

浮世絵表題の多言語検索をめざしたシソーラス構築法に関する研究

代表研究者 山 田 奨 治 国際日本文化研究センター研究部助教授
共同研究者 早 川 聞 多 国際日本文化研究センター研究部助教授

1. 研究調査の目的

最近の高度情報化社会の進展に伴って、美術工芸品等の文化財をデジタルコンテンツとして電子化し、ネットワークを通じて閲覧するといった、あたらしい利用法が可能となってきた。また、インターネットとWWWに代表される、広域性や利便性をもったメディアの登場により、国境・言語を越えた情報の交流の機会が格段に増加している。しかし、このような状況では、情報アクセス時における言語の壁が大きな問題である。問題は言語の壁のみならず、コンテンツ表題表現における文化的・専門的知識の欠落も、検索精度向上のためには障害となる。これらの問題に対処するために、ネットワーク上に存在する日本美術（とくに、浮世絵情報）を言語や文化的・専門的知識に依存せずに検索閲覧可能とする、「多言語情報ゲートウェイ」の構想を示す。当面想定するアクセス言語は英語と日本語で、このゲートウェイの主な機能的役割は、つぎの通りである。

- 1) コンテンツに付随するキャプション等の日本語テキストを英語に自動翻訳しデータベース化する。（自動翻訳には、用例に基づく翻訳手法が有力である。）
- 2) 利用者の検索要求を解釈して、上記のデータベース中から適切なものを検索して一覧を示し、利用者の選択により、画像付きのコンテンツに導く。

将来においてこのような機能を実現するためには、つぎのような観点から、対象となるコンテンツに関する語と語の類義関係を記述したシソーラスが必要である。

- ・用例にもとづく翻訳手法では、用例データベースのなかから翻訳対象表現と最も類似した対訳用例を見いだす必要があるが、そのさいに語の類義関係を示す情報が重要な手がかりとなる。
- ・利用者が指定した検索語のみで検索すると、所望するコンテンツのうちの一部しか得られない等の問題があり、検索目的を満足するためには、文化的・専門的知識にもとづいて指定された類義語または関連する語も含めて検索をおこなう処理が必要である。

このような要求を満たすシソーラスの構築方法について研究するのが、本研究の目的である。

本研究で目的とするような、検索および翻訳に役立つシソーラス構築は、「浮世絵」という比較的範囲の明確な適用対象で、実証的な研究をおこなうことにより、実用的および学術的な見地から、じゅうぶんな知見を獲得しうる研究テーマと位置づけることができる。

前述した情報流通の広域化・多様化を背景として、国際日本文化研究センターでは日本固有の優れた美術・文化財コンテンツを海外に発信するための方式を検討するため、「日本美術データベース」を構築し、内外の日本研究者に限って公開している。このデータベースに蓄積された浮世絵表題情報をベースに、今後は自動翻訳技術を活用できる方向性を探りたい。ほとんどが複合名詞である表題を自動翻訳するためには、用例にもとづく翻訳手法が適しているが、この手法では翻訳対象表現と最も類似した対訳用例が抽出できるかがおおきな課題である。この類似用例を見いだす処理では、既存の汎用的な類語辞典を用いるのが一般的であるが、翻訳精度の点で問題を抱えているのが現状である。また、検索においても、入力キーワードだけを用いて検索したのでは、検索結果にもれが生じるため、シソーラス等を用いて類義語に展開する必要がある。

日本美術データベースでの実践経験にもとづいて、検索や自動翻訳のために不可欠な知識源として、対象固有分野のシソーラスの構築を目指すに至ったものである。

情報検索の分野において、目的とする検索結果を導くために、シソーラスの参照が有効であることが指摘されているが、検索に特化したシソーラスはあまり例がない。たとえば国立民族学博物館で開発されたマルチメディア検索システムにおいては、収蔵品名表現のシソーラスが作成されているが、汎用性のあるものではない。

一般に入手可能な日本語シソーラスとして、(1) 類義語新辞典（角川書店）、(2) 分類語彙表（国立国語研究所）、(3) 日本電子化辞書（EDR）の概念辞書、(4) NTTの日本語語彙体系、がある。(1) および (2) は、人間が文書を書く際の手助けとなることを目的として作られていると考えられ、(3) および (4) は言語を計算機で扱った

めに作成されている。ただし、(4)は電子化された媒体として入手できない。

このように、本研究のような人文科学的な領域でのシソーラス作成の先行例は、ほとんど見られない。

2. 方法

本研究は、シソーラスの構築手法の研究である。シソーラスの作成では語の上位下位関係や類義関係などの決定をおこなう必要がある。この作業は、人間の考えによるところがおおきく、機械による自動化が困難な作業である。そのため、効率的にシソーラス作成をおこなうためには、まず基礎的検討をじゅうぶんにおこない、その後実際に応用した結果をフードバックさせてシソーラスの改良をおこなうといった手順をとる必要がある。われわれは、平成10年度からの2年間の研究期間でこれらの作業に取り組んできた。

シソーラス作成においては、用例から得られる語の類義関係を示す情報が重要な手がかりとなる。このような語の類義関係に対しては視点の問題を考慮する必要がある。すなわち、ひとつの語も別の視点から見れば別の意味の側面に焦点が当たることになる。例えば、「女形（おやま）」「立役（たちやく）」「おたふく」「ひょっとこ」という語は、「性別」という視点から見れば、「女形（おやま）」「おたふく」、「立役（たちやく）」「ひょっとこ」とそれぞれ分類され、「劇中における役割」という視点から見れば、「女形（おやま）」「立役（たちやく）」、「おたふく」「ひょっとこ」と分類されることが考えられる。このように、シソーラスは実際には単純な木構造とはならず、あるノードに複数の親ノードが存在するのが一般的であるが、汎用的になればなるほど、視点の問題まで考慮した高精度なシソーラスの構築は現実的には不可能となる。そこで、本研究では、「浮世絵」という比較的範囲の明確な適用対象を設定し、視点も考慮したシソーラスの構築手法に関して、実証的な研究をおこなっている。

本研究では、まず、浮世絵表題に関する対訳用例データを作成した。そして、シソーラスの作成に向けた浮世絵表題用語の抽出をおこない、用語分類をおこなった。

平成12年度は、浮世絵表題を対象にした用語辞書の拡充をおこなった。具体的には、国際日本文化研究センター作成による日本美術データベース（非公開分を含めて約10000件）を分析して、浮世絵表題で使用される用語を手作業で切り出し、それが地名であるのか人名であるのか、たとえば人名であるならば役者名なのか、劇中登場人物名なのか、遊女名なのかといったことを画題分析から決定し、用語の分類を決定するといった方法で、用語辞書の作成をおこなった。

3. 昨年度までの成果

国際日本文化研究センターでは、従前より日本美術作品の画像データベースを整備し、その一部を公開している(<http://www.nichibun.ac.jp/dbase/japan.htm>)。公開データベースには文字情報の英訳も掲載してあるが、当該英訳は日本研究者向けの英訳となっているため、翻訳困難な用語はあえて翻訳せず、日本語音のローマ字表記にとどめているケースが多い。たとえば「見立て」は”Mitate”と表記するなどである。多言語検索のためのシソーラス作成を視野に入れた場合、このままでは対訳用例としての使用に耐えない。そこで、シソーラス作成をにらんで、英訳をやり直す作業をおこなった。たとえば「見立て」は、”Representation of ~”といった風に、翻訳困難な日本語概念を積極的に英語に置き換えるという方針を貫くことにした。英訳対象とした浮世絵表題は、画像公開分の約5000件に非公開分に加えて約10000件である。

浮世絵表題は、「扇屋内花扇」「鬼もん喜兵衛河原崎権十郎」のように、名詞が連続する複合名詞の形態をとるものが多い。それらをただ単に字句通りに翻訳したのでは、原意を表現できないことがほとんどである。「扇屋内花扇」は、「扇屋」という揚屋の遊女「花扇」の意味である。その場合は、”The Courtesan Hanaoogi of Oogiya”と訳さなければならない。同様に、「鬼もん喜兵衛河原崎権十郎」は、役者の河原崎権十郎が「鬼もん喜兵衛」を演じている浮世絵なので、”Kawaharazaki Gonjuuroo in the Role of Kimon Kihei”と訳さなければ、原意を伝えることができない。同じように、「大星由良之介」は劇中登場人物なので、”The Actor in the Role of Ooboshi Yuranosuke”と訳すのが正しい。

これらのように、浮世絵表題の英訳は、すべてもとの浮世絵作品の画像を参照し、表題が伝えんとするところの意味を、できる限り正確に伝えるように配慮しながら翻訳作業を進めた。作品を参照しながらの作業であったので困難を極めたが、平成11年度に約10000件の浮世絵作品の表題について、英訳を付与することができた。今後、翻訳の正確さの検証を進めなければならないが、シソーラス作成のための対訳用例データとしては、ある程度の規模を確保できたのではないかと考える。

4. 今年度の成果

平成12年度は、浮世絵表題のシソーラス作成に向けた作業として、浮世絵表題に登場する単語の抽出と分類をおこなった。作業手順として、表題をなるべくおおきな概念で単語抽出をし、比較的に多くみられる種類の単語を集めて当座のラフな分類カテゴリーを作成して分類を進めた。分類カテゴリーは、つぎの37種類になっている。

場所 固有の地名 景観 店名 遊女名 遊女関係 屋号 役者名 歌舞伎その他
歌舞伎役名 力士 人名 その他の人物 風俗 動物名 植物名 品物 物語名・場面等
年中行事（歳時記） 年号（時間） 見立 寺社関係 歌詠み人名 職関係 屏風
遊楽図 美人図 講釈（講） 七福神 外国関係 合戦・戦争物 枕絵 競 楽器
婦人 乗り物 その他

分類カテゴリーごとの抽出単語数を、以下の表1に示す。抽出した単語数は、21509単語である。単語リストの一部を表2に添付した。

作業上の問題点として、以下のようなことがあった。第一に、作品の題目と内容が一致しないケースがあった。たとえば、題目に「A祭礼」とあっても、絵の内容は風景（祭礼ではない）としかみえないものがある。その場合Aという単語は祭礼名にするべきか、風景に分類するべきか簡単には判断できず、さしあたり両方のカテゴリーに重複して分類した。第二に、浮世絵に女性が描かれている場合、それが市井の女性なのか遊女なのか決めがたいという問題がある。これに関しても、作品に踏み込んだ調査をしながらひとつひとつ解決していった。第三に、種々の雑多な内容が「その他」のカテゴリーに残っている。将来にそれらの単語を分析して、サブカテゴリーに分類する必要がある。

表1 分類カテゴリーと抽出単語数

場所	689
固有の地名	1307
景観	1000
店名	325
遊女名	522
遊女関係	1038
屋号	30
役者名	343
歌舞伎その他	367
歌舞伎役名	2200
力士	97
人名	899
その他の人物	783
風俗	1779
動物名	627
植物名	480
品物	709
物語名・場面等	1635
年中行事（歳時記）	301
年号（時間）	297
見立	433
寺社関係	836
歌詠み人名	138
職関係	473
屏風	54
遊楽図	143
美人図	1623
講釈（講）	9
七福神	104
外国関係	376
合戦・戦争物	465
枕絵	29
競	30
楽器	52
婦人	92
乗り物	71
その他	1153

表2 浮世絵表題から抽出した単語（一部）

場所	固有の地名	景観	店名	遊女名	遊女関係
あつま	HIYAOGOKU	あさ日の出	あかねや	あげまき	■形様源氏絵合
あづま	あかさか	あふみ八けい	あざみ屋	あげ巻	■日の晴嵐
いせのうみ	あたけ	あらし山満花	あふきは	あつまや	あててみな
えびす屋店頭	あふひ坂	おちやのみづ	あふきや	あふむ	あふむ
オランダ屋敷	あふみ	かうつけ佐野ふ なはしの古づ	あふぎや	あやの	あやめ模様大振 袖の遊女
オランダ商館	アムリッサー	きはん	あふき屋	いく	いたそう 寛政年 間女郎の風俗
かいでや店先	あらし山	ぎやうとくしほ はまよりのほと のひかたをの ぞむ	あふぎ屋	いつとみ	いつとみ
かぎや店先	あら井	くさりじま秋月	あぶら屋	いつの	うかれ踊り
かげ	あら湯	くものかけはし	あふ起や	いつ富	うけれ妻
かこい	いしべ	こま形河岸	いせ屋	いと	うわきそう 高名 輪
かつしか渡場	いせ	さがみ川	いつミヤ	いな	おいらん
かべ	いちの谷	さみだれ	いづみ屋	いほざさ	おとずれ図
かりば	うしまち	しのぼづ	いづ文	うきふね	おのゝ道風
こま形河岸	ウダイプール	しばうら乃風景	いとや	おいと	かげかつ
さん海	うゑ野	しほ原あら湯の 秋	えびす屋	おうしう	かしまおどり
しほはま	おかざき	すほうの国きん たいはし	おびや	おかる	かむろ
ステンション	おかべ	すみだ川	おふぎや	おきた	かり宅
ステン所	おきつ	せいらん	おふき屋	おきた	かり宅狂哥附
せつちん	おちやのみづ	せき口	かいでや	おこん	かり宅八景遊君 之図
たいこばし	おつとも	せき口上水端 はせを庵椿や ま	かがや	おさん	きつねつり
ためいけ	お玉が池	はな盛	かぎや	おせよ	きぬぎぬの帰帆
てつぽふづ	お茶の水	びくにばし雪中	がくたは ら	おせん	きり火できよめる かみしんじんは芸 者のくせ
なにはや店先	かうつけ	ふじのゆき	かさ屋	おそめ	くつわや
にい宿のわた し	かうや	ぼせつ	かじ丁梅 松	おたつ	くるわ
のれん前	かけ川	やいかすみ	かな川や	おちよ	くるわのいきぢ
はなやしき	かじ丁	やぐら下の晩鐘	ぎおん井 筒	おちゑ	けいせい
はねだのわた し	かすみがせき	やみ	きく酒屋	おつま	ささやく遊女と禿
ひかた	かすみが関	よるの雨	きの国屋	おつゆ	すさき
ふところ	かつしか	よるの景	きやうや	おてふ	すだれのかげ
ふなはし	かなや	らくがん	くらはしや	おとり	そふまがき
ふね宿	かな川	ろうかの暮雪、 座敷の夕せう	こじまや	おなよ	そふまがきのよい の景

ほとり	かめみど天神 たいこばし	わたし場	こん屋	おの吉	たばこを吸う遊女
まつもと庭先	かめ山	阿波 鳴門の 風波	さのや	おはね	つく田嶋の晴
やかた	かんばら	阿波鳴門之風 景	すがや	おはん	つぼね
やくら	かん原	阿蘭陀海浜出 船之図	すみ安	おひさ	つぼね見世
やぐら	きが	愛宕山遠望	たちばな や	おふく	つり狐の落雁
やくら下	ぎやうとく	愛宕秋月	たち花や	おまん	つんとして
やぐら下	きりふりの滝	旭	たねや	お駒	てつぼう
やたい	きんたいはし	旭日	たまや	お仙	てつぼう
ろうか	きんたいばし	芦之湖之雨景	たま屋	お千代	てならひ子
わたし	くさつ	綾瀬の夕照	たるや	お染	てふどおあめの 盃で糸のきれる酒 めくりがあるとは 芸者の口くせ
わたし場	くさりじま	綾瀬川 浅草寺 遠景	つたや内	お長	ながれの水性、根 引の金性
芦の屋	くだんうしがふ ち	粟津の晴嵐	つた屋	お藤	にあいさう
芦辺	くわな	粟津晴嵐	つち屋	お半	ねむそう
伊豆の山中	こま形	安部河の富士	つの井	お筆	ねむそう 明治年 間娼妓乃風俗
威海衛附近 衣裳部屋	ごゆ さかの下	伊勢二見が浦 伊豆の山中	つるや つる屋	お縫 お房	ねり物姿 びくに、ふしミ、と きハ
井戸	さがみ川	夷港	てうしや	がくの小 さん	ふじのゆき
井戸水 井戸端	さくら川 さつた	衣紋坂渡雁 衣紋坂夜雨	てうじや てうじ屋	かしく か す み 野	ふじの雪 ふところに猫をい だく遊女
一帯	さつま	井の頭の池 弁 財天の社 雪の 景	てう子屋	か す 美 野	ふね宿の雪の景
宇治橋上 雨中	さる若 さる若町	一つ目千金 一月中院午前 九時写 薩太之 富士	てんまや なにはや	かたらい かたらゐ	ほととぎすと遊女 よしはら
厩 雲外	しのばす しのばす	隠田の水車 宇津之山 富士 眺望	なにハや なにわや	かつらき かつらぎ	よしはらの躰 よし原日本堤
雲間	しばうら	羽田大師の渡	なにわ屋	かつらぎ 太夫	よっぱらい
雲上	しほがま	羽田 中町にて (菓造之家)	ニューカ ルトン	かつ山	逢妓
永代ばし 英役館 駅	しほ原 しほ原あら湯 しまだ	雨 雨の日 雨景	びぜん屋 ひらの屋 ひら岩	かほる かほ世 かめ	逢妓八景 安永頃 伊勢おんどあこぎ

					網の手踊
駅路	しらすか	雨後	ひら野屋	かめの	伊勢音頭おどり
越後屋	しんばし	雨後のハヶ岳	ふしや	かよひぢ	伊勢音頭図
越前三国	しんよしはら	雨後のハヶ岳 (駒ヶ岳の石室 より)	ふじや	からうた	亥ノ刻 藝者
円窓	すさき	雨後の夕	ふじ屋	からさき	磯千鳥模様の遊 女
園	すざき	雨後観虹	まつがね や	から歌	引手茶屋

4. 今後の予定

本研究で作成した分類カテゴリーは、作業者の便宜的な理由によって定めたものなので、シソーラスのカテゴリーとしてそのまま使用することはできない。今後、概念レベルの再検討を進め、有効な分類方法を開発しなければならない。またその作業を進めるうえで、抽出した単語からもとの浮世絵をたどることができるツールの開発なども必要となろう。

今後、これらの単語をシソーラスとして構築していくうえにおいて、既存の日本語シソーラスを参考にしながら、できればある程度それと整合性を保ちながら浮世絵表題シソーラスを作成できることが理想的である。既存の日本語シソーラスとしては、国立国語研究所の分類語意表のようなものがモデルとして考えられる。今後は分類語意表を参考にしながら、浮世絵表題シソーラスの構成を構想していきたい。