

# 階層型クラスタリングを利用した文脈によるオノマトペの分類

古宮嘉那子, 小谷善行 東京農工大学 {kkomiya, kotani}@cc.tuat.ac.jp

## Introduction

・オノマトペとは **音を表すことで感覚や感性を表す言葉**

- ・意味的な近さと音韻的な近さ
- ・言葉の選び方によってニュアンスに大きな差

にこにこ、にこりと: 感じのよい笑み( ^\_^ ) ☺  
 にやにや: あまり感じのよくない笑み( -\_- )

語義の観点から 分類してその類似関係を可視化したい

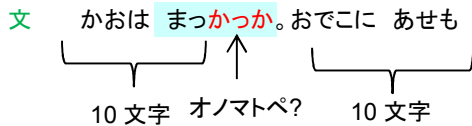
## Method

### オノマトペのクラスタリング手順

1. コーパスからオノマトペの用例を集める
2. それぞれの用例について素性を生成する
3. オノマトペの分布間の距離を測る
4. オノマトペ間の距離をもとにクラスタリングを行う

### Step1 オノマトペの用例収集

オノマトペの用例はコーパスから抽出(「ひまわり」を利用)  
 オノマトペとマッチする文字列とその左右10文字ずつ  
 オノマトペにマッチするがオノマトペそのものではない表現も  
 抽出されている



### Step2 クラスタリングのための素性作成

大抵、語義曖昧性解消に用いられる素性を作成

オノマトペの語義的な分類を行う

- オノマトペの前後二語までの形態素の表記(4素性)
- オノマトペの前後二語までの品詞(4素性)
- オノマトペの前後二語までの品詞の細分類(4素性)
- オノマトペの前後二語までの意味分類コード(4素性)
- 係り受け: オノマトペが修飾する動詞 (1素性)
- ChaSenとCaboCha、分類語彙表を利用

文 「に流れる。しかし外の**しとしと**雨の降る秋の朝は冷た」

…しかし **外** の **しとしと** **雨** の 降る 秋…

- ・システムはChaSen により用例がオノマトペであるかどうかをチェック
- ・文字列としてオノマトペを含んでもオノマトペではない用例を除く

### Step3 オノマトペのクラスタリングのための距離

・Jensen-Shannon divergence (JSD) を利用

$$D_{JSD}(P||Q) = \frac{1}{2} \sum_i P(i) \log \frac{P(i)}{M(i)} + \frac{1}{2} \sum_i Q(i) \log \frac{Q(i)}{M(i)} \quad \text{where } M = \frac{1}{2}(P+Q)$$

$P(i)$  は分布  $P$  の出力  $i$  の確率

二つのオノマトペ(OA とOB)間のJSD を以下のように定義

$$D_{JSD}(O_A || O_B) = \sum_j D_{JSD}(O_{A,j} || O_{B,j})$$

where  $O_A, O_B$  はオノマトペ A, B の分布で  $j$  は素性番号

### Step4 オノマトペのクラスタリング

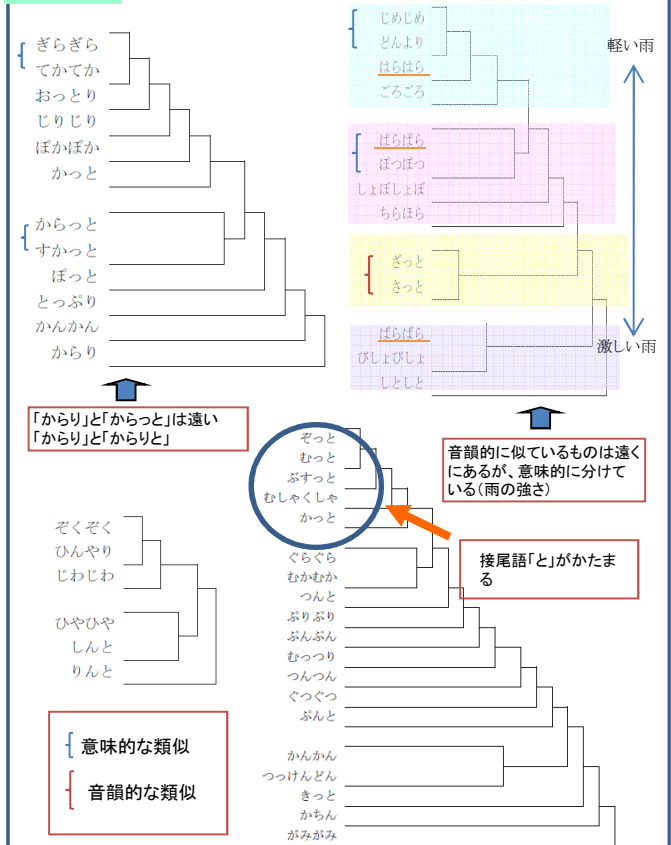
- ・階層型のシングルリンク・クラスタリングを利用
- ・オノマトペ間の関係は階層的と仮定

## Data

- ・「日本語オノマトペ辞典」から4種類を選んで使用
- ・文脈の分布をもとにクラスタリングを行うためには、ある一定以上の用例が必要
- ・10回以上出てくるオノマトペだけを採用し、それ以下のものは最終的な実験の対象外とした
- ・現代日本語書き言葉均衡コーパス(BCCWJ)を利用

オノマトペの種類	辞書	最終実験	Min	Max	Ave
照る・晴れる	17	12	12	228	79
雨・雪・氷	28	13	13	1175	230.46
寒い・冷たい	17	6	24	243	139.83
怒る・不機嫌・無愛想	40	19	15	5894	390.58

## Results



## Conclusion

- ・日本語のオノマトペを、語義の観点から関係を分類、可視化するために、オノマトペの文脈の分布をもとにしたクラスタリングを提案・実行した
- ・素性には、通常、語義曖昧性解消に用いられる、周りの単語や品詞、意味のような文脈の素性を利用
- ・音韻的な素性から抽出できないような意味的な関係が見られた
- ・オノマトペは形態素解析で正しく切るのは難しく、接尾語などの問題が残る。
- ・オノマトペの多義性解消も今後の課題