Linguistic Features of Sarcasm and Metaphor Production Quality

Stephen Skalicky and Scott A. Crossley

Department of Applied Linguistics
Georgia State University
scskalicky@gmail.com, scrossley@gsu.edu

Abstract

Using linguistic features to detect figurative language has provided a deeper insight into figurative language. The purpose of this study is to assess whether linguistic features can help explain differences in *quality* of figurative language. In this study a large corpus of metaphors and sarcastic responses are collected from human subjects and rated for figurative language quality based on theoretical components of metaphor, sarcasm, and creativity. Using natural language processing tools, specific linguistic features related to lexical sophistication and semantic cohesion were used to predict the human ratings of figurative language quality. Results demonstrate linguistic features were able to predict small amounts of variance in metaphor and sarcasm production quality.

1 Introduction

Computational approaches to figurative language identification and classification are becoming increasingly more sophisticated (e.g., Khodak et al., 2017). While these studies have produced computational models capable of predicting figurative from non-figurative language, these models typically have little to say regarding the *quality* of figurative language. However, it is important to consider the potential ways that linguistic features differ based on higher or lower quality examples of figurative language to better understand the linguistic nature of figurative language. Thus, the purpose of this study is to test whether linguistic features can be used to predict the quality of metaphor and sarcasm production, which are two types of figurative language. Specifically, this study investigates whether linguistic features related to lexical sophistication and semantic cohesion are predictive of human ratings of metaphor and sarcasm production quality. Because our purpose is not to develop models capable of differentiating between figurative and non-figurative language, we do not take a traditional classification approach that is commonly seen in computational figurative language research.

Creativity and Figurative Language. Creativity can be operationalized as an effective and original solution to a problem (Runco and Jaeger 2012), and figurative language is an example of linguistic creativity (Gerrig and Gibbs 1988). One method to operationalize the quality of figurative language is to consider the creativity of individual examples of figurative language. Because language associated with more creative ideas has been linked to greater conceptual distance via semantic network modeling (Acar and Runco 2014; Dumas and Dunbar 2014), as well as greater lexical sophistication via more diverse vocabulary and lower word frequency (Skalicky et al., 2017), it follows that figurative language (e.g., metaphors and sarcasm) quality may also be predicted using linguistic measures related to lexical sophistication and semantic cohesion.

Metaphor Quality. Although conceptual metaphors are defined as the mapping of one conceptual domain onto another, this mapping must also be apt and meaningful (Gibbs 1994; Glucksberg 2001). Moreover, metaphors do not need to include large gaps in conceptual domains in order to be defined as a metaphor. Indeed, the ability to create descriptive links between seemingly disparate concepts is fundamental to metaphor production (Kintsch 2008; Kintsch and Bowles 2002), and therefore metaphors with greater conceptual distance may also be more effective.

Sarcasm Quality. Sarcasm is best defined as specific instances of verbal irony which serve to provide ironic criticism or praise that is somehow contrary to reality (Colston 2017). Sarcasm naturally involves some sort of incongruity between what is said and the situation in which sarcasm is used.

Proceedings of the Workshop on Figurative Language Processing, pages 7–16
New Orleans, Louisiana, June 6, 2018. ©2018 Association for Computational Linguistics

Thus, one way to measure the effectiveness of sarcasm is to determine how incongruent a sarcastic statement is within a respective context.

Participants. A total of 61 participants were recruited for this study (46 females and 15 males). Participant age ranged from 17 to 63 (*M* = 25.56, *SD* = 8.341*)*. The participants were recruited from the undergraduate and graduate student population at a large public university in the southeastern United States. Participants were compensated for their participation in the experiment.

We opted to recruit our own set of participants and create a new corpus of sarcasm and metaphor for several reasons. First, doing so allowed us to gather additional measures from the participants, including measures of individual differences, linguistic features, and language background. Secondly, we were also able to capture behavioral information, such as how long it took participants to produce their metaphorical and sarcastic answers. Finally, we were able to ensure the participants were aware that their task was to provide metaphor and sarcasm, and provided definitions for doing so, which in turn allowed us to focus on the main purpose of this investigation (i.e., measuring differences in figurative language quality).

Metaphor Production Items. Two different metaphor production tasks were developed from previously used metaphor stimuli (Beaty and Silvia 2013; Chiappe and Chiappe 2007). First, a conventional metaphor task was designed containing 22 different items. Each item consisted of a Topic and a Description. All of the Topics were nouns (e.g., *her family*), and all of the Descriptions were descriptions or properties of those nouns (e.g., *something that keeps her stable and prevents her from drifting into danger*). Participants were instructed to use the Description of the Topic to write a metaphor reflective of the same meaning in the Description, but without reusing any of the words from the Description. In addition, a novel metaphor task was used, where participants were presented with two scenarios: the most boring class they have attended, and the most disgusting item they have ever eaten or drunk. For each scenario, participants were instructed to produce a metaphor that described their feelings during that scenario and were also provided with an example of how to start their metaphors (e.g., *Being in that class was like* ____).

Sarcasm Production Items. Twelve different drawn cartoons were adapted or created to serve as sarcasm production prompts. Four of these items were black and white cartoons used by Huang et al. (2015) to prompt sarcastic responses, each taken from the Rosenzweig Picture Frustration Study, originally designed to assess patient responses to frustrating situations in order to diagnose aggression (Rosenzweig 1945). Each of the black and white cartoons is a single-panel cartoon which depicts a frustrating situation with more than one speaker (e.g., one person's car breaks down and thus two people missed their train). The person responsible for the frustration is shown saying something, whereas the victim of the frustration is presented with a blank speech bubble. Four additional items were created by revising four single-panel *Bizarro!* comics. *Bizarro!* is a single-panel comic strip created by Dan Piraro that is syndicated daily in print newspapers across the United States. *Bizarro!* comics typically depict absurd or otherwise unlikely situations for the purpose of humor, social commentary, or both (www.bizzaro.com). The specific *Bizarro!* comics used in this study were four desert island comics, which each depicted two people stranded on a small desert island in the middle of an ocean. The original cartoons all contained a single speech bubble for one of the speakers, which was made blank for the purposes of this study. Finally, an additional four sarcasm production items were developed by creating original comics each comprised of three panels with two speakers. In each comic, the first two panels set up an initial situation (e.g., a young man is recruited to join the army and is guaranteed to travel the world in an exciting manner by a military recruiter), while the final panel includes one of the speakers with an empty speech bubble in a situation designed to prompt a sarcastic response (e.g., the young man ends up peeling potatoes instead of traveling the world). For each of the twelve comics, participants were instructed to imagine they were the speaker with the empty speech bubble and to write something sarcastic they would say if they were in that situation.

1.1 Procedure

Participants were recruited to complete the metaphor and sarcasm production tasks in a single laboratory session. The researcher briefly described the procedure of the experiment. Participants then began the production test and were randomly assigned to take the metaphor or the sarcasm production task first.

Metaphor Production. During the metaphor production task session, participants were first provided with a definition of metaphor: *A metaphor is a comparison between two things in order to help describe something*. Then, during each trial, the screen displayed the Topic and Description in clearly marked areas, with a blank text box for the participants to type their metaphor using the keyboard. After completing all 22 conventional metaphor prompts, participants then completed the two novel metaphor situations in a randomized order.

Sarcasm Production. During the sarcasm production task, participants were provided with a definition of sarcasm: *Sarcasm is a form of indirect language. When someone is being sarcastic, they mean something different than what they literally said*. Each trial involved one of the 12 comics randomly displayed above a text box, with a reminder asking participants to supply a sarcastic comment for the situation depicted in the comic. After typing their sarcastic statement into the answer box, participants pressed the Enter key to move on to the next comic until they completed all 12 comics (in a random order).

Each participant completed all of the metaphor and all of the sarcasm prompts in a random order within each block. Any answers that were indicative of a lack of attention or were not direct responses to the prompt (e.g., the participant did not attempt to create a metaphor) were discarded, leaving a total of 1304 metaphors and 716 sarcastic responses.

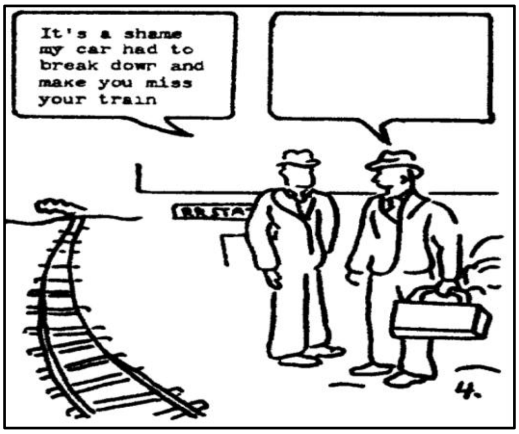
It's a shame my car had to break down and make you miss your train
RR STA
4.

Figure 1. Example sarcasm production item

Human Ratings. An analytic rubric was created in order to obtain measures of figurative language production quality for the metaphors and sarcastic responses provided by the participants. The rubric contained separate sections for metaphor and sarcasm, and was comprised of three separate subscales designed to capture metaphor or sarcasm quality based on participants' ability to develop accurate, effective, and original examples of metaphor and sarcasm. Accuracy was related to theoretical definitions of metaphor (conceptual distance) and sarcasm (incongruity), while effectiveness and originality were related to theoretical definitions of creativity (i.e., novelty and mirth). Accordingly, the metaphor section included the subscales *Conceptual Distance, Novelty,* and *Mirth*, and the sarcasm section included the subscales *Incongruity*, *Novelty*, and *Mirth*. Novelty refers to originality. Mirth is an emotional reaction typically associated with humor, wherein one can experience slight amusement to intense hilarity arising from humorous or playful stimuli (Martin 2007).

Each subscale was measured using a range of one through six, with a score of one meaning the example of figurative language did not meet the criterion in any way and a score of six meaning the answer met the criterion in every way. Two human raters were recruited to provide ratings of the participants' metaphor and sarcastic responses using this analytic rubric. After initial ratings, a third rater (i.e., the first author) adjudicated any disagreements of two points or greater for all of the subscales, resulting in the following adjudicated kappa levels of .872 for metaphor conceptual distance scores, .854 and .855 for metaphor novelty and metaphor mirth, .835 for sarcasm incongruity, and .783 and .777 for sarcasm novelty and sarcasm mirth. After adjudication, the raters' scores were averaged to provide a single score per subscale per item.

1.2 Linguistic Features

The metaphors and sarcastic responses produced by the participants were analyzed for lexical sophistication and semantic cohesion using two text analysis tools: The Tool for the Automatic Analysis of LExical Sophistication (TAALES; Kyle et al., 2017) and the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley et al., 2016), respectively. These tools read in raw text files and use existing taggers (e.g. Stanford

CoreNLP) and dictionaries (e.g., Corpus of Contemporary American English frequency values, MRC Psycholinguistic Database, WordNet Lexical Database) to provide a comprehensive output for a broad range of NLP features. Details regarding the construction and validation of these tools can be found in their respective citations.

Lexical Sophistication. Lexical sophistication is a measure of how complex a text is. For instance, texts with more diverse vocabulary, lower frequency words, and words that take longer to process in the mental lexical all contribute to a text's level of lexical sophistication. To date, very few studies have investigated lexical sophistication in the context of figurative language, aside from one study reporting that satirical product reviews were less concrete than non-satirical product reviews (Skalicky and Crossley 2015). Thus, there is a need to perform more investigation into lexical sophistication and figurative language in order to better determine if these features interact with perceptions of figurative language quality. This study includes broad measures of lexical sophistication related to lexical frequency, psycholinguistic properties of words, and word exposure in order to investigate and report any initial links between figurative language production quality and lexical sophistication.

From TAALES, several indices representative of lexical sophistication were calculated. First, measures of psycholinguistic properties of words were gathered because these measures represent cognitive representations of lexical items and can be used to assess the relative sophistication of lexical items (Kyle and Crossley 2015). Specifically, these measures were word Familiarity, Concreteness, Imageability, and Meaningfulness. Word Familiarity represents how familiar one is with a specific word, with more familiar words being words that are also more commonly encountered, making familiarity similar to word frequency. Word Concreteness refers how perceptible an entity associated with a particular word is (Brysbaert et al., 2014). For example, the word *dog* is more concrete than the word *music*. Word Imageability represents the ease of conjuring a mental image of a word, with words like *tree* being more imageable than words such as *abatement* (Salsbury et al., 2011). Word Meaningfulness represents how many different associations to other words a particular word has. For example, a word such as *tree* has more associations (e.g., *branch*, *leaf*, *wood*) than a word such as *savant*, which activates fewer associations (Salsbury et al., 2011). Measures of word Imageability, Familiarity, and Meaningfulness were all calculated based on the MRC Psycholinguistics Database norms (Coltheart 1981), which is a curated compilation of previous rating studies for these features. Word Concreteness values were calculated using the Brysbaert Concreteness norms (Brysbaert et al., 2014), which were derived from human ratings of word concreteness using online crowdsourcing.

In addition to those indices, linguistic features related to word exposure and use were also collected, as these represent the relative frequency of occurrence and use for certain words. These indices were spoken word frequency, semantic diversity, and age of acquisition. Spoken word frequency was calculated using counts from the spoken portion of the Corpus of Contemporary American English (COCA; Davies 2008). Semantic Diversity represents the number of different words contexts a particular word typically occurs in, and thus represents specificity of word meanings. Semantic Diversity was calculated for each word using the norms published by Hoffman et al. (2013). To calculate Semantic Diversity, Hoffman et al. (2013) separated the British National Corpus into chunks of 1,000 words, and then analyzed the total number of these 1,000 word contexts any particular word occurred in, as well as the semantic similarity of each word to all of the other words in those contexts. The end result is that words with higher Semantic Diversity can be used in more contexts and have more variable meanings than those with lower Semantic Diversity. Finally, Age of Acquisition (AoA) values represent human intuition regarding the age when they first learned a particular word. AoA values based on Kuperman et al., (2012) were used, which were collected using a large number of human raters via online crowdsourcing. All of these linguistic indices were calculated based on content words only.

Cohesion. TAACO was used in order to calculate semantic overlap between prompts and participant answers for the metaphors only. Distance between concepts used in metaphors has been accurately modeled using measures of semantic association, such as Latent Semantic Analysis (Kintsch 2008; Kintsch and Bowles 2002), and therefore a measure of semantic distance was included in this study in order to determine if distance between concepts influences human perceptions of metaphor production quality. To do so, the participants' metaphors were grouped by prompt and analyzed separately using the source text analysis option in TAACO. This option allows the user to load in a source text as a reference text for other texts to be compared against for semantic and cohesive similarity or differences. For each group of metaphors, the Description provided to the participants was loaded as the source text, and the participant's metaphor were analyzed to gather the amount of semantic overlap between participants' answers and the prompts using the word2vec measure in TAACO. Word2vec models the semantic direction and magnitude of words as they relate to other words (known as *vectors*). By modeling words as vectors, word2vec assumes words more closely grouped together are more semantically related than those that are further apart and employs predictive modeling in order to calculate the semantic relations among words in a text.

1.3 Statistical Analysis

The human ratings of figurative language production quality were first analyzed using Principle Component Analysis (PCA) in order to obtain weighted component scores of figurative language production quality for both the metaphors and the sarcastic responses. Afterwards, a series of linear mixed effects (LME) regression models were fit to determine if any of the linguistic features were predictive of figurative language production quality scores. For each LME model, the figurative language production quality score was entered as the dependent variable and the linguistic features were added as the independent predictor variables (also known as fixed effects). For metaphors, metaphor type (novel vs. conventional) was also added as a fixed effect, and for sarcastic responses, sarcasm prompt type was added as a fixed effect (black and white, desert island, or three-panel comics). Subjects and items were entered as crossed random effects, with a random slope of metaphor type or sarcasm prompt type fit on subjects where appropriate. Interactions were tested among the metaphor types and sarcasm prompt types and the linguistic features, with only significant interactions retained. The linguistic features were controlled for multicollinearity using Pearson correlations and variance inflation values (VIF), and were also z-scored before being entered into the models.

2 Results

2.1 Metaphor and Sarcasm Quality Ratings

The human ratings of metaphor and sarcasm for the three subscales (*Conceptual Distance/Incongruity, Novelty, and Mirth*) were analyzed using two separate PCAs for the remaining 1304 metaphors and 716 sarcastic responses after adjudication. Both of the PCAs reported that the Novelty and Mirth subscales loaded into a single component, which explained 71% of the variance in the PCA for metaphor production scores and 62% of the variance in the PCA for sarcastic response scores. For the metaphor PCA, the Conceptual Distance scores loaded into a separate component (from novelty/mirth) explaining 26% of the variance in ratings, and for the sarcastic responses PCA, the Incongruity subscale loaded into a separate component (from novelty/mirth) explaining 33% of the variance in ratings. Therefore, the ratings for Novelty and Mirth were averaged for both metaphors and sarcasms, and the ratings for Conceptual Distance and Incongruity were retained in their original manner, resulting in two dependent variables for the metaphors and sarcastic responses per item.

2.2 Predicting Metaphor Quality

Metaphor Conceptual Distance. An LME model with metaphor conceptual distance as the dependent variable and linguistic features related to lexical sophistication and source overlap (word2vec), along with metaphor type (conventional vs. novel) as predictor variables reported three linguistic indices as significant predictors of the conceptual distance ratings (Table 1).

First, metaphors containing words with higher average Age of Acquisition (AoA) scores received significantly lower conceptual distance ratings. Words with a higher AoA are those that are self-reported to be learned later in life based on human judgments, and therefore represent less frequent and more sophisticated vocabulary.

This suggests that more sophisticated language in terms of AoA scores was not necessary in order to construct metaphors with higher conceptual distance between the entities being described in the metaphors. For example, the following metaphor had an average AoA of 8.9 and a conceptual distance score of one: *Some professors are geniuses like a supercomputer*. The prompt for this metaphor was *Some professors are very smart*. The word *genius* has an AoA of 7.21 and the word *supercomputer* has an AoA of 12.44, and these two words contributed significantly to the higher AoA score. Moreover, the word *genius* is conceptually similar to the prompt (i.e., *very smart*), and does not allow for any alternative conceptual interpretations. Indeed, *genius* is essentially a synonym of *smart*, and thus represents the same concept, and the inclusion of *supercomputer* also contains concepts related to intelligence, further amplifying the notion of smartness evoked by the word *genius*. Conversely, the following metaphor has an average AoA of 3.5 and a conceptual distance score of five: *That book is worth my arm and leg* in response to the prompt *Some property is very valuable*. In this metaphor, the words *arm*, *leg*, and *book* all have AoA scores of less than four, and thus contribute to a relatively low AoA rating. Furthermore, there is greater conceptual distance between a variety of concepts in this metaphor, with the words *arm* and *leg* perhaps conceptualized as *high value currency*, but only if one is aware of the idiomatic use of the expression *costs an arm and a leg*. Unlike the *genius* metaphor with high AoA, the words *arm* and *leg* are also not more sophisticated synonyms of any words in the prompt.

In addition to AoA, metaphors with higher Semantic Diversity scores also received significantly lower conceptual distance scores. Words with higher Semantic Diversity are words with less specific and more ambiguous meanings, which may suggest that metaphors containing more semantically ambiguous words may not be directly referencing specific concepts to make an apt metaphorical comparison.

In a similar fashion, metaphors with higher average Word Concreteness received significantly higher conceptual distance scores. These findings suggest that the human raters' perceptions of conceptual distance in the metaphors were influenced by the use of specific words in the metaphors. This may be because metaphors with more specific word usage were better able to evoke conceptual comparisons that were more distantly related, making it easier for the raters to identify the size of the conceptual comparison in the metaphor. Conversely, metaphors with higher AoA scores may have tended to use conceptual synonyms with the same overall semantic meaning (e.g., the use of *genius* to describe a *smart* professor), leading to lowered perceptions of conceptual distance among the human raters.

The model explained a total of 4.1% of the variance in conceptual distance scores, suggesting that these linguistic features account for a relatively small amount of the variation in conceptual distance scores and that they did not play a strong role in the human raters' conceptual rating decisions.

Metaphor Novelty and Mirth. An LME model with the averaged metaphor novelty/mirth score of human ratings the dependent variable and the same linguistic features related to lexical sophistication and source overlap used in the previous model as predictor variables reported three linguistic indices as significant predictors of metaphor novelty/mirth ratings (Table 2).

First, MRC Imageability was a significant, negative predictor of the novelty/mirth ratings, suggesting that metaphors including more imageable words resulted in lower ratings of novelty/mirth. Second, word2vec source similarity was also a significant, negative predictor of novelty/mirth, suggesting that metaphors containing higher semantic overlap with the source text received lower ratings of novelty/mirth.

Third, COCA spoken word frequency was also a significant, negative predictor of novelty/mirth ratings, suggesting that metaphors containing words with higher spoken word frequency resulted in significantly lower ratings of novelty/mirth. There were no other significant main effects or interactions. These results cohere to suggest that metaphors received higher novelty/mirth ratings if they included more sophisticated language and also included less semantic overlap with the metaphor prompt.

From a lexical perspective, higher levels of both Spoken Word Frequency and Word Imageability resulted in significantly lower ratings of novelty/mirth for metaphors. The direction of their influence on the novelty/mirth ratings indicates that more lexically sophisticated metaphors received higher novelty/mirth scores.

In terms of cohesion, metaphors that contained greater semantic overlap with the metaphor prompt (as measured through word2vec) received significantly lower novelty/mirth scores. This finding makes intuitive sense because metaphors that were more closely related to the metaphor prompt were most likely those that were more cliché or did not make more distant comparisons.

The word2vec measure may also capture the extent to which participants relied on the language from the metaphor prompt. For example, the metaphor *Some relationships are like working in a research lab and having a project fail* received a novelty/mirth score of five and a semantic overlap score of -0.17. The only words repeated in this metaphor from the prompt are *some relationships*, while the rest of the metaphor includes words outside of the prompt.

Conversely, the metaphor *The earth is full of people working like bees* received a novelty/mirth score two and a semantic overlap score of 0.68. Unlike the previous metaphor, this metaphor almost completely repeats the metaphor prompt word for word (i.e., *the earth is full of busy people*) and only includes three original words.

Much like the model predicting metaphor conceptual distance ratings, the linguistic features predicting the metaphor novelty/mirth scores explained a relatively small amount of variance in rater scores (7.5%), suggesting that linguistic features were just one small influence on the human ratings of novelty and mirth.

	Estimate	***SE***	***t***	***p***
(Intercept)	4.559	0.089	51.179	< .001
Metaphor Type: Novel	-0.228	0.399	-0.571	0.575
Source Similarity (word2vec)	0.010	0.032	0.324	0.746
MRC Familiarity	0.015	0.027	0.569	0.570
MRC Imageability	-0.011	0.039	-0.277	0.782
MRC Meaningfulness	-0.034	0.034	-0.999	0.318
Age of Acquisition*	-0.123	0.035	-3.533	< .001
Brysbaert Concreteness*	0.102	0.039	2.610	0.009
COCA Spoken Word Frequency	0.027	0.031	0.877	0.380
Semantic Diversity*	-0.106	0.035	-2.993	0.003
* = Significant predictor. SE = Standard Error. Baseline for Metaphor Type = Conventional.				

Table 1. LME predicting metaphor conceptual distance scores

	Estimate	***SE***	***t***	***p***
(Intercept)	3.292	0.101	32.604	< .001
Metaphor Type: Novel	0.165	0.388	0.425	0.676
Source Similarity (word2vec)*	-0.127	0.041	-3.127	0.002
MRC Familiarity	0.064	0.035	1.830	0.068
MRC Imageability*	-0.106	0.050	-2.120	0.034
MRC Meaningfulness	0.003	0.043	0.064	0.949
Age of Acquisition	-0.065	0.045	-1.451	0.147
Brysbaert Concreteness	-0.067	0.050	-1.347	0.178
COCA Spoken Word Frequency*	-0.314	0.041	-7.660	< .001
Semantic Diversity	-0.040	0.045	-0.895	0.371
* = Significant predictor. SE = Standard Error. Baseline for Metaphor Type = Conventional.				

Table 2. LME predicting metaphor novelty/mirth scores

2.3 Predicting Sarcasm Quality

Sarcasm Incongruity. An LME model predicting incongruity ratings of the sarcastic responses using linguistic features (MRC Familiarity, MRC Meaningfulness, Age of Acquisition, Brysbaert Concreteness, COCA Spoken Word Frequency, and Semantic Diversity) reported that MRC Meaningfulness was a significant, negative predictor of incongruity ratings, suggesting that sarcastic responses with more average associations to other words resulted in lower ratings of incongruity (Table 3). This model only accounted for 2% of the variance in incongruity scores, suggesting that this linguistic feature played a small role in raters' perceptions of incongruity in the sarcastic responses.

Sarcasm Novelty and Mirth. An LME model predicting novelty/mirth ratings of the sarcastic responses using the same linguistic features as the previous model included one significant main effect and two significant interactions (Table 4).

The main effect demonstrated that sarcastic responses containing higher levels of average AoA received significantly higher novelty/mirth ratings. This finding provide some evidence suggesting that sarcastic responses which are more lexically sophisticated are perceived as more creative, because higher amounts of AoA tend to suggest higher levels of lexical sophistication.

For example, the sarcastic reply of *at least we have water* for one of the desert island comics received a novelty/mirth score of 2.25 and had an average AoA score of 3.04, whereas the sarcastic reply *you have surgical precision behind the wheel* in response to the puddle splash comic received a novelty/mirth score of 4.75 and had an average AoA of 7.45. The second example's use of *surgical precision* represents less frequent words when compared to the first example, which in turn provides a higher likelihood that the author of the second sarcastic response coined an answer that was unique when compared to the other participants, subsequently increasing perceptions of novelty and perhaps mirth among the human raters. Thus, the AoA results suggest that using more lexically sophisticated language could be one strategy for producing more creative sarcastic responses.

	Estimate	***SE***	***t***	***p***
(Intercept)	4.396	0.099	44.278	< .001
Sarcasm Prompt: Black and White	0.216	0.129	1.676	0.128
Sarcasm Prompt: Desert Island	0.175	0.129	1.352	0.209
MRC Familiarity	0.063	0.035	1.806	0.071
MRC Meaningfulness*	-0.067	0.032	-2.079	0.038
Age of Acquisition	0.034	0.030	1.130	0.259
Brysbaert Concreteness	0.027	0.034	0.780	0.436
COCA Spoken Frequency	-0.003	0.033	-0.103	0.918
Semantic Diversity	-0.026	0.036	-0.729	0.466
* Significant effect. SE = Standard error. Baseline for Sarcasm Prompt = Three Panel Comic.				

Table 3. LME predicting sarcasm incongruity scores

	Estimate	***SE***	***t***	***p***
(Intercept)	2.965	0.125	23.661	< .001
MRC Familiarity	0.007	0.044	0.154	0.877
Sarcasm Prompt: Black and White	0.190	0.162	1.175	0.272
Sarcasm Prompt: Desert Island	0.377	0.162	2.328	0.046
Age of Acquisition*	0.114	0.038	3.013	0.003
Brysbaert Concreteness	0.038	0.063	0.607	0.544
COCA Spoken Frequency	0.030	0.039	0.763	0.446
Semantic Diversity	-0.065	0.043	-1.507	0.132
MRC Meaningfulness	0.001	0.039	0.018	0.985
Significant Interactions				
MRC Familiarity: Sarcasm Prompt: Black and White	0.196	0.120	1.638	0.102
MRC Familiarity: Sarcasm Prompt: Desert Island*	0.428	0.128	3.351	0.001
Concreteness: Sarcasm Prompt: Black and White	0.057	0.088	0.652	0.515
Concreteness: Sarcasm Prompt: Desert Island*	0.210	0.085	2.472	0.014
* Significant effect. SE = Standard error. Baseline for Sarcasm Prompt = Three Panel Comic.				

Table 4. LME predicting sarcasm novelty/mirth scores

Additionally, two lexical features interacted with prompt type in that there were significant differences between the desert island prompt and the three-panel comic prompt for both features. These interactions demonstrated that increasing levels of MRC Familiarity and Brysbaert Concreteness significantly increased perceptions of novelty/mirth for sarcastic replies made in response to the desert island prompts when compared to the three-panel comic prompts. Higher levels of both MRC Familiarity and Brysbaert Concreteness suggest less lexically sophisticated language, because words that are more familiar correlate with more frequently used words, and words that are and more concrete represent concepts that are more easily retrieved due to their encoding as both a lexical item (e.g., car) as well as the visual concept of that same item (e.g., a concept of a car). Because there was less contextual information available in the desert island prompts, it may be that sarcastic responses including less sophisticated language (i.e., more concrete concepts that are more familiar) were better able to index specific ideas indicative of sarcastic meaning for the desert island prompts when compared to the three-panel comic prompts, where contextual information could fill in semantic gaps for the raters. Much like the other models, these features accounted for a relatively small amount of variance in the raters' scores (6.8%), again suggesting that linguistic features played a small yet significant role in raters' perceptions of creativity among the sarcastic responses.

3 Discussion

The purpose of this study was to investigate whether differences in figurative language quality could be predicted using linguistic features related to lexical sophistication and semantic cohesion. Overall, the findings suggest that variables representative of lexical sophistication (and semantic cohesion for metaphors) played a small yet significant role in explaining variance among rater perceptions of figurative language quality, and also that perceptions of quality included both theoretical constructs related to metaphor and sarcasm (i.e., conceptual distance and incongruity) as well as to more generalized constructs of creative ability (i.e., novelty and mirth).

In regards to the theoretical components, greater conceptual distance scores were predicted by more sophisticated and specific language, perhaps because more specific words are better able to encode specific concepts, allowing for a more direct metaphorical comparison between two entities. For sarcastic responses, greater incongruity was marked by language with a lower number of word associations, which may have been a result of the use of more conversational language in sarcastic responses (e.g., *thank you*). As for the novelty and mirth scores, overall the results demonstrated that greater levels of lexical sophistication led to greater perceptions of novelty and mirth for both metaphors and sarcastic responses, although this effect was mediated by the different prompts for sarcastic responses.

Linguistic features were better able to predict variance in the novelty and mirth scores when compared to the conceptual distance or incongruity scores, suggesting that the raters may have attended more strongly to linguistic features when considering the creativity of the metaphors and sarcastic responses when compared to the conceptual distance or incongruity. This suggests that linguistic features related to lexical sophistication may be more suitable for measuring general measures of creativity, which are but one component of figurative language quality.

Finally, the linguistic features explained more variance in the metaphors when compared to the sarcastic responses, which is most likely a result of the linguistic context in which metaphors operate. Specifically, the understanding of a metaphor requires the possessing of conceptual information encoded in the metaphor. However, in order to understand a sarcastic reply, one must be more aware of the surrounding social and pragmatic context. Echoing contextual information linguistically is not necessary in many sarcastic responses, as it is known knowledge already available to those within the situation. For example, a simple *thank you* can be taken as sarcastic in the right contexts, which would be difficult to differentiate through linguistic means alone. Therefore, the contextual nature of sarcasm quality may make it more difficult to define using quantitative linguistic features when compared to other types of figurative language, such as metaphor.

4 Conclusion

One limitation present in this data is that the answers produced by the participants were generally short, which in turn could easily bias some of the lexical measurements used, as all of them reported average scores for all the content words in an answer. Nonetheless, this study has shed further light on linguistic features of figurative language by investigating connections between figurative language quality, lexical sophistication, and cohesion using theoretical definitions of creativity, metaphor, and sarcasm and demonstrating that linguistic features of figurative language quality may in part be related to generalized notions of creativity. Future work employing classifiers designed to discriminate figurative language from non-figurative language may want to consider the quality of figurative language, and one method for doing so may lie in linguistic features related to creativity in the examples under investigation.

Acknowledgments

We would like to thank Tori Morrison and Oni Nistor for the figurative language quality ratings.

References

Acar, Selcuk and Mark A. Runco. 2014. Assessing associative distance among ideas elicited by tests of divergent thinking. *Creativity Research Journal* 26(2). 229–238.

Beaty, Roger E. and Paul J. Silvia. 2013. Metaphorically speaking: cognitive abilities and the production of figurative language. *Memory & Cognition* 41(2). 255–267.

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46(3). 904–911.

Chiappe, Dan L. and Penny Chiappe. 2007. The role of working memory in metaphor production and comprehension. *Journal of Memory and Language* 56(2). 172–188.

Colston, Herbert L. 2017. Irony and sarcasm. In Salvatore Attardo (ed.), *The Routledge handbook of language and humor*, 234–249. New York, NY: Routledge.

Coltheart, Max. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology* 33(4). 497–505.

Crossley, Scott A., Kristopher Kyle, and Danielle S. McNamara. 2016. The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods* 48. 1227–1237.

Davies, Mark. 2008. *The corpus of contemporary American English*. BYE, Brigham Young University.

Dumas, Denis and Kevin N. Dunbar. 2014. Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity* 14. 56–67.

Gerrig, Richard J. and Raymond W. Gibbs. 1988. Beyond the lexicon: Creativity in language production. *Metaphor and Symbol* 3(3). 1–19.

Gibbs, Raymond W. 1994. *The poetics of mind: Figurative thought, language, and understanding*. Cambridge, MA: Cambridge University Press.

Glucksberg, Sam. 2001. *Understanding figurative language: From metaphor to idioms*. Oxford University Press.

Hoffman, Paul, Matthew A. Lambon Ralph, and Timothy T. Rogers. 2013. Semantic diversity: A measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods* 45(3). 718–730.

Huang, Li, Francesca Gino, and Adam D. Galinsky. 2015. The highest form of intelligence: Sarcasm increases creativity for both expressers and recipients. *Organizational Behavior and Human Decision Processes* 131. 162–177.

Khodak, Mikhail, Nikunj Saunshi, and Kiran Vodrahalli. 2017. A large self-annotated corpus for sarcasm. *arXiv preprint arXiv:1704.05579*. https://arxiv.org/abs/1704.05579 (15 June, 2017).

Kintsch, Walter. 2008. How the mind computes the meaning of metaphor. In Raymond W. Gibbs Jr (ed.), *The Cambridge handbook of metaphor and thought*, 129–142. Cambridge, MA: Cambridge University Press.

Kintsch, Walter and Anita R. Bowles. 2002. Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol* 17(4). 249–262.

Kuperman, Victor, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods* 44(4). 978–990.

Kyle, Kristopher and Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly* 49(4). 757–786.

Kyle, Kristopher, Scott Crossley, and Cynthia Berger. 2017. The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior Research Methods*.

Martin, Rod A. 2007. *The psychology of humor: An integrative approach*. Burlington, MA: Elsevier Academic Press.

Rosenzweig, Saul. 1945. The picture-association method and its application in a study of reactions to frustration. *Journal of personality* 14(1). 3–23.

Runco, Mark A. and Garrett J. Jaeger. 2012. The standard definition of creativity. *Creativity Research Journal* 24(1). 92–96.

Salsbury, Tom, Scott A. Crossley, and Danielle S. McNamara. 2011. Psycholinguistic word information in second language oral discourse. *Second Language Research* 27(3). 343–360.

Skalicky, Stephen and Scott Crossley. 2015. A statistical analysis of satirical Amazon.com product reviews. *The European Journal of Humour Research* 2(3). 66–85.

Skalicky, Stephen, Scott A. Crossley, Danielle S. McNamara, and Kasia Muldner. 2017. Identifying creativity during problem solving using linguistic features. *Creativity Research Journal* 29(4). 343–353.