様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　元 年　 ６ 月 ２２ 日現在

機関番号：２７３０１
研究種目：基盤研究(C)（一般）
研究期間：2015～2018
課題番号：１５Ｋ００４７２

研究課題名（和文）ロンゴロンゴのデジタルアーカイブと言語情報の抽出

研究課題名（英文）Linguistic information extraction and digital archive on Rongorongo

研究代表者
山口　文彦（Yamaguchi, Fumihiko）

長崎県立大学・情報システム学部・教授

研究者番号：６０３３９１２４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：本研究は木製品に彫られた記号の並びを対象とする。遺物のデジタルアーカイブに向け、実際の遺物と同程度の大きさの、より明瞭な記号が彫られた模型をSfMで形状測定する実験を行った。彫られた記号を識別するために、形状の局所平均的な表面からの高さの差を用いる方法を試した。言語情報の抽出としては、記号を文字に分類する問題に取り組んだ。未解読文字の解読は正解が不明なために評価が難しい課題であり、手法自体を既知の言語を使って評価する研究手法をとる。そこで既知言語の文字画像から言語情報を抽出する実験を行うために必要なデータの整備を行った。また失われた言語に対する計算機指向の研究を知ってもらうための発表を行った。

研究成果の学術的意義や社会的意義
本研究課題は、未解読言語の解読という、従来人間の考古学者が行ってきた知的活動を部分的に自動化しようとする試みである。情報機器を使うことの利点は、ネットワークを介した情報の共有や膨大な計算を遂行できることだけでなく、明確に定義された手順であれば、対象を変えても同じように遂行できることにもある。対象が変わっても問題の質が変わらないことを慎重に検討する必要があるが、従来人間の手で行われていた作業を自動化することによって、判断の揺れや恣意の混入を避けることができる。そういう方向での計算機(言い換えれば人工知能）利用を進めるきっかけになるものと考えている。

研究成果の概要（英文）：The target of this study is the sequence of symbols engraved on wood artifacts. For digital archivement, an experiment to measure the 3D shape of a more clear symbol engraved on a model with SfM is conducted. In order to recognize the engraved symbols, a method using the difference in height from the local average surface of the shape is tried. As a linguistic information extraction, the problem of classifying symbols into letters is tackled. Decoding undeciphered characters is a difficult task to evaluate, because the correct answer is unknown, and we use research methods to evaluate the method itself using a known language. Therefore, we prepared the data necessary to conduct an experiment to extract linguistic information from character images of known languages. We also made a presentation in order to let computer-oriented research on lost languages be known widely.

研究分野：計算機科学，自然言語処理

キーワード：ロンゴロンゴ　自然言語処理　失われた言語

様　式　C－１９，F－１９－１，Z－１９，CK－１９（共通）

１．研究開始当初の背景

現在はチリ領となっているイースター島では，かつて図１に見るような緻密な記号の列を彫りつけた木製品が作られた．この記号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，未解読である．

図 1. ロンゴロンゴの例，Mamari の A 面 2 行目

イースター島に西欧人が到達したのは 1722 年のことであるが，ロンゴロンゴについては 1864 年よりも前の記録がなく，失われた言語の中では製作された年代が比較的新しいものと考えられる．そのため，現存する木製品や記号の総数は少ないものの，ロンゴロンゴには連続した記号列（文書）全体が比較的明瞭に残されているという特徴がある．この特徴は，記号が連続して出現する頻度の情報を用いる統計的自然言語処理にとって有利に働く．

近年，失われた言語を対象に計算機を用いて情報抽出しようとする研究が盛んである．これらの多くは使う者がいないという意味で失われた言語を対象とするものの，多くの知見が得られて解読が進んだ言語を対象としており，未解読とされる言語を対象とする研究は少ない．未解読な言語を対象とする場合，結果をどのように評価するかが課題となる．

未解読文字を解読しようとする研究においては，まず記号列を記号に分離し，各記号を分類と同定を行っている．これまでの研究では，Barthel が提案した記号の分類法にしたがってロンゴロンゴを記号列として扱ってきたものが多い．しかし，いくつかの記号の中には研究者によって分類や同定に意見の相違がある．記号化は解読における最初の段階なので，この段階での判断の揺れは，後段の情報抽出において，結果を大きく変えてしまう可能性がある．また，考古学的な視点からは，記号の画像情報だけでなく，ロンゴロンゴが彫りつけられた木製品そのものの3 次元データからも，有用な情報が得られるのではないかと指摘を受けている．例えば，記号を彫るときに使用された工具の痕から，製作者の意図を汲み取ることができるものと考えられる．

言語情報抽出の課題として，記号を文字に分類する問題がある．記号の一つ一つは人の手で彫られたものであり，それぞれ微妙に異なる．もちろん，同じ文字であれば似た記号であると予想されるが，似た記号が同じ文字であるかどうかは，特に未解読文字の場合は不明である．ロンゴロンゴについては，記号に番号付けをした先行研究はあるものの，この意味での文字への分類には広く認知された結論がない．

２．研究の目的

ロンゴロンゴ記号列に対し，画像情報や3次元データなどの実物を計測した情報を得る．そうした記号の形状の情報から，文字分類などの言語情報を抽出しようとする．言語情報抽出の手法については、同じ手法を別の既知言語に適用した結果をみることで，手法自体の評価を行う．このような手法自体の評価を行うという研究の方法について，特に考古学を専門とする研究者からの意見をもらうことも研究の目的である．また情報の抽出手法を未解読言語に適用した場合には，正解が未知であるために，結果の評価が難しい．すなわち，計算機科学の範囲では解読の手かがりを提供することはできるが，解読そのものについては，考古学的な考察や評価が不可欠となる．そこで，得られた結果について，計算機科学的な評価だけでなく，考古学的な見地からも評価する．

３．研究の方法

遺物の所有者・所有機関に協力を要請し，3 次元計測を行おうとした．残念ながら協力が得られなかったため，3 次元形状計測については，実物と同程度の大きさの記号列が彫られた木製のモデルを対象として，表面に浅く彫られた文字が形状として認識できる形で計測できるかどうかを確認する実験にとどまった．なお、本課題の期間の間に、形状の 3 次元計測の技術が急速に普及し、多くの考古遺物が 3 次元計測されるようになった。

計測して得られた 3 次元データから彫られた線を識別するためには，局所的に平均的な表面と計測点との高さの差を用いる．

一連の線の情報を個々の記号に分割する問題については，今回の研究では人間の手作業に依った．記号を文字に分割するには，記号の特徴量を，手書き文字認識の手法を参考に取得し，特徴量の階層的な分類から，Zipf 則などを指標に非階層的な分類を得ようとした．

こうした言語情報について，未解読文字の場合は正解が分からないので結果の評価が難しい．そこで，計算の結果ではなく，手法そのものを，既知言語を用いて評価するという方法をとる．そのために，現代日本語の手書き文字，江戸時代の日本語手書き文字，エジプトの神官文字などの既知の文字についてデータを整理し，同じ手法で文字分類をしてその性能を測った．分類手法の性能評価については，本課題が採択される直前に提案した方法を用いる．

４．研究成果

結果として，3 次元計測については成果が出ていない．遺物の計測などの情報収集については，今後も所蔵機関などに対する働きかけを行っていく予定である．その点で何人かの研究者・遺物の所有者とコンタクトが持てたことは一定の成果であると考えている．

3 次元計測によって得られたデータから彫られた線を識別する問題については，平均的な表面を局所的に定義することによって，一定の成果が得られることを確認した．

記号を文字に分類する問題については，記号特徴量の距離をもとに階層的な分類を行い，それ以外の（Zipf 則などの）指標を使って非階層的な分類をしようとしたが，良い結果を得られていない．むしろ，非階層的な分類にも特徴量の距離を使う方が好ましい結果であり，既知言語に対して F 値（適合率と再現率の調和平均）が 0.4 程度であった．

本研究は，従来人間の考古学者が行ってきた手順の一部を自動化しようとするものであると見ることができる．このような自動化は，人間が作業する場合に比べて，判断の揺れや恣意の混入といった問題を避けることができる点で，学術研究における客観性の確保にとって有意義である．こうした観点は，情報技術に興味を持つ一部の考古学者に受け入れられているものと感じている．

なお，失われた言語に対する計算機指向の研究を，多くの研究者に知ってもらうための発表も行った．

５．主な発表論文等

〔雑誌論文〕（計　1　件）

山口 文彦, ``未解読言語の解析技術'', 人工知能学会誌, Vol. 31, No. 6, pp. 775-779, (Nov., 2016)

〔学会発表〕（計　1　件）

山口 文彦, ``未解読言語に挑戦する人工知能'', 情報処理学会 SPT-EIP-DSP 合同研究会 招待講演, 情報処理学会研究会報告 Vol.2016-SPT-21,2016-EIP-74,2016-DSP-168, No. 8, 2 pages, (Nov., 2016)

〔図書〕（計　0　件）

〔産業財産権〕
○出願状況（計　0　件）
○取得状況（計　0　件）

〔その他〕
ホームページ
http://sun.ac.jp/prof/yamagu/Rongorongo/

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし