

韓国語における比喩的素描手法に関連する一考察

Study on Analysis of Figurative Descriptions for Korean Language
Ha Heon-myeong* 榊井文人 Pham Xuan Quyen, Michal Ptaszynski
(北見工業大学)[†]

1 はじめに

日本と韓国は歴史的にも地理的にも多くの知識や問題を共有している。特定の語に対する解釈も同様であり、両国において共通のイメージを持つ語も多い。例えば、「米(쌀)」という語を例に挙げると、韓国においても日本においても、「主食(주식)とする穀物(곡물)」といった共通イメージが連想される。

これに対し、互いに相容れないイメージを持つ語も存在し、その典型例として「靖国神社(야스쿠니 신사)」が挙げられる。日本における「靖国神社(야스쿠니 신사)」のイメージは「神聖な場所(신성한 장소)」や「死者を祀るところ(죽은자를 모시는 장소)」、「宗教(종교)」、「歴史的な神社(역사적인 장소)」など比較的ポジティブなイメージが連想される傾向にあるが、韓国において「靖国神社(야스쿠니 신사)」は 세계전쟁 주범참배(世界戦争主導者の参拝)や「불길한 장소(不吉な場所)」、「식민지 역사(植民地の歴史)」といったネガティブなイメージが連想され、両国間で大きく異なることがわかる。

上で述べたような語に対するイメージの相違は日韓両言語間のコミュニケーションに少なからず影響を与えると考えられる。例えば、「靖国神社を参拝する」という文を直訳した場合「야스쿠니 신사를 참배하다」となり、表層上の翻訳には問題は生じない。しかし、文の解釈においては、日本語では通常の神社参拝に近い意味として解釈されるのに対し、韓国語においては非常にネガティブな行為として理解されてしまう。

このように「靖国神社, 야스쿠니 신사」だけではなく、「竹島, 독도」「慰安婦, 위안부」などの日韓両国が対立する問題を解消するためには、別の語や表現への換言や、語に対する補足情報を付与するといった対応が必要となる。

榊井ら[1]は, WWW から収集した知識断片(descriptor)を用いて語の意味を比喩的に素描する手法を提案している。彼らは評価実験を通して、彼らの手法が有効なイメージ知識を自動構成できることを検証し、さらにユーザが未知語の意味を連想的に把握する際の支援策としても有効であることを確認している[2]。

また、彼らが提案した比喩的素描手法は、比喩的な関係を持つ語と語の関係を利用するという点で言語依存性が低い。すなわち、比喩的素描手法を韓国語に応用することは可能である。この手法を日韓両言語に適用して関連付けることで、両言語間においてイメージの齟齬が発生する語を発見し、適切な換言処理や補足付与を適用することが期待できる。そこで本研究では、日韓両言語間における語解釈の齟齬解消を目的として、比喩的素描手法システムを利用した知識の比較とギャップの判定について検討する。具体的には、まず韓国語に対して比喩的素描手法を適用し、その有効性と課題について考察する。さらに、同じ query 語に対して日韓両言語の知識

源から比喩的素描を行い、それぞれの知識断(descriptor)を比較する。比較においては、被験者を用いた心理学実験を実施し、その結果を分析する。

2 韓国語の特徴

現在韓国では韓国語(朝鮮語)を公式に表記する文字としてハングルが用いられているが、それ以前は漢文が使用されていた経緯から、韓国語の基本単語の多くは漢字に由来している。また、20世紀前半には日本により植民地統治されたことにより、日本の言語や文化の影響を受けた言語表現も多い。分類上は、アルタイ諸語との関連性から、日本語と同系統に類別されることもある。統語的には SOV 型の語順を持つこと、膠着言語であること、助詞が主題を特徴付けるなど、日本語との類似性は極めて強いことが特徴である。

しかしながら、語義やイメージについては異なる点も多い。例えば、「犬(개)」という単語を提示してみると、日韓両国とも「動物」とであると認識する点で共通するものである。しかし、日本では「獣」や「ペット」などのイメージが強く、「回し者」や「スパイ」といった解釈は第二義的である。これに対し韓国の「犬(개)」は、「悪いイメージ」や「侮辱」を比喩的に表現する場合が多く、日韓で異なるイメージが解釈されることになる。

3 比喩的素描手法

榊井ら[1]は, WWW から収集した descriptor を使って比喩的に素描する手法を提案し、日本語を対象とした実装システム Murasaki (以下, J-Murasaki と呼ぶことにする)を構築している。榊井らの手法は、直喩表現を生成する指標パターンを利用して, WWW から大量の名詞句の関係を抽出する。そしてこれらの情報に基づいて, descriptor 候補と呼ばれる query 語を素描する断片知識を獲得する。各 descriptor 候補の一般性と局所性を考慮してその記述力を推定し、これに基づきランキングを行ない視覚化する。その結果 query 語の定義がイメージでき、理解を促すのに十分な補足知識を提供する。

J-Murasaki は 9 つの処理過程から構成される。

(STEP1) query 語の取得, (STEP2) 比較表現(A)の生成, (STEP3) 比較表現(A)検索, (STEP4) descriptor 候補の抽出, (STEP5) 比較表現(B)の生成, (STEP6) 比較表現(B)検索, (STEP7) スコア計算, (STEP8) descriptor 候補 集合構築, (STEP9) descriptor 候補の分類・整理と可視化からなる。Murasaki の構成を図 1 に示す。

我々は、上述した比喩的素描手法を韓国語にも適用する。ただし、ハングル文字に対応した高性能な形態素解析器は入手できなかったため、STEP2, 4, 5 における単

*m1252408011@ialab.cs.kitami-it.ac.jp

[†]北海道北見市公園町 165 番地

語認識は人手で実施した。簡単のために、本手法を K-Murasaki と呼ぶことにする。

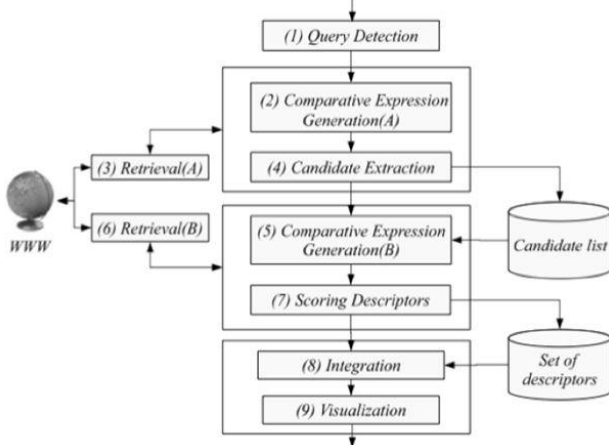


図 1 : Murasaki の構成

4 評価実験

K-Murasaki の処理能力を検証するために評価実験を行った。以下、評価環境と評価結果について述べる。

4.1 評価環境

20 個の query 語を用意し、これらに対して日本語と韓国語の両言語に対して比喩的素描手法を適用して descriptor を収集した。実験に用いた query 語を以下に示す。

バイク(바이크), 桜(벚꽃), ペンギン(펭귄),
蓮(연꽃), マグロ(참치), ベトナム(베트남),
サッカー(축구), 砂(모래), 犬(개), 韓国(한국),
台湾(대만), 中国(중국), 梅(매실), 夢(꿈),
富士山(후지산), 雪(눈), 日本(일본),
リンゴ(사과), 野球(야구), ブドウ(포도)

処理結果として得られた descriptor は、日本語については平均 23.5(最大 32, 最小 7)個, 韓国語については平均 25(最大 41, 最小 16)個であった。

上述した descriptor と query 語のセットを刺激語として日本語または韓国語を母国語とする被験者 14 人(日本語 14 人, 韓国語 14 人)に見せ, descriptor の妥当性を評価してもらった。評価にあたり, 以下の基準を設定した。

- (十分妥当) : query 語を説明／連想できる。
- × (妥当でない) : query 語を説明／連想できない。

また, 両言語間におけるイメージの齟齬の発生状況を調べるために, query と descriptor を相互に翻訳(日本語

の結果を韓国語へ, 韓国語の結果を日本語へ) し, 上記の被験者に再度評価してもらった。

4.2 評価結果

前節で述べた評価実験の結果について述べる。実験において, 日本人被験者と韓国人被験者の評価結果のばらつきを考慮するために被験者間の判定結果の κ 値を計算した。その結果, 韓国人被験者間の平均 κ 値として 0.67, 日本人被験者間の平均 κ 値として 0.64 が得られ, 両者とも被験者間の一致度は強いことが確認できた。さらに, 日本語 descriptor の翻訳結果を評価した韓国人被験者間の κ 値を計算したところ, 0.73 が得られた。

次に, K-Murasaki の抽出性能について述べる。図 2 は, 日韓両言語における descriptor 抽出性能を適合率で示したものである。対象とした 20 個の query に対する平均適合率は 79.25%であった。適合率のスコア上位は「桜」「バイク」「韓国」「夢」「中国」に対する 100%であり, スコア下位は「ペンギン」「リンゴ」に対する 40%, 38%であった。

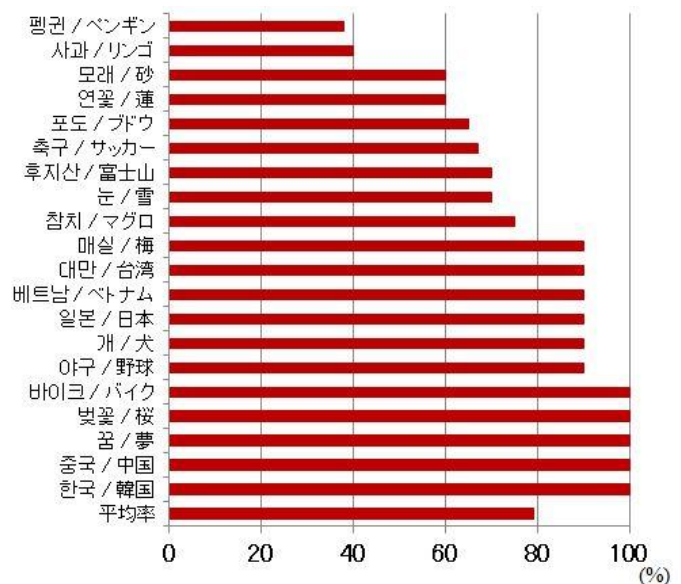


図 2 : K-Murasaki の適合率

さらに, 両言語間で descriptor がどの程度一致しているか(あるいは齟齬が生じているか)を調べた結果について述べる。具体的には, 処理対象とした 20 個の query 語の中から無作為に 10 個を選び, それらに関して日韓両言語の descriptor を比較し, 相互に相対再現率を計算した。

試行を 7 回繰り返し, 相対再現率平の均値をその結果をグラフ化したものを図 3 に示す。グラフでは, descriptor に付与されたスコアに基づいて 1~5 位, 6~10 位というように 5 位毎の階級を設定し, 各階級における descriptor の相対再現率を示している。青色のバーは J-Murasaki の相対再現率であり, 紫色のバーは K-Murasaki の相対再現率である。

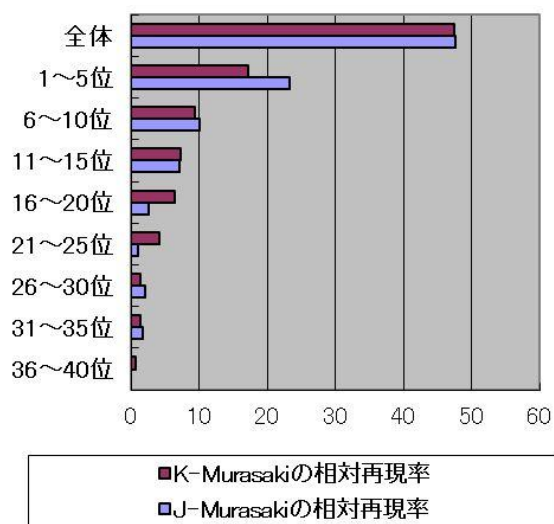


図 3 : K-Murasaki と J-Murasaki の相対再現率

この結果を詳しくみると、青色のバーでは、全体の相対再現率が 47.5%であり、階級別では 1~5 位が 17.1%, 6~10 位が 9.4%, 11~15 位が 7.3%, 16~20 位が 6.3%, 21~25 位が 4.1%, 26~30 位が 1.4%, 30~35 位が 1.3%, 30 位未満は 0.6%であった。紫色のバーで示す相対再現率は全体で 47.7%であり、階級別にみると、1~5 位が 23.3%, 6~10 位が 10.1%, 11~15 位が 7%, 16~20 位が 2.6%, 21~25 位が 1%, 26~30 位が 2.1%, 30~35 位が 1.6%, 30 位未満は 0%であった。

全体的には両者とも上位階級ほど相対再現率が高い傾向がみられた。さらに、J-Murasaki の結果が K-Murasaki の結果を上回っており、1~5 位の階級で 6.2 ポイントと最も顕著な差が得られた。16~20 位が 3.7 ポイント、21~25 位が 3.1 ポイントと、K-Murasaki の相対再現率が J-Murasaki をわずかに上回った。これは、日韓の言語に対して比喩素手法のランキングが機能しており、上位ほど精度よく descriptor が抽出されていることを示している。

表 1 は query 語毎に日韓両言語で意味が一致する descriptor を示している。descriptor 右側の数字は付与されたスコアに基づく順位である。

5 考察

評価結果に対する考察を行なう。

まず、図 2 を見ると K-Murasaki よに得られた descriptor の平均適合率は 79.3%と、J-Murasaki の結果 (68.1%) を上まわっている。このことから K-Murasaki の descriptor 抽出性能は一定の有効性を示しているといえる。国々の query 速についてみると適合率が高かった「桜」「韓国」「中国」「夢」「バイク」については、日韓とのイメージが交通であることが示唆され、適合率が低い「ペンギン」「リンゴ」は日韓でのイメージにギャップがあることが示唆される。表 1 に、上記の query に関する J-Murasaki の結果と K-Murasaki の結果を示す。

表 1 : J-Murasaki と K-Murasaki の共通 descriptor

query 語	共通 descriptor	
축구 サッカー	운동 (1), 스포츠 (2)	スポーツ (2)
	게임 (3)	ゲーム (3)
	구기 (4)	ボール (12)
	격렬한운동 (16)	激しい (19), 激しいスポーツ (29)
	단체경기 (18)	チームスポーツ (13), 団体スポーツ (34)
	놀이 (24)	遊び (17)
포도 ブドウ	과일 (2)	フルーティ (5), 果実 (13)
	열매 (5)	実 (2)
	맛 (7)	味 (4)
	새콤한과일 (8)	酸味 (9)
	눈 (11)	目 (11)
	당분 (17), 단맛 (20)	甘み (15)
후지산 富士山	옥녀봉 (1), 302 봉 (23)	独立峰 (2)
	활화산 (2)	火山 (3)
	산 (6)	山 (1)
	원뿔형 화산 (9)	コニーデ型火山 (9)
	높은산 (15)	高山 (7)
	모양 (16)	姿 (4), 形状 (5)
펭귄 ペンギン	동물 (2)	動物 (26)
	모습 (4), 모양 (15)	形状 (1), 格好 (9), 容姿 (21), 姿 (3), 模様 (34), 外見 (32)
	사람 (7)	人間 (28)
	새 (13), 조류 (4)	鳥類 (12), 鳥 (7)
	워킹 (18), 걸을걸이 (21)	歩き方 (2)
	동작 (34)	動き (17)
중국 中国	나라 (1), 국가 (3)	国 (13)
	큰나라 (6), 대국 (18), 거대한나라 (30)	巨大国家 (10)
	신흥국 (12), 신흥경제대국 (21), 신흥강국 (31)	新興国 (6)
	동물 (17)	獣 (5)
	이야기 (13)	人のお話 (10)
	열매 (1), 열매류 (2)	実 (12)
매실 梅	나무 (9)	木 (11)
	나라 (1)	国 (1)
	친일국가 (17)	親日国 (5), 親日的 (7)
	스포츠 (1), 운동 (2)	スポーツ (1)
	야구 野球	集団スポーツ (8), 団体スポーツ (5), チームプレイ (3)
	단체종목 (16)	
사과 リンゴ	내피부 (15)	皮 (2)

次に、図 3 の相対再現率についてみると、J-Murasaki と K-Murasaki とともに 47.7%, 47.5%とほぼ同等である。両言語ともほぼ半数の descriptor が共通であることがわかる。

対象とした query 数が少ないため厳密な議論は避けるが、語のイメージについていえば、両言語とも約半数の共通イメージを持っている可能性がある。イメージが共通する傾向の強い結果として、「サッカー (축구)」「ぶどう (포도)」がある。例えば「サッカー」については、日韓両国において非常にメジャーなスポーツとして認

知されており、ルールや道具なども共通していることで、共通 descriptor の一致率が高くなったと思われる。

これに対し、両者の間で descriptor が一致しない場合も半数存在する。例えば、「야구, 野球」「사과, リンゴ」「매실, 梅」「대만, 台湾」「개, 犬」などがある。

query 語「사과, リンゴ」「매실, 梅」の場合、日本で大きくイメージが異なる。「リンゴ (사과)」は、韓国では子供向け歌謡曲の曲名および歌詞として「リンゴのような私の顔」という表現が広く認知されているため、これらに関する descriptor が大量に取得されたことが不一致の主要因であると思われる。

同様に、query 語「매실, 梅」は、両国における梅の食べ方に違いがあることが descriptor の齟齬に繋がっていると思われる。

「台湾 (대만)」について考えると、日本では台湾は親日国として知られているためポジティブな descriptor が多数得られているが、韓国では台湾は敵対的なイメージが強いためネガティブな descriptor が多くなったことが相対再現率を低下させる原因であると考えられる。

「犬 (개)」の場合、両国の文化の違い強く影響していると考えられる。韓国では、従来から「悪人」「蔑み」の比喩として「犬」という語が多用されてきたためネガティブなイメージが強い。日本においても同様のイメージは存在するが、「ペット」や「愛犬」といったポジティブなイメージの方が先行するため、相対再現率が低下したと考えられる。

被験者間の κ 値をみると、韓国人被験者の κ 値は母国語に対するより日本語の結果に対する評価でより高かった。これは、両言語間でイメージに齟齬が生じる query 語については連想可能な許容範囲を超えているため、判断が容易であったためではないと思われる。このことを確かめるためには、同様の実験規模を拡大して κ 値を計算してみればよい。

以上のことから、相対再現率は日韓両言語におけるイメージの共通性や齟齬をよく反映しているといえ、一致率を利用すれば誤解の生じ易い語を推測できる可能性がある。

一方、J-Murasaki の手法を韓国語に適用した場合、同音異義語が問題となる。ハングルは母音と子音の組み合わせを表す表音文字であるため、漢字のように文字単体で意味を表すことがない。よって、同じ文字列の単位が複数の単語となりうるという曖昧性が頻発する。例えば、「배」という単語には「船」「梨」「腹」という3つの意味を持つ可能性があり、いずれの意であるかを断定するには単語が出現する箇所の前後の文脈を読まなければならない。

このような問題は query 語の検索と descriptor 候補の抽出から検出される確率がかなり高い。同音異語は現在にも韓国で大きな課題で残っている。そこで、次のような対応を考える。同音異語のような問題を減らすために、query 語の検索や descriptor 候補の抽出する過程で分流カテゴリを構築することを具象している。分流カテゴリを

繋がって検索することで同音異語の完全な除去は無理であるが、同音異語を持つ descriptor 候補の意味を分別することで、同音異語の解決策としても期待できる。

6 おわりに

本研究では、日韓二言語間における語解釈の齟齬解消を目的とした議論を行った。まず、榊井らが提案した日本語テキストを対象とした比喩的素描手法を拡張し、韓国語に対応する比喩的素描手法を提案した。次に、二つの比喩的素描手法を利用して両言語の素描知識を収集し、二言語間の知識の比較と知識のギャップについて議論した。

提案手法の有効性を検証するために、20 個の query 語を提案手法に適用し、評価実験を行なった。その結果、79.25%の適合率が得られ、提案手法が一定の性能を発揮することが確認できた。

さらに、query 語を日本語に翻訳して J-Murasaki にも適用し、提案手法の結果と J-Murasaki の結果を比較した。その結果、得られる descriptor のほぼ半数が両言語間で共通していることがわかった。

J-Murasaki と K-Murasaki の共通性や K-Murasaki の評価結果が上記のような結果が得られることで、K-Murasaki は韓国語の意味を比喩的に素描する方法により高い効力を持つと期待できる。また、日韓両国で、言葉の意味差が大きくないことが判断され、K-Murasaki と J-Murasaki が直接的に比べることができれば、両国の言葉から起きる意味の差を減少させることができると思われる。

今後は、Fuji らの研究[3]を利用して、日韓両言語で得られた descriptor を相互に翻訳し、その妥当性を評価することで、クロスリンガル比喩的素描の可能性について検討する予定である。

参考文献

- [1] 榊井文人, ジェブカ・ラファウ, 木村泰知, 福本淳一, 荒木健治: "WWW 活用による語の比喩的素描手法", 知能と情報, Vol. 22, No. 6, pp. 707-719(2010)
- [2] 川村佳史, 榊井文人, 河合敦夫, 井須尚紀: "WWW からの descriptor 抽出システムの開発とその評価", 言語処理学会第 13 回年次大会発表論文集, pp. 1133-1136(2007)
- [3] Atsushi Fujii, Tetsuya Ishikawa, Jong-Hyeok Lee: "Term Extraction from Korean Corpora via Japanese", Proceedings of 3rd International Workshop on Computational Terminology (CompuTerm2004), pp. 71-74(2004)