

科学研究費助成事業 研究成果報告書

平成 28 年 6 月 23 日現在

機関番号：27301

研究種目：基盤研究(C) (一般)

研究期間：2012～2015

課題番号：24500313

研究課題名(和文) ロンゴロンゴ記号列からのノンパラメトリックな情報抽出

研究課題名(英文) Non parametric information extraction from Rongorongo symbol sequences

研究代表者

山口 文彦 (Yamaguchi, Fumihiko)

長崎県立大学・その他部局等・教授

研究者番号：60339124

交付決定額(研究期間全体)：(直接経費) 3,100,000円

研究成果の概要(和文)：モアイ像で有名なイースター島では、かつて緻密な記号の列を彫り付けた木製品が制作された。この記号列はロンゴロンゴと呼ばれ、文字かもしれないと言われているが、読める人が現存せず未解読である。ロンゴロンゴは歌を記憶する助けとして刻まれたとする説があるが、具体的にどの記号列がどの歌を表しているかは分かっていない。また、記号を形状で分類した既存研究があるが、この分類が文字としての分類を表しているとは確認されていない。そこで、現地の歌の中の音節の出現順序と記号列の中の記号の出現順序の比較および記号画像の文字への分類について計算機を用いた実験を行った。結果として、対応する可能性のある歌と記号列が見つかった。

研究成果の概要(英文)：Rongorongo is sequence of symbols which is engraved on the surface of wooden artifacts in the Easter Island. Though these symbols may be characters, they are undeciphered. Rongorongo is said to be engraved in order to assist a person in memorizing chants. However, correspondence between symbol sequences and chants is unknown. A previous work classified the symbols by the shapes. However, it's unconfirmed that the classification express the character class. In this research task, occurrence orders of syllables in the chants and symbols in Rongorongo are compared. And the images of symbols are classified by Zipf's law which is a rule of thumb in statistics. In deciphering unknown language, we don't have the correct answer. Therefore, the methods should be evaluated. In our case, the comparing method and the classification method and are evaluated by using known language such as Japanese. As a result, some possible corresponding pairs of chant and Rongorongo sequence are found.

研究分野：計算機科学

キーワード：未解読文字 ロンゴロンゴ 自然言語処理

1. 研究開始当初の背景

モアイ像で知られるイースター島では、かつて緻密な記号の列を刻み付けた木製品が作られた。この記号列はロンゴロンゴと呼ばれ、文字であるかも知れないと言われているが、現在は読める人がいない未解読文字である。ロンゴロンゴが刻まれた木製品は 20 個余りがのこされており、現地の言葉や所蔵する博物館の所在地に由来する名前が付けられている。多くは板状で、両面に記号が彫りつけられている。それぞれの面は A/B または Recto/Verso という名前で区別される。

図 1 は Aruku-Kurenga と呼ばれるロンゴロンゴ板の Recto 面の冒頭にある記号列である^[1]。ロンゴロンゴ記号のそれぞれは、自然物や人を象形しているように見える。

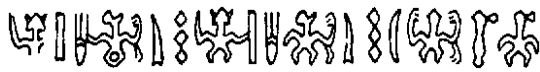


図 1. ロンゴロンゴ記号列の例

ロンゴロンゴの解読を目的とするさまざまな先行研究がある。Fischer はロンゴロンゴだけでなくイースター島の歴史や文化的背景などを一冊の本にまとめている^[2]。19 世紀にタヒチに伝道に赴いた Jaussen がロンゴロンゴが刻まれた木製品をいくつか所有しており、イースター島出身者に読ませたという記録がある。ロンゴロンゴは歌うように読まれたが、その内容は正しく読めたといえるものではなかったという。これはロンゴロンゴを読む様子をまねただけであると考えられている。Métraux はロンゴロンゴ文字の異なり数を 120 くらいであるとし、これは表音文字と考えるには多過ぎ、表意文字と考えるには少な過ぎると指摘している。Barthel はロンゴロンゴの記号をその形状の特徴で分類し、各分類に 3 桁の数字からなるコードを付した^[3]。このコードによって、木製品の表面に彫られた線から成る記号を、計算機上のテキスト情報として扱うことが可能になる。しかし、Barthel のコード化は形状で分類したものであって、文字への分類ではないことに注意が必要である。一方でロンゴロンゴの解読を示した仮説の中には、記号とその読みや意味との対応づけにおいて論理的な根拠に乏しい説も多い。ロンゴロンゴについて仮説が得られたとしても、未解読の言語であって正解が不明であるため、その仮説を検証することは困難である。

近年、失われた言語を対象に計算機を用いて情報を抽出しようとする研究が盛んである。こうした研究の背景には、統計的自然言語処理研究の発展がある。統計的自然言語処理は、辞書や文法などの言語の種類に依存する情報をできるだけ使わずに、文書そのものから得られる文字の出現頻度や、連続して出現する頻度をもとに言語や文書についての情報を得ようとする手法であり、Web 検索などの分野で用いられている。こうした手法そ

のものは言語に依存しないため、さまざまな言語に適用できるものと期待できる。しかしそれぞれの言語において良い性能を得るためには、言語に依存して適切にパラメータを調整する必要がある。通常は既知の言語を対象としており、その場合は期待する出力か否かという意味での正解を設定できるため、機械学習の手法を用いてパラメータを学習する方法が使われることが多い。自然言語処理の手法を失われた言語に適用する研究においても、現在は使われていないものの解読されている言語を対象とすることが多い。

2. 研究の目的

本研究では、統計的自然言語処理をはじめとする人工知能分野や計算機科学分野で培われた技術を用いて、ロンゴロンゴの解読に資する情報の抽出を試みる。

ロンゴロンゴの解読を困難にしている原因の一つは、地理的・文化的に孤立した場所で作られた文字であるために、同じ内容を別の言語で記した文書が存在しないことである。ロンゴロンゴが歌を記録するために作られたとする説があり、歌については、現地にある博物館が聞き取りによって収集してアルファベットで記録するという仕事をしている。そこで、ロンゴロンゴ記号の出現順序と歌詞におけるアルファベットの出現順序から、対応する可能性のあるロンゴロンゴと歌の組を見つけることができなかと考えた。

また、Barthel の分類は記号の形状にもとづいており、文字に分類しているわけではない。そのまま文字への分類として用いるには、Barthel の分類は詳細すぎるのではないかと指摘もある。もちろん似た形状の記号であれば同じ文字を表していると期待されるが、違う文字の形状が類似していることや、同じ文字を表す記号の細部が異なることも考えられる。そこで、記号の形状の特徴だけでなく、出現頻度などの情報から、文字としての分類を見つけることができなかと考えた。

3. 研究の方法

ロンゴロンゴの記号列とアルファベットの列(歌詞)が同じ歌を表しているとき、各記号にはその読みを表すアルファベットがあつて、それぞれの列において同じ順序で対応する記号とアルファベットが登場するものと仮定する。ロンゴロンゴの各行と、歌詞について、そのような対応づけが存在するか否かについて、計算機を用いて網羅的に調べた。ここで、歌詞についてアルファベット(表音文字)をそのまま用いるか、音節や単語などの単位に分けるかという選択肢がある。ポリネシアの言葉においては子音が連続することはほとんどなく、日本語と同様に母音または子音一つと母音で音節を構成する。イー

スター島で収集された歌の歌詞に現れる音節の異なり数は約 120 で Métraux の見積もった文字の異なり数と符合した。一方で、表音文字と表意文字が混在する可能性もある。記号が意味を持った単語を表す場合、一つの記号が複数の音節からなる読みを持つと推測される。本研究では、一つの Barthel コードが 0 個以上の音節に対応すると仮定した場合に、ロンゴロンゴの行が音節の列（歌詞）と一致する可能性があるか否かを調べた。網羅的に調べることによって、一致する可能性が無い場合に、無いと分かる。ただし、一致する可能性があるかと判定されたとしても、そのロンゴロンゴの行と歌詞が同じ歌を表している、ただちに判断することはできない。

このような記号の出現順序を比べる手法によって対応付けが見つけれられるか否かについては検証が必要である。しかしロンゴロンゴと歌詞から得られた結果のみでは、正解（対応の有無）が不明であるため、検証できない。そこで、既知の言語である日本語を用いて、同様の問題を設定し、同じ手法で対応づけを見つけられるか否かを調べた。この実験によって、手法そのものの確からしさを調べることができる。

記号と音節の出現順序によるロンゴロンゴと歌詞の対応付けを発見しようとする試みにおいては Barthel コードを用いたが、Barthel コードが文字よりも詳細な分類となっているために、対応づけを見逃している可能性がある。そこで、記号を文字に分類する説得力のある方法が必要である。記号の形状を特徴量として抽出する手法については、手書き文字認識の分野で研究されている。本研究では、輪郭線の方向成分特徴と呼ばれるものを用いた。この記号の形状特徴を用いて、似た文字を同じクラスに分類することができるが、文字の異なり数が不明であるために、いくつかのクラスに分類すればよいのかが分からない。そこで、言語学上の経験則である Zipf 則を用いて、分類の仮説のうちから、もっともらしいものを選ぶという方法を用いた。これは多くの組合せの中から、ある評価値を最大化する問題として一般化でき、そのような問題を解くための手法として、遺伝的アルゴリズムを採用した実験を行った。

文字クラスへの分類についても、未解読言語では正解が分からないために評価ができないという問題がある。そこで前述の歌詞との対応づけを見つけようとする研究と同様に、既知言語での実験を行って手法そのものの評価を行った。既知言語としては日本語を用い、データとして手書き文字認識の認識器の学習に用いられる手書き文字データベース⁴⁾を用いた。

4. 研究成果

ロンゴロンゴの行と現地の歌の歌詞との対応付けについては、4 つの大きなロンゴロンゴ木製品に含まれる 83 の行と、372 編の歌

詞の組合せ 30,876 通りのすべてについて可能性を調べ、618 通りについて可能性があることが分かった。逆に、残りの 30,258 通りについては、網羅的に調べているために、Barthel コードと音節の対応を見るかぎり、ロンゴロンゴと歌詞が対応する可能性がないと分かった。見つかった対応付けの例として、Aruku-Kurenga の Verso 面 1 行目と E

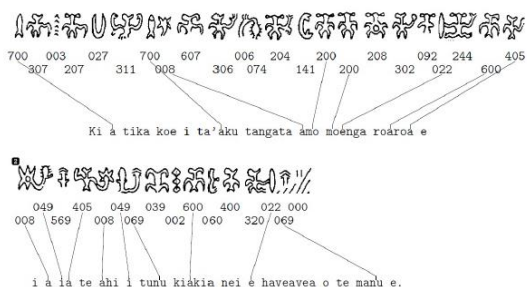


図 2. 見つかった対応付けの例

Ira e, e Raparenga e #5 の歌詞の対応を図 2 に示す。この成果は国際会議 CICLING2013 で発表され、Research in Computer Science 誌に掲載された。なお、同じ手法を日本語のかな漢字混じり文とカタカナ文の対応付けに用いて、手法の評価を行った。無作為に選んだ 20 個のかな漢字混じり文とそれをカタカナで表記したものの組合せ 400 通りについて、対応しているもののうちで、対応していると正しく判定されたものの割合（再現率）は 100%であり、対応していると判定されたもののうちで、正しく対応しているものの割合（適合率）は 11%であった。この適合率は、かな漢字混じり文の中に複数回出現する文字の種類が多いほど高くなり、7 種類の場合には 53%となる。日本語による考察から、この手法では、対応していないものを対応していると誤判定することがあるが、その可能性は頻出する文字の種類が多いほど少ないと分かった。そこで、ロンゴロンゴの列の中で頻出する記号の種類が 7 種類以上のものに限定すると、対応すると判定された 618 通りのうちの 39 通りが該当した。

統計的自然言語処理の分野では、文字や単語が連続して出現する頻度である N-gram を用いることが多い。暗号文の各記号が一つの文字を表す換字暗号では、こうした頻度情報によって解読できるものもある。そこで、ロンゴロンゴにおける Barthel コードおよび歌詞の音節の出現について N-gram を調べて比較するという研究を行った。結果として、Barthel コードの方が頻度の分布が平らであって、Barthel コードと音節が 1 対 1 に対応すると考えるのは不自然であることが確認された。この成果は国際会議 PACLING2013 で発表された。

ロンゴロンゴ記号の文字への分類を試みるのに先立って、記号の形状特徴を手書き文字認識の手法で抽出し、その類似度によって階層的にクラスタリングする研究を行った。

結果として、Barthel の分類を包含するようなクラスタリングが得られた。ただしこの結果は階層的クラスタリングであるため、文字クラスへの分類ではない。この研究の意義は、形状特徴を機械的に抽出することによって、人間が手作業で分類することによる判断の揺れなどの影響を小さくする道筋が得られたことにある。

上記の研究で得られた階層的なクラスタリングをもとに、頻度と頻度の順位が両対数グラフ上で直線にのるという経験則 (Zipf 則) に従う度合いを評価値として、非階層的なクラスタリングを求める実験を行った。これによって文字クラスへの分類が見つかることを期待したが、結果として得られた分類は、文字のクラスを表していると考えには不自然なものであった。言語においては、単語の出現頻度は Zipf 則によく従うが、文字の出現頻度は単語ほど従わないことが分かっており、Zipf 則以外の評価が必要と考えられる。

記号の文字への分類についても、未解読言語のみでは評価できないため、手法を既知言語で評価する必要がある。日本語の手書き文字データを用いて、同手法を評価した。結果として、Zipf 則がクラス数(文字の異なり数)を小さくする方向にのみ働く様子が明らかになり、Zipf 則以外にクラス数を増やす方向に働く評価値が必要であることが確認できた。

これまでに述べたように、未解読言語に適用しようとする手法を提案するとき、既知言語を用いて、その手法自体の評価を行う必要がある。このとき、言語や手法の特徴にできるだけ依存しない、手法そのものの評価軸が必要である。そこで、記号の画像を文字へ分類する問題について数理的な考察を行った。分類を見つけることは同値関係を見つけることであり、分類の正解の度合いは、有限な集合の同値関係の類似度に帰着できる。このことから、文字画像分類手法を評価する指標を提案した。指標を提案することで、明確な評価が可能となり、今後、手法の提案や他の言語への適用といった研究を加速させることが期待できる。

これらの、ロンゴロンゴ記号を文字へ自動的に分類しようとする研究の諸成果は、日本情報考古学会ならびに情報処理学会全国大会において発表された。

参考文献

- [1] "Rongorongo or the Hieroglyphs of the Easter Island Tablets", rongorongo.org, <http://www.rongorongo.org/>, (2003年9月参照), (現在は <http://kohautomotu.org/> に移設されている)
- [2] S.R. Fiesher, "Rongorongo -the Easter Island Script-", Clarendon Press, Oxford Studied in Anthropological Linguistics, vol. 14, 1997

[3] T. S. Barthel, "Talking Boards of Easter Island", Scientific American, 198, Jun., 1958, pp. 61-68

[4] M. Nakagawa, K. Matsumoto, "Collection of on-line handwritten Japanese character pattern database and their analysis", International Journal on Document Analysis and Recognition, vol.7, no.1, Springer, 2004, pp.69-81

5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文] (計 5 件)

- ① 山口 文彦, "文字画像分類手法の評価指標の提案", 日本情報考古学会第36回大会講演論文集, 査読あり, 2016, pp. 60-65
- ② 山口 文彦, "輪郭線方向成分と Zipf 則を用いた文字画像の自動分類", 情報処理学会第78回全国大会講演論文集, vol. 4, 査読なし, 2016, pp. 487-488
- ③ 山口 文彦, "Zipf 則を指標とするロンゴロンゴ記号の分類", 日本情報考古学会第35回大会講演論文集, 査読あり, 2015, pp. 80-85
- ④ 山口 文彦, "手書き文字認識手法を用いたロンゴロンゴ記号の類似度", 日本情報考古学会第33回大会講演論文集, 査読あり, 2014, pp. 78-83
- ⑤ Fumihiko YAMAGUCHI, "Comparison between Rongorongo and the syllable sequence of ancient chants from the Easter Island", Research on Computing Science, (ISSN: 1870-4069), vol. 70, Advances in Computational Linguistics, 査読あり, 2013, pp. 9-18

[学会発表] (計 6 件)

- ① 山口 文彦, "文字画像分類手法の評価指標の提案", 日本情報考古学会第36回大会, 2016/03/27, 九州国立博物館(福岡・太宰府)
- ② 山口 文彦, "輪郭線方向成分と Zipf 則を用いた文字画像の自動分類", 情報処理学会第78回全国大会, 2016/03/10, 慶應義塾大学(神奈川・横浜)
- ③ 山口 文彦, "Zipf 則を指標とするロンゴロンゴ記号の分類", 日本情報考古学会第35回大会, 2015/10/04, 共立女子大学(東京・神田)
- ④ 山口 文彦, "手書き文字認識手法を用いたロンゴロンゴ記号の類似度", 日本情報考古学会第33回大会, 2014/09/28, 東京農工大学(東京・小金井)
- ⑤ Fumihiko YAMAGUCHI, "N-grams on Rongorongo and chants in the Easter Island", Conference of the Pacific Association for Computational Lin-

- guistics 2013 (PACLING2013), 2013/09/02, Keio Univeristy (Tokyo/Mita)
- ⑥ Fumihiko YAMAGUCHI, "Comparison between Rongorongo and the syllable sequence of ancient chants from the Easter Island", 14th International Conference on Intelligent Text Processing and Computational Linguistics 2013, 2013/03/24-30, Multiplex 45 hall (Greek/Samos)

[その他]

ホームページ等

山口文彦, "Yamagu's Rongorongo Page",
<http://sun.ac.jp/prof/yamagu/Rongorongo>

6. 研究組織

(1) 研究代表者

山口 文彦 (YAMAGUCHI, Fumihiko)
長崎県立大学・教育開発センター・教授
研究者番号：60339124

(2) 研究分担者

(3) 連携研究者