

駄洒落データベースを用いた面白さの評価及び分析

Evaluation and Analysis of Fun Using 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 Araki¹ Yuzu Uchida² Kohichi Sayama³ Motoki Yatsu⁴

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: 現在広く普及しているスマートフォンや AI スピーカー上の対話エージェントは、雑談システムとしての能力が不十分であり、話していて楽しいシステムとはなっていない。この問題を解決するためには、ユーモア処理の高度化が必須である。その第一段階として、我々は 68,000 件の駄洒落データベースの開発を行った。しかし、この駄洒落データベースには面白さの度合いを示すスコアが付与されていなかった。そこで、3 名の評価者により面白さのスコアの付与を行った。次に、この駄洒落データベースを用いて、面白さのスコアの平均値の推移、評価者ごとのスコアの分布、評価者間の一致度、平均スコアの上位のものなどの分析を行った。この結果、駄洒落の面白さの感じ方には個人差が大きいことや対話形式で表現されている場合に面白さを感じるなどが明らかとなった。これらのことより文脈情報、背景知識が駄洒落の面白さの感じ方の大きな要因となっていることが確認された。

1 はじめに

近年、生涯独身者の急増や社会風習の変化により、高齢者単身世帯が急増している。そのため、孤独感を感じさせないための心のケアの需要も急激に増大している。孤独感を感じさせないためには、行政が高齢者単身世帯を定期的に訪問するための人員を雇用することが最善の方法である。しかし、この方法は高齢化社会の進展に伴い急増する医療費などのため現状の予算では困難である。そこで、この問題を解決するために訪問者を代替する存在として、人間と同等の対話能力を持つ対話システムの実現が緊急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や AI スピーカー上の Siri や OK Google や Alexa などの対話エージェントの普及が急激に進み、これらの対話エージェントと接する機会が増大している。しかし、これらのシステムは基本的に何らかの仕事を行うためのもの（タスク指向型システム）であり、話をするにより面白さや楽しさを感じることを主目的として開発されたものではない。会話による面白さや楽しさを主目的として開発された対話システムは非タスク指向型システム（雑談システム）と呼ばれる。

雑談システムの完成度は高いものではなく、5 回話

すと飽きてきて 20 回話するのが限界というのが現状である。このような問題に対して最近では対話システムシンポジウム¹という雑談システムの性能をコンペ形式で競う大会も実施されてきている。しかし、雑談システムとの対話の面白さはゲームや映画やドラマや漫画などといったエンターテインメントの面白さに遠く及ばない。このことは、ゲームをしたり映画などを見て朝まで徹夜してしまったということはあるが、雑談システムと話すのが楽しくて朝まで徹夜してしまったという話は聞いたことがないことから明らかである。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、依然として最終的な目的である話すのが楽しくて朝まで話込んでしまったというレベルには達していない。

この問題を解決するためには、人間並みの精度の高い応答が行えるということはもちろんであるが、それだけでは問題の解決にはならず、いわゆる話し上手な雑談システムを開発する必要がある。話し上手な雑談システムの要件は、相手の興味をひく話題について詳しく話すことができる、適切なタイミングで相槌を打つことができるなど、実に様々なものがあるが、その中の一つとしてユーモアを理解して笑い、適切なタイミングで相手を飽きさせないようにユーモアを発するというのも大きな要因であると考えられる。

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

¹ <https://jsai-slud.github.io/sig-slud/87th-sig.html>

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、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、これらの研究で用いられているユーモアはいずれも少数である。例えば、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。このことは、一般的なユーモアの性質、特性などを議論することが困難であることを意味する。

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 [11] などの研究を行ってきた。これらの研究を行う際に問題となったのが、評価を行うための大規模な標準的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。

このようなこれまで行われたきた研究の数々の問題を解決するための第一段階として、我々はユーモアに関する大規模なデータベースの構築を行った。まず、初めに 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]。さらにこの駄洒落データベースを 68,000 件まで拡張した [16]。また、この拡張された駄洒落データベースを用いて駄洒落の面白さの認知科学的見地からの考察 [17]、オノマトペに着目した駄洒落の面白さの分析 [18] などの研究を進めてきた。これら研究は一定の成果をあげている。

しかし、これらの研究においては、大規模な駄洒落データベースを用いた複数評価者による駄洒落の面白さの評価及び分析を行っていない。駄洒落の面白さを科学的に解明し、面白さを評価する関数を構築することは、雑談システムがユーモアを理解・生成するためには必須のものである。そこで、本稿では駄洒落データベースを用いて行った 3 名の評価者による大規模な面白さの評価結果とその分析結果について詳細に述べる。

2 駄洒落の面白さの評価

2.1 評価方法

駄洒落の面白さの評価に用いたものは、文献 [16] で述べた 68,000 件の駄洒落データベースである。この駄洒落データベースをクラウドソーシングで雇用した 3 名の評価者に各自 68,000 件の駄洒落を評価していただいた。評価者の属性を表 1 に示す。表 1 で「認定」とはクラウドソーシングの会社が返信速度・納品実績・獲得額・高評価などの一定の条件を満たしている登録者に与えるもので、良質な作業者を選択する際の目安となるものである。今回の評価者 3 名は全員この認定を受けているので、一定レベルの能力を満たしているものと考えられる。

表 1: 評価者一覧

作業者名	年代	経歴	認定
評価者 1	30 代後半男性	その他職種	○
評価者 2	30 代後半男性	その他職種	○
評価者 3	20 代後半男性	その他プログラマー・エンジニア	○

評価者には、タグ付けされていない駄洒落の原型を提示し、可能限り直感的に表 2 に示す 5 段階で面白さの評価の依頼を行った。

2.2 分析対象の駄洒落データベースについて

駄洒落データベースの開発に関する論文 [12, 14] を読んだ方々から駄洒落データベースを公開して欲しいという要望があり、公開用の駄洒落データベースの作

表 2: 面白さのスコア

スコア	説明
5	とても面白い
4	面白い
3	普通
2	面白くない
1	非常に面白くない (駄洒落ではないと思うものも含む)

成を行うこととした。今回、面白さの評価結果の分析を行なったのは、この公開用に開発された駄洒落データベースである。

公開用駄洒落データベースのフォーマットを図 1 に示す。図 1 に示すように、1 レコードは、通し番号から始まり、駄洒落文に一切のタグ付けを行っていない原型、単語ごとに形態素解析を行い、種表現や変形表現を記号で示したタグ付きのもの、駄洒落の種類別、評価者の面白さのスコア、スコアの平均の順に表記し、それぞれの要素を”,”で区切っている。

また、タグ付けされた駄洒落は事前に形態素解析ツール MeCab[20] を用いて、形態素解析を行なっている。単語間はスペースで分割されている。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、形態素解析結果については、変更を行っていない。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。

通し番号,原型,タグ付,種別,スコア1,スコア2,スコア3,平均スコア

図 1: 公開用駄洒落データベースのフォーマット

駄洒落の種類を表 3 に示す。図 1 の種別は、表 3 の種類に書かれている番号を記載したものである。収録された駄洒落の例を表 4 に示す。表 4 に示すように、先頭に通し番号を付与し、次に駄洒落の原型、タグ付けされた駄洒落を記載する。タグ付けの方法については、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。その次に駄洒落の種類別、面白さのスコアを記載している。

例えば、以下に示す表 4 の 2 では、駄洒落の原型は「高菜。あったかな」であり、タグ付けされたものは「(高菜)、[あったかな]?」となるので、種表現が「高菜」であり、変形表現は「あったかな」となる。ま

た、種別は併置型 Perfect を表す 1 であり、評価者 1, 2, 3 のスコアが全て 3 であり、平均値は 3.00 となる。

- 2, 高菜、あったかな?, (高菜)、[あったかな]? , 1, 3, 3, 3, 3.00

また、以下に示す 73 では、1 件の駄洒落の中に複数の変形表現が存在する例である。種表現「カレンダー」に対して変形表現「だれんだー」と「俺んだー」が対応している。このような場合には、種表現、変形表現の後ろに空白なしで対応する数字を書くことにより、対応関係を明確にしている。73 のように「カレンダー：だれんだー」, 「カレンダー：俺んだー」の複数の種表現、変形表現のペアが存在する場合には、その出現順に駄洒落の種類を表す番号を種別として「22」とタグ付け駄洒落の後に記載している。また、292 は、種表現と変形表現が複数ある場合の例であり、2820 は種表現のみ複数ある場合の例である。

- 73, カレンダーだれんだー俺んだー, (カレンダー)12 [だれんだー]1 [俺んだー]2, 22, 4, 2, 3, 3.00
- 292, イカとゾウが喧嘩 いかんぞう, (イカ)1 と (ゾウ)2 が 喧嘩 [いかん]1 [ぞう]2, 11, 4, 4, 3, 3.67
- 2820, 支那のビル萎びる, (支那)1 の (ビル)2 [萎びる]12, 11, 3, 2, 3, 2.67

2.3 公開に向けて行った修正

これまでも実験で使用する駄洒落データベースについては、種々の修正を行い使用してきたが、誤りなどを全て無くすることは不可能であるため一般公開に向けて再度種々の修正を行った。削除を行ったものは、以下のものである。

- 文字化けなど
- 助詞から始まるもの
- 性的表現を含むもの

このような観点から削除を行った結果を表 5 に示す。まず、1 点目の文字化けなどの削除というのは、文献 [14] で述べた拡張された駄洒落データベース中には文字化けや URL や記号だけのもが含まれていたため、これを第一著者が手作業で削除を行った。また、2 点

表 3: 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。

表 4: 収録された駄洒落の例

通し 番号	原型	タグ付	種別	スコア			
				1	2	3	平均
2	高菜、あったかな？	(高菜)、[あったかな]？	1	3	3	3	3.00
73	カレンダーだれんだー俺んだー	(カレンダー)12 [だれんだー]1 [俺んだー]2	22	4	2	3	3.00
292	イカとゾウが喧嘩 いかんぞう	(イカ)1 と (ゾウ)2 が 喧嘩 [いかん]1 [ぞう]2	11	4	4	3	3.67
2820	支那のビル萎びる	(支那)1 の (ビル)2 [萎びる]12	11	3	2	3	2.67

表 5: 削除を行った駄洒落の内訳

種類	説明	件数
1	文字化けなど	593
2	助詞で始まるもの	121
3	性的表現を含むもの	285
合計		999

目の助詞から始まるものについても削除を行った。これらの作業を行った際の具体的手順としては、種別が 4 (不明) で全員のスコアが 1 (平均スコアが 1.00) のものをリストアップし、それらを第一著者が見て上述したものに該当するものの削除を行った。これらのものは表 5 に示す 1 の 593 件と 2 の 121 件の合計 714 件である。

また、3 点目の性的表現を含む駄洒落、いわゆる下ネタ駄洒落の削除についてであるが、どこまでが性的な表現かは判断が難しいため第一著者が見て直接的に性的表現を含むもののみを削除することとした。厳密には、単語そのものに性的な表現が含まれていなくても通常の単語の組み合わせで婉曲的に性的なものを表現した駄洒落が存在するが、それらをすべて抽出することは、非常に難しいので明確に性的表現と判断できるものだけを削除することとした。具体的手順としては、第一著者がある程度の量の駄洒落を見て、性的表現を含む駄洒落と判断した基準となる単語の抽出を行った。このようにして抽出された単語は 58 語であった。この 58 語を含む駄洒落を検索し、285 件の駄洒落を検出し、削除を行った。

この結果、削除された駄洒落は 999 件であったので、

文献 [14] の 68,000 件の駄洒落データベースは、67,001 件まで減少した。ここから最後の 1 件を削除し、67,000 件を収録したものを公開用の駄洒落データベースとして公開を行った。

この駄洒落データベースを研究用に使用したい方は、公開用のサイト²に書かれている条件をお読みの上、第一著者までメールでご連絡ください。

3 評価結果の分析

2.3 で述べたような方法で開発された公開用駄洒落データベースに収録されている 67,000 件の駄洒落を用いて、面白さの評価結果の分析を行った。

3.1 平均値の推移

各評価者のスコア、および全評価者の平均スコアの 1,000 件ごとの平均値を算出した。その結果を図 2 に示す。図 2 を見ると評価者 1 のスコアが、およそ 36,000 件を境に急減に低下している。これに伴い、平均値も低下している。これは 2 章で述べたように文献 [14] で述べた 68,000 件の駄洒落データベースをクラウドソーシングを用いて面白さの評価をお願いした際に、最初は 1,000 件から初めて 5,000 件、10,000 件、20,000 件と依頼件数を増加させ、最後に残りの 32,000 件の依頼を行ったために 36,000 件以降で一度に行う作業量が膨大になったことも一因であると考えられる。また、すでに 36,000 件の駄洒落を見てきたことにより飽きたと

²<http://arakilab.media.eng.hokudai.ac.jp/araki/dajare.htm>

いう状態になり駄洒落に面白さを感じなくなったということも考えられる。このようなことを考えると今回、3名の評価者を雇用し、各自 68,000 件の駄洒落の評価をしていただいたが、飽きるということを考えると多くの評価者に少しずつ評価していただいた方が良かったのではないかと考えている。これは、今後の課題である。

また、評価者2の方は 58,000 件を境に少々スコアの低下が見られるが、これも飽きたことに起因している可能性がある。また、評価者3の方は、最初はほぼ同じスコアであるが、15,000 件でスコアの低下が見られる。これは、元々駄洒落を面白いと感じる感性が低く、15,000 件ですでに駄洒落の面白さに飽きてきているものと考えられる。

表6に評価者ごとのスコアの平均値を示す。評価者1と2は 2.54 と 2.62 で比較的近い値になっているが、評価者3は 2.17 と非常に低い値である。これは図2を見ても明らかで、評価者3は一貫して低い値である。そもそも単独の駄洒落に対して面白さを感じる人と感じない人がいて、3名を任意に選ぶとそのような人が入る可能性がある。つまり、駄洒落に対する面白さの感じ方には個人差が大きいということである。

表 6: 評価者ごとのスコアの平均値

評価者	平均
1	2.54
2	2.62
3	2.17
総合	2.44

3.2 スコアの分布

5段階評価の評価者別のスコアの分布を図3に示す。図3を見るとわかるように評価者1と2はスコアの分布が類似しているが、3.1でも述べたように評価者3は大きく異なっている。評価者3の場合、実にスコア2が8割を占めている。5点はわずかに1例であり、この点からも一般的に単独の駄洒落そのものに面白さを感じていないことがわかる。

これに対して、評価者1と2のスコアの分布は、2点が30%から40%であり、3点が約50%であるので、2点と3点が合わせて80%から90%を占め全体的には類似している。一方、評価者2は5点が249件であり4点は6,638件と合わせて1割を占めているが、評価者1は5点が26件で、4点が526件しかないので合わせても1%未満である。また、1点についても評価者1は5,880件と8%を占めているのに対し、評価者2は1点が868件しか存在せず1.3%である。した

がって、評価者1と2についても高評価と低評価については、約10倍の差がある。

このように評価者1と2は表6に示すように平均値が近い値を取り、図3に示すように2点、3点の分布は類似しているが、高評価や低評価では異なっているため、評価者1と2についても駄洒落が単独で存在する場合の面白さの感じ方には大きな違いがあり、個人差が大きいものと考えられる。

また、評価者3名全員のスコアを合計したものの分布を図4に示す。図4を見るとわかるように評価者3名のスコアの92.1%が2点か3点であることがわかる。また、4点は3.6%で1点は4.1%である。5点は、0.1%しか存在しない。このため表6で総合の平均スコアが2.44となっている。このように比較的低い値になった原因としては、そもそも駄洒落単独で面白さを表現することに限界があることが考えられる。これは、対話中のように何らかの文脈の中で駄洒落を発することで解決できる部分もあるものと考えられるが、これは今後の課題である。

前述したように、5段階評価にした場合、個人差が大きい結果となった。そこで、スコアを3段階評価に換算した場合のスコアの分布について考察を行った。スコアの換算は以下に行った。

- 5点および4点 → 3点
- 3点 → 2点
- 2点および1点 → 1点

このような換算を行い3段階評価にした場合の評価者別のスコアの分布を図5に示す。図5を見るとわかるように、5段階評価での5点と1点が全体に占める割合が少ないため、3段階評価に換算しても、全体的な傾向は5段階評価の場合と変わらず、評価者3が1点の評価が約80%と極端に多くなった。評価者1と2の1点、2点のスコア分布も同様に類似していたが、評価者2の5段階評価の際の5点と4点の占める割合が1割程度で評価者1の1%未満と比べて多かったため、評価者2の3段階評価の際の3点の割合も同様に異なっている。

3.3 評価者間の一致度について

3.3.1 k 係数による一致度の評価

次に評価者間の面白さのスコアの一致度について考察する。表7にコーエンの k 係数を用いて2名の評価者間の一致度を評価した結果を示す。5段階評価の場合には、いずれも0.1未満であることからわずかな一致という結果になった。その中でも、評価者1と2の k 係数

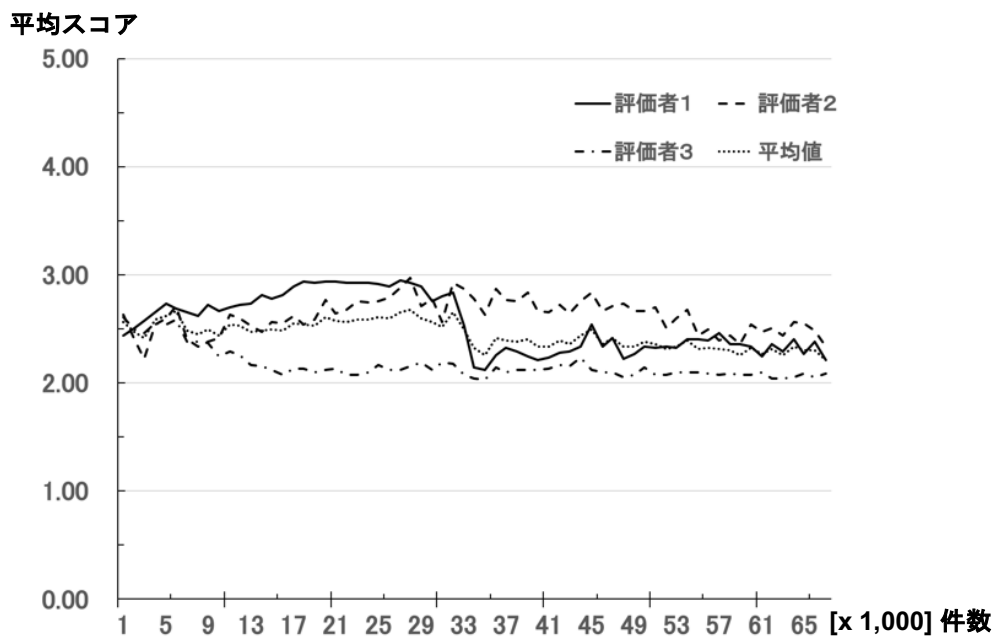


図 2: 平均値の推移

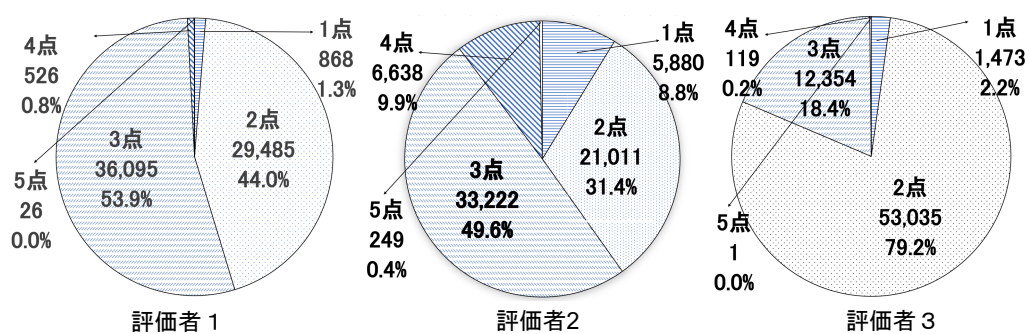


図 3: 5段階評価の評価者別のスコア分布

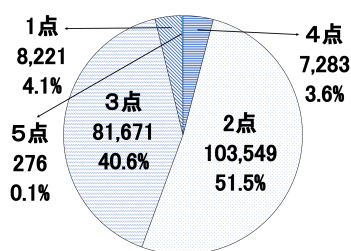


図 4: 5 段階評価の総合のスコア分布

が 0.079 と評価者 2 と 3，評価者 1 と 3 の 0.022, 0.046 に比べて，比較的高いことから評価者 1 と 2 の評価結果が比較的類似していて，評価者 3 が異なっていることがわかる．3 段階評価にした場合には，さらに評価者 1 と 2 が 0.100 と高くなり，評価者 2 と 3，評価者 1 と 3 が -0.001, 0.048 であることからこの傾向がさらに強くなるという結果となった．

評価者	k 係数	
	5 段階	3 段階
1 と 2	0.079	0.100
2 と 3	0.022	-0.001
1 と 3	0.046	0.048

表 7: 評価者間の一致度

3.3.2 評価者間の一致数と割合

評価者間の一致数とその割合を表 8 に示す．また，評価者間の一致状況を図 6 に示す．

まず，5 段階評価の場合について考察する．図 6 からわかるように 5 段階評価の場合の 2 者間の一致が最も多く 48,017 ペアとなっている．中でも評価者 1 と 3 の一致数が 19,077 ペアと評価者 1 と 2 の一致数 17,737 ペアより高い値となっている．図 3 を見るとスコアの分布は全く異なっていることから，評価者 1 と 3 は面白さを感じる感性が近く評価者 3 が 8 割を 2 点としたものを評価者 1 も 2 点と評価したものが多かったことを示している．これに比べて評価者 2 と 3 では，評価が一致した件数が 11,203 ペアと低いことから面白さを感じる感性に違いがあったものと考えられる．

また，評価者 3 名が一致したものが 12,657 組あるので，一致度の k 係数は低いものの，このような全く主観的な評価においては，比較的一致したものが多かったのではないかと考えられる．一方，評価者 3 名がすべて異なるスコアを付与したものが 6,326 組存在した．これは評価者 3 名が一致した場合の約半数であり，評価者 3 名がそれぞれ全く異なる評価をしたものが少なかったことを示しているものと考えられる．

また，評価者 3 名ともに 5 点としたものは 0 であり，1 点としたものは，126 組であった．そもそも評価者 3 が 5 点としたものは 1 件であり，3 名とも 5 点としたものが 0 であったのは当然の結果である．1 点が 126 組と少なかったことについては，2.3 で述べたように記号や文字化けなど明らかに駄洒落でないものを分析用に使用した公開用の駄洒落データベースから削除していることから評価者 1 の 1 点が 526 件と少なく，このことが 3 名ともに 1 点としたものが少なかった要因となっていると考えられる．

次に，3 段階評価の場合について考察する．図 6 からわかるように 3 段階評価に換算した場合も 5 段階評価と同様の傾向を示している．異なる点としては，評価者 3 名が一致したものが 15,887 組と約 3,000 組増加していることと 3 名とも異なるスコアを付与したものが 3,877 組と約 2,500 組減少していることである．これは，評価の段階を 5 段階から 3 段階に大まかにしたことによるものである．

また，3 段階にした場合でも 3 名とも 3 点と評価したものは 3 組であった．図 5 から 3 点とした数は，評価者 1，2，3 でそれぞれ 552 件，6,887 件，120 件であることを考えるとこれは低い値である．3 段階評価にしたとしても面白いものの感じ方には大きな個人差が存在することが確認された．

一方，3 名とも 1 点と評価したものは 11,961 組と 5 段階評価の場合の 126 組から約 10 倍と急増している．これは，評価者が 5 段階評価で 1 点あるいは 2 点と評価したものは，比較的一致していたということになる．これは，評価者 3 の評価が 3 段階評価の場合でも 1 点が約 8 割と偏っていたことが原因であると考えられる．

3.3.3 評価者全員のスコアが一致したものと異なったものの分析

図 6 に示すように 5 段階評価の場合，評価者 3 名全員のスコアが一致したものが，12,657 件あり，評価者 3 名ともに異なるスコアを付与したものが 6,326 件存在した．これらの一致したものと異なったものがどのようなスコアだったのかを示したものが図 7 である．図 7 の左側のグラフから評価者 3 名のスコアが一致した場合には，2 点だった場合が 68.0 % と最も多く，次に 3 点の 31.0 % である．したがって，2 点と 3 点が 99.0 % を占め，それ以外の 5 点，4 点，1 点で一致したものは，4 点が 3 組，1 点が 126 組で 1.0 % のみであった．このように評価者 3 名が一致する場合は，ほぼすべて 2 点の面白くない，あるいは 3 点の普通であったということがわかる．

次に，評価者 3 名全員の評価がすべて異なった場合のスコアの分布についてであるが，図 7 の右側を見ると 3 点以上 4 点未満の場合が 60.3 % であり，2 点以上

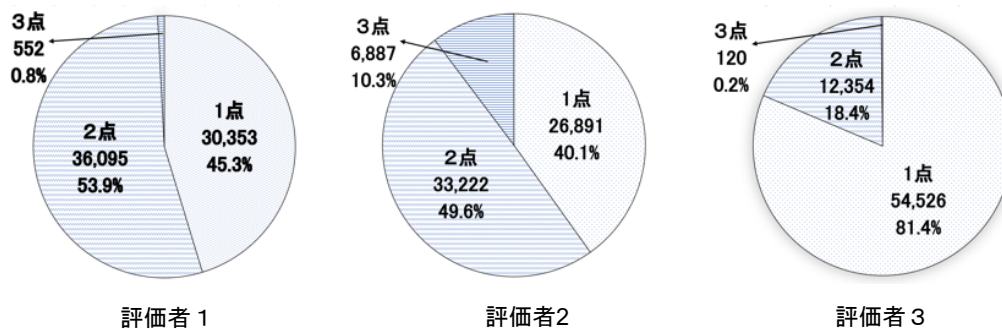


図 5: 3段階評価の評価者別のスコア分布

表 8: 評価者間の一致数と割合

	5 段階					
	3 者 一致	2 者一致				3 者 不一致
		12 一致	23 一致	13 一致	合計	
組数	12,657	17,737	11,203	19,077	48,017	6,326
割合 [%]	18.9	26.5	16.7	28.5	71.7	9.4
	3 段階					
	3 者 一致	12 一致	23 一致	13 一致	合計	3 者 不一致
	割合 [%]	割合 [%]	割合 [%]	割合 [%]	割合 [%]	割合 [%]
組数	15,887	17,946	12,108	17,182	47,236	3,877
割合 [%]	23.7	26.8	18.1	25.6	70.5	6.0

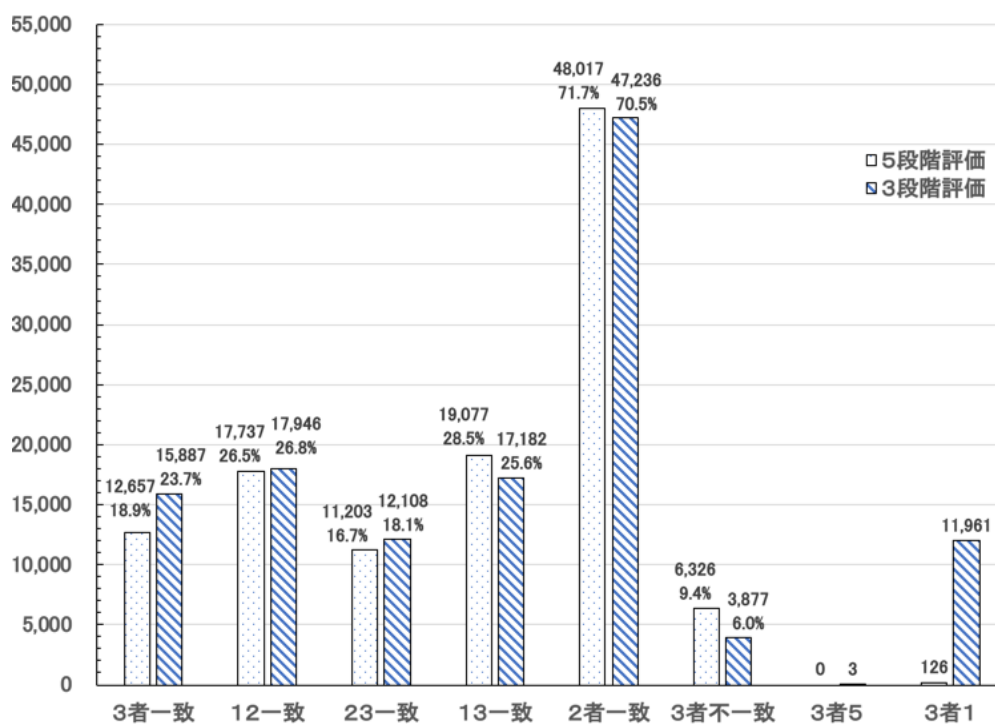


図 6: 評価者間の一致状況

3点未満の場合が、39.5 %と合計で 99.8 %を占めている。それ以外の4点以上5点未満は11組だけで0.2 %である。これも3名一致した場合と同様の傾向を示していて、3点の普通と2点の面白くない部分で3名の評価が分かれていることがわかる。このような結果になった理由については、3.2でも述べたように評価者3名のスコアを大半を3点と2点が占めていることが原因であると考えられる。

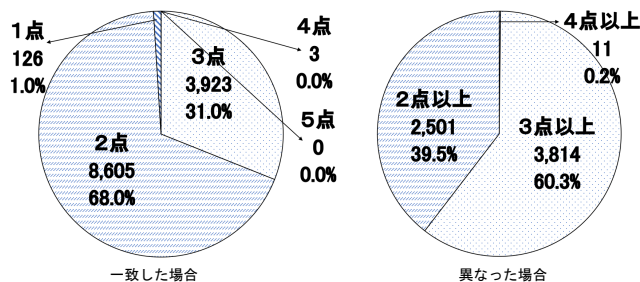


図 7: 評価者全員が一致したものと異なったもののスコアの分布

3.3.4 評価者全員のスコアが一致したものの例

表 9 に評価者 3 名のスコアが一致したものの例を示す。以下に示す表 9 の 8639, 26680, 39947 は、評価者 3 名全員が 4 点を付けたものである。これらはいずれも 1 件の駄洒落であるが、対話形式になっている。このことは、短いものであるが、対話の中で駄洒落を発した場合、その文脈に合致したのであれば、評価者が一致して面白いと評価することを示している。この結果からも面白い駄洒落を生成するためには文脈の中でその文脈に合致した駄洒落を生成することが重要であることがわかる。

- 8639, 「あのテニスの選手が持っているもの何?」 「…手にスプレーや!!」, 「あの (テニスの選手) が持っているもの何?」 「… [手にスプレーや] !!」, 2, 4, 4, 4, 4.00
- 26680, 『ジュース買ってこようか? お金ある?』 『Nomoney! <飲まねえ>』, 『ジュース 買ってこようか? (お金ある)?』 『[No money]! <飲まねえ>』, 2, 4, 4, 4, 4.00
- 39947, ニーチェ「神は死んだ!」 父「髪は死んだ!」, ニーチェ「(神)は死んだ!」 父「[髪]は死んだ!」, 1, 4, 4, 4, 4.00

また、以下に示す 27 は、評価者 3 名が全員 3 点を付けたものである。3 点は普通という意味であるが、「秋田はあきた」のようによく知られたもので容易に推測のつくものは、特に面白いということはなく普通という評価になるということである。これは、佐山らの文献 [13] に示すように、駄洒落を理解する場合の「処理負荷が少なすぎて面白くない場合」に相当するものと考えられる。

- 27, 秋田はあきた, (秋田) は [あき た], 1, 3, 3, 3, 3.00

また、以下に示す 8661 は、「要港」という単語がよく知られたものではないことが、評価者全員 2 点となった原因であると考えられる。そもそも「要港」という単語を知らないで意味がよくわからず、面白くないということである。このように駄洒落を言われた人がどのような背景知識を持っているかということも面白さを感じる際の重要な要因となる。

- 8661, 横須賀は要港すか?, (横須賀) は [要港 すか] ?, 2, 2, 2, 2, 2.00

これと同様に以下に示す 47709 は吉田拓郎の「人間なんて」という歌の一節を種表現とした駄洒落であるが、評価者 3 名が 30 代、20 代という年齢であることから、この歌を知らなかったため、何を元にした駄洒落であるかわからず評価者全員 1 点となったものと考えられる。これも背景知識の問題である。

- 47709, インゲンなんて、ララー、ラー、ラララ、ララー, [インゲン なんて、ララー、ラー、ラララ、ララー], 3, 1, 1, 1, 1.00

また、以下に示す 46698 は、数字だけの駄洒落であるが、「4173」が例えば「良い波」であることを推測することができず、1 点となった。これは、単独の駄洒落が数字だけで存在しても解釈が多様であり、駄洒落と理解できないためであると考えられる。

- 46698, 4173, [4 1 7 3], 3, 1, 1, 1, 1.00

3.3.5 評価者全員のスコアが異なったものの例

表 10 に評価者 3 名のスコアが異なったものの例を示す。以下に示す表 10 の 655 では、評価者 1 が 5 点を付けたにも関わらず評価者 2 が 3 点、評価者 3 が 2 点を付けたものである。種表現である名詞の「アルミカン」

表 9: 評価者 3 名のスコアが一致したもの例

通し 番号	原型	タグ付	種別	スコア			
				1	2	3	平均
8639	あのテニスの選手が持っている もの何?」「…手にスプレーや!!」	「あの(テニスの選手)が持っているも の何?」「…[手にスプレーや]!!」	2	4	4	4	4.00
26680	『ジュース買ってこようか?お金 ある?』『Nomoney!<飲まねえ>』	『ジュース買ってこようか?(お金ある)?』『[No money]!<飲まねえ>』	2	4	4	4	4.00
39947	ニーチェ「神は死んだ!」 父「髪は死んだ!」	ニーチェ「(神)は死んだ!」 父「[髪]は死んだ!」	1	4	4	4	4.00
27	秋田はあきた	(秋田)は[あきた]	1	3	3	3	3.00
8661	横須賀は要港すか?	(横須賀)は[要港すか]?	2	2	2	2	2.00
47709	インゲンなんて、ララー、ラー、 ラララ、ラーラ	[インゲンなんて、ララー、ラー、 ラララ、ラーラ]	3	1	1	1	1.00
46698	4173	[4 1 7 3]	3	1	1	1	1.00

を変形表現として動詞の「ある」と名詞の「ミカン」に分解したことを面白いと感じるかどうかで評価が分かれたものと考えられる。評価者には可能な限り直感的に面白いと感じるかどうかで評価して欲しいとお願いしたが、このように駄洒落の作り方がうまいと感じる場合に高評価を付ける評価者もいて、評価の基準が異なってしまう場合が存在した。

- 655, アルミカンの上にあるミカン♪, (アルミ カン) の 上 に [ある ミカン] ♪, 1, 5, 3, 2, 3.33

次に、以下に示す 16513, 29172 については、評価者 1 と評価者 2 が 5 点あるいは 4 点という高い評価をしたにも関わらず評価者 3 が 2 点という低い評価をしたものである。16513 については、種表現「アキバ」に対して「萌え」を連想語と捉え、そこから「萌え尽き症候群」という変形表現を作り出す過程を佐山らの文献 [13] に示す駄洒落を理解する場合の「処理負荷が大きすぎて面白くない」と感じるか「処理負荷がちょうどよく面白い」と感じるかの違いであると考えられる。評価者 3 については、全体的にスコアが低いので、そもそも単独の駄洒落を面白いものと感じる意識が低く、処理負荷が少しでも大きいと面倒なので面白くないと感じてしまった可能性がある。また、29172 については、対話型になっていて、これを短い文脈と捉えたことや全く同じ読み方で文節区切りを変更することにより面白さを感じたものである。これを面白いと感じるかどうかについての意見が分かれたものと考えられる。文節区切りの変更による駄洒落の面白さについては金久保の文献 [21] で確認されている。

- 16513, アキバで流行ってる病気は、萌え尽き症候群、(アキバ) で 流行っ て る 病 気 は 、 [萌え 尽き 症候群], 2, 4, 5, 2, 3.67
- 29172, 「オラフ、武器は?」「おら、吹雪!」, 「(オラフ、武器) は ?」 「[おら、吹雪]!」, 2,

5, 4, 2, 3.67

以下に示す 21472 は、一つの種表現が二つの変形表現に対応している場合で、これをうまいと感じて高い評価を付けるか直感的にみて面白くないと感じるかの違いである。

- 21472, ロシアの殺し屋、恐ろしや, (ロシア)12 の [殺し屋]1、[恐ろしや]2, 22, 4, 3, 2, 3.00

以下に示す 19071 については、「蛍の光」の駄洒落である。1 点は駄洒落ではないものも含むので、非常に低い評価となり、駄洒落であることはわかるが特に面白くないものは普通の 3 点を付与することを想定していたが、容易に推測がつくものと考え、評価者 2 は 1 点を付けている。これも評価基準の認識に相違があったものと考えられる。このように評価基準が揺れるものと考え、できるだけシンプルに「直感的に面白いと感じるもの」という条件だけを付けたが、逆に細かな条件を付けることで評価基準の揺れを防ぐことができたものと考えられる。一方、詳細な条件を付けると本当に面白いものではなく、うまいと感じるかどうかという評価になってしまう可能性が高くなってしまう。このようなことのないように評価方法の条件を適度に詳細に設定することは今後の課題である。

- 19071, 小樽の光、窓の雪～♪, [小樽 の 光、窓の雪～♪], 3, 3, 1, 2, 2.00

以下に示す 19860 は、評価者 2 は「折り紙付きの折り紙だよ」を通常の文とみなし、1 点としている。

- 19860, 折り紙付きの折り紙だよ, (折り紙付き) の [折り紙] だ よ, 2, 3, 1, 2, 2.00

「折り紙付き」の場合、多義語であり、以下に示す二つの意味がある。

1. 「折り紙」がついていること
2. 絶対に間違いないと保証できること

1の意味と解釈すれば駄洒落ではなく通常の文となり、2の意味と解釈すれば駄洒落となる。このように多義語に対する解釈の違いでスコアが異なる場合がある。

また、以下に示す66885については、「1個IN」をどう解釈するかである。「1個入っている」というように解釈できれば意味がわかるが、理解できなければ1点となる。

- 66885, コインがここに1個IN,(コイン)がここに[1個IN],1,2,3,1,2.00

3.4 平均スコア上位の例

表11に平均スコア上位10位までを示す。表11に示すように平均スコアの1位は4.33である。2位の4.00のものは14件あったので14件中の9件を示している。

平均スコアが1位のものは以下に示す99である。99は評価者2と評価者3が5点を付けていて、評価者1は3点である。特に図3に示すように評価者3が5点を付与したものは67,000件中1件だけだったので、評価者3にとっては非常に面白かったということになる。この99の駄洒落の面白さの要因を考察する。種表現の「みかん」が名詞であるのに対して、変形表現の「見つかんない」が動詞と品詞が変化していることや変形表現に促音の「っ」が挿入されていることがあげられる。これに加えて、「3日間」という語も読みが「みっかかん」ということで「みっか」という部分は読みが変形表現と一部一致していることも語感の良さを高め面白さを倍増させているということが考えられる。

- 99, 私のみかんが3日間、見つかんない・・・, 私のみかん)が3日間、[見つかんない]・・・,2,3,5,5,4.33

次に、2位の平均スコア4.00について見てみると以下に示す132,667,1698,2134や表9の8639,26680,39947の7件が対話形式になっている。これは、駄洒落が単独で存在するよりも何らかの文脈で発せられた方が面白さを感じることを示しているものと考えられる。

- 132, 「インドア派のあなたがインド? あは♪」, 「(インドア派)のあなたが[インド? あは♪]」,1,5,4,3,4.00

- 667, 男「イブの日は空いています?」女「相手います。」, 男「イブの日は(空いています)?」女「[相手います]。」,1,3,5,4,4.00

- 1698, 二人でうまいもの食ったって? グルめっ,(二人)でうまいもの食ったって?[グルめっ],2,3,5,4,4.00

- 2134, 囲碁愛好家の初対面の挨拶は、「囲碁よろしく!」,(囲碁)愛好家の初対面の挨拶は、「[囲碁よろしく]!」,2,4,5,3,4.00

また、以下に示す24,162,315については、種表現と変形表現の間で品詞の変化と文節区切りの移動がある点が面白さの要因と考えられる。24は、種表現「抱いた遺骨」の品詞が動詞、名詞から変形表現「大腿骨」という名詞に変化し、162についても、種表現の「辛子高菜」と変形表現の「枯らしたかな」の間で品詞の変化と文節区切りの移動がある点が面白さの要因と考えられる。これは、315についても同様である。

- 24, 抱いた遺骨が大腿骨,(抱いた遺骨)が[大腿骨],1,3,5,4,4.00
- 162, 辛子高菜を枯らしたかな?,(辛子高菜)を[枯らしたかな]? ,1,3,5,4,4.00
- 315, 内股を打ちしまったー,(内股)を[打ちしまったー],2,4,5,3,4.00

また、以下に示す3164については、種表現「栗」と変形表現「クリック」の間で、促音の挿入と名詞からサ変名詞への品詞の変化、4438については種表現「髪型」と変形表現「毛根な時間」のように連想による駄洒落が面白さの要因であると考えられる。

- 3164, 栗をクリックすると痛い,(栗)を[クリック]すると痛い,1,5,4,3,4.00
- 4438, 髪型を気にしてたら、毛根な時間,(髪型)を気にしてたら、[毛根な間],2,3,5,4,4.00

4 おわりに

現在広く普及しているスマートフォンやAIスピーカーなどの対話エージェントは、雑談システムとして

表 10: 評価者 3 名のスコアが異なったものの例

通し 番号	原型	タグ付	種別	スコア			
				1	2	3	平均
655	アルミカンの上にあるミカン♪	(アルミ カン) の 上 に [ある ミカン] ♪	1	5	3	2	3.33
29172	「オラフ、武器は?」「おら、吹雪!」	「(オラフ、武器) は?」 「[おら、吹雪] !」	2	5	4	2	3.67
16513	アキバで流行ってる病気は、萌え尽き症候群	(アキバ) で 流行 っ て る 病 気 は、[萌え 尽 き 症 候 群]	2	4	5	2	3.67
21472	ロシアの殺し屋、恐ろしや	(ロシア)12 の [殺 し 屋]1、[恐ろし や]2	22	4	3	2	3.00
19071	小樽の光、窓の雪～♪	[小樽 の 光、窓 の 雪～♪]	3	3	1	2	2.00
19860	折り紙付きの折り紙だよ	(折り紙付き) の [折り紙] だ よ	2	3	1	2	2.00
66885	コインがここに 1 個 I N	(コイン) が こ こ に [1 個 I N]	1	2	3	1	2.00

表 11: 平均スコア上位 10 件

通し 番号	原型	タグ付	種別	スコア			
				1	2	3	平均
99	私のみかんが 3 日間、見つからない...	私 の (みかん) が 3 日 間、[見 っ か ん ない] . . .	2	3	5	5	4.33
24	抱いた遺骨が大腿骨	抱 い (た 遺 骨) が [大 腿 骨]	2	3	5	4	4.00
132	「インドア派のあなたがインド? あは♪」	「(インドア 派) の あ な た が [イン ド ? あ は] ♪」	1	5	4	3	4.00
162	辛子高菜を枯らしたかな?	(辛子 高 菜) を [枯 ら し た か な] ?	1	3	5	4	4.00
315	内股を打ちしまったー	(内 股) を [打 っ ち ま っ た ー]	2	4	5	3	4.00
667	男「イブの日は空いています?」女「相手います。」	男 「イブ の 日 は (空 い て い ま す) ? 」 女 「[相 手 い ま す] 。」	1	3	5	4	4.00
1698	二人でうまいもの食ったって? ニグルめっ	(二 人) で う ま い も の 食 っ た っ て ? [グ ル め っ]	2	3	5	4	4.00
2134	囲碁愛好家の初対面の挨拶は、「囲碁よろしく!」	(囲 碁) 愛 好 家 の 初 対 面 の 挨 拶 は、 「[囲 碁 よ ろ し く] !」	2	4	5	3	4.00
3164	栗をクリックすると痛い	(栗) を [ク リ ッ ク] す る と 痛 い	1	5	4	3	4.00
4438	髪型を気にしたら、毛根な時間	(髪 型) を 気 に し て た ら、[毛 根 な 時 間]	2	3	5	4	4.00

注) 平均スコア 4.00 はこれ以外に表 9 の 3 件を含む 5 件存在した。

はその能力が不十分で話していて楽しいシステムとはなっていない。このような問題を解決するためにはユーモア処理の高度化が必須であり、その第一段階として68,000件まで拡張を行った駄洒落データベースの開発を行った。しかし、この駄洒落データベースには面白さの度合いを示すスコアが付与されていなかったため、3名の評価者をクラウドソーシングを用いて雇用し、面白さのスコアを付与した。さらに、面白さを付与された駄洒落データベースを一般公開するために修正を行った。この修正により公開用の駄洒落データベースは67,000件となった。

次に、面白さのスコアが付与された公開用駄洒落データベースを用いて行った面白さの評価結果の分析を行った。まず、最初に平均値の推移を分析したところ評価者ごとに単独の駄洒落に対する面白さの感じ方や飽き方には個人差が大きいことが明らかになった。また、評価者ごとのスコアの分布の分析を行った、この結果、評価者1と評価者2は類似した分布であるが、評価者3は大きく異なっていることが明らかとなった。

次に、評価者間のスコアの一致度について k 係数を用いて評価を行った。スコアの分布から推測されたように k 係数は、いずれも0.1未満と低く、わずかな一致が見られる程度であった。その中でも評価者1と評価者2は0.079と評価者3に対する評価者1と評価者2の0.022,0.046に比べて高い値になっていることから評価者1と評価者2のスコアの一致度は評価者3との間での一致度に比べて比較的高く、評価者3が面白さの感じ方が大きく異なっていることが明らかとなった。

さらに、3者一致、2者一致、3者不一致についても分析を行った。その結果、2者一致しているものが約7割を占め、3者一致しているものは約2割、3者不一致のものが約1割であることが明らかとなった。これらの分類の具体的な例について面白さの要因の分析を行った。また、平均スコアが上位のものについても具体的な例を用いて分析を行った。その結果、対話形式で書かれている駄洒落が短い文脈を持っていることから面白さを感じることや、背景知識の違いによって面白さの評価が分かれることが明らかとなり、面白さにおいて文脈情報や背景知識が重要な役割を担っていることが確認された。

今後の課題としては、今回の分析により対話形式で発せられる駄洒落が面白さを感じるのに大きな要因となっていることが確認されたので、単独の駄洒落しか収録されていない駄洒落データベースをさらに拡張し、駄洒落が発せられる際の対話中の文脈情報も収録することにより、どのような文脈でどのような駄洒落を発することが面白さを増加させるのかの分析を行う予定である。また、駄洒落以外のストーリー性のあるユーモアや漫才やコントのシナリオなどの他のユーモアについても研究を進める予定である。

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: 日本語駄洒落などなど生成システム「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, ジェブカ・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, ジェブカ 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 ことば工学会研究会資料, 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 ことば工学会研究会資料, 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? 一駄洒落の面白さを認識する理解の手続き, 人工知能学会第2種研究会 ことば工学会研究会資料, 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 ことば工学会研究会資料, 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] 荒木健治, 佐山公一, 内田ゆず, 谷津元樹: 駄洒落データベースの拡張及び分析, 人工知能学会第2種研究会 ことば工学会研究会資料, SIG-LSE-B803-1, pp.1-15, 2018.
- [17] 佐山公一, 荒木健治: 駄洒落理解過程のモデル化-駄洒落の事例を手続きで分類する- 人工知能学会第2種研究会 ことば工学会研究会資料, SIG-LSE-B803-2, pp.17-27, 2018.
- [18] 内田 ゆず, 荒木 健治: オノマトペに着目した駄洒落の面白さの分析-駄洒落の自動生成に向けて-, 第35回ファジィシステムシンポジウム 講演論文集 (FSS2019), pp. 332-336, 2019.
- [19] 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.
- [20] 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.
- [21] 金久保正明: 形態素解析手法と通俗的単語群に基づく類音文変換システム, 情報処理学会論文誌, Vol.54, No.7, pp.1937-1950, 2013.