日本語メタファーコーパス作成のためのガイドライン

宮澤 彬　　吉田 奈央　　宮尾 祐介
総合研究大学院大学 情報学専攻 / 国立情報学研究所
{miyazawa-a, naouyoshida, yusuke}@nii.ac.jp

1 はじめに

言語学，特に認知言語学において，メタファーは人間の思考過程を反映するものとして精力的に研究されてきた．メタファーは抽象的な概念の理解や新しい概念の命名などにおいて不可欠であり，その仕組みを明らかにしていくことは極めて重要である．自然言語処理においても，メタファー的な表現を正しく理解することができれば，それを語義曖昧性解消や文章の平易化などへ応用することが考えられる．このような理由からメタファー表現を収集して検索や計量的な分析を可能にすることは，言語学の理論的側面から自然言語処理における応用まで広く役立つと考えられる．しかし現在，日本語においてこれらの用途で必要となるメタファーの情報付きのコーパスはまだない．

そこで本稿では，日本語のメタファーコーパス作成のためのガイドラインを提示する．さらに本ガイドラインを用いて，京都大学テキストコーパスの目的語・動詞に対して行ったアノテーションについて説明する．最後に，アノテーション結果について一致率等の分析を行い，ガイドラインとアノテーションについての評価を行う．

2 関連研究

メタファーの情報が付与されたコーパスとして，英語ではVU Amsterdam Corpus [1] がある．メタファーとなる表現の単位は，形態素から複数の文まで考えられるが，ここでは単語の単位でメタファーか否か[1]がマークアップされている．このコーパスはBNC Babyコーパスの一部約50,000語から成り，学術・ニュース・フィクション・会話の計4つのレジスターを含んでいる．作成に使用されたガイドラインはMIPVU[2]と呼ばれ，[1] に詳細にまとめられている．

自然言語処理の分野では，コーパスではないがTroFi Example Base [2, 3] という用例集がある．これは元々メタファーの識別システムTroFi (Trope Finder) のために作られたものである．50個の動詞のリストに含まれる動詞を含む約3,700文について，動詞の用法がメタファー的か否かのアノテーターによる判断とTroFiの識別結果が付与されている．

はじめに述べたように，日本語ではまだメタファーの情報が付与されたコーパスは利用可能でない．少し視野を広げると，文学作品の直喩の用例を集めた [4] や，メタファーを含む様々な修辞技法について例を用いて解説している [5] などがある．しかしこれらは実際の言語使用の研究や機械学習の訓練データとしての用途を見据えて作られたものではなく，また紙媒体でしか利用できない．

3 ガイドライン

3.1 対象

MIPVUでは限定詞や前置詞に分類される語に対してもアノテーションを行っているが，今回はコストを抑える目的で目的語・動詞についてのみアノテーションを行う．動詞を対象に含めた理由は，動詞が他の品詞と比較してメタファーの割合が多く [1]，アノテーションがしやすいと考えたからである．目的語にあたる名詞も対象に含めたのは，動詞と共起する語が何か，またそれがメタファー的か否かといった情報が，その動詞がメタファー的か否かに影響を与えており，今後メタファーの自動判別等へ応用する上で目的語へのアノテーションが有用になると考えたからである．

ここで言う目的語は，格助詞「を」によって格標示されている名詞であり，詳細な照応解析を必要とする省略などは対象としていない．また「こと」や「もの」のような形式的な名詞，「する」や「やる」のような形

[1] 正確にはメタファーはさらにいくつかの下位分類に分けられている．
[2] Metaphor Identification Procedure VU University Amsterdam

式的な動詞は，単独ではほとんど実質的な意味を持っておらず，メタファーか否かの判断ができないため除外している．

アノテーターは表 1 で示されているように，目的語と動詞それぞれにメタファー的 (metaphoric) であることを表す記号 m，または字義通り (literal) であることを表す記号 l のどちらかを付与する．

3.2 資料

次項で説明する手順の中に語義を特定する作業があるが，その作業では「岩波国語辞典第五版タグ付きコーパス 2004」とその閲覧ソフト「GDA コーパスブラウザ」を用いる．複合動詞，特に語彙的複合動詞 [6] は意味の合成性が成り立つとは限らないものの，辞書に収録されていない場合が多いため，その場合は「複合動詞レキシコン」[7] を用いる．

3.3 手順

以下が m か l を判断するための主な手順であるが，その内容は MIPVU をほぼそのまま踏襲している．

1. 文脈語義を特定する．
2. 辞書を引き，基礎語義を特定する．
3. 基礎語義と文脈語義が十分に異なるか判断する．もし異なっているならば次の手順へ進む．異なると判断できない場合は，その語は字義通りと結論付ける．
4. 基礎語義と文脈語義の間にある種の類似性が存在するか判断する．存在すればそれをメタファー的であると見なす．

ある語の文脈語義とは，その語を含む文脈によって定まる語義である．基本的に文脈語義は，辞書のその語彙項目に記載されている語義の中から選ぶ．しかし場面特有の語義などは辞書に記載されていないため，その場合は備考欄にその語義を記載する．

例えば表 1 の 1 行目『香港移民は「チャイナタウン」の枠を飛び越えてしまった。』における「枠」という語について，辞書を引くと以下のように説明されている．

> **わく【枠】**
> ① まわりをふちどって囲むもの。「めがねの—」「記事を点線の—で囲む」
> ② 物事をふちどるような、定まった型。
> ㋐ 制約。「法律の—」「—にはまった (=型通りで新鮮味がない) 表現」
> ㋑ 限度となる規準。「新たに予算の—を獲得する」

ここでの文脈語義は②となる．なお岩波国語辞典の語義には大・中・小の粒度があるが，基本的には中区分① ②…を用いる．このことについては第 3.4 項で詳しく述べる．

基礎語義とは次のような性質を持つものである．

- 具体的であり，外見・音・手触り・臭い・味を想像しやすい．
- 身体的な動作に関連している．
- 詳細であり，漠然としていない．
- 歴史的に古い．

語義が 1 つしかない場合，それ以上基礎的な語義はないということであるから，その語義が基礎語義となる．また主な関心は現代の言語使用にあるため，既に廃れた語義，つまり現代語の辞書に掲載されていない語義は基礎語義としない．上で例として挙げた「枠」について，その基礎語義は①である．これは基礎語義の定義の，外見が想像しやすいという性質を満たしているからである．

十分に異なるかどうかは，辞書中で語義が別の項目になっているかどうかで判断する．上述の「枠」は基礎語義と文脈語義が辞書中で異なる項目になっているため，次の手順である類似性の検証の対象となる．

ある種の類似性とは，見た目や機能において何らかの共通の属性を持っていることである．典型的なのは実体のあるものが，抽象的なものに拡張されるような場合に見られる．例えば「覆す」という動詞は「ちゃぶ台」のような具体的なものに対して使われるときと，「常識」のような抽象的なものに対して使われる場合がある．これらは両方とも「通常の状態から異常な状態へ変化させる」という機能を持っているため，これらの語の間にはある種の類似性が成立しているとする．例文の「枠」についてこの規準を適用すると，まず①の語義は異なる 2 つの物質の境界であり，②の語義は現在の状態と別の状態を分ける境界のような存在を指している．よって「境界」として共通の機能を持っており，ある種の類似性が成り立つ．以上の手順から最終的にこの「枠」の用法はメタファー的であると結論付け，記号 m を付与する．

表 1: アノテーションの例

		目的語			動詞	
… 香港移民は「チャイナタウン」の	m	枠を		m	飛び越えてしまった。	
… 香港町が	m	低空飛行を		l	続ける	カナダ・トロント経済を …
… リッチモンドヒルなどに	l	住宅を		l	購入、	ショッピングセンターを …

3.4 辞書の特性の考慮

岩波国語辞典の特徴として，3 種類の語義区分（語義の粒度）が存在することが挙げられる．これは粗い物から順に大区分㊀ ㊁…，中区分① ②…，小区分㋐ ㋑…と表される．すべての語に存在するのは中区分であるため，基本的にはこれを用いる．しかし小区分の中で具体的な事物を表す語義と，それを抽象的・比喩的に拡張した語義が存在した場合は，それらは十分に異なるとみなせることにする．

岩波国語辞典では，語義の説明の中で「転じて」や「比ゆ的に」といった語のあとに比喩的な用法を紹介している箇所が多くある．例えば「乗り越える」という語を引くと 1 つ目の語義として，「高いものの上を（踏み）超えて向こう側におりる。「塀を—」。転じて、むずかしい局面を切り抜ける。」とある．この場合，アノテーション対象となる語の基礎語義が前半の記述に合致し，文脈語義が後半の説明に合致するならば，その語の用法はメタファー的であるとみなせることにした．

3.5 写像の理論について

認知言語学にはメタファーは領域間の写像 [8] として定義されることが多い．しかし領域の特定はしばしば困難であり，このような定義をそのままメタファーの判定に用いるは難しい．もし領域を特定できたとしても抽象度の恣意性の問題や，領域間に重なりが生じてしまう問題がある [9]．しかし第 3.3 項で述べた手順ではこれらの問題を直接扱わずに済む．MIPVU では領域間の写像が見出されるとき ‘MRW, direct’ (MRW は metaphor related word の略) という記号を付与すると定めている．しかし，これに該当する語は第 3.3 項の手順でカバーされると考えられるため，今回のガイドラインには写像の理論を用いた手順は含めていない．このことにより，アノテーターに対して期待される認知言語学的な知識が軽減されると考えられる．

表 2: 結果

	O	V
事例数	764	725
#m(A)	60 (7.9%)	214 (29.5%)
#m(B)	75 (9.8%)	205 (28.3%)
$\#\mathtt{m}_{\mathrm{agree}}$	35 (4.6%)	148 (20.4%)
Fleiss’ κ	0.477	0.587
精度	0.467	0.722
再現率	0.583	0.692
F-値	0.519	0.706

4 アノテーション作業

4.1 対象・作業者

前節で説明したガイドラインに基づき，京都大学テキストコーパス Version 4.0 のうち 1,100 文に対してアノテーションを行った．このコーパスを選んだ理由としては，第一に，文法に関する情報が付与されているため，アノテーション用のフォーマットの作成や，将来的な結果の利用に便利であると考えたことが理由である．加えて，新聞記事から成るために崩れた表現が少なく，ガイドラインの骨子を修正していく過程で過度に詳細な分析を避けられると考えたことも理由である．作業は著者 2 名で行った．

5 評価

結果を表 2 に示す．事例数は形式的な名詞・動詞を除いた，実際に m/l の判断がなされた語の数である．また項目 #m(A), #m(B) は，それぞれアノテーター A, B が m と判断した事例数であり，$\#\mathtt{m}_{\mathrm{agree}}$ は，両者がともに m と判断した事例数である．括弧内はそれぞれが総数に占める割合を表している．精度，再現率，F-値は，アノテーター 2 名のうちガイドライン作成者である方のアノテーター (A) の判断を正解データとみなして計算している．[1] では MIPVU の最終的な評価

として，4 人のアノテーターによる判断の一致率等を提示しており，これによるとニュースのレジスターでは Fleiss' κ が 0.96，全体では 0.85 となっている．今回の試みにおいて，MIPVU よりも大幅に低い一致率となった原因として，目的語では抽象的な名詞で意見があまり一致しなかったことがある．例えば「会発足の意義を力説する」の「意義」は以下の 2 つの語義がある．

> **いぎ【意義】**
> ① その言葉によって表される内容。意味。
>
> ② 行為・表現・物事の、それが行われ、また、存在するにふさわしい価値。「意義のある事業」

これらはどちらも抽象的な内容を指しており，どちらか一方を基礎語義と認定するのが難しい．このような場合は「基礎語義なし」として `m/l` の判断の対象から外すことが考えられる．他の不一致の原因としては，語義の説明において「転じて」等のメタファー的用法であることを表す表現なしに，通常の用法とメタファー的用法が併記されている場合の判断基準を決められていなかったことがある．例えば「南北統一の実現を呼びかけた」の「呼びかける」の語義は

> **よびか-ける【呼 (び) 掛ける】**〖下一他〗人に声を掛ける。「警察官がメガホンで―・けた」。また、説いて誘う。訴える。「大衆に―」「協力を―」

となっており，後半をメタファー的用法ととるかどうかはアノテーターの判断に任されていた．このような場合に明確な判断基準を定めることで一致率や F-値が改善されると期待できる．今回アノテーションを行ったのは，連続的な文章であり同じ単語が頻出する．よっていくつかの頻出語について判断を確定できれば，それらの数値は大きく向上するだろう．

6 おわりに

本稿ではまず日本語のメタファーコーパス作成のためのガイドラインを提示した．次に実際にそれを用いて京都大学テキストコーパスに対してアノテーションを行い，その評価を示した．今後はデータ，ガイドラインの整備を進め，メタファーの自動検出等への応用を行う予定である．

参考文献

[1] Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing, 2010.

[2] Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[3] Julia Birke and Anoop Sarkar. *Proc. of the Workshop on Computational Approaches to Figurative Language*, chapter Active Learning for the Identification of Nonliteral Language, pp. 21–28. Association for Computational Linguistics, 2007.

[4] 中村明. 比喩表現辞典. 角川書店, 1995.

[5] 佐藤信夫, 佐々木健一, 松尾大. レトリック事典. 大修館書店, 2006.

[6] 影山太郎. 文法と語形成. ひつじ書房, 1993.

[7] 神崎享子. 『複合動詞レキシコン』ver. 1 ―形態的・統語的・意味的情報付与―. 言語処理学会第 19 回年次大会発表論文集, pp. 761–764, 2013.

[8] George Lakoff. The contemporary theory of metaphor. *Metaphor and thought*, Vol. 2, pp. 202–251, 1993.

[9] Ekaterina Shutova and Simone Teufel. Metaphor corpus annotated for source - target domain mappings. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010.