感性評価のための事象説明文からの オノマトペ想起支援に関する研究

髙橋雅仁[†1]　田辺利文[†2]　首藤公昭[†3]

概要：近年，オノマトペ（擬音語，擬態語）を用いた感性評価に関する研究が多方面で活発に行われている．しかしながら，このような研究で用いるオノマトペ語彙については，必ずしも十分なデータ整備は行われていない．本研究では，我々の自然言語処理分野の研究開発成果である
・エントリー数約 38,800 のオノマトペ共起表現レキシコン
・意味ネットワーク上の単語の活性度の変化を用いたテキストセグメンテーション手法
を組み合わせ，感性評価の対象分野に関する事象説明文を入力すると，パラグラフ毎にその意味内容と関連するオノマトペ群を自動的に提示することにより，当該分野で用いられるオノマトペの収集・選択作業を効率的に行うことができるオノマトペ想起支援システムを開発することを目指す．

キーワード：オノマトペ，感性評価，複単語表現，日本語レキシコン，テキストセグメンテーション

Supporting Japanese Onomatopoeia Recall for Sensitivity Evaluation from Texts Describing an Evaluation Target

MASAHITO TAKAHASHI[†1]　TOSHIFUMI TANABE[†2]
KOSHO SHUDO[†3]

Abstract: In the Japanese language, various onomatopoeias can be used to convey delicate and sometimes sophisticated state of sound or other things intuitively and accurately. In recent years, studies have been conducted to evaluate the sensitivity associated with the use of these onomatopoeias. However, there is insufficient data to prepare an onomatopoeia vocabulary for such studies. To solve this problem, we used our previously developed Japanese multiword expression lexicon of onomatopoeias, which includes 38,800 entries. We additionally developed a method for segmenting text based on changes in the activity levels of words on a semantic network. Here, we aim to develop an onomatopoeia-recall-support system that automatically presents onomatopoeias related to each paragraph of the input text such as text describing a product.

Keywords: onomatopoeia, sensitivity evaluation, multiword expression, Japanese lexicon, text segmentation

1. はじめに

日本語に豊富に含まれるオノマトペ（擬音語，擬態語）は，ものごとを感覚的に的確に伝えることができる特性をもっている．たとえば，雨の降り方を，「雨がぱらぱら降ってきた.」，「雨がざぁざぁ降ってきた.」というように，オノマトペによって的確に表現できる．近年，これらのオノマトペを用いた感性評価等に関する研究が活発に行われている. たとえば, 味覚や食感[1], 物のテクスチャの質感[2], 商品の使用感[3]，都市の質感[4]，介護現場でのコミュニケーション[5]等幅広い分野において研究が行われている．
しかしながら，このような研究で用いるオノマトペ語彙データは，研究者が過去の関連研究の成果やオノマトペ辞典などから採取，あるいは，被験者の発したオノマトペ表現をもとに自ら作成しており，感性評価等を行うための最重要データであるにもかかわらず，必ずしも十分なデータ整備は行われていない．また，オノマトペを用いた研究分野の広がりとともに様々な分野のオノマトペ語彙データの必要性が高まっていると思われるが，現在のところ，広範囲の研究分野に適用可能な感性評価用オノマトペ語彙データベースは構築されていないようである．我々は，対象分野毎のオノマトペ語彙データの整備を効率的に行うためのオノマトペ想起支援システムの開発を通して，上記の課題の解決に取り組みたいと考えている．
本研究では，我々の自然言語処理分野の研究開発成果である
・エントリー数約 38,800 の人手で開発した構文情報を含むオノマトペ共起表現レキシコン（語彙目録）[6],[7],[8],[9]
・意味ネットワーク上の単語の活性度の変化を用いたテキストセグメンテーション手法[10],[11],[12],[13]
を組み合わせ，感性評価等の研究対象分野に関する事象説明文を入力すると，パラグラフ毎にその意味内容の表現に関連するオノマトペ群を対応するパラグラフの話題を示す

†1 久留米工業大学
Kurume Institute of Technology
†2 福岡大学
Fukuoka University
†3 福岡大学名誉教授
Professor Emeritus of Fukuoka University

キーワード群とともに自動的に提示するオノマトペ想起支援システムを開発することを目指す．ここで，テキストセグメンテーションとは，文書（テキスト）を話題の境界でパラグラフに分割することである．このシステムを用いることにより，感性評価等に関する研究を行う研究者が，評価対象についての機能や性質の説明やユーザや被験者によるその評価等を記述した必ずしもオノマトペを含まないテキストデータを用いて，それらのテキストデータのパラグラフ毎に提示される関連するオノマトペ群と対応するパラグラフの話題を示すキーワード群を参照しながら，当該分野の感性評価等に用いるための網羅性の高いオノマトペ語彙データの収集・選択作業を効率的に行うことができる．本稿では，このオノマトペ想起支援システムの構想について報告を行う．以下，2 章では，上記のオノマトペ共起レキシコンの概要，3 章では，上記のテキストセグメンテーション手法を用いたオノマトペ想起支援システムの構想について述べる．

2. オノマトペ共起表現レキシコン

2.1 オノマトペ共起表現レキシコンとは

今世紀に入り，日常の言語にはコロケーション，決まり文句，慣用表現等の特異表現が予想外に多種，多量に使われていることが重視されるようになり，自然言語処理や言語学の分野において，これらの複単語表現（Multiword Expression，MWE）[14]に関する種々の研究が進められている．筆者の一人である首藤は，1960 年代からこの種の日本語複単語表現の総括的なレキシコン Japanese MWE Lexicon（JMWEL）の開発を進めてきた[6], [7]. 本章では，JMWEL の一部をなす日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic[8]（以後，本レキシコンと記す）の概要を紹介する．

本レキシコンの主な特徴は，以下の 3 点である．

ⅰ　オノマトペと他語の共起表現中にギャップ（内部修飾句）が介在する可能性を記載している
ⅱ　オノマトペ（単体），オノマトペ共起表現中の語彙に漢字・片仮名異表記を与えている
ⅲ　オノマトペの連体，連用，動詞化用法を体系化して記載している

2.2 オノマトペ収録表現

本レキシコンの見出しは，

（1）　オノマトペ（単体）3,274 種
　（オノマトペ単体は複単語表現ではないが，便宜上，本レキシコンに含めている．）
（2）オノマトペと他語が共起した日常よく現れる句
　（以後，オノマトペ共起表現と記す）35,517 種

であり，新聞，雑誌等の記事，小説，テレビ，ラジオの放送文から内省によって抽出したものを基本として既存の不特定の辞典類[20]，[21]を使って補強したものである．

2.3 記載情報

本レキシコンは，Microsoft Excel で作成した xlsx ファイルに纏められており，1 行に割り当てた 1 個の見出しに対して，A～I 欄に下記の情報を与えている．たとえば，「クンクンと犬が鳴く」という表現に対して与えた情報を A～I 欄の順に列挙すれば以下のようになる．（空データをφで示す．）

A 欄	クンクン
B 欄	くんくんといぬがなく
C 欄	くんくん-と-いぬ-が-なく
D 欄	クンクン-と-犬-が-鳴く
E 欄	VP
F 欄	[[0to]*[[*Nga]*V30]]
G 欄	φ
H 欄	φ
I 欄	音

（ⅰ）オノマトペ：A 欄

B 欄の見出し表現中に使用されているオノマトペを片仮名表記で与える．オノマトペからその共起表現を検索する際に利用できる．

（ⅱ）見出し：B 欄

オノマトペ（単体），オノマトペ共起表現ともに平仮名べた書きで見出しを与える．同音異義，同音異機能オノマトペは原則として別見出しとする．たとえば，「ぱらぱら」は擬音と擬態で別見出し，「こんこん」では，擬音とは別に擬態の多義でも別見出しとする．

（ⅲ）分かち書き：C 欄

オノマトペ共起表現に対し，その分かち書きを平仮名表記上にハイフン「-」で区切って与える．分かち書き単位は，単語，接頭語，接尾語，接頭造語要素，接尾造語要素とし，活用語尾は形容動詞語尾「な」，「に」，「たる」，「と」以外は切り離していない．造語要素とは造語機能を持つ拘束形態素であり，多くの場合，「緊張-感」の「感」のように音読みの一漢字である．複合語は基本的にアンダースコア「_」で要素語に区切っている．

（ⅳ）異表記：D 欄

オノマトペ共起表現に対して，漢字，カタカナなど，異表記可能な部分には，C 欄の分かち書きの上で，正規表現類似の形式で選択肢を与える．たとえば，「ポッチャリ-と-し-た-(身)体_付き」という D 欄の記載は「ポッチャリ-と-し-た-身体_付き」，「ポッチャリ-と-し-た-体_付き」の可能性を表す．ハイフンやアンダースコアで区切られた C，D 欄の記載から種々の表記形を簡単に生成できる，たとえば，

C 欄の「からだ_つき」と D 欄から得られた「身体_付き」,「体_付き」から「からだつき」,「身体つき」,「身体付き」,「からだ付き」,「体付き」,「体つき」の 6 通りの表記形が得られる.

（ｖ）構文的機能：E 欄

収録したオノマトペ（単体）は，形態・構文上の機能により，

1　単純オノマトペ，　2　連用オノマトペ（副詞的オノマトペ），　3　接頭オノマトペ，　4　接尾オノマトペ，　5　名詞性オノマトペ

に分類される．1 は，格助詞「と」を後接して連用修飾機能を持つものである． 本レキシコンでは，「ころっと」などは，オノマトペ「ころっ」と格助詞「と」の共起表現とみなしている．末尾促音型オノマトペの殆どは 1 に分類される．2 は，そのままでも「と」を後接しても連用修飾機能を持つもの，3, 4 は，他語に連接して造語する機能を持つものである．表 1 に本レキシコンにおける 1〜5 の分布と例を示す．E 欄には表 1 の記号が記載されている． 1, 2 の機能は，後述する H 欄の情報でより詳細化される．

いっぽう，本レキシコンに収録しているオノマトペ共起表現には，

ｉ 名詞句，ii 動詞句，iii 形容詞句，iv 形容動詞（語幹）句，ｖ 連用修飾句，vi 連体修飾句，vii 名詞文形式

がある．表 2 に本レキシコンにおけるｉ〜viiの分布と表現例を示す．オノマトペ共起表現の E 欄には表 2 の記号が記載されている．

（vi）構文構造と内部修飾可能性表示：F 欄

オノマトペ共起表現に対して C 欄のハイフンによる分かち書きに基づき，係り受け構造を修飾子，被修飾子の対をカッコ[ ]で括って記載する．すなわち，句αの主辞が句βの主辞を修飾して出来た句α βの構造記述をα，βの構造記述 a, b を使って[ab]と記載する．

ここで，要素単語の構造記述は，以下の英記号とする．

・単純，連用，名詞性オノマトペ：O,
・接頭オノマトペ：Op,　・接尾オノマトペ：Os,
・接頭語：P,　・接尾語：S,　・接頭造語要素：Q,
・接尾造語要素：R,　・名詞：N,
・動詞：V（未然形 V11,V12，連用形 V22,V23，終止形 V30，連体形 V40，仮定形 V50，命令形 V60），

表 1　採録したオノマトペの分布と例

種別，記号	見出し数	例
単純オノマトペ，O	1,875	ツルリ，ホワッ，ドッカーン，グネッ，ピューン，ゴロリ，ヒヤッ，ドブン
連用オノマトペ，AdvO	1,143	ドッカリ，フラフラ，ミッチリ，フワフワ，チャリチャリ，ゴリゴリ
接頭オノマトペ，Op	150	ドタ，ジリ，グラ，ゴタ，ソヨ，ビリ
接尾オノマトペ，Os	32	タップリ，タラタラ，モリモリ，ピカ
名詞性オノマトペ，NPO	74	ブツブツ，コリコリ，フリフリ，デコボコ
計	3,274	

表 2　収録オノマトペ共起表現の分布と例

種別，記号	見出し数	例
名詞句，NP	4,133	サッパリ-と-し-た-性格，カリッ-と-し-た-口_当り，ホクホク-し-た-食感，サラサラ-し-た-肌_触り，ギリギリ-の-妥協
動詞句，VP	24,153	ドタッ-と-音-が-する，鼻-先-に-人参-を-ブラ-(下/提)げる，肌-が-パサパサ-に-乾く，馬-が-ヒィーン-と-嘶く，ニャーン-と-(ネコ/猫)-が-(鳴/啼)く，フッ-と-胸-に-浮かぶ，カツカツ-と-靴音-が-する，クラクラ-と-(眩暈/目眩)-が-する，キリキリ-痛む，(深(深/々)/シンシン)-と-夜-が-(更/深)ける
形容詞句，AP	610	モチモチ-と-柔らかい，ポンポン-と-威勢-が-良い，ガンガン-と-痛い
形容動詞（語幹）句，AdjVP	245	愛嬌-タップリ，フンワリ-と-柔らか，ツンツン-と-無_愛想
連用修飾句，AdvP	4,019	キョトン-と-し-て，ガラガラ-音-を-立て-て，熟(熟/々)-思う-に，ワイノワイノ-と
連体修飾句，AdnP	2,345	シドロモドロ-の，グチョグチョ-し-た，ガチガチ-な
名詞文形式など，NPS，INC，OP	12	英語-が-ペラペラ，予定-が-ビッシリ，収支-が-トントン
計	35,517	

・形容詞 : A（未然形 A13, 連用形 A22，A23, 終止形 A30, 連体形 A40，仮定形 A50，命令形 A60），
・形容動詞（語幹）：K00，　・副詞：D，　・連体詞：T,
・接続詞：C，　・機能語及び機能性自立語：活用形も含め英小文字ローマ字綴り
文節内の語接続も便宜上，左 2 分岐句構造とみなして上記と同様の記述を行っている．

たとえば, オノマトペ共起表現「クンクン-と-犬-が-鳴く」の構造記述は「クンクン」＝オノマトペ O,「と」＝格助詞 to,「犬」＝名詞 N,「が」＝格助詞 ga,「鳴く」＝動詞終止形 V30 であることから, [[Oto]*[[*Nga]*V30]] と記載する．図 1 にその意味する構文木と係り受け構造を示す.

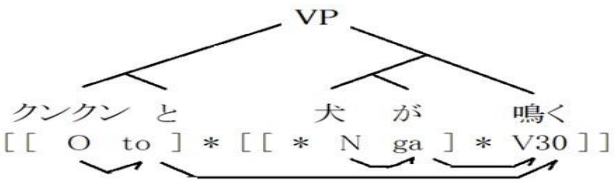

図 1 「クンクン-と-犬-が-鳴く」の構造記述

この例のように F 欄の構造記述内には適所にアスタリスク「*」が含まれており，その位置に，直後の句の主辞に対する修飾句が入り得ることを意味している．したがって，図 1 の構造記述 [[Oto]*[[*Nga]*V30]] は，たとえば,「クンクン-と-朝-から-隣-の-犬-が-寂し-そう-に-鳴く」のような拡張表現の可能性を示している．JMWEL のこのようなギャップ付き構造記述は，定形表現が持ち得る部分的な柔軟性を構文・意味解析機に反映させるための重要な仕組みである．

（vii）後方文脈条件：G 欄
オノマトペ（単体），オノマトペ共起表現に対し，文末側に呼応する語句がある場合にその情報を与える. たとえば,「オチオチと」に対しては，文末側に「休んではいられない」のような否定句が要求されることを<negation>と記す.

（viii）連体化，連用化，動詞化情報：H 欄
オノマトペ（単体）に対し，E 欄の構文機能情報を詳細化して与える．オノマトペを連体修飾，連用修飾に使用する場合と動詞化して使用する場合に通常使われる後接語句を以下のように整理した.
・連体修飾 :「な」,「の」,「たる」
・連用修飾 :「に」,「と」,「ε」
・動詞化 :「する」,「になる」,「とする」
ここで，ε は空列を表し，オノマトペが後接語句なしで連用修飾できる場合を表す．たとえば，オノマトペ「フラフラ」は,「フラフラの（…状態）」で連体修飾,「フラフラと（…歩く）」,「フラフラ（…歩く）」で連用修飾,「フラフラする」,「フラフラになる」,「フラフラとする」と動詞化すること,「フッ」の場合は,「フッと（…気が付く）」で連用修飾する以外には考えにくいことを，それぞれ，後接する語句集合の三つ組によって {no}-{to, ε}-{suru, ninaru, tosuru}, Φ-{to}-Φ と記載する．Φ は空集合を表わす．三つ組のパターンは 100 種程度である.

（ix）擬音，擬態の別：I 欄
オノマトペ（単体）に対し, 擬音, 擬態の別を「音」,「態」と記載する.

3. オノマトペ想起支援システムの構想

3.1 意味ネットワークを用いたテキストセグメンテーション手法

本節では，まず，従来のテキストセグメンテーション手法について概要を記し，続いて，我々が，過去に提案した名詞と動詞の共起情報を基に擬似的に構築される意味ネットワークを用いたテキストセグメンテーション手法について説明を行う.

従来のテキストセグメンテーション手法としては，テキストの結束性に関わる情報のうち,「同一語句の反復」や「類似性に基づく語句の反復」の情報を主として用いる方法 [15], [16], [17]が一般的である．ここで，類似性に基づく関連語句とは，類義関係や上位，下位の関係になる語句を意味する．テキストの結束性に関わる情報としては，この他に「近接性に基づく関連語句の反復」の情報がある．ここで，近接性に基づく関連語句とは，共起関係や因果関係にある語句を意味する．しかしながら，従来の研究では，近接性に基づく関連語句の反復はあまり利用されていない [18], [19]．これは，近接性に基づく関連語句（「バナナ」と「食べる」,「天才」と「発明」など）を網羅的に収集することが非常に困難なためである．Ferret[19]は，新聞記事データから構築した意味ネットワークを用いて 39 の新聞記事を連接させたテキストの境界を求める実験を行い，テキストの境界数と同数の境界候補を出力した場合，50%の正解率を得たが，同一語句の反復を用いる Hearst の方法と比較したところ, Hearst の方法では, 67%の正解率が得られ, Hearst の方法よりも境界認定精度が劣っていたことを報告している．我々が提案した手法[10], [11]も「近接性に基づく関連語句の反復」の情報を用いているが, 新聞記事や Web 上のテキスト等の大量のテキストデータから比較的容易に収集可能な単文内での名詞と動詞の共起関係のみを用いて疑似的に意味ネットワークを構築するため，本手法は，近接性に基づく関連語句の反復の情報を用いた従来の手法と比較して実現性が高いという特徴をもっている.

本手法では，意味的に関連のある単語がネットワーク状に結合された意味ネットワークにおいて，入力テキスト中の単語を順次入力することによって意味ネットワーク上の単語群が活性化された状態，すなわち，入力テキストに対する文脈情報を，単文中の名詞と動詞の共起関係データを

大量に格納した共起辞書を用いて動的に構築する．このようにして疑似的に構築された意味ネットワーク上の文脈情報について，特に話題の変化に応じて活性度が変化しやすい単語群（以後，「キーワード」と呼ぶ）に着目して，それらの活性度の変化を観察し，テキストのパラグラフ間の話題境界候補を出力する．

ここで，上記の共起情報を用いた文脈情報の生成方法の有効性について，意味ネットワークを用いた文脈情報の生成方法との比較を行いながら説明する．以下の２つの文からなる入力文を例にとり説明を行う．

“交響曲・が・演奏された.”

“聴衆・は・拍手した”

まず，意味ネットワークを用いた文脈情報の生成方法について説明する．図 2 は入力文中の第１文に対する文脈情報の生成処理が終了した直後の意味ネットワークの状態を示している．図 2 より，第１文中の単語“交響曲”および“演奏する”に対応する単語と，それらの単語と直接に，あるいは，比較的近い距離で間接的に結合している単語が活性化された状態になっており，特に，第２文中の単語“聴衆”に対する単語が活性化していることがわかる．これは，“交響曲“および，演奏する”という単語と“聴衆”という単語の間の結束性を示していると言える．次に，共起情報を用いた文脈情報の生成方法について説明する．図 3 は，入力文中の第１文における単語“交響曲”に対する文脈情報を表す集合Σの生成過程を示している．このとき，集合Σの中に，第２文中の単語“聴衆”が加えられることがわかる．図 3 における文脈情報を表す集合Σに加えられる単語と，図 2 の意味ネットワークにおいて活性化された状態にある単語とを見比べると，両者がよく似ていることがわかる．したがって，共起情報を用いた文脈情報の生成方法は，意味ネットワークを用いた文脈情報の生成方法における単語の活性化の振る舞いを近似したものと考えることができる．続いて，図 4 を用いて本研究におけるテキストセグメンテーションの原理を説明する．図 4 の上段のグラフは，段落Ａと段落Ｂからなる入力テキスト中の単語位置に対する段落Ａのキーワード Na，段落Ｂのキーワード Nb の累積刺激の変化を示している．ここで，累積刺激とは，疑似的に構築された意味ネットワーク上の名詞に対して入力テキスト中の名詞の入力の度に高められる刺激値が累積したものである．また，中段のグラフは，累積刺激を１階微分したものであり，意味ネットワーク上の名詞の活性度を示している．下段のグラフは，累積刺激を２階微分したものであり，意味ネットワーク上の名詞と関連する話題の開始位置では極大点が生じ，話題の終了位置では極小点が生じる．この性質を利用してテキストセグメンテーションを行う．実際のテキストセグメンテーションでは，図 4 に示すような理想的なキーワードを得ることは難しいと考えられるが，文脈依存性の高い十分な量の単語を予めキーワードとして選定しておき，これらのキーワード群に着目して，累積刺激の２階微分の極値の分布を調べることにより，話題境界の判定を行うようにする．本手法のアルゴリズムは後述するが，２つのパラグラフからなる 600 字程度の新聞の政治面の記事について，本手法を用いてテキストセグメンテーションを行った結果を図 5 に示す．図 5 において，パラグラフの境界位置付近で累積刺激の２階微分の極値が密集しており，パラグラフの境界が捉えられていることがわかる．

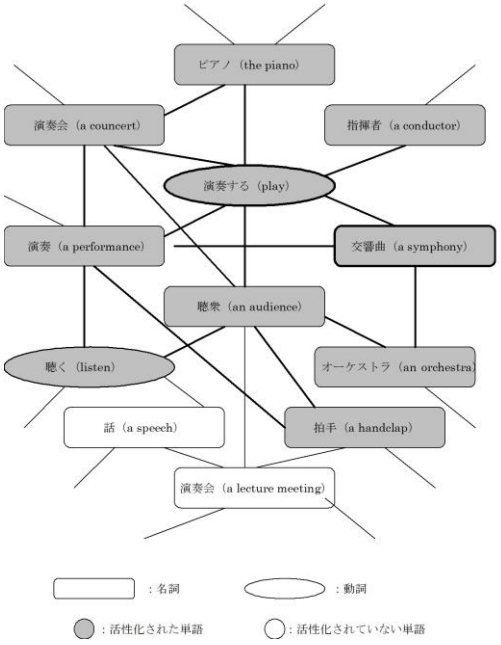

図 2　意味ネットワークによる文脈情報生成

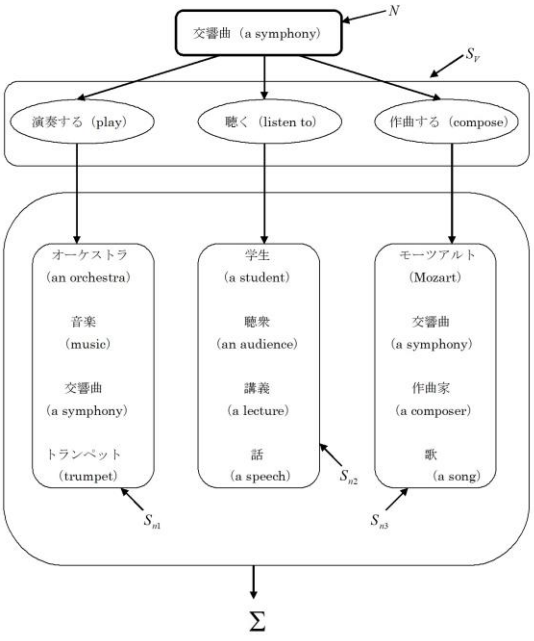

図 3　共起情報による文脈情報生成

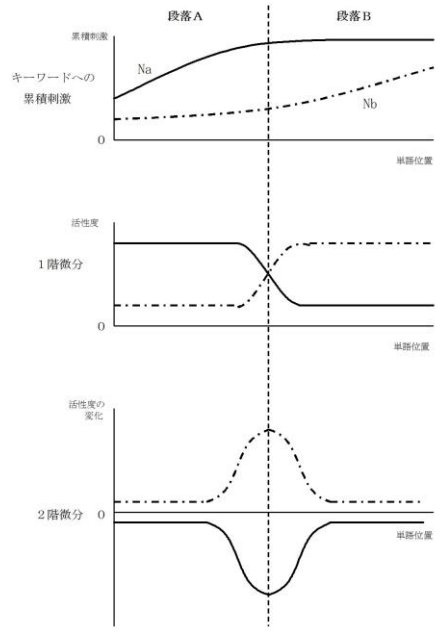

図 4　テキストセグメンテーションの原理

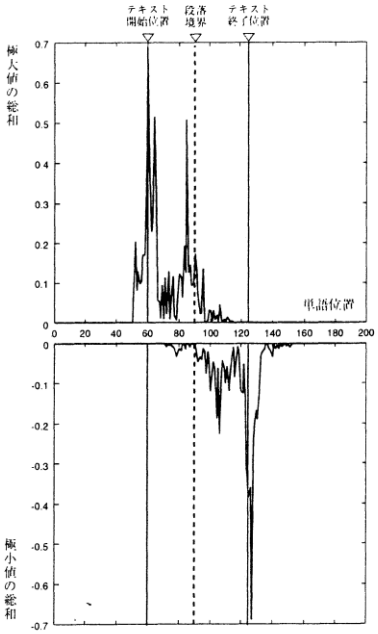

図 5　テキストセグメンテーションの実行結果の例

以下に，共起辞書とキーワード辞書の構築，および，テキストセグメンテーションアルゴリズムの詳細を述べる．

3.1.1 名詞と動詞の共起辞書の準備

文脈情報の生成に用いる言語知識として，一つの名詞と単文内でそれと共起する動詞の集合，および，一つの動詞と単文内でそれと共起する名詞の集合を用い，それぞれ，２項組(N,S_V), (V,S_N)で表す．ここで，N は名詞，V は動詞，S_V は N を格要素としてとる動詞 V_i とその出現頻度 m_i の対の集合，S_N は V の格要素となる名詞 N_i とその出現頻度 m_i の対の集合である．たとえば，名詞「雪」に対する共起情報は，(雪, {(降る,m_1),(積もる,m_2),(警戒する,m_3),…})となる．このような共起情報を大量に格納した共起辞書を事前に準備しておく．

3.1.2 文脈情報の生成アルゴリズム

文脈情報を名詞 N とその累積刺激 k からなる２項組(N,k)の集合 $\Sigma=\{(N_1,k_1),(N_2,k_2),\cdots,(N_i,k_i)\}$ によって表し，以下の手順で求める．集合Σは，入力テキスト中の名詞を順次入力する度に変化する．

ステップ１　$\Sigma=\phi$ とする．

ステップ２　入力テキストの先頭から単語を順次読み込み，品詞が名詞である単語を読み込む度に，以下の文脈情報の更新処理を行う．

１．読み込んだ名詞 N について，共起辞書より共起情報(N,S_V)を取り出す．

２．S_V 中のすべての動詞 $V_i(i=1,2,\cdots,m)$について，共起辞書より共起情報$(V_i,S_{Ni})(i=1,2,\cdots,m)$を取り出す．

３．S_{Ni} 中のすべての名詞 $N_j(j=1,2,\cdots,n_i)$を集合Σに加える．ただし，名詞 N_j の累積刺激 k_j は，N・V_i 間の共起の出現頻度 $a_i(i=1,2,\cdots,m)$と V_i・N_j 間の共起の出現頻度 $b_j(j=1,2,\cdots,n_i)$の関数 $f(a_i,b_j)$（たとえば，$k_j=a_i\cdot b_j$）で与える．なお，同じ名詞がすでに集合Σに存在する場合は，集合Σの要素は増さずにその単語の累積刺激の加算のみ行う．

3.1.3 話題境界判定用のキーワード辞書の準備

3.1.2 で示した文脈情報を表す集合Σに含まれるある名詞 N_i について，入力単語毎の累積刺激の変化を示すグラフを作成しそれを微分すると，意味ネットワーク上の名詞 N_i の活性度を近似できる．そこで，話題境界の判別に有効な名詞（キーワード）の集合を予め学習し，話題階層の粒度に応じた複数のキーワード辞書を作成する．キーワード学習は，**3.1.1** で定義した共起辞書中のすべての名詞について，キーワード学習用テキストに対する意味ネットワーク上の活性度の変化を記録し，話題境界でよく反応する名詞を集めることによって行う．このようにして，キーワード辞書を事前に準備しておく．なお，テキスト中のパラグラフは階層性を持つため，話題階層の粒度に応じて複数のキーワード辞書を用いてもよい．

3.1.4 テキストセグメンテーションのアルゴリズム

以下の手順で入力テキストに対するテキストセグメンテーションを行う．なお，ここでは，話題階層の粒度に応じて大分類用 D1，中分類用 D2，小分類用 D3 の 3 種類の

キーワード辞書を用いるものとする．

ステップ１　**3.1.2** で示したアルゴリズムに従って，入力テキストに対して集合Σの生成を行う．このとき，入力テキストの先頭から見て t 番目の名詞 $N_t(t=1,2,\cdots,u)$に対する文脈情報の更新処理が終了する毎に，キーワード辞書 D1, D2, D3 に含まれるすべてのキーワード $X_i(i=1,2,\cdots,v)$ に対する文脈情報の累積刺激 $k_{it}(i=1,2,\cdots,v, t=1,2,\cdots,u)$の値を集合Σから読み出し，記録していく．

ステップ２　各キーワード X_i に対して入力テキスト中の各名詞 N_t の位置における文脈情報の累積刺激 k_{it} の２階微分値を計算する．なお，２階微分を行う際は，事前に累積刺激 k_{it} の移動平均をとり平滑化を行う．

ステップ３　入力テキスト中のすべての文の境界位置 $b_j(j=1,2,\cdots,p)$と対応する入力単語列中の境界位置について，それらの境界位置の前後 q 単語の範囲に出現した各キーワード X_i に対する累積刺激 k_{it} の２階微分の極大値（キーワード X_i に関する話題の開始を示す）および極小値（キーワード X_i に関する話題の終了を示す）の絶対値の総和をキーワード辞書 D1，D2，D3 毎に求め $S1_j$, $S2_j$, $S3_j(j=1,2,\cdots,p)$とする．すなわち，X_i がキーワード辞書 D1 に含まれる場合に X_i に対する２階微分の極値を S1 に加算する．S2，S3 についても同様である．なお，累積刺激の２階微分の極値は，設定された閾値を越えない場合はその値を 0 とする．

ステップ４　ステップ３で求めた極値の絶対値の総和 $S1_j$, $S2_j$, $S3_j(j=1,2,\cdots,p)$をそれぞれ値の大きなものから順に並べ替え，設定された閾値を超えた値について対応する文の境界位置 b_j を話題の粒度のラベルを付与して話題境界候補として出力する．

3.2 オノマトペ想起支援システムにおけるオノマトペ提示アルゴリズム

本研究では，**2** で紹介したオノマトペ共起表現レキシコンと **3.1** で説明したテキストセグメンテーション手法を用いて，感性評価等の研究対象分野に関する事象説明文を入力すると，パラグラフ毎にその意味内容の表現に関連するオノマトペ群を対応するパラグラフの話題を示すキーワード群とともに自動的に提示するオノマトペ想起支援システムの開発を目指している．本システムにおけるオノマトペ提示の基本的なアルゴリズムを以下に示す．なお，必要に応じて，Web 上のテキストデータや後述の大規模日本語 n-gram データ等を用いてオノマトペ共起表現レキシコンのエントリーを増強することも考えられる．

ステップ１　**3.1** のテキストセグメンテーション手法で得られた話題境界候補データを用いて話題の階層構造を求め，得られた話題区間 i 毎に活性度が高かった名詞の集合$\Sigma N_i=\{N_1,N_2,\cdots,N_{mi}\}$を抽出する[12], [13].

ステップ２　ステップ１で求めた名詞の集合ΣN_i の要素となるすべての名詞 N_i に対して，入力テキストの該当区間 i で名詞 N_i を含む文 S_{ji} を抽出し，その格構造パターン（動詞およびその動詞と格関係をもつ名詞と格助詞の組）C_{ji} を取り出す．

ステップ３　ステップ２で求めた格構造パターン C_{ji} と 2 のオノマトペ共起表現レキシコンの各エントリーO_k との構造的なマッチングを行い，類似度の高いオノマトペ共起表現エントリーO_x 中のオノマトペ o_x を抽出し，オノマトペ o_x を入力テキストの該当区間 i と意味的に関係のあるオノマトペ候補の集合$\Sigma O_i=\{o_1,o_2,\cdots,o_{ki}\}$の要素に加える．

ステップ４　ステップ２において，入力テキストの話題区間 i に出現しなかった名詞$N_i \in \Sigma N_i$については，オノマトペ共起表現エントリーO_x のうち，名詞 N_i と共起するものから O_x 中のオノマトペ o_x を抽出し，オノマトペ o_x を入力テキストの該当区間 i と意味的に関係のあるオノマトペ候補の集合$\Sigma O_i=\{o_1,o_2,\cdots,o_{ki}\}$の要素に加える．

ステップ５　入力テキストのパラグラフ毎に，ステップ１で得たパラグラフの話題を示す名詞の集合$\Sigma N_i=\{N_1,N_2,\cdots,N_{mi}\}$とステップ３，ステップ４で得たオノマトペ候補の集合$\Sigma O_i=\{o_1,o_2,\cdots,o_{ki}\}$を出力する．

4. おわりに

本稿では，我々の自然言語処理分野の研究開発成果である

・エントリー数約 38,800 の人手で開発した構文情報を含むオノマトペ共起表現レキシコン
・意味ネットワーク上の単語の活性度の変化を用いたテキストセグメンテーション手法

を組み合わせ，感性評価等の研究対象分野に関する事象説明文を入力すると，パラグラフ毎にその意味内容の表現に関連するオノマトペ群を対応するパラグラフの話題を示すキーワード群とともに自動的に提示するオノマトペ想起支援システムの開発構想について概要を記した．本システムの開発に用いるオノマトペ共起表現レキシコンは，人手によりオノマトペ共起表現を採取しており，品質が高く，また，網羅性も高いと思われるが，既存のエントリーに加え，Web 上のテキストデータや Google 社の大規模日本語 n-gram データ等を用いてオノマトペ共起表現レキシコンのエントリーをさらに増強することが可能と思われる．また，近年，自然言語処理の研究において word2vec[24]等の機械学習を用いたアプローチが盛んに行われている．オノマトペを含む大量の文や n-gram データを学習データとし，機械学習により，任意の文から関連するオノマトペを出力したり，あるいは，文中に適切なオノマトペを挿入したりする機能を実現できる可能性がある．また，これらの手法を本稿で述べたオノマトペ想起支援システムに導入し，入力テ

キストに対するオノマトペの提示の精度を高めることもできるのではないかと思われる．

謝辞 本研究は平成 30 年度久留米工業大学学長裁量経費の助成を受けたものです．

参考文献

[1] 渡辺知恵美，中村聡史．オノマトペロリ：味覚や食感を表すオノマトペによる料理レシピのランキング．人工知能学会論文誌,2015,vol.30,no.130,p.340-352.
[2] 權 眞煥，吉野 淳也，高佐原 舞，中内 茂樹，坂本 真樹．質感を表現するオノマトペからみた自然感と高級感の関係．基礎心理学研究,2017,vol.36,no.1,p.40-49.
[3] 新里圭司，益子宗，関根聡．オノマトペを利用した商品の使用感の自動抽出．情報処理学会論文誌,2015,vol.56,no.4,p.1305-1316.
[4] 北雄介．オノマトペを用いた街歩きによる都市の様相の記述と分析．日本建築学会計画系論文集,2018,vol.83,no.749,p.1285-1295.
[5] 上村初美．介護のオノマトペは自然習得が可能なのかーEPA 候補者へのヒアリングから探るー．日本語教育方法研究会誌,2017,vol.23,no.2,p.46-47.
[6] Toshifumi Tanabe, Masahito Takahashi and Kosho Shudo. A lexicon of multiword expressions for linguistically precise, wide-coverage natural language processing. Computer Speech and Language(Elsevier),2014,vol.28,no.6,p.1317-1339.
[7] 高橋雅仁，田辺利文，首藤公昭．日本語複単語表現レキシコン（JMWEL）の概要と現状ー動詞性複単語表現を中心としてー．言語処理学会第 24 回年次大会発表論文集,2018,p.428-431.
[8] 首藤公昭，田辺利文，高橋雅仁．日本語オノマトペ共起表現レキシコン JMWEL_onomatopoeic．国立国語研究所言語資源活用ワークショップ 2018 発表論文集,2018,p.510-517.
[9] 首藤公昭．日本語処理研究工房ことばの森. http://jefi.info/
[10] Masahito Takahashi, Shin'ichiro Morisawa, Kenji Yoshimura, Kosho Shudo. Text Segmentation Using a Change of Keywords' Activation Levels. Proceedings of 5th Natural Language Processing Pacific Rim Symposium,1999,p.519-522.
[11] 高橋雅仁，森澤慎一郎，吉村賢治，首藤公昭．キーワードの活性度の変化を用いたテキストセグメンテーション. 2000 年情報学シンポジウム論文集,2000,p.145-152.
[12] 高橋雅仁，吉村賢治，首藤公昭．キーワードの活性度の変化を用いたテキストからの話題構造抽出．第 52 回電気関係学会九州支部連合大会講演論文集,1999,p.674.
[13] 高橋雅仁，吉村賢治，首藤公昭．キーワードの活性度の変化を用いたテキスト中の単語と話題の対応付け．言語処理学会第 6 回年次大会発表論文集,2000,p.324-327.
[14] I. A. Sag, T. Baldwin, F. Bond, A. Copestake and D. Flickinger. A Pain in the Neck for NLP. Proc. of the 3rd CICLING,2002.
[15] M. A. Hearst. TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. Computational Linguistics,1997,vol.23,no.1,p.33-64.
[16] 望月源，本田岳夫，奥村学．複数の知識の組合せを用いたテキストセグメンテーション．情報処理学会研究報告,1995,95-NL-109,p.47-54.
[17] J. Morris, G. Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics,1991,vol.17,no.1,p.21-48.
[18] 山本和英，増山繁，内藤昭三．手がかり語および語の類縁性を併用した段落分け．情報処理学会研究報告,1992,93-NL-92,p.41-48.
[19] O. Ferret. How to thematically segment texts by using lexical cohesion?. Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics,1998,vol.2,p.1481-1483.
[20] 阿刀田稔子，星野和子．擬音語擬態語使い方辞典第 2 版．創拓社出版,2004.
[21] 小野正弘．日本語オノマトペ辞典．小学館,2007.
[22] グーグル株式会社. GSK2007-C Web 日本語 N グラム第 1 版. 言語資源協会,2007, https://www.gsk.or.jp/catalog/gsk2007-c/
[23] 京都大学 黒橋・河原研究室．日本語形態素解析システム JUMAN. http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN
[24] 齋藤康毅．ゼロから作る Deep Learning ② - 自然言語処理編．オライリー・ジャパン,2018．“Word のスタイルの基礎”. https://support.office.com/ja-JP/article/d38d6e47-f6fc-48eb-a607-1eb120dec563, (参照 2016-02-20).