"A fork is a food stabber": Linguistic creativity in English L1 and L2 speakers

Stephen Skalicky (stephen.skalicky@vuw.ac.nz)
School of Linguistics and Applied Language Studies, Victoria University of Wellington
Wellington, New Zealand

Nancy D. Bell (nbell@wsu.edu)
Department of English, Washington State University
Pullman, WA, USA

Mihai Dascalu (mihai.dascalu@upb.ro)
Computer Science and Engineering, University Politehnica of Bucharest
Bucharest, Romania

Scott A. Crossley (scrossley@gsu.edu)
Applied Linguistics & ESL, Georgia State University
Atlanta, GA, USA

Abstract

Knowing more than one language provides a speaker with an increased pool of linguistic experiences and concepts. This expanded language knowledge is thought to benefit bilingual speakers on standardized tests of creative ability. However, relatively little research has explored bilingual performance on tests of *linguistic* creativity. In this study we compare the production of creative, attenuated descriptions produced by English L1 and English L2 speakers. Using computational measures of text similarity, we find that English L2 answers were significantly less similar than L1 answers, suggesting a greater number of concepts and topics were used by the L2 participants. Additionally, unsupervised cluster analysis found no strong differences in the number of cluster topics between the L1 and L2 data. As such, the L2 answers contained more breadth, whereas the L1 answers contained more depth. The results may reflect fundamental differences in the storage and use of L1/L2 language knowledge.

Keywords: bilingualism; linguistic creativity; attenuated descriptions; semantic similarity

Introduction

Linguistic creativity is rooted in a speaker's knowledge of what is possible in a language. Native or first language (L1) speakers possess rich knowledge of their L1, spanning across vocabulary, syntax, pronunciation, word meaning, and pragmatics. These different components of language all offer fertile grounds for linguistic creativity and play (Cook, 2000). An L1 speaker can emphasize similar sounding words and phrases to create puns, draw connections among disparate concepts through metaphor, subvert a listener's expectations through verbal irony, and more (Colston, 2015; Gibbs & Colston, 2012). Using these forms of linguistic creativity is not a special ability only held by gifted individuals, but rather represents the gamut of options available for speakers to accomplish specific communicative goals (Carter, 2016; Colston, 2015).

L1 knowledge influences second language (L2) production through lexical and conceptual mediation at higher levels of L2 proficiency (Jarvis, 2011). This is because language knowledge is integrated into a shared mental lexicon and co-activated during language use (Kroll & Ma, 2017). Evidence for this integrated lexicon comes in part from psycholinguistic experiments which have catalogued cross-linguistic priming for a variety of lexical features, including word meanings, word forms, and word associations such as polysemy (Tokowicz, 2014). Moreover, bilinguals experience a greater range of cultural customs and values through language. This increased amount of lexical knowledge and lived experiences is thought to positively contribute towards general creative ability (van Dijk et al., 2019), as evidenced by studies which have found bilinguals tend to outperform monolinguals on standardized tests of creativity (Hommel et al., 2011; Kharkhurin, 2007; Leikin, 2013).

At more advanced L2 proficiencies, the increased storage of lexical items, concepts, and experiences should provide an even greater linguistic toolbox from which a bilingual language user can draw. This is certainly the case when it comes to bilingual language play involving more than one language, often seen in multicultural cities such as Hong Kong (Luk, 2013) and Montréal (Lamarre, 2014).

However, relatively few studies have directly investigated bilinguals' production of linguistic creativity. Of the studies that do exist in this area, most report a positive relation between L2 development and linguistic creativity production (Bell et al., 2014; Wang & Cheng, 2016). Nonetheless, it is rare for a study to directly compare the linguistic creativity of L1 and L2 speakers. One such study found no differences in perceptions of creativity towards L1 and advanced L2 speakers' elicited production of metaphors and sarcasm (Skalicky, 2020). More research in this area is needed because it is still unclear whether the lexical and conceptual differences attributed to L2 speakers influence the production of linguistic creativity when compared to L1 speakers.

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

The increased storage of lexical items, concepts, and experiences which influence L2 production should provide a larger and deeper linguistic toolbox from which a language user can draw. Whether this translates into differences in the production of linguistic creativity is still an open question, although some evidence exists to suggest increased multilingual experiences facilitate the comprehension of creative language, such as novel metaphors (Werkmann Horvat et al., 2021). As such, the purpose of this study is to compare L1 and L2 linguistic creativity produced under controlled conditions. Specifically, we asked participants to invent attenuated descriptions for common objects and then compared the diversity of answers provided by L1 and L2 speakers using a computational measure of text similarity.

Attenuated Descriptions

In this study we examine attenuated descriptions of everyday objects. Attenuation is a linguistic strategy wherein a speaker purposefully underspecifies an entity or event through omission of a direct referent and instead uses an unnecessarily lengthy and roundabout description, known as underlexicalization (Fowler, 1986; Simpson, 2003). An attenuated description typically highlights general and/or specific characteristics of the entity or event being described. Potential strategies include referring to superordinate categories a noun belongs to (e.g., calling a human an animal) or describing a set of shared, indistinguishable features for a specific referent (e.g., a 2002 reference to Prince Harry as a "16-year old boy" to highlight accurate yet less salient characteristics of a member of the British Royal Family; Simpson, 2003, p. 133). Attenuation thus purposefully misdirects the inferences made for any one referent, serving as a strategy to shift focus away from the specific and towards the general features of an entity or event (Simpson, 2003).

Attenuated descriptions feature prominently in at least two contemporary forms of creative media. In his webcomic *xkcd*, Randall Munroe published an entry entitled *Up Goer Five*[1], which presented a blueprint of the Saturn V rocket used by NASA in the late 1960s and early 1970s. The blueprint is annotated with descriptions of various components and functions of the rocket using only the first 1000 most frequent words in English. Thus, the command module is a *people box*, the rocket fuel is *stuff to burn to make people go fast*, and the rocket itself is named the *up goer*. The second example also comes from a webcomic, Nathan Pyle's *Strange Planet*[2] which provides attenuated observations of everyday human behavior from the perspective of aliens. For example, the aliens refer to the process of sleeping as *being unconscious*, to an umbrella as a *sky shield*, and to a stuffed toy bear as a *fabric predator creature*. The humor and creativity in both the *xkcd* and *Strange Planet* examples are rooted in the refocusing of the entities, objects, and processes through attenuation.

The goal of the current study is to compare attenuated descriptions of objects produced by English L1 and English L2 speakers. Attenuation places these common things in a new light, prompting a reconsideration of typical concept and category memberships. As such, attenuated descriptions may be one form of linguistic creativity where L1 influence on L2 production provides a boost in creativity. If this is the case, the range of answers provided by L2 participants should be more diverse when compared to L1 participants. The following research question guides our study:

RQ1. What differences, if any, exist in the diversity and range of attenuated descriptions provided by English L1 and L2 speakers?

Method

Creativity Production Task: *Glossalia*

We designed an original creativity task for participants to generate new names for a series of everyday objects. The task asked participants to pretend they were hired to work for a fictional television show named *Glossalia*, a show designed to help teach English words to young L2 English learners. One of the methods for helping teach English words on the show involves a character who mistakenly refers to objects using unique and silly words and phrases. The task provided two examples of these mistakes (which were actually attenuated descriptions): referring to bread as *raw toast* and referring to an iron as a *wrinkle remover*. We then explained to the participants their goal: to generate similar names for a new set of objects to be used by the character on the show. The instructions encouraged the participants to be as creative as possible. Three rules were stipulated to promote the production of attenuated descriptions. First, the new name must describe the object in a meaningful way. Second, the new name must use English words which can be found in the dictionary. And third, the new name could not reuse the actual name of the object.

We used a random list generator from the website www.randomlists.com/things to locate copyright free pictures of common objects. We chose nine objects in total with accompanying color photos of actual objects (i.e., not drawings). These objects were: (1) an inflated green balloon, (2) a brown cardboard box with the lid flaps open, (3) several bunches of orange carrots, (4) a purple sofa, (5) a black plastic fork, (6) a pair of yellow-tinted sunglasses, (7) a metal shovel, (8) a white toilet, and (9) a brown leather wallet. We chose these objects because we felt they were diverse in their form and function, but also easily recognizable.

Participants

We gathered data from L1 and L2 English speakers. The L1 participants (n = 99) were recruited from the crowdsourcing platform Amazon Mechanical Turk (M_{age} = 39.7, SD_{age} = 12.3, 64.6% female). The Mechanical Turk participants spoke English as their first language and resided in the United States. The L2 participants (n = 78) were recruited from

[1] https://xkcd.com/1133/

[2] https://www.instagram.com/nathanwpylestrangeplanet/

undergraduate and graduate programming courses at the University Politehnica of Bucharest in Romania (M_{age} = 19.8, SD_{age} = .46, 30.8% female). The English L2 participants spoke Romanian as their L1 and had been learning English for an average of 11.2 years (*SD* = 2.71). When asked about their average daily English use, $20_{(25.6\%)}$ English L2 participants reported using English for less than one hour per day, $30_{(38.5\%)}$ used English for 1-3 hours per day, $17_{(21.8\%)}$ for 3-5 hours per day, and $11_{(14.1\%)}$ for more than 5 hours per day.

Procedure

We uploaded our *Glossalia* task to the survey platform *Qualtrics*. Participants first read the instructions (described above) and then provided new names for the nine objects by typing their answer into a text box below each picture. The pictures for the nine objects were presented one at a time and were accompanied by the prompt *what unique and silly name can you provide for this object*? The objects were presented in the same order as described above for every participant. The English L1 participants were forwarded to the survey through the Amazon Mechanical Turk website and were provided $0.30 USD for the task. The English L2 participants were recruited through their lecturer and were provided with extra credit in their course for their participation. There were no time limits imposed on any parts of the task.

Data Analysis

We gathered a total of 1,295 valid responses from our participants. We then removed any valid responses which violated any of the three rules for new names stipulated in the task instructions. This removed 53 responses (4.1%), leaving a final data set of 1,242 responses. Any spelling mistakes or punctuation errors in the responses were manually corrected.

Measuring Answer Creativity. Standardized measures of creativity assess both the total number of ideas produced (fluency), as well as how similar one answer is to a specific prompt when taking into consideration all other answers (originality; Kaufman et al., 2008). Because our participants only provided one name per object, we focused on measuring the originality of answers. One method to assess originality which has gained increased use is through the comparison of semantic distance between answers provided on tests of creativity (Kenett, 2019). These comparisons are done using computational measures of word distributions, such as Latent Semantic Analysis (Dumas & Dunbar, 2014), which allows for an objective, scalable, and automatic assessment of originality (Beketayev & Runco, 2016; Kenett, 2019).

Answers with greater semantic distance (i.e., lower similarity) are more original for two reasons. First, answers which are completely or partially repeated have higher similarity values, and thus higher similarity becomes a marker of lower originality. Second, answers with lower similarity use concepts with more remote associations to the other answers, marking themselves as more distinctive in terms of the ideas and content in the answer (Acar & Runco, 2014).

In addition to originality, it is also important to consider the effectiveness of creative products (Runco & Jaeger, 2012). When compared to originality, effectiveness is more difficult to define and less amenable to automatic assessment methods. In the case of our task, an effective attenuated description could be defined as one which follows our prompt criteria and is also evocative of the item being described. While we controlled for the first aspect by removing any answers which violated any of our instructions, we did not obtain any measures of how evocative the descriptions were. We plan to address this additional aspect of creativity in future research.

Calculating Similarity Scores. We analyzed our data using the Natural Language Processing library spaCy (v3.0)[3]. For any one answer, we calculated the semantic similarity of that answer to all other answers provided for the same object. Instead of using the default spaCy word vectors and methods for calculating similarity (i.e., taking the vector similarity of each word in a sentence and then averaging this similarity), we applied the Google Universal Sentence Encoder (USE) for spaCy (v0.4.3)[4]. The Google USE for spaCy is a custom pipeline in spaCy which incorporates the Google Universal Sentence Encoder sentence embeddings (Cer et al., 2018) that provide contextualized representations for the entire input. Thus, rather than calculating similarity as an averaged value across all the words in a sentence, we measured similarity between sentences as single units using the large USE model with Transformer architecture. Also, when considering other and newer sentence embedding models like Sent-BERT (Reimers & Gurevych, 2019), we selected USE given its pretraining on several different tasks (QA/Forums, SNLI, wiki/news) which are more suitable to reflect sentence similarity[5]. The cosine similarity values range from 0 (not similar) to 1 (identical); all values below 0 were truncated to 0.

Similarity values of 1 occurred when the answer was identical to another answer, which allowed for us to simultaneously model the level of direct repetition for any one answer, as well as gradients in similarity to other answers. After obtaining pairwise similarities between an answer and all others, we then averaged these values to obtain a single average similarity score per answer. We calculated these scores separately for the English L1 and L2 data. Table 1 provides an example of how answer similarity was affected by repetition. Specifically, in response to the picture of the fork, ten participants provided the answer "*food stabber*" (7 English L1, 3 English L2). The overall average semantic similarity for "*food stabber*" is higher for the L1 data, in part due to the higher repetition (Table 1).

[3] https://spacy.io/

[4] https://github.com/MartinoMensio/spacy-universal-sentence-encoder

[5]https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html

Table 1. Sample answer similarities (sim) in ascending order for the "*food stabber*" answer. Averages are from the entire set of answers.

A fork is a "food stabber"			
English L1	sim	English L2	sim
masochist's shoe horn	.07	*the little mermaid's comb*	.09
metal teeth	.16	*a smaller trident*	.14
noodle catcher	.25	*little hay mover*	.23
eating tool	.44	*food forklift*	.48
food jabber	.60	*food enemy*	.69
L1 average	.44	L2 average	.39

Mixed Effects Regression. We then constructed a mixed effects regression model in R (v4.0.2; R Core Team, 2020) using the *lme4* package (v1.1.27.1; Bates et al., 2015) to model the average similarity scores as a function of different object pictures and language group. Both object type and language group were entered as dummy-coded categorical variables, with balloon and English L1 representing the baselines. These variables were fit as an interaction, and intercepts were allowed to vary per participant. Post-hoc comparisons between the two language groups for all nine objects were calculated using the *emmeans* package (v1.6.3; Lenth, 2020).

Cluster Analysis. Our aim was also to model the spread of answer types for each object. To do so, we conducted an unsupervised *k*-medoid cluster analysis using the *cluster* package (v2.1.0; Maechler et al., 2019) to locate groupings of similar answers for each object and language group combination, based on the averaged similarity values. This enabled us to automatically locate exemplars representing different sets of answers. For each set of clusters, we let *k* range from 2-20 and chose the number of clusters with the largest average silhouette size (Levshina, 2015). This method has also been effectively employed to model the number of exemplars for linguistic constructions, such as denominal verb constructions (Watson et al., 2021). We have provided our code and data on the OSF Repository[6].

Results

Similarity Scores

The total distribution of answer similarity values for each object are displayed in Figure 1. As can be seen, the distributions of answer similarity values differed as a function of both the object being named and the L1/L2 language groups. In terms of language group, the average similarity among answers was generally lower for the English L2 participants when compared to the English L1 answers. This was true for all objects aside from names created for the carrot, with the greatest visual differences observed in similarity values for names created for the box and toilet objects. This suggests that, in general, the English L2 participants provided a greater diversity of answers for almost all objects. In terms of differences among objects, names created for the wallet had the overall highest average answer similarity, with a right skew towards greater similarity observed in both the L1 and L2 data.

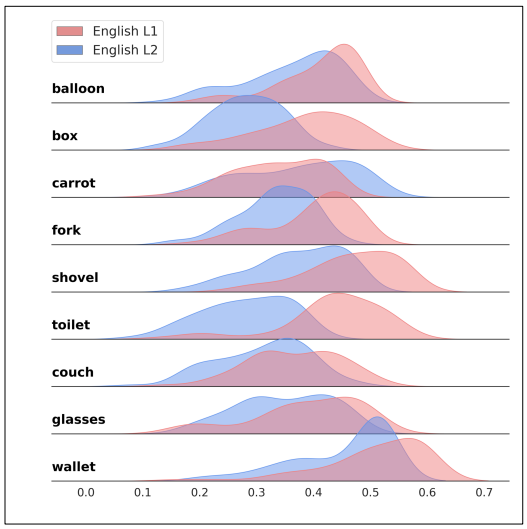

Figure 1. Distribution of similarity scores.

Results: Mixed Effects Regression. The interaction between language group and object type was significant ($F_{8,\ 1081.90} = 14.511, p < .001$). The conditional R^2 of the model (variance from predictors and random intercepts) was .457, and the marginal R^2 (variance from predictors alone) was .334, suggesting this interaction accounted for a good portion of the variance in similarity values. Pairwise comparisons using adjusted *p*-values indicated the similarity values were significantly different between the two language groups for all nine objects. Specifically, names provided for all objects aside from the carrot were significantly less similar when provided by English L2 participants when compared to English L1, confirming the visual differences noted in Figure 1 to also be statistically significant. Table 2 reports the descriptive statistics, estimates, confidence intervals, and *t* and *p* values for each comparison.

Results: Cluster Analysis. The total number of clusters and average cluster size are reported in Table 3. The L2 data contained a higher number of clusters for five of the nine objects, with the largest differences observed for the toilet and wallet clusters. In contrast, the L1 data contained an equally greater number of clusters for the couch object. The mean number of answers per cluster was relatively stable between both the L1 and L2 data, with approximately 5 items per cluster.

[6] https://osf.io/3657e/?view_only=e4b5960e752648a8a6c01cb438ccb7dc

Table 2. Results from regression contrasts. Average similarity values were significantly lower for L2 answers compared to English L1 for all objects aside from the carrot, where a significant contrast in the reverse direction was observed.

	English L1				English L2				Pairwise Contrasts		
item	*n*	*M (SD)*	5%CI	95%CI	*n*	*M (SD)*	5%CI	95%CI	*estimate*	*t*	*p*
balloon	63	.41 (.07)	.39	.43	75	.36 (.08)	.34	.38	.049	3.55	< .001
box	63	.38 (.09)	.36	.40	74	.28 (.06)	.26	.30	.100	7.18	< .001
carrot	65	.34 (.08)	.32	.36	74	.37 (.10)	.35	.39	-.032	-2.30	.022
couch	64	.37 (.08)	.35	.39	74	.31 (.08)	.29	.33	.058	4.21	< .001
fork	65	.40 (.08)	.38	.42	74	.33 (.06)	.31	.35	.069	4.97	< .001
glasses	67	.39 (.10)	.37	.41	75	.35 (.08)	.33	.37	.042	3.04	.002
shovel	64	.47 (.08)	.44	.48	74	.37 (.08)	.35	.39	.090	6.51	< .001
toilet	63	.43 (.09)	.41	.45	69	.28 (.08)	.26	.30	.149	10.57	< .001
wallet	64	.51 (.09)	.48	.52	76	.45 (.09)	.43	.47	.054	3.90	< .001

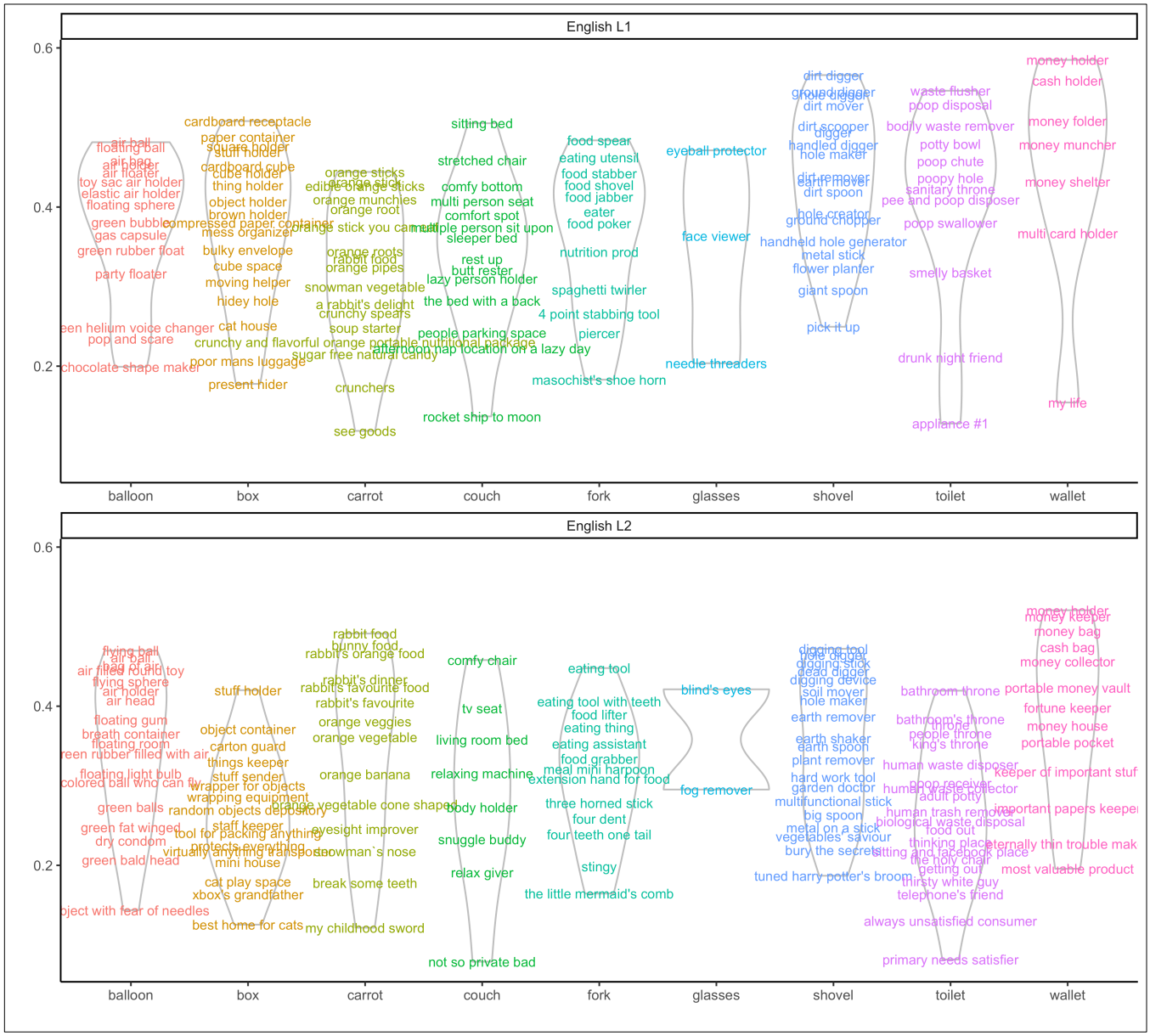

Figure 2. Cluster medoid exemplars plotted against their averaged semantic similarity values.

Exemplars identified in each cluster are plotted in Figure 2 as points over violin plots which reflect their distributions. As expected, based on the raw distributions plotted in Figure 1, the distribution of cluster exemplars was skewed more strongly towards higher similarity scores for the L1 data, when compared to the L2 data.

The largest difference was observed for names provided for the picture of the toilet. In the L1 data, many of the exemplars referred to the toilet's primary function as a method to dispose of bodily waste. The L2 exemplars contained similar answers, but also ranged into answers comparing toilets to thrones and made more personifications of the toilets (e.g., *always unsatisfied customer*). This suggests the L1 participants created answers restricted to fewer categories when compared to the L2 data.

Table 3. Cluster information.

	Mean Cluster Size		Total Clusters	
item	L1	L2	L1	L2
balloon	4.2	4.17	15	18
box	3.5	4.63	18	16
carrot	3.82	5.29	17	14
couch	4.57	9.25	14	8
fork	5.33	5.69	12	13
glasses	22.33	37.50	3	2
shovel	3.56	3.89	18	19
toilet	5.25	3.45	12	20
wallet	9.14	5.85	7	13
overall	4.97	5.41	116	123

Discussion

The purpose of this study was to explore the originality of attenuated descriptions provided by English L1 and L2 speakers. We operationalized originality as a function of semantic distance between attenuated descriptions provided for the same item. Results from our analysis indicate the average similarity among attenuated description was significantly lower for the English L2 data for eight of the nine objects, suggesting the L2 attenuated descriptions were more original, when compared to the L1 descriptions. Additionally, a cluster analysis identified no real trends in the number of clusters between our L1/L2 data.

We draw two main conclusions from these results. First, the semantic similarity analysis provides evidence to suggest that attenuated descriptions from L2 participants were more diverse than the L1 data. This could be a function of the combined influence of L1 experiences on L2 production, resulting in more disparate answers across a wider range of topics. With more options come more choices, and with more choices comes more originality. However, these options may be constrained by the nature of each object – L2 answers for the carrot object were narrower in topic, when compared to the L1 data, which may reflect cultural experiences associated with this object.

Second, although the L1 data was more homogenous in terms of semantic similarity, the cluster analysis suggests that the L1 participants still demonstrated a diversity of answer types. This likely reflects the different linguistic knowledge afforded to an L1 speaker. Knowledge of conventionalized and formulaic expressions is stronger for L1 speakers, resulting in more homogenous answers through convergence of similar collocational choices. L2 speech lacks the same level of conventionality, which in turn would affect the semantic similarity and ultimately our interpretation of originality and creativity. Combined with a greater range of options to choose from, one could expect the L2 data would *seem* more creative, because the L2 data cannot help but be more original. If this is the case, our data suggests the L1 and L2 speakers drew from different strengths when completing this task, manifesting not in different *degrees* of linguistic creativity, but rather in different *strategies*. This insight could only be gleaned through a combined approach using semantic scores and the cluster analysis.

As a between-subjects design, we acknowledge that our findings may also be related to other individual differences among our participants. Our data indicates the English L1 participants were older than the English L2 group, which may have further strengthened knowledge of conventionalized formulae among the English L1 speakers based on more years of exposure to English. Moreover, the English L2 participants were likely more homogenous in terms of socioeconomic factors such as level of education, but we did not collect similar information from the English L1 participants. One suggestion to address these limitations in future research is to conduct a within-participants design, where bilinguals perform a similar task in their L1 and L2.

In conclusion, we hesitate to make any claims that the L2 data was any more creative than the L1 data. Rather, we argue that our results reflect fundamental differences in the storage and use of L1/L2 language knowledge. The integration of lexical items from more than one language means that the bilingual participants had access to a greater range of word associations, which includes knowledge of polysemy, category membership, and other lexical relations (Tokowicz, 2014). These participants thus had a wider search field when thinking of their attenuated descriptions, which in turn would result in answers which are more original. A logical next step in this research is to incorporate additional aspects of creativity, such as effectiveness of the answers.

Acknowledgments

Many thanks to the three anonymous reviewers who provided invaluable feedback and suggestions for this project.

References

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beketayev, K., & Runco, M. A. (2016). Scoring divergent thinking tests by computer with a semantics-based algorithm. *Europe's Journal of Psychology*, *12*(2), 210.

Bell, N., Skalicky, S., & Salsbury, T. (2014). Multicompetence in L2 language play: A longitudinal case study. *Language Learning*, *64*(1), 72–102. https://doi.org/10.1111/lang.12030

Carter, R. (2016). *Language and creativity: The art of common talk* (Second Edition). Routledge.

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). Universal Sentence Encoder. *ArXiv:1803.11175 [Cs]*. https://doi.org/arXiv:1803.11175v2 [cs.CL]

Colston, H. L. (2015). *Using figurative language*. Cambridge University Press.

Cook, G. (2000). *Language play, language learning*. Oxford University Press.

Dumas, D., & Dunbar, K. N. (2014). Understanding fluency and originality: A latent variable perspective. *Thinking Skills and Creativity*, *14*, 56–67.

Fowler, R. (1986). *Linguistic criticism*. Oxford University Press.

Gibbs, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning*. Cambridge University Press.

Hommel, B., Colzato, L. S., Fischer, R., & Christoffels, I. K. (2011). Bilingualism and creativity: Benefits in convergent thinking come with losses in divergent thinking. *Frontiers in Psychology*, *2*, 1–5.

Jarvis, S. (2011). Conceptual transfer: Crosslinguistic effects in categorization and construal. *Bilingualism: Language and Cognition*, *14*(01), 1–8. https://doi.org/10.1017/S1366728910000155

Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. John Wiley & Sons.

Kenett, Y. N. (2019). What can quantitative measures of semantic distance tell us about creativity? *Current Opinion in Behavioral Sciences*, *27*, 11–16. https://doi.org/10.1016/j.cobeha.2018.08.010

Kharkhurin, A. V. (2007). The role of cross-linguistic and cross-cultural experiences in bilinguals' divergent thinking. In I. Kesckes & L. Albertazzi (Eds.), *Cognitive aspects of bilingualism* (pp. 175–210). Springer.

Kroll, J. F., & Ma, F. (2017). The Bilingual Lexicon. In E. M. Fernández & H. S. Cairns (Eds.), *The Handbook of Psycholinguistics* (1st ed., pp. 294–319). Wiley. https://doi.org/10.1002/9781118829516.ch13

Lamarre, P. (2014). Bilingual winks and bilingual wordplay in Montreal's linguistic landscape. *International Journal of the Sociology of Language*, *2014*(228), 131–151. https://doi.org/10.1515/ijsl-2014-0008

Leikin, M. (2013). The effect of bilingualism on creativity: Developmental and educational perspectives. *International Journal of Bilingualism*, *17*(4), 431–447. http://journals.sagepub.com/doi/abs/10.1177/1367006912438300

Lenth, R. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (1.7.2) [Computer software]. https://CRAN.R-project.org/package=emmeans

Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins.

Luk, J. (2013). Bilingual language play and local creativity in Hong Kong. *International Journal of Multilingualism*, *10*(3), 236–250. https://doi.org/10.1080/14790718.2013.808200

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2019). *cluster: Cluster analysis basics and extensions* [Manual].

R Core Team. (2021). *R: A language and environment for statistical computing* (4.1.2) [Computer software]. https://www.R-project.org/

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *ArXiv:1908.10084 [Cs]*. http://arxiv.org/abs/1908.10084

Runco, M. A., & Jaeger, G. J. (2012). The standard definition of creativity. *Creativity Research Journal*, *24*(1), 92–96.

Simpson, P. (2003). *On the discourse of satire: Towards a stylistic model of satirical humour*. John Benjamins Publishing.

Skalicky, S. (2020). Exploring perceptions of novelty and mirth in elicited figurative language production. *Metaphor and Symbol*, *35*(2), 77–96. https://doi.org/10.1080/10926488.2020.1820763

Tokowicz, N. (2014). *Lexical processing and second language acquisition*. Routledge.

van Dijk, M., Kroesbergen, E. H., Blom, E., & Leseman, P. P. M. (2019). Bilingualism and creativity: Towards a situated cognition approach. *The Journal of Creative Behavior*, *53*(2), 178–188. https://doi.org/10.1002/jocb.238

Wang, H., & Cheng, Y. (2016). Dissecting language creativity: English proficiency, creativity, and creativity motivation as predictors in EFL learners' metaphoric creativity. *Psychology of Aesthetics, Creativity, and the Arts*, *10*(2), 205–213. https://doi.org/10.1037/aca0000060

Watson, J., Samir, F., Stevenson, S., & Beekhuizen, B. (2021). Coin it up: Generalization of creative constructions in the wild. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*, 770–776.

Werkmann Horvat, A., Bolognesi, M., & Kohl, K. (2021). Creativity is a Toaster: Experimental evidence on how multilinguals process novel metaphors. *Applied Linguistics*, *42*(5), 823–847. https://doi.org/10.1093/applin/amab002