What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation

Littlemore, Jeannette; Perez-Sobrino, Paula; Houghton, David; Shi, Jinfang; Winter, Bodo

DOI:
10.1080/10926488.2018.1434944

Document Version
Peer reviewed version

Citation for published version (Harvard):
Littlemore, J, Perez-Sobrino, P, Houghton, D, Shi, J & Winter, B 2018, 'What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation', *Metaphor and Symbol*, vol. 33, no. 2, pp. 101-122. https://doi.org/10.1080/10926488.2018.1434944

What makes a good metaphor?

A cross-cultural study of computer-generated metaphor appreciation

Jeannette Littlemore[1], Paula Pérez Sobrino[2], David Houghton[1], Jinfang Shi[3], and Bodo Winter[1]

[1] University of Birmingham, United Kingdom

2Universidad Politécnica of Madrid

[3] East China Jiaotong University

Abstract

Computers are now able to automatically generate metaphors, but some automatically-generated metaphors are more well-received than others. In this paper, we showed participants a series of 'A is B' type metaphors that were either generated by humans or taken from the Twitter account '*MetaphorIsMyBusiness*' (*@MetaphorMagnet*), which is linked to a fully automated metaphor generator. We used these metaphors to assess linguistic factors that drive metaphor appreciation and understanding, including the role of novelty, word frequency, concreteness and emotional valence of the topic and vehicle terms. We additionally assessed how these metaphors were understood in three languages, including English, Spanish and Mandarin Chinese, and whether participants thought they had been generated by a human or a computer. We found that meaningfulness, appreciation, speed in finding meaning and humanness ratings were reliably correlated with each other in all three languages, which we interpret to indicate a more general property of 'metaphor quality'. We furthermore found that in all three languages, conventional metaphors and those that contained an 'optimal' (intermediate) degree of novelty were more likely to be perceived to be of higher quality than those that were extremely creative. Further analysis of the English data alone revealed that those metaphors that contained negatively valenced vehicle words and infrequent vehicle terms (in comparison with the topic terms) were more likely to be considered high-quality metaphors. We discuss the implications of these findings for the (improvement of)

automatic generation of metaphor by computers, for the persuasive function of metaphor, and for theories of metaphor understanding more generally.

1. Introduction

In recent years, in the field of Artificial Intelligence (AI) and Natural Language Processing (NLP), there has been an interest in the automatic generation of creative language by computers (McCormack & d'Inverno, 2012; Veale, 2015a). The over-arching aim of this research is to explore the mechanics of the creative process, and to demonstrate that creativity does not necessarily require high levels of 'human' prowess and specialist knowledge. In order to demonstrate that computers can indeed 'do' creativity, Veale (2015b) developed an automatic metaphor generator, which is linked to a Twitter feed '*@MetaphorMagnet*' (called '*MetaphorIsMyBusiness*'). The bot populates the copula frame "X is a Y" with random word choices every two minutes. Veale reports that many of these metaphors are 'liked' and re-tweeted by his followers, whilst others are simply dropped. What *@MetaphorMagnet* cannot do is work out why some of the tweeted metaphors are successful and more widely shared, whilst others are less successful. In other words, we do not yet know what makes a 'good' metaphor in this environment.

The question of what constitutes a 'good' metaphor within AI and NLP research is an important one. On a practical level, information on what makes a good metaphor can be fed back into metaphor generation algorithms in order to further refine the quality of the metaphors that are produced. On a theoretical level, looking at 'good' and 'bad' computer-generated metaphors helps us understand how metaphor works, and it helps us to understand the process of generating, understanding and appreciating creative figurative language.

In this paper, we compare human-generated metaphors to computer-generated metaphors, looking to see whether English, Spanish and Mandarin Chinese speakers can tell which metaphors were created by a human or not, thus developing a kind of 'Turing test' for

metaphor (Turing, 1950). In addition, we are looking to establish a measure of 'metaphor quality' that incorporates considerations of meaningfulness, appreciation, humanness and speed of processing. Finally, we test several factors that drive metaphor quality in human and computer-generated metaphors, including novelty (are conventional metaphors 'better' than highly creative metaphors?), concreteness (are metaphors with concrete vehicles 'better' than metaphors with relatively more abstract vehicles?), word frequency (are metaphors with frequent words 'better' than metaphors with infrequent words?) and emotional valence (are metaphors with positive words 'better' than metaphors with negative words?). Together, our findings showcase ways of improving existing metaphor generation algorithms and they showcase which linguistic and conceptual factors determine perceived metaphor quality.

2. What makes a 'good' metaphor? Predictors for metaphor quality

Although there is a wealth of literature showing that metaphors in general may be appealing to readers and listeners (e.g., Charteris-Black, 2011; Kittay, 1990), only a small number of studies have sought to identify specific features that render some metaphors more appealing than others. The first study to do so was Katz et al.'s (1988) investigation, which involved the collection of norms for 204 literary and 260 non-literary metaphors on 10 psychological dimensions: comprehensibility, ease of interpretation, metaphoricity, metaphor goodness, metaphor imagery, subject imagery (i.e. topic), predicate (i.e. vehicle) imagery, felt familiarity, semantic relatedness and number of possible interpretations. The average interscale correlation for these psychological dimensions was .76 and the highest correlation was between ease of comprehension and metaphor goodness, which suggests that people are more likely to appreciate those metaphors that they can understand easily. Their findings have since been validated in a replication study conducted by Campbell and Raney (2016), who found even stronger correlations between the different psychological dimensions

(average = .94). The correlation that they identified between felt familiarity and metaphor goodness has also been found in other studies, for example, the conventionality of non-literary metaphors (as measured by corpus frequency) has been found to relate to aptness ratings (Jones & Estes, 2006) and the familiarity of proverbs has been found to contribute to beauty ratings (Bohrn et al., 2013).

The most extensive investigation to date into the pleasure-evoking characteristics of metaphors was conducted by Jacobs and Kinder (2017). They took the literary metaphors from Katz et al.'s (1988) study and attempted to identify the strongest predictors of 'metaphor goodness', as measured in Katz et al.'s study. They began by looking at other variables in the Katz et al. study (such as aptness and the number of interpretations offered); they then moved on to consider a range of lexical and semantic features, such as word concreteness, valence and arousal. In a third stage, they considered interlexical aspects of the metaphors, considering, for example, the relative levels of concreteness of the topic and vehicle terms. In this final stage, they took a constructional approach in which they considered the phrase rather than individual words in the topic and vehicle. They found that 'metaphor goodness' was predicted by a wide range of features, with a high degree of interaction between them. Key features influencing perceived metaphor goodness were ambiguity (i.e. the metaphors that provoked a high number of different interpretations were ranked as being better than those that provoked a small number of interpretations), positive valence in the vehicle in comparison with the topic, high levels of arousal, and a high level of concreteness in the vehicle in comparison with the topic.

In our study, we looked at two of the predictors studied by Jacobs and Kinder (emotional valence and word concreteness). However, as we see below, our hypotheses were somewhat different from theirs, and we investigated additional predictors of metaphor

quality, including word frequency and optimal innovation. Our study also differs from theirs in that we factored in response times.

An additional difference between our study and Jacobs and Kinder's is that we investigated metaphor quality in the context of computer-generated metaphors. Our rationale for doing this is motivated by a study by Gibbs, Kushner and Mills (1991). These authors found that metaphorical expressions are perceived as more meaningful when participants are told that they come from a poet, rather than a computer. Our study further explores the extent to which "humanness" is a dimension relevant to metaphor quality, and it additionally relates "humanness" and other indicators of metaphor quality to four different predictors (novelty, concreteness, frequency, and emotional valence). In the following section, we review each predictor of metaphor quality in turn.

2.1. Predictor 1: Optimal Innovation

The first predictor of metaphor quality we consider is the degree to which metaphors are innovative. The metaphor literature is replete with studies that purport to measure the differences between people's ability to comprehend 'conventional' and 'novel' metaphor (e.g., Lai et al., 2009). However, apart from a small number of studies (e.g., Thibodeau & Durgin, 2008) the category of 'novel' metaphor tends to be very broadly defined and often conflates metaphors that involve novel extensions of existing conceptual mappings with metaphors whose potential meanings cannot easily be traced back to any pre-existing conceptual mapping. For example, the metaphorical expression *'A blue flame shot out of his eyes'*, although 'novel', can be traced back to an underlying conventional mapping ANGER IS HEAT. In contrast, the novel metaphorical expression *'the wall is a purple pen'* cannot easily be traced back to an extant conceptual mapping. This crucial distinction, which is introduced and discussed by Barnden (2015), does not tend to be considered in many studies of metaphor novelty.

A specific proposal which states that novelty should affect metaphor goodness is the 'optimal innovation' hypothesis (Giora et al., 2004), according to which a piece of creative language or art needs to trigger novel, but not *too* novel, inferential activity in order to be appreciated. According to Giora and colleagues (2004: 116), a stimulus is 'optimally innovative' if it involves

(a) a novel—less or non-salient—response to a given stimulus, which

differs not only quantitatively but primarily qualitatively from the salient response(s) associated with this stimulus and

(b) at the same time, allows for the automatic recoverability of a salient response related to that stimulus so that both responses make sense (e.g., the similarity and difference between them can be assessable)

In their study, Giora et al. (2004) found that 'optimally innovative' puns and pieces of arts were consistently rated as being more pleasurable than both conventional and highly innovative stimuli. She also found that understanding 'optimally innovative' stimuli required an intermediate amount of time compared to what was needed to process conventional and highly innovative stimuli. Thus, high innovation is less pleasurable than optimal innovation, and it takes longer to process. 'Optimal innovation' is, in many respects, similar to the notion of 'minimally counterintuitive concepts' put forth by Norenzayan and colleagues (2006). These researchers showed that cultural narratives such as myths and folktales are more likely to achieve cultural stability if they correspond to a minimally counterintuitive cognitive template that includes mostly intuitive concepts (that is, "conventional") combined with a minority of counterintuitive ones (i.e., unexpected or creative).

2.2. Predictor 2: Concreteness versus abstractness

The second predictor of metaphor quality we consider is the degree to which metaphorical vehicles (sources) are concrete, compared to their topics (targets). One of the core tenets of Conceptual Metaphor Theory is that metaphorical mappings are by and large asymmetrical (Jäkel, 1999; Goschler, 2005; Casasanto & Boroditsky, 2008; for discussion, see Winter, Marghetis, & Matlock, 2015). Most often, this asymmetry is characterized in terms of concreteness, with even the earliest formulations of Conceptual Metaphor Theory emphasizing that source domains generally tend to be more concrete than target domains (Lakoff & Johnson, 1980, 1999). Brysbaert, Warriner and Kuperman (2014: 904) define concreteness at the "degree to which the concept denoted by a word refers to a perceptible entity". Given that metaphors are commonly seen as being characterized by concrete-to-abstract mappings, language users may find metaphors that have *relatively* more concrete vehicles (compared to topics) to be easier to understand. Following on from this, they may also find metaphors that fit the concrete-to-abstract principle to be more quickly, pleasing, as well as more 'human'. As we saw above, Jacobs and Kinder (2017) found that literary metaphors were more strongly appreciated when they involved the use of concrete vehicles to describe abstract topics. We were interested in building on this finding by investigating whether the relationship also holds for non-literary, computer-generated metaphors.

2.3. Predictor 3: Word frequency

The third predictor of metaphor quality we consider is the degree to which vehicle and topic terms differ in word frequency. A number of researchers (e.g., Littlemore & Low, 2006; Semino, 2008) have emphasized the evaluative functions of metaphor. In many metaphors, particularly those that appear in the 'A is B' format, the role of the vehicle is to provide some sort of evaluative commentary on the topic, which also means that the vehicle 'B' should provide new information about the topic 'A'. The "Given-New Contract" (see e.g., Clark & Haviland, 1974) states, among other things, that language users expect new information, and

part of language understanding involves incorporating new information with existing knowledge. With this in mind, a good metaphor is likely to be one which offers new information in the vehicle term, i.e., some sort of specific information about the topic, so that it narrows down the range of things one might think about that topic, pinpointing a particular aspect of it to which the producer of the metaphor would like to draw attention. If this were the other way round, the vehicle would not tell us much about the topic. Specific, informative or novel concepts tend to be expressed through lower frequency vocabulary (Shannon & Weaver, 1949; Noble, 1953; Finn, 1977; Brown & Watson, 1987), which leads us to hypothesise that those metaphors in which the topic is expressed through a higher frequency word than the vehicle may be more strongly appreciated than those with the opposite word frequency pattern. In particular, 'A is B' type metaphors where the vehicle 'B' is highly frequent compared to the topic 'A' may not be seen as saying much new about the topic and hence be appreciated less.

2.4. Predictor 4: Emotional valence

Emotional valence refers to the degree to which a word excites pleasant or unpleasant emotions (Warriner, Kuperman, & Brysbaert, 2013), i.e., it is a measure of whether a word is overall positive (*vacation, sunset, happiness*) or overall negative (*murder, warplanes, disease*). Emotionally valenced adjectives clearly change the evaluation of the noun, i.e., a noun's valence out of context is different from the noun's valence in context (Liu, Hu, & Peng, 2013). Crucially though, this modulation effect is strongest if the noun is positive and the modifying adjective is negative (see also Jiang et al., 2014). Other studies have identified a human bias to give greater weight to negative entities (Rozin & Royzman, 2001), with people paying more attention to and remembering negative entities and events more than positive entities and events. This leads to the hypothesis that metaphors that exhibit negative valence in the vehicle will be more strongly appreciated than those that exhibit positive

valence. However, it has also been observed that metaphors with a positively valenced vehicle were more likely to receive a higher 'goodness' rating than metaphors with a negatively valenced vehicle (Jacobs & Kinder, 2017). In light of these findings, we assumed that vehicle valence would affect the 'goodness' ratings of the human-generated and computer-generated metaphors but we were unsure of the direction this relationship would have.

2.5. Metaphor quality and human judgments

Besides assessing how different factors influence metaphor quality, we sought to assess different indicators of metaphor quality and whether they are related to each other, similar to Katz et al. (1988). In our experiment, participants were asked for each 'A is B' type metaphor whether it made sense, whether it was appreciated, and whether it was generated by a human. We hypothesised that these three measures were related to each other. In particular, in the context of a task that featured both human-generated metaphors and computer-generated metaphors, those metaphors that were judged to make more sense and that were also appreciated more should also be more likely to be judged as human, presumably because participants think that computers are simply not as good in something that is as creative as generating metaphors. Viewing "humanness" as a dimension of metaphor quality is also motivated by the study of Gibbs and colleagues (1991), which found that metaphorical expressions were perceived as more meaningful when participants were told that they were generated by a human, rather than by a computer.

2.6. Ease of processing

There appears to be a relationship between appreciation and ease of comprehension. In their study of metaphors in advertising, Pérez-Sobrino, Littlemore and Houghton (submitted) found that well-liked advertisements were understood more rapidly than less liked

advertisements. This may be because participants were more likely to appreciate the advertisements that they understood more quickly, or alternatively, that they understood them more quickly because they liked them. This finding appears to mirror what has been found in the literature on cognitive fluency, where easy-to-process stimuli are generally liked more than difficult stimuli (Zajonc, 1968; Reber, Winkielman, & Schwarz, 1998; Oppenheimer, 2008). Based on these findings, we thus expect 'better' metaphors to be processed more quickly.

In addition, there are reasons to expect the above-mentioned optimal innovation hypothesis to play a role in speed of processing. It is a known fact that familiar stimuli are processed more quickly (e.g., Solomon & Postman, 1952; Postman & Conger, 1954; Jescheniak & Levelt, 1994). Given this, in conjunction with the above-mentioned optimal innovation hypothesis, we expect highly unconventional metaphors (those that are beyond optimal innovation) to take more cognitive processing effort, and thus, more processing time.

2.7. Cross-cultural variation

A final question that is of interest when exploring the acceptability ratings of metaphorical tweets is the extent to which such ratings vary cross-culturally. It has been shown that metaphor use varies across different languages and cultures (Kövecses, 2005), but we do not know whether the factors that affect metaphor quality are similar across languages and cultures. In our study we were interested in exploring the issue of cross-cultural variation between two languages that are linguistically related (English and Spanish) and situated geographically close to each other, and one (Mandarin Chinese) which is not related to the Indo-European languages and geographically more distant relative to the other two languages.

3. Research questions and hypotheses

In order to break down our general question, 'What makes a 'good' metaphor?', we formulated a subset of dimensions along which *goodness* can be measured, similar to the approach taken in Katz et al. (1988). The first one was whether people find *meaning* in the expression ("sense")[i]. Second, whether they rate it to be *pleasant* ("appreciation"). Third, whether people can tell *human* metaphors from non-human metaphors in a set of expressions that contains both ("humanness").

Therefore the first question that was of interest to us was: 'To what extent are metaphors that are judged to make sense, also appreciated and thought to be human?' Having established the nature of these relationships, our next question was: 'To what extent and in what ways do optimal innovation, word frequency, concreteness of the source and target terms and emotional valence impact one's assessment of whether a metaphor constitutes a 'good' metaphor, as well as the speed with which one makes that decision?' Finally, we aimed to investigate the extent to which all of these factors vary across the three languages, English, Spanish and Mandarin Chinese.

Given our discussion above, our specific hypotheses were formulated as follows:

H1. There will be a positive relationship between ease in finding meaning, appreciation and tendency to think that the metaphor was created by a human.

H2. Metaphor quality, as constituted by sense, appreciation and humanness, will be positively related to speed in finding meaning.

H3. Optimally innovative metaphors will be more strongly appreciated and more rapidly understood.

H4. Metaphors that combine a concrete vehicle term with a more abstract topic term will be more strongly appreciated.

H5. Metaphors that combine a low frequency vehicle term with a high frequency topic term will be more strongly appreciated.

H6. The valence of the vehicle will affect levels of appreciation but the direction of this relationship is uncertain.

To summarise, our experiments sought to establish that different measures of metaphor quality were meaningfully related to each other (H1) and that these are related to speed of processing (H2). We did this for English, Spanish and Mandarin Chinese. We furthermore looked specifically at different predictors of metaphor quality. For all three languages, we assessed the role of optimal innovation (H3). For English alone (due to data availability, see below), we assessed the effects of concreteness (H4), word frequency (H5) and emotional valence (H6) on our conglomerate measure of metaphor quality.

4. Method

4.1. Participants

21 British English, 18 Spanish, and 22 Mandarin Chinese speakers were recruited for the study (N=61). All participants were native speakers of these languages. English-speaking and Mandarin Chinese-speaking participants were recruited and tested in Birmingham during April and May 2014, and Spanish-speaking participants were recruited and tested in Logroño (Spain) during June and July 2014. Care was taken to ensure that Mandarin Chinese-speaking participants had not been in the UK for more than eight months. Table 1 shows the demographics of the participants by nationality, gender and age group is as follows:

Age group	**English** **(12 female, 9 male)**	**Spanish** **(13 female, 5 male)**	**Mandarin Chinese** **(18 female, 4 male)**

18-30	5	11	13
31-45	7	5	7
46+	9	2	2
Total	**21**	**18**	**22**

Table 1. Demographics of the participants recruited for this study

4.2. Selecting the metaphor stimuli

We selected our computer-generated metaphors from the '*@MetaphorMagnet*' database (Veale, 2015b). This database contains over 650 million words of Tweets, containing a variety of different types of metaphor. In order to minimise the effect of phraseological features we selected those that contained 'A is B' type metaphors where the vehicle was always an adjective-noun pair, such as "love is a comforting fire". We randomly selected fifty of these 'A is B' metaphors. For human-generated metaphors, we selected 10 non-literary metaphors from the aforementioned lists of normed metaphors by Katz et al. (1988) and Campbell and Raney (2016), all of which have been generated by humans. Metaphors were chosen that reflected a range of scores for comprehensibility, ease of interpretation and metaphor goodness as these were the variables that were of most relevance to our study. We added modifiers to these metaphors in order to make them match the syntactic pattern of the computer-generated metaphors. We did this in such a way as to ensure minimal meaning changes to the original metaphors. For example, the metaphor 'danger is a spice' became 'danger is a hot spice' and the metaphor 'wisdom is a foreigner' became 'wisdom is a distant foreigner'. All the metaphors were then translated into Spanish and Mandarin Chinese. This was first done using Google Translate, and the translations were subsequently checked and

corrected for idiomaticity by native speakers of Spanish and Mandarin Chinese. A full list of the metaphors in English, Spanish and Mandarin Chinese can be found in Appendix A.

4.3. Coding the metaphors for optimal innovation

We began by coding the metaphors into three levels of novelty: those that were conventional (1), those that displayed 'optimal' (i.e. intermediate) innovation (2), and those that displayed high innovation (3). Following Giora and colleagues (2004), we predicted that optimally innovative metaphors would be the most strongly appreciated.

We operationalised optimal innovation with respect to Conceptual Metaphor Theory (see Barnden, 2015). To do this, we took the *Master Metaphor List* as a basis to identify conventional metaphors (Lakoff, Espenson, & Schwartz, 1991; now extended and available through MetaNet https://metanet.icsi.berkeley.edu/metanet/). Those metaphors that involved a mapping that could be traced back to an established conceptual metaphor present in the list (e.g. *education is a glowing lantern*, which involves the conceptual metaphor KNOWLEDGE IS LIGHT) were labelled 'conventional' and given a score of 1. Conventional linguistic metaphors (e.g., *death is relaxing sleep*) were also given a score of 1.

Metaphors that exploited a conventional source domain in a creative way were labelled as being 'optimally innovative' and given a score of 2. An example of one such metaphor is '*indecision is a calm whirlpool*'. This expression draws on the conceptual metaphors MENTAL CONTROL IS PHYSICAL CONTROL and PSYCHOLOGICAL FORCES ARE PHYSICAL FORCES. The expression uses the term 'whirlpool' to represent some form of chaos or lack of control, which is a common pattern in the British National Corpus (35 instances of this usage were found in the first 100 citations). However, the fact that the whirlpool is described as '*calm*' means that it contains an element of innovation. In fact, this example also

aligns with the above-mentioned idea of a minimally counterintuitive idea (Norenzayan et al., 2006), as there is a somewhat counterintuitive element of contradiction between calmness and a whirlpool (which is inherently not calm).

Metaphors where the source domain was used in combination with a target domain that does not evoke a conventional mapping were labelled as being 'highly innovative' and given a score of 3. An example of one such metaphor is '*power is pure water*'. Although water (or at least fluid) is used as a source domain in some conceptual metaphors, it is usually used to connote movement or purity, and neither of these is easily mapped onto the domain of power.

Using these criteria, three coders independently coded the sixty English metaphors selected for the study. We then met to discuss cases of disagreement, relying on the above-mentioned *Master Metaphor List* (Lakoff, Espenson, & Schwartz, 1991) to check for conventional metaphors, as well as on the BNC and the NOW corpus to identify authentic usages. The agreed levels of innovation for the English and Spanish versions of the metaphors were identical to one another, but for the Chinese versions levels of innovation were deemed to be different in 12 cases. These differences were due to variation in the cultural connotations of some of the metaphors, in particular, those whose meanings depended on attitudes towards religion, the government, and alcohol. For example, the metaphor '*Government is a warming light*' was scored as 'highly innovative' in English and Spanish but 'conventional' in Chinese.

It should be noted here that metaphors produced by a human rather than a computer only received innovation ratings of 1 or 2. That is to say, none of the human-generated metaphors were rated as 'highly innovative'. This may reflect a human tendency to rely on known A is B mappings and then to innovate within these mappings.

4.5. Experimental procedure

The task was computerised delivered via E-Prime 2.0. Participants received the following instructions:

> *A metaphor is a statement which is not literally correct, but which establishes a relationship between two parts of a sentence. The ease with which this relationship can be interpreted can vary.*
>
> *For example, the statement 'snow is a winter coat' is an obvious metaphor – snow is not a winter coat, but the idea of a winter coat provides relevant information about snow – it covers everything, it keeps you warm, it's thick, and so on.*
>
> *It is also fairly easy to make sense of the statement 'memories are a slippery snake', even though interpretations might vary: memories can be unreliable (i.e. hard to pin down), long, they can go round in circles, they can be dangerous, and so on.*
>
> *On the other hand, it is more difficult to find a metaphorical meaning for an expression such as 'a piano is a soup spoon'; it is difficult to see what kind of information a soup spoon can give about a piano.*
>
> *Most of the metaphors in the experiment have been generated by a computer program called 'Metaphor Magnet', and have been posted on Twitter. We also have some expressions in the selection that have been generated by humans.*

Following these instructions, participants were first asked: 'Does this metaphor make sense?' (categorical yes/no response) Then: 'If you said yes, how meaningful is the expression?' (on a scale from 0 "I didn't say that it made sense!" to 3 "Completely meaningful") Then: 'How would you rate the quality of the metaphor you have just read?' (With a five-step continuum of smileys ranging from a frowny face to a smiley face). Finally,

we asked participants: 'Who or what generated this metaphor?' (With a two forced choice response option, a silhouette of a human and a vector image of a computer)[ii]

4.6. Data analysis

All analyses were conducted with the R programming language (R Core Team, 2016) and the packages described in Appendix B. The full datasets and analysis script are accessible under the following publicly accessible GitHub repository:

https://github.com/bodowinter/good_metaphors

The statistical analyses proceeded as follows. For each language, we first looked at correlations between the three dependent measures (sense, quality, and humanness ratings). From these three dependent measures we extracted a conjoined measure that we called 'metaphor quality' which was used in further analyses (H1). We then used this conglomerate measure to test for the factor novelty in all three languages (H3), assessing the optimal innovation hypothesis (see details below). For all languages we furthermore looked at the correlation between metaphor quality and speed (H2). Finally, for English only, we had additional data available that allowed us to test the effects of additional factors on metaphor quality. In particular, we incorporated concreteness ratings (Brysbaert et al., 2014), word frequencies from the British version of the SUBTLEX movie subtitle corpus (Keuleers, Lacey, Rastle, & Brysbaert, 2012) and emotional valence ratings (Warriner et al., 2013) into our analysis. We chose word frequencies from the SUBTLEX corpus because these are widely used in psycholinguistics and have been argued to be highly predictive of linguistic processing, more so than frequencies from other corpora (e.g., Brysbaert & New, 2009). For each of these analyses we performed a simple linear regression of 'metaphor quality' on concreteness, word frequencies and emotional valence. We used an asymmetry score which

measured the extent to which A was more or less concrete / frequent / positive than B (by computing the difference of A minus B). To compute this asymmetry score, for vehicle terms which contained adjective-noun pairs, we took the average concreteness / frequency / valence of the adjective and noun, which was then compared to the corresponding values of the topic.

In all analyses that follow, we focus on an items analysis (Clark, 1973). We deliberately ignored subject variation by averaging over subjects and obtaining one data point per item. This is in line with what has been done by Katz et al. (1988) and others, and it is commonly done in the analysis of norming (rating) studies (e.g., Lynott & Connell, 2012; Kuperman, 2015). Focusing on an items-analysis is justified as our research questions are focused on 'What makes a good metaphor?', for which the individual item is the relevant unit of replication. To the extent that the core results of our analyses replicate in three different languages the influence of subject variation is less of a concern.

5. Results

5.1 Correlations between sense, quality, and humanness

Following our items-based approach, average sense, appreciation and humanness scores were computed for each of the 60 metaphors. Figure 1 shows a matrix of scatterplots of the three dependent variables. The figure shows that there are high correlations between all the dependent variables that we hypothesised to be related. Spearman's rank correlations show that sense ratings were reliably correlated with appreciation ratings ($rho = 0.94$, $p < .001$) and humanness ratings ($rho = 0.86$, $p < .001$). Moreover, appreciation ratings and humanness ratings were reliably correlated with each other ($rho = 0.86$, $p < .001$).

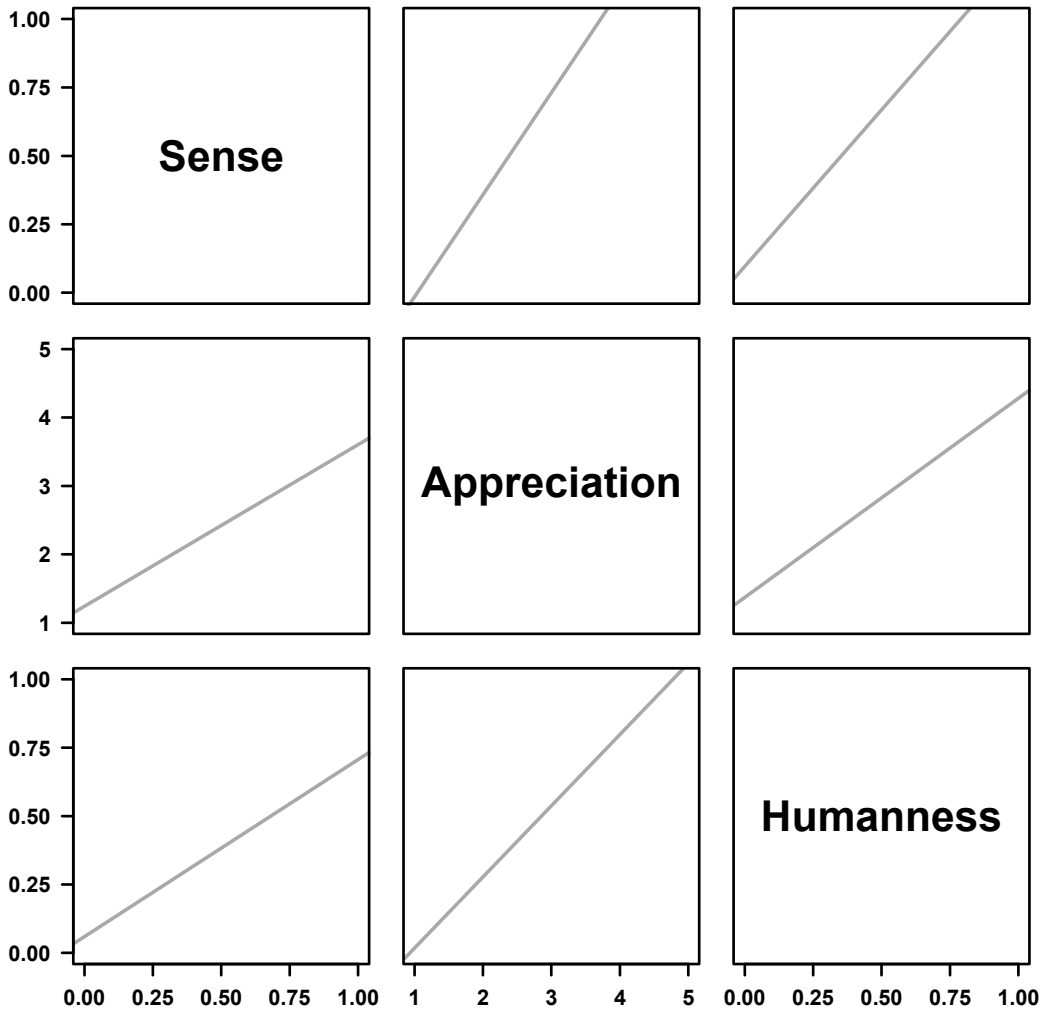

Figure 1. Scatterplot matrix showing correlations between the three main dependent measures (each data point represents one metaphor stimulus); the first row shows sense ratings (from 0% to 100% of participants who said "make sense"); the second row shows appreciation ratings (average from 1 to 5 rating scale); the third row shows humanness ratings (from 0% to 100% of participants who said "is human"); superimposed lines are simple linear regression fits

The strong correlation between all three dependent variables licenses considering them in a conjoined fashion. A Principal Component Analysis (with singular value decomposition and scaled variables) shows that all three variables load strongly onto one dimension (the first component), which alone already explains 93% of the overall variation in responses to the stimuli. A high value on this dimension indicates that the metaphor is judged high in sense, appreciation, and humanness ratings; conversely, a low value on this dimension indicates that the metaphor is judged low in sense, appreciation and humanness ratings. Thus, the conjoined measure can be seen as capturing what we here call 'metaphor quality'.

5.2. The relationship between novelty and metaphor quality

5.2.1. English

We performed a linear regression analysis in which the dependent variable 'metaphor quality' (first principal component of the Principal Component Analysis, see section 5.1.) was modelled as a function of several different predictor variables, i.e., what affects metaphor quality? The linear regression performed in this section contained the factors 'human' (whether the metaphor was actually human- or computer-generated) and 'novelty', following the coding procedure outlined above (with values ranging from 'conventional' = 1, through 'optimally innovative' = 2, to 'highly innovative' = 3). The factor novelty was added as a linear and a quadratic effect. The linear effect measures the extent to which metaphor quality increases or decreases linearly with higher values of novelty, i.e., whether 3 is more than 2 is more than 1 (or conversely, less than). The quadratic effect measures the extent to which there is a nonlinear relationship between novelty and metaphor quality. Assessing the quadratic effect of novelty can be taken as a direct test of the optimal innovation hypothesis,

as this hypothesis would state that optimally innovative metaphors (with the value 2) would be higher in metaphor quality than entirely conventional (1) or novel (3) metaphors.

Overall there was a reliable overall effect of novelty (linear and quadratic effect together, $F(2, 55) = 36.40$, $p < .001$). Both the linear ($F(1, 56) = 58.67$, $p < .001$) and the quadratic ($F(1, 55) = 7.40$, $p = .009$) novelty effect separately had a statistically reliable influence on metaphor quality. Whether the metaphor was ultimately human- or computer-generated missed the mark of statistical significance ($F(1, 55) = 3.94$, $p = .052$). Looking at R^2 values shows that the human effect described only about 3% of the variance in metaphor quality, compared to the quadratic novelty effect, which described 5% of the variance in metaphor quality. The linear novelty effect described 46% of the variance. Figure 2 shows the effect of novelty for English. As can be seen, there is a sudden drop between those stimuli that received a novelty rating of 2 and those that received a novelty rating of 3, which is the reason for the quadratic effect being statistically reliable in the regression model. While an inverted U-shaped curve would have been an even stronger confirmation of the optimal innovation hypothesis, this data suggests a somewhat weaker version of the hypothesis for our data, with a plateau between highly conventional (=1) and optimally innovative (=2) metaphors and a sudden drop in quality for highly novel (=3) metaphors.

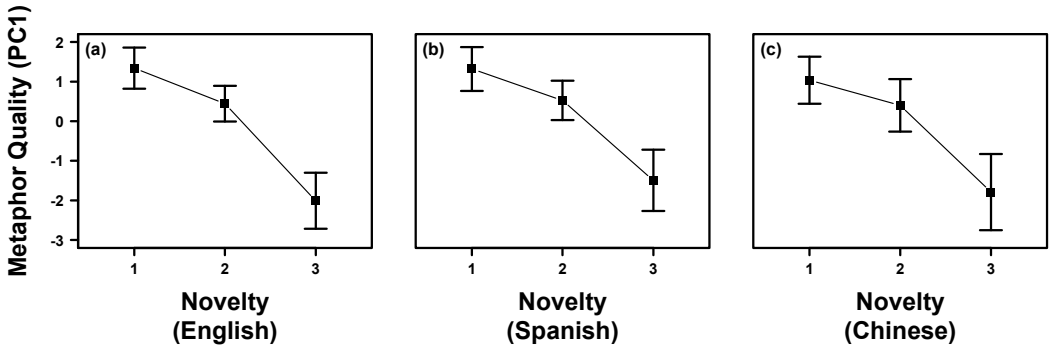

Figure 2: Predicted metaphor quality as a function of novelty (including linear and quadratic effect) for (a) English, (b) Spanish and (c) Mandarin Chinese; values taken from the

regression model described in the text; error bars show 95% confidence intervals over the predictions

5.2.2. Spanish

For Spanish participants, the three dependent measures were also reliably correlated with each other. Sense judgments were correlated with appreciation ratings (*rho* = 0.93, $p < .001$) and human judgments (*rho* = 0.79, $p < .001$). Also, quality ratings were reliably correlated with human judgments as well (*rho* = 0.84, $p < .001$). The Principal Component Analysis yielded very similar results to the English data, with the first component explaining 90% of the variance. This again suggests that the three variables behave in a highly similar fashion, which warrants conjoining them into a conglomerate measure of 'metaphor quality'.

Just as we did for English, this metaphor quality variable was subjected to a linear regression analysis with the factors 'novelty' (linear and quadratic) and 'human' (yes or no). Overall there was a reliable effect of novelty ($F(2, 56) = 21.63$, $p < .001$). The linear novelty effect had a statistically reliable influence on metaphor quality ($F(1, 57) = 37.48$, $p < .001$). This time, the quadratic effect missed the mark of significance ($F(1, 55) = 3.88$, $p = .054$). Whether the metaphor was ultimately human- or computer-generated was statistically reliable ($F(1, 56) = 7.32$, $p = .009$). The human effect described about 6% of the variance in metaphor quality. The quadratic novelty effect described 3% of the variance in metaphor quality; the linear novelty effect described 34% of the variance. As seen in Figure 2b, there was a rapid drop-off in metaphor quality for the highly novel metaphors, which again provides partial support for the optimal innovation hypothesis.

5.2.3. Mandarin Chinese

Just as was the case with the other two languages, the three dependent measures were also reliably correlated with each other in Mandarin Chinese. Sense judgments were correlated with quality ratings (*rho* = 0.93, $p < .001$) and human judgments (*rho* = 0.93, $p < .001$). Also, quality ratings were reliably correlated with human judgments as well (*rho* = 0.92, $p < .001$). The Principal Component Analysis yielded very similar results, with the first component explaining 95% of the variance. This again suggests that the three dependent measures behaved primarily in a highly correlated fashion, which warrants conjoining them into a conglomerate measure of 'metaphor quality', which was used for the subsequent regression analyses.

Overall there was a reliable effect of novelty on metaphor quality ($F(2, 55) = 13.99$, $p < .001$). Both the linear ($F(1, 56) = 22.43$, $p < .001$) and the quadratic ($F(1, 55) = 4.3$, $p = .04$) novelty effect had a statistically reliable influence on metaphor quality. Whether the metaphor was ultimately human- or computer-generated had no statistically reliable influence on metaphor quality ($F(1, 55) = 0.27$, $p = .61$). The human effect described close to 0% of the variance in metaphor quality. The quadratic novelty effect described about 5% of the variance in metaphor quality, and the linear novelty effect about 28%. As seen in Figure 2c, there was a rapid drop-off in metaphor quality for the highly novel metaphors, which again at least partially supports the optimal innovation hypothesis.

5.2.4. Interim summary

In sum, for three separate languages—English, Spanish, Mandarin Chinese—we found highly similar results, with the three dependent measures of sense, appreciation and humanness correlating very strongly with each other in each language, which suggests that combining them into a unitary measure of 'metaphor quality' is justified. This measure of metaphor quality was furthermore influenced by novelty in all three languages in the same way: More novel metaphors were appreciated less, but there was a reliable nonlinear relationship for

each language with a sudden drop-off for highly novel metaphors. Conventional and optimally innovative metaphors were more similar to each other. Finally, the effect of human versus non-human metaphors was marginal in all cases. It was statistically reliable for Spanish only, with non-reliable effects for English and Mandarin Chinese.

5.3. Novelty and metaphor quality in relation to speed of processing

We were interested in whether speed of processing is influenced by novelty (as per Giora et al., 2004) or metaphor quality (as was found for visual advertisements and multimodal metaphor by Pérez-Sobrino, Littlemore and Houghton, submitted). Because reaction times are something that relate to the individual, we performed a statistical analysis that actively models both subject and item variation, using linear mixed effects models with subjects and items as random effects (Baayen, Davidson, & Bates, 2008), including by-subject-varying random slopes for the main fixed effect in question (in this case either novelty or metaphor quality). For all reaction time analyses, the reaction times were log-transformed. Following this, log reaction times that were 1.96 standard deviations above or below a participant's mean response were excluded.

In the first model we test whether metaphor quality had an effect on speed of processing. This was not the case for English ($\chi^2(1) = 0.14$, $p = .71$), Spanish ($\chi^2(1) = 0.42$, $p = .51$) or Mandarin Chinese ($\chi^2(1) = 1.66$, $p = .2$). In other words, there was no statistically reliable overall effect of metaphor quality on speed of processing. Thus, there was no statistical support for H3.

Still, the role of novelty in speed of processing remained to be tested. Did novelty affect speed of processing? In the following linear mixed effects model, we included novelty as linear and quadratic effects (see section 5.2), as well as response ("makes sense" versus

"no sense") and the interaction of novelty and response. We additionally included the factor "human" versus "not human" into the model (no interactions)[iii].

For the English data, there was a statistically reliable interaction between novelty and response (linear and quadratic effect together, $\chi^2(2) = 25.03$, $p < .001$). The predicted log reaction times are shown in Figure 2a. As can be seen, "sense" responses were about equally fast for metaphors that were conventional (novelty = 1) and somewhat novel (novelty = 2), but they were slower for highly novel metaphors (novelty = 3), in line with the idea that highly novel metaphors take more time to process. On the other hand, responses that indicated that the metaphor made "no sense" were fastest for the maximally innovative metaphors. This indicates that participants found it easier to decide that unconventional metaphors made no sense, compared to conventional metaphors. Descriptive averages for the "makes sense" responses are about 3,850 ms in responding to conventional metaphors (novelty = 1), followed by 3,570 ms (novelty = 2) and 4,290 ms (novelty = 3). For the "no sense" responses, responses were slowest for the least novel metaphors, with about 5,100 ms, followed by 4030 ms (novelty = 2), and fastest for the highly novel metaphors (about 3,500 ms). There additionally was a barely reliable main effect of novelty ($\chi^2(2) = 6.07$, $p = .048$), with responses being slowest for conventional metaphors (novelty = 1). There was a reliable effect of response ($\chi^2(1) = 13.90$, $p < .001$), with participants being slightly faster in responding "makes sense" than in responding "makes no sense".

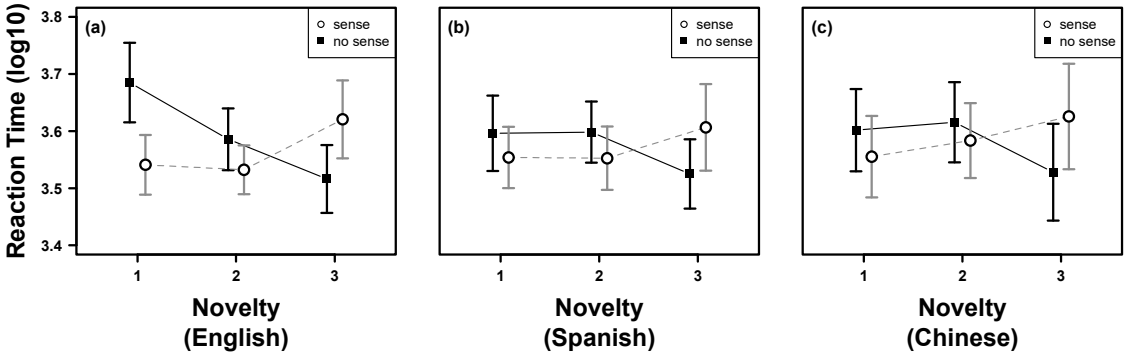

Figure 3. Predicted log reaction times of the linear mixed effects model analysis as a function of novelty for (a) English, (b) Spanish and (c) Mandarin Chinese; solid black line with black squares indicates "no sense" responses; dashed grey line with white circles indicates "makes sense" responses; error bars indicate 95% confidence intervals

The picture was somewhat similar for Spanish (see Figure 3b). Again, there was a reliable interaction between novelty and response ($\chi^2(2) = 9.90$, $p = .007$). This time, however, there was no reliable main effect for novelty ($\chi^2(2) = 0.78$, $p = .68$) or response ($\chi^2(1) = 0.52$, $p = .47$). This is also apparent in Figure 3b. In contrast to the English speakers, Spanish participants were not overall slower or faster in performing a "sense" or "makes no sense" judgment, and they equally were not overall faster as a function of novelty. However, like the English speakers and in line with the optimal innovation hypothesis, processing speed depended on the conjoined effect of response type and novelty (interaction). In particular, "makes sense" judgments were slowest for highly novel metaphors (novelty = 3), and for these metaphors, "makes no sense" judgments were the fastest. In terms of descriptive averages, "sense" judgments took 3,720 ms, 3,860 ms and 3,910 ms in order of increasing novelty. The "no sense" judgments took 3,950 ms, 4170 ms and 3,410 ms.

Finally, for the Chinese participants (Figure 3c), there again was a reliable interaction between novelty and response type ($\chi^2(2) = 9.09$, $p = 0.01$) and, like the Spanish speakers,

there were no main effects of novelty ($\chi^2(2) = 2.32$, $p = 0.31$) or response type ($\chi^2(1) = 0.93$, $p = 0.33$). Figure 3c shows that the pattern of response types is quite similar to the Spanish speakers, and it is also similar to the English speakers in that only for highly novel metaphors are "no sense" responses much faster and "makes sense" responses much slower. In terms of descriptive averages, "sense" judgments took 4,160 ms, 4,270 ms and 4,840 ms in order of increasing novelty. The "no sense" judgments took 4,320 ms, 4,690 ms and 4,070 ms.

Were participants faster in responding to human or in responding to computer-generated metaphors? The effect of humanness was statistically reliable for English ($\chi^2(1) = 10.57$, $p = .001$) and Spanish ($\chi^2(1) = 8.2$, $p = .004$), but not for Mandarin Chinese ($\chi^2(1) = 1.59$, $p = .21$). For the English participants, computer metaphors were responded to with an average of 3,800 ms, human metaphors were slower with an average of 4,000 ms. The same pattern held for the Spanish participants (3,660 ms computer versus 4,530 ms human) and in a weaker fashion also for the Chinese participants (4,310 ms computer versus 4,490 ms human).

Altogether, these results show a clear effect of novelty on speed of processing metaphors, with conventional metaphors and optimally innovative metaphors being processed in a similar fashion, compared to highly unconventional metaphors (novelty = 3).

5.4. Perceived humanness

To assess the effect of humanness, a simple logistic regression model was fitted on the subset of stimuli for which both human and computer-generated metaphors exist (novelty = 1 and novelty = 2). The logistic regression model fitted the proportion of human judgments as a function of whether the metaphor had actually been created by a human (yes versus no), while simultaneously controlling for novelty (simple linear effect). There was indeed a reliable effect of humanness (log odds estimate: +0.45, $SE = 0.16$, $z = 2.8$, $p = .0049$), with

metaphors that had been created by humans being judged as human 1.57 more likely then computer metaphors. The same effect held for Spanish (log odds: 0.65, *SE* = 0.17, *z* = 3.78, *p* = .0001), with human metaphors being judged as human 1.8 times more likely than computer metaphors. The same effect was not found for Mandarin Chinese (log odds: -0.11, *SE* = 0.18, *z* = -0.61, *p* = .54). Thus, only for Mandarin Chinese did we fail to find an effect of humanness.

5.5. Additional analysis for English: word frequency, valence, and concreteness

To further explore what makes a good metaphor, we hone into English, for which there are multiple databases that can be used to investigate factors driving metaphor quality. In this section, we look at word frequency, emotional valence and concreteness asymmetries between the topic ("A") and the vehicle ("B"). It needs to be kept in mind that all metaphors have the format "A is B", with B always consisting of an adjective-noun pair. In this analysis, we computed the average across the adjective and noun. In the first analysis, we created a difference between the log frequency (British SUBTLEX) of the vehicle minus the log frequency of the topic, i.e., the extent to which B is more frequent than A. This relationship is visualised in Figure 4a. The figure shows that participants preferred metaphors of the A is B type more if the A (vehicle) is a relatively frequent word, compared to the B (topic). This is in line with the idea that the vehicle is supposed to say something specific and novel about the topic. The relationship between frequency asymmetry and metaphor quality was statistically reliable ($F(1, 57) = 12.87$, $p < .001$, adjusted $R^2 = .17$).

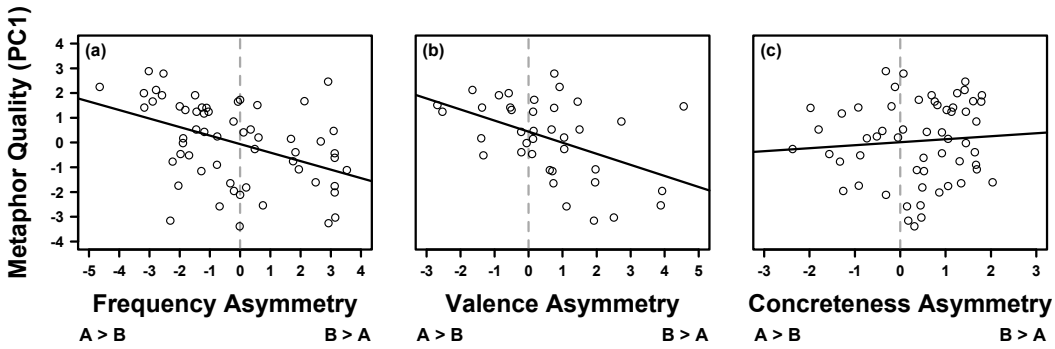

Figure 4. Metaphor quality (first principal component) as a function of (a) frequency asymmetry (to the right the vehicle is more frequent than the topic), (b) valence asymmetry (to the right the vehicle is more positive) and (c) concreteness asymmetry (to the right the vehicle is more concrete); lines indicate simple linear regression fits with superimposed 95% confidence region

On the other hand, participants preferred metaphors where A was more positive than B, or in other words, where a relatively more negative vehicle was used to describe a relatively more positive topic. The effect of emotional valence asymmetry was statistically reliable ($F(1, 36) = 9.90$, $p = .003$, adjusted $R^2 = .19$). Unfortunately, the valence data from Warriner et al. (2013) was only available for 38 of the 59 stimuli (64%). To assess whether this result was robust and to also provide converging evidence using a different approach, the norms from Snefjella and Kuperman (2016) were used. These authors computed the average "context valence" by looking at the valence of the five content words preceding or following a head word, based on the idea that words have what is sometimes called "semantic prosody" (Sinclair, 1991; Louw, 1993; Hunston, 2007; Stewart, 2010), i.e., a collocational profile that preserves consistent evaluation (positive / negative). This measure existed for all words in our stimuli, and there also was a reliable effect on metaphor quality ($F(1, 57) = 9.32$, $p = .003$, adjusted $R^2 = .13$) in the same direction. This shows that metaphors were preferred if the

vehicle contained words that are found to have more negative semantic profiles innaturally occurring corpus data.

These two analyses (emotional valence and context valence) support the idea that English speakers prefer A is B type metaphors where B is relatively more negative than A, broadly in line with the view that negative information is more salient (Rozin & Royzman, 2001) and able to modify a topic more strongly than positive information (Jiang et al., 2014). This shows that emotional valence does affect metaphor appreciation, as was also found by Jacob and Kinder (2017), however, we found an effect in the opposite direction. Whereas Jacobs and Kinder found that more *positively* valenced vehicles (in comparison with the topics) correlated with metaphor goodness ratings, here it is more *negatively* valenced vehicles. One possible reason for this difference could be that the participants in their study were judging literary metaphors whereas we were focussing on non-literary metaphors. The expectations of the participants in the two studies may therefore have been different (compare Gibbs et al., 1991). Jacobs and Kinder (in press) found marked differences between the literary metaphors and the non-literary metaphors in Katz et al.'s study. In particular, they found that literary metaphors were rated as more familiar and more difficult to interpret, and that they were more likely to contain dissimilar topic and vehicle terms. However crucially, the participants in Katz et al.'s study were not told whether the metaphors were literary or non-literary. As we saw above, metaphor processing and appreciation have been found to be affected by what the raters are told about the origin of the metaphors.

Finally, there was no statistically reliable effect of concreteness asymmetry (see Figure 4b) ($F(1, 55) = 0.35$, $p = .56$, adjusted $R^2 = -.01$). Thus, it did not matter whether B was relatively more or less concrete than A. Our analyses suggest that at least for A is B type metaphors, 'good' metaphors do not necessarily need concrete vehicles and abstract targets, as Conceptual Metaphor Theory would suggest. Instead, the metaphorical asymmetry appears

to be more strongly driven by emotional valence and word frequency. Again, this finding contrasts with Jacobs and Kinder's (2017) finding that high levels of vehicle concreteness, in comparison with the topic, *did* affect metaphor goodness ratings. As was mentioned above, this might be explained by the differences in the stimuli and what the participants had been told about their origin. However, we can conclude that at least for this dataset, what makes a 'good' metaphor is more strongly affected by asymmetries in word frequency and emotional valence than by the concrete-to-abstract principle.

The findings presented above did not change if metaphor quality was simultaneously regressed on the three factors (frequency, valence and concreteness) within the same regression model. This conjoined model altogether described 30% of the variance in metaphor quality (variance inflation factors indicate no problem with collinearity).

6. Discussion

The findings from our study provide broad support for the majority of our hypotheses. There were strong positive relationships between ease in finding meaning, appreciation and tendency to think that the metaphor was created by a human. This allowed us to consider these three outcome measures together as one unitary variable, which we called 'metaphor quality'. Our main finding, which was consistent across all languages, was that participants were more likely to find meaning in conventional and moderately innovative metaphors. In turn, they were much less likely to find meaning in highly innovative metaphors. These results do partially support the optimal innovation hypothesis (Giora and colleagues, 2004), which predicts that novelty does not have an across-the-board effect on metaphor quality but a nonlinear one, which is exactly what our results showed. However, our results do not fit the optimal innovation hypothesis completely: While there was a sudden drop-off for highly

novel metaphors, compared to optimally innovative metaphors, there was no "boost" for optimally innovative metaphors. Instead, conventional and optimally innovative metaphors behaved similarly in our study.

Our reaction time findings indicated that only the highly innovative metaphors showed a strong processing difference between 'sense' and 'no sense' judgments. It was easier for participants to reject a highly innovative metaphor as nonsensical. If they answered that a highly innovative metaphor made sense, they took much more time to do so. For Spanish and Mandarin Chinese, there were no differences in response times between moderately innovative and highly innovative metaphors, which again could be taken as loosely in line with the optimal innovation hypothesis. For English the pattern was similar, but English speakers were also much slower to say that a highly conventional metaphor made 'no sense'. To sum up, metaphors were preferred and processed more quickly if they were conventional or optimally innovative. If they were too innovative then people did not appreciate them. This finding held for all three languages studied.

Both the English and the Spanish participants were reliably able to distinguish those metaphors that had been generated by a computer from those that had been generated by a human. Interestingly, the Chinese participants were less able to spot the difference. One reason for this may have been that the metaphors used in the study had originally been created in English and were subsequently translated into Mandarin Chinese. The linguistic and geographical differences between these two languages may have meant that the human-generated metaphors may have sounded more alien to the Mandarin Chinese participants, even though they were based on apparently 'universal' conceptual metaphors. This finding extends other work suggesting that culture plays an important role in determining the ultimate meaning of linguistic expressions that reflect underlying conceptual metaphors (Kövecses, 2005).

In addition, we performed several more detailed analyses of the factors driving metaphor quality in English. These additional analyses for English showed that if the vehicle was relatively more frequent than the topic, metaphors were liked less. They also showed that participants preferred metaphors where the vehicle was more negative than the topic, i.e., where something bad was said about the topic. This means that the comment offered by the vehicle should be specific and negative, i.e., pointed. Word concreteness did not affect appreciation. This suggests that at least in A is B type metaphors, the topic and vehicle do not have to differ in concreteness in order to qualify as 'good' metaphors. In contrast, emotional valence and word frequency matter more. Word frequency has also been found a predictor of metaphorical and metonymical asymmetries in crosslinguistic data (Winter, Thompson, & Urban, 2013). Future research needs to look more into the extent to which word frequency can be used to explain patterns in topic/vehicle asymmetry, and whether word frequency is only a correlate of conceptual patterns (such as concept familiarity) or might perhaps play a more causal role (see Harmon & Kapatsinski, 2017).

The fact that emotional valence was a factor predicting metaphor quality is interesting for several reasons. First, it shows that metaphors need to be considered with respect to their evaluative functions and their meaning in terms of emotional valence (Kemp, 1999; Sakamoto & Utsumi, 2014; Semino, 2008; Winter, 2016: Ch. 8). Second, the results support findings from the general literature on emotion processing. Specifically, the fact that metaphors where the vehicle is relatively more negative than the topic is broadly in line with the idea that there is a negativity bias in processing (Rozin & Royzman, 2001) and in language (Jing-Schmidt, 2007), and that negative information has a stronger modulating effect than positive information (Jiang et al., 2014). It has already been shown that metaphor is more likely to evoke an emotional response than literal language (Citron & Goldberg,

2014). Here we have shown that people prefer it when the emotion being conveyed is negative.

Another important contribution of our study relates to the role of conceptual metaphor in shaping people's response to linguistic metaphor, including responses to A is B type metaphors. The underlying presence of a known conceptual metaphor was predictive of both speed of processing and perceived metaphor quality for A is B type metaphorical expressions, suggesting that individuals have an awareness of and preference for conceptual metaphor, even if it is at a subconscious level: expressions that contained conceptual metaphors were preferred compared to those that contained no known conceptual metaphors.

One possible limitation of our study is the fact that all of the metaphors appeared in the copula A is B construction, which is rare in everyday language (Cameron, 2003). This means that our findings have potentially limited application to real-world data. On the other hand, several leading researchers in the field have used A is B constructions in order to develop, test and confirm theories of metaphor that are highly influential (e.g., Glucksberg, McGlone, & Manfredi, 1997; Katz et al., 1988; Campbell & Raney, 2016). Our study follows this tradition in that we have used the A is B construction to explore the relationship between linguistic and conceptual metaphor, focusing in particular on the impact that the latter has on speed of processing and appreciation. As such, our results are important for understanding psycholinguistic work that is based on A is B type metaphors.

These findings have several implications for the automatic generation of metaphors by computers. It is relatively easy for algorithms to produce metaphors. However, prioritising metaphors that are conventional or optimally innovative presents a greater challenge. This would require the development of databases showing domains that are conventionally linked conceptual metaphors in the languages involved and the vocabulary items that tend to be associated with those conceptual metaphors. MetaNet

(https://metanet.icsi.berkeley.edu/metanet/) could serve as a starting point in this respect, but more extensive, detailed information would need to be provided, in order to allow for the process of optimal metaphor generation to be fully automated. Our findings with respect to frequency and valence are more promising as databases containing words that are normed according to these criteria are readily available and could be easily incorporated into metaphor generation algorithms. If a consideration of these variables was factored into the algorithm, then people may find it harder to distinguish between those metaphors that have been automatically generated and those that have been produced by a computer, thus increasing the probability that these metaphors will pass the metaphorical Turing test.

We have found that a good computer-generated metaphor is one which draws on relationships with which the perceiver is familiar but that connects them in a somewhat creative way, one which describes general entities in specific ways thus narrowing down the focus, and one which contains negative evaluations. We have also seen, by comparing our results to those of previous studies, that peoples' perceptions of a metaphor's origin may shape their evaluation of the metaphor (see Gibbs et al., 1991). More research is needed to investigate this issue further. In addition to being of value to computer scientists, our findings are of value to those working in professions that require them to employ language for persuasive purposes. These professions include, but are not limited to, advertising, speech-writing, and journalism. Moreover, the use of metaphor extends well beyond language; it has been shown to be prominent in art, film, architecture, music, and gesture, where it serves a range of expressive and evaluative functions. The next step is to ascertain whether the features we have discovered here are also true of metaphor in these other forms of human expression.

Acknowledgements

This research has been funded thanks to a Marie Curie Individual Fellowship (Project ref. EMMA-658079) and the national project FFI2013-43593-P (Ministry of Innovation and Competitiveness, Spain).

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390-412.

Barnden, J. (2015). Open-ended elaborations in creative metaphor. In T.R. Besold, M. Schorlemmer, & A. Smaill (eds.), *Computational Creativity Research: Towards Creative Machines* (pp. 217-242). Amsterdam: Atlantis Press.

Bartoń, K. (2016). MuMIn: Multi-model inference. R package version 1.15.6.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bohrn, I.C., Altmann, U., Lubrich, O., Mennighaus, W., & Jacobs, A.M. (2013). When we like what we know – a parametric fMRI analysis of beauty and familiarity. *Brain and Language, 124*, 1-8.

Brown, G. D., & Watson, F. L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, *15*(3), 208-216.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41,* 977-990.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46,* 904-911.

Campbell, S.J., & Raney, G.E. (2016). A 25-year replication of Katz et al.'s (1988) metaphor norms, *Behaviour Research Methods, 48*(1), 330-340.

Cameron, L. (2003). *Metaphor in educational discourse: Advances in Applied Linguistics.* London, UK: Continuum.

Charteris-Black, J. (2011). *Politicians and Rhetoric: The Persuasive Power of Metaphor,* Basingstoke: Palgrave MacMillan.

Casasanto, D., & Boroditsky, L. (2008). Time in the mind: Using space to think about time. *Cognition*, *106*(2), 579-593.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359.

Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. *Cognition, 125*(3), 452-465.

Citron, F. & Goldberg, A. (2014). Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, *26*(11): 2585-2595.

Deignan, A., Littlemore, J. ,& Semino, E. (2013). *Figurative Language, Genre and Register,* Cambridge: Cambridge University Press.

Finn, P. J. (1977). Word frequency, information theory, and cloze performance: A transfer feature theory of processing in reading. *Reading Research Quarterly*, *13*(4), 508-537.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Giora, R., Fein, O., Kronrod, A., Elnatan, I., Shuval, N., & Zur, A. (2004). Weapons of mass distraction: Optimal innovation and pleasure ratings, *Metaphor and Symbol, 19*(2), 115-141.

Glucksberg, S., McGlone, M.S. &Manfredi, D. (1997). Property attribution in metaphor comprehension. *Journal of Memory and Language, 36*(1), 50-67.

Goschler, J. (2005). Embodiment and body metaphors. *Metaphorik, 9*, 33-52.

Gries, S. (2011). Phonological similarity in multi-word units. *Cognitive Linguistics, 22*(3), 491-510.

Harmon, Z., & Kapatsinski, V. (2017). Putting old tools to novel uses: The role of form accessibility in semantic extension. *Cognitive Psychology, 98*, 22-44.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics, 12*(2), 249-268.

Jacobs, A.M. and Kinder, A. (2017). "The brain is the prisoner of thought": a machine learning assisted quantitative narrative analysis of literary metaphors for use in neurocognitive poetics. *Metaphor and Symbol, 32*(3), 139-160.

Jacobs, A. M., Kinder, A. (in press). What makes a metaphor literary? Answers from two computational studies. *Metaphor and Symbol.*

Jäkel, O. (1999). Is metaphor really a one-way street? One of the basic tenets of the cognitive theory of metaphor put to the test. In L. de Stadler, & C. Eyrich (eds.), *Issues in Cognitive Linguistics* (pp. 367-388). Berlin/New York: de Gruyter.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824-843.

Jiang, Z., Li, W., Liu, Y., Luo, Y., Luu, P., & Tucker, D. M. (2014). When affective word valence meets linguistic polarity: behavioral and ERP evidence. *Journal of Neurolinguistics, 28*, 19-30.

Jones, L. and Estes, Z. (2006). Roosters, robins, and alarm clocks: aptness and conventionality in metaphor comprehension. *Journal of Metaphor and Language, 55*(1), 18-32.

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 Symbolic Activity*, *3*(4), 191-214.

Kemp, E. (1999). Metaphor as a tool for evaluation. *Assessment and Evaluation in Higher Education, 24*(1), 81-89.

Keuleers, E., Lacey, P., Rastle, K., & Brysbaert, M. (2012). The British Lexicon Project: Lexical decision data for 28,730 monosyllabic and disyllabic English words. *Behavior Research Methods, 44*(1), 287-304.

Kittay, E. (1990). *Metaphor: Its cognitive force and linguistic structure.* Oxford: Oxford University Press.

Kövecses, Z. (2005). *Metaphor and culture.* Cambridge: Cambridge University Press.

Kuperman, V. (2015). Virtual experiments in megastudies: A case study of language and emotion. *The Quarterly Journal of Experimental Psychology, 68*(8), 1693-1710.

Lai, V.T., Curran, T., & Menn, L. (2009). Comprehending conventional and novel metaphors: An ERP study. *Brain Research, 1284*, 145-155.

Lakoff, G., Espenson, J., & Schwartz, A. (1991). Master Metaphor List (2nd draft copy). Retrieved on 14th June 2017 for the last time from: http://araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf

Littlemore, J., & Low, G. (2006). *Figurative thinking and foreign language learning.* Basingstoke: Palgrave MacMillan.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & T. Tognini-Bonelli (eds.), *Text and Technology:* In honour of John Sinclair (pp. 157-176). Amsterdam: John Benjamins.

Liu, H., Hu, Z., & Peng, D. (2013). Evaluating word in phrase: the modulation effect of emotional context on word comprehension. *Journal of Psycholinguistic Research, 42*, 379-391.

McCormack, J., & d'Inverno, J. (eds.) (2012). *Computers and Creativity.* UK: Springer.

McGlone, M. & Tofighbakhsh , J. (1999). The Keats heuristic: Rhyme as reason in aphorism interpretation. *Poetics, 26*(4), 235-244

Noble, C. E. (1953). The meaning-familiarity relationship. *Psychological Review*, *60*(2), 89-98.

Norenzayan , A., Atran, S., Faulkner, J., & Schaller, M. (2006). Memory and mystery: The cultural selection of minimally counterintuitive narratives. *Cognitive Science, 30*, 531-53.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237-241.

Pérez-Sobrino, P.and Littlemore, J. (submitted). Cross-cultural variation in the appreciation of advertisements, Submitted to *Applied Linguistics.*

Postman, K., & Conger, B. (1954). Verbal habits and the visual recognition of words. *Science, 119*, 671-673.

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

Radden, G. & Kövecses, Z. (1999). Towards a theory of metonymy. In K.-U. Panther & G. Radden *(*eds.*)*, Metonymy in Language and Thought (pp. 17-59). Amsterdam: John Benjamins.

Reber, R., Winkielman, P., & Schwarz, N. (1998). Effects of perceptual fluency on affective judgments. *Psychological Science, 9*(1), 45-48.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296-320.

Sakamoto, M., & Utsumi, A. (2014). Adjective metaphors evoke negative meanings. *PloS one, 9*(2), e89008.

Semino, E. (2008). *Metaphor in discourse.* Cambridge: Cambridge University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2016). afex: Analysis of factorial experiments. R package version 0.16-1.

Snefjella, B., & Kuperman, V. (2016). It's all in the delivery: Effects of context valence, arousal, and concreteness on visual word processing. *Cognition, 156*, 135-146.

Solomon, R. L., & Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology, 43*, 195-201.

Stewart, D. (2010). *Semantic prosody: A critical evaluation.* London: Routledge.

Sweetser, E. (1990): *From etymology to pragmatics: Metaphorical and cultural aspects of semantic structure,* Cambridge: Cambridge University Press.

Thibodeau, P., & Durgin, F. (2008). Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language, 58*(2), 521-540.

Turing, A. (1950). Computing machinery and intelligence. *Mind 59*(236), 433-460.

Veale, T. (2015a). *Exploding the creativity myth: The computational foundations of linguistic creativity.* London: Bloomsbury.

Veale, T. (2015b). Game of tropes: Exploring the placebo effect in computational creativity. In H. Toivonen, S. Colton, M. Cook, & D. Ventura (eds.), *Proceedings of the Sixth International Conference on Computational Creativity* (pp. 78-85). Provo, Utah: Brigham Young University.

Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*, 1191-1207.

Wickham, H. (2017a). stringr: Simple, consistent wrappers for common string operations. R package version 1.2.0. https://CRAN.R-project.org/package=stringr

Wickham, H. (2017b). tidyverse: Easily install and load 'tidyverse' packages. R package version 1.1.1. https://CRAN.R-project.org/package=tidyverse

Winter, B. (2016). The sensory structure of the English lexicon. PhD thesis*,* University of California, Berkeley.

Winter, B., Marghetis, T., & Matlock, T. (2015). Of magnitudes and metaphors: Explaining cognitive interactions between space, time, and number. *Cortex*, *64*, 209-224.

Winter, B., Thompson, G., & Urban, M. (2013). Cognitive factors motivating the evolution of word meanings: Evidence from corpora, behavioral data and encyclopedic network structure. In E. A. Cartmill, S. Roberts, H. Lyn, & H. Cornish (eds.), *10th International Conference on the Evolution of Language* (pp. 353-360). New Jersey: World Scientific.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology, 9*, 1-27.

Appendix A

Metaphors used in the study[iv]

Metaphor in English	**Metaphor in Spanish**	**Metaphor in Mandarin Chinese**
Speech is a comforting song.	La palabra es una canción reconfortante.	话语是一首安慰的歌。
Religion is a relaxing music.	La religión es una música relajante.	宗教是一首轻松的音乐。
Business is a pure water.	Negocio es agua pura.	生意是一杯纯净水。
Fear is a sweet love.	El miedo es un dulce amor.	恐惧是一种甜蜜的爱情。
Music is a comforting fire.	La música es un fuego reconfortante.	音乐是一团安慰的火。
Art is a gentle love.	El arte es un amor tierno.	艺术是一种温柔的爱。
Language is a celebrated art.	El lenguaje es un arte célebre .	语言是一种著名的艺术。

War is a tender love.	La guerra es un amor tierno.	战争是温柔的爱。
Power is a comforting fire.	El poder es un fuego reconfortante.	权力是一团安慰的火。
Security is a warming light.	La seguridad es una luz cálida.	安全是温暖的光。
Love is a comforting fire.	El amor es un fuego reconfortante.	爱情是一把安慰的火。
Life is an enjoyable comedy.	La vida es una comedia agradable.	生活是一场愉快的喜剧。
Life is an appealing wine.	La vida es un vino atractivo.	生命是一杯诱人的葡萄酒。
Language is a gentle love.	El lenguaje es un amor tierno.	语言是一种温柔的爱。
History is a modern science.	La historia es una ciencia moderna.	历史是一门现代科学。
Power is a warming light.	El poder es una luz cálida.	权力是一种温暖的光。
Design is a warming light.	El diseño es una luz cálida.	创作是温暖的光。
Religion is a celebrated art.	La religión es un arte célebre.	宗教是一种著名的艺术。
Disorder is a calming sleep.	El desorden es un sueño tranquilo.	障碍是平静的睡眠。
War is an exciting love.	La guerra es un amor apasionante.	战争是令人兴奋的爱。
Teaching is a warming light.	La enseñanza es una luz cálida.	教学是温暖的光。

Medicine is a tender love.	La medicina es un amor tierno.	药是一种温柔的爱。
Truth is soothing music.	La verdad es una música relajante.	真理是一首舒缓的音乐。
Marriage is a hot hell.	El matrimonio es un infierno candente.	婚姻是炙热的地狱。
Food is enjoyable music.	La comida es una música agradable.	食物是一首愉快的音乐。
Business is a tasty wine.	El negocio es un vino sabroso.	财务是一杯美味的葡萄酒。
Art is a beautiful love.	El arte es un amor hermoso.	艺术是一种美丽的爱。
Love is a sparkling rainbow.	El amor es un arco iris brillante.	爱情是闪闪发光的彩虹。
Business is relaxing music.	Negocio es una música relajante.	交易是一首轻松的音乐。
Music is a warming light.	La música es una luz cálida.	音乐是温暖的光。
Song is a comforting fire.	La canción es un fuego reconfortante.	歌声是一团安慰的火。
Technology is a warming light.	La tecnología es una luz cálida.	技术是温暖的光。
Programming is a cherished art.	La programación es un arte apreciado.	编程是一个珍贵的艺术品。
Destiny is a lovely heaven.	El destino es un cielo precioso.	命运是一个可爱的天堂。
Power is pure water.	El poder es agua pura.	权力是一杯纯净水。

Love is soothing music.	El amor es una música relajante.	爱情是一首舒缓的音乐。
Love is a beautiful painting.	El amor es una pintura hermosa.	爱情是一副美丽的图画。
Love is a beneficial medicine.	El amor es un medicamento beneficioso.	爱是一种有益的药。
Government is a charming painting.	El gobierno es una pintura adorable.	政府是一副迷人的画卷。
Pain is an appealing beauty.	El dolor es una belleza atractiva.	疼痛是一种诱人的美。
Death is a relaxing sleep.	La muerte es un sueño relajante.	死亡是轻松的睡眠。
Writing is a soothing healing.	La escritura es una curación calmante.	写作是一种舒缓的愈合。
Disease is a warming light.	La enfermedad es una luz cálida.	疾病是温暖的光。
Resistance is a warming light.	El movimiento de resistencia es una luz cálida.	阻力是变暖的光。
Security is an invigorating energy.	La seguridad es una energía vigorizante.	安全是令人振奋的能量。
Imagination is an attractive beauty.	La imaginación es una atractiva belleza.	想象是一种有吸引力的美。
Learning is a rewarding investment.	El aprendizaje es una inversión gratificante.	学习是一项有价值的投资。
Government is a warming	El gobierno es una luz cálida.	政府是一束温暖的光。

light.		
Blood is an appealing wine.	La sangre es un vino atractivo.	血液是一杯诱人的葡萄酒。
Truth is a stunning beauty.	La verdad es una belleza impresionante.	真理是一个绝色美女。
Wisdom is a distant foreigner.	La sabiduría es un extranjero distante.	智慧是一个遥远的外国人。
Danger is a hot spice.	El peligro es una especia picante.	危险是一种辣味调料。
Evolution is an ongoing lottery.	La evolución es una lotería en curso.	时间是一个很好的医生。
Time is a good physician.	El tiempo es un buen médico.	睡眠是一片汪洋大海。
Sleep is a vast ocean.	El sueño es un vasto océano.	睡眠是一片汪洋大海。
Discipline is a strong fertilizer.	La disciplina es un potente fertilizante.	纪律是一种强化肥。
Indecision is a calm whirlpool.	La indecisión es un remolino en calma.	优柔寡断是一个平静的旋窝。
History is cracked mirror.	La historia es un espejo agrietado.	历史是一面破碎的镜子。
Anger is a cold blizzard.	La ira es una fría tormenta de nieve.	愤怒是一场寒冷的暴风雪。

Education is a glowing lantern.	La educación es una linterna brillante.	**教育是一个发光的灯笼。**
Practice Items	**Practice Items**	**Practice Items**
Knowledge is a comforting fire.	El conocimiento es un fuego reconfortante.	知识是一团安慰的火。
Leadership is exciting fighting.	El liderazgo es una lucha emocionante.	领导力是令人兴奋的战斗。
Culture is a celebrated art.	La cultura es un arte célebre.	文化是一种著名的艺术。
Creation is a gentle love.	La creación es un amor tierno.	创造是温柔的爱。
Business is exciting fighting.	El negocio es una lucha emocionante.	商场是令人兴奋的战场。
Empire is a celebrated art.	El imperio es un arte célebre.	帝国是一种著名的艺术。
Dance is a gripping battle.	La danza es una batalla apasionante.	舞蹈是一场扣人心弦的战斗。

Appendix B

The R packages `stringr` version 1.2.0 (Wickham, 2017a) and `tidyverse` version 1.1.1 (Wickham, 2017b) were used for data carpentry. The `car` package version 2.1.3 (Fox & Weisberg, 2011) was used to compute variance inflation factors for regression models to assess the presence of collinearity. The packages `lme4` version 1.1.13 (Bates, Maechler, Bolker, & Walker, 2015) and `afex` version 0.16.-1 (Singmann, Bolker, Westfall, & Aust,

2016) were used for linear mixed effects models. The package `MuMIn` version 1.15.6 (Bartoń, 2016) was used to compute R^2 values for mixed models.

Acknowledgement

We would like to thank Beinan Zhou for her help in developing the reaction time study. We thank Sarah Duffy and Marcus Perlman for their helpful comments and suggestions.

[i] We also included an additional question, in which we asked the participants to rate on a scale from 1 to 3 how easy it was to understand the expression. Participants were only presented with this question if they had provided a positive answer to Question 1. Because a large number of participants answered 'no' to Question 1 and were therefore not presented with this question, we ended up with a large number of missing values which rendered this question uninformative. We therefore decided to leave this question out of the final analysis.

[ii] These questions appeared on the screen at the same time as the metaphors. They disappeared once they were answered, to be replaced with the following question whilst the metaphors remained.

[iii] As random effects, we included subject and items, with by-subject random slopes for novelty and response, as well as by-item random slopes for response. In lme4 syntax, the fitted model was: LogRT ~ (Novelty + Novelty_Squared) * Response + Human + (1 + Novelty + Novelty_Squared + Response|Subject) + (1 + Response|Item). *p*-values are reported based on likelihood ratio tests of the full model with the fixed effect in question against the null model without the fixed effect. Estimation was performed using maximum likelihood.

[iv] It should be noted that, due to a coding error, there were only 59 metaphors in the Mandarin Chinese version of the test and that in all three version of the test, the item "war is an exciting love" was accidentally included even though it was ungrammatical.