Automated scoring of creative metaphor production with deep learning

Ricardo Primi
Universidade São Francisco &
Edulab21, Instituto Ayrton Senna, Brazil
rprimi@mac.com

Abstract

The 21st century skill of creativity has been emphasized for its contribution to personal, professional, and social achievement. Its assessment remains challenging, however. Correcting divergent productions is one of them. Usually, this involves training a large number of raters to score each response before you can calculate a reliable score for an individual. In large scale assessment, this shortcoming becomes problematic. To promote the use of creativity assessment in large-scale data collection, it is necessary to develop a reliable automated correction system. In this study we tested the effectiveness of deep learning in assessing students' divergent metaphor productions for a “fill in the blank task” like "The camel is the ________ of the desert.".

1 Introduction

1.1 Rumelhart model for analogical reasoning

David E. Rumelhart is well known for his work on backpropagation algorithms that enabled neural networks to learn. He is perhaps less known for his work on analogies. In 1973, Rumelhart and Abrahamson proposed a model for analogical reasoning, which explained information retrieval as a process that depended more on the structure of memory than its content.

As an example, the authors illustrate this when answering the question: Who is the father of Deep Learning? You may execute two distinct cognitive processes: (a) when you have information stored in your memory, you can remember that information, that is, access and retrieve the information: "Geoffrey Hinton"; (b) when you do not know the answer, you may reason thinking of the words and how they are related and drawing an answer from them.

A core aspect of Rumelhart and Abrahamson (1973) model is that “retrieval depends to a much greater extent on the form of the relationship among the words” (p. 2). They proposed that memory should be viewed as a multidimensional Euclidean space. Concepts would be represented by points in this space. The degree of similarity between two concepts is inversely proportional to the Euclidean distance between them. For instance, they demonstrated that 30 mammals could be represented in 3D Euclidean space (latent vectors) based upon *ferocity*, *anthropomorphism*, and *size*. Each animal was represented by a numerical vector indicating its intensity in each of the three dimensions. A gorilla, for example, had a high value in all three dimensions; a mouse had a low value in all three dimensions.

Memory structure is the key to differentiate types of semantic similarity between concepts, which form geometric structures in Euclidian space of latent attributes: (a) Serial order (seriation) could be shown by a straight line with similar distances between the points (b) classification (class membership) could be shown by all the points gravitating towards a central point (c) analogy/metaphors could be represented by a parallelogram in which each concept is a vertex and the distances A and B are balanced with distances C and D.

1.2 Sternberg’s domain interaction theory for quality of metaphors

How do people understand metaphors like: The camel is a boat or a taxi in the desert? Tourangeau and Sternberg (1981) studied the appropriateness and comprehensibility of metaphors using the logic of Rumelhart and Abrahamson (1973). They proposed two properties for metaphor quality, equivalence and remoteness. In the Camel metaphor, for example, there are two domains, one related to deserts, the other to seas. Equivalence is a function of the common semantic dimensions of number and parallelism with which concepts are related within a domain. Camel and boat, for example, are modes of transportation in their respective environments. Distance between domains is referred to as remoteness, as a greater distance, up to a certain point, is associated with originality, while a large distance renders metaphors incomprehensible. According to Tourangeau and Sternberg (1981), the interaction between domains can lead to the development of new meanings for an idea. At the end of the day, both domains are interpreted differently when they are combined to form a metaphor.

A number of experiments were conducted by the authors to create metaphors, represent concepts in a multidimensional Euclidian space, and predict what ideas will complete metaphors, as well as the ratings of the quality of metaphors based on the distances between vectors and geometric structures predicted by the model. In order to conduct these studies, researchers needed to know beforehand the latent dimensions underlying the concepts on which they were working and vector representations of each concept on the latent dimensions. Tourangeau and Sternberg considered eight domains—birds, land mammals, sea creatures, ships, aircraft, land vehicles, U.S. historical figures, and modern world leaders—that could be represented by a two-dimensional model of two variables *power/aggression* and *prestige*. It was a costly aspect of the experiment to have researchers rate similarities among concepts in order to determine the latent dimensions.

Rumelhart and Abrahamson (1973) used a previous study where animals were represented in a three-dimensional space based on ferocity, anthropomorphism, and size. On the one hand, they were successful in demonstrating that multidimensional Euclidian distances and geometric structures could predict good metaphors. On the other hand, these results were narrowly focused on the domain for which they had defined latent dimensions and concept representation.

1.3 Vector space models for efficient word representations.

Several years later Mikolov, Sutskever, Chen, Corrado, & Dean, (2013) proposed a method for discovering latent dimensions and very efficient vector representations for words which capture syntactic and semantic relationships among them. This method identifies latent dimensions of semantic similarity based upon the co-occurrences of words in large corpora of natural texts. It was a remarkable result of this work that a geometric operation on vectors on Euclidian space could solve analogies in exactly the manner predicted by Rumelhart and Abrahamson (1973). A classic example is the analogy: “King - Man + Woman” results in a vector very close to “Queen.” (p. 746). Interestingly Mikolov et al. (2013) refer to an earlier paper Rumelhart, Honton & Williams (1986) but not to his work on analogies when providing the basis for his study.

Mikolov's method for determining vector representations for words in natural language overcomes the limitations of earlier studies with limited applications and opens up new possibilities for cognitive modelling analogical reasoning/metaphors.

1.4 Creativity assessment

The **monkey** is a/the ______________ of the **forest**

Metaphors	*Explanation*
1) *clown*	*In the forest monkey is playful and fun as well as the clown in the circus..*
2) *child*	*Because in the park a child is hyperactive, playful and naive as the monkey in the forest*
3) *alpinist*	*Because the monkey climb the trees like the alpinist Who climbs the mountains*
4) *Member of the fores't rooters (or cheerers)*	*Because he hang on with gangs screaming and making a mess.*

The **camel** is a ______________ of the **desert**

Metaphors	*Explanation*
1) *boat*	*In the sea the boat is a means of transport walking swinging like a camel in the desert*
2) *motorcycle*	*Because motorcycle is a transport for one or two people and need only a few amount of fuel like a camel in the desert that needs little water*
3) *slug*	*Because his walking is slowly, marking the floor and swung her butt like the camel*
4) *Barrichello*	*Because when he is not stopped he is walking slowly*

Figure 1. Metaphor creation task example.

Creativity is considered a 21st century skill that contributes to personal, professional, and social success. However, its assessment remains difficult. Correcting divergent productions is one of them.

One way to evaluate creativity potential is via divergent thinking tasks, where individuals are asked to come up with as many ideas as possible from a given stimulus, such as the Metaphor Creation Test (MCT, Primi, 2014) which asks participants to generate metaphorical ideas from words such as "The **door** is the _______ of the **house**.". Subjects were asked to generate one to four ideas for each prompt. Then, trained raters evaluated each idea on a scale of 0 to 3. The scoring rubrics were developed by operationalizing the notions of semantic equivalence (within domain similarities) and remoteness (between domain similarities) from the domain interaction model of Tourangeau and Sternberg (1981). Examples of two items with four responses are present in Figure 1.

For this task, it is usually necessary to train a large number of raters to assign points to each answer before you can calculate a reliable score for each respondent. The process is known as subjective rating (Primi, Silvia, Benedek, & Jauk, 2018). In large scale assessments, this shortcoming is problematic. To encourage the use of creativity assessment in large-scale data collection, it is necessary to develop a reliable automated correction system. It poses a challenge for the feasibility of creative assessment in large-scale assessments (OECD will include creativity assessment in PISA 2021 and will have to address this). We could potentially solve this problem if we develop an automated scoring system that emulates the behavior of raters.

Several researchers have begun to examine the use of Glove word vector representations to score divergent thinking tasks and to correlate them with human ratings, with promising results (correlations of approximately .80, Dumas, et al., 2021; Ichien, et al., 2021; Johnson, et al., 2021; Selcuk et al., 2021). However, none of these studies used more recent contextual representations from BERT. They do not explore divergent metaphor production nor do they explore the cognitive model of metaphors.

We tested and compared different models, from bag of words to complex contextual representations from transformers in order to evaluate the effectiveness of scoring students' divergent metaphoric productions for "fill in the blank" tasks, such as "The camel is the ____ of the desert.".

2 Experiments

2.1 Hypotheses

The primary focus of this study is to determine whether deep learning models can reliably predict the quality of metaphors produced in a divergent thinking test. Can it be comparable to the benchmarks of inter-rater reliability required for human raters? As a general rule, a kappa coefficient of reliability r = .70 between two raters is considered a very good agreement. Our main hypothesis is that attentional models with rich word vector representations, such as BERT, can achieve this benchmark (H1).

A second test will be conducted using the analogical model of reasoning developed by Rumelhart and Abrahamson (1973). Analogy/metaphors can be represented by a parallelogram in which each concept is represented as a vertex and the distances between concepts A (*camel*) and B (*desert*) are balanced with distances between concepts C (*boat*) and D (*sea*). When responding to the Metaphor Creation Test, participants are required to create metaphors such as blank space in space C: "The *camel* (A) is the *boat* (C) of the *desert* (B). As stated by Rumelhart and Abrahamson, there exists an optimal distance between C and A/ B that will make a concept original and understandable. If they are close, it is likely to be too common to be considered original, but if they are too far apart, it can be difficult to understand. Therefore, we predict that, when using word vectors to represent ideas, the distance between word vectors A/B and C will be nonlinearly related to the quality score of metaphors (H2).

2.2 Datasets

Dataset consisted of a sample contained 974 middle-school children, adolescents, and adults from Brazil (N=651) and Portugal (N=187)— 63.3% women and 36.7% men— from five samples of previous studies that answered to the Metaphor Creation Test. The participants' ages ranged from 9 to 77 years (M = 20.6, SD= 10.2; 88.6% between 9 and 30 years). The rater sample consisted of 18 graduate students who collaborated to complete the ratings as part of their research activities.

The database consisted of 12,174 responses. The majority, 8,050 (69.6%), of the responses were scored by two raters; 1,498 (13%) were scored by three raters; and the remaining responses were scored by four to nine raters. Rasch-Many Facet Partial Credit model was used to score responses (Primi, 2014). In this design raters are considered as items of a test. Each idea received an estimated standardized score accounted for rater differences in leniency-severity dimension.

From the total of 12,174 responses, we randomly split 9,983 (82%) responses for training and 2,191 responses for validation. All models were trained using this scheme. The size of the vocabulary was 8,524 unique words. Each response contained 1 to 29 words with an average of 7.9 words (SD=2.6).

We used pre-trained word embeddings from two sources: (a) *static word embeddings*: database from Hartmann et al., 2017 from the Interinstitutional Center for Computational Linguistics (NICL, http://nilc.icmc.usp.br/embeddings, University of São Paulo, Institute of Mathematics and Computer Sciences). The authors trained word embeddings based on large corpus of Portuguese texts. They trained word vectors using four types of algorithms producing latent representations formed of 50 to 1000 dimensions. We used Glove 600-dimensional vectors because of their best test results in solving semantic analogies in intrinsic evaluation tasks; (b) *contextual word embeddings*: BERTimbau a BERT model trained in Portuguese language (Souza, Nogueira & Lotufo, 2020).

2.3 Metrics

Raters scored ideas on a scale from 0 (not a metaphor), 1 (a metaphor that is appropriate), 2 (a metaphor that is appropriate and remote), and 3 (a metaphor that is both appropriate and outstanding). Macro average F1 score will be used to evaluate the model quality. We will also calculate Kappa to compare automated scoring produced by each model with industry benchmarks. Considering judges may differ in their level of severity/leniency, we use the Many Faceted Rasch Model (MFRM) in order to equate raters' scores. As a result, for each idea we have a continuous MFRM score that separates rater-related differences from the scores. Before training the models, we will predict a score for each idea from this continuous score to have a common metric across different group of raters.

The first analysis is based on data from each idea. But creativity tests are intended to obtain a subject creativity score. Therefore, we will aggregate each scored idea of a particular subject in order to calculate a creativity score for each subject. Afterwards, we will correlate the automated aggregated score with the human ratings. A minimum correlation of .70 is expected between the automated system and human scored responses.

2.4 Models

As a baseline, we will use a word representation based on a bag of words and pointwise mutual information (PMI). We will then test (a) the logistic regression model and (b) the fully connected neural network. By combining word representations and models, we will produce four models.

A bi-directional LSTM will then be tested with two types of pre-trained word vectors that will be frozen before training (a) static Glove vectors and (b) contextual vectors from BERT. In order to obtain contextual word vectors, we will pass subject responses to BERTimbau and save word representations from the final layers.

Our last step will be to use a BERTimbau model to predict metaphor scores based on a fine-tuned [CLS] token.

We will therefore test three models during this phase. In our H1 prediction, we expect that LSTM with contextual vectors or the BERT model will perform better than baseline models and LSTM using static word vectors.

To obtain item representation for the second hypothesis (H2), we will calculate the average of the word vectors for A and B words. Next, we will calculate the Euclidian distance between this point and the word vector of each response. We will then explore the relationship between this distance measure and human scores. A smoothing spline will be used to predict human scores from this distance. The same analysis will be conducted twice, once with static vectors and once with contextual vectors, and the coefficient of determination of the two types of representations will be compared.

A summary of all the models tested can be found in Table 1. We will vary the complexity of models (logistic regression, simple neural networks, LSTM and attentional models - BERT), as well as the complexity of word representations (bag of words, PMI, static and contextual dense word embeddings).

First, we will determine what model and word representation approximate the benchmark of inter-rater reliability in order to qualify for field test use in large scale assessments. H1 proposes that more complex models and contextual representations of words will perform better.

In order to accomplish the second objective, we intend to explore cognitive model-specific predictions regarding the structure of memory representations using word vectors. We will examine whether this relationship is non-linear and how well this simple distance measure can predict metaphor quality.

Table 1. Summary of the design

Hypoth.	Models	Word representations
Baseline	Logistic regress	Bag of words
Baseline	Fully connected NN	Bag of words
Baseline	Logistic regress	PMI
Baseline	Fully connected NN	PMI
H1	Bidirectional LSTM	Static word vectors (Glove)
H1	Bidirectional LSTM	Contextual vectors (BERTimbau)
H1	BERT finetuned	Contextual vectors (BERTimbau)

H2	Exploratory analysis with Smoothing spline	Static word vectors (Glove)
H2	Exploratory analysis with Smoothing spline	Contextual vectors (BERTimbau)

3 Results

In Table 2, we present the main conclusions of the experiments related to the first hypothesis. The first two columns indicate the model and the word representation used. In the last three columns we show the macro-averaged F1 score, kappa and the correlation between the model's predicted score and actual raters' equated score at the level of ideas and aggregated by subjects. There are a few general points worth noting. First, none of the models approached the desired benchmark of .70 for Kappa or machine-rater correlation. Kappa values ranged from .26 to .50 and correlations from .31 to .59. A second point is that contextual word vectors from BERT had a better performance than baseline, supporting our first hypothesis (H1). Our best model was when we finetuned the token [CLS] from BERT, which represents the entire sentence. Third, models that use pre-trained distributed representations, such as Glove and BERT, perform better than baseline models.

In order to test the second hypothesis, we examined the relationship between semantic distance of the subject's response to the item stem and the quality of the metaphor (Rumelhart and Abrahamson, 1973

Table 2. Metrics for each combination of model and representation.

Model	**Word representations**	**Details**	**Macro agv. F1 Score**	**Kappa**	***r* (idea/subj.)**
Baseline: logistic regression	Bag of words		0.424	0.264	*r* =.31/.39
Baseline: Fully connected Neural Network (FCNN)	Bag of words	2 layers of 80 and 40 units L2 and .50 dropout	0.331	0.327	*r* =.36/.45
Baseline: logistic regression	PMI		0.373	0.332	*r* =.36/.44
Baseline: Fully connected Neural Network (FCNN)	PMI	2 layers of 80/40 units, L2 and .50 dropout	0.311	0.261	*r* =.32/.42
Bidirectional LSTM	Static Vectors Glove	80 unities and .50 dropout	0.401	0.415	*r* =.45/.56
Bidirectional LSTM	Static Vectors BERT layer 0	80 unities and .50 dropout	0.387	0.380	*r* =.41/.52
Bidirectional LSTM	Static Vectors BERT Layer 12	80 unities and .50 dropout	0.326	0.274	*r* =.31/.39
Bidirectional LSTM	Aggregated Vectors BERT Layer 12	80 unities and .50 dropout	0.406	0.411	*r* =.45/.55
BERT finetuned with a Fully connected Neural Network (FCNN) on top of [CLS] token	Contextual Vectors BERT	300 unities	0.368	0.358	*r* =.39/.49
GPT3	Contextual Vectors		0.432	0.043	*r* =.12/.37

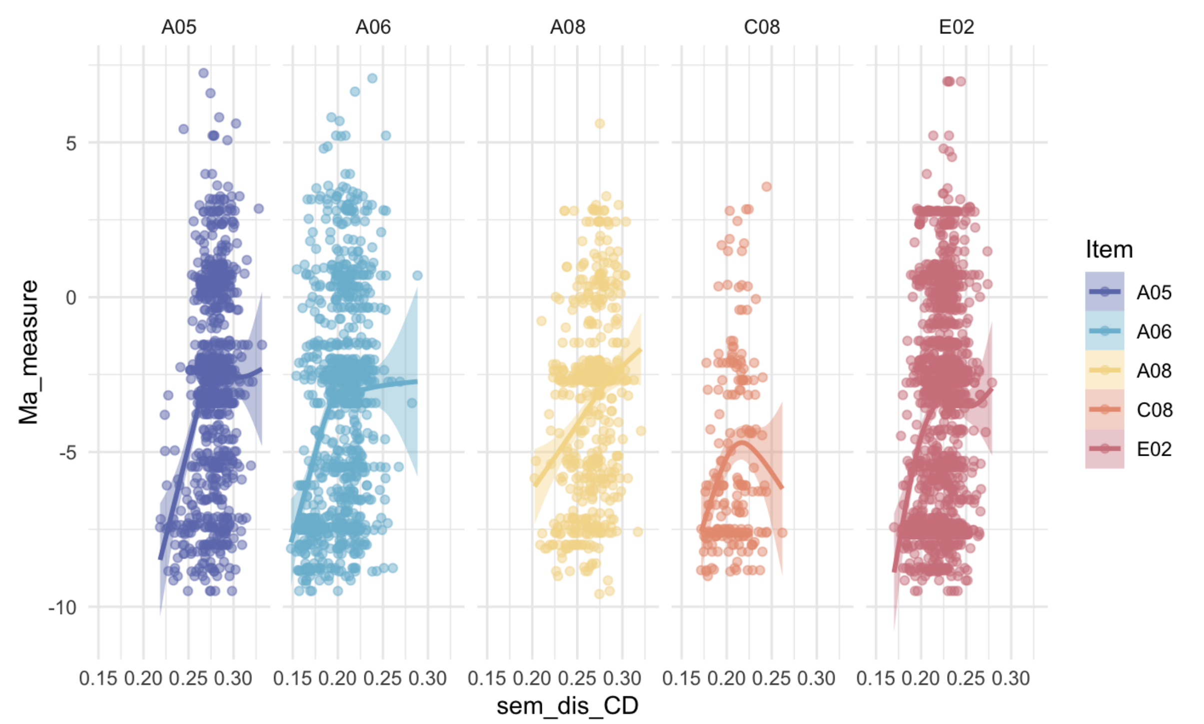

Figure 2. Relationship between metaphors scores and semantic distance of AB to CD for items A05: planets ____ sun, A06: bus_____city, A08: love _____feelings, C08: door____house, and E02 grass_____ backyard.

hypothesis). Our first step was to compute the Euclidian squared distance between the mean of the Glove word vectors between items A and B and the mean of the Glove word vectors used by the subjects in their responses. This relationship is illustrated in Figure 2. There were 11 items for which there were more than 100 responses. Correlations of this semantic distance with raters' scores ranged from 0.01 to .34 (although it must be noted that these coefficients assume a linear trend). This relationship suggests that there is an optimal distance beyond which metaphors are not considered good metaphors, supporting our second main hypothesis (H2).

4 Analysis and conclusions

In this paper, we investigated whether deep learning models can reliably predict the quality of metaphors produced in a large-scale divergent thinking test. This first attempt did not meet the high standards of inter-rater reliability required for practical use. In contrast to earlier studies, our results are lower (.59 versus .80). This may be the result of the metaphor task being more complex than the alternative use task tested in previous studies. By applying the same method for computing semantic distance between word vectors of the stimulus and the response, a method used by Dumas et al., 2021; Ichien et al., 2021; Johnson et al., 2021; Selcuk et al., 2021, we found correlations of small magnitude .03 to .34. These data suggest that metaphor task is more complex than alternate use task.

An important aspect of this study was to demonstrate support for our first hypothesis that contextual representations of BERT are significantly superior to the base-line model representations used in previous studies. This brings us to our second hypothesis. In this study, we are able to demonstrate that the relationship between semantic distance response metaphor candidates and metaphor quality is non-linear as predicted by Rumelhart and Abrahamson (1973) and Tourangeau and Sternberg (1981). It may be that this is part of the task's complexity. The use of more complex representations of context and cognitive models proved to be a promising approach for future efforts to improve the automated scoring of complex tasks such as metaphor generation. To ensure accurate predictions, it will be important to perform qualitative error analysis as well as use ensembled models.

5 References

Roger E. Beaty and Dan R. Johnson. 2021. Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods*, 53(2):757–780.

Dan Richard Johnson, James C. Kaufman, Brendan Baker, Baptiste Barbot, Adam Green, Janet van Hell, Evan Kennedy, Grace Sullivan, Christa Taylor, Thomas Ward, and Roger Beaty. 2021. Extracting Creativity from Narratives using Distributional Semantic Modeling. December. DOI: 10.31234/osf.io/fmwgy

Denis Dumas, Peter Organisciak, and Michael Doherty. 2021. Measuring divergent thinking originality with human raters and text-mining models: A psychometric comparison of methods. *Psychology of Aesthetics, Creativity, and the Arts*, 15(4):645–663.

Nathan Hartmann, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jessica Rodrigues, and Sandra Aluisio. 2017. *Portuguese Word Embeddings: Evaluating on Word Analogies and Natural Language Tasks*. arXiv:1708.06025 [cs], August. arXiv: 1708.06025.

Nicholas Ichien, Hongjing Lu, and Keith J. Holyoak. 2021. Predicting patterns of similarity among abstract semantic relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*:No Pagination Specified-No Pagination Specified.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October.

David E Rumelhart and Adele A Abrahamson. 1973. A model for analogical reasoning. *Cognitive Psychology*, 5(1):1–28, July.

Selcuk Acar, Kelly Berthiaume, Katalin Grajzel, Denis Dumas, Charles "Tedd" Flemister, and Peter Organisciak. 2021. Applying Automated Originality Scoring to the Verbal Form of Torrance Tests of Creative Thinking. *Gifted Child Quarterly*:00169862211061874, December.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs]*, September. arXiv: 1301.3781.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Ricardo Primi. 2014. Divergent productions of metaphors: Combining many-facet Rasch measurement and cognitive psychology in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 8(4):461–474.

Ricardo Primi, Paul J. Silvia, Emanuel Jauk, and Mathias Benedek. 2019. Applying many-facet Rasch modeling in the assessment of creativity. *Psychology of Aesthetics, Creativity, and the Arts*, 13(2):176–186.

Roger Tourangeau and Robert J. Sternberg. 1981. Aptness in metaphor. *Cognitive Psychology*, 13(1):27–55, January.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. Portuguese Named Entity Recognition using BERT-CRF. *arXiv*:1909.10649 [cs], February. arXiv: 1909.10649.

Appendix1. Metaphor Creation Test (MCT): Creativity Assessment Using Metaphor Production

The main goal of the Metaphor Creation Test (MCT) is to measure individual differences in creative potential. It is based on the underlying cognitive processes of creativity derived from research on analogical reasoning (Sternberg & Nigro (1983), Tourangeau & Sternberg (1981, 1982). From the cognitive perspective, creativity involves specific processes of idea production based on knowledge that already exists, that is, it involves the unique and unusual reorganization or recombination of knowledge. The metaphorical thinking is a particular kind of analogical reasoning that uses known ideas to create new meanings for other ideas. Some examples of metaphor are: "The camel is a ship on the desert" "The hanger is the clothes' spinal cord", The mustache is the antenna of the cat" and "The horse is tick's pasture.

Reclassification of information is an underlying basic cognitive component of the creativity process. Metaphor and analogies are thinking processes that are used to reclassify information and are central activities in producing new ideas. If so creative individuals would be more able to produce metaphors and conversely the easiness of metaphor productions would be an indicator of creativity potential.

Based on this model a test was constructed requiring subjects to create metaphors. Below it is presented the instructions for answering the test:

Instructions

In this test we want you to invent metaphors to complete sentences. See the examples below:

The **monkey** is a/the __________________ of the **forest**

Metaphors	*Explanation*
1) *clown*	*In the forest monkey is playful and fun as well as the clown in the circus..*
2) *child*	*Because in the park a child is hyperactive, playful and naive as the monkey in the forest*
3) *alpinist*	*Because the monkey climb the trees like the alpinist Who climbs the mountains*
4) *Member of the fores't rooters (or cheerers)*	*Because he hang on with gangs screaming and making a mess.*

The **camel** is a __________________ of the **desert**

Metaphors	*Explanation*
1) *boat*	*In the sea the boat is a means of transport walking swinging like a camel in the desert*
2) *motorcycle*	*Because motorcycle is a transport for one or two people and need only a few amount of fuel like a camel in the desert that needs little water*
3) *slug*	*Because his walking is slowly, marking the floor and swung her butt like the camel*
4) *Barrichello (ex-formula 1 driver)*	*Because when he is not stopped he is walking slowly*

In the following items complete the fields left blank with the metaphor that you invented and explanation as in the examples above

Some useful tips:

- First try to find out the relations between the two words presented (monkey / jungle, camel / desert).
- Try to avoid conventional ideas

- Create as many metaphors as you can (up to four per item).
- In order to explain your ideas try think in this way: "Camel" is related to "Desert" as well as "Boat" is to ..."?" or "Monkey" is related to "Forest" as well as "Clown" is for " ?

Items

Then for each item subject is required to create one to four metaphors based on the relationship presented on each item:

E01. The **horn** is the __________________ of the **car**
E02. The **grass** is the ________________________ of the **land**
A03. The **stars** are the____________________ of the **night**
A04. The **ball** is the ________________________ of the **players**
A05. The **planets** are the ________________________ of the **sun**
A06. The **bus** are the ________________________ of the **city**
A07. The **hanger** is the ________________________ of the **clothes**
A08. The **dor** is the ________________________ of the **house**
A09. The **fish** are the ________________________ of the **sea**

The booklet includes four spaces for ideas and explanations like the example bellow:

1. The **horn** is a/the __________________ of the **car**

Metaphor	*Explanation*
1)	
2)	
3)	
4)	

Scoring Quality of the Metaphors

Basic principles of scoring are based on Sternberg’s domain interaction theory in which proposes that in metaphors reclassify the meaning of one object/event/idea seeing it thought the lens of another domain (and its relationships) based on similarity, although the relationship changes its meaning when we use it to see something else outside it’s on domain, therefore, it is called a domain interaction. In this theory good metaphors have equivalent or parallel relationships across domains (equivalence) and the domains are semantically distant (remoteness). These two basic principles are used to score the ideas that are produced.

Consider the metaphor: "The camel is a ship on the desert”

Consider the analogy: The camel is to the desert as the ship is to the sea

In the analogy we have this structure A is to B as C is to D

In metaphor we have this structure A is the C of B

If we consider the terms A (camel) and B (desert) we can imagine that these two terms exist on a shared semantic space. Therefore we can discover various relationships between these terms in this space, for example, that camels are means of transport in the desert, camels are animals that live in the desert, etc. Likewise the terms C (boat) and D (Sea) also share a common semantic space with their own relationships. That said the criteria for judging equivalence involves identifying the parallel relationship between C: D (that is implied and may be inferred from the subjects explanations of the metaphor) with A: B. The rater needs to examine if the relationships that the idea C - proposal by the subject – with its domain are equivalent to the relations A:B. In the example boat (B) and camel (D) share at least an equivalent relation with their semantic universe (sea and desert) that is being modes of transport.

The criteria for judging remoteness is related to the distance between the semantic universes AB / CD. In one hand, there is a semantic domain implied in the item (AB), in the other hand, there is the domain implied by the subjects answer (C and its implied D). The more distant, different or remote the domains the more interesting the metaphor may be. Therefore the rater needs to examine the novelty of the relationship based on the distance of the semantic domains involved. But sometimes the domains are too distant that metaphor lacks comprehensibility.

The general criteria for judging good metaphors is that they are equivalent and remote, that is preserve a clear structure of relations between the terms with domanin and the relationships between domains are far apart. Many ideas presented (metaphor candidates) fail at some point of the aspects mentioned above. In this sense we have created a system of gradual score described bellow. Each idea receives a score 0, 1 , 2 or 3 according to the rules:

Score	Criteria/Examples
0	**Not a metaphor** An idea C that is concrete/factual idea. *Ex: The camel is the means of transportation in the desert* An idea C that is an adjective A. *Ex. The Camel is the brown animal in the desert* An idea that C represents only an association with any of the terms of the item *Ex. The stars are the clarity of the night* An unintelligible idea.
1	**Correct metaphor** An idea C that is equivalent (r (A: B) = r (L: D)) and moderately remote. *Ex: The horn is the voice of the car*
2	**Correct and remote metaphor** An idea that C reaches the criterion for scoring 1 but reaches higher level of novelty (in terms of the distance of the semantic domain) *Ex. The bus is the cholesterol of the city.* Errors in languages should not play a role in determining the scoring. Also ideas don't need to correspond to reality.
3	**Correct and outstanding metaphor** An idea that C reaches the criterion for scoring 2 but reaches advanced level of novelty An idea that has humor characteristics A response that is more elaborate that needs to be considered in the context of the explanations given by the subject. Usually these answers are understood after reading the explanations since they present unusual associations that are not obvious like a response that is scored 1. *Ex. The hanger is the botox of the clothes (explantion: because it prevents the wrinkles on clothes)*