# 機器の不具合を記述した日本語と英語のコーパスにおけるオノマトペ

那須川 哲哉　　海野 裕也　　村上 明子
日本アイ・ビー・エム株式会社 東京基礎研究所

## 1. はじめに

機器の不具合を報告する際、特に日本語においては、音を示すオノマトペを用いて状況を表現することがある[1,2]。例えば、筆者らがテキストマイニング[3]の対象とした経験のあるデータのうち、PC ヘルプセンターにおけるコンタクト記録[4]においては、『電源投入すると「ギュッ」という蛙を踏んだような音が鳴り起動し始めます。』のように不具合が伝えられ、『もし電源投入時の「ギュイーン」という音であれば、電源のノイズであるため、特に問題はないと説明。』と対応したという報告を記録したデータが存在し、オノマトペを問題判別に活用している状況が示されている。また、自動車の不具合に関しても、筆者らの経験上、エンジニアはオノマトペを重要視しており、例えばマフラー周辺の異音が「ガラガラ」と表現される場合と「カラカラ」と表現される場合では推測される不具合のタイプが異なるという話を聞いている。

オノマトペは、感覚的な表現であり、日本語を母語としない日本語学習者にとっては習得が困難な表現と言われている点で、言語表現としての特殊性を有している。このオノマトペの使われ方の特徴やその処理の可能性を調査する一手法として、我々は、テキストマイニングを試みた。対象データは、音を示す擬音語オノマトペが機器の不具合を表現するために用いられているコーパスである。まず、コーパス内から擬音語オノマトペを抽出し、各オノマトペが擬音語として用いられているデータを特定した上で、個々のオノマトペ表現の出現傾向や各表現を含むデータの特徴を分析した。本稿ではその分析内容と分析結果を示し、オノマトペの特徴と、その処理の可能性に関して考察する。

## 2. 分析対象データと分析手法

機器の不具合が記述されているテキストのコーパスとして、日米において自動車の不具合を報告している日本語と英語のデータ、及び PC ヘルプセンターにおいて応対の概要が記録された日本語のデータを用いた。

テキストマイニングのツールとしては、IBM TAKMI ® [5] の機能を日本語と英語向けに製品化した IBM® Content Analyzer[1]を用いた。

### 2.1. 分析対象データ

自動車不具合に関する日本語のデータとしては、国土交通省自動車交通技術安全部審査課が収集公開している「自動車不具合情報 [2]」のデータのうち、2001 年 4 月から 2009 年 3 月までの 24,458 件を用いた。米国における英語のデータとしては米国政府組織に属する National Highway Traffic Safety Administration (NHTSA) が収集公開している “Consumer Complaints[3]” のデータのうち、2010 年 1 月 5 日までの 587,623 件を用いた。日米のデータとも基本的に運転手から報告された不具合の概要が自由記述形式のテキストで入力されており、そこに車名や走行距離、不具合装置といった定型データが紐付いている。テキスト部分には、例えば「スイッチを入れてもミラーが全く動かなくなった」「エンジンからキーンという音がする。」“WHILE DRIVING AT ANY SPEED BRAKES MADE A GRINDING SOUND.[4]” のように具体的な不具合の内容が記述されており、個々の自動車会社が収集・蓄積している顧客の声のデータとの類似性が高い。

PC ヘルプセンターにおける日本語のデータとしては、1997 年 7 月から 1998 年 4 月にかけて、当時の日本 IBM（株）お客様相談センターの PC ヘルプセンターへ寄せられた[4]と同様の問合せ記録のうち 324,677 件を対象とした。テキスト中には「Q：しばらく前からスピーカーからキーンという音がする。A：コンセントの位置や電化製品を遠ざけて下さい。」のように、顧客からの問合せ内容の概要が「Q：」から、対応の概要が「A：」から始まる形で、どのような問合せにどう対応したか記述されている。

### 2.2. 音を示す表現の抽出

#### 2.2.1. 日本語データからのオノマトペの抽出

日本語データから音を表現するオノマトペを抽出するためテキストマイニングツールのパタン抽出機能を用いた。まず、

- 「○○」という音
- □□という音

という表現に含まれる○○及び□□の部分の文字列を抽出し、その内容を確認した上で、オノマトペ以外の表現の抽出を避けるため、

- ○○の先頭が平仮名もしくはカタカナ
- □□を構成する全形態素[6]の先頭がカタカナ

の場合のみ抽出するという制約を加えた。さらに、抽出対象を増やすため、

- 「○○」と音

というパタンを加えると共に、「音」の代わりに、末尾が「音」である形態素（「異音」や「雑音」など）にもマッチするようにした。

こうして抽出された表現をオノマトペ候補として辞書登録した。その際、オノマトペ以外の意味でも使われる可能性が高い一部の表現（例えば鍵を示す「キー」など）は除くようにした。

次に、各オノマトペ候補の表現がデータ中に出現する際、

- 後続する文字が「に」でない

[1]現時点では、対応言語を 11 言語に拡張した IBM Cognos ® Content Analytics という製品がこの機能を継承している。

[2] http://www.mlit.go.jp/jidosha/carinf/rcl/index.html

[3] http://www-odi.nhtsa.dot.gov/complaints/

[4] NHTSA の“Consumer Complaints”のデータにおいて、英語のテキスト部分は全て大文字で記述されている。

（擬音語のオノマトペに「に」には接続しないという特徴があるため[7]）

- 前後にカタカナの形態素が存在しない（例えば「ピー」という表現が形態素解析時に未登録語の「ピアツーピア」のような表現から抽出されるのを避けるため）

という制約を充たした場合に、擬音語オノマトペとして認識し、抽出するようにした。

#### 2.2.2. 英語のデータからの音を示す表現の抽出

英語のデータにおいて、音がどの様に表現されるかを分析するため、日本語データと同様にテキストマイニングツールのパタン抽出機能を用いて

- make XXX sound
- hear XXX sound
- make XXX noise
- hear XXX noise

という表現に含まれるXXXの部分に相当する単語列を抽出した。ここで make, hear, sound, noise は part-of-speech (POS) tagger で処理した語幹表現にマッチする。XXX は5単語以内の単語列であり、そこから先頭の冠詞を除いた単語列を抽出した。例えば、“THE BRAKES WERE MAKING A VERY LOUD GRINDING NOISE.”という文から、“VERY LOUD GRINDING”が抽出される。

また、実データを処理した抽出結果をふまえ、

- XING sound
- XING noise

のように、sound もしくは noise の直前に出現し、語尾に ING を含む単語も抽出するようにした。但し、MAKING, HEARING, CAUSING, CREATING, SOUNDING, EXPERIENCING, EVERYTHING の7語は対象外とした。

| キーワード | 頻度 |
|---|---|
| ガタガタ | 93 |
| カタカタ | 44 |
| カラカラ | 29 |
| ガラガラ | 29 |
| キーキー | 22 |
| ガリガリ | 19 |
| ギシギシ | 17 |
| ゴトゴト | 14 |
| ガクン | 13 |

**図 1: 日本語自動車不具合データから抽出された擬音語のオノマトペとその出現データ数**

| サブカテゴリ／キーワード | ガタガタ 93 | カタカタ 44 | カラカラ 29 | ガラガラ 29 | キーキー 22 | ガリガリ 19 | ギシギシ 17 | ゴトゴト 14 | ガクン 13 | ゴトゴト 11 | ガックン 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| エンジン 7562 | 21 0.4 | 17 0.7 | 13 0.7 | 8 0.3 | 2 0.0 | 5 0.2 | 1 0.0 | 1 0.0 | 0 0.0 | 1 0.0 | 2 0.0 |
| 動力伝達 3464 | 13 0.5 | 6 0.3 | 5 0.3 | 4 0.2 | 1 0.0 | 6 0.6 | 1 0.0 | 3 0.2 | 12 3.3 | 0 0.0 | 6 1.1 |
| 制動装置 2438 | 21 1.3 | 2 0.0 | 1 0.0 | 2 0.0 | 14 3.3 | 3 0.2 | 0 0.0 | 3 0.3 | 0 0.0 | 0 0.0 | 1 0.0 |
| 保安灯火 1710 | 3 0.1 | 1 0.0 | 1 0.0 | 2 0.1 | 2 0.1 | 1 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 |
| 電気装置 745 | 0 0.0 | 1 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 |
| 緩衝装置 650 | 6 0.7 | 0 0.0 | 0 0.0 | 0 0.0 | 1 0.0 | 0 0.0 | 7 5.0 | 3 1.1 | 0 0.0 | 2 0.5 | 0 0.0 |
| 排ガス・騒音 481 | 0 0.0 | 3 0.5 | 7 3.9 | 9 6.7 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 | 0 0.0 |

**図 2: 擬音語のオノマトペと不具合装置との相関関係[7]**

## 3. 分析結果

### 3.1. 日本語自動車不具合データにおける擬音語オノマトペ

24,458 件のデータ中、オノマトペ候補の抽出パタンを含むデータが 217 件存在し、121 種類の表現が抽出された。そのうち、「キー」「ガタ」「シュー」「バン」という4表現はオノマトペ以外の意味で用いられるケースが多いと判断して除外し、残る 117 表現を辞書登録した。辞書ベースの抽出の結果、抽出パタンを含む 217 件以外の 304 件のデータにおいても擬音語オノマトペが含まれており、合計で 521 件のデータにおいて何れかの表現が擬音語として用いられているという結果が得られた。その頻度分布を頻度順に上位9件まで示したのが図1である。

次に、抽出された各オノマトペと、それを含むデータに紐付いている不具合装置との相関関係を図 2 に示す。縦軸左端のセルに、不具合装置名称と、各装置に関する不具合データの全体的な件数が示されている。最上段のセルの数値は各オノマトペを含むデータの総数である。それ以外のセルにおいて上にある数値は、左端の不具合装置に関するデータのうち上段のオノマトペを含むデータの件数である。各セルの下にある数値は、 左端の不具合装置と上段のオノマトペとの相関を示す指標[8]である。図 2 においては、相関指標が 2 以上のセルがハイライトされており、動力伝達（装置）の不具合は「ガクン」と、制御装置の不具合は「キーキー」と、緩衝装置の不具合は「ギシギシ」と、排ガス・騒音（装置）の不具合は「カラカラ」及び「ガラガラ」と相関が強く、保安灯火（装置）や電気装置の不具合では擬音表現が少ないことが示されている。

---

[7] 図中、保安灯火の行と電気装置の行の間は表示スペースの都合上 6 行省略されている。

[8] 相関指標としては、各オノマトペを含む文書集合をA、各不具合装置に紐付いた文書集合を B として、以下の値に補正を行った値を用いている。 ここで D は 24,458 件の全データ集合、#は文書集合中の文書数を表す。

$$\frac{\#(A \cap B)/\#D}{(\#A/\#D)\,(\#B/\#D)}$$

これは、基本的には、分析対象となる文書集合 B において A が出現する割合と、全体文書集合において A が出現する割合との比を取ったものであり、1を超える値であれば相関が強いということになる。

また、排ガス・騒音（装置）の不具合で「カラカラ」もしくは「ガラガラ」を含むデータにおいては、上記の相関指標が高く、特徴的に出現している表現として「触媒」「音」「マフラー」「内部」が得られた。原文を読むと、「マフラーの触媒が内部で割れて、ガラガラ騒音を発するようになった。」のように、具体的な不具合現象に対して特定の擬音表現が共通して用いられる傾向が見出された。

擬音語オノマトペが使われていると判断された 521 件のデータのうち、93 件においては、「音」という文字が含まれていない。すなわち、「音」という表現を伴わずに、「ベルトがキュルキュルと鳴く。」「止まる時ガタガタいう。」のようにオノマトペのみで音の発生が表現されていた。こういったケースに対応するパタンをオノマトペ抽出用に追加することで、オノマトペをより網羅的に収集できるようになると考えられる。しかし、図 1 に示されているとおり、高頻度の表現は限定されるため、パタンを追加して網羅性を上げても、新たに獲得されるのは低頻度の表現ばかりとなる。したがって、本研究の目的である傾向分析には、今回用いたパタンのみでも充分な量のオノマトペが取られていると判断した。

### 3.2. PC ヘルプセンターの日本語データにおける擬音語オノマトペ

324,677 件のデータ中、オノマトペ候補の抽出パタンを含むデータが 367 件存在し 138 種類の表現が抽出された。この 138 表現を辞書登録した結果、合計で、1,030 件のデータにおいて何れかの表現が擬音語として見出された。その頻度分布の一部を図 3 に示す。図 1 に示された自動車の不具合情報におけるオノマトペとは異なる表現が頻度上位を占めている。

| キーワード | 頻度 |
|---|---|
| ピー | 133 |
| ツー | 53 |
| キーン | 43 |
| ピッ | 37 |
| ピポパ | 36 |
| ジー | 28 |
| ピーガー | 28 |
| ブーン | 27 |
| ブー | 26 |
| カタカタ | 26 |

**図 3: PC ヘルプセンターの日本語データから抽出された擬音語のオノマトペとその出現データ数**

次に、抽出された擬音語オノマトペと共起しやすい表現を図 2 と同様の相関分析で調査したところ、擬音語オノマトペと相関の高い表現は、「スピーカ」「コンセント」「マウス」「キーボード」など、主にハードウェア関係の表現であった。また、「スピーカ」と相関の高い擬音語オノマトペを相関の高い順に並べると「ギュイーン」「ギューン」「ギュイン」「ブーン」「キーン」「バリバリ」「ビー」「ザーザー」「ザー」「ビーン」のようになった。「ギュイーン」「ギューン」「ギュイン」はいずれも「電源投入」と相関が高く、「キーン」という音は「ハウリング」と相関が高い。すなわち、自動車不具合データにおける出現傾向と同様に、各オノマトペは特定の不具合に関連している。この性質により、「ギュイーン」「ギューン」「ギュイン」のようなオノマトペの意味的な近さを、共起表現から評価できる可能性がある。

### 3.3. 英語の自動車不具合データにおける音の表現

587,623 件のデータ中、sound もしくは noise を含むデータは 41,004 件存在し、そのうち、調査に用いた

(make OR hear) XXX (sound OR noise)

というパタンを含むデータは 11,780 件存在した。そこから抽出された XXX の部分にくる（冠詞を除いた）単語列の頻度分布を図 4 に示す。

| キーワード | 頻度 |
|---|---|
| LOUD | 1954 |
| GRINDING | 920 |
| POPPING | 539 |
| KNOCKING | 392 |
| CLUNKING | 323 |
| THUMPING | 216 |
| LOUD POPPING | 205 |
| SQUEAKING | 192 |
| LOUD GRINDING | 189 |
| CLICKING | 182 |
| STRANGE | 169 |
| SAME | 163 |

**図 4: 英語の自動車不具合データから抽出された音を形容する表現とその出現データ数**

音の大きさ、異常性、同一性を示す LOUD、STRANGE、SAME を除くと、音の種類を形容する高頻度の表現はいずれも ING 形をとっている。そこで、2.2.2 で示したとおり、sound もしくは noise の直前に出現する ING 形を抽出した。その結果、587,623 件のデータ中、191,93 件のデータから抽出された。この ING 形の表現と、それを含むデータに紐付いている不具合装置（COMPONENT DESCRIPTION）との相関関係を図 5 に示す。図 2 と同様に特定の表現が特定の装置の不具合と強い相関関係にある様子が示されている。

| サブカテゴリ／キーワード | GRINDING 3518 | POPPING 1874 | KNOCKING 1603 | CLUNKING 1560 | SQUEAKING 823 | THUMPING 646 | WHINING 564 | TICKING 534 | CLICKING 526 |
|---|---|---|---|---|---|---|---|---|---|
| POWER TRAIN:AUTOMATIC TRANSMISSION 37905 | 378 1.5 | 121 0.8 | 145 1.1 | 242 2.0 | 35 0.4 | 49 0.8 | 154 3.5 | 46 0.9 | 32 0.6 |
| SERVICE BRAKES, HYDRAULIC:ANTILOCK 32292 | 585 2.7 | 79 0.6 | 56 0.5 | 55 0.5 | 98 1.7 | 36 0.7 | 9 0.1 | 20 0.4 | 46 1.1 |
| ENGINE AND ENGINE COOLING:ENGINE 27997 | 83 0.4 | 101 0.9 | 396 4.6 | 62 0.6 | 20 0.3 | 18 0.3 | 23 0.5 | 174 5.7 | 24 0.6 |
| VEHICLE SPEED CONTROL 19409 | 56 0.3 | 29 0.3 | 16 0.2 | 24 0.3 | 5 0.0 | 8 0.1 | 7 0.1 | 10 0.3 | 9 0.2 |
| AIR BAGS:FRONTAL 18370 | 17 0.1 | 13 0.1 | 4 0.0 | 0 0.0 | 2 0.0 | 4 0.0 | 0 0.0 | 1 0.0 | 4 0.0 |
| ELECTRICAL SYSTEM 18123 | 74 0.5 | 53 0.6 | 25 0.3 | 32 0.4 | 17 0.4 | 14 0.4 | 20 0.7 | 28 1.1 | 55 2.4 |

**図 5: 音を形容する ING 形表現と不具合装置との相関関係**

### 3.4. 日本語と英語の自動車不具合データにおける音の表現の対応付け

日米の自動車不具合データにおいて、日本語の擬音語オ

ノマトペも、英語の ING 形表現も、特定の不具合に相関が高い。すなわち、特定の不具合を示す文脈に特徴的な表現が使われる傾向が認められた。日米のデータは、対訳コーパスでないが、同じような内容(ここでは自動車不具合)を記述したコーパスである。そのため、出現文脈において辞書登録されている語の対訳表現を比較して対訳候補を抽出する手法[8]によって、日本語の擬音語オノマトペと、英語の音を形容する ING 形表現の意味的対応性を見出せる可能性がある。

そこで実際に、出現頻度の高い「ガタガタ」「カタカタ」「カラカラ」「ガラガラ」「キーキー」の 5 語に対する対訳候補を抽出してみた。その結果、英語の音を形容する ING 形表現としては、唯一 RATTLING が、最も頻度の高い「ガタガタ」の対訳候補の中に含まれていた。sound もしくは noise の直前に出現する RATTLING を含むデータの件数は 229 件で、データ全体における頻度の順位は 19 位であり、決して高くない。従って、文脈内容の同等性の比較がうまく機能して、「ガタガタ」と RATTLING が結び付けられた可能性が高い。出現頻度の低い他の表現では、対訳候補に ING 表現が含まれなかったものの、「カラカラ」及び「ガラガラ」の対訳候補のトップが exhaust system であり、「キーキー」の対訳候補に brake が含まれていた。文脈内容の同等性が多少は捉えられていると考えられる。

## 4. 考察

自動抽出パタンを用いた分析のため、必ずしも網羅的ではないものの、器具の不具合を表現した日本語のテキストにおいては、特定の不具合に特徴的な擬音語オノマトペが使われる傾向を見出すことができた。自動車不具合のデータと PC ヘルプセンターの問い合わせ記録における擬音語オノマトペの分布は大きく異なっており、分野や不具合によってオノマトペが使い分けられている。

全体的に件数が少ないものの、図 1 及び図 3 に示された頻度分布が Zipf の法則に近い分布となっていることは興味深い。オノマトペは感覚的な表現でありながらも、使われ方に共通性が見られ、高頻度の表現は慣用的な表現となっている可能性が高い。例えば、PC ヘルプセンターの記録における「ギュイーン」「ギューン」「ギュイン」「ぎゅーん」といった表現は、どれかが突出して多用されるようになると、表記が統一される可能性も考えられる。

今回分析した日本語の自動車不具合データにおいては、「音」という表現が含まれるデータが約 2 千件存在した。その中で、擬音語オノマトペが出現するのは 2 割程度であり、残りの大半は「大きな音」「異音」といった音の大小や異常性を示すのみで、「金属音」「破裂音」「爆発音」のように音の性質を示すケースや「空気の漏れるような音」「金属の擦れた音」といった例え方をしているケースは合わせても 1 割に満たなかった。一方、英語のデータでは sound もしくは noise を含む約 4 万 1 千件の半数近い約 1 万 9 千件のデータにおいて、ING 形の表現により音が形容されていた。日本語データでは音の性質の表現におけるオノマトペの役割が大きいことから、英日の対応付けを通じて擬音語オノマトペの意味分類ができるようになれば、その有用性は高そうである。

## 5. おわりに

近年、WEB 全体からオノマトペを収集し、概念辞書を構築する試み[9]や用例事典を開発する試み[10]が増えてきている。本研究では機器の不具合情報を含む限定されたコーパスを対象とし、テキストマイニングツールを用いて各オノマトペが出現する文脈(同じテキスト中の共起表現及びそれに紐付けられた提携情報)の傾向を調査することで、擬音語オノマトペの特徴をつかみ、その応用可能性を探ることを試みた。

結果的に、少なくとも出現頻度の高い表現は使われ方に共通性が見受けられ、慣用的な使われ方をしている可能性が認められた。今回分析対象としたようなコーパスが今後増え、さらに大量のデータを対象とした処理が可能になれば、多様な応用が期待できると共に、感覚的な表現が慣用的な表現になっていく過程を見出せる可能性がある。



## 謝辞

日本語自動車不具合データ中のオノマトペに関しては、2008 年 7 月、筑波大学大学院の坂本明子氏に予備調査をしていただきました。ここに記して感謝致します。

## 参考文献

[1] 高田正幸. 擬音語を利用した音質評価 ー「ドタン」「バタン」から何が分かる?ー. 騒音制御, Vol.26, pp.30-34, 2002.

[2] 田中元八郎, 松原謙一郎, 佐藤太一. 機械の異常音の擬音語表現. 日本音響学会誌, No.53, pp.477-482, 1997.

[3] 那須川哲哉. テキストマイニングを使う技術/作る技術—基礎技術と適用事例から導く本質と活用法. 東京電機大学出版局, 2006.

[4] 那須川哲哉. コールセンターにおけるテキストマイニング. 人工知能学会, Vol.16 No.2, 2001.

[5] T. Nasukawa and T. Nagano. Text analysis and knowledge mining system. IBM Systems Journal, Volume 40, Issue 4, pp.967-984. 2001.

[6] 丸山宏, 荻野紫穂. 正規文法に基づく日本語形態素解析. 情報処理学会論文誌, Vol.35, No.7, pp.1293-1299, 1994.

[7] 黄慧. 日本語のオノマトペに後続する助詞について ー「と」および「に」をめぐってー. コーパスに基づく言語学研究報告, No.1, pp.267-285, 2009.

[8] 那須川哲哉, Daniel Andrade, 海野裕也, 村松祐希, 山本和英. 言語横断テキストマイニングのための翻訳対抽出. 言語処理学会第 15 回年次大会, pp.108-111, 2009.

[9] 奥村敦史, 齋藤豪, 奥村学. Web 上のテキストコーパスを利用したオノマトペ概念辞書の自動構築. 情報処理学会研究報告, 自然言語処理, 154-10, 2003.

[10] 浅賀千里, 渡辺知恵美. Web コーパスを用いたオノマトペ用例辞典の開発. 電子情報通信学会 第 18 回データ工学ワークショップ, DEWS2007 D9-2, 2007.