Ⅲ

実験

文化財関係用語シソーラスの構築と実践活用例：
文化財多言語事業への展開を見据えて

高田 祐一 • 奈良文化財研究所

Supporting the Translation of Cultural Heritage Information with the CDASRJ Thesaurus

Takata Yuichi • Nara National Research Institute for Cultural Properties

語彙／vocabularies　発掘調査報告書／fieldwork reports　シソーラス／thesauri

1　はじめに

文化財を多言語化する際には、一定の専門用語対訳が必要である。対象とする読者や文脈によって、専門用語の使用は検討されるべきであるが、刊行物や展示解説において、用語対訳にバラつきがでることは、利用者に混乱が生じる可能性がある。翻訳者においても、すべての解説等を同一の翻訳者が永久的に翻訳することはない。仮に翻訳者がかわっても一定の基準や用語対訳集に基づいて翻訳することで品質を標準化することができる。用語対訳の基礎情報として、語彙情報を整備していくことが有効である。本稿は、文化財関係用語シソーラスについて、必要性、構築方法、実践例、今後の可能性について報告する。

2　考古学用語データベース

奈良国立文化財研究所埋蔵文化財センター事業として、1976年度に考古学関係用語シソーラスの作成作業を進めている（岩本 1977）。約3万種類の専門用語の出現頻度を整理し、構造分析や用語の体系づけを実施した。

1982年度科学研究費補助金「考古学遺物・遺跡データベースの作成と利用法の確立」（課題番号57123118、代表者：及川昭文）において、考古学用語データベース（考古学専門用語のシソーラス）の構築が計画された。田中琢が担当したようである（田中 1988）。田中は、考古学用語が混乱しているが、用語の使用は研究者の考えの根底とつながっているため、統一できないしすべきでないとした。しかし、研究成果を普及させる面からすれば、用語の関係性を整理し、シソーラスが必要であると考えたようだ。しかし、「専門家でも、特別の才能と相当な学識を備えた人が必要なことが実際に始めてよくわかった。ここで一頓挫。」ということ

で、中断したらしい。

残念なことに 2021 年現在においてこれらの成果は引き継がれていない。そのため、2016 年度から新たに文化財関係用語シソーラスを構築し、現在 19 万の専門用語を収集している。

3　文化財関係用語シソーラスの構築

文化財関係用語シソーラスは、全国遺跡報告総覧（以下、遺跡総覧）の内部システムとして実装している。シソーラス構築の手順としては次の通りである。

まず考古学を中心に 23 の辞書・事典類の見出し語を収集した。『日本考古学用語辞典』『旧石器考古学辞典』『国史大辞典』等である。さらに『発掘調査のてびき』や『石垣整備のてびき』からも収集した。戦争遺跡関連用語などは、既存の辞書に用語が採録されていないことから、報告書を実際に閲覧し用語を採録した。奈文研が保持するデータや研究員からの提供もあり、語彙は約19万となっている。権利関係に問題のない辞書については、日英・日中・日韓の用語対訳も登録した。石切場・石切り場・石丁場・採石場など、ほぼ意味が同義のものは類義語として辞書内部で関係づけている。

印刷物の報告書を電子化する際には、画像データを OCR 処理することによってテキストデータ化する。その際、一部に誤認識が生じる課題がある。例えば、石と右、文と丈である。遺跡総覧は、そのテキストデータをデータベース化するため、誤認識されたテキストを登録することになってしまう。誤認識の用語では、全文検索で検索結果に漏れが生じることになる。そこで誤認識されやすい漢字をとりまとめ、専門用語と突合することによって、表記ゆれ専門用語約 6 万語を生成し、システムに組み込んだ。

現在のデータ量は以下の通り（2021 年 2 月 21 日時点）。

語彙数：190,230
英語用語数：8,501
韓国語用語数：694
中国語簡体字用語数：695
よみ数：64,801
類義語数：5,098
関連語数：13,389
説明数（典拠）：126,947

表記ゆれ数：59,736

4　報告書テキストデータの分析

遺跡総覧には、報告書の本文テキストが21億文字登録されている（2021年2月時点）。このテキストを解析するには、辞書が必要となるが、文化財専門用語は特殊であるため、用語として認識できず十分な解析ができない。しかし、文化財関係用語シソーラスであれば、適切に単語を抽出できる。その実践活用例を述べる。

4.1　全国の頻出語と地域的な特徴語

遺跡総覧に登録されているテキストを対象に自然言語処理にて図化した結果を示す。図１の報告書ワードマップ（頻出用語俯瞰図）は、遺跡総覧に登録されている報告書に対し、考古学関係用語の出現回数を集計し、図化した。用語については桃色：遺物に関する用語、黄色：遺構に関する用語、水色：その他　で分類した。図では遺物に関する用語の割合が多いことを確認できる。その中でもナデ・口縁部・底部など土器に関する用語が多いのが特徴である。発掘調査では、土器が出土し、遺跡の時代特定や評価のために土器を観察し、その際には口縁などを重点的に観察し、その成果を報告書に記載することが多いことから、経験としても違和感のない結果といえる。最新の図は全国遺跡報告総覧「報告書ワードマップ」で閲覧できる。
（https://sitereports.nabunken.go.jp/ja/visualization/term）

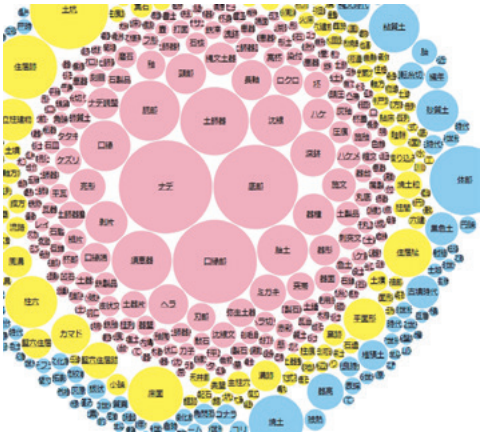

図1　報告書ワードマップ（頻出用語俯瞰図）

4.2　地域的な特徴語

全国遺跡報告総覧に登録されている報告書に対し、都道府県ごとに考古学関係用語の特徴語を図化した（図2）。当該都道府県内にて頻出する用語（よく使われる用語は重要）かつ他都道府県では出現頻度が低い用語（希少用語は重要）であることを勘案するため、当該都道府県の強い特徴を示す用語を可視化できる。自然言語処理技術のベクトル空間モデルの TF（索引語頻度）と IDF（逆文書頻度）を組み合わせた TF-IDF にて算出した。データ数が少ない場合、結果が偏る可能性があるため、PDF が登録されている報告書類（遺跡総覧の報告書種別 a01）について、都道府県下で 300 件を超えた場合にワードマップを作成し、毎週更新対象となる。その地域で特徴的な用語を提示することで、キーワードの入力に頼らない検索方法を提供している。専門用語を知らなくとも検索する事が可能になる。最新の図は「報告書ワードマップ」（https://sitereports.nabunken.go.jp/ja/visualization/term）にて都道府県を選択すると閲覧できる。

4.3　用語の変遷

用語の使用については、時代的傾向がある。例えば、縄文土器、縄文式土器、縄紋土器の使用頻度に特徴がある。各年ごとのそれぞれの語が何冊の報告書で使われたかを示したのが図 3 である。当然、報告書が多数発行される年ほど、頻度が多くなると想定されるため、報告書発行数も加えた。

70 年代前半は、縄文土器と縄文式土器の頻度が拮抗するものの 70 年台後半からは縄文土器の頻度が圧倒する（図4）。この傾向は現在まで続く。縄文式土器については、90 年代前半までは一定の使用頻度があったものの 90 年代半ばから、急

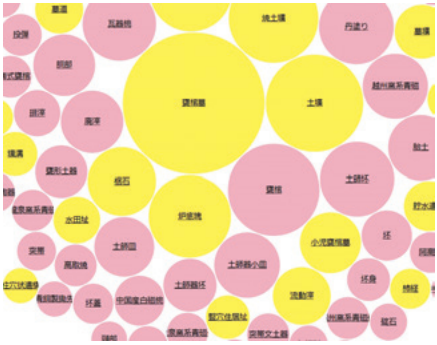

図2　報告書特徴語ワードマップ-福岡県

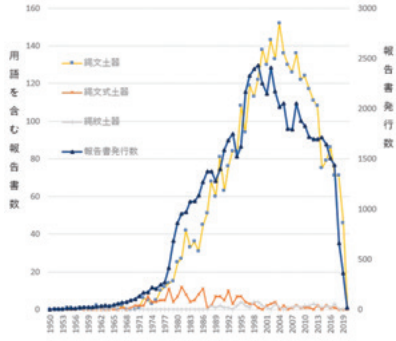

図3　縄文土器関係用語の使用推移

減している（図 5）。縄紋土器については、少数ながら一定の使用がみられる（図 6）。研究史を整理する際に参考となるだろう。

5　文化財関係用語シソーラスの活用事例

5.1　報告書ごとの頻出用語と類似報告書

調査研究のために類似例を探す場合、関連する全ての報告書を探し出し通読することは困難である。遺跡総覧では、本文内容が類似している報告書を自動提示する機能がある（図 7・8）。当該報告書のテキストに対し、シソーラスの用語をもって本文頻出用語を抽出することで、用語のグループを作成する。同様の作業を全ての報告書にも適用し、それらの個々の用語群の構成に類似している報告書を自動提示している。

5.2　文化財イベントの頻出用語と類似イベント・類似報告書

遺跡総覧では、各機関が文化財関係イベント情報を登録可能である。イベント情報には、企画趣旨が記載され、その文中には文化財関係用語が含まれる。自動でその趣旨文から用語を抽出し、イベントごとに用語群を構成する。この用語群

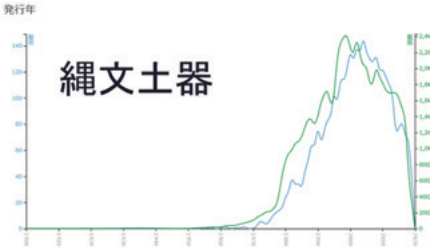

図4　報告書における「縄文土器」の使用推移

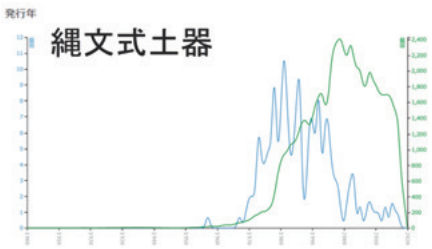

図5　報告書における「縄文式土器」の使用推移

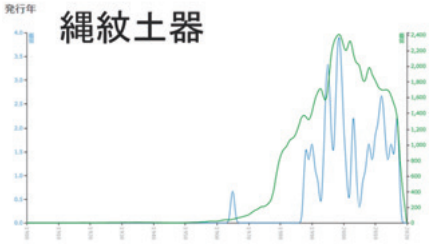

図6　報告書における「縄紋土器」の使用推移

を活用することで、イベント同士・イベントと報告書などの組み合わせで類似イベントあるいはイベントに類似する報告書を自動提示している。イベント参加前の予習、参加後の復習などに活用できる。遺跡総覧の利用ユーザはコアな文化財ファンが多いと想定されることから、特定の関心をもった利用者に効果の高い周知が可能である。

5.3　クロスリンガルおよび類義語検索機能

日本考古学の成果に関心を示す海外の研究者には、言語の壁や報告書を手にとって閲覧できないという、情報アクセスの問題がある。日本語を習熟しても、日本の専門用語には多くの類語がある。遺跡総覧内部のシソーラスを活用することで、英語の用語を投入するだけで日本語用語に自動変換し、類義語を自動で付与する（図9）。簡単に言語の壁を超えて網羅的な検索が可能である。

日本語キーワード検索で、シソーラスに表記ゆれ・類義語の登録がある場合は、検索ボックスにチェックを入れるだけでOR検索が可能である（図10）。

図7　報告書本文頻出用語

図8　類似報告書の自動提示

図9　英語用語の日本語化

図10　キーワード検索時の表記ゆれ・類義語の付与

6　今後の整備方法

6.1　未知語の収集

現在の文化財関係用語シソーラスは、既存の辞書類がベースである。新たな用語の登場に対し、辞書の採録はタイムラグが発生する。近年発行の報告書に含まれる用語について、抜け落ちている可能性がある。そのために、まだ辞書に採録されていない専門用語を収集する必要がある。報告書テキストについて、ひらがな区切りによる分かち書きを実施したうえで、漢字３字以上７文字以下に絞り込んだ未知語候補 9,938,663 語を抽出済みである。今後この中から未知語を採録する予定である。

6.2　word2vecによる類似用語抽出

word2vec とは単語をベクトル化する技術で、関連用語や類似用語の計算が可能となる。例えば四隅突出型墳丘墓の処理結果は以下となる。数字はスコア。

墳丘墓：0.7719687223434448
西谷墳墓群：0.6747199892997742
前期古墳：0.6649181246757507
弥生墳丘墓：0.6608040928840637
方形周溝墓群：0.6515164971351624

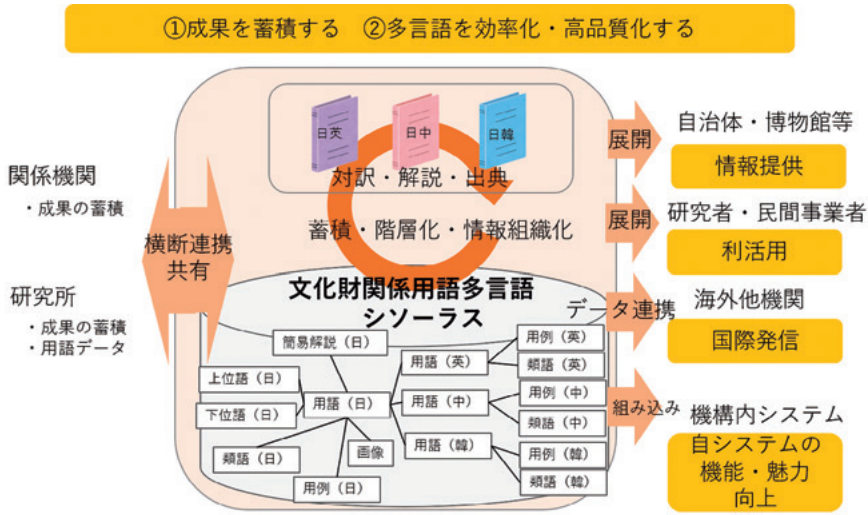

図11　文化財関係用語シソーラスの利活用図

集団墓：0.6483949422836304
四隅突出：0.642971396446228
前方後方墳：0.6399630308151245
弥生墳墓：0.6326301097869873
貼石墓：0.6310558319091797
方形台状墓：0.6215837597846985

これは、関連語等を設定する際に参考情報にできる。膨大な用語を自動で属性分析することによって、シソーラス構築を加速させることができる。

7　文化財多言語解説事業の成果蓄積

2017年、文化財の多言語解説等による国際発信力強化の方策に関する有識者会議は「文化財に関する国際発信力強化の方策について（提言）」を示した。文化財の国際発信力強化等に必要な事項を取りまとめ、特に多言語解説整備の加速のために必要な事項として次の点をあげている。

> 専門用語や共通用語の多言語化データベース
> 多言語化に当たっては，文化財解説で頻出する用語や他の言語では表現しにくい日本文化の専門用語などを，専門家がどのように解説，表現しているかが参考となろう。こうした情報に，多くの人々がアクセス可能となることが望ましい。例えば，日本特有の歴史文化に関する専門用語などの共通の多言語データベースをウェブサイト等で公開するシステムを国として整備していくことで可能となると考えられる。これにより，前提知識や文化的な背景なども踏まえた魅力が伝わる説明の在り方や，多用される表現，専門家が使う対訳の傾向などを共有する仕組みが構築される。その際，例えば誰もが自由に活用できるインターネットサイト上に，専門家やプロの人材等が対訳を登録，更新できる仕組みも想定されうる。

国立文化財機構では、文化財の多言語発信をするために、専任の研究職を配置している。日々成果が蓄積しており、これらの成果から語彙情報を集約することで、提言で必要とされたシソーラスを整備できる。自治体や民間翻訳事業者、研究者にとって有用な語彙情報となると想定される。

おわりに

文化財の多言語化は、丁寧に推進する必要がある。例えば、各国語で同じ表記をする場合でも、意味が異なるケースやニュアンスが微妙に異なるケースがある。また、ある用語やテーマを多言語化する際には、そのテーマが内包していた曖昧さなどが、表面化する場合がある。都度、専門家と翻訳者でコミュニケーションを取りながら、解決を図るほかない。そして、その知見を蓄積していくことが重要である。効率的かつ漏れなく蓄積し、関係者で共有していくことで、議論を積み重ねていくことができる。

本稿は、科学研究費 16H05881「日本考古学国際化のための考古学関係用語シソーラス構築と自動英語化の研究」代表：高田祐一の成果の一部である。

参考文献

岩本圭輔（1977）「埋蔵文化財関係用語の収集と整理」『奈良文化財研究所年報』奈良文化財研究所
高田祐一（2019）「発掘調査報告書の電子公開による情報発信とその新たな可能性」『デジタル技術による文化財情報の記録と利活用』奈良文化財研究所
田中琢（1982）「考古学、みかけだけのはなやかさ」『同朋』同朋舎出版
田中琢（1988）「ある考古学研究者のパーソナルなコンピュータ史」『人文科学データベース研究』人文科学データベース研究刊行会