Number of shared topic-vehicle significant features affects speakers' preference for metaphorical expressions

Oka, Ryunosuke; Kusumi, Takashi

Number of shared topic-vehicle significant features affects speakers' preference for metaphorical expressions

Ryunosuke Oka[1*], Takashi Kusumi[2].

[1] Information Technology R&D Center, Mitsubishi Electric Corporation, Kamakura, Kanagawa 247-8501, Japan.

[2] Graduate School of Education, Kyoto University, Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan.

[*]Corresponding author at: Mitsubishi Electric Corporation, Kamakura, Kanagawa 247-8501, Japan.

E-mail address: Qualia1006@gmail.com (R. Oka)

Abstract

This study examined whether the number of shared topic-vehicle significant features affects speakers' preference for the use of metaphorical rather than literal expressions. Across five experiments, participants were asked to choose one expression that best paraphrased a given sentence from a list of options. The results of Experiments 1 and 5 showed that participants' choice of metaphorical expression increased with greater numbers of shared topic-vehicle significant features in a given sentence. In Experiments 2 and 4, we found that the effect of the number of unshared features was smaller than that of shared significant features. Experiment 3 replicated the findings of Experiment 2 when metaphors were replaced with similes. Our results suggest that the number of topic-attributed features affects participants' preference in the use of metaphorical expressions. Our results support the fundamental tenets of the inexpressibility hypothesis in the context of metaphor form preference.

Keywords: metaphor; shared topic-vehicle significant features; metaphorical versus literal preference

Number of shared topic-vehicle significant features affects speakers' preference for metaphorical expressions

In our daily lives, we communicate many aspects of events using literal expressions. In this context, "literal expression" refers to an expression whose meaning can be directly inferred from the meanings of its components (Gibbs, Buchalter, Moise, & Farrar, 1993). For example, we often use literal expressions such as, "The party last night was exciting." Alternatively, we sometimes use metaphorical expressions to communicate our ideas. In this context, "metaphorical expression" refers to an expression in which one word (the topic) is understood in terms of a second word (the vehicle) that belongs to a different category from the first word (Gibbs et al., 1993). For example, "The party last night was a roller coaster," is a metaphorical expression. Cameron (2003) reported that 50 metaphors were used per 1000 words in ordinary conversations. Thus, we can use both literal and metaphorical expressions to describe the excitement of last night's party.

This discussion raises a question: When do we prefer literal expression and when do we prefer metaphorical expression? Though some theories have tried to explain why we prefer metaphorical forms in some situations (Ortony, 1975) and empirical investigations of theories (e.g., Fainsilber & Ortony, 1987; Fussell & Krauss, 1989a), the context in which we prefer metaphorical expressions needs further investigation.

The present study examined whether the number of significant features shared by both the topic and the vehicle (i.e., shared topic-vehicle significant features) affects speakers' preference for metaphorical rather than literal expression. The number of shared topic-vehicle significant features is determined by counting the number of features semantically shared with the vehicle that are attributed to the topic. For example, in the sentence, "The party last night was fun," only one feature (i.e., fun) is attributed to the topic. In another sentence, "The party last night was fun, vigorous, and flowing," three features (i.e., fun, vigorous, flowing) are

attributed to the topic. The relationship between the number of shared topic-vehicle significant features and form preference (metaphorical or literal) remains unclear in current research. Therefore, this study explored how the number of features attributed to a topic—especially when meaning is shared between the topic- and vehicle-concepts and captures an important property of the vehicle-concept—influences the preference for metaphorical over literal expression.

When we prefer metaphor: The inexpressibility hypothesis and compactness hypothesis

When do we prefer literal expression and when do we prefer metaphorical expression? Ortony (1975) proposed two hypotheses regarding this question: the inexpressibility hypothesis and compactness hypothesis.

The inexpressibility hypothesis posits that some topics (or topic-related features) cannot be articulated using literal expression alone. For example, when we read a sentence such as, “The thought slipped my mind like a squirrel behind a tree,” it is difficult for readers to translate the ideas (e.g., swiftness, suddenness, “ungraspableness”) that this metaphor evokes into literal language. Some studies (Fainsilber & Ortony, 1987; Fussell & Moss, 1998; Williams-Whitney et al., 1992) have tested this hypothesis. In these studies, participants were asked to recall and describe a past emotional event (e.g., the happiest event they had experienced). Participants were required to describe these events based on the group that they had been assigned to. Specifically, participants were assigned to one of two description type-conditions. In the “Feelings condition,” participants were asked to describe an emotion they had felt during their recalled events, and in the “Behaviors condition,” participants were asked to describe how they had behaved during their recalled events. The researchers counted the metaphors embedded in participants’ descriptions. The inexpressibility hypothesis predicted that the number of metaphors would be higher in the feeling condition than in the behavior condition; while the subjective quality of feelings is difficult to describe using literal

language, metaphorical expressions enable us to communicate these kinds of descriptions. The results of these studies, whereby more metaphors were used by participants in the feeling condition than in the behavior condition, supported the inexpressibility hypothesis. These results suggest that the frequency of metaphor use differs across topics (i.e., feelings or behaviors).

The compactness hypothesis posits that metaphorical expression enables us to communicate many features in a few words. For example, when we read a sentence such as, "My job is a jail," it transmits the idea in fewer words than when explaining the same thing using literal expression alone (e.g., "My job is tough, confining, and does not allow advancement"). Even though they did not refer to the compactness hypothesis, some studies (Fussell & Krauss, 1989a, 1989b) have provided supporting evidence for this hypothesis. In these studies, participants were asked to verbally explain abstract line drawings. Results showed that participants used more metaphorical expressions than literal expressions in their explanation (Fussell & Krauss, 1989a). Moreover, when metaphorical expressions were used in explanations, sentence lengths were shorter than when the explanation did not use metaphorical expressions (Fussell & Krauss, 1989b). These results suggest that when we explain abstract topics like line drawings, metaphorical expressions are more preferred, and that this relates to the sentence length of the explanation. As the compactness hypothesis posits, when we explain a topic with many topic-attributed features (i.e., something abstract), metaphorical expressions are preferred. The compactness hypothesis can apply not only to abstract line drawings, but to a variety of topic areas. For example, when a speaker explains a topic to a listener unfamiliar with it using analogy, the topic can be explained without redundant literal explanation (Glucksberg, 1989).

The inexpressibility hypothesis and compactness hypothesis are interrelated, and both might be conditions for using metaphorical expressions. For example, though Fussell's studies

provided supporting evidence for the compactness hypothesis in the context of abstract line drawings, it is difficult to explain the whole picture of a line drawing only by explaining each feature literally. To convey the whole picture of an abstract line drawing, metaphorical expression with a concrete vehicle may be helpful. In another example, Fainsilber and Ortony's (1989) results can be interpreted as showing that because a transient feeling has multiple features (i.e., it is difficult to explain using single literal feature), feeling was better explained using metaphorical expressions.

In summary, the inexpressibility hypothesis and compactness hypothesis suggest that we prefer metaphorical expressions when the topic has many features that have no corresponding literal expressions. Where previous studies focused on free descriptions of emotional experiences and line drawings, we focus on metaphor form preference, which has sometimes been discussed in comparison to simile form preference (Chiappe & Kennedy, 1999, 2001).

There are two merits to investigating metaphor form preference over literal expression. First, unlike studies of metaphor form preference over simile form, there have been few studies of metaphor form preference over literal expression. To our knowledge, the study most pertinent to ours is Schraw, Trathen, Reynolds, and Lapan (1988). In their Experiment 2, they examined whether lexicalization (controlled by participants' mother language: native, non-native), context (a preceding context that facilitates idiomatic meaning: context, no context), and familiarity (how often they hear and read the sentence: high, average, literal control) affect the preference for idiomatic over literal interpretations. In their study, participants were asked to paraphrase a sentence that can be interpreted both idiomatically and literally (e.g., in context condition, "The politician believed in his views./The man took a stand."). They showed the effects of context and familiarity on idiomatic preference: The more familiar the expression and the more suitable the context, the

more strongly preferred idiomatic interpretations were over literal ones. Though Schraw et al. (1988) showed that context and familiarity were important in idiomatic interpretation preference, it was still unclear whether the number of shared significant features (one form of the context investigated in this study) affected nominal metaphor form preference. Answering this question might give us some hints regarding the contexts in which we prefer (or even use) metaphor form over literal expression.

Second, by using a metaphor form preference task, we can easily control the context of the paraphrased sentence. Though previous studies clarified the contexts in which we prefer metaphor form, it is still unclear what aspects of emotional experiences (or line drawings) affect the use of metaphorical expression. Using a form preference task, we can easily control the context (e.g., the number of shared significant features attributed to the topic) and specify how each factor affects the metaphor form preference.

The Present Study

We examined whether the number of shared topic-vehicle significant features affects the preference for metaphor. We hypothesized that the higher number of message features, the greater will be the preference for metaphor use.

In this research, we conducted five experiments. Participants in each experiment were asked to choose from four options the one that best paraphrased a given sentence (e.g., “Her cat is cute, cherished, and selfish”). There were two critical options: metaphorical (i.e., “Her cat is a princess”) and literal (i.e., “Her cat is lovely”). We controlled topic-attributed features (e.g., in the aforementioned example: “cute, cherished, and selfish”) and compared the proportion of metaphorical responses between conditions.

In Experiment 1, we examined whether the number of shared topic-vehicle significant features affected preference for metaphorical over literal expression. In this experiment, we designed three conditions (i.e., one, two, and three shared significant features). In the one-

shared-significant-feature condition, participants were presented with one topic-attributed feature (e.g., “Her cat is cute”). In the two-shared-significant-feature condition, participants were presented with two topic-attributed features (e.g., “Her cat is cute and cherished”). In the three-shared-significant-feature condition, participants were presented with three topic-attributed features (e.g., “Her cat is cute, cherished, and selfish”). All these topic-attributed features were shared topic-vehicle feature(s) that were embedded in the metaphor (i.e., in this example: “Her cat is a princess”). If the shared topic-vehicle significant features affected speakers’ preference for metaphor use, the proportion of metaphorical responses should have been higher in the three-feature condition than in the one- and two-shared-significant-feature conditions.

In Experiment 2, we examined whether the number of shared and unshared features affected preference for metaphor. Even if the results of Experiment 1 supported our hypothesis, the effect of the number of topic-vehicle shared significant features on speakers’ preference for metaphorical over literal expression would remain unclear. This is because in Experiment 1, we controlled not only the number of shared topic-vehicle significant features but also the number of topic-attributable features. In Experiment 2, we examined the differential effects of these two variables. Specifically, we replaced the two-shared-significant-feature condition with a one-shared-two-unshared-feature condition. In the one-shared-two-unshared-feature condition, one feature was the same as that used in the one shared-significant-feature condition. However, there were two additional features that entailed meanings that were relevant to the topic (i.e., her cat) but irrelevant to the vehicle (i.e., princess). For example, in the sentence, “Her cat is cute, small, and round,” “cute” is a shared significant feature, whereas “small” and “round” are unshared features. Therefore, in Experiment 2, even though the number of features attributed to the topic was the same (i.e., three), the type of features attributed to the topic was different. Specifically, all features were

shared significant features in the three-shared-significant-feature condition; some features in the one-shared-two-unshared-feature condition were unshared features.

As posited by the compactness hypothesis, unshared features could activate a preference for metaphor rather than for literal form because it is difficult for a literal expression to capture many to-be-attributed features (Ortony, 1975). However, because unshared features only activate topic-related concepts, the effect of metaphor form preference must be weaker than that of shared significant features; shared significant features activate not only the topic concept but also the vehicle concept that is of significance to the topic. In other words, the proportion of metaphorical responses should be higher in the three-shared-significant-feature condition than in the one-shared-two-unshared-feature condition.

In Experiment 3, we examined whether the number of shared and unshared features affected the preference for simile. Some previous studies reported differences in the interpretation of the meanings that imbue metaphors and similes (Glucksberg, 2008; Glucksberg & Haught, 2006; Hasson, Estes, & Glucksberg, 2001; Haught, 2013). This form difference (i.e., the presence or absence of a hedge “like” or “as”) is evident even in speakers’ preference for metaphorical over literal expression use. To test this possibility, we replaced metaphor options (e.g., “Her cat is a princess”) with simile options (e.g., “Her cat is like a princess”) in this experiment. All other conditions were the same as in Experiment 2. We hypothesized that the overall results would be similar to those observed in Experiment 2 because both metaphors and similes are metaphorical expressions that encompass more types of meaning than literal expressions. In addition, we hypothesized that the overall effect of the number of shared significant features would be larger on simile use than on metaphor use. According to the literal base theory of figurative language (Chiappe & Kennedy, 2001), metaphors should include many shared significant features, whereas similes may include fewer shared significant features (Chiappe & Kennedy, 1999). In accordance with their

hypothesis, Chiappe and Kennedy (1999) showed that aptness (i.e., the extent to which a comparison manages to capture salient properties of the topic) is correlated with metaphor form preference. This result suggests that simile use requires only a few shared significant features.

In Experiment 4, we reexamined whether the number of shared and unshared features affected the preferences for metaphor. In this experiment, we used not only the list from Experiment 2 but also a counterbalanced list. In Experiments 1, 2, and 3, the one shared-significant-feature condition was always the same. Thus, the results obtained in the three experiments might have been due to the three-shared-significant-feature condition including the central property, which was not present in the one-shared-significant-feature condition. This might reflect the effect of a specific feature—for example, because some features attributed to the topic are more prototypical than others, we thus might have obtained the results in Experiment 2. A counterbalanced list would help in dealing with this problem. If we replicated the results obtained in Experiment 2 even with a counterbalanced list, this would show that it was not the differences in the feature presented in the one-feature condition but the number of shared significant features attributed to the topic that affected the preference for metaphorical over literal expressions.

In Experiment 5, similar to Experiment 4, we reexamined how the number of topic-vehicle shared features affected the preference for metaphor. In this experiment, we used not only the list from Experiment 1 but also the two counterbalanced lists. If we replicated the results of Experiment 1 with the counterbalanced list, our hypothesis would be supported more strongly.

The datasets and/or analyzed during the current studies are available from the corresponding author upon reasonable request.

Experiment 1

Method

Participants. We recruited 120 participants (68 males and 52 females) between the ages of 21 and 66 years ($M = 40.1$, $SD = 9.8$) through Crowdworks, a crowdsourcing service in Japan. All participants were recruited anonymously. Each participant was paid 100 yen as compensation. In this experiment, only people of Japanese origin were asked to participate.

Design. Experiment 1 involved a 3 (number of shared significant features: one, two, and three shared significant features) × 3 (list: list A, B, and C) experimental design. The number of shared significant features was treated as a within-participants and within-items variable; on the other hand, the list was a between-participants variable.

Stimuli. We used three types of stimuli: metaphors, shared significant features, and literal expression. Table 1 shows some examples of the stimuli used in Experiment 1.

Table 1

Examples of stimuli used in Experiment 1 (English translation with original Japanese text)

One shared significant feature	Two shared significant features	Three shared significant features	Metaphor	Literal	Nonsense-metaphor	Nonsense-literal
Her cat is cute. (彼女の子猫はかわいい)	Her cat is cute and cherished. (彼女の子猫はかわいく、大切にされている。)	Her cat is cute, cherished, and selfish. (彼女の子猫はかわいく、大切にされており、わがままだ)	princess (王女だ)	lovely (愛くるしい)	voyage (旅だ)	long (長い)
That job is tough. (あの仕事は辛い)	That job is tough and does not allow advancement. (あの仕事は辛く、逃れられない。)	That job is tough, does not allow advancement, and is confining. (あの仕事は辛く、逃れられず、閉じ込められる)	prison (牢獄だ)	hard (苦しい)	sheep (羊だ)	bright (明るい)
That butterfly is beautiful. (あの蝶は美しい)	That butterfly is beautiful and fluttering. (あの蝶は美しく、舞う)	That butterfly is beautiful, fluttering, and gorgeous. (あの蝶は美しく、舞い、華やかだ)	dancer (踊り子だ)	pretty (きれいだ)	prison (牢獄だ)	painful (苦しい)

To collect metaphor stimuli, we referred to 120 Japanese similes anthologized by

Nakamoto and Kusumi (2004). In their collection, all similes were in the following format: "NOUN is like a NOUN" ([in Japanese, "*MEISHI ha MEISHI no youda*"]. Many of these similes have been used in previous studies on metaphor (or simile) comprehension among Japanese participants (Kusumi, 1995; Nakamoto, 2003). Because we wanted to use metaphors, we changed each simile (e.g., "A smile is like a flower") to a metaphor (e.g., "A smile is a flower").

With regard to shared significant feature stimuli, we prepared three typical interpretations for each metaphor collected by Oka, Ohshima, and Kusumi (2019). In their study, participants ($N = 50$) were asked to generate a maximum of three interpretations for each of the 120 similes collected by Nakamoto and Kusumi (2004). They grouped the generated interpretations using the following steps: first, nonsense responses were excluded; second, similar interpretations were grouped using a common token; third, any token with only one response was excluded; finally, two tokens were merged into one if they shared the same feature, as listed in a Japanese dictionary.

With regard to literal stimuli, we prepared a synonym for each of the most frequently generated tokens. Most of the synonyms were identified using Japanese WordNet. To examine the validity of the literal stimuli, we conducted a pilot study in which 10 participants rated the extent to which the meaning of a paraphrased synonym was apt for the original feature on a six-point Likert scale; none of these participants participated in the main experiment. Based on the results of this pilot study, we selected 45 metaphors and 45 literal expressions from the items with ratings between the mean and −1 SD that were higher than the midpoint (3.5) as the stimulus set for the main experiment.

In summary, in the main experiment, we used 45 metaphors with one to three shared significant features per metaphor and 45 literal expressions with meanings that were perceived to be the same as the most frequently generated tokens in Oka et al.'s (2019) study.

Procedure. We collected data using the Qualtrics platform (Qualtrics, Provo, UT, available at https://www.qualtrics.com). Participants could respond to this survey only using a personal computer. First, participants read information about this study and provided informed consent. Second, participants read detailed instructions about this study; they were required to read a short sentence describing a given topic (e.g., "Her cat is cute, cherished, and selfish"). This sentence was composed of a topic (i.e., "Her cat") and some features (i.e., cute, cherished, and selfish). The participant's task was to choose an option that best paraphrased the given sentence. There were four options: (i) metaphorical ("Her cat is a princess"), (ii) literal ("Her cat is lovely"), (iii) nonsense-metaphorical ("Her cat is a voyage"), and (iv) nonsense-literal ("Her cat is long"). The first author chose nonsense-metaphor and nonsense-literal options by selecting a vehicle and literal expression that did not share any given feature with the topic. The presentation orders of the options and stimuli were randomized. Once participants had completed the questionnaire, they were debriefed about the study. This task took approximately 10 minutes to complete.

Results

We first examined nonsense-metaphor and nonsense-literal responses as filler responses. Table 2 shows the proportion of metaphorical, literal, and filler responses for each condition. As predicted, metaphorical responses increased with the number of shared significant features. To clarify the pattern of this result, we used a mixed-effects logistic regression model via the lmer program of the lme4 package (Bates, Maechler, Bolker, & Steve, 2015) in the R (version 3.3.2) environment for statistical computing (R Development Core Team, 2016). We entered the number of shared significant features as a fixed factor with one shared feature as the default level. In addition, we specified intercepts for participants and items as random factors. We first ran a model with responses coded as one of the following: metaphorical responses and non-metaphorical responses (i.e., literal and filler responses).

Both the two- versus one-shared-feature contrast (β = 1.22, SE = 0.09, z = 12.33, p < .001) and the three- versus one-shared-feature contrast (β = 1.87, SE = 0.10, z = 18.80, p < .001) yielded significant results. In addition, we ran a model that was the same as the first model, but with two shared significant features as the default level. Both the one- versus two-shared-significant-feature (β = −1.22, SE = 0.09, z = −12.33, p < .001) and the three- versus two-shared-significant-feature conditions (β = 0.65, SE = 0.08, z = 7.67, p < .001) showed significant contrasts.

Table 2

Proportion of responses across the shared-significant-feature conditions in Experiments 1, 2, 3, 4, and 5

Condition	Metaphorical	Literal	Filler
Experiment 1 (Metaphor; N = 120)			
One shared significant feature	.16	.84	.00
Two shared significant features	.31	.68	.02
Three shared significant features	.41	.57	.02
Experiment 2 (Metaphor; N = 120)			
One shared significant feature	.13	.87	.01
One shared significant and two unshared features	.22	.74	.04
Three shared significant features	.37	.61	.02
Experiment 3 (Simile; N = 120)			
One shared significant feature	.31	.68	.01
One shared significant and two unshared features	.38	.57	.04
Three shared significant features	.62	.36	.02
Experiment 4 (Experiment 2 + counterbalanced list; N = 240)			
List used in the Experiment 2			

One shared significant feature	.17	.82	.01
One shared significant and two unshared features	.23	.71	.06
Three shared significant features	.42	.56	.02
Counterbalanced list			
One shared significant feature	.18	.79	.02
One shared significant and two unshared features	.28	.64	.07
Three shared significant features	.43	.54	.03
Experiment 5 (Experiment 1 + 2 counterbalanced lists; $N = 216$)			
List used in the Experiment 1			
One shared significant feature	.18	.80	.01
Two shared significant features	.35	.65	.00
Three shared significant features	.50	.50	.01
Counterbalanced list A			
One shared significant feature	.23	.76	.00
Two shared significant features	.41	.57	.01
Three shared significant features	.43	.56	.01
Counterbalanced list B			
One shared significant feature	.19	.80	.01
Two shared significant features	.50	.48	.02
Three shared significant features	.60	.38	.02

Discussion

In Experiment 1, we observed the main effect of the number of shared topic-vehicle significant features on the speaker's preference for metaphorical over literal expression use.

This result supports our hypothesis. Interestingly, our result showed a linear trend, whereby an increase in metaphorical responses also increased shared topic-vehicle significant features. Although we only tested one-, two-, and three-feature conditions, this result suggests that the greater the number of shared topic-vehicle significant features, the more apt the use of metaphorical over literal expressions is likely to be.

In addition, we conducted a logistic mixed-effect model with a familiarity measure. Familiarity was considered an important predictor for metaphor processing in previous studies (e.g., Bowdle & Gentner, 2005; Roncero & de Almeida, 2015). Though there have been few studies to clarify the relationship between familiarity and metaphor preference compared to literal expressions, we investigated this possibility. As reported in S2, though the familiarity measure had a positive effect on metaphor preference, so did the number of shared significant features.

Furthermore, we confirmed that the number of shared significant features also affected literal preference over non-literal. This suggests that the number of shared significant features increases the metaphor preference and decreases the literal preference. Details of analysis are reported in S4.

In Experiment 2, we distinguished the effects of the numbers of shared topic-vehicle significant features and topic-attributed features. As discussed in the earlier sections of this article, we compared the one-shared-two-unshared-feature condition with the three-shared-significant-feature condition. These two conditions had the same number of topic-attributed features (i.e., three) but different numbers of shared topic-vehicle significant features. By comparing these conditions, we could separately evaluate the effects of the numbers of shared topic-vehicle significant features and topic-attributed features on speakers' preference for metaphor use.

Experiment 2

Method

Participants. We recruited 120 participants (63 male and 57 female) between the ages of 20 and 59 years (M = 39.8, SD = 8.7) in the same way as in Experiment 1. Because we recruited participants anonymously, 31 participants overlapped with those who constituted the sample used in Experiment 1. Analysis of the overlapping participants is reported in S3.

Design. Experiment 2 involved a 3 (number of shared significant features: one shared significant features, one shared significant and two unshared features, and three shared significant features) × 3 (list: list A, B, and C) design. The number of shared significant features was a within-participants and within-items variable; on the other hand, the list was a between-participants variable.

Stimuli. Similar to Experiment 1, we used three types of stimuli: metaphorical, shared significant features, and literal. We changed the shared-significant-feature stimuli in the two-shared-significant-feature condition. Specifically, instead of the two-shared-significant-feature condition, we used a one-shared-two-unshared-feature condition in Experiment 2. In this condition, there were three features (e.g., “Her cat is cute, small, and round”); among these features, one was the same as the feature used in the one-shared-significant-feature condition (i.e., cute). The others were features with meanings that were relevant to the topic (e.g., her cat) but irrelevant to the vehicle (i.e., princess). Irrelevant features were selected by the first author after consulting a dictionary; most of these irrelevant features were listed in the dictionary. The first and second authors checked and chose two irrelevant features that were to serve as unshared features for each metaphor. Table 3 shows some examples of stimuli that were used in the one-shared-two-unshared-feature condition in Experiments 2 and 3.

Table 3

Example stimuli in the one-shared-two-unshared-feature condition (English translation with original Japanese text)

Her smile is beautiful, friendly, and constant. (彼女の笑顔は美しく、友好的で、いつも絶えない)
That job is tough, progressing nicely, and nerve-wracking (あの仕事は辛く、はかどり、神経を使う)
His love is changeable, a blessing, and true. (彼の愛は移り変わり、祝福され、打ち明けられる)

Procedure. The procedure was identical to that of Experiment 1.

Results

Similar to the analytic strategy used in Experiment 1, we examined nonsense-metaphorical and nonsense-literal responses as filler responses. Table 2 shows the proportion of metaphorical, literal, and filler responses in each condition. Two emergent results merit attention. First, as in Experiment 1, the proportion of metaphorical responses was higher in the three-shared-significant-feature condition than in the one-shared-significant-feature condition. Second, and more importantly, the proportion of metaphorical responses was higher in the three-shared-significant-feature condition than in the one-shared-two-unshared-feature condition. To clarify the pattern of these results, we again used a mixed-effects logistic regression model. We entered the number of shared significant features as the fixed factor with three shared significant features as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses that were coded using the following binary variable: metaphorical and non-metaphorical responses (i.e., literal and

filler responses). Results pertaining to both the three-shared-significant-feature versus the one-shared-two-unshared-feature (β = 1.08, SE = 0.09, z = 11.89, p < .001) and the three-shared-significant-feature versus one-shared-significant-feature contrasts (β = 2.00, SE = 0.11, z = 18.96, p < .001) were significant. More importantly, (i) the three-shared-significant-feature versus one-shared-two-unshared-feature condition and (ii) the three-shared-significant-feature versus the one-shared-significant-feature condition contrasted significantly ($\chi^2(1)$ = 74.32, p < .001) via the linear Hypothesis program of the car package (Fox & Weisberg, 2019). In addition, we ran a model that was the same as the first model but with the one shared significant feature as the default level. The one-shared-two-unshared-feature versus one-shared-significant-feature contrast (β = 0.92, SE = 0.11, z = 8.62, p < .001) yielded a significant result.

Discussion

In Experiment 2, we replicated the results of Experiment 1: The three-shared-significant-feature condition showed a higher proportion of metaphorical responses than the one-shared-significant-feature condition. In addition, the one-shared-two-unshared-feature condition showed a higher proportion of metaphorical responses than the one-shared-significant-feature condition. These two results imply that the numbers of shared and unshared features affect the speaker's preference for metaphor use. More importantly, in this experiment, participants chose more metaphorical responses in the three-shared-significant-feature condition than in the one-shared-two-unshared-feature condition. Taken together, these results suggest that even though both shared features and unshared features affect speakers' preference for metaphor, the effects of the number of features were stronger for shared features than for unshared features.

In Experiment 3, we explored the effect of the numbers of shared topic-vehicle significant features and topic-attributed features on speakers' preference for simile over literal expression use. We replaced the metaphors used in Experiments 1 and 2 with similes; all the

other conditions were the same as those used in Experiment 2. If we replicated the results obtained in Experiment 2, this would indicate that topic-attributed features are the key factors that determine a speaker's preference for simile over literal expression use.

Experiment 3

Method

Participants. We recruited 120 participants (69 men and 51 women) between the ages of 20 and 64 years ($M = 39.5$, $SD = 9.0$) using the same methodology employed in Experiments 1 and 2. Because we recruited participants anonymously, 56 participants overlapped with either Experiment 1 or Experiment 2.

Design. The design was identical to that of Experiment 2.

Stimuli. We used three types of stimuli: simile, shared significant features, and literal. In contradistinction to Experiment 2, we changed metaphorical stimuli and nonsense-metaphorical stimuli: Instead of metaphorical stimuli (i.e., "NOUN is a NOUN" [in Japanese, "*MEISHI ha MEISHI da*"]), we used simile stimuli ("NOUN is like a NOUN" [in Japanese, "*MEISHI ha MEISHI no youda*"]) in Experiment 3. We also replaced nonsense-metaphorical stimuli with nonsense-simile stimuli. For example, when a participant responded to a paraphrased version of the sentence, "Her cat is cute, cherished, and selfish," four options were presented: (i) simile (i.e., "Her cat is like a princess"), (ii) literal (i.e., "Her cat is cute"), (iii) nonsense-simile (i.e., "Her cat is like a voyage"), and (iv) nonsense-literal (i.e., "Her cat is long").

Procedure. The procedure was identical to that of Experiments 1 and 2.

Results

Similar to Experiments 1 and 2, we examined nonsense-simile and nonsense-literal responses as filler responses. Table 2 shows the proportions of simile, literal, and filler

responses in each condition. Two results merit further discussion. First, similar to the results of Experiments 1 and 2, the proportion of simile responses was higher in the three-shared-significant-feature condition than the one-shared-significant-feature condition. Second, the proportion of simile responses was higher in the three-shared-significant-feature condition than in the one-shared-two-unshared-feature condition. To clarify the pattern of these results, we used a mixed-effects logistic regression model. We set the number of shared significant features as a fixed factor with three shared significant features as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model on responses that were coded using the following binary: simile responses versus non-metaphorical responses (i.e., literal and filler responses). Both the three-shared-significant-feature versus the one-shared-two-unshared-feature contrast (β = 1.38, SE = 0.08, z = 16.35, $p < .001$) and the three-shared-significant-feature versus one-shared-significant-feature contrast (β = 1.89, SE = 0.09, z = 21.26, $p < .001$) showed significant differences. As in Experiment 2, (i) the three-shared-significant-feature versus the one-shared-two-unshared-feature contrast, and (ii) the three-shared-significant-feature versus one-shared-significant-feature contrast showed a significant difference ($\chi^2(1)$ = 74.32, $p < .001$). In addition, we ran a model that was the same as the first model but with one shared significant feature as the default level. The one-shared-two-unshared-feature versus one-shared-significant-feature contrast (β = 0.50, SE = 0.08, z = 6.00, $p < .001$) also showed significant differences.

In addition, we explored whether the proportion of simile responses (Experiment 3) was higher than that of metaphorical responses (Experiment 2). To address this objective, we used a mixed-effects logistic regression model. We entered the number of shared significant features contrasts (i.e., three-shared-significant-feature vs. one-shared-two-unshared-feature and three-shared-significant-feature vs. one-shared-significant-feature contrast) and expression type (i.e., metaphor, simile) with three shared significant features and metaphors

as the default level. In addition, we specified intercepts for participants and items as random factors. The main effect of expression type (β = 1.32, SE = 0.08, z = 16.65, p < .001) was statistically significant. In addition, both the three-shared-significant-feature versus one-shared-two-unshared-feature contrast (β = 0.92, SE = 0.08, z = 11.03, p < .001) and the three-shared-significant-feature versus one-shared-significant-feature contrast (β = 1.70, SE = 0.09, z = 18.00, p < .001) showed significant differences.

Discussion

In Experiment 3, we replicated the results of Experiment 1: The three-shared-significant-feature condition showed a higher proportion of simile responses than the one-shared-significant-feature condition. In addition, the one-shared-two-unshared-feature condition showed a higher proportion of simile responses than the one-shared-significant-feature condition. These two results imply that the number of shared and unshared features affects the speaker's preference for simile use. More importantly, in this experiment, participants chose more simile responses in the three-shared-significant-feature condition than in the one-shared-two-unshared-feature condition. Taken together, these results suggest that even though both shared and unshared features affect speakers' preference for simile use, the effects of the number of features were stronger for shared features than for unshared features. Furthermore, comparisons between preferences in the use of metaphors and similes in each condition showed that participants preferred simile use to metaphor use. This result suggests that, as suggested by literal base theory of figurative language (Chiappe & Kennedy, 2001), one shared significant feature is enough to use this form.

Furthermore, as in Experiment 2, we have confirmed that the number of topic attributed features also affects literal preference over non-literal. This suggests that the number of topic attributed features increases metaphor preference and decreases literal preference. Details of analysis are reported in S4.

In Experiment 4, we replicated Experiment 2 with a counterbalanced list. If we replicated the results of Experiment 2, this would indicate that it was not the differences in the features presented in the one feature condition but rather the number of shared significant features attributed to the topic that affects the preference for metaphorical expressions over literal expressions.

Experiment 4

Method

Participants. We recruited 240 participants between the ages of 20 and 64 years (M = 40.3, SD = 8.7) using the same methodology that was employed in Experiments 1, 2, and 3. Six of the participants were excluded owing to errors in data collection. Because the participants were recruited anonymously, six participants overlapped with those who were included in Experiments 1, 2, and 3.

Design. Experiment 4 involved a 3 (number of shared significant features: one shared significant feature, one shared significant and two unshared features, and three shared significant features) × 6 (list: list A, B, C, D, E, F) design. The number of shared significant features was a within-participants and within-items variable, while the list was a between-participants variable.

Stimuli. In Experiment 4, we not only used the list from Experiment 2 (list A, B, C) but also the counterbalanced lists (list D, E, F). To prepare the counterbalanced lists, we prepared a synonym for each of the tokens in the three shared significant features. Most of the synonyms were identified using Japanese WordNet. To examine the validity of the literal stimuli, we conducted a pilot study (N = 60), as in Experiment 1. Based on the results of this pilot study, we selected 36 metaphors and 36 literal expressions. Items with ratings between the mean and −1 SD that were higher than the midpoint (3.5) were set as stimuli for the main experiment. For

example, in the list that was used in Experiment 2, for "that butterfly is a dancer," participants were presented with "that butterfly is beautiful" as one shared feature and "that butterfly is appealing" as literal. In the counterbalanced list, participants were presented with "that butterfly is gorgeous" as one shared feature and "that butterfly is splendid" as literal.

Procedure. The procedure was identical to that employed in Experiments 1, 2, and 3.

Results

Similar to the analytic strategy used in Experiments 1, 2, and 3, we examined nonsense-metaphorical and nonsense-literal responses as filler responses. Two results obtained in this experiment merit attention. First, there was no statistical difference between the list used in Experiment 2 and the counterbalanced list. To clarify the pattern of the results, we used a mixed-effects logistic regression model. We entered the list as the fixed factor along with the list used in Experiment 2 as the default level. In addition, we specified intercepts for participants and items as random factors. We used a model with responses that were coded using the following binary: metaphorical and non-metaphorical responses (i.e., literal and filler responses). Results showed that there was no significant difference between the list conditions (β = 0.17, SE = 0.32, z = 0.53, p = .60.).

Second, similar to Experiments 2 and 3, we obtained both the number of the features and the shared-significant features; the proportion of metaphorical responses was higher in the three-shared-significant-feature condition (.42) than in the one-shared-two-unshared-feature condition (.26) and the one-shared-significant-feature (.18) conditions. In addition, there was a significant difference between the three-shared-significant-feature versus the one-shared-two-unshared-feature conditions, and the three-shared-significant-feature versus one-shared-significant-feature contrast.

Discussion

In Experiment 4, we replicated Experiment 2 with the counterbalanced list. Results

confirmed that even if we changed the type of the shared significant feature presented to the participants in the one shared-significant-feature condition, the effect of the number of shared significant features was preserved. These results suggested that it was not the difference of the feature presented in the one feature condition (e.g., feature prototypicality) but the number of shared significant features attributed to the topic that affects the preference for metaphorical expressions over literal expressions.

Furthermore, as in Experiments 2 and 3, we confirmed that the number of topic attributed features also affected the preference for literal over non-literal. This suggests that the number of topic attributed features increases metaphor preference and decreases literal preference. Details of the analysis are reported in S4.

In Experiment 5, we replicated Experiment 1 with two counterbalanced lists. If we replicated the results obtained in Experiment 1, then it was not the differences in the feature presented in the one feature condition but the number of shared significant features attributed to the topic that affects the preference for metaphorical expressions over literal expressions.

Experiment 5

Method

Participants. We recruited 216 participants between the ages of 20 and 69 years (M = 40.5, SD = 9.1) using the same methodology that was employed in Experiments 1, 2, 3, and 4. Because the participants were recruited anonymously, some participants may have overlapped with those who were included in Experiments 1, 2, 3, and 4.

Design. Experiment 5 involved a 3 (number of shared significant features: one shared significant feature, two shared significant features, and three shared significant features) × 9 (list: list A, B, C, D, E, F, G, H, I) design. The number of shared significant features was a within-participants and within-items variable, while the list was a between-participants variable.

Stimuli. In Experiment 5, we not only used the list from Experiment 1 (list A, B, C) but also the counterbalanced lists (list D, E, F, G, H, I). To prepare the counterbalanced lists (namely, counterbalanced list A and counterbalanced list B), we prepared a synonym for each of the tokens in the three shared significant features. Most of the synonyms were identified using Japanese WordNet. To examine the validity of the literal stimuli, we conducted a pilot study (N = 30), as conducted in Experiment 1. Based on the results of this pilot study, we selected 24 metaphors and 24 literal expressions. Items with ratings of the aptness (the extent to which the meaning of a paraphrased synonym was apt for the original feature on a six-point Likert scale) higher than 4 were set as stimuli for the main experiment. For example, in the list used in Experiment 1, for "her smile is a flower," participants were presented with "her smile is beautiful" as one shared feature and "her smile is appealing" as literal. In counterbalanced list A, participants were presented with "her smile is bright" as one shared feature and "her smile is radiant" as literal. In counterbalanced list B, participants were presented with "her smile is gorgeous" as one shared feature and "her smile is splendid" as literal.

Procedure. The procedure was identical to that employed in Experiments 1, 2, 3, and 4.

Results

Similar to the analytic strategy used in Experiments 1, 2, 3, and 4, we examined nonsense-metaphorical and nonsense-literal responses as filler responses. Two results obtained in this experiment merit attention. First, there are some statistical differences between the list used in Experiment 1 and the counterbalanced lists. To clarify the pattern of the results, we used two mixed-effects logistic regression model. In one model, we entered the list as the fixed factor along with the list used in Experiment 1 as the default level. In addition, we specified intercepts for participants and items as random factors. We used a model with responses that were binary-coded as follows: metaphorical versus non-metaphorical responses

(i.e., literal and filler responses). Results showed that there was no significant difference for the counterbalanced list A (β = 0.12, SE = 0.23, z = 0.52, p = .60.) but that there was a significant difference for the counterbalanced list B (β = 0.52, SE = 0.23, z = 2.29, p < .05).

Second, though we observed an effect of the list, more importantly, we found an effect of the number of shared significant features; the proportion of metaphorical responses was higher in the three-shared-significant-feature condition (.51) than in the two-shared-significant-feature (.42) and the one-shared-significant-feature (.20) conditions. In addition, there was a significant difference between the three-shared-significant-feature versus the one-shared-two-unshared-feature condition, and in the three-shared-significant-feature versus one-shared-significant-feature contrast. To verify this pattern, we entered the number of shared significant features as a fixed factor with one shared feature as the default level and the list as the fixed factor, along with the list used in Experiment 1 as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses coded as one of the following: metaphorical responses and non-metaphorical responses (i.e., literal and filler responses). In addition to the results the positive effect of the list as explained above, both the two- versus one-shared-feature contrast (β = 1.52, SE = 0.10, z = 15.99, p < .001) and the three- versus one-shared-feature contrast (β = 2.04, SE = 0.10, z = 21.06, p < .001) showed significant differences. In addition, we ran a model that was the same as the above model but with two shared significant features as the default level. In addition to the results regarding the positive effect of the list as explained above, both the one-shared-significant-feature versus two-shared-significant-feature contrast (β = −1.52, SE = 0.10, z = −15.99, p < .001) and the three-shared-significant-feature versus two-shared-significant-feature contrast (β = 0.52, SE = 0.08, z = 6.25, p < .001) were significant.

Discussion

In Experiment 5, we replicated Experiment 1 with two counterbalanced lists. Though

the pattern slightly changed between the original list used in Experiment 1 and a counterbalanced list, the overall results confirmed that even if we changed the type of shared significant features presented to the participants in the one-shared-significant-feature condition, the effect of the number of shared significant features was preserved. These results suggest that it was not the differences in the feature presented in the one feature condition but the number of shared significant features attributed to the topic that affects the preference for metaphorical expressions over literal expressions.

Furthermore, as in Experiment 1, we confirmed that the number of shared significant features also affects the preference for literal over non-literal, suggesting that the shared significant features increase metaphor preference and decrease literal preference. Details of this analysis are reported in S4.

General Discussion

Across five experiments, we tested whether the number of shared topic-vehicle significant features affects speakers' preference for metaphorical expression use. Participants were asked to choose the option out of four that best paraphrased a given sentence. The main finding of Experiment 1 is that the greater the number of shared significant features, the greater the speaker's preference for metaphor use tended to be. Experiments 2 and 3 yielded two significant findings. First, the three-shared-significant-feature condition recorded higher metaphor and simile responses than the one-shared-two-unshared-feature condition and one-shared-significant-feature condition. Second, the one-shared-two-unshared-feature condition provoked a greater metaphor and simile preference than the condition with only one shared significant feature. Moreover, the effect of the number of shared topic-vehicle significant features was larger for similes than for metaphors. Furthermore, in Experiments 4 and 5, by utilizing a counterbalanced list, we showed that the results obtained in Experiment 1 and 2 were not due to the feature presented in the one-shared-significant-feature condition but because of the number of shared significant features.

In all five experiments, we found that the number of topic-attributed features affected participants' preference for metaphorical expression use; both the three-shared-significant-feature condition and the one-shared-two-unshared-feature condition showed a greater preference for metaphorical expression use than the one-shared-significant-feature condition. As Ortony (1975) suggested in the compactness hypothesis, it is difficult for literal expressions to uniquely and simultaneously explain many to-be-attributed features. More importantly, in Experiments 2, 3, and 4, we controlled shared topic-vehicle significant features (i.e., three-shared-significant-feature condition) and topic-vehicle unshared features (i.e., one-shared-two-unshared-feature condition); we found that shared topic-vehicle significant features had a greater effect on speakers' preference for metaphorical expression

use than unshared features. These results suggest that there is greater activation of metaphor form preference when there are many shared significant features between the topic and the vehicle.

Our results showed the effect of context (i.e., number of shared significant features) on metaphorical paraphrase over literal paraphrase using a form preference task. Though Schraw et al. (1988) showed the effect of context and familiarity on the preference for idiomatic interpretations over literal interpretations, they did not examine whether context affects metaphor form preference over literal form. As explained in the Introduction, though there have been many studies of metaphor form preference over simile form, few studies have investigated the context that affects metaphor form preference over literal form. This study showed the importance of the number of shared significant features in metaphor form preference.

Our results support the fundamental tenets of the inexpressibility hypothesis and compactness hypothesis. As the inexpressibility hypothesis and compactness hypothesis imply, participants could not explain the topic with three or two topic-attributed features at the same time in a single literal option in Experiments 1 and 5. Though previous studies tested these two hypothesis based on the descriptions collected from the speaker/writer's speaking/writing (Fainsilber & Ortony, 1987; Fussell & Moss, 1998; Williams-Whitney et al., 1992) and writer's explanation (Fussell & Krauss, 1989a, 1989b), this study supports the inexpressibility hypothesis based on the reader's preference for metaphorical expression. In addition, our study broadens the scopes of the inexpressibility hypothesis and compactness hypothesis. Though previous studies limited the topic domain to emotional experience (Fainsilber & Ortony, 1987; Fussell & Moss, 1998; Williams-Whitney et al., 1992) and abstract line drawings (Fussell & Krauss, 1989a, 1989b), our study tested these hypotheses using a sentence completion task.

Our results showed that trope type (metaphor vs. simile) had a significant effect on metaphor preference in Experiment 3. According to the literal base theory of figurative language (Chiappe & Kennedy, 2001) and aptness theory (Chiappe & Kennedy, 1999), similes require fewer shared significant features than metaphors do. This result can also be interpreted based on the observations of metaphor/simile frequency count via the Google search engine (Roncero, Kennedy, & Smyth, 2006; Roncero, De Almeida, Martin, & De Caro, 2016). According to these study findings, when metaphors (e.g., “Crime is a disease”) or similes (e.g., “Crime is like a disease”) were searched using Google, more similes were accompanied by explanations than metaphors (e.g., “Crime is like a disease because it spreads by direct personal influence”; Roncero et al., 2006). Moreover, in simile, many different features were mentioned in these explanations. This situation, whereby similes were accompanied by different features, was similar to the task requirements that our study entailed. In our tasks, participants were presented with the topic and its attributed features when they chose the best paraphrase from the list of response options. In this situation, participants could more easily access the simile option (e.g., “Her cat is like a princess”) than the metaphor option (e.g., “Her cat is a princess”). Because to-be-attributed features were embedded in the question (e.g., participants were asked to paraphrase the sentence, “Her cat is cute, cherished, and selfish”), it might have been easy for participants to choose a paraphrased option that presents the given sentence in terms of similes accompanied by features (e.g., “Her cat is like a princess because it is cute, cherished, and selfish”).

Even in the three-shared-significant-feature condition in Experiment 1, the average number of metaphorical responses did not reach 50%. This raises the question: How many features are needed to reach a metaphor preference rate of 50%? Three interesting results from the present study address this question. The first is that no metaphor reached a 50% response rate in any of the shared-significant-feature conditions. The highest metaphorical-

response rate in the one-shared-significant-feature condition was found for the metaphor "That riot is a storm" (i.e., 40% for the metaphor, "That riot is fierce"). The second interesting result is that there were some metaphors for which the number of metaphorical responses reached 50% in the three-shared-significant-feature condition. For example, the metaphor "Her cat is a princess" evidenced a metaphorical-response rate that was higher for the three-shared-significant-feature condition (i.e., 60% for the sentence, "Her cat is cute, cherished, and selfish") than the two-shared-significant-feature condition (i.e., 20% for the statement, "Her cat is cute and cherished") and the one-shared-significant-feature condition (i.e., 0% for the statement, "Her cat is cute").

The third interesting result is that there were other metaphors for which the number of metaphorical responses reached 50%, even in the two-shared-significant-feature condition. For example, with regard to the metaphor, "That job is a jail," the proportion of metaphorical responses was higher in the three-shared-significant-feature condition (i.e., 83% for the statement, "That job is tough, does not allow advancement, and is confining") than the two-shared-significant-feature condition (i.e., 63% for the statement, "That job is tough and does not allow advancement") and the one-shared-significant-feature condition (i.e., 18% for the statement, "That job is tough"). These three results suggest that (i) the number of shared significant features did not evidence a clear cut-off point at which participants' preference for metaphorical responses reached 50%; (ii) at least two shared significant features are required to reach a 50% metaphorical-response rate; and (iii) there could be another extraneous factor that affects the achievement of a 50% metaphorical-response rate. The potential factor could be an aptness of the feature to the metaphor. If the feature attributed to the topic captures salient property to the topic and the vehicle, metaphor form could be more preferred than literal form. Future research should investigate these specific factors that might result in a 50% metaphorical-response rate.

In conclusion, our results supported the fundamental tenet of the inexpressibility hypothesis and compactness hypothesis in the context of metaphor form preference. We found that the number of topic-attributed features affects participants' preference for metaphorical expression use. Moreover, shared topic-vehicle significant features have a greater effect on speakers' preference for metaphorical expression than topic-vehicle unshared features.

Acknowledgements

The work was supported by the Program for Leading Graduates Schools "Collaborative Graduate Program in Design" (Program Number K02) by the Ministry of Education, Culture, Sports, Science and Technology, Japan for data collection.

References

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed effects models using lme4. *Journal of Statistical Software*, 67, 1-48.

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

Cameron, L. (2003). *Metaphor in educational discourse*. London, England: Continuum.

Chiappe, D. L., & Kennedy, J. M. (1999). Aptness predicts preference for metaphors or similes, as well as recall bias. *Psychonomic Bulletin & Review*, 6, 668–676.

Chiappe, D. L., & Kennedy, J. M. (2001). Literal bases for metaphor and simile. *Metaphor and Symbol*, 16, 249–276.

Fainsilber, L., & Ortony, A. (1987). Metaphorical uses of language in the expression of emotions. *Metaphor and Symbol*, 2, 239–250.

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression*. Sage Publications.

Fussell, S., & Krauss, R. (1989a). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology*, 25, 203–219.

Fussell, S., & Krauss, R. (1989b). Understanding friends and strangers: The effects of audience design on message comprehension. *European Journal of Social Psychology*, 19, 509–525.

Fussell, S., & Moss, M. (1998). Figurative language in emotional communication. In S. Fussell, & R. Kreuz (Eds.), *Social and Cognitive Approaches to Interpersonal Communication*, 113–141. NY: Psychology Press.

Gibbs, R. W., Buchalter, D. L., Moise, J. F., & Farrar, W. T. (1993). Literal meaning and figurative language. *Discourse Processes*, 16, 387–403.

Glucksberg, S. (1989). Metaphors in conversation: How are they understood? why are they

used? *Metaphor and Symbol*, *4*(3), 125–143.

Glucksberg, S. (2008). How metaphors create categories—Quickly. In R. W. Gibbs Jr. (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 67–83). Cambridge, UK: Cambridge University Press.

Glucksberg, S., & Haught, C. (2006). On the relation between metaphor and simile. *Mind and Language*, 21, 360–378.

Hasson, U ., Estes, Z . & Glucksberg, S. (2001). Metaphors communicate more effectively than do similes. *Abstracts of the Psychonomic Society 42d Annual Meeting*, Vol. 6 , pp. 103. Austin, TX. Psychonomic Society Publications .

Haught, C. (2013). A tale of two tropes: How metaphor and simile differ. *Metaphor and Symbol*, 28, 254–274.

Kusumi, T. (1995). *Hiyu no shori katei to imi kouzou [The semantic structure and processes of metaphorical expressions]*. Tokyo, JP: Kazama Shobou.

Nakamoto, K. (2003). Semantic priming effect of metaphor constituent terms. *Perceptual and motor skills*, 96(1), 33-42.

Nakamoto, K., & Kusumi, T. (2004). A classification of 120 Japanese metaphorical expressions on the basis of four psychological dimensions. The Science of Reading, 48, 1-10.

Oka, R., Ohshima, H., & Kusumi, T. (2019). Development and validation of an item set of simile interpretations for metaphor research [In Japanese with English abstract]. *Japanese Journal of Psychology*, 90, 53-62.

Ortony, A. (1975). Why metaphors are necessary and not just nice. *Educational Theory*, 25, 45–53.

R Development Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing (URL http://www.R-project.org/).

Roncero, C., Kennedy, J. M., & Smyth, R. (2006). Similes on the internet have explanations. *Psychonomic Bulletin & Review*, 13, 74–77.

Roncero, C., & de Almeida, R. G. (2015). Semantic properties, aptness, familiarity, conventionality, and interpretive diversity scores for 84 metaphors and similes. *Behavior Research Methods*, 47, 800–812.

Roncero, C., De Almeida, R. G., Martin, D. C., & De Caro, M. (2016). Aptness predicts metaphor preference in the lab and on the internet. *Metaphor and Symbol*, 31, 31–46.

Schraw, G., Trathen, W., Reynolds, R. E., & Lapan, R. T. (1988). Preferences for idioms: Restrictions due to lexicalization and familiarity. *Journal of Psycholinguistic Research*, 17, 413–424.

Williams-Whitney, D., Mio, J., & Whitney, P. (1992). Metaphor production in creative writing. *Journal of Psycholinguistic Research*, 21, 497–509.

Supplemental material

S1. Descriptive Statistics and Correlations for Study Variables (Number of metaphor = 48)

Measure	*M*	*SD*	1	2	3	4	5	6	7	8	9
1. Familiarity	3.19	1.14	-								
2. Conventionality[a]	6.52	1.48	.55 ***	-							
3. Comporehensibility[b]	5.57	1.38	.77 ***	.67 ***	-						
4. Similarity[b]	4.14	1.41	.81 ***	.61 ***	.93 ***	-					
5. Uniqueness[b]	4.53	0.36	-.51 ***	-.27 †	-.45 **	-.57 ***	-				
6. Funniness[b]	4.37	0.55	.54 ***	.46 **	.72 ***	.61 ***	-.02	-			
7. Metaphor preference one[c]	0.16	0.11	.48 **	.59 ***	.49 **	.50 **	-.32 *	.33 *	-		
8. Metaphor preference two[c]	0.31	0.19	.39 **	.53 ***	.30 *	.32 *	-.32 *	.07	.63 ***	-	
9. Metaphor preference three[c]	0.41	0.20	.45 **	.35 *	.36 *	.33 *	-.29 †	.17	.45 **	.69 ***	-

[a]Oka, Ohshima, & Kusumi (2019)

[b]Nakamoto & Kusumi (2004)

[c]Results of Experiment 1

†p <.10, *p < .05, **p < .01, ***p < .001

S2. Mixed-effects logistic regression model with number of features and familiarity

In this supplement, we report the relationship between average familiarity measure and metaphor preference.

First, we collected a familiarity measure for each metaphor. We asked participants (N = 24) to rate how much they had heard/read the presented expressions on a 9-point Likert scale (1 = *not at all familiar* to 9 = *very familiar*). We selected all 45 metaphors used in Experiment 1 as stimuli for this questionnaire. We calculated the average familiarity rating across participants for each metaphor and used these scores as a familiarity measure.

Second, to clarify the relationship between familiarity measure and metaphor preference, we conducted a mixed-effects logistic regression model. We entered two fixed factors: the number of shared significant features with on -shared significant feature as the default level and the familiarity measure as fixed factor. In addition, we specified intercepts for participants and items as random factors. We first ran a model with responses coded as one of the following: metaphorical responses and non-metaphorical responses (i.e., literal and filler responses). Three significant effects merit attention: (a) the two- versus one-shared-feature contrast (β = 1.22, SE = 0.10, z = 12.33, p < .001), (b) the three- versus one-shared-feature contrast (β = 1.22, SE = 0.10, z = 12.33, p < .001), and (c) the familiarity measure (β = 0.47, SE = 0.13, z = 3.70, p < .001). In addition, we ran a model that was the same as the first model but with two shared significant features as the default level. The three- versus two-shared-significant-feature contrast (β = 0.65, SE = 0.08, z = 7.66, p < .001) was significant. These results suggests that though the familiarity measure had a positive effect on metaphor preference, the number of shared significant features did so as well.

S3. Crowdsourcing service used and participant overlap between experiments

In this section, we report some properties of the crowdsourcing service we use and the overlapping participants between experiments.

First, the subject pool was drawn from a crowdsourcing service in Japan (Crowdworks). Crowdworks had more than 2,000,000 memberships in 2018 (press release in Japanese in 2018; https://crowdworks.co.jp/news/0007748/), and thus was sufficiently large.

Second, we have checked the number of overlapping participants based on their IP address. In addition, we compared the response pattern between overlapping participants and non-overlapping participants. First, there were 31 participants in Experiment 2 who overlapped with Experiment 1. We confirmed there were no significant effects of this group difference on the pattern of metaphor selection (β = 0.45, SE = 0.28, z = 1.62, *n.s.*). Second, there were 56 participants in Experiment 3 who overlapped with either Experiment 1 or Experiment 2. We confirmed there were no significant effects of this group difference on the pattern of metaphor selection (β = 0.26, SE = 0.22, z = 1.19, *n.s.*). There were only six participants in Experiment 4 who overlapped with Experiments 1, 2, or 3. These results suggest that there was little effect of participant overlap on metaphor preference.

S4. The effect of number of features on literal preference

In this section, we report the effect of number of shared significant features on literal preference in each experiment. All the statistical analyses reported below were performed with a mixed-effects logistic regression model via the lmer program of the lme4 package (Bates, Maechler, Bolker, & Steve, 2015) in the R (version 3.3.2) environment for statistical computing (R Development Core Team, 2016). Also, nonsense-metaphor and nonsense-literal responses were coded as filler responses.

In Experiment 1, we entered the number of shared significant features as a fixed factor with one shared feature as the default level. In addition, we specified intercepts for participants and items as random factors. We first ran a model with responses coded as one of the following: literal responses and non-literal responses (i.e., metaphor and filler responses). Both the two- versus one-shared-feature contrast ($\beta = -1.31$, $SE = 0.10$, $z = -13.29$, $p < .001$) and the three- versus one-shared-feature contrast ($\beta = -1.96$, $SE = 0.10$, $z = 19.68$, $p < .001$) showed significant differences. In addition, we ran a model that was the same as the first model but with two shared significant features as the default level. Results showed that the three- versus two-shared-significant-feature contrast ($\beta = -0.65$, $SE = 0.08$, $z = 7.72$, $p < .001$) was significant. These results confirmed that in the three-shared-significant-feature condition (.57) fewer literal responses were reported than in the two- (.68) and one-shared-significant-feature conditions (.81). These results suggest that the number of shared significant features has a negative effect on literal preference over metaphorical and filler expressions.

In Experiment 2, we entered the number of shared significant features as a fixed factor with one-shared significant features as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses that were binary-coded as follows: literal and non-literal responses (i.e., metaphorical and filler responses). Results pertaining to both the one-shared-significant-feature versus the one-shared-two-unshared-feature contrast ($\beta = -1.15$, $SE = 0.10$, $z = -11.09$, $p < .001$) and the one-shared-significant-feature versus three-shared-significant-feature contrast ($\beta = -2.05$, $SE = 0.10$,

$z = -19.58$, $p < .001$) were significant. In addition, we ran a model that was the same as the first model but with one shared significant and two unshared features as the default level. Results showed the one-shared-two-unshared-feature versus three-shared-significant-feature contrast ($\beta = -0.90$, $SE = 0.09$, $z = -10.12$, $p < .001$) was significant. These results confirmed that in the three-shared-significant-feature condition (.61), fewer lower literal responses were recorded than in the one-shared-two-unshared-feature (.74) and one-shared-significant-feature conditions (.87). These results suggest that the number of shared significant features has a negative effect on literal preference over metaphorical and filler expressions. These results suggest that, even though both shared features and unshared features affect speaker's preference for literal expression, the effects of the number of features were stronger for shared features than for unshared features.

In Experiment 3, we entered the number of shared significant features as the fixed factor with one shared significant feature as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses that were binary-coded as follows: literal and non-literal responses (i.e., metaphorical and filler responses). Results pertaining to both the one-shared-significant-feature versus the one-shared-two-unshared-feature contrast ($\beta = -0.68$, $SE = 0.08$, $z = -8.19$, $p < .001$) and the one-shared-significant-feature versus three-shared-significant-feature contrast ($\beta = -1.91$, $SE = 0.09$, $z = -21.53$, $p < .001$) were significant. In addition, we ran a model that was the same as the first model but with one shared significant and two unshared features as the default level. Results showed the one-shared-two-unshared-feature versus three-shared-significant-feature contrast ($\beta = -1.23$, $SE = 0.08$, $z = -14.65$, $p < .001$) was significant. These results confirmed that the three-shared-significant-feature condition (.36) showed fewer literal responses than one-shared-two-unshared-feature (.57) and one-shared-significant-feature conditions (.68). These results suggest that the number of shared significant features has a negative effect on the literal preference over metaphorical and filler expressions. These results suggest that, even though both shared features and unshared features affect speaker's preference for literal expression, the effects of the number of features were stronger for shared features than for unshared features.

In Experiment 4, two results were obtained. First, there was no statistical difference between the list used in Experiment 2 and the counterbalanced list. To clarify the pattern of the results, we used a mixed-effects logistic regression model. We entered the list as the fixed factor along with the list used in Experiment 2 as the default level. In addition, we specified intercepts for participants and items as random factors. We used a model with responses binary-coded as follows: literal and non-literal responses (i.e., metaphorical and filler responses). Results showed no significant difference between the list conditions ($\beta = -0.22$, $SE = 0.30$, $z = 0.71$, *n.s.*). Second, similar to Experiments 2 and 3, we examined both the number of features and number of shared-significant features; the proportion of literal responses was lower in the three-shared significant features condition (.55) than in the one-shared-two-unshared-feature (.68) and the one-shared significant feature (.81) conditions. In addition, there was a significant difference between the three-shared-significant-feature versus one-shared-two-unshared-feature, and the three-shared-significant-feature versus one-shared-significant-feature contrasts. To clarify these pattern, we entered the number of shared significant features as the fixed factor with one shared significant features as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses that were binary-coded as follows: literal and non-literal responses (i.e., metaphorical and filler responses). Results pertaining to both the one-shared-significant-feature versus one-shared-two-unshared-feature contrast ($\beta = -1.00$, $SE = 0.08$, $z = -13.23$, $p < .001$) and the one-shared-significant-feature versus three-shared-significant-feature contrast ($\beta = -1.80$, $SE = 0.08$, $z = -23.51$, $p < .001$) were significant. In addition, we ran a model that was the same as the first model but with one shared significant and two unshared features as the default level. Results showed the one-shared-two-unshared-feature versus three-shared-significant-feature contrast ($\beta = -0.80$, $SE = 0.07$, $z = -11.80$, $p < .001$) was significant.

Finally, in Experiment 5, two results were obtained. First, there were some statistical differences between the list used in Experiment 1 and the counterbalanced lists. To clarify the pattern of the results, we used two mixed-effects logistic regression models. In one model, we entered the list as a fixed factor along

with the list used in Experiment 1 as the default level. In addition, we specified intercepts for participants and items as random factors. We used a model with responses that were binary-coded as follows: metaphorical and non-metaphorical responses (i.e., literal and filler responses). Results showed that there was no significant difference for counterbalanced list A (β = −0.12, *SE* = 0.22, *z* = −0.57, *n.s.*) but a significant difference for counterbalanced list B (β = −0.54, *SE* = 0.22, *z* = −2.51, *p* < .05).

Second, though we observed an effect of the list, more importantly, we measured the effect of the number of shared significant features; the proportion of literal responses was lower in the three-shared-significant-feature condition (.48) than in the two-shared-significant-feature (.57) and one-shared-significant-feature (.79) conditions. In addition, there was a significant difference between the three-shared-significant-feature versus the one-shared-two-unshared-feature, and the three-shared-significant-feature versus one-shared-significant-feature contrast. To verify this pattern, we entered the number of shared significant features as a fixed factor with one shared feature as the default level and the list as the fixed factor along with the list used in Experiment 1 as the default level. In addition, we specified intercepts for participants and items as random factors. We ran a model with responses binary-coded as follows: literal responses and non-literal responses (i.e., metaphorical and filler responses). In addition to the results regarding the positive effect of the list as explained above, both the two- versus one-shared-feature contrast (β = −1.48, *SE* = 0.09, *z* = −15.92, *p* < .001) and the three- versus one-shared-feature contrast (β = −1.99, *SE* = 0.09, *z* = −21.06, *p* < .001) showed significant effects. In addition, we ran a model that was the same as the above model but with two shared significant features as the default level. The result showed that the three- versus two-shared-significant-feature contrast (β = −0.51, *SE* = 0.08, *z* = −6.24, *p* < .001) was significant.