

メタファーを理解するとはどういうことか  
ー心理学的アプローチおよび工学的アプローチー

# メタファー表現の自動生成

総合研究大学院大学

宮澤 彬

2018 年 12 月 16 日

# はじめに

メタファーには非メタファー的な表現では得られない視覚的な効果などを狙って、小説や演説などで好んで使われる。

**懐中電灯のように他人を照らす（批判する）だけで、自分を照らさないはいけない**

**大きな海（中国）は風雨に翻弄されることはない。中国は5千年以上の苦難を乗り越えて、変わらずここにある。**

習近平国家主席の演説を、朝日新聞 DIGITAL 「『世界の市場』見せつけた中国 輸入博で『爆買い外交』」(<https://digital.asahi.com/articles/ASLC54W7XLC5UHBI01W.html>) より引用。

日本語を対象として「よいメタファー表現」を自動生成する

## 生成の目的（メタファーの利点）

- ▶ 政治家が説得力のある演説をするのに効果的  
([Charteris-Black, 2011](#)).
- ▶ メタファー的な表現は、非メタファー的な表現と比較して、感情に及ぼす影響が大きい ([Mohammad et al., 2016](#)).
- ▶ 新たな（実在しない）事物や現象の名付けに不可欠。  
(例)「炎上」「燃料の投下」「延焼」

→ 小説や演説原稿の執筆支援といった応用

## 関連研究

（広義の）メタファーまたは**比喩**は、レトリックの一部門として古代ギリシャの時代から研究されてきた。

▶ **直喩**

「ような」といった比喩を明示する言葉を使う表現

（例）「りんごのような頬」，「風のように走る」

▶ （狭義の）メタファーまたは**隠喩**

中心的な意味とは異なる意味で使われている表現

（例）「愛情を**注ぐ**」，「**澄んだ**気持ち」

▶ **メトニミー**

「隣接性」をもつ別の語を用いる表現

（例）「筆を折る」，「漱石を読む」

Steen et al. (2010) は *BNC Baby* の各語に *metaphor-related word* (MRW) などのラベルを付与したコーパス *VU Amsterdam Metaphor Corpus* を構築した.

MRW の判定基準 (MIPVU) は以下のようなものである.

1. 語の**文脈語義**を特定する.
2. **より基礎的な語義**<sup>[1]</sup>が現代語の辞書に存在するか確かめる.
3. そのような語義が存在して、かつ文脈語義がその語義との関連で理解される場合、それを MRW と見なす.

---

<sup>[1]</sup> より具体的であったり、五感に結び付きやすかったり、身体的な動作に関係しているような語義.

例えば以下の文中の “struck” は MRW である.

*His speech **struck** me as the feeblest of the day.*

**strike** /straɪk/ VERB

1. to hit against, or to crash into, someone or something ← **より基礎的な語義**
2. to make someone have a particular opinion or feeling ← **文脈語義**

*Macmillan Dictionary*

## 関連研究 IV (自然言語処理)

自然言語処理においてメタファーを扱うタスクは主に以下の3つである.

**メタファー識別** (Tsvetkov et al., 2014)

(例) “prisoner **escaped** jail” → Lit,  
“blunder **escaped** notice” → Met

**メタファー理解** (岩山他, 1991; Shutova et al., 2013)

(例) 「りんごのような頬」  
{ "色": { "肌色": 0, "赤": 1, "青白": 0 },  
"外形": { "球状": 1, "平面状": 0 },  
"表面": { "滑らか": 1, "ざらざら": 0 } }  
“**hold back** (the) truth” → “**conceal** (the) truth”

**メタファー生成** (Abe et al., 2006)

(例) “character” + “young, innocent, and fine”  
→ “the character like a puppy”



Tsvetkov らは、複数の言語に対応したメタファー識別モデルを提案した ([Tsvetkov et al., 2013, 2014](#)).

- ▶ 抽象度と想像可能性

*MRC psycholinguistic database* ([Wilson, 1988](#))

(例) think: 346|384, cat: 485|551

- ▶ 上位概念

*WordNet* ([Miller, 1995](#))

(例) cat.n.01 ∈ noun.animal

- ▶ 単語のベクトル表現

[Huang et al. \(2012\)](#), [Faruqui and Dyer \(2014\)](#)

- ▶ 固有表現

(例) 京都 ∈ location

## タスク定義

## 入力

主語・目的語・動詞からなる表現「 $s$  が  $o$  を  $v$ 」(これを  $(s, o, v)$  と表記)で,  $v$  が非メタファー的に使われているもの

(例)  $(s, o, v) = (\text{“彼”}, \text{“気持ち”}, \text{“考慮する”})$

## 出力

表現  $(s, o, v')$  が後述する評価指標で高く評価されるような  $v'$  の集合. 例えば入力と同義であることや, メタファー的であることが評価される.

(例)  $\{\text{“汲み取る”}, \text{“掬い取る”}, \text{“理解する”}\}$

# タスク定義 II

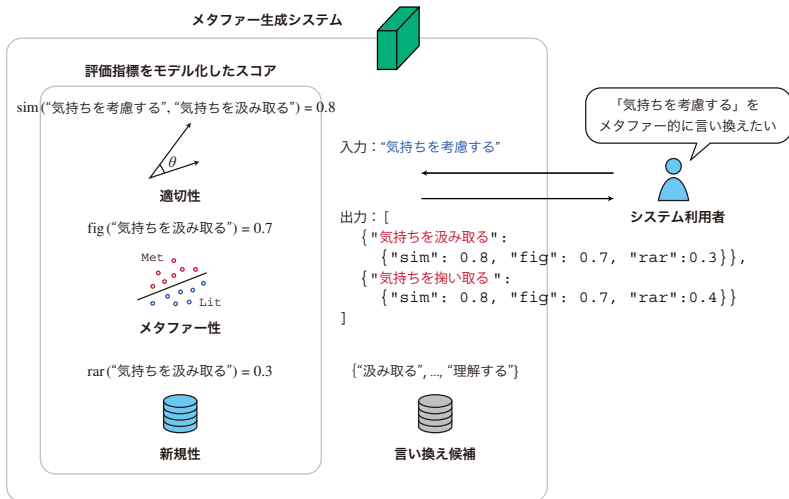


図 1: システムの概要図

なぜ動詞とその目的語の組を対象とするのか

- ▶ 動詞はメタファー的に使われやすい<sup>[2]</sup>
- ▶ メタファー検知・理解のタスクで SVO を扱う研究が多い  
([Shutova et al., 2013](#); [Tsvetkov et al., 2014](#))

なぜ日本語を対象とするのか

- ▶ 自ら評価を行いやすい
- ▶ 評価者を募りやすい

---

<sup>[2]</sup> ニュースにおける各品詞の MRW の割合は、名詞 13.2%, 形容詞 21.0%, 動詞 27.6%である ([Steen et al., 2010](#)).

## 評価指標・評価方法

# 評価指標 I

各出力は表現として以下の指標について人手で評価される。各例は入力を「彼が気持ちを考慮する」とした場合。

- ▶ **適切性**

入力表現と比較して、どの程度意味が似ているか。

- ▶ **メタファー性**

どの程度メタファー的か。

- ▶ **新規性**

どの程度新しいと感じるか。

- ▶ **理解可能性**

どの程度理解しやすいか。

- ▶ **相対的な好ましさ**

ある別な表現と比較して、どちらをより使いたいと感じるか。

## 評価指標Ⅱ（生成・評価）

Miyazawa and Miyao (2017) はメタファー生成タスクに必要な評価指標として**メタファー性・新規性・理解可能性**を導入した。

また、それらがクラウドソーシングで評価できること、全指標で高評価の表現はより使いたいと感じられやすいことを示した。

表 1：理解可能性の評価結果（抜粋）。

	水	不満	愛	希望	嫉妬	ネコ
X が溢れる	3.6	3.8	3.7	3.3	3.5	2.1
X が満ちる	3.8	3.2	3.4	3.6	2.2	1.0
X を撒き散らす	3.5	3.9	2.9	2.2	2.6	0.4
X が漏れる	3.9	3.8	2.1	0.9	1.9	0.4
X が濁る	3.7	1.4	1.5	0.8	0.4	0.1
X が沸騰する	4.0	2.1	2.0	1.2	2.0	0.3



# 評価指標 III

また Miyazawa and Miyao (2017) では、メタファー性・新規性・理解可能性が**すべて高い表現は、好ましい**（より使いたいと感じさせる）**メタファー表現**であることを示した。

表 2：評価の和による順位と好ましさ。

上位 10% の表現	下位 90% の表現	上位が好ましい
不満を飲む (23)	油を汲み取る (1087)	✓
怒りがこぼれる (6)	岩に溺れる (1117)	✓
羞恥心が溜まる (44)	羞恥心を注ぐ (856)	✓
?情報が濁る (106)	空気を撒き散らす (212)	✓
悲しみがしみる (32)	理解が流れる (721)	✓
楽しさが渦巻く (81)	不満に漬かる (1241)	-
言葉が滲む (14)	恐怖が流れる (307)	-
感情を注ぐ (44)	意図に漬かる (654)	✓
不安が流れ出る (44)	情熱を汲み取る (165)	✓
情報に溺れる (23)	油が溜まる (1241)	✓

# 評価方法 I

評価はクラウドソーシングによって行う。各表現、各指標について例えば以下のような設問を  $N$  人に提示する。

以下の表現は新しいですか。5 段階で評価してください。

「新聞社が結果を見送る」

5. 使ったこと、見聞きしたことがない新しい表現である

1. 広く使われている慣用的な表現である

選択肢は 1 から 5 の順で段階的に変化するものとします。

得られた回答  $(a_i)_{i=1}^N$  を次式で集約し、最終的なスコアとする。

$$\sum_{i=1}^N \frac{a_i - 1}{4}. \quad (1)$$

## 評価方法 II

相対的な好ましさの評価では、表現  $\varepsilon$  と  $\varepsilon'$  のどちらをより使いたいと感じるか尋ね、Bradley-Terry モデル (Bradley and Terry, 1952) で実数のスコアを得る。

表現  $\varepsilon$  が  $\varepsilon'$  より好ましい (これを  $\varepsilon \succ \varepsilon'$  と表記) 確率を

$$p(\varepsilon \succ \varepsilon') = \frac{\pi_\varepsilon}{\pi_\varepsilon + \pi_{\varepsilon'}} \quad (2)$$

とモデル化する。各  $\varepsilon \succ \varepsilon'$  が  $n_{\varepsilon\varepsilon'}$  回ずつ観測される尤度は

$$L(\boldsymbol{\pi}) = \log \prod_{\varepsilon \neq \varepsilon'} \left( \frac{\pi_\varepsilon}{\pi_\varepsilon + \pi_{\varepsilon'}} \right)^{n_{\varepsilon\varepsilon'}} \quad (3)$$

である。これを最大化する  $\boldsymbol{\pi} = (\pi_\varepsilon)_\varepsilon$  は一意に求められるため、その各要素  $\pi_\varepsilon^*$  を  $\varepsilon$  の相対的な好ましさのスコアとして用いる。

## 提案手法

各評価指標に対応した自動計算可能なスコア（**内部スコア**）を導入し，これらが大きくなるような候補を出力する．

表 3：評価指標と内部スコアの対応関係．

評価指標	内部スコア
適切性	類似性
メタファー性	比喩らしさ
新規性	希少性
理解可能性	（普遍性）
相対的な好ましさ	総合スコア

表現同士の意味の近さを直接測ることは難しいため、動詞のベクトル表現間のコサイン類似度で近似する。すなわち入力  $(s, o, v)$  と出力  $(s, o, v')$  の類似性を

$$\text{sim}(v, v') = \frac{1}{2} \left( 1 + \frac{\gamma(v) \cdot \gamma(v')}{\|\gamma(v)\| \|\gamma(v')\|} \right) \quad (4)$$

で定義する。ここで  $\gamma$  は語にそのベクトル表現を対応させる写像である。アフィン変換は 0 以上 1 以下の値をとるようにするためのものである。

はじめに表現をメタファー的か否か (Met/Lit) に分類する識別モデルを構築する. 表現  $(s, o, v')$  の意味に関する特徴量を  $\phi(s, o, v')$  と表す. 比喩らしさを以下で定義する.

$$\text{fig}(s, o, v') = p(\text{Met} \mid \phi(s, o, v')). \quad (5)$$

特徴量は基本的に [Tsvetkov et al. \(2014\)](#) に基づく.

表現  $(s, o, v)$  の希少性  $\text{rar}$  を以下で定義する.

$$\text{rar}(o, v) = \left( 1 + \frac{p(g(o), v)}{p(g(o)) p(v)} \right)^{-1}. \quad (6)$$

ここで  $p$  はある語 (unigram) またはある連続する 2 語 (bigram) の生起確率で,  $g(o)$  は  $o$  の属するクラスターを表す.

これは 2 語の関連の度合いを測る自己相互情報量 (PMI)

$$\text{PMI}(o, v) = \log \frac{p(o, v)}{p(o) p(v)} \quad (7)$$

の変種である.



PMI を直接使わないのは

1.  $\mathbb{R} \setminus [0, 1]$  の値をとることがある
2. コーパス中で共起しない  $(o, v)$  はすべて  $-\infty$  になってしまうからである.

1 つ目の解決策として, 変換  $x \mapsto \text{sigmoid}(-x)$  を適用する.

2 つ目の (部分的な) 解決策として, 各名詞のベクトル表現から作られたクラスター  $\{\mathcal{C}_i\}_{i=1}^K$  と各動詞  $v$  の共起確率  $p(\mathcal{C}_i, v)$  を用いることにする.

総合スコアは、類似性・新規性・比喩らしさの調和平均である.

$$\begin{aligned} \text{os}(s, o, v, v') \\ = 3 \left( \frac{1}{\text{sim}(v, v')} + \frac{1}{\text{rar}(o, v')} + \frac{1}{\text{fig}(s, o, v')} \right)^{-1}. \quad (8) \end{aligned}$$

重要度や値の分布について検証不足のため重み付けなどを行っていない.

データセット

# データセット I

必要なデータセットは以下である.

- ▶ 動詞に Met/Lit のラベルが付いた表現のリスト. 入力と, メタファー識別器の訓練・テストに用いる.  
(例)  $\mathcal{D} = [(\text{“取材陣が息を殺す”}, \text{Met}), \dots, (\text{“<person/>が小説を朗読する”}, \text{Lit})]$
- ▶ 言い換えの候補となる他動詞の集合.  
(例)  $\mathcal{V} = \{\text{“開ける”}, \dots, \text{“考慮する”}\}$
- ▶ 比喩らしさの特徴量として用いる, 語の意味に関する情報.  
(例) 抽象度, 想像可能性, 上位概念
- ▶ 単語のベクトル表現. 類似性・新規性・比喩らしさの計算に用いる.

## データセット II (入力と訓練・テストデータ)

ラベル付きの表現には宮澤他 (2016) のデータセットを用いた。これは京大コーパス (Kawahara et al., 2002) の OV 対の一部に MIPVU の簡略版に基づくラベル付けをしたものである。

表 4: データセットの評価。Fleiss'  $\kappa$ <sup>[3]</sup> が 0.4 を超えており、アノテーター A1 と A2 の間に緩やかな合意があることが分かる。

動詞の総数	A1 の #Met	A2 の #Met	Fleiss' $\kappa$
725	214 (29.5%)	205 (28.3%)	0.587

---

<sup>[3]</sup>Fleiss'  $\kappa$  (Fleiss, 1971) はアノテーター間の一致率の指標の一つで、偶然による一致の影響を考慮したもの。

「現代日本語書き言葉均衡コーパス」(BCCWJ)の動詞から「京都大学格フレーム」([Kawahara and Kurohashi, 2006a,b](#))を用いて他動詞のみを抽出した.

### 飲む / のむ:動1

<ヲ格> 酒 87526, 薬 43330, ビール 40683, 水 36182, ...

## データセット IV (比喩らしさの特徴量)

- ▶ **抽象度・想像可能性**

「日本語 WordNet」(Isahara et al., 2008) で英語に翻訳し  
MRC psycholinguistic database を参照

- ▶ **上位概念**

日本語 WordNet で英語に翻訳し *WordNet* (Miller, 1995)  
を参照

- ▶ **単語のベクトル表現**

CBOW アルゴリズム (Mikolov et al., 2013) を BCCWJ に適  
用して得た 100 次元ベクトル

- ▶ **固有表現**

京都大学テキストコーパスのアノテーションを利用

# 実験



本研究では3つの実験を行う。実験2と3ではクラウドソーシングを利用する。

▶ **実験1**

メタファー識別器の性能を計測する

▶ **実験2**

各内部スコアと人手評価と相関しているか検証する

▶ **実験3**

総合スコアが、相対的な好ましさをモデル化できているか検証する

# 実験1：メタファー識別器の性能

精度，再現率ともに 0.5 以上になっており，メタファーの識別器として機能していることが確認できた。

データセットが異なるため単純には比較できないが [Tsvetkov et al. \(2014\)](#) の F 値 0.86 と比べると低い。データサイズの拡大，アノテーションの見直しなどで改善できる可能性がある。

表 5：メタファー識別器の性能。

正解率	精度	再現率	F 値
0.83	0.78	0.50	0.61

## 実験2：人手評価と内部スコアの相関

10 個の入力表現から得られた出力について、内部スコアとクラウドソーシングで得られた評価値を比較した。これらの間には弱い相関が成り立つことが分かった。

表 6：内部スコアと人手評価の相関係数。

内部スコア	評価指標	相関係数
類似性	適切性	0.27
比喩らしさ	メタファー性	0.38
新規性	新規性	0.35
普遍性	理解可能性	0.35

## 実験2：人手評価と内部スコアの相関

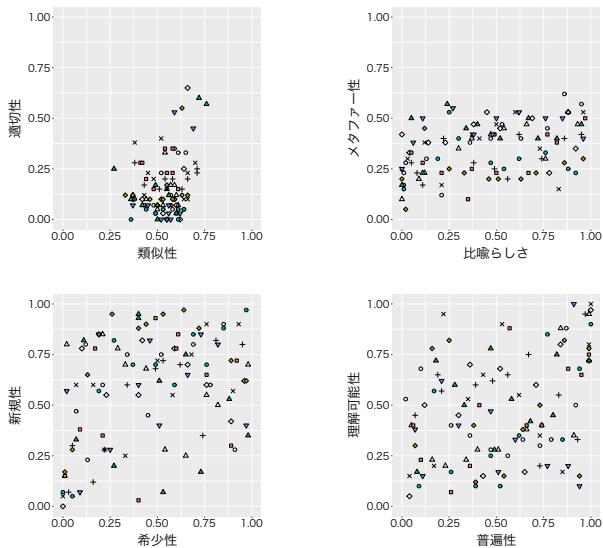


図2：内部スコアとクラウドソーシングで得られた評価の関係。

## 実験2：人手評価と内部スコアの相関

### 相関が弱くなった原因

- ▶ 評価者はわかりにくい表現のメタファー性を高く評価する傾向  
(例)「Xさんがデータを漕ぐ」  
(メタファー性：0.5, 比喩らしさ：0.05)
- ▶ 目的語と動詞が整合していない表現がある  
(例)「Xさんが株を出す」(新規性：0.85, 希少性：0.21)
- ▶ 常に自動詞として使う動詞が候補に含まれてしまっている  
(例)「\*圧政者が漁民を負ける」

## 実験3：総合スコアと相対的な好ましさ

表7：Kendall の  $\tau$  を用いた内部スコアと相対的な好ましさの比較。  
記号 \* は有意水準 5% で帰無仮説「 $\tau = 0$ 」が棄却されることを表す。

入力表現	sim	fig	rar	os
新聞社が結果を集計する	*0.64	-0.02	0.42	0.42
〇〇さんが株を購入する	0.33	-0.20	-0.07	-0.20
〇〇さんが小説を朗読する	-0.16	*0.56	-0.24	*0.69
〇〇さんが会社を設立する	-0.38	0.47	-0.38	*0.51
軍団が記録を塗り替える	-0.24	-0.24	-0.07	-0.29
〇〇さんが本を読む	0.29	0.07	-0.20	0.07
〇〇さんが水を注ぐ	0.20	-0.07	-0.38	-0.67
圧政者が漁民を襲撃する	-0.42	0.29	0.11	0.33
〇〇さんがデータを示す	-0.33	0.02	-0.33	0.22
私が村を訪れる	0.07	-0.07	0.33	-0.20

## 実験3：総合スコアと相対的な好ましさ

### 近い事例

「〇〇さんが会社を支える」はシステム、人手評価ともに最上位であった（類似性：0.60，比喻らしさ：0.94，新規性：0.77）.

### 離れている事例

「〇〇さんが水を授受する」は人手評価では最も好まれたものの、システムでは最下位であった（類似性：0.52，比喻らしさ：0.0015，新規性：1.0）.

比喻らしさと新規性は、極端に大きい、または極端に小さい値をとりやすい.

## まとめと今後の課題



## まとめ

1. メタファー生成のタスクの定義を行った
2. 新しい評価指標を導入した
3. ベースライン手法を実装した
4. クラウドソーシングで評価を行った

## 今後の課題

- ▶ （文脈を考慮した）言い換えとしての好ましさの評価
- ▶ 理解可能性に対応した内部スコアの定式化
- ▶ 他の言語や統語構造（Adj-N など）への応用

- Abe, K., K. Sakamoto, and M. Nakagawa (2006) "A computational model of metaphor generation process," in *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pp. 937–942.
- Bradley, R. A. and M. E. Terry (1952) "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, Vol. 39, No. 3/4, pp. 324–345.
- Charteris-Black, J. (2011) *Politicians and rhetoric: The persuasive power of metaphor*. Springer.
- Faruqui, M. and C. Dyer (2014) "Improving Vector Space Word Representations Using Multilingual Correlation," in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 462–471. Association for Computational Linguistics.
- Fleiss, J. L. (1971) "Measuring nominal scale agreement among many raters.," *Psychological Bulletin*, Vol. 76, No. 5, p. 378.

- Huang, E., R. Socher, C. Manning, and A. Ng (2012) “Improving Word Representations via Global Context and Multiple Word Prototypes,” in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 873–882. Association for Computational Linguistics.
- Isahara, H., F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki (2008) “Development of the Japanese WordNet,” in *Proceedings of the Sixth International Conference on Language Resources and Evaluation*.
- Kawahara, D. and S. Kurohashi (2006a) “Case Frame Compilation from the Web using High-Performance Computing,” in *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. European Language Resources Association.

- (2006b) “A Fully-Lexicalized Probabilistic Model for Japanese Syntactic and Case Structure Analysis,” in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*.
- Kawahara, D., S. Kurohashi, and K. Hasida (2002) “Construction of a Japanese Relevance-tagged Corpus,” in *Proceedings of the Third International Conference on Language Resources and Evaluation*. European Language Resources Association.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013) “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*.
- Miller, G. A. (1995) “WordNet: a lexical database for English,” *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41.

- Miyazawa, A. and Y. Miyao (2017) “Evaluation Metrics for Automatically Generated Metaphorical Expressions,” in *Proceedings of the 12th International Conference on Computational Semantics (Short papers)*.
- Mohammad, S. M., E. Shutova, and P. D. Turney (2016) “Metaphor as a Medium for Emotion: An Empirical Study,” in *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics (\*Sem)*, Berlin, Germany.
- Shutova, E., S. Teufel, and A. Korhonen (2013) “Statistical Metaphor Processing,” *Computational Linguistics*, Vol. 39, No. 2.
- Steen, G. J., A. G. Dorst, J. B. Herrmann, A. Kaal, T. Krennmayr, and T. Pasma (2010) *A Method for Linguistic Metaphor Identification: From MIP to MIPVU*. John Benjamins Publishing.

## 参考文献 V

- Tsvetkov, Y., E. Mukomel, and A. Gershman (2013) “Cross-Lingual Metaphor Detection Using Common Semantic Features,” in *Proceedings of the First Workshop on Metaphor in NLP*, pp. 45–51. Association for Computational Linguistics.
- Tsvetkov, Y., L. Boytsov, A. Gershman, E. Nyberg, and C. Dyer (2014) “Metaphor Detection with Cross-Lingual Model Transfer,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 248–258. Association for Computational Linguistics.
- Wilson, M. (1988) “MRC psycholinguistic database: Machine-usable dictionary, version 2.00,” *Behavior Research Methods, Instruments, & Computers*, Vol. 20, No. 1, pp. 6–10.
- 宮澤彬・吉田奈央・宮尾祐介 (2016) 「日本語メタファーコーパス作成のためのガイドライン」, 『言語処理学会第 22 回年次大会発表論文集』, 言語処理学会.

岩山真・徳永健伸・田中穂積 (1991) 「比喻を含む言語理解における顕現性の役割」, 『人工知能学会誌』, 第 6 巻, 第 5 号, 674-681 ページ.