

Web コーパスを用いたオノマトペ用例辞典

浅賀 千里 (指導教員: 渡辺 知恵美)

1 はじめに

オノマトペとは「どきどき」や「しっかり」などのいわゆる擬態語・擬音語のことである。オノマトペは感覚的なものであることなどから、日本語学習者にとってオノマトペの用例を習得するのは難しい。

オノマトペの意味・用法を理解するには、複数の適切な用例からオノマトペが文中でどのように使われているのかを知ることが有効である。

また、オノマトペは時代と共に意味が変わっていくことから常に新しい用例を知ることが重要となる。

先行研究 [2] にて香林らは「窓際のトットちゃん」からオノマトペの含まれている文章を手動で抽出して翻訳し、3ヶ国語以上に対応したオンライン多言語辞書を構築したが最新のオノマトペの用例まで網羅することはできず、用例の抽出を手動で行っているため開発コストがかかる。

そこで、本研究では Web コーパスから数多くのオノマトペの最新の用例を自動抽出し、日本語学習者のためのオノマトペ用例辞典を開発する。

2 オノマトペ用例辞典

本研究で開発するシステムはユーザがオンライン上で検索したオノマトペを含んでいる文章を Web から抽出し画面上に表示することで、日本語学習者にオノマトペの用例を提示するシステムである。

本辞典では、Web サイトから数多くの最新の用例を抽出するが、用例として適切ではない文章も多く抽出されてしまうのでそれを補うために青空文庫¹からも用例を抽出する。

2.1 システムの流れ

まず、オノマトペリストからユーザが選択したオノマトペを含む Web ページ、あるいは青空文庫のページを Yahoo!API²を用いて検索し自動取得する。そして、これらのページの中からオノマトペを含む文章を抽出し用例とする。また、日本語係り受け解析器 cabocha³を用いてオノマトペに係る語句とその品詞を求める。抽出した文章と情報から用例としての適正を判定し、適正と判断したものをデータベースに格納する。用例提示用の Web サーバを設置し、データベースへの検索インタフェースを提供する [1]。

2.2 スキーマ定義

本システムでは以下のようなテーブルを定義した。

```
onomatopes
(id int, onomatope varchar)
sentences
(id int, onomatope_id int, url varchar,
cache varchar, sentence text,
modword varchar, modpart varchar,
attach varchar, reftitle varchar)
```

onomatopes は ID(id) を付与したオノマトペ (onomatope) を手動で入力したオノマトペリストである。

¹<http://www.aozora.gr.jp/>

²<http://developer.yahoo.co.jp/>

³<http://chasen.org/taku/software/cabocha/>

sentences は自動抽出してきた用例 (sentence)、オノマトペが用例内で係っている語 (modword) とその品詞 (modpart)、付属語 (attach)、抽出元の URL(url) とキャッシュ(cache)、そしてそのページのタイトル (reftitle) を格納するためのテーブルである。それぞれの用例には ID(id) を与える。

3 用例として適正な文章の抽出

本辞典では用例として適正な文章のみを抽出するためにオノマトペの文法的性質を辞典に取り入れる。

3.1 オノマトペの文法的性質を利用した用例の抽出

オノマトペは、語尾に付属語をつけることによって様々な品詞の役割を果たす。オノマトペの品詞と付属語の関係を表 1 に示す。また、オノマトペは「ぼかぼか陽気」など、他の名詞と結合した複合名詞として用いられる場合も多い。

我々は、抽出された文章の用例としての適正度を向上させるために以下の 2 点の手法を提案する。

1) オノマトペに付属語を付けたものを見出し語として検索し、用例を抽出

2) 1) の文章の中で、オノマトペが用言または名詞を修飾するもののみを抽出する。ただし「～だ」と「～する」はその限りでない。

付属語	オノマトペの品詞	例
と	副詞	髪がさらさらと揺れる
な、の	連体詞	さらさらな髪
だ	形容動詞	髪がさらさらだ
する	動詞	髪がさらさらする

表 1 オノマトペと付属語の関係

3.2 文法的性質を用いた用例抽出の精度を調べる実験

本辞典システム用いて、3.1 節の手法を適用した場合、Web ページ、青空文庫から抽出した文章がそれぞれのかなりの割合で用例として適切かを調査するため、「ぼかぼか」や「しっかり」など任意のオノマトペ 10 個に対して以下の手法で用例を抽出し、その割合を調べた。用例を抽出する対象は Yahoo!検索での上位 50 ページである。

1) 付属語をつけて検索した文章の内、オノマトペを含む文章。

2) 付属語をつけずに検索した文章の内、オノマトペを含む文章。

3) 付属語をつけて検索した文章の内、オノマトペが用言又は名詞に係るもの。ただし、係り受け関係の判断は手動で行う。

4) 付属語をつけずに検索した文章の内、オノマトペが用言又は名詞に係るもの。ただし、係り受け関係の判断は手動で行う。

5) 付属語をつけて検索した文章の内、オノマトペが用言又は名詞に係るもの。ただし、係り受け関係の判断は cabocha で行う。

6) 付属語をつけずに検索した文章の内、オノマトペが用言又は名詞に係るもの。ただし、係り受け関係の判断は cabocha で行う。

3.2.1 実験の結果

Web ページから抽出した結果を表 2, 青空文庫から抽出した結果を表 3 に示す。結果はオノマトペそれぞれについて用例として適正, 不適正, また複合名詞であったものの数を出しまとめたものである。

	付属語	品詞限定	適正	複合名詞	不適正
1)	あり	なし	2284	0	762
2)	なし	なし	228	791	564
3)	あり	手動	2236	0	0
4)	なし	手動	228	791	0
5)	あり	cabocha	656	0	151
6)	なし	cabocha	123	109	197

表 2 Web ページからの抽出結果

	付属語	品詞限定	適正	複合名詞	不適正
1)	あり	なし	41	0	0
2)	なし	なし	16	0	3
3)	あり	手動	39	0	0
4)	なし	手動	16	0	0
5)	あり	cabocha	12	0	4
6)	なし	cabocha	9	5	0

表 3 青空文庫からの抽出結果

3.2.2 実験の考察

表 2 の 2) では, オノマトペ単体を文章としているものと複合名詞が多く抽出された。今回抽出された複合名詞は「ぼかぼか日記」など用例として不適正なものばかりであったので, 1) で付属語をつけ, それらの文章を外すのは効果的と考えられる。

表 2 の 3) と 4) はそれぞれ 1) と 2) で抽出された文章の内, オノマトペの係っている語の品詞で用例を限定したもの(係り受けの判断は手作業で行った)であるが, 1), 2) で抽出された「のんびりと。」(タイトル)やオノマトペの係っている語が省略されている文章などが 3), 4) では除去され, 不適正な文章は全く抽出されなかった。このことからオノマトペの係っている語の品詞で用例を限定することは有効であると考えられる。

本辞典では cabocha を用いてオノマトペの係っている語とその品詞を抽出する。表 2 の 5), 6) は cabocha で係り受けの抽出を行ったものである。手動で係り受けを抽出した表 2 の 3), 4) と比べ, 5), 6) では不適正な文章が抽出されてしまっている。その原因は cabocha の係り受けの誤判断であることが多い。例えば「ぼかぼかは天然鉱石を精製した製品で...」という例では, 「ぼかぼか」は「製品」に係ると誤判断された。cabocha ではいくつかのルールをベースに係り受けを判断しており, 誤認識の例にあるような固有名詞としてのオノマトペの利用などは想定されていないことが原因として考えられる。また, Web ページから正確な文章のみを抽出できているわけではないので, cabocha が対処できないことがある。そこで, これらの対策として, 用例を手動で修正することを考えている。

表 2 では「ぼかぼかな親子」など文法的には適正であるがあまり一般的には使われないような文章がいくつか抽出された。今回はこれらの文章も適正な文章の内に入れて, 「ぼかぼかな陽気」など万人が理解できるようなものと区別するため, 本辞典システムでは抽出された文章のオノマトペ, 付加した付属語, オノマトペの係っている語を 1 つの文章として再度検索し用例を集め, その用例の件数でランキングすることを考

オノマトペを選択: ぼかぼか 暖かい 暖かい

「ぼかぼか」の用例(87件)

用例	係る語	出典	編集
ぼかぼかと暖かい日差しが辺りを照らす	暖かい	URL	Edit
ぼかぼかと暖かい日差しが降り注ぐ縁側で, 雅は自分の引っかかっていた桜の木を見上げながらぼんやりとしていた	暖かい	URL	Edit
ぼかぼかと暖かくなり, 風に透明感を感じるようになったら, 桜の花も春の訪れを祝いだす	暖かい	URL	Edit
ぼかぼかと朝から暖かく, 道すがら紅梅の匂いが鼻についてくる	暖かい	URL	Edit
今日はぼかぼかと暖かい日になりました	暖かい	URL	Edit
今日はぼかぼかと暖かい横浜です	暖かい	URL	Edit

図 1: 用例表示画面のイメージ

えている。これにより, 一般的な用例と一般的でない用例との差が大きくなり, 学習者の理解をより深められるのではないかと考えている。

表 3 より, 青空文庫からはかなり高い割合で適正な文章を抽出できていることがわかる。だが, 用例数がとても少ない。それは学習者が理解しやすいよう, 抽出元を現代仮名遣いのページに限定しているために 50 ページ全てを抽出できていないためである。また, 現代仮名遣いのページに限定しても外国人が理解するのに難しい用例も抽出される。また, 青空文庫からは最新の用例が抽出されないこともあり, 一般の Web ページからの抽出も必要となる。

3.3 検索インタフェース

Web ページから抽出した用例の表示画面のイメージを図 1 に示す。ユーザが用例を知りたいオノマトペを選択するとその用例が一括表示される。用例は, 上からオノマトペに「と」「の」「な」「だ」「する」をつけて抽出した用例, 付属語を付けて抽出した用例の順で表示される。また, そのそれぞれについて, オノマトペの係っている語の数でソートされている。1 つの用例でしか係られていない語を含む文章は用例から削除している。用例にはそれぞれ「URL」がついておりクリックするとその用例の抽出元のキャッシュページが参照できるようになっている。また, 用例として不適切だと思われる文章はユーザが編集できるようになっている。

4 まとめと今後の課題

本稿では, オノマトペの用例を自動抽出し, それらを日本語学習者に提示するというオノマトペ用例辞典について記した。

今後は「陽射しがぼかぼかと暖かい」の「陽射し」と「ぼかぼか」のパターンなどオノマトペが係っている語以外の語との用例パターンを抽出することも考えている。また, それを使ってオノマトペ同士を関連付けることも考えている。

また, 利用者参加型の手法をシステムに組み込み, ユーザがコメントを残したりできるようにすることも考えている。

参考文献

- [1] George Chang, Marcus J.Healey, James A.M.McHugh and Jason T.L. Wang.: “Webマイニング,” 共立出版, 197p.
- [2] 香林 隆子.: “オノマトペのオンライン多言語辞書の構築,” DEWS2002 論文集, DEWS2002 A4-4, March 2002.