地質・言語情報の数値化・統計解析処理に関する研究

地質・言語情報の統計解析（課題番号 127102）
研究期間：平成 24 年 7 月 26 日～平成 27 年 3 月 31 日
研究代表者：石原与四郎　研究員：乙武北斗

①研究成果

1. はじめに

本研究テーマでは，地質情報，言語情報という異なった対象に対してそれぞれの対象に適したアプローチによってこれらを定量的に評価することを目的とした．地質情報の解析においては，一般には肉眼によって判定される葉理境界について画像を利用した定量的かつ迅速な測定方法を試み，その有効性を検討した．一方，言語情報の解析においては，形態素解析における方言・口語的表現の影響，日本語敬語表現を対象とした自動添削モデルの構築を試みた．以下では，それぞれ詳細を説明する．

2.1　地質情報の解析

野外で得られる地質情報は，認定や測定誤差に起因するデータの不確かさを含むことが多い．たとえば最も基本的な地質情報の一つである地層の厚さの測定には，地層の境界の認定に加え，その測定誤差が含まれる．このような地質データから定量的な情報を得るためには様々なデータ取得方法が検討されているが，特に大量の情報を取り扱う必要のある場合にはデジタルデータの取得や統計処理が積極的に行われる．

本研究テーマのうち，地質情報に関わるものとしては，縞状堆積物の解析を行った．縞状堆積物の解析においては，岡山県真庭市の蒜山原層，栃木県那須塩原市の宮島層に認められる縞状珪藻土を対象に，年縞堆積物の自動的で定量的な層厚計測，認識方法を検討した．

2.1.1 縞状堆積物

縞状珪藻土は，それらを構成する珪藻の種類や季節的な流入物の違い等によって，層厚1 ～ 2 mmの年縞（ねんこう）堆積物を形成する．年縞堆積物は，年代のプロキシとしてだけではなく，その層厚から珪藻の生産量が見積もられたりすることで，堆積速度と環境変動との関連性が議論されてきた．年縞の構成は構成物や形成環境によって異なるので，実際の地層を解析するにあたっては，その最初の段階でそれらの検討が必要である．

岡山県真庭市の蒜山原高原に分布する蒜山原層は，中部更新統の湖成層で厚さ20 mほどの縞状珪藻土が認められる．縞状珪藻土は，厚さ1 ～ 2 mm程度の淡緑色と濃緑色の1枚のセットからなり，洪水流起源および湖底斜面崩壊の起源の重力流堆積物を挟む．年縞は約8000年分が認められることがわかっており，これらには湖成堆積物にしばしば認められる太陽黒点周期や気候変動に対応する周期的な層厚変動も認められていた．これらの層厚は露頭で撮影された写真画像を用いて，顕微鏡下で計測されており，人為的な誤差や再現性に乏しいことが推定される．一方，栃木県の那須塩原市に分布する宮島層は，中部更新統の湖成堆積物である塩原層群の主体を成す地層である．蒜山原層と同様に年縞と推定される細かい縞状珪藻土からなるが，相対的に多くの重力流堆積物を狭在する．宮島層は，重力流堆積物を引き起こしたイベントの解析をする上で極めて有効な地層である．

2.1.2 解析方法

縞状珪藻土の解析においては，年縞を成す葉理の認定および重力流堆積物の認定そして層厚の計測が重要である．年縞の認定においては，年縞を構成する季節性葉理の総合的な判定が必要なことから，目視で行われる場合や，かさ密度に置き換えられる物性値として軟X線透率を利用する場合が認められるが，前者は再現性が維持できない可能性が高く，後者は手間がかかる上に試料の条件が限定される．地質情報の解析では，野外でも室内でも容易に得られる写真画像を用いた．これらを用いて，蒜山原層では年縞葉理の認定と層厚計測，宮島層では，これらに加え重力流堆積物認定に判別分析を試み，その有効性を検討した．

解析においては，8 bitのグレー画像の濃淡（図2.1）の

変化率および平均層厚から得られた計算ウィンドウ内における中間値を基に年縞の季節性葉理を認定することとした．すなわち，淡緑色，濃緑色の葉理のそれぞれを認定した．重力流堆積物についてはこれらの認定に加え，24bitのRGB画像を用いて標準化した色調が年縞葉理とは特徴が異なることを利用して識別を試みた．厚さに関しては，年縞の葉理に対して垂直な方向へのpixel数をカウントして得た．

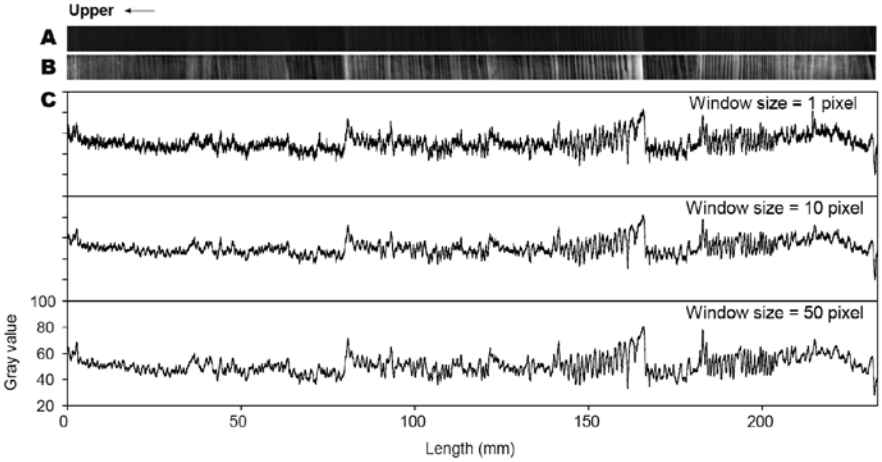

図2.1 年縞の8 bit グレー画像とプロファイルの例（Sasaki et al., 2015）

(2) 解析結果・問題点

上述のような解析手法を画像に適用した結果，年縞の認識は肉眼とほぼ同様に行われることがわかった．既存研究では年縞の変化率を利用したものがあるが，それらの結果よりもより肉眼での判定に近くノイズに強いという傾向が認められた．一方，この解析においては適切なウィンドウの設定が重要となり，濃淡プロファイルの事前の解析が必要となる．重力流堆積物は，主として塊状・均質のものであれば比較的正確に認定を行うことが可能であることがわかった．しかしながら，実際には緩やかな粒度変化を示す例等でうまく認識ができないものも認められた．より歴史記録としての重要性の高い現世の湖成堆積物を解析する場合，このような認定基準や解析手法の開発はより重要性が増すと考えられる．

2.2 言語情報の解析

2.2.1 言語的特徴の統計解析

言語的特徴の統計的解析の一環として，Web上に公開されている地方自治体の議会会議録を対象とした方言を含む発言の出現傾向の分析を行った．本分析においては，プログラムで自動処理によって収集，データベース化された168自治体の地方議会会議録を利用した．しかしながら，公開されている地方議会会議録すべてを対象に発言を分析することは，規模が非常に大きいため困難である．発言に方言が含まれるかどうかは，その発言者が方言を含む口語的な表現を普段から多用するかどうかという点に強く影響されると仮定し，本分析では地方議会会議録中の各発言者から1つの発言文を抜き出し，それらに方言が含まれているかどうかの判断を行った．また，方言が含まれると判断された発言文字列に対しては，形態素解析の結果，どのような解析結果が得られるかについても分析を行った．

発言における方言の有無の調査は，出身地の異なる３名が分担して行った．方言の判断は表2.2.1に示す基準に沿って，該当するラベルを各発言者の発言に付与することで行った．表2.2.1におけるラベル2は口語的表現がある場合，または判断者が意味を特定できない未知の単語がある場合に付与される．

都道府県別に発言のラベル付与状況をまとめたものを図2.2.1に示す．ラベルが１，すなわち方言と思われる表現が発言に含まれる割合が最も高い都道府県は福井県で，4.9%となった．ラベル２も含めた割合では，大分県が12.0％と最も高い地域となった．全体の傾向として，西日本の方が方言を含む割合が高いことがわかった．また，各ラベルが付与された発言文の表層的特徴を分析した結果，方言の可能性がある口語的表現を含む（ラベル１と２）と判断された発言文は，それらを含まない（ラベル０）と判断された発言文と比較して，ひらがなの含有率が10ポイント以上高いことが確認された．

以上の結果から，特に西日本の地方自治体の会議録をコンピュータで自動解析する場合，発言文に含まれる方言の考慮が必要であると考えられる．方言や口語的表現は形態素解析誤りを引き起こす原因になりやすいためである．方言や口語的表現を含む発言はひらがな含有率が高い点と，そのような表現自体がひらがなで構成されることが多い点から，ひらがなをヒントに方言の有無の自動推定を行い，解析精度を向上できる可能性がある．

表2.2.1　ラベルとその意味

ラベル	意　　味
0	方言・口語的表現を含まない
1	方言が含まれる
2	口語的表現または見慣れない単語がある
9	誤って収集された文（目次，記号等）

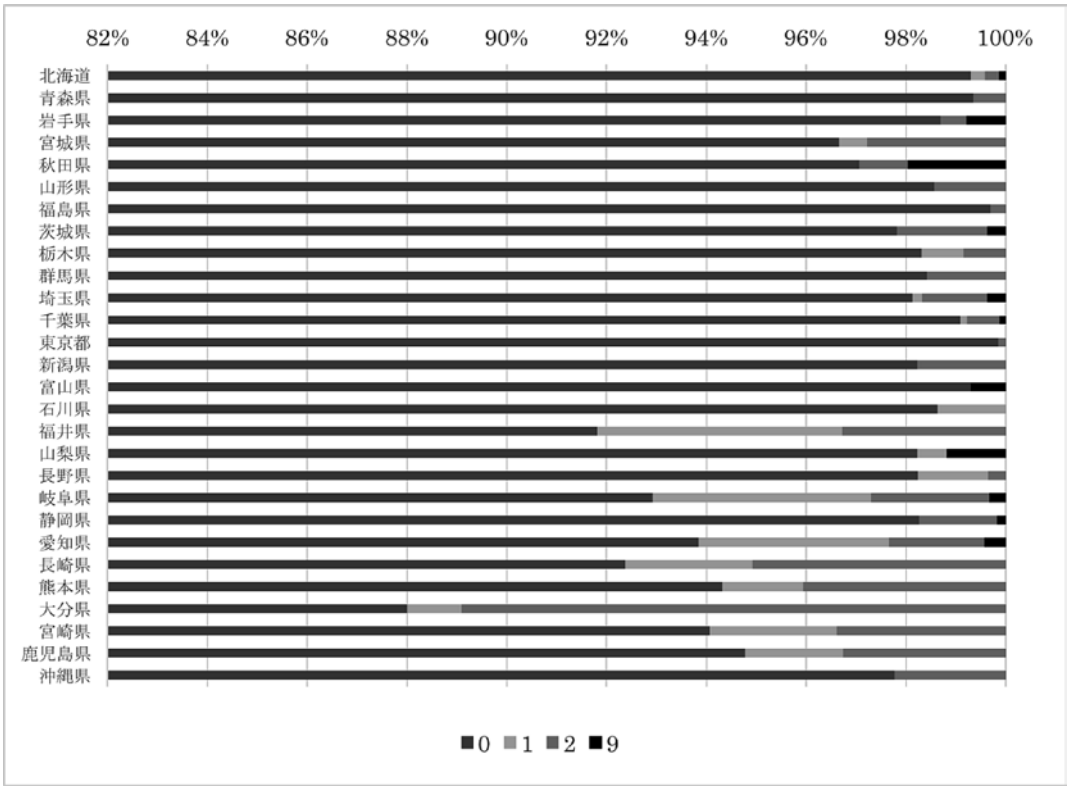

図2.2.1 年縞の8 bit グレー画像とプロファイルの例（Sasaki et al., 2015）

2.2.2 言語的特徴を用いた自動添削モデルの構築

言語的特徴を用いた自動添削モデルの構築の一環として，日本語敬語表現を対象とした自動添削モデルの構築を試みた．敬語に関しては，日本語学習者のみならず，日本語ネイティブであっても多種多様な誤用をすることが報告されている．そのため，関連する先行研究がいくつか存在するが，本研究では地方議会会議録データを利用した機械学習による敬語の自動添削モデルの構築と評価を行った．

対象とする敬語の種類は尊敬語，謙譲語，丁寧語の3つとし，用言のみを対象とした．また，モデルの構築に先立ち，元データとなる福岡市の2001年から2003年までの市議会会議録に対して統計的な分析を行った．図2.2.2は用言を対象とした敬語の種類の分布を表している．図2.2.2から，議会では意見の発言が多いため相手に対する尊敬を表す尊敬語が少なく，逆に謙譲語や丁寧語が多いことが確認された．統計量に基づく機械学習では，学習データの量が性能を左右することが多いため，本自動添削モデルは謙譲語や丁寧語の推定精度が尊敬語のものよりも高い可能性がある．

今回作成する自動添削モデルで使用した素性は以下の箇条書きで示す通りである．

- 用言の終止形
- 文末に位置するかどうか
- 用言の品詞，時制，モダリティ（疑問形，依頼，意見，推測など）
- 表層格（助詞「が」が付属する単語など）

自動添削モデルは線形分類器の一つである最大エントロピー分類器を用いて作成した．また，評価実験の対象データとして福岡市の市議会会議録（約29万の用言を含む）を用い，10分割交差検定によって評価を行った．図2.2.3に実験結果を適合率と再現率の調和平均であるF値

で示す．ベースラインは提案手法との比較のために用意したもので，各用言で最も出現確率の高い敬語タイプを出力するものである．図2.2.3より，単に確率の高いものを選択して出力するベースラインよりも総じて提案手法が優れていることが確認された．ただし実験に用いたデータのサイズは十分とはいえないため，より大規模なデータを用いた評価が必要であると考えられる．

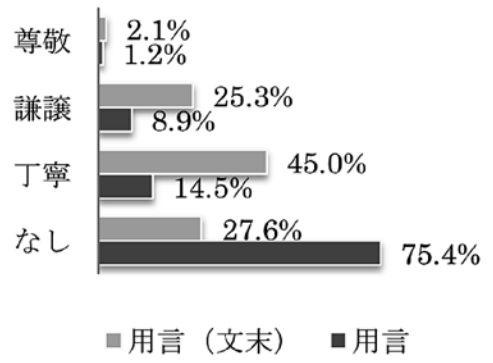

図2.2.2　用言の敬語の分布

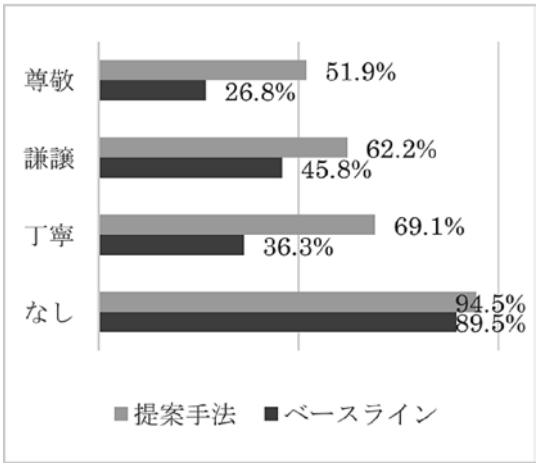

図2.2.3　評価実験の結果（F値）

2.2.3 今後の課題

地質学の論文・報告書の電子データとして十分な量のPDFファイルを準備することができたが，言語情報解析のためのテキスト化処理を十分に行うことができていない．今回の研究成果から得られた言語情報の統計分析と機械学習の知見を用いて，論文・報告書を対象に分析を進めることが今後の課題である．

②研究業績

［論文］

弓　真由子・石原与四郎，2012，重力流堆積物基底の侵食痕の特徴化：特にフルートマーク形成に関する流れの持続時間による影響．堆積学研究，71，173-190.

Ototake, H. and Yoshimura, K., 2012, Development and Evaluation of a Model for Japanese Honorific Expressions Using Assembly Minutes, Proc. of 2012 International Conference on Asian Language Processing, 73-76.

田辺　晋・石原与四郎，2013，東京低地と中川低地における沖積層最上部陸成層の発達様式：“弥生の小海退”への応答．地質学雑誌，119，350-367.

Yumi, M., Ishihara, Y. and Komatsubara, J, 2013, Digitalization of scallop-microtopography using soft-X ray images. Journal of Speleological Society of Japan, 37, 41-54.

Uchida, S., Kurisaki, K., Ishihara, Y., Haraguchi, S., Yamanaka, T., Noto, M., Yoshimura, K., 2013, Anthropogenic impact records of nature for past hundred years extracted from stalagmites in caves found in the Nanatsugama Sandstone Formation, Saikai, Southwestern Japan. Chemical Geology, 347, 59-68.

石原与四郎・宮崎友紀・江藤稚佳子・福岡詩織・木村克己，2013，東京港湾地域のボーリング情報を用いた浅層３次元地質・地盤モデル．地質学雑誌，119，554-566.

木村克己・花島裕樹・石原与四郎・西山昭一，2013，埋没地形面の形成過程を考慮したボーリングデータ補間による沖積相基底面モデルの３次元解析：東京低地北部から中川低地南部の沖積相の例. 地質学雑誌，119，537-553.

石原与四郎，2013，地盤の３次元モデル化－福岡平野を例として－．地盤工学会誌，61，16-19.

吉村和久・内田章太・栗崎弘輔・石原与四郎・原口　聡・山中寿朗・能登征美，2013，七釜鍾乳洞の石筍から明らかになった長崎県西海市中浦地区の数百年間の土地利用変遷．月刊地球，35，628-635.

村上崇史・石原与四郎・藤川将之・無名穴学術調査団，2013，山口県秋吉台無名穴洞口部テラスに認められるリムストーン充填層の堆積相．洞窟学雑誌，38，52-60.

石原与四郎・高清水康博・松本　弾・宮田雄一郎，2014，日南海岸沿いの深海堆積相と重力流堆積物．地質学雑誌（補遺），120，41-62.

佐々木華・石原与四郎・吉村和久，2014，石筍の年縞の

自動認定とその課題―長崎県西海市七釜鍾乳洞龍王洞石筍への適用．洞窟学雑誌，39，53-66.

Yasutomo Kimura, Fumitoshi Ashihara, Arnaud Jordan, Keiichi Takamaru, Yuzu Uchida, Hokuto Ototake, Hideyuki Shibuki, Michal Ptaszynski, Rafal Rzepka, Fumito Masui and Kenji Araki, 2014, Using Time Periods Comparison for Eliminating Chronological Discrepancies between Question and Answer Candidates at QALab NTCIR11 Task, QALab workshop of NTCIR 11, 550-555.

Sasaki, H., Saito-Kato, M., Komatsubara, J. and Ishihara, Y., 2015, Application of a method for detecting varve characteristics in sediments for time series analysis: an example using a soft-X ray image of varve from the Hiruzenbara Formation. Journal of Sedimentological Society of Japan, 74, 31-43.

佐々木華・石原与四郎・佐々木泰典・齋藤めぐみ・成瀬　元，2015，中期更新統蒜山原層の湖成年縞堆積物に狭在する洪水・斜面崩壊堆積物の堆積相と挟在頻度．堆積学研究，74，45-53.

Tanabe, S., Nakanishi, T., Ishihara, Y. and Nakashima, R., 2015, Millennial-scale stratigraphy of a tide-dominated incised valley during the last 14 kyr: Spatial and quantitative reconstruction in the Tokyo Lowland, central Japan. Sedimentology, 10.1111/sed.12204.

Sasaki, H., Sasaki, Y., Saito-Kato, M., Naruse, H., Yumi, M. and Ishihara, Y., 2015, Lacustrine sediment gravity-flow deposits and stratigraphic changes intercalated in varved diatomite: an example from the Hiruzenbara Formation, Okayama Prefecture, southwest Japan. Quaternary International, 10.1016/j.quaint.2015.08.032

高丸圭一・内田ゆず・乙武北斗・木村泰知，2015，地方議会会議録コーパスにおけるオノマトペ－出現傾向と語義の分析－，人工知能学会論文誌，30/1，306-318.

［報告書］

石原与四郎，2013，平成24年度国土政策関係研究支援事業 研究成果報告書「地盤の３次元モデルの構築とその共有に関する研究－地盤・防災情報のユビキタス化－」．86p.

水野清秀・風岡　修・田辺　晋・宮地良典・石原与四郎・安原正也・小松原純子・中島善人・小松原　琢・石原武志・稲村明彦・吉田　剛・香川　淳・森崎正昭・野崎真司・菅野美穂子・古野邦雄・酒井　豊・木村満男・古賀千裕，2013，地形および地質学的手法による液状化調査．地質調査総合センター速報 63「巨大地震による複合的地質災害に関する調査・研究中間報告」．179-231.

高丸圭一・乙武北斗・渋木英潔・木村泰知・森辰則，2013，形態素N-gramを用いた地方議会会議録コーパスの地域変異検出の試み―文末表現を例に―，言語処理学会第19回年次大会発表論文集，737-740.

乙武北斗・渋木英潔・高丸圭一・木村泰知・森辰則，2013，地方議会会議録コーパスの学際的応用を目的としたn-gramデータの構築およびウェブUIの試作，言語処理学会第19回年次大会発表論文集，733-736.

乙武北斗・吉村賢治，2013，英文における固有名詞を対象とした定冠詞の自動付与手法の提案，第29回ファジィシステムシンポジウム講演論文集，756-757.

高丸圭一・内田ゆず・乙武北斗・木村泰知，2014，地方議会会議録におけるオノマトペの出現傾向に関する基礎的検討-少数の自治体に高頻度で出現するオノマトペについて－，言語処理学会第20回年次大会発表論文集，566-569.

小松哲也・乙武北斗・吉村賢治，2014，音声対話におけるフィラーの自動検出について，D-40.

川田真史・乙武北斗・吉村賢治，2014，雑談システムにおける知識源の構造化について，2014年度電子情報通信学会九州支部学生会講演会，D-35.

仲村幸樹・乙武　北斗・吉村　賢治，2014，雑談システムにおける発話生成について，2014年度電子情報通信学会九州支部学生会講演会，A-24.

佐々木　啓・乙武　北斗・吉村　賢治，2014，雑談システムにおける対話のコントロールについて，2014年度電子情報通信学会九州支部学生会講演会，D-36.

折舘直樹・久留間嵩之・乙武北斗・吉村賢治，2014，地方議会会議録における方言の自動推定について，2014年度電子情報通信学会九州支部学生会講演会，D-39.

高丸圭一・内田 ゆず・乙武北斗・木村泰知，2014，地方議会会議録コーパスを用いたオノマトペの分析，第6回コーパス日本語学ワークショップ講演論文集，83-92.

乙武北斗・折舘直樹・吉村賢治，2014，地方議会会議録の方言を含む発言における形態素解析誤りの分析，第30回ファジィシステムシンポジウム講演論文集，660-663.

木村泰知・渋木英潔・内田ゆず・乙武北斗・高丸圭一・森　辰則，2014，地方議会会議録におけるオノマトペの自動抽出手法の提案，第30回ファジィシステムシンポジウム講演論文集，638-641.

［著書］

岡村行信・尾崎正紀・松本　弾・西田尚央・松島紘子・

木村克己・中村洋介・加野直巳・駒澤正夫・大熊茂雄・花島裕樹・水野清秀・康　義英・池原　研・石原与四郎・山口和雄・上嶋正人・中塚　正・金谷　弘, 2013, 海陸シームレス地質情報集「福岡沿岸域」. 数値地質図, S-3（DVD）, 産業技術総合研究所地質調査総合センター.