外来語の由来元を利用した作家の西洋かぶれ度判定

石川　京太郎† 　執行　健人†††‡ 清光　英成‡

†神戸大学　国際文化学部　〒 658-8501 神戸市灘区鶴甲 1-2-1

†††The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong

‡神戸大学大学院　国際文化学研究科　〒 658-8501 神戸市灘区鶴甲 1-2-1

E-mail: swanky.street@outlook.jp, 185c124c@stu.kobe-u.ac.jp, kiyomitu@kobe-u.ac.jp

あらまし　明治, 大正期の文学作品で用いられる外来語には現代小説ではみられない特徴的な用法が多く, 作家の西洋文化への傾倒具合を分析する上で重要な要素となりうる. 外来語の語彙や使用頻度の情報をもとに作家がどの程度西洋文化に傾倒しているかを定量化した指標「かぶれ度」と, それを用いた可視化について議論する.

キーワード　外来語　小説　自然言語処理 オープンデータ

1. はじめに

西洋文化からの影響に染まった言動をとる日本人を「西洋かぶれ」と表現することがある. 母国ではない文化からの影響を受け, 個人の使用語彙や所作, 思考法などが変化する現象は非常に興味深く, 個人がどのように異文化を受容しているのかを分析する上で重要な要素となりうる.

森鴎外や夏目漱石に代表される明治, 大正期の日本の作家の作風には西洋文化の影響が色濃くみられる. 中でも外来語に関しては,「プロレタリア」や「ドッペルゲンゲル」など現代小説では用いられない様な単語の使用例も多く, 作家がどの文化圏からどういった影響を受けているかという特徴を分析する際に重要な要素となりうる. 文学的なアプローチから作家への西洋文化からの影響を研究する事例は存在したが, 作家の留学経験や作品の舞台となっている場所の地理的な情報など, メタ情報をもとにある程度大胆な仮説を立てて検証していくというスタイルが少なくない. しかしながら「西洋にかぶれている状態」は語彙や所作に現れうると考えられ, 定量化された作者の語彙や単語の使用頻度のデータが分析に利用できれば, 西洋文化への傾倒度をより客観的な数値で明らかにすることが可能になると期待できる.

そこで, 作家が作品中で使用している外来語に着目し, 工学的アプローチから自然言語処理で用いられる技術を利用することで, 作家がどの程度西洋文化に傾倒しているかを「かぶれ度」という指標で数値化, 可視化する手法を提案する. これにより, これまで主に文学的アプローチから提示されてきた作家に対する西洋文化の影響についての仮説の検証, 及び新たな仮説の構築において, 新しいアプローチの可能性を示すことができた.

2. 用語の定義

本稿では以下の用語について, 下記のように定義し利用する.

[かぶれ度]

かぶれ度は作家が対象の国にどの程度文化的に傾倒していたかを示す指標である. 本研究では, 当時の日本の文化に対して特に大きな影響を及ぼしていたと考えられるイギリス, フランス, ドイツ, イタリア, スペインの 5 カ国を算出の対象とした. この 5 カ国は現代の各国の国境を想定したものではなく, あくまで明治時代初期のヨーロッパ各国の勢力図を考慮し, 現在のオランダやオーストリア由来と推定される単語はドイツ由来に, ポルトガル由来とされる単語はスペイン由来にといった変換を行なっている. 作品から外来語を抽出し, 各語の語源と推測される国, 単語の出現回数をもとに, 作家の各国への文化的傾倒度を定量化する.

[語源タグ]

後述する各単語の語源となる国を推測するプロセスで用いる, 語源国の判別材料となる単語, 語源となる国, 集計の際の重みの 3 つの要素の組みである. 試行錯誤の結果「英語」や「フランス語」等, 特に語源国の推測にあたって重要と判断できる単語については集計の際に重み付けを実施している. 対象テキストにおける各語源タグの出現回数を, 語源国ごとにグループ化して集計することで語源国を推測する.

3. 関連研究

3.1 文学的アプローチからの研究

これまでにも文学的なアプローチから, 夏目漱石や森鴎外など個別の作家に焦点を当てた海外文化からの影響や異文化受容についての研究は存在した. 大前[1]

は夏目漱石の作品の多くが他の作品の構成や文体を参考に執筆されている点に着目し，とくに先行研究が少ない『坊っちゃん』という作品について，イギリスの小説家ディケンズの『ニコラス・ニクルビー』と比較しながら，主に作品の構成や使用語彙の観点からディケンズの影響を考察している．Scott[2]は，ドイツの中編小説に見られる当時のロマン主義(Romantics)，リアリズム小説(Literary realism)，象徴主義(Symbolism)等のフレームワークが森鴎外の作風に与えた影響について論じている．

3.2 工学的アプローチからの研究

対照的に，工学的なアプローチからの文学研究の例としては金子[3]らの単語ベクトルの類似度に着目した英米文学の作品解釈の試みが挙げられる．金子らはハーマン・メルヴィルの『タイピー』を対象として，Word2ec を利用して求めた作家の特定作品の単語ベクトルと一般人の単語ベクトルの類似度の差から，特定の単語が持つ意味合い，含意を解釈する手法を提案している．花畑[4]らは，文学作品の作家の語彙の特徴を，単語の出現回数をもとにベクトル化する Bag of Words の手法を用いて定量化し，同時に文脈の特徴を単語の分散表現をもとにベクトル化する fastText の手法を用いて定量化，それぞれの出力をアンサンブルすることで作家の推定に利用している．しかしながら，著者が確認した範囲ではこれまでに作家間で統一した指標を用いて西洋文化への傾倒度を定量化する研究は存在しない．

4. 提案手法

図 1 に提案するシステムの概要図を示す．入力テキストにはインターネット上で公開されている大規模コーパスである青空文庫を用いる．青空文庫の全作家のうち，作品数が 10 以上存在する作家を対象とする．また，対象とする作家の各作品について，文字遣い種別が新字新仮名版，ファイルサイズが 5 キロバイト以上という条件を満たす作品のみ処理対象としている．

図 1　システムの概要図

4.1 外来語の抽出

図 2 に外来語抽出のイメージを示す．対象作品のテキストファイルから正規表現を用いて外来語を抽出する．外来語の多くがカタカナで表現されている点に着目し，任意のカタカナ 2 文字以上が連続する単語を外来語の候補とみなし，抽出の対象とした．抽出後の各語に対して，別途作品本文を形態素解析した結果から品詞情報を取得し，品詞が副詞であるものを処理対象から除外した．これは外来語ではない擬音語や擬態語を取り除くためである．形態素解析器として Mecab を，形態素解析用の辞書については新語や固有表現に強いとされる mecab-ipadic-NEologd を利用した．

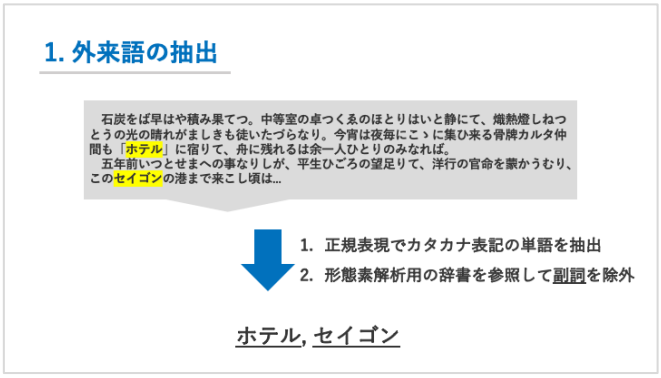

図 2　外来語抽出のイメージ

4.2 抽出した各単語の語源国推定

抽出に成功した各外来語について，オンライン上の情報資源をもとに各単語の語源となった国を推測する．図 3 に語源国推測のイメージを示す．語源国推測のもととなる情報には，日本語版ウィキペディアの情報を構造化データとして取得可能なサービスである DBpedia Japanese を利用し(以下 DBPedia)，DBpedia 上でのクエリの実行には RDF(Resource Description Framework)問い合わせ言語の 1 つである SPARQL を利用した．

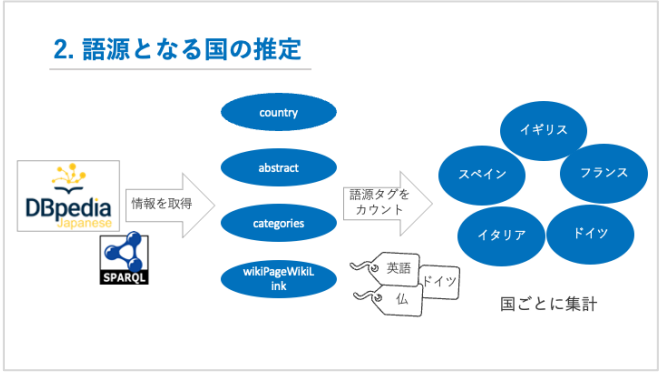

図 3　語源国推測のイメージ

まず 4.1 で取得した各外来語をキーとして subject, predicate, object の三つ組(RDF トリプル)の一覧を取得するクエリを DBPedia 上で実行する. subject, predicate, objectはそれぞれ主語, 述語, 目的語に対応し, 検索したい外来語を主語としてクエリを実行することで, 関連するプロパティの値を「subject の predicate は object である」というデータ構造で一挙に取得することが可能である. 取得したクエリ結果から語源国の情報の抽出に有効だと考えられる 4 種類の構造化データを取り出す. 対応するトリプルが 1 つも存在しない単語に関しては語源の推定は困難と判断し, 処理対象から除外する.

次に取得したテキストデータから, 語源タグと一致する部分を正規表現で抽出し, 語源タグごとにその出現回数を集計する. その後語源国によってタグをグループ化し, 重み付けを踏まえた各国の値をそれぞれ累計する. 累計後の各国のスコアを比較し, もっともスコアの大きい国をその外来語の語源国とみなす. 複数国のスコアが同率で一位となった場合, 同率一位となった国全てを語源国と判断する. 表 1 に本項で用いた語源タグの一例を示す.

単語	語源国	重み
英語	イギリス	10
イギリス	イギリス	5
英	イギリス	1
フランス語	フランス	10
フランス	フランス	5
仏	フランス	1

表 1　語源タグの例

4.3 かぶれ度の定義と算出

対象とする全作家に対して、これまでのプロセスで取得した外来語と語源国の一覧をもとに, 語源国ごとに単語数を集計する. この際イギリス由来と推測される単語が想定より多くなってしまったが, これは語源国の推測に現代語の情報資源を利用しているためだと考えられる. この問題に対処するため, 対象となる全作家のデータに対して標準化を行い, 各国間での分布のばらつきを調整する.

4.4 データの可視化

レーダーチャートを用いて算出した作家のかぶれ度を可視化する. 作家の外来語の語彙と出現回数, 語源国の情報については WordCloud を用いて可視化する.

レーダーチャートの詳細は以下のとおりである. まず, 今回かぶれ度の算出対象とした 5 ヵ国について, グラフの原点から放射状に伸ばした軸と一対一でそれぞれ対応させる. 次に特定作家の各国のかぶれ度を, それぞれ対応する国の軸の上にプロットし, 隣接した項目同士を線分で結んで多角形を表示する. 本項では各軸の最小値(原点)は-1, 最大値は 1 とした. プロットされた値が原点に近いほど, 作家の対応する国への文化的傾倒度は小さく, 反対に原点から遠く外側にあるほど, 作家の対応する国への文化的傾倒度は大きいものとみなす.

外来語の語彙と出現頻度については WordCloud を用いて可視化した. WordCloud はある文章中に出現する単語を, その出現頻度に応じた大きさのフォントサイズで一画面に敷き詰める様に表示する可視化手法である. 本項では単語のフォントサイズで出現頻度を, 単語の色で語源国を表現した.

5. 結果と考察

森鴎外と夏目漱石を対象に得られた分析結果から生成したレーダーチャートと WordCloud について考察し, 考察結果をもとに本稿の提案手法について評価する. この二人を考察対象として選出した理由は, 文学的アプローチからの先行研究での成果を踏まえ, 二人には特定の国への文化的な傾倒が存在することが予想できたためである. 森鴎外についてはドイツへの留学経験があること, また自身のドイツ滞在について扱った小説作品が存在することからドイツのかぶれ度が大きくなると仮説を立てた. 夏目漱石についてはイギリスへの留学経験があること, 英語, 英文学に造詣が深いという指摘が存在することからイギリスのかぶれ度が大きくなると仮説を立てた. 以下それぞれの仮説を検証する.

5.1 かぶれ度について

図 4 は森鴎外, 図 5 は夏目漱石のかぶれ度をプロットしたレーダーチャートである. 森鴎外については各国間で比較した際に特にドイツのかぶれ度が大きいことが窺える. 夏目漱石については各国間でかぶれ度の差異は森鴎外と比較すると小さいが, 5 ヵ国の中ではイギリスのかぶれ度が最も大きくなっている. 二人の特徴の表出についての仮説は概ね正しいことが検証できたといえる. また, 森鴎外と夏目漱石でグラフの面積を比較した際に大きな差があることがわかる. 二人の語源国の推測に成功した総単語数に 4~5 倍程度の差があったことを考慮すれば, その特徴を想定通りに可視化できているといえる.

5.2 外来語の語彙と使用頻度について

図 6 は森鴎外, 図 7 は夏目漱石の使用語彙を可視化した WordCloud である. 森鴎外についてはベルリン, ドレスデンなどドイツの地名が多く出現し, かつ各語の語源国が想定通りドイツ由来として色分けされてい

る．この結果と，森鴎外の作品にドイツを舞台としたものがあることから，作品中にドイツの地名が多数出現していたことが予想できる．夏目漱石の WordCloud からは，総単語数が多いこと，かつ色分けに着目すると 5 カ国の色が満遍なく出現していることが読みとれ，レーダーチャートでの分析結果と概ね合致しているといえる．また，シャツやウィリアムなど特定の単語のフォントサイズが突出して大きくなっており，作中での出現回数の多い重要な単語であることが予想できる．実際に夏目漱石 の作品中では登場人物の名前として「赤シャツ」と「ウィリアム」が用いられている．このように，従来の文学研究では手間をかけて作品を読むことで研究の手がかりを見つけていたところを，この WordCloud では一枚の画像を一見することで研究の手がかりとなりうる情報を取得することができ，WordCloud の利点を活かすことができているといえる．

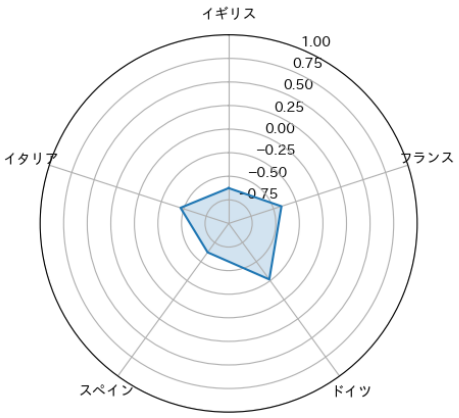

図 4　森鴎外のかぶれ度

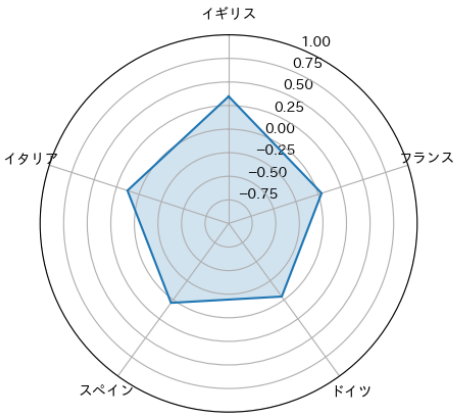

図 5　夏目漱石のかぶれ度

図 6　森鴎外の WordCloud

図 7　夏目漱石の WordCloud

6. まとめ

本稿では，作家の外来語の語彙と使用頻度からその作家がどの国にどの程度文化的に傾倒していたかを示す指標，かぶれ度という概念を示し，その算出方法を複数の可視化手法とともに提示した．結果，これまで文学的なアプローチから指摘されていたような作家の特定の国から受けている文化的影響について，各作家間で共通の指標を用いて可視化することに成功した．

今後の展望として，表記揺れへの対策によって語源国を推測するプロセスにおける精度の向上が期待できる．これは現代語のデータベースに登録されている表記と作品中での表記が異なる場合に有効である．また，フォクシェネアヌ[5]が指摘している様に同じ作家の中でも同一の語を表すと考えられる単語について複数種類の表記が存在しうる．作家が意図的に使い分けをしている可能性も考えられ，こうした単語を一つの単語とみなすか，それぞれ別の独立した単語としてみな

すかは精査が必要である.

また, 現状かぶれ度は使用している外来語の語源国の情報をもとに算出している. これに加えて外来語の意味的性質に着目し, 指標に反映することが可能だと考えている. 具体的には, 単語をベクトル化することで単語の意味情報を定量化し, 指標に反映する等の手法が考えられる.

参 考 文 献

[1] 大前義幸. 『ニコラス・ニクルビー』と『坊つちゃん』ー漱石作品におけるディケンズの影響を迫ってー. 岩手県立大学宮古短期大学部研究紀要, 第31巻, pp.1-10, 2021.

[2] Jennifer Scott. The Influence of the German Novelle on the Works of Mori Ogai. 就実論叢, 43号, pp.157-162, 2014.

[3] 金子淳, 大槻恭士, 坂口隆之. テキストマイニングと単語ベクトルを援用した英米文学研 究の試み. The 35th Annual Conference of the Japanese Society for Artificial Intelligence, 2021.

[4] 花畑圭佑, 青野雅樹. 語彙と文脈に着目した文学作品の著者推定. 言語処理学会 第 25 回 年次大会 発表論文集, pp. 1347-1350, 2019-03.

[5] アンカ・フォクシェネアヌ. 明治時代の文学作品における外国語・外来語の使用. アルザス日欧知的交流事業日本研究セミナー「明治」報告書. 2014-04.