自 然 言 語 処 理 154−10
（２００３．３．６）

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

奥村 敦史[†]　齋藤 豪[‡]　奥村 学[‡]

概　要：感性を表す言葉であるオノマトペ (擬音語・擬態語) は新語・造語が多く，既存の辞書には語彙が不足している．また，既存の自然言語処理用コーパスにもオノマトペはあまり出現しない．そこで本研究では，自動生成したオノマトペ候補語をクエリとしてWeb上のテキストを検索し，候補語を含む用例を取得することでこれをコーパスとみなす．次に得られたコーパスを解析し，候補語がオノマトペかどうかの判定を行う．オノマトペと判断された語については，係り受け解析結果の頻度情報などを利用し，その語義や用法を得る．最後に，複数の候補語の語義を照らし合わせて，語義間の距離を定義したオノマトペ概念辞書を構築する．

Automatic construction of a Japanese onomatopoeia dictionary using text data on the WWW

Atsushi OKUMURA[†]　Suguru SAITO[‡]　Manabu OKUMURA[‡]

Abstract: Onomatopoeias which express sensibility include many new words and coined words, and the existing dictionaries are insufficient of their vocabularies. Furthermore, onomatopoeias seldom appear in the existing corpus for natural language processing. In this work, we generate candidate words of onomatopoeias automatically and search the text on the Web with a search engine using the candidates as a query. Therefore we can acquire a corpus containing examples of the candidates. Then, we process the corpus and judge whether each candidate is onomatopoeia or not. If a candidate is judged to be an onomatopoeia, we give its sense and usage from results of syntactic analysis, and construct a concept dictionary of onomatopoeias.

1　はじめに

近年，認知科学や心理学などの分野が活発になり，感性を表現する言葉であるオノマトペ (擬音語・擬態語の総称を指す) が注目されるようになってきた．これに伴い，NLP の分野でもオノマトペを含む表現を正しく解析したいという要求が高まっている．

NLP での解析には辞書やシソーラスといった知識源が不可欠であるが，既存の知識源にはオノマトペに関する詳細かつ網羅的な記述はなされていないのが現状である．しかし，大規模な知識源を人手で構築することは非常にコストが高く，さらにオノマトペには新語や用法の変化が多いという性質があるため，オノマトペ概念辞書は自動的に構築するのが望ましい．

知識の自動獲得は1980年代後半からコーパスベースが中心になっている．コーパスから特定の語彙の用法を収集し単語の共起データなどを統計的に処理することで辞書を構築できる．ところが，現在用意されている NLP 用の大規模コーパスは，多くが新聞記事であり，オノマトペの出現頻度や出現語数が少なく，特に新語はほとんど現れない．以上のような背景から，本研究では Web 上のテキストをコーパスとみなし，オノマトペ概念辞書を自動構築する．

Web 上のテキストをコーパスとして用いることの有効性は既に [1, 2] が示しているように語彙数や言い換えの種類が多いという点が指摘されやすい．しかし本研究ではさらに，Web 上のテキストは不特定多数の個人よる文書であるということから，Web 上に存在する大量のテキストをコーパスとして用い，実際に用いられているオノマトペの用例を統計的に解析することにより，網羅的かつ新語などを反映した「生きた」オノマトペ概念辞書の自動構築を目指す．

2　提案手法

我々の提案する手法は，まずオノマトペの候補語を生成し，次に Web 検索エンジンで候補語をクエリとして Web を検索し，候補語を含む用例集を獲得する．さらにその用例を分析し，候補語がオノマトペであるかどうかを判定し，オノマトペであると判断されたものに対して辞書を構築する．

2.1　オノマトペ候補語

オノマトペの候補語生成には，音韻論による言語学的観察 [3] からわかっている語構成の分類を参考にする．オノマトペには「もぐもぐ」,「がったん」,「ちらりと」などいろいろな形のものがあるが，多

[†]東京工業大学大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology
aokumura@lr.pi.titech.ac.jp
[‡]東京工業大学 精密工学研究所
Precision and Intelligence Laboratory, Tokyo Institute of Technology
{suguru,oku}@pi.titech.ac.jp

くのオノマトペは語基とパタンの組み合わせで構成されている．例えば「もぐもぐ」なら語基は「もぐ」であり，パタンは「ABAB」で表される．

本研究では，[3] によって語数が多いとされている，以下の 10 のパタンに当てはまる候補語を生成する．

- ABAB
  がたがた，もぐもぐ，だらだら
- AっBり
  がっかり，ひょっこり
- AんBり
  しょんぼり，すんなり
- ABっと
  がたっと，ぱくっと
- AっBん
  がったん，どっかん
- ABりと
  かたりと，ひらりと
- ABんと
  すとんと，ぽつんと
- AB と
  さっと，はたと
- ABりABり
  ゆらりゆらり
- ABんABん
  ぐるんぐるん

パタンの ‘A’ には 1 文字の平仮名 (「あ」「い」「う」…「か」「き」「く」…「が」「ぎ」「ぐ」…) に加え，2 文字で 1 音節となる「きゃ」「きゅ」「きょ」などが当てはまる．パタンの ‘B’ には，‘A’ に撥音 (「ん」)，促音 (「っ」)，長音 (「ー」) を加えたものが当てはまる．日本語として正しくない文字列「っっ」「んん」などを含む語は生成しない．また「ABAB」のみ，繰り返しの先頭が濁音化する連濁語 (「しみじみ」など) も生成する．

文献 [3] によると，上記パタンの中でも語数が最も多いのは「ABAB」でこれはオノマトペの典型的パタンと考えられる．

また，オノマトペの品詞は多い順に**副詞**，**形容動詞**，**サ変名詞**となっている．本研究ではこれ以外の品詞で用いられる語はオノマトペではないと考える．

2.2 Web からの用例抽出

生成した候補語をクエリとして，検索エンジンを用いて Web 上のテキストを検索する．そしてまず該当件数を調べ，ある閾値以上の件数が得られた候補語については，該当ページの URL を取得する．

取得した URL のページから，候補語を含む用例を抽出する．候補語の共起単語を得る方法としては，単純な文字列ベースの N-gram 手法と，より高度な係り受け解析を行う手法の 2 通りが考えられる．本研究では精度を上げるために，後者の係り受け解析を行う．また，今回は候補語を含む単一文内の係り受け関係によって語義を獲得し，その前後の文 (文脈) は使用しない．

係り受け解析は厳密な処理であるため，入力が正しい文でないと，解析結果にも誤りが出てしまう．また入力文が長い場合にも誤りが起きやすい．したがって，係り受け解析を正しく行うためには，入力文が非文にならないよう維持しながらできるだけ文長を短くすることが望ましい．

多くの Web ページは著者の意図のみによって書かれ，文長に制限がない．そこで本研究では，この多様な Web 上のテキストを正しく解析するために，Web ページから正しく文を切り出すためのフィルタを構築する．フィルタは以下のアルゴリズムで文を切り出す．

1. まずファイルが HTML ファイルであるかプレインテキストであるかを判定する．
   (a) HTML ファイルの場合は，ヘッダやフッタを取り除いたあと，文が跨ぐことのない特定のタグ (<TABLE>, <HR>, <H1>, など) で段落分けをする．また，HTML ファイル内で<PRE>タグ (ソースファイルに書いたとおりに表示する) があった場合は，その中の部分はプレインテキストと同じ扱いをする．
   (b) テキストファイルは空行や罫線を境界に段落分けをする．
2. 段落分けが行われたら，候補語を含まない段落は削除し，候補語を含む段落の中でさらに文を特定する．文の特定とは，段落内での文区切りを特定することであるが，段落内に句読点や「！」「？」などの記号がない場合には，1 行 1 文 (つまり改行が文区切りである) と仮定する．

この他，機種依存文字や半角カナを取り除いたり，掲示板やチャットのログなどによく見られる，次のような引用符を取り除く処理も行う．

```
nobody> 彼は流暢な英語ですらすらと
nobody> 話し始めた。
```

2.3 形態素解析

獲得した用例文を入力として形態素解析を行う．形態素解析器には JUMAN[5] を用いる．JUMAN は入力文を内部辞書と照らし合わせて形態素を特定するが，辞書構築を目的とする本研究において，ツールの内部辞書 (既存辞書) に含まれる語と含まれない語との間で解析の差が生じるのを避けるため，形態素解析器に頼らずに候補語の形態素と品詞を特定する必要がある．候補語に形態素区切りと品詞を与える手順の流れを図 1 に示す．

まず入力文中の候補語の前後に形態素区切りを付与してから形態素解析を行う．JUMAN は空白を 1 形態素とみなし，その空白は透過処理をする (前後の形態素の品詞決定に影響しない)．したがって，候補語の前後に空白を挿入することで強制的に形態素区切りを与えることができる．しかしながらこれだ

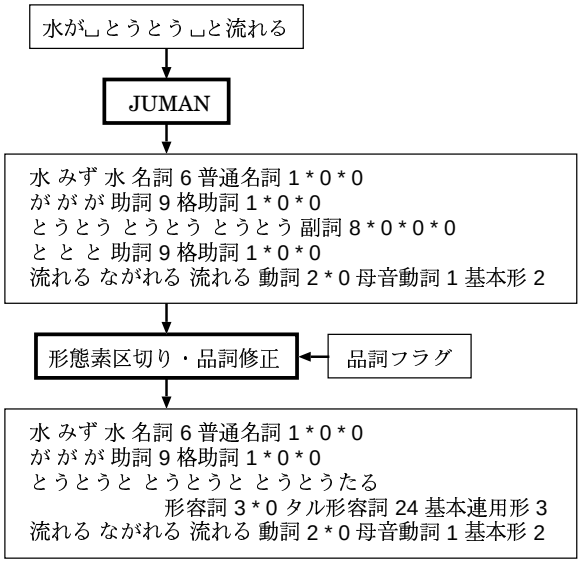


図 1: 候補語の形態素・品詞決定の流れ

けでは正しい形態素や品詞は得られず，オノマトペが形容動詞またはサ変名詞の場合は活用語尾までを 1 形態素としなければならない．

これには候補語の品詞を推定する必要がある．候補語がどのような品詞をとる語なのかを調べるために，ある品詞の場合に典型的に後接すると考えられる文字を候補語に付与し，Web 上のテキストと獲得した用例文の両方に対し，再度検索を行う．その結果，ある閾値以上の用例がある場合には，候補語はその品詞になり得ると推定する手法を採用する．

以降，本節では JUMAN 品詞体系に基づいて形容動詞を「タル形容詞」「ナ形容詞」「ナノ形容詞」に分けて考える．

品詞と検索する後接文字の種類の対応は，次の通りである．

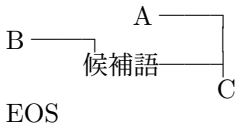
- サ変名詞
  「候補語+**する**」「候補語+**した**」「候補語+**して**」を検索し，該当件数の合計が閾値を超える候補語に**サ変名詞フラグ**を与える．
- タル形容詞
  「候補語+**たる**」を検索し，該当件数が閾値を超える候補語に**タル形容詞フラグ**を与える．
- ナ形容詞・ナノ形容詞
  「候補語+**だ**」「候補語+**な**」「候補語+**に**」「候補語+**の**」の検索を行う.「だ」「な」「に」の該当件数がある閾値を超える候補語に対してその 3 つの合計と「の」の該当件数の比を取り,「の」が閾値を超える割合を占めるなら**ナノ形容詞フラグ**，超えないならば**ナ形容詞フラグ**を与える．ただし,「に」「の」は助詞である場合も考えられるので,「だ」「な」の該当件数をより重視する．
- 名詞
  「候補語+**が**」「候補語+**は**」「候補語+**を**」を検索し，該当件数の合計が閾値を超える候補語に**名詞フラグ**を与える．また，それよりもさらに高い閾値も設け，合計値がその閾値を超える場合にはその候補語は名詞にしかならないと判断し，**非オノマトペフラグ**を与える．
- 副詞
  上記いずれのフラグも与えられず，かつ特に「候補語+**と**」の検索該当件数がある閾値を超える場合，候補語に**副詞フラグ**を与える．
- 感動詞
  上記いずれのフラグも与えられず，かつ「候補語+**，**」の検索該当件数がある閾値を超える場合は感動詞の可能性が高いので，候補語に**非オノマトペフラグ**を与える．ただしこの検索に限っては，Web 検索エンジンで読点の検索をするのが不可能なため，獲得した用例文に対する検索のみを行う．

ここで名詞フラグと非オノマトペフラグを分けたのは，オノマトペにも名詞的用法があるからである (「ぎざぎざが痛い」など)．

次に，非オノマトペフラグが付与されなかった候補語について，用例を一文ずつ形態素解析する．上記の検索で候補語に与えた品詞フラグと形態素解析の出力の後接する形態素とを照らし合わせて，個々の用例文においての候補語の品詞を決定し，形態素区切りの修正と品詞の付与を行う (図 2)．また，後ろの形態素が「さん」「くん」「様」などの名詞性接尾辞だった場合は，直後の形態素が助詞「が」「を」などだった場合と同様に，図 2 の一番下の分岐で候補語の品詞を名詞とする．

2.4 係り受け解析

本研究では，係り受け解析に KNP[6] を使用する．KNP の出力結果の一部が



となっていた場合を考える．候補語が副詞または形容動詞連用形の場合は動詞節 C と名詞節 A の共起を取り出し，候補語がサ変名詞の場合は名詞節 B，候補語が形容動詞連体形の場合は名詞節 C を取り出す．

この操作を全ての用例について行い，統計的な共起情報から辞書を生成する．

2.5 辞書生成

次の 2 つの用例文を獲得した場合を考える．

a) 弱火で 15 分間ことこと煮込む
b) 野菜をとろ火でことこと煮る

これらの文で「ことこと」と共通して共起する単語は存在しないが,「弱火」と「とろ火」,「煮込む」と

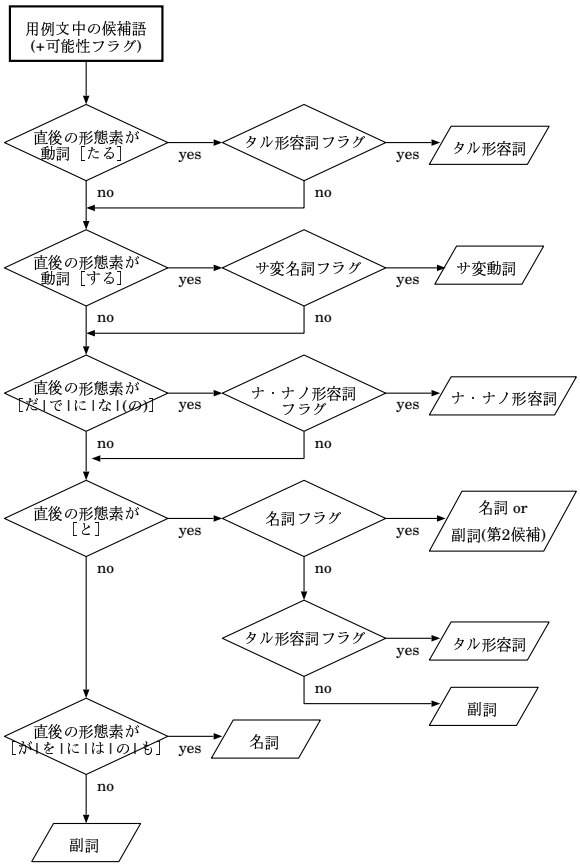


図 2: 品詞割り当ての流れ

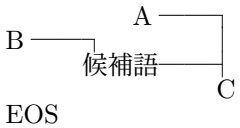
「煮る」はそれぞれ意味が近い．したがって，これらを

c) 野菜を (弱火/とろ火) で 15 分間ことこと (煮込む/煮る)

とすることで共起情報をまとめあげることができる．

以上の操作は単語概念間の距離を定義したシソーラスを用いて実現し，単語概念間のパス長がある閾値以下の場合に組み上げ操作を行う．この操作の際には用例数を合計する．つまり,「煮る」が 10 回出現し「煮込む」が 5 回出現した場合,「煮る/煮込む」は 15 回出現したと考える．この組み上げ操作後の出現回数に対しても閾値を設け，その出現回数を超えない共起語はノイズとみなす．

各候補語について全ての用例を解析し，得た共起情報を以下の手順に従って処理し，辞書を生成する．用例中の係り受け関係が



EOS

のようになっているとして考える．

1. まず名詞節 C が獲得されている場合，つまり形容詞フラグが付与されており，その係り先の名詞がある程度有意に集められている場合を考える．このとき，連用形として用いられる用例によって，名詞節 A+動詞節 C の組み合わせも獲得できているはずである (できていないなら共起が有意ではない)．名詞節 A は，名詞+「が/は」「を」「に」「で」となっているものを獲得しておく．連体形用法から獲得した名詞節 C と連用形用法から獲得した名詞節 A の主格「が/は」の名詞との距離が閾値よりも近いならば，これを有意な共起とみなし，形容詞のオノマトペとして辞書のエントリに加える．
   - 「がらがらの電車」「電車が/空く」→「電車ががらがらに空く」
2. 次に，名詞節 B がある場合 (候補語がサ変名詞の場合) を考える．名詞節 B も上と同様に，名詞+「が/は」「を」「に」「で」を獲得する．ここで，候補語が副詞である用例が存在する場合には名詞節 A+動詞節 C の共起が取れているので，名詞節 B の格要素と名詞節 A の格要素とのマッチングを取り，一致する場合にこれをサ変名詞の取りうる格要素とし，同義表現とみなす．サ変名詞のオノマトペとして辞書のエントリに加える．
   - 「口を/あんぐりする」「あんぐりと/口を/開ける」→「あんぐりする=口を開ける」
   - 「本を/ぱらぱらする」「ぱらぱらページをめくる」→「ぱらぱらする=ページをめくる」
3. 最後に候補語が副詞の場合は，名詞節 A+動詞節 C を獲得しているので，多く共起する動詞 A について，さらに多く見られる A+C の組み合わせを獲得し，副詞のオノマトペとして辞書のエントリに加える．

擬音語にも擬態語にも存在するような多義語 (ex.「かんかん鐘が鳴る」vs「かんかんに怒る」) に対応するため，異なる用法 (つまり，異なる品詞であったり異なる動詞と共起する) 場合には，それらは別エントリとして加える．

また，意味分類は全てシソーラスを用いているため，人間が見ると意味が近いと感じられる動詞でもシソーラスの分類において距離が離れている場合には，それらは別概念として，別エントリとして加えられる．

最終的な辞書の書式は次の通りである．

1. 見出し語
2. 語幹
   - サ変名詞・形容動詞 (JUMAN 品詞分類のタル・ナ・ナノ形容詞にあたる) の語幹
   - 副詞の場合は活用しないため，見出し語と同一

3. 品詞
   - EDR 辞書の表記に習った
   - JD1(副詞)，JN1;JVE(サ変名詞)，形容動詞 (JAM)
4. 擬音語 (SO)/擬態語 (MI)
   - 「(音/声) が (聞こえる/する)」または「(鳴る/鳴く)」または「(音/声) を (立てる/出す)」と共起する場合のみ擬音語とし，それ以外は全て擬態語
5. 係る動詞
   - 見出し語が副詞の場合のみ記載
6. 同義表現 (動詞)
   - 見出し語がサ変名詞の場合のみ記載
   - ex.「じくじくする」に対して「傷が痛む」
7. 用例
   - 元の用例文から直接取ったものではなく，共起した動詞及び格要素から生成する

3 実験

3.1 Web 検索・用例抽出

3.1.1 候補語生成 (1)

まず第一に，オノマトペの典型である「ABAB」パタンの候補語を生成した．生成した語数は 30,867 である．これをクエリとして，検索エンジンを用いて Web 上のテキストから候補語を検索した．Web 検索エンジンは，今回特別に Google* に許可をもらい，これを用いた．

検索結果から，用例を含む Web サイトの URL を取得した．用例抽出に用いるページは，Google のキャッシュではなくオリジナルのものとした．

獲得する用例文数が少ないと統計処理で有意な結果が得られない．獲得 URL 数が少ないと当然用例文数も少なくなるので，今回の実験では時間節約のためにこの段階でフィルタリングを行った．ヒット件数が 100 件未満のものは URL の獲得を行わず，さらに獲得できた URL 数が 100 未満のものは，用例抽出を行わない．

「ABAB」パタンの候補語 30,867 語のうち，Google 検索のヒット件数が 100 件以上だったものは 2,812 語，URL が 100 件以上獲得できたのは 2,148 語あった．

3.1.2 候補語生成 (2)

典型的なオノマトペである「ABAB」パタンの候補語数は，Web 検索によるフィルタリングの結果，15 分の 1 程度にまで削減された．次に，Google で 100 件以上ヒットした「ABAB」パタンの候補語に基づき，他のパタンの候補語を生成した．

*http://www.google.com/

次に，獲得した URL の Web ページから用例文を抽出した．Google は Web ページ中の記号 (句読点や括弧など) や空白を無視して検索するため，実際には文字列としてクエリ語が含まれていないページもヒットする．そこで，前節で取得した URL のページに第 2.2 節で述べたフィルタを適用し，クエリ語である候補語が文字列として正しく含まれる文だけを抽出した．抽出した結果，文数が 100 未満しかなかったものはオノマトペ候補から除外する．

各パタンの生成語数から用例文獲得までの語数を表 1 に示す (「ABりABり」「ABんABん」の 2 パタンの候補語は生成した時期が異なり，Google で 1 件以上ヒットした「ABAB」パタンの語基に基づいているため，他の拡張版よりも語数が多い)．

表 1: 各候補語パタンの語数 (1)

候補語	生成語数	ヒット数 100 以上	獲得 URL 100 以上	獲得文数 100 以上
ABAB	30,867	2,812	2,148	1,732
AっBり	3,534	315	193	178
AんBり	2,006	218	91	70
ABっと	2,432	1,235	608	417
AっBん	2,333	448	218	167
ABりと	2,433	521	247	203
ABんと	2,334	956	396	250
AB と	2,433	2,047	—	—
ABりABり	5,712	101	27	16
ABんABん	5,609	189	47	29

ここで,「AB と」パタンの語は (他のパタンと異なり 3 音節で短い文字列のために) ヒット件数が概して多く，用例獲得に要する時間の都合から今回は実験を見送った．

3.2 用例解析

用例文を 100 以上獲得できた候補語に対し，活用検索を行った．この結果，非オノマトペフラグ (「名詞にしかならないフラグ」を含む) が付与されたものは候補から除かれる．ただし,「ABっと」「ABりと」「ABんと」の 3 パタンの候補語については，パタン内に (元々は助詞の)「と」を含んでいて活用せず，常に副詞になることから，活用検索は行っていない．

非オノマトペフラグが付与されなかった語数を表 2 に示す．これらの候補語に対し，形態素解析及び係り受け解析を行い，その結果から辞書を出力した．

3.3 結果・評価

自動構築した辞書に含まれる，各パタンと品詞の分布を表 3 に示す．ここで「uniq 見出し語」の項は，語義を無視した見出し語の異なり数,「uniq 語幹」の項は活用語の語幹 (「もぐもぐする」の場合は「もぐもぐ」) の異なり数である．

この自動構築辞書を，既存の辞書と比較して評価する．本研究では既存のオノマトペ辞書として EDR 日本語単語辞書 [4]，現代擬音語擬態語用法辞典 [8]，

表 3: 自動構築辞書のパタン・品詞分布

	副詞	サ変名詞	形容動詞	エントリ	uniq 見出し語	uniq 語幹
ABAB	2,948	322	80	3,350	1,025	691
AっBり	365	75	0	440	163	125
AんBり	63	21	0	84	40	29
ABっと	617	0	0	617	239	239
AっBん	9	1	2	12	8	8
ABりと	374	0	0	374	129	129
ABんと	211	0	0	211	88	88
ABりABり	24	0	0	24	12	15
ABんABん	16	2	0	18	14	27
合計	4,627	421	82	5,130	1,718	1,351

表 2: 各候補語パタンの語数 (2)

候補語	品詞フラグ付与
ABAB	1,355
AっBり	164
AんBり	48
ABっと	417
AっBん	147
ABりと	203
ABんと	250
ABりABり	14
ABんABん	28

英辞郎 [7] に収録の音辞郎の 3 つを利用する．EDR 辞書は NLP 用の電子化辞書であり，擬音語と擬態語はエントリの「用法」という項目にそれぞれ “SO”, “MI” と記述されている．現代擬音語擬態語用法辞典は機械用辞書ではなく電子化もされていない．そこで評価に用いるにあたって，見出し語のみを電子データ化して用いた．音辞郎は擬音・擬態表現の和英辞書であり，機械用ではないが電子データとして提供されている．

まず，収録されている語数を本研究の自動構築辞書と既存 3 辞書とで比較した．結果を表 4 に示す．

表 4: 語数比較

辞書	エントリ	uniq 見出し語	uniq 語幹	生成パタン	その他
本研究	5,130	1,718	1,351	1,351	0
EDR	3,077	1,626	1,357	961	396
用法辞典	1,075	1,075	1,216	743	473
音辞郎	8,656	1,049	1,251	920	331

ここで,「uniq 語幹」は活用語の語幹の異なり数であるが，本研究の自動構築辞書は「ABっと」「ABりと」「ABんと」を語幹として扱っている．そこで既存 3 辞書については,「AB(っ)」「ABり」「ABん」というエントリがあった場合に，それぞれ「ABっと」「ABりと」「ABんと」という語幹を補足した．そのため表 4 では，現代擬音語擬態語用法辞典と音辞郎の 2 つにおいて見出し語の異なりよりも語幹の異なりの方が多くなっている．また「生成パタン」の項は，本研究で用いたオノマトペ生成パタンに合致する語幹の数,「その他」は合致しない数である．

以下，既存 3 辞書はオノマトペ生成パタンに合致する語のみを考え，語の一致と書く場合には語幹の一致を指すものとする．

まず本研究の辞書と既存 3 辞書の 4 辞書が相互にどれほど一致しているかを調査した．結果を表 5 に示す．この結果を見ると，自動構築した辞書と既存

表 5: 語の異なり数

	本研究	EDR	用法	音辞郎
本研究のみ	—	649	793	734
EDR のみ	259	—	412	405
用法辞典のみ	185	194	—	245
音辞郎のみ	303	364	422	—

3 辞書との一致は，既存辞書同士の一致数よりもわずかに多い．また，既存辞書に含まれていない語を獲得できていることがわかる．しかし，既存辞書同士でもかなり収録されている語に異なりがあるために，自動構築した辞書と個々の既存辞書との比較をしても正しい考察が得られないことが考えられる．そこで既存 3 辞書を統合したものを評価基準の既存辞書とみなして自動構築辞書との比較を行った．

既存辞書の異なりは 1,447 語で，既存辞書に含まれず自動構築辞書のみに含まれる語は 487 語，逆に既存辞書には含まれるが獲得できなかった語は 583 語，自動構築辞書と一致する語は 864 語あった．既存辞書に含まれる語の再現率は，$864/1,447 = 59.7\%$ である．

既存辞書に含まれず自動構築辞書のみに含まれる 487 語について，人手で精度を求めた．その結果は表 6 のようになり，既存辞書に含まれない新語を 266 語獲得することができた．新語の例を表 7 に示す．また，既存辞書に含まれる語は全て正しいと仮定すると，全体の精度は表 8 のようになる．

3.4 考察

3.4.1 非オノマトペが含まれる問題

前節で述べたように，既存辞書に含まれない 487 語のうちの約 45%はオノマトペではなかった．正しくない語は大きく 3 つに分けられた．

- 名詞+と
  - 「うどんと」,「きゅうりと」
- 助詞 (他の単語の一部)+オノマトペ

表 6: 自動構築辞書のみに含まれる語の精度

	正	誤	計
語数 (uniq 語幹)	266	221	487
割合 (%)	54.6	45.4	100

表 7: 獲得した新語の例

新語	用例	既存の類似語
うぞうぞ	虫がうぞうぞ動く	うじょうじょ
きらんと	目がきらんと光る	きらりと
げしげし	げしげしと蹴る	—
てこてこ	てこてこ歩く	とことこ
にへにへ	にへにへ笑う	にへらにへら
ぱさりと	髪をぱさりと下ろす	ぱさっと
ほてほて	ほてほて歩く	—

表 8: 自動構築辞書の精度

	正	誤	計
語数 (uniq 語幹)	1,130	221	1,351
割合 (%)	83.6	16.4	100

  - 「がどっと」(〜が/どっと),「てはっと」(〜して/はっと)
- 名詞や動詞 (の一部)+助動詞
  - 「あさんと」(おかあさんと),「ねったり」(つねったり，ひねったり，等)

これらはいずれも特定の表現や単語と共起する傾向が見られ，そのためフィルタリングで取り除くことができなかった．

3.4.2 既知のオノマトペが獲得できない問題

既存辞書に含まれるが獲得できなかった 583 語について，どの段階でフィルタリングされたのかを調査した．結果を表 9 に示す．

表 9: フィルタリングされたオノマトペ

フィルタされた段階	語数	割合 (%)
生成 (未生成)	126	21.6
ヒット数 100 未満	45	7.7
URL 数 100 未満	110	18.9
用例文数 100 未満	59	10.1
非オノマトペフラグ	40	6.9
辞書構築	203	34.8
合計	583	100

意外なことに，最終的に辞書構築の段階で取り除かれている語が 203 語もあった．この段階は詳細な言語的解析を行っているため，他の段階でのフィルタリングよりは妥当性が高いはずである．

本研究は，既存辞書に含まれている語であってもWeb 上での頻度が少なければ獲得しない手法を取っている．実際に用例が少なかったためにフィルタリングされた，所謂「死語」の例を表 10 に示す．

これらの語は獲得できなくても誤りではないと考えられるため，獲得できなかった 583 語が本当に

表 10: 獲得できない死語の例

死語	既存辞書による概念説明
おじおじ	おそるおそる行うさま
くらりくらり	繰り返し大きく揺れるさま
じゃぶんと	じゃぶんという音
ちょきりと	はさみで切るときに音が出るさま
まじりまじり	じっと見つめるさま
みしりみしり	きしんで音をたてるさま

「獲得すべき語」なのかを推定するために，評価に用いた既存 3 辞書のうち単一の辞書にしか収録されていない語を調査した．その結果，既存 3 辞書のうちのどれか 1 つにしか出現しない語は 583 語のうち 450 語もあり，それらが表 9 の各段階で占める語数を表 11 に示す．

表 11: 各フィルタリングの妥当性

フィルタ	語数	単一辞書	差	差 (%)
生成	126	102	24	19.0
ヒット数	45	42	3	6.7
URL 数	110	97	13	11.8
用例文数	59	47	12	20.3
非オノマトペ	40	35	5	12.5
辞書構築	203	127	76	37.4
合計	583	450	133	(22.8)

この結果，単一の辞書にしか収録されていない語を一般的ではないオノマトペだと仮定すると，フィルタリングすべき語と同時にフィルタリングすべきではない語が削除されている割合 (表 11 の「差 (%)」) が多いのは，やはり最終段階の辞書構築と，用例文数不足によるフィルタ，候補語生成の段階である．逆にフィルタリングされるべき語だけが多く削除されているのはヒット数のフィルタであり，これは直感的にも正しい．

これらの時間的制約によるフィルタリングを行わずに，全ての候補語を生成し，その用例を最終段階の解析まで行うこともできる．しかし，今回の実験では最終段階の辞書構築手法のフィルタリング精度が最も悪かったため，途中段階のフィルタリングを行わなかったとしても全体のフィルタリング精度は上がらないだろう．今回の実験では，最終段階でのフィルタリングのパラメタ調整をヒューリスティックスにより行ったため，今後はこの点についてより良いパラメタを求めるための実験が必要になるだろう．

3.4.3 多義語に関する問題

これまでは，自動構築辞書で獲得できた語の中で既存辞書に含まれる語は正しいという過程に基づいた考察を行ってきた．しかし，自動構築辞書の中の一部にはオノマトペでないものも含まれる．

例えば,「うとうと」は「うとうと眠る」の他に「うとうと流れる」が獲得されていた．これは「とうとうたる」(「水がとうとうと流れる」) に起因する誤りである．ところが今回の実験では「とうとうたる」は獲得されていなかった．この原因の一つに

は,「とうとう」にはオノマトペの用法以外に,「結局」「ついに」という意味の (オノマトペではない) 副詞が存在することがある．後者の副詞はオノマトペの用法に比べて一般的に非常に多く出現し，そのため「とうとうと流れる」「とうとうたる流れ」がほとんど獲得できていなかった．

この問題に対応するには，品詞決定のための後接文字検索をより重視する必要がある.「と」を付けた「とうとうと」での検索ヒット数が多い場合に,「とうとう」で検索して抽出した用例ではなく「とうとうと」で検索を行って用例を抽出するのである．これは他の形容動詞やサ変名詞の場合にも当てはまる．

さらに「うとうと流れる」が誤りであることを判断するには,「うとうと」の末尾の「と」が「とうとうたる」の活用語尾であることを判断できればよい．これ以外にも「が」「は」「も」などの文字が候補語の末尾に当てはまる場合があるため，助詞や活用語尾と同じ文字を末尾に含む候補語に対しては例外処理を行う必要がある．

同様に，オノマトペではない語義を持つオノマトペに「しばしば」がある．今回の実験では，頻度を表す副詞の「しばしば順番を待つ」などと共に「目をしばしばする」というオノマトペも正しく獲得できていた．この「とうとう」と「しばしば」の結果の差は Web コーパス中の頻度によるものではなく (Google 検索で「とうとう」は 457,000 件,「しばしば」は 248,000 件ヒットした),「しばしば」にはサ変名詞フラグが付与されており，それによってわずかな「目をしばしばする」という用例が辞書構築時に評価されたためである．

3.4.4 より高精度を目指して

全体の精度をより向上させる簡単な方法が一つある．それは最終的に候補語の語義を獲得した段階で，候補語と係り先の動詞から短いフレーズを生成し，それを再び Web 検索エンジンで検索することである．そのヒット数でその語義の尤もらしさを簡単に計ることができる．

また,「和気あいあいと」「どうどう巡り」などの成句として用いられる表現は，第 2.4 節で述べた共起情報だけでは足りず，他の節の共起情報も獲得する必要がある．あるいは N-gram による統計処理も有効かもしれない．

4 おわりに

本研究では，Web 上のテキストデータをコーパスとして用いることで，現在一般に広く用いられているオノマトペの概念辞書を自動構築する手法を提案した．その結果，5,130 エントリ，オノマトペの語幹にして 1,351 異なりを含むオノマトペ概念辞書を構築した．また，既存の 3 つのオノマトペ辞書を統合した，オノマトペ語幹 1,447 異なりを含む辞書と比較して，既存辞書に含まれる語のうち 864 語を獲得し，再現率は 59.7%であった．獲得した語で既存辞書に含まれない新語 487 語について人手で評価したところ，その精度は 54.6%であり，獲得した語が既存辞書に含まれているものを全て正解と仮定すると，自動構築辞書全体の精度は 83.3%であった．既存辞書に含まれるが今回の実験では獲得できなかった語について，辞書構築手法のどの段階でフィルタリングされたのかを調査した．その結果，最終的な用例解析を行い自動構築する段階で候補から除かれているものが 37.4%あり，時間的制約のために行った他の各フィルタリングと比べてもっとも精度が低かった．オノマトペを正しく獲得できるようにするためには，この部分の改良が必要であることが判明した．

本手法はオノマトペの候補語を生成し，Web 検索エンジンを利用して用例文を抽出し統計的に解析することで，候補語の品詞や用法を獲得する．この手法はオノマトペ以外の語彙にも適用できるため，未知語の概念獲得などへの応用が考えられる．また，このように Web から抽出した用例を解析して構築した辞書は当然ながら Web 文書の解析に有利であり，今後ますます増えるであろう Web コーパス関連の研究にとって非常に有意義であるに違いない．本研究の成果はオノマトペ概念辞書の自動構築に止まらず，即時性の強い Web コーパスの利用可能性を示す意味でも大変興味深いものであると自負している．

参考文献

[1] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin, and Andrew Ng. Web question answering: Is more always better? In *Proceedings of SIGIR '02*, pp. 291–298, 2002.

[2] Deepak Ravichandran and Eduard Hovy. Learning surface text patterns for a question answering system. In *Proceedings of the ACL Conference 2002*, 2002.

[3] 田守育啓．日本語オノマトペの音韻形態．筧壽雄, 田守育啓（編）, オノマトピア 擬音・擬態語の楽園, pp. 1–15. 勁草書房, 1993.

[4] (株) 日本電子化辞書研究所．EDR 電子化辞書 2.0 版 仕様説明書, 1999.

[5] 京都大学大学院情報学研究科．日本語形態素解析システム JUMAN version 3.61, 1999.

[6] 黒橋禎夫. 日本語構文解析システム KNP version 2.0 b6. 京都大学大学院情報学研究科, 1998.

[7] 道端秀樹． 英辞郎． 株式会社アルク, http://www.alc.co.jp/, 第 1 版, 2002.

[8] 飛田良文, 浅田秀子．現代擬音語擬態語用法辞典．東京堂出版, 2002.