When Do Vehicles of Similes Become Figurative? Gaze Patterns Show that Similes and Metaphors are Initially Processed Differently

Frank H. Durgin (fdurgin1@swarthmore.edu)
Swarthmore College, Department of Psychology, 500 College Avenue
Swarthmore, PA 19081 USA

Rebekah Gelpi (rgelpi1@swarthmore.edu)
Swarthmore College, Department of Psychology, 500 College Avenue
Swarthmore, PA 19081 USA

Abstract

Recent emphases on differences between metaphors and similes pose a quandary. The two forms clearly differ in strength, but often seem to require similar interpretations. In Experiment 1 we show that ratings of comprehensibility are highly correlated across simile and metaphor sentences differing only in the presence or absence of "like". In Experiment 2 we show that comprehensibility ratings for figurative forms predict both early (first pass) and late (second pass) fixation durations for metaphor vehicle, but only late fixation durations for vehicles in similes. Simile vehicles appear to initially be processed similarly to literal comparisons, with figurative interpretation occurring later. These observations are consistent with the different pragmatic strengths, and similar interpretations of the two forms.

Keywords: simile; metaphor; analogy; career of metaphor, implicature, eye-movements

Introduction

Theories of figurative speech differ in emphasizing either (abstract) categorization (e.g., Glucksberg & Keysar, 1990) or analogical comparison across domains (e.g., Tourangeau & Sternberg, 1982). Bowdle and Gentner (2005) proposed that unfamiliar figurative usages tend to be preferred in simile form ("Moonlight is like bleach") whereas familiar ones are preferred in metaphor form ("Alcohol is a crutch"). They suggest that the surface form of a simile mirrors the cognitive processing (analogical reasoning) needed for an unfamiliar figurative meaning.

But it isn't the case that mere comparison captures analogical comparison. Aristotle is sometimes described dismissively as a comparison theorist (e.g., Tourangeau & Sternberg, 1982), but Israel, Harding and Tobin (2004) rightly point out that Aristotle is a metaphor-first theorist. Aristotle has also been characterized as considering metaphors to be implicit analogies (Levin, 1982). Similes, on this view, might be figurative to the extent that they require cross-domain implicit analogical reasoning.

Glucksberg and Haught (2006) observed that adding an off-category adjective (e.g., "Moonlight is (*like) romantic bleach") disrupts the preference scheme identified by Bowdle and Gentner (2005). From this Glucksberg and Haught seek to argue that figurative categorization is the only plausible account of metaphor. They argue that their adjectival noun-phrases are dis-preferred in simile form because similes refer to literal referents. But non-existing categories like "romantic bleach" block comparison generally, and do not differentiate literal from figurative comparison. Moreover, they seem to violate the need for distinct domains required for analogy to work.

In this paper we will adopt the strategy of contrasting online comprehension of similes both with metaphor and with literal comparisons (as recommended by Israel et al., 2004). By this means we will test whether similes are simply interpreted literally, as Glucksberg and Haught (2006) argued, or if they also differ from literal comparisons in ways that clarify their normal designation as figurative.

In Experiment 1 we will show how similar the comprehensibility ratings of associated metaphor and simile forms are. Despite previously recognized differences in strength (e.g., Glucksberg & Keysar, 1990), aptness (e.g., Kennedy & Chiappe, 1999), and preference (e.g., Bowdle & Gentner, 2005), we find that comprehensibility judgments are relatively similar across both simile and metaphor forms for simple vehicles without adjectival modification. This suggests similar interpretations are reached for the figurative meaning of the vehicle in each figurative form.

In Experiment 2, we will use measures of gaze during reading to show that initial reading of similes more closely resembles that for literal comparisons than that for metaphors. Metaphors show reliable effects of figurative comprehensibility during initial first-pass reading; similes and literal comparisons do not show such comprehensibility effects early on. However, during second-pass reading, fixation durations on simile and metaphor vehicle both show strong correlations to ratings of figurative comprehensibility. This suggests that the figurative interpretation of a simile vehicle may simply be computed later than that of a metaphor.

Experiment 1: Comprehensibility Ratings

Methods

Materials (for Experiments 1 and 2) Seventy-five metaphors were gathered or updated from previous studies (Katz, Paivio, Marschark & Clark, 1988; Bowdle & Gentner, 2005; Thibodeau and Durgin, 2011). For gaze analysis (Experiment 2), each sentence was extended with a few words intended to be largely neutral with respect to

interpreting the sentence (e.g., "A smile is (like) a magnet for people."). A set of literal comparison statements was developed using the same vehicles (e.g., "A black hole is like a magnet in space."). To balance the number of sentences that did not include "like" and were not figurative, we also included 25 literal categorization statements filler sentences unrelated to the 75 experimental items. Three lists of 100 items were constructed in which one third of the 75 items appeared in each of the three forms (metaphor, simile, literal comparison) along with the 25 fillers.

Participants and Task A total of 91 adults were recruited through Amazon's Mechanical Turk to make ratings of comprehensibility of the critical word in each sentence. A third were assigned to each of the three lists of stimuli.

Results and Discussion

As shown in Figure 1, comprehensibility ratings of metaphor and simile vehicles were highly correlated across items, $R_{73} = 0.85$, $p < .001$. A common factor, produced by averaging the figurative rating sets by item accounted for 92% of the variance in the simile ratings and 93% of the variance in the metaphor ratings. There was, of course, no correlation between this common rating measure for figurative vehicles and the ratings given for the same words in literal comparisons with the same vehicles, $R_{73} = 0.01$, ns.

It appears that ratings for both figurative sentence forms are typically based on similar figurative meaning, defined in relation to the topic.

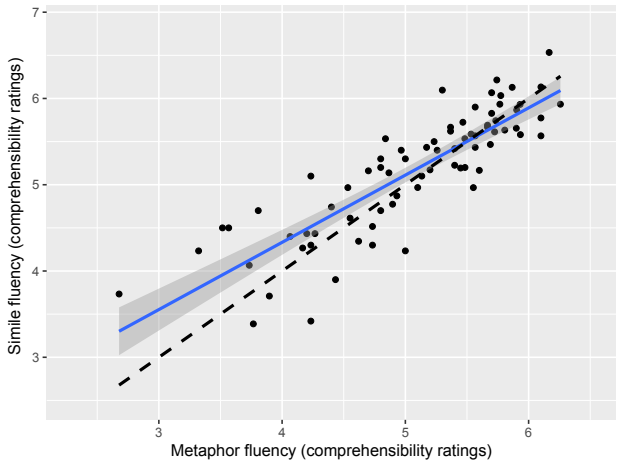

Figure 1. Correlation between comprehensibility ratings for each figurative vehicle across figurative sentence forms.

Experiment 2: Gaze Measures when Reading Comparisons, Similes and Metaphors

The rating data of Experiment 1 are consistent with the assumption of many scholars that there is good reason to suppose that similar figurative meanings are often achieved by sentences in simile and metaphor forms, even if they may sometimes diverge in understanding. However the rating data reflect post-interpretive evaluations and do not bear on the question of whether the initial cognitive encounter with the vehicle (e.g., during reading of the metaphor) is substantially different in similes and metaphors. In order to better understand how comprehension may unfold differently for the two figurative forms, we next measured gaze patterns during reading of these same sentences.

Methods

Participants Thirty-six Swarthmore College undergraduate students who were native English speakers participated in partial fulfillment of an introductory course requirement.

Design and Procedure The linguistic materials were identical to those used in Experiment 1, except that 6 practice items were developed to allow participants time to adapt to the task. One third of the 75 experimental items appeared in each of three forms (metaphor, simile, literal comparison) for each participant, according to one of three lists, and were randomly ordered and intermixed with 25 (filler) literal categorization sentences. Sentences were presented one at a time on a monitor in front of the participant after first establishing gaze on a fixation point just to the left of the presentation of the sentence. Participants were to read the sentence and respond by pressing a key when they had comprehended it. The sentence was then removed, and an easy multiple-choice comprehension test followed, asking which of four terms was most relevant to the meaning of the sentence they had just read (e.g., for "A smile is (like) a magnet for people", the correct answer was "attract"). Subjects made their choice using a game-pad with an appropriate spatial mapping to the choices on the screen. The entire experimental session took about 20 minutes.

Gaze Recording Gaze was tracked at 1000 Hz, using an Eyelink 1000 (SR Research). The experimental code was implemented in Experiment Builder (SR Research), and gaze parameters during reading were extracted using custom software in conjunction with the area-of-interest definition tools provided by Experiment Builder.

Results and Discussion

Analysis Strategy Our principal interest was to compare the immediate processing of similes to the processing of metaphors, as measured by gaze parameters, and to secondarily use literal comparisons as an alternative comparison condition for similes. We used item-wise rating data from Experiment 1 as a predictor in LMER models. For the figurative items these ratings were averaged across figurative conditions. For gaze variables, we first used model comparisons to test whether the ratings had predictive value that differed between figurative sentence types. For example, if metaphor vehicles are treated as figurative words, while simile vehicles are treated as literal words, we might expect ratings of figurative interpretability from Experiment 1 to predict only the metaphor forms, and not the simile forms. To test for this we compared LMER models that included an interaction term (between ratings

and sentence type) with those that did not. Such model comparisons produce a Chi-square statistic. We also used LMER modeling to compare similes with literal comparisons. For overall comprehension time, we conducted a single overall analysis since comprehension time was expected to be correlated with comprehensibility ratings for all sentences.

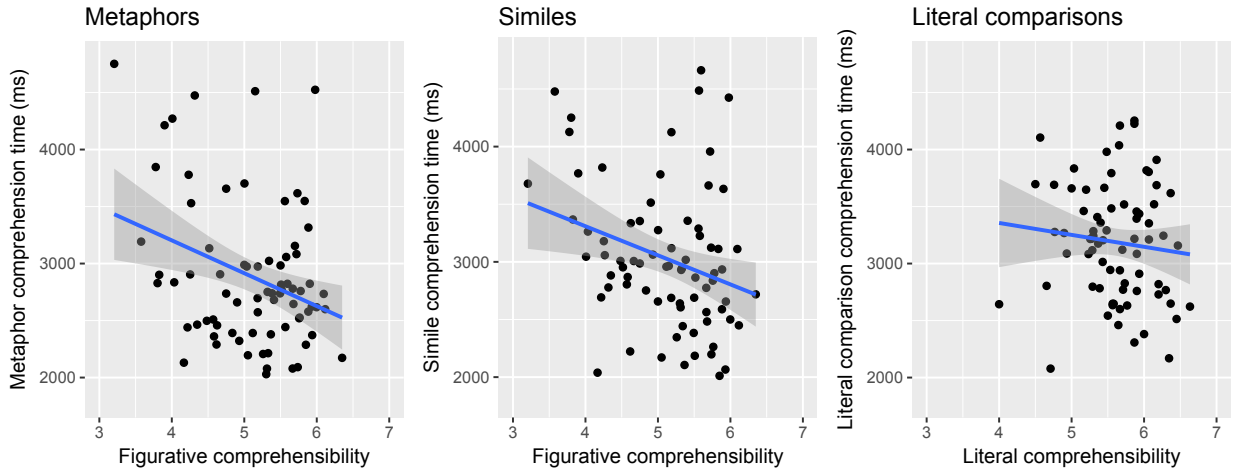

Figure 2. Mean comprehension times (mean response latency computed in log space by item) for metaphors (left), similes (middle), and literal comparisons (right) as a function of item comprehensibility (ratings). Best fit and SE shown.

Comprehension Time Participants' main task was simply to indicate comprehension of the sentence after reading it by pressing a key. The distribution of times was skewed, so centered log transformations of these times were used for statistical modeling. The main LMER model included sentence type (metaphor, simile and comparison) and comprehensibility ratings from Experiment 1 as predictors.

Error terms included subjects and item as well as the slopes for sentence form by subjects and by items, and the slopes of ratings by subjects. Model comparisons showed that including the interaction between sentence type and rating did not explain reliably more variance than a model without the interaction, $X^2(2) = 0.27$, $p = .875$, indicating that the relation to ratings did not differ by sentence type. Rather for all three sentence types, ratings predicted comprehension time, $t(117.3) = 3.54$, $p < .001$ (Satterthwaite approximations of *df* will be reported throughout; see Luke, 2006). However, compared to the similes, response times were reliably longer for the literal comparison sentences, $t(48.6) = 3.94$, $p < .001$. Consistent with effects previously observed for apt or conventional figurative vehicles (e.g., Bowdle & Gentner, 2005; Glucksberg & Haught, 2006), comprehension time was marginally shorter for the metaphors than similes, $t(35.1) = 1.93$, $p = .062$. The observed relationship of rated comprehensibility and comprehension time is shown in Figure 2 separately for each sentence type. These data reflect the expected relationships between comprehensibility ratings and comprehension time.

Gaze Behavior: Overview. The analysis of gaze behavior data is organized into five reading events relative to the target word (i.e., the vehicle) region: (1) Duration of initial fixation(s) on the critical word, (2) the subsequent frequency of regressions back out to the left (3) the duration of time between first fixating the critical word and finally reading past the word, (4) the likelihood of refixating the word after passing it, and (5) if refixation occurred, the total time taken fixating the word after first passing it.

Our primary interest is in differences associated with the presence of "like" in advance of the figurative word, since the simile and metaphors forms used are otherwise identical. Comparisons between the literal comparisons and similes are also of interest, however, given that we are asking whether the figurative vehicle words in similes are initially processed literally.

Gaze Data Transformation and Truncation Duration data associated with gaze patterns were also log transformed to reduce skewing and were centered for analyses. Transformed durations that were more than 4 standard deviations above the transformed mean for that measure were truncated to 4 SDs. The proportion of data affected by this method was less than 1% each measure discussed. Ratings were centered (by subtracting off the mean) prior to analysis.

Gaze Behavior 1: First Fixation Duration. If participants initially seek to understand metaphor vehicles figuratively, but simile vehicles literally, the duration of their first fixation on the critical word might correlate with ratings of comprehensibility for figurative items only for metaphors.

Consistent with this hypothesis, comparison of LMER models of the figurative sentence data, with and without interactions terms, showed that the relationship of FFD to ratings differed for the two figurative forms, $X^2(1) = 7.32$, p = .007. As expected, separate models of the two condition indicated that FFD for the figurative vehicle was related to

the rated comprehensibility in the metaphor form, $t(35.6) = 4.06$, $p < .001$, and not in the simile form, $t(68.0) = 0.10$, ns. Models of FFD contrasting the two comparison sentence forms (literal comparisons and similes), found no differences in FFD between the two forms ($X^2(2) = 2.74$, $p = .254$. The data for each form are plotted in the top panels of Figure 2. This pattern is consistent with the idea that the figurative vehicle in a simile is initially treated as a literal referent, as argued by Glucksberg and Haught (2006a). In contrast, initial encounters with metaphor vehicles produced effects consistent with an immediate search for a figurative interpretation.

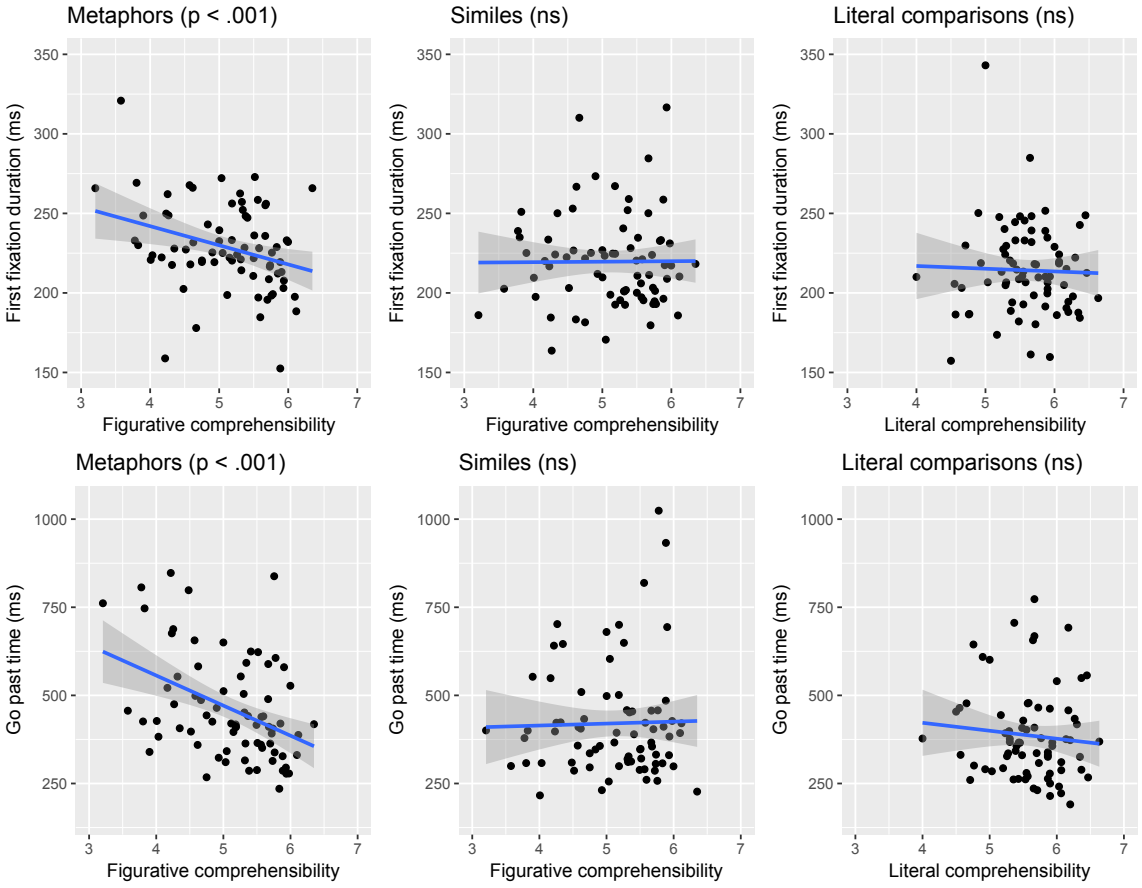

Figure 3. Top row: First fixation duration (FFD). Bottom row: Go Past Time (GPT). Geometric means (by item and sentence type) for the figurative vehicle (or equivalent) are shown as a function of mean rated comprehensibility.

Gaze behavior 2: Go past time. Go past time (GPT) is defined as the entire duration from first entering the critical word until finally passing to the right of the critical word. GPT for each sentence form is shown in bottom panels of Figure 3 as a function of comprehensibility rating. Because GPT includes FFD, FFD was included as a covariate in LMER analyses of GPT (the analyses come to the same conclusions without the covariate).

Again, we first sought to test whether there were reliable relationships between GPT and rated comprehensibility for figurative items, and, if so, whether these differed between the various forms. Indeed, model comparisons indicated that the relation of GPT to rated comprehensibility differed reliably between simile and metaphor forms, $X^2(1) = 9.8$, $p = .002$. In these models, GPT was also reliably longer for metaphors than similes, $t(37.6) = 2.84$, $p = .007$. Separate LMER models showed that, for metaphor sentences, GPT was reliably related to figurative comprehensibility ratings, $t(64.4) = 2.60$, $p = .011$. This was not the case for similes, $t(74.1) = 1.49$, $p = .141$ (where the trend was in the opposite direction, consistent with delays for highly conventional metaphors presented as similes).

Models comparing GPT for the similes and literal comparison sentences5 indicated that including the interaction of ratings and sentence type in the models made no reliable difference, $X^2(1) = 2.75$, $p = .097$. A separate model of literal comparisons confirmed that here was also no reliable relationship between GPT and rated comprehensibility, $t(76.9) = 0.70$, $p = .486$. Thus, prior to exiting the critical word to the right in the simile form, the figurative vehicle may still be treated primarily as a referent to the literal comparison category as the reader passes on the rest of the sentence.

Gaze Behavior 3: Returns to the Target Word From the Right Participants often returned to the critical word after they had already read beyond it. Indeed, this happened in roughly 59% of trials (60% of metaphor trials, 58% of literal comparison trials and 57% of simile trials). Is the likelihood of such *Regressions In* (RI: if the critical word was refixated on a given trial after exiting to the right) related to rated comprehensibility? LMER models of RI for the figurative sentences showed a reliable relationship existed between rated comprehensibility and the likelihood of RI, $t(59.9) = 2.29$, $p = .026$, and that this effect did not reliably differ between metaphors and similes, $X^2(1) = 0.003$, *ns*. But LMER models of RI for comparison sentences also showed a reliable main effect of rated comprehensibility, $t(38.4) = 2.27$, $p = .029$, and no evidence of an interaction, $X^2(1) = 0.04$, *ns*. Thus, RI was more likely, for all sentence forms, as comprehensibility decreased.

Gaze Behavior 4. Second Pass Total Time (P2TT) Given that a reader had refixated the critical word after having read more of the sentence, was the total time spent refixating it before responding related to rated comprehensibility? Total comprehension time was included as a covariate because it was highly correlated with P2TT. Whereas FFD and GPT both distinguished similes from metaphors, understanding a simile typically requires reaching a similar understanding to the understanding required for a metaphor. When during reading comprehension might this happen?

LMER models of P2TT for the figurative sentences indicated that the relation between P2TT and comprehensibility ratings was highly reliable for these two forms, $t(63.0) = 2.77$, $p = .007$), but did not differ between the figurative sentence forms, $X^2(1) = 0.37$, p = .548. Thus, P2TT appears to be similarly related to comprehensibility ratings for similes and metaphors, as shown in Figure 4. In contrast, LMER models of the comparison statements (i.e., similes and literal comparisons analyzed together) failed to show reliable relationship between ratings and P2TT, $t(39.9) = 1.71$, $p = .095$, but also failed to detect reliably different effects of ratings for similes and literal comparisons, $X^2(1) = 0.36$, p = .548.

To resolve the mixed evidence regarding similes in these two analyses, we modeled the effect of ratings on each sentence type separately, both with and without the total comprehension time as a covariate. For literal comparisons, there was no evidence that P2TT was related to comprehensibility ratings either with the covariate included, $t(64.3) = 0.56$, $p = .580$, or without it, $t(59.9) = 0.99$, $p = .327$. Conversely, consistent with overall analyses of figurative sentences, in individual analyses of each of the figurative forms P2TT was similarly, but weakly related to comprehensibility when the covariate was included (metaphor: $t(49.3) = 1.82$, $p = .075$; simile: $t(37.5) = 1.94$, $p = .060$) and highly reliably related to comprehensibility when the covariate was not included in the models (metaphor: $t(34.4) = 3.13$, $p = .004$; simile: $t(58.0) = 2.92$, $p = .005$). Recall that the overall relationship of comprehensibility and P2TT for figurative items in our combined analyses was reliable even with total response time included as a covariate (i.e., p = .007, above). A similar analysis without the covariate also provided strong evidence of an overall relationship, $t(52.4) = 2.99$, $p = .004$.

Does P2TT help to explain overall response time differences for figurative items? To test whether P2TT, itself, can account for longer overall overt comprehension responses, a new analysis of overall response time was conducted (for trials displaying RI) with and without P2TT as a covariate. Without P2TT included, there was strong evidence of a relationship between ratings of comprehensibility and comprehension time for these trials, $t(23.7) = 3.08$, $p = .005$, but when P2TT was included as a covariate, no evidence of the relationship between response time and comprehensibility remained, $t(58.1) = 1.50$, $p = .138$. For figurative items, then, it appears that P2TT likely represents time used for computing a figurative interpretation of the sentence.

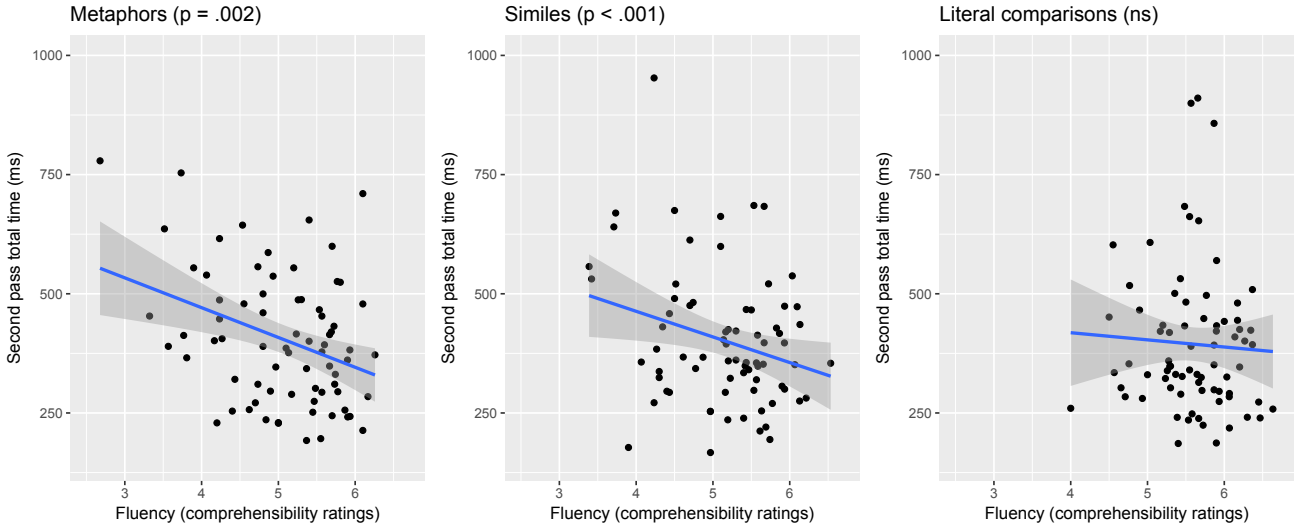

Figure 4. Geometric mean second pass total gaze duration (P2TT) by item as a function of rated comprehensibility for each sentence form. Only the 1431 trials where regression into the critical word occurred are included.

General Discussion

In Experiment 1, we observed that ratings of comprehensibility for vehicles in simile or metaphor form were highly correlated. The average figurative ratings from Experiment 1 were used in Experiment 2 to try to predict gaze variables related to the figurative vehicle during reading. We reasoned that the similar comprehensibility ratings of similes and metaphors observed in Experiment 1 reflected similar ease or difficulty with deriving the appropriate figurative meaning of the vehicle; we sought evidence of when this might unfold during reading.

Gaze patterns for metaphor vehicles, but not simile vehicles, reflected rated figurative comprehensibility from the very first fixation. Metaphor vehicles that were judged less comprehensible were subjected to longer initial periods of analysis. In contrast, similarly-rated similes, did not show immediate effects. For similes, as for literal comparisons, initial measures of processing time for their vehicles in similes were unrelated to rated comprehensibility.

But simile processing resembled metaphor processing during second-pass reading of the sentence. Both simile and metaphor vehicles showed comprehension-related durations of fixation. These second-pass durations were related to the rated comprehensibility of the word in the sentence both for metaphors and for similes. This pattern was not found for literal comparisons.

The difference between similes and metaphors at first fixation might be regarded as reflecting the weaker pragmatic assertion involved in declaring that something is *like* something rather than that it *is* something (Rubio-Fernández, Geurts & Cummins, 2016). To say that something is like something else implies that it is also unlike it. In this sense similes are sensibly experienced as weaker than metaphors at first pass, even if the ultimate interpretation of what is being said about the topic ultimately requires accessing a figurative or abstract interpretation of the vehicle.

Overall, these data suggest that similes are initially treated similarly to literal comparisons, consistent with the arguments of Glucksberg and Haught (2006). However, the second pass data and the ratings data both suggest that sentence comprehension for simile forms still requires identifying a figurative interpretation. We think this supports Aristotle's assertion of the figurative nature of simile that is embedded within his longer discussion of metaphor. Aristotle (400 BC/1991) wrote "A simile is also a metaphor, for there is little difference." This quote clearly implies that metaphor (literally a "carrying-over" of meaning) is the larger category.

Our study has used similes derived from metaphors. The data show that reading such similes differs substantially from reading their corresponding metaphors. However, the data also support the idea that the interpretive demands for the figurative vehicle may ultimately be similar in both forms. This distinguishes simile from literal comparison.

Acknowledgments

This research was supported by a faculty research grant to FHD from Swarthmore College. RG was supported by a Frances Velay Fellowship.

References

Aristotle (1991). *On rhetoric: A theory of civic discourse* (G. A. Kennedy, translator). New York: Oxford University Press. (Original work published ~400 BC).

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

Carston, R., & Wearing, C. (2011). Metaphor, hyperbole and simile: A pragmatic approach. *Language and Cognition*, *3*, 283-312.

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. (2000). Are metaphors elliptical similes? *Journal of Psycholinguistic Research*, *29*, 371-398.

Glucksberg, S., & Haught, C. (2006). On the relation between metaphor and simile: When comparison fails. *Mind & Language*, *21*, 360-378.

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

Israel, M., Harding, J. R., & Tobin, V. (2004). On simile. In M. Achard and S. Kemmer (eds) *Language, Culture, and Mind*, (pp123-135). CSLI Publications.

Katz, A. N., Paivio, A., Marschark, M., & Clark, J. M. (1988). Norms for 204 literary and 260 nonliterary metaphors on 10 psychological dimensions. *Metaphor and Symbol*, *3*, 191-214.

Levin, S. R. (1982). Aristotle's theory of metaphor. *Philosophy & Rhetoric, 15*, 24-46.

Luke, S. G. (2016). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, 1-9. DOI:10.3758/s13428-016-0809-y

Rubio-Fernández, P., Geurts, B., & Cummins, C. (2016). Is an apple like a fruit? A study on comparison and categorisation Statements. *Review of Philosophy and Psychology*, 1-24. DOI: 10.1007/s13164-016-0305-4

Thibodeau, P. H., & Durgin, F. H. (2011). Metaphor aptness and conventionality: A processing fluency account. *Metaphor and Symbol*, *26*, 206-226.

Tirrell, L. (1991). Reductive and nonreductive simile theories of metaphor. *The Journal of Philosophy*, *88*, 337-358.

Tourangeau, R., & Sternberg, R. J. (1982). Understanding and appreciating metaphors. *Cognition*, *11*, 203-244.