の [シュート 見 とく]1 と いっ て [シュート 見 てる]2 11
10154	(雷) の 数 は 3 6 本 。 [サンダース] だ から 2
18258	(カービィ)1 が (カビゴン)2 に ぶつかっ た 。 [カービィゴン]12 12
19813	(オーボエ) の [応募へ] 行く 2
21624	(タラ)1 が [[足ら ん]1 ち ゅら]2 13
24223	((猫)1 カフェ)2 で 、 [ニャンと]1 [猫 が 屁]2 22
26133	(牛)1 の (フン)2 、 [踏ん じゃっ た]2 ！ [も ー]1 21
44186	[超 ショック] 3
46750	(木)2345 は 言い まし た 「 (さっき)1 、 [殺気]1 感じ た … [ガキ か]2 ？ … … いや 、 [き の せい]3 だ な … … [気 に し ない]4 、 [気 に し ない]5 」 11111

- 8184 (くも)1 を [食らう ど]1 。 え ？ (くも)2 は [酸っぱい だ]2 22

8184 番は，種表現が同音異義語として複数出現し，それに応じた変形表現が存在している例である．種表現の「くも」には「雲」と「蜘蛛」の二つの意味がある．この例では，さらにこれらを英語にした場合の発音を基に変形表現である「食らうど」(クラウド),「酸っぱいだ」(スパイダー）を作成している．

- 10021 (周 富 徳)1 が (周 富輝)2 の [シュート 見 とく]1 と いっ て [シュート 見 てる]2 11

10021 番は，固有名詞としての名前を動詞を含む一つの文に展開している例である.「周富徳」を「シュート見とく」のように 1 文に展開している．

- 10154 (雷) の 数 は 36 本 。 [サンダース] だ から 2

10154 番は,「雷の数は 36 本」という駄洒落がすぐには理解できないため「サンダース」だからという理由を説明している．この場合種表現が「雷」で変形表現は「サンダース」として処理している．

- 18258 (カービィ)1 が (カビゴン)2 に ぶつかっ た 。 [カービィゴン]12 12

18258 番は，二つの種表現が，一つの変形表現にかかる例である． 種表現である「カービィ」と「カビゴン」が変形表現である「カービィゴン」にかかっている．前者は同音なので Perfect であるが，後者は音が異なるので Imperfect となる．

- 19813 (オーボエ) の [応募へ] 行く 1

19813 番は，文字としては「オーボ」と「応募（おうぼ)」で異なるが音としては同じである場合，併置型の Perfect とすべきか Imperfec とすべきかということが問題となる例である．駄洒落は，本来音で聞くものであることを考え，今回の駄洒落データベースにおいては，音として同じであれば Perfect として扱うものとした．したがって，この場合の駄洒落の種類は併置型 Perfect を表す 1 となる．

- 21624 (タラ)1 が [[足ら ん]1 ち ゅら]2 13

21624 番は，２つの変形表現が重複している例である． 種表現である「タラ」が変形表現である「足らん」になっていて，さらに「タランチュラ」を種表現とする変形表現「足らんちゅら」が存在している．前者は同音なので併置型 Perfect であるが，後者は種表現が存在しないので重畳型となる．

- 24223 ((猫)1 カフェ)2 で 、 [ニャンと]1 [猫 が 屁]2 22

24223 番は，２つの種表現が重複している例である．「猫」と「猫カフェ」が種表現として重複して存在し，「猫」に対する変形表現として「ニャンと」が存在し，

さらに「猫カフェ」を種表現とする変形表現「猫が屁」が存在している．

- 26133 (牛)1 の (フン)2 、 [踏ん じゃっ た]2 ！ [もー]1 21

26133 番は，種表現の出現順とは逆に変形表現が出現する例である．「牛」と「フン」が種表現として出現し，その後で「フン」に対する変形表現「踏んじゃった」,「牛」に対する変形表現として「もー」が出現している．

- 44186 [超 ショック] 3

44186 番は，種表現が文内に存在せず，文脈や背景知識として存在する重畳型の例である．変形表現「超ショック」に対して，種表現である「朝食」を容易に推測することができる．

- 46750 (木)2345 は 言い まし た 「 (さっき)1 、 [殺気]1 感じ た … [ガキ か]2 ？ … … いや 、 [き の せ い]3 だ な … … [気 に し ない]4 、 [気 に し ない]5 」 11111

46750 番は，駄洒落が５つ存在し，そのうちの 4 つが種表現「木」から生成されている例である．種表現「木」に対して「ガキか，きのせい，気にしない，気にしない」の 4 つが変形表現として出現している．さらに種表現「さっき」に対して変形表現「殺気」が出現している．

5　駄洒落データベースの分析結果

本章では，駄洒落データベースの総語彙数，種別語彙数などの統計的な分析結果について述べる．また，種表現，変形表現で頻出するもの上位についても考察する．

5.1　統計的な分析結果

表 4 に示すように併置型の駄洒落が 96.3 %を占めている．このことからほとんど全ての駄洒落が種表現を文内に含んでいることがわかる．また，種表現と変形表現が同一の読みであるかという点で言うと Perfect と Imperfect の割合がほぼ同率であることから，半数が同一の読みであることがわかる．

また，駄洒落の件数は 51,000 件であるが，1 件の駄洒落の中に複数の駄洒落が含まれている場合があるので，駄洒落数は 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 (種別４の不明）も 1,102 件存在していたので，これを引くと実際に駄洒落の含まれていた件数は，49,898 件となる．また，総分類数は 51,685 個であるが，この中には 1,102 件の駄洒落ではないものも含まれていたので，実際の総駄洒落数は 50,583 個となる．したがって，1 件あたりの平均駄洒落数は，50,583/49,898=1.01 個となる．これは，1 件の駄洒落中にほぼ 1 個の駄洒落が存在したということになる．

また，総単語数は 455,276 語，総文字数は 1,499,160 文字であった．平均単語数は 8.93 文字，平均文字数が 29.4 文字であったことから，1 件の駄洒落は約 9 単語から構成され，1 単語を構成する文字数が 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．

表 4: 駄洒落の収録数及び割合

種類		数 [個]	割合 [%]
併置型	1.Perfect	24,718	47.8
	2.Imperfect	25,081	48.5
	合計	49,799	96.3
3. 重畳型		784	1.5
4. 不明		1,102	2.1

総分類数 [個]	51,685
総駄洒落数 [個]	50,583
総駄洒落件数 [件]	51,000
総駄洒件数（不明を除く）[件]	49,898
総単語数 [語]	455,276
総文字数 [文字]	1,499,160
平均駄洒落数 [個]	1.01
平均単語数 [語]	8.93
平均文字数 [文字]	29.40
1 語当たりの平均文字数 [文字]	3.29

5.2　タグ付け作業における誤りについて

駄洒落データベースへのタグ付けは，人手で行ったため誤りが発生する．それらの誤りについては，すべて除去することは極めて困難であるが，プログラムを作成することによりほとんどのものを除去することができた．本節では，これらの誤りについて分析した結果について述べる．

これらの誤りの種類別の内訳を表 5 に示す． 表 5 に示すように種別を表す数字の誤りが全体の 54.9 %を占めていた．特に併置型の Perfect と Imperfect の不統一による誤りが多かった．これは，発音が同じ場合には Perfect，異なる場合には Imperfect としたが,「公費」と「コーヒー」のように長音が入った場合に，同じとするか違うものとするかで 3 名の作業者の間で意志統一されていない部分があったことによるものと考えら

れる．なお，前述したようにこの場合は音が同じということから Perfect となる．

また，2 つ目に多い誤りである種別前の空白が抜けているものについては，空白という明示的でないものを区切り記号として使用したために起きた誤りであると考えられる．したがって,「:」などの記号を空白の代わりに用いることも考えられるが，このような記号を区切り記号として用いると駄洒落本体の中で「：」を使用できないことになるので，可能な限り駄洒落本体に出現しない記号を用いる必要がある．なお，この誤りには，空白が 2 つ以上あるものも含まれている．

次に多い誤りは，記号の誤りであった．種表現を ( ), 変形表現を [ ] で囲むこととしたが，これを誤るものである．また，元々駄洒落本体に ( ) や [ ] の記号が入っていた場合には，< >に置き換えることとしていたが，この中に検索漏れがありそれが置換されずに残ったものの中で人手による修正漏れが存在した．また，これ以外に種別の数字を書き忘れたもの，対応する数字の誤り，記号の抜けなどの誤りが存在した．

今回のタグ付け作業では，合計で 639 個の誤りが存在した．駄洒落件数は 51,000 件であったが，１文中に複数の駄洒落を含む場合が存在したので，総分類数は表 4 に示すように 51,685 個である．したがって，誤り率は 639/51,685 = 1.2 %となる．

これらの誤りは，可能な限り除去しているが，完全にすべての誤りを除去できたわけではない．

次に，各作業者別の誤りの割合を表 6 に示す．一人の作業者が，17,000 件の駄洒落のダグ付けを行い，さらに他の人がタグ付けした 17,000 件の駄洒落のクロスチェックを行っている．したがって，一人の作業者が直接関与した駄洒落は 34,000 件となる．このうちでいくつ誤ったのかを示したものが表 6 に示す誤り数である．それを 34,000 で割ったものが各人の誤り率となる[1]．各人の誤り率を見ると 1.0 %から 1.3 %であり，これは人手で行った場合の誤り率として妥当な範囲であると考えられる．

表 6: 作業者別の誤り率

作業者	誤り数 [個]	誤り率 [%]
A	439	1.3
B	341	1.0
C	411	1.2

[1]実際には，クロスチェックで 2 人の作業者の意見が分かれたものについては 3 人目の作業者が決定しているが，すべてを見てはいないのでこの部分は誤り率の分母から除外している．

5.3 種表現，変形表現の頻出表現について

表 7 に種表現と変形表現のペアの出現回数別の頻度とその割合を示す．また，図 2 にそのグラフを示す．表 7 で占有率は，出現回数を駄洒落総数で割ったものである．総駄洒落数は，表 4 に示す総駄洒落数 50,583 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，変形表現の指定とその駄洒落の種類の指定が行われたことを示している．また，種表現の語彙数は 43,010 語であり，変形表現の語彙数は 43,787 語であった．これは，種表現の存在しない重畳型の変形表現の種類が 777 件であることを示している．表 4 から重畳型の駄洒落が 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．重畳型は，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られる．

表 7: 種表現と変形表現のペアの出現回数別頻度

駄洒落数	出現数	総出現数	占有率 [%]
1	38,904	38,904	76.91
2	3,713	7,426	14.68
3	768	2,304	4.55
4	227	908	1.80
5	85	425	0.84
6	48	288	0.57
7	20	140	0.28
8	14	112	0.22
9	5	45	0.09
10	2	20	0.04
11	1	11	0.02
総駄洒落数		50,583	100.00
種表現語彙数	43,010		
変形表現語彙数	43,787		

図 2 に示すように，１回から４回出現するもので，95 %以上を占めていて，ほぼ全てのものが４回以下しか出現しないことがわかる．つまり，同じ種表現と変形表現を何度も使用することはほとんどないということである．これは，種表現と変形表現が駄洒落の本質であり，これらが決まるとその他の表現のバリエーションはそれほど大きくないということを示しているものと考えられる．

表 8 に種表現と変形表現のペアの頻度ランキング上位１０組を示す．また，表 8 で出現頻度上位３ペアについて含まれる駄洒落を表記したものが表 9 である．出現頻度が最大のものは種表現が「ドイツ」，変形表現が「どいつ」で 11 回であった．

表 5: タグ付け誤りの種類

誤りの種類	例	数 [個]	割合 [%]
種別の数字の誤り	誤：2888 (公費) で [コーヒー] 飲む 1 正：2888 (公費) で [コーヒー] 飲む 2	351	54.9
種別の前の空白の抜け	誤：12177 (二酸化炭素) を [信用 中]2 正：12177 (二酸化炭素) を [信用 中] 2	103	16.1
記号の誤り	誤：1396 (セイロン 島] で [正論] 2 正：1396 [セイロン 島] で [正論] 2	79	12.4
種別の数字がない	誤：眼鏡 は 少々 (近視) で も 使用 [禁止] 正：眼鏡 は 少々 (近視) で も 使用 [禁止] 1	50	7.8
対応する数字の誤り	誤：1741 (伊藤)12 の [解答]1 は [正答]1 11 正：1741 (伊藤)12 の [解答]1 は [正答]2 11	38	5.9
記号の抜け	誤：(改造) する 間 、 介添 <かい ぞ> え] し て 2 正：(改造) する 間 、 [介添 <かい ぞ> え] し て 2	12	1.9
同じものが 2 つ存在	誤：7888 (菅 さん) が [閑散] と し た 場所 で [換算] 1 正：削除	5	0.8
種表現の範囲の誤り	誤：48382 (武田 信玄) が [進言] し た 1 正：48382 武田 (信玄) が [進言] し た 1	1	0.2
合計		639	100.0
誤り率 = 総誤り数 (639)/総分類数 (51,685)			1.2

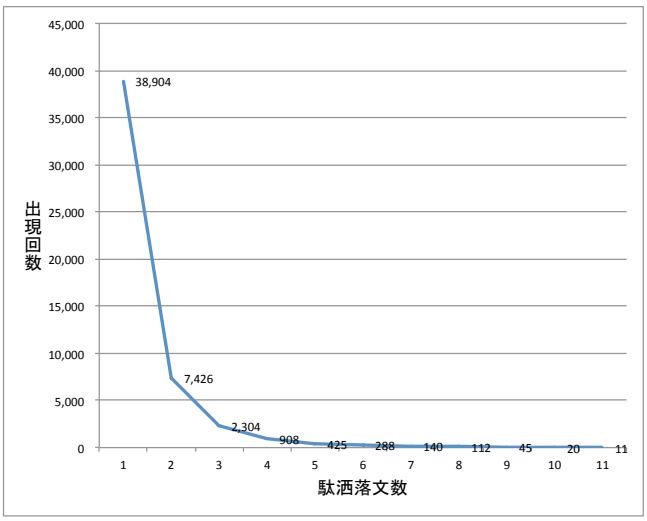

図 2: 種表現と変形表現のペアの出現回数別頻度の推移

この最大頻度の「ドイツーどいつ」の具体的な駄洒落を表 9 で見ると変形表現の「どいつ」という用語が，様々な表現に使用可能であることがわかる．そのような観点から出現頻度上位の「帽子ー防止，理科ー理解」を見ると同様に変形表現の「防止，理解」という用語が様々な表現に使用可能である．このように変形表現から多様な表現が容易に思い起こされる場合に，駄洒落の多様性が増大し，出現頻度が大きくなるものと考えられる．

表 8: 種表現と変形表現のペアの頻度ランキング

順位	種表現	変形表現	頻度
1	ドイツ	どいつ	11
2	帽子	防止	10
2	理科	理解	10
4	秋田	飽き た	9
4	ケーキ	景気	9
4	医院	いい ー ん	9
4	鮭	酒	9
4	朝食	超 ショック	9
9	胃	いい	8
9	韓国	勧告	8

注) 頻度が 8 回のものについては 14 組あるのでそのうち 2 組を示す．

次に，一つの種表現がどれだけ多様な変形表現を作成できるのかを調査するために種表現から作成される

表 9: 種表現と変形表現のペア出現頻度上位 3 位までの駄洒落

番号	駄洒落
順位：1，種表現：ドイツ，変形表現：どいつ，出現頻度：11	
4805	(ドイツ)12 と [同一]1 って 言っ た の 、 [どい つ]2 だ ？ 21
6289	「 邪魔 な (ドイツ)12 人 は [どい つ]1 だ ？ 」 「 全員 [邪魔 <ん>]2 です 。 」 12
6672	(ドイツ) の 悪口 を 言う の は 、 どこ の [どい つ] だ ？ 1
6701	(ドイツ) に 行っ た の は [どい つ] だ ？ 1
6758	(ドイツ)12 人 は [どい つ]1 だ ？ 『 [じゃ 〜 真ん中]2 』 12
15316	(ドイツ) に 行く の は [どい つ] ？ 1
18381	(ドイツ) 語 が 好き だ という の は 、 どこ の [どい つ] だ ？ 1
18728	(ドイツ) 人 って [どい つ] ？ 1
32893	(ドイツ) 人 は [どい つ] 1
33551	(ドイツ) が 好き な の は [どい つ] だ ？ 1
45430	(ドイツ) 人 は 、 どこ の [どい つ] だ 1
順位：2，種表現：帽子，変形表現：防止，出現頻度：10	
15736	(帽子) の 盗難 [防止] 1
16046	(帽子) の ずれ [防止] 1
19627	(帽子) の 色 落ち [防止] 1
23523	(帽子) で 日焼け を [防止] 1
32361	(帽子) で [防止] する 1
33438	(帽子) で 熱中 症 を [防止] 1
41865	(帽子) かぶっ て ボケ [防止] 1
45076	(帽子) かぶれ ば ○ ○ を [防止] できる 1
45694	熱中 症 を (帽子) で [防止] 1
48445	(格闘 家)1 の (帽子)2 で [核 投下]1 を [防止]2 11
順位：3，種表現：理科，変形表現：理解，出現頻度：10	
12343	(理科) を [理解] する 1
12370	(理科) の 授業 は あまり [理解] でき ない 1
22193	(理科)12 が [理解]1 でき なく て [わり ー か]2 12
25576	(理科) は [理解] でき ませ ん 1
26571	(理科) を [理解] しろ 1
28604	(理科) 、 [理解] でき ん 1
34346	(理科) が [理解] でき ない 1
35010	(理科) を [理解] しよ う 1
44964	(理科) の 先生 を [理解] しろ 1
49658	(理科) の 授業 は [理解] でき ない 1

変形表現の種類数のランキングを行った．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 43,787 で割ったものである．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を示す．

表 10 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 を見るとわかるように「ドウ」の部分が文末表現としての「どう」として変形表現で使用されるので，多様性が大きくなっている．2 位の「サメ」について見てみると，英訳の「シャーク」やサメの映画で有名な「ジョーズ」の 3 語から変形表現が作成されていることにより多様性が大きくなっている．また，3 位の「手紙」では，英訳である「レター」が文末表現の「れた」と同音であることから 1 位の「ブドウ」と同様な仕組みで変形表現の多様性が増大している．

表 10: 一つの種表現から生成される変形表現数のランキング

順位	種表現	種類数	占有率 [%]
1	ブドウ	35	0.080
2	サメ	34	0.078
3	手紙	32	0.073
4	魚	31	0.071
5	レタス	30	0.069
5	イカ	30	0.069
5	猫	30	0.069
5	メジャー	30	0.069
9	仙台	29	0.066
9	内容	29	0.066

次に，一つの種表現がどれだけ多くの多様な駄洒落を作成できるのかを知るために種表現から作成される駄洒落の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，駄洒落数を総駄洒落数 50,585 で割ったものである．表 12 で注目すべき点は，上位 5 位までのうち 4 語がカタカナ語である点である．これは，表 10 に示す種類数でも同様の傾向があり，上位 5 位までの 8 語のうち 5 語がカタカナ語である．しかも，表 10 に示す種類数では漢字語である「手紙」も「レター」としての駄洒落が多いので，ほとんどカタカナ語で占めていると考えることができる．

カタカナ語は，比較的短い語が多く，しかも文末表現と似た音になることがあると駄洒落の表現の自由度が増し，比較的容易に駄洒落を思い付くことができるものと考えられる．

表 13 に作成された駄洒落の数が最も多い種表現「メジャー」の変形表現と駄洒落の例を示す．表 13 に示すように変形表現「駄目じゃー」に対しては，6 種類の駄洒落が存在するが,「9868 (メジャー) じゃあ [駄目 じゃー] 1」と「17090 (メジャー) じゃ 、 [駄目 じゃー] 1 」のようにその差異が微々たるものであったものも存在した．一方,「4618 (メジャー) 買う の が [夢 じゃー] 1」と「32758 (メジャー) なんて 夢 の また [夢 じゃー] 1」のように同じ変形表現でも全く異なる文脈で使用されているものも存在した．

表 12: 一つの種表現から生成される駄洒落数のランキング

順位	種表現	駄洒落数	占有率 [%]
1	メジャー	51	0.101
2	マスター	46	0.091
3	魚	45	0.089
4	サメ	40	0.079
4	イカ	40	0.079
6	時計	39	0.077
7	カッター	38	0.075
8	火星	37	0.073
8	内容	37	0.073
10	ブドウ	36	0.071
10	バッタ	36	0.071
10	手紙	36	0.071

表 13 をみるとわかるように,「メジャー」の「ジャー」の部分が文末表現の「じゃー」と同音のことから駄洒落の数が多くなったものと考えられる．

このように文末表現と同音となる部分を含む種表現が，変形表現の種類や作成できる駄洒落の数が多くなる傾向があり，駄洒落の作成においては種表現の中に文末表現と同様の音を含むことが重要であると考えられる．

6 おわりに

本稿では，まず始めに少子高齢化社会における心のケアを補うこと及び医療的な効果の観点からユーモアを処理する技術の確立が必要であることを述べ，研究を進めるために標準的な駄洒落データベースの構築が必要であることを述べた．次に具体的な駄洒落データベースの構築方法として，収集方法，タグ付け方法などについて述べた．

また，複数の駄洒落が 1 文に存在するもの，一つの種表現に複数の変形表現が存在するもの，異なる種類の駄洒落が 1 文に存在するもの，種表現や変形表現が入り子状に重複して出現するものなどについて，具体例を挙げて説明を行なった．

また，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結果及びタグ付け作業における誤りの分類と量的な考察を行なった．また，種表現，変形表現における頻出表現について，統計的

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

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

注) 変形表現の後ろの数字は出現頻度を表す.

表 13: 一つの種表現から生成される駄洒落数 1 位「メジャー」の変形表現と駄洒落の例

順位：1，種表現：メジャー，駄洒落数：51（変形表現種類数：30）
駄目 じゃー 6：夢 じゃー 3：お め ー じゃ 3：お め ー じゃー 3：不明 じゃー 3：指名 じゃー 2：感銘 じゃー 2： メジャー 2：判明 じゃー 2：有名 じゃー 2：３ 人 目 じゃー 2：使命 じゃー 2：メジャ せ 2：何 人 目 じゃー 1： 任命 じゃー 1：明治 や 1：説明 じゃー 1：目 じゃ ない 1：無理 じゃー 1：目 じゃー ない 1：本命 じゃー 1： 宿命 じゃー 1：目指し て 1：２つ 目 じゃー 1：メジャ し て 1：究明 じゃー 1：１ 年 目 じゃー 1：ダメ じゃー 1： うめ ぇじゃん 〜 1：オメエ じゃ ねぇ ー 1

番号	駄洒落
9868	(メジャー) じゃあ [駄目 じゃー] 1
17090	(メジャー) じゃ 、 [駄目 じゃー] 1
21374	(メジャー) で 測る の は [駄目 じゃー] 1
25892	あいつ は (メジャー) じゃあ [駄目 じゃー] 1
41631	その (メジャー) は [駄目 じゃー] 1
50454	(メジャー) 貸し て ー ！ [駄目 じゃー] 1
4618	(メジャー) 買う の が [夢 じゃー] 1
32758	(メジャー) なんて 夢 の また [夢 じゃー] 1
38266	(メジャー) に 行く の が [夢 じゃー] 1
12406	(メジャー)12 、 [お め ー じゃ]1 、 [無理 じゃー]2 22
44501	(メジャー) 行き て ー ！ [お め ー じゃ] 、 無理 じゃー 2
45137	その (メジャー) 、 [お め ー じゃ] 、 無理 じゃー 2
20265	(メジャー) 使う の は 、 [お め ー じゃー] 2
26220	(メジャー) に 行く の は [お め ー じゃー] 2
39273	(メジャー) で 計る の は 、 [お め ー じゃー] 2

注) 変形表現の後ろの数字は出現頻度を表す.

な分析を行なった．その結果，出現頻度が４回以下のものが 95 %以上を占め，同じ種表現と変形表現が出現することが極めて少ないことが確認された．

一方，何度も出現する駄洒落の特徴として，変形表現がカタカナ語であり，かつ文末表現と同音の部分を含むものが数多く出現することが明らかとなり，文末表現を同音として駄洒落を生成することが容易であることが確認された．

今後の課題としては，形態素解析結果の基準の作成と修正及びさらなるデータベースの拡張が挙げられる．形態素解析については，現状では形態素解析ツールの解析結果をそのまま利用しているが，変形表現が未知語であることが多く，誤りが存在する．そもそも人間が見てもどのように分割すべきか不明な場合もあるので，まず変形表現に出現する未知語に対する形態素解析の正しい分割方法および品詞を決定し，その後に機械学習を用いた形態素解析ツールに正しい形態素解析結果を学習させる必要がある．また，現在は 51,000 件を収録しているが，駄洒落におけるオノマトペの解析を行うためには，量的に不十分なため 10 万件程度まで拡張する必要があると考えられる．

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] Mihalcea, R., Strapparava, C. and Pulman, S.: Computational Models for Incongruity Detection in Humor, Linguistics and Intelligent Text Processing, Lecture Notes in Computer Science, Vol.6008, pp.364-374, 2010.

[6] Mihalcea, R. and Pulman, S.: Characterizing Humour: An Exploration of Features in Humorous Texts, Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Text Processing, pp.337-347, 2007.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.