Aptness is more important than comprehensibility in preference for metaphors and similes

Dan L. Chiappe, John M. Kennedy, Penny Chiappe

Abstract

Figurative comparisons can be expressed as metaphors (e.g., “politics is a circus”) or similes (e.g., “politics is like a circus”). What determines the form in which a comparison is expressed? We examine two potential factors—aptness and comprehensibility. To be apt is to capture important features of a topic. Comprehensibility means being relatively easy to understand. We show both of these judgments are related to errors in a recognition memory test (i.e., remembering a simile as a metaphor or a metaphor as a simile). However, aptness was a better predictor of the errors than comprehensibility. Furthermore, while both aptness and comprehensibility predicted preference for the metaphor or simile form of comparisons in a direct test of preference, aptness explained unique variance, while comprehensibility did not. We argue that although comparisons have to be comprehensible to be proper metaphors or similes, aptness is more important in determining whether a comparison is preferred as a metaphor or as a simile.

1. Introduction

Metaphors and similes are devices used by poets and lay people to express figurative comparisons. Non-literal assertions are variations on literal assertions and, at least on the surface, metaphors such as “man is a wolf” take the form of a literal assertion of a categorical relation between two concepts, much as in “conifers are...
trees.” In contrast, similes such as “my love is like a red, red rose” superficially assert a relation of likeness between two items, much as in “Chevys are like Fords” (Chiappe and Kennedy, 2001). Here we ask what determines the preference for expressing a comparison between two concepts as a simile or a metaphor.

First, let us define some terms. We will refer to the two concepts involved in figurative comparisons as the “topic” and the “vehicle”. The topic is the subject of a figurative statement (i.e., “man” in “man is a wolf”), and the vehicle is the concept that we are using to say something new about the topic (i.e., “red, red rose” in “my love is like a red, red rose”).

Given that any figurative comparison can be expressed as a metaphor or as a simile, why do people choose one form of expression over the other? Is a person just as likely to say “trees are like umbrellas” as they are to say “trees are umbrellas”? Is “crime is a disease” preferred to “crime is like a disease”? If so, what determines the best way of expressing the comparison between trees and umbrellas, or crime and disease?

Here, we will examine two features of comparisons that might be used in their evaluation. These are the virtues of aptness and comprehensibility. We define aptness as “the extent to which a comparison captures important features of the topic” (Chiappe and Kennedy, 1999; Gibbs, 1993, 1994; Katz, 1989, 1992; Malgady and Johnson, 1976; Tourangeau and Sternberg, 1981, 1982). To take an example, consider the metaphor “life is a valuable gift”. This metaphor captures some important features of life, such as the fact that it is precious and that we are lucky to have it. However, perhaps it does not capture as many important features of life as does “life is a journey” (Lakoff and Johnson, 1980). According to the “conventional metaphor” view, the latter reflects a longstanding conceptual mapping between the domains of “life” and “journeys” that can be elaborated in many different ways. As a result, it captures many important features associated with life—that it often has a goal and destination, that it can be long and arduous, that one can lose one’s way, that it can be undertaken with fellow travelers, and so on. Understood in this fashion, the comparison between life and journeys may be more apt than that between life and a valuable gift.

We define comprehensibility as “how easy it is to understand a comparison” (Gerrig and Healy, 1983; Gibbs, 1993, 1994; Katz, 1989, 1992; Malgady and Johnson, 1976; Tourangeau and Sternberg, 1981, 1982). A comparison is highly comprehensible if one can construct an interpretation relatively easily. Certainly, participants find some comparisons easier to comprehend than others. Many might find “marriage is a zero-sum game” to be more difficult to comprehend than “marriage is a fortress”. In some cases, ease of comprehension may arise because the statements are instances of more general metaphoric mappings, such as the mapping between “arguments” and “war” (e.g., “she defended her arguments against vigorous attacks”), and between “ideas” and “food” (e.g., “his lecture was a three course meal”) (Lakoff and Johnson, 1980; Gibbs, 1994). However, we maintain that the main factor determining comprehensibility is how readily people can identify relevant properties of the vehicle that are attributable to the topic.

It is likely that comprehensibility and aptness are related but distinct. A comparison can be easy to comprehend but nonetheless not be very appropriate. For
example, consider describing a friend who lay out in the sun too long by saying “Jane is a lobster”. This statement may be easy to understand, but it does not strike us as particularly apt. The properties it captures are very few in number (e.g. color), and thus may fail to bring out many of Jane’s features (e.g., her carelessness). Moreover, a comparison can be very appropriate, but be difficult to comprehend. This is particularly the case with literary, or poetic metaphors (Steen, 1994). For example, Kundera (1984) describes the music of J.S. Bach’s time as being like “a rose blooming on a boundless, snow-covered plain of silence”. Though this strikes one as quite apt upon reflection, arriving at an interpretation is a difficult process.

It has been suggested that aptness and comprehensibility judgments reflect different stages in the processing of figurative comparisons. As Gibbs (1994) reminds us, “[all] language interpretation takes place in real time ranging from the first milliseconds of processing to long-term reflective analysis” (p. 116). In particular, Gibbs (1994) argues comprehension and appreciation operate at two temporal extremes. Comprehension emphasizes the initial stages of processing. In contrast, appreciation is an assessment of the merits of a comparison that is likely to be done, if at all, at a later stage, via a conscious, optional, and reflective judgment undertaken on the products of comprehension (Gibbs, 1994).

Comprehension may indeed emphasize initial reactions to statements and aptness judgments may often be relatively late. We do not wish to claim, however, that aptness judgments cannot be simultaneous with comprehension at times (as it may be with some humorous remarks that are easy to grasp). Nor do we wish to argue that all of comprehension occurs in less than a second. Some remarks take time to fathom. Rather, we take the central point of Gibbs’ account to be that there may be influences on judgments of aptness that are not part and parcel of comprehension.

Gerrig and Healy (1983) found that certain factors could affect comprehensibility without affecting aptness judgments. Specifically, the order in which the literal and figurative components of sentences appear (e.g., “the night sky was filled with drops of molten silver” versus “drops of molten silver filled the night sky”), affected the time that it took participants to state that they had comprehended the metaphors. (The former order was more difficult.) However, this factor did not have an effect on participants’ judgments of the goodness of the metaphors. As a result, they concluded, “comprehending and appreciating a metaphor are not equivalent processes” (Gerrig and Healy, 1983, p. 669).

Gerrig and Healy (1983), however, did not specify the processes involved in appreciation and comprehension. It is possible the two share some important processes brought to bear in a different manner in each case. Specifically, both surely involve a search for relevant properties of the vehicle that are attributable to the topic (Chiappe, 1998). When judging comprehensibility of a comparison, a person likely proceeds by searching for these relevant properties, some highly salient (Giora et al., 1998; Glucksberg, 2001) and some perhaps obscure. If readily accessible, the person might judge the comparison to be very comprehensible. Likewise, when judging for aptness, a person will proceed by searching for properties shared by the topic and the vehicle. If the commonalities are substantial and pertinent, a person might judge the comparison to be very apt, even if the features are of low salience.
The two judgments, however, may differ in the way the search for relevant properties is carried out. When comprehension is the goal, we will likely seek enough properties to make the utterance of the comparison consistent with the assumption that the speaker was trying to say something relevant. We can stop when we have reached that goal (Grice, 1975; Sperber and Wilson, 1995). When we determine the aptness of a comparison, however, a more extensive assessment of the set of relevant properties might be undertaken. As a result, aptness may take longer to judge than comprehensibility.

If aptness and comprehensibility are distinct, what role does each play in determining whether a person expresses a particular comparison as a metaphor or simile? Is one more central?

If metaphors and similes are related to literal categorization and similarity claims (Chiappe and Kennedy, 1999, 2001), the metaphoric form of a comparison might be preferred when a comparison is highly apt. This is because categorization claims generally imply that an object has more properties associated with a particular category than do similarity claims. For example, the statement “*x* is an apple” implies the object has all of the properties needed to be a member of the category “apple,” while the statement “*x* is like an apple” implies the object only has some of them. As a result, literal similarity claims imply fewer common properties than literal categorization claims.

Some of what holds for literal claims about categorization and similarity may also hold for metaphors and similes, because they have the same form as the literal claims (Chiappe and Kennedy, 2001). Consider the comparison between a person’s job and a jail. What determines whether people will express this as “my job is like a jail” or “my job is a jail”? The answer, say Glucksberg and Keysar (1990, 1993), has to do with the extent to which the topic and the vehicle share common properties (see also Roberts and Kreuz, 1994). Specifically, one chooses the metaphor if one feels “the job in question has all the properties of the…category…‘jail’” (1993, p. 413). One chooses the simile if one feels that one’s job does not have “all the properties of the class ‘jail’…but only some of them” (1993, p. 413).

In our terms, people may use the metaphorical form of a comparison when they feel a comparison is particularly apt, capturing many properties. They may use the simile form when they do not feel a comparison is quite as apt. What about comprehensibility? If aptness involves a more extensive and reflective evaluation of the ground of a comparison, it is possible that aptness plays a more important role in predicting the form in which a comparison will be preferred. This is because comprehensibility judgments reflect having enough properties to be understood, but would not reflect an assessment of the pertinence of the similarity of the topic and vehicle, as the aptness judgments do.

At this point we need to introduce a cautionary note. The foregoing discussion assumes that metaphors involve more properties than similes because they take the form—on the surface—of categorization statements. Our caution is that they are metaphorical, not literal, categorization claims. Consider the metaphor “my job is a jail”. Treating this statement as if it were a literal categorization statement entails a category mistake. A person’s job cannot be a jail. Glucksberg and Keysar (1990,
resolve this problem by claiming that in metaphors the vehicles are taken to be prototypical exemplars of particular ad hoc categories, not the categories labeled by the literal use of the terms. Thus, people use the name of the prototypical example of an ad hoc category to refer to that category. For instance, a jail is the prototypical example of things that are unpleasant, constraining and confining, where people are held against their will. As a result, we can allude to this category using the word “jail.” When a person says “my job is a jail” they are claiming their job belongs in the ad hoc category exemplified by jails. Likewise, on hearing a new metaphor, the listener searches for a novel ad hoc category, which would be named by the vehicle term.

The class-inclusion view contrasts with the comparison theory, which goes back to Aristotle (1926; see also Fogelin, 1988; Gentner, 1983, 1989; Gentner and Clement, 1988; Gentner and Markman, 1997; Kintsch, 1974; Koen, 1965; Miller, 1979; Ortony, 1979a,b; Tversky, 1977). The theory holds metaphors are similes with the term of comparison (the “like”) suppressed. Due to the omission of the term “like”, the statement is relatively concise and pithy—matters of style rather than substance. On this account, “my job is a jail” is not a class-inclusion statement. It is merely an abbreviation for “my job is like a jail”. This theory does not predict any relation between the comprehensibility or aptness of a comparison and the form in which it is preferred. Since metaphors and similes mean the same thing, a comparison is expressed as a metaphor rather than as a simile simply if there are reasons to be brief.

Chiappe and Kennedy (1999) found support for the class-inclusion theory. They found aptness predicted the form in which comparisons were preferred. Participants were asked to judge whether particular comparisons are best expressed as metaphors or similes. This generated a range of preferences, with highly apt comparisons being more likely to be preferred as metaphors than as similes.

Chiappe and Kennedy (1999), however, did not examine if comprehensibility predicts preferred form of expression. The present experiment thus examines whether comprehensibility plays a role in this regard, and whether it is as important as aptness. We gave participants metaphors and similes and had them judge aptness and comprehensibility. We also measured the latencies to make those judgments. Preferred form of expression was examined in two ways. One was explicit, by relying on people’s judgments when asked whether the metaphor or the simile form is the best way of expressing the relation between a topic and a vehicle. A second way was implicit, by examining errors in a recognition memory test. A memory test was used because errors in recalling the form of comparisons may be predicted from preference judgments. Comparisons preferred as metaphors (e.g., “genes are blueprints”) might be misremembered as metaphors more often than as similes (“genes are like blueprints”), while comparisons preferred as similes (e.g., “highways are like snakes”) might be misremembered as similes more often than as metaphors. Thus, when it is difficult to remember the form of comparisons, people asked to remember the forms might rely on their preference for the metaphor or simile forms of comparisons. As a result, memory errors might indicate preference for the form of comparisons. Here, we asked whether the errors are related to comprehensibility or aptness judgments.
Finally, the present study tested our claim that aptness involves a more thorough assessment of the merits of a comparison. If this claim is true, then the latency to make aptness judgments should be greater than the latency to make comprehensibility judgments.

2. Method

2.1. Participants

Thirty-four participants (27 female and 7 male), with a mean age of 25 years (S.D. = 3.40), participated in this experiment. Participants were from the Department of Education at the University of British Columbia, and they received $5.00. The participants were tested individually.

2.2. Stimuli and apparatus

We used 30 comparisons (see Appendix) in both their metaphor and simile form taken from Chiappe and Kennedy (1999). This experiment was conducted using SuperLab software running on an IBM ThinkPad computer.

2.3. Procedure

In the first part, participants were shown 30 comparisons. They saw 15 as metaphors and 15 as similes, mixed, not blocked. The items were presented one at a time and in a different random order for each subject. Half of the participants were asked to read each statement and, as their first task, to rate each one for its aptness. Aptness was defined as “the extent to which a comparison captures important features of the topic.” Participants had to indicate, by key-press, the aptness of each statement, using a scale ranging from “1” (very inappropriate) to “10” (very appropriate). The other half of the participants was also asked to read each of the statements. However, as their first task they were asked to rate each statement for its comprehensibility. This was defined as “how easy it is to understand a statement”. The participants had to give their estimate of the comprehensibility of each statement by key-press, using a scale from “1” (very incomprehensible) to “10” (very comprehensible).

Participants made their responses in two stages. In particular, they were told to press the space bar once they had made their decision on the aptness or comprehensibility of the item. Once they pressed the space bar they were asked to type into the computer the number from 1 to 10 that they had decided on. Latency was measured from the time that the items appeared on the screen to the time when the subject pressed the space bar to indicate that they had made their decision. Participants were told that it was very important that they press the space bar only once they had arrived at a decision on the aptness or comprehensibility of the comparison.
Once participants completed their first task, they were given a surprise recognition test. They were shown the 30 comparisons they had seen in the first part, but in a different random order. Moreover, the comparisons were shown in both their metaphor and simile form, as a pair of statements, one presented above the other. Each pair was presented in the center of the screen. The order in which the metaphors and similes appeared was randomized. Half of the time the metaphors appeared as the top member of the pair and half of the time the similes appeared on top. Participants were asked to indicate which form of the comparison they had seen in part one. They had to press the “/” key (labeled “M”) if they had seen it as a metaphor and they had to press the “Z” key (labeled “S”) if they had seen a comparison expressed as a simile.

In the third part, participants were shown the same sentences they had seen in the first, but in a different random order. As before, the sentences appeared one at a time. Participants who rated the items for their aptness in the first part were now asked to rate for comprehensibility, and participants who first rated the items for comprehensibility were now asked to rate for aptness. Once again, the judgments and the latencies to make them were recorded.

3. Results

We will begin by examining whether the form of the comparisons (i.e., metaphor or simile) had an effect on the aptness and the comprehensibility judgments.

Consider first aptness. When the comparisons were rated for aptness first, the mean aptness rating for metaphors was 4.76 (S.D. = 1.32), and the mean for similes was slightly higher (4.93, S.D. = 1.47). When the subject’s last task was to rate the comparisons for aptness, the mean aptness rating for metaphors was 6.33 (S.D. = 1.86), and for similes it was again slightly higher (6.91, S.D. = 1.46). Collapsing across order, the correlation between the aptness ratings for metaphors and similes was 0.81, $P < 0.001$. However, a 2×2 repeated-measures ANOVA with order (aptness rating first vs. aptness rating last) and comparison type (metaphor and simile) as repeated-measures factors revealed a main effect of comparison type, with similes being more apt than metaphors, $F_1 (1, 32) = 4.60, P < 0.05; F_2 (1, 29) = 5.68, P < 0.05$. There was also a main effect of order with aptness being higher when judged last, $F_1 (1, 32) = 37.31, P < 0.0001; F_2 (1, 29) = 92.41, P < 0.001$. There was no interaction between order and comparison type, $F_1 (1, 32) = 1.68$, n.s.; $F_2 (1, 29) = 1.68$, n.s.

Let us now consider the comprehensibility ratings. When comprehensibility was rated first, the mean comprehensibility rating for the metaphors was 7.03 (S.D. = 1.63), and for the similes was slightly higher (7.35, S.D. = 1.59). When the subject’s last task was to rate comprehensibility, the mean rating for the metaphors was 6.83 (S.D. = 1.61), and for the similes was slightly higher (7.02, S.D. = 1.85). Collapsing across order, the correlation between the comprehensibility ratings for the metaphors and similes was 0.86, $P < 0.001$. A 2×2 repeated-measures ANOVA with order and comparison type as the repeated-measures factors revealed no main
effect of comparison type, $F_1 (1, 32) = 2.32$, n.s.; $F_2 (1, 29) = 2.89$, n.s. There was also no main effect of order, $F_1 (1, 32) < 1$; $F_2 (1, 29) = 3.34$, n.s., and no interaction between order and comparison type, $F_1 (1, 32) < 1$; $F_2 (1, 29) < 1$. Evidently, the comparisons were assessed for comprehensibility in a stable fashion, which follows from the argument that assessing a comparison for its comprehensibility is not an option.

Consider now the differences in latencies for aptness and comprehensibility judgments. Do aptness judgments take longer? Because analyses revealed no differences in latencies for making the judgments for metaphors and similes, the following analyses will collapse across statement type. When aptness was rated first, the mean latency for rating aptness for each topic-vehicle pair was 3.6 s (S.D. = 0.60) and the mean latency for rating comprehensibility was 2.7 s (S.D. = 0.55). When comprehensibility was rated first, the mean latency for rating aptness was 3.1 s (S.D. = 0.64), and the mean latency for rating comprehensibility was 2.8 s (S.D. = 0.62). A 2×2 repeated-measures ANOVA with order (aptness first vs. comprehensibility first) and judgment type (aptness rating vs. comprehensibility rating) as the two repeated-measures factors revealed a significant main effect of judgment type, with aptness judgments taking longer than comprehensibility judgments, $F_1 (1, 32) = 9.93$, $P < 0.005$; $F_2 (1, 29) = 18.66$, $P < 0.001$. The mean latency to make the aptness ratings was 3.3 s (S.D. = 0.57), while for the comprehensibility ratings, the latency was 2.8 s (S.D. = 0.46). None of the interactions reached significance. In short, the key finding here is that aptness judgments took longer to make than comprehensibility judgments.

So far judgments and latencies have been considered separately. To compare the two, we calculated mean aptness ratings and mean comprehensibility ratings for each comparison, as well as mean latencies for making the judgments, collapsing across order and statement type. The mean aptness rating was 5.73 (S.D. = 1.32); the mean comprehensibility rating was 7.06 (S.D. = 1.52), with comprehensibility ratings being greater than the aptness ratings, $t (29) = 13.32$, $P < 0.001$. Naturally, a claim can be comprehensible (“Sunburnt Jane is a lobster”) without being particularly apt (lobsters are not careless). Comprehension requires suitable features. Aptness involves weighing the suitability of the features.

We found a correlation of 0.94 between the comprehensibility judgments and the aptness judgments, $P < 0.001$. As aptness increased, so too did comprehensibility. This is consistent with earlier findings by Trick and Katz (1986) of strong positive correlations between aptness and comprehensibility. However, the correlation between the latencies was only 0.14 (n.s.). Furthermore, there was a correlation of $−0.14$ (n.s.) between the mean aptness judgments and the latencies to make those judgments, so higher aptness did not lead to a decrease in amount of time required to make the aptness judgments. In contrast, the correlation between the comprehensibility judgments and the comprehensibility latencies was $−0.49$, $P < 0.01$. Thus, what is easy to understand was understood quickly.

The correlation between the comprehensibility latencies and the aptness judgments was $−0.55$ ($P < 0.01$). This means that those items that were comprehended faster also tended to receive higher aptness ratings. A similar relation did not hold
for the comprehensibility judgments and aptness latencies. The correlation between
the comprehensibility ratings and the latencies to judge aptness was \(-0.13\) (n.s.).
Two factors are likely involved here: A comparison can be easy to understand, but
we may still have to weigh the pertinence of its features, just as we do for some less-
comprehensible comparisons. Further, if the comparison is hard to understand we
may quickly decide it is not especially apt.

In short, this analysis revealed a telling dissociation between comprehensibility
and aptness. Although the two judgments are highly correlated, the latencies to
make those judgments are not. Furthermore, the aptness of the items is not reflected
in the latency to make the aptness judgments, though the comprehensibility of the
items is reflected in the latency to make the comprehensibility judgments. To be
highly comprehensible is to be comprehended faster (which supports the claim that
people’s judgments of comprehensibility are reliable). The results also suggest com-
prehensibility is an important component determining the aptness of an item, as
revealed by the significant correlation between the comprehensibility latencies and
the aptness judgments.

The dissociation permits us to continue to seek effects of the differences, and to
pursue our suggestion that judging aptness may involve more assessment of key
features than does judging comprehensibility. Judging comprehensibility may be a
more practiced and automatic process than judging aptness. Judging aptness may
take into consideration many of the same features and evaluative dimensions as
judging comprehensibility plus others besides, such as low-salient features and
degree of pertinence.

Next we turn to the results of the recognition memory test. This required partici-
pants to pick out the form in which the comparisons were presented. Overall, the
mean proportion of items correctly recognized was 0.65 (S.D. = 0.12), significantly
greater than chance, \(t_1 (33)=9.36, P<0.001; t_2 (29)=6.78, P<0.01\) (but, helpfully,
lower than ceiling). Of particular interest were the errors people made when recog-
nizing the metaphors and similes. These are (a) recognizing a metaphor as a simile,
and (b) recognizing a simile as a metaphor. These two kinds of errors tell us about
any bias in remembering the comparisons as metaphors and as similes. When parti-
cipants rated for comprehensibility prior to aptness, the mean proportion of meta-
phors falsely recognized as similes was 0.24 (S.D. = 0.20), and the mean proportion
of similes falsely recognized as metaphors was 0.47 (S.D. = 0.21). Likewise, when
participants performed the aptness ratings prior to the comprehensibility ratings, the
mean proportion of metaphors falsely recognized as similes was 0.29 (S.D. = 0.21)
and the mean proportion of similes falsely recognized as metaphors was 0.41
(S.D. = 0.21).

Were the errors related to the aptness and comprehensibility of the comparisons?
To address this issue, we began by creating mean aptness and mean comprehensi-
bility ratings for each comparison. This is justified since the correlations between the
ratings for metaphor and simile on each judgment are very high, as noted above,
combining the two increases the power of the analysis, and in principle the theore-
tical interest lies in the content of the comparisons (i.e. “crime” with “disease”).
Mean aptness ratings for the comparisons were the averages of the aptness ratings
for the comparisons expressed as metaphors and expressed as similes, collapsing across order. Mean comprehensibility ratings were calculated in an analogous fashion.

The 30 comparisons were first ranked in terms of their mean aptness ratings, and were then split into two groups. The groups were formed from the 12 items rated highest in aptness (the top 40%) and the 12 items rated lowest in aptness (the bottom 40%). (If the 12 items were selected on the basis of the aptness ratings of metaphors, 8 of the 12 high-apt items would still be on the list, and 9 of the 12 low apt. If the 12 items were selected on the basis of aptness ratings of similes, there would also be 9 of the 12 high apt and 9 of the 12 low apt on the list.) One group was labeled the “high aptness” group, the other the “low aptness” group. The mean aptness rating for the “high aptness” group was 6.94 (S.D. = 0.40), with a range from 6.30 (for the comparison between “nature” and “laboratory”) to 7.48 (for the comparison between “crime” and “disease”). The mean aptness rating for the “low aptness” group was 4.43 (SD = 0.98), with a range from 2.85 (for the comparison between “treetrunks” and “straws”) to 5.64 (for the comparison between “jobs” and “jails”).

Aptness level was then related to the proportion of similes remembered as metaphors and to the proportion of metaphors remembered as similes. The data are summarized in Table 1. A $2 \times 2 \times 2$ repeated measures ANOVA was used to examine the results, with aptness level (high vs. low), order (comprehensibility followed by aptness vs. aptness followed by comprehensibility) and error type (metaphors recognized as similes vs. similes recognized as metaphors) as the factors of interest. The results revealed a significant main effect of error type, with the proportion of similes recognized as metaphors being greater than the proportion of metaphors recognized as similes, $F_1 (1, 32) = 21.67, P < 0.001; F_2 (1, 22) = 23.25, P < 0.001$. There was no main effect of order, $F_1 (1, 32) = 2.60$, n.s.; $F_2 (1, 22) < 1$ and no main effect of aptness level, $F_1 (1, 32) = 1.71$, n.s.; $F_2 (1, 22) < 1$. Of particular interest in this study, however, was the significant interaction between aptness level and error type, $F_1 (1, 32) = 14.61, P < 0.001; F_2 (1, 22) = 7.09, P < 0.05$. The key finding is as follows: The proportion of metaphors recognized as similes was greater for the low aptness items than for the high aptness items, while the proportion of similes recognized as metaphors was greater for the high aptness items than for the low aptness items. Evidently, aptness is related to the preference for the metaphor or simile form of a comparison.

Table 1

<table>
<thead>
<tr>
<th>APTNESS</th>
<th>Metaphors recognized as similes</th>
<th>Similes recognized as metaphors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.20 (0.16)</td>
<td>0.58 (0.26)</td>
</tr>
<tr>
<td>Low</td>
<td>0.31 (0.25)</td>
<td>0.40 (0.28)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COMPREHENSIBILITY</th>
<th>Metaphors recognized as similes</th>
<th>Similes recognized as metaphors</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>0.28 (0.18)</td>
<td>0.55 (0.29)</td>
</tr>
<tr>
<td>Low</td>
<td>0.30 (0.25)</td>
<td>0.42 (0.29)</td>
</tr>
</tbody>
</table>

The numbers in parentheses are standard deviations. The interaction between aptness and error type was significant in the subject analysis ($P < 0.001$) and in the item analysis ($P < 0.05$). The interaction between comprehensibility and error type was significant in the subject analysis ($P < 0.05$) but not in the item analysis.
In a similar fashion, let us turn now to the comprehensibility ratings. Is comprehensibility also able to predict the biases in remembering the comparisons? The 30 comparisons were ranked in terms of their mean comprehensibility ratings, and were then split into two groups. The groups were again formed from the 12 items rated highest in comprehensibility and the 12 items rated lowest. (If the 12 items were selected on the basis of comprehensibility ratings of metaphors, 10 of the 12 high-comprehensibility would still be on the list, and 8 of the 12 low. If the 12 items were selected on the basis of the comprehensibility ratings of similes, there would also be 10 of the 12 high and 8 of the 12 low on the list.) The mean comprehensibility rating for the “high comprehensibility” group was 8.35 (S.D. = 0.49), with a range from 7.72 (for the comparison between “sermons” and “sleeping pills”) to 9.30 (for the comparison between “life” and “journey”). The mean comprehensibility rating for the “low comprehensibility” group was 5.67 (S.D. = 1.37), with a range from 3.04 (for the comparison between “smiles” and “razors”) to 6.99 (for the comparison between “alcohol” and “crutch”).

Comprehensibility level was then related to the proportion of similes remembered as metaphors and the proportion of metaphors remembered as similes. This was done using a 2×2×2 repeated measures ANOVA, with comprehensibility level (high vs. low), order (comprehensibility followed by aptness vs. aptness followed by comprehensibility) and error type (metaphors recognized as similes vs. similes recognized as metaphors) as the factors of interest.

The results revealed a significant main effect of error type, with the proportion of similes recognized as metaphors being greater than the proportion of metaphors recognized as similes, $F_1 (1, 32) = 9.05, P < 0.005; F_2 (1, 22) = 11.60, P < 0.005$. There was no main effect of order, $F_1 (1, 32) = 1.21$, n.s.; $F_2 (1, 22) < 1$ and no main effect of comprehensibility level, $F_1 (1, 32) = 3.95$, n.s.; $F_2 (1, 22) = 1.66$, n.s. The interaction between comprehensibility level and error type was significant when analyzed by subject, $F_1 (1, 32) = 4.72, P < 0.05$, but not by item $F_2 (1, 22) < 1$.

Collapsing across order, the proportions of the two kinds of errors for each level of comprehensibility were as follows: The proportion of similes recognized as metaphors was 0.55 (S.D. = 0.29) for the high comprehensibility items and 0.42 (S.D. = 0.29) for the low comprehensibility items. The proportion of metaphors recognized as similes was 0.28 (S.D. = 0.18) for the high comprehensibility items and 0.30 (S.D. = 0.25) for the low comprehensibility items. Thus, the proportion of similes recognized as metaphors was greater for the high comprehensibility items than the low comprehensibility items, but the proportion of metaphors recognized as similes was only slightly higher for the low comprehensibility items than for the high comprehensibility items. These results are summarized in Table 1.

In short, the results of the memory portion of the experiment suggest the aptness of the items appears to be a better predictor of the memory errors than comprehensibility. This is also revealed by estimates of the effect size. The effect size, measured using the eta squared statistic, was greater in the case of aptness than in the case of comprehensibility. The eta squared was 0.13 for the interaction between comprehensibility level and error type when the results were analyzed by subject and 0.07 when analyzed by item. These effect sizes are markedly smaller than the eta
squared for the interaction between aptness level and error type. In this case, the eta squared was 0.31 in the subject analysis, and 0.24 in the item analysis. Thus, aptness can predict the pattern of errors observed in the recognition experiment, and it plays a more robust role than comprehensibility in this regard.

To further examine the issue of the relative importance of aptness and comprehensibility in determining the form in which comparisons are expressed, we related the results of the present experiment to those of a previous study. In Chiappe and Kennedy (1999), participants were shown the same comparisons used in here. However, their participants were asked to judge whether each of the comparisons is best expressed in their metaphor form or in their simile form.

We related the comprehensibility and aptness ratings from the present study to preferences for the metaphor form of the comparisons from Chiappe and Kennedy (1999). An analysis revealed a correlation of 0.75 ($P < 0.001$) between preference for metaphors and mean aptness ratings of the comparisons—as aptness increased so too did preference for the metaphor form. The analysis revealed a correlation of 0.72 ($P < 0.001$) between preference for metaphors and mean comprehensibility ratings, showing that as comprehensibility increased, so too did preference for metaphors.

The correlation analyses demonstrate that both aptness and comprehensibility are important in accounting for preference for the metaphorical form of the comparisons. However, the correlation analyses do not tell us whether one factor is more important in this regard. This was examined further using a commonality analysis (Kerlinger and Pedhazur, 1973). The use of a commonality analysis permits the examination of the unique and common variance that aptness and comprehensibility contributed to the preference for expressing comparisons as metaphors.

Aptness ratings displayed a squared multiple correlation with the mean preference for metaphors of 0.56 (as indicated by the 0.75 correlation above). The 56% of variance in metaphor preference explained by the aptness ratings is decomposed into 5% unique variance ($P < 0.05$) and 51% ($P < 0.001$) variance shared with comprehensibility ratings.

Comprehensibility ratings displayed a squared multiple correlation with the mean preference for metaphors of 0.52 (as indicated by the 0.72 correlation above). The 52% ($P < 0.001$) of variance in metaphor preference explained by the comprehensibility ratings can be decomposed into 0.2% unique variance (n.s.), and 51.4% ($P < 0.001$) variance shared with aptness ratings.

In sum, much of the explained variance in metaphor preference was shared between aptness and comprehensibility ratings, revealing that both factors are important predictors of preference for form of expression. However, aptness ratings contributed variance independent of comprehensibility ratings, although the converse was not true.

4. Discussion

The present study suggests that aptness and comprehensibility both predict preferences for metaphors and similes, but aptness is the more important factor. The
biases in errors in the recognition task, reporting the metaphor or simile forms, could be better predicted from the aptness of the comparisons than from their comprehensibility. The commonality analysis found that although both factors were important predictors, aptness could predict metaphor preference better than comprehensibility. Aptness judgments were also more sensitive to the form of the comparisons than comprehensibility judgments.

The findings also support other important claims about the roles of aptness and comprehensibility. To begin, the aptness judgments took significantly longer to make than the comprehensibility judgments. Moreover, the latency to make the comprehensibility judgments was positively related to the ratings of comprehensibility, but this was not the case for aptness. Furthermore, the comprehensibility judgments were generally higher than aptness judgments, suggesting comprehensibility is a criterion that is easier to satisfy.

The results are consistent with the view that the processes determining aptness require establishing comprehensibility, but that aptness involves something extra. To judge the comprehensibility of a comparison we consider relevant common properties of the vehicle and topic. These properties are also involved in determining the aptness of a comparison. (No doubt, this similarity accounts for the substantial shared variance in metaphor preference accounted for by both factors.) However, the aptness ratings are likely to involve an evaluation of these properties. Is their selection insightful? Are they of consequence? To what extent do the properties matter to our beliefs about the topic? A comparison between cigarettes and time-bombs might not only be understandable, it might capture properties that the subject deems central to the topic. If so, it likely would be deemed not just comprehensible, but highly apt. The aptness ratings would reflect a more extensive analysis of the shared properties than the comprehensibility ratings (Gibbs, 1994). Comparing highways and snakes might be relatively easy, and both are certainly typically long rather than wide, but the selection of these features may not be especially insightful. Instead, the differences may strike one: Road perils are not chiefly due to the road itself, but danger from a snake arises directly from the snake.

The present study found that aptness judgments were higher when made subsequent to rating for comprehensibility and the recognition task. (They increased from mean ratings of 4.76 for metaphors and 4.93 for similes to 6.33 and 6.91 respectively, while comprehensibility ratings did not increase when made second.) Thus, greater familiarity with the items led to higher aptness judgments, consistent with previous findings of positive correlations between these two factors (e.g., Blasko and Connine, 1993; Katz et al., 1988). It could be that requiring subjects to study the tropes in other tasks before making the aptness judgments allowed more features to come into play. For example, “sermons are sleeping pills” may suggest sleepiness of individuals on first examination, and later make connections with broader issues such as “the opium of the people”, keeping the masses quiescent. If so, the increase in aptness when rated second may reflect some new and perhaps even enlightening features coming into play that were not evident when it was judged first. This issue deserves further study.
Furthermore, we found that the similes generally received higher aptness ratings than the metaphors. We believe this is likely due to the particular selection of tropes. While some of our items were virtually unanimously preferred as metaphors (e.g., “life is a journey”) and some virtually unanimously preferred as similes (e.g., “highways are like snakes”), there was a general bias in favor of the simile form of the statements as a group (Chiappe and Kennedy, 1999). Although subjects were asked to judge aptness by considering the extent to which a statement captures important features of a topic, it is possible that in selecting a rating, participants also took into account whether an item was expressed in its most adequate form. Thus, an item that captures few features of the topic and is expressed as a simile might receive higher aptness ratings than if expressed as a metaphor. Conversely, an item that captures many features of a topic and is expressed as a metaphor might receive higher ratings than if expressed as a simile. Because the simile form was generally regarded as most adequate for these items, this would lead to overall higher ratings given to similes. Of course, this is compatible with our finding that greater aptness is associated with a preference for the metaphor form.

4.1. Theoretical implications

The results of the present experiment have implications for theories of metaphor comprehension. In particular, the finding that higher aptness, and to a lesser extent higher comprehensibility, leads to a greater preference for metaphors than similes is inconsistent with the comparison theory. This theory holds metaphors are abbreviated similes (Aristotle, 1926; Fogelin, 1988; Kintsch, 1974; Miller, 1979). As a result, it predicts that preference for metaphors and similes should be constant across levels of aptness and comprehensibility. The only reason a person would choose to express a comparison as a metaphor rather than as a simile is for brevity.

In contrast, the results are consistent with class-inclusion theories (Glucksberg, 2001; Glucksberg and Keysar, 1990, 1993). Metaphors are in the form of categorization statements (“Smiths are apples”), which are stronger than similarity claims (“limes are like apples”). As a result, metaphors are often used when a topic has many pertinent properties associated with the vehicle, and similes are often used when the topic does not have as many of those properties. At first blush, the Glucksberg and Keysar (1990, 1993) version of the categorization theory of metaphor also seems to explain an interesting finding of our experiment—an overall bias to remember the comparisons as metaphors. (It should be noted that Chiappe and Kennedy (1999) found a similar bias using a recall task rather than a recognition task, suggesting that this is a stable effect.) Glucksberg and Keysar (1993) say “Even though the simile ‘cigarettes are like time bombs’ has the surface form of a comparison, it is understood as an implicit categorization” (1993, p. 412). They claim similes are understood by being converted into metaphors and “S is like P will . . . be converted . . . into the original metaphor form, S is P” (p. 408). If similes are converted into metaphors during comprehension this could explain the bias in remembering the comparisons as metaphors.
However, there is reason for skepticism. Although the assumption that similes are converted into metaphors can account for our finding an overall bias, it cannot account for other findings. For example, if true, then we would expect metaphors to be more comprehensible than similes. As Glucksberg and Keysar (1993) say, “The simile...poses a more difficult comprehension problem for a hearer. Hearers must recognize that the comparison...is an implicit categorization...If so, similes may be more difficult to understand than their corresponding metaphors...” (p. 423). We did not find this to be the case. Similes did not receive lower comprehensibility ratings than metaphors, and it did not take participants longer to comprehend similes than metaphors. As a result, the reason why the items tended to be remembered as metaphors remains unclear and further research is required to uncover the source of this bias. For example, it could be a preference for brevity is relevant.

The present research could also be expanded usefully by testing the force of the account in richer discourse contexts than individual tropes appearing in isolation. Context can readily affect the comprehensibility and aptness of figurative claims. For instance, although Malgady and Johnson (1976) found a high correlation between metaphor goodness and the similarity of topic-vehicle pairs when presented alone, other research suggests that there is no such relationship when these same items are embedded in a context (McCabe, 1983). Thus, expressions that might be difficult to understand on their own might be considered easy to understand and quite apt if enough contextual support is provided (Gibbs, 1994). The question arises, then, whether high aptness remains related to preference for the metaphor form of a trope in discourse contexts. Indeed, if so an intriguing ironic reversal is possible: Some poets might deliberately reverse this use of aptness—using similes for highly-apt comparisons and restricting metaphors to low-apt claims. Undermining expectations in fashions such as this is vital for literary liveliness, and can be especially effective when a device is pertinent to a text’s function.

Certainly, the present account needs to be related to poetic and literary uses of metaphor and simile, for both tropes are mainstays of literature. In “Spanish Dancer,” for example, Rilke uses a simile to describe the arms of a flamenco dancer, making peril relevant when he claims they are “like startled rattlesnakes” (Mitchell, 1989). And, of course in “Romeo and Juliet”, Shakespeare uses the metaphor, “Juliet is the sun”. What determines the form that is used in poetic settings, and the effect on the audience? Although aptness may play a crucial role in determining form of expression, it is sure that other factors are in play, notably matters of style, including considerations of rhythm and meter.

No doubt the factors involved are subtle, for within particular cases poets readily switch from one form to the other. For example, in the poem mentioned above, Rilke compares the dancer’s movement to a flame. He writes “As on all its sides a kitchen-match darts white flickering tongues before it bursts into flame: with the audience around her, quickened, hot, her dance begins to flicker in the dark room... One upward glance and she ignites her hair and, whirling faster and faster, fans her dress into passionate flames” (Mitchell, 1989, p. 43). In this case, Rilke introduces the vehicle using an explicit assertion of analogy, a form of similarity, and further develops it using metaphor, by talking about the dance as a flame. What is
the reason for this alternation? Clearly, matters of aptness cannot be the whole story here. But once a simile is used it can set the stage for a second comparison to be offered as a metaphor, even if on its own it might be best presented as a simile. The reason could be that aptness is increased by suitable contexts, much as comprehension is made easier by advance organizers. In short, effects of one comparison on another deserve examination, in addition to studies on comparisons in isolation.

To conclude, although aptness and comprehensibility predict the preferred form of expression of comparisons, and the two judgments are related, aptness plays a more critical role. Our evidence involves preference assessed indirectly via a memory test, and directly through explicit judgments of the best way of expressing a comparison. The results are most consistent with a class-inclusion theory of metaphor, which notes that metaphors have the form of literal categorization claims while similes take the form of assertions of likeness. However, future research will have to determine the role that contextual information plays in metaphor-simile preference, including factors involved in poetic uses of these two tropes.

Acknowledgements

The authors wish to thank three anonymous reviewers for their helpful comments on an earlier draft of this manuscript.

Appendix. Metaphors and similes

Life is (like) a journey; cigarettes are (like) time-bombs; genes are (like) blueprints; education is (like) a stairway; crime is (like) a disease; encyclopedias are (like) gold-mines; nature is (like) a laboratory; alcohol is (like) a crutch; friends are (like) anchors; roosters are (like) alarm-clocks; ideas are (like) diamonds; indecision is (like) a whirlpool; surgeons are (like) butchers; mosquitos are (like) vampires; arguing is (like) war; deserts are (like) ovens; sermons are (like) sleeping pills; words are (like) daggers; salesmen are (like) bulldozers; hearts are (like) closets; desks are (like) junkyards; man is (like) wolf; billboards are (like) warts; jobs are (like) jails; giraffes are (like) skyscrapers; smiles are (like) razors; treetrunks are (like) straws; trees are (like) umbrellas; people are (like) doors; highways are (like) snakes.

References


Kintsch, W., 1974. The Representation of Meaning in Memory. Erlbaum, Hillsdale, NJ.


Dan Chiappe received his PhD in Psychology from the University of Toronto in 1997. Currently he is an Assistant Professor in Psychology at the California State University, Long Beach. He has published numerous articles on the role of comparison and categorization processes in the comprehension of figures of speech. His other research interests include foundational issues in evolutionary psychology, and he is presently working on a book examining the evolution of domain general mechanisms in intelligence and learning.

John M. Kennedy obtained his PhD at Cornell, with James and Eleanor Gibson, whose work on realism and distinctive features still influences his current work on metaphor, symbolism, pictures and other forms of representation. He is currently Full Professor at the University of Toronto in Psychology. His work on pictures, perspective and the blind was singled out by the NY Times and the London Times in December 2002 as a major innovation of the year.

Penny Chiappe received her PhD in Applied Developmental Psychology from the Ontario Institute for Studies in Education/University of Toronto in 1997. Currently she is an Assistant Professor in the Reading Department at the California State University, Fullerton. Her other research interests include the cognitive and linguistic processes involved in reading disabilities and reading acquisition for native English speakers and linguistically diverse children.