

読み句の構造に基づく郷土かるた読み句生成手法の提案

神澤 義人[†] 佐藤 哲司^{††}

[†] 筑波大学 情報学群 知識情報・図書館学類 〒305-8550 茨城県つくば市春日 1-2

^{††} 筑波大学 図書館情報メディア系 〒305-8550 茨城県つくば市春日 1-2

E-mail: [†]{kamisawa,satoh}@ce.slis.tsukuba.ac.jp

あらまし 対象地域の事象を端的に表現する郷土かるた読み句の生成手法を提案する。提案法では、既存の郷土かるたである上毛かるたから、読み句が持つ音数と品詞構造についてのテンプレートを作成する。読み句生成に使用する語句は、入力として与える名所や特産品などの特徴語と、Web 文書群から取得した関連語である。関連語は特徴語との共起度で絞り込みを行う。読み句テンプレートの構造情報に基づいて特徴語と関連語を配置することで、特徴語について表現する読み句候補を生成する。生成の結果、語句の採用基準に対してテンプレートを基にした音数の制約を与えた場合に、出力読み句候補数を著しく絞り込めることが確認できた。

キーワード 文書生成, 郷土かるた, 語句選出

1. はじめに

かるたとは、平安時代の貝合せに端を発する、ある事柄について記した読み札と、それに対応する絵札からなる玩具である。かるたを使用する遊びは幾つか存在する。一般的には取り札と読み札を1つの組とし、読み札に記された読み句の読み上げ内容から、場に並んだ取り札の中から対応する1枚を取得することで、最終的な取得枚数を競うものが、全国的によく知られている。このように、かるた遊びのルールは明快かつ単純である。かるた遊びとは、言葉の理解がある程度進んでいる子供から大人まで、幅広い年代が親しむことができる遊戯である。

かるたに採用される読み句の特徴としては、音数制約がないことが挙げられる。例外として題材に万葉集のような詩歌を採用する場合には特定の音数で揃うこともあるが、短歌や俳句を詠む時のような明確な音数制約は存在していない。そのため、かるたは、読み句ごとに異なる音数で表現されることが数多くある。ただ、あくまでも厳格な規則が存在していないだけであり、日本語として発話する以上、全体の長さや音が切れる位置には、緩やかであっても一定の傾向は存在する。この傾向は、かるたの種類によって異なる。

かるたの種類は、短歌の上の句と下の句を組とした百人一首や、ことわざを読み句としそれを表現する絵札と合わせて組としたいろはかるたなど多岐にわたる。郷土かるたは、主に地方自治体、あるいは地域活性化を活動目的とする人物や団体が主導して作られた、特定地域を題材としたかるたである。日本全国各地で地域活性化を目的とした活動が盛んに行われているが、中でも郷土かるたは、次世代を担う子供達が郷土のことを学ぶことができる教材として注目されている[2]。郷土かるたで採用される題材には、地域の特色や地名、縁のある著名人などがある。遊戯を通じ、若い世代に対する郷土についての知識の伝達、郷土愛の涵養、世代を越えた交流を行い地域活性を促進することが、郷土かるたが作られる主な狙いとなっている。

郷土かるたを新規に作るようになった場合、対象とする郷土

の規模にもよるものの、年単位で取り組まれることが多い。多くの郷土かるたの製作フロー中に、題材や採用する読み句を一般公募して、選出する行程があるためである。そのため、郷土かるたの完成には多くの人材と時間が必要となり、発案から即座に教材として使用することは難しい。

一般的に郷土かるたは他のかるたと同様に、句頭が50音順に「あ」から「わ」の音で始まる読み句を1句ずつ採用しており、1つの郷土かるたは44句の読み句^(注1)を持つことが多い。

1つの郷土かるた読み句群とは、特定地域について多角的に述べた短文群であり、地域が持つ複数の要素を一括りのコンテンツとしてまとめることが可能な媒体である。

本研究では、既存の郷土かるたの音数や品詞の配置構造について解析し、これを基にして新規地域を題材とした郷土かるたの読み句を生成する手法の提案を行う。郷土かるたの読み句の生成手法を確立することにより、簡潔な日本語表現で地域の特徴を表現することができるようになる。これにより、外国人観光客へのPR素材としての活用や、地域の埋もれた魅力を再発見することに繋がると期待している。

本論文の構成は次の通りである。第2章では、関連研究について概観し、本研究の立ち位置を示す。第3章では、郷土かるたとして広く知られている群馬県の上毛かるたについて解析した結果について記す。第4章で、調査結果で得られたかるたの特徴を再現する郷土かるた生成手法を提案し、5章では実際の生成を行った結果とその考察について述べる。第6章で本論文のまとめ、今後の課題について述べる。

2. 関連研究

山口、原口は郷土かるたについて、『郷土を代表するような様々な事象を詠んだいろはかるたの一種』[1]と定義している。ここでの郷土とは、学区や市町村、都道府県などの様々なレ

(注1): 「あ」「ゑ」などを除いた44音分。かるたの種類によっては濁音の句頭を採用するなどの理由で読み句の数が増える。

ベルの範囲を指し、内容については自然や歴史、産業、文化などを扱うとしており、多岐にわたる。同文献内において、原口は、2008年4月30日現在までに製作された郷土かるたは1,450種を超えるとしている。

教育的な側面にも着目しており、原口らは郷土かるたの持つ社会的有用性について『「郷土かるた」は単なる遊びではなく、郷土認識の育成という観点から見て社会科教育にとって注目すべき存在である』と指摘[2]している。

郷土かるた研究では、群馬県の上毛かるたを用いることが多い。上毛かるたは、戦後の群馬県の復興を願い、昭和22年12月に誕生した戦後日本初の郷土かるたである。かるたに用いる題材を上毛新聞にて一般公募し、18人の編纂委員がまとめた。2010年現在までに130万部以上の発行部数を記録[1]しており、最も発行部数が多い。上毛かるたが発行された翌年の昭和23年から毎年上毛かるたの群馬県大会が開かれ、近年は全国大会が開催されるまでに活発な活動が行われている[3]。また、1998年には英語版の上毛かるたが発行されている。

内藤らの研究[4]では、郷土かるたの1つである長者町かるたを用いたかるた大会で活用する目的に端を発して、カルタ遊び支援システムを提案している。これは、音声認識によって読み上げられた読み手の音声から読み札を判断し、画像処理を通して対応する場の絵札を発見し、プロジェクタによって場の絵札の位置を提示するシステムで、48枚の絵札の内41枚を正確に認識できたとしている。また、このシステムを別の郷土かるたである小牧かるたへと適用し、小学生を対戦相手としてかるた遊び実験を行ったところ、7枚の読み上げ中1枚をシステムが先に認識できたと報告している。

音響音声学的な側面においては、大澤らの研究[5]によって、上毛かるたが持つ独特な韻律特徴について、かるた経験者と非経験者間における違いが公表されている。かるた経験者は一定の時間枠内に収める割合が大きく、前半と中盤の読み上げ時間の合計と後半の読み上げ時間の割合がほぼ同じになるなど、読み方を調整する傾向が見られた一方で、かるた非経験者にはそのような傾向が現れなかったことを示した。

壇らの研究[6]では、コミュニティに投稿された、英語俳句が含まれるメールから、単語成分と品詞成分の二種類の文書ベクトルを構成し、サポートベクターマシンを用いた学習と分類によって特徴抽出を試みている。その結果、英語俳句を含む正解文書と英語俳句を含まないノイズ文書との分類において、品詞成分による文書ベクトルを用いた場合に適合率と再現率が改善されたが、文書中の作者名という固有名詞抽出の効果は期待できないように思えると述べている。

Ajiによる研究[7]では、英語俳句コーパスやWord Vector Modelなどを用いた英語俳句の生成に取り組まれている。1つの題材について、人が製作した読み句とシステムが生成した読み句の中から、それぞれ良いとされるものとランダム抽出したものを合わせた計4句を実験協力者に提示し、チューリングテストとアンケート調査を実施した。その結果、システムによる生成句の一部が人間によるものと区別できない結果が得られた。

かるたという題材に対して、機械によるかるた読み句の生成

という視点でアプローチした研究はこれまで知られていない。郷土かるたの読み句は、人が持つ伝えたい地域の事柄についての知識と、それを反映させて読み句を構成する技術によって制作されている。本研究では、郷土かるたの読み句が持つ端的に地域の事柄を表現しているという特徴に注目し、それを機械によって再現することを目的として設定する。郷土かるたの読み句は俳句とは異なる音数構造を持つことが特徴の1つであるが、本研究ではそれを考慮することにより、俳句とは異なる形式を持ったかるた読み句の生成手法を提案する。

3. 郷土かるたの解析

読み句生成のために、上毛かるたを例に挙げて特徴を説明した後、解析した結果について述べる。

3.1 かるたの読み句について

郷土かるたの読み句構造を解析するにあたり、かるたの読み句について各名称を定義し、基礎的事項を整理する。

かるたの読み句は、単語や文節が持つ音数により、音の切れ目が暗黙的に示された日本語の短文である。本研究では、読み句が持つ音の切れ目のことを区切れと呼び、区切れによって区切られた句の各部を、初句、二句、…結句と、句頭から順に呼称する。

かるたの読み句は、日本古来の詩歌である俳句や短歌とは似て非なるものである。俳句や短歌には『五、七、五』や『五、七、五、七、七』のように、詩歌の制作に際して基準となる音数の規則が存在する。このように、5音と7音を基調として文章を構成することを、七五調と呼ぶ。この一方で、かるたの読み句には、読み句全体の音数を定めるような規則は存在していない。例としては、「犬も歩けば棒に当たる」「論より証拠」のように、ことわざを読み句として採用したいろはかるたを挙げることができる。

3.2 上毛かるたの読み句について

2章で述べたように、上毛かるたは郷土かるたの中でも特によく知られた存在である。上毛かるたの読み句は「あ」から「わ」まで、外箱が未開封の段階ではいろは順に並べられた44句が収録されており、群馬県の名所、歴史上の人物、特産品、群馬県が持つ特徴などを題材としている。上毛かるたの読み句については、短歌や俳句のように季語を必要としないこと、結句の最後には名詞が現れる、という構造的な特徴がある。

3.1節で述べたように、上毛かるたの読み句にも、一意に定まった音数構造は存在していない。しかし、山口らの研究[1]によれば、上毛かるたは、七五調を採用していたいろはかるたの伝統を踏まえているため、7+5の12文字を基本としているという。山口らは七五調について、12文字の制約の中で物事の本質を的確に表すことは非常に難しいとする一方で、本質を短い中に盛り込んで作られた読み句を、五七五の俳句調と比較して無駄がなく冗長でないと評価している。また、7音と5音を合わせた12音が、日本人の一呼吸間に平均して読み上げる音数であると論文で紹介[2]し、呼吸に馴染む七五調の文句に繰り返し触れることにより、群馬県民は知らず知らずのうちに読み札を記憶するのだと述べている。

表 1 上毛かるた読み句の音数

総音数	音数配置	件数
12	4,8	1
	7,5	13
	8,4	1
13	7,6	5
	8,5	4
14	7,7	2
	8,6	1
15	7,8	5
	8,7	7
16	8,8	2
17	5,7,5	1
18	5,8,5	1
	6,7,5	1

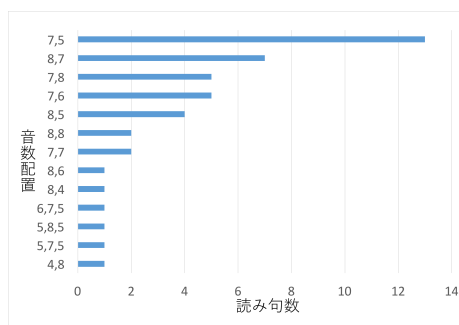


図 1 上毛かるた読み句の音数配置の傾向

上毛かるたの読み句の音数について、構造の内訳を示したのが表 1 である。音数については実際の読み札画像を参考に七五調を意識しながら、筆者の手によって区切りをいれた結果によるものである。結果、上毛かるたの読み句は 4 音から 8 音ごとに区切れが入る。読み句は「初句、結句」あるいは「初句、二句、結句」のいずれかの構造を取る。つまり、上毛かるたにおいては区切れが 3 つ以上出現することはない。

そして先述の通り、上毛かるたは七五調の概念を下敷きにして制作されている。そのためか、最も多い音数配置は計 12 音で構成されたもので、44 句中 13 句がこれにあたる。

更に、図 1 に読み句の音数配置の傾向を示す。最も多い七五調は 13 句あるが、13 種類の音数配置があり、かるたの読み句が短歌や俳句のように、一意に定まった音数の枠組みがないことと、採用される音数構造にはある程度の傾向があることがわかる。

4. かるた読み句生成手法の提案

本研究における郷土かるた読み句生成の流れを図 2 に示す。3 章で解析したかるたの特徴からテンプレートファイルを作成する。テンプレートファイルから取得した構造情報に対して、Web から収集した語句を適用し、読み句を生成する。

4.1 読み句テンプレート

読み句生成システムは、xml 形式のテンプレートファイルと新規に対象とする地域と特徴語を入力とする。テンプレートファイルには、上毛かるたの読み句 44 句の構造が階層的に表

記されている。図 3 にテンプレートファイルの一部を抜粋して示す。

図 3 にある xml 構造について、各要素の意味は以下の通りである。

poem 読み句。id 属性によって 1 から 44 の番号が振られている。

text 読み句本体

rubi よみがな

syllabic 音数構造

morpheme 形態素。形態素の種類を示す type 属性を持つ。

sub 形態素の属性。要素を持たない

word 形態素の語句

sounds 形態素の持つ音数

この内、morpheme 要素は 1 つの poem 配下に複数存在し、また、morpheme 要素配下には 1 つ以上の sub 要素、word 要素、sounds 要素がある。

44 句ある上毛かるたの読み句を、それぞれ形態素解析し、各形態素の持つ属性を取得する。それぞれの形態素は別々の morpheme 要素とし、形態素の属性については morpheme 配下に sub 要素として与える。これらに加え、読み句の中心となる固有名詞を選別し、characteristic 属性を持つ sub 要素を与えた。characteristic 属性を与えられた属性を、本研究では『特徴語』と呼称することにする。システムはこのテンプレートから必要な要素を選択して、読み句生成時の枠組みを構成する。

4.2 関連語の収集

新規に読み句を生成するには、読み句の中心にするための『特徴語』と、特徴語について表現・説明する『関連語』を用意する必要がある。本節では、Web 検索エンジンを用いて関連語を収集する過程について述べる。新規ご当地かるたの題材とな

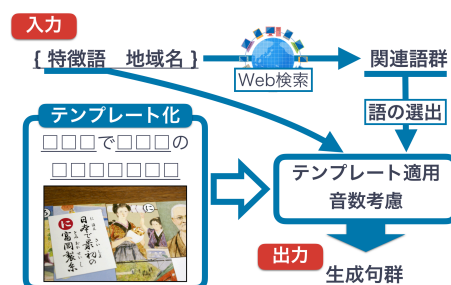


図 2 読み句生成の流れ

```
<poem id = 1 ~ 44 >
<text> 読み句本文 </text>
<rubi> 読み句のルビ </rubi>
<syllabic>音数構造 </syllabic>
<morpheme type=形態素の種類 >
  <sub type=語句の持つ属性 />
    <word> 単語 </word>
    <sounds> 単語の音数 </sounds>
  </morpheme>
  :
</poem>
```

図 3 読み句テンプレートの xml 構造

る地域にある名所や観光地、名産品、人物名のような名詞を特徴語とする。地域名と特徴語 1 つとの語句間に空白を挟んだ文字列をクエリとして Web 検索エンジンに与え、AND 検索を行う。得られた検索結果の内、上位 100 件を該当地域の特徴語について特に記述している Web ページとみなす。この Web ページについての説明や該当箇所の抜粋が検索結果画面に表示される。この文書はスニペットと呼ばれる。検索結果上位 100 件のスニペットを収集する。各スニペットに対して、形態素解析を行い、出現語句の一覧と、スニペット 100 件中の出現頻度を取得する。ここで得られた語句の内、複数の Web ページにおいて入力した地域・特徴語とともに出現する語句は、特徴語と関連可能性のある語句である。これが特徴語の関連語であり、生成句へ適用する候補群である。

4.2.1 特徴語と関連語の共起度

特徴語を使った Web 検索により収集した関連語は、全て特徴語に紐付いた語句であるとは限らない。そのため、特徴語に対して真に関連する語句を選出する必要がある。そこで本研究では、特徴語と各関連語間で共起度を算出し、関連度の選出を行うこととする。共起度の算出に用いた式を以下に示す。

$$\text{共起度} = \frac{|x_r + y_c|}{\min((|x| + y_c), (|y| + x_r))} \quad (0 \leq \text{共起度} \leq 1)$$

ここで、 x は特徴語 A と地域名で AND 検索を掛けて出現した上位 100 件のスニペット群、 y は関連語と地域名で AND 検索を掛けて出現した上位 100 件のスニペット群である。 x_r は文書集合 x における関連語が出現する文書数、 y_c は文書集合 y における特徴語が出現する文書数である。これら出現回数は形態素解析結果を基にしている。そのため、入力した特徴語が別の固有名詞の一部となっているなどしている場合があるが、本研究ではそれぞれ別の語句としてカウントしている。以上の計算を特徴語と収集した各関連語について行い、共起度に基づいた共起度ランキングを作成する。

4.3 テンプレートに基づく語句配置

ここまでの説明で、既存かた読み句の枠組み情報を持つテンプレートファイルとそれに当てはめる語句群が用意された。xml 形式で作成したテンプレートファイルから、品詞配列や音数配列のような情報を生成各要素ごとに取得して、csv 形式で扱えるようにする。語句の適用時、第一に参照するのは品詞配列の情報である。既存かた読み句の構造を引用することで、極端な品詞の誤用を防止することを期待するためである。基となった読み句が、句頭から名詞、動詞、名詞と並んでいるのなら、そのテンプレートを用いて作られる新規読み句の品詞配列も先頭から名詞、動詞、名詞の順となる。まず、品詞配置テンプレートの名詞部分の一箇所に対し、特徴語を適用する。特徴語の適用は総当たりとし、適用可能な限りのパターンを生成する。例として、名詞が a, b, c の三箇所に配置されている構造がテンプレートに記述されている場合、a 部に特徴語があるもの、b 部に特徴語があるもの、c 部に特徴語があるものと、三種類のパターンが用意されることとなる。特徴語を適用した後、残りの枠に関連語を適用する。この時、特徴語と関連語の共起度に基づいて基準値を設けることにより、ランキング下位、つまり共起度が

低い語句が適用されないようにフィルタリングすることが可能である。基準値の設定は、得られた共起度の値によって適宜設定を変更する必要がある。そして、関連語は各品詞ごとに総当たりでテンプレートへの適用をする。適用の際に、morpheme 要素が持つ単語ごとの音数や sub 要素を考慮に入れることにより、読み句として違和感のない音数の読み句、語関係の維持を期待できる読み句が生成可能となる。

5. 読み句生成実験

前章で述べた手法に基づき、実際に読み句を生成した。実験概要を説明したのち、各条件のもと生成されたそれぞれの句を比較する。最後に生成句に対する考察をまとめる。

5.1 実験構成

5.1.1 テンプレート作成

本手法の基盤となる、かるた読み句のテンプレートファイルを作成する。テンプレートを作成するにあたり、既存のかるた読み句として群馬県の上毛かるたを採用した。今回の読み句の解析には、形態素解析ソフト MeCab [8] を使用した。句頭が「あ」から「わ」で始まる上毛かるたの読み句 44 句をそれぞれ形態素解析し、その結果に基づき xml 形式でテンプレートファイルに記述した。この時、形態素解析により上手く取得できなかった固有名詞の修正、特徴語の選定が筆者により行われた。また、各 morpheme 配下の sub 要素として、都市名、人名などの属性を手動で付与した。

5.1.2 特徴語と関連語の収集

実際に地域と特徴語を入力として Web から関連文章と関連語を収集するため、今回は Web ブラウザ自動操作ソフトウェア Selenium [9] を用いた。Selenium によって Web ブラウザアプリケーション Firefox を操作し、Web 検索エンジン Google 検索に対し特徴語と地域名を空白で繋いだ文字列をクエリとして与え、AND 検索を行う。この行程の事前に Firefox のユーザ設定を行い、一画面上に表示される検索結果数の上限を 100 件にしている。

検索実行後は、検索結果画面の html ファイルを取得し、スクレイピングを行い、スニペットを収集する。収集したスニペットを MeCab により形態素解析する。この時、固有表現辞書である mecab-ipadic-NEologd [10] を使用することにより、新出の固有表現を取得できるようにした。

5.2 読み句生成

5.2.1 条件設定

今回の読み句生成実験では、音数を考慮したか否かで生成句にどのような違いが生まれるかを検証する。テンプレートファイルから読み句における品詞の配置構造のみを取り出して生成した場合と、品詞の配置構造と音数構造について取り出して生成した場合とでの結果を比較する。

本研究の生成手法においては、テンプレートから取り出す情報の最低限は品詞配列である。しかし、これだけの情報で新規に読み句を生成することは難しい。例として、テンプレートへ名詞を総当たりで適用するとする。上毛かるたの読み句構造では、1 句につき最大 4 つの名詞が含まれる。これに対

表2 「つくば」「筑波山」で収集した文書から取得した関連語候補
上位 20 件

関連語	文書数
つくば	45
筑波	37
山	37
筑波山神社	20
筑	20
波山	19
情報	19
つくば駅	13
2016 年	12
つくばエクスプレス	10
茨城県つくば市	10
シャトルバス	10
つくば市	10
温泉	10
宿泊	10
こと	10
茨城県	9
女体山	9
山頂	9
周辺	9

し、新規の名詞を 50 個用意し、総当たり、つまり順列を用いて当てはめていった場合、1 つの読み句テンプレートにつき $50!/(50-4)! = 5,527,200$ 件の文字列が生成されることとなる。これに加え、形容詞、副詞まで考えると、さらに生成数が上昇してしまう。

そこで、特徴語と関連語の共起度が一定以上あり、テンプレートの音数構造に嵌まる語句のみに絞り込むことによって、読み句として価値の低い文字列の生成を防ぐことができると考える。今回は、品詞配列の内、名詞部のみに着目して、固有名詞である特徴語と、特徴語の関連語を組み合わせることによって、読み句の生成を行った。動詞、副詞などその他の品詞については、テンプレートファイルの morpheme 配下の words 要素に記述した、既存かるた読み句で用いられていた語句をそのまま使用した。

5.2.2 関連語の収集・選別

今回は茨城県つくば市を対象地域とし、筑波山を特徴語として読み句を生成した。対象地域として「つくば」、特徴語を「筑波山」として検索を行った結果、1270 語の関連語候補を収集した。その内、出現文書数が上位 20 件の語句を表 2 に示す。AND 検索を行った対象地域である「つくば」が 45 件と最も多く取得でき、その次にはつくばの漢字表記である「筑波」が登場する。そこからは「山」、「筑波山神社」、「筑」、「波山」と続く。

続いて、取得した関連語候補について、「つくば」と組み合わせた AND 検索により、語句収集を行った。今回は、特徴語について関連語候補を収集した際に、出現文書数が上位 100 件の語句を対象として収集を行った。

取得したスニペット群中に「筑波山」が出現する文書数を数えた結果、上位 20 件の語句を表 3 に示す。表 2 以上に「つく

表3 「つくば」「筑波山の関連語」で収集した文書中で
「筑波山」が出現した関連語 上位 20 件

関連語	文書数
筑波山神社	90
筑	90
直行	88
ケーブルカー	82
ロープウェイ	72
中腹	70
男体山	67
つくば	66
女体山	65
山頂	64
つつじヶ丘	62
山	61
東の	57
温泉	52
877m	52
日本百名山	51
一望	50
関東平野	47
日帰り温泉	45
グランドホテル	40

ば」や「筑波山」という語句を含んだ語句を用いて検索した際に、特徴語「筑波山」が出現していることがわかるが、それ以外にも「直行」や「ケーブルカー」、「ロープウェイ」、「中腹」、「男体山」、「女体山」のような、筑波山に関係するか、対象地域であるつくばにおいて、筑波山と一緒に用いることが多い名詞が上位に登場している。

これらの結果を用いて、筑波山と各関連語の共起度を算出した結果上位 20 件が表 4 である。概ね表 3 に近い語句の順序が取得できた。読み句を生成するにあたり、生成に使用する関連語との共起度の下限値を設定することによって、生成について共起が薄い関連語を無視し、不必要な読み句の生成を回避することが可能である。

5.2.3 音数を考慮しない読み句生成

共起度ごとに、実際に生成した読み句の数と、使用した関連語句数を表 5 に示す。共起度によってフィルタリングをしない限り、関連語の出現件数上位 100 件と変わらない個数の関連語群が使用される。その結果、直積で考えうる組み合わせ全てで読み句がテンプレートの名詞部に適用され、莫大な数の読み句が生成された。ここでの 43,196,900 件ある生成句群は、名詞部に対して新規語句を当てはめた結果生成することとなる全ての読み句の組み合わせを持つ文書集合である。この生成に対して適切な条件と設定を与えることによって、有用な読み句を絞り込んでいく。

まずは共起度下限値を 0.1 に設定し、使用する関連語数を凡そ半数までに絞り込み、読み句の生成を行った。その結果、生成句数は 7,500,000 程度に絞り込むことができた。共起度下限値を 1 に近づければ近づけるほど使用する関連語を絞り込むことができる。次に 0.3 まで共起度下限値を押し上げると、生成

表 4 特徴語:「筑波山」の関連語との共起度 上位 20 件

関連語	共起度
直行	0.929
筑波山神社	0.924
筑	0.924
ケーブルカー	0.880
ロープウェイ	0.768
つくば	0.766
中腹	0.740
男体山	0.735
山頂	0.723
女体山	0.718
山	0.715
つつじヶ丘	0.667
東の	0.6421
温泉	0.585
877m	0.566
日本百名山	0.556
関東平野	0.545
一望	0.535
筑波	0.492
日帰り温泉	0.480

表 5 共起度に基づく音数を考慮しない筑波山読み句生成数

共起度下限値	0.0	0.1	0.3	0.7
使用関連語数	100	56	27	11
生成句数	43,196,900	7,476,952	808,866	49,522

される句の数は 800,000 程度に落ち着く。共起度が 0.7 以上の単語のみを使用するよう設定した場合、49,522 語まで生成句の数を抑えることができた。共起度下限値 0, 0.1, 0.3, 0.7 について、生成した読み句をランダムで 10 件ずつ抽出した結果を表 6 に示す。生成句は、形態素ごとにカンマによって分割されており、区切れ部は空白として表現している。

5.2.4 音数を考慮した読み句生成

次に、テンプレートファイルが持つ音数構造を参照して、読み句の生成を行った。ある位置の形態素の種類を参照する際に、その形態素が持つ音数と特徴語や関連語が持つ音数と一致するのかを判断し、語句の取捨を行う。特徴語と関連語の音数については、特徴語は入力時に読み仮名を同時に入力し、手で与える。関連語は、形態素解析の結果から読み仮名を取得する。

その後読み仮名から、「ア」、「イ」、「ウ」、「エ」、「オ」、「ヤ」、「ユ」、「ヨ」、「ワ」の小書き文字を考慮し、読み上げた際の音数へと変換する。音数を考慮した語句選択には、この音数を利用する。

5.2.3 節と同様、表 7 に共起度下限値の設定別に行った生成結果をまとめた。表 5 と比較して、表 7 で示した生成句数は著しく減少している。共起度下限値の設定を行わない時点で、読み句生成数は 19,711 句に抑えられた。共起度下限値の設定値を上昇させて、使用する共起度の絞り込みを行うことで、さらに読み句の生成数は絞り込まれた。最終的には共起度 0.7 以上の値を持つ関連語 11 語のみを使用するように絞り込んだ結果、その中で適切な音数構造で読み句を作った場合に 36 句生成す

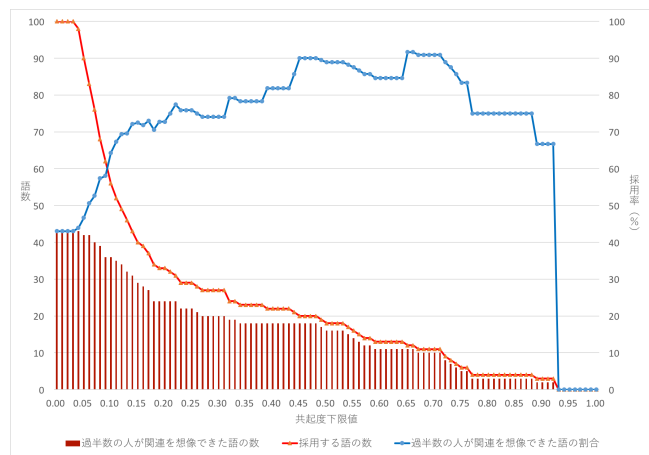


図 4 共起度による関連語選別実験の結果

ることができた。

5.2.3 節と同様に、生成した読み句群の中から無作為に 5 句を選出し、その結果を表 8 に示す。共起度下限値を 0 としていた段階と比較して、値を上昇させることで筑波山と直接の関係が薄い「等」「口コミ」のような語句が選択されなくなった。

5.3 人手による共起度の有効性評価

生成実験における共起度により語を選別する手法の有効性を検証するため、今回の読み句生成に使用した関連語 100 語について、実験協力者による判定実験を行った。実験協力者は 7 名で、筑波大学に在籍し、筑波山への来訪経験や Web 検索や書籍による調査知識を持つ、いずれも 20 代の学生あるいは研究員である。実験協力者は、5.2.2 で収集した筑波山の関連語候補 100 語に対し、筑波山と各関連語候補間で

- (1) 関係が想像できる
- (2) 関係が想像できない

の二択で判定をした。この時、5 秒を目安とし、判断がつかなかったものについては『関係が想像できない』を選択するように指示した。

評価実験の結果を図 4 に示す。関連語候補中から採用する語の数は、共起度下限値を上昇させることで初期値の 100 から急激に減少していき、0.94 を超えたところで 0 になる。関連語候補の内、実験協力者の過半数が特徴語との関係を想像できた語の数を棒グラフで表している。この採用する語の数と棒グラフとの隙間部分が、半数未満の実験協力者が関係を想像できないとした語の数を表している。そして、採用する語の内、実験協力者の過半数が関係を想像できた語の割合を表しているのが、もう 1 つの折れ線グラフである。共起度下限値の設定によって急激に割合が増加し、なだらかな上昇を見せた後は、分母の減少により割合そのものが低下し、0 に収束する。共起度下限値の値を引き上げていくことによって、採用する語の内、人手によって特徴語との関係を想像できるような語を最大で 91.7% 含めることが可能だと示している。

5.4 考 察

今回の生成実験では、特徴語と共起する語を優先して適用し、単語ごとの音数要素を考慮に入れることによって、郷土かるた

表 6 音数規則を含まない「筑波山」に関する生成句例

共起度下限値	基となった読み句	生成句
0.0	草津よいとこ / 葉の温泉	男体山 よい サイト / バス の 筑波山
	理想の電化に / 電源群馬	筑波 の 日本百名山 に / 8 案内
	水上、谷川 / スキーと登山	筑波山 日帰り温泉 / 人気 と 露天風呂
	浅間のいたづら / 鬼の押し出し	標高 の 多く / バス の 筑波山
	仙境尾瀬沼 / 花の原	情報 宿泊 / 筑波山 の 茨城県つくば市筑波
0.1	理想の電化に / 電源群馬	筑波山 の つくば に / 866 筑波山神社
	理想の電化に / 電源群馬	中腹 の グランドホテル に / つくばね 筑波山
	仙境尾瀬沼 / 花の原	つつじヶ丘 筑波山 / アクセス の ロープウェイ
	仙境尾瀬沼 / 花の原	グランドホテル ケブルカー / 日 の 筑波山
	太田金山 / 子育て呑竜	湯 筑波山 / つつじヶ丘 日
0.3	草津よいとこ / 葉の温泉	筑波山 よい 筑 / 波山 の つつじヶ丘
	仙境尾瀬沼 / 花の原	ケーブルカー 日本百名山 / 一望 の 筑波山
	繭と生糸は / 日本一	つくば と つつじヶ丘 は / 筑波山 露天風呂
	繭と生糸は / 日本一	筑波山 と 男体山 は / 日帰り温泉 茨城県つくば市筑波
	水上、谷川 / スキーと登山	山 筑波山 / 日帰り温泉 と つくば
0.7	三波石 / と共に名高い / 冬桜	直行 / と共に 名高い / 筑波山
	登る榛名の / キャンプ村	登る 女体山 の / 筑波山
	中山道 / しのぶ安中 / 杉並木	筑波山 / しのぶ ロープウェイ / 筑波山神社
	利根は坂東 / 一の川	直行 は 筑波山神社 / 筑波山 の ケブルカー
	浅間のいたづら / 鬼の押し出し	直行 の 筑 / つくば の 筑波山

表 7 共起度と形態素音数を考慮した筑波山読み句生成数

共起度下限値	0.0	0.1	0.3	0.7
使用関連語数	100	56	27	11
生成句数	19711	2092	119	36

の読み句として不適な読み句の生成を抑制することが可能であるか検証した。

本研究中での不適な読み句とは、

- ・ 特徴語について記述されていない
- ・ 音数が多く、冗長である
- ・ 日本語表現として誤りがある

の 3 点とする。

特徴語について記述されていない読み句は、Web 上の文書中において、特徴語と共起する語、関連語を収集し、それを特徴語と組み合わせることによって特徴語と無関係な語句を避ける。共起度について選択の下限を設定することで、特に特徴語と関係の強い語句を組み合わせることができることを期待して、その通りに実行した。その結果、筑波山においては「直行」「筑波山神社」「筑」「ケーブルカー」「ロープウェイ」の五つの関連語が上位に現れた。筑波山神社は、その名の通り筑波山を神体山とする神社であり、筑波山との縁が非常に深い。ケーブルカー、ロープウェイについては、筑波山の麓から頂上までを繋ぐ、筑波山の名物である。また、つくば市は交通網の整備により、つくば駅のバスターミナルを発着駅として、筑波山を始めとした各所への直行バスが数多く運用されていることから、直行という語が筑波山と共起したと考えられる。「筑」については、形態素解析の過程で出現した、「筑波」を含む語句の一部であると考えられる。文脈により、「筑波山」が「筑」と「波山」に分解される箇所とそのまま「筑波山」と辞書に認識される箇所と

が、あるスニペット群中に同居していたことが、「筑波山」「筑」の共起に繋がったのではと考える。表 4 にある通り、「波山」が共起度上位 30 件に出現していることも、この仮説を裏付ける要素の 1 つである。

そして、共起度によって語の選定をすることは、普遍的に使われる一般名詞や、形態素解析によって獲得された筑波山とは直接関係のない単漢字、単数字のようなノイズを排除する手段としても有効である。今回提案した手法では、特徴語と採用する複数の関連語を組み合わせる読み句を生成する。不要な語を取り除き、人が関係を想像できるような特徴語と共起する語を採用することによって、計算コストを下げながら、よりかるとらしい読み句の生成の実現に近づくことができる。

音数が多く冗長な読み句の生成を避けるには、七五調に則って構成された既存の郷土かるたに倣うように、音数の制約を語句選択時に与える手段を実行した。音数制約を与えることにより、無差別に生成された読み句から、人が作った郷土かるた読み句の構造にふさわしくない語句配置のものを一括で取り払うことができた。

以上の結果から、表 8 のような読み句を生成することができた。声に出して読み上げて違和感のあるような、無理な音数配置の読み句は生成されず、入力した地域と特徴語と全く関係のない語句は取り除かれている。しかし、日本語意味的な不和は大きな問題として抱えたままである。上毛かるたから引用した動詞や、周辺語句との接続の良し悪しを、今回の生成においては考慮に入れておらず、結果として無理のある内容となった読み句が数多く出力されている。

生成した読み句の中には、「つくばに映える 筑波山」「ロープウェイ と共に名高い 筑波山」のような、読み句として意味が通っているようなものも存在している。このような読み句の

表 8 音数構造を用いた「筑波山」に関する生成句例

共起度下限値	基となった読み句	生成句
0.0	桐生は日本の / 機どころ	店舗 は 利用 の / 筑波山
	雷と空っ風 / 義理人情	風呂 と 筑波山 / 地図 周辺
	雷と空っ風 / 義理人情	こと と 筑波山 / 7 観光
	雷と空っ風 / 義理人情	7 と 筑波山 / 0 今回
	雷と空っ風 / 義理人情	得 と 筑波山 / 山 宿泊
0.1	碓井峠の / 関所跡	つつじヶ丘 の / 筑波山
	登る榛名の / キャンプ村	登る 筑波 の / 筑波山
	中山道 / しのぶ安中 / 杉並木	ケーブルカー / しのぶ 直行 / 筑波山
	雷と空っ風 / 義理人情	筑 と 筑波山 / 山 アクセス
	雷と空っ風 / 義理人情	筑 と 筑波山 / 神 直行
0.3	しのぶ毛の国 / 二子塚	しのぶ 山頂 / 筑波山
	桐生は日本の / 機どころ	つくば は 筑波 の / 筑波山
	関東と / 信越つなぐ / 高崎市	中腹 と / 直行 つなぐ / 筑波山
	関東と / 信越つなぐ / 高崎市	東の と / 直行 つなぐ / 筑波山
	関東と / 信越つなぐ / 高崎市	東の と / 温泉 つなぐ / 筑波山
0.7	碓井峠の / 関所跡	ケーブルカー の / 筑波山
	三波石 / と共に名高い / 冬桜	ロープウェイ / と共に 名高い / 筑波山
	しのぶ毛の国 / 二子塚	しのぶ 中腹 / 筑波山
	銘仙織り出す / 伊勢崎市	中腹 織り 出す / 筑波山
	三波石 / と共に名高い / 冬桜	筑波山 / と共に 名高い / 女体山

取得を可能としていくことが、本研究における次の段階として不可欠な要素であり、単純に語句配置の条件で絞り込むだけでなく、今後は、日本語としての意味に着目して読み句の選出を行っていくべきだと考える。

6. 結 論

本論文では、郷土かるたに注目し、新規地域について読み句を生成手法を提案した。

提案手法では、郷土かるたが持つ独特な構造を新規読み句でも再現するため、形態素解析を用いて郷土かるたのテンプレートファイルを作成した。そこへ題材とする地域における特徴的な事象を指す語句「特徴語」と、特徴語と Web 文書中で共起する「関連語」を適用することによって、読み句の生成する。その過程で、共起の度合いを算出し、より強く特徴語と共起する関連語を読み句へ採用することにより、特徴語と無関係な語句を取り除けることを示した。さらに、テンプレートが持つ形態素あたりの音数を参照することによって、既存かるたの音数構造の枠を外れるような読み句の生成を防ぐことに成功した。今後は、テンプレートから参照する読み句の特徴について精緻化を行うとともに、語句の選出手法の改良、前後の品詞との共起を参照するなどの改良を行うことにより、より自然的に特徴語について表現し、郷土かるたとして実用的な読み句の生成機構の構築を目指す。

謝 辞

本研究は JSPS 科研費 JP16H02904 の助成を受けたもので、ここに記して謝意を示します。

文 献

- [1] 原口美貴子, 山口幸男. 郷土かるた、上毛かるたの魅力と意義—郷土かるた王国「群馬」からの発信. 群馬大学教育学部紀要 人文・社会科学編, Vol. 59, pp. 9–20, Feb. 2010.
- [2] 原口美貴子, 山口幸男. 「上毛かるた」の札の分析: 社会科郷土学習の基礎資料として. 群馬大学教育学部紀要 人文・社会科学編, Vol. 45, pp. 197–214, 1996.
- [3] 上毛かるた日本一へ!! - 上毛かるた日本一決定戦 "king of jmk" オフィシャルサイト. <https://www.kingofjmk.jp/>, 参照 Dec. 21, 2016.
- [4] 内藤将司, 水野慎士. カルタを使ったデジタルコンテンツの提案. マルチメディア、分散協調とモバイルシンポジウム 2013 論文集, Vol. 2013, No. 2, pp. 2151–2156, Jul. 2013.
- [5] 大澤由香里, 福盛貴弘. 上毛かるたのプロソディー研究. 北海道言語文化研究, Vol. 9, pp. 47–62, 2011.
- [6] 檀裕也, 墨岡学, 和田武. 英語俳句の特徴抽出と文書ベクトルの構成. 情報処理学会全国大会講演論文集, Vol. 73, No. 1, pp. 515–516, Mar. 2011.
- [7] Alham Fikri Aji. Automated haiku generation based on word vector models. Master's thesis, University of Edinburgh, 2015.
- [8] Kudo Taku. Mecab: Yet another part-of-speech and morphological analyzer, 2016. <http://taku910.github.io/mecab/>, 参照 Dec. 19, 2016.
- [9] Selenium - web browser automation. <http://www.seleniumhq.org/>, 参照 Dec. 19, 2016.
- [10] Toshinori Sato. neologd/mecab-ipadic-neologd: Neologism dictionary based on the language resources on the web for mecab-ipadic. <https://github.com/neologd/mecab-ipadic-neologd>, 参照 Nov. 1, 2016.