Computational Exploration of Metaphor Comprehension Processes

Akira Utsumi (utsumi@se.uec.ac.jp)
Department of Systems Engineering, The University of Electro-Communications
1-5-1 Chofugaoka, Chofushi, Tokyo 182-8585, Japan

Abstract

A central question in metaphor research is what processes are involved in metaphor comprehension, especially which of comparison and categorization processes governs metaphor comprehension. In this paper, I attempt to provide a computational solution to this problem using comparison and categorization algorithms based on word vectors in a multidimensional semantic space constructed by latent semantic analysis. These algorithms receive word vectors for the topic and the vehicle of a metaphor and compute a vector for the metaphorical meaning. The resulting vectors can be evaluated on the degree to which they mimic human interpretation of the same metaphor. Using this simulation framework, I tested five competing views of metaphor comprehension: the categorization view (Glucksberg, 2001), the comparison view (Gentner, 1983) and three hybrid views — the conventionality view (Bowdle & Gentner, 2005), the aptness view (Jones & Estes, 2005) and the interpretive diversity view (Utsumi & Kuwabara, 2005) — which claim that vehicle conventionality, aptness and interpretive diversity, respectively, determine a shift between both processes. The simulation result was that the interpretive diversity view outperformed the other four views on two different measures. This result can be seen as computational evidence in favor of the interpretive diversity view.

Keywords: Computational modeling; Latent semantic analysis (LSA); Metaphor comprehension

Introduction

How metaphors are psychologically comprehended is one of the most central topics for metaphor research on which a considerable number of studies have been made. Nevertheless, these studies are divided on this issue; some studies (e.g., Gentner, 1983; Gentner, Bowdle, Wolff, & Boronat, 2001) have proposed that metaphors are processed as *comparisons* or analogical mappings, while others (e.g., Glucksberg, 2001; Glucksberg & Keysar, 1990) have argued that metaphors are processed as *categorizations*.

The comparison view by Gentner and her colleagues (Gentner, 1983; Gentner et al., 2001) argues that metaphors (and analogies) are processed as comparisons consisting of a process of structural alignment between representations of the topic and the vehicle followed by a process of projection of aligned features or relations into the topic. For example, in comprehending the metaphor "A rumor is a virus", two concepts *rumor* and *virus* are aligned, salient alignments such as ones about contagion or infection prevention are found, and features and relations inferred from such alignments are projected into the topic. Note that the initial alignment process is symmetric (i.e., the products of the process do not change even if the topic and the vehicle are reversed), while the subsequent projection process is asymmetric (i.e., directional). Hence, the intuition that the reversed metaphor "A virus is a rumor" seems meaningless can be attributed to the projection process, not to the alignment process.

On the other hand, the categorization view by Glucksberg and his colleagues (Glucksberg, 2001; Glucksberg & Keysar, 1990) claims that metaphors are seen as categorization (i.e., class-inclusion) statements expressing that the topic is a member of an abstract superordinate category exemplified by the vehicle. For example, the metaphor "My job is a jail" is comprehended so that the topic *my job* is categorized as an ad hoc category like "unpleasant and confining things" to which the vehicle *jail* typically belongs. Note that the topic also plays an important role in metaphor comprehension; it constrains dimensions by which it can be characterized. In the case of the above metaphor, *my job* facilitates attribution of features related to tasks and jobs, while blocking out irrelevant features such as ones related to jail building.

Very recent studies have tried to reconcile these two opposite views of metaphor comprehension into a coherent metaphor theory. However, they disagree on how both views are reconciled, in other words, what property of metaphor determines a shift between both processes. Bowdle and Gentner (2005) claim that it is vehicle conventionality that determines such shift; their career of metaphor theory argues that, although metaphors are basically processed as comparisons, conventional metaphors are processed as categorizations by accessing stored categories, which are conventionalized by repeated figurative use. Jones and Estes (2005, in press) argue against the career of metaphor view and advocated that metaphor aptness mediates both processes by empirically demonstrating that apt metaphors were more likely to be processed as categorizations than less apt metaphors. Glucksberg and Haught (in press) also reported that novel but apt metaphors were easy to comprehend in categorization form than in comparison form, and concluded that the aptness or the quality of metaphors determines the choice of comprehension strategy.

Against these views, Utsumi and Kuwabara (2005) and Utsumi (2006) claim that *interpretive diversity* determines whether metaphors are processed as comparisons or categorizations. Interpretive diversity is a measure of the semantic richness of literal or figurative utterances including metaphors; it is high to the extent that more features constitute the utterance meaning and that their relative saliences are more evenly distributed. The interpretive diversity view then argues that diverse metaphors are comprehended by the process of categorization, but less diverse metaphors with only a few features require the process of comparison because the process of categorization is often expected to attribute many features of a category to the members of that category.

The purpose of this paper is to provide a convincing answer to the question of which of these metaphor views is most plausible by means of computer simulation. For this purpose, this study presents a computational model of comparison and categorization processes using a semantic space constructed by latent semantic analysis (LSA) (Landauer & Dumais, 1997). This study then examines how well a computational model embodying each metaphor view mimics human comprehension by comparing the interpretations of metaphors obtained by the computer simulation with human interpretations of the same metaphors obtained in a psychological experiment (Utsumi, 2005). The metaphor view that achieves the best simulation performance can be seen as the most plausible view.

My study essentially differs from other computational studies of metaphors (e.g., Fass, 1991; Martin, 1992; Thomas & Mareschal, 2001) in that they did not test the validity of their models in a systematic or exhaustive way, nor did they make a new contribution to the psychological or cognitive theory of metaphor. Kintsch (2000) proposes a computational model of metaphor comprehension based on LSA. His predication algorithm is also used in this study as a model of categorization, but he did not test its psychological validity as a model of metaphor comprehension. In addition, his study does not allow for the fact that some metaphors are comprehended as comparisons. Lemaire and Bianco (2003) also employ LSA to develop a computational model of referential metaphor comprehension, but they do not address how well it mimics human interpretations; they only showed that it mimics processing time difference between when supporting context is provided and when it is not provided. Moreover, their model is theoretically less well motivated.

Computational Model

LSA Model

LSA is a method for automatically constructing a high-dimensional semantic space from the statistical analysis of huge corpus of written texts. LSA was originally proposed as a document indexing technique for automatic information retrieval, but several studies (Landauer & Dumais, 1997) have shown that LSA successfully mimics many human behaviors associated with semantic processing.

The basic idea behind LSA is that semantically similar words are likely to occur in semantically similar paragraphs and semantically similar paragraphs are likely to include semantically similar words. This mutual relation is represented by a word-paragraph matrix, and the dimension of the matrix is reduced, often to 200-400, by means of a linear algebra technique known as singular value decomposition.

In a constructed semantic space, each word x is represented as a vector $\boldsymbol{v}(x)$. Hence similarity $sim(x, y)$ between words x and y can be computed as the cosine $\cos(\boldsymbol{v}(x), \boldsymbol{v}(y))$ of the angle formed by two word vectors. For example, using a semantic space of Japanese words used in this paper, similarity between *computer* ("*konpyuta*" in Japanese) and *Windows* ("*uindouzu*" in Japanese; Microsoft's OS) is computed as .63, while similarity between *computer* and *window* ("*mado*" in Japanese; glass in the wall) is computed as –.02.

Comparison and Categorization Models

In an LSA model, a vector representation $\boldsymbol{v}(s)$ of a piece of text s (e.g., phrase, clause, sentence, paragraph) consisting of constituent words $w_1, \cdots, w_n$ can be defined as a function $f(\boldsymbol{v}(w_1), \cdots, \boldsymbol{v}(w_n))$.[1] In accordance with this formalization, metaphor comprehension is modeled as computation of a vector $\boldsymbol{v}(M) = f(\boldsymbol{v}(w_T), \boldsymbol{v}(w_V))$ representing the meaning of a metaphor M with the topic w_T and the vehicle w_V.

Then what we have to think next is what function is appropriate for better simulating the target behavior, in this case the categorization process and the comparison process. In the following description, I use the phrase "n neighbors of a word x" to refer to words with n highest similarity to x, and denote a set of n neighbors of x by $N_n(x)$.

Comparison The algorithm of computing a metaphor vector $\boldsymbol{v}(M)$ by the process of comparison is as follows.

1. Compute a set of k words (i.e., alignments between the topic w_T and the vehicle w_V) as $N_i(w_T) \cap N_i(w_V)$ by incrementing i by 1 until $|N_i(w_T) \cap N_i(w_V)|$ reaches k.
2. Compute a metaphor vector $\boldsymbol{v}(M)$ as the centroid of $\boldsymbol{v}(w_T)$ and k vectors computed at Step 1.

Step 1 corresponds to the alignment process and thus it is symmetric, while Step 2 corresponds to the projection process and thus it is asymmetric.

Categorization The algorithm of computing a metaphor vector $\boldsymbol{v}(M)$ by the process of categorization is as follows.

1. Compute $N_m(w_V)$, i.e., m neighbors of the vehicle w_V.
2. Selects k words with the highest similarity to the topic w_T from $N_m(w_V)$.
3. Compute a vector $\boldsymbol{v}(M)$ as the centroid of $\boldsymbol{v}(w_T)$, $\boldsymbol{v}(w_V)$ and k vectors selected at Step 2.

This algorithm is identical to Kintsch's (2000) predication algorithm. As Kintsch suggests, this algorithm embodies the categorization view in that a set of k words characterizes an abstract superordinate category exemplified by the vehicle.

Simulation Experiment

Method

Human experiment The materials used in the experiment were 40 Japanese metaphors of the form "An X is a Y". They were created from 10 groups each of which consisted of two topic words and two vehicle words. For example, from the two topics *life* ("*jinsei*") and *love* ("*ai*") and the two vehicles *journey* ("*tabi*") and *game* ("*gemu*"), the following four metaphors were created: "Life is a journey" ("*Jinsei ha tabi da*"), "Life is a game" ("*Jinsei ha gemu da*"), "Love is a journey" ("*Ai ha tabi da*") and "Love is a game" ("*Ai ha gemu*

[1] This formalization can be easily extended to involve contextual effects on semantic processing; the vector $\boldsymbol{v}(s)$ can be defined as $f(\boldsymbol{v}(w_1), \cdots, \boldsymbol{v}(w_n), \boldsymbol{v}(c))$ where $\boldsymbol{v}(c)$ represents a vector for context of s. This paper, however, does not address contextual effects of metaphor comprehension, because examining the mechanism of metaphor comprehension without context accentuates processing difference, as the existing empirical studies have done.

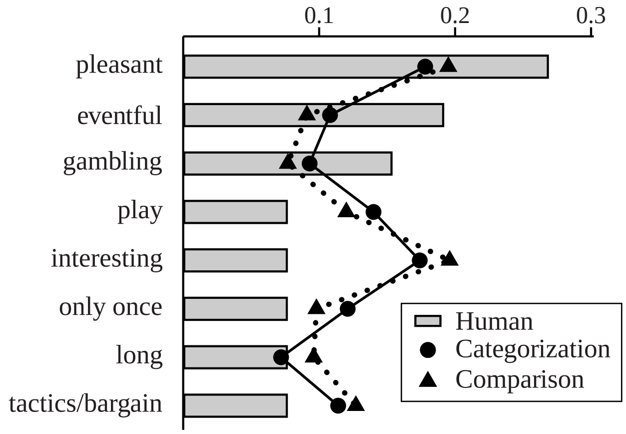

Figure 1: "Life is a game" metaphor

da"). Topic and vehicle words were selected from an experimental study on Japanese metaphor (Kusumi, 1987) and a list of words frequently used for Japanese metaphors (Nakamura, 1995).

Eighty undergraduate students of Japan Women's University, who were all native speakers of Japanese, were assigned to 10 metaphors that shared neither vehicles nor topics. They were asked to consider the meaning of each metaphor and to list three or more features of the topic that were being described by the vehicle of the metaphor. For the listed features of each metaphor, closely related words or phrases were combined into the same feature, and any feature listed by only one participant was dropped. For each feature w_i of a metaphor M, the degree of salience $sal(w_i, M)$ is then assessed as the number of participants who listed that feature, i.e., the number of tokens. These features were used as landmarks with respect to which model's interpretation and human interpretation were compared for evaluation. For example, as shown in the bar graph of Figure 1, eight features were listed for the metaphor "Life is a game", and the feature *pleasant* had the highest salience, i.e., the number of participants who listed it was largest.

Interpretive diversity of each metaphor M was then calculated as Shannon's entropy H defined by the following formula (Utsumi, 2005).

$$H = -\sum_{i=1}^{n} p_i \log p_i \tag{1}$$

$$p_i = \frac{sal(w_i, M)}{\sum_{j=1}^{n} sal(w_j, M)} \tag{2}$$

For example, the interpretive diversity of the metaphor "Life is a game" in Figure 1 is calculated as 3.13, given that the bar length for a feature w_i corresponds to p_i. The mean interpretive diversity across 40 metaphors was 3.01 (SD = 0.42).

For vehicle conventionality and metaphor aptness, an additional 144 Japanese undergraduate students at the University of Electro-Communications were recruited and assigned to 10 metaphors. Half of them were asked to rate how apt each metaphor was on a 7-point scale ranging from 1 (*not at all apt*) to 7 (*extremely apt*), and the other half of them were asked to rate how conventional the most salient meaning of each metaphor (e.g., *pleasant* for "Life is a game") was as an alternative sense of the vehicle term (e.g., *game*) on a scale of 1 (*very novel*) to 7 (*very conventional*). These ratings were then averaged across participants for each metaphor. The mean aptness rating across 40 metaphors was 3.70 (SD = 1.07) and the mean conventionality rating was 4.46 (SD = 1.19).

Computer simulation The semantic space used in the simulation was based on a Japanese corpus of 251,287 paragraphs containing 53,512 different words, which came from a CD-ROM of Mainichi newspaper articles (4 months) published in 1999. The number of dimensions of the semantic space was set to 300, and thus all words were represented as 300-dimensional vectors.

In the computer simulation, for each of the 40 metaphors, two metaphor vectors were computed: one by the comparison algorithm and another by the categorization algorithm. After that, for all the features $w_i, \cdots, w_n$ listed for that metaphor M in the human experiment, similarity to the metaphorical meaning $sim(w_i, M)$ was computed separately for two metaphor vectors. Features with higher similarity to the metaphorical meaning can be seen as more relevant to the interpretation of the metaphor. In Figure 1, for example, the word *pleasant* has the highest similarity to the metaphor vector by categorization, while the word *interesting* has the highest similarity to the metaphor vector by comparison.

Evaluation measures To evaluate the ability of the model to mimic human interpretations, I use the following measures.

- *Kullback-Leibler divergence (KL-divergence)*:

$$D = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i} \tag{3}$$

$$q_i = \frac{sim(w_i, M) - \min_x sim(x, M)}{\sum_{j=1}^{n} \{sim(w_j, M) - \min_x sim(x, M)\}} \tag{4}$$

It measures how well a model simulates the salience distribution of features relevant to human interpretation, or in other words, the degree of dissimilarity between human interpretation and computer's interpretation. Hence lower divergence means better performance. In Figure 1, for example, KL-divergence between the salience distribution of human interpretation and the similarity distribution of computer interpretation is 0.190 for the categorization model and 0.227 for the comparison model, thus suggesting that, in this case, the categorization model better mimics human interpretation than the comparison model.

- *Spearman's rank correlation*:

$$r = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n^3 - n} \tag{5}$$

$$d_i = rank(sim(w_i, M)) - rank(sal(w_i, M)) \tag{6}$$

It measures how strongly the computed similarity of relevant features is correlated with the degree of salience of those features. A higher correlation means that the model yields better performance. In Figure 1 the categorization model yields a higher correlation ($r = .21$) than the comparison model ($r = -.07$), which again indicates that the categorization model is superior to the comparison model.

Table 1: KL-divergences and rank correlations of the categorization model and the comparison model

	All metaphors[a]	Interpretive diversity		Conventionality		Aptness	
		High[b]	Low[c]	High[d]	Low[e]	High[f]	Low[g]
KL divergence							
Categorization (m=250, k=5)	**0.260**	$\boxed{\mathbf{0.185}}$	0.344	$\boxed{\mathbf{0.295}}$	**0.214**	$\boxed{\mathbf{0.275}}$	**0.248**
Comparison (k=3)	0.270	0.219	$\boxed{\mathbf{0.327}}$	0.310	$\boxed{0.216}$	0.283	$\boxed{0.260}$
Rank correlation							
Categorization (m=250, k=7)	**0.222**	$\boxed{\mathbf{0.237}}$	0.206	$\boxed{\mathbf{0.139}}$	**0.334**	$\boxed{\mathbf{0.262}}$	**0.189**
Comparison (k=5)	0.197	0.154	$\boxed{\mathbf{0.244}}$	0.122	$\boxed{0.298}$	0.252	$\boxed{0.152}$

Note. Boldfaces indicate that the corresponding model (i.e., categorization model or comparison model) achieved better performance over the other model. Boxed numbers indicate that the hybrid view in question predicts that the corresponding model should achieve better performance. [a]n=40. [b]n=21. [c]n=19. [d]n=23. [e]n=17. [f]n=18. [g]n=22.

Result

For each of the 40 metaphors, KL-divergences and rank correlations were computed using the metaphor vector by the categorization model with any of the combination of $m = 100, 150, \cdots, 500$ and $k = 1, \cdots, 10$, and using the metaphor vector by the comparison model with $k = 1, \cdots, 10$. These values were then averaged across metaphors. Concerning KL-divergence, the categorization model achieved the best performance when $m = 250$ and $k = 5$, and the comparison model did the best performance when $k = 3$. Concerning rank correlation, the combination of $m = 250$ and $k = 7$ was optimal for the categorization model and $k = 5$ was optimal for the comparison model. Table 1 lists mean divergences and correlations calculated using these optimal parameters.

As the second column of Table 1 shows, the categorization model, on average, outperformed the comparison model on both measures. It suggests that the categorization view may be more plausible than the comparison view.

Furthermore, in order to examine the plausibility of the three hybrid views, I divided the 40 metaphors into two groups according to their mean interpretive diversity (i.e., high-diversity and low-diversity), their mean conventionality rating (i.e., high-conventionality and low-conventionality), or their mean aptness rating (i.e., high-aptness and low-aptness). And I calculated mean KL-divergences and mean rank correlations for each of these groups, which are shown in the third through the last columns of Table 1.

The simulation result concerning interpretive diversity was that for high-diversity metaphors the categorization model yielded better performance (i.e., lower divergence and higher correlation) than the comparison model, while for low-diversity metaphors the comparison model achieved better performance. This result is fully consistent with the interpretive diversity view (Utsumi & Kuwabara, 2005; Utsumi, 2006). (Note that boxed boldfaces are the sign of consistency between the simulation result and the theoretical prediction.) In addition, Figures 2 and 3 show that this simulation result is general, not specific to the particular value of the parameters. For example, Figure 2 (a) illustrates that, in the case of high-diversity metaphors, the comparison model (denoted by filled triangles connected by a dotted line) showed higher KL-divergence and thus worse performance than the categorization model (denoted by other small figures), regardless of values of the parameters m and k. On the other hand, Figure 2 (b) shows the reverse pattern; the comparison model achieved lower divergence and thus better performance than the categorization model for any combinations of m and k

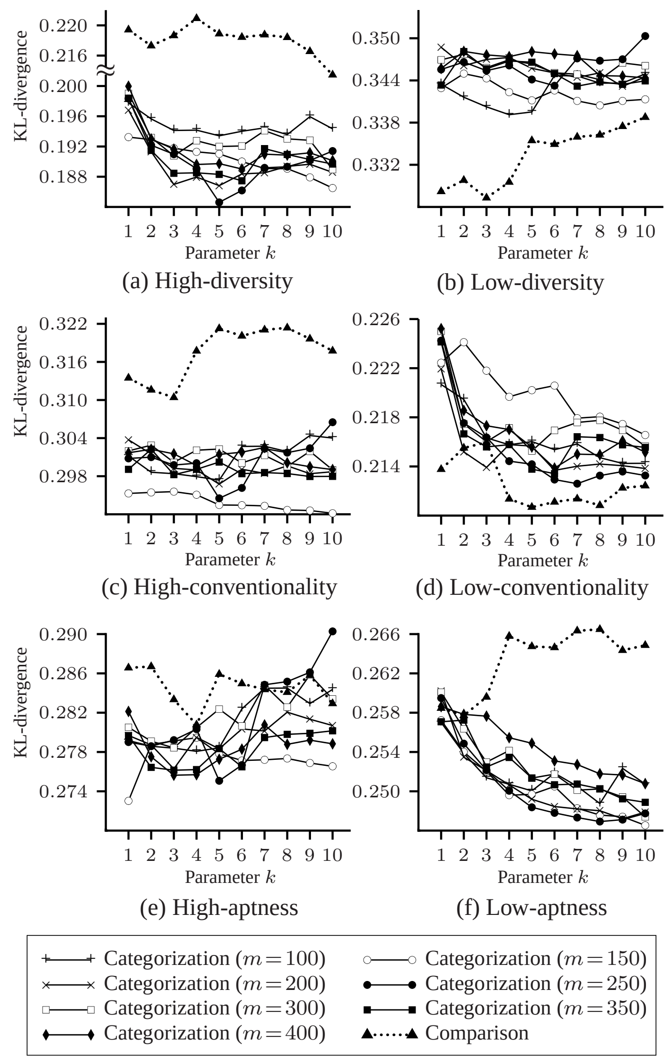

Figure 2: Simulation results in terms of Kullback-Leibler divergence for various values of the parameters m and k

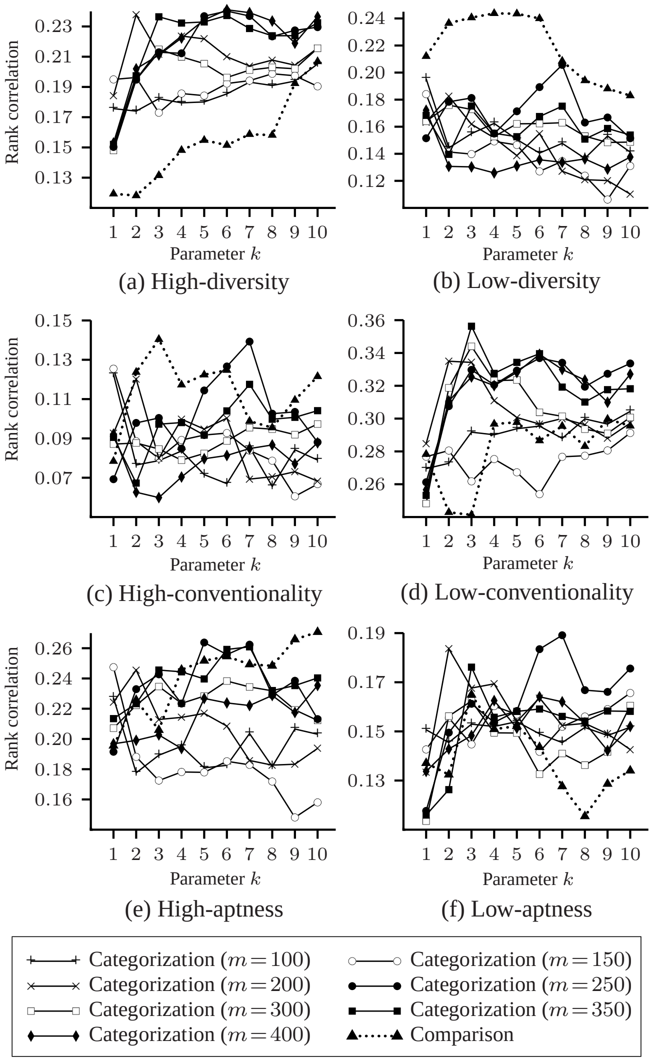

Figure 3: Simulation results in terms of rank correlation for various values of the parameters m and k

values. Figure 3(a)-(b) shows that this result holds true for rank correlations.

Simulation results concerning vehicle conventionality and aptness are not consistent with the theoretical predictions. The conventionality view and the aptness view predict that the process of comparison is required for comprehending novel metaphors and less apt metaphors, respectively, but as the fifth through the last columns of Table 1 show, the categorization model outperformed the comparison model even when metaphors were less conventional or less apt. Figures 2 and 3 also shows that for almost all the combinations of parameters m and k, the categorization model yielded lower divergence for low-aptness metaphors (Figure 2(f)) and higher correlation for low-conventionality metaphors (Figure 3(d)) or low-aptness metaphors (Figure 3(f)), which contradicts the predictions. What is worse is that, as Figures 3(c) and 3(e) show, many rank correlations of the categorization model were lower than those of the comparison model when conventional or apt metaphors were considered.

To quantitatively compare the performance among the five

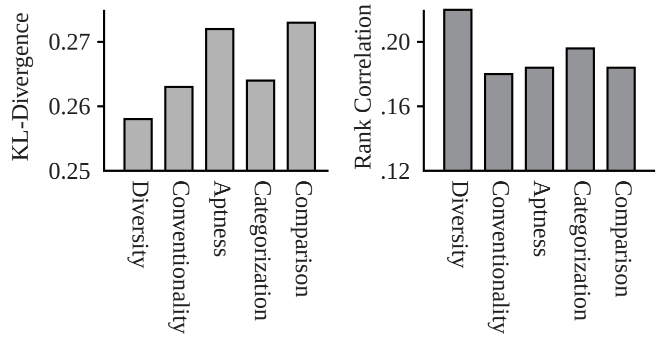

Figure 4: Comparison among the competing metaphor views

competing views, I calculated a mean KL-divergence and rank correlation for each hybrid view by summing up the scores of the categorization model for metaphors that are predicted to be processed as categorizations by that view and the scores of the comparison model for those that are predicted to be processed as comparison. For example, the mean KL-divergence of the interpretive diversity view was calculated from the divergences of the categorization model for high-diversity metaphors and the divergences of the comparison model for low-diversity metaphors.

Figure 4 shows mean divergences and correlations of the five competing views that are averaged over the parameter k ($m = 250$). The interpretive diversity view achieved the best performance on both measures, thus suggesting that the interpretive diversity view is the most plausible theory of metaphor comprehension. The conventionality view was the second-best one on the measure of divergence, but it was the worst on the measure of rank correlation. The aptness view is much less plausible given that its performance fell below the performance of the categorization view on both measures.

Discussion

The LSA simulation result presented in this paper can be seen as computational evidence in favor of the interpretive diversity view, but here it is worth discussing the validity of the simulation result in more detail.

An important issue to discuss is the psychological plausibility of the categorization model and the comparison model that produced the simulation result. In general, a categorization statement "An X is a Y" is processed so that, due to default inheritance, features characterizing *Y-ness* are highlighted unless they are obviously irrelevant to X, and at the same time other salient features of X are downplayed. Kintsch (2001) showed with many examples that the predication algorithm (i.e., the categorization model in this paper) works in this way. For example, the vector for "a pelican is a bird" computed by this algorithm becomes more similar to the features related to the *birdness* such as *sing beautifully* than the original vector of *pelican*, and less similar to irrelevant features to the *birdness* such as *eat fish* and *sea*. On the other hand, a comparison statement "An X is like a Y" is reasonably assumed to be comprehended so that only the features shared by X and Y are highlighted without other *Y-ness* features being highlighted. Indeed the comparison algorithm proposed in this paper works better than the categorization algorithm for literal comparison statements. For example, as Figure 5 shows, shared features such as *swim* and *fish* are more similar

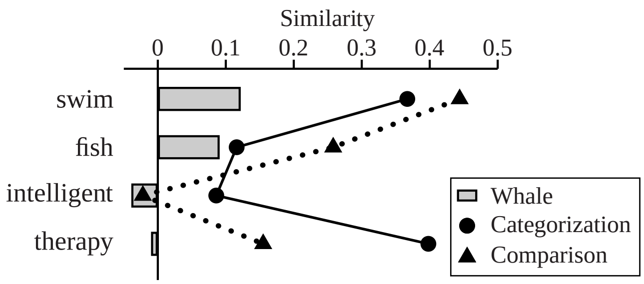

Figure 5: Literal comparison “A whale is like a dolphin”

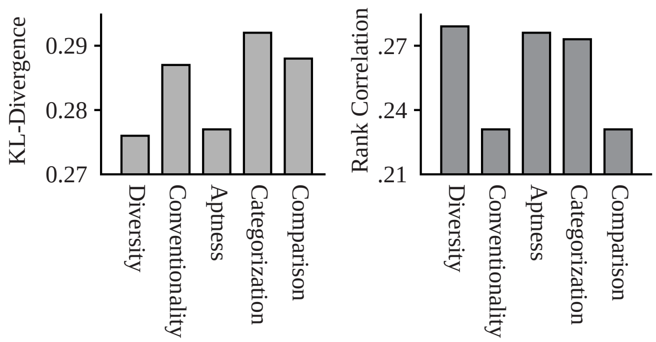

Figure 6: Simulation results by the different semantic space

to the vector for “a whale is like a dolphin” computed by the comparison algorithm than to the vector by the categorization algorithm. In contrast, *dolphinness* features like *intelligent* and *therapy* which are not shared by whales are less similar to the vector by the comparison algorithm than to the vector by the categorization algorithm.

Another issue on the plausibility of the model is the generality of the simulation result, i.e., whether other LSA semantic spaces derived from different corpora produce the same simulation result. To address this issue, I constructed a different 300-dimensional semantic space from a corpus of Japanese famous novels “Shincho Bunko No 100 Satsu” and conducted the same simulation experiment. Figure 6 depicts the simulation result in the same way as in Figure 4. As in the case of the newspaper corpus, the interpretive diversity view again achieved the best performance on both measures among the five views when mean divergences and correlations were calculated using the optimal values of the parameters.

These consistent results strengthen the validity of the conclusion of this paper that the interpretive diversity view is the most plausible theory of metaphor comprehension. My research group is also trying to apply this computational experimentation methodology to metaphor-simile distinction and other types of metaphors like predicative metaphors and adjective-noun metaphors.

Acknowledgments

This research was supported by Grant-in-Aid for Scientific Research(C) (No.17500171) from Japan Society for the Promotion of Science.

References

Bowdle, B., & Gentner, D. (2005). The career of metaphor. *Psychological Review*, *112*(1), 193–216.

Fass, D. (1991). Met*: A method for discriminating metonymy and metaphor by computer. *Computational Linguistics*, *17*(1), 49–90.

Gentner, D. (1983). Structure mapping: A theoretical framework for analogy. *Cognitive Science*, *7*, 155–170.

Gentner, D., Bowdle, B., Wolff, P., & Boronat, C. (2001). Metaphor is like analogy. In D. Gentner, K. Holyoak, & B. Kokinov (Eds.), *Analogical mind: Perspectives from cognitive science* (pp. 199–253). MIT Press.

Glucksberg, S. (2001). *Understanding figurative language: From metaphors to idioms*. Oxford University Press.

Glucksberg, S., & Haught, C. (in press). On the relation between metaphor and simile: When comparison fails. *Mind & Language*.

Glucksberg, S., & Keysar, B. (1990). Understanding metaphorical comparisons: Beyond similarity. *Psychological Review*, *97*, 3–18.

Jones, L., & Estes, Z. (2005). Metaphor comprehension as attributive categorization. *Journal of Memory and Language*, *53*, 110–124.

Jones, L., & Estes, Z. (in press). Roosters, robins, and alarm clocks: Aptness and conventionality in metaphor comprehension. *Journal of Memory and Language*.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, *7*(2), 257–266.

Kintsch, W. (2001). Predication. *Cognitive Science*, *25*(2), 173–202.

Kusumi, T. (1987). Effects of categorical dissimilarity and affective similarity between constituent words on metaphor appreciation. *Journal of Psycholinguistic Research*, *16*, 577–595.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lemaire, B., & Bianco, M. (2003). Contextual effects on metaphor comprehension: Experiment and simulation. In *Proceedings of the 5th International Conference on Cognitive Modeling (ICCM’2003)*.

Martin, J. (1992). Computer understanding of conventional metaphoric language. *Cognitive Science*, *16*, 233–270.

Nakamura, A. (1995). *Hiyu hyogen jiten* [Japanese dictionary of metaphorical expressions]. Tokyo: Kadokawa Shoten.

Thomas, M., & Mareschal, D. (2001). Metaphor as categorization: A connectionist implementation. *Metaphor and Symbol*, *16*(1 & 2), 5–27.

Utsumi, A. (2005). The role of feature emergence in metaphor appreciation. *Metaphor and Symbol*, *20*(3), 151–172.

Utsumi, A. (2006). *Interpretive diversity explains metaphor-simile distinction.* Manuscript submitted for publication.

Utsumi, A., & Kuwabara, Y. (2005). Interpretive diversity as a source of metaphor-simile distinction. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 2230–2235).